COST ESTIMATION
Methods and Tools

Gregory K. Mislick
Daniel A. Nussbaum

WILEY

Cost Estimation

Wiley Series in
Operations Research and Management Science

A complete list of the titles in this series appears at the end of this volume.

Cost Estimation

Methods and Tools

Gregory K. Mislick

Department of Operations Research
Naval Postgraduate School
Monterey, California

Daniel A. Nussbaum

Department of Operations Research
Naval Postgraduate School
Monterey, California

WILEY

Contents

Foreword

What will be the cost of something we might want to buy? This is a question we frequently ask in our everyday lives. It is a question that can often be answered simply by determining the seller's price – and maybe with a little haggling. But what if we are a government agency that wants to buy something that has never been made before? Or what if we want a newer version of something we already have, but with greater capabilities and never-yet-seen-or-built equipment? And what if we need to make resource allocation decisions today about something we will pay for in the future? Price is not always available – therefore, we must make a cost estimate. As the Government Accountability Office (GAO) explains: "Cost estimates are necessary for government acquisition programs . . . . to support decisions about funding one program over another, to develop annual budget requests, to evaluate resource requirements at key decision points . . . . Having a realistic estimate of projected costs makes for effective resource allocation, and it increases the probability of a program's success."[1]

A cost estimate is the determination of the likely future cost of a product or service based on an analysis of data. Estimating that future cost involves employing inter-disciplinary quantitative analysis techniques. It is partly science, art and judgment. Two of the best scientists, artists and judges in the cost estimating field have joined forces to present this practical guide to cost estimating, primarily (although not exclusively) for Defense practitioners. Dr. Daniel A. Nussbaum is regarded as one of the Navy's premier experts in cost estimating, having been the Director of the Naval Center for Cost Analysis as a member of the Senior Executive Service (SES), and as the International President of the Society for Cost Estimating and Analysis (SCEA), (now renamed the International Cost Estimating and Analysis Association, ICEAA). LtCol Gregory K. Mislick, USMC (Ret.) has significant practical experience in cost estimating from performing cost analyses as a Marine Corps officer, from over 13 years of teaching cost analysis at the Naval Postgraduate School (NPS), and from leading student research as a thesis advisor at NPS. He has been a finalist for the Richard W. Hamming Teaching Award at NPS. Together they have the knowledge and experience in cost estimating for acquisition programs, and are the right team to bring this book to the public. An example of this is that they have recently been awarded – as part of a NPS Operations Research Department team – the Military Operations Research Society's (MORS) Barchi prize award, awarded to recognize the best paper given at the MORS Symposium.

---

[1] U.S. Government Accountability Office, *GAO Cost Estimating and Assessment Guide*, (Washington, GAO-09-3SP, March 2009), p.15.

The timing is perfect for such an endeavor for at least three reasons. First, DoD's cost estimating capability has been the subject of severe criticism. According to Senator Carl Levin, chairman of the Senate Armed Services Committee: "The GAO found that report after report indicated that the key to successful acquisition programs is getting things right from the start with sound systems engineering, cost-estimating, and developmental testing early in the program cycle."[2] Secondly, enactment of the Weapons Systems Acquisition Reform Act (WSARA) in 2009 and signed by the President has placed added emphasis on cost analysis, cost estimating and cost management in DoD. Third, fiscal pressures are forcing hard choices inside DoD. Planned budget reductions, lower funding for overseas contingency operations, discretionary budget caps under the Budget Control Act, and short-term sequestrations all mean that DoD will have to work harder to get costs right at the beginning in order to make the optimal resource allocation decisions in a tight fiscal environment.

In such an environment, what is *not* needed is a book based on theories of cost estimating, but rather a practical book for the community with useful applications for doing the difficult work of cost estimating. That is the purpose of this *how-to* book – to provide a useful guide that the cost estimator can utilize when engaged in the science, art and judgment of cost estimating.

The chapters in the book follow a logical sequence drawing the reader ever more deeply into the practice and processes of cost estimating. It begins in Chapters 1 and 2 with useful introductory material presenting key definitions, describing the DoD policy, statutory and regulatory environments, and setting the frameworks within which cost estimating is done. Chapter 3 first explains the non-DoD acquisition process, and then takes the reader into the reality of beginning a cost estimating assignment by discussing cost analysis requirements, methodologies and processes. Chapters 4 and 5 explain the sources and uses of data, emphasizing the importance of methods for working with different types of data from disparate sources and different time periods. Chapters 6 through 9 will perhaps become the most worn pages of the book, explicating, as they do, the statistical methodologies and regressions that are essential to cost estimating. The concept of learning curves, perhaps the most misunderstood aspect of cost estimating, is addressed in Chapters 10–12. Chapters 13 through 15 then discuss specific topics in cost estimating: use of analogous costing, single-variable cost factors, step-down functions, allocation of profit and overhead costs, and software cost estimating. Chapter 16 addresses the critical challenges associated with risk and uncertainty, and considerations when conducting a Cost Benefit Analysis. The textbook concludes in Chapter 17 with a summary to tie any loose ends together. Taken together, the chapters of this book provide a one-source primer for the cost estimating professional. It is both a textbook for learning and an easy access reference guide for the working practitioner.

A final word about the broad applicability of this book: cost estimating is a discipline interrelated to the other important processes and functions of Defense management. It is required at key milestones of the acquisition processes, of course, but cost estimating is, or should be, a key element of the processes of requirements determination, programming and budgeting, and program performance evaluation. Ultimately, cost considerations must be part of the toolkit for DoD's policy and management leaders. While they

---

[2]"Summary of the Weapons Systems Acquisition Reform Act of 2009," available at http://www.levin.senate.gov/newsroom/press/release/?id=fc5cf7a4-47b2-4a72-b421-ce324a939ce4, retrieved 20 August 2013.

may never conduct an actual cost estimate themselves, their capabilities for making decisions will be enhanced by an understanding of how cost estimating is done, its strengths and limitations, and its language and art. This book needs to be on their bookshelves, as well.

Dr Douglas A. Brook

About the Authors

Reproduced with permission of Javier B. Chagoya.

Gregory K. Mislick is a Senior Lecturer at the Naval Postgraduate School in the Operations Research (OR) Department in Monterey, California. His expertise is in life cycle cost estimating and modeling, probability and statistics, regression analysis, data analysis, and optimization. He is a retired Marine Corps Lieutenant Colonel who flew the CH-46E helicopter and served for 26 years after graduation from the US Naval Academy. He has served as an Associate Dean at NPS and is the Program Manager for the Masters in Cost Estimating and Analysis degree program at NPS. He is also a member of numerous professional societies, including the Military Operations Research Society (MORS) and the International Cost Estimating and Analysis Association (ICEAA). His education includes a BS in Mathematics from the US Naval Academy and an MS in Operations Research from the Naval Postgraduate School. He served as a Board member for Leadership Monterey Peninsula for six years, has been an Associate Board member for the Big Sur Marathon for 13 years, and is Race Director for the JUST RUN Kids 3k event held twice per year in Monterey. He is an active runner, having completed 28 marathons with four marathons under 2:30.

Daniel A. Nussbaum is the Chair of the Energy Academic Group and a faculty member in both the Operations Research Department and the Business School at Naval Postgraduate School (NPS). He has worked in the private sector and in government providing support on a broad range of cost, financial, and economic analyses. He served as the Director, Naval Center for Cost Analysis, during which he was the chief advisor to the Secretary of Navy on all aspects of cost and schedule estimating. He also directed all Navy Independent Cost Estimates as required by Congress and senior Defense leadership on ships, aircraft,

missiles, electronics, and automated information systems. His education includes a BA in Mathematics and Economics, Columbia University; Ph.D. in Mathematics, Michigan State University; a Fellowship from National Science Foundation in Econometrics and Operations Research at Washington State University; National Security Management, National Defense University; Employment of Naval Forces, US Naval War College; and Senior Officials in National Security (SONS) Fellowship, Harvard University, Kennedy School of Government.

Preface

How did this journey to write a textbook begin? After graduating from the US Naval Academy with a degree in Mathematics, I flew helicopters and served in the US Marine Corps for 26 years, piloting the mighty CH-46E Sea Knight (the "Battle Phrog") during four overseas deployments. Along the way I earned my primary Masters Degree in Operations Research from the Naval Postgraduate School. One of my favorite classes while there was in cost estimation, and in particular I liked the usefulness and practicality of the subject matter. After graduation, I was fortunately assigned to the Marine Corps Research, Development and Acquisition Command (MCRDAC, now MARCORSYSCOM) in Quantico, Virginia for three years in Program Support Analysis (PSA), where I exclusively concentrated on Cost Estimation and Test and Evaluation. Those three years calculating costs, utilizing numerous quantitative methods, briefing seniors on the analysis and results found, very much stoked the fire in me for this area of study. I returned to the fleet and flew and deployed a few more times over the next nine years, and then returned to NPS as a military faculty member in the Operations Research Department in August 2000. When I arrived, the cost estimation class was being taught by Tim Anderson. Within a year, Tim retired from the US Navy and the course was seeking a new instructor. I eagerly volunteered to teach the class and have been doing so since 2001. At this time, approximately 250 students take the class each year, either as a resident student or as a distance learning, non-resident student.

While the cost estimation course had been taught with very detailed PowerPoint slides for many years before Dr. Daniel A. Nussbaum and I arrived, numerous students requested a textbook because they wanted either more detail or some "gaps" filled in and thus the idea for this textbook was born. Fortunately, publishers were very interested because little information like this existed on the market! The primary goal of this book is to help provide the tools that a cost estimator needs to predict the research and development, procurement, and/or operating and support costs or for any elements in a work breakdown structure in those phases.

While this textbook took 1.5 years to complete, many people have helped to provide information for it, either directly or indirectly. Anyone who has ever taught the cost estimation class here at NPS has had a hand in developing the slides that have been used for many years. The list of instructors at NPS includes Tim Anderson, Dan Boger, CDR Ron Brown (USN), CAPT Tom Hoivik (USN), CDR Steven E. Myers, (USN), and Mike Sovereign. In addition, a few others who provided significantly to the slides include: Major Tom Tracht (USAF), Dr Steve Vandrew of DAU (now NAVAIR), and the late Dr. Steven Book of the Aerospace Corporation. Thanks to all of these individuals and our apologies in advance if we have missed anyone! A few of the examples in the textbook are exactly

from the original slides, while most others have been revamped or upgraded to align with the times.

Our goal when writing this textbook was to make it sound more like a "conversation" than a textbook, as if we were in the same room together and we were speaking to you about the topic. We hope that we have achieved that goal! We have also tried to make this textbook applicable to both DoD and non-DoD entities and organizations, to their cost estimators, and to those involved in Cost Engineering.

When I was a student, a pet peeve of mine was when an instructor would say something like "It is intuitively obvious that....." and would then proceed from Step A to Step D without explaining what was happening in Steps B and C! Well, of course, it is intuitively obvious if you have been working in the field for twenty-five years! But for those of us who are reading about a subject for the first time and are just being introduced to the subject matter, maybe it is not so intuitively obvious. Consequently, in the quantitative chapters in this textbook, you will see calculations provided to almost every problem. If the calculations are not provided, it is only because a previous example already went through the same steps.

It was also our intent to try to stay away from really technical math terms like "degrees of freedom" and describing a term as an "unbiased estimator for the variance." Instead, our goal is to provide explanations in layman's terms in easy-to-understand language that "just makes sense." The book is organized as follows:

- Chapter 1: A historical perspective from Dr. Daniel A. Nussbaum on the past thirty years in cost estimation and changes he has observed along the way.

- Chapter 2: Background information on what cost estimation is and why it is important, the DoD acquisition process and the phases and acquisition categories of developmental programs and the key terminologies used in the acquisition environment.

- Chapter 3: The non-DoD acquisition process and the key terminologies, processes and methods used in the cost estimating field.

- Chapter 4: Once you are assigned a cost estimate, where do you get your data from? This chapter discusses many data bases, websites and the most significant cost reports used in cost estimating, as well as providing an introduction to Earned Value Management (EVM).

- Chapter 5: Once you have your data, how do you normalize it so you can accurately use it? You must normalize for content, quantity, and inflation, with normalizing for inflation being the most important of the three.

- Chapters 6–14: The "meat" of the book, involving the myriad quantitative methods utilized in the field of cost estimation, including regression analysis, learning curves, cost factors, wrap rates, and the analogy technique.

- Chapter 15: An overview of software cost estimation.

- Chapter 16: An overview of both Cost Benefit Analysis/Net Present Value, and Risk and Uncertainty.

- Chapter 17: Summary.

There are a number of people whom I would like to thank for their help in this textbook endeavor. First, I would like to thank Anne Ryan and Jae Marie Barnes for volunteering to review a number of the chapters in this text and offering ideas, suggestions, and recommendations on what to keep and what to change; Kory Fierstine for keeping me updated on DoD policy and always being there to listen and help when I had questions; Dr. Sam Buttrey for patient answers to my statistical questions; and I would like to thank

many of my former students whom I have had in class along the way – your questions, suggestions, and ideas have helped me learn as much from you as I hope you learned from me. I would also like to thank my two sisters, Hope and Barbara, and my mom for their love and support over the years. I wish my dad were still alive to read the book. He may not have understood too much of this subject, but he still would have been one of my biggest supporters along the way!

I owe my biggest thanks to my co-author Dr. Daniel A. Nussbaum, for being such a great friend and mentor over the past ten years that we have worked together. Dan has never hesitated to answer a question, share his knowledge in any subject area that I perhaps needed help in, and provide sanity checks to ideas. We also started the Masters in Cost Estimating and Analysis (MCEA) degree program here at NPS in 2010. It has been a fun journey together.

Gregory K. Mislick

After completing a Ph.D. in theoretical mathematics and teaching mathematics to under-graduates, I was invited to attend a postdoctoral Fellowship in applied mathematics, and within one year I was tackling operations research projects for the US Army Concepts Analysis Agency (now Center for Army Analysis). A few years later, I transferred to the headquarters of US Army Europe, in Germany, in order to head a division called the Economic Analysis and Systems division. The issues facing a major command often required projections of the impact of budget cuts, cost estimates in the ever-changing multiple currency environment, and many other demands for cost estimation and analysis.

Shortly after returning to the US, I began working for the US Navy. Ever since, I have been specializing in cost estimation projects and problems inside the Department of Defense and outside the DoD with other US government agencies, other nations, and with private organizations. In 2004, I was invited to return to teaching at the Naval Postgraduate School, where, together with Greg Mislick and others, we developed the first dedicated Master's degree program in cost estimating and analysis.

It has been clear to me for many years that a textbook would be helpful to students who are learning this discipline, along with its underlying mathematical and analytical principles and methods. I was delighted to join Greg Mislick in developing this text and I appreciate the hard work he has put into it. He and I have also teamed up success-fully, with others, to be awarded the prestigious Cost Analysis and Management Sciences Community Award presented by the Assistant Secretary of the Navy for Financial Management and Comptroller in 2010, and the 2014 Military Operations Research Society's Barchi prize, awarded to recognize the best paper given at their Symposium.

There are numerous professional heroes who provided me guidance and wisdom through the years and I would like to briefly recognize and thank them for all they have done while I have this opportunity to do so. First, I would like to reach back to August O. Haltmeier, my high school mathematics teacher, an unpretentious, professional man who fired my interest in mathematics, ultimately leading to my interests in econometrics, operations research, and cost estimating. The list also includes my thesis advisor, Dr. John Rankin Kinney, who showed much patience and confidence in me and whose kindnesses I strive to repay indirectly. I also thank Leonard S. Freeman, who showed me that analysis could have impact inside the Department of Defense in order to make better, and more defensible, recommendations about large and important national security issues. Finally, I thank Dr. Richard ("Dick") Elster and Dr. Douglas A. ("Doug") Brook, both of whom have been at the Naval Postgraduate School, who showed me by example that one can combine careers that serve both the public interest and the academic world.

There are also my personal and familial heroes. Rosa Meyer Nussbaum, Aaron Nussbaum, Ethel Arac Nussbaum, and Fernand Nussbaum, who form the best quartet of grandparents and parents that anyone could ask for. They lavished love and guidance, conscience and practicality, and unwavering support for family and education. I am strengthened daily by my daughters, Elise Jacobs and Amy Craib, and their families, and I take great delight and pride in watching them shape their own visions, values, professions, and ways of life, creating stability in an ever-more complex world. Their examples solidify and validate my optimism for the future. Closest to home, and therefore of the greatest consequence, is my wife, Beverly Davis Nussbaum, who has lovingly shared all of my adult life and experiences. Her insight, calming words, editorial prowess, and unending love make my achievements possible.

Daniel A. Nussbaum

Acronyms

ACAT	Acquisition Categories
ACWP	Actual Cost of Work Performed
AFCAA	Air Force Cost Analysis Agency
AFIT	Air Force Institute of Technology
AFTOC	Air Force Total Ownership Cost
ANOVA	Analysis of Variance
AoA	Analysis of Alternatives
AUC	Average Unit Cost
B$	Billions (of Dollars)
BCE	Baseline Cost Estimate
BCWP	Budgeted Cost of Work Performed
BCWS	Budgeted Cost of Work Scheduled
BRAC	Base Realignment and Closure
BY$	Base Year Dollars
CAE	Component Acquisition Executive
CAIG	Cost Analysis Improvement Group
CAIV	Cost as an Independent Variable
CARD	Cost Analysis Requirements Description
CBA	Cost Benefit Analysis
CBO	Congressional Budget Office
CCDR	Contractor Cost Data Report
CDD	Capabilities Development Document
CDF	Cumulative Distribution Function
CEBOK®	Cost Estimating Book of Knowledge
CER	Cost Estimating Relationship
CES	Cost Element Structure
COA	Course of Action
COTS/NDI	Commercial Off-the-Shelf/Non-Developmental Item
CPR	Contract Performance Report
CAC	Cumulative Average Cost
CV	Cost Variance (or) Coefficient of Variation
CY$	Constant Year Dollars
DAB	Defense Acquisition Board
DAES	Defense Acquisition Executive Summary
DAMIR	Defense Acquisition Management Information Retrieval
DASA-CE	Deputy Assistant Secretary of the Army for Cost and Economics
DAU	Defense Acquisition University

DCARC	Defense Cost and Resource Center
DoD	Department of Defense
DoDCAS	Department of Defense Cost Analysis Symposium
DOE	Department of Energy
DRPM	Direct Reporting Program Manager
EMD	Engineering and Manufacturing Development
EVM	Earned Value Management
FFRDC	Federally Funded Research and Development Centers
FP	Function Points
FV	Future Value
FY	Fiscal Year
FY$	Fiscal Year Dollars
FYDP	Fiscal Year Defense Plan
GAO	Government Accountability Office
IC	Intelligence Community
ICE	Independent Cost Estimate
ICEAA	International Cost Estimating and Analysis Association
ICD	Initial Capabilities Document
IT	Information Technology
JIC	Joint Inflation Calculator
K$	Thousands (of Dollars)
LCCE	Life Cycle Cost Estimate
LLF	Lost Learning Factor (for production breaks)
LMP	Lot Midpoint
LN	Natural Log
LRIP	Low Rate Initial Production
M$	Millions (of Dollars)
MAIS	Major Automated Information System
MC	Multi-Collinearity
MCEA	Masters in Cost Estimating and Analysis (at NPS)
MDA	Milestone Decision Authority
MDAP	Major Defense Acquisition Program
MILCON	Military Construction
MIL-STD	Military Standard
MTBF	Mean Time Between Failure
NCCA	Naval Center for Cost Analysis
NPS	Naval Postgraduate School
NPV	Net Present Value
O&M	Operations and Maintenance
OSD	Office of the Secretary of Defense
OSD CAIG	Office of the Secretary of Defense Cost Analysis Improvement Group
OSD CAPE	Office of the Secretary of Defense Cost Assessment and Program Evaluation
OSMIS	Operating and Support Management Information System (Army)
PDF	Probability Density Function
PDRR	Program Definition and Risk Reduction
PEO	Program Executive Officer
PM	Program Manager
POE	Program Office Estimate
POL	Petroleum, Oil and Lubrication

PPBES	Planning, Programming, Budgeting and Execution System
PV	Present Value
QA	Quality Assurance
RDT&E	Research, Development, Test, and Evaluation
RFP	Requests for Proposals
R-HW	Recurring Hardware
ROI	Return on Investment
ROM	Rough Order of Magnitude
SAR	Selected Acquisition Report
SCEA	Society of Cost Estimating and Analysis
SCP	Service Cost Position
SD	Standard Deviation
SE or SEE	Standard Error or Standard Error of the Estimate
SE/PM	Systems Engineering/Program Management
SLOC	Software Lines of Code
SME	Subject Matter Expert/Expertise
SRDR	Software Resource Data Report
SV	Schedule Variance
SYSCOM	Systems Command
TY$	Then Year Dollars
UFC	Unified Facilities Criteria
USD (AT&L)	Under Secretary of Defense for Acquisition, Technology and Logistics
VAMOSC	Visibility and Management of Operation and Support Costs
WBS	Work Breakdown Structure
WBSE	Work Breakdown Structure Element
WSARA	Weapons Systems Acquisition Reform Act (2009)

CHAPTER ONE

"Looking Back: Reflections on Cost Estimating"

We are delighted that you have chosen to learn about a vibrant career field that few people know about in any significant depth: the field of cost estimating. However, before we discuss the background, terminologies, statutory requirements, data sources, and the myriad quantitative methods involved and introduced within this textbook to help you become a better informed or better cost estimator, I would like to first discuss with you the idea of cost estimation as a profession. There are two facts that are most important in this field that are in seeming contradiction of each other: first, that cost estimating is "ubiquitous," always existing either formally or informally in every organization; and second, that it is often "invisible," or at least many times overlooked. My goal in this chapter is to provide some personal observations from my 30+ years of experience to shed light on the many significant purposes and roles played by cost estimation. These experiences also provide the opportunity for me to thank the many leaders, mentors, coworkers, and others who have provided me skills and insights throughout my career and whose contributions are reflected in this book. Lastly, I will comment on important changes that have been occurring in the profession within the last 30 years and some forecasts of future changes.

In the past, nobody went to school to become a cost estimator. To illustrate this point, I studied mathematics and economics in school, while my co-author, Greg Mislick, also studied mathematics and flew helicopters for the U.S. Marine Corps in his previous career. By different routes, we became practitioners of operations research and specialists in addressing the issue of "What will it cost?" In recent years, however, there have been graduate-level certificate and master's degree programs introduced to hone the skills of, and establish professional standards for, cost estimators. This textbook is our attempt to pass those skills and lessons learned on to you and to increase the knowledge and experience of those working in this field.

Every organization – from a typical household to the greatest nation – relies upon the disciplines and processes of this profession. The one question that typically comes up in a conversation about most topics is "What does it (or what will it) cost?" It is a question that you and I will ask numerous times in both our professional and personal lives. The most frequent response that we get to this question (especially from those who do not really want to give us an answer) is "I can't tell you exactly," as if this were a useful or satisfactory response. The answer is just a dodge. We were not expecting an *exact dollars*

and cents answer. Rather, we were looking for an *approximate* answer to help us plan for the expense while we sort through the various options that we may be considering.

The characteristics of a useful answer to the "What does it cost" question that we look for in the circumstances of our daily lives are remarkably similar to those answers in commercial and government applications of cost estimating, although complexity and scale may camouflage this similarity. The essential characteristics of any good cost estimate are completeness, reasonableness, credibility, and analytic defensibility. Note that this list does not include the need to make sure the answer has plenty of *precision*. While the budgets we develop and the cost-benefit analyses that we construct require specific numbers in them, our professional work as cost estimators does not rely on getting answers "correct to the penny." A cost estimator should not agonize over the lack of a narrow range of possible costs for the cost estimate. If the ranges are overly large, the user of the estimate, such as your sponsor, consumer, boss, or other person who asked for the cost estimate, may tell you that they need a tighter range of costs, and you will then need to seek additional data or another methodological approach to support the refinement of your estimate. However, it may also be that a wide-ranging estimate that meets the criteria of completeness, reasonableness, credibility, and analytical defensibility is all that is required in the case of rapidly changing conditions and immature technologies. In fact, in the absence of data, it may be *all* that is possible.

This textbook is designed specifically to provide context to those cost estimating objectives of completeness, reasonableness, credibility, and analytical defensibility. Moreover, it will teach those mathematical techniques and procedures that are relevant to develop cost estimates and it will provide you with significant guidance through that development process.

I would like to share with you now six examples/stories that illustrate some of the lessons that I have learned while holding various positions in the corporate and government worlds. This includes positions while at a headquarters and during duty in the field, while inside the United States and abroad, and mostly (but not exclusively) while serving in defense-related positions. These examples are diverse, indicating the broad applicability of the tools of cost estimating. I hope that you will find them helpful while tackling programs and assumptions of your own in your present (or possibly future) career in cost estimating.

Example 1.1 Cost Estimation in Support of a Major Ship Program As the Navy proceeded to build the inaugural (lead) ship in a new class of ships, large cost growth began to challenge the project. I was asked to figure out why this was occurring. After much analysis, I found that there were several reasons for the growth. One reason, in particular, that is useful to discuss for the education of new cost estimators, was that one of the cost driving assumptions made during the original cost estimate was simply, and significantly, incorrect. The original cost estimate had assumed that when the Navy's lead ship was being built in the shipyard, there would be another *commercial* ship under construction at the same time, and these two ship programs would share the shipyard's overhead costs. This would relieve the Navy's ship from carrying the full burden of the shipyard's overhead costs by spreading these costs over the two ships. The cost estimators who produced the original cost estimate had relied upon credible information and had exercised appropriate due diligence. They had confirmed that the shipyard had the work for the second ship on its order books, and they received confirmation from the Defense Contract Audit Agency (DCAA) of this order, as well as DCAA's satisfaction that this commercial ship would be built in accordance with the contract. It was thus reasonable to assume that the shipyard's overhead rates would be split between the two ships. However, when the Navy ship was

being built, the commercial firm canceled its ship building order, for its own business reasons. Consequently, it turned out that there was no second ship in the yard with which to share overhead rates, and these overhead costs had to be covered entirely by the US Navy!

Naturally – and through no fault of the US Navy – there were indeed significant (and legitimate) cost increases. They did not occur due to inexperience or due to naïve program management by either the government or the shipyard. Thus, the first lesson learned in this example is that *assumptions always drive cost*. The second lesson learned is that *assumptions can be fragile*. While a key assumption in your program may be accurate at one moment in time, it may not be accurate at a later time. This was indeed the case here. The third lesson learned in this example is that the one constant you can always rely on is change. *Change will always occur*! Even a cost estimate completed with due diligence and reasonableness can be wrong. After all, in this example, all the evidence pointed to the fact that there would be a second ship in the yard. In conclusion, be aware that plans and circumstances do change during the life of your program, and they most likely will. Ultimately, these changes will affect your estimated and actual costs.

Example 1.2 Cost Estimation in Support of a Major Missile Program   A frequent root cause of underestimation in a cost estimate (and therefore a strong invitation to a cost overrun) is the omission of the cost of a significant component of the necessary work. In cost estimating, we refer to such a component of the necessary work as a work breakdown structure element (WBSE). Whether the omission occurs purposely or accidentally makes no difference, for as the project is executed, labor costs and material costs associated with the missing WBSE will still accrue, the bills will be presented for payment, and a cost overrun will then occur.

A cost estimate on a missile that was an expensive new major defense weapon system provides an example. Let's call our missile program that we are costing Program A. Experience with such weapon systems that were part of Program A had taught me that a sensor would be part of this weapon system; in fact, I expected that the development of this sensor would be one of the *major* cost elements in the research and development (R&D) phase of the lifecycle cost estimate. The program office confirmed that a new sensor was indeed part of the design of this weapon and that the new sensor was integral to the weapon's successful performance. However, the program manager's R&D cost estimate *did not include the sensor*!

When I asked the PM why there were no sensor costs included, he stated that a separate program (which we will call Program B) was developing the sensor and that his Program A would do a "technology lift" from Program B, thereby avoiding any sensor development R&D cost to his program. While I understood this argument, I also knew that there was no guarantee that Program B would be able to complete its development of the sensor and make it available within the timelines of our program, and I was skeptical that there would be no sensor-related R&D charges in our program. The key problem, however, was that if the sensor development in Program B was delayed, it would then delay Program A, extending our schedule until the sensor technology was in fact completed. Any extension would then cause additional costs. Consequently, I argued for the identification of contingency funds in the program to cover this possibility. Fortunately, the program manager agreed, which proved to be fortuitous when Program B (that was developing the sensor) was ultimately canceled. Thus, we had to restructure Program A to incorporate the sensor development project within our own budget.

The major lesson learned here is that technology that is not developed or "mature" always presents the very real possibility that it just may not work, or may be delayed in its development, and a dependent program will then also be delayed, with corresponding

increases in costs. For instance, when a program is delayed, personnel costs will increase since the workers still need to be paid, but now over a much longer period of time. There is also a more general lesson, which is that it is important to identify all technological and other risks to any program and consider their cost impacts, and to then develop contingency cost estimates under the assumption that these risks may unfortunately come to pass.

Example 1.3 Cost Estimation in Support of a Major Ship Program   For years, the US Navy knew that it needed to modernize its combat logistics fleet (CLF). However, during those years, the shipbuilding appropriations were being used nearly in their entirety to develop and procure surface combatant ships instead. To work around this funding problem, a clever idea suggested that a commercial shipyard build the next generation of CLF ships, with the intention that upon completion of each ship, the Navy would then enter into a "long-term lease" for the ships. This would thus allow the CLF to be funded from the operations and maintenance account (the O&M, N appropriation) of the Navy, rather than funding it from the shipbuilding appropriations, as was the norm. I was asked to analyze whether this arrangement made financial sense, while others were examining the financial and legal implications of this potential arrangement. My analysis was to be a "cash flow" and/or "net present value" cost benefit analysis, comparing the cost of the conventional method of procuring this ship from shipbuilding appropriations with the proposed "build-to-lease" option using operations and maintenance dollars. I also needed to include the many "what-if" analyses to test the sensitivity of the bottom line cost to variations in the assumptions and values of variables being used in the analysis.

   After significant study, we found that under a wide variety of reasonable circumstances, the proposed idea of "build-to-lease" made financial sense. If you were to consider only the financial metrics of the analysis, a reasonable person would be led to propose the course of action which leveraged savings from the "build-to-lease" option. This cost beneficial proposal, however, was not brought to the attention of the Secretary of the Navy despite its cost and benefits, since it was deemed to be "poor public policy and practice" for a variety of reasons. In other words, no matter how financially attractive or analytically defensible this proposal was, it was matters of public policy that trumped the analysis and drove the decision to not "build-to-lease." The lesson learned here is that cost issues are always a *major* concern, but they are almost never the *only* concern. Cost estimating is a function that informs and supports decision-making. An analyst should not assume that decision-makers will inevitably follow the recommendations of his or her analysis, regardless of how complete, reasonable, credible, analytically defensible, and even elegant that it may be.

Example 1.4 Cost Estimation in Support of a Major Automated Information System (MAIS)   Often, it is important to do *rough order of magnitude* (ROM) cost estimates to help senior defense personnel sort through the cost implications of alternative courses of action for complicated projects. Knowing whether a program is going to cost *roughly* $75 Million or *roughly* $200 Million helps decision-makers distinguish between those options that are potentially viable and those that are not. A memorable example of this was the idea to develop an automated defense-wide system to support civilian and military personnel and pay functions across all of the military services and all "components" (that is, the Active forces, the Reserve Forces and the National Guard). This was intended to be the largest enterprise resource planning program of its kind ever implemented. We were tasked to "develop an estimate of what this new MAIS would cost and to compare that cost with the cost of maintaining approximately 90 legacy personnel and pay systems which

this program would replace." We were tasked long before any specific requirements (other than the broad description given in the previous sentence) had been fully thought out or discussed. At this time, there was certainly no estimate of the size of this program, and size is often a critical variable for developing a credible cost estimate. We recognized that this personnel and pay system would be a very large software and hardware effort, and to reasonably capture the cost of the software program, we needed an estimate of its size, measured either in source lines of code or in function points. The program manager certainly had no estimate of either. We were trying to "bound" the problem by saying (in effect) that this effort was going to be "bigger than a breadbox and smaller than a barn," or, as we decided, "bigger than any Microsoft product, but smaller than the largest missile-defense project." Obviously, the estimates we developed had a wide range of outcomes. Was this range of estimates useful to the decision-makers, or did they need exact answers in order to make their decisions? The important lesson learned here was that at the front end of a project, when many unknowns still exist, rough order of magnitude estimates with a wide range of possible cost outcomes may still be sufficient for senior decision-makers to move the decision process forward.

Example 1.5 Cost Estimation in Support of a Major Policy Decision A major overseas US Army command was having difficulty with retention and maintaining its force structure within the enlisted ranks. One significant part of the problem was identified as low re-enlistment rates among the junior enlisted members. Based on US Army regulations at the time, the command's policy allowed a service member's family to accompany him or her overseas. The Army would pay for the overseas move and also support the family overseas with a housing allowance and base exchange privileges, but only if the service member had attained at least a minimum rank. The junior enlisted personnel who were not re-enlisting at sufficient rates were precisely those who were below the rank necessary to have their families accompany them and to receive those elements of family support. Consequently, their families usually remained in the US while the service member completed the overseas tour. In that way, retention suffered due to hardships caused by this separation. We proposed a policy that would extend this family support to these junior enlisted members. Our analysis addressed whether the estimated costs of implementing this policy outweighed the benefits of the estimated lower recruiting and training costs due to higher retention rates. The rough order of magnitude estimates that we provided were sufficient to convince the Army to implement this policy and the results of this policy change did indeed increase enlisted retention for the U.S Army. *For the cost analysts, it is highly satisfying to see crunched numbers turn into effective policy.*

Example 1.6 Cost Estimation in Support of a MAIS Program This example involves analyzing the issue of when to insource or outsource goods and services. While this particular case is taken from the US Navy, it has broad applicability in all services and businesses. The Navy was considering outsourcing its ashore (i.e., its "non-ship") information technology (IT) infrastructure and operations, including all of the computer hardware, software, training, and help desks. Even before the Request for Proposal (RFP) was developed or published, the Navy required a cost benefit analysis to address the important issue of the Return on Investment (ROI) of such an enterprise. ROI is simply a fraction. The numerator is the savings estimated from the proposed outsourced system when compared to the existing (or status quo) system, while the denominator is the estimated cost of the investment required to transition and maintain the proposed outsourced system. The challenge in this analysis was that we had little data to estimate either the numerator or denominator! The Navy did not have the costs for either the infrastructure or operations of the existing

system, nor did the Navy have any insight into which vendors might bid for a comprehensive outsourced set of goods and services and therefore, what it would cost to capitalize and operate the proposed outsourced option. Given those significant restrictions, we still made reasonable assumptions to characterize these unknowns, and, from these assumptions, we developed our cost estimates. Subsequently, we conducted extensive sensitivity analyses in order to understand the relationship between ROI and the different values of important independent variables. We identified a critical variable in these computations, which was the fraction of best commercial practices which the Navy would be likely to achieve if it transitioned from the existing insourced option to the proposed outsourced option. For example, if the Navy was able to harvest only a small fraction of the benefits of the commercial best practices (say something in the 0 to 20% region), then the hoped-for savings would not be achieved, and the ROI would be negative or unattractive. On the other hand, if the Navy was able to harvest a larger fraction of the benefits of the commercial best practices (say something in the 80 to 95% region), then the hoped-for savings would be achieved, and the ROI would be positive and attractive. Therefore, we could present a two-dimensional graph to senior decision-makers. The horizontal-axis of the graph was the "percentage of best practices achieved," and the vertical-axis of the graph was the "ROI of investing in the new system." This graph compresses a great deal of the assumptions, methodology, and analysis of the cost and benefits into a single visualization tool. Of course, there is a corresponding loss of analytical nuance and texture in this simplification. Nevertheless, it was precisely the simplicity and transparency of this tool that permitted very senior decision-makers to make a reasoned decision, grounded in credible and defensible analysis. The lesson learned here was it is not necessary – in fact it is almost *never* necessary – to say to the decision-maker that "This is the correct answer, and here is the analysis that supports that assertion." Rather, the objective of the cost estimator should be to think of him or herself as a "guide" to the decision-maker, piloting him or her through the decision space that underlines the problem at hand. In that way, the analysis is left in the hands of the analyst, and the decision is left in the hands of the decision maker, both as it should be.

Hopefully, these six examples helped you to see the difficulties that can be encountered throughout your program, and the lessons learned from these examples. At this point, I want to transition from talking about specific examples that provide lessons learned in cost estimating to reviewing some of the changes that have been occurring in the cost estimating community over the past 30 years and changes that I feel are likely to occur in the future. It is useful to consider these changes within three dimensions: *people, processes, and the data*. The following are descriptions of these three dimensions, complete with some personal observations.

People: When discussing "People," the following questions need to be considered:

- Who are the people who enter the cost estimating profession?
- What are their academic and professional backgrounds?
- Where do they obtain their education and training in the intellectual, technical, methodological, and ethical requirements of the profession?

Thirty years ago, government cost estimators entered the profession primarily with bachelor's degrees in the engineering sciences, followed by degrees in mathematics, statistics, economics, and operations research. With the exception of a master's degree

program at the Air Force Institute of Technology (which was by and large for young Air Force officers, Captains and below), there were no formalized, degree-granting courses of instruction in cost estimating at civilian colleges, universities, or military schools, either at the undergraduate or graduate level. New entrants into the field learned their skills largely in an apprenticeship mode; that is, they were assigned to work on a project under the supervision of someone who already had experience and knowledge, and the new entrants continued in that mode until they had learned enough to become the mentors to the next entrant. To combat this lack of formal education in the cost estimation field, the Office of the Secretary of Defense, Cost Analysis Improvement Group (OSD, CAIG) commenced an annual Department of Defense Cost Analysis Symposium, called DODCAS. DODCAS served as a forum for the exchange of information and skills within the Department of Defense's extended cost estimating community, and remarkably is still going strong, although there has been a recent hiatus due to the exigencies of sequestration and other recent budget stressors! This is not just for attendance by members of DoD organizations, but also for DoD contractors as well. (For more information, conduct an internet search for "DODCAS").

Fortunately, the landscape for education within the cost estimating community is changing. Beginning cost estimators are now better educated, as many new entrants already have masters degrees in the previously mentioned disciplines. Formal and informal internship programs have also been developed by organizations for their new entrants. We are entering the era of more advanced education now available specifically in the cost estimating field.

To amplify this last point on education, there are now three major repositories of cost estimating intellectual capital, with two of them leading to certification in cost estimating and the third to a master's degree in cost estimating and analysis. These three include the following:

- The first major certification source is the *International Cost Estimating and Analysis Association* (ICEAA; http://www.iceaaonline.org/), formerly known as the "Society of Cost Estimating and Analysis" (SCEA), which has developed sophisticated training and certification programs. While these programs are used by some people and organizations within the DoD, the primary customers are those contractors and consultants who provide goods and services to the government, as well as commercial firms whose work is not oriented to government needs. Commercial firms that wish to have uniform standards across their enterprise have adopted the ICEAA certification process as their standard. More is said on this topic in Chapter 3.

- The second major certification source is the *Defense Acquisition University*, which primarily supports personnel who work in various areas related specifically to the Defense Department's processes and the business of acquiring goods and services. This includes training and certification in cost estimating. Numerous training modules are available in a wide variety of subject areas.

- While the first two sources provide certifications, the Naval Postgraduate School (NPS) and the Air Force Institute of Technology (AFIT) developed a joint distance learning *Master's Degree program in Cost Estimating and Analysis*. The first cohort of thirty students commenced their studies in late March 2011, and that same cohort proudly graduated in March 2013. The second cohort with 26 students graduated in March 2014. Cohorts commence annually each spring for this two-year program. The program is open to all who meet its entrance requirements. The program is unique in granting a Master's degree specifically in cost estimating and analysis.

- Further information is available on the NPS website at: http://www.nps.edu
  /Academics/DL/DLPrograms/Programs/degProgs_MCEA.html

It is hard to predict the future. However, in a limited, quantitative way, that is precisely what the profession of cost estimating does. I expect the future personnel mix will include many more people with formal training in cost estimating from the various sources now available, especially the distance learning *Master's Degree program in Cost Estimating and Analysis* described in the previous paragraph. The timing is apt as these well-trained personnel will be needed to replace a population of baby boomers who are turning toward retirement.

Having discussed the people involved in the cost estimating field, I next turn my attention to the *Processes* involved in cost estimating.

Processes: When discussing "Processes," the following questions need to be considered:

- What are the intellectual underpinnings of the profession that permit cost estimates to be made?
- What are the processes by which cost estimates are developed, validated, and inserted into the larger decisions that affect acquisition, budgeting, and analysis of options?

Thirty years ago, the main methodologies for developing cost estimates were cost factors, cost estimating relationships (aka CERs, which are equations that express cost as a function of technical or performance variables), and learning curves, all of which will be explained fully in this text. These methodologies were underpinned by "actuals," cost data from analogous, historical programs. Risk and uncertainty in a cost estimate were addressed by the standard procedure of sensitivity analysis, which observes how the baseline cost estimate behaves when important assumptions are varied through reasonable ranges. The results of sensitivity analyses permit the analyst to provide a more nuanced and robust cost estimate than one that merely provides a point estimate. For example, rather than stating that "The baseline cost estimate for Program X is $1.2B (in FY13$)," the cost estimator can instead state that "The baseline cost estimate for Program X is $1.2B (FY13$), and the range of cost outcomes is expected to be $1.0B to $1.3B as the quantity procured is increased/decreased within a range of ±10%."

Thirty years ago, we were also just beginning to automate all of the computational aspects of our estimates. Prior to this time, it was quite normal for estimates to be developed on big paper spreadsheets, with arithmetic done manually, or for more complicated computations such as learning curves, computed on an electro-mechanical calculator. The advent of the laptop-based program called VISICALC (an early antecedent of today's ubiquitous EXCEL-type spreadsheets), caused quite a stir in the cost estimating community! Its consequences far exceed the simple cost savings in the number of pencils and erasers that we no longer need to use. More importantly, we are now able to do many more sensitivity analyses within the timeframe allotted to do the study, thereby providing a richer set of supports to the decision-makers.

As a new program wound its way through the hierarchical decision process, the associated cost estimate traveled with it in a review cycle that checked for the completeness of the following areas:

- Work breakdown structure elements (WBSE)
- The credibility of the analogous program data
- The correctness in accounting for all assumptions

- Applying appropriate and accurate methodologies
- Checking for other possible errors

Cost estimates were briefed to several levels of decision makers, ultimately ending at very high management levels. This process was a good indication that cost estimates, especially for major projects, were taken seriously. Back then, the Army, Navy, and Air Force already had Service Cost Centers to perform these functions, and OSD CAIG (now OSD CAPE) reviewed the largest programs. (For a history of the CAIG, see Reference 1 at the end of this chapter). These reviews were held under the auspices of regulation, meaning that each service had developed its own regulations for proper handling of cost estimates. Statutory authorities came later, when laws were developed that mandated that the cost estimates be accomplished in accordance with certain standards.

In 2015, the review processes remain largely intact. Cost factors are still derived from analogous, historical programs, and CERs are still derived using statistical regression techniques and historical data. Moreover, sensitivity analyses are still done. But now, with our capability to do Monte Carlo simulation analyses with great rapidity on automated spreadsheets, we are instead easily able to generate probability distributions for our cost estimates, thereby providing richer statistical insight for decision makers.

Three important changes to the decision-making process occurred to aid decision makers. These include the following:

- The first change is that there are now statutory requirements (mandated by law and Congress) that govern important phases of the cost estimating process. One example is the requirement for all major programs (typically, programs whose cost estimates exceed certain monetary thresholds) to have two cost estimates completed on them. One estimate is completed by the responsible program office (called the Program Office Estimate, or POE), and the other one is an Independent Cost Estimate (known as an ICE) to check and ensure the completeness and reasonableness of the POE.

- The second change is that the ICE must be considered by the Secretary of Defense in making the Milestone decisions. Milestones are periodic reviews performed as a program progresses through its life cycle from conceptualization, to research and development, to procurement, to operations and maintenance, and finally, to disposal. The fate of a program – whether to be continued or restructured or cancelled – is decided at these milestone reviews and the ICE is one of the big contributors to that decision. This requirement only calls for consideration of the ICE; it does not require the adoption of any particular cost estimate.

- The third change is that both the POE and the ICE must include the cumulative probability distributions that are generated by Monte Carlo simulations, as previously discussed.

The passage of the Weapon Systems Acquisition Reform Act in 2009 (WSARA 2009) heightened the visibility of the cost estimator and the cost estimation process within the Department of Defense and this intense focus on costs and on the education of increased numbers of cost estimators, plus the processes by which costs are estimated, will continue. WSARA 2009 mandated the creation of the Director of Cost Assessment and Program Evaluation (OSD CAPE), thus elevating the purpose and visibility of the cost estimation field. The director position requires Senate confirmation, and the office has very broad execution and reporting responsibilities in both cost estimation and cost-benefit analysis.

Data: Last, I turn to the Data used in cost estimating. Data includes the various measurements that are taken of the project for which a cost estimate is being developed. These measurements are often numerical values, but they may also be categorical variables. Examples of these measurements include cost metrics, physical measurements such as size and performance descriptions, and programmatic descriptors, such as the quantities being bought, the duration of the R&D phase, and the number of vendors to whom procurement contracts will be awarded. As you will explore in depth later in this book, data derived from analogous historical programs (aka "actuals") are the indispensable core of every professionally developed cost estimate. Therefore, the identification, collection, normalization, organization, storage, and analysis of data underpin everything that we do as cost estimators.

Thirty years ago, the collection of historical data was an "ad hoc," paper-based effort. The Department of Defense asked vendors to provide cost reports, showing time-phased costs incurred on a program that are allocated to the various elements in the project's work breakdown structure. Contractors were generally under no pressure to make these reports available, and consequently, to a large extent, they often did not provide these reports. The reports that *were* provided were accumulated in a library, and they were tended to by a very small cadre of dedicated personnel. Cost estimators often kept their own files of particular reports, and they protected them as valuable sources of analogous actuals for future cost estimates. I often think that this stage of cost estimating data-keeping was comparable to the medieval period in human history, (before printed books became available), in which hand-produced manuscripts were rare, valuable, and had very limited distribution.

Today's data resources situation is extraordinarily different from what it was 30 years ago. Data is provided by vendors in the form of cost reports, which are collected, normalized, subjected to error-searching routines, and filed in accessible web-based databases with due regard to proprietary data protection and security requirements. These reports and databases are described in Chapter 4 on Data Sources. Storage and accessibility of historical data continues to create a greater depth to the available data and easy access for those with the approved needs.

In closing this chapter and these reflections, it is hoped that these introductory remarks will provide you with an appreciation of the scope, applicability, difficulties, and utility of cost estimating and perhaps inspire you to master the material in this text and fine tune the diverse skills we apply to the complex question of "What will it cost?" Chapter 2 will begin this journey: it will discuss what cost estimating is; what the characteristics of a good cost estimate are; why we do cost estimating and the importance of cost estimating in the Department of Defense (DoD) and in Congress; how and when cost estimates are created and used; and cost estimating terminology.

Reference

1. Don Shrull, ed. *The Cost Analysis Improvement Group: A History*. Logistics Management Institute (LMI), McLean, VA, 1998, ISBN 0-9661916-1-7, Library of Congress Catalog Card Number 98-65004.

Introduction to Cost Estimating

2.1 Introduction

In this chapter, we will introduce the basics and the background associated with the field of cost estimating. We will discuss the purpose of cost estimating, what the characteristics of a good cost estimate are, why we do cost estimating, what the importance of cost estimating is in the Department of Defense (DoD) and to Congress, explain the DoD 5000 series acquisition process and regulations to include the milestones and the acquisition categories, and also discuss how and when cost estimates are created and used in the acquisition process. We will end this detailed chapter by explaining key terminologies used extensively in this career field, and we will support them with substantive examples. This chapter will help you "talk the talk," helping you understand why we must do cost estimating, and under what guidelines we must operate while doing so.

2.2 What is Cost Estimating?

Danish physicist, Niels Bohr, who won the Nobel Prize in Physics in 1922, wrote "Prediction is very difficult, especially when it's about the future." He could very well have been speaking about the art and science of cost estimating! We commence our overview of this career field by providing a definition of cost estimating. You will note that it includes several important italicized words that we will explain one at a time:

> "Cost estimating is the process of collecting and analyzing *historical data* and applying *quantitative models*, techniques, tools, and databases in order *to predict* an estimate of the future cost of an item, product, program or task. Cost estimating is the application of the art and the technology of approximating the probable worth (or cost), extent, or character of something *based on information available at the time*." [1]

- *Historical data*: Like all other scientific endeavors, cost estimating is grounded in available data. With tongue in cheek, we point out that data from the future are

obviously hard to find. That leaves us, inevitably and necessarily, with the need for all relevant and available data from the past, which are known as historical data. A fundamental and major component of the process of the cost estimating profession is the search for historical data. Once found, you will need to work on the collection, organization, normalization, and management of that historical data.

- *Quantitative models*: In addition to the use of historical data, the profession of cost estimating is scientifically grounded by using transparent, rationally defensible and reviewable quantitative models. At any university where a class in cost estimating is offered, it is usually offered in the operations research, business, or systems engineering departments, attesting to its quantitative underpinnings.

- *To predict*: The profession of cost estimating is the business of predicting the future. Many people say "you cannot predict the future" but that is patently false. If I say "Tomorrow it is going to rain," I might be right, or I might be making a silly or incorrect statement. In either case, whether I am correct or incorrect, I am attempting to predict the future. Moreover, many of the things that we do as humans – from taking an umbrella to work to planning for our children's college education – are based on attempting to predict the future. Prediction, also known as forecasting, is a useful and even noble undertaking.

One often hears objections to the use of historical data to underpin estimates of future costs that sound something like this: "*When you are examining historical data, you are looking at the results of all the errors and mistakes and historical misjudgments that previous managers made, and you are using those errors as the basis to project the outcome of a new program with judgments and predictions that the current decision-maker must make. Therefore, you are predicting that the current Program Manager will make the same mistakes as the previous Program Manager, and there is no reason to believe that!*"

The answer to this complaint is that it is *not* true that you will make the same errors and misjudgments that your predecessors made. There are two reasons for this. First, it was rarely misjudgments or incompetence on the part of the previous PMs that led them to stray from the path of their project plans; rather it was often external events, ones beyond their control, to which they adjusted their project's plans. Second, you will face a different set of circumstances and a different set of external events that will inevitably be different from your project's initial plans, and that will force you to adjust your project's plans. A shorter, and more facetious, response would be that "You won't make the same mistakes that your predecessors made: you will make your own!"

- *Based on information available at the time*: When we are developing a cost estimate, we want to know the circumstances that will pertain to the time when the project will be executed. However, we can only use the information that is available to us now, and then attempt to estimate the conditions that will pertain later when the project is executed. Understandably, we cannot anticipate every condition or possible change of condition that may occur, especially if the execution occurs far into the future. For example, if we are estimating the cost of buying 2,000 F-35 aircraft, and the DoD eventually decides to buy 1,800 (or 2,500) of these aircraft instead of 2,000, then surely our estimate for 1,800 (or 2,500) will not accurately predict the cost for this project. Our estimate will be "wrong," but our estimate will have been in line with the "then" current and "best" program assumptions. In any case, the cost estimator needs to consider contingencies and stand ready to incorporate future changes in the circumstances and program plans into a revised and updated cost estimate.

2.3 What Are the Characteristics of a Good Cost Estimate?

There are numerous characteristics that need to be evident in a cost estimate, and as we stated in Chapter 1, the essential characteristics of any good cost estimate are completeness, reasonableness, credibility, and analytic defensibility. The following list provides some of the most significant operating characteristics:

- Good cost estimating practices are anchored in historical program performance. As previously stated, a cost estimate must be grounded in data from prior and analogous (similar or related) experiences. These experiences should be specifically cited as data sources, so that the user of the cost estimate can have confidence that the cost estimate is grounded in sound information that is based on appropriate prior experiences.

- A good cost estimate must also reflect the current and potential future process and design improvements. At first glance, this statement appears to partially contradict the previous one but indeed it does not. While historical data necessarily incorporate the processes of the past, it is also true that the new project for which we are developing a cost estimate will be accomplished with the benefit of updated design improvements and manufacturing processes. Although we do not have historical cost data for these updates and improvements, we still have to account for and estimate their impacts in our new cost estimate. This is most often done using "professional judgment," which is also known as "subject matter expertise (SME)." Like historical data, SME data adjustments and their sources and justifications should be cited and transparent.

- A good cost estimate must be understandable by program and business leaders. This pragmatic standard speaks to the need to have simpler-rather than more complex-approaches underlying a cost estimate. Those who are the recipients and users of cost estimates in their decisions are often very senior people, with little time to read the details of what we do. We are lucky when they read the executive summary of what we have done in its entirety! Therefore, you have a diminished window in which to convince those who use the cost estimate that it is reasonable and credible. That is a task more readily accomplished by simple approaches than by complex ones.

- A good cost estimate identifies its ground rules and assumptions. A friend of mine used to say "You let me make the assumptions and I'll let you do the analysis." His point was that it is the assumptions that drive the bottom lines of all analysis and therefore we have to pay close attention to what they are. It would be very nice if there were a set of assumptions that was accepted by all those who are going to use the cost estimate. This is probably not possible, given the diversity of audiences for a particular cost estimate. The best that we can do is to incorporate sensitivity analysis into our cost estimates, in order to accommodate variations in the baseline ground rules and assumptions.

- A good cost estimate addresses the risks and uncertainties inherent in the program plan. While it is true that the end result of a cost estimate may end up in a budget display as part of a point estimate (i.e., a single number representing what the cost is estimated to be), it is also true that this point estimate is sensitive to the underlying assumptions that were made as part of the analysis. In fact, there are several numbers, probably even a range of numbers that could have been designated as the point estimate, if we had made different assumptions. It is, therefore, important

that in presenting a cost estimate, that we point out these sensitivities, along with their impacts on the point cost estimate. We will develop this theme more fully in Chapter 16 when we discuss Risk and Uncertainty.

At a technical level, there are additional attributes, which increase our confidence in the reliability and credibility of the estimate, and which increase our trust that the work represents a sound analysis of the available data and, therefore, may be used as an estimate of future costs.

- A major attribute of a good cost estimate is that it is driven by requirements. This means, at a minimum, that all programmatic and system requirements are documented. We would not think of asking for a cost estimate on renovating the kitchen in our house without having a clear idea of what the renovation is intended to accomplish and include. This is also true with cost estimates for major commercial and defense systems. There are several documents that fulfill this need, including those used by DoD in its "requirements process." These include the Initial Capabilities Document (ICD), Capabilities Development Document (CDD), and others. There also may be project or system specifications, such as those laid out in Requests for Proposals (RFP). A very important document used for cost estimating, which will be discussed in Chapter 3, is the Cost Analysis Requirements Description (CARD). Similarly, in the nongovernmental world, companies usually develop and document specific requirements for new projects and systems.

- A good cost estimate is also based on the description of a project that has well-defined content, and which also has identifiable risk areas. These provide the technical basis for using the estimating methods that we bring to bear. This topic will be taken up when we address various cost estimating methodologies later in the text.

- A good cost estimate can be validated by independent means. For large government projects, which tend to have large anticipated expenditures, an accepted "best practice" within the cost estimating profession is to have an independent cost estimate, as specified in the Government Accountability Office's "GAO Cost Estimating and Assessment Guide," [2] as a check on the original cost estimate, which was performed by the Program Manager. In addition to the GAO document, a second source that attests to the importance of an independent cost estimate is the website for the Congressional Budget Office (CBO). CBO supports Congress' legislative-making responsibilities by "producing independent analyses of budgetary and economic issues to support the Congressional budget process." [3]

- A good cost estimate is traceable and auditable. This means that the cost estimate can be re-created from the data sources, ground rules and assumptions upon which it is based. An informal standard in the cost estimating community is that a cost estimate should be transparent enough for a reasonably smart high school student with a modest degree of numeracy to follow the arguments used in the development of the cost estimate, apply the data and assumptions, and be able to reproduce the estimate.

2.4 Importance of Cost Estimating in DoD and in Congress. Why Do We Do Cost Estimating?

There are three broad decision-making processes in which cost estimating plays a central role: long-term planning, budgeting and choosing among alternatives.

- *Long-term planning*: Long-term planning is part of strategic planning. Cost estimating fills the critical role of providing affordability analyses. It is true in all organizations – both government and nongovernment – that strategic changes are made only over the course of multiple years and it is necessary to know whether the costs associated with the change are "affordable." (Note: there are many ways to define affordable!) It is the cost estimating community that provides these initial cost estimates, and then it is others who decide whether these estimates can be "afforded." Nevertheless, it is the cost estimating profession that provides the estimates of the resources necessary to embark upon and pursue these strategic changes.

- *Budgeting*: As an intrinsic part of building and refining budgets, cost estimating supports a series of activities that are aligned with the budgeting process. These activities include developing initial cost estimates for budget preparation, justifying cost estimates, and amending the estimates in the light of changing/changed circumstances.

- *Choosing among alternatives*: In support of decision-makers who must explore options and choose among alternatives, cost estimating supports the process by providing cost estimates and comparisons among the costs of alternative options for achieving a particular goal. It is applied to choosing among options in many walks of life:

  - Options among goods: should I rent or buy?

  - Options among systems: which database system should I use for my organization?

  - Options among processes: should a contracting action be sole-sourced or subject to open competition?

The process of sorting and choosing among options, including analysis of their comparative costs is often called a cost-benefit analysis, an economic analysis, or a business case analysis.

An example of the use of cost estimating to support decisions among alternatives is the requirement to develop and analyze cost estimates on whether a function that is performed by a particular governmental agency should be maintained within the government, or should it be outsourced to a commercial (or non-governmental) provider. Other examples include:

- The use of the tools and thought processes of cost estimating and cost-benefit analysis to provide analytical structure to DoD's decision to move to (or away from) an All-Volunteer Force.

- The US Army's decisions to outsource base operations support at various posts, camps, and installations. These services would include building maintenance, grounds maintenance, and/or fire department support.

- The Air Force's decision to buy or lease the next-generation Air Force Tanker aircraft.

- The Navy's decision to outsource parts of its information technology (IT) infrastructure.

At the core of each of these analyses were cost estimates that enabled quantitative analysis and comparisons of alternative ways to provide the services, including whether it should be provided organically by the government or procured from a commercial source. Although the comparison of options may be full of complexities and subtleties, it is critical that the analytical procedures followed adhere to best practices within the professional cost estimating community, so that the conclusions and recommendations of the analyses can be understood by those who read them, have the virtues of credibility and consistency, and ultimately, they can be used with confidence to support public policy decisions. An

example of the procedures that meets this high standard for developing and these analyses can be found in the Federal Activities Inventory Reform (FAIR) Act of 1998 [4]. This act provides processes for identifying those functions of the Federal Government that are and are not inherently governmental functions. Only those functions that are deemed not inherently governmental may be analyzed for potential outsourcing to commercial sources.

2.4.1 IMPORTANCE OF COST ESTIMATING TO CONGRESS

In addition to the internal functional uses of cost estimating and the best practices reasons for their development and use, there are statutes within the US legislative codes that require the development of Life Cycle Cost Estimates (LCCE) at major milestones within DoD's acquisition process, as well as the requirement to develop Independent Cost Estimates (ICE). These requirements can be found in:

- *USC Title 10, Section 2432:* This section requires the Secretary of Defense to report a full life cycle cost for each Major Defense Acquisition Program (MDAP). The importance of this statutory section is that it mandates in law the necessity to do full life cycle costs for MDAPs, which are DoD programs and projects that are sufficiently large, as determined by estimated cost. You may see the phrase "whole life costing," which is often used in the United Kingdom. For decision-making purposes (at least for MDAPs), it is insufficient to project costs from anything less than the cost of the full life cycle – that is, R&D costs plus procurement costs plus operating and support costs plus disposal costs.

- *USC Title 10, Section 2434:* This section states that the Secretary of Defense may not approve System Development and Demonstration (SDD), or the production and deployment of a Major Defense Acquisition Program unless an independent estimate of the full life-cycle of the program … [has] been considered by the Secretary. Moreover, this section of Title 10 defines the requirement of an ICE as follows: an "*independent estimate … [shall] – (a) be prepared by an office or other entity that is not under … the military department … directly responsible for … [developing or acquiring] the program; and (b) include all costs . . . . without regard to funding source or management control.*" This statutory section introduces the important concept of risk management in the development processes of MDAPs, namely the idea of a second and independent life cycle cost estimate (ICE), and it specifies organizational characteristics required by the organization eligible to develop the ICE. Some program managers are not enthusiastic about having an outside team coming in to do an ICE, because it can easily be misunderstood as an external intrusion and "second-guessing" the work of the program manager's internal team. However, senior managers who are responsible for shepherding portfolios of projects welcome ICEs and find them very useful, especially in support of their risk management processes. An ICE should assure an unbiased estimate and a way to check the assumptions of the original cost estimate. As the GAO Cost Assessment Guide states:

  "*The ability to generate reliable cost estimates is … necessary to support [Office of Management and Budgets] capital programming process. Without this ability, there is a risk of experiencing cost overruns, missed deadlines, and performance shortfalls – all recurring problems that our program assessments too often reveal. Furthermore, cost increases often mean that the government cannot fund as many programs as intended or deliver them when promised.*" [5]

Congress has emphasized the importance of high-quality cost estimates as part of the requirement to develop cost-benefit analyses in support of decision-making, as mandated by the Weapons Systems Acquisition Reform Act (WSARA of 2009, *Public Law 111-23, 22 May 2009)*. One measure of the increased emphasis on the role of cost estimating in this Act is that it created a new position: the Director of Cost Assessment and Program Evaluation (CAPE), which requires Senate confirmation. There are many functions tasked to this new office; the vast majority of them are related to the improved development of cost estimates and their uses in making decisions within the DoD.

Moreover, within the DoD, there are two important instructions to establish the priority and structure of cost estimating as a part of all major acquisitions, budgetary estimates and decisions with financial implications:

- DoD Directive 5000.4M requires that the Office of the Secretary of Defense, Cost Analysis Improvement Group (OSD CAIG) be chaired by the Deputy Director, Resource Analysis, in the Office of the Director of the CAPE.

- DoDI 5000.02 identifies the CAPE as the agency that fulfills the requirements of the statute to prepare an independent life-cycle cost estimate. This same instruction also requires the CAPE to prepare an ICE report for all defined milestone reviews after the program has gone through the technology development phase, also known as Milestone A or the program initiation phase.

Within the Army, Navy, Air Force, and Marine Corps, regulations and instructions provide detailed and up-to-date implementation of the statutes. Other federal agencies that have formalized their internal cost estimating functions (such as NASA, FAA, the Department of Energy, the Department of Commerce, the Department of Homeland Security, and the Intelligence Community) also support and extend the statutory requirements with the implementation of regulatory guidance.

There are also good governance and best practices reasons to develop sound cost estimates. We have already mentioned some, such as ensuring affordability within the budgeting process, and providing decision-makers with unbiased, analytically based ramifications of the projects that they are choosing and the budgets that they are developing. This is not only for internal management purposes, but also for presentation to the administration and/or Congress. Within government, as well as to CEOs and Boards of Directors outside government, this same reasoning manifests itself in the cost estimates that support the analyses that the Congressional Budget Office (CBO) does in quantifying the cost consequences of proposed and current congressional legislation. These include the budgetary impact of allocations to agencies of government, and trade-offs in allocations for acquisitions, maintenance, defense initiatives, social initiatives, and so on. It should be noted that the different allocations that may occur from these processes are often more about politics and scale than about the substance of the analytical procedures.

2.5 An Overview of the DoD Acquisition Process

In 1971, the Deputy Secretary of Defense David Packard signed the first DoD 5000.1 Series regulations concerning how the DoD acquisition process should operate. (Interestingly enough, this was the same Packard, along with William Hewlett, who co-founded

the Hewlett–Packard IT company). So given these regulations, how and when are cost estimates used? We shall start this explanation with an introduction to the DoD acquisition process, which involves a program from its very beginning until its complete end, and is called the program's "life cycle." The life cycle of a product can be compared to that of "the cradle to grave" human timeframe. As an overview, it covers the time from the birth of an idea for a product, system or process; continues through its design and specifications; then fabricates and tests prototypes, followed by full-scale manufacturing; follows it through its operating life and maintenance; and, finally, ends at its disposal. Life Cycle Cost Estimates (LCCEs) provide a cost estimate for the totality of the resources that will be necessary throughout the product's life cycle. It is important to stress that cost estimating and analysis, as applied to financial management throughout government and industry, provides for the structured collection, analysis and presentation of LCCEs to assist in decision making *based on a product's full life cycle.* The LCCE concept is relevant whether we are talking about an automated information system, an organizational restructuring, a weapon system or any other endeavor that will require resources.

The life cycle of a product, such as a new vehicle, starts when an automobile company does Research and Development (R&D), chooses its design, builds and tests prototypes, and then produces and markets that vehicle. This phase is followed by the operations and maintenance phase, which the owner of the vehicle has to fund and provide throughout the life of the vehicle. Life cycle costs are usually structured into four phases: R&D, Investment (also called Production, Procurement, or Acquisition), Operating and Support (O&S), and Disposal. These phases are depicted in the life cycle cost model in Figure 2.1:
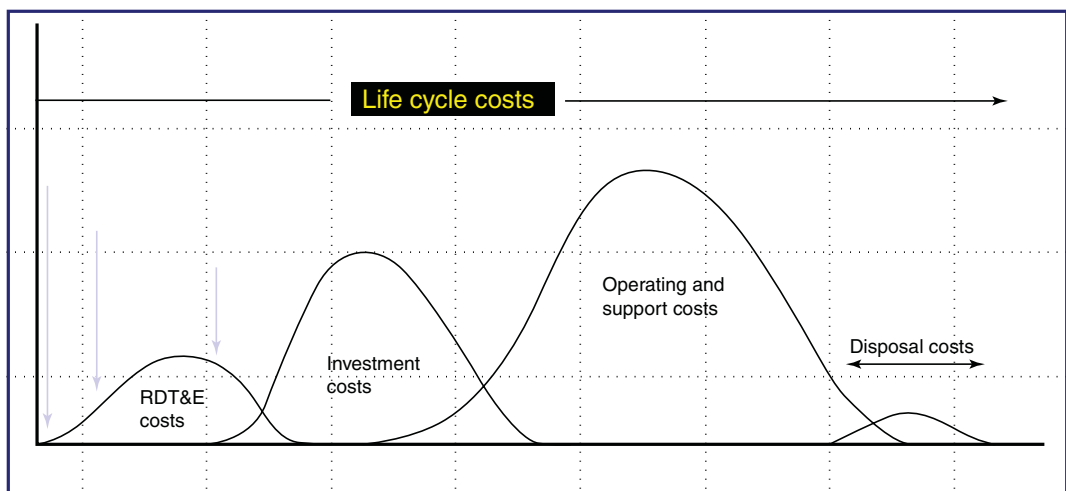
FIGURE 2.1 The Four Phases of the Life Cycle Cost Model.

Let's discuss each phase, as well as discussing the kinds of activities associated within each phase:

1. *Research and Development (R&D):* This phase includes those program costs primarily associated with research and development initiatives, including the development of a new or improved component or capability, to the point where it is ready for operational use. R&D costs include the equipment costs necessary for, and funded under, Research, Development, Test, and Evaluation (RDT&E) appropriations, as well as related Military Construction (MILCON) appropriation costs. They exclude costs that appear in the Military Personnel, Operations and Maintenance, and Procurement appropriations. The formal name for R&D is RDT&E costs.

2. *Production/Investment*: These are the estimated costs of the investment phase, including total cost of procuring the prime equipment; related support equipment; training, both individual and unit; initial and war reserve spares; preplanned product improvements and military construction. The studies associated with this phase include performance, scale, and process studies.

3. *Operations and Support (O&S)*: These are the estimated costs of operating and supporting the fielded product or system, including all direct and indirect costs incurred in using the system. This includes personnel, maintenance (both unit and depot), and sustaining investment (replenishment spares). In most programs, and especially those that consume fuel and need spare parts, the bulk of life-cycle costs occur in the O&S category.

4. *Disposal*. These are the estimated costs to dispose of the system after its useful life. This may include demilitarization, detoxification, long-term waste storage, and environmental restoration and related costs. A product or system may be retired or disposed of prior to the end of its useful life, in which case it may be sold, with the proceeds returned to the Treasury. Alternatively, components and materials of the product or system may have residual value that can be captured to offset disposal costs. While we list Disposal as a fourth phase here, the costs of Disposal are sometimes included in the O&S cost phase. Regardless of which phase you categorize them in, the important thing is to ensure that you account for them!

Additional information and definitions in these areas can be found on the International Cost Estimating Analysis Association (ICEAA) website. [1]

While Figure 2.1 shows the four phases pictorially, it actually shows much more than that. Other "take-aways" include:

- There is a time phasing in the life cycle. An indication of this, for example, is that the investment costs occur before the operating and support costs, but the investment phase is not completely finished prior to commencing the O&S phase. For example, a number of units will have been produced and sent out to the fleet for use, at the same time that others continue to be produced.

- There is a relative sizing of the total costs within each phase. For example, total R&D costs tend to be less than total procurement costs, which, in turn, tend to be less than total operating and support costs. It turns out that this general statement is true for many projects in the Department of Defense, but especially for those projects that need manpower to operate them, fuel to power them, and spare parts to keep them in good repair. Products like a vehicle or an aircraft are very good examples. On the other hand, a missile does not get refueled (as a vehicle or an aircraft does), nor does it incur regular repair costs, so the procurement phase might incur more costs than the O&S phase in a missile program. Another example is software, which incurs major development costs, but then very little production costs and relatively low (but not zero) operations and maintenance costs. In some parts of high technology, R&D may be the most expensive of the life cycle phases if the product itself is relatively inexpensive to produce. Therefore, for some projects, the relative sizing argument may be useful, and for other areas, like software – which will be covered in Chapter 15 on Software Cost Estimating – it is less useful.

- While the previous paragraphs describe the phases of an LCCE, cost estimates are developed, updated and used (or at least they *should* be) in almost every decision-making process through which a product passes during its life cycle. In

Section 2.4, we discussed the three broad decision-making processes in which cost estimating plays a central role: long-term planning, budgets, and choosing among alternatives.

Within the context of the final two of these major processes, cost estimates are developed as part of the analysis section of the following applications:

- Budget development, justification, and execution
- Planning and program development and justification
- Science and technology portfolio formation and development
- Selection of equipment through "Analyses of Alternatives (AOAs)." AoAs are similar to a cost-benefit analysis, and are very common in the weapon systems analysis arena
- Efficient management through equipment life cycle

These applications may demand different types of cost estimates, generated with the different levels of detail that are available at the relevant stage in the project's life cycle.

Many people think that the question of "What will it cost?" is a straightforward and self-evident question. Sadly this is not the case; in fact, the question is laden with ambiguity, requiring answers to many questions concerning the *capabilities* and the *acquisition environment* of "it" before attempting to quantify the *cost* of "it." The answer to "What will it cost?" is also inextricably linked to "For what purpose will this estimate be used?" For example, if the cost analyst needs input for developing next year's one-year budget, that requires a different answer – or level of analysis – than if the estimate is to prepare for a weapon systems milestone review, which, by law, requires a detailed LCCE that spans many years. Furthermore, within the context of an LCCE, one needs to know whether the cost estimate is to include particular product improvements over the proposed life cycle, and whether these improvements are "preplanned" product improvements or not. The answers to these questions define and significantly impact the scope, the complexity, the difficulty, and the results of the cost estimate.

To state this concept more succinctly, context counts, as in the story in which a person asks another "What time is it?," and the second person proceeds to explain how to build a clock! This begs the question of whether you just needed to know what time of day it was, or did you actually need to know how the clock was built?!

Within the Department of Defense, a critical application of life cycle cost estimates is in the acquisition process, and in particular, for the "milestones" within the acquisition process. For MDAPs as well as for the many not-so-major programs, the life cycle trajectory (from conceptualization to technology development to prototyping and testing to production), is punctuated by a process of formalized reviews at each stage called "milestones." It is important to know a few things about this process. The purpose of these periodic milestone reviews is to provide risk assessment and risk management and mitigation tools within the life cycle of these very complex systems. Each of these milestone reviews seek to understand the following about the project being reviewed:

- What is the current status of the program with regard to the classic program management dimensions of cost and schedule and performance?
- What are the goals the program seeks to accomplish in the next phase? That is, what is expected to be achieved in the next phase of the program, and what standards or

metrics or tests will be used to determine whether the program has achieved these objectives? These metrics are often stated as "What are the exit criteria to this milestone review?" with the implication that failure to meet defined standards will result in termination of the program.

- What are the risks that the program may encounter as it enters the next phase, and what are the strategies proposed to mitigate these risks? Again the evaluation categories of this analysis are taken from the classic program management dimensions of cost and schedule and performance.

The decision maker at these milestone reviews is called the Milestone Decision Authority (MDA). From 1971 to 1991, the MDA for all Major Defense Acquisition Programs (MDAPs) was the Secretary of Defense. Since 1991, that position is held by the Under Secretary of Defense for Acquisition, Technology and Logistics, or USD (AT&L).

Figure 2.2, taken from the Defense Acquisition University (DAU) website, shows the placement of the milestone reviews (listed as A, B, and C) and how they are embedded within the larger structure of the DoD 5000 Series acquisition process.
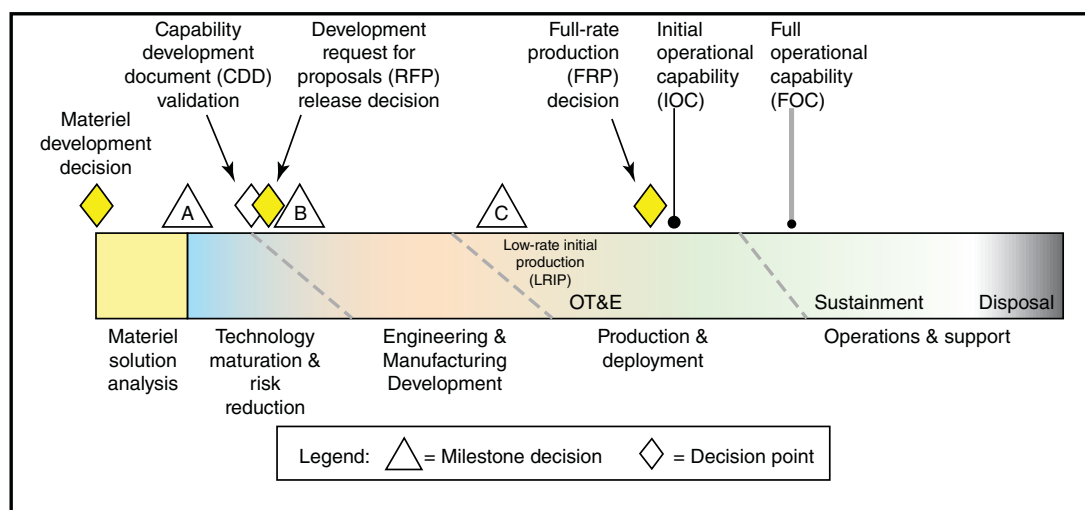
FIGURE 2.2 The Defense Acquisition Management System.

The names of the phases in Figure 2.2, as well as the number of milestones and their placements within this framework have continually changed over the years as revisions occur, but the purpose of the overall process – that of being a risk mitigation tool – has always remained the same. Much more information on this topic can be obtained from the Defense Acquisition University website [6].

The following sub-paragraphs provide an overview of what a program should have accomplished prior to the respective milestone. For readers who have worked in the acquisition world for some time and are already familiar with this information and structure, you know that the system undergoes regular revisions. For newer analysts who want more information, there are numerous books available that cover this topic in significant detail. For all readers, we are just trying in this section to provide an overview for you. Use Figure 2.2 as a visual guide to the following explanations.

- *Milestone A*: Concept and Technology Development: The purpose of this phase is to examine alternative concepts, including cooperative opportunities and procurements or modification of Allied systems or equipment, to meet the stated mission need. This milestone ends with a selection of the new system architecture. The MDA will consider all possible technology issues and identify possible alternatives before making the Milestone A decision.

- *Milestone B*: System Development and Demonstration: The purpose of this phase is to develop a system, reduce program risk, ensure operational supportability, design for producibility*, ensure affordability* (the cost estimator's part), and demonstrate system integration, interoperability, and utility. All subsystems will be integrated into a complete system. Once the integration has been completed, there will be a demonstration for the first flight, the first shot, the first drive, or the first data flow across systems to show interoperability, etc. Overall, this phase is intended to integrate the subsystems and to reduce system-level risk.

- *Milestone C*: Production and Deployment: The purpose of this phase is to achieve an operational capability that satisfies mission needs. The system must be demonstrated for technological and software maturity, and demonstrate no significant manufacturing risks. It must also be demonstrated that the system is affordable throughout the life cycle, that it is optimally funded, and that it is properly phased for rapid acquisition. At Milestone C, authorization to enter into low rate initial production (LRIP) for MDAP's and major systems can occur, as well as authorization into production or procurement for nonmajor systems that do not require LRIP. This phase will also include an independent cost estimate, an economic analysis, a manpower estimate, and an acquisition strategy.

Navigating through these three milestones can take a significant number of years to accomplish. One possible way to shorten a production process is not to develop the new item as a new production program. Perhaps there is an item that is already available commercially, or a previously developed item that is in use by the government or by one of our allies, that can be modified to meet your needs. Items like these are known as Commercial Off-the-Shelf (COTS) or Non-Developmental Items (NDI). Potential benefits of COTS/NDI items include:

- A lower life cycle cost
- More rapid deployment
- A capability already proven
- A broader industrial base
- Possible access to state-of-the-art technology

A simple example to illustrate an NDI item would be if you had a cell phone that needed to also have a Global Positioning System (GPS) (disregard the fact that there are many out there already!) If you started to develop a new combined phone and GPS system, it would take a significant amount of time and money to do so. But if you had a phone already in existence on the open market, and were able to modify it by adding a GPS to it, then theoretically it would cost much less to produce and it would be completed in a much shorter period of time. But if that option is not available, then you will need to complete the development through the normal DoD acquisition process.

2.6 Acquisition Categories (ACATs)

A technology project or acquisition program is required to be validated through the milestone review process when it surpasses certain monetary thresholds. Table 2.1 summarizes each of the acquisition categories, the monetary thresholds required at each ACAT, and who the Milestone Decision Authority is at each threshold. Similar to the Defense Acquisition Management System and the milestones, these authorities and the monetary thresholds for each acquisition category are revised periodically. As can be seen, the largest and most expensive programs are ACAT I programs, followed by ACAT II, then ACAT III, etc. Notes and clarifications concerning Table 2.1 thresholds and Authorities are provided below the table.

TABLE 2.1   Summary of Acquisition Categories and Monetary Thresholds for Each

Acquisition Category (ACAT)	Criteria (FY2014$)	Milestone Decision Authority (MDA)
ID MDAP	> $480M RDT&E > $2.79B Procurement or so designated by MDA	Defense Acquisition Executive (DAE)
IC MDAP	Same as ID MDAP thresholds	Head of DoD component or the Component Acquisition Executive (CAE)
IAM MAIS	> $520M Lifecycle Costs > $165M Procurement > $40M Single Year	DAE or as delegated
IAC MAIS	Same as IAM MAIS thresholds	Head of DoD Component or the CAE
ACAT II	> $185M RDT&E > $835M Procurement or so designated by the MDA	CAE or as delegated
ACAT III	Less than ACAT II thresholds	As designated by the CAE
Other	As needed by each service	As designated by the CAE

Notes from Table 2.1 include:

- The largest acquisition category is an ACAT ID program, which is also known as a Major Defense Acquisition Program (MDAP). The "D" in ACAT ID stands for the Defense Acquisition Board (DAB), the Board that oversees these programs. The criteria for inclusion in this category are that the RDT&E costs must exceed $480M, or procurement costs must exceed $2.79B, both in FY14$. An ACAT ID program will sometimes involve a number of the uniformed services, such as with the F-35 Joint Strike Fighter. A program can also be designated an ACAT I program by the Defense Acquisition Executive (DAE) even if it does not reach these monetary thresholds if it is deemed important enough or "high visibility" enough to be categorized as such.

- As discussed in Section 2.5, the MDA for an ACAT I program is the person who conducts the milestone reviews, and is that person who gives the final "thumbs up

or thumbs down" as to whether your program can continue to the next phase and to the next milestone. This decision is critical to the determination of whether to continue to fund the program or to cancel it. The Decision Authority is the person who designates who the MDA is for a particular program. As the estimated cost threshold of a program decreases from ACAT I to ACAT II to ACAT III, so does the perceived importance of the program, and therefore the person designated as the MDA will begin to decrease slightly in rank.

- An ACAT I "C" program is a program for which the milestone decision has been delegated to one of the services, or in DoD vernacular, to a "Component." An ACAT IC program has the identical monetary thresholds as an ACAT ID program, but the designated Service has been delegated the authority to make the milestone decision. Examples of a single service program include the Army's Guided Multiple Launch Rocket System (MLRS), the Navy's Tactical Tomahawk, and the Air Force's Joint Direct Attack Munition (JDAM).

- A MAIS is a Major Automated Information System. What is a MAIS? Think of computer and IT-related systems. An ACAT I MAIS has its own monetary thresholds that differ from a standard ACAT I program. If an IT program is less than the thresholds for a MAIS, then it is just called an Automated Information System (AIS).

- In the Milestone Decision Authority column, you will see the term Component Acquisition Executive (CAE). The CAE is synonymous with the Service Acquisition Executive (SAE). They are responsible for all acquisition functions within their service, or "Component." For the three services, these positions are held by the:

    - Assistant Secretary of the Army for Acquisition, Logistics and Technology (ASA (AL&T))

    - Assistant Secretary of the Navy for Research, Development and Acquisition (ASN (RD&A))

    - Assistant Secretary of the Air Force for Acquisition (SAF/AQ)

- An ACAT II program has the minimum thresholds as shown, with its ceilings being those for an ACAT I program. Thus, if the totals reach $480M in RDT&E, or $2.79B in Procurement, it would then become an ACAT I program. The MDA for an ACAT II program in the Navy is the Assistant Secretary of the Navy for Research, Development and Acquisition, ASN (RD&A), or the respective position in the other uniformed services.

- The thresholds and designations of an ACAT III program (and lower) are used as needed and required. They will differ from service to service. The MDA's for these programs include the Commander of the service's Systems Command (SYSCOM); the Program Executive Officer (PEO); and the Direct Reporting Program Manager (DRPM).

Having discussed the DoD acquisition process and the acquisition categories involved in that process, we will next discuss some of the primary terminologies used in the cost estimating career field.

2.7 Cost Estimating Terminology

Every profession has its own unique language, with numerous terms known to those who work within that profession. Understanding these terms is a necessary requirement, as

they can provide distinctions among ideas and concepts that need distinguishing for practitioners. This is important so that their analysis will be contextually accurate, technically sound, and maximally useful.

A complete list of DoD terminologies and acronyms can be found at the DAU website and is updated periodically. As of the time of this writing, the latest edition was "Glossary of Defense Acquisition Acronyms and Terms", 15th edition, published in December 2012. Check for the latest edition possible. This glossary concentrates on numerous terms used within the Department of Defense and can prove very useful to you; however, we will attempt here to define those terms that are most important in the cost estimation career field.

First we would like to discuss the differences between Cost and Price. The simple phrase "What does it cost?" appears very clear, explanatory, and unambiguous. There are, however, two major ambiguities here in just these four words, as discussed in the following two paragraphs:

- The first ambiguity is the *distinction* between cost and price. *Cost* is a quantitative measurement of the resources needed to produce an item; or rephrased, it is the amount of money needed to actually produce that item. *Price*, on the other hand, is what you and I must pay for that item in the marketplace. There are two primary differences between cost and price. First, there is profit. Profit can be either negative or positive. A negative profit results in a loss to the seller, while a positive profit will result in a gain to the seller. Second, the market's assessment of the value of a product that determines its ultimate price in the marketplace may have little to do with the cost of making that product. While both profit and market value will also influence costs (e.g., for basic materials and labor), they are less determinative of the final price at which the item is sold. A key fundamental in the profession of cost estimating is that cost estimating focuses on costs, not price! The previous sentence is so important and so fundamental to the profession of cost estimating that it bears repeating:

    "*The profession of cost estimating focuses on costs, not price.*"

- The second ambiguity is equally substantive. If you walk into a grocery store and ask "How much does that head of lettuce cost?," the employee to whom you address the question understands the meaning as "What do I have to pay, in dollars and cents, to buy this head of lettuce and walk out of the store with it?" In turn, you understand the dollars and cents answer when it is given to you. However, when used within the profession of cost estimating, the same phrase of "How much does that head of lettuce cost?" has a number of different meanings. Is it the cost to plant the lettuce? Is it the cost to irrigate it and tend to it while it is in the field? Does it involve the cost of harvesting that head of lettuce, and then taking it to the processing plant to be packaged? Or the cost of the vehicles involved that transport it to the store? These are questions that the cost estimator must consider when determining "How much does that head of lettuce cost?"

While answering this question would not be an easy answer, an even more involved question would be "What does an F/A-18 cost?" One has to address the numerous ambiguities in this question before one can even begin to answer this question. For example, does the question mean the cost of the very first aircraft that was produced, or the cost of the 50th one? Or, perhaps one wants to know the unit cost based on the average of all of the aircraft that have been produced so far? Or maybe it is referring to the cost of maintenance and fuel for an hour of flight time or a mission? Additionally, does the computation

require us to include the research and development costs in the estimate, and do we want the response to be in today's dollars or in some future budget year dollars? These questions represent a small subset of the questions that might be asked in an effort to understand our "simple" question, and we are led to the conclusion that "cost estimating ain't about the price of lettuce!" More properly, we can see that we need to know the language and terminology of the cost estimating career field and the concepts that they represent. We need to distinguish, as a minimum, among different notions and categorizations of costs, because when we mention cost, we realize that *cost is not a unique term* – there are many types of costs out there! Some examples of different categories of costs that we must consider are as follows:

- Recurring vs. Nonrecurring Costs
- Direct Costs vs. Indirect Costs
- Fixed Costs vs. Variable Costs
- Overhead Costs
- Sunk Costs
- Opportunity Costs
- Life Cycle Costs

We will attempt to define each of these terms and give a numerical example of this type of cost where possible.

- *Recurring Costs* are costs that are repetitive and occur *each time a company produces a unit*. This is in contrast to the business world or in your personal life, in which a fixed cost such as mortgage or rent that is paid on a monthly basis would be considered a recurring cost. This, however, is different in the world of acquisition and production. In this environment, a recurring cost is a cost that occurs every time that a unit is produced. If you make 1, 10, 100, or 1000 units, a recurring cost would occur *each time a unit is produced*. This would include any cost that happens because of the activities on a production line, such as installing the drive train into a vehicle or configuring the avionics package on an aircraft. That cost will occur for every item that is produced. These are predominantly costs that are incurred in material, labor, and machinery usage.

- *Nonrecurring Costs* are those costs that are *not* repetitive and cannot be tied (nor are they proportional to) the quantity of the items being produced. Nonrecurring costs generally include start up costs, such as developing or establishing your capacity to operate, and may also include the purchasing of a specialized machine for a production line in a plant. Other nonrecurring costs can include test and evaluation, tooling, and costs incurred in the design phase, such as design engineering costs. A statistician might say that nonrecurring costs are those that are not correlated with the quantity of items that are produced.

- *Direct Costs* are those costs that can be reasonably measured and allocated to a specific output, product or work activity. Typical direct costs include the labor and material costs directly associated with a product, a service, or a construction activity. They also include the labor involved to attach the landing gear, the costs to prepare blueprints of a design, or the labor costs for the time it takes to conduct Quality Assurance (QA) inspections.

- *Indirect Costs* are those costs that *cannot* be attributed or allocated to a specific output, product, or work activity. Typical indirect costs include the costs of common tools, general supplies, electricity, and equipment maintenance that are needed and shared with other projects. Since these costs are shared with other projects, the indirect costs can be difficult to attribute (or involve too much effort) to allocate directly to a specific output. So how should they be allocated among the shared projects? Accounting rules are permissive in these allocations, and therefore no single, universally accepted rules about how to do this are available to the cost estimator. Probably, the most widely used procedure is to allocate the indirect costs as a proportion or percentage of a "basis," such as direct labor hours or direct material dollars in a project. We will provide an example of this in the upcoming definition on Overhead Costs.

- *Fixed Costs* are costs that are unaffected by the output quantity of 1, 10, 100 or 1000 units. Typical fixed costs include insurance and taxes on facilities, general management and administrative salaries, monthly mortgage and/or rent, depreciation on capital, and interest costs on borrowed capital and loans. The key point is that these costs do not vary with the number of units produced. Fixed Costs are generally associated with nonrecurring costs.

- *Variable Costs* are those costs that *are* associated with production and that do vary with the quantity of units produced. Typical variable costs include material and labor, as those are the two areas most affected by a change in output and quantity. Variable costs are generally associated with recurring costs. They are also the primary costs that should be considered when analyzing the economic impact of a proposed change to an existing operation, such as a change in quantity to be produced, or the annual rate of production.

- *Overhead Costs* consist of plant operating costs that are not direct labor or direct material costs. (Note: Indirect costs, overhead, and "burden" are terms that are sometimes used interchangeably). Typical overhead costs include electricity, general repairs, property taxes, and supervision. Various methods are used to allocate overhead costs among products, services, or activities. Indeed, the profession of cost accounting is most closely associated with the activities of recording, allocating, and managing this type of data. The most commonly used methods involve an allocation of overhead costs that are in proportion to a basis, such as direct labor costs, direct labor hours, direct materials costs, the sum of direct labor and material costs, or machine hours. Consider the following example:

Overhead Costs Example: Suppose that your production plant is producing the same item in varying quantities for three different companies: Company A, B, and C. Table 2.2 shows the totals for direct labor and material costs at your production plant for those three contracts, listed in FY13$. The total cost of the three contracts was $120M:

TABLE 2.2 Direct Labor and Material Costs for the Overhead Costs Example

Contract #	Direct Labor and Material Costs (FY13$)
A	$20M
B	$60M
C	$40M
Total	$120M

At the end of FY13, the following overhead costs at your production plant were calculated, as shown in Table 2.3:

TABLE 2.3   Overhead Costs Accrued During FY13

Overhead Category	Overhead Costs (FY13$)
Taxes	$16M
Utilities	$2.5M
Repairs	$6M
Admin support	$3.5M
Total	$28M

As can be seen, a total of $28M in overhead costs occurred during FY13, and these costs must be allocated fairly to each of the three contracts. To do so, we will allocate them proportionally to the direct labor and material costs of each contract. From Table 2.2, we can observe that the total amount of the three contracts is $120M.

Calculations for the percentage of overhead costs to each contract are as follows:

- Contract A: $20M / $120M = 16.67% of the total of the three contracts
- Contract B: $60M / $120M = 50% of the total of the three contracts
- Contract C: $40M / $120M = 33.33% of the total of the three contracts

Since the total overhead cost was $28M, we must now take the proper percentage of overhead costs and allocate them to each program. Doing so reveals the following:

- Contract A: 16.67% × $28M = $4.667 M of overhead costs are allocated to Contract A

- Contract B: 50% × $28M = $14M worth of overhead costs are allocated to Contract B

- Contract C: 33.33% × $28M = $9.333 M of overhead costs are allocated to Contract C

TABLE 2.4   Final Allocation of Overhead Costs to Each Contract

Contract #	Direct Labor and Material Costs (FY13$)	Percentage of Total Cost of Contracts (FY13$)	Allocation of Overhead Cost (FY13$)
A	$20M	0.1667	$4.667M
B	$60M	0.5000	$14.000M
C	$40M	0.3333	$9.333M
Total	$120M		

Table 2.4 shows these calculations in a combined table. As can be seen, the given overhead costs were fairly allocated to each contract in proportion to the amount of work being done during the fiscal year. An unfair way to allocate these costs would have been to just assign 33.33% to each of the contracts. In that case, Contract A would be paying an unfair share, while Contract B would be getting a great deal and paying much less than it should!

- *Sunk Costs* are those costs that have occurred in the past and have no relevance to estimates of future costs and revenues for alternative courses of action. Since sunk costs are not retrievable, nor can they be gathered by the decision maker and redirected to another purpose, they are not part of prospective cash flows in the future. An interesting mis-employment of sunk costs is that they are sometimes used in decisions on whether to cancel a program that may be poorly run and facing significant cost overruns. For example, a decision maker might say "We can't cancel that program. We have already spent $50M on it!" The $50M is considered a sunk cost at this point, since it has already been spent. But if the program is being poorly managed, then sometimes taking the $50M loss and cancelling the program now is the more prudent decision, rather than having losses continue to mount in the future. Numerous examples of cancelled programs are available for review, and the total spent on each is considered a sunk cost. Another use of sunk costs can be found in the following example.

  Sunk Costs Example: This example examines whether remaining with the "status quo" program or moving to Option A is the more desirable option. Consider the following given information:

	Sunk Cost (M$)	To Go (M$)	Total (M$)
Status Quo	200	100	300
Option A	0	250	250

  Overall, we can observe that the status quo program is the more expensive on a total basis of the two options: $300M to $250M. So is it better to move to the less expensive option, Option A? The answer depends on whether we include or ignore the sunk costs.

  Including sunk costs: It must be observed that $200M has already been spent on the status quo program, and this amount represents our sunk cost. No money has yet been spent on Option A. Only $100M is remaining to be spent on the status quo program, while Option A would require an additional $250M outlay. But if we include the sunk costs in our decision and just seek the lowest total, we would still choose Option A, $250M to $300M, since that is overall the least expensive of the two options.

  Ignoring sunk costs: If we ignore the sunk costs, we minimize our future expenditures by choosing the status quo program, since we only owe $100M more, rather than the $250M we would owe if Option A were chosen.

  Conclusion: If the status quo program and Option A are equally effective, then the correct choice in this example is to remain with the status quo. The general principle in financial decision-making is to ignore sunk costs.

- *Opportunity Costs* occur when there are a few options that need to be considered. Each of these options has a cost or a "reward." The opportunity cost is a measure of the lost value when you do *not* choose what turns out to be the optimal solution or alternative. Since the best alternative was not selected, the opportunity to use that best alternative was foregone or rejected. In many cases, these opportunity costs are often hidden or implied. This point will be illustrated in the following example:

  Opportunity Costs Example: This financial scenario provides an excellent example of opportunity costs. Let's assume that you had $50,000 to invest, and your two choices were to invest the money into (1) stocks and mutual funds, or (2) real estate. After three years, the following results were calculated:

- The stocks and mutual funds investment you were considering would have risen to a value of $80,000, providing a $30,000 increase in value.

- The real estate investment you were considering would have risen to a value of $120,000, providing a $70,000 increase in value.

Clearly, had you invested in the real estate, you would have made $70,000 and that would have been your best investment. In this case, your opportunity cost would have been $0, since you had invested in the best/optimal choice. However, if you had invested in the stocks and mutual funds instead, you would have only made $30,000 instead of $70,000, so your opportunity cost is $40,000, which represents the difference between the optimal/best solution and the solution that you chose. Thus your opportunity cost of $40,000 represents the amount that you "lost in opportunity" due to not choosing the optimal alternative.

In the world of the Department of Defense, one major use of opportunity costs is in Base Realignment and Closures (BRACs), during which analysts compare and contrast a number of different alternatives that consider whether certain military bases should be kept open or should be closed, and decide which alternatives are the most cost effective in the long run. Deciding whether to keep Base A open, or to move the personnel from Base A to another location is partly about the direct costs associated with these changes, but also partly about analyzing the opportunity costs available among the different alternatives.

Summary

In this chapter, we discussed numerous topics while introducing the background and terminology of the career field of cost estimating. While discussing what cost estimating is, the definition included the most important aspects of the task at hand. Mainly, this includes that predicting the future cost of an item, product, program, or task is accomplished by applying quantitative models, techniques, tools, and databases in order to predict an estimate of the future cost, and that it is necessary to use historical data to underpin the quantitative models. Moreover, the idea that our prediction is based not only on data from the past but also on information available at the current time and assumptions that needed to be made are also key aspects to this definition.

A good cost estimate is anchored in historical program performance. It must also reflect current and potential future process and design improvements, be understandable by program and business leaders, and address the risks and uncertainties inherent in the program plan.

We do cost estimating in order to support three major and critical decision-making processes: (1) Long-term planning, especially in providing affordability analyses; (2) Building and refining budgets, including developing an initial cost estimate for budget preparation, justifying cost estimates, and amending the estimates in the light of changed circumstances; and (3) Supporting decision-makers who must explore options and choose among alternatives. Cost estimating plays a critical role in the Department of Defense (DoD) in all three of these areas, and the necessity of providing these analyses is underscored by statutes and regulations.

The everyday question of "What will it cost?" is neither straightforward nor self-evident, but rather one that is subject to ambiguity. Therefore, it is necessary to understand the basic terminology of cost estimating. We found that cost is not a unique

term! In fact, there are many types of cost, including recurring costs, nonrecurring costs, direct costs, indirect cost, opportunity costs, and others.

References

1. ICEAA Website Glossary. http://www.iceaaonline.com/?s=glossary
2. "GAO Cost Estimating and Assessment Guide", http://www.gao.gov/new.items/d071134sp.pdf
3. Congressional Budget Office. http://www.cbo.gov/about/overview
4. Federal Activities Inventory Reform (FAIR) Act of 1998. http://www.whitehouse.gov/omb/procurement_fairact
5. GAO Cost Estimating and Assessment Guide, http://www.gao.gov/new.items/d093sp.pdf, page3.
6. Defense Acquisition University (DAU) website: www.dau.mil/default.aspx

Applications and Questions:

2.1 (T/F) Good cost estimating practices are anchored in historical program performance.

2.2 What are the four phases in a program's life cycle cost model?

2.3 Name the three broad decision-making processes in which cost estimating plays a central role.

2.4 Why can't we be certain that the initial cost estimate that we develop is ultimately the "right answer?"

2.5 Why is the U.S. Congress involved in the cost estimating process?

2.6 Why do we do a cost-benefit analyses (CBA) when choosing among several public policy or program choices?

2.7 What is the purpose of having milestones in the acquisition process? Can't we just trust our program managers to do the right thing?

2.8 Should milestones be applicable only to acquiring weapon systems and automated information systems, or can we apply them to service contracts such as base operations, meal planning and service, water and fuel delivery as well? Why or why not?

Non-DoD Acquisition and the Cost Estimating Process

3.1 Introduction

Much of what is covered in this text has roots in the US Department of Defense's practices, procedures, and nomenclature. This is because DOD's weapon systems acquisition process was an early adopter of the formal thought processes that have come to be known as the cost estimating profession. For example, learning curves – whether the unit theory variant or the cumulative average theory variant – both arise from observations in wartime aircraft production facilities. Additionally, the early writings of people's observations about the behavior of the cost of systems came from the Department of Defense. Naturally, though, there are other environments in which cost estimating takes place. As stated in Chapter 1 " … *cost estimating is ubiquitous, almost always existing either formally or informally in every organization …* " In this section, we identify some of the organizations and other environments in which cost estimating is practiced, and we then discuss the cost estimating practices within these environments. First, we address "other-than-DoD" (or non-DoD) US government executive agencies that have formal cost estimating organizations within them, and then we identify and address aspects of the commercial world, especially those that support government functions. These organizations include government and commercial firms, and those that fall "in-between" the two, called Federally Funded Research and Development Centers (FFRDCs). After we identify these organizations, we then describe the role that cost estimating plays within the organization, the guidance under which the cost estimating organization operates, and its cost estimating practices. Once that is complete, we delve into the cost estimating process and the myriad terms that the cost estimating field encompasses.

3.2 Who Practices Cost Estimation?

The US Government Agencies which have Cost Estimating Organizations include (but is not limited to):

- The General Accountability Office (GAO)
- The Intelligence Community (IC)
- National Aeronautics and Space Administration (NASA)
- Federal Aviation Administration (FAA)
- Department of Energy (DOE)
- Department of Commerce
- Department of Homeland Security (DHS)
- US Census Bureau

The commercial firms which have cost estimating organizations include:

- The large defense and aerospace corporations, such as The Boeing Company and Lockheed Martin
- The consultancies which support the US Government, such as Science Applications International Corporation (SAIC) and Booz Allen Hamilton

The Federally Funded Research and Development Centers (FFRDC's), which include:

- The Institute for Defense Analysis (IDA)
- The Center for Naval Analysis (CNA)
- The RAND Corporation
- The MITRE Corporation

Let's describe each of these US government agencies that have cost estimating organizations.

3.3 The Government Accountability Office (GAO) and the 12-Step Process

The GAO is the audit, evaluation, and investigative arm of the US Congress. It is part of the legislative branch, and it is charged with examining matters relating to the receipt and payment of public funds. Their work often involves judging the correctness of cost estimates developed by other executive agencies of the US government. In pursuing this objective, the GAO has published the "GAO Cost Estimating and Assessment Guide" (Richey/Cha/Echard) [1]. This guide provides a 12-step process that acts as a guide for the proper development of a cost estimate, and one that will withstand GAO's scrutiny. More specifically, the 12-step process strives to achieve multiple objectives:

- To guide the analyst to create a comprehensive, accurate, credible, well-documented and well-estimated cost analysis.
- To allow overall for a more realistic cost estimate.
- To allow a program's cost estimates to be validated both internally and externally.

• To place importance on each of the 12 steps. The ultimate goal of the 12-step process is to ensure high-quality cost estimates that are delivered in time to support important decisions.

The GAO's 12-Step Estimating Process is shown in Figure 3.1, and consists of the following steps:

Intiation and research	Assessment	Analysis	Presentation
Your audience, what you are estimating, and why you are estimating it are of the utmost importance	Cost assessment steps are iterative and can be accomplished in varying order or concurrently	The confidence in the point or range of the estimate is crucial to the decision maker	Documentation and presentation make or break a cost estimating decision outcome

Analysis, presentation, and updating the estimate steps can lead to repeating previous assessment steps

FIGURE 3.1 The GAO 12-Step Estimating Process.

The 12 steps in the GAO process can be defined as follows:

1. Define the estimate's purpose
2. Develop the estimating plan
3. Define the program characteristics and the technical baseline
4. Determine the estimating structure
5. Identify the ground rules and assumptions
6. Obtain the data
7. Develop the point estimate and compare to the independent cost estimate (ICE)
8. Conduct sensitivity analysis
9. Conduct risk and uncertainty analysis
10. Document the estimate
11. Present the estimate to management for approval
12. Update the estimate to reflect actual costs and changes

Each of the steps is its own process, and we will describe each below:

Step 1: Define the estimate's purpose. This step incorporates the following:

• A thoroughly documented estimate, which should include explanations of the method, data used, and the significance of the results.

- Since there are two main functions of a cost estimate, the first goal of the cost estimate is to help managers evaluate and select alternative systems and solutions. The second goal is to support the budget process by providing estimates of the funding required to efficiently execute the program.
- Link the program's purpose to the agency's missions, goals, and strategic objectives. Make clear what the benefits that the program intends to deliver are, along with the appropriate performance measures for benchmarking progress.

Step 2: Develop the estimating plan. This step includes:

- Understanding the customer's needs. This is necessary since the estimating plan serves as an agreement between the customer and the cost estimating team.
- Matching the data availability with the estimate's ultimate use.
- Understanding that all costs concerning development, procurement, construction, and operation and support (without regard to funding source or management control) must be provided to the decision maker for consideration.

Step 3: Define the program characteristics and the technical baseline. This third step includes:

- Data being used in support of the cost estimate must be traceable to its original use.
- Determining the technical baseline definition and description, which provides a common definition of the program in a single document.
- Determining the acquisition strategy, the system characteristics, the system design features, and the technologies which are to be included in the design.

Step 4: Determine the estimating structure. This step entails performing the following:

- Defining the WBS (work breakdown structure) and describing each of its elements. This is important since the WBS is the cornerstone of every program. It describes in detail the work that is necessary in order to accomplish a program's objectives.
- Creating the structure to describe not only what needs to be done, but also how the activities are related to one another.
- Creating the structure to provide a constant framework of planning and assigning responsibility, and tracking technical accomplishments.

Step 5: Identify the ground rules and assumptions. Step 5 is important for:

- Creating a series of statements that define the conditions that the estimate is to be based on, recognizing that cost estimates are often based on limited information.
- Creating ground rules and assumptions that accomplish the following:
  - Satisfy requirements for key program decision points.
  - Answer detailed and probing questions from oversight groups.
  - Help make the estimate complete and professional.
  - Provide useful estimating data and techniques to other cost estimators.
  - Provide for later reconstruction of the estimate when the original estimators may no longer be available.

Step 6: Obtain the data. Step 6 allows the analyst to:

- Match the data collection objectives with an understanding of what needs to be estimated.
- Understand that data collection is a lengthy process, not a "one time and we are finished" process. Data collection continues throughout the development of the cost estimate.
- Collect relevant technical and cost data, and the analyst must document all steps that are used to develop the estimate.
- Collect the data through interviews, surveys, data collection instruments, and focus groups.
- Use historical data to underpin the cost estimate.
- Ensure that the data are applicable and credible by performing checks of reasonableness.

Step 7: Develop the point estimate and compare to the Independent Cost Estimate (ICE). The cost analyst must ensure that:

- Once all supporting data have been collected, normalized, checked for reasonableness and analyzed, the methodologies which will be used for developing the cost estimate are considered and selected.
- The point estimate is unbiased, neither overly conservative nor overly optimistic.
- The cost model develops a cost estimate for each WBS element.
- All estimating assumptions are included in the cost model.
- All costs are expressed in constant year dollars.
- All costs are spread into the years in which they are expected to occur.
- Once all of the aforementioned steps are complete, that the cost estimate be compared against the independent cost estimate, in order to find where, and determine why, there are/may be differences.

Step 8: Conduct sensitivity analyses. This step is necessary in order to:

- Identify what cost elements represent the greatest risks.
- Analyze how the original cost estimate is affected by changes in a single assumption by recalculating the cost estimate with different quantitative values for selected input factors. Likely candidates for these input factors include:
  - Potential requirement changes
  - Changes in performing characteristics
  - Testing requirements
  - Acquisition strategy
  - Configuration changes in hardware, software, or facilities

Step 9: Conduct risk and uncertainty analysis.

- Uncertainty analysis should be performed to capture the cumulative effect of risk factors.
- Monte Carlo simulation is the usual methodology used for risk and uncertainty analysis.

Step 10: Document the Estimate. This step is extremely important since documentation will be an essential tool when:

- Validating and defending a cost estimate.
- Reconciling differences with an independent cost estimate or understanding the bases for these differences.
- Developing future cost estimates.

Other tasks necessary in this step include:

- Identifying all data sources, including how the data were normalized.
- Striving to make the documentation sufficiently detailed and providing enough information so that someone unfamiliar with the program can readily and independently re-create the estimate.
- Remembering that a poorly documented cost estimate can cause a program's credibility to suffer.

Step 11: Present the Estimate to Management for Approval. When executing this step, the analyst must remember the following:

- Ensure that the briefing is crisp and complete and presented in a consistent format.
- Describe the methodologies used, plus the main cost drivers and the outcome.
- Include the following items in the presentation:
  - The title page, briefing date, and the name of the person being briefed.
  - A top-level outline
  - The estimate's purpose
  - The program overview
  - The ground rules and assumptions
  - The cost estimate results
  - The process used to develop the estimate
  - Sensitivity analysis
  - Risk and uncertainty analysis
  - The comparison to the independent cost estimate
- Discuss any concerns or challenges encountered
- Discuss the conclusions and recommendations

Step 12: Update the estimate to reflect actual costs and changes. This final step in the GAO estimating process is necessary since:

- Programs can (and often do) change significantly throughout the life cycle due to scope changes, risk, quantity, or schedule changes, etc.
- It is necessary to update program costs and examine differences between estimated and actual costs.
- An analyst must link cost estimating and earned value management analysis, to update program costs and to again examine differences between estimated and actual costs.

In summary, the GAO Cost Estimating Guide and its 12-step cost estimating process provide a formal, step-by-step approach for achieving a comprehensive, accurate, credible,

and well-documented and well-estimated cost analysis. It provides a structure so that a realistic cost estimate will be developed, one which allows a program's cost estimates to be validated both internally and externally. They should also ensure that high-quality cost estimates are delivered in time to support important decisions.

3.4 Cost Estimating in Other Non-DoD Agencies and Organizations

Numerous non-DoD organizations practice the art and science of cost estimating. An overview description of each follows:

3.4.1 THE INTELLIGENCE COMMUNITY (IC)

There are seventeen separate US government organizations that form the Intelligence Community. Information on all of them can be found at the Intelligence.gov website [2]. Overall, the IC is governed by Intelligence Community Directives, and the area of cost estimating is governed by Intelligence Community Directive 109 (ICD 109), which speaks to the issue of Independent Cost Estimates (ICEs) and Agency Cost Positions (ACPs). This ICD is available for review at the "Office of the Director of National Intelligence" website [3]. From page 1 of that directive, the Associate Director of National Intelligence for Systems and Resource Analyses (ADNISRA) may direct the Cost Analysis Improvement Group (CAIG) of the Office of the Director of National Intelligence (ODNI) to prepare an ICE, or an Intelligence Community (IC) element to prepare an ACP to support the development, determination, and effective execution of the National Intelligence Program (NIP) budget [4].

The section of US laws entitled 50 USC§ 415a refers to a particular part of the US laws. The "50 USC" means U.S. CODE: TITLE 50 – WAR AND NATIONAL DEFENSE, while the notation "§ 415a" means Article 415a. This section specifically requires that the IC develop, present, and use LCCE ICEs for their major systems, and that a specified organization be identified for each. For collection programs, ICEs have to include certain WBSEs that are not used on the DoD side. These WBSEs include:

- The cost of new analyst training
- The cost of new hardware and software for data exploitation and analysis
- Any unique or additional costs for data processing, storing, and power, space and cooling throughout the life cycle of the program.

ICE updates are required in the Intelligence Community for three reasons: (1) when there is a completion of any preliminary design review associated with the major system; (2) if there is any significant modification to the anticipated design of the major system; or (3) when any change in circumstances renders the current independent cost estimate for the major system inaccurate.

3.4.2 NATIONAL AERONAUTICS AND SPACE ADMINISTRATION (NASA)

NASA is the agency of the United States government that is responsible for the nation's civilian space program and for aeronautics and aerospace research, and NASA conducts its

work through four mission directorates, all of which require cost estimates. As the developer and procurer of major pieces of space equipment in support of these directorates, NASA has set up the Headquarters Cost Analysis Division (CAD) specifically to make use of the cost estimating discipline. " … (CAD) is a part of the Office of Evaluation and supports the Office of the Administrator with cost estimating policy, cost estimating capabilities, and decision support . . . .and is responsible for ensuring that NASA's estimates are continually improving and increasing in accuracy. . . . ..CAD provides cost estimates for possible future programs, … and develops cost analysis policy for the Agency, and is available to perform quick turn-around cost [analyses]" [5].

3.4.3 THE FEDERAL AVIATION ADMINISTRATION (FAA)

The FAA is the national aviation authority of the US government. Its mission is "to provide the safest, most efficient aerospace system in the world," [6], and it has the authority to oversee and regulate all aspects of American civil aviation. The FAA has three important operational roles, which include:

- Developing and operating a system of air traffic control (ATC) and navigation for both civil and military aircraft
- Researching and developing the National Airspace System and civil aeronautics
- Developing and carrying out programs to control aircraft noise and other environmental effects of civil aviation

As the developer and procurer of major pieces of the US ATC system, FAA has a formal cost estimating organization to support these objectives as part of its Investment Planning and Analysis organization. Its cost organization's major functions are to (1) support investment analysis by building databases; (2) determine costs using estimating tools and techniques, and (3) develop standard agency-wide estimating guidelines. Additionally, they have published costing information including *Guidelines for FAA Cost Estimating, Guidelines for Documenting Cost Basis of Estimate,* and the *Basis of Estimate Briefing Template.* More detailed information is available at their website [7].

Additionally, the Department of Energy (DOE), the Department of Commerce, the Department of Homeland Security (DHS), and the US Census Bureau all have a cost division within their organizations whose purpose is to estimate costs and optimize all resources available for analyses and system acquisition.

3.4.4 COMMERCIAL FIRMS

As was stated in Chapter 1 and at the beginning of this chapter " … *Cost estimating is ubiquitous, almost always existing either formally or informally in every organization …* " It follows then that it exists in the large commercial, defense, and aerospace corporations, exemplars of which are The Boeing Company, Lockheed Martin, Raytheon, General Dynamics, and Northrop Grumman. In the late 1990s and early 2000's, there was extensive roll-up and consolidation within the defense aerospace industry, and these consolidations forced these companies to begin focusing on the standardization of processes within their newly enlarged enterprises. For example, the "legacy" parts of a company developed cost estimates or wrote contracts or performed inventory tracking (or choose your favorite enterprise-wide process) in one way, while the "newly added/consolidated" pieces of the company were developing these processes or estimates in a different manner. Such non-standardization proved extremely inefficient for the enterprise at large, and these companies needed to remedy this emerging problem.

The general solution – with some company-to-company variations – was that these companies approached the Society for Cost Estimating and Analysis (SCEA) (now called the International Cost Estimating and Analysis Association (ICEAA)) in search of cost estimating intellectual capital that they could use as an "enterprise-wide" standard. The result was their adoption of ICEAA's Cost Estimating Book of Knowledge (CEBoK®). A full explanation of CEBoK can be found at the ICEAA website, while a summary description is included here:

3.4.5 COST ESTIMATING BOOK OF KNOWLEDGE (CEBOK)

"CEBoK is ICEAAs official training course material. It is a user-friendly cost estimating and analysis training resource, with information organized into 16 interactive modules within five general subject areas, designed to cover all of the topics that represent the body of knowledge that ICEAA promotes and tests for in the Certified Cost Estimator/ Analyst (CCEA) exam. The modules are designed to facilitate self-study and study in small groups, or can be used effectively in a classroom environment." [8] The five sections and 16 modules in CEBoK are:

- Section 1: Cost estimating
  1. Cost estimating basics: introduction, overview, cost products, and the cost estimating process
  2. Costing techniques: using costing techniques and comparison of techniques
  3. Parametric estimating: the basics of parametrics and the parametric estimating process
- Section 2: Cost analysis techniques
  4. Data collection and normalization: the importance of data, key principles, collection considerations, collection process, collection techniques, sources, and data normalization
  5. Inflation and index numbers: inflation and cost estimating, concepts and definitions, indices, tables and escalation procedures
- Section 3: Analytical methods
  6. Basic data analysis principles: the types of data, univariate data analysis, scatter plots, data validation, and visual display of information
  7. Learning curve (LC): LC theory, application, and advanced topics
  8. Regression analysis: bivariate data analysis, regression models, preliminary concepts, linear regression, nonlinear models, and how to select models
  9. Cost and schedule risk analysis: types of risk, risk modeling, risk management, cost growth analysis, and schedule risk analysis
  10. Probability and statistics: statistical measures, probability distributions, Monte Carlo Simulation considerations and process, techniques, sources, and normalization
- Section 4: Specialized costing
  11. Manufacturing cost estimating: process overview, functional cost elements, labor and material estimating and issues
  12. Software cost estimating: the software development process, approaches, drivers, estimating techniques, challenges, data collection, and models

- Section 5: Management applications

  13. Economic analysis: principles, process, and special cases

  14. Contract pricing: process, types, basis of estimate (BOE) documentation, and standards

  15. Earned value management: EVM components, analysis, risk management integration, rules of thumb, and tools

  16. Cost management: total ownership cost (TOC), cost as an independent variable (CAIV), target costing, activity-based costing (ABC), initiatives, and cost estimating

Additionally, these companies encourage professional development to further enhance the skill-sets of its workforce, and most offer a tuition assistance program for certification courses as well as continuing education (CE) at both the Bachelor and Master levels. They encourage external training through distance-learning extension programs at select universities and through professional affiliations such as ICEAA. Furthermore, they provide their employees with a variety of opportunities to participate in internal and external seminars, conferences and training, including webinars.

3.4.6 FEDERALLY FUNDED RESEARCH AND DEVELOPMENT CENTERS (FFRDCs)

FFRDC's conduct research for the US Government. They are administered in accordance with US Code of Federal Regulations, Title 48, Part 35, Section 35.017 by universities and corporations. There are 39 recognized FFRDCs that are sponsored by the US government and a complete list can be found at the National Science Foundation website in reference [9].

What does an FFRDC do? An FFRDC "meets some special long-term research or development need which cannot be met as effectively by existing in-house or contractor resources. FFRDCs enable agencies to use private sector resources to accomplish tasks that are integral to the mission and operation of the sponsoring agency. An FFRDC, in order to discharge its responsibilities to the sponsoring agency, has access beyond that which is common to the normal contractual relationship to Government and supplier data, including sensitive and proprietary data, and to employees and installations equipment and real property. The FFRDC is required to conduct its business in a manner befitting its special relationship with the Government, to operate in the public interest with objectivity and independence, to be free from organizational conflicts of interest, and to have full disclosure of its affairs to the sponsoring agency." [10]

In the following sections, we discuss the cost estimating capability of three FFRDCs, namely the Institute for Defense Analysis (IDA), The MITRE Corporation (MITRE), and the RAND Corporation (RAND).

3.4.7 THE INSTITUTE FOR DEFENSE ANALYSIS (IDA)

IDA assists the United States government in addressing important national security issues, particularly those requiring scientific and technical expertise. IDA only works for the US government, and, to ensure freedom from commercial or other potential conflicts of interest, IDA does not work for private industry. IDA manages three centers, two of which

serve primarily the Department of Defense and the third that serves the Executive Office of the President. The Studies and Analysis Center is the largest of IDA's three FFRDCs. It supports the Office of the Secretary of Defense, the Joint Chiefs of Staff, and the unified military commands, and it includes the Cost Analysis and Research Division (CARD), which houses IDA's cost analysis capability.

IDA is sponsored to do its work by either OSD or the Joint Staff. In some cases, that sponsorship is due to a Congressional requirement for an independent analysis. Additionally, IDA sometimes performs cost estimates or develops cost estimating tools to support the Services, but OSD or Joint Staff sponsorship is still normally required on these estimates even when the funding is coming from the Services. IDA's cost estimating work typically does require access to sensitive and proprietary data and certainly requires IDA to maintain objectivity and independence at both the individual and organizational levels [10]. For additional information on IDA, visit their website, shown in reference [11].

3.4.8 THE MITRE CORPORATION

The mission of MITRE is to work in partnership with the US government, applying systems engineering and advanced technology to address issues of critical national importance in the areas of scientific research and analysis, development and acquisition, and systems engineering and integration. MITRE's cost analysis capability lies within the acquisition effectiveness organization. MITRE can support their sponsors with capabilities in cost analysis, cost engineering, system assessment, trade-off analysis, and acquisition decision support. Additional information on MITRE can be found at their website, shown in reference [12].

3.4.9 RAND CORPORATION

RAND is a nonprofit global policy think tank formed to offer research and analysis to assist in developing objective policy solutions. It has numerous locations, including a few in Europe, with its headquarters in Santa Monica, CA [13]. It is composed of over 15 major sub-organizations. From the cost estimating perspective, the most important sub-organization is RAND Project Air Force, and within that, the Resource Management Program (RMP). The goal of this program is to maximize the efficiency and effectiveness of Air Force operations in a resource-constrained environment, and to that end it provides studies, supports analyses and conducts research in many areas including weapon-systems cost estimating. For more information on RAND, visit their website at http://www.rand.org.

In conclusion, the intent of this part of the chapter was to identify the broad organizational reach of the cost estimating profession, as well as to identify standard processes and best practices that can help these organizations as they develop cost estimates. We make the point that there are numerous organizations that support a cost estimating capability within their organization. Some of these organizations are described in this chapter, in order to display the great diversity that these organizations represent. In Chapter 2, we defined the DoD acquisition process in great detail, and found that the DoD 5000 Series provides a regulatory requirement that requires program managers and government agencies preparing cost estimates to subject their programs to progress reviews called milestones, and that they must also prepare numerous documents

providing system descriptions, specifications, necessary quantities, etc, for completion and dissemination at prescribed times. But while many organizations are not government organizations and thus are not subject to the same regulatory regime, they do have access to the standard processes and best practices identified so far in this chapter.

How is this accomplished? What structure should they follow to complete a given developmental or acquisition project? These non-DoD organizations and FFRDC's can gravitate to one of a few systems that help them achieve this goal. First, the GAO Cost Estimating and Assessment Guide provides a 12-step process that serves as a guide that analysts can use to create a comprehensive, accurate, credible, and well-documented and well-estimated cost analysis. The structure of the 12-step process offers the opportunity for a program's cost estimates to be validated both internally and externally as well. Second, the ICEAA has developed and maintains the Cost Estimating Book of Knowledge (CEBoK), which can provide a template and structure for them to follow as well.

Now that we have completed describing both the DoD acquisition process (Chapter 2) and the non-DoD acquisition process (Chapter 3), let us now focus on the terminologies and concepts that are used in the cost estimation discipline.

3.5 The Cost Estimating Process

In this section, we will describe the key terms and concepts used in the cost estimating process. While many of these terms and concepts are representative of the process followed by the US Department of Defense, the more salient fact is that they have broad and general applicability across any organization, even if a name may be different from organization to organization. When commencing a cost estimate, the following four steps are pertinent guides in the general cost estimating process, and we will describe each step in detail:

1. Definition and planning, which includes:
    a. Knowing the purpose of the estimate
    b. Defining the system
    c. Establishing the ground rules and assumptions
    d. Selecting the estimating approach
    e. Putting the team together
2. Data collection
3. Formulation of the estimate
4. Review and documentation

3.6 Definition and Planning. Knowing the Purpose of the Estimate

When commencing a cost estimate, it is important to first know the purpose of the estimate. Why is this important? Aren't all cost estimates the same? Recall from Chapter 2 that two of the major reasons for conducting a cost estimate are to support the budget process (both the formulation and justification of budget proposals), as well as to help in choosing among alternatives ("comparative studies"). If the purpose for developing the

cost estimate is to support the budget process, we will also want to know what the estimate will be used for, because that purpose will dictate the scope of the estimate, the amount of time the estimate takes to develop, the level of detail in which we need to do the estimate, and which estimating techniques we will use in the development of the cost estimate. Surely an estimate of the one-year budget for research and development requires a different level of effort than one for a milestone review, which will cover all phases of the life cycle.

There are numerous types of cost estimates and all have different names. A list of the primary cost estimates that are associated with budget formulation and justification are listed here:

- Life cycle cost estimates (LCCEs)
- Program office estimates (POEs)
- Baseline cost estimates (BCEs)
- Independent cost estimates (ICEs)
- Service cost positions (SCPs)
- Rough order of magnitude estimates (ROMs), and
- "What-if" exercises

- A *Life Cycle Cost Estimate (LCCE)* is a cost estimate that encompasses *all* phases of the product's life cycle, including R&D, production, operations/support, and if needed, disposal. This is the primary estimate completed in the cost estimation field, and an LCCE is needed at each milestone. It considers all costs in a program from its inception until it is completely disposed of. A popular outlook for an LCCE is considering all costs in a program "from cradle to grave."

- The *Program Office Estimate (POE)* is the Program Manager's estimate of the resources required for his/her program. It will be continually updated throughout the life of the program. History suggests that a POE is often an "optimistic" estimate, as the PM will want to think that his/her program will run smoothly with minimal delays or technical issues encountered.

- A particular POE is the *Baseline Cost Estimate (BCE),* which is usually the first POE completed on an acquisition program. The BCE is important because as the program progresses, it is used to measure program cost growth. For example, if the BCE (the first POE) is for $100M at the Milestone A review, and the POE at the Milestone B review is for $120M (adjusted for inflation), then comparing the Milestone B estimate with the BCE, you can observe that there has been a $20M and 20% program cost growth since the first milestone review.

- The *Independent Cost Estimate (ICE)* is an estimate developed by an organization *independent* of those responsible for developing, acquiring and operating the system, and its purpose is to check the completeness and credibility of the POE. It is the estimate that is legislatively mandated for the large ACAT I programs, and all ACAT I programs (and some ACAT II programs) will have an ICE completed on them. The ICE is usually prepared by OSD CAPE on a DOD or Joint program. An ICE can also be developed by a Service Cost Agency as a "component" estimate, and this estimate is sometimes referred to as a Component Cost Analysis (CCA). While POE's are generally optimistic, ICE's are generally more pessimistic in nature, although the word used within the cost estimating profession is "conservative."

- What if the POE and ICE are not similar to each other? Each of the services has a process which melds the POE and Service Cost Agency's cost estimate (whether it is the official ICE or the CCA) into what is called a *Service Cost Position (SCP)*. In this process, the two estimates are compared to each other. If the estimates are similar in cost, then the SCP can be either one of the estimates. But if the estimates are not "close" to each other in cost, then efforts must be made to reconcile and resolve these differences and the SCP emerges *after* this reconciliation. Note that, as often happens in the discipline of cost estimating, the word "close" is not defined. More detailed information on the SCP can be found at the DAU website [14].

- There are times when a cost estimate needs to be just an approximation, such as identifying when a solution or alternative may be approximately $5M or approximately $75M. In the first case, the $5M may actually turn out to be $4.2M or $6.3M as an actual cost, but the $5M at least gives you a "rough" starting point as to what the cost may be. The same applies for the $75M estimation. It may turn out that the actual cost for that project might be $68M or $82M after a detailed cost estimate is completed, but the $75M gives the decision maker at least a starting point for the cost of a particular alternative. This might be useful when an initial idea of the functionality of the product is known, but well before the design of the product has been completed in any detail. There is therefore very little specific information available about the product on either its technical detail or the timeline over which it will be developed. When a cost estimate is desired at an early stage, the resulting estimate is called a *Rough-Order-of-Magnitude (ROM)* estimate.

- Another type of cost estimating is known as a "*what-if*" exercise. This is an analysis in which one programmatic aspect of a program is being changed, and there is a need to estimate the financial consequence of this change. The change might be an adjustment to the quantity of the items being procured, or for an extension of the timeline during which research and development takes place, or it might be a change to the mean time between failure (MTBF) of an important component. In all cases, a good program manager wants to know the financial impact of the proposed change, and cost estimation is the discipline that provides that insight. This is many times referred to as "sensitivity analysis," as well.

Author's Note: For the U.S. Army personnel reading this, the USA Training and Doctrine Command has produced a succinct two-sided summary of numerous cost terms and Army points of contact called the "TRADOC Analysis Center Cost Analysis Smart Card", version 1.6 dated 01 OCT 2013, for use by its Army employees.

Having discussed the types of cost estimates that are associated with budget formulation and justification, we will now identify those cost estimates that are associated with comparative studies. These estimates occur most often in the process of comparing costs and benefits among alternative courses of action in order to achieve a particular objective. These comparative studies have several names in the literature, including economic analyses, analyses of alternatives (AoAs), cost of operations and effectiveness analysis (COEA), cost-benefit analysis (CBA), business case analysis (BCA), and trade-off studies. Regardless of their name, they all seek to compare alternatives, seeking the best value and "bang

for the buck" among competing alternatives, while highlighting relative advantages and disadvantages of the alternative courses of action, and also considering the sensitivity of the results of the analysis to changes in the assumptions of the analysis. These analyses are also used to support source selections, in-sourcing vs. out-sourcing decisions, and base realignment and closure (BRAC) decisions.

One of the desired outcomes of cost estimates is the ability to identify cost growth in a program. A management technique for establishing cost goals and containing cost growth is called "Cost as an Independent Variable" (CAIV). CAIV's main focus is containing cost growth. CAIV centers on the trade space between performance, schedule, cost and risk, as it searches for the "best value for a given price" solution, rather than achieving the greatest performance or lowest cost possible. To understand CAIV, let's contrast it with the traditional process of cost estimating. Generally, a cost estimate is developed from a design and the cost is dependent upon the technical and performance characteristics of that program. In CAIV, the process is reversed, so that a cost goal – usually an aggressive cost goal – is set (say at $10M), and then a "best value" of combinations of performance and technical design parameters that meet this cost goal are sought. CAIV contains growth by essentially saying "Whatever you build, it cannot exceed this cost." You can utilize CAIV in your personal life when you shop for a new car. If you buy a $30,000 car, it will usually price out closer to $33,000–$34,000 once taxes and delivery charges, etc, are incorporated. But using CAIV while shopping, you would tell the car salesman that you have no more than $30,000, so whatever car you get, it cannot cost out to any higher than that amount. Using CAIV in this fashion, you are containing any cost growth that might occur during the purchase, though perhaps at the expense of the performance of that vehicle as options may be removed to ensure price containment.

While discussing CAIV, we should also cover a topic called the P, D, T criteria. This criteria represents a trade-off analysis between competing priorities in the areas of Performance (P), Dollars (D), and Time (T). Each of these criteria has a "good" outcome and a "bad" outcome:

- *Performance*: High performance is "good" while low performance is 'bad"
- *Dollars*: Inexpensive is "good" while expensive is "bad"
- *Time*: A short wait is "good" while waiting a long time is "bad"

It is a commonly accepted practice within the PM community that the interaction and relationship of these three variables is such that you can have any two of the "good" outcomes at the expense of the third, implying that the third outcome will be the "bad" one. For example, if your priorities are such that you want a new item to possess great handling qualities (high performance) and you want it quickly (short wait), then you have decided what your two "good" outcomes will be, so the third variable (dollars) will be the "bad" outcome – in this case, it will be expensive to purchase. If you don't want to spend much money on the item or project ("good") and you want it quickly ("good"), then it will not be a high performance item. This trade-off dilemma can be critical in the execution of a program, as the Program Manager must decide and "balance" his/her program issues between the performance, schedule, and the cost of the program as variables change. No matter what your desires are, you can only get two of the good outcomes at the expense of the third.

3.6.1　DEFINITION AND PLANNING. DEFINING THE SYSTEM

No one could possibly develop a cost estimate without having definitions and descriptions of the system to be estimated. This system description will provide the basis upon which the system cost will be estimated, including physical and performance characteristics, plus development, production and deployment schedules. The Cost Analysis Requirements Description (CARD) provides this information that you need. When you are assigned a cost estimate, the first question you will probably ask is "When is it due?" so that you will know how much time you have to complete it. The second and next immediate question should then be "Where is the CARD?" Since their introduction into the cost estimating discipline, CARDs have become the basis for sound cost estimates and for tracking program changes, and is the baseline from which the life cycle cost estimate is produced. It describes all of the salient features of the acquisition program and of the system itself, and includes a milestone schedule, the Work Breakdown Structure (WBS) and the Cost Element Structure (CES). The DoD 5000 series acquisition process requires that a CARD be developed as part of the decision review of each key milestone. Figure 3.2 displays the top level Table of Contents for a CARD (on the left), as well as a lower level breakout (on the right) for three selected chapters.

DoD CARD Table of Contents

Table of Contents
1.0　System Overview
2.0　Risk
3.0　System Operational Concept
4.0　Quality Requirements
5.0　System Manpower Requirements
6.0　System Activity Rates
7.0　System Milestone Schedule
8.0　Acquisition Plan and/or Strategy
9.0　System Development Plan
10.0　Facilities Requirements
11.0　Track to Prior CARD
12.0　Contractor Cost Data Reporting Plan

1.0 System Overview
1.1 System Characterization
1.2 Technical Characteristics
1.3 System Quality Factors
1.4 Embedded Security
1.5 Predecessor and/or Reference System

3.0 System Operational Concept
3.1 Organizational Structure
3.2 Basing and Deployment Description
3.3 Security
3.4 Logistics

9.0 System Development Plan
9.1 Development Phases
9.2 Development Test and Evaluation
9.3 Operational Test and Evaluation

FIGURE 3.2 Cost Analysis Requirements Description.

　　The CARD is the common program description used to develop both the POE and the ICE, and if reconciliations are required, then it is also used for the SCP. Once a requirement is determined, representatives from both the program office and the prime contractor will meet to discuss and determine the programmatic and technical data that is required for this program to create the CARD. Once the CARD is created, if the POE and ICE differ, these differences will be reconciled to then create the Service Cost Position. Figure 3.3 displays this process:
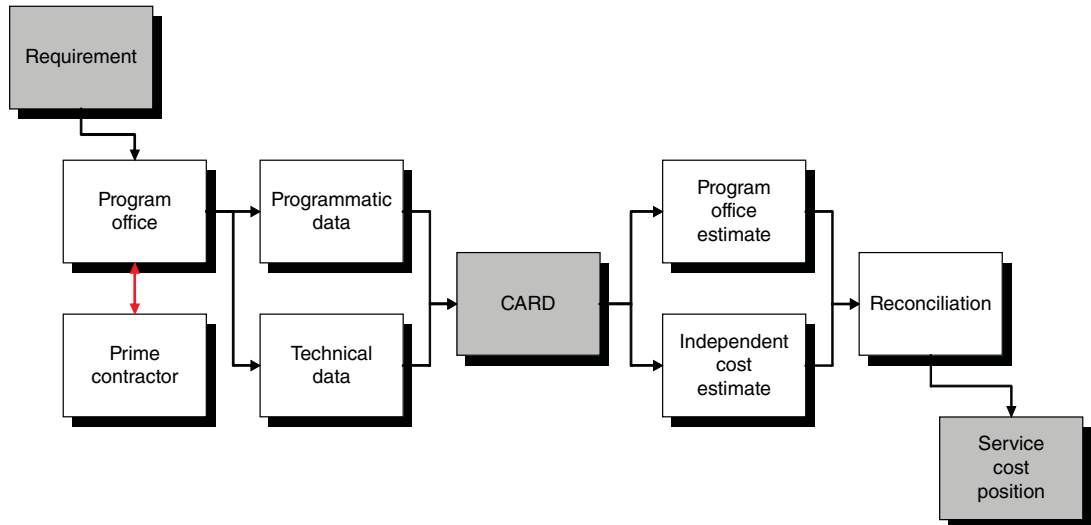
FIGURE 3.3 The CARD Reconciliation Process.

3.6.2 DEFINITION AND PLANNING. ESTABLISHING THE GROUND RULES AND ASSUMPTIONS

There are many ground rules and assumptions necessary to develop a cost estimate. As was previously discussed in Chapter 1, the ground rules and the assumptions that you make will greatly affect the outcome of your cost estimate. These ground rules and assumptions are dynamic and thus are always changing, from the beginning of an estimate until its completion. The following list encompasses just a few of the myriad ground rules and assumptions necessary in a typical program:

- Determining the base year of the dollars in your estimate
- Specific inflation indices to be used
- Participants, including contractors
- Determining the timing of initial operating capability (IOC) and full operating capability
- Scope and possible limitations of the program
- Schedule and time phasing
- Timetable for transition to each phase in the life cycle
- Determining any foreign participation assumptions
- Determining technology levels of readiness
- Determining relationships to, and dependencies on, other programs
- Determining maintenance and logistics support concepts

Be aware that it is highly unlikely that all of the ground rules and assumptions that your cost estimate is initially based on will turn out to be true as the program history unfolds. It is important to understand that they will need to be continually updated throughout the life of your program.

3.6.3 DEFINITION AND PLANNING. SELECTING THE ESTIMATING APPROACH

When determining how you will want to formulate your cost estimate, keep in mind that the goal is to provide an estimate that is realistic and credible. To achieve this goal, the cost estimator should always strive to follow these good cost estimating practices:

- The estimate should be anchored in historical program performance.
- The cost drivers that underpin our estimates must be intuitively appealing, in the sense that an engineer would agree that the cost driver is a reasonable explanatory variable for cost. We often begin our analysis using weight as a cost driver, since weight often correlates with cost, and because weight is an engineering variable for which we have an estimate very early in the life of a program, but this is not the only variable that we can use!
- The estimate should be accompanied by good statistics. This would include the significances of the F- and t-statistics for the underlying regressions.
- The estimate should be requirements driven, to include programmatic and system requirements.
- The estimate should be well-defined in terms of its content and risk areas.
- The estimate must reflect future process and design improvements. A word of caution here is that the estimator must be convinced that these future processes and design improvements will indeed take place, and that they are not fanciful wishes
- The estimate needs to be validated by independent means, and within estimating/modeling accuracy.
- The estimate should be traceable and auditable and able to be recreated from the bases of estimates.

At its conclusion, the estimate should be explainable to management and to program and business leaders. This is a variation of the "keep it simple" principle, which is not meant to speak down to your audience, but rather to recognize that your audience is (as we all are) very busy in their lives, and that we are lucky when they read even the executive summary of our documentation. Rarely does our audience read the entire details of our cost estimates.

The methodologies we use should be one of the three recognized methodologies, which include the following:

- Analogy estimating
- Parametric estimating
- Engineering build-up estimating

While we will cover these three briefly here, detailed descriptions of the first two methodologies are provided in later chapters.

- The *Analogy* cost estimating approach is often characterized by the phrase "It's Like One of These," because it is based upon using a single analogous historical data point. It is generally used when there is only *one* historical program to compare your program to, or when you do not have the time for a full analysis of numerous systems

and you need a quick answer. It can also be useful when you need a quick "ballpark" estimate.

Overall, this technique is used to compare the new system that you are trying to estimate with one historical system. The key to this approach is selecting that historical system that most closely resembles the new system. In doing so, you will need to determine the major cost drivers similar to each program; determine the similar characteristics of the systems being compared; assess the design and production implications of the proposed system relative to the analogous system(s); and estimate the cost based on these design/production considerations.

TABLE 3.1   Analogy Example

Engine	Thrust (hp)	Cost (FY14$M)
Historical	2500	200
New	3500	

Algebraically, the analogy approach can be explained by establishing a ratio of the value of the explanatory variable in the new program to the value of the explanatory variable in the historical program. For example, in Table 3.1, the thrust of the historical engine is 2,500 horsepower (hp), while the thrust of the new engine is 3,500 hp. When you establish a ratio between the two engines (new to old), calculations show that 3500/2500 = 1.4. Thus, the new engine has 1.4 times the thrust of the old engine. If this were the only explanatory variable available to us for comparison purposes, we could then estimate the cost of the new engine as 1.4 times the cost of the old engine. In this case, our cost using the analogy technique would be $200M \times 1.4 = $280M (FY14$).

The downside of using the analogy technique is that it is an estimate based upon only a single historical data point, which may not be representative of the problem being addressed. This technique is discussed in greater detail in Chapter 14.

- The *Parametric* cost estimating approach is often characterized by the phrase "This Pattern Holds," because it is based on the idea that a collection of data points, when displayed as a scatter plot, appear to have a particular pattern. Once this pattern is established, we will then assume that this historical pattern will hold in the future. We use the word "parametric" in the sense of a "characteristic," and we are asserting that cost is a function of physical and performance characteristics, which are also called "explanatory variables." Using functional notation, we get that

$$\text{Cost } = f(\text{physical and performance and technical characteristics of a program})$$

This equation is interpreted as "Cost is a function of the physical and performance and technical characteristics of a program." Underlying the parametric cost estimating approach are "Cost Estimating Relationships" (CERs), which rely on explanatory variables such as weight, power, speed, frequency, thrust (and so on) to estimate costs. The procedure for doing this consists of "statistically fitting a line or function to a set of related historical data, and then substituting the value of the parameter of the new system into the resulting equation." This is in essence the definition of Regression Analysis and is the fundamental tool used in this estimating approach. Thus, when the dependent variable in a regression is cost, the ultimate equation is known as a

CER. It is important that the CER is developed using analogous historical data. That means that the typical delays, problems, mistakes, redirections, and changing characteristics that occurred during the development of the historical programs will recur in some rough sense in the new system, as well. Regression Analysis will be covered in great detail in Chapters 7–9.

In order to develop a parametric cost estimate, the cost estimator needs a data base of historic cost and performance data. For example, if you are estimating the cost of a new building, explanatory variables that you might consider would be the number of square feet in the building, or the number of floors, or perhaps the construction schedule (time). If we are estimating the cost of an aircraft, explanatory variables might be weight, speed, wing area, or aircraft range.

A special case of a parametric approach is the Learning Curve, which postulates that the recurring cost (measured either in dollars or in hours) of doing or producing something diminishes with repetition. You get "better" as you go along, as you "learn" the process better. The specific structure of the learning curve is Cost $= A \times X^b$, where the intercept "A" and the slope "b" are parameters that are estimated by using regression analysis. We will cover learning curves in great detail starting in Chapter 10.

- The *Engineering Buildup* cost estimating approach is often characterized by the phrase "It's Made Up of These," because it is a detailed, "bottom-up" application of labor and material costs. In this approach, many detailed estimates are summed together to form the total estimate. A significant characteristic to this approach is that it is in fact what people *outside* the cost estimating profession believe is the best cost estimating approach due to its great detail. The downside of the approach, however, is that it is very data intensive and time consuming and therefore expensive to produce. The authors of this text do not believe that the increased expense of producing an engineering approach cost estimate is justified, as it rarely provides any significantly greater accuracy than other methodologies. In fact, many small errors can combine to produce a large error in the overall cost estimate.

In conclusion, there are three primary cost estimating approaches: Analogy estimating; Parametric estimating; and Engineering build-up estimating. Some in the cost estimating field consider there to be a fourth approach called "Extrapolation from Actuals." This method is used after production has already begun and a number of items have already been produced in order to estimate the cost of continued production. It can also be used to estimate costs after a major change in quantity or performance of an existing system. However, we feel that this is just an application of learning curves, in which the values of the independent variable come not from analogous programs but from the program itself. That makes this method a particular kind of learning curve, which in turn, is a particular kind of parametric approach. Therefore, we do not feel that "Extrapolation from Actuals" is its own unique cost estimating approach.

3.6.4 DEFINITION AND PLANNING. PUTTING THE TEAM TOGETHER

The last step in the Definition and Planning phase is putting the team together. It is rarely the case that a cost estimate is developed by a single person. In truth, cost estimating is a "team sport," in the sense that many functional disciplines cause costs to accrue in a program. It is not possible for the cost estimator to understand the details of each of

these functional disciplines, and therefore we include numerous representatives on the cost estimating team. Members of the team will include engineers, logisticians, contracting officers, and program management personnel, to name just a few.

3.7 Data Collection

Data collection is one of the most important and time-consuming activities in developing a cost estimate. The type of data that we collect will depend upon the estimating methodology that we intend to use. You will need to determine what your cost drivers are, what raw cost data you will need, and what analogous program costs exist. The availability of needed data may force a change in the estimating methodology that you use, as well.

Chapter 4 addresses Data Sources and the data collection issue more fully, but suffice it to say that the data we collect is driven by two contradictory impulses. On the one hand, we want to collect as much performance and technical and programmatic data as possible for the program whose cost estimate we are developing, as well as data from other analogous programs. But on the other hand, we are many times limited by the general scarcity of analogous data, especially at the lower ends of the WBSEs. In order for us to access the data we seek, it must have been collected previously and then made available to us in a timely fashion. Thus, there can be a tension between wanting to collect more data – since more data generally appeals to analysts – and the difficulties of actually collecting more data. Moreover, more data generally means more work, for which you may not have the time if your estimate is due quickly.

While performance and technical data are almost always available, it is unfortunately a fact that cost data from a program is not always documented, making data collection more difficult.

3.8 Formulation of the Estimate

It is in this process that we develop the factors and analogies and CERs, including learning curves, to provide a cost estimate of each of the individual work breakdown structure elements. It is important to remember that each of these estimates and the methodologies that provide these estimates, is underpinned by the data that you collected in the data collection phase. So, we can observe that if the estimate is carried out at a very detailed level – that is, with many WBSE's – then the data collection effort is commensurately larger. If this is the case, the cost estimator wants to be careful not to spend too much time on the task of data collection so that there is no time left for the analysis and development of the cost estimate. Additionally, in this phase of the cost estimating process, the estimates are aggregated into the major phases of the life cycle, namely research and development, production, and operating and support costs. It should be noted that the final cost estimate is usually completed in a particular constant year dollar and is shown as a point estimate. However, if this estimate is to support the budget process, then there are two more steps that need to be taken. Those two steps are to: (1) spread the LCCE across the relevant fiscal years, and (2) apply the appropriate inflation indices in order to transform the constant year estimates into Then-Year estimates. These two steps result in the production of a time phased, inflation adjusted life cycle cost estimate.

3.9 Review and Documentation

A significant amount of work goes into the many steps outlined earlier, and it is unlikely that a process involving all of these steps will be mistake-free. Even though most estimates are developed in a spreadsheet environment, errors of omission and commission regularly occur, so that some quality assurance to check that the estimate is reasonable and realistic and complete is in order, before the cost estimator "goes public" with the estimate. A good way to cross-check the results of your estimate is to consult with those coworkers who assisted you in a particular area. Does your professional colleague think that your estimate is reasonable and realistic and complete? Did you interpret what they said correctly? It is also true that a single point estimate is almost assuredly going to turn out to be incorrect. This is because, as we have stated before, it is highly unlikely that all of the ground rules and assumptions underlying the cost estimate will turn out to be true as the program history unfolds. Therefore, sensitivity analyses are in order to understand the financial consequences of various technical and programmatic risks and uncertainties. Finally, documentation of the cost estimate that was developed is a "best practice" within the profession. It provides a means for other analysts today to reproduce what we are working on, as well as to use it for future cost estimating. While you are using historical data to work on your program today, the study that you are working on now will be part of the historical data used in a few years. So document well and document often! A guide to the required level of completeness and complexity of the documentation is that it should be written so that it is understandable by a reasonably intelligent liberal arts major.

Sections 3.5–3.9 detailed the four sections and steps of the cost estimating process. Having discussed numerous aspects of the process and data collection, we need to discuss how we handle and categorize the data that we are able to retrieve for our cost estimate. We will place each data set into a functional grouping "family tree" called a Work Breakdown Structure.

3.10 Work Breakdown Structure (WBS)

A WBS decomposes a project into smaller components for ease of management control. It is best described as "a product-oriented family tree composed of hardware, software, services, data, and facilities which results from systems engineering efforts during the development and production of a defense material item." It displays and defines the products being developed and produced, and it relates the elements of work to be accomplished to each other and to the end product. By displaying and defining the efforts to be accomplished, the WBS becomes a management blueprint for the product. Its relational aspects, including the time-phasing, duration and "what-gets-done-first," communicate management's plan for how a program is to be completed.

3.10.1 PROGRAM WORK BREAKDOWN STRUCTURE

This is the structure that encompasses an entire program at a summary level. It is used by government cost estimators and contractors to develop and extend a *contract work breakdown structure*.

A program WBS consists of at least three levels of the program with associated definitions:

- Level 1: The entire material item such as an aircraft system, ship system, space system, surface vehicle system, etc.
- Level 2: Major elements of the material item found in Level 1
- Level 3: The elements subordinate to the Level 2 major elements

A sample Program WBS with three levels of WBS elements can be found in Figure 3.4.

Program WBS		
1	2	3
Aircraft system	Air vehicle	Air frame Propulsion Communications/identification Navigation/guidance Fire control Automatic flight control Central computer Electronic warfare suite Weapon delivery equipment Armament
	System test and evaluation	Development test and evaluation Operational test and evaluation Mockups Fire control Test facilities
	Systems engineering/program management	Systems engineering Program management Integrated logistic support
	Common support equipment training	Maintenance trainers Aircrew training device Training course materails
	Data	Technical publications Engineering data Management data Support data Data depository
	Operational/site activation	Contractor technical support
	Initial spares and repair parts	

FIGURE 3.4 Top Level Program WBS.

Let's interpret this WBS. First, we can observe that the Level 1 material item is the aircraft system. Let's suppose that the cost of this aircraft system in Level 1 is $200M. In addition, we can observe that the aircraft system is comprised of seven different WBS elements, all found in Level 2: the air vehicle, system test and evaluation, systems engineering/program management, common support equipment training, data, operational/site activation, and initial spares and repair parts. Each of these seven major elements of the aircraft system will have a cost associated with them and these seven costs will add up to $200M.

Moving to Level 3, we can observe that the Air Vehicle consists of ten separate WBSEs: the airframe, propulsion, communications/identification, navigation/guidance, fire control, automatic flight control, central computer, electronic warfare suite, weapon

delivery equipment, and the armament. The System Test and Evaluation major element in Level II consists of five areas found in Level 3: development test and evaluation, operational test and evaluation, mockups, fire control, and test facilities. The cost of the ten elements in Level 3 for the Air Vehicle sum to the total cost for the Air Vehicle, just as the next five items in Level 3 will sum to the cost of system test and evaluation in Level 2, and both contribute to the total cost of Level 1. The same goes for all of the other areas in Levels 2 and 3, as well.

While Figure 3.4 is a sample WBS and contains elements specific to that program, there are elements common to most types of systems. These include the following:

- Integration, assembly, test and checkout
- System engineering/program management
- System test and evaluation
- Training
- Data
- Peculiar support equipment
- Operational/site activation
- Industrial facilities
- Initial spares and repair parts

Figure 3.4 displays a Top Level Program WBS that consists of three levels; but sometimes additional information is needed. An Expanded WBS consists of an additional level of elements, making four levels total. An example of an Expanded WBS is shown in Figure 3.5.

Program WBS			
1	2	3	4
Aircraft system	Air vehicle	Air frame	Wing Fuselage Empennage Flight control Hydraulic system Environmental control Crew station system Landing/arresting gear system Integ, assembly, test, chkout
		Propulsion Communications/identification	Radio system Data link Communications system S/W
		Navigation/guidance Fire control	Radar Computer Controls and displays System software
	System test and evaluation	Automatic flight control Central computer Electronic warfare suite Weapon delivery equipment Armament Development test and evaluation	Wind tunnel articles and test Static articles and test Fatigue articles and test

FIGURE 3.5 An Expanded Four Level Program WBS.

Note that in the expanded Four Level Program WBS, the first three levels are identical to the Top Level Program WBS, but an additional level has been added. In Figure 3.5, we can notice that the Air Frame in Level 3 consists of nine WBS elements, starting with the wing and fuselage and ending with the integration, assembly, test and checkout costs. The costs of all nine elements are summed to calculate the total cost of the Air Frame shown in Level 3.

Now that we understand what a WBS is and what its prime components are, let's examine the regulations provided for a government program concerning WBSs. This regulation is applicable to all DoD entities as well as by defense contractors.

3.10.2 MILITARY-STANDARD (MIL-STD) 881C

The Department of Defense published MIL-STD 881C on 3 October 2011, entitled "Work Breakdown Structures for Defense Materiel Items." This standard is approved for use by all Departments and Agencies of the Department of Defense (DoD) as a guide and direction for all WBSs in government acquisition and/or development programs. The primary objective of this Standard is to achieve a consistent application of the WBS for all programmatic needs, including performance, cost, schedule, risk, budget, and contractual. This Military Standard is applicable to all defense materiel items (or major modifications) established as an integral program element of the Future Years Defense Program (FYDP), or otherwise designated by the DoD Component or the Under Secretary of Defense (Acquisition). This Standard is mandatory for all ACAT I, II, and III programs, and it should be included as a contract requirement [15].

MIL-STD 881C presents direction for effectively preparing, understanding, and presenting a Work Breakdown Structure. It provides the framework for Department of Defense (DoD) Program Managers to define their program's WBS and also to defense contractors in their application and extension of the contract's WBS. MIL-STD 881C is divided into four sections: Section 1 defines and describes the WBS. Section 2 provides instructions on how the WBS is applied as well as how to develop a Program WBS in the pre-award timeframe. Section 3 provides direction for developing and implementing a Contract WBS and Section 4 examines the role of the WBS in the post-award timeframe. Additional information on Mil-STD 881C can be found in Reference [15].

3.11   Cost Element Structure (CES)

While a WBS is used in the Research and Development and Production phases of the life cycle, the WBS equivalent for Operating and Support (O&S) Costs is called the cost element structure (CES). The CES establishes a standard matrix for identifying and classifying system O&S costs. It is designed to capture as many relevant O&S costs as practical, and should be tailored to meet each specific system's needs. Its purpose is the same as the WBS's purpose. A generic CES can be found in Figure 3.6.

Note that the aforementioned cost element structure consists of seven distinct areas. A description of each of these follows.

1.0	Mission Personnel	
	1.1	Operations
	1.2	Maintenance
	1.3	Other
2.0	Unit-level Consumption	
	2.1	POL/Energy Consumption
	2.2	Consumable Material/Repair Parts
	2.3	Depot-Level Repairables
	2.4	Training Munitions/Expendable Stores
	2.5	Other
3.0	Intermediate Maintenance (External to Unit)	
	3.1	Maintenance
	3.2	Consumable Material/Repair Parts
	3.3	Other
4.0	Depot Maintenance	
	4.1	Overhaul/Rework
	4.2	Other
5.0	Contractor Support	
	5.1	Interim Contractor Support
	5.2	Contractor Logistics Support
	5.3	Other
6.0	Sustaining Support	
	6.1	Support Equipment Replacement
	6.2	Modification Kit Procurement / Installation
	6.3	Other Recurring Investment
	6.4	Sustaining Engineering Support
	6.5	Software Maintenance Support
	6.6	Simulator Operations
	6.7	Other
7.0	Indirect Support	
	7.1	Personnel Support
	7.2	Installation Support

FIGURE 3.6 Generic Operating and Support Costs CES.

1.0 Mission Personnel: This section includes the cost of pay and allowances of officer, enlisted, and civilian personnel required to operate, maintain, and support an operational system or deployable unit. Based on a composite rate, mission personnel costs include the following:

- Basic pay
- Retired pay accrual
- Incentive pay
- Special pay
- Basic allowance for quarters
- Variable housing allowance
- Basic allowance for subsistence

- Hazardous duty pay
- Reenlistment bonuses
- Family separation allowances, etc.

2.0 Unit-Level Consumption includes the following:

- Cost of fuel and energy resources
- Operations, maintenance, and support materials consumed at the unit level
- Stock fund reimbursements for depot-level repairables
- Munitions expended in training
- Transportation in support of deployed unit training
- TAD/TDY pay
- Other costs such as purchased services

3.0 Intermediate Maintenance includes the cost of labor and materials and other costs expended by designated activities/units in support of a primary system and associated support equipment

- Calibration, repair, and replacement of parts, components, or assemblies
- Technical assistance

4.0 Depot Maintenance includes the cost of labor, material, and overhead incurred in performing major overhauls or maintenance on a defense system, its components, and associated support equipment at centralized repair depots, contractor repair facilities, or on site by depot teams.

- Usually portrayed on an annual basis

5.0 Contractor Support includes the cost of contractor labor, materials, and overhead incurred in providing all or part of the logistics support to a weapon system, subsystem, or associated support equipment.

6.0 Sustaining Support includes the cost of replacement support equipment, modification kits, sustaining engineering, software maintenance support, simulator operations. War readiness material is specifically excluded.

7.0 Indirect Support includes the costs of personnel support for specialty training, permanent changes of station, medical care, base operating support and real property.

This concludes describing the seven distinct areas found in the CES from Figure 3.6.

Summary

In the first half of this chapter, we discussed the non-DOD acquisition process, in contrast to the DOD acquisition process described in Chapter 2. The primary means to do so is by using the 12 step process found in the GAO Cost Estimating and Assessment Guide, and we covered each step in detail. We then discussed several non-DOD organizations who practice the art and science of cost estimating, including the Intelligence Community, NASA, the FAA, commercial firms, and FFRDC's who support the United States government with scientific and technical expertise. Many of them use the Cost Estimating Book of Knowledge (CEBoK), written by, and available at, the International Cost Estimating and Analysis Association (ICEAA) website. The remainder of the chapter then covered the numerous terminologies and concepts used in the cost estimating discipline. We described the types of cost estimates that are created, expounded upon concepts such

as CAIV and the "P, D, T" criteria, and then reviewed the document used as the basis for all cost estimates called the CARD. We finished the chapter by describing the three recognized methodologies used in cost estimating, and with a discussion of work breakdown structures and cost element structures, including the MIL-STD 881C guide. Now that we understand the processes and some of the terminologies used in cost estimating, Chapter 4 will describe the many resources available to help you find the data you need to develop your estimate, including several websites and reports useful in the field of cost estimating.

References

1. "GAO Cost Estimating and Assessment Guide", Richey/Cha/Echard, March 2009. http://www.gao.gov/assets/80/77175.pdf

2. Intelligence.Gov website. http://www.intelligence.gov/mission/member-agencies.html

3. Office of the Director of National Intelligence website. http://www.dni.gov/index.php/intelligence-community/ic-policies-reports/intelligence-community-directives.

4. Intelligence Community Directive 109 (ICD 109), Independent Cost Estimates, page 1.

5. NASA.gov website. http://www.nasa.gov/offices/ooe/CAD.html

6. FAA website, www.faa.gov.about/

7. FAA website Methodology Topics: Cost Analysis. http://www.ipa.faa.gov/Tasks.cfm?PageName=Cost%20Analysis

8. ICEAA website. http://www.iceaaonline.com/

9. National Science Foundation website. http://www.nsf.gov/statistics/nsf05306/

10. Private note from Dr. David Nicholls, chief of Cost Analysis Division at the Institute for Defense Analysis to Dr Dan Nussbaum, January 2014.

11. The Institute for Defense Analysis (IDA) website. https://www.ida.org/

12. Mitre website: http://www.mitre.org

13. RAND website. http://www.rand.org.

14. DAU website. https://acc.dau.mil/CommunityBrowser.aspx?id=347875

15. MIL-STD 881C 3 Oct 2011.pdf

Applications and Questions:

3.1 The GAO 12 step process can be found in what publication?

3.2 NASA, the FAA and many commercial firms have organized cost divisions, and many use CEBoK® for their official cost estimating training course material. (T/F)

3.3 When commencing a cost estimate, what four steps are pertinent guides in the general cost estimating process?

3.4 Considering the P, D, T criteria: If you want to field a system quickly and you do not want to pay a lot of money for it, you can expect the system's performance to be high or low?

3.5 The melding of the POE with the ICE produces what document?

3.6 What are the three recognized methodologies for cost estimating?

3.7 Which document provides the system description and is considered the basis upon which the system cost will be estimated, including physical and performance characteristics, plus development, production and deployment schedules?

3.8 The WBS is used in what phase(s) of a programs life cycle?

3.9 The CES is used in what phase(s) of a programs life cycle?

Data Sources

4.1 Introduction

In the previous chapter, we learned about the non-DoD acquisition process, the cost estimating process and some of the additional key terminologies used in this career field. Now that we understand *why* we do cost estimating, what do you do once you are assigned a cost estimate? One of the first things you will need to do is to find the data you need to develop your estimate. So, where is data available to you? This chapter will explain the many resources that you can use from which to acquire data and we will also guide you through the necessary steps in the data collection process: what to collect and from where plus data considerations, including problems that you may encounter with your data or data sources. We will cover the numerous cost databases available on the Internet and discuss the primary significant cost reports that are required by both contractors and program managers. One of the key cost reports is the Contract Performance Report (CPR), which contains a method for the program manager to monitor his/her program's progress and performance using an analysis system called Earned Value Management (EVM), which we will discuss in some detail. The overall purpose of this chapter is to guide you in the proper directions to search for the data that you need to create a reasonable and credible cost estimate.

4.2 Background and Considerations to Data Collection

Data collection is typically one of the most difficult, costly, and time consuming activities in cost estimating. There are many reasons for this:

- It is not always clear what data you need at the beginning of your study
- Your data requirements will most likely change and evolve
- It can be hard to gain access to some data sources
- Your data sources are sometimes difficult to use or sort

- The data that you find may not be exactly what you need
- There is typically a "story" behind the data that's important to understand
- There may be significant adjustments to the data necessary in order for it to support your needs
- Data may be missing (note: this point highlights the importance of documenting your work)

To collect the data that you need, select systems that are *relevant* to the system that you are costing. Choose analogous systems or components based upon elements identified and defined in your work breakdown structure (WBS). Typical cost drivers (those variables that *most* affect the cost) include both *physical* and *performance* characteristics.

- *Physical characteristics*: These characteristics include weight, volume, the number of holes drilled, the number of parts to assemble, the composition of materials, etc.
- *Performance characteristics*: These characteristics include power, thrust, bandwidth, range, speed, etc.

You will need to identify relevant historical cost, technical, and programmatic data that needs to be collected. These include:

- Program schedule, development quantity, production quantity
- Physical and performance data from operating manuals (such as NATOPS), manufacturer's specifications, and test data
- Improvements in technology are also an extremely important consideration. Measures of technology include percentage of composite material and radar cross section, among others

Where is this data found? Based on the difficulties of finding appropriate data to underpin your analysis, the short (and somewhat facetious) answer to this question is "Anywhere you can!" There are, however, data sources available to the analyst and the purpose of this chapter is to help you become aware that these sources exist. Numerous website examples (all with different purposes) include:

- The three so-called "Service Cost Agencies" associated with the Navy, Air Force and the Army:
  - Naval Center for Cost Analysis (NCCA)
  - Air Force Cost Analysis Agency (AFCAA)
  - Deputy Assistant Secretary of the Army for Cost and Economics (DASA-CE)
- Defense Cost and Resource Center (DCARC) at OSD CAPE
- Operating and Support Cost Data Bases for the Navy/Marine Corps, Air Force, and the Army:
  - Visibility and Management of Operation and Support Costs (VAMOSC, US Navy)
  - Air Force Total Ownership Cost (AFTOC, US Air Force)
  - Operating and Support Management Information System (OSMIS, US Army)

- Defense Acquisition Management Information Retrieval (DAMIR) within OSD and maintained by USD (AT&L)

Data can also be found at the Program Office of the program about which you are seeking information. Other sites to consider include the facilities for the prime contractor and their subcontractors, as well as past DoD, Department of Homeland Security (DHS), or Army Corps of Engineers projects.

In addition, you can look at previous reports, such as:

- Contractor Accounting Records
- Contractor Cost Data Reports (CCDR)
- Contract Performance Reports (CPR)
- Software Resource Data Reports (SRDR)
- Selected Acquisition Reports (SAR)
- Cost proposals/bids, or other sources within industry and government

Lastly, you can check catalog prices from places like RadioShack and engineering and parts warehouses as sanity checks to a cost estimate on a particular item. We will cover a number of these data sources mentioned earlier in detail in this chapter.

When collecting data, the data that we need will be both *quantitative* and *qualitative*. For example, cost data is generally quantitative in nature and so is the required technical data. However, programmatic data is generally qualitative in nature. Program risk data can be either quantitative or qualitative.

4.2.1 COST DATA

Cost data is data collected specifically to support the development of an estimate for a particular cost element. Examples include:

- Physical and performance characteristics of systems, subsystems and components
- Products such as propulsion systems and airframes and valves
- Functional areas such as engineering and program management
- Activities such as intermediate maintenance and depot level repair
- Information supporting normalization of analogous cost data, such as quantities, fiscal year references, and content differences.

4.2.2 TECHNICAL DATA

Technical data is data collected from analogous systems. Examples include:

- Systems of similar size
- Systems with similar performance standards
- Systems with similar technology
- Consumption rates of analogous systems
- Operating personnel levels of analogous systems

- Material quantities used in previous systems
- Logistics information and details from pervious systems

4.2.3 PROGRAMMATIC DATA

Programmatic data is higher level data collected at the program level that has applicability to multiple elements in the LCCE. The focus is on program planning documentation, not necessarily cost and technical data. Examples include:

- Cost or financial data in the form of prior year expenditures (sunk costs)
- Future year budget profiles
- Performance goals and strategies, operating concepts
- The Acquisition Environment: Is this a sole source contract or a competitive bid among a few contractors?
- The Acquisition Schedule: Are we buying all items in one lot or will it be a long production run?

4.2.4 RISK DATA

Risk data comes from all sources, and is collected concurrently with cost, technical and programmatic data. Risks from one type of data can impact, or have a correlation to, other types of data.

Once you receive these four types of data, you need to ensure that you consider these three steps:

1. Review all of your data collected to ensure homogeneity. You need to ensure that you are working with standard quantities, constant year dollars, and adequate coverage of all the WBS elements that you need considered. Additionally, there needs to be at least a rough analogy between the complexity of the historical systems and the complexity of the proposed system.

2. Allocate your data to the proper WBS elements. Organize your data on a consistent basis from system to system, contractor to contractor, WBS element to WBS element, etc. Ideally, you would like to distinguish between recurring and nonrecurring costs, direct and indirect costs, support costs, and profit as well.

3. Identify problems with your data, and if problems exist, do your best to resolve those problems! Problems that you may encounter with your data or data sources include:

   a. Differences in categories or accounting methodologies among contractors

   b. The information is provided in the wrong format

   c. The information is in different units (pounds vs. kilograms, or peak power vs. average power)

   d. Gaps in the data

   e. Differences in types of program

   f. Sometimes previous program managers did not document correctly, or did not "buy" the data from the contractor, so it is incomplete

4. Program changes can (and usually do) occur over time, too, such as manufacturing methods and technological changes and/or increases in capabilities. Other problems

include major failures in the development and testing phase, or a strike by the contractor work force causing a break in production. The effects of a break in production are covered in Production Breaks/Lost Learning in Chapter 12.

4.3 Cost Reports and Earned Value Management (EVM)

There are many, many cost reports that need to be made throughout the course of a program. Fortunately, we are not going to cover them all here, but we will focus on the most important ones! Cost reports have two fundamental purposes: one for prospective (that is, forward-looking) analysis and the other for retrospective (that is, backward-looking) analysis. In the prospective mode, these cost reports are a *source of cost data* and some of the major reports are the primary databases used for cost estimating. In the retrospective mode, cost reports are used for *cost monitoring*. They can help to provide an early warning of cost growth, which will hopefully allow a program manager ample time to be able to mitigate the risks, re-baseline the program, fix the program, or perhaps abandon it early enough to avoid major losses of investment capital.

There are two major contract management reports that come from the prime contractor:

- Contractor Cost Data Report (CCDR), and
- Contract Performance Report (CPR)

There are two major Cost, Schedule, and Performance reports from the program manager that must go to others:

- Selected Acquisition Report (SAR), and the
- Defense Acquisition Executive Summary (DAES)

We will now cover these reports in detail. First we will discuss the two reports that come from the prime contractor to the program manager and that are also provided to DCARC (to be explained later in this chapter).

4.3.1 CONTRACTOR COST DATA REPORTING (CCDR)

The CCDR is the DoD's primary means of collecting periodic and time-phased data on costs incurred by DoD contractors. The CCDR is *critical* to establish reasonable cost estimates as it provides historical costs from the prime contractors accounting system ("actuals)." These actuals are displayed by major work breakdown structure elements, which permit the cost estimator to use this analogous data to develop the cost estimate for the program at hand. While the CCDR provides detailed costs, it is not really a cost monitoring system. That task falls to the Cost Performance Report.

The CCDR is comprised of four major sections, which are all Excel based. The four sections are as follows:

- DD Form 1921: Cost Data Summary Report
- DD Form 1921-1: Functional Cost Hour Report

- DD Form 1921-2: Progress Curve Report
- DD Form 1921-3: Contract Business Base Report

DD Form 1921: Cost Data Summary Report. This form displays the nonrecurring and recurring costs, both "To Date" and "At Completion," for all WBS elements that are included in the program's CSDR plan. It also includes contract totals and indirect costs such as general and administrative costs, undistributed budget totals, management reserve, facilities, capital cost of money, and profit/fee.

DD Form 1921-1: Functional Cost Hour Report. This form reports select WBS elements, and also breaks them out in the nonrecurring and recurring cost "To Date" and "At Completion" format. It includes detailed breakout of all resource data including direct labor hours, direct labor dollars, material, and overhead dollars. It reports by four functional work categories: engineering, tooling, quality control, and manufacturing, as well as displaying the price from the direct-reporting subcontractors.

DD Form 1921-2: Progress Curve Report. This form reports select hardware WBS elements reported by unit cost or by lot cost. It provides direct recurring costs and hours "To Date," a detailed breakout of direct costs, and it also reports by the four functional work categories: engineering, tooling, quality control, and manufacturing. This report and its data are used for modeling learning curves and for projecting future units.

DD Form 1921-3: Contract Business Base Report. This form is an annual report that contains aggregate figures on other business that is occurring at the contractor's site. This report and its data are used to facilitate the overhead cost analysis at a specific contractor's site.

The CCDR is indeed one of the most necessary and widely used cost reports, as it establishes historical costs from the prime contractor's accounting system for the primary areas in a program. Figure 4.1 and Figure 4.2 show a copy of two of the forms in a CCDR for visual familiarity: DD Form 1921 and DD Form 1921-1.

The second report from the contractor to the PM is the Contract Performance Report.

4.3.2 CONTRACT PERFORMANCE REPORT (CPR)

The CPR is primarily for the program manager and provides extensive information in the area of Earned Value Management (EVM) as it relates to this particular program. It is used to obtain individual contract cost and schedule performance information from the contractor. This information is intended for the Program Manager's use in determining the financial and schedule health of his/her program. It is also used for making and validating program management decisions, and it provides early indicators of contract cost and schedule problems, as well as the effects of management action previously taken to resolve such problems.

The general CPR process uses what is known as the "Earned Value Management (EVM)" analysis to help the PM monitor the progress and performance of the current contract and the program. It is also used to ensure that the contractor's internal cost and schedule control systems are sound and are producing valid, timely progress information.

The CPR allocates the program's budget to WBS elements. Contractors are required to plan their workload in accordance with detailed work packages, and EVM provides the

COST DATA SUMMARY REPORT

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 8 hours per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Executive Services Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR COMPLETED FORM TO THE ABOVE ORGANIZATION.

1. MAJOR PROGRAM		a. NAME:								

b. PHASE/MILESTONE
☐ Pre-A   ☐ B   ☐ C-FRP
☐ A   ☐ C-LRIP   ☐ O&S

2. PRIME MISSION PRODUCT

3. REPORTING ORGANIZATION TYPE
☐ PRIME / ASSOCIATE CONTRACTOR   ☐ DIRECT-REPORTING SUBCONTRACTOR   ☐ GOVERNMENT

4. NAME/ADDRESS (Include ZIP Code)
a. PERFORMING ORGANIZATION   b. DIVISION

5. APPROVED PLAN NUMBER

6. CUSTOMER (Direct-reporting subcontractor use only)

7. CONTRACT TYPE

8. CONTRACT PRICE

9. CONTRACT CEILING

10. TYPE ACTION
a. CONTRACT NO.:
b. LATEST MODIFICATION:
c. SOLICITATION NO.:
d. NAME:
e. TASK ORDER/DELIVERY ORDER/LOT NO.:

11. PERIOD OF PERFORMANCE
a. START DATE (YYYYMMDD):
b. END DATE (YYYYMMDD):

12. APPROPRIATION
☐ RDT&E
☐ PROCUREMENT
☐ O&M

13. REPORT CYCLE
☐ INITIAL
☐ INTERIM
☐ FINAL

14. SUBMISSION NUMBER

15. RESUBMISSION NUMBER

16. REPORT AS OF (YYYY MMDD)

17. NAME (Last, First, Middle Initial)

18. DEPARTMENT

19. TELEPHONE NUMBER (Include Area Code)

20. EMAIL ADDRESS

21. DATE PREPARED (YYYY MMDD)

WBS ELEMENT CODE A	WBS REPORTING ELEMENTS B	NUMBER OF UNITS TO DATE C	COSTS INCURRED TO DATE (thousands of U.S. Dollars)			NUMBER OF UNITS AT COMPLETION G	COSTS INCURRED AT COMPLETION (thousands of U.S. Dollars)		
			NONRECURRING D	RECURRING E	TOTAL F		NONRECURRING H	RECURRING I	TOTAL J

22. REMARKS

DD FORM 1921, 20110518	PREVIOUS EDITION IS OBSOLETE.	SECURITY CLASSIFICATION Unclassified

FIGURE 4.1 DD Form 1921: Cost Data Summary Report.

FUNCTIONAL COST-HOUR REPORT

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 16 hours per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Executive Services Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR COMPLETED FORM TO THE ABOVE ORGANIZATION.

1. MAJOR PROGRAM   a. NAME:

b. PHASE/MILESTONE	**2. PRIME MISSION PRODUCT**	**3. REPORTING ORGANIZATION TYPE**	**4. NAME/ADDRESS**(Include Zip Code)	**5. APPROVED PLAN NUMBER**
☐ Pre-A ☐ B ☐ C-FRP ☐ A ☐ C-LRIP ☐ O&S		☐ PRIME / ASSOCIATE CONTRACTOR ☐ DIRECT-REPORTING SUBCONTRACTOR ☐ GOVERNMENT	a. PERFORMING ORGANIZATION b. DIVISION	

6. CUSTOMER (Direct-Reporting Subcontractor Use Only)

7 TYPE ACTION
a. CONTRACT NO.:
b. LATEST MODIFICATION: c. SOLICITATION NO.:   e. TASK ORDER/DELIVERY ORDER/LOT NO.:
                                           d. NAME:

8. PERIOD OF PERFORMANCE
b. END DATE *(YYYYMMDD)*:
a. START DATE *(YYYYMMDD)*:

9. REPORT CYCLE
☐ INITIAL
☐ INTERIM
☐ FINAL

10. SUBMISSION NUMBER

11. RESUBMISSION NUMBER

12. REPORT AS OF (YYYYMMDD)

13. NAME (Last, First, Middle Initial)

14. DEPARTMENT

15. TELEPHONE NO. (Include Area Code)

16. EMAIL ADDRESS

17. DATE PREPARED (YYYYMMDD)

18. WBS ELEMENT CODE

19. WBS REPORTING ELEMENT

20. NUMBER OF UNITS
a. TO DATE   b. AT COMPLETION

21. APPROPRIATION
☐ RDT&E
☐ PROCUREMENT
☐ O&M

FUNCTIONAL DATA ELEMENTS	COSTS AND HOURS INCURRED TO DATE (thousands of U.S. Dollars or thousands of hours)			COSTS AND HOURS INCURRED AT COMPLETION (thousands of U.S. Dollars or thousands of hours)		
	A. NONRECURRING	B. RECURRING	C. TOTAL	D. NONRECURRING	E. RECURRING	F.TOTAL
ENGINEERING						
(1) DIRECT ENGINEERING LABOR HOURS						
(2) DIRECT ENGINEERING LABOR DOLLARS						
(3) ENGINEERING OVERHEAD DOLLARS						
(4) TOTAL ENGINEERING DOLLARS						
MANUFACTURING OPERATIONS						
(5) DIRECT TOOLING LABOR HOURS						
(6) DIRECT TOOLING LABOR DOLLARS						
(7) DIRECT TOOLING & EQUIPMENT DOLLARS						
(8) DIRECT QUALITY CONTROL LABOR HOURS						
(9) DIRECT QUALITY CONTROL LABOR DOLLARS						
(10) DIRECT MANUFACTURING LABOR HOURS						
(11) DIRECT MANUFACTURING LABOR DOLLARS						
(12) MANUFACTURING OPERATIONS OVERHEAD DOLLARS *(Including Tooling and Quality)*						
(13) TOTAL MANUFACTURING OPERATIONS DOLLARS *(Sum of rows* 6, 7, 9, 11, and 12)						
MATERIALS						
(14) RAW MATERIAL DOLLARS						
(15) PURCHASED PARTS DOLLARS						
(16) PURCHASED EQUIPMENT DOLLARS						
(17) MATERIAL HANDLING OVERHEAD DOLLARS						
(18) TOTAL DIRECT-REPORTING SUBCONTRACTOR DOLLARS						
(19) TOTAL MATERIAL DOLLARS						
OTHER COSTS						
(20) OTHER COSTS NOT SHOWN ELSEWHERE (Specify in Remarks)						
SUMMARY						
(21) TOTAL COST *(Direct and Overhead)*						

22. REMARKS

DD FORM 1921-1, 20110518

FIGURE 4.2 DD Form 1921-1: Functional Cost Hour Report.

capability to monitor these work packages in a time-phased way. It tracks not only the "Current Time Period" (i.e., this month or quarter), but it also then adds progress to a "Cumulative to Date" tracking for the program. There are numerous metrics used in the EVM system, but the primary three metrics are the following (Note: all three of these metrics are in dollar or hour units):

Budgeted cost of work scheduled (BCWS): This is the amount budgeted for the totality of the work packages that we *scheduled* to accomplish in a certain period of time.

Budgeted cost of work performed (BCWP): This is the amount budgeted for the work that we actually *accomplished* in that period of time.

Actual cost of work performed (ACWP): This is what it actually *cost* us to accomplish the work while completing these work packages.

Once these three metrics are known, the PM can now calculate the variance in both schedule and cost:

- Cost Variance: CV = BCWP – ACWP
- Schedule Variance: SV = BCWP – BCWS

Cost variance is a measure of how much more (or less) the cost is from what was originally planned. Schedule variance is a measure of how much more (or less) work has been accomplished in a certain time frame from what was originally planned. You can also convert this variation to time units by dividing by $/time. Additionally, the Cost Performance Index (CPI) and the Schedule Performance Index (SPI) are efficiency indices of the cost and schedule performance indicators. These are important because without them, it is difficult to compare projects of different sizes to one another. For example, you cannot tell whether a $1M variance is good or bad. In a $1B program, a variance of only $1M is very good, but for a $2M program, a variance of $1M is *not* good! This concept is similar to that of standard deviation, which will be discussed in more detail in Chapter 6 on Statistics. CPI and SPI are defined as follows:

- Cost Efficiency: CPI = BCWP / ACWP Favorable is > 1.0, Unfavorable is < 1.0
- Schedule Efficiency: SPI = BCWP / BCWS Favorable is > 1.0, Unfavorable is < 1.0

Note that for CPI, if the budgeted cost for a task was greater than what the actual cost for that task was, you will calculate a CPI of greater than 1.0, and this is desirable. If the CPI is less than 1.0, this implies that the actual cost for a task was greater than the budgeted cost for that task, thus implying a cost overrun for that task. Similarly, when considering the program schedule, SPI calculates whether you are ahead or behind schedule time-wise by using similar principles.

A sample copy of a CPR is displayed in Figure 4.3. Note the areas allocated for BCWS, BCWP, ACWP, SPI, and CPI for both "Current Period" and "Cumulative to Date," as well as Budgeted Cost and Estimated Cost at Completion and their Variance.

Now let's illustrate an EVM scenario by considering the following short example.

INTEGRATED PROGRAM MANAGEMENT REPORT
FORMAT 1 – WORK BREAKDOWN STRUCTURE   DOLLARS IN _____

PENDING
UPDATE TO OMB No.0704-0188

The public reporting burden for this collection of information is estimated to average 1.7 hours per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports, (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THIS ADDRESS. SUBMIT COMPLETED FORMS IN ACCORDANCE WITH CONTRACTUAL REQUIREMENTS.

1. CONTRACTOR	2. CONTRACT	3. PROGRAM	4. REPORT PERIOD
a. NAME	a. NAME	a. NAME	a. FROM (YYYYMMDD)
b. LOCATION (Address and ZIP Code)	b. NUMBER	b. PHASE	b. TO (YYYYMMDD)
	c. TYPE d. SHARE RATIO	c. EYMS ACCEPTANCE No YES (YYYYMMDD)	

5. CONTRACT DATA

a. QUANTITY	b. NEGOTIATED COST	c. ESTIMATED COST OF AUTHORIZED UNPRICED WORK	d. TARGET PROFIT/FEE	e. TARGET PRICE	f. ESTIMATED PRICE	g. CONTRACT CBLING	h. ESTIMATED CONTRACT CBLING	I. DATE OF OTBOTS (YYYYMMDD)

6. ESTIMATED COST AT COMPLETION   **7. AUTHORIZED CONTRACT OR REPRESENTATIVE**

	MANAGEMENT ESTIMATE AT COMPLETION (1)	CONTRACT BUDGET BASE (2)	VARIANCE (3)	a. NAME (Last, First, Middle, Initial)	b. TITLE
a. BEST CASE				c. SIGNATURE	d. DATE SIGNED (YYYYMMDD)
b. WORST CASE					
c. MOST LIKELY					

8. PERFORMANCE DATA

ITEM (1)	CURRENT PERIOD					CUMULATIVE TO DATE					REPROGRAMMING ADJUSTMENTS			AT COMPLETION		
	BUDGETED COST		ACTUAL COST	VARIANCE		BUDGETED COST		ACTUAL COST	VARIANCE		COST VARIANCE	SCHEDULE VARIANCE	BUDGET	BUDGETED	ESTIMATED	VARIANCE
	WORK SCHEDULED (2)	WORK PERFORMED (3)	WORK PERFORMED (4)	SCHEDULE (5)	COST (6)	WORK SCHEDULED (7)	WORK PERFORMED (8)	WORK PERFORMED (9)	SCHEDULE (10)	COST (11)	(12a)	(12b)	(13)	(14)	(15)	(16)
a. WORK BREAKDOWN STRUCTURE ELEMENT																
b. COST OF MONEY																
c. GENERAL AND ADMINSTRATIVE																
d. UNDISTRIBUTED BUDGET																
e. SUBTOTAL (PERFORMANCE MEASUREMENT BASELINE)																
f. MANAGEMENT RESERVE																
g. TOTAL																
9. RECONCILIATION TO CONTRACT BUDGET BASE																
a. VARIANCE ADJUSTMENT																
b. TOTAL CONTRACT VARIANCE																

UPDATED FROM DD FORM 2734/1, MAR 05, PENDING APPROVAL   LOCAL REPRODUCTION AUTHORIZED.

FIGURE 4.3 Cost Performance Report.

4.3.3 EVM EXAMPLE

Ringo Chinoy Enterprises has won the contract for producing the third generation HMMWV (the "Hummer") for the US Army. On October 1st, the following six work packages were scheduled to be completed for that month.

(1) Avionics: $15k	(2) Armor: $5k
(3) Drive Train: $30k	(4) Suspension: $20k
(5) Steering: $25k	(6) Stereo: $15k

With this given information, calculate the Budgeted Cost for Work Scheduled (BCWS):

- (Answer: summing up the costs for each of the six work packages, we calculate that BCWS = $110k)

At the end of October, Ringo huddled with his managers to determine what they had accomplished during the month, and what still needed to be done, if anything. The results of this meeting found that the first four work packages had been completed, but that the final two work packages were not completed (i.e., steering and stereo). With this given information, calculate the Budgeted Cost for Work Performed (BCWP):

- (Answer: summing up the costs for just the first four work packages that were completed, we calculate that BCWP = $70k)

Comparing the BCWS and BCWP, calculate the Schedule Variance (BCWP – BCWS). How far behind schedule are they in terms of budget? How far behind are they in terms of schedule?

- Answer: SV = BCWP − BCWS = $70k − $110k = −$40k in terms of budget. Note that the negative sign signifies being behind schedule. In terms of time, we divide SV/BCWS, and find that −$40k / $110k = −0.3636 months behind. This equates to −0.3636 months × 30 days / month = 11 days behind schedule.

Ringo then calculated how much it *actually* cost his company to accomplish the four work packages that had been completed, and it turned out to be $95k. Now calculate the Cost Variance (BCWP – ACWP).

- Answer: CV = BCWP – ACWP = $70k − $95k = −$25k.

Summary: The following can be concluded from the aforementioned calculations:

- Ringo thought that they would be able to complete $110K worth of work in the month of October.
- In reality, his company only completed $70k worth of work. Thus, after this month, they are $40k behind in terms of work scheduled and eleven days behind in terms of time.
- The amount of work that was completed was expected to cost $70k to complete. Instead, it actually cost $95k to complete that amount of work. Thus, they are $25k behind in terms of cost.
- CPI = BCWP/ACWP = $70k/$95k = 0.74, which is undesirable since the index is less than 1.0. More is being spent on the program than was planned, an indication of "cost inefficiency."
- SPI = BCWP/BCWS = $70k/$110k = 0.64, which is also undesirable since the index is less than 1.0. The program is taking longer to complete than was planned, an indication of "schedule inefficiency."
- All of these metrics/numbers would then be input onto the CPR, Figure 4.3, into the "Current Time Period." After the following month using the new calculations for BCWS, BCWP, etc., totals would then be summed and inputted into the "Cumulative to Date" area, as well as the Current Time Period. This is where a PM can then track whether his/her program is ahead or behind in terms of cost and schedule.

This simple example illustrates how EVM is used to track program progress, and it is used extensively throughout the acquisition process and by numerous organizations. One significant source for information on EVM can be found at the Defense Acquisition University (DAU) website and is called the "Gold Card." The Gold Card provides an excellent overview of EVM, its terminology, program variance metrics, and the equations that are most used in EVM. The front side of the Gold Card is shown here in Figure 4.4 [1].

The back side of the Gold Card predominantly shows EVM acronyms. For further information and for in-depth training on EVMS, visit the ICEAA, DAU, or DOE websites.

EARNED VALUE MANAGEMENT

'GOLD CARD'

Total Allocated Budget

Management Reserve

PMB

EAC

BAC

ScheduleVariance

ACWP$_{Cum}$

Cost Variance

Projected Slip?

$

BCWS$_{Cum}$

BCWP$_{Cum}$

Time

Time Now

Completion Date
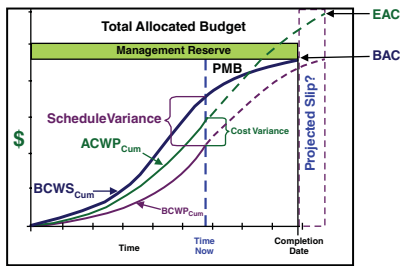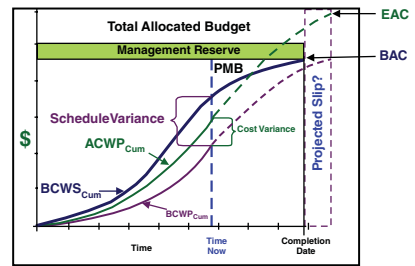
VARIANCES Positive is Favorable, Negative is Unfavorable

OVERALL STATUS

Cost Variance	CV	=	BCWP – ACWP
	CV%	=	(CV / BCWP) *100
Schedule Variance	SV	=	BCWP – BCWS
	SV%	=	(SV / BCWS) * 100
Variance at Completion	VAC	=	BAC – EAC
	VAC%	=	(VAC / BAC) *100

% Schedule = (BCWS$_{CUM}$ / BAC) * 100
% Complete = (BCWP$_{Cum}$ / BAC) * 100
% Spent = (ACWP$_{CUM}$ / BAC) * 100

EFFICIENCIES

Cost Efficiency   CPI = BCWP / ACWP   Favorable is > 1.0, Unfavorable is < 1.0
Schedule Efficiency SPI = BCWP / BCWS   Favorable is > 1.0, Unfavorable is < 1.0

BASELINE EXECUTION INDEX (BEI) & Hit Task %

BEI = Total Tasks Completed / (Total Tasks with Baseline Finish On or Prior to Current Report Period)

Hit Task% = 100 * (Tasks Completed ON or PRIOR to Baseline Finish / Tasks Baselined to Finish
within Current Report Period)

ESTIMATE @ COMPLETION = ACTUALS TO DATE + [(REMAINING WORK) / (PERFORMANCE FACTOR)]

EAC$_{CPI}$ = ACWP$_{Cum}$ + [(BAC – BCWP$_{Cum}$) / CPI$_{Cum}$]

EAC$_{Composite}$ = ACWP$_{Cum}$ + [(BAC – BCWP$_{Cum}$) / (CPI$_{Cum}$* SPI$_{Cum}$)]

TO COMPLETE PERFORMANCE INDEX (TCPI) § #

TCPI$_{Target}$ = Work Remaining / Cost Remaining = (BAC – BCWP$_{CUM}$) / (Target – ACWP$_{Cum}$)

§ To Determine the TCPI $_{for BAC, LRE, or EAC}$ Substitute TARGET with BAC, LRE, orEAC

To Determine the Contract Level TCPI forEAC, You May Replace BAC with TAB

FIGURE 4.4 Front Side of the DAU Gold Card.

Having discussed the two primary reports that come from the contractor to the PM, let's now turn to what reports the PM must accomplish once the CCDR and CPR are received. There are two major Contract Management reports that come from the program manager to others that we will discuss. They are the Selected Acquisition Report (SAR) and the Defense Acquisition Executive Summary (DAES).

- *Selected Acquisition Report (SAR)*: This report is prepared for submission by the program manager to Congress. SARs are an annual requirement for all MDAPs, and they are submitted in conjunction with the submission of the President's budget. The SAR provides the status of *total* program cost, schedule, and performance, including R&D, procurement, military construction, and O&S costs, as well as program unit cost information. It also includes a full "life cycle" analysis. *Quarterly reports are required when there has been an increase of 15% or more in baseline unit costs or there is a delay of six months or more in completing an acquisition milestone.* Variances between planned and current costs and schedule must be explained in terms of seven cost growth categories. These seven cost growth categories are economic changes, quantity changes, schedule changes, engineering changes, estimating changes, support changes, and other changes. The SAR evaluates a number of terms, such as the Acquisition Program Baseline (APB), the Program Acquisition Unit Cost (PAUC), and the Average Procurement Unit Cost (APUC), all key terms in an acquisition program.

- *Defense Acquisition Executive Summary (DAES)*: This report is prepared by the PM for internal DoD control and is required only if assigned to do so. The USD (AT&L) designates the ACAT I programs that must submit this report. Reports are designed to provide advance indications of both potential and actual program problems before they become significant. These reports are internal to DoD and are distributed from the Program Manager to USD (AT&L) and other OSD independent cost assessment offices, as necessary.

  The importance of the DAES is in recognizing that problems are expected to surface in these programs and that these reports will aid in communication and early resolution. They must be submitted quarterly, if assigned. "Out-of-cycle" or additional reports are required when unit cost thresholds are exceeded or excess unit costs are anticipated. The DAES presents total costs and total quantities for all years through the end of the acquisition phase. It encompasses DoD component quantity and cost projections for the total program. These projections cover the total program over the full program's life cycle. Thus, they are not limited to the amount funded in the current budget or to the total amount budgeted and programmed through the FYDP.

  Presentation of the total program is intended to provide a comprehensive understanding of total program requirements and performance. It displays a "traffic light" style assessment of performance using red, yellow, and green colors to immediately highlight how well (or not) an area of a program is doing. A "Green" highlight means that that particular area of the program is doing well, usually in the sense of being on track in cost, schedule, and performance; "Yellow" means that there are potential problems in that area; and "Red" indicates that the area highlighted is not doing well. Areas include cost, schedule, and performance as a minimum.

We have now discussed four of the major reports required by both the contractor and the PM. While there are many other reports (and we could fill up a few chapters on those reports), here is a short synopsis on just a few of the other reports, for you to get a feel for what types of program information are required at various times:

- *Contract funds status report (CFSR).* This report is used to obtain funding data on contracts in excess of six months duration. It is used to assist DoD components in several areas of fund management:
  - Developing funding requirements and budget estimates for that program
  - Updating and forecasting contract fund requirements
  - Planning and decision making on possible funding changes
  - Determining funds in excess of contract needs that may be available for de-obligation
- *Unit cost reports (UCR)*: This report is prepared for all acquisition programs for which SARs are submitted. They begin with submission of the initial SAR and terminate with the submission of the final SAR. These reports are submitted quarterly by the program manager to the component acquisition executive as part of the DAES submission. The following information is included in these reports:
  - A current estimate of either the program acquisition unit cost (PAUC) or the average unit procurement cost (AUPC)
  - Cost and schedule variances (in dollars) for major contracts
  - Any change to a program's schedule, milestones, or performance
- *Supplemental contractor cost report (SCCR)*: This report is a quarterly summary of CPR data prepared by the program manager for submission to OSD. The database format is specified by the Office of the Assistant Secretary of Defense (Comptroller) to facilitate preparation of automated Contractor Performance Measurement Analysis (CPMA). This allows analysts to identify contracts with negative variances or unfavorable trends, and also tracks management reserve, as well as cost and schedule trends. Contracts that are currently deviating by 10% or more, or which are *projected* to deviate by 10% or more, will require a special report.

So, as a cost analyst, now that you have learned about some of the important reports, where do I *find* all of these reports? The next section will cover where these reports are housed and how they can be found.

4.4 Cost Databases

The primary databases available to the government cost analyst include the following:

- *Defense Cost and Resources Center (DCARC)*: This website is owned by OSD-CAPE (DoD Level) and is used for the compilation of Research and Development and Procurement phase costs.
- *Operating and Support Costs Databases*: The US Army, the US Navy/Marine Corps, and the US Air Force each have their own O & S database.
- *Defense Acquisition Management Information Retrieval (DAMIR)*: This website is owned by OSD-AT&L (DoD Level) and is used for the compilation of Selected Acquisition Reports, SAR Baselines, Acquisition Program Baselines (APBs) and Assessments, and DAESs.

Let's look at and describe each of these websites individually.

4.4.1 DEFENSE COST AND RESOURCE CENTER (DCARC)

The DCARC is an organization within OSD-CAPE at the Pentagon that provides the larger cost community with its quality data needs. Their mission is to provide an accurate record of historical program cost data, within a searchable and easy-to-use system, to assist the government cost analyst in creating realistic budgets and credible cost estimates. The primary goals of DCARC include cost collection, planning, and the execution of the CSDR process. They seek to provide large scale data availability.

The DCARC database collects many different types of data and stores them in two separate systems. They contain predominantly those costs that are accrued in the *R&D and Procurement phases* (though as we write this they are commencing to gather O&S costs now, as well). The DCARC contains two main databases:

1. *Defense Automated Cost Information Management System (DACIMS)*, and the
2. *Earned Value Central Repository (EVCR)*.

- *Defense Automated Cost Information Management System (DACIMS)*: This database contains the Cost and Software Data Report (CSDR) Reporting System. The CSDR is a compilation of the Contractor Cost Data Reports (CCDRs) and the Software Resource Data Reports (CSDRs). Thus, Cost and Software Data Report (CSDR) = Contractor Cost Data Reports (CCDR) + Software Resource Data Reports (SRDR).

  Both the CCDRs and the CSDRs reside within the DACIMS. These data can be accessed via an account on the DCARC website. The CSDRs capture actual contractor-incurred cost data to provide the visibility and consistency needed to develop your credible cost estimate. They also capture the software resource data including size, effort, and schedule. These reports are required on all ACAT IA, IC, and ID programs.

- *Earned Value Central Repository (EVCR)*: This database contains the Earned Value Management data and is a compilation of the Contract Performance Reports previously discussed in Section 4.3.2. It contains all necessary information concerning EVM for each program. This database also contains the Integrated Master Plans (IMPs) and Integrated Master Schedules (IMSs) for each program.

4.4.2 OPERATING AND SUPPORT COSTS DATABASES

Operating and Support costs have received increasing attention over time. As discussed in Chapter 2, Section 2.5, the reason for this is because total O&S costs generally exceed total Research and Development and Procurement costs by a large amount, and they involve significant sums of money spent in the O&S phase of a program's life cycle. The need for O&S collection began in 1975 when the Deputy Secretary of Defense directed the Services to collect O&S costs, which was in response to Congressional criticism of the Services for their inability to predict and report O&S costs.

Correspondingly, and after many years, each of the uniformed services developed their own Operating and Support Costs (O&S) databases and all three can be found at the appropriate Service Cost Agency website. Since the US Marine Corps is part of the Department of the Navy, the Navy website also includes all of the US Marine Corps O&S costs. The three service O&S databases are named as follows:

- *Visibility and Management of Operation and Support Costs (VAMOSC, US Navy)*. This website is hosted and managed at the Naval Center for Cost Analysis (NCCA) website.

- *Air Force Total Ownership Cost (AFTOC, US Air Force).* The website is hosted at the Air Force Cost Analysis Agency (AFCAA) website and is managed by the Deputy Secretary of the Air Force for Cost and Economics (SAF/FMC).
- *Operating and Support Management Information System (OSMIS, US Army).* OSMIS is hosted and managed at the Deputy Assistant Secretary of the Army for Cost and Economics (ASA FMC) website.

While each of these databases is designed differently, they contain essentially the same type of information. The three websites are designed to provide a single source of data for their respective service, organized by a system, infrastructure, or category, which in turn significantly reduces the time needed for analysts to find pertinent data. They provide "one-stop shopping" for historical, current year, and FYDP cost information. Costs are in both Then-Year and Constant-Year dollars. The websites are used for use in programs concerning aviation, infrastructure, ships, USMC ground systems, weapons, military and civilian personnel, and logistics consumption, to name a few. The data from these websites and all of the websites that we are describing can generally be accessed by government personnel and government sponsored contractors.

4.4.3 DEFENSE ACQUISITION MANAGEMENT INFORMATION RETRIEVAL (DAMIR)

The final website that we will discuss is the DAMIR website, which is owned by OSD-AT&L. The primary goal of DAMIR is to streamline acquisition management and oversight by leveraging the capabilities of the net-centric environment. DAMIR identifies the various data sources that the Acquisition community uses to manage MDAP and MAIS programs, and provides a unified web-based interface through which to present that information. DAMIR is the authoritative source for Selected Acquisition Reports (SARs), SAR Baselines, Acquisition Program Baselines (APBs) and Assessments, and DAESs. DAMIR has both a Classified and an Unclassified website.

Summary

In this chapter, we discussed the data that you will need to conduct your estimating and analysis, and also where to find that data. We explained the many resources available to you to acquire data and also guided you through the necessary steps in the data collection process: what to collect, from where, and data considerations, including problems that you may encounter with your data or data sources. We discussed a few of the significant cost reports that are required by both contractors and program managers. The contractor must submit a CCDR and CPR to the PM; in return, the PM must issue a SAR and sometimes a DAES, if warranted. The CCDR provides historical costs from the prime contractor's accounting system ("actuals"), as well as major work breakdown structure elements to help determine the cost of the program, while the CPR contains the data that supports an analytical way for the Program Manager to monitor his/her program's progress and performance, using a system called Earned Value Management. EVM was discussed in detail, with an example provided as well. The chapter ended with a discussion of some of the primary websites that house the R&D, Production, and the O&S costs. The overall purpose of this chapter was to guide you in the proper directions to search for the data that you need to create a reasonable and credible cost estimate. The end of this chapter

also denotes a shift in the type of information in our book. We will now move away from the qualitative and background information used in cost estimating and commence with the quantitative portion of the textbook, starting with Data Normalization.

Reference

1. DAU website. "Gold Card"

Applications and Questions:

4.1 When collecting data, the data that you need will be both _____ and _____.

4.2 What are the four types of data we need to collect?

4.3 Each of the uniformed services has its own cost agency. Name these three:

4.4 The system calculated in the CPR used to help the PM monitor the progress and performance of his or her program is called?

4.5 What type of costs does the DCARC website primarily track, and name the two websites within DCARC?

4.6 Name the three service-specific websites that track Operating and Support costs.

4.7 Name the website that is the authoritative source for Selected Acquisition Reports (SARs).

Data Normalization

5.1 Introduction

In the previous chapter, we discussed Data Sources and where you would find the data that you will need once you are assigned a cost estimate. Once you get your data, however, what do you do with it?! That short question leads us to this lengthy chapter on Data Normalization. The reason this chapter contains such depth is because identification and normalization of cost data is one of the most challenging and perennial problems that will confront you as a cost analyst. Why? Because once you find and receive the data that you are seeking, it is rarely in the format or the base year that you need it in. Data normalization involves taking raw cost data and applying adjustments to that data to gain consistent, comparable data to be used in your estimates.

It is necessary to normalize your data in three ways: normalizing for content, quantity, and inflation. While we will cover normalizing for content and quantity, the majority of this chapter will be focused on normalizing data for inflation, which includes learning how to use the Joint Inflation Calculator (JIC). This chapter signifies a shift in information, in that the majority of background information needed in cost estimation has been discussed in the first four chapters. Chapter 5 is the commencement of the numerous quantitative areas of cost estimation.

5.2 Background to Data Normalization

Since historic cost data involves the purchasing of goods and services in different time periods, we need to know how to compare the dollar cost of goods and services in one period with the dollar cost of comparable goods and services in another period. For example, if you purchased an F-18 fighter aircraft (let's call it aircraft #1) in the year 2000 for $50 million, and then you purchased another F-18 (aircraft #2) in the year 2008 for $60 million, which aircraft was actually "more expensive?" At first glance, it would appear that aircraft #2 is $10 million more expensive than aircraft #1. However, an accurate answer is not possible yet, as you must first account for the eight years of inflation involved in this problem. You would most likely do one of two things to determine which aircraft was actually more expensive: (1) Inflate the $50 million from the year 2000 up to the

year 2008, so that you could then compare that value with the $60 million in 2008 to see which was greater; or (2) deflate the $60 million in 2008 down to the year 2000, and then compare that value with the $50 million purchase made in that year. By doing this, you have both dollar values in the same year, so a fair comparison can now be made. In actuality, you could also inflate/deflate both values to *any* common base year, such as 2004 or 2010, but that gets us into many permutations of possibilities and would involve changing both of these values! So for now, let us just consider the two options presented and think of converting just one of the values.

When data are received from different sources, they must be "normalized," so that your data base will contain consistent data on which we can perform statistical analysis. Normalization provides consistent cost data by neutralizing the impacts of external influences. Adjusting actual cost to a uniform basis has two desired objectives:

- To reduce the dispersion of the data points to give you more data comparability, and
- To expand the number of comparable data points to give you more homogeneity

Note that in Figure 5.1, the data in the "Not Normalized" graphing on the left is scattered and has no pattern to it. However, after normalizing the data for inflation, you can see in the right scatter plot that we now have data that are less scattered and displays an upward trend in its pattern, one where we could perform a regression on this data and hope for a useful outcome.

Normalize data to ensure data comparability
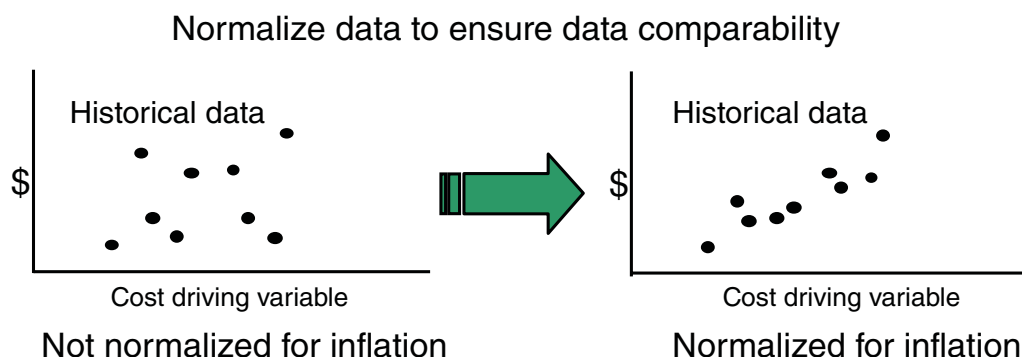
FIGURE 5.1 Comparison of "Not Normalized" vs. "Normalized" Data for Inflation.

As previously mentioned, there are three broad categories for normalizing data:

- Normalizing for Content
- Normalizing for Quantity
- Normalizing for Inflation

Normalizing for Content involves ensuring that you have identical Work Breakdown Structures (WBSs) between your new WBS and the historical WBS and that you are assigning/inputting your costs into the correct categories.

Normalizing for Quantity involves ensuring that you are comparing equal quantities of items or ensuring that you are at the same point on the learning curve when comparing two production lines. The usual standard is to compare items such as T1 (the theoretical first unit cost in a production scenario) or another quantity further along in the production process, such as Unit #100. Learning curves will be covered in great detail commencing in Chapter 10.

Normalizing for Inflation denotes removing the effects of inflation when comparing the costs of goods that are purchased in different years.

5.3 Normalizing for Content

Normalizing for content involves ensuring that you have identical WBS structures between your new WBS (for the system that you are costing) and the historical WBS from previous systems. Is there an "apples-to-apples" comparison between your categories? This is largely a problem of *mapping* different data sets. Figure 5.2 compares the WBS categories that you are expecting to use vs. the WBS that you acquire from your historical data sources.

My WBS	Historical Data
Air vehicle	Air vehicle
Airframe	Airframe
Powerplant	Propulsion
Communications	Comm / Nav
Navigation	Avionics
ECM	
Auto Flight Control	
Mission Subsystem	SE
SE/PM	PM
Data	Data

FIGURE 5.2 **Mapping of your Intended Aviation WBS Elements to Historical WBS Elements.**

In the left column of Figure 5.2 entitled "My WBS," we have developed a WBS for the aviation system that we are attempting to cost. But in reality, when you retrieve the historical data from your data sources discussed in the previous chapter, you will most likely find that the WBSs you retrieve are not exactly the same as the one that you envision and have developed. Clearly, the "Historical Data" WBS on the right is different from the "My WBS" column. These differences must be resolved. Let's discuss some of the differences between the two columns shown in Figure 5.2:

- Is the Airframe category under "My WBS" the same as the Airframe category in the "Historical Data?"
- Is "Powerplant" in the left column equivalent to "Propulsion" in the right column?
- "My WBS" lists "Communications" and "Navigation" in separate categories; however, in the historical data column, it is listed as "Comm/Nav," but there is also an "Avionics" category included that is not in your WBS.
- "My WBS" combines "SE/PM," but it is broken out separately in the historical data WBS.

This example illustrates the need to ensure that your WBS is comparable to the historical WBS and to ensure that you are inputting the costs that you need into the proper categories.
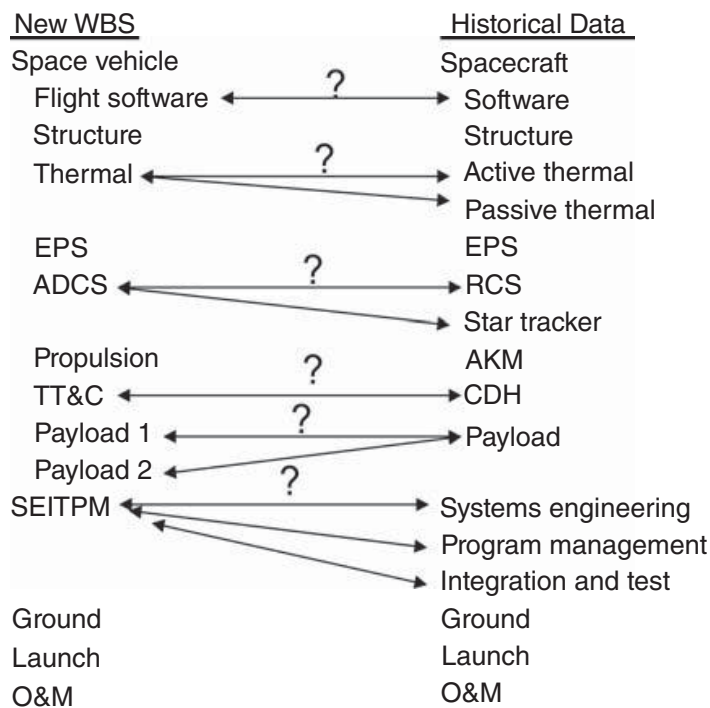
FIGURE 5.3 **Mapping of New Spacecraft WBS Elements to Old WBS Elements.**

Figure 5.3 is a second example of attempting to map different data sets. It is a difficult process, since no two contractors keep the same type of records and formats:

Again, you can see numerous differences (and some similarities) between the WBSs. The WBS that you created is found in the left column, and the Historical Data WBS is shown in the right column. Differences in the two columns include:

- Is "Flight Software" in your New WBS equivalent to "Software" in the right column?
- Your WBS lists "Thermal," while the Historical Data WBS breaks Thermal out between "Active Thermal" and "Passive Thermal." Is the "Thermal" in the left column just the sum of these two on the right?
- Is "ADCS" equivalent to "RCS + Star Tracker?"
- Is "TT&C" just a different name for "CDH," or are they different categories?

Numerous other differences exist, as can easily be seen in Figure 5.3. This example emphasizes the importance of Normalizing for Content when you receive your data.

5.4 Normalizing for Quantity

The second type of normalization we must accomplish is normalizing for quantity. Normalizing for quantity denotes ensuring that you are comparing equal quantities of items or ensuring that you are on the same point on the learning curve when comparing two production lines. The following two questions illustrate the importance of understanding quantity:

- How does quantity affect cost?
- Does Cost Improvement take place in our program? If so, at what rate is cost improving?

The easiest example of how quantity is affected by cost is by going to your local supermarket. You could purchase a box of cereal for say $3.50 per pound and you can buy it in a one pound box. However, if you go to Costco or Sam's Club, you can purchase the same product for perhaps only $2.25 per pound, but you are required to buy a three or five pound box. Depending on the size of your family and the ages of your kids, that may or may not be advantageous! But the reality is that the more quantity you buy – or when you buy in bulk – the lower the unit price should be. This is true not only in your supermarket, but also in the acquisition world, when we are purchasing a large number of items, whether the items are aircraft, flight suits, or maintenance tools. The more of these items that we purchase, the less we should pay per unit.

Normalization for quantity also ensures that we are comparing the same type of cost, whether it is the "Total Cost for N units," the "Lot Cost from Units 101 to 200," or merely a Unit level cost. When we collect production cost data, we usually receive it in terms of "Total Cost for X Units," or lot costs for units "X through Y."

In Chapter 10, we will learn about learning curves in great detail. But as an overview here in our discussion of quantity, the learning curve theory states that "As the quantity produced doubles, the unit cost, or cumulative average cost, decreases at a constant rate." This decrease is called the "rate of learning."

When we normalize for quantity, we try to find an "anchor point" to use as a data point for comparisons and for developing CERs. An example of an anchor point would be to use the cost of the first unit, or Unit #100, or Unit #1000, something that is at the same point in the production process in one system versus the same point in production of another system. Why is this important?

Example 5.1 We will use a sports example to illustrate this point. Let's consider the careers of three National Basketball Association (NBA) players who are very well known: Michael Jordan, Kobe Bryant, and LeBron James. I select these three athletes for a reason that will become apparent in the example. (Note: The points and seasons played are *approximated* for ease of comparison)

Player	# Seasons Played	# Points Scored	# Championships Won
Michael Jordan	20	38,000	6
Kobe Bryant	15	30,000	5
LeBron James	8	18,000	2

At first glance, it would appear to be an easy answer to say that Michael Jordan (MJ) was the best player of these three, because he has more points scored (38,000) and more championships won (6). However, what is the problem with making that statement? The problem is that in this example, MJ had 20 seasons with which to accrue these points and championships. At this time, Kobe Bryant has played five fewer seasons, but only has one less championship. Moreover, at this time, Lebron has only played for eight seasons, and therefore has had significantly less time to score points and win championships. What, then, would be a fairer comparison? You could find the average of points scored per season, but that would be a less effective methodology when comparing the number of championships won. Consequently, let us find an "anchor point" with which to use to make a fairer comparison. Since Lebron has only played for eight seasons, *let us use*

eight seasons as our anchor point, since they have all played at least that many seasons. We could then compare how many points and championships Michael Jordan had at the end of his first eight seasons, and also do the same for Kobe Bryant and LeBron James. Only then could any argument about who is the better player be made *fairly*. Of course, there are many other metrics that could be used as well – besides points scored and number of championships – but that will not be considered here!

Just as we have compared three players at the same point in their athletic careers, we must do the same when comparing costs from one historical system to another, as well. We must use the same "anchor point" in each of the production processes, in order for it to be a fair comparison.

Example 5.2 Consider this data set. We are trying to find the cost of a new fighter aircraft and we have found the following historical costs of:

- The cost of the 18th F-104
- The cost of the 300th F-4
- The cost of the 403rd F-14
- The cost of the 500th F-16
- The average cost of the last 100 F-18's

Why is this a bad data set? This is not a good data set because there is *no continuity* in the production process from aircraft to aircraft. We have acquired the costs of historical aircraft at various points in the production process for the first four aircraft and the average cost of 100 aircraft in our last data point. Unless you know the learning curve percentage for each of these aircraft, *the data are essentially useless to you and unusable.*

Now consider the following data set. We found the following:

- The cost of the 100th F-104
- The cost of the 100th F-4
- The cost of the 100th F-14
- The cost of the 100th F-16
- The cost of the 100th F-18

Clearly, this data set is much better! The reason is because we are at the identical point (unit #100) of the production process for each of these five aircraft. They are at the same point in the learning curve and thus the costs can correctly and fairly be compared to each other.

A second component of normalizing for quantity answers the following: Does cost improvement take place in our program? If so, at what rate is cost improving? We will cover this material in great detail in Chapter 10 on learning curves, but clearly cost improvement is more pronounced and significant when our rate of learning/rate of improving is higher.

5.5 Normalizing for Inflation

The final type of normalization we must account for is normalizing for inflation. While it is important that we normalize for content and quantity, we do most of our normalization in cost estimating to account for inflation. If System X costs $1M today, how

much would that same system cost five or ten years from now? While we do not know the exact answer yet, this is a reflection of the fact that a dollar spent today buys more than it will in the future, but it buys less than it did in the past. Thus, we will consider the increased effects of inflation over time. Notable exceptions, however, include computers and micro-electronics, as the price for these commodities decreases as technology improves. Of all the topics discussed in cost analysis, none will be encountered more frequently than inflation.

So what is inflation? Inflation is the *consistent rise in the price of a given market basket of goods produced by an economy*, which includes goods and services. It is usually measured by the rate of rise of some general product-price index and is measured in percent per year. Many different measures of inflation are required because prices do not rise evenly. So what are some of the most common market baskets in society? Excellent examples include the consumer price index (CPI), the product price index (PPI), the stock market, oil prices, fuel prices, stocks, bonds, mortgage rates, credit card rates, housing prices, etc. Each of these examples is considered a different market basket. A primary point to note is that prices for each of these categories do not rise evenly against each other. Fuel prices may be rising at the same time that housing prices might be falling (as was the case from about 2007–2012 in the United States) so each of these market baskets has their own distinct indices. Similarly, the Department of Defense uses different measures, and these will be discussed in great detail in the next section.

Let's begin our discussion of inflation by examining more closely what an index is. *An index* is simply a *ratio of one quantity to another*. It expresses a given quantity in terms of its relative value compared to a base quantity. Common examples in the finance world include the European Union's Euro (€) or the Japanese Yen (¥). When these currencies are discussed, their value is usually being compared against the US dollar ($). For the euro, in 2014 the conversion is approximately €1.4 euros for one US dollar. For the yen, there is approximately ¥100 yen for one US dollar. Both of these values (1.4 and 100) are considered an index, with the US dollar being the basis for comparison.

An inflation index is an index designed to *measure price changes over time*. It is a ratio of one price – or combination of prices – to the price of the same item or items in a different period of time. In order to create an inflation index, a *base period must first be selected* before an inflation index can be developed. The good news is that any year can be chosen! We will create an index here, so you can see how easy they are to produce. It will also assist you in your understanding of the many indices that we will look at in this chapter, especially the DoD inflation indices.

Example 5.3 In this example, let us learn how to construct an index. In order to make an index for any market basket, we must first select a base year, and the index for that year is always assigned a value of one (1.00). Price changes in that market basket, then, are always compared to the base year selected. The base period for a defense weapon system is often the fiscal year in which the program was initially funded. Here we have gathered historical data on fuel costs and these costs are shown in Table 5.1:

As previously stated, the first thing we must do is to select a base year. Again, *any year* can be selected to be the base year. In this example, we could select 2002 as the base year, since it is the first year of the data set, and if we do, then all of the costs per gallon would be compared against the value of $1.78 in 2002. We could also select 2007 since it is the latest year, and then all of the costs per gallon would be compared against the value of $2.90 in 2007. Or, we could select any other year in between, perhaps because you have a designated year needed in your cost estimate, or for a simpler reason such as the indices might be easier to calculate and visualize in a certain chosen year. In this

TABLE 5.1 Historical Fuel
Data Set for Example 5.3

Year	Cost Per Gallon
2002	$1.78
2003	$1.90
2004	$2.00
2005	$2.20
2006	$2.50
2007	$2.90

case, let's select 2004 as our base year because the cost of fuel per gallon in that year is an "easy-to-use-and-visualize-and-discuss" number @ $2.00 per gallon. Thus, since 2004 is the designated base year, the index for 2004 will be 1.00. Let's construct our complete set of indices for this data set based on the Base Year of 2004.

Since 2004 is our base year, the $2.00 per gallon will always be our *denominator* when calculating our index. To calculate the index for each of these years, we will take the cost per gallon in that year, and then divide it by the base year cost of $2.00 per gallon. Calculations to create our index are as follows:

- 2002: $1.78/$2.00 = 0.89
- 2003: $1.90/$2.00 = 0.95
- 2004: $2.00/$2.00 = 1.00
- 2005: $2.20/$2.00 = 1.10
- 2006: $2.50/$2.00 = 1.25
- 2007: $2.90/$2.00 = 1.45

Again, note that we divide by the $2.00 per gallon to calculate each index because that is the cost in the base year. This is the reason why your base year will *always* have an index of 1.00, since you are taking the cost for that year and dividing it by itself (in this case, $2.00/$2.00 = 1.00)

Thus our completed indices from Example 5.3 will look like the following, found in Table 5.2:

TABLE 5.2 Completed Index in Each Year from
Example 5.3

Year	Cost Per Gallon	Index
2002	$1.78	0.89
2003	$1.90	0.95
2004	$2.00	1.00
2005	$2.20	1.10
2006	$2.50	1.25
2007	$2.90	1.45

Let's interpret the results found in Table 5.2. For the year 2002, the cost of fuel ($1.78/gallon) created an index of 0.89 when compared to the $2.00/gallon cost in 2004.

This means that the cost for fuel in 2002 was 89% of what a gallon of fuel cost in the year 2004. In the year 2003, the cost of fuel ($1.90/gallon) was 95% of the cost in 2004, since it had an index of 0.95. Note that all of the fuel costs increased as the years went by, most likely due to the effects of inflation, and all of the costs are compared to the base year. As we look beyond the base year of 2004, we now see that the indices are *greater* than 1.00, since the cost of fuel in those years *exceeded* the cost of fuel in 2004 (@ $2.00/gallon). Since the cost of fuel in 2005 is $2.20/gallon, it is 10% more (index = 1.10) than what that same fuel cost in 2004. In 2006, the cost ($2.50/gallon) was 25% more (index = 1.25) than what that same fuel cost in 2004; and the cost in 2007 ($2.90) was 45% higher (index = 1.45) than in 2004. Note that all of these indices are compared only to the base year of 2004. This is important to remember as we make conversions from one year to another that may not involve ending in the base year. Regardless of what year we start in, *we must always convert to the base year first*, and *then* convert to the year in which we are interested. Now that we have established a set of fuel indices from our historical data, let's practice some conversions using these indices.

Example 5.4 Using the fuel indices calculated in Table 5.2, calculate the answers to the following three questions:

- *Question #1:* Twenty gallons of fuel was worth $40.00 in 2004. How much did that same 20 gallons of fuel cost you in 2007?

- *Answer #1:* In this problem, you are converting the value of $40 from 2004 to its value in 2007. To make this conversion using the aforementioned index, it is merely a one step process, since we are commencing in the base year. The calculation would be $40.00 times the inflation index in the year that we would like to convert to. In this example, as we are converting to the year 2007, the index we use is 1.45. Thus, $40.00 x 1.45 = $58.00. Note that we *multiply* by the index because we need our final answer to be *greater* than the starting value, since we are going to a later period in time, and cost is increased due to inflation.

   To summarize Answer #1 in words, 20 gallons of fuel in 2004 would have cost you $40; three years later, in 2007, that same amount of fuel cost you $58, an increase of 45%.

- *Question #2*: A quantity of fuel was worth $2,500 in 2006. What was that same quantity of fuel worth back in 2004?

- *Answer #2*: In this problem, you are converting the value of $2,500 from 2006 to what its value was in 2004. To make this conversion using the aforementioned index, it is again a one step process, since we are calculating our conversion to the base year of 2004. Thus, the calculation would be $2,500 divided by the inflation index in the year from which we are converting. In this example, the index for 2006 is 1.25. Thus, $2,500 ÷ 1.25 = $2,000. Note that this time we *divide* by the index because we need our final answer to be *smaller* than the starting value, since we are going back in time and our cost will correspondingly be less expensive.

   To summarize Answer #2 in words, a quantity of fuel that cost you $2,500 in 2006 would have cost you only $2,000 in 2004.

   For Question #3, let's consider question number #2 again, but we will modify the question slightly. Instead of converting to the year 2004, let us instead convert from 2006 to the year 2002. How does this make the problem different?

- *Question #3*: A quantity of fuel was worth $2,500 in 2006. What was that same quantity of fuel worth back in 2002?

- *Answer #3*: In this problem, you are converting the value of $2,500 from 2006 to what its value was back in 2002. To make this conversion using our index, it is no longer a one step process, as we are calculating our conversion to a year *other than the base year*. It is now a two-step process: we must first convert from 2006 to 2004 (as we have previously performed in question #2), and then we must make a second conversion, from 2004 to 2002. Note that your first calculation must *always* go through the base year *first*, because the indices are all indexed against that base year. The costs in 2006 are 125% of the costs in 2004, not 125% of the costs in 2002!

Thus, in question #3, the first calculation would be the same as in question #2, which was $2,500 ÷ 1.25 = $2,000 in 2004. But now we must next convert from 2004 to 2002. Note that the index for 2002 is 0.89. Since we desire our final answer in 2002 to be *smaller* than our value in 2004, we must now *multiply* by the second index to calculate our final answer in 2002. Thus, our calculation is $2,000 × 0.89 = $1,780.

To summarize Answer #3 in words, a quantity of fuel that cost you $2,500 in 2006 would have cost you only $1,780 in 2002.

5.6 DoD Appropriations and Background

The previous section discussed some of the "market baskets" that we find in our society. We then made a set of indices from historical data for one of those market baskets – fuel – and we were then able to convert dollar values from one fiscal year to another using those indices. Similarly, the Department of Defense also has a large number of "market baskets" to choose from, and each of these has its own set of inflation indices.

DoD inflation indices are developed *for each service* for a particular activity or type of procurement, called Appropriations. Appropriation examples from each of the Services include:

- Aircraft Procurement, Navy (APN)
- Shipbuilding and Conversion, Navy (SCN)
- Military Pay, Army (MPA)
- Weapons and Track Combat Vehicles Procurement, Army (WTCV)
- Operations and Maintenance, Air Force
- Reserve Personnel, Marine Corps (RPMC)
- Research, Development, Test and Evaluation, Navy (RDT&EN)
- Military Construction (MILCON)
- Fuel for all Services (OSD cost element)

Note that the first seven appropriations were service-specific, whereas the final two were DoD-wide, meaning that all of the services would use the same indices for those appropriations. Each of these appropriations (such as aircraft, spacecraft, weapons, ship construction, operations and maintenance, and manpower) has its own rate of inflation. These DoD inflation indices are used by analysts to convert between Constant Year dollars

(CY$), Fiscal Year dollars (FY$), Base Year dollars (BY$), and Then Year dollars (TY$). You *must* ensure that you are using the proper inflation indices for the project you are working on.

In addition to the military services, many other organizations such as NASA and government agencies like the Congressional Budget Office also use inflation indices. Inflation indices are important as they remove the effects of changes in the prices of goods/commodities we buy which are due merely to inflation, to be able to identify changes in the actual costs of the goods over time.

There are two perspectives for inflation indices. The first is for the "big picture view" overall. This helps organizations such as the Office of Management and Budget (OMB) (i.e., the White House) to understand the real cost of how the economy is doing. It implies a broad index, such as the Gross Domestic Product (GDP) deflator. The second perspective is more "micro-economy" oriented. It helps the service operating agencies to understand how many real resources they are capable of buying in the present economy. The DoD needs the inflation indices to calculate the real cost growth of what it buys, such as weapons systems, which are calculated in Constant Year dollars. They are also used to estimate future budget requirements, which are calculated in Then Year dollars.

Inflation issues are important to officials at the highest level of our government. The Weapon Systems Acquisition Reform Act of 2009 (WSARA 09) was signed by the President of the United States and it requires OSD-CAPE (within the DoD) to "*Periodically assess and update the cost (or inflation) indexes used by the Department of Defense to ensure that such indexes have a sound basis and meet the Department's needs for realistic cost estimation*" [1]. In addition, Title 31 of U.S. Code authorizes the OMB to supervise budget preparation. OMB annually issues "economic assumptions" for use by agencies in preparing their budget submissions, which includes forecasts of five price indexes: Military pay, civilian pay, fuel, health care, and other purchases. Inflation for other purchases is represented by the GDP deflator [2].

The Under Secretary of Defense (Comptroller) [USD(C)] issues inflation guidance to DoD components, including [2]:

- Distinct price indexes for each appropriation account for each Service
- Estimated price level changes based on data provided by OUSD (Comptroller)
- The "most likely" or "expected" full cost to be reflected in the budget

With the previous two paragraphs offered as a quick background on the requirements for inflation indices, there are three terms that we must discuss and understand that are used consistently in this area of study. They are Constant Year dollars, Base Year dollars, and Then Year dollars.

5.7 Constant Year Dollars (CY$)

CY$ reflect the purchasing power or value of the dollar in the specified constant fiscal year. In discussing an acquisition program, the dollar amount projected is discussed as if it were totally expended in that specified year. For example, if the Total Depot Maintenance for the Armored Vehicle Launched Bridge was calculated to be $4.77M in Fiscal Year 2005 (FY05$), the $4.77M is considered to be in constant year dollars. This does not imply that all of the $4.77M was actually *paid* to the contractor in FY05. But it does imply

that total depot maintenance *would* have cost $4.77M if all expenditures had occurred in FY05. Let's suppose that the actual expenditures were made over a three-year period, perhaps looking something like the following:

Year	Payments Made (= $5M)
2005	$2 Million
2006	$2 Million
2007	$1 Million

If you add up the payments made in these three years, it appears that we made $5 million in payments. But if we take out the effects of inflation from the 2006 and 2007 payments, the value of the payments made (in constant year dollars) look more like the following: (Note: these numbers are notional for the illustration of this example)

Year	Payments Made (= $4.77M)
2005	$2 Million
2006	$1.9 Million (after removal of one year of inflation)
2007	$0.87 Million (after removal of two years of inflation)

The payment in 2005 is not subject to inflation, so it is worth the full $2 million that was paid. The payment made in 2006, however, is subject to *one year* of inflation. Thus, the $2M paid in 2006 was actually only worth $1.9M once the effects of inflation were removed. The payment made in 2007 is subject to *two years* of inflation. Thus, the $1M payment in 2007 was only worth $0.87M after the effects of two years of inflation have been accounted for. Therefore, when the three payments are all reflected in terms of constant year 2005 dollars, the total payment was worth $4.77M in Constant Year 05$. Moreover, the $4.77M reflects the total cost *if* all expenditures had occurred in FY05.

To expound upon this point with a different example, let's say that an aircraft program is going to cost $300M over a 10 year period. This $300M is (most probably) given in Constant Year dollars, and let's suppose that this program is in constant year CY13$. This does not imply that all $300M will be paid to the contractor in the year 2013. In reality, the payments would be spread out over the full ten years, from CY13 to CY22. Perhaps we will pay $50M in the first year (2013), and then spread the final $250M equally over the last nine years. At the end of the ten years, we will have paid significantly *more* than $300M, due to the effects of inflation incurred in each year over the final nine years of payments. But, the key point to remember is that *the program will still be referred to as a $300M program*! The fact that we are actually paying more than $300M (and perhaps significantly more) is generally the concern of the Comptroller, who is the one who has to make the payment to the contractor and needs to know the exact amount to make the payment for.

A relevant parallel to this in our everyday life would be if you purchased a new car. Let's suppose that you bought a new Honda for $30,000. If you paid the full cash amount of $30,000 when you purchased the car, the car cost you exactly $30,000 to buy. However, what if you only paid $10,000 as a down payment and then financed the remaining $20,000? If you financed that $20,000 over a three-year period, you would pay a certain amount of interest; if you financed it over a five-year period, your amount of interest paid would be *more* than over a three-year period; and if you financed it over a seven-year period, you would pay even *more* interest. But the bottom line is that it is still only a $30,000 car, despite the extra $3,000 or $5,000 that you paid in interest! If someone asked you "How

much did the car cost?," you would still say that it was a $30,000 car. You would not add in the interest charges to the price when revealing what the car cost you. Correspondingly, it works the same in the $300M example. Even though the program is a $300M program, you will pay more for it because you have spread the costs over a total of ten years. But overall, despite the fact that you paid significantly more than the $300M, it is still regarded as a $300M program. The only difference in these examples is that government contracts have to deal with the effect of *inflation*, while the consumer (i.e., you) must deal with the effects of *interest* (but they work essentially the same way!)

5.8 Base Year Dollars (BY$)

BY$ are a subset of CY$. The base year is formally defined as *the fiscal year in which the program was initially funded*. This is important as it enables a decision maker or program manager to distinguish between a change in the cost of a program (perhaps due to potential problems), and a change in the purchasing power of the dollar due to the effects of inflation. The following example will demonstrate the importance of this concept.

Example 5.5 Base Year Dollars   Let us assume that three cost estimates were performed on an ACAT I program for the Milestone A, B, and C reviews over a five year period. Table 5.3 is the data set for this example:

TABLE 5.3   Data Set for Base Year Example 5.5

Year of Report	CY Estimate	BY Estimate
2008 (MS A)	$450M (CY08$)	$450M (BY08$)
2010 (MS B)	$467M (CY10$)	$450M (BY08$)
2013 (MS C)	$501M (CY13$)	$476M (BY08$)

The cost estimate for the Milestone A review back in 2008 was for the program to cost $450 million. Since it was the first year of funding for the program, we will also consider 2008 to be the Base Year as well. Two years later, the Milestone B review occurred, and the cost estimate in 2010 dollars was now for $467M. While it appears that the program has increased by $17 million (from $450M to $467M), when you take out the effects of the two years of inflation and convert your $467M cost back into 2008 dollars, it turns out that your program is still estimated to cost $450M (BY estimate). Thus, by converting back to the base year of 2008, we see that we have *no cost growth* in that program at the present time. So far, so good! The $17 Million increase was due merely to inflation and not due to an increase in the cost of the program.

Milestone C occurs three years later, in 2013. Your cost estimate for this milestone is now for $501 million in CY13$. It appears that your program has now increased by $51 million, an increase of just over 11%. Yikes! But when you take out the effects of five years of inflation and convert your 2013 estimate back to 2008, you find that your program has only increased by $26 million to $476M. In addition, by converting your cost estimates back to the base year, we see what our actual cost growth has been in the program. This enables the program manager to determine exactly how much his/her program has increased in real dollars, instead of seeing increases due to inflation. Therefore, in Example 5.5, the program has increased by only 5.8% over the five year period from 2008 to 2013.

5.9 DoD Inflation Indices

 In Section 5.5, we practiced some dollar conversions using an index we made from historical fuel data. Now that we have discussed Constant Year and Base Year dollars, let us learn how to make conversions from one year to another year using inflation indices from an actual DoD appropriation table. Table 5.4 displays the inflation indices for the DoD appropriation category of Aircraft Procurement, Navy (APN) with a Base Year of 2012. These indices were generated from the Joint Inflation Calculator (JIC) at the NCCA website. The top label on Table 5.4 displays the appropriation table, which service the indices are for, the base year of the indices, and the date that the inflation indices were generated (i.e., APN/Navy/2012/4 May 2013). While the actual indices for APN generated by the JIC range from the year 1970 to year 2060, we have truncated the table slightly to the years covering 1980–2017 for space considerations, and the examples will cover costs only between those years. (Note: Detailed directions on how to use the JIC will be discussed in Section 5.11 in this chapter.)

Analyzing Table 5.4, we see that the first column reflects the *Fiscal Year* that the data are from: 1980 to 2017. The second column displays the *Inflation Rate* that was encountered by the aircraft procurement economy in that fiscal year. *The Raw Index* column are the indices used for CY/BY and FY dollar conversions. Note that the base year of 2012 has a Raw Index of 1.0000. This is exactly equivalent to the index of 1.00 that we calculated in the fuel example in Section 5.5, Example 5.3. The *Weighted Index* column is used when you have Then-Year dollar conversions. The *Budget Year Index* is only used for Then-Year to Then-Year conversions. The last column is the *Budget Year Inflation Rate* for the *Budget Year Index* column.

Since only costs that are computed using the same base year (or constant year) are comparable to each other, we adjust costs using DoD inflation indices to reflect the effects of inflation for three reasons:

- To adjust historical costs to the same standard (CY\$ or BY\$)
- To submit budget requests to Congress (TY\$)
- To calculate "escalation" for contractors, which adjusts reasonable profits if inflation is less (or more) than expected.

To understand the DoD inflation indices a little more, Table 5.5 is a five-year snapshot of the APN indices shown in Table 5.4, from the fiscal years 2010 to 2014, and it encompasses the base year for the indices of 2012.

Note that the Raw Index for the year 2012 is 1.0000, because it is the base year. Now let us look at the year 2013. Because there was an inflation rate of 2.1% in that year, the raw index of 1.000 in 2012 is now 1.0210. This is calculated by merely inflating the previous index (=1.000) by 2.1%. Since the Inflation Rate for the Base Year + 1 = 2.1%, the new Raw Index (for 2013) is $(1.000) \times (1 + 0.021) = 1.0210$. Correspondingly, in 2014, there was an inflation rate of 1.90%, so we must inflate the Raw Index of 1.021 from 2013 by 1.9%, thus becoming 1.0404 (=1.021 × 1.019). This is how the Raw Indices are calculated for each year. You can do the same for the raw indices *prior* to 2012 by *deflating* by the inflation rate in the given year. For example, if you deflate the Base Year index of 1.000 by 1.80% (the inflation rate in 2012), you will get the raw index of 0.9823, found in 2011 (=1.000 ÷ 1.018 = 0.9823). We can now practice how to make conversions from one fiscal year to another using the APN inflation indices.

TABLE 5.4   APN Inflation Indices, With a Base Year of 2012

APN = Aircraft Procurement, Navy				
NAVY	Base Year = 2012			4-May-13
Fiscal Year	Inflation Rate %	Raw Index	Weighted Index	Budget Year Index	Budget Year Inflation Rate %
1980	11.80%	0.3569	0.4205	0.4039	11.56%
1981	11.60%	0.3983	0.4681	0.4497	11.33%
1982	14.30%	0.4553	0.5085	0.4885	8.63%
1983	9.00%	0.4963	0.5409	0.5196	6.38%
1984	8.00%	0.5360	0.5626	0.5405	4.02%
1985	3.40%	0.5542	0.5789	0.5562	2.90%
1986	2.80%	0.5697	0.5966	0.5731	3.05%
1987	2.70%	0.5851	0.6175	0.5932	3.51%
1988	3.00%	0.6027	0.6444	0.6190	4.35%
1989	4.20%	0.6280	0.6702	0.6438	4.00%
1990	4.00%	0.6531	0.6933	0.6660	3.45%
1991	4.30%	0.6812	0.7132	0.6851	2.87%
1992	2.80%	0.7002	0.7292	0.7005	2.25%
1993	2.70%	0.7192	0.7428	0.7136	1.86%
1994	2.00%	0.7335	0.7566	0.7268	1.85%
1995	1.90%	0.7475	0.7687	0.7385	1.61%
1996	2.00%	0.7624	0.7795	0.7489	1.40%
1997	1.80%	0.7761	0.7862	0.7553	0.86%
1998	0.70%	0.7816	0.7954	0.7641	1.16%
1999	0.80%	0.7878	0.8056	0.7739	1.28%
2000	1.40%	0.7989	0.8163	0.7842	1.33%
2001	1.80%	0.8132	0.8260	0.7935	1.19%
2002	0.80%	0.8198	0.8364	0.8035	1.26%
2003	1.00%	0.8279	0.8531	0.8196	2.00%
2004	2.00%	0.8445	0.8757	0.8412	2.64%
2005	2.80%	0.8682	0.9003	0.8649	2.82%
2006	3.10%	0.8951	0.9253	0.8889	2.77%
2007	2.70%	0.9192	0.9469	0.9096	2.33%
2008	2.40%	0.9413	0.9612	0.9234	1.51%
2009	1.50%	0.9554	0.9749	0.9366	1.43%
2010	0.80%	0.9631	0.9971	0.9579	2.28%
2011	2.00%	0.9823	1.0210	0.9809	2.40%
2012	1.80%	1.0000	1.0409	1.0000	1.95%
2013	2.10%	1.0210	1.0610	1.0193	1.93%
2014	1.90%	1.0404	1.0812	1.0386	1.90%
2015	1.90%	1.0602	1.1017	1.0584	1.90%
2016	1.90%	1.0803	1.1226	1.0785	1.90%
2017	1.90%	1.1008	1.1440	1.0990	1.90%

TABLE 5.5 Five years of the Table 5.4 APN Inflation Indices, With a Base Year of 2012

Year	Inflation Rate (%)	Raw Index
2010	0.80	0.9631
2011	2.00	0.9823
2012	1.80	1.0000
2013	2.10	1.0210
2014	1.90	1.0404

TABLE 5.6 Initial Data Set to Example 5.6 using APN Inflation Indices

Helicopter Program	Development Costs	Base Year Cost (CY12$M)	Final Cost (CY10$M)
MH-60S	$512M (CY05$)		
CH-46D	$465M (CY98$)		
TH-57	$235M (CY03$)		
MH-53E	$723M (CY01$)		

Example 5.6 Using the APN inflation indices in Table 5.4, let's normalize the following helicopter development costs found in Table 5.6 from their various years to CY10$. Since the base year for the APN indices is 2012, these conversions will require a two step process: first, from the year that you are beginning in to the base year of 2012; then, a second conversion from the base year 2012 to our target year of 2010. Recall that the Raw Index column is used for all CY$, BY$, and FY$ conversions.

Note that each of the historical costs come from different years: CY05, CY98, CY03, and CY01, respectively. Therefore, we must normalize this data to one particular year so that we may compare them properly. In this case, we have chosen to normalize the data to the year 2010. Therefore, for each conversion, we must first convert from the original year to the base year of 2012, and then convert that base year value back to the year that we desire, 2010.

MH-60S conversion: We are converting $512M from CY05 to CY10. The two-step process is to convert from CY05$ to CY12$ first, and then from CY12$ to C10$. We will write this process as

CY05$ → CY12$ → CY10$

The two indices we will need are from the years 2005 and 2010, and the Raw Indices for these two years are 2005 = 0.8682 and 2010 = 0.9631. Since we want the costs from 2005 to get *larger* when converting to 2012 (due to inflation in the later years), we will divide in the first calculation:

Convert from CY05$ → CY12$:
$512M ÷ 0.8682 = **$589.73M (CY12$)**

We now have our answer in the base year of 2012. Next, we will convert from 2012 to our target year of 2010. (CY12$ → CY10$). Note that since we want the value to get *smaller* as we convert back to 2010, we will need to multiply this time, since the inflation indice is less than 1.000:

Convert from CY12$ → CY10$:

$589.73M × 0.9631 = **$567.96 (CY10$)**

Thus, what cost us $512M in 2005 will now cost us $567.96M in 2010.

The calculations for the remaining three conversions are as follows:

- *CH-46D conversion*: CY98$ → CY12$ → CY10$

    Indices needed are: 1998 = 0.7816 and 2010 = 0.9631

    First, convert from CY98$ → CY12$:

    $465M ÷ 0.7816 = **$594.93M (CY12$),** then convert from CY12$ → CY10$:

    $594.93M × 0.9631 = **$572.98 (CY10$)**

- *TH-57 conversion*: CY03$ → CY12$ → CY10$

    Indices needed are: 2003 = 0.8279 and 2010 = 0.9631

    First conversion: $235M ÷ 0.8279 = **$283.85M (CY12$),** then

    Second conversion: $283.85M × 0.9631 = **$273.37 (CY10$)**

- *MH-53E conversion*: CY01$ → CY12$ → CY10$

    Indices needed are: 2001 = 0.8132 and 2010 = 0.9631

    First conversion: $723M ÷ 0.8132 = **$889.08M (CY12$),** then

    Second conversion: $889.08M × 0.9631 = **$856.27 (CY10$)**

The four conversions in Example 5.6 are shown in chart form in Table 5.7. The format shown using Excel makes it easy to see what costs we began with, the indices that were used for our conversions, and the final answers that were made in your documentation. *Note that you always divide by the first index to get to your base year, and then multiply by the second index to get to your target year!*

TABLE 5.7 Final Answers to Example 5.6 Conversions Using APN Raw Indices

Helicopter Program	CY$	Development Costs ($M)	Indice to BY12	Cost in 2012$M	Indice to 2010	Final Cost (CY10$M)
MH-60S	2005	512	0.8682	589.7259	0.9631	567.9650
Ch-46D	1998	465	0.7816	594.9335	0.9631	572.9804
TH-57	2003	235	0.8279	283.8507	0.9631	273.3766
MH-53E	2001	723	0.8132	889.0802	0.9631	856.2731

As a side note, and using the final answer for MH-60S in Table 5.7 as an illustration, cost answers can be written in one of two ways:

- $567.965M (CY10$), or
- $567.965 (CY10$M)

Either way is considered acceptable. Do not forget the dollar units, either!

- K$ (=Thousands)
- M$ (=Millions)
- B$ (=Billions)

5.10 Then Year Dollars (TY$)

In reality, do all expenditures for a program occur within one year? Almost never! TY$ represent the amount of money needed when the *expenditures* for goods and services are made. They reflect the actual flow of expenditures during an acquisition program and they include an expenditure profile.

Consider an acquisition program that is expected to cost $100M in constant dollars over five years. Perhaps you arrange to make payments to the contractor in the following manner:

2010	2011	2012	2013	2014
$20M	$20M	$20M	$20M	$20M

If the government pays the contractor $20M for five straight years in a $100M contract, who is getting the better deal, the government or the contractor? Clearly, in this example, the *government* would be getting the better deal in that example. This is because there is no inflation included in any of the four final payments from 2011 to 2014. In reality, the expenditures will look more like this (whole numbers used for simplicity):

2010	2011	2012	2013	2014
$20M	$21M	$22.5M	$24M	$26M

The payout in 2010 is not subject to inflation. But the payouts from 2011 through 2014 will be, and all are considered to be in Then-Year dollars. The payout in 2011 is subject to one year of inflation; 2012 two years; 2013 three years; and 2014 is subjected to four years of inflation. Thus, these numbers represent the amount of money actually needed when the expenditures are made to pay the contractor.

Let's discuss how to make Then-Year conversions from one year to another year. Most importantly, TY dollar conversions require you to use the *Weighted Index* in the inflation indices. There are three types of conversions: TY$ to CY$, CY$ to TY$, and TY$ to TY$ conversions. We will once again use the APN inflation indices found in Table 5.4 to practice these.

Example 5.7 Perform the following five conversions using the *APN Inflation Indices* from Table 5.4 (with BY12$):

TY$ → CY$ Conversions:

(A) 630K TY06$ to CY04$

(B) 2.4M TY10$ to CY08$

CY$ → TY$ Conversions:

(C) 523M CY04$ to TY07$

(D) 375.4M CY07$ to TY13$

TY$ → TY$ Conversions:

(E) 24M TY08$ to TY14$ (two ways)

TY$ to CY$ conversions: Examples A and B are conversions from Then-Year dollars to Constant Year dollars. The two-step process in Example A is thus from TY06$ → CY12$ → CY04$. No matter what conversion you perform (TY to CY, CY to TY, or CY to CY), the first step *always* converts you to CY12$, since that is your base year.

Since the first number is in TY06$ (=$630K), we will need to use the *Weighted Index* for that conversion, which is the fourth column in the inflation indices. The *Weighted Index* for the year 2006 = 0.9253, and the conversion will leave us in CY12$. The second conversion is then just a CY to CY conversion as we have already seen, from CY12$ to CY04$. Thus, use the 2004 *Raw Index* = 0.8445 to get to CY04$. Let's give this a try.

- TY06$ → CY12$ → CY04$

  First, we will convert from TY06$ → CY12$:

  $630K ÷ 0.9253 = **$680.86K (CY12$)**

  Next, we will convert from CY12$ → CY04$:

  $680.86M × 0.8445 = **$574.98K (CY04$)**

  Thus, what cost us $630K in 2006 when we made a payment would have only cost $574.98K in 2004.

  The calculations for the TY to CY conversions in Example B are as follows:

- TY10$ → CY12$ → CY08$. Indices needed are: *Weighted Index* for 2010 = 0.9971 and the *Raw Index* for 2008 = 0.9413

  TY10$ → CY12$: $2.4M ÷ 0.9971 = **$2.407M (CY12$),** then

  CY12$ → CY08$: $2.407M × 0.9413 = **$2.266M (CY08$).**

  This completes the conversions from TY$ to CY$ in Examples A and B.

CY$ to TY$ conversions: Examples C and D are conversions from Constant Year dollars to Then-Year dollars. Looking at Example C, the two indices we will need are from the years 2004 and 2007. The two-step process is from CY04$ → CY12$ → TY07$. Since the first conversion is CY$ to CY$, we will need to use the *Raw Index* for that conversion, and the result will leave us in CY12$. The *Weighted Index* will then be used in the second conversion from CY12$ to TY07$. The *Raw Index* for 2004 = 0.8445, and the *Weighted Index* for 2007 = 0.9469.

- CY04$ → CY12$ → TY07$:

  First conversion of CY04$ → CY12$: 523M ÷ 0.8445 = **$619.301M (CY12$)**, then

  Second conversion of CY12$ → TY07$: 619.301M × 0.9469 = **$586.416M (TY07$).**

  The calculations for the CY to TY conversions in Example D are as follows:

- CY07$ → CY12$ → TY13$. Indices needed are: *Raw Index* for 2007 = 0.9192 and the *Weighted Index* for 2013 = 1.0610.

  First conversion of CY07$ → CY12$: $375.4M ÷ 0.9192 = **$408.398M (CY12$),** then

  Second conversion of CY12$ → TY13$: $408.398M × 1.0610 = **$433.311M (TY13$)**

This completes the conversions from CY$ to TY$ in Examples C and D.

The final problem in Example 5.7, Example E, involves TY$ to TY$ conversions. *These conversions are from TY08$ to TY14$, and they can be accomplished in two ways:*

- Conversion Method 1: TY08$ → CY12$ → TY14$
- Conversion Method 2: TY08$ → TY12$ → TY14$

Conversion Method 1 combines what we did in the TY$ to CY$ and CY$ to TY$ conversions, going through the base year of CY12. Method 2 utilizes the fifth column on the inflation indices, the *Budget Year Index*, and all answers remain in TY dollars. Either method will calculate an identical answer (within three significant digits accuracy).

For Example E, Method 1, you would need to use the *Weighted Index* for 2008 = 0.9612, which would convert you to CY12$. You would then need the *Weighted Index* column again to get you to TY14$, and that indice is equal to 1.0812.

The calculations for this method are as follows:

- TY08$ → CY12$ → TY14$

First conversion of TY08$ → CY12$: $24M ÷ 0.9612 = **$24.969M (CY12$)**

Second conversion of CY12$ → TY14$: $24.968M × 1.0812 = **$26.996M (TY14$)**

Example E, Method 2, demonstrates a second method to calculate the same problem, this time using the *Budget Year Index*. In this method, you will only use the *Budget Year Index* column and do not use either the *Raw* or the *Weighted Index*. For this example, you would need to use the *Budget Year Index* for 2008 = 0.9234, which would convert you to TY12$. You would then use the *Budget Year Index* column again to get you to TY14$, and that indice is equal to 1.0386.

The calculations for this method are as follows:

- TY08$ → TY12$ → TY14$

First conversion of TY08$ → TY12$: $24M ÷ 0.9234 = **$25.991M (TY12$)**

Second conversion of TY12$ → TY14$: $25.991M × 1.0386 = **$26.995M (TY14$)**

Either method will produce a similar answer, subject to round off on significant digits. This completes the two conversion methods from TY$ to TY$ in Example 5.7E.

All six conversions from Example 5.7 are shown in chart form in Table 5.8. Using this format makes it easy to see what indices were used, and the conversions that were made in your documentation.

5.11 Using the Joint Inflation Calculator (JIC)

Learning how to use the indices and making conversions like we just did is an excellent way to understand the concept that you must *always* go through the base year while making conversions. However, a quicker way to make a number of conversions is to use the Joint Inflation Calculator (JIC), an Excel-based automated tool that is easy to use.

To use the JIC, go to the Naval Center for Cost Analysis (NCCA) website, and download the calculator from their homepage to your desktop. The NCCA

TABLE 5.8　Final Answers to Example 5.7 Conversions Using APN Indices

Example	Initial Cost	Type Dollars	Indice to CY12	Cost in CY12$	Indice to Desired Year	Final Cost	Year
A	630	TY06$K	0.9253	680.860	0.8445	574.986	CY04$K
B	2.4	TY10$M	0.9971	2.407	0.9413	2.266	CY08$M
C	523	CY04$M	0.8445	619.301	0.9469	586.416	TY07$M
D	375.4	CY07$M	0.9192	408.399	1.061	433.311	TY13$M
E(1)	24	TY08$M	0.9612	24.969	1.0812	26.996	TY14$M
E(2)	24	TY08$M	0.9234	25.991 (TY$)	1.0386	26.995	TY14$M

website can be found at: www.ncca.navy.mil, or just conduct an internet search for "NCCA."

To use the JIC, step-by-step instructions are offered in Example 5.8. We will practice using the JIC with previously used numbers, again using the APN inflation indices with Base Year 2012. While using this tool, you will note that there are many other Appropriations for each service.

Example 5.8 Perform the following conversions using the Joint Inflation Calculator (JIC):

(1) *CY$/FY$ → CY$/FY$ Conversion*: $512M CY05$ to CY10$

(2) *TY$ → CY$ Conversion*: $630K TY06$ to CY04$

(3) *CY$ → TY$ Conversion*: $375.4M CY07$ to TY13$

(4) *TY$ → TY$ Conversion*: $24M TY08$ to TY14$

Step by step instructions for using the JIC:

- You must first "Enable" the Excel-based macro. Click on "Options" at the top left corner, then click on "Enable this Content."

To Make Conversions From One Year to Another:

- Click on "Query."
- At 1, select the desired *Service* you want the indices from.
- At 2, select the desired *Appropriation category* (APN, WPN, etc)
- At 3, enter the year of the data you are converting *from.*
- At 3A, select type of conversion *(FY to FY, FY to TY, etc)*
- At 3B, enter the year you are converting *to.*
- Under 3C, the "Quick Look" box below is now complete with your entered instructions.
- In the first blue box in the Input column in the "Quick Look," enter the value of the data that you are converting. Answer will now appear in the last column (Output/Result) in the same row as your data. You can enter six entries at a time.

- Note: The "Print Quick Look" link will print out the answers in the entire "Quick Look" box.

   To Make Appropriation Tables (such as APN, SCN, etc):

- Click on "Query."
- At 1, select the desired *Service* you want the indices from.
- At 2, select the desired *Appropriation category* (APN, WPN, etc).
- At 3, enter the Base Year you desire for the Appropriation Table you are generating.
- Click on "Generate Inflation Table."
- The Inflation Table will provide the actual inflation indices from 1970 to the base year you designated, and then what is projected after this base year through the year 2060.

TABLE 5.9   Final Answers to JIC Conversions in Example 5.8 Using APN Indices

Starting Cost	JIC Conversions	JIC Indice	Final Cost	Year
512	CY05$M to CY10$M	1.1093	567.97	CY10$M
630	TY06$K to CY04$K	0.9127	574.998	CY04$K
375.4	CY07$M to TY13$M	1.1542	433.292	TY13$M
24	TY08$M to TY14$M	1.1248	26.996	TY14$M

   Table 5.9 compiles the four questions/answers from Example 5.8 using the JIC. Note that there is only one indice used, instead of the two when using the inflation indices. This is because the JIC is combining both of the indices you used from the inflation indices into one index.

   Looking at the first example of $512M, the JIC combines the two indices that you found in Example 5.6 for the MH-60S problem. The 2005 *Raw Index* was .8682, and the 2010 *Raw Index* was .9631. If you take the 0.9631(second indice) and divide by 0.8682 (the first indice), you get an answer of 1.1093, which is exactly what the JIC indice is in the aforementioned answer. As a rule of thumb, always take the second index used and divide by the first index to find the single index used by the JIC. The answers using the inflation index table and the answers using the JIC will be very close, but there may be a slight variation due to round off errors of the inflation indices. The inflation indices use four significant digits after the decimal point, while the JIC uses significantly more as an Excel-based spreadsheet.

5.12   Expenditure (Outlay) Profile

In a given year, why are the Weighted Indices different from the Raw Indices? To answer that question, we must first discuss an Expenditure (or Outlay) Profile. As background, each program manager is allotted a certain amount of money by Congress to pay for his or her program, and he/she is only authorized to spend that amount of money. This amount is known as the Total Obligation Authority (TOA). An Expenditure (or Outlay) Profile is the rate at which a given year's money that a PM was authorized to spend was actually

expended, or is expected to be expended. Therefore, an expenditure profile can be either historical or future looking. It is calculated by the Office of the Secretary of Defense (OSD) based on Then Year dollars. The following Expenditure Profile begins in 2010, and outlays to pay for the program's development costs will occur in the following percentages over a four year period, from 2010 to 2013:

FY	Development (%)
2010	53.2/34.5/8.8/3.5

This expenditure profile can be interpreted in two different ways:

- *Historical*: "Out of the total money appropriated for development in FY10, 53.2% *was* expended in FY10, 34.5% was expended in FY11, 8.8% was expended in FY12, and 3.5% was expended in FY13."

- Or the expenditure profile can be *forward-looking*: "Out of the total money appropriated for development in FY10, 53.2% *is expected to be* expended in FY10, 34.5% is expected to be expended in FY11, 8.8% is expected to be expended in FY12, and 3.5% is expected to be expended in FY13."

The Expenditure Profile will help you to determine what your *actual* annual payments will be to the contractor in your program. This is accomplished when the expenditure profile is combined with the inflation rate for the years in your program – from that knowledge, you are then able to build your own weighted index for your program or the program you are costing. *It is important to note that each program will have its own unique set of Weighted Indices, based on the inflation encountered in the years of that program combined with the percentages encountered in the expenditure profile*. Example 5.9 will illustrate this procedure by building a Weighted Index from scratch by combining the Raw (compound) Inflation Index and inflation rates with the expected Expenditure Profile. The final outcome is actually a similar concept to compound interest.

Example 5.9 Assume a historical program had a five year outlay profile and the contract had originally been signed for $100M in 2005 constant year dollars (CY05$). How much was actually paid to the contractor over those five years and what were the five annual payments? Table 5.10 shows the initial data set for this example:

TABLE 5.10 Initial Data Set for Example 5.9

A	B	C	D
FY	**Inflation Rate**	**Raw Index**	**TY Outlay Profile (%)**
2005	2.5%	1.0000	34.50
2006	2.8%	1.0280	25.75
2007	2.9%	1.0578	20.00
2008	3.0%	1.0895	15.50
2009	3.0%	1.1222	4.25
			Total 100.00

Based on the information in Table 5.10, we know the following information:

- *Column A*: The program was for five years, from FY2005 to FY2009
- *Column B*: This column shows the inflation rate encountered in each fiscal year from FY05–09
- *Column C*: This column reveals that the initial year for funding in this program was 2005, and 2005 was considered to be the Base Year (since the Raw Index = 1.000). It is a common practice to make the first year of the funding the base year. The remainder of the column shows the Raw Index for each of those years derived from that year's inflation rate.
- *Column D*: The Expenditure (or Outlay) Profile reveals that 34.50% of the funds were intended to be paid (expended) in 2005; 25.75% was to be expended in 2006; 20.00% was to be expended in 2007; 15.50% was to be expended in 2008; and the final 4.25% was to be expended in 2009 for a total of 100% of the funds. These percentages represent the percentage of the total expected to be expended each year *in TY dollars*, since TY dollars represent the amount needed when the expenditures are actually made.

Given this expenditure profile and the inflation expected in each fiscal year, what will the five payments be in each respective year? The first thing to consider is that since the outlay profile in Column D is in TY Dollars, these percentages do not reflect what the inflation rate is for each year, so we will need to account for that inflation first. The easiest way to understand the necessary steps in this procedure is to display the final calculations in a summarized table and describe each step that was taken. The final results, then, are displayed in Table 5.11. Note that the first four columns in Table 5.11 are the four columns in Table 5.10.

TABLE 5.11 Amount Needed per Year to Pay for a Five Year Program with the Given Outlay Profile

					F	G	H
A	B	C	D	E	Normalized	TY Outlay	$103.66 Million
	Inflation	Raw	TY Outlay	CY Outlay	CY Outlay	Weighted	Payment by
FY	Rate	Index	Profile (%)	Profile (%)	Profile (%)	Index	Year
2005	2.5%	1.0000	34.50	34.500	0.35763	0.35763	35.7629
2006	2.8%	1.0280	25.75	25.049	0.25966	0.26693	26.6926
2007	2.9%	1.0578	20.00	18.907	0.19599	0.20732	20.7321
2008	3.0%	1.0895	15.50	14.226	0.14747	0.16067	16.0674
2009	3.0%	1.1222	4.25	3.787	0.03926	0.04406	4.4056
			100.00	96.469	1.000	**1.03660**	**103.6605**

- *Column A:* The five years of the program
- *Column B*: The inflation rates during the respective years
- *Column C:* The raw indices based on the inflation rates (Base Year = 2005)
- *Column D:* How the funds were expected to be expended in each year, by % of total (in TY dollars)

- *Column E*: Results after removing inflation from the TY Outlay Profile { = Column D ÷ Column C }
- *Column F*: Results after normalizing Column E to sum to 100% { = each Column E percentage ÷ 96.469}
- *Column G*: Results after factoring the raw indices inflation back in to the Normalized CY Outlay Profile { = Column C * Column F }. The weighted indices in Column G now sum to 1.0366. Thus, the total amount needed to pay the $100M contract in full is approximately $100M * 1.0366 = $103.66M, in Then Year dollars. This represents what the actual TY payment percentages were, instead of the initial proposed TY Outlay Profile in Column D.
- *Column H*: Payment by Year { = $100M * Column G }

Summarizing Table 5.11, once you determine how much you are authorized to pay to the contractor each year (percentage-wise) via the TY Outlay Profile in Column D, you must first take into account the effects of inflation into those TY percentages, and adjust the percentages accordingly. The results are found in Column E, where the initial TY Outlay Profile percentages had the Raw Index percentages removed from them, producing the Expenditure Profile. This represents how much the percentages (and thus the money) in Column D was actually worth after inflation was removed. But now, if you sum the five percentages in Column E, we find that the five percentages only sum to 96.469%, which means that the money that we are expending is really only paying 96.469% of the total that we need to, as we "lost" 3.531% of our buying power due to inflation (100% − 96.469% = 3.531%). To make up for this 3.531%, we must then normalize our data and adjust our amounts paid back to 100% by dividing each of the percentages in Column E by 96.469%. The results of that normalization are found in Column F, and these results represent the outlay profile in CY05$. Finally, in Column G we multiply the raw indices for each year back in again, producing a Weighted Index for this program to determine the annual payments needed.

Note the difference between Column D (the original *intended* payments schedule) and Column G (the *actual* payments schedule). Column G can be interpreted as the percentage of the total payments that you will now pay in each year. Originally, from Column D, the PM thought that he/she would be paying 34.5% of the total needed in 2005, but in actuality, the 2005 payment turned out to be 35.763% of the total needed; in 2006, it was 25.75% vs. 26.693%, etc. Table 5.12 summarizes these differences between the initial TY Outlay Profile (Column D) and the final TY Outlay Profile (Column G), and also provides the total amount paid in each year to the contractor (in millions):

TABLE 5.12   Differences in Initial and Final TY Outlay Profile

A	D	G	H
FY	**Initial Per Year TY Outlay profile (%)**	**Final Per Year TY Outlay profile (%)**	**Payment by Year (M$)**
2005	34.50	35.7629	35.76
2006	25.75	26.6926	26.69
2007	20.00	20.7321	20.73
2008	15.50	16.0674	16.07
2009	4.25	4.4056	4.41
	100.00	**103.6605**	**103.6605**

Ultimately, in Example 5.9, we ended up paying a total of $103.66M, broken down by years in Column H. Note that paying a higher percentage of costs in the earlier years of the program will save a significant amount of money in the later years. This is akin to when you purchase a car – the larger the down payment you pay of the initial cost up front, the less money you have to finance (or be loaned), thus reducing the total interest that you will have to pay over the course of your loan. Example 5.9 demonstrates how to make your own weighted index, based on combining the inflation rates for each year of your program along with the expected outlay profile for the program. As a program manager, this is an important skill to know; as a cost estimator, you may rarely have the need to do this, but if you do, Example 5.9 provides a complete guide on how to do so. If you do not know the outlay profile of a program, or if the program is still in its infancy or the outlay profile is as yet undetermined, then the Weighted Indices found in the DoD inflation indices or the JIC provide a perfectly good starting point for your estimate.

Summary

In this chapter, we commenced with the first of numerous quantitative areas of cost estimation. Data Normalization involves taking raw cost data and applying adjustments to that data, to gain consistent, comparable data to be used in your estimates. It is necessary to normalize data for content, quantity, and inflation. While we discussed briefly the normalizing required for content and quantity, the majority of the chapter was spent looking at the number of ways that the cost estimator must normalize cost data for inflation. We discussed what an index is, showed how to make an index (e.g., fuel), and calculated a number of examples using it. We discussed how the DoD has many different Appropriation categories/tables, what types of dollars there are (FY$, CY$, BY$, and TY$) and then learned how to convert these dollars from one type of dollars to another. The four conversions are CY$ to CY$, CY$ to TY$, TY$ to CY$ and TY$ to TY$. The key point to remember is that whenever making conversions, *you must always convert to the base year first*, and *then* convert to the year in which we are interested. We also learned how to perform these conversions using the Joint Inflation Calculator (JIC). We concluded the chapter with an example showing how an Expenditure Profile is combined with the annual inflation rates and the raw index to produce a Weighted Index, and given those numbers, how to calculate what you will need to pay/budget for your program annually. This chapter also signified a shift in information, in that the majority of background information needed in cost estimation has already been discussed in the first four chapters. Chapter 6 will cover a brief review of statistics, and which of the primary statistical metrics that cost estimators use.

References

1. Defense Acquisition University (DAU), Weapon Systems Acquisition Reform Act of 2009 (WSARA 09), DTM 09-027.

2. DAU, Defense Acquisition Guidebook, Chapter 1: DoD Decision Support Systems.

Applications and Questions:

5.1 There are three broad categories for normalizing data. What are these three categories?

5.2 When normalizing for quantity, we always want to find an _____ _____ to use as a data point for comparisons.

5.3 What is inflation?

5.4 When developing an index, the base year *always* has an index of what number?

5.5 In Section 1.5, Example 5.3, what would the inflation indices look like if the Base Year selected was 2002 instead of 2004? (Index for 2002 is now 1.000)

5.6 In Section 1.5, Example 5.3, what would the inflation indices look like if the Base Year selected was 2007 instead of 2004? (Index for 2007 is now 1.000)

5.7 DoD inflation indices are developed *for each service* for a particular activity or type of procurement and are called _____?

5.8 Which index do the CY$, BY$, and FY$ calculations and conversions use in the DoD Inflation Indices?

5.9 Which index do TY$ conversions use in the DoD Inflation Indices?

Statistics for Cost Estimators

6.1 Introduction

In Chapter 5 on Data Normalization, we finally got the chance to work with numbers and do some mathematics, and in this chapter, we will continue that trend. We will first discuss how cost estimators use statistics, and we will provide background and examples for the mean, median, variance, standard deviation, coefficient of variation and other statistics used when analyzing and describing a data set. This is important material for background, as this information will be used to pave the way for the statistics that we will encounter in the next chapter on Regression Analysis such as standard error, the F-Statistic and t-statistics. We will define what these terms really mean, and more importantly, why we should care! We will then discuss the measures of central tendency and the dispersion statistics that are important in our field. This is not intended to be an all-inclusive or detailed chapter on every statistical method and probability distribution that exists. Rather, it is intended to be an overview of statistics and a common sense approach to how cost estimators can use them. We will provide background and examples on how statistics are used to support a claim and a review of these key terms.

6.2 Background to Statistics

In today's world, we are constantly being bombarded with statistics and statistical information. Examples include: customer surveys, medical news, political polls, economic predictions, military equipment performance, and personnel status. How can we make sense out of all this data? More importantly, how do we differentiate valid from flawed claims? Statistical analysis plays an important role in virtually *all* aspects of business and economics, and it is present throughout numerous modeling approaches that impact the military as well. We will discuss three examples that all use statistics in different ways to display the power and purposes of using statistics.

6.3 Margin of Error

Example 6.1 Politics I: Let us consider the following political example and discuss the statistics used during an upcoming political election. Let's say that the US presidential election will be held in one week. On the news today, the TV analyst says that a recent poll revealed that "If the election were held today, Candidate A would win with 55% of the vote." But what is the next comment that the analyst will always add? He or she will mention that Candidate A will receive 55% of the vote "but with a margin of error of ±3% or ±5%," or something similar to that. Why do we have this "margin of error," and what is the difference between a 3% margin of error and a 5% margin of error? Let's explore these two questions.

In the United States, there are roughly 146 million registered voters in a national Presidential election [1]. When the poll in the previous paragraph was conducted on which candidate a registered voter would vote for, were all 146 million voters actually interviewed? The answer, of course, is a resounding "No." Why not? The reason is that it would take a significant amount of time and money to be able to poll every single voter in the United States, and it is not feasible, either time-wise or cost-wise, to do so. Instead of interviewing all 146 million voters, the pollsters will interview a "sample" of voters. This sample is merely a subset of the population, and the information gained from the sample will be used to represent the views, or statistics/information, of the entire population of voters. Let's say that the pollsters interviewed 10,000 registered voters. In this case, the size of the "entire population of voters" is 146 million voters, and the size of the "sample" is 10,000 voters. The key idea to remember is that because we did not interview all possible voters and only took a sample, there is going to be *uncertainty* in our estimation of the true portion of the population that favors Candidate A. One of the primary uses of statistics is to quantify this uncertainty. This leads us to two primary areas to discuss: (1) ensuring that the sample we take is "unbiased," so that our conclusions will be representative of the population that we are trying to describe, and (2) quantifying the "margin of error," so that we have a better idea of just how "good" our answer is.

In order to be certain that our sample is "unbiased," it is important to ensure that the voters who were polled are representative of the population as a whole. If you were to interview all 10,000 voters from a city such as San Francisco, CA, this would be a very "biased" sample, because San Francisco is known to be a very liberal city and most of the voters polled would vote for the Democratic nominee. Conversely, if you were to interview all 10,000 voters from the state of Texas – which is generally regarded as a conservative state – then most of the voters polled would vote for the Republican nominee. Neither of these two samples could be considered "unbiased," and the resulting conclusion of your poll would not be representative of the population as a whole. Therefore, to ensure an "unbiased" sample, you would need to conduct polling that would interview a wide cross-section of registered voters from many different cities and states.

Now let us answer the question about the difference between a margin of error of ±3% or ±5%. This difference is quantified by the size of the sample that you took. Let's say that the 10,000 voters that you interviewed gave you the margin of error of ±3%. But if you had only interviewed 1,000 voters, a much smaller sample, then you would have a larger margin of error, say ±5%, due to the fact that you interviewed fewer people. The closer your sample size is to the actual population size, the smaller the margin of

error that you will have. As your sample number approaches your population number, the uncertainty in your answer decreases.

The numbers in the previous paragraph were for illustrative purposes only. In actuality, the difference between a 3% and 5% margin of error involves much smaller numbers. There are numerous calculators that you can find on the Internet that will calculate the sample size needed for any given population size and desired margin of error. (Search for "Sample Size versus Margin of Error calculator" [2]. The following example illustrates the actual numbers for population and sample size on the example in the previous paragraph. These are based on a Confidence Level of 95%, which means that on average, 95% of the confidence intervals (i.e., the set of intervals generated by each sample percentage, plus or minus its 3% margin of error) will include the true population value.

$$\text{Population size: } 150,000,000 \text{ voters, and a confidence level} = 95\%$$

- If you desire a Margin of Error of 1%, then the necessary sample size is: *9,604 voters*

- If you desire a Margin of Error of 2%, then the necessary sample size is: *2,401 voters*

- If you desire a Margin of Error of 3%, then the necessary sample size is: *1,068 voters*

- If you desire a Margin of Error of 4%, then the necessary sample size is: *601 voters*

- If you desire a Margin of Error of 5%, then the necessary sample size is: *385 voters*

Thus, if you desire a margin of error of 3%, it is necessary to interview/poll 1,068 voters. If you are happy with a 5% margin of error, then only 385 voters are necessary to poll. So you can see that as sample size increases, the Margin of Error decreases. This means that you are closer to predicting correctly how the total population of voters would vote.

As a side note, the Margin of Error calculator reveals that sample sizes required for desired margins of error change very little once your population exceeds a size of 20,000.

Example 6.1 demonstrated one way that statistics can be used to determine the sample size required for polling and predicting outcomes at a certain level of confidence. Now consider a second example, one significantly different in purpose.

Example 6.2 Politics II: This election example will illustrate another way to use statistics. Let's suppose that Candidate A has been the Mayor of a city for four years and is now running for re-election. During the election campaign, the opponent, Candidate B, states that the present mayor is very "soft on crime." This can be illustrated by the fact that "*the burglary rate has doubled in the four years that the mayor has been in office.*" This statistic may be factually true! But is it a statistic that we should be concerned and alarmed about? What other details do we not have and that Candidate B is not providing that would give us a clearer picture as to whether the mayor is indeed "soft on crime?"

What you are missing is the actual number of burglaries that have occurred in those four years to determine if we should be concerned or not. If the burglary rate has increased from 300 to 600 burglaries, then the burglary rate has doubled and is certainly a situation that we *should* be concerned about, and maybe the present mayor is *indeed* "soft on crime." But what if the number of burglaries had been a total of one, and increased to a total of two during the four years of Candidate A being mayor? Once again, the statistic that the burglary rate has doubled is correct, but is increasing from one to two over a four year period actually a "doubling rate" that we should be concerned with? Clearly, we would be concerned with an increase from 300 to 600 burglaries or homicides, but much less so if it was merely from one to two. Again, the statistic of "doubling" is factually correct in both

of these examples, but is only troubling in the first of the two scenarios. This illustrates that the statistic was used to help the claim by Candidate B, but further investigation into that statistic was necessary to deem whether it was one with which to be concerned and accurately telling us the whole story. The statistic that Candidate B was *not* telling us was the one that we needed to do more investigation on.

Example 6.3 Nutrition   A third use of statistics can be found in the following nutrition example. Let's say that you are out shopping at the local grocery store, and one of the vendors is selling a product that has "only 5 grams from fat per serving." Is this good or bad?? How bad can it be for you, given that it has only five grams from fat? The answer of course is that … …..."*It depends.*" Let's look inside the numbers and do the math.

Numerous nutritionists recommend that your diet consist of approximately 50% of your calories from carbohydrates, 15–25% from protein, and 20–30% from fat. These numbers can vary slightly depending on whether your goal is trying to build lean muscle while lifting weights, or if you are trying to just lose some weight. But overall, numerous nutritionists will confirm these approximate numbers [3]. Thus, consuming products where over 30% of the calories are from fat should be kept to a minimum.

Going back to the product that has only five grams from fat, the first thing we need to do is look at the nutrition label. In order to make sense of this label that is required by the FDA on every product, one needs to know the following conversions:

1 gram of fat $=$ 9 calories

1 gram of protein $=$ 4 calories

1 gram of carbohydrates $=$ 4 calories

Focusing just on the fat intake of this product, since this product has five grams from fat, it actually means that it has:

5 grams from fat $\times$ 9 calories/gram of fat $=$ 45 calories from fat per serving.

Determining whether 45 calories from fat is "good" or not depends on the number of calories in a serving. Let's assume that this product is 65 calories per serving. Then doing the conversion, this product is actually:

45 cal from fat $\div$ 65 calories total per serving $=$ 69.23% calories from fat per serving!

Clearly, this is a product that is not good for you in terms of being "healthy." So while "Only 5 grams from fat" does not sound too bad by itself in the original claim, when you calculate the percent fat in a serving, you find that 69% of the calories are from fat in each serving. This statistic would more often than not steer you away from purchasing this product.

However, if the total calories in a serving were 450 calories instead of 65 calories, then the percent of calories due strictly from fat is only:

45 cal from fat $\div$ 450 calories total per serving $=$ 10% of calories from fat per serving

Clearly, this is much better for you, fat-wise, though it is now 450 calories total per serving instead of 69 calories total!

Why do we bring up Examples 6.2 and 6.3? The answer is that it shows you why you should be wary of *any* claims using statistics. You can always find a few good statistics on any product, or any program, but the person who is "selling" the product or program is only going to use the statistics that help them. In fact, they will use the statistic that helps their cause the most. So when listening to someone selling a product, what you really need to figure out is "Which statistics are they NOT telling us?"

"How to Lie With Statistics" was a classic book written by Darrell Huff. My favorite quote from that book is the following: "The crooks already know these tricks. Honest men must learn them in self-defense." [4] Believe it or not, this book was written in 1954! That message is as true today as it was back then. Okay, so hopefully wiser and more skeptical about statistics with these three examples behind us, we proceed with the chapter material.

A *statistic* is a number, such as a mean or a variance, computed from a set of data. They can be expressed in numerous ways, such as a real number, a percentage, or a ratio, as shown in Examples 6.1 through 6.3. The objective of statistics is to describe a data set and make *inferences* (predictions and decisions) about a *population* based upon information contained in a *sample*. In Example 6.1, we inferred that Candidate A would win the election due to the results of the sample that we conducted.

Rarely do we use the whole population. Example 6.1 considered a voting population of 146 million while taking a sample of 10,000 voters. When do we ever actually poll the entire population? One of the times would be for the US Census, which is conducted every 10 years to determine many demographics about all people living in the United States, including the actual total population. But it takes much time and money and personnel to conduct the Census, and thus the Constitution requires that it be conducted only once every ten years. For emphasis, consider that the 2010 Census cost $13 billion to conduct [5].

The remainder of this chapter will discuss important terms to describe data sets such as the measures of central tendency (mean, median, and mode), the dispersion statistics (range, variance, and standard deviation), and the coefficient of variation. To illustrate these terms succinctly, a home buying example will be provided that we will use to follow throughout the remainder of this chapter.

6.4 Taking a Sample

Consider a scenario where you are changing jobs and moving to a new state where you would like to purchase a new home. Let's say that the town has a population of 30,000 people. To house those 30,000 people, with an average of 2 to 3 people per home, let's say that there will be 12,500 homes in that town.

Since you would like to purchase one of these homes, you will of course need to know the prices. To find out what the *exact* cost a home would be, you could comb the Internet and find a real estate website that would contain the cost of every home in that town. However, determining the exact cost of each of the homes and then adding them up to be able to get an average price for the town would take a significant amount of time, money, and effort to do. So, instead of considering all 12,500 homes in the population, we will take a sample of say 1,000 homes. Since we have only considered the cost of 1,000 homes in a population of 12,500, there will be some *uncertainty* in our final answer.

(In this case, using a sample error calculator, with a population of 12,500 homes and a 95% confidence level, a sample of 984 homes would provide a margin of error of ± 3%) [2].

After surveying the sample of 1,000 homes, you derive that the average cost of a home is $350,000, with a standard deviation of $25,000 (to be discussed). Thus the following statistics can be summarized:

- *Population of Homes (N)*: 12,500
- *Sample Size of Homes Used (n)*: 1,000
- *Sample Mean/Average*: $350,000
- *Sample Standard Deviation*: $25,000

Since statistical inference is the process of making an estimate, prediction, or decision about a *population based on a sample*, then the average cost of all 12,500 homes is inferred to be $350,000, based on the 1,000 homes that were surveyed. This is because the *sample* mean is an unbiased estimator of the *population* mean.

Once statistics are computed, methods of organizing, summarizing, and presenting data can be performed in a numerous different ways. These methods include graphical techniques and numerical techniques. The actual method used depends on what information we would like to extract. Are we seeking measures of central location, like the mean, median or mode? Or are we seeking the measures of variability and dispersion? These terms will now be defined.

6.5 Measures of Central Tendency

These statistics are ones that we are most familiar with, and describe the "middle region" of our data set.

- Mean (Average): The arithmetic average of the data set.
- Median: The "middle" of the data set.
- Mode: The value in the data set that occurs most frequently.

Mean: The Mean, or Average, is the arithmetic average of a data set. It is calculated by taking the sum of the observed values (y_i) divided by the number of observations (n).

Example Data Set #1: 12, 8, 4, 15, 17, 21, 7 ($n = 7$)

The sum of $12 + 8 + 4 + 15 + 17 + 21 + 7 = 84$. Therefore, the mean of this data set is: $84 \div 7 = 12$.

Median: The Median is the *middle* observation of a data set. Numbers must first be *ordered* from the lowest to highest value to determine the median.

Example Data Set #1: 12, 8, 4, 15, 17, 21, 7

After ordering this data set from lowest value to highest value, we get the following:

4, 7, 8, 12, 15, 17, 21

Since the number of data points in this set is an *odd* number ($n = 7$), you disregard the first three numbers and the last three numbers and the remaining middle observation is 12. Thus, the median of this data set is 12.

If your data set has an *even* number of data points, you must take the average of the two observations at the center. Consider the following data set. Note that the number of data points is an even number now ($n = 8$), as we added the number 13 to the original data set.

Example Data Set #2:   4, 7, 8, 12, 13, 15, 17, 21

Here, there is no "middle" observation, so we remove the first three and the last three numbers and what remains is the 12 and 13. We must now average those two numbers, so $(12 + 13) \div 2 = 12.5$. Thus, the median of this data set is 12.5.

Mode: The Mode is the value of the data set that occurs most frequently. Consider this data set:

Example Data Set #3:   4, 7, 8, 12, 12, 15, 17, 21

Here the mode is 12, since 12 occurred twice and no other value occurred more than once. Data sets can have more than one mode, while the mean and median have one unique value. Data sets can also have NO modes. For example, the original data set was:

Example Data Set #1:   4, 7, 8, 12, 15, 17, 21

Since there is no value that occurs more frequently than any other, no mode exists in this data set. (Note: You could also argue that this data set contains seven modes, since each value occurs as frequently as every other, but we consider there to be no modes).

As a general rule, if the data are symmetric and uni-modal, the mean, median, and mode will be the same. If the data are skewed, the median may be a more reliable statistic to use because the mean could be unrepresentative of the data, due to outliers. The mean is greatly influenced by outliers, as shown in the following mean/median example.

Mean/Median Example: If you are purchasing a home, your real estate agent will most likely discuss the prices of homes in terms of median instead of the mean. This is due to the outliers that can occur in real estate. Consider the following data set of five home prices from a neighborhood that you are interested in:

- Home #1: $200,000
- Home #2: $250,000
- Home #3: $300,000
- Home #4: $350,000
- Home #5: $400,000

The mean and the median for these five homes are the same. They are both = $300,000. But let's add a new home that just went on the market and the asking price is $1.5M.

- Home #6: $1,500,000

Now the median = ($300,000 + $350,000) ÷ 2 = $325,000, and the mean is now $500,000! The mean has changed significantly to $500,000, and no longer accurately reflects the average cost of the homes in that neighborhood anymore, due to the effect of the significantly higher price of Home #6. However, the median is still very representative of the neighborhood at $325,000. This is why your real estate agent will most likely discuss the prices of homes in terms of median instead of the mean. Home #6 would be considered an *outlier,* since it is significantly different (in this case, much higher) from the rest of the data points in the set. This outlier will skew your data to the right and, again, can make the mean unreliable as a predictor.

These three statistics are almost never the same, unless you have a symmetric, uni-modal population (i.e., "just one hump" or "peak" in your data, an example of which is the "bell curve," also called the standard normal distribution).

Figure 6.1 is representative of the original data set of five homes in the Mean/Median example. However, data can be "skewed" toward one side or another, which would look more like the shape found in Figure 6.2. Figure 6.2 is representative of the Mean/Median example after Home #6 (the outlier) has been added.
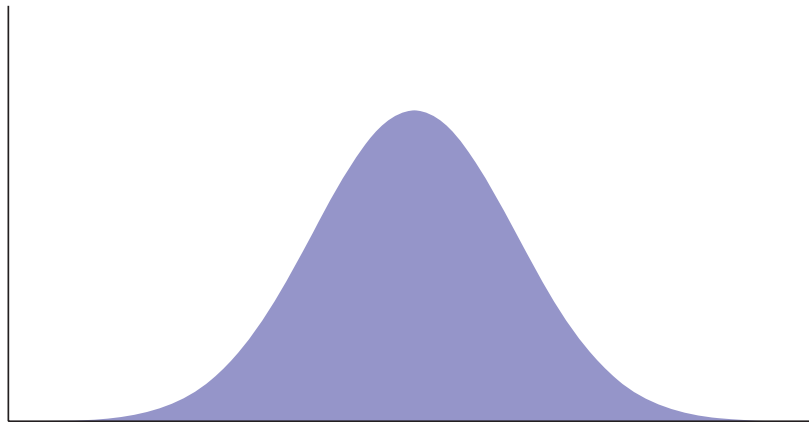
FIGURE 6.1 Symmetric "Bell Curve" Distribution of Data.
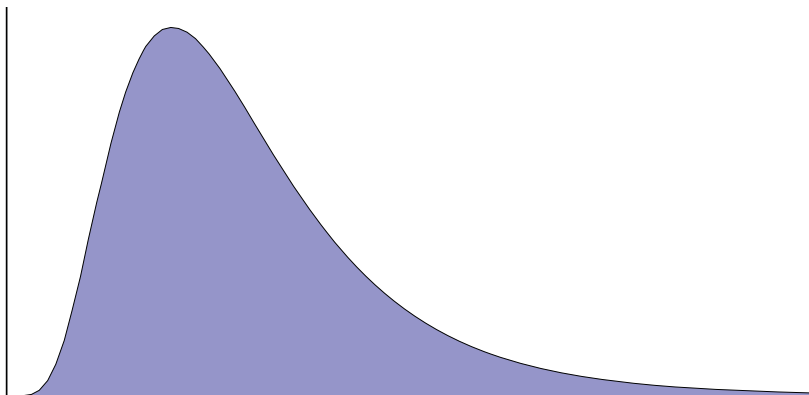
FIGURE 6.2 Non-symmetric "Skewed" Distribution of Data.

6.6 Dispersion Statistics

The mean, median, and mode by themselves are not sufficient descriptors of a data set. The reasons are shown in the following example:

- Data Set 1: 48, 49, 50, 51, 52
- Data Set 2: 5, 15, 50, 80, 100

Note that the mean and median for both data sets are identical at 50, but the data sets are glaringly different! The difference is in the *dispersion* of the data points. The dispersion statistics that we will discuss are the Range, Variance, and Standard Deviation.

Range: The range is simply the difference between the smallest and largest observation in a data set. Using the same data set as above,

- *Data Set 1*: 48, 49, 50, 51, 52: the range is: 52 − 48 = 4, and
- *Data Set 2*: 5, 15, 50, 80, 100: the range is: 100 − 5 = 95.

So, while both data sets have the same mean and median, the dispersion of the data, as depicted by the range, is much smaller in Data Set 1 than that in Data Set 2. While the range can be helpful while analyzing your data set, the most important measures of dispersion are the variance and standard deviation. To discuss these, recall the Section 6.4 "Taking a Sample" home example, and the following statistics that were calculated:

- *Population of Homes (N)*: 12,500
- *Sample Size of Homes Used (n)*: 1,000
- *Sample Mean/Average*: $350,000
- *Sample Standard Deviation*: $25,000

Variance: There are two types of variances we will discuss: Population Variance and Sample Variance. The Population Variance, σ^2, is a measure of the amount of variability of the *entire population* of data relative to the mean of that data. As shown in Equation 6.1, it is the average of the squared deviations of the observations about their mean.

$$\sigma^2 = \frac{\sum (y_i - \mu)^2}{N} \tag{6.1}$$

From our example, the population variance is the variance calculated from the entire population of 12,500 homes. The numerator in this equation is merely the sum of the "squared differences" between the actual costs of each of the 12,500 homes from the average of the 12,500 homes, if we calculated all of these squared differences. For the calculations in our example, y_i is the actual cost of a home; the mean, μ, is the average cost of all 12,500 homes. Thus, there will be 12,500 homes (or 12,500 y_is) that will be included in

this calculation. So what we will have is the following:

$$(\text{Cost of home } Y_1 - \text{population mean})^2 + (\text{Cost of home } Y_2 - \text{population mean})^2$$

$$+ (\text{Cost of home } Y_3 - \text{population mean})^2 + \dots + (\text{Cost of home } Y_{100}$$

$$- \text{population mean})^2 + \dots + (\text{Cost of home } Y_{1000} - \text{population mean})^2$$

$$+ \dots\dots + (\text{Cost of } Y_{12,500} - \text{population mean})^2$$

This, then, is the "price of each home minus the mean," then that difference is squared, for each home. After these numbers are summed, we divide by the entire population of 12,500 homes ($N = 12,500$) to get the *population variance*.

Note: Why do we "square" the differences in the numerator in Equation 6.1? If we did not square the differences in the numerator, the sum of all of these differences (also called "residuals") will always be equal to zero. Some of the costs of the homes will be greater than the mean, while others will be less than the mean, and correspondingly, we will have differences that are positive and differences that are negative. The sum of these positive and negative differences will be equal to zero. (This is not hard to prove, but we do not do it here). Therefore, by squaring each difference first, and then summing, we are ensuring a non-negative number in the numerator of the variance. The numerator would be equal to zero only in the special (and unusual) case where all of the original numbers are the same as the mean.

Another way to look at why we square the differences in the numerator is to assume that we "penalize" those data points that are further away from the mean. If the differences between the actual data points and the mean are small, then squaring that number also results in a small number. But if the difference in the actual cost of the home minus the mean is a large number, then by squaring that large number, we produce an even *greater* number, which adds significantly to the variance. So in essence, you are penalizing those numbers that are further away from the mean, those which add greatly to the variance and standard deviation.

The Sample Variance, s^2, shown in Equation 6.2, is used to *estimate* the *actual* population variance, σ^2, by using a sample data set. In this example, the sample is the data for the 1,000 homes that we are using to estimate the 12,500 homes.

$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n - 1} \tag{6.2}$$

Therefore, we have $n = 1,000$, and the mean, denoted by $\overline{Y}$, was calculated in our example to be \$350,000. The sample variance, s^2, is calculated by:

$$(\text{Cost of home } Y_1 - \$350,000)^2 + (\text{Cost of home } Y_2 - \$350,000)^2$$

$$+ \dots + (\text{Cost of home } Y_{100} - \$350,000)^2$$

$$+ \dots + (\text{Cost of home } Y_{500} - \$350,000)^2$$

$$+ \dots\dots + (\text{Cost of home } Y_{1000} - \$350,000)^2$$

This, then, is the sum of "the price of a home minus the mean of \$350,000, then squared," for each of the 1,000 homes.

There is one significant difference between Equation 6.1 and 6.2. Note that the denominator for the population variance in Equation 6.1 was $N = 12{,}500$, whereas in the sample variance equation it is no longer for the entire sample of $n = 1{,}000$, but rather one minus that number, $n - 1 = 999$. Why do we divide by the entire population of 12,500 in the *population variance*, yet in the *sample variance*, our denominator is actually one less than the sample size? The mathematical answer has to do *with Degrees of Freedom*, but in practical terms, how does the sample variance compare to the population variance? The short answer, for which a proof is available in any mathematical statistics text, is that the long-run average of the sample variances is the population variance. As an illustration, consider the following example:

Let's suppose that we have the cost of all 1,000 items in a population. With those 1,000 items, we can calculate a population variance, σ^2, using the computations shown earlier. Now let's suppose that we take a sample of 50 items from that 1,000. We will then be able to calculate a sample variance, s^2, from those 50 items. If we then take a large number of samples of size 50, we will be able to calculate a sample variance for each sample. Some of those sample variances will be above the population variance, and others will be below the population variance. After taking a large number of samples then, the average of all of the sample variances is theoretically going to be equal to the value of the population variance, and in practice it would be very, very close to that number the more samples we take. Thus, when you use the $n - 1$ denominator in the sample variance, the expected value of s^2, or $E(s^2) = \sigma^2$.

Mathematically, we are raising the value of s^2 (from what it would have been had we divided by n) so that it is an unbiased estimator of σ^2. Note that when n is small, (e.g., a sample size of $n = 10$), subtracting one to get $n - 1 = 9$ is a significant event, as it raises your sample variance greatly! This reinforces the idea that the smaller your sample size is, the greater the variance you will have – again, due to increased uncertainty. However, as the sample size increases, subtracting the 1 from n units in the denominator is of much smaller significance.

Standard Deviation: Understanding the concept of variation is essential to understanding uncertainty, but we still rarely use variance as a statistic. Why not? The reason is that variance is not a "common sense" statistic, since it describes the data in terms of squared units. In the previous example, we have the cost of the home minus the mean cost of the home, but then we square that difference, so the result is in "squared dollars." It is hard to spend "squared dollars" in your local Home Depot! So to convert the variance to a unit that we can actually use, we simply take the square root of the population variance, and the result is called the Population Standard Deviation, denoted by σ. This formula is shown in Equation 6.3 and is merely the square root of Equation 6.1.

$$\sigma = \sqrt{\frac{\sum (y_i - \mu)^2}{N}} \tag{6.3}$$

Similarly, the Sample Standard Deviation, s, found in Equation 6.4, is used to estimate the *actual* population standard deviation, σ. The sample standard deviation, s, is measured in the same units as the original data from which the standard deviation is being calculated (in this case, dollars). Equation 6.4 is merely the square root of Equation 6.2.

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}} \tag{6.4}$$

Why is standard deviation important? It shows the amount of dispersion from the mean that there is in the data set. When data are normally distributed, recall from your probability courses that you would expect the following (approximately):

- ±1 standard deviation from the mean captures ~68% of the data
- ±2 standard deviations from the mean captures ~95% of the data
- ±3 standard deviations from the mean captures ~99.7% of the data
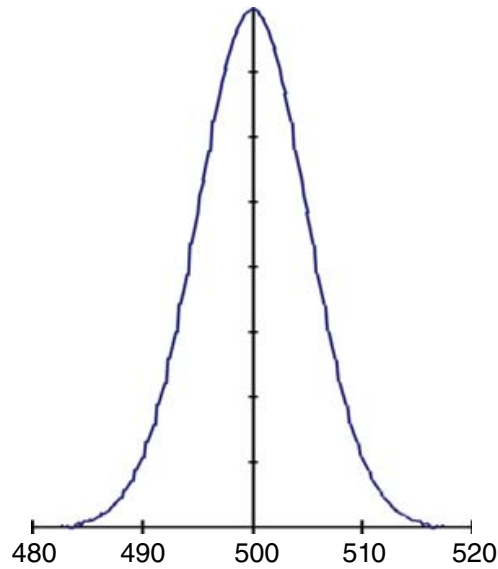
FIGURE 6.3 Graph of Small Standard Deviation (Mean = 500, Std Dev = 5).

Figure 6.3 is an example of what a graph looks like with a mean and a small standard deviation (in this case: mean = 500 and standard deviation = 5). Figure 6.4 is an example of what a graph looks like with the same mean but with a larger standard deviation (in this case: mean = 500, standard deviation = 10). Note how much wider and "flatter" Figure 6.4 is than Figure 6.3. The larger standard deviation implies a much greater uncertainty.
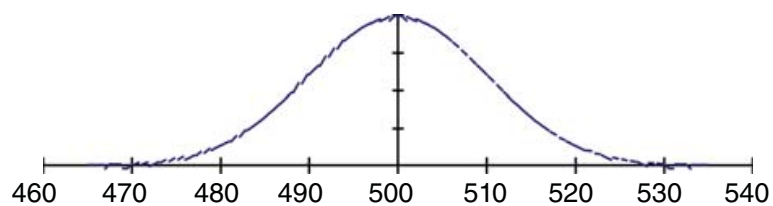
FIGURE 6.4 Graph of Large Standard Deviation (Mean = 500, Std Dev = 10).

Example 6.4 Let us examine two scenarios concerning standard deviation. In Scenario 1, we have the following from the home buying example data:

Scenario 1: Sample Mean = $350,000
           Sample Standard Deviation = $25,000

So in this 1,000 home example, we would expect the following:

- 68% of the homes ($n = 680$) would fall between: $350,000 \pm (1 \times \$25,000) =$ *$325,000 to $375,000*

- 95% of the homes ($n = 950$) would fall between: $350,000 \pm (2 \times \$25,000) =$ *$300,000 to $400,000*

- 99.7% of the homes ($n = 997$) would fall between: $350,000 \pm (3 \times \$25,000) =$ *$275,000 to $425,000*

Why is a smaller standard deviation better than a larger standard deviation? If we increase the standard deviation in this example from $25,000 to $50,000, we would now expect the following results:

Scenario 2: Sample Mean = $350,000
Sample Standard Deviation = $50,000

- 68% of the homes ($n = 680$) would now fall between: $350,000 \pm \$50,000 =$ *$300,000 to $400,000*

- 95% of the homes ($n = 950$) would now fall between: $350,000 \pm (2 \times \$50,000) =$ *$250,000 to $450,000*

- 99.7% of the homes ($n = 997$) would now fall between: $350,000 \pm (3 \times \$50,000)$ = *$200,000 to $500,000*

It is easy to observe that the data in *Scenario 1* has less dispersion and that the ranges of cost are smaller. In *Scenario 2*, with the larger standard deviation, the costs have a greater dispersion and the ranges are larger. Note that the data for one standard deviation in *Scenario 2* is identical to the two standard deviation data in *Scenario 1*.

A military example that would drive home the importance of a smaller standard deviation is in the accuracy of weapons. If an indirect fire weapon has a range of 500 meters (m) and a standard deviation of 25m, that is far more accurate than one with an identical mean of 500m and a standard deviation of 50m. If you do the math here similarly as in Example 6.4, the 25m standard deviation in an indirect fire weapon system clearly offers a significantly greater level of accuracy, thus providing significantly more safety to the infantry and artillery troops involved in the battle or exercise.

What does the standard deviation actually represent? In Example 6.4, we would say that the sample standard deviation of $25,000 represents the *average estimating error* for predicting subsequent observations. In other words: *On average, when estimating the cost of homes that belong to the same population as the 1,000 homes above, we would expect to be off by $\pm\$25,000$ in future calculations.*

6.7 Coefficient of Variation

The last term that we will discuss in this chapter is the Coefficient of Variation. If you had a data set and calculated the standard deviation for that data set to be $100,000, would that be good or bad? Does $100,000 sound high? The answer here – as in many other scenarios – is ...... "*It depends!*" But what does it depend on?

 (a) A standard deviation of $100,000 for a task estimated at $5 Million would be very good indeed.

 (b) A standard deviation of $100,000 for a task estimated at $100,000 is clearly useless!

In (a) above, in normally distributed data, we would thus conclude that 68% of the time, our cost estimate would fall between $4.9 Million and $5.1 Million; 95% of the time it would fall between $4.8 Million and $5.2 Million, and 99.7% of the time it would fall between $4.7 Million and $5.3 Million. In (b), well, we need new data!

So what constitutes a "good" standard deviation? The "goodness" of the standard deviation is not its actual value, but rather what percentage the standard deviation is of the estimated value it is compared against. For us, the estimated value is almost always the mean of the cost data.

The *coefficient of variation* (CV) is defined as the "average" percent estimating error when predicting subsequent observations within the representative population. The CV is the ratio of the standard deviation (s_y) to the mean, denoted as $\overline{Y}$.

$$CV = \frac{s_y}{\overline{y}} \qquad\qquad (6.5)$$

- In (a), the Coefficient of Variation is: $100,000 \div $5M = 2\%$ (Awesome)
- In (b), the Coefficient of Variation is: $100,000 \div $100,000 = 100\%$ (Yikes!)

Since both of these values are unit-less and in percentage form, they can now be readily compared. The CV is the "average" percent estimating error for the population when using the mean as the estimator, and it can also be regarded as the "average" percent estimating error when estimating the cost of future tasks.

Let's calculate the CV from Example 6.4:

Scenario 3: Standard deviation = $25,000 and $\overline{Y}$ = $350,000
                     CV = $25,000 ÷ $350,000 = *7.14%*

Scenario 4: Standard deviation = $50,000 and $\overline{Y}$ = $350,000
                     CV = $50,000 ÷ $350,000 = *14.28%*

Therefore, for subsequent observations *we would expect our estimate to be off on "average" by 7.14%* for Scenario 1*, or 14.28%* for Scenario 2 when using $350,000 as the estimated cost. Clearly, the lower the CV, the better the estimate!

Author's Note: Having discussed how to solve for the mean, median, mode, range, variance, and standard deviation by hand, there is indeed an easier way to solve for these quickly using the Excel Analysis Tool Pak. In Excel, select Data (above the tool bar); then select Data Analysis; and you will see a variety of Analysis Tools that drop down. Select Descriptive Statistics; highlight the data that you want analyzed, put a check in the box on Summary Statistics, select ok, and voila! You will still have to manually calculate the Coefficient of Variation: CV = Standard Deviation/Mean.

If you select Data and do not have Data Analysis along the top, the toolpak may need to be installed. To do so, open to an Excel worksheet, click on the Microsoft Office button in the top left corner; click on Excel Options at the bottom; select Add-ins on the left and a "Manage Add-Ins" will appear on the bottom, so select "Go"; check the Analysis Toolpak (note: not the Toolpak VBA); hit OK, and it is now installed. Different versions of Microsoft will probably have slightly different directions, but these should get you close regardless of the version. We will show a printout of Descriptive Statistics in Table 7.2 of Chapter 7.

Summary

In this chapter, we discussed how cost estimators use statistics, and we provided background and examples for the mean, median, mode, range, variance, standard deviation, and coefficient of variation when describing a data set. We stressed that when we take a sample, there is going to be *uncertainty* in the answer because we did not look at the entire population. We quantified that uncertainty by using a margin of error table. This chapter then concentrated on the measures of central tendency and the dispersion statistics that are important in our field. Overall, it was intended to be an overview of statistics and how cost estimators can use them, as well as how others can use statistics to support their claims in areas such as politics or when selling a product.

It is important to understand that in the cost estimation field, sample sizes are generally much smaller than what a statistician would encounter if he/she was doing a study on the effects of, say, a new medical drug being taken by numerous volunteers. In the cost estimation field, you may consider yourself lucky if you have more than ten historical programs to choose from, and not thousands in a data base. We generally work more to the 80% confidence level, not 90%, 95%, or 99.9% confidence levels, as others might be used to. We want to give the comptroller paying the bills at least an 80% chance of having sufficient funding to cover the expenses. Armed with this knowledge of statistics and many of its key terms, we are now ready to take on the powerful topic of Regression Analysis.

Author's Note: There are numerous textbooks available on probability and statistics that can be used as a general reference. In the Operations Research Department at the Naval Postgraduate School, a widely used textbook on the subject is written by Jay L. Devore. You may use any probability and statistics textbook that you prefer to reinforce the subject area. We will list the Devore textbook as a "General Reference" below in case you need it.

References

1. United States Census Bureau: People/Voting and Registration
2. Margin of Error Calculator in Example 1
3. Dr Richard Collins, The Cooking Cardiologist blog. "Balance Carbohydrates, Protein and Fat for Better Health", June 23, 2011.
4. Darrell Huff. "How to Lie With Statistics." Introduction, page 9. Copyright 1954.
5. United States Census Bureau

General Reference

1. Devore, Jay L. Probability and Statistics for Engineering and the Sciences, Sixth Edition, Thomson Brooks/Cole, 2004.

Applications and Questions:

6.1 In statistics, we rarely have the time or money to look at the statistics of an entire population. Consequently, we take a subset of that population, which is called a
_____.

6.2 Since we only take a sample of the entire population, there is going to be _____ in our answer.

6.3 This uncertainty can be quantified (e.g.) as "±3%" or "±5%" and is called the _____ ___ _____.

6.4 Your local high school football team played ten games during this past season. The following data set summarizes the number of points scored in each of those ten games:

Game #	Football scores
1	17
2	34
3	10
4	42
5	7
6	24
7	24
8	21
9	10
10	24

What are the three measures of central tendency for this data set (i.e., solve for the mean, median, and mode)?

6.5 What are the three measures of dispersion for the same data set in Question 6.4 (i.e. solve for the range, variance, and standard deviation)?

6.6 What is the Coefficient of Variation for this data set? Would you consider this CV good or bad?

Linear Regression Analysis

7.1 Introduction

In this chapter, we will discuss the robust topic of Regression Analysis. By definition, regression is used to describe a "statistical relationship between variables," and we will discuss in detail why using regression analysis may provide you a better estimate for your system or program, rather than using the descriptive statistics such as the mean and standard deviation that we learned in the previous chapter. Regression depends on analogous, applicable historical data to make its prediction. We will also describe the statistical measurements used in regression analysis, such as standard error, the F-Statistic, the t-Statistics, and coefficient of variation, and how to compare one regression output to another and decide which is "better." We will provide background, examples, and a review of many key terms. We begin the chapter by showing the power and importance of regression analysis with an easy example.

7.2 Home Buying Example

Example 7.1 Consider a scenario where you are changing jobs and moving to a new state (similar to last chapter) and you would like to purchase a home in your new hometown. You've narrowed your choices down to just a few neighborhoods where you would like to live due to preferences such as location, distance to work, and the school district for your kids. Table 7.1 is a list of the prices of the last fifteen homes that were sold in those neighborhoods in which you are looking:

Given the data set in Table 7.1 on these fifteen homes, what type of statistics and information can we gain from them? If we use the Descriptive Statistics package (either in the Excel Analysis Toolpak or another data package), we can calculate the useful statistics found in Table 7.2:

From Table 7.2, we can see that the mean/average home in these neighborhoods is priced at $430,666.67, and we can also see the variance, standard deviation, and the range

---

TABLE 7.1 Prices of the Fifteen Homes Used in Example 7.1

Home #	Price ($)
1	300,000
2	400,000
3	350,000
4	800,000
5	450,000
6	250,000
7	225,000
8	450,000
9	550,000
10	400,000
11	220,000
12	350,000
13	365,000
14	600,000
15	750,000

TABLE 7.2 Descriptive Statistics for the Fifteen Homes in Table 7.1

Descriptive Statistics	
Mean	430666.6667
Standard error	45651.0227
Median	400000
Mode	400000
Standard deviation	176805.6506
Sample variance	31260238095
Kurtosis	0.1949423
Skewness	0.9237138
Range	580000
Minimum	220000
Maximum	800000
Sum	6460000
Count	15

of the prices, from the lowest priced home (minimum) of $220,000 to the highest priced home (maximum) of $800,000.

But what does this price data tell us about the *size* of these homes? How big a house are we getting for the given price? What if you were single and only needed a 1,200 square foot home? What if you were married with two kids and needed a home that was a minimum of 2,000 square feet? The aforementioned data, and specifically the mean of $430,666.67, do not provide any insight about the size of a home to be able to answer these questions. All you know at this point is what the average price of a home is in this sample of fifteen,

TABLE 7.3 Prices and Square Feet of the
Fifteen Homes in Example 7.1

Home #	Price ($)	Square Feet
1	300,000	1400
2	400,000	1800
3	350,000	1600
4	800,000	2200
5	450,000	1800
6	250,000	1200
7	225,000	1200
8	450,000	1900
9	550,000	2000
10	400,000	1700
11	220,000	1000
12	350,000	1450
13	365,000	1400
14	600,000	1950
15	750,000	2100

because we are missing the information that relates the *size of the home* to its *price*. To correct this deficiency, we go back to Table 7.1 and find the size (in square feet) of each of these fifteen homes. This updated information is provided in Table 7.3, now showing each of the homes with its price and its size.

Armed with this new information, let's make a scatter plot of Price vs. Square Feet to see what this data looks like, and what it might be telling us.

From the graph in Figure 7.1, we see that there is a positive correlation between the two variables; that is, as the square footage of the home increases, the price of the home increases, which is the result that we would expect. But if we were to make a prediction on
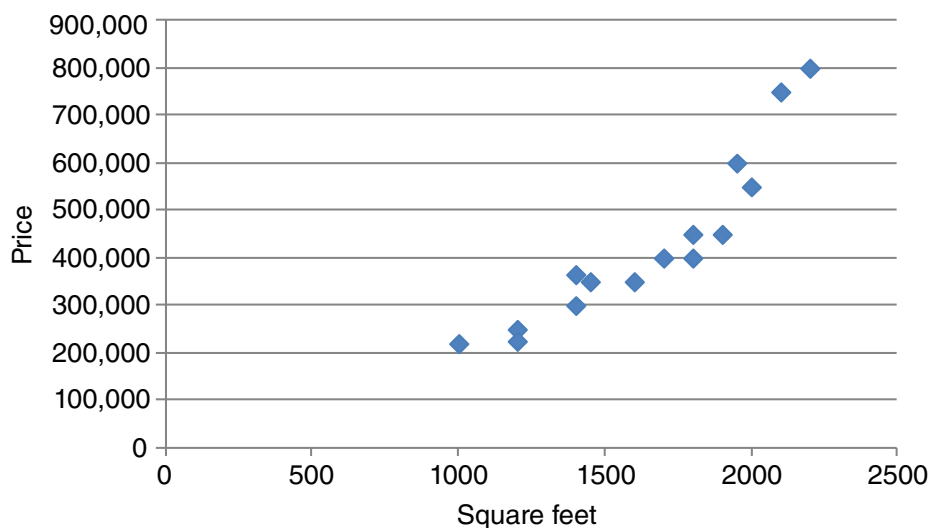
FIGURE 7.1 **Scatter Plot of Price versus Square Feet from Data in Table 7.3.**

the price of any size home using this data, what would that prediction be, and how would we get it? We can do this by using regression analysis. Using a software regression tool, such as the Regression option in the Excel Analysis Toolpak, or the software package JMP, let's perform a Regression Analysis on the data set in Table 7.3. When entering the data, Y is your cost data and is considered the dependent variable; X is your square feet data and is considered the independent or explanatory variable. The results of that regression are found in Table 7.4.

TABLE 7.4 Regression Result of Price versus Square Footage of the Fifteen Homes in Example 7.1

A	B	C	D	E	F
Regression Statistics					
Multiple R	0.91936		**Price vs. Square Feet**		
R Square	0.84523				
Adjusted R Square	0.83332				
Standard Error	72183.1553				
Observations	15				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.69908E+11	3.69908E+11	70.9941	1.26178E−06
Residual	13	67735302725	5210407902		
Total	14	4.37643E+11			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	**−311221.8767**	90000.5658	−3.45800	0.00424	−505656.2777
Square Feet	**450.5396**	53.4714	8.42580	0.00000	335.0216

Disregarding the majority of the statistics involved in this regression for the moment, let's focus first on the outcome that the regression provides. In the bottom left-hand corner of the regression printout in Table 7.4 in Columns A and B, you will see the two terms "Intercept" and "Square Feet," and the "Coefficients" column immediately to their right. In this example, then, the intercept equals −311,221.8767 and the coefficient for the square footage (which represents the slope of the regression line) is 450.5396. Therefore, the regression equation can be written (rounded off) as the following:

$$\widehat{Y} = -311,221.87 + 450.539 \times X \tag{7.1}$$

where $\widehat{Y}$ = estimated price of the home, and X = square feet of the home. In other words, what Equation #1 says is:

"The estimated price of a home" = $ − 311,221.87 + $450.539 × (# of square feet)

Regression Equation 7.1 now provides you with what we like to call a "prediction" regarding the fifteen homes in these neighborhoods. A few pages ago, we asked the ques-

tions about what a $1,200$ ft.^{2} home would cost or what a home that was at least $2,000$ ft.^{2} would cost, in these neighborhoods and with this data set. We can now answer those questions! The 1,200 square feet and the 2,000 square feet represent the value for X in Equation 7.1. So using Equation 7.1 and inserting those numbers, we now have the following "predictions":

- For a 1,200 square foot home:
  - Predicted price $= \$-311,221.87 + 450.539 \times (1,200) = \$229,425.93$
- For a $2,000$ square foot home:
  - Predicted price $= \$-311,221.87 + 450.539 \times (2,000) = \$589,856.13$

Thus, our prediction for a $1,200$ ft.^{2} home is \$229,425.93 and our prediction for a home that is $2,000$ ft.^{2} is \$589,856.13. These are the prices that you estimate that you would pay for homes in those neighborhoods for the given square footage.

Figure 7.2 displays the Equation 7.1 regression line in our data for this example. Note that it is a "straight line" prediction, with both an intercept and a slope, and is merely the equation of a line that is trying to "best fit" within the data that we have.

A quick analysis of the regression line in Figure 7.2 reveals the following information. The coefficient for square feet, 450.54, is also the slope of the regression line and can be interpreted as \$450.54 for each additional square foot of home. An economist would call this the "marginal cost per square foot." Note also that the intercept is a negative number at $-\$311,222$. Being a negative number is not a concern to us because we are only concerned about the numbers and prices that are within the range of our data. In this case, the minimum size of a home is $X = 1,000$ ft^2 and the largest home is $X = 2,200$ ft^2. Since the intercept occurs where the regression line crosses the Y-axis at $X = 0$, it is far to the left and well outside the range of our data (which is between $1,000$ and $2,200$ sq ft), and thus is not a concern to us.

With this example highlighting the possibilities that regression analysis offers in statistical analysis, let's discuss some background and nomenclature before moving on to the statistics that are involved in regression analysis.
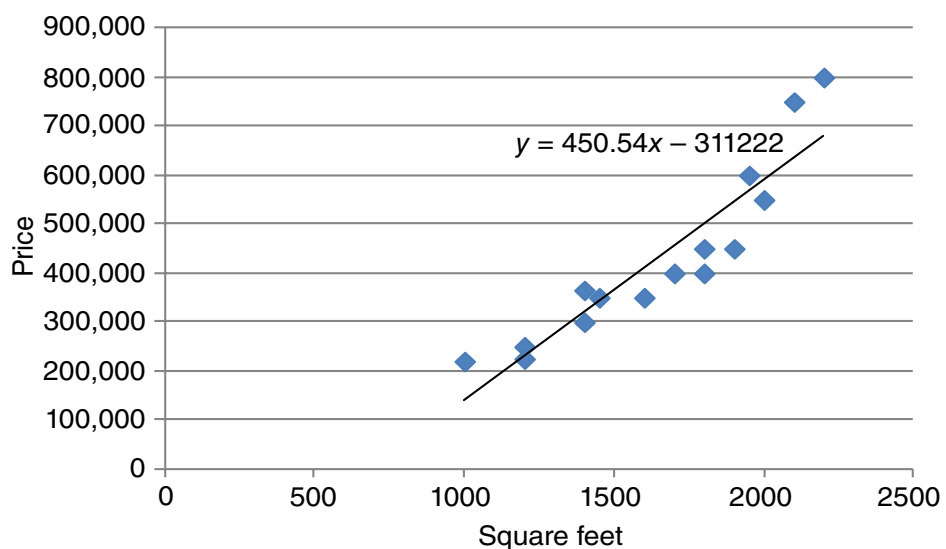
FIGURE 7.2 Regression Line Added to Scatter Plot from Figure 7.1.

7.3 Regression Background and Nomenclature

As previously stated, Regression Analysis is used to describe a *statistical* relationship between variables. Specifically, it is the process of estimating the "best fit" parameters that relate a dependent variable to one or more independent variables. In our field of cost estimation, cost is the dependent (or unknown, or response) variable. It is generally denoted by the symbol Y, and it is what we are trying to solve for. Cost is *dependent* upon the system's physical or performance characteristics, which are the model's *independent* (or known or explanatory) variables.

Since the dependent variable is a cost, the regression equation is often referred to as a *Cost Estimating Relationship,* or *CER.* The independent variable in a CER is often called a *cost driver.* A CER may have a single cost driver or it may have multiple cost drivers. If a CER has a single cost driver, it may look like one of the following:

TABLE 7.5   Potential Cost Estimating Relationships Cost Drivers

Cost	Cost Driver
Aircraft design	Number of drawings
Software	Lines of code
Power cable	Linear feet

Thus, from Table 7.5, we see that the cost of the aircraft design is *dependent upon* (or caused by, or explained by) the number of drawings; the cost for software is *dependent upon* the number of lines of code; and the cost of the power cable is *dependent upon* the number of linear feet of that cable. We will find out in Chapter 8 that cost may be dependent on not just a single independent variable, but perhaps on many/multiple independent variables, but this chapter will focus entirely on describing and using one single independent variable.

There are three symbols that you will see continually in this chapter and throughout the regression chapters:

Y_i = *any of the data points in your set, and there are "i" of them.*

$\widehat{Y}$ = *Y*(*hat*) = *the estimate of Y provided by the regression equation*

$\overline{Y}$ = *Y*(*bar*) = *the mean or average of all i cost data points*

In words, these three symbols are described as Y"*i*", Y"hat", and Y"bar."

The linear regression model takes the following form, as shown in Equation 7.2:

$$Y_i = b_0 + b_1 X_i + e_i \tag{7.2}$$

where b_0 is the Y intercept (or where the regression line crosses the Y-axis), and b_1 is the slope of the regression line. Both b_0 and b_1 are the unknown regression parameters and e_i is a random error term.

We desire a model of the following form, which is our regression equation:

$$\widehat{y}_x = b_0 + b_1 x \tag{7.3}$$

This model is estimated on the basis of historical data as:

$$y_i = b_0 + b_1 x_i + e_i \qquad (7.4)$$

where

$$e_i \sim N(0, \ \sigma_x^2), \text{ and iid}$$

In words, Equation 7.4 says:

- Actual cost (y_i) = Estimated cost (the regression) + Error of estimation (e_i)

Thus, the "Actual cost" will equal "what you estimate the cost to be" plus "some amount of error."

Your "errors" must be normally distributed, have a mean of zero, must have equal variances, and must be independent and identically distributed, or "iid." We will discuss these terms, errors, and their importance in the "Residual Analysis" section of this chapter.

When performing regression analysis, the most common method used is the "least squares best fit (LSBF)" regression method, where b_0 and b_1 are chosen such that the "sum of the squared residuals" is minimized. From Equation 7.4, we rearrange to solve for the errors, e_i, as seen in Equation 7.5, and then we seek b_0 and b_1 such that Equation 7.6 is true:

$$e_i = y_i - (b_0 + b_1 x_i) = y_i - \widehat{y} = \text{residuals} \qquad (7.5)$$

$$\sum (y_i - \widehat{y})^2 = \text{minimum} \qquad (7.6)$$

Equations 7.5 and 7.6 represent the keys to understanding what regression analysis is ultimately trying to accomplish. Equation 7.5 states that the error term e_i – the error in your prediction – is equal to the actual data point (y_i) minus what your prediction says it should be, $\widehat{Y}$. The difference between the actual cost and your predicted cost is your error or *residual*. We would like this error to be small, which would make our prediction close to the actual cost. We informally describe that Equation 7.6 defines the goal in regression analysis as "minimizing the errors," or minimizing the difference "between the actual cost and your predicted cost," and that we want the "sum of all of these squared errors" to be as small as possible. When we have that sum as small as it can possibly be – given all of the combinations of $b_{0's}$ and $b_{1's}$ possible – then we have our best regression equation, since we have achieved the smallest sum of squared errors possible.

Example 7.2 "Sum of Squared Errors:"   To visualize this concept of minimizing our errors, let's look at the following two figures. Figure 7.3 is a copy of Figure 7.2. Note that the regression line included is the "straight line" prediction, with both an intercept $(= -311,222)$ and a slope $(= 450.54)$. It is the equation of a line that is trying to "best fit" within the data that we have. The equation of that line is shown at the top of the chart.

You can see that the regression line indeed fits the data well, as it bisects and "mirrors" the data closely and it will closely reproduce the actual values for most of the data points
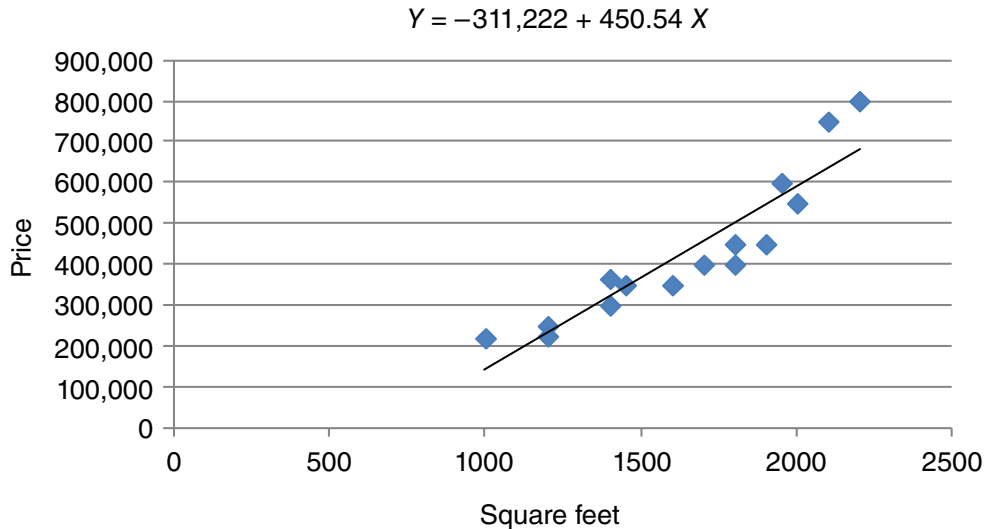
FIGURE 7.3 A Good Fit Regression/Prediction Line from the Data Set in Table 7.3.

(i.e., the actual cost vs. our predicted cost from the regression will be similar). In fact, some of the points go *exactly* through the regression line, like the data point for the home that is $1,450$ sq ft ($= \$350,000$). In this case, the actual cost and our predicted cost for that size home are exactly the same!

But now let's look at Figure 7.4. What if our regression/prediction line looked like this?
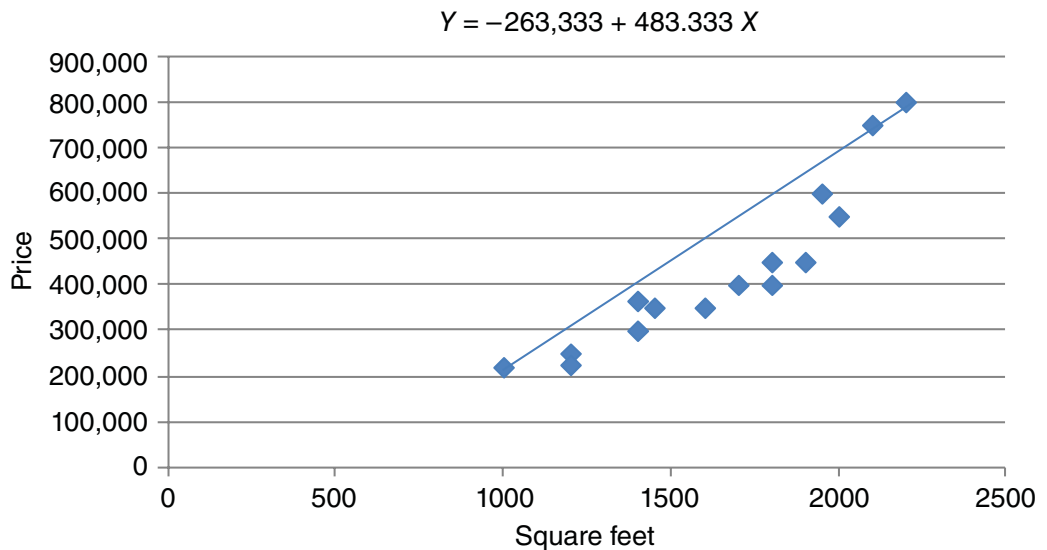
FIGURE 7.4 A Poor Fit Regression/Prediction Line from the Data Set in Table 7.3.

While the data set remains the same as in Figure 7.3, we can see that the regression line now used (and manually drawn in here) to "predict" the price of a home vs. the square feet of that home will *not* be as good a fit to the underlying data as the one in Figure 7.3, as the prediction line is higher (above) most of the data points – in fact, *it merely connects the end points*, while the remaining data points fall below and to the right (note: calculations for this line are shown in a few paragraphs). How will this affect our predictions? If you

look at the data point in Figure 7.4 for the one home that is $2,000$ ft.^{2}, the actual price for that home was $550,000, yet our "prediction" for the $2,000$ ft.^{2} home would be for about $700,000. Our prediction is thus significantly higher ($= \$150,000$) than the actual price! While we can clearly see *visually* that this will be the case for almost all of the data points, let us examine what occurs *mathematically,* as well. To do this, we will concentrate again on the goal of minimizing the sum of our squared errors, from Equation 7.6: $\sum (y_i - \widehat{y})^2$.

This equation says that we are trying to minimize the sum of the squared differences between the actual cost and the predicted cost (these differences are called "the residuals"). For the one data point at 2,000 sq ft, the *actual* cost was $550,000, the *predicted* cost was for $700,000, and so our difference (or residual) is $150,000. We can observe that the differences between the actual costs and the predicted costs will be significantly large for *every* data point using this regression line (except for the end points). We then "square" the differences (making the values even larger), and then sum these squared residuals together. That will eventually produce a very, very large sum of squared errors in our regression result.

Table 7.6 shows the extreme differences in the "sum of the squared errors" from the two regression/prediction lines shown in Figures 7.3 and 7.4.

Part A shows the sum of squared errors calculations for the regression line of $Y = -311,222 + 450.54 \times X$, as seen in Figure 7.3. Part B shows the sum of the squared errors calculations for the regression line in Figure 7.4. The calculations for the slope and intercept of this regression line in Part B are just slightly more involved, as they needed to be found manually from the values of the end points. The two end points in Figure 7.4 are $(X_1 = 1,000; Y_1 = 220,000)$ and $(X_2 = 2,200; Y_2 = 800,000)$. Consequently, using these values we can solve for the slope of the regression line:

- Slope $= m = b_1 = \frac{\text{"Rise"}}{\text{"Run"}} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{(800,000 - 220,000)}{(2,200 - 1,000)} = \frac{580,000}{1,200} = 483.333 = $ slope

- Next, to find the equation of a line given two points, we use the equation $Y - Y_1 = m(X - X_1)$. Using the point $(X_1 = 1,000; Y_1 = 220,000)$, we now have $Y - 220,000 = 483.333(X - 1,000)$. Multiplying through and simplifying this equation yields the equation of the line in Figure 7.4 as $Y = -263,333 + 483.333 \times X$.

The sum of the squared errors in Part A is 67,735,302,730, compared to 225,414,579,056 in Part B. Recall that our goal from Equation 7.6 is to *minimize* the sum of these squared errors, and we will see in a few sections that this sum is the numerator in the Standard Error equation. The smaller this sum is, the smaller the standard error is in your regression. The calculations in Table 7.6 indicate that the prediction line in Figure 7.4 will not be a good predictor, and this is clearly not what we desire in our prediction model. An interesting note is that the second line (the one in Figure 7.4) had two data points that were *exactly* on the prediction line (both of the end points) resulting in a residual of zero at these points, while the actual regression line in Figure 7.3 had only one. However, this does not mean that having a few exact points is the better indicator/predictor. It is the line that most closely fits ALL of the data in the sense of minimizing the sum of squared errors from the actual data that is the most important.

Now let's look at Figure 7.3 once again. As previously discussed, we can see that the regression line appears reasonable, and informally we would say that this is because the differences between the actual costs and the predicted costs are extremely small. In fact, mathematically the intercept (b_0) and slope (b_1) that were calculated for that regression

TABLE 7.6   Sum of the Squared Errors Calculations in Figures 7.3 and 7.4 for Example 7.2

Part A, results from Figure 7.3: Using the Equation of the Line $= -311,222 + 450.54 \times X$

Home #	Price($)	Square Feet	Predicted Cost Using the Regression	Difference of Predicted from Actual	Squared Differences
1	300,000	1400	319,534	−19,534	381,577,156
2	400,000	1800	499,750	−99,750	9,950,062,500
3	350,000	1600	409,642	−59,642	3,557,168,164
4	800,000	2200	679,966	120,034	14,408,161,156
5	450,000	1800	499,750	−49,750	2,475,062,500
6	250,000	1200	229,426	20,574	423,289,476
7	225,000	1200	229,426	−4,426	19,589,476
8	450,000	1900	544,804	−94,804	8,987,798,416
9	550,000	2000	589,858	−39,858	1,588,660,164
10	400,000	1700	454,696	−54,696	2,991,652,416
11	220,000	1000	139,318	80,682	6,509,585,124
12	350,000	1450	342,061	7,939	63,027,721
13	365,000	1400	319,534	45,466	2,067,157,156
14	600,000	1950	567,331	32,669	1,067,263,561
15	750,000	2100	634,912	115,088	13,245,247,744
				Sum =	**67,735,302,730**

Part B, results from Figure 7.4: Using the Equation of the Line $= -263,333 + 483.333 \times X$

Home #	Price($)	Square Feet	Predicted Cost Using the Regression	Difference of Predicted from Actual	Squared Differences
1	300,000	1400	413,333	−113,333	12,844,414,222
2	400,000	1800	606,666	−206,666	42,711,000,889
3	350,000	1600	510,000	−160,000	25,599,936,000
4	800,000	2200	800,000	0	0
5	450,000	1800	606,666	−156,666	24,544,360,889
6	250,000	1200	316,667	−66,667	4,444,435,556
7	225,000	1200	316,667	−91,667	8,402,765,556
8	450,000	1900	655,000	−205,000	42,024,877,000
9	550,000	2000	703,333	−153,333	23,511,008,889
10	400,000	1700	558,333	−158,333	25,069,370,556
11	220,000	1000	220,000	0	0
12	350,000	1450	437,500	−87,500	7,656,223,750
13	365,000	1400	413,333	−48,333	2,336,098,222
14	600,000	1950	679,166	−79,166	6,267,310,972
15	750,000	2100	751,666	−1,666	2,776,556
				Sum =	**225,414,579,056**

line produce the smallest sum of squared errors of any combination of $b_{0's}$ and $b_{1's}$, and that is why it is our best regression line. In conclusion, we have shown why the regression line in Figure 7.3 is preferred to the one in Figure 7.4, as it produces the smallest sum of squared residuals, *which is our goal in regression analysis.*

In Chapter 6 on Statistics, Equation 6.1 was the equation used when computing the variance of a data set. The note after that equation explained why we "square" the terms in the numerator for variance. The same principle applies in Equation 7.6 mentioned earlier. We once again have cost residuals from "the actual cost versus the predicted cost," and some of these residuals will be positive, while others will be negative. If you added all of these residuals together, the positives and negatives would sum to zero. Therefore, in Equation 7.6, as in the variance equation, we square the cost differences and then sum them to get the total errors. You will see Equation 7.6 again as part of the standard error equation in Section 7.5.

To compute an actual regression, we seek to find the values of b_0 and b_1 that minimize $\sum (y_i - \hat{y})^2$. To do so, one may refer to the "Normal Equations" in (7.7) mentioned subsequently, the proofs of which can be found in numerous textbooks but is not a goal here: [1]

$$\sum Y = nb_0 + b_1 \sum X$$

$$\sum XY = b_0 \sum X + b_1 \sum X^2 \tag{7.7}$$

With two equations and two unknowns, we can solve for b_1 first, and then for b_0:

$$b_1 = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{\sum XY - \dfrac{\sum X \sum Y}{n}}{\sum X^2 - \dfrac{(\sum X)^2}{n}} = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$b_0 = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n} = \overline{Y} - b_1 \overline{X} \tag{7.8}$$

Fortunately, we do not have to calculate the intercept and slope by hand anymore, since there are numerous statistical packages that will do regression for you!

Recall Example 7.1 when we were calculating the prices of homes in a few neighborhoods. After computation of Descriptive Statistics, we found that the average cost of all homes in our data set was \$430,666.67. Thus,

$$\overline{Y} = \$430,666.67$$

Then we developed a cost estimating relationship (CER) between home price and square feet using LSBF regression in Table 7.4, and calculated the following:

$$\hat{Y} = -311,221.87 + 450.539 \times X$$

Then we estimated the purchase price of a 2,000 sq ft home:

$$\text{Predicted price} = -\$311,221.87 + \$450.539 \times (2,000) = \$589,856.13$$

What do these numbers mean?

- $\overline{Y} = \$430,666.67$ is the estimate of the average purchase price of all homes in those neighborhoods, based on the given sample
- $\hat{Y} = \$589,856.13$ is the estimate of the homes in the data set that are 2,000 square feet

A key point to understand is that if the regression statistics (which we will soon cover) for the regression are good and are in fact better than the Descriptive Statistics, then you will prefer to use the regression equation instead of the Descriptive Statistics to develop your cost estimate. But if the regression statistics are not very good, you can always go back to using the mean and standard deviation, since that data is still available and viable. A good regression means that you prefer the regression equation as an estimator, instead of using the mean. But what makes a good regression? We will now discuss the primary statistics used in regression and their importance, to determine what makes a regression "good" or "bad," and what makes one regression better (or worse) than another.

7.4 Evaluating a Regression

When evaluating your regression, there are numerous statistics that need to be analyzed. The primary ones that we will discuss include the standard error, the coefficient of variation (CV), the F-Statistic, the t-Statistics, and the Coefficient of Determination (R^2). Table 7.7 below is the same regression printout from Table 7.4, but this time with the *Standard Error*

TABLE 7.7 Regression Result of Price versus the Square Footage of the Fifteen Homes in Example 7.1

A	B	C	D	E	F
Regression Statistics					
Multiple R	0.91936		**Price vs Square Feet**		
R Square	0.84523				
Adjusted R Square	0.83332				
Standard Error	**72183.1553**				
Observations	15				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.69908E+11	3.69908E+11	70.9941	1.26178E−06
Residual	13	67735302725	5210407902		
Total	14	4.37643E+11			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	−311221.8767	90000.5658	−3.45800	0.00424	−505656.2777
Square Feet	450.5396	53.4714	8.42580	0.00000	335.0216

bolded and columns labeled A–F for emphasis and identification of those statistics in this section.

7.5 Standard Error (SE)

The first of the terms that we will cover is Standard Error. In Table 7.7, our standard error is found in Column B and is equal to 72,183.1553. In Descriptive Statistics, the standard deviation is the deviation of your data points *compared to the mean*. In regression analysis, the standard error is that deviation *compared to your regression (or prediction) line,* and thus in both cases smaller is better. In this example, since our dependent variable is price, the standard error is in the same units and is equal to $72,183.15. How was this number determined? The equation for determining standard error is shown in Equation 7.9.

$$SE = \sqrt{\frac{\sum (y_i - \widehat{y}_i)^2}{n - 2}} \tag{7.9}$$

Recall that in Equation 7.6, we were (informally speaking) trying to minimize the difference between the actual cost and the predicted cost. A close inspection here reveals that Equation 7.6 is the numerator in Equation 7.9 and thus we are trying to minimize this measure of the cost errors again. The "sum of all of the squared errors" is then divided by $n - 2$. The final procedure is to take the square root, thus bringing the "squared units" back to the original units that we can use once again (in this case, "dollars").

A quick note about the denominator in Equation 7.9. The denominator of $n - 2$ is actually "$n - k - 1$", where n is the total number of data points and k is equal to the *number of independent variables used in the regression*. In this example, one independent variable was used (X = square feet), and so $k = 1$, and the denominator thus becomes $n - k - 1 = n - (1) - 1 = n - 2$.

The sum of the squared errors in Example 7.2 was found in Table 7.6 to be 67,735,302,730. Since the denominator for Equation 7.9 is $n - 2$ and there are 15 data points (i.e., $n = 15$), then the denominator in Equation 7.9 is $= n - 2 = (15) - 2 = 13$.

Finally, solving for the Standard Error, we get $SE = 67,735,302,730 \div 13 = 5,210,407,902$. Since this is in "squared units," we take the square root of $5,210,407,902 = \$72,183.15$. This is our standard error.

Thus, for the home data in Example 7.1, the mean is $430,666.67 and the standard error is $72,183.15. This means that *on "average" when predicting the cost of future homes in this neighborhood, we will be off by* approximately $+/- \$72,000$.

As a note of interest, if we go back to the Descriptive Statistics calculated in Table 7.2 using our original data set, the standard deviation was found to be $176,805.65. Using the regression equation of Price vs. Square Feet just calculated, our standard error was $72,183.15. We can now see that our errors are lower when using the regression than when using the descriptive statistics and the cost mean and standard deviation. A conclusion from this is to say that "Using the regression is preferred to using the mean" in this case, which is what we had hoped for when attempting regression.

Two final notes about Standard Error:

1. When working with basic statistics, such as the data set in Table 7.1, we calculate and use standard deviation, which is the mathematical deviation *about the mean of*

your data. When performing a regression, we now use standard error, which is the measuring the mathematical deviation *about the regression line.*

2. Referring to Table 7.7, do not confuse the "standard error for the regression" found in Column B (= $72, 183.15) that we just discussed with the standard errors found in Column C. The standard errors in Column C are the "standard errors for the coefficients" found in the regression (for intercept and square feet), and we are not concerned with those metrics at this time.

7.6 Coefficient of Variation (CV)

Now that we have calculated a value for the standard error, how do we know if it is considered "good" or not? Recall from Equation 6.5 in the Statistics chapter that the CV is:

$$CV = \frac{s_y}{\bar{y}}$$

which is the standard deviation of the data divided by the mean. For standard error, we find that the CV is merely the standard error divided by the mean. So, in Example 7.1, we found that:

- Standard Error = $72, 183.15 and Mean = $430, 666.67

    Thus, our Coefficient of Variation for our standard error is found to be

    $$SE \div Mean = \$72, 183.15 \div \$430, 666.67 = 0.1676 = 16.76\%$$

This answer of 16.76% is actually quite good, as our CV is relatively low, and it says that *on average we will be off by 16.76% when predicting the cost of future homes in this neighborhood.* The smaller the CV the better. In the beginning of this chapter, Table 7.2 shows the descriptive statistics for this housing data. If we were to calculate the CV using that data, we would find the following:

- Standard Deviation = $176, 805.65 and Mean = $430, 666.67

    Thus, our CV for Example 7.1 using *Descriptive Statistics* is

    $$SD \div Mean = \$176, 805.65 \div \$430, 666.67 = 0.4105 = 41.05\%$$

This says that *on "average," we will be off by 41.05% when predicting the cost of future homes in this neighborhood* using just the cost data. It is easy to see that the CV using regression analysis is quite an improvement over using the CV in the descriptive statistics, 16.76 vs. 41.05%. The conclusion in this example is that using the regression is greatly preferred to using the mean as a predictor, the same conclusion that was derived when comparing the standard error vs. standard deviation of the same data set.

TABLE 7.8 Regression Result in Example 7.1 Highlighting the ANOVA Section

A	B	C	D	E	F
Regression Statistics					
Multiple R	0.91936		**Price vs Square Feet**		
R Square	**0.84523**				
Adjusted R Square	**0.83332**				
Standard Error	72183.1553				
Observations	15				
ANOVA					
	_df_	_SS_	_MS_	_F_	_Significance F_
Regression (SSR)	**1**	**3.69908E+11**	**3.69908E+11**	**70.9941**	**1.26178E−06**
Residual (SSE)	**13**	**67735302725**	**5210407902**		
Total (SST)	**14**	**4.37643E+11**			
	Coefficients	_Standard Error_	_t Stat_	_P-value_	_Lower 95%_
Intercept	−311221.8767	90000.5658	−3.45800	0.00424	−505656.2777
Square Feet	450.5396	53.4714	8.42580	0.00000	335.0216

7.7 Analysis of Variance (ANOVA)

The Analysis of Variance section of a regression covers a number of statistical measures, including the F-Statistic, and the Coefficient of Determination (R^2). Table 7.8 shows the regression results again from Table 7.4, but this time _bolding_ the statistics that are covered in this ANOVA section.

The ANOVA section can be seen in the middle of Table 7.8. We will cover these statistics briefly to show where the numbers come from, but not to prove each equation.

ANOVA Section: In Column A in the ANOVA section, there are three areas that are addressed: The Sums of Squares for the Regression (or SSR), the Sums of Squares for the Residual (or SSE), and the Sums of Squares for the Total (or SST). Column B discusses Degrees of Freedom (df) for each of these areas. While understanding these are important in theory, they are not important from a calculation standpoint, so we will skip over these and address the Sum of Squares (SS) in Column C. These totals are calculated in the following equations:

$$\text{SST} = \text{Total sum of squares} = \sum (Y_i - \overline{Y})^2$$

$$\text{SSE} = \text{Sum of squares errors} = \sum (Y_i - \widehat{Y}_i)^2$$

$$\text{SSR} = \text{Sum of squares regression} = \sum (\widehat{Y}_i - \overline{Y})^2$$

$$\text{SST} = \text{SSE} + \text{SSR}$$

The following statements are true (but not necessarily obvious) from the above equations.

- The SST is the total sum of squares, and represents the *total variation in the data that underlies the regression*. It is found by summing the squared differences of the *actual* cost of each home (the Y_i's), and the mean of the homes (where the mean = $\$430,666.67$). There will be 15 Y_i's ($Y_1, Y_2, \ldots Y_{10}, \ldots Y_{15}$). Thus, for each data point calculate the difference between the actual cost and the mean cost of $\$430,667.67$, square that difference and then add up all fifteen of these differences to get the sum. That sum is SST.
  - In Table 7.8, SST $= 4.37643 \times 10^{11}$

- The SSE is the sum of squares errors and represents the *Unexplained Variation in the regression*. This is what we would like to minimize. It is found by summing the squared differences in the actual cost of each home (the Y_i's) with the cost predicted by the regression, $\widehat{Y}$, of each home. Thus, for each data point calculate the difference between the actual costs and the predicted costs for each home, square that difference, and then add up all fifteen to get a sum. The goal of regression analysis is to minimize this sum. The sum is SSE.
  - In Table 7.8, SSE $= 67,735,302,725$

- The SSR is the sum of squares regression and represents the *Variation Explained in the regression*. It is found by summing the squared differences between the prediction and the mean. Thus, for each data point, calculate the difference between the predicted costs and the mean costs for each home, square that difference, and then add all fifteen to get a sum. We would like this number to be as large as possible. Since SST = SSE + SSR, maximizing SSR is the same as minimizing SSE.
  - In Table 7.8, SSR $= 3.699 \times 10^{11}$

Moving to Column D in the ANOVA section, MS is the mean squared error and is found by dividing the SS in Column C by its degrees of freedom in Column B, for the rows on regression and residual.

- For SSR: $3.699 \times 10^{11} \div 1 = 3.699 \times 10^{11}$
- For SSE: $67,735,302,725 \div 13 = 5,210,407,902$

The F-statistic is then found in Column E by dividing the "MS" for Regression (SSR) by the "MS" for Residual (SSE), both from Column D. While we will discuss what the F-Statistic represents in Section 7.9, in actuality we are less concerned with what the *actual* F-statistic is and are more concerned with what the *Significance* of the F-statistic is.

F-Statistic: MS for Regression $\div$ MS for Residual

$$= 3.699 \times 10^{11} \div 5,210,407,902 = 70.9941$$

Finally, in Column F, the significance of the F-Statistic is found by using hypothesis testing, where the *null hypothesis* (Ho) is that "the slope of the regression is equal to zero." The *alternative hypothesis* (Ha) is that "the slope of the regression is *not* equal to zero." The percentage in this column is referred to as the "Significance of the F-statistic," and represents the probability that the null hypothesis is true, which is to say that the slope used in the regression equation is equal to zero, and clearly, this is not a desirable outcome. If

the slope $b_1 = 0$, then in the equation for your regression, $y_x = b_0 + b_1 x$, you merely have a horizontal line at the value of b_0 for all values of x. You will now have that the predicted cost is always equal to the intercept, b_0, and does not change. This situation would not make for a useful prediction! Consequently, we would like the F-statistic significance to be as *small* as possible (the closer to zero the better) so that it falls into the F-distribution rejection region. If it does fall into that region, we can reject the null hypothesis and thus conclude that the slope of the regression is *not* equal to zero.

As a side note, if you are using Excel or another statistical package and are evaluating the numbers in ANOVA or elsewhere, you may see a number in a result that looks like either one of the following. We mention this because the second number listed below was the significance of the F-Statistic in Table 7.8 and was 1.26178E−06. Consider these two numbers:

- 1.26178E+06
- 1.26178E−06

Here is how to interpret both of these. The first number is a very *large* number, and the E + 06 is read as 10^6, so it is really 1.26178×10^6 which is equal to 1,261,780. The E + 06 is also interpreted as moving six significant digits to the *right* from the decimal in 1.26. The second number is a very *small* number, and the E−06 is read as 10^{-6}, so it is really 1.26178×10^{-6}, which is equal to .00000126178. The E−06 is also interpreted as moving six significant digits to the *left* from the decimal in 1.26.

7.8 Coefficient of Determination (R^2)

These statistics are found in Columns A and B near the top of Table 7.8. The Coefficient of Determination (R^2 or R-Square) is the percentage that Explained Variation (SSR) is of the Total Variation (SST), representing *the percentage of total variation explained by the regression model.* You want R^2 to be as large as possible and as close to 1.0 as you can get.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{SSR}}{\text{SST}}$$

In Table 7.8, our $R^2 = \text{SSR}/\text{SST} = 3.699 \times 10^{11} \div 4.37643 \times 10^{11} = 0.8452$. This number is calculated for you by the regression model and is found near the top of Column B, next to "R Square." So in Table 7.8, we can see that 84.52% of the total variation is explained by our regression model, which overall is quite good. The closer to 100% that R^2 is, the better.

There are two other statistics that we will consider when discussing R-Square:

- Above the R-Square, you will see a metric called "Multiple R." This is called the correlation coefficient and tests the strength (or correlation) between the two variables in the regression. If you square this correlation number R, you will get "R-Square."
- Underneath "R-Square" in our regression, you will see a metric called "Adjusted R-Square." The long-term phrase for "Adjusted R-Square" is "R-Square Adjusted for

Degrees of Freedom." Its value will *always* be less than or equal to "R-Square." Mathematically, it is not possible for Adjusted R-Square to be greater than R-Square, as is clear from equation 7.10:

$$\text{Adjusted } R^2 = 1 - \frac{(\text{SSE}/n - k - 1)}{\text{SST}/(n-1)} \tag{7.10}$$

The purpose for Adjusted R^2 requires a quick discussion on a single variable regression vs. a multiple variable regression.

If you perform regression analysis using a single independent variable, you will achieve a certain R-Square value. In our model, we achieved an R-Square of .8452, as seen in Table 7.8. If you then performed a second regression – but this time with one additional independent variable (so there are now *two* independent variables) – the R^2 value will *automatically and artificially increase, or at least not decrease,* due to that additional independent variable. Therefore, the purpose of "Adjusted R^2" is to enable you to compare the R^2 value in regressions that have an unequal number of independent variables, such as a regression with only one independent variable versus a regression with two independent variables, or a regression with two independent variables versus a regression with three. The Adjusted R^2 "adjusts" for the artificial increase in the R^2 value caused by the addition of independent variables. A meaningful comparison between the two regressions can then be made.

7.9 F-Statistic and t-Statistics

The F-statistic was described mathematically in Section 7.7 on ANOVA. The F-Statistic tells us whether the full regression model, $\widehat{Y}$, is preferred to the mean, $\overline{Y}$. It looks at the *entire* regression model and determines whether or not it is "a good model." In our example, does Price versus Square Feet make a good regression? The answer was a resounding yes. The Significance of the F-Statistic was 1.26178×10^{-6}, which is a very small number, and we want that significance to be as small as possible. It also concludes that the "probability that the slope of your regression is zero" is very small at that value, so the regression model would be preferred over just using the mean. When this number is small, it is falling in the rejection region for the F-Statistic hypothesis testing, as shown in Figure 7.5. The rejection region is in the dark tail to the right of the Fc (critical F) at whatever significance level that you have previously chosen. For this problem, let's assume
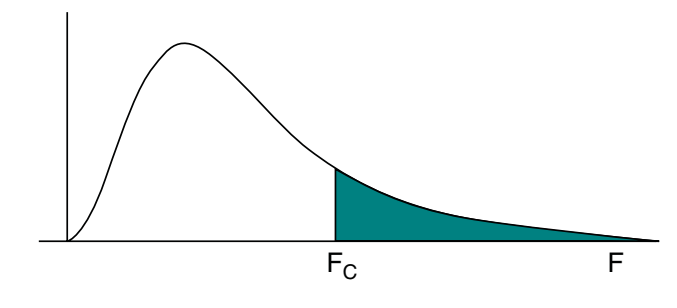
FIGURE 7.5 The F-Distribution and its Rejection Region (shaded).

that the significance level is at the 0.10 level. Remember that the null hypothesis Ho is that "the regression slope is equal to zero." The alternate hypothesis (Ha) would be that "the regression slope is *not* equal to zero," which is the outcome that we desire.

Since 1.26178×10^{-6} falls within the rejection region, well below the 0.10 significance level, we will reject H_0 and conclude that "the regression slope is not equal to zero," and thus *the full regression model is better than the mean* as a predictor of cost.

t-Statistic: Closely related to the F-Statistic is the t-statistic. There are many mathematical explanations on what the t-statistic does, such as: "testing the strength of the relationship between Y and X," which in our example was the relationship between Price and Square Feet; or, "testing the strength of the coefficient b_1;" or "testing the hypothesis that Y and X are *not* related at a given level of significance." While all of these explanations may be correct, let's look at the t-statistic in a different perspective.

While the F-Statistic is looking at the "entire" regression model and determines whether the model is "good" and usable or not, the t-statistic is looking at *each individual independent variable* and determining whether that particular independent variable is significant to the model or not. There will always be *only one* F-statistic in a regression printout, regardless of how many independent variables there are. As for the t-statistic, there will always be one t-statistic for *each independent variable.* Thus, in a single variable regression like the one in Table 7.8, there will be exactly one F-statistic and one t-statistic, because there is one model and one independent variable. As a side note, we usually disregard the t-statistic and the p-value for the *intercept*, because the intercept, which occurs at the *Y*-axis, is generally well outside the range of our data.

Like the F-statistic significance, the t-statistic significance is also a hypothesis test, where the null hypothesis Ho is that the "slope (or coefficient) for that independent variable is equal to zero." The alternative hypothesis (Ha) is that "the slope for that independent variable is *not* equal to zero," which is the outcome we prefer. In essence, we are testing whether the slope of the regression line differs significantly from zero, and clearly, we want that to be the case.

The values for the F-statistic significance and the t-statistic significance will be identical in value in a single variable regression model. If there is any difference at all between the two, it is merely because of the differences in calculation when using the F-distribution and the t-distribution. For both the F-statistic significance and the t-statistic significance, we want each to be as close to zero as possible.

In the next chapter on multi-variable regression, we will find that we will now have differences between the F-statistic significance and the t-statistic significances. This is because in a *two* independent variable regression model, there will again be *only one* F-statistic and thus one F-statistic significance, but there will now be *two* t-statistics and *two* t-statistic significances, one for each of the independent variables. In a *three* variable regression model, there will again be only one F-statistic and one F-statistic significance, but there will now be *three* t-statistics and *three* t-statistic significances, one for each independent variable. The t-statistic significance is also referred to as the p-value or probability value. One way to look at this p-value is that it is "the probability that this event occurred randomly." If the p-value is small, then the probability that the event occurred randomly is highly improbable, which is the desired result. Figure 7.6 shows a t-Distribution. It is similar to a normal curve, but "flatter" in height and with wider "tails" to the left and right.

If the t-statistic significance (*p*-value) test falls within the rejection (shaded) region, this indicates that the dependent variable *X* and independent variable *Y* are related. Thus,
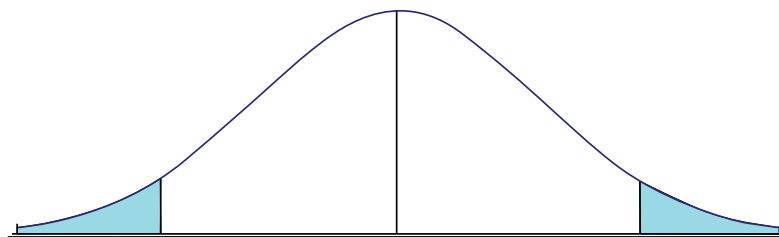
FIGURE 7.6 The t-Distribution and its Rejection Region (shaded).

we would reject H_0 that "the slope for that independent variable is equal to zero" and say that we prefer the model *with b_1* to the model *without b_1*, which is our desired outcome.

Having completed an analysis of the numerous statistics involved in a regression printout, the following list summarizes the desired outcome/result for each statistic:

- R^2 or Adjusted R^2: The bigger the better and as close to 1.0 as possible.
- Standard Error: The smaller the better and as close to 0 as possible.
- Coefficient of Variation: This involves dividing the standard error by the mean. The smaller the better and as close to 0 as possible.
- Significance of F-Statistic: If less than a desired significance level (say 0.10), then we prefer the regression model to the mean. The smaller the better and as close to 0 as possible.
- Significance of the t-statistic (or p-value): If less than a desired significance level (say 0.10), then we prefer the model *with b_1*, else we prefer it *without b_1*. The smaller the better and as close to 0 as possible.

Knowing these statistics in a single variable linear regression is important for understanding how "good" your regression is; but perhaps just as important, these statistics will be used to compare other linear models to each other, to determine which regression is the "best" of them all.

7.10 Regression Hierarchy

While there are many ways to look at and evaluate regression results, we offer the following hierarchy/guidelines to assist you. We will refer to this guideline as the "Hierarchy of Regression." The Hierarchy is broken down into two parts or phases.

Hierarchy of Regression

When reviewing a regression printout and comparing the results of one regression to another, use the following hierarchy in order to establish your preferred/"best" regression:

Phase I:

1. Does your regression pass the Common Sense test? In other words, does Cost increase/decrease as it should as the value of the independent variable

increases/decreases? (This tests whether the *slope of the regression line* is realistic and does not refer to the individual data points).

2. Is the F-stat *significance* below 20%**?

3. Is the t-stat *significance* (p-value) below 20%**? (Note: Again, this does not apply to the p-value of the intercept, but only to the independent variable)

- If the answer is YES to all three of the Phase I criteria, then *keep this regression for further consideration.* Do not eliminate it at this point for any reason, such as low R-square or high standard error values. At the bottom of the regression, we generally write "This regression passes the first three criteria of the Regression Hierarchy. Therefore, we will keep it for further consideration." Proceed to the other regressions and apply the same Phase I criteria to them as well.

- If the answer is NO to *any* of the three Phase I criteria, then REJECT the model due to its not passing that particular condition. We usually write "This regression fails the Regression Hierarchy due to (fill in the blank). Therefore, we will eliminate it from further consideration."

Phase II:

After performing the Phase I test for each regression, take the regressions that "passed" the Phase I criteria and make a chart to compare them to each other using the following metrics:

- Compare the R-Square Value (or Adjusted R-square, if required): (higher is better)
- Compare the Standard Errors: (lower is better)
- Compare the Coefficient of Variations: (lower is better) (Note: CV = SE/mean)

Using a table format for these comparisons works best because of the visual advantages provided for the reader and analyst. An example chart is included subsequently in Table 7.9:

TABLE 7.9 Hierarchy of Regression Comparison Table

Regression	*R*-square	Std Error	Coefficient of Variation
Cost vs. Weight			
Cost vs. Power			
Cost vs. Frequency			

After comparing the regression results in the Phase II chart, pick the model/regression that possesses the best statistics: this is your preferred regression equation. Then, answer any further questions using this preferred/"best" equation.

*Author's Note: We use this regression hierarchy while teaching as a way to get all of our students to look at the same metrics in a logical fashion in the short period that we have to teach them the subject area. While there are other metrics you can look at as well, especially if results are very close to each other, these are the primary ones to use when comparing results of one regression against another. Also, note the significance level of .20** for the significance of the F-statistic and the t-statistics. You may work to whatever significance level you deem*

appropriate (i.e., .10 or .05), but be aware that with very small data sets, working to a level tighter/more binding than 0.20 might not yield usable CERs.

7.11 Staying Within the Range of Your Data

When using a regression equation, ensure that the system or program or product that you are trying to predict the cost of falls within the range of the historical data that you are using. In the Example 7.1 data found in Table 7.3, the smallest value for home size was 1,000 square feet and the largest value was 2,200 square feet. Consequently, "staying within the range of your data" means that the size of the home that you are trying to predict by using your regression should fall within the range of homes between 1,000 and 2,200 square feet in size. The historical data that you used in this scenario to create the regression yields a prediction that is specifically *for that range of data*.

The following two examples discuss considerations for staying within the range of your data:

Example 7.3a You are tasked with predicting the cost of a home based on square feet, with historical square foot data values ranging between a smallest value of 1,000 square feet and a largest value of 3,000 square feet. Thus the range of your data in this case would be between 1,000 and 3,000 ft.[2] What if you needed to predict the cost of a 3,200 sq ft house with this data? Should you use the regression that you have or not?

Answer: In this example, while 3,200 sq ft falls slightly outside the range of your data, it would probably be acceptable to still use this regression for your prediction. The difference in a 2,800 sq ft home or a 3,000 sq ft home from your data set to a 3,200 sq ft home could be just merely the addition of a small laundry room or a back bathroom, so predicting outside the range of this data (maximum size was 3,000 sq ft) would probably not cause a problem nor affect the reliability of your prediction. This would probably also hold true for a home on the smaller end that was approximately 900 sq ft, as well. However, this might not be the case if the home became greater than say 3,500 or 4,000 sq ft, especially if it was located in a different neighborhood, or if it had a unique design, such as being circular. So in conclusion, Example 7.3a is an example where predicting outside the range of your given data might be acceptable.

Example 7.3b You are tasked with predicting the cost of a new aircraft and the historical data that you possess are for aircraft with maximum airspeeds between 500 knots and Mach 2. (Note: "Mach 1" is the speed of sound). What if you needed to predict the cost of an aircraft with a maximum airspeed of Mach 2.3? Should you use the regression that you have or not?

Answer: In this particular case, it would probably be *very unwise and inaccurate* to use the regression that you have calculated from the historical data. The reason is because the cost to increase an aircraft's maximum airspeed from Mach 2.0 to Mach 2.3 might be exponential in nature and thus could be very cost prohibitive. Why is this? In order to make the aircraft fly 0.3 Mach faster, you might need a larger engine with more horsepower. However, a larger engine may no longer fit into the space that was provided in the aircraft for the original engine and significant alterations (thus costs) may be necessary to make it fit. Moreover, the engine may be significantly heavier due to the increased horsepower, which in turn increases the aircraft weight, and that might cause center of gravity (CG) considerations and thus significant modifications for that aircraft. You may

also need to modify the outside of the aircraft in some way to reduce the "coefficient of drag" to be able to increase the airspeed, too. The only way that you would know for sure what modifications should be done is to consult engineers who specialize in that field, but any modifications would most likely increase the aircraft cost greatly. So in conclusion, Example 7.3b is an example for when predicting outside the range of your given data would probably be highly unacceptable, unwise, and inaccurate.

7.12 Treatment of Outliers

In general, an outlier is a residual that falls greater than two standard deviations from the mean. You can have outliers from the independent variable (X) data or from the dependent variable (Y) data. The standard residual generally takes on one of these three forms found in (7.11), where SE is the Standard Error of the regression and the S_X and S_Y represent the standard deviation of the X data and the Y data, respectively:

$$\frac{Y_i - \widehat{Y}}{\text{SE}} \text{ or } \frac{X_i - \overline{X}}{S_X} \text{ or } \frac{Y_i - \overline{Y}}{S_Y} \tag{7.11}$$

Recall that since 95% of the population falls within 2 standard deviations of the mean in normally distributed data, then in any given data set, we would expect 5% of the observations to be outliers. If you have outliers, you generally do not want to throw them out unless they do not belong in your population.

7.12.1 HANDLING OUTLIERS WITH RESPECT TO *X* (THE INDEPENDENT VARIABLE DATA)

All data should come from the same population. You should analyze your observations to ensure that this is so. Observations that are so different that they do not qualify as a legitimate member of your independent variable population are called "Outliers with respect to the independent variable *X*."

- To identify outliers with respect to *X*, simply calculate the mean of the *X* data, and the standard deviation, S_X. Those observations that fall greater than two standard deviations from the mean are likely outlier candidates.

- You *expect* 5% of your observations to be outliers – therefore, the fact that some of your observations *are* outliers is not necessarily a problem. You are simply identifying those observations that warrant a closer investigation.

Example 7.4 Consider the following example. The data in Table 7.10 is theoretical data from an Army indirect fire weapon that was fired/tested eight times and the values represent the distances that the rounds traveled down range, in meters:

- *Column A* represents the distance in meters that the round traveled down range on each shot

- *Column B* is the average distance for the 8 shots (mean = average = 823.125 meters)

- *Column C* is the difference (or residual) of the distance traveled minus the mean (Column A–Column B). Note that some of the rounds traveled less than the mean

TABLE 7.10   Data from Firing an Indirect Fire Weapon for the US Army in Example 7.4

A	B	C	D
X_i	$\overline{X}$		$\dfrac{(X_i - \overline{X})}{S_X}$
Range	Mean	$X_i - \overline{X}$	
600	823.125	−223.13	−0.5908
925	823.125	101.88	0.2698
450	823.125	−373.13	−0.9880
420	823.125	−403.13	−1.0675
1000	823.125	176.88	0.4684
800	823.125	−23.13	−0.0612
790	823.125	−33.13	−0.0877
1600	823.125	776.88	2.0571

(the negative numbers) and some of the rounds traveled further than the mean (the positive numbers). *An important note is that these negative and positive residuals will sum to zero.*

- *Column D* represents the standard residual for each observation. The standard residual is the residual found in Column C divided by the standard deviation calculated from Column A (found to be = 377.65). The standard deviation of 377.65 was found by merely using the descriptive statistics function on the Column A data.

Since Column D represents the standard residual for each observation, each data point can be described as follows. The standard residual of −0.5908 in Column D, Row #1 (for range = 600) can be interpreted as "−0.59 standard deviations to the left of the mean." Again, this was found by dividing the residuals in column C (= −223.13, or 223.13 meters short of the mean) by the standard deviation (= 377.65). The negative symbol in front of the 0.59 means that the value is *less* than the mean, or to the "left" of it on a scale (in this case, 600m vs. 823.125m). The standard residual in Row #2 (0.2698) can be interpreted as 0.2698 standard deviations to the *right* of the mean. It is a positive number since it is to the right and greater than the mean (in this case, 925m vs. 823.125m).

The first 7 observations all fall within 2 standard deviations of the mean. However, note that in Column D in the last row (for range = 1600 meters), the standard residual is 2.0571, which means that this data point is 2.0571 standard deviations to the right of the mean. This means that this observation is, by our definition, an outlier, as it is greater than 2 standard deviations from the mean and it should be examined as to whether it belongs in the current dataset. Outliers are not necessarily excluded from the dataset, but should be treated as useful information.

7.12.2 HANDLING OUTLIERS WITH RESPECT TO *Y* (THE DEPENDENT VARIABLE DATA)

There are two types of outliers with respect to the dependent variable *Y*:

- Those that are outliers with respect to Y itself (generally, the cost data) and
- Those that are outliers with respect to the regression model

Outliers with respect to *Y* itself are treated in the same way as those with respect to *X* (just shown) and are not of great concern. *But, outliers with respect to the regression model are of particular concern, because those represent observations that our model does not predict well.* When our regression gives us a prediction for a certain value of X, the outlier with respect to the regression model will be greater than two standard errors away from that prediction. These types of outliers with respect to the regression model $\widehat{Y}$ are identified by comparing the residuals to the standard error of the estimate (SE). This is referred to as the "standardized residual," similar to column D in Table 7.10 from Example 7.4. The correct computation for this is $\frac{(Y_i - \widehat{Y})}{\text{SE}}$ = Number of Standard Errors.

Remember: the fact that you have outliers in your data set is not necessarily indicative of a problem. Rather, it is important to determine *why* an observation is an outlier. Here are some possible reasons why an observation is an outlier:

- It was a random error, in which case it is not a problem
- It is not a member of the same population. If this is the case, you want to delete this observation from your data set
- You may have omitted one or more other cost drivers that should be considered, or your model is improperly specified with an incorrect cost driver. Engineers can help you resolve what the proper independent variables should be
- The data point was improperly measured (it is just plain wrong)
- Unusual event (war, natural disaster)

While the first four aforementioned reasons are fairly self-explanatory, let's discuss the final reason, an example of an "unusual event" that might lead to an outlier. Let's say that your local Home Depot sells 10 electric-producing home generators per month, on average.

The following is the number of home generators sold by Home Depot in the first eight months of a given year:

- January: 10
- February: 9
- March: 11
- April: 10
- **May: 257**
- **June: 153**
- July: 10
- August: 10

Notice the huge jump/increase in home generators sales in the months of May and June. What do you think happened to make this huge spike occur? Most likely there was a natural disaster such as a hurricane or an earthquake, and electric power was lost to a large number of homes in this area. Consequently, citizens of this town rushed to the local Home Depot to purchase a home generator to restore their electrical power during that timeframe. Also note that *after* the spike in May and June, sales returned to normal in July and August. This is a case where an outlier is not indicative of a problem and also should be treated as useful information. This would greatly enhance Home Depot's opportunity to predict future sales, both with and without a natural disaster.

Your first reaction should *not* be to throw out an outlier data point! In fact, since outliers provide useful information, here are two options to improve your regression results:

- Dampen or lessen the impact of the observation through a transformation of the dependent and or independent variables {e.g., using the natural log (ln)}
- Develop two or more regression equations (one with the outlier and one without the outlier)

If you are unfamiliar with the transformation of data using the natural log, this area will be covered in Chapter 9 on "Intrinsically Linear Regression."

7.13 Residual Analysis

The last section in this chapter deals with residual analysis. What if you perform a regression, and when looking at the statistics, you find that the F-statistic significance is high, the p-values for the t-statistics are high, the R-Square values are low, and the standard errors and CV's are high? Clearly this is not a good regression! The reason may be because you are trying to fit a straight line to data that are not linear. How can you tell if this is the case? The easiest way is to graph the original data in a scatter plot and see if it appears linear or not. Figure 7.2 was an example of data that is quite linear. But, there are many instances when data is not linear.

Another method – besides a scatter plot of the original data set – to determine if your data is nonlinear is to check the residual plots. Recall that the residuals (or errors) are the differences between the actual costs and your predicted costs or the actual costs and the mean.

If the fitted/regression model is appropriate for the data, there will be no pattern apparent in the plot of the residuals (e_i) vs. X_i. Note in Figure 7.7 how the residuals are spread uniformly across the range of the X-axis values. There are just as many residuals above the line as below the line. These residuals appear randomly and uniformly distributed:

But if the fitted model is *not* appropriate, your regression results will be poor and a relationship between the X-axis values and the e_i values will be apparent. Residual patterns will *not* look like what we see in Figure 7.7. To familiarize you with some of these
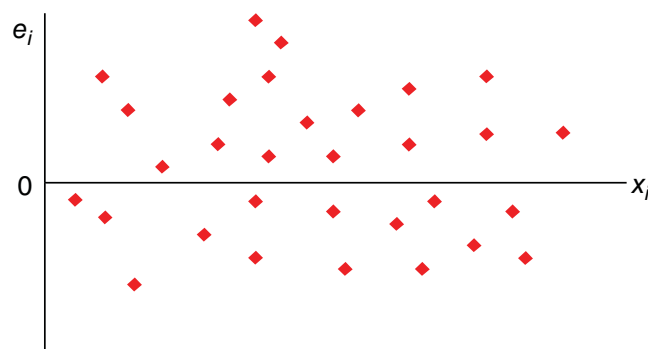
FIGURE 7.7 Residuals that are Spread Uniformly Across the Range of *X*-Values.

non-normal residual patterns, we will cover four models whose residuals are *not* uniformly distributed. They are:

- Non-normal distribution
- Curvilinear relation
- Influential relation
- Heteroscedasticity

Non-normal distribution: In Figure 7.8, note that the majority of the residuals are negative and are below the *X*-axis. These residuals are thus not normally distributed:
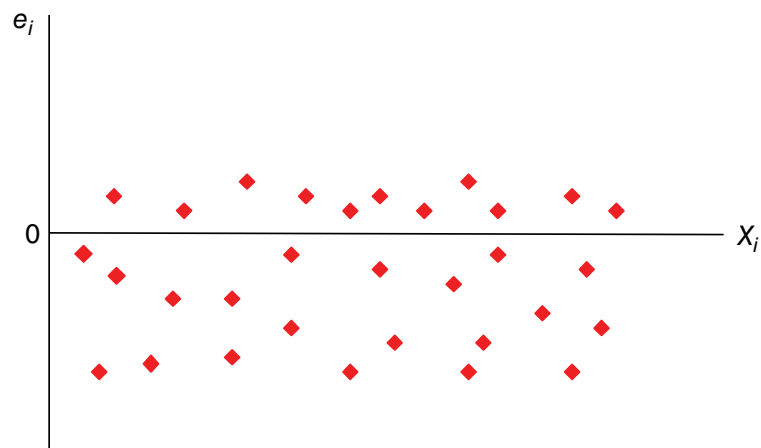
FIGURE 7.8 Residuals in a non-Normal Distribution.

Curvilinear relation: In Figure 7.9, note that the errors are small while *X* is small, increase in the middle of the data, and then get small again as *X* continues to increase.
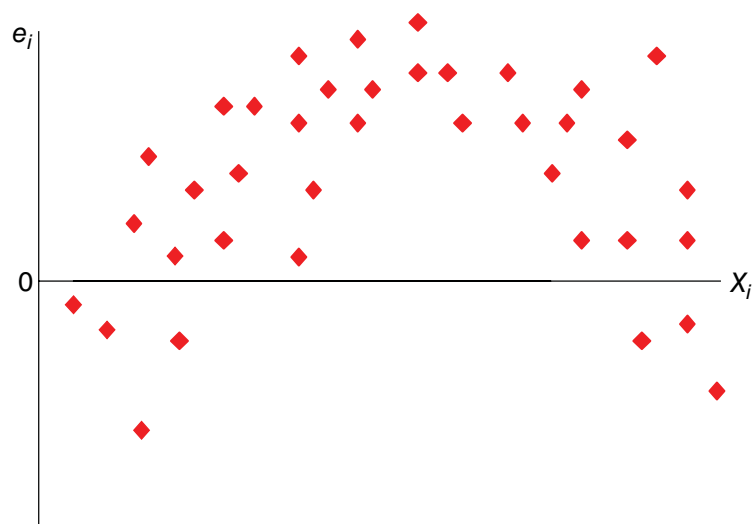
FIGURE 7.9 Residuals in a Curvilinear Relation.

This is what a plot of exponentially distributed data would look like if you tried to fit a straight line regression to it. Moreover, these residuals are clearly not normally distributed.

Influential relation: In Figure 7.10, note that the errors are negative at first and continue to rise and become increasingly positive as *Y* increases.
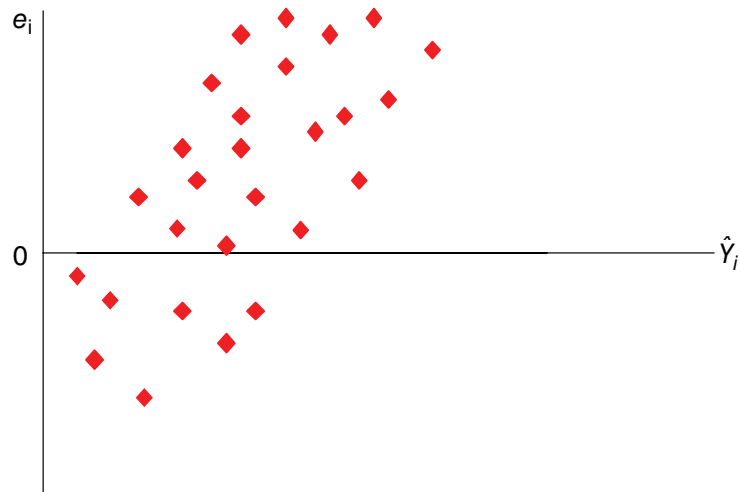
FIGURE 7.10 **Residuals in an Influential Relation.**

Heteroscedasticity: In Figure 7.11, note that the errors continue to increase both positively and negatively as t increases. Heteroscedastic properties are those of "unequal variance," as compared to homoscedasticity or "equal/similar variance."
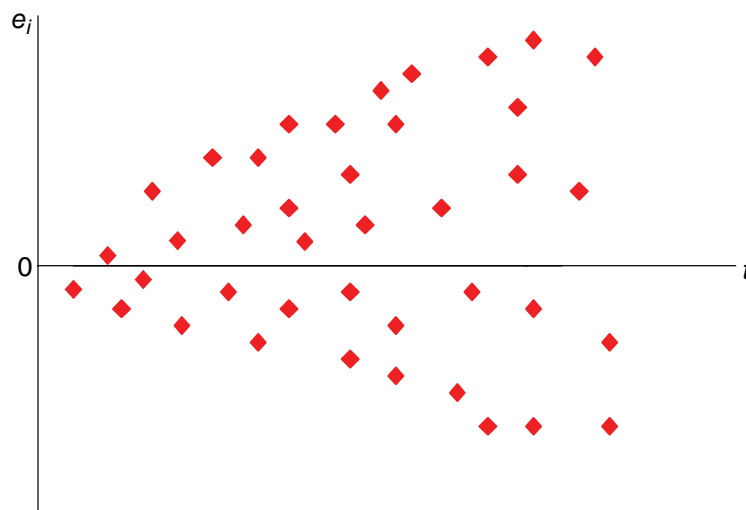
FIGURE 7.11 **Heteroscedastic Residuals.**

When performing residual analysis and any of the aforementioned four conditions exist in your residual plots, most likely your data is non-linear, or you need to add another,

possibly linear, independent variable to the model. Therefore, the standard properties of LSBF regression no longer apply. If this is the case, you may need to attempt data transformations to help your data "appear" more linear. Transformations may include the following:

$$X' = \frac{1}{X} \quad X' = \frac{1}{Y} \quad X' = \log X \quad Y' = \ln Y \quad Y' = \log Y$$

Log-linear transformation allows use of linear regression, but predicted values for Y are in "log dollars" or "ln dollars," which must be converted back to dollars. This transformation and conversion will be covered in greater detail in Chapter 9 on Intrinsically Linear Regression.

7.14 Assumptions of Ordinary Least Squares (OLS) Regression

The following is a list of the assumptions that must be followed when performing OLS regression:

- The values of the independent variables are known. For a fixed value of X, you can obtain many random samples, each with the same X values but different Y_i (cost) values due to different e_i values
- The errors (e_i) are normally distributed random variables with means equal to zero and constant variance (homoscedasticity). .... $e_i \sim N(0, s^2)$
- The error terms are uncorrelated

Summary

In this chapter, we introduced the topic of regression analysis and how regression is used to describe a *statistical relationship* between variables. We discussed the advantages of using regression over using the descriptive statistics of a data set and showed how using regression can improve your prediction in favor of just using the mean. In addition, we discussed the many statistics that are involved in regression analysis and the purpose for each one of them in the regression model. We used a housing example, regressing Price vs. Square Feet to highlight each of these statistics and followed that example through a majority of the chapter, especially trying to help you visualize the concept of standard error. We then introduced a "Hierarchy of Regression," which is a method to not only help you learn how to quickly and efficiently analyze a regression printout, but also to help you gain confidence in which metrics to look at and in what order. We will continue to use and expand this knowledge in Chapter 8, where we will advance from using a single independent variable in a regression to regressions with two or more independent variables. We do so to try to improve our ability to predict cost. We will also discuss the concept of multi-collinearity and the problems that it may cause in a multi-variable regression.

Author's Note: As noted in Chapter 6 on Statistics, there are numerous textbooks available on probability and statistics that discuss regression analysis that can be used as a general reference. In the Operations Research Department here at the Naval Postgraduate School, a widely

used textbook is written by Jay L. Devore. While we have used this text as our reference in this chapter, you may use any probability and statistics textbook that you prefer, as there are many good ones from which to choose.

One other note we would like to include is if your linear regression is not providing usable results and you have a large number of binary or categorical data in either your dependent or independent variables, you may want to consider using logistics regression instead of ordinary least squares regression. Logistics regression is also referred to as "logit regression." An example would be when your independent variables are basically 1/0 "Yes/No" answers along the lines of "Has the person completed this training?" or "Does this patient have a preexisting condition?" If this is the case, then your column of data will contain just two answers: "1 = Yes," and "0 = No." Even with this methodology, you still may not get a solid correlation, but it is always worth a try. A search online for "logistics regression" will yield numerous results if you need to learn more about this regression methodology.

Reference

1. Devore, Jay L. Probability and Statistics for Engineering and the Sciences, Sixth Edition, Thomson Brooks/Cole, Chapter 12.

Applications and Questions:

7.1 Regression Analysis is used to describe a _____relationship between variables.

7.2 While trying to derive a straight line prediction using regression analysis, the goal is to minimize the differences between the actual cost and the predicted cost (True/False)

7.3 You are assigned a project to cost a solar array panel to be delivered to the International Space Station. One of the popular metrics for the solar array panel is Beginning of Life (BOL) power. Historical data has been collected on ten programs similar to the one you are estimating and can be found in the following table:

Cost (2010$)	BOL Power (Watts)
1,200,000	800
2,300,000	1500
2,100,000	1300
1,600,000	900
1,700,000	1100
2,600,000	1600
2,100,000	1400
2,200,000	1450
2,100,000	1250
3,075,000	2400

Perform regression analysis on Cost vs. BOL Power. What is the regression equation produced?

7.4 Using the regression hierarchy, does this regression pass the Phase 1 criteria?

7.5 What are the values for R^2 and standard error for this regression?

7.6 What is the coefficient of variation for this regression? Is that value considered good or bad?

7.7 Are there any outliers in the BOL Power data? How do you determine that mathematically?

Multi-Variable Linear Regression Analysis

8.1 Introduction

In the previous chapter, we were introduced to regression analysis using a single independent variable. In this chapter, we will progress from having *one* independent variable in a regression to regressions with *two or more* independent variables. The purpose of considering and adding more variables is to try to *improve* our regression prediction, and to better explain our cost with multiple independent variables rather than with just a single variable. We will also discuss the widely misunderstood concept of Multi-Collinearity (MC), and the problems that it may cause in a multi-variable regression. We will discuss two ways to detect whether MC is present between your independent variables, and introduce you to a correlation matrix while doing so. If MC does exist, we will then determine whether or not it is causing a problem in your regression.

In your mathematics readings or in other textbooks, you may observe the title of this material written as Multiple Regression, Multi-variate regression or Multi-variable regression. They all mean the same thing: simply, a regression with more than one independent variable. Of the three terms, we have chosen to use Multi-variable regression.

8.2 Background of Multi-Variable Linear Regression

In the previous chapter, we predicted costs via regression analysis by using one independent variable. However, when you are predicting the cost of a product or program, you may have to consider more than one independent variable. The reason is that in any program you may have up to three *major* elements that could affect your cost:

- *Size*: This includes characteristics such as weight, volume, and quantity, etc …
- *Performance*: This includes characteristics such as speed, horsepower, and power output, etc …

- *Technology:* This includes characteristics such as gas turbine (versus, for example, turbofans), stealth, and composites, etc …

So far we've tried to select cost drivers that model cost as a function of *one* of these parameters, *X*:

$$Y_i = b_0 + b_1 X + e_i$$

But what if one variable is not enough? What if we believe that there are other significant cost drivers? In multi-variable linear regression, we will be working with the following model with multiple independent variables, X_i's:

$$Y_i = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + e_i$$

What do we hope to accomplish by considering additional independent variables? We hope to improve our ability to predict the cost of our system. We do this by reducing variation: not *total* variation (SST), but rather the *unexplained* variation (SSE), terms discussed in the previous chapter. Regardless of how many independent variables we bring into the model, we cannot change the total variation, because the total variation depends upon the differences between the actual costs and their mean:

$$SST = \sum (y_i - \bar{y})^2$$

But we *can* attempt to further minimize the unexplained variation, since that depends upon the differences between the actual cost and the predicted costs.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

The closer our prediction is to the actual costs, the smaller our errors, and thus the smaller the unexplained variation. Hopefully, adding the additional independent variables will accomplish this goal.

But what premium do we pay when we add another independent variable? We lose one degree of freedom for each additional variable, and if the sample size is small, it may be too large a price to pay. Why is this?

Recall Equation 7.9 for Standard Error (SE) in the single variable linear regression chapter, shown here again as Equation 8.1.

$$SE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} \tag{8.1}$$

The denominator is actually $n - k - 1$, where n is the number of data points, and k is the number of independent variables in that regression. In a single variable model such as in Equation 8.1 where $k = 1$, the denominator becomes merely $n - 2$ (since $n - k - 1 = n - (1) - 1 = n - 2$). But let's suppose that we have a multi-variable regression with four independent variables (thus $k = 4$). Here, the denominator for the standard error equation would now become $n - k - 1 = n - (4) - 1 = n - 5$. If we have a small data set with only 10 data points ($n = 10$), then our denominator in this equation would then become equal to $n - 5 = (10) - 5 = 5$. Since we are now dividing the numerator by only a factor of 5, our resultant answer for standard error would become correspondingly larger. This is

what is meant by "if the sample size is small, adding variables may be too large a price to pay," as it increases our standard error. If the sample size is much greater, though, at say $n = 100$, then the denominator of $n - k - 1 = n - 5 = 100 - 5 = 95$ would not cause as great an increase in the standard error, since n is large.

As in single variable linear regression, the same regression assumptions still apply in multi-variable regression:

- The values of the independent variables are known. For a fixed value of X, you can obtain many random samples, each with the same X values but different Y_i (cost) values due to different e_i values.

- The errors (e_i) are normally distributed random variables with means equal to zero and constant variance (homoscedasticity). You may see this indicated as follows in many textbooks: $e_i \sim N(0, s^2)$.

- The error terms are uncorrelated.

8.3 Home Prices

Example 8.1 Recall Example 7.1 from Chapter 7 on the "Price of a Home" versus its size using "Square Feet" as the independent variable. This original data and the regression results from this data are found in Table 8.1 and Table 8.2:

TABLE 8.1 Original Data Set of Prices and Square Feet of the 15 Homes in Example 8.1

Home #	Price ($)	Square Feet
1	300,000	1400
2	400,000	1800
3	350,000	1600
4	800,000	2200
5	450,000	1800
6	250,000	1200
7	225,000	1200
8	450,000	1900
9	550,000	2000
10	400,000	1700
11	220,000	1000
12	350,000	1450
13	365,000	1400
14	600,000	1950
15	750,000	2100

From Table 8.2, we can see that the results of this regression are quite good. Using the Hierarchy of Regression from the previous chapter as a guide, we can summarize the following:

Phase 1 Criteria:
- The regression passes the Common Sense test, as the slope of 450.539 is positive, so the "Price of Home" will increase as the "Square Feet" of the home increases, which makes intuitive sense. We are unconcerned that the intercept is negative!

TABLE 8.2 Regression Results from the Fifteen Homes in Example 8.1: Price of Home vs.Square Feet

A	B	C	D	E	F
Regression Statistics					
Multiple R	0.91936		**Price vs. Square Feet**		
R Square	0.84523				
Adjusted R Square	0.83332				
Standard Error	72183.1553				
Observations	15				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.69908E+11	3.69908E+11	70.99406372	1.26178E–06
Residual	13	67735302725	5210407902		
Total	14	4.37643E+11			
	Coefficients	*Standard Error*	*t Stat*	*P–value*	*Lower 95%*
Intercept	−311221.8767	90000.5658	−3.45800	0.00424	−505656.2777
Square Feet	450.5396	53.4714	8.42580	0.00000	335.0216

- The Significance of the F-statistic is very low at 1.26178×10^{-6}
- The P-value of the t-statistic is very low at 1.26178×10^{-6}

Therefore, this regression easily passes the first three (Phase I) criteria of the Regression Hierarchy. The estimated price is rising as the size of the home increases (as it should), and both the significances of the F-statistic and the t-statistic (p-values) are well below the .20 level threshold. Recall that these significances will be identical in a single variable regression. Moving to Phase 2 of the Regression Hierarchy, while we are not comparing this regression to another regression, we *can* look at the three metrics to determine whether we indeed have a good regression on its own merit:

Phase 2 Criteria:

- The *R*-Square value = 0.8452
- The Standard Error = 72,183.155, which equates to $72,183.15.
- The Coefficient of Variation (CV) = SE/mean = 72,183.15 / 430,666.67 = 16.76%

From the Phase 2 criteria, we can observe that our R-Square value is fairly high at .8452, and the standard error is reasonably low at $72,183.15. But the only way to really tell if the standard error is "good" or not is by calculating the coefficient of variation. In essence, the CV converts the actual standard error number (= $72,183.15) into a percentage. We had previously calculated the mean of the cost data by taking the average of the fifteen homes in Table 8.1, and this mean = $430,666.67. Therefore, when calculating the CV, we divide the standard error ($72,183.15) by the mean ($430,666.67), and we calculate the CV to be a value of 16.76%. This means that we assume that we will be off, on average, by ± $16.76% when using this regression.

Overall, these Phase 2 statistics are good, and the Price of Home versus Square Feet is a good regression. But there are other factors that also influence the price of a home besides

just Square Feet! Perhaps if we considered adding a few more independent variables, could we make an even better prediction? Would this help us to get even better statistics? Let us attempt this by adding two more independent variables. While there are many variables to choose from (such as location), for illustration we will now include both Number of Acres and Number of Bedrooms, in addition to the original independent variable of Square Feet.

Table 8.3 includes the original price data, as well as the data for all three independent variables: Square Feet, Number of Acres, and Number of Bedrooms.

TABLE 8.3 Data Set Revised to Include Two Additional Independent Variables

Home #	Y Price ($)	X1 Square Feet	X2 # of Acres	X3 # of Bedrooms
1	300,000	1400	0.25	2
2	400,000	1800	1	3
3	350,000	1600	0.25	2
4	800,000	2200	1.5	4
5	450,000	1800	0.5	3
6	250,000	1200	0.25	2
7	225,000	1200	0.5	2
8	450,000	1900	1	3
9	550,000	2000	0.5	4
10	400,000	1700	0.25	3
11	220,000	1000	0.25	2
12	350,000	1450	0.33	3
13	365,000	1400	0.5	2
14	600,000	1950	0.5	4
15	750,000	2100	2	4

Noting the ranges of all the data, we observe that:

- The home prices range from a minimum of $220,000 to a maximum of $800,000
- The square feet (size of the home) range from 1,000 sq ft to 2,200 sq ft.
- The number of acres range from 0.25 acres to 2 acres
- The number of bedrooms range from 2 to 4

Now let us perform a multi-variable regression on the Price of a Home versus all three independent variables. In your regression ToolPak, the Y value of "Price of Home" is the dependent variable. For the independent variables (your X's), highlight all three of the columns at once to be able to perform this regression. Results from this regression are found in Table 8.4:

Initially, let's use the Hierarchy of Regression to evaluate this regression.

Phase 1 Criteria:

- The regression passes the Common Sense test, as all three coefficients for slopes are positive for the independent variables, and the "Price of a Home" will increase as the "Square Feet," "Number of Acres," and "Number of Bedrooms" all increase.

TABLE 8.4   Regression Results from the Three Variable Data Set in Table 8.3

	Price of Home vs. Square Feet, Number of Acres and Number of Bedrooms				
Regression Statistics					
Multiple R	0.95879				
R Square	0.91929				
Adjusted R Square	0.89727				
Standard Error	56668.1703				
Observations	15				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	4.02319E+11	1.34106E+11	41.76103	2.65112E−06
Residual	11	35324096816	3211281529		
Total	14	4.37643E+11			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	−174904.2059	83169.5460	−2.1030	0.05929	−357959.1422
Square Feet	195.6465	98.6379	1.9835	0.07283	−21.4541
# of Acres	97399.4901	39914.5916	2.4402	0.03281	9548.0663
# of Bedrooms	77162.9897	39662.4018	1.9455	0.07771	−10133.3681

- The Significance of the F-statistic is very low at 2.6511×10^{-6}.
- All three of the p-values of the t-statistic are less than .20. The p-value for "Square Feet" is 0.0728; for "Number of Acres" it is 0.0328; and for "Number of Bedrooms" it is 0.0777. If any one of these three p-values had exceeded 0.20, we would have rejected this regression from further consideration.

Therefore, this regression easily passes the first three criteria of the regression hierarchy. But before moving to the Phase 2 criteria, we will offer a quick observation about the intercept being a negative number (= −174,904.2059), and why that is acceptable in this example. The reason that we are not concerned that the intercept is negative is because the range of the data in our independent variables does not include the y-axis or origin. X_1=Square Feet ranges from a low of 1,000 to a high of 2,200 sq feet; X_2= Number of Acres ranges from 0.25 to 2 acres; and X_3=Number of Bedrooms ranges from 2 to 4 bedrooms. None of the data for any of the independent variables that underlie our regression equation includes 0 (the origin), or even numbers close to 0. Such small numbers would have given us a negative output from the regression equation. However, even if we input the lowest value for each of the independent variables into the regression equation (= 1.000; 0.25; 2), we would still get a positive output, and therefore there is no concern that we have a negative number for our intercept. This is the same reason that we disregard evaluating the p-value for the intercept.

Moving to the Phase 2 criteria, we find the following:

Phase 2 Criteria:

- The R-Square value = 0.9193
- The Standard Error = 56,668.17, which equates to $56,668.17.
- The Coefficient of Variation (CV) = SE/mean = 56,668.17 / 430,666.67 = 13.16%

From the Phase 2 criteria, we find that we have a good regression with solid statistics. But is this three variable regression better than the single variable regression of "Price of Home" vs. "Square Feet?" Let's compare the Phase 2 criteria for the two regression results. These are found in Table 8.5.

TABLE 8.5   Comparison Chart between the Two Regressions in Example 8.1

	Adjusted R^2	**Std Error**	**CV**
Price vs.Square Feet	0.8333	72,183.15	16.76%
Price vs All Three	0.8973	56,668.17	13.16%

The chart reveals clearly that the regression using all three independent variables has results that are statistically superior to the single variable regression model. First, we compared the R-Square values, but actually needed to use the Adjusted R-Square values to do so, as there were a different number of independent variables in each regression (3 vs. 1). The comparison for the Adjusted R-Square was 89.73% versus 83.33%. The standard error was \$56,668.17 vs. \$72,183.15; and the CV was 13.16% vs. 16.76%, once again using the cost mean of \$430,666.67. Consequently, the conclusion would be that the regression with Price versus all three independent variables is statistically the superior regression, and the one that we would use for further calculations. To summarize these results, we would state the following:

- Adjusted R-Square: That 89.73% of the total variation is explained by the regression model.
- SE: That when using this regression, one standard error is $\pm$ \$56,668.17.
- CV: That when using this regression, one standard error is $\pm$ 13.16%.

Now that we can perform a multi-variable regression and evaluate its statistics, we will introduce the concept of Multi-Collinearity and discuss when, if, and how it can cause problems in a multi-variable regression.

8.4 Multi-Collinearity (MC)

What do the slope coefficients (b_1, b_2, ... , b_k) represent? In a simple linear model with one independent variable X, we would say b_1 represents the change in Y given a one unit change in X. Consider $Y = 2 \times X$. For every unit of increase in X, Y will change/increase by a factor of 2. So, $b_1 = 2$ (the slope) represents "the change in Y given a one unit change in X." But in the multi-variable model, there is more of a *conditional* relationship. Y is determined by the *combined* effects of *all* of the Xs. Thus, we say that b_1 represents the *marginal* change in Y given a one unit change in X_1, *while holding all of the other X_i's constant*.

An example of this can be demonstrated in the financial world of stocks vs. mutual funds. Let's say that you own stock in Company A. If the price of the stock of Company A doubles in one day, you have just doubled the amount of money that you own in that account! However, if you have a mutual fund that has 100 stocks in its portfolio (only one of which is Company A), then if the Company A stock doubles in one day, it does not mean that your *entire mutual fund* will double. The end result of how much your mutual

fund is worth at the end of that day *depends on*, or *is conditional on*, what the other 99 stocks did that day as well. Similarly, the final cost in a multi-variable regression *depends on*, or *is conditional* on, the contributions of *all* of the independent variables in the equation.

One factor in the ability of the regression coefficient to accurately reflect the marginal contribution of an independent variable is the *amount of independence* between the independent variables. If X_1 and X_2 are statistically independent, then a change in X_1 has no correlation to a change in X_2. There are very few examples of this in real life. One is perhaps the size of a house versus the color of that house or perhaps the geographical location (i.e., different states) of that house? But even that second example could conceivably have some correlation.

Usually, however, there is *some amount of correlation between variables*. When variables are highly correlated to each other, it means that they are very similar in some way to each other. Multi-collinearity occurs when X_1 and X_2 are highly related (thus similar) to each other. When this happens, there is an "overlap" between what X_1 explains about Y (the cost) and what X_2 explains about Y. This makes it difficult to determine the true relationship between X_1 and Y, and X_2 and Y, or which variable is really driving the cost.

For example, let's say that you own a car that has a 300 horsepower (hp) engine that can attain a top speed of 140 miles per hour (mph). Suppose that you want the car to go faster than 140 mph because you intend to race it on the local race track. To increase the speed, you decide to replace your 300 hp engine with a bigger and more powerful 400 hp engine. But when you do this, you are most likely *also* increasing the weight of the engine as well. Thus, the power of the engine and the weight of the engine are both increasing at the same time, and these two independent variables are most likely highly correlated to each other. In this scenario, there is a high probability that multi-collinearity will exist between power and weight. How do we know for sure if it exists?

There are two ways to detect multi-collinearity between independent variables:

- *Method 1* is by observing the regression slope coefficients in simple linear regressions for each variable independently, and then comparing the same slope coefficients when using both variables in a multi-variable regression. If the value of b_1 or b_2 changes significantly from one regression to another, then there may be a significant amount of correlation between X_1 and X_2.

- *Method 2* is by computing a pair-wise correlation matrix. This can be carried out by a number of software packages. In our opinion, this is the easier and more insightful of the two methods. We will explore the virtue of each method in the following section.

8.5 Detecting Multi-Collinearity (MC), Method #1: Widely Varying Regression Slope Coefficients

When X_1 and X_2 are highly correlated, their coefficients (b_1 and b_2, either one or both) may change significantly from a one-variable regression model to the multi-variable regression model. Consider the following regression equations calculated from a missile data set given the standard format:

$$Y_i = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + e_i$$

Example 8.2 Missile data set:

- Regression Equation #1: Cost $= -24.486 + 7.789 \times$ Weight
- Regression Equation #2: Cost $= 59.575 + 13.096 \times$ Range
- Regression Equation #3: Cost $= -21.878 + 8.317 \times$ Weight $+ 3.110 \times$ Range

   What conclusions can we gather from these three regression equations?

- In Equation #1, *Cost vs. Weight*: Note that the slope coefficient b_1 for the independent variable Weight $= 7.789$.
- In Equation #2, *Cost vs. Range*: Note that the slope coefficient b_2 for the independent variable Range $= 13.096$.
- In Equation #3, *Cost vs. Weight and Range*: In this multi-variable regression model, note that the slope coefficient (b_1) for Weight in Equation #3 is now 8.317, which is relatively unchanged from the slope in Equation #1 (from 7.789 to 8.317). However, notice how drastically the slope coefficient for Range (b_2) has changed, from 13.096 in Equation #2 to only 3.11 in Equation #3. This shows that the slope contribution from the independent variable Range is significantly different in the multi-variable regression model from what it contributed in the single variable model, while the contribution from Weight is virtually identical in both regressions. This signifies that MC is most likely present between the two independent variables, Weight and Range.

   One thing to keep in mind is that the presence of MC is not necessarily *or* automatically *a bad thing!*

8.6 Detecting Multi-Collinearity, Method # 2: Correlation Matrix

Producing a correlation matrix can be accomplished in a number of software statistical packages. In Excel, it would be found in the Data Analysis ToolPak Add-Ins/Correlation. Once the correlation matrix is computed, the values in the matrix represent the "r" values, or correlation, between the variables. A higher value for "r" in the matrix between independent variables represents a higher correlation (or similarity) between those two independent variables. Conversely, lower values for "r" represent lower correlations or similarity between those two independent variables. We will define variables as "multi-collinear," or highly correlated, when $r \geq 0.7$. The reader should note that this threshold is a matter of convenience, defined by the authors, rather than a universally accepted standard.

A general Rule of Thumb for determining if you have MC by using a correlation matrix is:

$r \leq 0.3$	Low Correlation: little to no MC present
$0.3 \leq r \leq 0.7$	Gray Area: may or may not have MC present
$r \geq 0.7$	High Correlation: you will need to perform an MC check

Table 8.6 displays a sample correlation matrix that was notionally calculated to show the correlation between Weight and Range from Example 8.2. Note that the correlation

TABLE 8.6   Sample Correlation Matrix for Method 2

	Weight	**Range**
Weight	1	
Range	0.9352	1

between Weight and Range is $r = .9352$. Using the general Rule of Thumb matrix from above, we see that these two independent variables are highly correlated to each other, since $r \geq 0.7$. Therefore, MC would be present in a regression of Cost vs. Weight and Range.

Correlation matrices will be discussed in greater depth in the next section, Section 8.7.

8.7 Multi-Collinearity Example #1: Home Prices

Let's look at attempting a multi-variable regression and providing an in-depth example of using a correlation matrix. To do this, let's consider the following "Price of Home" data on eight homes, displayed in Table 8.7. The data includes three independent/explanatory variables: Square Feet, Number of Acres, and Number of Bedrooms, to try to help predict the Price of a Home. We will use this data to demonstrate how to detect for possible multi-collinearity in your data set. If it is present, we will then determine whether or not it is causing a problem.

TABLE 8.7   Data set for Example 8.2 with Three Independent Variables

A Price of home	**B** Square feet	**C** # Acres	**D** # Bedrooms
$500,000	1,400	0.25	2
$600,000	1,800	0.50	5
$550,000	1,600	0.33	3
$1,000,000	2,200	3.00	4
$650,000	1,800	0.75	3
$450,000	1,200	0.20	2
$425,000	1,200	0.25	3
$650,000	1,750	0.50	4

Using this data in Table 8.7, let us now compute a correlation matrix using Excel's Data ToolPak (Data Analysis/Correlation, then highlight all of the data). This correlation matrix is found in Table 8.8.

Initially, note the column of 1's on the diagonal in Table 8.8. This is because the variables in those cells are being correlated (or compared) against themselves (i.e., Price of Home vs. Price of Home, or Square Feet vs. Square Feet), so the correlation is $r = 1$, since they are "exactly" equal or similar to themselves. The rest of the values shown are the correlations between one variable to another. If you look at column B, you can see that "Price of Home" is highly correlated to "Square Feet" ($r = 0.9363$) and "Number of Acres" ($= 0.9475$), but is not as correlated to "Number of Bedrooms" ($= 0.5276$). Since

TABLE 8.8   Correlation Matrix for the Data in Table 8.7

A	B	C	D	E
	Price of home	**Square feet**	**# Acres**	**# Bedrooms**
Price of home	1			
Square feet	0.9363	1		
# Acres	0.9475	0.7951	1	
# Bedrooms	0.5276	0.6907	0.3832	1

the "Price of Home" (*Y*) is the dependent variable, you *desire* for the correlation to be high between *Y* and the independent variables, since they are the variables that we are using to try to explain the Price!

Author's Note: These "r" values in Column B can also give you an initial indication of how good a single variable regression would be between these variables and Price. If the correlation value "r" is high, you will most likely have a very good single variable regression between Price and that variable. If the correlation value is low, however, you will most likely have a moderate-to-poor single variable regression between Price and that variable.

When testing for MC in a regression, though, we are only concerned about whether the *independent variables* are highly correlated ($X1$, $X2$, etc.). Thus, we should compute a correlation matrix for only the independent variables (i.e., exclude Price from this matrix). This time, let's calculate a correlation matrix that *only includes the independent variables.*

Table 8.9 is the correlation matrix from this example that only includes the independent variables. We know that MC is present when two *independent* variables are highly correlated ($r \geq 0.7$) to each other.

TABLE 8.9   Correlation Matrix for the Data in Table 8.7, now excluding Price of Home

A	B	C	D
	Square Feet	**# Acres**	**# Bedrooms**
Square Feet	1		
# Acres	0.7951	1	
# Bedrooms	0.6907	0.3832	1

From the table, we again note the 1's in the diagonal and can also notice in Column B that Square Feet and Number of Acres are highly correlated to each other at $r = 0.7951$ (since $r \geq 0.7$). Square Feet and Number of Bedrooms are correlated at $r = 0.6907$ (in the gray area and very near 0.7), and in Column C we find that Number of Acres and Number of Bedrooms are correlated at $r = 0.3832$, which is a fairly low correlation.

So does the fact that Square Feet and Number of Acres are highly correlated to each other at $r = 0.7951$ mean that we have a problem? The answer is … … … … perhaps, but we don't know yet … … ….. MC can be present and *not* be creating a problem. Or, it can be present, and it *can* be causing a problem.

So how do we know which is the case?

8.8 Determining Statistical Relationships between Independent Variables

In general, MC does not necessarily affect our ability to get a good statistical fit, nor does it affect our ability to obtain a good prediction, *provided that we maintain the proper statistical relationship between the independent variables*. If we do *not* maintain this statistical relationship, MC can create variability and thus instability in our regression coefficients, thus rendering our final answers unstable and inaccurate.

So how do we determine that statistical relationship? We do so by calculating a simple linear regression between the two highly correlated independent variables. In this case, since the correlation between Square Feet and Number of Acres is $r = 0.7951$, we will run a regression between Square Feet and Number of Acres, to find out what the statistical relationship is between those two variables. The results of this regression can be found in Table 8.10.

TABLE 8.10 Regression Result of Square Feet vs. Number of Acres

SUMMARY OUTPUT	Square Feet vs. Number of Acres				
Regression Statistics					
Multiple R	0.7951				
R Square	0.6321				
Adjusted R Square	0.5708				
Standard Error	224.1764				
Observations	8				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	518157.0675	518157.0675	10.3105	0.0183
Residual	6	301530.4325	50255.0721		
Total	7	819687.5000			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	1409.2105	102.6661	13.7262	9.295E–06	1157.9957
# Acres	290.0200	90.3207	3.2110	1.834E–02	69.0133

From the regression, we find that *Square Feet = 1409.2105 + 290.02 × Number of Acres*.

This is the statistical relationship between Square Feet and Number of Acres that we needed to derive. This relationship must be maintained in future calculations in order for us to be able to use this regression. An amplification of why this is important is found in MC Example #2.

8.9 Multi-Collinearity Example #2: Weapon Systems

Consider the following data set found in Table 8.11. We have been tasked to determine the cost of a new weapon system, using both its Power and Weight as the independent variables. All costs have been normalized to FY12$.

TABLE 8.11   Data Set for Multi-Collinearity Example #2 with Two Independent Variables

Cost	Power (hp)	Weight (lbs)
$8,000	250	600
$9,000	275	650
$11,000	300	725
$14,000	325	800
$16,000	350	810
$19,000	375	875
$21,000	400	925
$23,000	450	1,000

Let's try to determine if MC is present in this data set by using both methods. Using Section 8.5 Method 1, let's first generate the following three regressions using the data set in Table 8.11:

- Cost vs. Power
- Cost vs. Weight
- Cost vs. Power and Weight

Here are the results from these three regressions:

- Equation #1: Cost $= -13{,}115.28 + 82.907 \times$ Power
- Equation #2: Cost $= -17{,}317.92 + 40.649 \times$ Weight
- Equation #3: Cost $= -15{,}128.54 + 45.884 \times$ Power $+ 18.323 \times$ Weight

Analyzing these three equations, we find the following results:

- The slope coefficient for Power has changed significantly from 82.907 in Equation #1 to 45.884 in Equation #3
- The slope coefficient for Weight has changed significantly from 40.649 in Equation #2 to 18.323 in Equation #3.

Both of these results are an indication that Power and Weight are highly correlated. Thus, MC is most likely present, and we must then determine if it is creating problems or not.

We can also use Section 8.6, Method 2 to determine if MC is present by computing a correlation matrix. Results of doing so are found in Table 8.12. Both Method 1 and Method 2 establish whether multi-collinearity is present between independent variables.

TABLE 8.12 Correlation Matrix for the
Multi-Collinearity Example #2

	Power	Weight
Power	1	
Weight	0.9915	1

Using Method 2, we can see that the correlation is $r = 0.9915$ between Power and Weight. We know that when $r \geq 0.7$, MC is present. Using the data in Table 8.11, we must now regress Power vs. Weight (or Weight vs. Power) to see what the statistical relationship is between these two variables. Here are the results of those two regressions:

$$Power = -47.715 + 0.4865 \times Weight$$

or

$$Weight = 109.874 + 2.02 \times Power$$

It does not matter which of these two equations we choose to work with, but for the purpose of this example, let's use the first regression. Since Power $= -47.715 + 0.4865 \times$ Weight, we have established the statistical relationship between the two independent variables in our regression, and we must maintain this statistical relationship in the development program that we are trying to estimate, to ensure that our regression is giving us accurate results.

We could have also chosen the second regression and worked with Weight vs. Power, and again it does not matter which equation we choose to work with. Why does it not matter? Looking at the two regressions, you note that in the Power vs. Weight regression, the slope coefficient for Weight is essentially 0.5. In the Weight vs. Power regression, the slope coefficient for power is essentially 2.0. So overall, Power is approximately 1/2 of Weight, and Weight is conversely 2 times Power, merely the inverse of each other. Regardless of which regression you choose, the relationship should hold in either equation.

Let's examine two cases that test this statistical relationship between Power and Weight.

Case 8.1 The statistical relationship between Power vs. Weight has been established as:

$$Power = -47.715 + 0.4865 \times Weight$$

Suppose that in the system you are costing, the engineers have determined that Power will most likely $= 360$ hp, and Weight will most likely $= 950$ lbs. In this case, is the statistical relationship between Power and Weight being maintained? Or, in mathematical terms,

Does 360 (the likely value for Power)

$$= 47.715 + 0.4865 \times (950) \ ?? \ \ (Approximately?)$$

After inserting the values of power and weight into the equation and performing the calculations, does $360 = 414.46$ (approximately)?

Answer: Yes, in this case, there is only about a 15% difference between the actual and the projected costs. *The "rule of thumb" that we generally use is that this difference should be no greater than 30 – 35%.* In this case then, we would say that multi-collinearity is present and is *not* causing a problem. Why is this significant? The answer can be seen in the following three calculations.

Recall the following three regressions from the *Multi-Collinearity Example #2:*

- Equation #1*:* Cost $= -13{,}115.28 + 82.907 \times$ Power

- Equation #2: Cost $= -17{,}317.92 + 40.649 \times$ Weight

- Equation #3: Cost $= -15{,}128.54 + 45.884 \times$ Power $+ 18.323 \times$ Weight

Using the projected values of Power $= 360$ hp and Weight $= 950$lbs in these three regressions, we get the following results (all costs in FY12$):

- Cost $= -13{,}115.28 + 82.907 \times (360) = $ *$19,218.45*

- Cost $= -17{,}317.92 + 40.649 \times (950) = $ *$21,298.63*

- Cost $= -15{,}128.54 + 45.884 \times (360) + 18.323 \times (950) = $ *$20,173.07*

We can easily see that the final cost estimates in Case 8.1 are very similar and consistent in all three equations, and this is because Power and Weight *did* follow the required statistical relationship needed between them.

Case 8.2   But now suppose that in the system you are costing, the engineers have determined that Power will most likely $= 260$ hp and Weight will most likely $= 975$ lbs.

Does 260 (the likely value for Power) $= -47.715 + 0.4865 \times (975)$?? After performing the calculations, does $260 = 426.62$ (approximately)?

Answer: No, since the difference is about a 65% difference, which is much greater than the 30–35% Rule of Thumb. Observe the following, then, why this makes a difference. The cost estimates are very dissimilar (with a wide variance) in all three equations, since Power and Weight did not follow the required relationship (all costs in FY12$):

- Equation #1: Cost $= -13{,}115.28 + 82.907 \times$ Power

- Equation #2: Cost $= -17{,}317.92 + 40.649 \times$ Weight

- Equation #3: Cost $= -15{,}128.54 + 45.884 \times$ Power $+ 18.323 \times$ Weight

Using the values now of Power $= 260$ hp and Weight $= 975$lbs in these three regressions, we get the following results (all costs in FY12$):

- *Cost $= -13{,}115.28 + 82.907 \times (260) = $ $8,440.54*

- *Cost $= -17{,}317.92 + 40.649 \times (975) = $ $22,314.85*

- Cost $= -15{,}128.54 + 45.884 \times (260) + 18.323 \times (975) = $ $14,666.23

We can see that the final cost estimates in Case 8.2 are very dissimilar in all three equations, because the inputs for Power and Weight did not hold the required statistical relationship. Thus, our results become unstable and unusable.

8.10 Conclusions of Multi-Collinearity

In MC Example #2, when the statistical relationship was maintained between the two highly correlated variables – Power and Weight – the resultant answers were very consistent and stable. This was shown in Section 8.9, Case 8.1. But when the statistical relationship was not maintained between the two highly correlated variables, the resultant answers were inconsistent and unstable. This was the issue encountered in Section 8.9, Case 8.2. The result is that you are no longer able to use that regression and will instead have to revert to your second best regression.

Hopefully, MC will not be a problem in the second best regression, too! If the problem does persist, perhaps the engineers were incorrect in the determination of what the value of the independent variables would be (i.e., perhaps the estimate that Weight will equal 975 lbs is incorrect). If that is the case, perhaps the input values will change, and the statistical relationship between the two variables would once again be able to be maintained.

If multi-collinearity does exist, there are two possible final conclusions:

- That it exists, and it is *not* causing any problems. In that case, you may continue to use that regression.
- That it exists, and it *IS* causing a problem. In this case, you will not be able to use that regression, as MC is making the regression unstable, with unpredictable results. Answers in your MC check should be within approximately 30–35% when using the values provided for the new system. If not, then it is most prudent to use the next best regression.

Mathematically, when multi-collinearity is present, we can no longer make the statement that b_1 is the change in Y for a unit change in X_1, while holding X_2 constant. The two variables may be related in such a way that precludes being able to vary one while the other is held constant. For another example, perhaps the only way to increase the range of a missile is to increase the amount of the propellant, thus increasing the missile weight. So range and propellant weight will be highly correlated.

One other effect is that multi-collinearity might prevent a significant cost driver from entering the model during model selection. If this is the case, you should consult with your engineers on whether they still believe the variable is an important cost driver. If it is, should we then drop a variable and ignore an otherwise good cost driver? The short answer is "Not if we don't have to." It would be best to involve the technical experts and try to determine if the model is correctly specified. Perhaps you have included variables that are essentially the same thing (like weight in pounds as one independent variable and also weight in kilograms, so they will be very highly correlated). You can also try to combine the variables by multiplying or dividing them. Instead of using miles and gallons as two independent variables, combine them to make miles/gallon, if at all possible.

When the regression results are illogical in a multi-variable regression, such as when cost varies inversely with a physical or performance parameter, omission of one or more important independent variables may have occurred, or the variables being used may be

highly correlated. This does not necessarily invalidate a linear model, but additional analysis of the model is now necessary to determine if additional independent variables should be incorporated, or if consolidation or elimination of existing variables is necessary. It is important to apply common sense to the equations to ensure a good model. Work with your engineers to do this when your equations or your results do not make sense.

8.11 Multi-Variable Regression Guidelines:

Guideline #1 If you are performing a series of multi-variable regressions to find which regression gives you the best results, we offer the following template:

First, attempt the full regression model. That is, if we are costing out a new system and we have historical data on Length, Weight, and Range as our three independent variables, perform a regression using all three independent variables:

- Cost vs. Length, Weight and Range

But let's suppose that the results of that regression show that it *does not* pass the Phase 1 criteria of the Regression Hierarchy. Perhaps the p-value for Length or Weight exceeds the 0.20 significance level threshold, or it does not pass the common sense test for one of the slopes. Consequently, since you would reject that regression model, you would then need to attempt the following pair-wise regressions (or all three combinations of the two-variable regressions):

- Cost vs. Length and Weight
- Cost vs. Length and Range
- Cost vs. Weight and Range

Hopefully, one or more of those regressions will be considered a "good" regression. But if none of those regressions pass the regression hierarchy, then you can always perform the single variable regressions and choose one of these as your preferred model:

- Cost vs. Length
- Cost vs. Weight
- Cost vs. Range

Guideline #2 An issue that can arise when performing these regressions is a question that we have been asked in class on numerous occasions. The question has to do with a scenario when comparing regressions to each other. What if, in the course of doing your regressions, you get the following as the two best regression models from a statistical standpoint:

- Cost vs. Length and Weight
- Cost vs. Range

Let's suppose that it turns out that Cost vs Range, the single variable regression, has just a slightly higher Adjusted R^2 by a very slim margin, say 91.5 vs. 91.0%, and the standard error is lower also by a very small amount. (Note: Recall that you use Adjusted R^2 when comparing two regressions with an unequal number of independent variables). This now becomes a judgment call. Do you use Cost vs. Range because it has just a slight advantage in the statistical comparison? With the Cost vs. Length and Weight regression, however, you have *two* independent variables with which to help predict the cost of your system, whereas in Cost vs. Range, you are only using *one* independent variable to make that prediction. Our personal opinion is that in the majority of times, we would prefer to use the regression that has the two independent variables, compared to relying on one variable only, since it utilizes two variables and twice as much data. We realize that that would go slightly against the statistics. We would certainly solve for both regressions: putting the values for Length and Weight in the first regression, and then the value for Range in the second regression, to see if the cost estimates come out similar to each other. If they do, then you can feel confident that your answer is reasonably accurate, despite which one you select, since both regressions are yielding approximately the same answer. If they do not, then again it becomes a judgment call, and we would highly recommend you discuss the independent variables with the engineers to assist you in that decision.

Summary

In this chapter, we progressed from having *one* independent variable in a regression to using regressions with *two or more* independent variables, what we call multi-variable regression analysis. We added additional variables to improve our prediction and decrease our standard error, and we provided an example of a multi-variable regression model versus a single variable regression model, and then we compared the resultant statistics against each other. In this case, the multi-variable regression produced the better statistical results and would be the desired regression to use. But while attempting to use a multi-variable regression, one might encounter the effect of multi-collinearity. Multi-collinearity occurs when two independent variables are highly correlated (or similar) to each other. We discussed two ways in which to detect whether MC is present between your independent variables and introduced you to a correlation matrix. Once it was established that MC was present, we then demonstrated how to perform a regression between the two highly correlated independent variables, to determine the statistical relationship that needs to be maintained between those two variables in future calculations.

One key point to keep in mind is that the presence of MC is not *necessarily* or *automatically* a bad thing! But if multi-collinearity does exist, there are two possible final conclusions:

- That it exists, and when the statistical relationship between the independent variables is maintained, it is *not* causing any problems. In that case, you may continue to use that regression.
- That it exists, and when the statistical relationship between the independent variables is unable to be maintained, then MC *is* causing a problem. In this case, you will not be able to use that regression, as MC is making the regression unstable with unpredictable results.

Finally, we provided guidelines to assist in selecting the proper regression when a few scenarios are encountered.

We have now discussed single and multi-variable regression analysis using *linear* data sets. In Chapter 9, we will discuss how to handle data sets that are *nonlinear*.

Applications and Questions:

8.1 The purpose for using more than one independent variable in a regression is to try to *improve* our regression prediction, to better explain our cost with multiple independent variables rather than with just a single independent variable. (True/False)

8.2 By introducing additional variables into our regression, we cannot change the total variation in the model. However, we *can* attempt to further minimize the _____ _____, since that depends upon the differences between the actual cost and the predicted costs.

8.3 You are assigned a project to cost a solar array panel to be delivered to the International Space Station. As we saw in Chapter 7, one of the popular metrics for the solar array panel is Beginning of Life (BOL) power. We will now add an additional independent variable of Efficiency. Historical data has been collected on ten programs, and can be found in the following table:

Cost (2010$)	BOL Power	Efficiency (%)
1,200,000	800	14
2,300,000	1500	20
2,100,000	1300	21
1,600,000	900	17
1,700,000	1100	18
2,600,000	1600	24
2,100,000	1400	26
2,200,000	1450	22
2,100,000	1250	25
3,075,000	2400	28

Perform a multi-variable regression analysis on Cost vs. BOL Power and Efficiency. What is the regression equation produced?

8.4 Using the regression hierarchy, does this regression pass the Phase 1 criteria?

8.5 What are the values for R^2 and standard error for this regression?

8.6 What is the coefficient of variation for this regression? Is that value considered good or bad?

8.7 Compare the results from the single linear regression of Cost vs BOL Power in Chapter 7, Question 7.3 at the end of the chapter to your answers in 8.5 and 8.6 here. Which regression has the better statistics?

8.8 Create a correlation matrix for the data in Problem 8.3 above. Are the independent variables highly correlated to each other? How do you know if they are or not?

8.9 If you find that they are highly correlated, what would you do to check whether multi-collinearity was causing a problem or not in your preferred regression? What is the statistical relationship between BOL Power and Efficiency?

Intrinsically Linear Regression

9.1 Introduction

The first thing that you will likely notice is the unusual title of this chapter. Why not just use the more commonly known terms of "non-linear regression" or "curvilinear regression?" The term "Intrinsically linear regression" is used when your data set is not linear, but some transformation may be applied in order to make the data appear/become linear. Thus, intrinsically linear regression is a regression that contains transformed data. There are many methods of transformations including logs, square roots, and inverses. In this chapter, we will concentrate on handling data that is not linear and transform it using the natural log(ln).

After providing an example data set that initially reveals that the data is non-linear, we will apply a natural log transformation and show in a scatter plot that the data is now linear. We can then perform a "least squares best fit" (LSBF) regression to establish the intercept and slope of that line. However, the equation that we derive from that regression will have results that are in natural log units, such as "ln dollars" or "ln hours." Therefore, we will need to transform the regression equation back to its original units (i.e., dollars or hours), using the exponential function. After the transformation, we will have calculated the non-linear regression equation that best describes and fits the original data set.

9.2 Background of Intrinsically Linear Regression

In Chapters 7 and 8, we discussed a number of assumptions of the standard regression model. One of the assumptions was that the residuals are normally distributed. If this assumption does not hold, then we may have non-linear data, and if so, we will need to transform the data into a form that will make the data "appear" linear, so that ordinary least squares regression analysis can then be used on the transformed data. Sometimes even this

is not possible, as some data are just not able to be transformed to make it linear (e.g., data that follows the structure of a sine wave). But in this chapter, we will work with data that is non-linear but that *can* be transformed ("intrinsically linear") to appear linear after transformation. All the steps for linear regression may then be performed on this transformed data.

Three of the most common forms of non-linear models that can be transformed are the following: [1]

- Model #1: Logarithmic: $y = a + b \times \ln x$
- Model #2: Exponential: $y = a \times e^{bx}$
- Model #3: Power: $y = a \times x^b$

These three non-linear models are shown graphically in Figure 9.1.
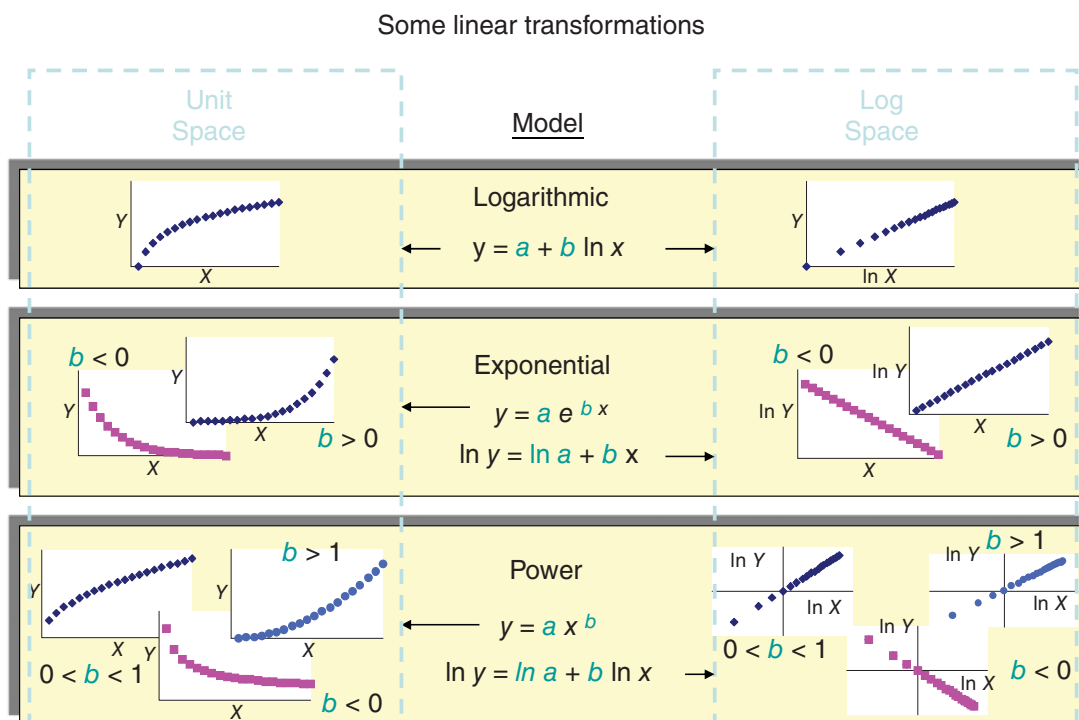
FIGURE 9.1 Three Examples of Non-Linear Transformations.

Note that the three non-linear models are shown graphically on the left side in Figure 9.1 in the column labeled "Unit Space." Once the transformation has been applied, they now appear linear and their scatter plots are found in the right column labeled "Log Space." Of the three models shown earlier, we will concentrate on the third one, the *Power* model, since we will use it extensively in the upcoming chapters on Learning Curves.

9.3 The Multiplicative Model

In a linear equation, $y = b_0 + b_1 x$, we have learned that a unit change in X causes Y to change by b_1. But in a multiplicative equation, a change in X causes Y to change

by a percentage that is *proportional* to the change in X. These differences are shown in Figure 9.2.

The Multiplicative Model

- *Linear equation* : A unit change in X causes Y to change by b_1

- *Multiplicative equation* : A change in X causes Y to change by a percentage *proportional* to the change in X
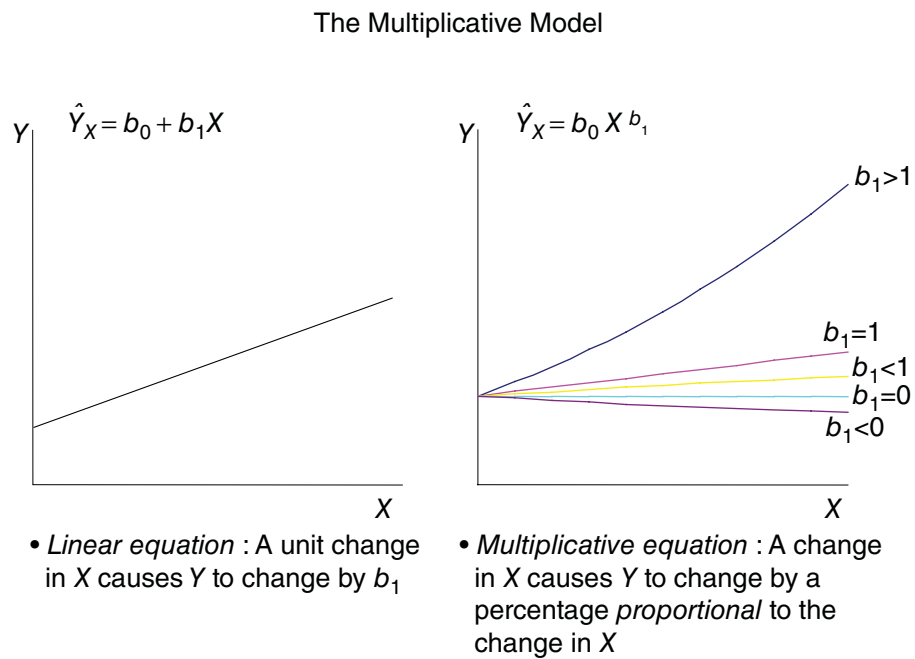
FIGURE 9.2 **Linear Model vs. a Multiplicative/Non-Linear Model.**

However, in order to produce a cost estimating relationship using the LSBF regression method previously discussed in Chapter 7, we must transform the multiplicative model into a linear model (at least temporarily). In doing so, the solution creates a log-linear equation:

$$\widehat{Y} = AX^{b_1} \iff \ln(\widehat{Y}) = b_0 + b_1 \ln(X) \tag{9.1}$$

The left side of Equation 9.1, $\widehat{Y} = AX^{b_1}$, is the original multiplicative power model. The right side of Equation 9.1, $\ln(\widehat{Y}) = b_0 + b_1 \ln(X)$, is the result after applying the natural log(ln) transformation. The equation now "appears" linear with an intercept and a slope, so we can perform a linear regression on $\ln(Y)$ vs. $\ln(X)$. After the regression, however, the result is in "ln dollars," and, therefore, it must be transformed back into its original (i.e., "dollar") units. This reverse transformation is accomplished by using the inverse of the ln function, namely, the exponential function.

Let's demonstrate this with an example.

9.4 Data Transformation

Example 9.1 Consider the following data set found in Table 9.1. It is representative of a production process with Y being the number of hours it takes to produce the specified item number, X:

If you plot the aforementioned data of Y vs. X (or, hours vs. unit number) to see what the data looks like, you get the following scatter plot in Figure 9.3:

TABLE 9.1 Data Set for Example 9.1

(**Y**) Hours	(**X**) Unit Number
60	5
45	12
32	35
25	75
21	125
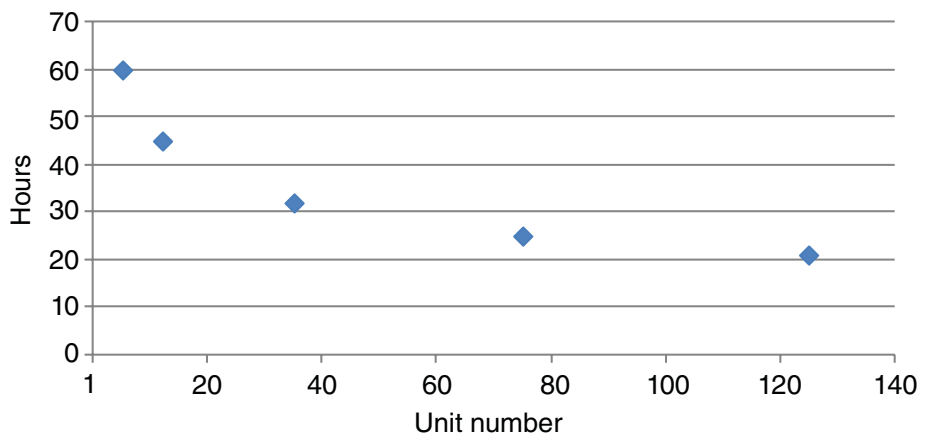
FIGURE 9.3 **Scatter Plot of the Original Data Set in Example 9.1.**

From Figure 9.3, you can observe that the data is non-linear. So, desiring to make the data appear more linear, let's apply the natural log(ln) function on the raw data. The calculations of this transformation are shown in Table 9.2.

TABLE 9.2 Raw Data Transformations in Example 9.1

Hours	**Unit Number**	**ln (Hours)**	**ln (Unit #)**
60	5	4.09434	1.60944
45	12	3.80666	2.48491
32	35	3.46574	3.55535
25	75	3.21888	4.31749
21	125	3.04452	4.82831

Now that we have transformed the data, let's plot the transformed data and see what it looks like. That result is found in Figure 9.4.

In Figure 9.4, we can clearly observe that the data is significantly more linear after the natural log transformation. Now that the data is in this form, we can perform a regression to calculate the slope and intercept of this line. The results of that regression are found in Table 9.3.
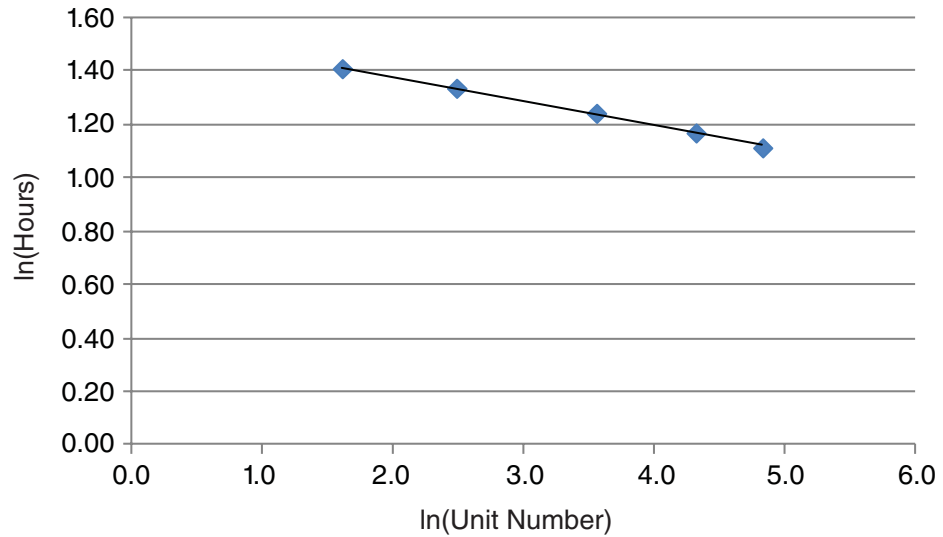
FIGURE 9.4 Scatter Plot of Transformed Data for Example 9.1 (Note How Much More Linear the Plot is).

TABLE 9.3 Regression Results of the Transformed Data from Example 9.1

		ln(Hours) vs. ln(Unit Number)			
Regression Statistics					
Multiple R	0.99996				
R Square	0.99992				
Adjusted R Square	0.99989				
Standard Error	0.00441				
Observations	5				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.73151	0.73151	37688.5729	3.0138E–07
Residual	3	5.82277E–05	1.94092E–05		
Total	4	0.73156			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	4.61650	0.00595	775.5429	4.72771E–09	4.597561122
ln(Unit Number)	–0.32463	0.00167	–194.1354	3.0138E–07	–0.329955405

From the regression printout, and looking at the coefficients for the intercept and ln (unit number), we get the following result:

$$\widehat{Y} = 4.6165 - 0.32463 \times X$$

But this regression result is in ln units, and this brings home the purpose of this chapter! Since we are in ln units, what we *actually* have is:

$$\ln(\widehat{Y}) = 4.6165 - 0.32463 \times \ln(X)$$

Certainly, we cannot leave the results in natural log units, because we operate in our daily lives using hours or dollars, not natural log hours or natural log dollars! So how do we convert this equation back to its original (or "regular") units?

Mathematically, the exponential function is the inverse of the natural log function. Therefore, if we take the exponential of each side of the equation, we will convert this regression that is in natural log units back into regular units (in this example, hours).

To accomplish this and convey the process, we will break this equation into three parts:

Part A	=	Part B	−	Part C
$\ln(\widehat{Y})$	=	4.6165	−	$0.32463 \times \ln(X)$

- Part A = $\ln(\widehat{Y})$
- Part B = 4.6165 (the intercept)
- Part C = $-0.32463 \times \ln(X)$ (the slope times $\ln(X)$)

Taking the exponential of both sides of the equation gets the ln units back into its original units, with one conversion (of the intercept A) remaining. Analyzing each part separately:

- Part A = Taking the exponential of $\ln(\widehat{Y})$ just leaves us with $\widehat{Y}$, since the exponential is the inverse of the natural log, and essentially "undoes" the ln. You could equate this loosely to taking the square root of X^2, leaving you with just X. "Taking the exponential of the ln" acccomplishes essentially the same thing.
- Part B = Taking the exponential of 4.6165 is just a mathematical calculation = $\exp(4.6165)$ = 101.1399.
- Part C = taking the exponential of $-0.32463 \times \ln(X)$ leaves you with $X^{-0.32463}$.

The calculations in Parts A and B are relatively straightforward. For Part C, though, the easiest way to calculate this is by understanding a property of the natural log. One property of the natural log is:

- Property #1: $\ln X^2 = 2 \times \ln X$, or equivalently, $2 \times \ln X = \ln X^2$

You can read Property #1 from left to right, or from right to left, because equality is a symmetric relation. Note that in reference to the $\ln X^2$, the "2" is the exponent for X. But equivalently, it can also become the slope in front of the ln, as seen that it is equal to $2 \times \ln X$. Conversely, the slope in front of the ln, $2 \times \ln X$, can also be written as the exponent for X, becoming equal to $\ln X^2$. So in Part C, we have:

- $-0.32463 \times \ln(X)$, and by Property #1, this is also equal to $\ln X^{-0.32463}$, by making the slope the exponent of X. Thus when we take the exponential of $\ln X^{-0.32463}$, we are merely left with the result of $X^{-0.32463}$, which is similar to the Part A transformation as well.

Looking at the three parts together as an entire equation, we get the following result after the exponential transformation:

$$\ln(\widehat{Y}) = 4.6165 - 0.32463 \times \ln(X)$$

becomes

$$\hat{Y} = 101.1399 \times X^{-0.32463}$$

Note that the final equation once again takes on the form of the Power Model, $y = a \times x^b$. This exponential transformation will be used numerous times in the Learning Curves portion of cost estimating, starting in the next chapter.

9.5 Interpreting the Regression Results

When looking at the regression results in Table 9.3, it is imperative to know that the statistics of the transformed data can be misleading. Most importantly, the Standard Error and the R^2 reported for a log-linear model cannot be compared to those for a linear model.

$$\text{SE}_{\text{unit}} = \sqrt{\frac{\text{SSE}}{n - k - 1}} = \sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n - k - 1}} = \$XXX$$

This is because both are functions of SSE. Recall that SSE is the error sum of squares for this regression which are in natural log units, and the standard error is expressed in terms of dollars. They are in different units!

The Standard Error in log space has a different meaning than that in unit space. It is expressed in a percentage, so an SE of 0.55 means it is 55% of the input value.

$$\text{SE}_{\text{log}} = \sqrt{\frac{\text{SSE}}{n - k - 1}} = \sqrt{\frac{\sum (\ln(Y_i) - \ln(\hat{Y}))^2}{n - k - 1}} = XXX$$

However, while we cannot compare log-linear statistics to linear regression statistics, we *can* compare log-linear statistics results between log-linear models.

For the reader who wants more on this topic, the information is available in numerous statistical textbooks, such as Devore, Jay L. Probability and Statistics for Engineering and the Sciences, Sixth Edition, Thomson Brooks/Cole, 2004.

Summary

In this chapter, we discussed "intrinsically linear regression," which simply is a regression that contains transformed data. When your original data set is not linear, in many cases you may utilize transformations to make your data set appear linear, at least for the purpose of subsequent analysis. While there are many methods of transformations including logs, square roots, and inverses, we concentrated on transforming data using the natural log(ln).

After providing an example data set that was non-linear, we applied a natural log transformation, which displayed a scatter plot that was now linear. We then performed a LSBF regression on this transformed data to establish the intercept and slope of that line. However, the equation that we derived from that regression was in natural log units, in this case "ln(hours)." We then transformed the regression equation back to its original units (hours) using the exponential function. Upon completion, we have now calculated the non-linear regression equation that best describes the original data set. This exponential transformation will be used numerous times in the following two chapters, Chapters 10 and 11, as we take on the important topic of learning curves.

Reference

1. Devore, Jay L. *Probability and Statistics for Engineering and the Sciences*, Sixth Edition, Chapter 13.

Applications and Questions:

9.1 The term "Intrinsically linear regression" is used when your data set is not linear, but some transformation may be applied in order to make the data appear/become linear. (True/False)

9.2 Consider the following data set in a production process. Note that cost is in FY13$:

Unit Number	(FY13$) Unit Cost
1	5,000
10	4,000
50	3,100
75	2,900
100	2,750

Plot this data set to determine if it is linear or non-linear data.

9.3 Transform the data to make it linear by using the natural log.

9.4 Make a scatter plot of the transformed data set.

9.5 Fit a trend line or perform a regression on the transformed data set. What is the equation of the line that best fits this data?

9.6 Since this equation of the line is in natural log dollars, use the exponential function to convert the equation back into dollar units. What is the resultant equation?

Learning Curves: Unit Theory

10.1 Introduction

Learning curve analysis is developed as a tool to estimate the recurring costs in an assembly or production process. Recall that *recurring* costs are those costs that are incurred on each unit of production. The dominant factor in learning theory is direct labor. It is based on the common observation that as a task is accomplished several times, it can be completed in shorter periods of time. Thus, each time you perform a task, you become better at it and accomplish the task faster than the previous time, due to an inherent improvement in efficiency.

There are two predominant theories on learning curves: Unit Theory and Cumulative Average Theory. Both theories suggest that as we repeat a task, we get better at it, and in manufacturing, it has been found that we get better at a constant rate. This chapter will introduce the concept of learning curves and then discuss Unit Theory principles in great detail. Numerous equations will be introduced and five examples will guide you through the necessary steps to calculate the decreasing recurring costs. The method of using transformed data in a regression that we discussed in Chapter 9 will be continued and used extensively in this chapter and in Chapter 11 on Cumulative Average Theory. Note that the "cost" of a unit can be expressed in dollars, labor hours, or other units of measurement.

10.2 Learning Curve, Scenario #1

Let's imagine that you live in a place where it rains often. Due to the rain, you decide that you want to purchase an automatic garage door opener, so that you can avoid getting wet when you get home and out of your car. After installation, you will now be able to drive straight into the garage without getting out of your car to open the door first. So let the task begin! You head down to the local hardware store or your Home Depot and search for the aisle where they sell automatic garage door openers. You eventually find the correct aisle and then look through the myriad products that are available, reviewing price and quality. You finally decide on the model that best fits your needs and budget.

Once you get home, you open the box and orient yourself with what is inside. The instructions say to "Take Nut A and combine it with Bolt B, and then slide onto Bracket C." You search for these parts and finally figure out how they fit together. But when you try to hang the main operating hardware unit in your garage, you realize that you need the "special tool" to do so and you have to go back to the hardware store for it. You come home and finally get the garage door opener installed, and after 20 hours of effort, you finally stand in your driveway proudly pushing the button and watching the door go up and down. Success! While you are doing this, however, your neighbor comes over and says that he has been thinking of installing one of them and would you please help him install his as well? You agree to do so and once again the task begins. But this time, as you go back to the hardware store, you know exactly what aisle the box is on, exactly which product that you want, and you find it with only minor delay. When you get home and begin putting it together, though, you still are not very good with "Taking Nut A and combining it with Bolt B, and then sliding onto Bracket C." But you do not have to go back to the store for the "special tool" this time, and you are a bit more familiar with the installation procedures, so by the time you complete the task, this time it only took you 16 hours. Once again, however, as you are admiring that the garage door is working, *another* neighbor emerges and says that they want one as well. As the day and week goes on, word spreads quickly, and you end up agreeing to install garage door units for the entire neighborhood and weeks later, by the time you complete the fifteenth one, you are now taking only 6 hours to complete the task. So let's contrast that to when you installed your first one, which took 20 hours. It is evident that you have gotten much better at the task. How much better? You have gotten 14 hours better, from 20 hours for the installation of the first unit down to 6 hours for the fifteenth unit $= 14$ hours of improvement – or you have "learned" 14 hours. You can never get to the point where you improve so much that it takes *zero* hours to accomplish the task, because it will always take you some finite amount of time to complete the work; plus, as you get better, your improvement occurs in smaller and smaller increments. On a graph, this would be when the slope of your learning line becomes "asymptotic," which is where your curve becomes significantly "flatter" or more horizontal.

Figure 10.1 is a graph of the learning curve from this example. Note that in this scenario, the *Y*-axis is in hours, and the *X*-axis is the quantity of garage doors installed.
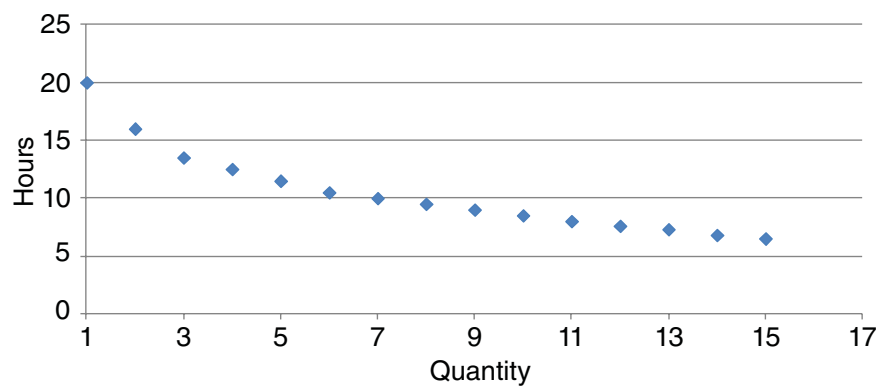
FIGURE 10.1 **Garage Door Installation Learning Curve.**

Of course, there are other possible factors to make us faster besides production workers learning their tasks better, such as:

- Redesign of a product for lower cost production
- An improved production facility

- Management learning
- Better layout/better efficiencies
- Engineering and production improvements
- Lower cost suppliers, and
- Better "make-buy" decisions

But in general, as we become more efficient and competent, we can perform the task faster.

As previously mentioned, there are two theories of learning curves: the first is Cumulative Average Theory and the second is Unit Theory. While this chapter will concentrate on Unit Theory to demonstrate the concept, we will briefly introduce Cumulative Average Theory, because historically it was the first learning theory developed.

10.3 Cumulative Average Theory Overview

"If there is learning in the production process, the cumulative average cost of some doubled unit equals the cumulative average cost of the un-doubled unit times the slope of the learning curve." This theory was first discovered by T. P. Wright back in 1936, and it was based on examination of World War I aircraft production costs [1]. Aircraft companies and the Department of Defense were interested in the regular and predictable nature of the reduction in production costs that Wright observed. It implied that a fixed amount of labor and facilities would produce greater and greater quantities in successive periods. We will discuss Cumulative Average Theory in greater detail in Chapter 11. But first, we will concentrate on Unit Theory.

10.4 Unit Theory Overview

"If there is learning in the production process, the cost of some doubled unit equals the cost of the un-doubled unit times the slope of the learning curve." This theory is credited to J. R. Crawford back in 1947 [1]. He led a study of World War II airframe production costs that was commissioned by the US Air Force to validate the previous learning curve theory. To describe the "doubled" vs. the "un-doubled" unit in words, if there is learning in the production process, the cost of the, say, 200th unit (the "doubled" unit) is equal to the cost of the 100th unit (the "un-doubled" unit) times the slope. The basic concept of unit theory is that as the quantity of units produced doubles, the "cost" of producing a unit is decreased by a constant percentage.

To illustrate this concept, for an 80% learning curve, there is a 20% decrease in unit cost each time that the number of units produced doubles. Thus

- The cost of unit 2 is 80% of the cost of unit 1
- The cost of unit 4 is 80% of the cost of unit 2
- The cost of unit 8 is 80% of the cost of unit 4
- The cost of unit 50 is 80% of the cost of unit 25
- The cost of unit 100 is 80% of the cost of unit 50, etc.
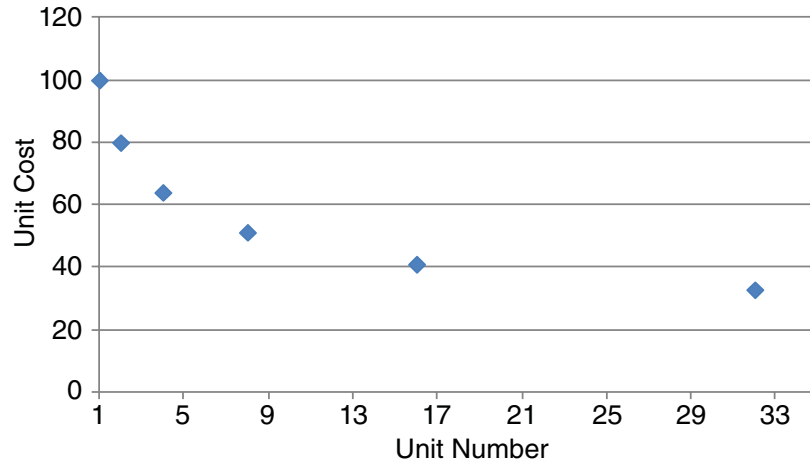
FIGURE 10.2 **80% Unit Theory Learning Curve.**

Figure 10.2 above is an example of an 80% learning curve. Note that in this case, cost is in dollars instead of hours. If cost is in hours, those hours will eventually need to be converted to dollars. We will learn how to make that conversion in Chapter 13 on Wrap Rates.

One note of interest is that an 80% learning curve is "better" than a 90% learning curve. In our regular lives, we are used to 90% being "better" than 80% in most things; for example, if you took an exam, you would rather get a 90% than an 80% for your exam grade. However, in learning curves, the opposite is true. Why is this?

We just learned that for an 80% learning curve, there is a 20% decrease in cost each time that the number of units produced is doubled. But in a 90% learning curve, there is only a 10% decrease in cost each time that the number of units produced is doubled. In a 95% learning curve, there is only a 5% decrease each time the number of units produced is doubled. Therefore, a 90% learning curve is better than a 95% learning curve; an 80% learning curve is better than a 90% learning curve, and so on. But as a rule, rarely do learning curves exceed 70% in any production process.

Unit theory is defined by the following equation:

$$Y_x = A \times x^b \tag{10.1}$$

where

Y_x =   the cost of unit X. This is the cost of the unit you are seeking.
A =   the theoretical cost of unit one (also known as $T1$)
X =   the unit number
b =   a constant representing the slope of the learning curve (where slope $= 2^b$)

Note that A is the *theoretical* (or mathematically calculated) cost of unit one, and not necessarily the *actual* cost of the first unit.

Let's now discuss the learning parameter, b. In practice, $-0.5 \leq b \leq -0.05$. Mathematically, since the slope of the learning curve $= 2^b$, the first number corresponds roughly with learning curve slopes just over 70%, since $2^b = 2^{-0.5} = 0.707$. This is a learning curve in a highly *manual* operation. The learning curve in highly *automated* industries is much flatter, corresponding more to the second number, $b = -0.05$. This is approximately 96%, since $2^b = 2^{-0.05} = 0.966$. The learning parameter is largely determined by the type of industry and the degree of automation in that industry. The more automated the industry, the less learning that usually occurs.

Note that for $b = 0$, the equation $Y_x = A \times x^b$ simplifies to $Y = A$, which means that any unit on the learning curve costs the same as the first unit. In this case, the learning curve is a horizontal line and there is no learning. This is not good in the business world! This is also referred to as a 100% learning curve. This is one of the few times where getting 100% is not a good thing!

Where does slope $= 2^b$ come from? We know that as the number of units produced doubles, the unit cost is reduced by a constant percentage, and this is referred to as the slope of the learning curve.

$$\text{Cost of unit } 2n = (\text{Cost of unit } n) \times (\text{Slope of learning curve})$$

Isolating for slope, you find

$$\text{Slope of learning curve} = \frac{\text{Cost of unit } 2n}{\text{Cost of unit } n} = \frac{A \times (2n)^b}{A \times (n)^b} = 2^b \qquad (10.2)$$

Taking the natural log of both sides, $\ln(\text{slope}) = \ln(2^b) = b \times \ln(2)$, and thus

$$b = \frac{\ln(\text{slope})}{\ln(2)} \qquad (10.3)$$

For a typical 80% learning curve: $b = \ln(\text{slope})/\ln(2)$, and thus $b = \ln(0.8)/\ln(2) = -0.3219$. This number "$b$" is the slope parameter for the learning curve and is the slope of your regression line. In addition, note that when using this equation $b = \ln(\text{slope})/\ln(2)$, that a slope of 80% is inserted into the equation as $\ln(0.80)/\ln(2)$, and not $\ln(80)/\ln(2)$.

General guidelines for slopes [1]:

- If an operation is 75% manual and 25% automated, slopes are generally in the 80% vicinity.
- If an operation is 50% manual and 50% automated, slopes are generally about 85%.
- If an operation is 25% manual and 75% automated, slopes are generally about 90%.
- Shipbuilding slopes are generally in the 80–85% range.

The average slope for the aircraft industry is about 85%. But departments within an organization can vary greatly from that. Assuming repetitive operations within an industry, typical slopes may include:

- Electrical: 75–85%
- Electronics: 90–95%
- Machining: 90–95%
- Welding: 88–92%

To use a learning curve for a cost estimate, a slope and first unit cost are required. The slope may be derived from analogous production situations, industry averages, historical slopes for the same production site, or historical data from previous production quantities. First unit costs may be derived from engineering estimates, Cost Estimating Relationships (CER's), or historical data from previous production quantities.

When historical production data is available, slope and first unit cost can be calculated by using the learning curve equation. But since $Y_x = A \times x^b$ is a "curved line" (see Figure 10.1 or 10.2), we must attempt to make it linear in order to make the expression amenable to LSBF regression, and we do so by using a natural log (ln) transformation. This procedure was discussed in the previous chapter on "Intrinsically Linear Regression."

Taking the natural log of both sides of $Y_x = A \times x^b$ yields $\ln(Y_x) = \ln(A) + b \times \ln(x)$. If we rewrite this equation as $Y' = A' + bX'$ (which now looks like a linear equation), we can solve for A' and b using simple linear regression.

Figure 10.3 is a graph of the garage door example data, after using a natural log transformation. Figure 10.1 was the graphing of the original data; Figure 10.3 is the graphing of the *transformed* data using the natural log. Note how the nonlinear data has been transformed into data that is much more linear. It is now possible to run a

$$y = -0.4098x + 3.0599$$
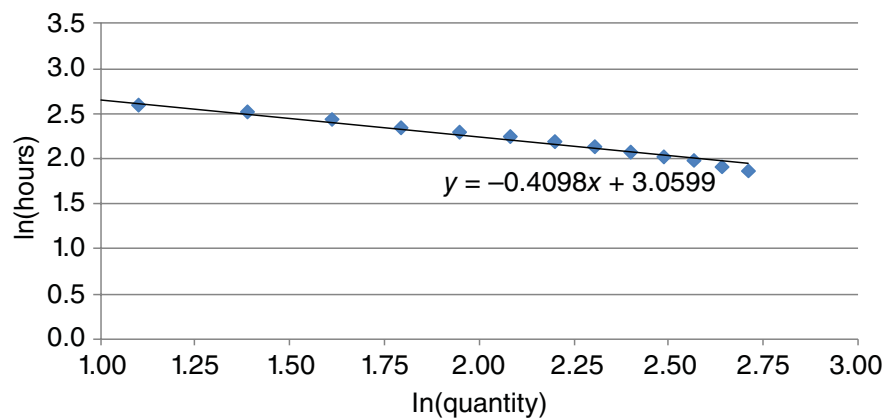
FIGURE 10.3 Garage Door Example: Transformed Data.

simple linear regression on this data to find the slope and intercept of this line and production process. The equation of the line is shown on the graph as $\widehat{Y} = 3.0599 - 0.4098X$. This equation can be derived by performing a regression on the original data points, or it can also be found in Excel by "right clicking" on any data point in the scatter plot/graph, selecting "Add Trendline," and then selecting "Display Equation on Chart." This will produce a line and equation such as the one seen in Figure 10.3, mentioned subsequently. A regression equation and a trend line are identical in a single variable regression. Note also that the slope of this trend line is negative, which is as expected. The negative slope shows that there is learning in the process, as costs get less expensive as you produce more quantities.

10.5 Unit Theory

Example 10.1 Let's attempt a unit theory learning curve example. Given the same historical data found in Table 10.1, find the Unit Theory learning curve equation that describes this production environment. Use this equation to predict the cost (in hours) of the 150th unit and find the slope of the curve.

TABLE 10.1   Data Set for Example 10.1

Hours	Unit #
60	5
45	12
32	35
25	75
21	125
??	150

If you plot the number of hours (y-axis) vs. the unit number (x-axis), you will see that you indeed have a curve that can be modeled with $Y_x = A \times x^b$, as shown in Figure 10.4.
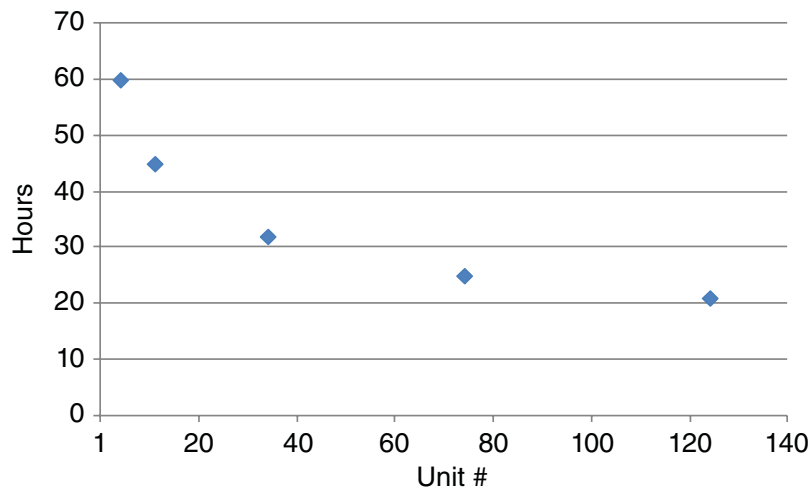
FIGURE 10.4 Unit Theory Learning Curve Graph, Example 10.1.

So to transform the scatter plot in Figure 10.4 and "straighten the curve," we will take the natural logs (ln) of each of the data points. This produces the following data chart found in Table 10.2.

TABLE 10.2   Transformed Data in Natural Log Units, Example 10.1

Hours	Unit #	ln (Hours)	ln (Unit #)
60	5	4.09434	1.60944
45	12	3.80666	2.48491
32	35	3.46574	3.55535
25	75	3.21888	4.31749
21	125	3.04452	4.82831

Plotting ln(hours) vs. ln(unit #), we find the following in Figure 10.5. Note that the graph line has become more linear and that the equation of that line contains a negative slope, showing learning in the process.
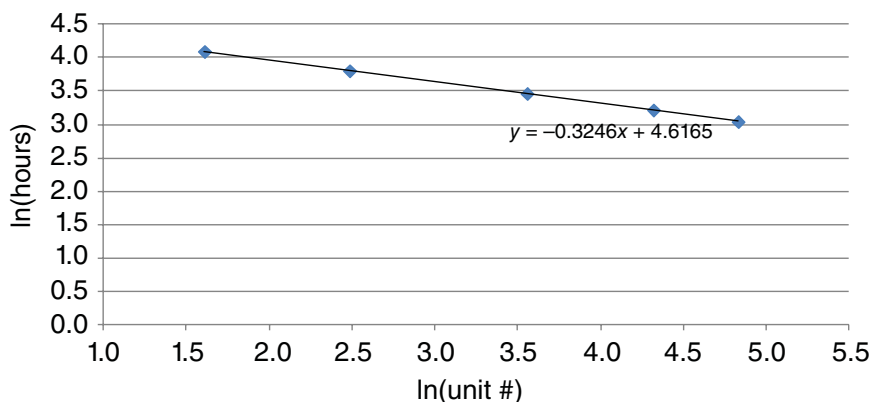
FIGURE 10.5 Graph of Transformed Data, Example 10.1.

The equation shown in Figure 10.5 is found by simply "adding a trend line" in Excel. This equation can also be found by taking a regression of this data set from Table 10.2. Doing so reveals the following result in Table 10.3:

TABLE 10.3   Regression Output for Example 10.1

SUMMARY OUTPUT					
	ln (Hours) vs. ln(Unit #)				
Regression Statistics					
Multiple R	0.99996				
R Square	0.99992				
Adjusted R Square	0.99989				
Standard Error	0.00441				
Observations	5				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.73151	0.73151	37688.5729	3.0138E–07
Residual	3	0.00006	0.00002		
Total	4	0.73156			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	4.61650	0.00595	775.542914	4.728E–09	4.59756
ln (Unit #)	–0.32463	0.00167	–194.13545	3.014E–07	–0.32996

Looking at the regression results, at first glance we observe the following equation:

$$\widehat{Y} = 4.6165 - 0.32463 \times X$$

(or in words) : Hours $= 4.6165 - 0.32463 \times$ Unit #

But actually, since the regression is in natural log units, we instead have:

$$\ln(\text{Hours}) = 4.6165 - 0.32463 \times \ln(\text{Unit \#})$$

Therefore, since we are in natural log units, we must transform the data back to "regular" or "original" units, and we do this by using the exponential function, since the exponential function is the inverse of the natural logarithmic function. Taking the exponential of each side reveals:

$$\widehat{Y} = \exp(4.6165) \times X^{-0.32463}$$

Since $A = e^{4.6165} = 101.139$ and $b = -0.32463 = $ slope coefficient, then the equation that describes this data (back in $Y = A^{*}X^{b}$ format) can be written:

$$Y_x = 101.139(X)^{-0.32463}$$

In words, this says that "The cost of the Xth unit equals 101.139 hours times the unit number we seek raised to the -0.32463 power." Note again that the slope parameter "b" is negative, which is a good thing since it denotes learning is occurring in the process. From this equation, $T1 = A = 101.139$ hours and the slope of the learning curve $= 2^{b} = 2^{-0.32463} = 79.85\%$.

Solving for the cost in hours of the 150th unit, and using $Y_x = 101.139(X)^{-0.32463}$, we find that

$$Y_{150} = 101.139(150)^{-0.32463} = 19.883 \text{ hours}$$

Thus, the cost of the 150th unit in this learning curve is projected to be (or will take, in hours) $Y_{150} = 19.883$ hours.

Therefore, the conclusions to Example 10.1 are the following:

- Time to complete the first unit: 101.139 hours (aka, A or $T1$)
- Time to complete the 150th unit: 19.883 hours

So we have "learned," or gotten "better" by, $101.139 - 19.883 = 81.25$ hours. In this production process, we have gotten better by over 81 hours while producing Units #1 – #150.

10.6 Estimating Lot Costs

After finding the learning curve equation which best models the production situation, the cost estimator will now use this equation to estimate the cost of future units. But rarely, as in the aforementioned example, are we asked to estimate the cost of a particular unit. Rather, we usually need to estimate the cost of a number of units, grouped in lots. Why is this?

In the execution of DoD weapon systems acquisitions, services do not purchase just one single item. For example, let's say that the Army, the Marine Corps, and the United Kingdom all would like to purchase High-Mobility Multipurpose Wheeled Vehicles (HMMWVs or "Humvees"). After negotiations and surveying present inventory, it is determined that a total of 2,000 HMMWVs will be needed. The US Army has decided that it needs 900 HMMWVs; the US Marine Corps needs 800; and the United Kingdom needs 300. Accordingly, based on budgets and priorities, the "buy" schedule might look like the following:

- The US Army would get the first 500 Humvees;
- The US Marine Corps would get the next 500;
- The US Army would get the next 400;
- The United Kingdom would get the next 300;
- The US Marine Corps would get the final 300 Humvees.

In this scenario, since the US Army is getting a total of 500 Humvees in the first lot, there would be no concern by the Army with what the cost of the 35th Humvee or the 73rd Humvee is, or even the cost of the 125th one. Instead, the Army would be concerned with the total cost of the first lot of 500 Humvees, because that is how they would "pay the bill," and how the vehicles would be delivered. Lots are generally referred to as Lot #1 then Lot #2; or Lot A then Lot B; or they can also be referred to in Blocks, as well. Note that given the above *military* scenario, you could change the scenario to *civilian industrial* issues, such as cars being made in Detroit, Michigan, or in Tokyo, Japan, with different destinations and buyers upon completion. The cars might be made in blocks of 500 or 5,000 and sent to designated locations or countries from that block. The same learning curve scenario would be valid.

So how would we calculate the cost of these first 500 Humvees for the US Army? This cost is calculated using a cumulative total cost equation and is found in Equation 10.4:

$$CT_N = A(1)^b + A(2)^b + \cdots + A(N)^b = A \left[\sum_{x=1}^{N} x^b \right] \tag{10.4}$$

where CT_N = the cumulative total cost of N units.

Note that since the cost of a single unit is $Y_x = A \times X^b$, the total cost here is merely a finite series summation of all of the single unit costs; in our Humvee example, $N = 500$ units that are summed.

This total cost, CT_N, may be approximated using the following equation:

$$CT_N \cong \frac{A(N)^{b+1}}{b+1} \tag{10.5}$$

The difference between Equation 10.4 and 10.5 is that Equation 10.4 is the *exact* cost (easily summed and calculated in Excel), while Equation 10.5 is an approximation of Equation 10.4. Observe that the approximation in Equation 10.5 is derived by taking the definite integral of the sum of the X^b's, from 1 to N, before passing to the limit. If we were using calculus, this is merely computing the *total area under the curve* from unit #1 to unit #500 when taking the integral of Equation 10.4.

Example 10.2 Recall the data set from Example 10.1:

Hours	Unit #
60	5
45	12
32	35
25	75
21	125

Instead of computing the cost of the 150th unit as we did in Example 10.1, let's now compute the cost (in hours) of the *first 150 units collectively*. We could also say that the first lot consists of 150 units. Since the equation that describes this data was found to be $Y_x = 101.139(x)^{-0.32463}$, we know that $A = 101.139$ hours; that $b = -0.32463$; so then $b + 1 = (-0.32463) + 1 = 0.67537$, and $N = 150$. Thus, the total cost for the first 150 units, using Equation 10.5, is found to be

$$\text{CT}_N = \frac{101.139 \times 150^{0.67537}}{0.67537} = 4,416.15 \text{ hours}$$

This answer of 4,416.15 hours was found using the approximation method, which is easy and convenient for "back of the envelope" calculations. If you wanted to be more precise and find the *exact* number of estimated hours, you could easily do so in Excel, using the equation $Y_x = A \times x^b$ (the cost for a particular unit), solving for Y_1, Y_2, and up to Y_{150}, and then adding all 150 costs together for one sum. This method would yield a final answer of 4,329.527 hours. The difference between the two methods in this case is approximately 2%.

Note again that this equation has calculated the cost for the first lot of 150 units. But how would we then calculate the cost of the second, third, or any later lot? To compute the exact total cost of a specific lot with first unit # F and last unit # L, Equation 10.6 is used:

$$\text{CT}_{F,L} = A \left[\sum_{x=1}^{L} x^b - \sum_{x=1}^{F-1} x^b \right] \tag{10.6}$$

Equation 10.6 is approximated using Equation 10.7.

$$\text{CT}_{F,L} \cong \frac{AL^{b+1}}{b+1} - \frac{A(F-1)^{b+1}}{b+1} \tag{10.7}$$

You will need four parameters/inputs to use either Equation 10.6 or 10.7. They are: A, b, F, and L, where:

- $A = T1 = $ Theoretical first unit cost (determined by the regression)
- $b = $ slope parameter (determined by the regression)
- $F = $ First Unit of the desired lot
- $L = $ Last Unit of the desired lot

Note what Equations 10.6 and 10.7 are actually accomplishing. In the first half of the equation, you are first calculating the cost of all of the units (1 through L), and then subtracting out the costs of the units that you do not care about in this calculation (the "$F - 1$" part). Therefore, the units that remain are units $F - L$. Note also that Equation 10.7 is merely the integral of Equation 10.6, as again we are merely measuring the area under the curve from units $F - L$. We will illustrate this point with an example.

Example 10.3 Compute the cost (in hours) of the lot containing units 26–75 from the previous example. What we are trying to solve for is shown graphically in Figure 10.6.

Since we are not calculating the *first* lot cost (which would be Lot #1 = Units 1 through 25), we can no longer use Equation 10.4 or 10.5, and instead must use the lot cost Equation 10.6 or 10.7. For ease of discussion and calculation, we shall use
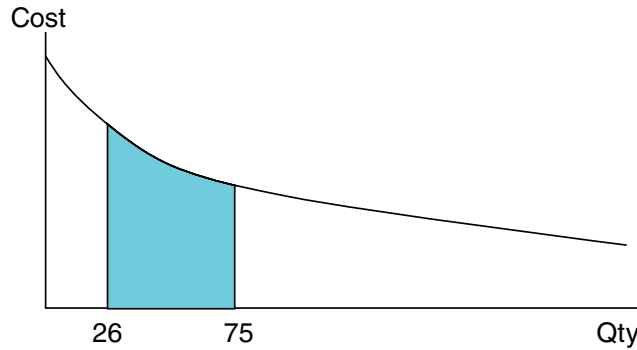
FIGURE 10.6 Learning Curve Depiction for Example 10.3.

Equation 10.7. Calculations for the lot cost of Units 26 to 75, using the values of A and b from Example 10.2, and $F = 26$ and $L = 75$, are shown here:

$$\mathrm{CT}_{F,L} \cong \frac{AL^{b+1}}{b+1} - \frac{A(F-1)^{b+1}}{b+1}$$

$$A = 101.3$$

$$F = 26$$

$$L = 75$$

$$b = -0.325$$

$$\mathrm{CT}_{26,75} \cong \frac{(101.3)(75)^{-0.325+1}}{-0.325+1} - \frac{(101.3)(26-1)^{-0.325+1}}{-0.325+1} = 1448.8 \; hrs$$

In addition, note that this equation is calculating the total cost of the first 75 units and then subtracting out the costs of the first 25 units, so what remains are Units 26–75. The reason that calculations must be performed this way is because of the "upper curve" in the learning curve. (note the curvature "up" to the left of Unit 26 in Figure 10.6, in Lot 1 from Units 1–25). Though we are indeed still just measuring the area under the curve, we cannot use a typical "length × width" or "length × height" equation to calculate that area, due to the curvature in the learning curve. Thus, in Example 10.3, the total cost in hours of the lot from Units 26 to 75 is calculated to be 1,448.8 hours.

10.7 Fitting a Curve Using Lot Data

Cost data is rarely, if ever, reported by the unit number of an item and its corresponding cost. Rather, the cost data is generally reported by the manufacturer for production lots by the total lot costs and the number of units in that lot, similar to the following in Table 10.4.

We must somehow fit a curve to this lot data, to solve for the learning curve in this production environment! Previously, in Example 10.1 we graphed this data using Unit Number as our X-axis, and Unit Cost (in hours) as our Y-axis. But that is not possible with lot data because we now have the number of units in a lot and the total cost for that lot. Therefore, lot data must be adjusted since learning curve calculations require a unit number and its associated unit cost. In essence, this means that we need an X and a Y value for each lot to be able to graph it and to fit a learning curve through these points.

TABLE 10.4   Typical Display of Lot Data Information

	Number of Units	Direct Labor Lot Costs
Lot 1	50	$10M
Lot 2	50	$8M
Lot 3	100	$14M
Lot 4	50	$6M

To accomplish this, the Unit Number and Unit Cost for a lot are represented by the algebraic lot midpoint (LMP) and average unit cost (AUC). For your calculations, the following now apply:

- Lot midpoint (LMP) = *X*-axis
- Average unit cost (AUC) = *Y*-axis

We will discuss how to derive each of these parameters for each lot.

10.7.1 LOT MIDPOINT

By definition, the Algebraic LMP is the theoretical unit whose cost is equal to the average unit cost for that lot on the learning curve. But what does that actually mean?! Let's consider a lot containing 50 units. The first unit will have a certain cost, the second unit will have a slightly lower cost, unit 30 will have an even lower cost, and the lot ends at unit number 50, which will have the lowest cost in that lot. Adding the costs of each of the 50 units and dividing that sum by 50 will yield an average unit cost for that lot. Let's say the AUC is $50,000 exactly. The LMP will thus be at that unit in the lot whose cost was equal to the AUC of $50,000 in that lot. Unit #1 will be much greater than $50,000; Unit #50 will be much lower than $50,000, but all units together will average to $50,000, and one unit will be the closest to that exact cost. The goal is trying to solve for that unit. But that still may not be overly clear to you. Therefore, we offer an alternative explanation, and so for us, we feel that the simplest way to look at LMP is that it is "*that exact unit which will divide the desired lot into two equal areas.*"

In Figure 10.7, consider the shaded area under the curve for the first 75 units.
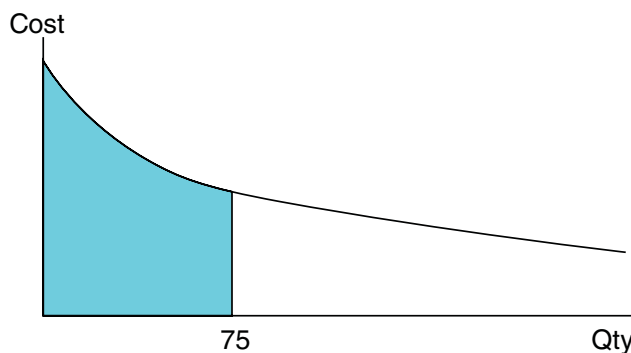
FIGURE 10.7 Learning Curve Scenario for Lot Midpoint.

The numerical midpoint for this lot would be at unit #37.5 (=75/2), assuming that we could work in half units. Let's draw this vertical line, now shown in Figure 10.8.
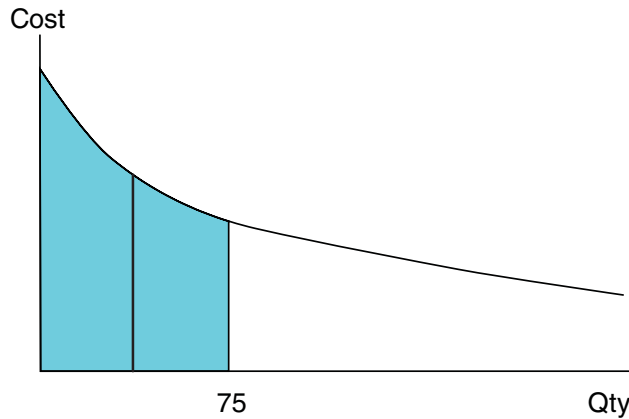
FIGURE 10.8 Learning Curve Scenario for Lot Midpoint.

Observe that at the vertical line drawn at unit #37.5, the two areas to the left and right of 37.5 would *not* be of equal size; clearly, the area to the left of 37.5 would be greater in area than the area to the right of 37.5. Therefore, 37.5 could *not* be the LMP for this scenario. Since the LMP is that unit which would divide the lot into *two equal areas*, the LMP clearly must be further to the left from unit 37.5. Knowing this, how do we calculate this number? How much further to the left is it?

Calculation of the exact LMP is an iterative process. Numerous software packages have developed algorithms that will solve for the LMPs. But if software is not available, estimating the LMPs can easily be completed by using a parameter-free approximation method, found in Equations 10.8 and 10.9 ([2], page 15)

- For the first lot *only* (the lot starting at unit 1):

$$\text{If lot size} < 10, \text{ then LMP} = \text{Lot size}/2 \tag{10.8}$$

$$\text{If lot size} \geq 10, \text{ then LMP} = \text{Lot size}/3$$

- For *all subsequent lots* (lot #2 and *every lot* thereafter):

$$\text{LMP} = \frac{F + L + 2\sqrt{F \times L}}{4} \tag{10.9}$$

Where

$$F = \text{ the first unit number in a lot, and}$$
$$L = \text{ the last unit number in a lot.}$$

Once you solve for the unit number that LMP represents, the LMP becomes the independent variable that you will utilize on your *X*-axis. Note: the aforementioned equations for determining LMP are valid for when you do not know what the learning curve is or do not know what the value is for the slope parameter, *b*. Equations to solve for LMP when the learning curve is known (or estimated) will be covered in Section 10.9.

10.7.2 AVERAGE UNIT COST (AUC)

The dependent variable (Y) to be used is the AUC for each lot, which can be found by Equation 10.10:

$$AUC = \frac{\text{Total lot cost}}{\text{Lot size}} \tag{10.10}$$

This is a much easier process than finding LMP. Once all values for LMP and AUC are calculated, both will have to be transformed logarithmically before we can use them in our regression equations. We will then regress ln(AUC) vs. ln(LMP).

Example 10.4 Let's use the data set provided at the start of Section 10.7 in Table 10.4 and learn how to solve for LMP and AUC. Moreover, the LMP (our independent variable) will be our *X*-axis for graphing and the AUC (our dependent variable) will be the data on our *Y*-axis. Table 10.5 contains the data set, along with the chart that we will need to complete:

TABLE 10.5 Given Data for Example 10.4, and Columns Needed to Solve for Lot Costs Using Unit Theory Principles

Units	Direct Labor Costs ($)	First	Cumulative/ Last	AUC	LMP
50	10,000,000				
50	8,000,000				
100	14,000,000				
50	6,000,000				

For our calculations to complete this chart in Table 10.5, we will first need to solve for the first and last units of the lot, and then solve for LMP and AUC. While L stands for the last unit in each lot, it also represents the cumulative quantity of units produced up to and including that lot, as well. For cumulative quantities, we have the following:

- Lot 1: the first unit F equals 1 and the last unit L equals 50;
- Lot 2: the first unit F equals 51 and the last unit L equals 100;
- Lot 3: the first unit F equals 101 and the last unit L equals 200;
- Lot 4: the first unit F equals 201 and the last unit L equals 250;

While these calculations are simple additions, they are *essential* to get correct, since all calculations using an incorrect F or L will yield an incorrect lot cost. Please be careful! For example, note that Lot 4 is for 50 units and ranges from units 201 to 250. It is not from 201 to 251, even though that would *appear* to be 50 units. You must count the first unit in each lot, so a lot from 201 to 250 is exactly 50 units.

With F's and L's complete, we will now address the LMP calculations. For brevity, LMP1 will refer to the LMP for Lot 1; LMP2 will be the LMP for Lot 2, and so on; correspondingly, AUC1 will be the AUC for Lot 1, AUC2 for Lot 2, etc.

When solving for LMP1, let's first remember Equation 10.8 for our calculation:

$$\text{If lot size} < 10, \text{ then LMP} = \text{Lot size}/2;$$

$$\text{If lot size} \geq 10, \text{ then LMP} = \text{Lot size}/3$$

In our example, since our first lot size is 50 units (which is ≥ 10), we have LMP1 $= 50/3 =$ 16.666.

For all other lots, we use Equation 10.9:

$$\text{LMP} = \frac{F + L + 2\sqrt{F \times L}}{4}$$

where

$$F = \text{the first unit number in a lot, and}$$
$$L = \text{the last unit number in a lot.}$$

- Thus, LMP2 $= \frac{51+100+2\times\sqrt{51\times100}}{4} = 73.457$. In the same manner,
- LMP3 $= \frac{101+200+2\times\sqrt{101\times200}}{4} = 146.313$, and
- LMP4 $= \frac{201+250+2\times\sqrt{201\times250}}{4} = 224.833$.

Author's Note: As a check, please ensure that your LMP falls in the interval between your First and Last! While this sounds very basic and logical, it is easy to make a math mistake on this. Note that LMP2 = 73.457 falls approximately in the middle between 51 and 100, and slightly to the left of the numerical midpoint of 75. If your calculation for LMP2 came out to, say, 253.54 due to a math error, there is no way that can be correct! The answer must fall between 51 and 100, since you are calculating the midpoint for that lot. The same applies for LMP3 and LMP4, as well. For LMP3 = 146.313, this value falls between 101 and 200, and is slightly to the left of the actual midpoint of the lot (approximately 150); and LMP4 = 224.833 is the lot midpoint between 201 and 250.

For the Average Unit Cost (AUC) calculations, we will use Equation 10.10:

$$\text{AUC} = \frac{\text{Total lot cost}}{\text{Lot size}}$$

- AUC1: $10M/50 = $200,000
- AUC2: $8M/50 = $160,000
- AUC3: $14M/100 = $140,000
- AUC4: $6M/50 = $120,000

Thus, the average cost for all 50 units in Lot 1 is $200,000 per unit. AUC decreases to $160,000 per unit in Lot 2; then to $140,000 per unit in Lot 3; and finally to $120,000 per unit in Lot 4. The average unit cost *should* decrease from one lot to another, since we "get better" and units cost less as we make more of them.

Now that we have completed all of our computations, the overall Unit Theory chart should look like Table 10.6 below:

TABLE 10.6 Example 10.4 Completed Calculations, Using Unit Theory Principles

Units	Direct labor Costs ($)	First	Cumulative/ Last	AUC	LMP
50	10,000,000	1	50	200,000	16.6667
50	8,000,000	51	100	160,000	73.4571
100	14,000,000	101	200	140,000	146.3134
50	6,000,000	201	250	120,000	224.8326

A scatter plot of AUC (Y) vs. LMP (X) reveals the following nonlinear curve in Figure 10.9:
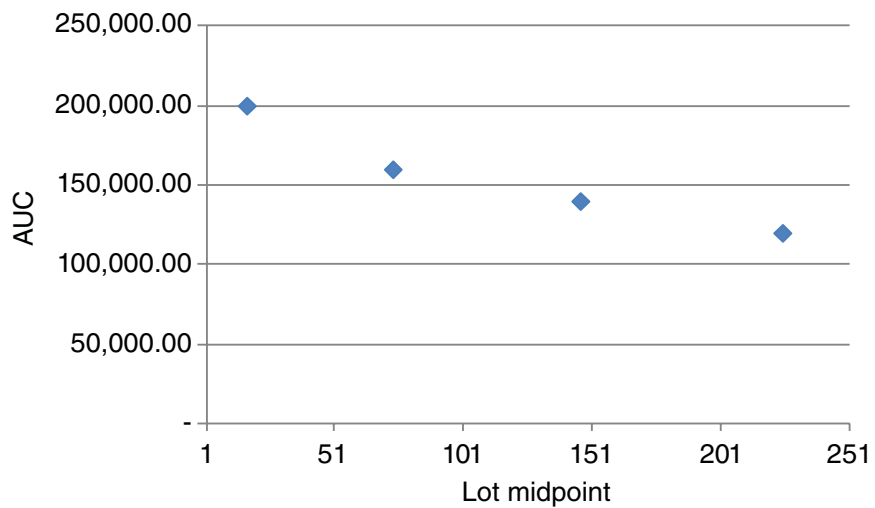
FIGURE 10.9 Scatter Plot of AUC (Y) vs. LMP (X) in Example 10.4.

To make the data more linear in order to perform a linear regression, take the natural logs of both LMP and AUC. Table 10.7 is the completed Unit Theory chart prior to calculating a regression.

TABLE 10.7 Example 10.4 Completed Calculations, With Natural Logs

Units	Direct labor Costs ($)	First	Cumulative/ Last	AUC	LMP	Y ln (AUC)	X ln (LMP)
50	10,000,000	1	50	200,000	16.6667	12.206	2.813
50	8,000,000	51	100	160,000	73.4571	11.983	4.297
100	14,000,000	101	200	140,000	146.3134	11.849	4.986
50	6,000,000	201	250	120,000	224.8326	11.695	5.415

The resultant regression of ln(AUC) vs. ln(LMP) is found in Table 10.8.

So, from the regression it appears that what we have is $\widehat{Y} = 12.75144 - 0.18686 \times X$, where $Y =$ AUC, and $X =$ LMP. But since these results are in natural log units, what we really have is:

$$\ln(AUC) = 12.75143 - 0.18685 \times \ln(LMP)$$

TABLE 10.8   Regression Output for Example 10.4

SUMMARY OUTPUT					
		ln(AUC) vs ln(LMP)			
Regression Statistics					
Multiple R	0.98421				
R Square	0.96867				
Adjusted R Square	0.95301				
Standard Error	0.04693				
Observations	4				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.13617	0.13617	61.83816	0.01579
Residual	2	0.00440	0.00220		
Total	3	0.14058			
	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	12.75144	0.10664	119.57606	0.00007	12.29261
ln(LMP)	−0.18686	0.02376	−7.86372	0.01579	−0.28910

Since this equation is in "ln dollars," we take the exponential of each side to eliminate the natural logs to transform the equation back into units of dollars. While doing so, you get:

$$\text{AUC} = \exp(12.75143) \times \text{LMP}^{-0.18685}, \text{ and so the final equation converts to}$$

$$\text{AUC} = 345,044.9578 \times \text{LMP}^{-0.18685}$$

In this equation, T1 = \$345,044.96, and the slope of the learning curve = 2^b = $2^{-0.18685}$ = 0.8785 = 87.85%. This is the equation that best models the production environments on the data set provided in Example 10.4. You can now solve for the Average Unit Cost of any future lot once you know the LMP of that lot, which is easily derived once you know the first unit (F) and last unit (L) of each lot, using Equation 10.9 for all subsequent lots.

10.8 Unit Theory, Final Example (Example 10.5)

Let's now try one final, and more involved, example. Given the following historical production data on a tank turret assembly, find the *Unit Theory Learning Curve* equation, which best models this production environment and estimate the cost (in man-hours) for the seventh production lot of 75 assemblies which are to be purchased in the next fiscal year.

With the given lot data information in Table 10.9, calculate a Unit Theory chart, as shown in Table 10.10.

Attempt to completely fill in this chart on your own. When it is completed, here are the calculations for each block:

TABLE 10.9   Historical Production Data for Example 10.5

Lot #	Lot Size	Cost (Man-Hours)
1	15	36,750
2	10	19,000
3	60	90,000
4	30	39,000
5	50	60,000
6	50	In process
7	75	??

TABLE 10.10   Unit Theory Learning Curve Chart for Example 10.5

Lot Size	Cost (Man-Hrs)	First	Cumulative/ Last	AUC	LMP	Y ln (AUC)	X ln (LMP)
15	36,750						
10	19,000						
60	90,000						
30	39,000						
50	60,000						
50	in process						
75	??						

For First and Last (Cumulative):

- Lot 1: the first unit F equals 1 and the last unit L equals 15;
- Lot 2: the first unit F equals 16 and the last unit L equals 25;
- Lot 3: the first unit F equals 26 and the last unit L equals 85;
- Lot 4: the first unit F equals 86 and the last unit L equals 115;
- Lot 5: the first unit F equals 116 and the last unit L equals 165;
- Lot 6: the first unit F equals 166 and the last unit L equals 215;
- Lot 7: the first unit F equals 216 and the last unit L equals 290;

Lot midpoint (LMP): using Equations 10.8 and 10.9:

- LMP1 = Since our first lot size is 15 units (which is ≥ 10), LMP1 = 15/3 = 5.00
- LMP2 = $\frac{16+25+2\times\sqrt{16\times25}}{4}$ = 20.25
- LMP3 = $\frac{26+85+2\times\sqrt{26\times85}}{4}$ = 51.26
- LMP4 = $\frac{86+115+2\times\sqrt{86\times115}}{4}$ = 99.97
- LMP5 = $\frac{116+165+2\times\sqrt{116\times165}}{4}$ = 139.42
- LMP6 = $\frac{166+215+2\times\sqrt{166\times215}}{4}$ = 189.71
- LMP7 = $\frac{216+290+2\times\sqrt{216\times290}}{4}$ = 251.64

Average Unit Cost (AUC): using Equation 10.10:
- AUC1: 36,750/15 = 2,450
- AUC2: 19,000/10 = 1,900
- AUC3: 90,000/60 = 1,500
- AUC4: 39,000/30 = 1,300
- AUC5: 60,000/50 = 1,200
- AUC for Lot 6 and 7 are unknown. In fact, we are solving for these numbers!

Table 10.11 is the completed calculations, including the natural logs of LMP and AUC.

TABLE 10.11 Completed Unit Theory Learning Curve Chart for Example 10.5, Including Natural Logs

Lot Size	Cost (Man-Hrs)	Cumulative/		AUC	LMP	*Y* ln (AUC)	*X* ln (LMP)
		First	Last				
15	36,750	1	15	2450	5	7.80384	1.60944
10	19,000	16	25	1900	20.25	7.54961	3.00815
60	90,000	26	85	1500	51.26	7.31322	3.93682
30	39,000	86	115	1300	99.97	7.17012	4.60491
50	60,000	116	165	1200	139.42	7.09008	4.93752
50	??	166	215	??	189.71	??	5.24549
75	??	216	290	??	251.64	??	5.52800

The regression will be between ln(AUC) vs. ln (LMP). Results are found in Table 10.12.

TABLE 10.12 Resultant Regression from ln(AUC) vs. ln(LMP) in Example 10.5

SUMMARY OUTPUT					
		ln(AUC) vs. ln(LMP)			
Regression Statistics					
Multiple R	0.99793				
R Square	0.99586				
Adjusted R Square	0.99449				
Standard Error	0.02168				
Observations	5				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.33942	0.33942	722.46721	0.00011
Residual	3	0.00141	0.00047		
Total	4	0.34083			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	8.16997	0.03076	265.62564	0.00000	8.07208
ln(LMP)	−0.21678	0.00806	−26.87875	0.00011	−0.24244

After completing the aforementioned regression, your results reveal an initial intercept $= 8.1699$ and $b = -0.21678$ and thus the equation:

$$\ln(\text{AUC}) = 8.1699 - 0.21678 \times \ln(\text{LMP})$$

After the natural log transformation, the Unit Learning Curve equation that we have calculated for this production scenario is:

$$Y_x = 3533.22x^{-0.21678}$$

The final equation reveals that $T1 = 3{,}533.22$ man-hours and the slope of the learning curve $= 2^b = 2^{-0.21678} = 86.05\%$. The original question in Example 10.5 asks for us to estimate the cost (in man-hours) of the tank turret assembly's seventh production lot. This lot contains the units from $F = 216$ to $L = 290$. Calculations below using the lot cost Equation 10.7 display the result that these 75 units in Lot 7 will take a total of 79,866.7 hours to produce.

$$\text{CT}_{F,L} \cong \frac{AL^{b+1}}{b+1} - \frac{A(F-1)^{b+1}}{b+1}$$
$$A = 3533.22$$
$$F = 216$$
$$L = 290$$
$$b = -0.217$$

$$\text{CT}_{216,290} \cong \frac{(3533.22)(290)^{-0.217+1}}{-0.217+1} - \frac{(3533.22)(216-1)}{-0.217+1} = 79{,}866.7 \text{ hrs}$$

Author's Note: Recall that the values for A and b were derived from the regression in Table 10.12. Once you have these two values from any regression – which tell you the intercept and slope of the regression line – you can calculate the lot cost for any subsequent lot. You just need to know the F and the L for that lot that you desire to cost.

10.9 Alternative LMP and Lot Cost Calculations

In Section 10.7, we learned the following parameter-free approximation equations to determine our LMP.

- For the first lot *only* (the lot starting at unit 1):

$$\text{If lot size} < 10, \text{ then LMP} = \text{Lot size}/2 \qquad (10.8)$$

$$\text{If lot size} \geq 10, \text{ then LMP} = \text{Lot size}/3$$

- For *all subsequent lots* (lot #2 and *every lot* thereafter):

$$\text{LMP} = \frac{F + L + 2\sqrt{F \times L}}{4} \qquad (10.9)$$

where

$$F \ = \ \text{the first unit number in a lot, and}$$
$$L \ = \ \text{the last unit number in a lot.}$$

These equations were necessary to calculate the LMP's and are used when we only have lot data and do not know the value of the learning curve or the value of the slope parameter, b. In fact, the slope (and intercept) is what we are ultimately trying to determine. Those values were found as a result of the regression between $\ln(\text{AUC})$ and $\ln(\text{LMP})$.

In some cases, however, you may actually know the value for b already or may have a good estimate of its value. Reasons for knowing or estimating the slope of the learning curve or the slope parameter, b, include the following:

- You have produced a certain number of items X, then experienced a production break and are now resuming production. The assumption is that the learning curve established prior to the production break remains the same as after production is resumed. The topic of production breaks will be covered in detail in Chapter 12.

- You may have an assumed slope, which is most likely taken from a similar historical program in a similar production environment, that you would use as a one-point analogy.

- You may have an assumed "base-case" slope. This might be necessary if you were performing a number of sensitivity analyses and a base-case number for the slope of the learning curve is necessary. It would be assumed from one historic program or a number of historical programs.

If any of these is the case, then equations using a slope can most likely provide a better approximation for the LMP. You will note that Equations 10.8 and 10.9 do not include any slope parameter. If you have a value for a slope, this changes the value of the LMP due to the changing shape of the curve and the area under the curve. Numerous studies have been conducted to help calculate a better LMP when a learning curve is known or assumed, and they include the following:

A popular LMP approximation method is H. Asher's (RAND Corporation) LMP approximation, where $F =$ the first unit in the lot and $L =$ the last unit in the lot. This approximation is highly accurate in most cases, except when the learning curve is very steep (in the 70% range) in the early lots: [3, page 6]

$$\text{LMP} = \left[\frac{(L + 0.5)^{1+b} - (F - 0.5)^{1+b}}{(L - F + 1) \times (1 + b)} \right]^{1/b} \tag{10.11}$$

(Note: If you have this scenario where the learning curve is steep in the early lots, you may consider using the "Asher's Approximation with Correction Terms," devised by Dr. David Lee. This long but accurate formula applies the Euler–Maclaurin summation formula, which approximates finite sums using integrals (and vice versa) to increase the accuracy of the approximation) [3, page 7].

One other popular alternate approximation method for the total lost cost equation found in Equation 10.7 is found in Equation 10.12.

$$\text{Total lot cost} = \frac{A \times (L + 0.5)^{b+1} - A \times (F + 0.5)^{b+1}}{b + 1} \tag{10.12}$$

Author's Note: As we end this introductory chapter on learning curves, we would like to point out a parallel that we find interesting. That is that the concept of learning curves is similar to decay models, such as the radioactive decay in a nuclear reactor. In a radioactive decay example, it is a radioactive item that has an initial known size that will decay and get smaller over time, declining exponentially. Similarly, in our learning curve scenarios, we begin with a known number of hours for our T1 (in dollars or hours) that will get smaller over time, declining not exponentially but rather via a power model. We offer this for the pictorial representation of the learning curve and its applicable, parallel concepts.

Summary

In this chapter, we introduced the theory of learning curves. Learning curve analysis is developed as a tool to estimate the recurring costs in an assembly or production process. The dominant factor in learning theory is direct labor, and the theory is based on the common observation that as a task is accomplished several times, it can be completed in shorter periods of time. Thus, each time you perform a task, you become better at it and accomplish the task faster than the previous time.

We discussed Unit Theory learning curve principles in detail and introduced numerous equations and five examples to guide you through the necessary steps to calculate how much these recurring costs are decreasing as you produce more lots and the total costs of those lots. Note that the "cost" of a unit can be expressed in dollars, labor hours, or other units of measurement. In the next chapter, we will continue to discuss learning curves and will discuss the second learning curve theory (second of 2 theories), called Cumulative Average Theory.

Author's Note: An excellent reference for additional learning curve background and the equations used in this chapter is "Cost Estimating," by Rodney D Stewart, Second Edition, 1991.

References

1. Stewart, Rodney D. "*Cost Estimating*", Second Edition. John Wiley and Sons, Inc., 1991, Chapter 4.
2. LCDR J. McCourt, USN, and Dr. Daniel A. Nussbaum. "How Good is the Folklore?" The Estimator Magazine, Society for Cost Estimating and Analysis (SCEA), Spring 1994, pages 14–21.
3. Lee, Jason T., "*Midpoint Formulas*," Technomics, Inc, 2007.

Applications and Questions:

10.1 The premise of learning curve theory is that each time you perform a task, you become better at it and accomplish the task faster than the previous time, due to an inherent improvement in efficiency. (True/False)

10.2 Learning curve analysis is developed as a tool to estimate the _____ costs in an assembly or production process.

10.3 What are the two predominant theories on learning curves?

10.4 In learning curve theory, a 95% learning curve is "better" than an 85% learning curve. (True/False)

10.5 If the slope parameter "b" $= -0.3154$, what is the slope of the learning curve?

10.6 In Unit Theory, if the cost of unit 100 is \$4M and the learning curve is 87%, what would you expect the cost of unit 200 to be?

10.7 In Unit Theory, if the cost of unit 100 is \$4M and the learning curve is 87%, what would you expect the cost of unit 300 to be?

10.8 Using the numbers calculated in question (7), what would the cost be for the first 100 units in that lot?

10.9 What would the cost of the second lot of 100 units be? (Units 101–200)

Learning Curves: Cumulative Average Theory

11.1 Introduction

In Chapter 10, we learned that the predominant idea behind learning curves is that as we repeat a task, we get better at it. This implies that a fixed amount of labor and facilities will produce greater and greater quantities in successive periods. In manufacturing, it has been found that we get better at a constant rate. In this chapter, we will focus on the second of the two learning curve theories, Cumulative Average Theory (CAT). While there are distinct differences between Unit Theory and Cumulative Average Theory, there are also many similarities as well. In this chapter, we will discuss those differences and similarities, as well as discussing under what conditions you would prefer to use one theory over the other.

11.2 Background of Cumulative Average Theory (CAT)

"If there is learning in the production process, the cumulative average cost of some doubled unit (say, unit 100) equals the cumulative average cost of the un-doubled unit (unit 50) times the slope of the learning curve." As mentioned in Chapter 10, this theory was first discovered by T. P. Wright back in 1936 and it was based on examination of World War I aircraft production costs. Aircraft companies and the Department of Defense were interested in the regular and predictable nature of the reduction in production costs that Wright observed. Indeed, it was found that aircraft manufacturers were able to produce greater and greater quantities using the same amount of labor in successive periods.

To illustrate this concept, for an 80% cumulative average learning curve, we observe the following:

- The average cost of 2 units is 80% of the cost of 1 unit;
- The average cost of 4 units is 80% of the average cost of 2 units;

- The average cost of 8 units is 80% of the average cost of 4 units; and
- The average cost of 50 units is 80% of the average cost of 25 units, etc.

Thus, for an 80% cumulative average learning curve, there is a 20% decrease in average cost each time that the cumulative quantity produced is doubled. This is in contrast to Unit Theory, which says that for an 80% learning curve there is a 20% decrease in the *unit* cost each time that the number of units produced is doubled.

Figure 11.1 is an example of an 80% Cumulative Average Theory learning curve. Note that in this case, the *Y*-axis is the Cumulative Average Cost, and the *X*-axis is Cumulative Quantity. It looks identical to the 80% learning curve in Unit Theory, merely with different units.
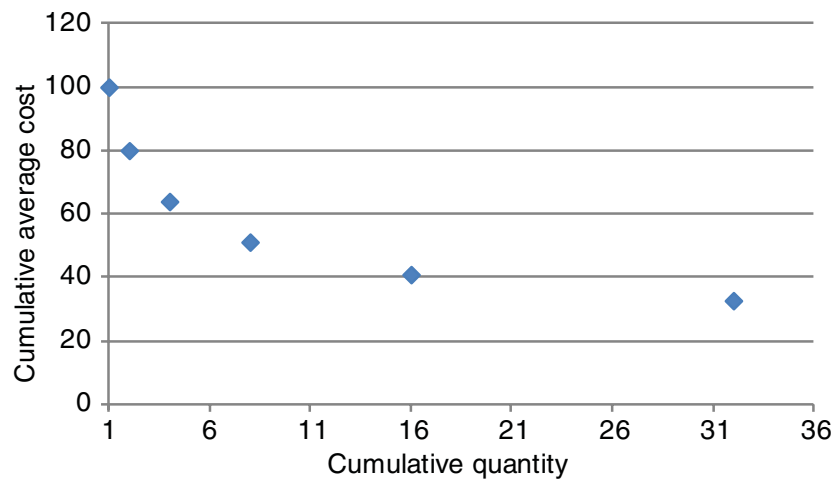
FIGURE 11.1 80% Cumulative Average Theory Learning Curve.

Cumulative Average Theory is defined by the equation:

$$\overline{Y}_N = AN^b \tag{11.1}$$

where

- $\overline{Y}_N$ = the cumulative average cost of N units
- A = the theoretical cost of unit one (aka T1)
- N = the cumulative number of units produced
- b = a constant representing the slope (where slope = 2^b)

Note that the cumulative average theory and the unit theory equations are quite similar and that A is again the *theoretical* (or mathematically calculated) cost of unit one and not necessarily the *actual* cost of the first unit. Note also that the A computed in Cumulative Average Theory will be different from the A computed for Unit Theory, even though the same symbol is used in both cases. The primary difference in the equations is that Unit Theory concentrates on each lot or unit *individually*, while Cumulative Average Theory calculates the cumulative average cost of all units to date, *collectively*.

Cumulative average theory is used in situations where the initial production of an item is expected to have large variations in cost due to:

- use of "soft" or prototype tooling
- inadequate supplier base established
- early design changes
- short lead times

This theory is preferred in these situations because the effect of averaging the production costs "smoothes out" initial unit cost variations.

The first similarity between Unit Theory and Cumulative Average Theory is that the learning parameter, b, is handled in exactly the same way in both theories. In words: "As the number of units doubles, the average unit cost is reduced by a constant percentage, which is referred to as the slope of the learning curve." The parameter "b" once again represents the steepness of that learning curve. Mathematically:

$$\text{Average cost of units 1 through } 2n = (\text{Average cost of units 1 through } n)$$

$$\times (\text{Slope of learning curve})$$

Isolating for slope, you find that:

$$\text{Slope of learning curve } = \frac{\text{Average cost of units 1 through } 2n}{\text{Average cost of units 1 through } n}$$

$$= \frac{A \times (2n)^b}{A \times (n)^b} = 2^b \tag{11.2}$$

Taking the natural log of both sides, you find that $\ln(\text{slope}) = \ln(2^b) = b \times \ln(2)$, and thus $b = \ln(\text{slope})/\ln(2)$. These equations are identical to the handling of "b" in Unit Theory.

To use a learning curve for a cost estimate, a slope and first unit cost are also required. As in Unit Theory, the slope may be derived from analogous production situations, industry averages, historical slopes for the same production site or historical data from previous production quantities. First unit costs may be derived from engineering estimates, CERs, or historical data from previous production quantities.

When historical production data is available, slope and first unit cost can be calculated by using the learning curve equation. But since $\overline{Y}_N = AN^b$ is not a straight line, we again attempt to make it linear by using a natural log transformation. By now, you are very familiar with this procedure.

11.3 Cumulative Average Theory

Example 11.1 Given the following historical lot data, find the Cumulative Average Theory learning curve equation for this production environment:

We start our example with the data set in Table 11.1 for four lots. The four lots contain 22, 36, 40, and 40 units, with total costs of $98M, $80M, $80M, and $78M, respectively, for the four lots. But, unlike in Unit Theory, there is no longer a *lot midpoint (LMP)* or an *average unit cost (AUC)* to calculate as our *X*-axis and *Y*-axis for graphing.

TABLE 11.1   Data Set for Example 11.1

Lot #	Lot Quantity	Lot Cost ($M)
1	22	98
2	36	80
3	40	80
4	40	78

Instead, we need to find *cumulative quantity* for the *X*-axis, and for the *Y*-axis we need to first calculate the *cumulative cost* after each lot, and then the *cumulative average cost* after each lot.

The *Y*-axis and the *X*-axis are now represented by these two terms:

- *Y*-Axis: Cumulative average cost (CAC)
- *X*-Axis: Cumulative quantity

While the parameters and inputs needed for building the Cumulative Average Theory chart differ from those in Unit Theory, you will find that this method is mathematically much easier to calculate.

Table 11.2 shows the columns that we need to solve for the Cumulative Average chart. Let's look at the mathematics of each column:

TABLE 11.2   Columns Needed to Solve for Lot Costs, Using Cumulative Average Theory Principles

Lot #	Lot Quantity	Lot Cost ($M)	Cum Quantity (N)	Cum Cost (M$)	Cum Avg Cost (M$)
1	22	98			
2	36	80			
3	40	80			
4	40	78			

Cumulative quantity:

- Since the first lot contains 22 units, the cumulative quantity after one lot is 22
- The second lot contains 36 units, so the cumulative quantity of items produced after two lots is 58 units ($22 + 36 = 58$)
- The third lot contains 40 units, so the cumulative quantity of items produced after three lots is 98 units ($58 + 40 = 98$)
- The fourth lot also contains 40 units, so the cumulative quantity of items produced after all four lots is 138 units ($98 + 40 = 138$)

Reminder: In Cumulative Average Theory, *Cumulative Quantity* is our *X*-axis for graphing now.

Cumulative cost:

In order to calculate the cumulative average cost, we must *first* calculate the cumulative cost for each lot.

- Since the first lot cost is $98M, our cumulative cost after one lot is $98M;
- The second lot cost is $80M, so the cumulative cost after two lots is $178M ($98M + $80M = $178M);
- The third lot cost is $80M, so the cumulative cost after three lots is $258M ($178M + $80M = $258M);
- The fourth lot cost is $78M, so the cumulative cost after all four lots is $336M ($258M + $78M = $336M).

Now that we have calculated the cumulative cost for each lot, we can now find the cumulative average cost for each lot.

Cumulative average cost:

- Since the first lot cost is $98M for 22 units, the cumulative average cost of the first lot is $4.455M ($98M/22 units = $4.455M cost per unit for 22 units).
- After two lots, we have invested a total of $178M for 58 units; thus, the cumulative average cost after the first two lots is $3.069M ($178M/58 units = $3.069M per unit for 58 units).
- After three lots, we have invested a total of $258M for 98 units; thus, the cumulative average cost after the first three lots is $2.633M ($258M/98 units = $2.633M per unit for 98 units).
- After four lots, we have invested a total of $336M for 138 units; thus, the cumulative average cost after all four lots is $2.435M ($336M/138 units = $2.435M per unit for 138 units).

In Cumulative Average Theory (CAT), *Cumulative Average Cost* is our *Y*-axis for graphing now. Note that in CAT, we are concerned with the *total cost* and *total average cost* after each lot *collectively*. For example, after three lots, we calculated that we had a total of 98 units produced, and we had spent a total of $258M. To contrast that, in Unit Theory, we were only concerned with what the lot cost was for each lot by itself (individually), not cumulative totals from the other lots, as well. Table 11.3 is the completed CAT chart:

If we were to graph Cumulative Average Cost (CAC) vs. Cumulative Quantity (Cum Qty), we would find that the data is nonlinear, so we will need to take the natural logs

TABLE 11.3 Example 11.1 Completed Calculations, Using Cumulative Average Theory Principles

Lot #	Lot Quantity	Lot Cost ($M)	Cum Quantity (N)	Cum Cost (M$)	Cum Avg Cost (M$)
1	22	98	22	98	4.455
2	36	80	58	178	3.069
3	40	80	98	258	2.633
4	40	78	138	336	2.435

of each of the four data points in order to make the graph appear linear. This is again necessary for us to be able to perform an OLS regression on this transformed data to find the equation of this learning curve. Table 11.4 is the fully completed Cumulative Average Theory chart, now including the natural logs:

TABLE 11.4 Completed Calculations for Example 11.1, With Natural Logs

Lot #	Lot Quantity	Lot Cost ($M)	Cum Quantity (N)	Cum Cost (M$)	Cum Avg Cost (M$)	(Y) ln (CAC)	(X) ln (Cum Qty)
1	22	98	22	98	4.455	1.494027	3.0910
2	36	80	58	178	3.069	1.121352	4.0604
3	40	80	98	258	2.633	0.968124	4.5850
4	40	78	138	336	2.435	0.889947	4.9273

With the ln calculations, we can now perform OLS regression to find out the equation of this line. Table 11.5 contains the results of that regression:

TABLE 11.5 Regression Output for Example 11.1

SUMMARY OUTPUT						
		ln(CAC) vs. ln(Cum Qty)				
Regression Statistics						
Multiple R	0.99514					
R Square	0.99031					
Adjusted R Square	0.98547					
Standard Error	0.03234					
Observations	4					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	0.21379	0.21379	204.45669	0.00486	
Residual	2	0.00209	0.00105			
Total	3	0.21588				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	
Intercept	2.50786	0.09851	25.45756	0.00154	2.08400	
ln(Cum Qty)	−0.33354	0.02333	−14.29884	0.00486	−0.43390	

Looking at the regression results, it appears that we have the following:

$$\hat{Y} = \text{CAC} = 2.5078 - 0.33354 \times \text{Cum Qty}$$

But actually, since the regression used natural log units, we instead have:

$$\ln(\text{CAC}) = 2.5078 - 0.33354 \times \ln(\text{Cum Qty})$$

Therefore, since we are in natural log units, we must transform the data back to the "original units" and we do this by using the exponential function. The intermediate transformation is $Y = \exp(2.5078) \times X^{-0.33354}$ and $b = -0.33354$, so the final equation (looking

like $\overline{Y}_N = AN^b$) becomes:

$$\overline{Y}_N = 12.278 \times (N)^{-0.33354}$$

The slope of this learning curve $= 2^b$; thus slope $= 2^{-0.33354} = 0.7936 = 79.36\%$.

Concluding Example 11.1, we have successfully completed calculating a Cumulative Average Theory learning curve equation from the given lot information. T1 = $12.278M, with a learning curve of 79.36%.

11.4 Estimating Lot Costs

The learning curve that best fits the production data must be used to estimate the cost of future units. As discussed in the previous chapter on Unit Theory, estimates are usually for the cost of units grouped into a production lot, say, "Lot *A* consisting of 50 units," or "Block C consisting of 100 units." Lot cost equations can be derived from the basic equation $\overline{Y}_N = AN^b$. Since $\overline{Y}_N = AN^b$ is the *average cost* of *N* Units, then the *total cost* of *N* units can be computed by multiplying the average cost of *N* units by the total number of units *N*. Thus, Equation 11.3 calculates the total cost of *N* units:

$$CT_N = AN^b \times N = A \times N^{b+1} \tag{11.3}$$

where $CT_N =$ the cumulative total cost of N units. Additionally, if you needed to find the *cost of unit N* (for some reason) while using Cumulative Average Theory, it can be approximated by Equation 11.4:

$$\text{Cost of Unit } N = (1 + b) \times AN^b \tag{11.4}$$

Note that Equation 11.4 is derived by taking the derivative of Equation 11.3, which is the *instantaneous slope* at that point in the curve.

In Cumulative Average Theory, Equation 11.5 allows us to compute the *total cost of a specific lot* with first unit #*F* and last unit #*L*:

$$CT_{F,L} = CT_L - CT_{F-1} = A \times [L^{b+1} - (F - 1)^{b+1}] \tag{11.5}$$

Note that the difference between the total cost Equation 11.5 here, and the total cost Equation 11.7 in Unit Theory, is that Equation 11.5 is *not* divided by $b + 1$. Note also that you are once again calculating the total cost of *L* units and then subtracting out the first $F - 1$ units that you are not interested in. You will need the same four parameters to solve this equation as in Unit Theory: A, b, F, and L.

11.5 Cumulative Average Theory, Final Example

Example 11.2 In this example, we will attempt the same problem that we completed at the end of the Unit Theory chapter (Section 10.8, Example 10.5), but this time we

will solve it using Cumulative Average Theory principles. Upon completion, we will then compare the two answers and discuss the differences in the results of the two theories.

Given the following historical production data on a tank turret assembly found in Table 11.6, find the *Cumulative Average Theory Learning Curve* equation that best models this production environment and estimate the cost (in man-hours) for the seventh production lot of 75 assemblies, which are to be purchased in the next fiscal year.

TABLE 11.6   Lot Data Information for CAT Example 11.2

Lot #	Lot Size	Cost (man-hours)
1	15	36,750
2	10	19,000
3	60	90,000
4	30	39,000
5	50	60,000
6	50	in process
7	75	??

With this given information, we will need to fill in the CAT chart with the following columns of data prior to performing our regression. Note that cost in this example is in man-hours (man-hrs) and not dollars, as it is in CAT, Example 11.1.

Attempt to fill in the Table 11.7 chart on your own, following the same methodology that was discussed in Section 11.3, CAT Example 11.1.

TABLE 11.7   Table Format for Example 11.2, using Cumulative Average Theory Principles

Lot #	Lot Size	Cost (man-hrs)	Cumulative Quantity	Cumulative Cost (man-hrs)	Cumulative Avg Cost (man-hrs)
1	15	36,750			
2	10	19,000			
3	60	90,000			
4	30	39,000			
5	50	60,000			
6	50	in process			
7	75	??			

Upon completion, check your work with the following calculations for Example 11.2 shown here:

Cumulative Quantity (CumQty): (units):
- After Lot #1, the cumulative quantity is: 15
- After Lot #2, the cumulative quantity is: (15 + 10) = 25
- After Lot #3, the cumulative quantity is: (25 + 60) = 85
- After Lot #4, the cumulative quantity is: (85 + 30) = 115
- After Lot #5, the cumulative quantity is: (115 + 50) = 165

- After Lot #6, the cumulative quantity is: (165 + 50) = 215
- After Lot #7, the cumulative quantity is: (215 + 75) = 290

Cumulative Cost: (man-hours):

- After Lot #1, the cumulative cost is: 36,750
- After Lot #2, the cumulative cost is: (36,750 + 19,000) = 55,750
- After Lot #3, the cumulative cost is: (55,750 + 90,000) = 145,750
- After Lot #4, the cumulative cost is: (145,750 + 39,000) = 184,750
- After Lot #5, the cumulative cost is: (184,750 + 60,000) = 244,750
- Unknown for Lots 6 and 7, and what we are solving for

Cumulative Average Cost (CAC): (= Cumulative Cost/Cumulative Quantity) (in man-hours):

- After Lot #1, the cumulative average cost is: (36,750/15) = 2,450
- After Lot #2, the cumulative average cost is: (55,750/25) = 2,230
- After Lot #3, the cumulative average cost is: (145,750/85) = 1,714.71
- After Lot #4, the cumulative average cost is: (184,750/115) = 1,606.52
- After Lot #5, the cumulative average cost is: (244,750/165) = 1,483.33
- Unknown for Lots 6 and 7, and what we are solving for

Note that CAC decreases after each lot, as it should in a learning environment. We are "getting better" after each lot.

Table 11.8 is the completed calculations in table format, including the natural logs of Cumulative Quantity (Cum Qty) and Cumulative Average Cost (CAC).

The regression will be between the dependent variable ln(CAC) vs. the independent variable ln (Cum Qty). Results from that regression are found in Table 11.9.

TABLE 11.8   Completed Calculations in Table Format for Example 11.2

Lot #	Lot Size	Cost (man-hours)	Cumulative Quantity	Cumulative Cost (man-hrs)	Cumulative Avg Cost (man-hrs)	(X) ln (Cum Qty)	(Y) ln (CAC)
1	15	36,750	15	36,750	2450	2.70805	7.80384
2	10	19,000	25	55,750	2230	3.21888	7.70976
3	60	90,000	85	145,750	1714.71	4.44265	7.44700
4	30	39,000	115	184,750	1606.52	4.74493	7.38183
5	50	60,000	165	244,750	1483.33	5.10595	7.30205
6	50	in process	215	??	??	??	??
7	75	??	290	??	??	??	??

TABLE 11.9   Regression Output for CAT, Example 11.2

SUMMARY OUTPUT					
		ln(CAC) vs. ln(Cum Qty)			
Regression Statistics					
Multiple R	0.99972				
R Square	0.99944				
Adjusted R Square	0.99925				
Standard Error	0.00594				
Observations	5				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.18800	0.18800	5320.70686	0.00001
Residual	3	0.00011	0.00004		
Total	4	0.18810			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	8.38010	0.01197	700.18825	0.00000	8.34201
ln(Cum Qty)	–0.21048	0.00289	–72.94318	0.00001	–0.21966

Our regression reveals an intercept = 8.3801 and a slope parameter $b = -0.21048$, so initially we have the following:

$$\text{CAC} = 8.3801 - 0.21048 \times \text{CumQty}$$

But since we are in log-linear units, what we really have is ln(CAC) and ln(CumQty), so we must take the exponential of this equation to convert back to regular/original units. In doing so, we calculate the following:

$$\exp(8.38010) = e^{8.3801} = 4359.43, \text{ and } b = -0.21048$$

and the final equation calculated for Cumulative Average Theory for this learning curve scenario is:

$$\overline{Y}_N = 4359.43 \times N^{-0.2105}$$

In this equation, T1 = 4,359.43 man-hours and the learning curve = $2^b = 2^{-0.21048} = 86.42\%$. Recall that in the original question, we were asked to estimate the cost (in hours) of the tank turret assembly's *seventh production lot*. Using Equation 11.5, we calculate that the seventh production lot, from $F = 216$ to $L = 290$, will take a total of 80,645 hours to complete (note: brackets used to more easily differentiate from the parentheses):

$$\text{CT}_{216,290} = A[L^{b+1} - (F-1)^{b+1}]$$

$$A = 4359.43$$

$$L = 290$$

$$F = 216$$

$$b = -0.2105$$

$$\text{CT}_{216,290} = 4359.43[290^{-0.2105+1} - (215)^{-0.2105+1}] = 80{,}645 \text{ hrs}$$

Let's compare the results from this problem, solved using Unit Theory principles in the previous chapter and solved using Cumulative Average Theory principles here. These results are found in Table 11.10.

TABLE 11.10   Comparison of Results Using Unit Theory and CAT Principles

	Number of man-hours	**Learning Curve**
Unit Theory	79,866	86.05%
Cumulative Average Theory	80,645	86.42%

You will note that the Cumulative Average total is higher by about 1%: 79,866 compared to 80,645, which will almost always be the case. In addition, the CAT learning curve is slightly higher as well: 86.05–86.42%, and recall that the *lower* the learning curve (in this case, 86.05% vs. 86.42%) the better, since you save more each time the number of units produced doubles. In Section 11.6, we will compare these two theories more closely.

11.6 Unit Theory vs. Cumulative Average Theory

A graphical comparison of the two theories is shown in Figure 11.2 and this graph reveals the following:

Note that for any given set of data, since cost generally decreases, the Unit Theory cost curve will roughly parallel the Cumulative Average Theory cost curve, but will *always* lie below it. The Unit Theory learning curve will also be steeper than the Cumulative Average Theory learning curve. There are a few reasons why this is the case:

- Since the Cumulative Average curve is based on the average cost of a production quantity rather than the cost of a particular unit or lot, it is less responsive to cost trends than the unit cost curve. *A sizable change in the cost of any unit or lot of units is required* before a change is reflected in the Cumulative Average curve.
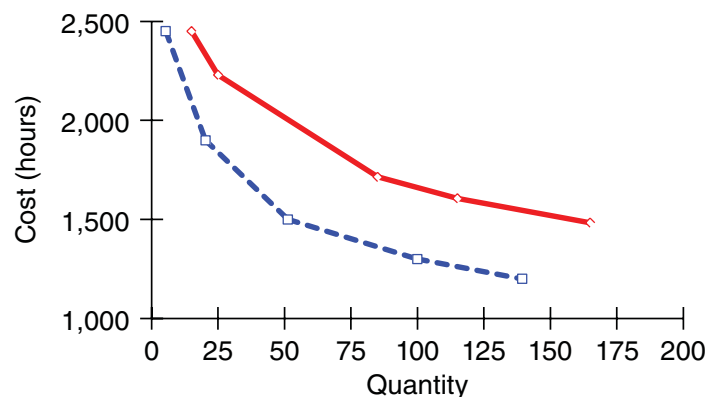
FIGURE 11.2 Comparison of Unit Theory (dashed line) vs. Cumulative Average Theory (Solid line).

- The Cumulative Average curve is "smoother" and always has a higher R^2. Moreover, it will have a smaller SSE, since the curve based on averages makes it is easier to get closer to the data points.

Since the Unit cost curve will always lie below the Cumulative Average cost curve, *government negotiators generally prefer to use Unit cost curves since it gives a smaller estimate and it is more responsive to recent trends.*

When graphing, differences in the two theories are:

- *Unit Theory*

   X-axis is lot midpoint (LMP);

   Y-axis is average unit cost (AUC)

- *Cumulative Average Theory*

   X-axis is cumulative quantity;

   Y-Axis is cumulative average cost

11.6.1 LEARNING CURVE SELECTION

Which type of learning curve to use is an important decision in your cost estimate, so how do you know when one theory is preferred over the other? There are many factors to consider when selecting the proper learning curve to use, including:

- *Analogous Systems*: Systems that are similar in form or function; or the development/production process being used may provide justification for choosing one theory over another
- *Industry Standards*: Certain industries gravitate toward one theory versus the other
- *Historical Experience*: Some defense contractors have a history of using one theory versus another because it has been shown to best model that contractor's production process
- Another factor to consider is "*Which model gives the better statistics?*"

Sometimes you may not have a choice. There are commands in the Department of Defense that *only* use Unit Theory cost curves in their cost estimates. Generally, however, your decision is based on the expected production environment, as certain environments favor one theory over another. Predominately, you will use one theory over the other when the following production environments are present:

- *Use Cumulative Average Theory*: When the contractor is starting production with prototype tooling, has an inadequate supplier base established, expects early design changes, or is subject to short lead-times.
- *Use Unit Theory*: When the contractor is well prepared to begin production in terms of tooling, suppliers, lead-times, etc. There is much less uncertainty in this type of production environment.

Summary

In this chapter, we continued the discussion of learning curve analysis and discussed in detail the second theory, Cumulative Average Theory (CAT). We introduced CAT equations and used two examples to guide you through the necessary steps to calculate the decreasing recurring cumulative costs. Since the equations used in each theory are different, one must ensure not to mix the two theories and their equations during a cost estimate or analysis.

Unit Theory focuses on the cost of each lot *individually*. For example, what is the cost of Lot #1? What is the cost of Lot #2? Lot #3? Each successive lot should be less expensive than the previous lot, if there is learning in the process. But in Cumulative Average Theory, we are calculating what the total (cumulative) costs and the cumulative average costs are after all of the lots are added together. After only one lot, the two theories are still similar: it is just the cost of that lot. But what is the cumulative quantity and cumulative average cost after 2 lots? We are now concerned with the total number of both lots added together and the average cost of those units up to that point. After three lots, we have added all of the three lot costs together and divided that total cost with the total number of units produced to that point. Therefore, you can see that the two theories have very different focuses. We also discussed when one of these theories is preferred over the other and how to make your learning curve selection during your cost estimate. If it is a well-defined program with little uncertainty as to tooling, suppliers, etc., you would most likely use unit theory learning curve principles. But if there was less certainty in any of these areas, then cumulative average theory may be preferred.

Now that we have learned the two primary theories, the next chapter discusses a detailed application of learning curves called Production Breaks/Lost Learning, in which Unit Theory principles must be used.

Applications and Questions:

11.1 The primary difference between Unit Theory and Cumulative Average Theory is that Cumulative Average Theory concentrates on each lot *individually,* while Unit Theory calculates the cumulative average cost of all units to date, *collectively*. (True/False)

11.2 Cumulative Average Theory is preferred over Unit Theory in some situations because the effect of averaging the production costs "smoothes out" initial unit cost variations. (True/False)

11.3 In Unit Theory, when calculating lot costs the dependent variable (*Y*-axis) is _____ and the independent variable (*X*-Axis) is _____. In Cumulative Average Theory, when calculating lot costs the dependent variable is _____ and the independent variable is _____.

11.4 In Cumulative Average Theory, if the slope parameter "*b*" $= -0.3154$, what is the slope of the learning curve?

11.5 In Cumulative Average Theory, if the average cost of 100 units is $4M and the learning curve is 87%, what would you expect the average cost of 200 units to be?

11.6 In Cumulative Average Theory, if the average cost of 100 units is $4M and the learning curve is 87%, what would you expect the average cost of 300 units to be?

11.7 Using the numbers calculated in question (5) and (6), what would the lot cost be for units 301 to 500, using cumulative average theory principles?

Learning Curves: Production Breaks/Lost Learning

12.1 Introduction

In Chapters 10 and 11, we learned about the two theories in learning curves: unit theory and cumulative average theory. In this chapter, we will discuss a detailed application that uses unit theory principles called production breaks, also known as lost learning. Production breaks occur when a program has produced X number of items, and production suddenly comes to a halt. Breaks can occur in a program for many reasons, including funding delays, strikes, or technical problems encountered during production. Breaks can last for just a few months or up to a few years. Generally, the longer the break, the more learning that will have been lost by production personnel, as workers "forget" the processes that they had learned. This will cause costs to increase from the level that they were at when the break occurred, since workers will need to be re-trained, assembly lines reconfigured and revalidated, and many other start-up costs incurred. This chapter will document the steps needed to determine how much learning was originally gained and then lost, what the increased cost of the first unit of production will be after the break is over, and to what unit do we need to go back to, "cost-wise," when restarting the production line (aka, resetting the learning curve). Production break analysis essentially measures the cost penalties associated with these breaks in production.

There have been numerous studies on production breaks and at least four methods have been devised to quantify the learning lost during them. A December 1988 thesis written at the Naval Postgraduate School in Monterey, California, by Captain Jeffrey Everest, USMC, discussed four of these methods: (1) the George Anderlohr Method, (2) the Defense Contract Audit Agency (DCAA) Method, (3) the Pinchon–Richardson Model, and (4) the Cubic Learning Curve Method. Everest discussed each method in detail and concluded that the Anderlohr Method was the most effective method to evaluate the loss of learning due to a break in production [1, page 3], and this method is still widely used today. This chapter will cover the Anderlohr Method and guide you through a "Lost Learning" example in detail.

12.2 The Lost Learning Process

When attempting to determine what the start-up costs will be after a production break and how much learning has been lost, two questions need to be explored:

1. How much of the learning that had been achieved already has now been lost (or forgotten) due to the break in production?
2. How will this "lost learning" impact the costs of future production items?

Question #1 can be answered by using the "Anderlohr Method," for estimating the amount of learning that has been lost due to the production break. Question #2 can then be answered by using the "Retrograde Method," which determines which unit on the original learning curve is the appropriate unit from where to start again "cost-wise."

We have discussed learning curves using both Unit Theory and Cumulative Average Theory principles. Since the ultimate aim in production breaks is to determine which unit we will need to go back to on the original learning curve, calculations in this area *must* use Unit Theory principles only.

12.3 Production Break Scenario

Let's suppose that the NeverCrash Aerospace Company has six different aircraft that it is producing. These aircraft include the NC #100, NC #200, NC #300, NC #400, NC #500 and the newest aircraft, the NC #600. One hundred (100) workers are presently working on the newest program, the NC #600, and so far they have produced 75 aircraft. But after the 75th aircraft was produced, funding delays precluded the program from continuing, thus a break in production will now occur.

Let us suppose that this production break lasts for a total of eight months. In this scenario, the management of NeverCrash Aerospace would not allow 100 personnel to just sit idly during this production break while waiting for the funding issues to be rectified. Instead, they would prefer to reassign these personnel to other programs that they are producing. Let's assume that 80 of the 100 personnel are reassigned and that management has decided that 10 personnel will be reassigned to the NC #100, 15 personnel will be assigned to the NC #200, 20 more each for the NC #300 and #400, and the final 15 personnel will work on the NC#500. That accounts for 80 of the 100 personnel. The final 20 personnel will remain on the NC #600, to handle any and all changes and issues that are ongoing on that program during the production break.

Eight months later, the funding issue is resolved, and management eagerly wants to restart the NC #600 production line to fulfill all contract orders. But while they are making plans to reopen the NC #600 production line, do you think that all of the original 80 personnel will return to the NC #600 program? The answer is almost certainly "No." Some of the personnel will probably be retained on the NeverCrash aircraft program that they were assigned to during the break, due to detailed involvement or major contributions in that program now. A few of the personnel may have been promoted from a "worker" position to a "supervisory" position. A handful may have accepted a position with another company, or a position within NeverCrash that did not involve the NC #600. A few may have retired, and a few may have taken maternity leave. So let us assume that a total of 20 personnel from the 80 reassigned will not be returning to the NC #600. This would mean

that 20 new personnel will need to be completely trained about all aspects of the NC #600. The other 60 personnel will have lost some skill and knowledge on the program due to the eight month delay, as well, so there will be some re-training necessary for those personnel. Therefore, those training and re-training costs will need to be added in to the cost of the first unit produced (unit #76, since 75 units were previously built) when production re-opens. Many other costs will need to be added in, as well. They include the cost to reassemble the assembly line, the cost to replace tools that were broken during the break while working on other programs, and the cost to update the manuals and instructions, as necessary. This would imply that the cost of the next unit to be produced (#76) will be *significantly more expensive* than the cost of the last unit produced (#75) prior to the production break.

So how do we determine what costs are necessary to include? What areas will cause the most cost? While each program is different, we will discuss standard procedures for determining what areas these costs will occur in.

12.4 The Anderlohr Method

To assess the impact on the cost of a production break, it is first necessary to quantify how much learning was achieved prior to the break, and then quantify how much of that learning was lost due to the break. When we discuss "How much learning was achieved," recall the garage door example in Chapter 10 on Unit Theory. The first garage door opener assembly took you 20 hours to accomplish. By the 15th door, you had improved to the point where it was only taking you 6 hours to accomplish the same task. Thus, you had "learned," or gotten better by, 14 hours. But if you took a long break between installing them, say a full year, then when you would attempt to install the next garage door (the 16th), there would be some learning that was lost, and the 16th door would no doubt take you more than 6 hours, since you would not be as competent as you were before the break. But how much longer would it take you now, due to the learning that you lost?

George Anderlohr, a Defense Contract Administration Services (DCAS) employee in the 1960s, divided the learning lost due to a production break into five categories [1, pages 33–34]:

- Personnel learning
- Supervisory learning
- Continuity of productivity
- Methods
- Tooling

Personnel learning: This area involves the workers who actually work on the production line. While some were retained on the program, the majority of the workers were transferred to other programs. The larger the personnel turnover or re-training that is necessary (due to the length of the delay), the higher the loss of skill and dexterity and knowledge that will occur in personnel learning.

Supervisory learning: This area involves the skills and dexterity and knowledge of the supervisors in the program. While a few may have been retained on the original program, the majority will most likely have been transferred to other programs. Therefore, there will be training required for the new supervisors, as well as re-training necessary for those

returning. There will also be more time required for the supervisors to train and supervise the new personnel.

Continuity of Productivity: This area involves the assembly/production lines. Whenever there is a break in production, costs occur in restarting an assembly/production line process. For example, if a program originally had four assembly lines open, at least two will most likely have been reassembled for a different program (say the NC #300) during the production break. Costs will be incurred when changing the assembly lines back to the original program, as well as the time needed for maintenance of equipment that has remained idle during the production break.

Methods: This area involves primarily the instruction manuals for the program involved. This will most likely be one of the areas least affected by the production break.

Tooling: The final category discussed by George Anderlohr involves replacing tools that were broken, lost, or that wore out during the break while working on other programs. Cost may also be incurred if the break involved the purchase of tools that will make assembly/production more efficient.

In analyzing these five categories, each production situation must be examined and a weight assigned to each category based on the situation. An example *weighting scheme* for an aircraft production line might be:

Category	Weight (%)
Personnel Learning	35
Supervisory Learning	20
Continuity of Production	20
Methods	10
Tooling	*15*
Total	100

Note immediately that the weights must sum to 100%. These weights imply the importance of each category and are dependent upon many issues, including the type of program, the amount of automation in that program (which will consequently greatly affect the personnel and supervisory learning categories), and the extent to which the final three categories have "losses" in this program. The longer the break is, the more learning that is usually lost. Note that these weighting factors are very subjective, and the weights above are estimated based on years of experience from management personnel.

To find the amount and the percentage of learning lost, we must find the learning lost in each of these five categories and then calculate a weighted average based upon the weighting scheme assigned. We will demonstrate this procedure in an example.

12.5 Production Breaks Example

Example 12.1 (Part 1) The Anderlohr Method. A contractor who assembles military trucks experiences a seven-month break in production due to a strike at the plant. During the break in production, the contractor transferred many of his resources to other programs. When the strike was settled, the contractor conducted a survey and provided the following information:

- 80% of the production personnel are expected to return to this program. The remaining 20% will be new workers or transfers from other programs.

- 85% of the supervisors are expected to return to this program. The remainder will be newly promoted personnel and transfers from other programs.

- During the production break, two of the four assembly lines were changed and converted to other uses. Thus, these two lines will have to be reassembled to their original configuration prior to commencing production.

- Also during the break, the contractor upgraded some capital equipment on the assembly lines, thus requiring modifications to 5% of the shop instructions.

- An inventory of tools revealed that 8% of the tooling will have to be replaced due to loss, wear and breakage.

- Finally, it is estimated that the assembly line workers will have lost 30% of their skill and dexterity and that supervisors lost 15% of their skills needed for this program, during the production break. (Note: this area is highly subjective! The longer the break and the higher the turnover of personnel that is occurring, the higher the percentage that will be lost in the Personnel and Supervisory areas).

The first calculation we need to accomplish is to establish how much learning has been lost. We will do so in each of the five categories, and will ultimately calculate a "lost learning factor (LLF)," which is the total percentage of learning that we have "forgotten" or "lost," and that we must "give back" to the process. From the given information, we can calculate the following percentages of learning that was lost in each of the five categories, found in Table 12.1:

TABLE 12.1   Calculations for Learning Lost in Each of the Five Anderlohr Categories

LLF calculation	A Percent Returning	B Percent Skill Retained	C Percent Learning Retained	D Percent Learning Lost
Personnel	0.80	0.70	0.56	0.44
Supervisors	0.85	0.85	0.7225	0.2775
Continuity of Production	XX	XX	0.50	0.50
Methods	XX	XX	0.95	0.05
Tooling	XX	XX	0.92	0.08

Let's discuss the results and calculations found in Table 12.1:

- *Column A* is the percentage of the personnel and supervisors that are returning to the program as it restarts. 80% of the original production personnel are returning to the program, as are 85% of the supervisors.

- *Column B* is the percentage of skill and dexterity retained by the returning personnel and supervisors. Personnel in this example are expected to have lost 30% of their skill and dexterity, therefore they retain 70% in this area. For the supervisors, since they are expected to lose 15% of their skill and dexterity, they will retain 85% in this area.

- *Column C* is the percentage of learning that was retained in the process. For personnel and supervisors, Column C is the product of Column A × Column B (i.e., for

personnel: $0.80 \times 0.70 = 0.56$; for supervisors: $0.85 \times 0.85 = 0.7225$). Therefore, personnel are expected to retain 56% of the learning that they had achieved prior to the production break, while supervisors are expected to retain 72.25% of their learning. For the other three categories, the values are given information, or "1 - given information." For Continuity of Production, since two of four assembly lines had remained intact, 50% of the learning in this area was retained; for Methods, since 5% of the shop instructions needed to be changed, 95% ($1 - 0.05 = 0.95$) of the shop instructions were retained. In the Tooling category, 92% of the tooling remained intact, since 8% had to be replaced.

- *Column D* is the percentage of Learning Lost in each of the five categories. This percentage is merely "100% minus the percentage of learning retained." Therefore, it is calculated by subtracting the learning retained percentage in Column C from 1.00 (i.e., for personnel, $1.00 - 0.56 = 0.44$; for supervisors, $1.00 - 0.7225 = 0.2775$, etc.)

Now that we have calculated the percentage of learning lost in each of the five categories, we must now combine those percentages with the *weighting scheme* (i.e., weights) that were assigned to each category to calculate the individual weighted losses. These five weighted losses are then summed to find the total weighted loss. This total weighted loss represents the total percentage of learning that was lost due to the break in production and is commonly referred to as the LLF. The calculations for the LLF are shown in Table 12.2:

TABLE 12.2 Lost Learning Factor (LLF) Calculations for Example 12.1

	A Weight	B Percent Lost	C Weighted Loss
Personnel	0.35	0.44	0.154
Supervisors	0.20	0.2775	0.0555
Continuity of Production	0.20	0.50	0.100
Methods	0.10	0.05	0.005
Tooling	0.15	0.08	0.012
		LLF =	0.3265

Results of Table 12.2:

- Column A is the weighting scheme assigned to each of the five categories
- Column B is the percentage of learning lost in each of the five categories (from Column D, Table 12.1)
- Column C is the weighted loss in each of the five categories. Column C is merely Column A × Column B for each category (i.e., for personnel, $0.35 \times 0.44 = 0.154$; for supervisors, $0.20 \times 0.2775 = 0.0555$, etc.)

Summing all five of the weighted losses in Column C reveals the LLF in this scenario, which is equal to 32.65%. The LLF represents the percent of learning that we will have to "give back" in the production process. This LLF will now be used to estimate the impact of the cost on future production using the Retrograde Method, a seven-step process.

This completes part 1 of Example 12.1 covering the Anderlohr Method.

12.6 The Retrograde Method, Example 12.1 (Part 2)

Continuing with Example 12.1, assume that 300 military trucks had been produced in Lot 1 prior to the seven month production break. The first truck had required 3,500 man-hours to complete (T1), and the learning curve slope for the first 300 units was calculated to be 92%. Using the LLF of 32.65% found in Table 12.2, estimate the number of hours required to produce the *next* 300 units (Lot 2), which are to be completed in the upcoming fiscal year.

Ultimately, we seek to estimate the number of hours required to produce the next 300 units (or Lot #2). Reference [2] presents a six-step process for the Retrograde Method to aid in this pursuit. We have added an additional step from this reference (between steps #5 and #6) and changed some of the wording for further amplification and clarification.

Therefore, the seven-step Retrograde method is as follows:

1. Find the amount of learning that had been achieved prior to the production break.
2. Estimate the number of hours of learning that were lost.
3. Estimate the cost of the first unit *after the break*.
4. Find the unit (*X*) on the original learning curve which is approximately the same as the estimated cost, in hours or dollars, of the first unit after the break.
5. Find the number of units that you need to retrograde.
6. Determine the First and Last units of the new lot that you want to calculate.
7. Estimate the cost of the new lot after the break.

Before continuing Example 12.1, let's consider the following scenario depicting pictorially what the intended process is in the Retrograde Method. Let's first assume that in a production scenario, 140 units have been produced before a production break occurs. This is shown in Figure 12.1.

In Figure 12.1, the cost of (or the time it took to complete) the 140th unit is depicted at Point A, and is for approximately 26 hours. After this 140th unit was produced, a break in production occurs for a total of 6 months.
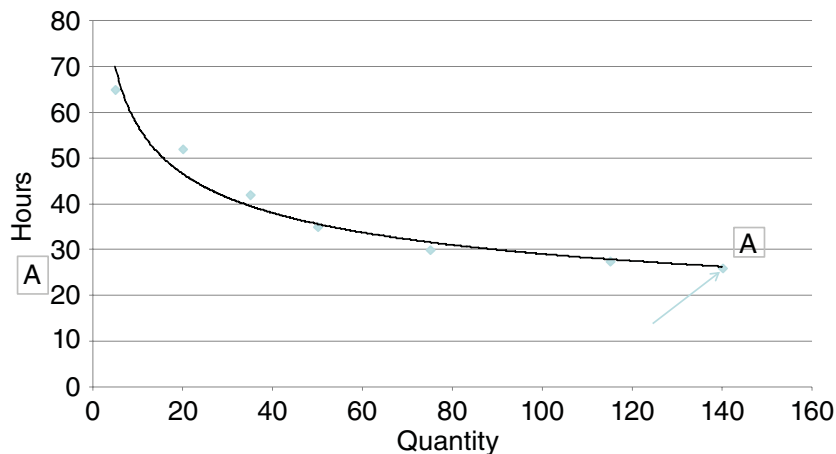
FIGURE 12.1 140 Completed Units in a Production Process Before a Production Break.

At the end of the six months, production resumes, at unit #141. Figure 12.2 displays the cost of this first unit (unit 141) after production resumes.
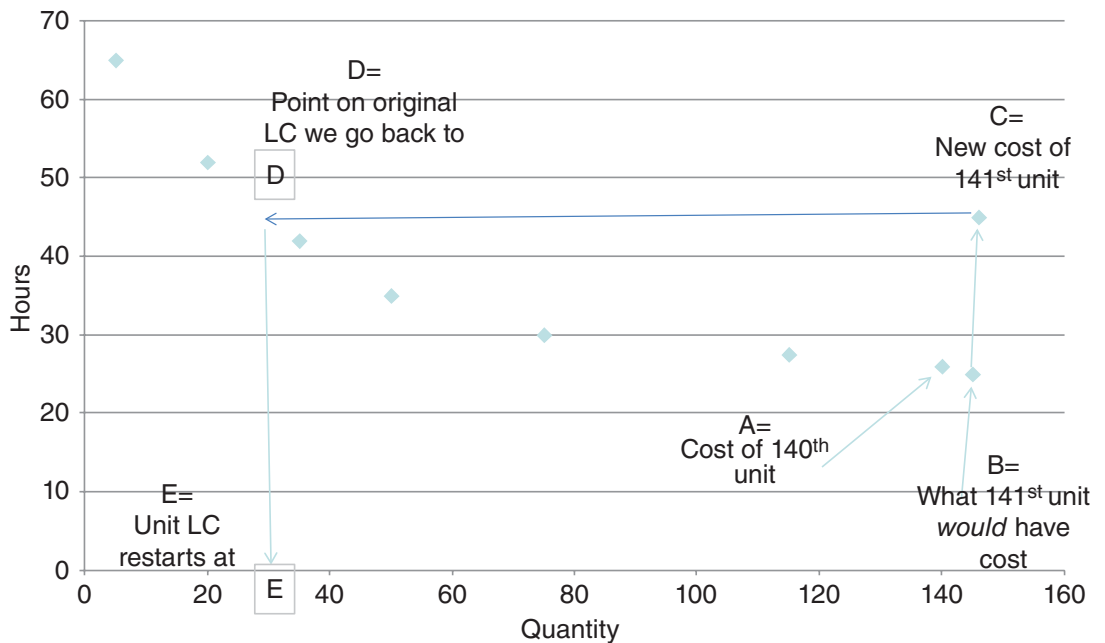
FIGURE 12.2 Calculating the Cost of the First Unit in a Production Process After Completion of a Production Break.

Figure 12.2 can be summarized as follows:

- *Point A*: The original production process had produced 140 units. The cost of unit 140 is 26 hours, which is shown as Point A.
- *Point B*: Without the production break, the cost of unit 141 would have been at Point B (approximately 25 hours).
- *Point C:* Since there *was* a break in production, however, a number of costs must be added to the first unit after production resumes, so unit 141 *now* costs the amount at Point C. We will call this unit "Unit 141*."
- *Point D:* This is where the cost for unit 141* (unit 141 after the break plus start-up costs) intersects with the original learning curve. In this scenario, the cost at Point C intersects the original learning curve at Point D and is approximately 45 hours.
- *Point E:* This is the unit on the original learning curve where we "cost-wise" resume production. In this scenario, the unit that took 45 hours to produce on the original learning curve (at Point D) corresponds approximately to unit 30, Point E.

Thus, in this scenario, the new unit 141 is equivalent in cost to unit 30 on the original learning curve and this is the point where we will re-commence "cost-wise" on the learning curve now that production has resumed. It was necessary for us to retrograde a total of $141 - 30 = 111$ units to get back to unit 30. If we were to calculate a new lot cost for the next 140 units, our First unit (F) would be $F = 30$, and our Last unit (L) would be $L = 169$, for the lot with units 30–169. Had there been no break, the new lot would have been costed from units 141 to 280. These calculations will be expounded upon in

Example 12.1. With this scenario in mind, let us now continue with our example and utilize the Retrograde method.

Continuing with Example 12.1, Part 2, recall that 300 military trucks had been produced in Lot 1 prior to the seven-month production break. The first truck had required 3,500 man-hours to complete (T1) and the learning curve slope for the first 300 units was calculated to be 92%. Using the LLF of 32.65% found in Table 12.2, estimate the number of hours required to produce the *next* 300 units (Lot 2), which are to be completed in the next fiscal year.

We will now calculate and discuss each of these seven steps of the Retrograde method for Example 12.1.

- *Step #1:* Find the amount of learning that had been achieved prior to the production break.
  The learning achieved is the difference between how long it took us to produce the first unit (Y_1) and how long it took us to produce the 300th unit (Y_{300}). From the Unit Theory chapter, Equation 10.3, we find that the learning curve of 92% for this example produces a value for the slope parameter $b = -0.1203$. We can then calculate that the 300th unit (Y_{300}) took 1,762.26 hours to produce. Therefore, since the first unit took us 3,500 hours to produce, and the 300th unit took us 1,762.26 hours to produce, we have gotten "better," or "learned," 1,737.74 hours after 300 units. This is a nice improvement from the number of hours it took us to complete the original first unit. Calculations include the following:

    Learning achieved (LA) $= Y_1 - Y_{300}$

    $LA = 3,500 - Y_{300}$

    $b = \ln(\text{slope})/\ln 2 = \ln(0.92)/\ln 2 = -0.1203$

    $Y_{300} = A \times X^b = 3,500 \times 300^{-0.1203} = 1,762.26 \text{ hours}$

    Therefore, $LA = 3,500 - 1,762.26 = 1,737.74 \text{ hours}$

- *Step #2*: Estimate the number of hours of learning that were lost.
  From Step #1, we found that we had "learned' a total of 1,737.74 hours by Unit 300. But we have now lost 32.65% of that total (= 567.37 hours), due to the break in production.

    Learning lost $= LA \times LLF$

    Learning lost $= 1,737.74 \times 32.65\% = 567.37 \text{ hours}$

- *Step #3*: Estimate the cost of the first unit *after the break* (in this example, $Y_{301}*$).
  The cost (in hours) of unit 301 "after the break" ($= Y_{301}*$) is estimated by adding the cost of what unit 301 *would* have been "without the break" (Y_{301}) plus the hours of learning that were lost (567.37, found in Step #2). In this example, the cost of unit 301 "before the break" (Y_{301}) would have been 1,761.55 hours. After the break, however, the cost of unit $301*$ is 1,761.55 hours plus the 567.37 hours of learning lost. Thus, the cost of unit $301*$ after the break is 2,328.92 hours.

    $Y_{301}*(\text{after break}) = Y_{301} + \text{Learning lost}$

    $Y_{301} (\text{before break}) = A \times X^b = 3,500 \times 301^{-0.1203} = 1,761.55 \text{ hours}$

$$Y_{301}{}^* = 1,761.55 + 567.37$$
$$Y_{301}{}^* = 2,328.92 \text{ hours}$$

- *Step #4*: Find the unit (X) on the original learning curve which is approximately the same as the estimated cost, in hours or dollars, of the first unit after the break.
  We have used the basic equation $Y_x = A \times X^b$ for the "cost of unit X" many times, but we have always known which unit that X corresponded to, and we were usually solving for the cost of that unit, Y_x. In Step 4, however, what we know are the values for Y_x, A and b, and we need to solve for the unit X that corresponds to the cost of $Y_{301}{}^* = 2,328.92$ hours. In the second and third equation mentioned subsequently, we are solving for X by dividing each side of the equation by A, and then raising both sides by the power of "1/b", which results in the third equation.

$$Y_x = A \times X^b$$

$$Y_x/A = X^b$$

$$X = [Y/A]^{1/b}$$

$$X = [2328.92/3500]^{1/-0.1203}$$

$$X = 29.55$$

Our calculations reveal that the original unit X on the learning curve that took 2,328.92 hours to produce is unit number 29.55. The first unit (T1) took 3,500 hours to make, and as we increased the quantity produced and moved down the learning curve, we eventually reached unit 29.55 (artificially, since we do not work in partial units) and it took 2,328.92 hours to produce that unit. But since we cannot work in partial units, we will round 29.55 down to unit number 29. Why round "down" to unit 29 instead of rounding "up" to unit 30? Rounding down is the safer and more conservative approach to take, since the first unit (F) in your lot will start one unit sooner (29 vice 30), and thus the estimated lot cost will be slightly greater. Usually in mathematics, rounding up occurs if a decimal of 0.5 or greater is calculated. In cost estimation, it would be more prudent and conservative to round down. While there is no definitive rule in cost estimation on this practice – so maybe we are creating one here! – we feel that you should round "down" until your decimal reaches 0.7 or greater. Therefore, in this example, if the unit we were going back to had been 29.70 or greater (instead of 29.55), we would have rounded up instead. When a unit is at least 70% completed, it is so close to completion that it is almost a whole unit and that is why we would recommend rounding up in that case.

In summary, Step #4 reveals that the cost of our new unit 301 ($Y_{301}{}^*$) is equivalent to the cost of unit number 29 on the original learning curve. Therefore, we will have to go back up the learning curve "cost-wise" to unit number 29 when restarting the production process. This will be the first unit (*F*) in our new lot.

An important assumption to point out at this time is that while *using this method, the learning curve slope is assumed to remain the same as it was before the break in production*. In this example, the previous learning curve was 92%. We will assume it will remain 92% unless otherwise proven. If the learning curve slope is *not* the same as

when producing Lot #1, due to perhaps significant improvement in the production process, then an entirely new learning curve will need to be assumed and it is advisable to start back at T1.

- *Step #5:* Find the number of units that you need to retrograde.
  The number of units of retrograde is how many units you need to go back up the curve to reach the unit found in Step #4. Since we *would* have started at unit 301, it is necessary for us to go back to unit 29 from unit 301. Therefore, we will need to retrograde a total of 272 units.

$$301 - 29 = 272$$

- *Step #6*: Determine the First and Last units of the new lot that you want to calculate. This step is accomplished by applying the number of units of retrograde to the First and Last units of the lot you originally wanted to cost. In this example, we originally wanted to calculate the lot from units 301 to 600. Since we are retrograding 272 units back up the learning curve, we must deduct 272 units from both the first and last units of the lot to solve for the *new* "First" and "Last" units of our lot. The first unit was supposed to be 301 and the last unit was supposed to be 600. The new First and Last unit are as follows:

  - First = 301 − 272 = 29
  - Last = 600 − 272 = 328

Therefore, the lot we *now* want to estimate is actually from Units 29 to 328, which will be considerably more expensive than what the lot from Units 301 to 600 would have been, had we not experienced a production break. *Note that the first unit F is always the same unit that you find in Step #4.*

- *Step #7*: Estimate the cost of the new lot after the break.
  We want to solve for the lot cost with units $F = 29$ to $L = 328$. Using the Unit Theory lot cost Equation 10.7, we calculate the following:

$$\mathrm{CT}_{F,L} \cong \frac{AL^{b+1}}{b+1} - \frac{A(F-1)^{b+1}}{b+1}$$

Given information:

$A = 3,500$
$b = -0.1203$
$b + 1 = 0.8797$
$F = 29, \ L = 328$

$$\mathrm{CT}_{29,328} = \frac{3,500 \times 328^{0.8797}}{0.8797} - \frac{3,500 \times 28^{0.8797}}{0.8797} = 575,441.13$$

Therefore, our estimated cost for the lot of the next 300 trucks will take *575,441.13 man-hours* to complete.

An interesting note is that if there had *not* been a break in production for seven months, the cost of Lot 2 from units 301 to 600 *would have* been 504,815.29 man-hours, using the lot cost equation with $A = 3,500$, $b + 1 = 0.8797$, $F = 301$

and $L = 600$. Instead, the new lot took 575,441.13 hours to produce. The difference is a total of 70,625.84 additional man-hours that needed to be used due to the seven month break in production. This of course equates to a much higher labor cost occurring due to the break.

Summary

Production breaks may occur in a program for many reasons and may last for just a few months or up to a few years. Generally, the longer the break, the more learning that will have been lost by both production personnel and by management supervisors, and this higher lost learning will cause higher costs during the restart process. In essence, production break analysis measures cost penalties that are associated with these breaks in production. This chapter discussed the Anderlohr Method for finding the percentage of learning that was lost and the Retrograde Method that determined which unit we needed to go back to on the original learning curve, in order to calculate the new lot costs once the break is over. The entire methodology must use unit theory learning curve principles and equations in its calculations.

An interesting article was written by John T. Bennett in TheHill.com in 2011. In his article, Bennett discussed a bipartisan group of 120 lawmakers who were warning that the US Army's plan to shut down a production line of Abrams tanks, and then re-opening the production line three years later, would not be cost-effective. "The cost of shutdown and restart of Abrams tank production appears to be more than the cost of continued limited production," according to the group. "Instead of reconstituting this vital manufacturing capability at a higher cost, it would seem prudent to invest those select resources in continued Abrams production." The article continues, and quotes Loren Thompson, a defense specialist at the Lexington Institute and an industry consultant, who said it is "hard to believe the Army's contention that shutting down the tank plant for three years and then restarting is cheaper than sustaining low-rate production." The problem, largely, is the impact on the current workforce, Thompson said. "You can mothball equipment, but you can't mothball people," he said. "Skilled workers will go elsewhere for jobs and suppliers will drift away" [3]. Bennett's contention is that it is less expensive to keep the production line open with limited production, rather than shutting it down for three years, due to the high cost in restart after the production break.

Author's Note: I read a study once by the RAND Corporation that found that there was a loss of (on average) 60% on a program with a two-year production break. I have tried to find the source of that study but am unable to reference it, other than that it was conducted by the RAND Corporation. But if a program experiences a two-year production break, you can expect that the lost learning factor (LLF) will be approximately 60%. For additional information on the topic of production breaks, see Reference [4].

In Chapters 10–12, we have discussed learning curves and how to calculate lot costs. Chapter 13 on Wrap Rates logically follows, as Wrap Rates is a method to assist you in converting the number of hours needed to produce a lot into dollars, as you attempt to cost out your program.

References

1. Everest, Jeffrey D. "Measuring Losses of Learning due to Breaks in Production". Masters Thesis, Naval Postgraduate School, December 1988.

2. Malashevitz Steve, Williams Bob, Kankey Roland. Retrograde Method. "Treatment of Breaks in Production", Defense Acquisition University (DAU), undated, pages 5–7.

3. Bennett John T., TheHill.com, "Lawmakers to Army: Keep Making Tanks". 10 May 2011

4. Delionback Leon M. "*A Prediction Model to Forecast the Cost Impact from a Break in the Production Schedule*". George C. Marshall Space Flight Center, Alabama, September 1977.

Applications and Questions:

12.1 When there is a break in production, the amount of learning that is lost by working personnel is highly correlated to the length of the break. (True/False)

12.2 After a break in production, the first unit produced after the break is significantly _____ than the last unit produced before the break.

12.3 The Anderlohr Method analyzes five categories where losses occur during a production break. Calculations using these five categories results in determining the_____.

12.4 When production resumes, the seven-step process to determine what unit we need to go back to on the learning curve is called the _____ _____.

12.5 A production that was halted for eight months is ready to recommence operations. Given the following information, calculate the percentage of Learning Lost in each of the five categories:

	Percent Returning	Percent Skill Retained	Learning Retained	Learning Lost
Personnel	70%	65%		
Supervisors	75%	75%		
Continuity of Production	XX	XX	33%	
Methods	XX	XX	94%	
Tooling	XX	XX	95%	

12.6 Given the following weighting factors, now calculate the Lost Learning Factor:

	Weighting Factor	Learning Lost	Weighted Loss
Personnel	0.30		
Supervisors	0.25		
Continuity of Production	0.15		
Methods	0.15		
Tooling	0.15		
		LLF =	

12.7 Continuing the same problem: Given that T1 = $55,000 (FY13$), the learning curve is 89%, and 500 items were made prior to the break in production, answer each of the seven steps of the Retrograde method. What would be the cost for the next lot of 500 items?

Wrap Rates and Step-Down Functions

13.1 Introduction

In the previous three chapters, we learned about calculating the cost of several items being produced and grouped into lots. We accomplished this by using two different learning curve theories. In Section 12.6, Example 12.1, we calculated that the new lot of 300 units would take 575,441.13 hours to complete. But how much money is needed to *pay* for that many labor hours? In this chapter, we will discuss a technique called the *Wrap Rate technique* which will help you convert the labor hours to a dollar figure. It is a method used to allocate profit and other overhead costs to actual labor costs. Wrap rate total costs equals the sum of three costs: direct labor costs, overhead costs, and other costs. This chapter will discuss each of these costs, provide a detailed example on how to calculate a wrap rate, and suggest what area you should focus on during contract negotiations that has the greatest potential to lower total costs in your program. Wrap Rate is also called the *Fully Burdened Labor Rate.* In the second half of this chapter, we will discuss a final application of learning curves called Step-Down Functions.

13.2 Wrap Rate Overview

Many of our cost estimating relationships use labor hours as a cost driving variable, and many programs predict costs in terms of a cumulative number of direct labor hours. There are two major categories of direct labor. *Manufacturing* is the "hands on" effort to produce a product and *Engineering* is the activity associated with the research, design, and development (or preparation) of products and procedures. Once we have determined the total labor hours required to produce an item (usually via learning curves or some CER), we need a means to convert those total labor hours to dollars. A Fully Burdened Labor Rate is a rate which includes *all* of the contractor costs and hours necessary to complete the task, and we then convert those required hours into dollars.

Example 13.1 True Story! About three years ago when I was leaving my home for work one day, I went into the garage and pushed the button for the garage door to open … . . . . and absolutely nothing happened. While I was ultimately able to open the door manually, further inspection revealed that the large spring that opens the garage door had broken, and was something beyond my capability to fix. So, I phoned the garage door repair company and made an appointment for the following day. When the repair man arrived the next day, he conducted his inspection and found that the spring was indeed broken, and went about his business to remove the old one and replace it with a new one. Two hours later, the job was complete and the repair man handed me a bill for $350. When I looked closer at the bill, it turned out to be divided into two sections: $180 for labor and $170 for parts. The $180 labor for two hours equates to $90 per hour, which is not bad money!

Let's look closer at this $90 per hour charge. Does the repair man actually make $90 per hour for his salary? The answer, of course, is no. Knowing that this is the case, let's inspect what this $90 per hour actually pays for.

First, I searched the internet to find what an average salary for a garage door repair technician might be. Not surprisingly, the salary will vary depending on the location within the United States, but the repair man does make a decent salary, and it appeared to be somewhere between $20 and $30 dollars per hour. For our example, let's take the average and call it $25 per hour. This amount would be considered the *Direct Labor Rate*.

So if the repair man is making $25 per hour, where is the remainder of the $90 per hour going? To answer this, let's think about the process. I had to make a phone call to the garage door company, and a receptionist had to answer the phone. Thus, part of the money will pay for the receptionist's salary and the phone that he/she is using. The company has its headquarters in an office space, and that office space needs heat, air conditioning, and electricity; plus there is a rent or a mortgage, and numerous office supplies, and this list can get very lengthy. While there are plenty of categories within that building that need to be paid for, there is also the truck that the repair man drove to my house that needs to be considered. Within the truck, there is also POL (petroleum, oil, and lubrication) costs, gasoline that was used to get there, the tools that he needed to conduct the repairs, and, of course, there is always an expensive "special tool" unique to that business that really drives costs up – in this case, the special tool was the tool needed to stretch and attach the large spring used. The owner of the company must also pay insurance for all of the workers, plus Social Security, workmen's compensation, etc. It turns out to be a lot of overhead costs that we need to consider! So, let's say that the overhead costs turn out to be $50 per hour, so we now have $25 in direct labor and $50 in overhead costs, so that accounts for $75 thus far. The final $15, then, is for the "other costs" such as profit that the company needs to make in order to stay in business.

Thus,

$$\text{Fully Burdened Labor Rate} = \text{Direct Labor Rate} + \text{Overhead Costs} + \text{Other Costs}$$
$$(13.1)$$

To simplify Equation 13.1, the sum of the direct labor cost and the overhead cost makes the company (theoretically) "break even," so in this case the final $15 per hour is the profit and fees that the company requires to make the business profitable.

13.3 Wrap Rate Components

Let's examine each of the wrap rate components (Direct Labor, Overhead, and Other costs) more closely.

13.3.1 DIRECT LABOR RATE

There are four general factors that impact the direct labor wage rate:

- *Variations in Geographical Locations.* This variation focuses on the cost of living by geographical location. A task accomplished in a higher cost of living area such as San Diego, CA, will generally cost more than if the job was accomplished in a comparatively less expensive market, such as Boise, Idaho. In this example, this will force the direct labor wage rate to be higher in San Diego.

- *Variations in the Contractor Labor Force:* This variation focuses on the Supply and Demand of workers and the work available. When a company has only a few positions available, they can generally hire workers for less than the normal salary which might be expected for such a position. This is because workers are willing to take a job that pays less when there are fewer jobs available – some salary is better than no salary! Conversely, a company will have to pay a higher salary when there are few workers available with certain skills, to entice that worker to sign with them and not with a competitor. Moreover, when a significant amount of work is available, there will be more salaries to pay, versus when there is little work available and thus fewer salaries to pay.

- *Variations in Skills:* Some workers have a high level of skill and education, while others have significantly less in both of those areas. Generally, those with higher skill sets are compensated more than those with lower skill sets to offer. Companies may need to look around the country if they are seeking a specific skill that few people possess. This will also cause the salary for that employee to be significantly higher.

- *Variations with Time:* These are variations that occur due to inflation and Cost of Living Allowances (COLA's) necessary in certain geographical locations.

The second component of a wrap rate is the Overhead Rate.

13.3.2 OVERHEAD RATE

Overhead is also known as "Indirect Costs" or "Burden." These costs are not directly identifiable with a *specific* cost objective, but rather are spread throughout the program. There are two distinct types of overhead costs:

- Those that are so general in nature they cannot be assigned to a specific cost objective. Examples include:
  - Plant and equipment maintenance
  - General and administrative
- Those that are so inconsequential per item that the cost of accounting for them exceeds the benefits derived from doing so. Examples include:
  - Consumables such as washers, sandpaper, lubricants.

Most firms collect indirect costs in *aggregate cost accounts* called overhead pools. Examples include:

- Manufacturing and engineering indirect pools

- Material overhead pools
- General and administrative pools
- Service center pools

These pools generally consist of the following:

- *Manufacturing overhead*: Indirect labor costs including supervision, inspection, and maintenance.
- *Engineering overhead*: Costs of directing and supporting the activities of the Engineering Department.
- *Material overhead*: Costs related to transportation, acquisition, inspection, and handling of materials.
- *General and administrative*: The costs of the company's general and executive offices.
- *Service center pools*: The costs associated with the activities of a service center.

How do we calculate what these overhead costs are? In order to recover overhead costs, it is necessary to allocate the indirect costs to a particular cost objective. This allocation results in charging the primary cost objective a share, or percentage, of the total indirect costs. Overhead recovery bases often include:

- Direct labor hours
- Direct labor dollars
- Number of production units
- Machine hours used

The overhead cost, then, will be a percentage of one of these recovery bases. For example, the overhead cost for a particular contract might be 120% of the direct labor hours in that program or perhaps it is 40% of the machine hours used to produce the product. This brings us to the final component needed to build a Fully Burdened Labor Rate, which is "Other Costs."

13.3.3 OTHER COSTS

"Other Costs" typically include:

- Profit (or fee)
- Cost of money
- General and administrative

The profit, or fee, is the most understood of these three inputs and needs little expounding upon. Clearly a company needs to make a profit in order to stay in business. But what is meant by the *cost of money?* Consider when you take an international trip. Once you land at your final destination and pick up your luggage from baggage claim, the first thing you generally do is exchange your US Dollars for the currency of the country that you are now located. Let's suppose that you have landed in the United Kingdom. When you give the attendant $100 of US currency in exchange for British Pounds, do

you receive the equivalent of $100 in British currency back? The answer is no; you will get less, most likely somewhere around $93 to $95 in value in return. The $5.00–$7.00 surcharge would be considered a "cost of money."

One other example of the cost of money is the following: let's suppose that a US-based company receives a contract from the United States government. There is generally a lag time, or delay, between when the contract is signed and when the company will actually receive their money. What if the first installment of the contract takes two to three months to finally arrive? It could even be longer than this if the contract is signed toward the end of any given fiscal year (or if there is a governmental Continuing Resolution in effect). Would the company remain idle for two to three months and commence no work because the money has not arrived yet? The answer would be no – they would actually begin to get the "wheels in motion" and to start working on the new program prior to any money arriving. This might require them to go to a bank to borrow money first, and the money that they borrow will accrue some percentage of interest that will have to be paid back. This is also considered a "cost of money."

The final input to "Other costs" is for any general and administrative costs that also need to be accounted for. So in our garage door example, the "other costs" accounted for a total of $15 per hour. These costs are usually derived as a percentage of the direct labor wage rate and the overhead costs. Having discussed the three elements of the Wrap Rate technique, let's consider the following example to show these cost calculations.

13.4 Wrap Rate, Final Example (Example 13.2)

Our example comes from a contract that involves work on a ship's hull. Recall first the garage door example, Example 13.1, and the $90 per hour "fully burdened labor rate" charged, broken down as follows:

- Direct labor wage rate: $25 per hour
- Overhead costs: $50 per hour
- Other costs: $15 per hour

Recall also that Equation 13.1 is:

$$\text{Fully Burdened Labor Rate} = \text{Direct Labor Wage Rate} + \text{Overhead Costs}$$
$$+ \text{Other Costs}$$

Task: Estimate the *fully burdened labor rate per hour* and then the *fully burdened contractor total support cost* on a ship's contract given the following information:

- Expected contractor support is 120 man-months
- One man-month equals 160 man-hours
- Contractor support wage rate is $52.50 per man-hour (CY13$)
- Overhead rate is 150% of Direct Labor dollars
- "Other Cost" rate is 15% of (Direct Labor dollars + Overhead dollars)

Solution: calculating the wrap rate in Example 13.2: (all costs in CY13$)

- Wage Rate = $52.50 per labor hour (i.e., the average salary that a worker makes)
- Overhead Rate = (150%)(Wage Rate) = (150%)($52.50) = $78.75 per labor hour
- Other Cost Rate = (15%)(Wage Rate + Overhead Rate) = (15%)($52.50 + $78.75) = $19.69 per labor hour
- Wrap Rate = Labor Wage Rate + Overhead Rate + Other Cost Rate = $52.50 + $78.75 + $19.69 = $150.94 *per labor hour*. This is also called the fully burdened labor rate.

Thus, while the salary of a worker is $52.50 per hour, it will actually cost the government $150.94 per labor hour. We must now calculate the total number of hours to complete this contract:

- Total Labor Hours = 120 man-months × 160 man-hours per man-month = 19, 200 man-hours

It will take a total of 19,200 man-hours to complete this contract. Therefore, the *fully burdened contractor total support cost is* $150.94 *per man-hour* × 19, 200 *man-hours* = $2, 898, 048 *(CY13$)*

Note how quickly we increased from $52.50 per labor hour to a total of $150.94 per labor hour once the overhead and the other costs are included!

If you are negotiating a contract, note that the area that you can *greatly* influence cost savings is if you are able to negotiate the overhead rate percentage down. This is because the overhead rate is used twice in the wrap rate calculations: first in the overhead rate itself, and second, as a percentage of the "other costs." Initially reducing this overhead rate percentage will save you costs in both of these areas.

13.5 Summary of Wrap Rates

In these first four sections, we discussed the *Wrap Rate technique* as a method used to allocate profit and other overhead costs to actual labor costs. It is also the method used to convert a total number of labor hours to labor dollars. This is a direct application from when we discussed lot costs previously in Chapters 10, 11, and 12, converting a total number of labor hours derived from a particular lot into labor dollars that can now be charged to a contract. A "real life" example was provided (the garage door example) to illustrate how this technique is used in daily living. The Wrap Rate total cost, also known as the *Fully Burdened Labor Rate,* equals the sum of three costs: direct labor costs, overhead costs, and other costs. Each of these costs was discussed and a detailed example on how to calculate a wrap rate was presented. Finally, a suggestion was offered to concentrate on the contractor's overhead rate as the area of greatest potential savings during contract negotiations, when attempting to lower total costs in your program.

In the second half of this chapter, we will discuss our final application of learning curves, a means by which to calculate first unit production costs from prototype costs. This method is called a Step-Down Function.

13.6 Introduction to Step-Down Functions

Chapters 10–12 discussed how you can determine production costs of a particular unit or a particular lot using either of the two learning curve theories. The first part of this

chapter then discussed how to convert the total labor hours required to produce a lot into dollars. However, instead of costing out a number of units in a lot, what if you need to calculate what the cost of a *prototype* might be? Do you use learning curves for that as well? Most often, the answer to that will be "No." Prototypes generally do not use a learning curve; instead, they accumulate an average unit cost (AUC) for the number of prototypes that were built, since there are usually so few of them. For example, if there were five prototypes of an unmanned aerial vehicle (UAV) produced, there would not be a learning curve calculation necessary for the five units, nor a calculation to determine that prototype #5 was less expensive than prototype #1. Rather, there would be a total cost and an average unit cost for the five prototypes collectively. Let's suppose that the total cost for the five prototypes was $15M. This total would then be divided by the five units to produce an AUC per prototype of $3M.

The remainder of this chapter will discuss step-down functions. Recall that there are three phases in the acquisition life cycle: the research and development (R&D) phase, the production phase, and the operating and support costs (O&S). Step-Down Functions occur between the R&D and production phases and is a method of estimating the theoretical first unit production cost based upon prototype cost data. Thus, there is a "step-down," or reduction, from the cost of a unit in development (the prototype) to the cost of the first unit in the production phase.

13.7 Step-Down Function Theory

It has been found, in general, that the average unit cost of a prototype is more expensive than the first unit cost of a corresponding production model. This is due to the many changes that may be necessary along the way during Test and Evaluation (T&E) to make the new systems operate as required. Let's suppose that a new system is being tested. Originally, engineers estimated that only 50 psi of hydraulic pressure would be necessary to operate this new system. But during the testing phase, it was found that the hydraulic pressure needed to operate properly, consistently, and safely was actually 75 psi. Thus, the system failed during testing due to the necessity for greater hydraulic pressure requirements. This would require the program manager to halt any further testing until the system could be brought up to the required 75 psi. Costs to complete this would be incurred not only in the hydraulic material and hoses used to retrofit the system, but also in the personnel and labor costs that would be required to make the engineering changes. Consequently, this would raise the unit cost for that prototype. Thus, this prototype cost would be greater than the eventual cost of the first unit in the production phase, once all of the "kinks" and deficiencies were worked out.

The ratio of the production phase first unit cost to the prototype average unit cost is known as a "Step-Down" factor. The actual cost difference between the average unit cost of the prototype and the production first unit cost is known as the Step-Down.

An estimate for the Step-Down factor for a given weapon system can be found by examining historical data on similar weapon systems and developing a cost estimating relationship (CER). Prototype average unit cost will be the independent variable; the dependent variable will be the first unit production cost (aka, A or T1), since that is what we are trying to solve for. Here is an example of where you are regressing one type of cost against another!

Once an appropriate CER is developed, it can be used with actual or estimated prototype costs to estimate the cost of the first unit of production. Let's illustrate this with an example.

13.8 Step-Down Function Example 13.1

We desire to estimate the first unit production cost for a new missile radar system. Let's call this new system the APGX-500. Historical data leads us to believe that the learning curve for the new APGX-500 will be 95%, *using unit theory principles*. The estimated total prototype cost in the development phase is expected be $28M for 8 prototype radars.

Historical data on four similar radar systems has been collected and can be found in Table 13.1. All historical costs have been converted to FY12$.

TABLE 13.1   Historical Data for Example 13.1

Radar system	Production Cost @ Unit 150	Number of Prototypes	Prototype Total cost	Prototype Avg unit cost
APG-100	0.995M	13	97.11M	7.47M
APG-200	0.414M	12	33.36M	2.78M
APG-300	2.5M	4	73.2M	18.3M
APG-400	1.852M	11	145.75M	13.25M

A scatter plot of this data is shown in Figure 13.1. The *Y*-axis is Production Cost (at Unit 150 on the learning curve), since that is the cost we are solving for. The *X*-axis is Prototype Average Unit Cost (P-AUC).
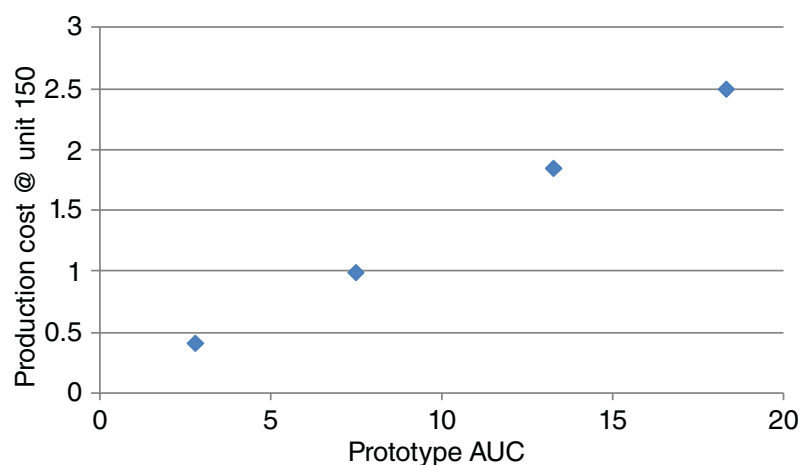

FIGURE 13.1 Scatter Plot of Data in Example 13.1.

In looking at the Figure 13.1 scatter plot, it is clear that the data is already linear, *so no natural log conversion is necessary*. The regression results of Production Cost (@ Unit 150) vs Prototype AUC are found in Table 13.2.

TABLE 13.2 Regression Output for Example 13.1

SUMMARY OUTPUT					
		Production Cost vs. Prototype AUC			
Regression Statistics					
Multiple R	0.99943				
R Square	0.99886				
Adjusted R Square	0.99829				
Standard Error	0.03808				
Observations	4				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.54114	2.54114	1752.43157	0.00057
Residual	2	0.00290	0.00145		
Total	3	2.54404			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	0.01793	0.03895	0.46026	0.69053	–0.14965
Prototype AUC	0.13611	0.00325	41.86205	0.00057	0.12212

As the historical data gave production costs for Unit #150 on the learning curve, we need to develop our CER based on this Unit #150 cost. We will then have to go "back up the learning curve" to calculate the theoretical first unit cost (T1), using the Unit Theory learning curve equation $Y_x = A^*X^b$.

From the regression results in Table #13.2, we calculate the regression equation to be:

$$\text{Production Cost (@ Unit 150)} = 0.0179 + 0.1361^*(\text{P-AUC})$$

Since the Prototype total cost is $28M for eight prototypes, we have a P-AUC = $28M / 8 = $3.5M. Thus:

$$\text{Production Cost (@ 150)} = 0.0179 + 0.1361^*(3.5M) = .49425M = \$494,250 \text{ FY12\$}$$

Now that we have calculated the estimated cost of the 150th unit in production, we can now find an estimate for our T1 in production, going "back up" the learning curve using our unit theory cost equation and the (assumed) 95% slope as follows:

$$Y(150) = A^*X^b = \$494,250$$

$$b = \ln(0.95)/\ln(2) = -0.0740$$

Substituting, we get:

$$Y(150) = \$494,250 = A^*(150)^{-0.0740}$$

and solving, we find:

$$A = \$716,106.19 \text{ FY12\$}$$

This result gives an estimate for the first unit production cost to be $716,106.19 (FY12$)

Since prototype AUC was \$3.5M and the first unit production cost is estimated to be \$716,106, we have a significant step-down from P-AUC to the first unit production costs.

- The step-down factor is: \$716, 106 / \$3.5M = 0.2046, meaning that the cost of the prototype is only 20.46% of the AUC of the prototypes. This implies a reduction in cost of 79.54% from the prototype AUC to the cost of the first unit in production.
- The total step-down in actual dollars is \$3.5M − \$716, 106 = \$2, 783, 894. (FY12\$)

A study on the actual step-down in hardware costs from the Research and Development phase to the production phase was conducted in 1994 by Paul Hardin and co-author Dr. Dan Nussbaum, then of NCCA. Historical costs in the areas of missiles, shipboard electronics, arrays, tracked vehicles, general electronics, and specific radar systems were examined. The typical step-down factor was found to be in the range of 0.47 to 0.67, proving a significant reduction in cost for the first unit of production from its prototype development cost [1].

13.9 Summary of Step-Down Functions

In the second half of this chapter, we discussed Step-Down Functions, which is a method of estimating the theoretical first unit production cost based upon prototype cost data in the development phase. While production costs of a particular unit or lot can be calculated via learning curve theory, prototypes generally do not use a learning curve. Instead, they compute an average unit cost for the number of prototypes that were built, since there are usually very few of them. We then provided an example on how to use historical data to create a CER to determine what the production first unit cost would most likely be.

This was the final chapter that discusses learning curves. In the next few chapters, we will explore a few other methodologies that cost estimators can use to derive costs. Specifically, Chapter 14 will examine two new areas of study: cost factors and the analogy technique.

Reference

1. Hardin, Paul L and Nussbaum, Daniel A. "Analyses of the Relationship Between Development and Production Costs and Comparisons with Other Related Step-up/Step-down Studies". Naval Center for Cost Analysis, January 7, 1994, page 13.

Applications and Questions:

WRAP RATES:

13.1 "Wrap Rate total costs" equals the sum of what three costs?

13.2 The labor rate which includes *all* of the contractor costs and hours necessary to complete the task is called the _____ _____.

13.3 The Direct Labor Rate is influenced by what four factors?

13.4 Most firms collect indirect costs in aggregate cost accounts called _____ _____?

13.5 Given the following FY13$ costs in your program, calculate the contractor's Fully Burdened Labor Rate per hour:

- Direct Labor Wage Rate: $45 per hour

- Overhead rate is 130% of Direct Labor dollars

- "Other Cost" rate is 20% of (Direct Labor dollars + Overhead dollars)

13.6 Using the Fully Burdened Labor Rate per hour found in (5), calculate the total labor cost to your program given the following information:

- Expected contractor support is 80 man-months

- One man-month equals 160 man-hours

STEP-DOWN FUNCTIONS:

13.7 Prototypes generally use unit theory learning curves to determine the cost of each prototype and also their total costs.(True/False)

13.8 Instead of lot costs, what two costs are generally associated with prototypes?

13.9 The ratio of the production phase first unit cost to the prototype average unit cost is known as?

13.10 Your program had seven prototypes that cost a total of $450,000. What was the AUC for your prototypes?

13.11 If the first unit production cost in your program is $35,000, what was the step-down and the step-down factor from your answer in question 13.10?

Cost Factors and the Analogy Technique

14.1 Introduction

In this chapter, we will leave the topic of learning curves and discuss other methodologies and techniques that cost estimators can use to derive necessary costs. Until this point, we have regarded cost estimating relationships (CERs) as complex equations with a number of independent variables, similar to the regressions that we have covered already. However, a CER can be as simple as a ratio between two variables. A CER in which cost is directly proportional to a single independent variable is known as a *cost factor*. In the first half of this chapter, we will discuss how cost factors can be both calculated and utilized and highlight their importance in the field of cost estimation. In the second half of this chapter, we will discuss the Analogy Technique for estimating necessary costs.

14.2 Cost Factors Scenario

Your work supervisor requests that you attend a two-day conference in a city that is 150 miles from your hometown. Since the conference is not too far away, you drive your own personal vehicle (instead of flying or renting a car) to the work site and return at the conclusion of the conference. When you file for your travel claim, your employer will generally reimburse you for the number of miles that you drove, plus possible per diem for food. But considering just the driving portion of your trip, let us assume that your finance folks will reimburse you at $0.52 per mile for POL (Petroleum, Oil and Lubrication) costs. Round-trip mileage for the trip turned out to be exactly 300 miles. Thus, your reimbursement would be 300 miles × $0.52 per mile = $156. The $0.52 per mile is a *cost factor*.

In our example, you have Travel costs = (Miles traveled) × (POL factor) where the POL factor = 0.52. A cost factor can be expressed as either a ratio or a percentage and is

used as a multiplier of an independent variable (in this case, miles traveled). It is assumed that the relationship is *linear*, and that the y-intercept is zero. Therefore, your reimbursement is actually an equation of a line, where $Y = m \times x + b$, and the intercept $b = 0$, so $Y = m \times x$. In this case, the slope of the line $m = 0.52$.

Other examples include:

- Systems Engineering/Program Management (SE/PM) Costs = Recurring hardware (R-HW) × (SE/PM Factor)
- Software costs = (Lines of code) × (Software factor)

The key in developing a cost factor is identifying the primary cost drivers:

- Miles traveled *determines* the total amount of POL cost
- Recurring hardware costs *determines* the total amount of SE/PM cost
- Number of lines of software code *determines* the total amount of software cost

Cost factors can be historical or forward looking. Examples include:

- Tooling costs for the Joint Strike Fighter are expected to be 10% of the recurring hardware costs in that program. (forward looking)
- Systems Engineering/Program Management (SE/PM) costs were found to be 40% of the Recurring Hardware costs in the Arleigh Burke-class Destroyer program. (historical)
- Initial spares and repair parts were calculated to average 15% of the aircraft systems costs on five helicopter programs. (historical)

During data collection, ensure that you collect *actual* cost data and not *budgeted* data. There is usually no need to normalize for inflation, unless an individual data point spans multiple years, since the cost factor is represented as a percentage. You still, however, may need to normalize for quantity (i.e., T1), especially if working with recurring hardware costs on a learning curve. Let's illustrate this technique with the following example:

14.3 Cost Factors

Example 14.1 Suppose we want to estimate the Systems Engineering/Program Management (SE/PM) cost for the new APG-700 radar. We are currently in the initial developmental phase of acquisition and have no detailed description of the SE/PM costs other than that the APG-700 will be similar to six previous systems. For this example, let's assume that the six analogous systems are the APG-100, APG-200, APG-300, APG-400, APG-500, and the APG-600.

Consultation with technical experts has led us to believe that the SE/PM cost is driven by the *recurring hardware costs* of the system. This means the experts feel that there will be a fairly constant cost factor (or constant percentage) of SE/PM to recurring hardware (R-HW) from program to program and recurring hardware will be the basis (or denominator) for comparison.

Developmental cost data was collected on the six radar systems (the APG-100 to the APG-600). Table 14.1 is the data on the first of these systems, the APG-100.

TABLE 14.1 Work Breakdown Structure for the APG-100 in Example 14.1

Item	Budgeted Cost of Work Performed (BCWP)	Latest Revised Estimate ($K)
Recurring hardware	**102555**	**132900**
Central processor	15600	17400
Peripheral subprocessors	5205	5705
Antenna subsystem	35900	38900
Other subsystems	6450	6800
System software	62900	65750
Integration, assembly and test	9500	9970
Platform integration	5010	5200
SE/PM	**8205**	**10635**
System test and evaluation	3090	3200
Training	5250	5450
Data	1500	1650
Peculiar support equipment	9580	13560
TOTAL	270745	317120

Based on the bolded data in Table 14.1, and using the latest revised estimate (LRE) column of costs, the SE/PM cost factor is calculated as follows:

$$\text{SE/PM factor} = \frac{\text{SE/PM cost}}{\text{R-HW cost}} \times 100$$

Thus, in the first analogous system (the APG-100), the SE/PM factor $= 10,635 / 132,900 = 0.08$, using the recurring hardware costs as the basis for our comparison.

There were a total of six analogous systems, and we just calculated that the APG-100 had an SE/PM to R-HW factor of 8%. Suppose the APG-200 was calculated to have an SE/PM to R-HW factor/ratio of 5.7%. Then, the APG-300 was found to have a ratio of 9.1%. We do this same ratio on the data for all six systems. As a summary, the six analogous ratios are captured in Table 14.2 for Example 14.1:

TABLE 14.2 Data Set #1 of SE/PM Ratios for All Six Analogous Systems

System	PSE/R-HW Factor (%)
APG-100	8.0
APG-200	5.7
APG-300	9.1
APG-400	6.5
APG-500	7.7
APG-600	8.2

After gathering the data and looking at this data set, the question we must now answer is: "Were the technical experts correct? Do we agree with their initial assessment that the SE/PM costs are driven by recurring-hardware?"

A rephrasing of this previous question is: "Do we like this data set? Why or why not?" Leveraging the knowledge gained in the Statistics chapter, we perform Descriptive Statistics on this data set and find the following values:

Mean: 7.53%

Variance: 1.51%

Standard Deviation: 1.23%

Coefficient of Variation: 16.33%

Range: 3.4% (min = 5.7%, max = 9.1%)

It is safe to conclude that Table 14.2 is indeed a good data set and that the ratios are very consistent, since the standard deviation about the mean is just 1.23%, the CV is only 16.33%, and the range of the data (from minimum to maximum value) is only 3.4%. Thus we would agree that the technical experts were correct when assuming that SE/PM costs were driven by (or a consistent factor of) recurring hardware costs. Ultimately, you would feel confident in the ratio and costs found in the APG-700 between SE/PM and R-HW costs. But what if the results of the six analogous systems observed are the values found in Table 14.3 instead? Would your conclusion change or remain the same?

TABLE 14.3 Data Set #2 of SE/PM
Ratios for All Six Analogous Systems

System	PSE/R-HW Factor (%)
APG-100	8.0
APG-200	15.7
APG-300	3.1
APG-400	12.5
APG-500	7.7
APG-600	18.2

The Descriptive Statistics on the Table 14.3 data set are as follows:

Mean: 10.87%

Variance: 31.67%

Standard Deviation: 5.63%

Coefficient of Variation: 51.79%

Range: 15.1% (min = 3.1%, max = 18.2%)

Do you like this data set? Clearly, this is not a very good data set as the standard deviation is very high relative to the mean at 5.63%, the CV = 51.79%, and the range of the data is large as well. With this data set, you would have very little confidence in whether the cost ratio in the APG-700 was going to be low like the 3.1% or high like the 18.1%.

The data sets in Tables 14.2 and 14.3 are very different data sets that would foster opposite conclusions. And your conclusion, in this case, would be that the technical experts were *incorrect* in their assumption concerning SE/PM and R-HW if using Table 14.3 data. If this occurs, you can choose to query the technical experts again and see if they may have another opinion or option for you to explore, or you may need to attempt to derive a cost factor on your own.

14.4 Which Factor to Use?

Let's go back to the Table 14.2 data set found in Example 14.1. If you did encounter results like this, you would be happy to use the data. But what value would you ultimately use for your cost factor?

System	PSE/R-HW Factor (%)
APG-100	8.0
APG-200	5.7
APG-300	9.1
APG-400	6.5
APG-500	7.7
APG-600	8.2

- Would you use the average of all of the percentages for your cost factor? (7.53%)
- Would you select an individual factor, perhaps from the system that you felt was most analogous? (for example, perhaps the 6th system, the APG-600, 8.2%?)
- Would you average the last two factors of 7.7% and 8.2%, since they were the two most recent systems? (i.e., 7.95%)
- Maybe you would use the mean plus one standard deviation? (7.53% + 1.23% = 8.76%)

There is no correct answer as to which cost factor you should use. You will need to exercise your own judgment, or the collective judgment of your team, in choosing one of the aforementioned methods or even derive another method/metric not mentioned here. You will just need to defend your assumptions on how you derived that cost factor. Remember, you will not prove that your answer is correct, but you want to show that your answer is *reasonable and credible* and based on sound mathematical methods and reasoning.

14.5 Cost Factors Handbooks

Sometimes, it is not necessary to calculate your own cost factors. Each military service has its own Cost Factors Handbooks, and numerous organizations have derived their own for internal use, as well. These handbooks are used to estimate a variety of costs and cost activities associated with that military service or those necessary at a specific organization. Cost

factors are used to calculate costs in a myriad of categories. For example, when discussing military personnel costs, numerous cost categories are associated within the personnel category, such as active duty, retirement, special pay, travel allowances, etc. Cost factors are also used in support costs, training costs, logistics, materiel, facilities construction, and production factors, to name just a few.

An example would be calculating the cost of a new aircraft hangar. An appropriate cost factor that you could find might be $275 per square foot. So if you knew that your aircraft hangar would be 2,000 square feet, then its cost would be calculated to be $2,000$ square feet $\times$ 275 per square foot $= \$550,000$.

There are numerous examples of Cost Factors handbooks. The following list provides a sample of what your service or organization has available for you. Please check with your supervisor to see what cost factors handbooks are being used and required by your organization, if any. An example list of cost factors handbooks include:

- The Marine Corps Cost Factors Manual
- The US Army Cost Analysis Handbook
- The DoD Inflation Handbook
- Air Force Cost and Planning Factors
- The Historical Air Force Construction Cost Handbook

As mentioned, some organizations within the services have derived their own cost factors handbooks, as well:

- The Acquisition Support Cost Factors and Estimating Relationships Handbook, Electronic Systems Center/Acquisition Cost Division, Hanscom AFB
- The Military Sealift Command's Port Engineer Cost Estimating Guide

These handbooks will be updated every few years to keep in step with the recent effects of inflation.

Cost factors may be a single point estimate, but they may also include a mean, median, standard deviation, and range of the data, from the lowest to highest value. This is certainly more desirable information, if available, to help understand the possible bounds of the range of your uncertainty.

14.6 Unified Facilities Criteria (UFC)

The "Unified Facilities Criteria (UFC)" is known as the "DoD Facilities Pricing Guide," and is *the* pricing guide used by the primary services when the costing of building facilities is necessary. They are used extensively during Military Construction (MILCON)-type and engineering-type projects. This type of costing is significantly different than the costing used in the production of military weapon systems, which uses learning curves when numerous systems will be built. When developing a facility, it is usually a single, unique system that is being developed; perhaps it is a new gymnasium at your organization or a new gasoline facility at the airbase. The UFC's were previously called the "DoD Facilities Cost Factors Handbook." The UFC program is administered, signed, and endorsed by four service entities: HQ, US Army Corps of Engineers; the Naval Facilities Engineering

Command (NAVFAC); the office of the Air Force civil engineer; and the Department of Defense [1].

The Unified Facilities Criteria were formed in 2002 when the Department of Defense (DoD) and the military services initiated a program to unify all technical criteria and standards pertaining to planning, design, construction, and operation and maintenance of real property facilities. The objective of the Unified Facilities Criteria (UFC) program is to streamline the military criteria system by eliminating duplication of information, increasing reliance on private-sector standards, and creating a more efficient criteria development and publishing process. UFC documents provide planning, design, construction, sustainment, restoration, and modernization criteria for all services and DoD [1].

Both technical publications and guide specifications are part of the UFC program. Previously, each service had its own publishing system resulting in criteria being disseminated in different formats. UFC documents have a uniform format and are identified by a number, such as UFC 1-300-1. Management of the UFC program is by an Engineering Senior Executive Panel comprised of the senior engineer executive from each military service and DoD. If you need additional information on the UFC program, it can be found in Mil Std 3007, "Department of Defense Standard Practice for Unified Facilities Criteria (UFC) and Unified Facilities Guide Specifications (UFGS)." (Ref 1)

14.7 Summary of Cost Factors

In this chapter, we introduced cost factors with the conclusion that sometimes complex CER's are not necessary to be able to estimate desired costs. Instead, we can use cost factors to estimate those costs. An example for calculating a cost factor was provided and included both good and poor cost factors. The results from the first data set were consistent and supported a conclusion about the ratio of SE/PM to R-HW that the technical experts felt would exist. In the second data set, however, the numbers were erratic and inconsistent and we would have rejected the technical experts' opinion concerning the basis for the given ratios. It would then have been necessary to compare the SE/PM costs to a different basis besides recurring hardware.

There are often times when we have limited visibility into the task that we are trying to estimate and sometimes we just need to "fill in a hole" in our estimate. Cost factors can fill this need. Another use of cost factors is as a "sanity check" of the primary estimating methodology. All of the military services and numerous organizations have their own set of cost factors for use in their field.

The Cost Factor methodology has sometimes been criticized for its simplicity. But, remember that all else being equal, using the simplest technique available is not always a bad thing. Moreover, remember that this is but *one* of several tools in the cost estimator's toolbox. In the second half of this chapter, we will learn another new technique that ties in well with cost factors: the Analogy technique.

14.8 Introduction to the Analogy Technique

Having just learned that a cost factor is merely a CER in which cost is directly proportional to a single independent variable, the second half of this chapter will continue to use cost factor-type processes and statistics to help us make good decisions, as we introduce the

Analogy technique. So far, we have focused on CERs that utilize a number of historical data points from which to make an estimate. But what if we only have *one* data point with which to help predict our cost? The Analogy technique is the method used when only one data point is available. It is characterized by using a single historical data point that serves as the basis for your cost estimate.

14.9 Background of Analogy

What if we are costing a second generation system and only have the first generation system to use for historical perspective? Analogy estimates are usually characterized by use of a *single historical data point* serving as the basis for our cost estimate. As you might imagine, use of this methodology can be considered "risky" because the historical data is too limited to allow any useful statistical analysis.

The analogy is an estimate based on a relative scaling of a historical data point:

$$\text{New program cost} = (\text{scaling factor}) \times (\text{historical program cost})$$

The scaling factor should not simply be a point estimate, but rather a "most likely" range, to provide uncertainty information, if at all possible. The analogy technique is most useful when the new system is primarily a new combination of existing subsystems for which recent historical cost data are available. It is very useful in early milestone, ill-defined programs, and as a check on estimates used by other methods. It is also the method that you would most likely use if your boss says "I have to see the Admiral in three hours and need a quick "back-of-the-envelope guess-timate" on a new system or a part of that system." Since you will not have the time to do extensive research on databases, the best you will be able to do in the short period of time that you have is to find that one system that most closely resembles (or is the most "analogous" to) the system that you are attempting to cost.

The advantage of the analogy technique is that it is relatively quick to do. Many new programs consist of modified or improved versions of existing components, combined in a new way to meet a new need. In the analogy technique, we break the new system down into components, usually via a work breakdown structure (WBS) that can be compared to similar existing components. The basis for comparison can be in terms of:

- Capabilities
- Size
- Weight
- Reliability
- Material composition, and/or
- A less well-defined, but often used term: *Complexity*.

When development costs and production cost estimates are needed, the analogy technique offers several approaches. Primarily, you can separate development and production estimates, each based on data related specifically to the development and production acquisition phases. You can base your estimate of future costs of the new system on the historical cost of the previous system. In addition, production estimates are based on production

data. You can then use historical production ratio factors to estimate development costs, if needed. This is similar to what we learned in Chapter 13 on Step-Down Functions, though in this case, it would be considered a "step-up" factor instead of a "step-down" factor, since you are now estimating prototype development costs from the production costs, instead of the other way around.

14.10 Methodology

Using the known item's value, apply quantified adjustments to that item which measure the differences between the old and the new system. Good actual data is essential! The historical or analogous system should be similar not only in performance characteristics, but also similar from the standpoint of manufacturing technology. For example, if for some reason I was tasked to determine the cost of a 2014 Lexus, it would be most advantageous if I could find data from the 2013 Lexus of the same model. That way, I would be comparing cars with similar performance characteristics and manufacturing technologies. But what if I were only able to find data for a 1964 Volkswagen Beetle? While they are both automobiles, the performance characteristics and manufacturing technologies would be significantly different, thus making a comparison of the two significantly more subject to disbelief about its relevance.

Questions that are relevant to ask when assessing the relative differences between the old system and the new system include the following:

- How much different is the new system compared to the old system?
- What portion (i.e., how many components) of the old is just like the new?
- What is the factor of complexity between the two systems?

An essential concept to grasp is that analogy is a cost estimating method by which we *assume that our new system will behave "cost-wise" like the historical system*. We will define the new system in terms of design or physical parameters, performance characteristics, and known similar systems between the two.

The historical data that you will need includes what fiscal year (FY$) that the analogous system costs are in to ensure that you are using a single base year for comparison. You will also need to know what the historical system's learning curve was, and also determine the pertinent data categories based on data availability.

14.11 Example 14.1, Part 1: The Historical WBS

Your first step is to develop a WBS with the historical data of the analogous system and fill in the data/costs in that WBS. A sample format WBS for the analogous system might look like the following, as seen in Table 14.4. Note that the Non-Prime Mission Equipment cost is the sum of four subcategories: SE/PM, Training, Data and Spares.

TABLE 14.4   Sample WBS for the Historical System in
Your Analogy

Old system				
Recurring hardware		_____		
Non-recurring hardware		_____		
Non-prime mission equipment		_____		
		SE/PM		_____
		Training		_____
		Data		_____
		Spares		_____
Subtotal			_____	
Overhead			15% x subtotal	
Fee			10% x (subtotal+overhead)	

Now that you have a template for your historical system, input your data into that template. An example of a completed WBS is shown in Table 14.5. These costs were accrued from a total of $N = 500$ units and are in FY10$K.

TABLE 14.5   Completed WBS for the Historical System in Your Analogy

Old System	**All Costs in FY10$K**	
500 Units		
Recurring hardware	549.158	
Non-recurring hardware	218.600	
Non-prime mission equipment	404.400	
SE/PM		180.34
Training		54.65
Data		76.51
Spares		92.9
Subtotal	1172.158	
Overhead: 15% × subtotal	175.824	
Fee: 10% × (subtotal + overhead)	134.798	
Total =	1482.780	

Your *Subtotal* is comprised of the recurring hardware costs plus the nonrecurring hardware costs plus the non-prime mission equipment costs and this total equals $1,172.158 (FY10$K). The non-prime mission equipment total is the sum of the SE/PM, training, data, and spares costs and equals $404.40 (FY10$K). The total cost for the historical system including overhead and the profit/fee was $1,482.78 (FY10$K).

But the key to the Analogy technique is that while we have the *actual* costs in our historical WBS, the *percentages* that these numbers represent is really what is important,

as we will assume that the new system will act "percentage-wise" (for all elements) like the analogous system did. So let's calculate what these percentages are, and for the calculations we will use the recurring hardware (R-HW) costs as our basis (or denominator), since they are always a significant portion of the costs in a program and it tends to be the primary cost driver in most areas (in this example, R-HW = $549.158K). These percentages can be found in Table 14.6:

TABLE 14.6   Percentages of WBS Elements vs. Recurring Hardware

Old System	All Costs in FY10$K	
500 Units		***
Recurring Hardware	549.158	Percentage of R-HW
Non Recurring Hardware	218.600	0.39806
Non-prime mission equipment	404.400	
SE/PM	180.34	0.32839
Training	54.65	0.09952
Data	76.51	0.13932
Spares	92.90	0.16917

Let's examine what Table 14.6 is telling us. If you look in the final column (denoted by ***), you can see that the Nonrecurring Hardware costs (= $218.60k) were 39.806% of what the recurring hardware costs were (= $549.158k). This is calculated as $218.60 / $549.158 = 0.39806. By the same procedure, we find the following:

- SE/PM costs were 32.839% of the recurring hardware costs
- Training costs were 9.952% of the recurring hardware costs
- Data costs were 13.932% of the recurring hardware costs
- Spares costs were 16.917% of the recurring hardware costs

Having calculated these percentages, you are almost done with the WBS of the old and analogous system. But our final step is to calculate the value for the T1 in this historical system. This will be needed in Part 2 of the process, so we might as well calculate it now. To do so, we will use the total lot cost equation that we learned in Chapter 10 on Unit Theory. The equation is reprinted here again as Equation 14.1:

$$\mathrm{CT}_N \cong \frac{A(N)^{b+1}}{b+1} \qquad (14.1)$$

From Table 14.6, we found that the recurring hardware costs = $549.158K. We also know that $N = 500$ units and we found out from historical data that the learning curve was 85%. Therefore our calculations show that $b = \ln(0.85) / \ln(2) = -0.23447$ and $b + 1 = 0.76553$. Rearranging Equation 14.1 and using these inputs and solving for A, we find that $A = \mathrm{T1} = \$3.61$ (FY10$K). This is the T1 for the historical system that we are using as our analogous system.

14.12 Example 14.1, Part 2: The New WBS

Now that we are done with the *historical (or analogous)* WBS, we turn our attention to developing your *new* system WBS template so that it looks similar in content to the historical WBS, as shown in Table 14.7:

TABLE 14.7   Your New WBS Template, Similar to Table 14.1 and the Historical WBS

	New system				
Recurring hardware			_____		
Non-recurring hardware			_____		
Non-prime mission equipment			_____		
		SE/PM		_____	
		Training		_____	
		Data		_____	
		Spares		_____	
Subtotal			_____		
Overhead			_____		
Fee			_____		

In order to fill in the data for your new WBS template from Table 14.7, the key now is for you to determine the recurring hardware costs of the *new* system. To do so, you will most likely need to calculate the cost of your new T1. Once you have that value for T1, you can calculate the total recurring hardware cost for your new system.

In order to calculate the T1 for your new system, Equation 14.2 will need to be used:

$$\text{T1 (new)} = \text{T1 (old)} \times (\text{Complexity factor}) \times (\text{Prod Imp Factor})$$

$$\times (\text{Miniaturization factor}) \tag{14.2}$$

Equation 14.2 quantifies the difference between the historical system's T1 and the new system's T1, and requires you to consider the following three factors:

1. Determine a *Complexity factor* between the historical and new system's recurring hardware:
   - This factor represents the cost ratio due simply to differences in complexity (i.e., the new system is 20% more complex)
   - It is based on design and performance differences.
   - It may require conversations with technical specialists and engineers.

2. Determine a *Miniaturization factor* between the historical and new system's recurring hardware:

- This factor represents the cost ratio due simply to miniaturization
- If it is smaller/larger, then it should cost less/more based on size only

3. Determine a *productivity improvement factor* between the historical and new system's recurring hardware:

- This factor represents the cost ratio due simply to improved productivity
- Are there significant technological improvements?

4. Moreover, consider any *other* improvements that may have been made. Are there any other possible factors, such as "ruggedization" (for military: tools, cots, etc)?

Given these four considerations, during our comparison of the historical system to the new system we determined the following:

- The new system is 20% more complex
- The new system has a 30% productivity improvement
- The new system is 15% smaller in size

Note that these are subjective assessments, so it is essential to get proper input from your technical specialists and engineers. We are now ready for our calculations and since the historical system's T1 was previously determined (in Example 14.1, Part 1) to be T1 = $3.61 (FY10$K), our calculations for the new T1 are as follows:

$$\text{T1 (new)} = \text{T1 (old)} \times (\text{Complexity factor}) \times (\text{Prod Imp factor})$$

$$\times (\text{Miniaturization factor})$$

$$\text{T1(new)} = \$3.61 \times 1.2 \times 1.3 \times 0.85 = \$4.787 \text{ (FY10\$K)}$$

Note that if the new system is 20% more complex or has a 20% productivity improvement, you multiply by 1.2. If it is 20% less complex or 20% smaller in size, you multiply by 0.80. This applies to all three factors. The new T1 = $4.787 is now the T1 (or A) in your total lot cost equation for the new system you are trying to cost.

With the aforementioned calculations complete, we can now determine the recurring hardware costs for your new system, once again using Equation 14.1. Inputs are T1 = $4.787; learning curve = 85% (thus $b = -0.23447$ and $b + 1 = 0.76553$); and $N = 500$ once again. We will assume that the learning curve will remain the same (85%) as the historical system until there is proof to the contrary. Using Equation 14.1 again, we find:

$$\text{CT}_N \cong \frac{A(N)^{b+1}}{b+1} = \frac{(4.787)(500)^{.76553}}{0.76553} = \$728.1883 \text{ (FY10\$K)}$$

Thus, the recurring hardware costs for the new system = $728.1883 (FY10$K). With this total, we can now determine *all* of the other WBS element costs in Table 14.7 using the percentages calculated from the historical system in Table 14.6! Table 14.8 shows the completed WBS for the new system. Each cost was calculated by multiplying the R-HW costs of $728.1883 (FY10$K) times the "percentage of R-HW" for each WBS element

TABLE 14.8   Completed WBS for the New System in Your Analogy

New System		All Costs in FY10$K	
500 Units			***
Recurring hardware	728.1883		Percentage of R-HW
Nonrecurring hardware	289.8653		0.39806
Non-prime mission equipment	536.2376		
SE/PM		239.1323	0.32839
Training		72.4663	0.09952
Data		101.4529	0.13932
Spares		123.1861	0.16917
Subtotal	1554.2913		
Overhead: 15% × subtotal	233.1437		
Fee: 10% × (subtotal + overhead)	178.7435		
Total =	1966.178		

found in the *** column. For example, Nonrecurring Hardware costs equals $728.1883 × 0.39806 = $289.8653 (FY10$K) (with small round off error).

The key to this methodology was finding what the cost of the recurring hardware was. After that, we simply applied the percentages (as seen in the column labeled ***) that we calculated from the historical system and assumed that the new system would behave "cost-wise" like the historical system:

- Nonrecurring Hardware is 39.80% of recurring hardware = $289.865
- SE/PM is 32.84% of recurring hardware = $239.132
- Training is 9.95% of recurring hardware = $72.466
- Data is 13.93% of recurring hardware = $101.453
- Spares is 16.92% of recurring hardware = $123.186

Once you add these costs up, you are done! The advantages of the Analogy technique is that it is quick, it is easy, and it takes advantage of historical cost behavior. The disadvantages are that there is no knowledge of uncertainty and it is a point estimate only! (i.e., $1,966.178 FY10$K +/− ??)

14.13 Summary of the Analogy Technique

In the second half of this chapter, we introduced the Analogy technique for costing. It is characterized by using a single historical data point that serves as the basis for our cost estimate. Just one data point may be available if there is only one historical program completed prior to your program or when time does not permit you any other option. Analogies and cost factors are very similar in that they both use statistics, ratios and percentages that are compared to other systems. Use of this methodology can be considered "risky" because the historical data is too limited to allow any useful statistical analysis. After gathering

your data, it is up to the cost estimator to properly apply the ratios and/or percentages calculated. This was shown in Example 14.1 provided in Sections 14.11 and 14.12. In the next chapter, we will commence a completely different topic, as we begin to discuss software cost estimation.

Reference

1. Whole Building Design Guide (WBDG) DOD website, Unified Facilities Criteria Guide http://www.wbdg.org/references/pa_dod.php)

Applications and Questions:

COST FACTORS:

14.1 A cost factor is a CER in which cost is directly proportional to just _____ independent variable.

14.2 Cost Factors can be either historical or forward looking (True/False)

14.3 The pricing guide used by the primary services when costing out engineering and military construction projects is called the _____ _____.

14.4 You are tasked with determining the tooling costs on a new system. Technical experts are estimating that these tooling costs should be approximately 33% of its prototype manufacturing costs. You were able to find two similar systems as your historical data. Prototype manufacturing costs for System #1 was $166,190 and System #2 had prototype manufacturing costs of $119,110. The tooling costs for these two systems were $39,730 and $24,960, respectively. All costs are in FY09$.

   Task: Is 33% a pretty good estimate for this? Explain if you agree or disagree with the tech experts and support your answer mathematically. If you do not agree, what would your estimate be and why?

ANALOGY TECHNIQUE:

14.5 The analogy technique is the method used when only _____ historical system point is available for comparison.

14.6 When using the analogy technique, we assume that our new system will behave "cost-wise" like the historical/analogous system. (True/False)

14.7 In the analogy technique, we break the new system down into components, usually via a WBS, that can be compared to similar existing components. The basis for comparison between the two systems can be in terms of (name 4 of the 6):

14.8 When calculating your new T1, it is necessary to quantify differences between the historical program and the new system with three distinct factors. Name those three factors:

14.9 After completing the historical system's WBS with its historical costs, the key is to calculate the_____ of those costs against the recurring hardware costs. These will be used to calculate the costs in the new system's WBS.

Software Cost Estimation

15.1 Introduction

In this chapter, we will discuss an entirely new topic: software cost estimation. This chapter is intended to be just an overview of this unique area of cost estimation, as costing software development is a very complex task, and numerous articles and books have been written on the topic. It is an area of development that is very hard to estimate accurately. One of the many unique aspects of software cost estimation include only having development and maintenance costs, with no costs accrued in a production phase as we see in hardware cost estimation. There are no learning curves associated with software development. Moreover, estimating software cost is accomplished by the size of the effort and not using independent variables such as power, weight, or frequency to determine the cost of the software used in the new system. The intended take-away from this chapter is that *estimating the cost for software development is extremely difficult.*

15.2 Background on Software Cost Estimation

There are many areas that need to be considered when attempting to determine the cost of a new software developmental project, including broad categories such as the attributes of the product, the platform, the personnel, and the project. A finite list of an almost infinite number of possible tasks includes the following: [1]

- Making investment or other financial decisions involving a software development effort
- Setting project budgets and schedules as a basis for planning and control
- Deciding on or negotiating tradeoffs among software cost, schedule, functionality, performance, or quality factors
- Making software cost and schedule risk management decisions
- Deciding which parts of a software system to develop, reuse, lease, or purchase
- Making legacy software inventory decisions: what parts to modify, phase out, or outsource

---

- Setting mixed investment strategies to improve organization's software capability, via reuse, tools, process maturity, outsourcing, etc.
- Deciding how to implement a process improvement strategy

Any need to answer any of the aforementioned tasks at your workplace would require significant training and a textbook or manual specifically discussing techniques used in software cost estimating. Realizing how many tasks are possible using software, we will begin your familiarization with the subject by defining and discussing the following seven areas of interest in software:

1. What is Software?
2. What are the work breakdown structure elements (WBSE) in a typical software cost estimating task?
3. Software costing characteristics and concerns
4. Measuring software size: source lines of code (SLOC) and function points (FP)
5. The software cost estimating process
6. Problems with software cost estimating: cost growth
7. Commercial software availability

15.3 What is Software?

Software usually refers to one of four premises/domains:

- A generic term for *computer programs*, including systems programs, which operate the computer itself.
- The instructions that tell the computer what to do with *application programs,* which control the particular task at hand.
- A set of *advanced computer modules* that allow the user to plan efficient surveys, organize, and acquire satellite navigation data, verify and download data, process and analyze measurements, perform network adjustments, and report and archive final results.
- The entire set of programs, procedures, and related *documentation* associated with a computer system

But overall, we can generalize/summarize that there are two major categories of software:

- "System Software" needed to run and operate your computer. Examples include software such as Windows 7 or Mac OS or Linux.
- "Application Software" needed to accomplish your tasks. Examples include software such as Microsoft Excel/Word/Power Point, "Recycle" (the sample editor for a Mac), or JMP

15.4 The WBS Elements in a Typical Software Cost Estimating Task

Having defined and given examples of existing software, we will discuss the work break-down structures elements in a typical software cost estimate. These will of course vary from program to program, but a typical WBS can be broken down into the following five categories, as shown in Figure 15.1:
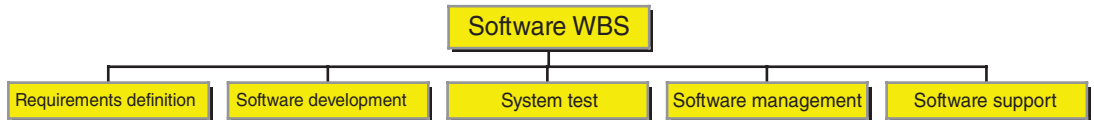
FIGURE 15.1 **A Typical WBS for Software Cost Estimation.**

These five categories typically consist of the following subcategories:

- *Requirements definition (RD):* This WBSE typically consists of the following four areas:
  - Interface definition
  - Requirement specification
  - Operating concept development
  - Algorithm development
- *Software development (SD)*: This WBSE typically consists of the following four areas:
  - Requirements analysis
  - Architecture development
  - Design, code, and unit test software
  - Integration and testing
- *System test (ST)*: This WBSE typically consists of the following four areas:
  - Test planning
  - Test development
  - Integration and testing
  - Acceptance testing
- *Software management (SM)*: This WBSE typically consists of the following four areas:
  - Project management
  - Supplier management
  - Personnel management
  - Team building
- *Software support (SS)*: This final WBSE typically consists of the following five areas:
  - Configuration management
  - Quality assurance
  - Software environment readiness

- Test benches
- Security and network administration

Throughout this textbook, we have discussed in great detail that hardware cost estimation has three phases in its life cycle: Research and Development; Production/Investment; and Operating and Support costs. However, in software cost estimation, there are only *two* phases to the software life cycle: (1) Development costs and (2) Maintenance costs. The reason for this is explained in the next section.

15.5 Software Costing Characteristics and Concerns

The following paragraphs cover software metrics and topics of concern while considering the cost of software:

- The standard cost estimating means used to estimate the development cost for hardware programs are not applicable to estimating software programs. There are only research and testing costs. This is because software development tasks are all *Nonrecurring Development costs!* The largest cost occurs of course in the development phase. Once the software is developed and tested and is working properly – all accomplished in the developmental phase – then it is merely the task to make the appropriate number of copies of that software for the number of items being produced. This is why there are no costs accrued in, or attributed to, a production phase. There are also no learning curve theories that can be applied, either, as it is a one-time development.

- *Programming and coding is the Easy Part!* Figuring out what the software solution is for any technical problem encountered, or what software language should be used for a particular task, is what is most difficult. This is the task for a good software engineer.

- Software requirements cannot be fully captured in any finite list. The true list of requirements is virtually infinite, and these requirements will generally change significantly over the course of a program.

- There are no "technical" characteristics such as Weight, Power, or Frequency that are utilized in hardware cost estimating that play the role of "Cost Driver" in a software program. The primary *measurable* cost driver is the *size* of the software program. The two main ways to cost via size are "Number of Lines of Code" and "Function Points." These will be discussed in the next section.

- Personnel tasks in software development are *uniquely* personnel-intensive. Within the same company or work group, productivity (measured, e.g., by number of lines of code completed in a given time) will vary greatly among programmers due to individual skills.

- A software engineers' self-esteem and traditional optimism generally can cause them to underestimate how much code will be needed. This is significant because it affects not only the number of lines of code that needs to be produced, but also affects the timeline to complete the development of that software.

- The initial delivered code – "right out of the box" – often performs inadequately and fundamental modification costs are often prohibitive, thus greatly increasing the cost.

- Hardware deficiencies that cannot be fixed for various reasons during the later stages of a project are circumvented by re-tasking software, thus generating new development costs and adding to maintenance costs as these modifications are made.

We will now discuss the important topics of source lines of codes and function points.

15.6 Measuring Software Size: Source Lines of Code (SLOC) and Function Points (FP)

As previously mentioned, there are no "technical" characteristics such as Weight, Power, or Frequency (as utilized in hardware cost estimating) that play the role of "cost driver" in a software program. The primary measurable cost driver is the *size* of the software program. How big is the software application or the system that is being evaluated? Current sizing of a software program includes the following two methods:

- *Source lines of code (SLOC)*
  - SLOC is the oldest and most widely used method for software cost estimating
  - Large programs are sized in K-SLOC (thousands)
- *Function points (FP)*
  - Estimates size based on user-defined functionality
  - Function point theory was established in late 1970s

15.6.1 SOURCE LINES OF CODE: (SLOC)

There are numerous arguments for and against using SLOC as a means to determine the cost of software. The arguments *for* using SLOC include:

- It is easy to find how many lines of code there are in a program, after it is written. Then, to estimate the cost, you simply multiply the number of lines of code in that program times the cost per line of code. Simple! Plus . . . . .
- There is plenty of historical data available for comparisons and analogies. Plus . . . . .
- SLOC is supported by most commercial cost estimating tools (to be discussed in Section 15.9)

However, there are also arguments *against* using SLOC, as well. These arguments include:

- *There is no standard as to what counts as a line of code*. Moreover, the number of lines of code is heavily dependent on (1) the language in which the software is written and (2) the programmer who is writing the software.
- *SLOC rewards inefficient design* and penalizes tight, efficient design. For example, if you are paid by the number of lines of code, why write efficient coding in your program in 1,000 lines of code, when you will make three times more money for writing the same exact program in 3,000 lines of code! Thus, there is *little to no*

incentive for a software developer to reduce the number of lines of code in a program, since they will make less money for being efficient.

- *Cross-language estimates are inconsistent.* What it takes to perform a task in one language may take another language significantly more/fewer lines of code to accomplish. Large programs can utilize *many* different languages, since the strength of one language is to perform a certain function well, while another language can perform another function better.

Example 15.1 Cross-language estimates are inconsistent   To illustrate the point that "cross-language estimates are inconsistent," let us consider the following example. A particular task was written in three different software languages: Macro Assembly, Ada 83 and C++. All three tasks had the identical outcome regardless of the language used. In analyzing each task by its coding software language, Macro Assembly is the oldest/earliest language of the three, followed in order by Ada 83 and then C++. Details, such as the source lines of code and the number of months required to perform each task were tabulated and the results are found in Table 15.1:

TABLE 15.1   Results of the Identical Task Using Three Different Software Languages

	Macro Assembly	Ada 83	C++
SLOC required	10,000	3,500	2,500
Effort per activity, staff months			
Requirements	1.0	1.0	1.0
Design	3.0	2.0	0.5
Coding	5.0	1.5	1.0
Testing	4.0	1.5	1.0
Documentation	2.0	2.0	2.0
Management	2.0	1.0	0.5
Total project, months	18.0	10.0	7.5
Total project, SLOC per staff month	556	350	333

Analyzing Table 15.1, we find the following:

- Macro Assembly required 10,000 lines of code to complete the task
- Ada 83 required 3,500 lines of code to complete the task
- C++ required 2,500 lines of code to complete the task
- Macro Assembly required 18 months to complete the task
- Ada 83 required 10 months to complete the task
- C++ required 7.5 months to complete the task

But while the number of lines of code required and the number of months required to complete the task both favor C++, note the following metric:

- Macro Assembly completed 556 lines of code per month ($10,000 / 18 = 556$)

- Ada 83 completed 350 lines of code per month
- C++ completed "only" 333 lines of code per month

Note that these metrics are thus moving in the "wrong" directions when analyzed against modern, high-level languages. It appears that Macro Assembly is accomplishing more per month, yet it took much longer to accomplish the task in that language! Moreover, if you were to pay each developer by the number of source lines of code, the company using Macro Assembly for its program would be paid significantly more than the other two companies. This illustrates that if the developer was evaluated merely on its SLOC written per month, then the company using a higher level language would be penalized for being more efficient, as it appears that they are doing "less" each month. This example highlights the SLOC paradox confronting the software cost estimator.

In conclusion, however, while there are three significant arguments against using source lines of code to determine the size of a program, *it is nevertheless a better-understood and better-defined cost driver than any other*. A second way to evaluate the size of a program is by using Function Points.

15.6.2 FUNCTION POINT (FP) ANALYSIS

FP analysis is an alternate method for using SLOC to measure the size of a software program. Rather than counting the number of lines of code used to perform a function, function points are the weighted sums of five different factors that relate to user requirements. These five factors include:

- Internal logical files
- External interface files
- External inputs
- External outputs
- External inquiries

These five factors perform the following functions:

- *Internal logical files (ILF)*: These include logical groupings of data in a system, maintained by an end user (i.e., databases and directories)
- *External interface files (EIF)*: These include groupings of data from another system used only for reference purposes (i.e., shared databases or shared mathematical routines)
- *External inputs (EI)*: These involve unique data or control inputs that cross the system boundary and cause processing to occur (i.e., input screens and tables)
- *External outputs (EO)*: These involve unique data or control outputs that cross the system boundary after processing has occurred (i.e., output screens and reports)
- *External inquiries (EQ)*: These involve unique transactions that cross the system boundary to make active demands on the system via direct retrieval of information contained on the files (i.e., prompts and interrupts)

To summarize these five functions, the following overview of the functions is provided:

- The software for the system queries and "reaches out" for data
- When it receives the data, it keeps it, stores it, or performs a task
- All tasks accomplished will be classified as one of the five mentioned function points; thus, some will be *internal logical files* tasks; some will be *external interface files* tasks, etc
- One of the internal systems will add up how many times each of these five functions are being performed and it will calculate a total sum for each function
- The cost of each of these five functions must then be determined and multiplied by the number of times each function was performed, to get total cost

According to some analysts, FPs may be more useful as a metric, as it offers a definitive way to "standardize" software costing. However, it has not been used as often as SLOC as a cost driver, so there is less historical experience regarding how well it works. In response, some commercial models are producing "conversion factors;" that is, they are converting function points into an equivalent number of lines of code for various languages.

Some cost estimation models include expert judgment models that estimate SLOC based on the opinions of one or more experts. Parametric models, like regression, use inputs consisting of numerical or descriptive values to compute program size. Estimates are developed through mathematical formulas that use statistical relationships between the size and software characteristics. Additional information on Function Points can be found in [2].

15.7 The Software Cost Estimating Process

This process can be summarized as following these general guidelines to establish a rough order of magnitude (ROM) cost of the software program:

1. Estimate the project size, either via source lines of code or via function points.
2. Develop productivity measures. These will most likely include how many lines of code or function points that you can accomplish in a month:
   a. # SLOC/month
   b. # FP/month
3. Estimate the schedule in "person-months" or "man-months" to determine the amount of months or years it will take to complete the task.
4. Apply the labor rates of your company of the workers, such as $30/hour.
5. Estimate the costs for the program by multiplying the salary per hour of your workers with the number of hours necessary to complete the project. This will include both total costs and time-phased costs.

Following this general process will provide a ROM estimate on the cost of your software program. However, developing credible and defensible estimates in steps 1 and 2 are always the difficult parts of the process.

15.8 Problems with Software Cost Estimating: Cost Growth

In Section 15.5, we discussed that occasionally a software engineers' self-esteem and traditional optimism can cause them to underestimate how much code will be needed in a program. This is significant because it affects not only the number of lines of code that needs to be produced, but also ultimately affects the timeline to complete that software. But certainly there are *many* reasons why a software program may experience significant growth. The following is a scenario that demonstrates how cost growth can arise in software cost estimating:

Example 15.2 Cost Growth Example   *Original Assumptions*

1. The contractor estimates that it should take 100,000 lines of code to develop the new program.
2. The contractor also estimates that their programmer productivity can produce approximately 3,600 lines of code per developer-month.
3. Therefore, this would result in a development program of $100,000 / 3,600 = 27.77$ months.

Unfortunately, in this project, the original assumptions did not hold. Instead, the following occurred:

1. The number of lines of code grew from the estimated 100,000 to 175,000.
2. Productivity inevitably slipped from 3,600 lines per developer-month to 2,600 lines per month instead, as the project moved forward.
3. Thus, the number of lines of code grew by a factor of 1.75, and.....
4. The length of time for coding grew by $3,600 / 2,600 = 1.38$, and so … …
5. Total growth for this software program became 1.75 more lines of code × 1.38 more time for coding = 2.42.

Thus, the estimated cost grew by a factor of 2.42 times the original cost estimate!

Figure 15.2 is a graphical depiction of four historical programs from one sector of production that utilizes a significant amount of software. The lighter columns on the left represent the estimated source lines of code for projects 1 though 4 at the beginning of the program. The darker columns on the right display the actual number of lines of code it took to complete each project. As can be calculated from the lines of code displayed, the number of lines of code in these programs increased by a factor of 1.99, 2.63, 3.32, and 2.0, respectively.

Table 15.2 also represents data on historical programs and it displays the difference between the estimated source lines of codes and the source lines of codes required upon completion for sixteen US Air Force and US Navy projects from a number of different type programs. The last column represents the growth factor in SLOC that occurred in each program. Only two of these programs exceeded a cost growth factor of 2.0.
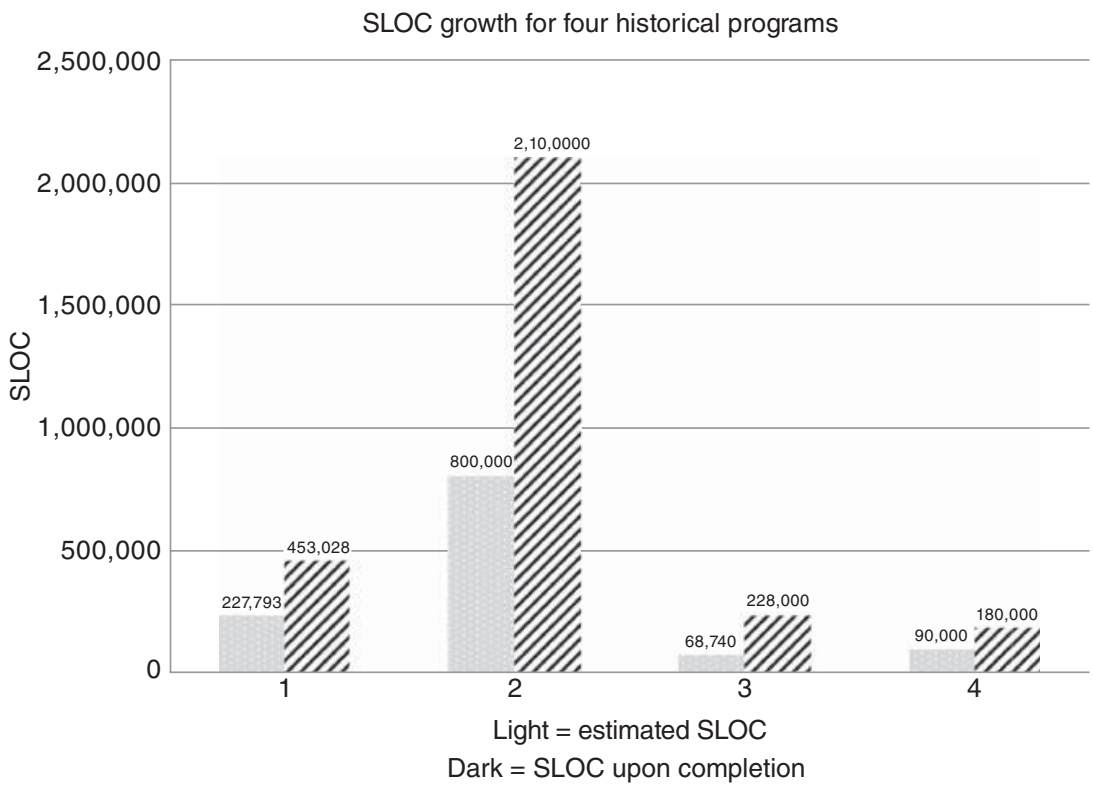
FIGURE 15.2 Software Growth Experience in Four Historical Programs.

TABLE 15.2   Estimated SLOC vs. Actual SLOC at Program Completion

Project	Estimated SLOC	Actual SLOC Upon Completion	Growth Factor
1	618,000	709,000	1.15
2	23,599	25,814	1.09
3	14,000	70,143	5.01
4	41,800	46,303	1.11
5	45,000	45,000	1.00
6	39,294	119,400	3.04
7	22,000	30,000	1.36
8	15,500	26,513	1.71
9	100,000	122,000	1.22
10	532,000	877,129	1.65
11	206,650	394,309	1.91
12	74,000	82,930	1.12
13	213,800	261,800	1.22
14	153,000	185,000	1.21
15	83,900	108,850	1.30
16	1,246,272	1,272,200	1.02

15.9 Commercial Software Availability

In the "old days," no one knew how to estimate software costs. But, where there is a need in a free enterprise system, products will come to market to satisfy a need! For software cost estimating, some such products are PRICE-S, SEER-SEM, and COCOMO, among others.

A problem arises, however, when it comes to the inputs that are required by the software estimating products. While there are manuals and documents that will fully explain what each of the inputs mean and how to use them to build an input file for the products, some products are proprietary "Black Boxes" into whose workings the analyst has little to no insight. In those cases, the problem that this creates is that we not only do not know *how* the software program works, but also cannot fully explain to anyone else why our cost estimate is what it is. This is discomforting when you have to brief your boss to rationalize your costs, or perhaps even brief Congress. All you can do is display your set of inputs, but you will be unable to explain precisely how the inputs affect the estimate. Sensitivity Analysis can help, but even that will still cause substantial speculation on your part. Certainly, the equations that they develop will employ statistically correct methodologies, but the black box syndrome creates a vacuum, in that the population data and resulting estimating equations are not revealed to the end user. Therefore, when using a software cost estimating package, you are in essence "buying" the parametric estimating equations to be used.

The following is a sampler of models that can be used for software cost estimating:

- Gartner TCO Manager: Total Cost of Ownership for computer systems
- SEER-SEM: Software development cost estimating
- COCOMO 2000: Software development cost estimating
- True Planning: Estimates the scope, cost, effort, and schedule for software projects
- Revised Intermediate COCOMO (REVIC): Estimates the cost of software development projects
- Software life cycle management (SLIM): Describes the *time* and *effort* required to finish a software project of specified *size*.
- Automated Cost Estimating Integrated Tools (ACEIT): an integrated suite of analysis tools for the desktop to automate cost analysis, to include software cost estimation. Produced by Tecolote Research, Inc.

The software package called COCOMO ("Constructive Cost Model") was developed by Barry W. Boehm and is hosted on the website at the University of Southern California (USC) School of Software Engineering and offered free of charge. COCOMO is a model that allows one to estimate the cost, effort, and schedule when planning a new software development activity and can be found at the following URL: [1] (or search for COCOMO + USC)

http://sunset.usc.edu/csse/research/COCOMOII/cocomo_main.html

Earlier, one difficulty identified in Section 15.6.1 is that there is no standard to tell us what constitutes a line of code. Is a line of text describing what the next section of coding

will accomplish considered a line of code? To assist in this area, a software program devised within COCOMO assists in attempting to determine how many lines of code are actually in a program that you are reviewing. This program can be found at the following URL: (or search for CODECOUNT + USC)

http://csse.usc.edu/research/CODECOUNT/

15.9.1 COTS IN THE SOFTWARE ENVIRONMENT

Commercial-off-the-shelf (COTS) models/tools are available that can provide CERs and parametric equations for various types of cost estimating. The common thought while estimating program costs in the *hardware* environment is that it is cheaper to use COTS than it is to create a program from scratch, and this is generally true. But does this assumption hold true in the software environment, as well?

Findings by The MITRE Corporation conclude that there are many risks associated with COTS-related software products. The conclusions from their study on a COTS software product include [3]:

- It will not be reliable
- It will not meet response time requirements
- It will consume too many resources
- COTS integration efforts are less well understood than software development efforts
- COTS interfaces are often not well described in the documentation
- That interfaces between COTS products and other pieces of software may need rewriting each time the product is upgraded
- That all interfaces must be tested throughout the upgrade cycle of each COTS product

MITRE also found that debugging is difficult due to the "Black Box" situation since:

- The user can make inferences about the product *only* by observing its behavior
- If vendor support is even available, vendors will tend to "blame" each other when compatibility problems arise
- The "Commercial Rush to Market" means that the end users become the testers!

Therefore, The MITRE Corporation study recommends that performance expectations *be scaled back* if COTS must be accommodated in a program.

15.10 Post Development Software Maintenance Costs

While hardware programs generally have a life cycle of 15–20 years (and many times much longer than that), a software program is generally assumed to have a maximum life cycle of about ten years, due to the quickly changing field of IT and how quickly

computers and applications improve and are developed and processors increase in speed. Therefore, maintenance for a software program is required for just a ten-year life cycle. The maintenance budget is typically set at 100% to 150% of the total software development cost. Therefore, this breaks down to an annual budget for software maintenance in the range of 10%–15% of the total software development cost.

Summary

In this chapter, we introduced the complex area of software cost estimating. While we provide just an overview of this unique area of cost estimation, we discuss what software is and the typical WBS elements in a typical software cost estimating task, as well as software costing characteristics and concerns, and the process used to estimate the cost. A major part of the chapter discussed source lines of code and function points, the two primary means to determine the size of a software project, rather than using the traditional independent variables such as power, weight, or frequency that we use in hardware programs. In addition, one of the many unique aspects of software cost estimation includes only having development and maintenance costs phases and that there are no learning curves associated with software development. Lastly, we discussed problems with software cost growth, and we discussed some software cost estimating tools that are available on the commercial market. The intended take-away from this chapter is that estimating the cost for software development is an area that is very hard to perform accurately.

References

1. USC Viterbi School of Engineering, Center for Systems and Software Engineering website. Barry W. Boehm, COCOMO 81.
2. International Function Point Users Group (IFPUG). http://www.ifpug.org/
3. J. Clapp, A. King, and A. Taub, "COTS – Commercial Off-the-Shelf: Benefits and Burdens," The Edge Perspectives, Vol. 2, No. 1. The MITRE Corporation, March 2001, 10 + ii pages.

Cost Benefit Analysis and Risk and Uncertainty

16.1 Introduction

This final chapter covers two important topics: cost benefit analysis (CBA) combined with net present value (NPV); and risk and uncertainty analysis. While everything that we have learned to this point has been necessary to develop the final point estimate in your analysis, there is some uncertainty embedded within each of the inputs in all work breakdown structure elements. Risk and uncertainty will allow you to provide not just a point estimate but also to apply a "most probable" range of dispersion from that point estimate. For example, if our final cost estimate is for $80M, we would prefer to provide the decision-maker with a most probable range such as $80M ± $10M, so our range of costs would be between $70M and $90M, with a desired level of confidence. We can accomplish this by using a technique known as Monte Carlo simulation. But first, we start with a short look at considerations for conducting a Cost Benefit Analysis and calculating Net Present Values for various courses of actions (COAs).

16.2 Cost Benefit Analysis (CBA) and Net Present Value (NPV) Overview

A CBA is an important and classical application of cost estimating, namely making choices among several COAs. It is often the case that we need to compare more than one COA, capturing the relevant costs and benefits associated with each, in order to determine the most cost effective means of meeting a stated objective. Implicit in this statement is the recognition that there may be alternative ways of meeting an objective and that each alternative requires different resources and produces different results. This type of problem has several different names, sometimes being called an economic analysis, sometimes an analysis of alternatives, sometimes a cost and operational effectiveness analysis (COEA), and sometimes a cost-benefit analysis. For purposes of simplicity, let's call each of these analyses a "Cost-Benefit Analyses" in this chapter. Regardless of its name, however, each

analysis strives to accomplish the same purpose and objective. A reasonable definition of this kind of analysis is:

"A systematic approach to the problem of choosing the best method of allocating scarce resources to a given objective"

The types of problems that could benefit from these analyses include the following:

- Comparing alternatives
- Evaluating cash flows over time
- Choosing between leasing and buying
- Choosing between out-sourcing or providing services "in-house"
- Computing return on investment metrics to optimize portfolio of projects

Some real life examples of such problems include the following:

- Buy a house, buy a condo, rent an apartment, rent a house, or live with your parents
- Repair your car, purchase a new car, lease a new car, purchase a used car, or use public transportation
- Choose between manned and unmanned systems (e.g., UAV's or satellites)
- The next generation US Air Force tanker: should we lease or buy?

The importance of performing a CBA cannot be understated in this day and age. The following quote from a memorandum that was signed by both the Vice Chief of Staff of the Army and the Under Secretary of the Army discusses their importance in the US Army. The quote was taken from the "US Army Cost Benefit Analysis Guide," April 2013.

"We must make the best possible use of our limited funds and ensure that no significant resource-related issue is decided without a thorough review of its costs, its projected benefits, and the trade-offs that might be required to pay for it. In our decision-making, we need to supplement professional experience and military judgment with solid data and sound analytical techniques. Toward this end, we are directing that each unfunded requirements and new or expanded program proposals submitted to the Secretary of the Army . . . .be accompanied by a thorough CBA. This analysis must identify the total cost of the proposal, the benefits that will result, the bill-payers that will be used to pay for it, and the second and third level effects of the funding decision. The net result of the CBA should be a strong "value proposition" – a clear statement that the benefits more than justify the costs and required trade-offs." [1]

In addition, the US Army guide describes a CBA process that comprises the following eight major steps:

1. Define the problem/opportunity. Include the background and circumstances
2. Define the scope and formulate facts and assumptions
3. Define and document alternatives (including the status quo, if relevant)
4. Develop cost estimates for each alternative (including status quo, if relevant)
5. Identify quantifiable and nonquantifiable benefits
6. Define alternative selection criteria

7. Compare alternatives

8. Report results and recommendations [2]

We have all used the concept of a CBA in our daily lives, though perhaps not as detailed as this eight-step guide. An easy example of this would be when you have decided to purchase a new car instead of leasing one or repairing your present automobile. Let's suppose that you have narrowed your choice of new cars down to three options:

- A standard, "no-frills" car that would only cost $15,000 to purchase.
- A solid, mid-level medium-priced car that would cost $30,000 to purchase.
- A premium car that is "flashy" and would cost $45,000 to purchase.

Each of these cars provides different benefits for its cost. While the premium car has a bigger engine, drives very well, and would be the most fun to drive, if you are married and have a family, perhaps the standard "no-frills" car would be the most practical. You could actually buy *two* of them for the same cost as one mid-level car and that purchase would still be $15,000 less than the premium car. In essence, in a CBA you are comparing not only the costs to purchase, but also the benefits and the maintenance costs and the myriad other ramifications that each of the options will include.

Implicit in the name CBA is the requirement to perform a cost estimate of the alternative COA, and then to compare them in terms of the differences in many areas, such as cost or revenue in a certain timeframe. So how do we make these comparisons?

For each COA, we will need to collect data concerning our time-phased costs, which have been estimated through standard cost estimating methodologies. This data may or may not have been normalized for inflation, so ensure you normalize as necessary. Consider the following example in Table 16.1, in which cell entries are in FY13$M and payments are spread over six fiscal years:

TABLE 16.1   Course of Action (COA) Example (Costs in FY13$M)

	FY1	FY2	FY3	FY4	FY5	FY6
COA #1	30	20	5	5	5	5
COA #2	15	15	15	15	15	15

When analyzing the numbers in Table 16.1, COA #1 could represent buying a product with higher initial/upfront costs, and then lower operating and support costs in the later years. COA #2 could represent a long-term lease, with constant year-to-year costs. Two possible resolutions to the question "Which COA is better?" come to mind:

- A simple addition of the total costs prefers COA #1 ($70M to $90M)
- If the buyer cannot afford the higher upfront costs of COA #1, then the buyer will prefer COA #2. This decision might depend on the status of the organizational budget when the decision is made.

There are other possible solutions and outcomes, as well, but the two COA's mentioned earlier make clear the nature of the problem, namely, how to compare these two "time-phased" products of the cost estimating process. The resolution to this question comes from the recognition of two principles: