



Joint Agency Cost Estimating Relationship (CER) Development Handbook

9 February 2018

Preface

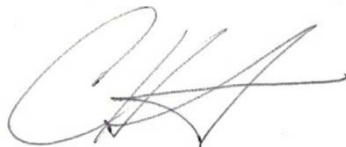
There are many valid analytic methods for correlating potential technical or other program characteristics with cost. Analysts may find certain methods more suited to specific or niche applications than others – or have understandable preferences for particular methods. However, we recognized an overarching need within the Government cost analysis community to provide a handbook to serve as a fundamental reference for parametric Cost Estimating Relationship (CER) development. This handbook provides a variety of statistical techniques to help equip the analysts' toolboxes for attacking a range of estimating problems and data issues. Additionally, it presents tried-and-true processes to help analysts choose the appropriate method, understand the mathematics behind several of the available regression and curve-fitting techniques, validate and select the most representative CER, and properly document their efforts.



Wendy P. Kunc
Director
Naval Center for Cost Analysis



Stephen G. Barth
Deputy Assistant Secretary of the Army
(Cost & Economics)



Charles Hunt
Agency Programmatic Analysis and Research
Capability Division
NASA Headquarters



Paul E. Titrault
Director, Cost Analytics and Parametric Estimating
Missile Defense Agency


Acknowledgements

This handbook was conceptualized, developed and edited under the guidance of Mr. Duncan Thomas, Mr. John Fitch, and Mr. Bruce Parker of the Naval Center for Cost Analysis (NCCA). Principal support was provided by Mr. Nick Lanham of NCCA, and Mr. Alfred Smith of Tecolote Research.

Creating this handbook has been a very rewarding experience, more so as it was developed through a joint effort with talented analysts representing multiple cost agencies across the Federal government. The work presented on the following pages is the product of dedicated individuals who committed countless hours in meticulously reviewing every paragraph, figure, and table, and providing insightful suggestions and material. This handbook is a far better product because of their generous and collaborative contributions. In particular, I wish to thank the following individuals and their organizations:

Mr. Todd Andrews, NCCA	Mr. Bruce Kraft, USAF
Ms. Anna Bobkoskie, NAVAIR 4.2	Ms. Katherine McCormack, DASA-CE
Mr. Benjamin Breaux, NCCA	Mr. Justin Moul, NCCA
Ms. Heather Brown, NCCA	Mr. Praful Patel, NCCA
Mr. Dane Cooper, NAVSEA	Mr. David Proctor, NAVAIR 4.2
Mr. Mike Dorko, SPAWAR	Mr. Kenneth Ragland, DASA-CE
Mr. Marc Greenberg, NASA	Mr. Daniel Schluckebier, NCCA
Mr. David Henningsen, DASA-CE	Ms. Mary Anne Scully, AFCAA
Ms. Pamela Johnson, NCCA	Dr. Christian Smart, MDA
Mr. Jay Jordan, NRO	Mr. Daniel Strickland, MDA
Mr. Charles Kelley, DASA-CE	Mr. Steve VanDrew, NAVAIR 4.2
Mr. Raymond Kleinberg, TACOM	Ms. Corinne Wallshein, NCCA

Our goal was to develop a handbook that can be easily understood and applied by junior analysts, and serve as a well regarded reference for senior analysts. We hope you find it useful.



Duncan D. Thomas
Technical Director
Naval Center for Cost Analysis

TABLE OF CONTENTS

TABLE OF CONTENTS 4

TABLE OF FIGURES..... 8

TABLE OF TABLES..... 10

INTRODUCTION..... 11

HOW TO USE THIS HANDBOOK 13

1.0 STEP 1: PURPOSE, SCOPE, COLLECT, VALIDATE, & NORMALIZE DATA..... 14

 1.1 Introduction..... 14

 1.2 Preparing to Collect Data..... 14

 1.3 Cost Estimate Purpose and Scope..... 15

 1.3.1 Cost Estimate Purpose 15

 1.3.2 Cost Estimate Scope and Work Breakdown Structure..... 15

 1.3.3 Obtain Subject Matter Expert Guidance to Help Identify Potential Cost Drivers 17

 1.3.4 Define Viable Hypothesis..... 18

 1.4 Sources of Data 19

 1.5 Collect and Validate the Raw Data 21

 1.6 Cost Data Normalization..... 23

 1.6.1 Content over Time 23

 1.6.2 Accounting Changes over Time..... 23

 1.6.3 WBS/CES mapping 24

 1.6.4 Escalation / Inflation..... 25

 1.6.5 Adjust for Quantity 27

 1.6.6 Cost Per Unit Characteristic..... 29

 1.6.7 Other Normalization Considerations 29

 1.7 Linking Cost to Schedule..... 30

 1.8 Summary and Introducing the Electronics Example Dataset..... 30

2.0 STEP 2: ANALYZE NORMALIZED DATA 32

 2.1 Overview..... 32

 2.2 Cost Estimating Methods..... 33

 2.3 Choosing Between Analogy, Straight Average or a CER..... 34

 2.3.1 Assess Number of Observations (**n**)..... 34

 2.3.2 Analogy Estimate..... 35

 2.3.3 Estimating with Very Small Data Sets..... 35

 2.3.4 Straight Average 36

 2.4 Univariate Data Analysis 38

 2.4.1 Significant Digits 38

 2.4.2 Descriptive Statistics..... 39

 2.4.3 Generate a Histogram 40

 2.5 Measure Correlation between Dependent and Independent Variables 41

 2.5.1 Correlation Types 42

 2.5.2 Identify Redundant Variables and Potential Multicollinearity 43

 2.6 Scatter Plot of the Most Promising Cost Drivers..... 44

 2.7 Identify Potential Variable Sets 46

 2.8 Hypothesize Functional Form and Error Term 46

 2.8.1 Linear Functional Form 50

 2.8.2 Power Functional Form..... 51

 2.8.3 Exponential Functional Form 53

 2.8.4 Logarithmic Functional Form..... 54

2.8.5 Triad Functional Form	55
2.8.6 More General Functional Forms	55
2.8.7 Caution When Regressing Transformed Data	55
2.8.8 Error Terms	56
3.0 STEP 3: GENERATE CER	57
3.1 A Guide to Regression Methodology Selection	58
3.1.1 Using OLS as a Regression Method Baseline	58
3.1.2 Choosing Where to Go When One or More OLS Statistical Assumptions is Violated	59
3.2 Select Variable Set	60
3.2.1 Value of Prior Information	60
3.2.2 Overview	61
3.2.3 Dummy Variables	61
3.3 Regression Methodologies Detail	63
3.3.1 Ordinary Least Squares (OLS)	65
3.3.2 Generalized Least Squares (GLS)	73
3.3.3 Transformable Linear and the Log-Linear Model	81
3.3.4 Generalized Linear Model (GLM)	85
3.3.5 Non-linear Least Squares (NLS)	86
3.3.6 Ridge Regression	93
3.4 Estimation with Prior Information	98
3.4.1 Types of Prior Information	98
3.4.2 Exact Prior Information on Parameter Relationships	98
3.4.3 Pseudo-Exact Prior Information on Parameter Values	99
3.4.4 Inexact Prior Information on Parameter Values	104
4.0 STEP 4: VALIDATE CER	105
4.1 Graph CER	107
4.1.1 Visualizing the Simple Regression (Single Predictor) CER	108
4.1.2 Visualizing the Multiple Regression (Single Predictor) CER	108
4.2 Model Assumptions	109
4.2.1 Ordinary Least Squares (OLS)	112
4.2.2 Weighted Least Squares (WLS)	127
4.2.3 Transforms and the Log-Linear Model	132
4.2.4 Generalized Linear Model (GLM)	132
4.2.5 Non-linear Least Squares (NLS)	132
4.2.6 Ridge Regression	133
4.2.7 Restricted Least Squares (RLS)	133
4.3 Model Diagnostics	133
4.3.1 Influential Points	134
4.3.2 Multicollinearity	141
4.4 Model Significance	145
4.4.1 Statistical Significance of CER	147
4.4.2 Validate Variable Set	148
4.5 Model Quality	151
4.5.1 Assess Metrics of Fit	151
4.5.2 Assess Metrics of Prediction	157
4.6 Model Selection	160
4.6.1 Variable Selection	161
4.6.2 Functional Form Selection	165
4.7 CER Responsiveness	167
5.0 STEP 5: CHARACTERIZE UNCERTAINTY	168

5.1 Adjust Point Estimate.....	170
5.1.1 Overview.....	170
5.1.2 Adjusting the Log-Linear Regression Result.....	170
5.2 Generate Confidence Interval	171
5.2.1 Overview.....	171
5.2.2 Extension to Other Model Forms.....	173
5.3 Generate Prediction Interval	173
5.3.1 Overview.....	173
5.3.2 Example	174
5.4 Generate CER S-Curve and Histogram.....	176
6.0 STEP 6: DOCUMENT CER.....	178
6.1 Scope/Purpose of the Recommended Cost Estimating Relationship	179
6.2 Data Documentation	180
6.2.1 Data Sources	180
6.2.2 Raw Data.....	181
6.2.3 Data Normalization.....	182
6.3 CER Development	183
6.3.1 Identify Cost Drivers	183
6.3.2 Document Regression Method Selection and CER Functional Forms	185
6.3.3 Document the Selected CER.....	187
6.3.4 Characterize CER Uncertainty.....	191
APPENDIX.....	195
APPENDIX A GENERAL THEORY.....	195
A.1 Arithmetic	196
A.1.1 Basic Operations	196
A.1.2 Weights	196
A.1.3 Linear Algebra	197
A.2 Probability	197
A.2.1 Foundations of Probability.....	197
A.2.2 Probability Distributions.....	200
A.3 Statistics	200
A.3.1 Descriptive Statistics.....	201
A.3.2 Inferential Statistics.....	212
A.3.3 Data Analysis Challenges	221
A.3.4 Data Mining	224
A.4 Regression Analysis.....	225
A.4.1 Ordinary Least Squares (OLS).....	225
A.4.2 Generalized Least Squares (GLS).....	225
A.4.3 Log-Linear Regression.....	225
A.4.4 Generalized Linear Model (GLM).....	227
A.4.5 Non-linear Least Squares (NLS).....	233
A.4.6 Ridge Regression	235
A.4.7 Mathematical/Numerical Techniques	235
A.4.8 Minimum-Unbiased-Percentage-Error (MUPE).....	236
A.4.9 Advanced Regression Methodologies.....	237
A.5 Influence Diagram.....	239
APPENDIX B MAXIMUM LIKELIHOOD ESTIMATION FOR REGRESSION OF LOG-NORMAL ERROR (MRLN) SUMMARY.....	241
APPENDIX C CER DEVELOPMENT CHECKLIST	244
APPENDIX D CORRELATION CRITICAL VALUE TABLES	245

APPENDIX E PARTIAL REFERENCES	247
APPENDIX F DATA SETS	248
F.1 Electronics Example	248
F.2 Cost Improvement Curve Example.....	248
F.3 Power Density Example.....	249
F.4 Pseudo-Exact Prior Information Example	249
APPENDIX G ACRONYMS	250
G.1 General	250
G.2 Multicollinearity.....	250
G.3 Cost Estimating and Regression Methods.....	251
G.4 Advanced Regression Methods.....	251
G.5 Influence Points.....	251
G.6 Regression Statistics.....	251
G.7 Assumption Tests	252
G.8 Fit/Predictive Statistics.....	252
G.9 DoD Terminology	252
G.10 Tools.....	254

TABLE OF FIGURES

Figure 1: CER Development Process..... 12

Figure 2: User Requirement Translated to Cost..... 17

Figure 3: Simplified Influence Diagram Example 18

Figure 4: Consolidate Raw Cost Data into a Summary Table 22

Figure 5: Notional WBS Mapping 24

Figure 6: Step 2 - Analyze Normalized Data 32

Figure 7: Confidence and Prediction Interval For the Straight Average of the Electronics Cost Data..... 38

Figure 8: Histogram of Electronics Cost Data 41

Figure 9: Scatter Plot of Cost vs. Power 44

Figure 10: Scatter plot of Cost vs Aperture 45

Figure 11: Scatter Plot of Cost vs Cost per Kilowatt 45

Figure 12: Demonstration of Multicollinearity between related variables 46

Figure 13: Selecting a Functional Form and Error Term 48

Figure 14: Concave Down Patterns 49

Figure 15: Concave Up Patterns 50

Figure 16: Linear Functional Form Example..... 50

Figure 17: Power Functional Form Examples 51

Figure 18: Power Equation in Unit and Fit Space..... 52

Figure 19: Example Cost Improvement Curve 53

Figure 20: Exponential Functional Form in Unit and Fit Space 54

Figure 21: Logarithmic Functional Form in Unit and Fit Space..... 55

Figure 22: Step 3: Generate CER..... 58

Figure 23: Dummy Variables Linear Example 62

Figure 24: Dummy Variables CIC Example..... 63

Figure 25: Simple Linear Regression Output 67

Figure 26: Simple Linear Regression Model Scatter Plot..... 68

Figure 27: Multiple Linear Regression Matrix Math..... 71

Figure 28: Multiple Linear Regression Output 72

Figure 29: Multiple Linear Regression Model Predicted vs. Actual Plot 73

Figure 30: Weighted Least Squares Scatter Plot..... 79

Figure 31: Weighted Least Squares Regression Output 80

Figure 32: Common Linear Transformations 81

Figure 33: Log-Linear Regression Model Scatter Plot 84

Figure 34: Log-Linear Regression Output 85

Figure 35: Local versus Global Solution 88

Figure 36: NLS Regression Model Scatter Plot..... 90

Figure 37: NLS Regression Output..... 90

Figure 38: Ridge Specific Regression Output..... 96

Figure 39: Ridge Trace (Perturbation) Plot..... 97

Figure 40: Fixed Coefficient Example Prediction Interval Comparison..... 103

Figure 41: Step 4: Validate CER..... 106

Figure 42: Anscombe’s Quartet 107

Figure 43: Graphical View of Simple CER 108

Figure 44: 3-D Visualization of Data..... 109

Figure 45: Step 4.2 Model Assumptions..... 111

Figure 46: Independence of Errors Residual Plots..... 114

Figure 47: Scatter Plots For Assessing the Homoscedasticity Assumption..... 116

Figure 48: Residual Plots For Assessing the Homoscedasticity Assumption 116

Figure 49: Histogram of Standard Residuals	118
Figure 50: Normal Q-Q Plot Examples.....	119
Figure 51: Non-linear Residual Plots.....	121
Figure 52: Non-linear Behavior of Residuals	121
Figure 53: Linearity – Predicted versus Actuals Example.....	122
Figure 54: Residual Plots For Linear Fits on Nonlinear Data.....	123
Figure 55: Residual Plot For the Example Linear Fit	123
Figure 56: Standardized (Internally Studentized) Residual Plot.....	124
Figure 57: Durbin-Watson Test Statistic	124
Figure 58: Histogram of Standardized Residuals	125
Figure 59: Normal Probability-Probability Plot from CO\$TAT.....	126
Figure 60: OLS Actual vs. Predicted Plot.....	127
Figure 61: Residual Plot, WLS Example (OLS).....	129
Figure 62: Residual Plot, WLS Example (Method 3).....	129
Figure 63: P-P Plot, WLS Example (OLS).....	131
Figure 64: P-P Plot, WLS Example (Method 3).....	131
Figure 65: Error versus Residual.....	135
Figure 66: Comparison of Residual Types	136
Figure 67: Diagnostic Plots for Section 3.3.1.2 OLS Example	139
Figure 68: Weak and Strong Correlation Between Predictors	141
Figure 69: Scatter Plot Matrix.....	144
Figure 70: Predicted versus Actual Plot.....	156
Figure 71: Step 5: Characterize Uncertainty.....	168
Figure 72: Bias versus Variance	169
Figure 73: OLS Example 95% Confidence/Prediction Intervals	174
Figure 74: Compare OLS CER PI to a Straight Average PI.....	175
Figure 75: OLS Example Prediction Interval Output	175
Figure 76: OLS Example Prediction Interval CDF.....	177
Figure 77: OLS Example Prediction Interval PDF	177
Figure 78: Documenting the Estimating Relationship.....	179
Figure 79: Electronics Data Scatter Plots	185
Figure 80: Electronics Data Normal Probability Plot	187
Figure 81: Documenting Fit Statistics and ANOVA	189
Figure 82: Documenting Outlier Analysis and Predictive Measures.....	190
Figure 83: Representative Charts to Document the CER.....	191
Figure 84: Documenting Representative Prediction Intervals	192
Figure 85: GLM Regression Model Scatter Plot	230
Figure 86: GLM Regression Output	231
Figure 87: Example of Data Normalization and Hypothesized Relationships Between Variables	240

TABLE OF TABLES

Table 1: Generic Primary and Secondary Data Sources 20

Table 2: Integrated Technical and Programmatic Data Table 22

Table 3: Notional Escalation Table..... 26

Table 4: Adjusting Collected Cost to FY2016..... 27

Table 5: Rate Affected CIC Notional, Normalized Dataset..... 29

Table 6 Notional Electronics Data Set..... 31

Table 7: Assessing the Accuracy of the Electronics Univariate Analysis 37

Table 8: Descriptive Statistics Summary 40

Table 9: Normalized and Composite Variable Correlation Matrix..... 43

Table 10: Notional Data to Demonstrate Functional Forms 49

Table 11: Summary of Regression Methodologies..... 59

Table 12: Summary of Methodology Properties..... 60

Table 13: Simple Linear Model Data Example 67

Table 14: Multiple Linear Model Data Example 70

Table 15: Weighted Linear Squares Data Example 78

Table 16: Example WLS Methodology Comparison..... 80

Table 17: Power Model Data Example 84

Table 18: OLS Results Before Applying Ridge Regression..... 96

Table 19: Before and After Ridge Example Coefficients with Ridge Parameter = 0.125 97

Table 20: Fixed Coefficient Example – Full Model 102

Table 21: Fixed Coefficient Example – Fixed $\beta_3 = 3$ 102

Table 22: Fixed Coefficient Example – Fixed $\beta_3 = 5$ 103

Table 23: Multivariate OLS Example Outlier Analysis Table..... 134

Table 24: Multicollinearity Data Example..... 142

Table 25: Multicollinearity Data Example Correlation Matrix..... 143

Table 26: Multicollinearity Data Example Analysis..... 143

Table 27: ANOVA Table – Section 3.3.1.3 Example..... 147

Table 28: Coefficients Table – Multiple Linear Regression Example..... 149

Table 29: Simple Linear Regression Formula Summary 152

Table 30: AIC Relative Probability Example 157

Table 31: Variable Selection Metrics..... 163

Table 32: Variable Selection Table..... 165

Table 33: Functional Form Selection Metrics..... 166

Table 34: Raw Cost Data and Notes 181

Table 35: Technical and Programmatic data..... 181

Table 36: Deriving First Unit Cost 182

Table 37: Converting Raw Cost to a Base Year 2016 Cost 182

Table 38: Key Normalized Electronics Data 183

Table 39: Pearson Product Moment Correlation..... 184

Table 40: Spearman Rank Correlation..... 184

Table 41: Summary Results: Fit Statistics 186

Table 42: Summary Results: Predictive Statistics 186

Table 43: Summary of Graphs to Include in Documentation 188

Table 44: Methods to Determine Histogram Bin Width..... 203

Table 45: Pearson Product Moment Critical Values..... 245

Table 46: Spearman’s Rho Critical Values..... 246

INTRODUCTION

Cost estimators forecast the amount of resources, time, and effort it takes to implement and execute requirements defined to support the needs of the user and functional communities. To meet the forecasting needs, the cost estimator draws heavily from historical data to develop cost estimates that:

- are relevant, defensible and objective;
- facilitate what-if analysis;
- support the identification of cost drivers; and
- provide the basis for assessing the risk and uncertainty associated with the estimate

A primary cost estimating method is to use historical data to develop parametric cost estimating relationships (CERs) through regression analysis. Regression analysis is a statistical process for estimating the relationships between a dependent variable (the element estimated) and one or more independent variables (variables that influence cost). Parametric cost estimating models are used throughout the life cycle, but are particularly useful tools for preparing early conceptual estimates when there is little technical detail. They are also useful for quickly examining the cost impacts of a range of alternative options. While this handbook uses cost estimating examples to demonstrate the regression process, the guidance is equally relevant for estimating duration (schedule), labor hours, a technical characteristic or any other item of interest.

There are many valid ways to approach, perform, and use regression analysis. The main process flow focuses on the actions taken to develop and validate the CER, and the rationale for doing so, and provides examples when possible. The goal of this document is to provide the cost analyst:

- guidance on how to collect and prepare data for the regression process
- a comprehensive resource describing the most widely used regression methods in our industry
- reasons to employ a given regression method and guidance on how to apply the selected method
- an objective basis for selecting the CER to be used in the cost estimate
- a demonstration on how to fully document the selected CER

The CER development process begins by fully defining the purpose of the estimate. The cost analyst must then perform the necessary literature search and consult with stakeholders, program office authorities, and technical experts to develop the ground rules, constraints, assumptions, boundaries, and a full description of the item to estimate.

Figure 1 shows the six basic steps of the CER development process. Analysts rarely perform this process in a linear manner. The dotted lines illustrate the most common iterative steps in the process, for instance:

- 2 back to 1: Unable to identify any meaningful drivers
- 3 back to 2: Desired regression method fails to converge
- 4 back to 3: CER fails to validate, investigate other regression methods
- 4 back to 2: CER fails to validate, no other regression methods to try, look for other drivers
- 5 back to 3: Uncertainty assessed to be unacceptable (too large or too narrow, often a subjective assessment)

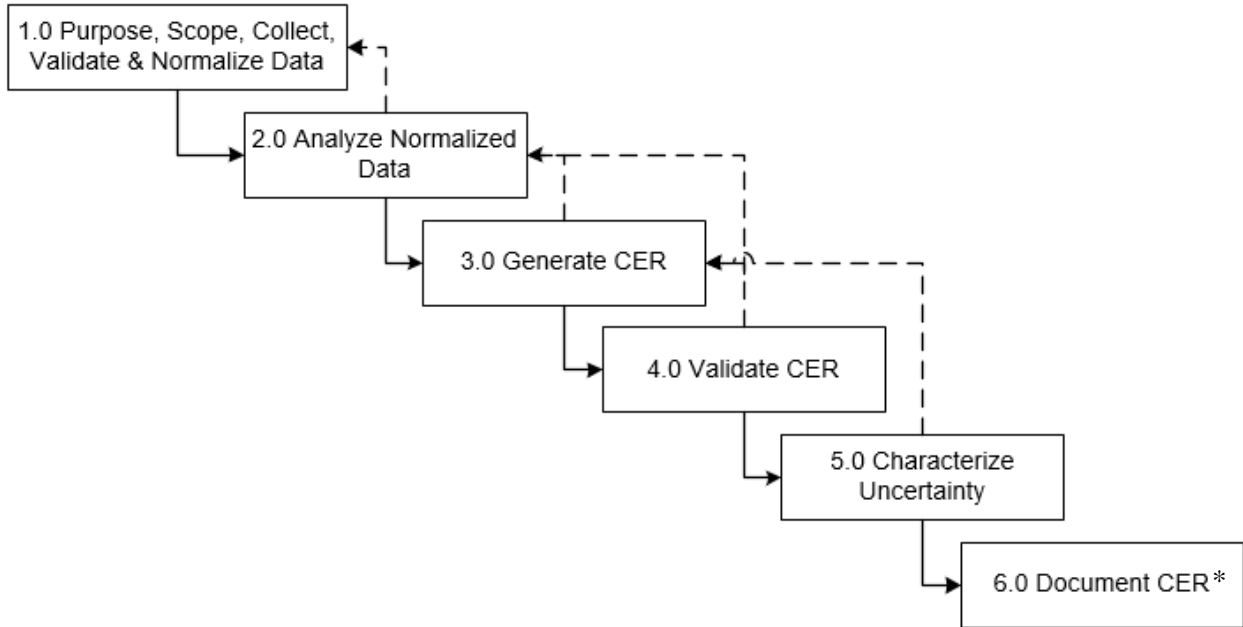


Figure 1: CER Development Process

The CER development process is iterative and equally valid for estimating cost, duration (schedule), labor hours, technical or any other quantitative aspect of a well-defined project.

**The final step should be completion of the documentation, since documentation should be completed at each step.*

HOW TO USE THIS HANDBOOK

This handbook provides a summary of the processes, mathematics and guidance for the development of CERs using regression analysis. The handbook functions more as an electronic resource than as a comprehensive textbook, enabling the reader to navigate through hyperlinks instead of leafing through physical pages.

The material in this handbook helps the analyst choose the appropriate regression method, understand the mathematics behind several of the most popular regression methods, validate and select the best CER, establish the range around the CER result and document the entire process.

Throughout the handbook, there are flow charts that detail the process flow at each stage. In addition, there are files that complement the handbook. The complete set of files includes:

- **1. CER Handbook:** This handbook published as a Portable Document Format (PDF) file.
- **2. CER Handbook Examples:** MS Excel file containing Figures and Tables in this handbook.
- **3. CER Handbook Documentation:** MS Excel file containing the Calculations, Figures and Tables used in [Step 6: Document CER](#).
- **4. CER Handbook StatSoftwareMatrix:** MS Excel file comparing the features of the following software products that support statistical analyses in CER development:
 - **MS Excel:** Contains several statistical functions and a data analysis add-in
 - **CO\$TAT:** An MS Excel add-in developed to support cost statistical analysis
 - **Minitab:** Comprehensive statistical package
 - **R:** Public domain tool delivering a programing language for statistical computing
 - **SAS:** Statistical Analysis System, advanced analytics
 - **JMP¹:** Developed by SAS to provide a simpler user interface for advanced data analysis with large datasets
 - **STATA:** Statistics and Data, data analysis and statistical software.
- **5. ZIP File Containing CER Handbook Flow Charts:** Several Visio files with the flowcharts used in this handbook, including a file that contains the complete process.

The CER Handbook is not a substitute for textbooks and papers, some of which are listed in [Appendix E PARTIAL](#) References nor does it serve as an alternative to formal cost analysis training.

¹ Pronounced “jump”, the name was derived from its inventor “John’s Macintosh Project.”

1.0 STEP 1: PURPOSE, SCOPE, COLLECT, VALIDATE, & NORMALIZE DATA

1.1 Introduction

The purpose of the estimate drives the scope and approach. The estimator needs to build a structure of Work Breakdown Structure (WBS) elements that will provide the required insight into the cost of the program. The program elements are populated with methodologies that will best support the estimate purpose and any anticipated what-if analysis.

To estimate each WBS element, four general methodologies are available: analogy, factor, parametric, and engineering build-up. This guide focuses on parametric (i.e., CERs), which uses statistical techniques to form an equation showing the relationship between dependent (cost and schedule), and associated independent variables (technical, programmatic, etc.).

Regardless which methodology is used, the analyst needs to understand the underlying factors that drive the cost and schedule for the individual program elements, as this drives how the analyst will model the costs for these elements. This knowledge comes from training, experience, and extensive discussions with those who have management or engineering experience with similar programs. Analysts should develop, maintain, and modify as necessary a working hypothesis of the underlying set of drivers for each element of cost and schedule. This hypothesis drives data collection, and in the case of CERs, the functional form of the CER. Influence diagrams are useful tools for developing a set of underlying relationships.

1.2 Preparing to Collect Data

Historical data are needed to support a robust cost estimating research. The main data types to be collected include:

- **Cost:** Recurring and Non-recurring, and further subdivided into categories such as labor, material, overhead and fee are some examples of cost.
- **Programmatic:** Total quantities, quantity profiles, contract type, sole source or competitive, quantities, production rate requirement, initial operation dates and maintenance concept are some examples of programmatic.
- **Performance and Technical:** Speed, range, depth, survivability, noise reduction are some examples of performance data. Weight, frequency, power, volume, and density are some examples of technical data.
- **Schedule:** Hours, months, and years are some examples of schedule data.

There is a substantial amount of effort required prior to beginning the actual data collection process. The analyst must clearly define the purpose and project description (scope). From this information, the analyst can hypothesize plausible functional forms (i.e., linear, exponential, etc.) and potential cost drivers. Validated data from a variety of different projects performed at different times and under different circumstances must be normalized, that is, rendered consistent with and comparable to each other. This section addresses each of these steps.

1.3 Cost Estimate Purpose and Scope

1.3.1 Cost Estimate Purpose

Potential uses for a cost estimate include:

- Compare the benefit of a particular project relative to its cost
- Investigate the cost impact of alternative ways to satisfy requirements
- Determine if building a new item is more cost effective than buying the item
- How to shape a solution to fit into a specific cost (cost as an independent variable)
- Support for a milestone review
- How to allocate funds across a variety of project elements
- Assess the impact of the timing of when work is performed
- Independent assessment of a cost estimate
- Basis for a budget request
- Support contract negotiations

The analyst also needs to understand the nature of the alternatives to be investigated to determine the level at which the data are needed. The data will dictate the level at which the model can generate costs. The purpose of the estimate drives the scope of the estimate.

Defining the estimate purpose is the foundation for determining the estimate scope and the basis for how to construct the model.

1.3.2 Cost Estimate Scope and Work Breakdown Structure

The cost estimate scope identifies the bounds of the estimate. The analyst needs to approach every estimate with a clear understanding of the estimate scope, assumptions and ground rules.

To gain an understanding of program scope, analysts typically refer to the program baseline or other programmatic documentation. The definition and content of a program baseline vary from organization to organization. In generic terms, most program baseline descriptions include²:

- Program's purpose and its system and performance characteristics and all system configurations
- Any technology implications
- Its program acquisition schedule and acquisition strategy
- Its relationship to other existing systems, including predecessor or similar legacy systems
- System quantities for development, test, and production
- Deployment and maintenance plans
- Support (manpower, training, etc.), security needs and risk items

Examples of program baseline descriptions include the NASA Cost Analysis Data Requirement (CADRe)³ and Department of Defense (DoD) Cost Analysis Requirements Description (CARD).

² These are from Chapter 7 of the Government Accountability Office (GAO) Cost Estimating and Assessment Guide (GAO-09-3SP), March 2009

CADRe is a three-part document that describes a NASA project at each milestone, contains key technical parameters, and captures the estimated and actual costs in a WBS structure. NASA's CADRe system provides historical record of cost, schedule, and technical project attributes so that estimators can better estimate future analogous projects. The final CADRe containing actual costs rather than estimates is the best source.

The CARD⁴ succinctly describes the program baseline. This includes the key technical, programmatic, operational, and sustainment characteristics of a program, along with supporting data sources, and provides all of the program information necessary to develop a cost estimate. Use of the CARD enables different organizations preparing cost estimates to develop their estimates based on the same definition of the program requirements. As a program evolves and its costs and funding needs change, the CARD, as a living document, evolves with it. If a CARD does not exist, is incomplete, or is not current, the analyst must work with the program office to develop something that provides the proper foundation for the estimate.

A well-defined WBS should be part of the program baseline documentation. A WBS contains three pieces of information: WBS Number, WBS Element (name), and WBS Definition. The WBS ensures that all components of the system are addressed with no inadvertent omissions of subsystems or functions. The appropriate WBS structure and respective data collection effort varies depending on the life-cycle phase: Development, Production, Operating and Support (O&S), or System Disposal. For defense Development and Production contracts, MIL-STD-881C⁵ (or the most current version) is the authoritative guidance document. The Cost Assessment and Program Evaluation (CAPE) O&S Cost Estimating Guide⁶ is the authoritative source for O&S Cost Element Structure (CES).

CERs are often developed at lower levels than the WBS structures defined in MIL-STD-881C or the CES structures defined in the CAPE O&S Cost Estimating Guide⁷. These details are usually found in the program baseline or other program documentation. They are critical pieces of information to provide the analyst context for the CER under development. In particular, the cost analyst should examine the WBS/CES to determine if other elements in the WBS/CES may influence the element(s) they are estimating. For example, items produced during Development may be manufactured on the same production line used for Production. Additionally, the project WBS/CES structure is a key consideration for selecting (and adjusting) analogous historical projects to use as the basis for estimating the new item.

A well-defined WBS/CES describes the project scope and is a key consideration for selecting (and adjusting) analogous systems.

³ https://www.nasa.gov/offices/ocfo/functions/models_tools/CADRe_ONCE.html

⁴ For more detail on the CARD, see the Cost Assessment Data Enterprise (CADE) public website. <http://cade.osd.mil/policy/card>

⁵ "Work Breakdown Structures for Defense Materiel Items," MIL-STD-881C, 3 October, 2011.

⁶ "Operating and Support Cost-Estimating Guide," Office of the Secretary of Defense (OSD) Cost Assessment and Program Evaluation (CAPE), March 2014.

⁷ An O&S WBS is planned for the next version of MIL-STD-881, currently under development.

1.3.3 Obtain Subject Matter Expert Guidance to Help Identify Potential Cost Drivers

Figure 2 illustrates how the user requirement becomes a cost estimate of alternatives (4.0) or a budget cost estimate (6.0). Along the way, operators, engineers and program subject matter experts (SMEs) construct the necessary documentation to identify performance specification, design alternatives, and the final system configuration. Apart from identifying the program baseline and assembling other program documentation, the SMEs provide valuable insight and context for the performance, technical and programmatic factors that may drive cost.

Rarely is the process in **Figure 2** performed in a linear manner. The dotted lines illustrate common iterative steps, for instance:

- 2 back to 1: Unable to define performance parameters that meet the user needs
- 3 back to 2: Unable to specify an alternative that will meet the performance requirements
- 4 back to 3: Cost for all alternatives are unacceptable
- 5 back to 3: Unable to convert an alternative into a detailed specification
- 6 back to 5: Product configuration must change to meet budget constraint

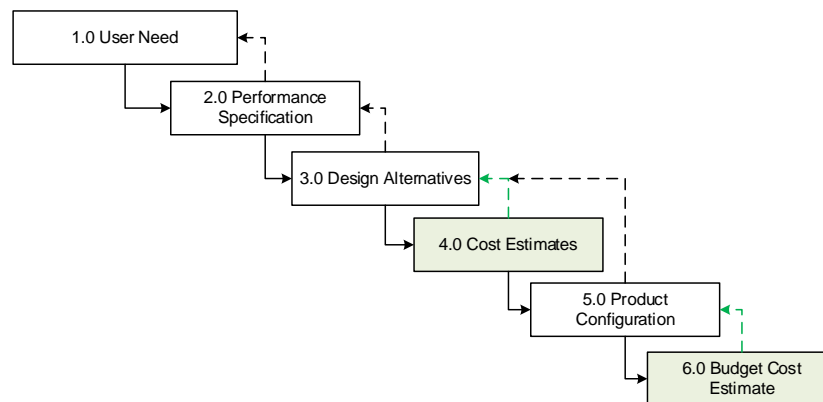


Figure 2: User Requirement Translated to Cost

The analyst is now ready to develop an initial hypothesis.

The program baseline, WBS/CES, and SME insights provide a basis for hypothesizing how to estimate a WBS/CES element.

The hypothesis can begin with development of an influence diagram. In the context of regression analysis, an influence diagram is a compact graphical representation of potential variables and how they may influence cost. The diagram also helps visualize how the variables may interact with each other. **Figure 3** provides a simplified example of an influence diagram illustrating some of the variable information gathered in support of the electronics example estimate. A thick black arrow indicates a hypothesized positive correlation between a variable that may influence cost and the cost element. That is, there is an expectation that as the variable increases in magnitude, the item cost will increase as well. The

dashed (green) arrows denote presumed negative correlation. Technology Readiness Assessment (TRA)⁸ is an example of possible negative correlation. The greater the assessment number, the more mature the technology, leading to a more confident estimate (not necessarily lower cost, but probably lower risk).

The influence diagram will often include categorical variables. In the context of cost regression analysis, categorical variables take on one of a limited, and usually fixed, number of possible values. In **Figure 3**, several categorical examples are illustrated. See [3.2.3 Dummy Variables](#) for guidance on how to employ categorical variables in a regression analysis. See Appendix [A.5 Influence Diagram](#) for another example of an Influence Diagram.

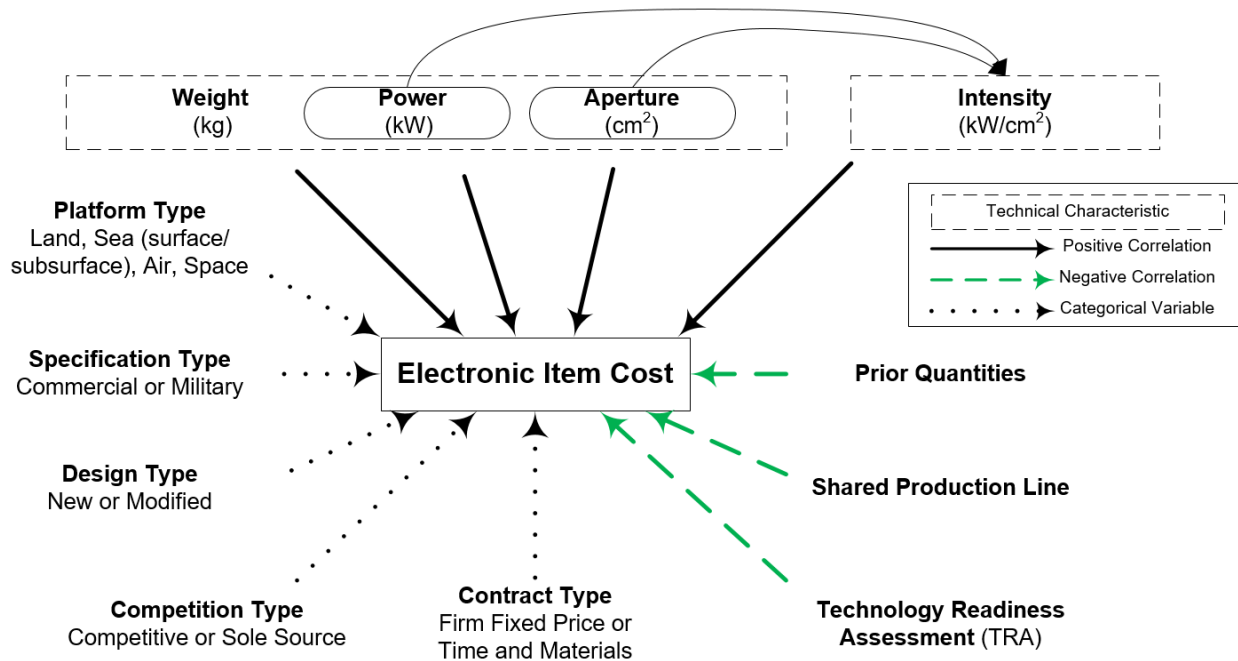


Figure 3: Simplified Influence Diagram Example

A thorough understanding of the estimate purpose, scope, WBS/CES, and operational/engineering insight is the basis for hypothesizing variables that influence the cost of interest.

1.3.4 Define Viable Hypothesis

Analysts are recommended to formally define a hypothesis and function form that relates the independent and dependent variable(s). Whether or not formally defined, recognize from a logical perspective, all CERs implicitly assume a hypothesized relationship between the dependent and independent variables. For example, we may hypothesize a linear relationship between cost (y) and system output power (x): $y = ax + b$ or a non-linear relationship: $y = ax^b$.

⁸ Technology Readiness Assessment (TRA) Guidance, April 2011 (revision posted 13 May 2011), Assistant Secretary of Defense for Research and Engineering (ASD(R&E)). Also known as Technology Readiness Level as defined by NASA in: Technology Readiness Levels, 5 April 1995, Advanced Concepts Offices, NASA

This hypothesis will serve as the starting point for the CER data collection process and supporting analysis. At the conclusion of this process, the resulting CER must make sense from a technical perspective.

Working with SMEs not only helps identify potential cost drivers, but also helps with hypothesizing the shape of the relationship (does the driver move with or counter to cost, is the rate of change constant, are there logical thresholds, etc.). The goal of this step is to prioritize the data to be collected prior to searching for sources of data. If more extensive data are readily available than required by the hypothesis, it should be collected as well, as this may provide insights necessary for resolving problems later, and may lead the analyst to recast their hypothesis.

Many tools can search through mountains of data to identify similar relationships to test for statistical significance. While there is a place for such tools, effective and efficient cost analysis relies on serious thought about cost-to-cost driver relationships that reflect the underlying engineering, operational, and programmatic relationships. This is even more important given the limited sample sizes and the noisiness of cost data. A CER showing that total fuel consumption decreases as hours of flight increases does not make sense and may require additional analysis to confirm CER realism. Analysts are encouraged to form one or more hypotheses based on what they expect should work and to proceed through the data collection and regression challenges with the hypothesis in mind.

The resulting hypotheses should guide the data collection process.

1.4 Sources of Data

At this point, we should have a good idea how to identify analogous historical programs and the data to be collected. [1.2 Preparing to Collect Data](#) introduced the type of data to be collected: cost, programmatic, performance, technical and schedule. A broader breakdown of data types includes:

- **Quantitative:** historical data on cost, programmatic, performance, technical and schedule
- **Qualitative:** often subjective in nature and often provided by SMEs
- **Primary data:** data collected from an original source
 - The contractor is a major source of primary data
 - The test center is the source of primary data on flight and weapons testing
 - Depots are the source for data on overhauls, shop replaceable units (SRUs), modifications, etc.
- **Secondary data:** data collected from a source other than the original data source
 - Documented cost estimates, factors books, studies, audit reports, and industry standards are examples of secondary data.

The program baseline (introduced in [1.3.2 Cost Estimate Scope and Work Breakdown Structure](#)) describes what we must estimate. **Table 1** provides a list of generic primary and secondary data sources⁹ to consider when searching for historical program data.

Table 1: Generic Primary and Secondary Data Sources

Data type	Primary	Secondary
Basic accounting records	x	
Data collection input forms	x	
Cost reports	x	x
Historical databases	x	x
Interviews	x	x
Program briefs	x	x
Subject matter experts	x	x
Technical databases	x	x
Other organizations	x	x
Contracts or contractor estimates		x
Cost proposals		x
Cost studies		x
Focus groups		x
Research papers		x
Surveys		x

Source: DOD and NASA.

Before data collection starts, the cost analyst should identify what data source is authoritative for each data field and why. During data collection, analysts often find that multiple data sources yield conflicting data, and the analyst needs a paradigm to resolve these conflicts and prioritize data collection efforts. The best data sources are traceable without alteration to primary data sources.

For DoD analysts, there is a variety of data sources. The Cost Assessment Data Enterprise¹⁰ (CADE) is a DoD initiative to collect, organize and display program cost, programmatic, technical and schedule data in an integrated single web-based application. CADE provides the government analyst an authoritative source for Cost and Software Data Reports (CSDR), Earned Value Management (EVM) data, and Visibility And Management of Operating and Support Costs (VAMOSOC) to support the cost estimating process.

Program offices maintain a variety of other documents which may provide useful data, to include :

- Initial Capabilities Document (ICD) and Capability Development Document (CDD)
- Acquisition Plan (AP) / Acquisition Strategy (AS)
- Deployment Plan
- Software Requirement Specification (SRS)
- Life Cycle Sustainment Plan (LCSP)

⁹ Table 10 from Government Accountability Office (GAO) Cost Estimating and Assessment Guide (GAO-09-3SP), March 2009

¹⁰ <http://cade.osd.mil/>

- Test and Evaluation Master Plan (TEMP)
- Integrated Logistics Support Plan (ILSP)
- Integrated Master Plan / Schedule (IMP/IMS)
- Contracts and Proposals
- Program and Milestone Review Briefings

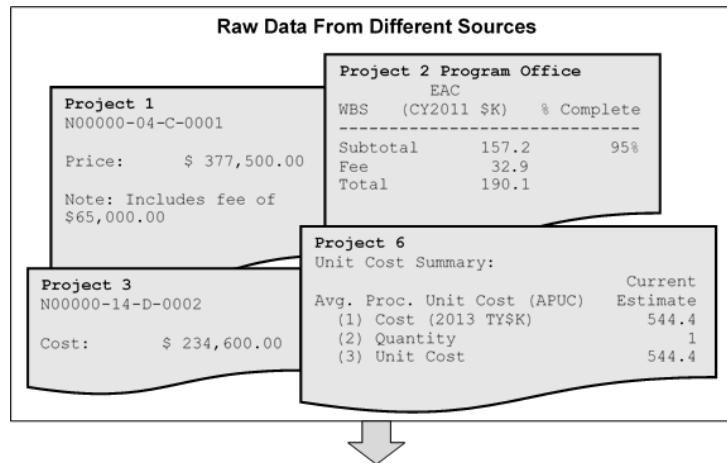
The analyst should talk to other experienced analysts to determine if other data sources may be available to support the CER development process.

There are many potential sources of data. The handbook provides generic sources. The agencies charged with performing cost estimates should identify the sources of authoritative data.

1.5 Collect and Validate the Raw Data

Raw data are unaltered records from primary or secondary sources. At this point in the process, the goal is to capture all of the information (data, context, documentation, etc.) required. **Figure 4**, an electronic data set, illustrates raw data collected from a variety of different projects with different content and arranged in different formats. Raw data drawn from these disparate formats should be organized into a common format. By organizing the data using a consistent format, detecting missing data becomes easier. Before this information can be subjected to any statistical analysis, the data must be organized, consolidated, and normalized.

CER Development Handbook



Observation	Price (\$K)	Fee Included	Type	Dollar Type	Year	Source	Comments
Project 1	\$377.5	\$65.0	Unit Cost, 90% Complete	Constant	2004	SAR - 2014	Fee provided by PMO
Project 2	\$190.1	\$32.9	Unit 10 EAC, 95% Complete	Constant	2011	PMO	PMO data in constant dollars (BY11)
Project 3	\$234.6		Unit CostC, 100% Complete	Constant	2014	SAR - 2015	
Project 4	\$343.4	\$79.0	EAC Total, 99% Complete	Budget	2007	CSDR (1921)	as of 31 May 2007
Project 5	\$521.7	\$108.0	FFP Values	Budget	2008	Contract	
Project 6	\$544.4		EAC, 98% Complete	Budget	2013	SAR	as of 30 Sep 2013
Project 7	\$782.4	\$198.0	EAC, 95% Complete	Budget	2005	CPR	as of 30 April 2005
Project 8	\$944.8	\$192.0	FFP Value	Budget	2011	Contract	
Project 9	\$479.0		T&M Ceiling Value	Budget	2012	Contract	
Project 10	\$1,000.0	Unknown	Estimate	Constant	2007	Foreign	Source in US\$, unknown Foreign to US adjustment
Project 11	\$861.0	\$145.0	Actual	Constant	2015	PMO	Not a comparable system

Figure 4: Consolidate Raw Cost Data into a Summary Table

Documenting all relevant information is important to understand the context of observations. **Table 2** shows an example of technical, categorical, and programmatic data from **Figure 4**.

Table 2: Integrated Technical and Programmatic Data

Observation	Power (kW)	Aperture (cm ²)	Unit Number	Learning Slope	Service	Contract Type
Project 1	10.00	8.70	1		Air Force	FFP
Project 2	5.00	8.00	10	0.95	Air Force	T&M
Project 3	5.20	8.20	1		Air Force	FFP
Project 4	7.00		1		Air Force	T&M
Project 5	12.00	9.00	1		Air Force	FFP
Project 6	17.80	9.50	1		Air Force	T&M
Project 7	21.00	9.20	1		Air Force	T&M
Project 8	25.00	9.70	1		Air Force	FFP
Project 9	18.00		1		Air Force	T&M
Project 10	6.20	8.20	1		N/A	FFP
Project 11	13.00	9.25	1		Army	FFP

Figure 4 and **Table 2** are simplified examples demonstrating key steps in CER development. Individual agency policy and guidelines define the format, content, and tools used to collect and document raw data.

Organizing raw data using a consistent format will make it easier to identify gaps.

1.6 Cost Data Normalization

The process of normalization reduces the noise in the data. The following sections discuss the most common normalization procedures in support of a cost estimate.

The goal of data normalization is to ensure data are comparable in content across the observations. The observations should capture “noise” that represents different technical, schedule, management, contractor, risk, policy, and related challenges. The goal is not to remove all scatter from the data. The goal is to have scatter that supports a defensible estimate and the ability to construct a realistic range around that estimate.

It is sometimes appropriate (especially when the number of available data observations is small) to include physically different observations, such as air- and ground-based systems, in a single stratified data set. These stratifications are addressed with categorical variables and discussed further in [1.6.1 Content over Time](#). Other times, the data may be accurate but simply unexpected. Do not turn objective, data driven models into ones that are subjective and driven by assumptions. The following are items to consider when performing cost data normalization.

1.6.1 Content over Time

Programs evolve and change over time, often because of requirements creep, and just as often because they outlive the design life span. Normalizing for time assumes a change will be made to a portion of the raw data with the introduction of a dummy variable or a factor.

An example is noting a large change in lot cost from one lot to the next that requires an adjustment. This can happen when the contractor changes the production line or chooses to buy rather than make a particular item. The presence or absence of customer furnished parts (or Government Furnished Equipment – GFE) is a consideration as is the absence of vendor unpriced effort in the cost record.

Normalizing for time includes adjusting raw data for life cycle phase and/or acquisition strategy. Using data fields such as program start year, program phase (Development, Production, and O&S), and simultaneous vs. sequential development allows the analyst to adjust raw data to reflect the important context between observations and account for programmatic differences.

1.6.2 Accounting Changes over Time

Each contractor has a unique accounting structure that may change according to program requirements or change over time in response to changing policy, regulations or efforts to improve productivity. If a company makes a change to their accounting system, two observations collected at different points in time may need an adjustment to be comparable. For example, a major accounting change that moves manufacturing support costs into overhead could be mistaken for learning effects. Documenting definitions for each WBS element may require the analyst to make an adjustment (e.g., re-map) to the provided labor categories.

Accounting details can be found in the vendor’s financial disclosure statements or by contacting the appropriate Defense Contract Management Agency (DCMA) office. Other accounting details to understand are overhead and fee. It is important to understand how the accounting system reports

overhead and fee and the collected data adjusted accordingly to generate consistent data across projects. Fee tends to be removed from collected cost data, while overhead tends to remain.

1.6.3 WBS/CES mapping

The WBS/CES, and supporting definitions, will vary between project offices and between contractors. An early step in the normalization process is to map data into a standard WBS ([1.3.2 Cost Estimate Scope and Work Breakdown Structure](#)). Mapping a particular data item, record or contractor WBS/CES element to a specific element in a standard WBS is often subjective. The goal is to make the best effort to ensure the standard WBS/CES element is populated with all the cost associated with that element. Well-documented subjective assumptions that include a data point rather than eliminating a data point are preferred.

Figure 5 below illustrates a notional WBS mapping challenge. On the left is the contractor WBS illustrating how they collected cost. On the right is the standard WBS, used as the basis to estimate cost. Even if the contractor provides cost data in the standard WBS, a best practice is to understand how the contractor performed the mapping. In other situations, the analyst must perform the mapping.

Figure 5 also illustrates two mapping techniques to address the challenge noted above: many-to-one and one-to-one. Specific examples are:

- Many-to-one, the source WBS
 - Need to move Integration, Assembly, Test and Checkout (IAT&C) costs from the four Air Vehicle sub elements to a single, separate IAT&C element.
 - Need to move the separately reported Project Management and System Engineering costs into one element for System Engineering and Project Management (SEPM)
- One-to-one, the source WBS
 - Did not report System Test and Evaluation costs (it is missing). Further research is required.

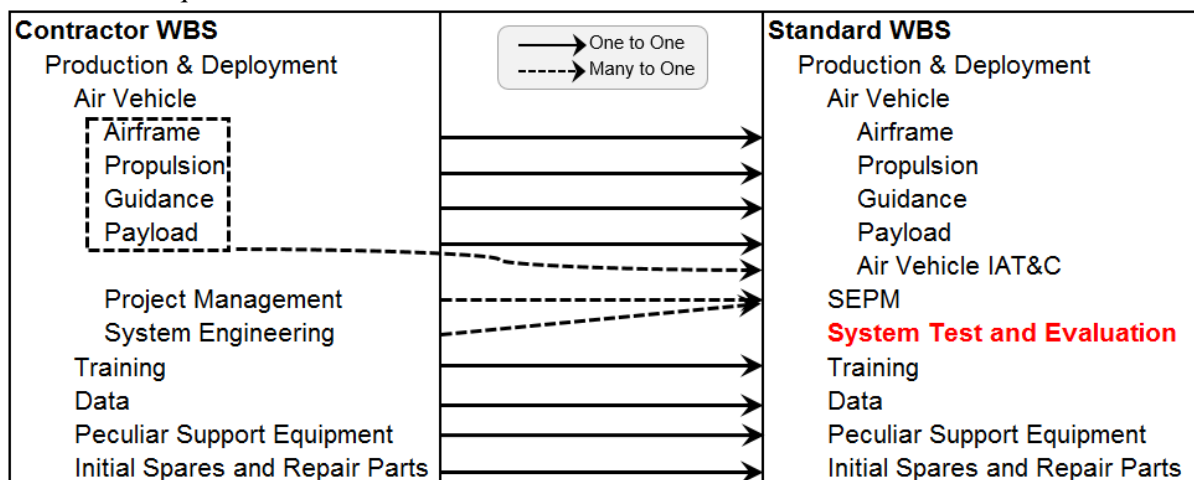


Figure 5: Notional WBS Mapping

Mapping historical project source WBS/CES elements to a standard WBS/CES is a first step to make historical project costs comparable.

1.6.4 Escalation / Inflation

Escalation and inflation correct for the buying power of funding over time. To estimate the cost of a new system properly, the analyst needs to understand the relationships between historical costs, technical characteristics, quantity orders and purchase timings. In order to derive defensible relationships to estimate the final cost, the analyst must account for the effects of persistent underlying cost increases or decreases that occur regardless of individual programmatic decisions.

Two forms of underlying cost increases that the analyst must consider are inflation and escalation. Inflation, a subset of escalation, is defined as “the proportionate rate of change in the general price level, as opposed to the proportionate increase in a specific price.”¹¹ In other words, inflation measures the change in the value of the dollar over time. The Department of Defense uses the term “escalation” to refer to price changes of particular goods (a specific commodity such as steel ships) and services in specific sectors of the economy.¹²

When converting cost data to constant dollars in support of CER development, it is recommended that an escalation index measuring cost changes of analogous items be applied, and not an inflation index. To convert CER results to a final Then Year (TY) cost estimate¹³ it is recommended that escalation, not inflation based rates be used. Examples of data that should be adjusted using commodity-based escalation include contractor labor rates, aircraft unit costs, and fuel costs. Sources for escalation rates include the Bureau of Labor Statistics (BLS). There are many others.

Historical cost data represents either expenditures or obligations. The key distinction is important because raw indices are used to convert expenditures (transactions at a specific point in time) and weighted indices are used to convert obligations (dollars which will be spent over an outlay period).¹⁴ A raw index is used when funds are obligated (guaranteed) and expended (paid for) in a single year. When funds are obligated in one year, but expended over a number of years, a weighted index is used to account for price change that occurs in those subsequent years. Concepts for raw and weighted indexes apply equally to both inflation as well as escalation indexes.

*Normalizing using inflation rates yields constant year dollars.
The recommended approach is to normalize using escalation rates yielding constant prices.*

¹¹ Office of Management and Budget (OMB) Circular A-94

¹² Inflation and Escalation Best Practices for Cost Analysts, OSD CAPE, April 2016

www.cape.osd.mil/files/InflationandEscalationBestPracticesforCostAnalysisforWebsiteForPubRelReview.pdf

¹³ Then Year” refers to dollars required to make payments on goods or services over a specified number of fiscal years. The outlay profile is a function of the commodity estimated. The term “Base Year” refers to dollars required to make a payment over a 12-month fiscal year. Costs normalized to a Base Year (BY) with an inflation index are called “Constant-Year (CY) dollars,” or constant dollars. Costs normalized to a BY with an escalation index are called “Constant Prices (CP).”

¹⁴ DOD 7000.14 – R Vol. 2, 1-14.

For demonstration purposes, **Table 3** illustrates the data required to normalize cost data to a given constant price. Raw factors to convert from one fiscal year (FY) to another are derived from the Rate information. The Weighted factors are derived from the Outlay Profile and the Raw factors¹⁵.

Table 3: Notional Escalation Table

Source:		Notional Electronics Escalation								
Revision Date:		3-Mar-16								
GFY	Rate	Raw	Weighted	Outlay Profile						
2000	1.014	0.75790	0.77759	23.7%	22.2%	21.2%	19.5%	8.0%	3.7%	1.8%
2001	1.018	0.77154	0.78986	23.7%	22.2%	21.2%	19.5%	8.0%	3.7%	1.8%
2002	1.008	0.77771	0.80364	23.7%	22.2%	21.2%	19.5%	8.0%	3.7%	1.8%
2003	1.010	0.78549	0.79350	66.84%	23.95%	7.18%	0.66%	0.35%	1.02%	
2004	1.020	0.80120	0.81094	69.77%	21.43%	5.32%	2.35%	0.36%	0.77%	
2005	1.028	0.82363	0.83479	65.9%	24.5%	7.6%	0.8%	0.4%	0.8%	
2006	1.031	0.84916	0.85954	65.9%	24.5%	6.6%	1.3%	0.5%	0.3%	0.9%
2007	1.027	0.87209	0.88122	62.0%	25.0%	8.0%	3.0%	0.7%	0.3%	1.0%
2008	1.024	0.89302	0.89928	65.0%	25.0%	7.0%	1.5%	0.7%	0.3%	0.5%
2009	1.015	0.90642	0.91153	65.0%	25.0%	7.0%	1.5%	0.7%	0.3%	0.5%
2010	1.008	0.91367	0.92377	62.0%	25.0%	8.0%	3.0%	0.7%	0.3%	1.0%
2011	1.020	0.93194	0.94105	62.0%	25.0%	8.0%	3.0%	0.7%	1.3%	
2012	1.018	0.94872	0.95810	60.0%	25.0%	8.0%	3.0%	2.0%	1.0%	1.0%
2013	1.015	0.96295	0.97221	60.0%	25.0%	8.0%	3.0%	2.0%	1.0%	1.0%
2014	1.015	0.97739	0.98526	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2015	1.011	0.98814	0.99722	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2016	1.012	1.00000	1.01154	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2017	1.018	1.01800	1.02998	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2018	1.018	1.03632	1.04929	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2019	1.020	1.05705	1.07027	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2020	1.020	1.07819	1.09168	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2021	1.020	1.09976	1.11351	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2022	1.020	1.12175	1.13578	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2023	1.020	1.14419	1.15850	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2024	1.020	1.16707	1.18167	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%
2025	1.020	1.19041	1.20530	62.0%	26.0%	5.0%	3.0%	2.0%	1.0%	1.0%

Table 4 illustrates notional values required to adjust the collected costs to a specific base year, in this case FY2016. When the collected costs are in a specific FY, the adjustment is derived by dividing the Raw factor for that FY by the Raw factor for the year of the collected data. For example, Project 1:

$$\text{FY2004 to FY2016 Adjustment} = 1.00000/0.80120 = 1.24813$$

¹⁵ The data in Table 3 is derived from Office of the Under Secretary of Defense, “Revised Inflation Guidance President’s Budget.” January or February 2005 through 2016 <https://www.ncca.navy.mil/tools/inflation.cfm>

Be cognizant of significant digits when deriving escalation and inflation indices. Precision displayed in tables and images are generally not the same as the precision used in calculations. Take care to be consistent with precision.

For DoD projects, collected cost data could be in terms of Then Year (TY) dollars. The process for adjusting collected TY costs is similar, except that the denominator is the Weighted factor rather than the Raw factor.

There are only nine projects listed in **Table 4**, not the eleven observations captured in [Figure 4: Consolidate Raw Cost Data into a Summary Table](#). Project 10 and 11 were dropped from the dataset going forward because they were determined to be too different from the estimated project to be useable (project 10 was foreign, project 11 was deemed not comparable). Project 10 and 11 remain in the raw data collected to support the estimate. They would be the first candidates to revisit should there be a need to discover additional data.

Table 4: Adjusting Collected Cost to FY2016

Observation	First Unit Cost	Fiscal Year	Dollar Type	FY to FY16	TY to FY16	Normalized Cost
Project 1	\$312.5	2004	Constant	1.248132		\$390.0
Project 2	\$186.4	2011	Constant	1.073029		\$200.0
Project 3	\$234.6	2014	Constant	1.023132		\$240.0
Project 4	\$264.4	2007	Budget		1.134794	\$300.0
Project 5	\$413.7	2008	Budget		1.111997	\$460.0
Project 6	\$544.4	2013	Budget		1.028587	\$560.0
Project 7	\$584.4	2005	Budget		1.197909	\$700.0
Project 8	\$752.8	2011	Budget		1.062648	\$800.0
Project 9	\$479.0	2012	Budget		1.043737	\$499.9

Normalizing can yield significantly different results depending on the escalation indices used. Take caution when developing labor CERs using labor costs vice labor hours.

1.6.5 Adjust for Quantity

Adjusting for quantity simply means putting the dollars (cost) on a per unit scale that allows the cost of any number of units to be estimated. If there is no change in the cost per unit regardless of the quantity, a simple division of the total cost by the number of units is sufficient. If the cost of a unit is affected by quantity, if possible, attempt to differentiate between fixed costs (e.g., program management, systems engineering) versus variable costs (e.g., material, direct labor).

The context of each data source must be established to normalize for quantity. Considerations include:

- How many units are captured in the source data?
- Is the source data a specific part of a production line or the entire production line?

- Is the cost data associated with a specific unit or lot (total or average)?
- Are there fixed costs that should be separated from the quantity-based cost data?
- Is there a cost improvement curve¹⁶ (CIC) based on quantity that should be considered?

If CIC analysis is required, there are two key considerations: 1) is the cost improvement explained by the number of units only; 2) is the rate at which units are produced also influencing the cost? Brief descriptions of both are presented in the next two sections.

1.6.5.1 Basic Cost Improvement Curve (CIC) (Learning) Theory¹⁷

When touch labor is involved, a more elaborate treatment is often necessary to capture the impact of cost improvement, rate of production and/or related fixed costs. CIC analysis is a methodology to estimate unit cost based on known unit or lot-based cost data from a production line. CIC theory is based on the observation that unit (or average unit) cost is reduced by a constant factor (slope) each time the number of units, Q ¹⁸, is doubled¹⁹.

In the context of cost normalization, document the Q assumption and utilize CIC analysis to estimate the cost, T_x , of a specific unit where x is the unit number. For example, T_1 is the first unit cost and T_{100} is the unit cost at the 100th unit across the data set. If necessary to estimate a theoretical T_x , a slope must be selected. There are at least two ways to select a slope and thereby approach the T_x calculation:

- **Use historical cost data from prior production to derive slope and calculate T_x :** Use this approach if the estimate is a continuation of prior production and the historical production process is analogous.
- **Use cost data from an analogous project to derive slope and calculate T_x :** Use this approach when the production processes are analogous.

The choice is ultimately subjective, but important. In the case of the Electronics data, the collected cost for Project 2 is unit 10. All other costs collected represent the first unit cost. Project 2 cost must be adjusted for cost improvement. In this case, the theoretical first unit cost is the relevant analogy to the estimated system. Therefore, the CIC slope from the source data are used to derive the theoretical first unit cost. Given unit theory, as opposed to cumulative average theory [need to document theory], and a source data CIC slope of 95%, the exponent for the CIC equation is calculated:

$$b = \frac{\ln(0.95)}{\ln(2)} = -0.074$$

¹⁶ Also known as a learning curve.

¹⁷ CIC theory is a wide-ranging topic that includes unit and cumulative average theory and is not covered in detail in this handbook. See footnote 20 for additional detail.

¹⁸ Q is commonly used to denote the unit number in the CIC equation $\text{UnitCost} = T_1 * Q^b$, where T_1 is the first unit cost and b is $\text{LN}(\text{Slope})/\text{LN}(2)$.

¹⁹ Goldberg, Andrew S., Touw, Anduin (March 2003) "Statistical Methods for Learning Curves and Cost Analysis", CIM D0006870.A3/1 Rev

Given that the EAC, at 95% complete, of unit 10 (price – fee from **Figure 4**) is \$157.2K, the theoretical first unit cost (T_1) is calculated as:

$$T_1 = \frac{190.1 - 32.9}{10^{-0.074}} = 186.40$$

1.6.5.2 Production Rate-Affected CIC Theory

When production rates decrease, personnel-related expenses per unit tend to increase as fixed costs are spread over fewer units. Additionally, material costs per unit may increase because volume discounts are reduced or eliminated. When production rates increase, there may be the opposite effect driving the unit cost down. **Table 5** illustrates a normalized, notional CIC dataset that is affected by rate when referencing “Lot Quantity.” This dataset is used to demonstrate several aspects of regression analysis.

Table 5: Rate Affected CIC Notional, Normalized Dataset

Collected, Validated and Normalized Data				Calculated From Normalized Data		
	Lot Total Cost FY2016\$K	Lot QTY	Low Rate Initial Production	Ave Unit Price FY2016\$K	First Unit	Last Unit
Year	LotTotCost	Qty	LRIP	AUP	First	Last
2004	18.182	8	1	2.273	1	8
2005	24.975	20	1	1.249	9	28
2006	52.003	35	1	1.486	29	63
2007	37.751	29	1	1.302	64	92
2008	40.240	35	1	1.150	93	127
2009	34.302	35	0	0.980	128	162
2010	27.763	29	0	0.957	163	191
2011	37.289	36	0	1.036	192	227
2012	35.329	38	0	0.930	228	265
2013	36.291	38	0	0.955	266	303
2014	42.899	43	0	0.998	304	346
2015	18.955	18	0	1.053	347	364

Use cost improvement curve (CIC) analysis to normalize data where unit cost, or average unit cost, decreases with quantity.

1.6.6 Cost Per Unit Characteristic

Cost data in terms of total dollars will rarely be directly comparable across projects. This is the reason we perform regression, to determine if there is a statistically significant performance, technical, or time characteristic that helps explain cost. Weight-based CERs are a common estimating method for many commodities. These CERs normalize for weight (e.g., dollars per pound or kilogram) and provide a means to estimate the cost for different sized options. In the context of normalization, deriving cost per unit relationships such as dollars per unit weight, dollars per unit power, dollars per month, etc. can provide some initial insight into the comparability of the data.

1.6.7 Other Normalization Considerations

There are many ways seemingly similar projects have significantly differing characteristics that need to be addressed in the normalization procedure. Two of many possible examples are discussed in this section.

Effectively Different Products: When normalizing data from different projects, significant portions of the products may be treated differently. In the event products differ, recommend normalizing at a lower WBS level.

Joint Programs: When multiple partners participate in the design, development, and production of a program, understand the relationship of each partner and the funding provided. If there is duplicity of work due to multiple parties, normalize for it. Document all assumptions to alleviate any misrepresentations.

More information on data normalization techniques are found in [A.1 Arithmetic](#), both [A.1.1 Basic Operations](#) and [A.1.2 Weights](#). Additionally, [A.3.1.2.1 Mean](#) addresses simple evaluation of data to ensure systems or components are similar enough to include in the initial dataset, even if excluded by further analysis.

1.7 Linking Cost to Schedule

A potentially important parameter to capture during the data collection process is duration. This may not be a simple task since consistently defining the start and end points may not be possible. Characterizing the CER with a duration could provide useful insight into its application to the estimate. To extrapolate a CER associated with programs averaging 12 months in length to a project 18 months in length, there may be a need to adjust the CER. Linking cost to schedule explicitly is beyond the scope of this handbook²⁰.

1.8 Summary and Introducing the Electronics Example Dataset

The data collection process begins with a clear understanding of the [purpose](#), [scope](#), and [expert operational/engineering context](#). From that basis, [data sources](#) can be identified and the [collected](#) raw data can be organized in a consistent and traceable manner. Not all [normalization](#) concepts in this section apply to every data set and further normalization may be required beyond what has been discussed. The goal of data normalization involves taking unlike properties or characteristics and adjusting them in a consistent way for comparison and statistical analysis.

Table 6 contains the normalized data introduced in **Table 2** and **Figure 4**. This is the dataset used throughout the remainder of the handbook.

²⁰ The Joint Agency Cost Schedule Risk and Uncertainty Handbook (JA CSRUH) does provide some guidance on how to convert CERs that estimate total cost to methods that are sensitive to duration.

Table 6 Notional Electronics Data Set

Observation	Collected, Validated and Normalized Data				Calculated Data	
	Cost (FY16\$K)	Power (kW)	Aperture (cm ²)	FFP or T&M 1 or 0	Cost per Unit Power (\$K/kW)	Intensity (kW/cm ²)
Project 1	\$390	10.0	8.7	1	39.000	1.149
Project 2	\$200	5.0	8.0	0	40.000	0.625
Project 3	\$240	5.2	8.2	1	46.154	0.634
Project 4	\$300	7.0		0	42.857	
Project 5	\$460	12.0	9.0	1	38.333	1.333
Project 6	\$560	17.8	9.5	0	31.461	1.874
Project 7	\$700	21.0	9.2	0	33.333	2.283
Project 8	\$800	25.0	9.7	1	32.000	2.577
Project 9	\$500	18.0		0	27.778	

Two calculated independent variables were added to the dataset:

- **Cost per Unit Power:** This is an example of [1.6.6 Cost Per Unit Characteristic](#). In this case, the cost per unit power is quite different across the projects, meaning a simple factor (linear) relationship between cost and power may not explain enough of the variation in the cost data.
- **Intensity:** Intensity is the power per unit area, in this case kilowatts per square centimeter of the aperture. It is an example of creating a potentially useful variable from two collected parameters.

With the data collected, validated, organized and normalized, the next step is to perform statistical analysis on the data.

2.0 STEP 2: ANALYZE NORMALIZED DATA

2.1 Overview

The goals of this section are:

- Determine the cost estimating approach
- Understand the normalized data through descriptive statistics
- If the parametric approach is selected:
 - Identify potential cost drivers
 - Hypothesize functional form
 - Hypothesize error term (additive or multiplicative)

There are many ways to approach the goals of Step 2. The flowchart (where n is the sample size) in **Figure 6** is one.

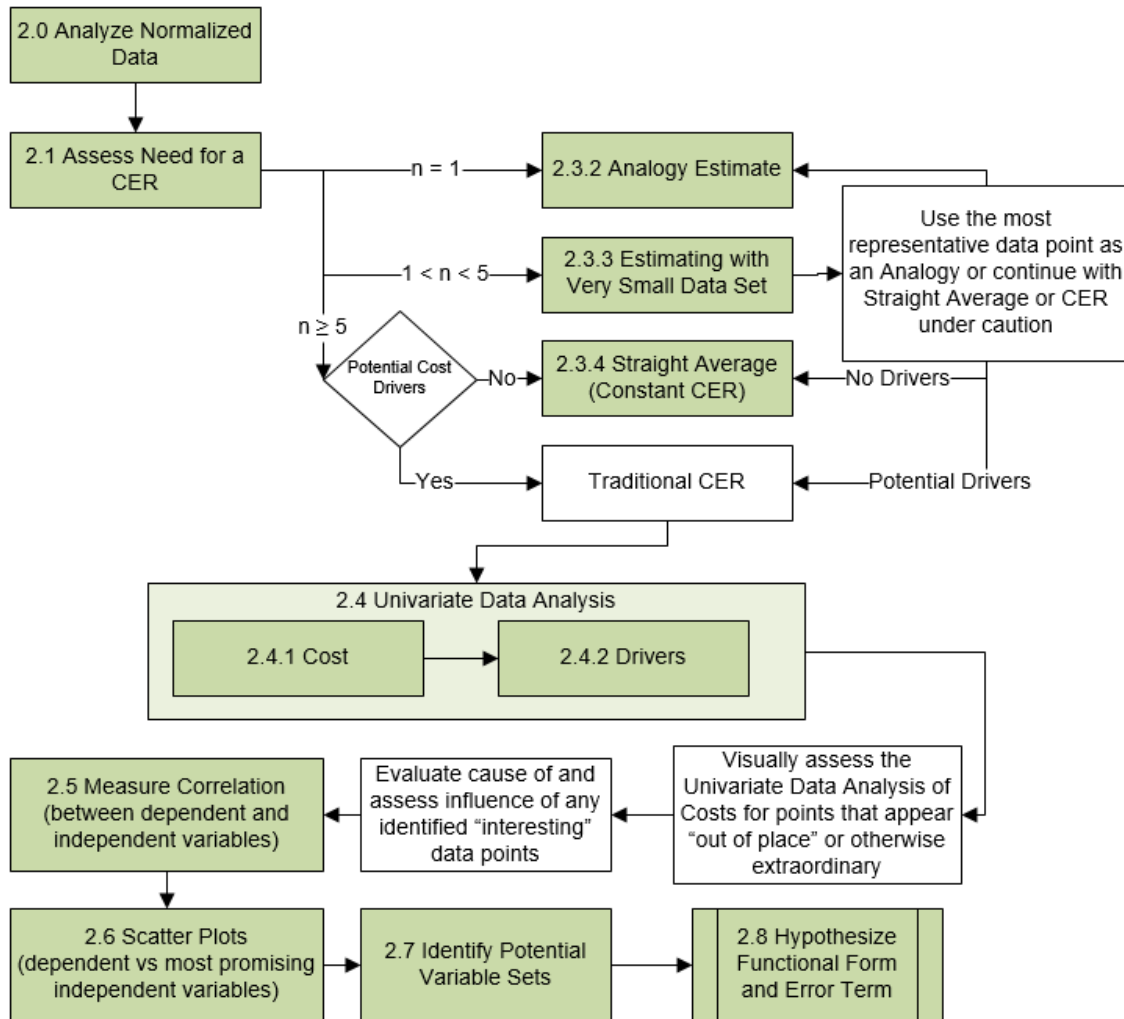


Figure 6: Step 2 - Analyze Normalized Data

Though presented as a logical series of steps, the CER development process tends to be iterative. If any step of the investigation or analysis reveals the need for additional data, return to the [data collection step](#).

2.2 Cost Estimating Methods

Initial analysis of the data allows the analyst to select the appropriate methods from the following:

- **Analogy** estimates by scaling or adjusting historical costs from a similar program to account for key physical, performance, or programmatic metrics. Typically adjustment, or scaling, factors are obtained from the engineering and/or program management teams based on expert opinion of complexity, etc.
- **Straight Average** is the simplest of the statistical estimating methods. This method is simply the average of the historical projects' cost and other statistics.
- **Parametric** analysis uses comparable historical data to develop CER models based upon performance, technical, programmatic and/or schedule data. It extends the analogy process to multiple observations and is the focus of this handbook.
- **Engineering Build-Up** is feasible when significant design information exists to support this method. Two major components of an engineering build-up include:
 - **Labor** estimates based on industrial engineering standards
 - **Material** estimates based on catalog prices, vendor quotes, or comparable items
- **Extrapolation from Actuals** is an estimating technique implemented as program execution proceeds and as program-specific cost data are accumulated.

Remember to document rationale and source for chosen methods.

Engineering Build-up and Extrapolation from Actuals are applicable in the later stages of a new project. They are not discussed further in this handbook. One or more of the following qualitative or statistical reasons may drive the choice between Analogy, Straight Average, or Parametric:

- **Item value:** Focus estimating effort on the larger items, rather than spending a lot of time estimating an item of relatively small value. A simple analogy, even a poor one, may be sufficient for a low value item that is a tiny fraction of the program cost.
- **Difficulty in collecting cost driver²¹ data:** Obtaining quality data may not be possible for hypothesized cost drivers.
- **What-If analysis requirements:** Parametric CERs facilitate a rapid and consistent way to investigate the cost impact of different solutions. They also provide a way to identify project parameters that have the most influence on cost (Section [4.7 CER Responsiveness](#)).
- **Reduce variance of the estimate:** A reason to build a CER is to reduce the prediction interval around the estimated value when compared to a straight average. Even when a straight average result matches the more complex CER point estimate, the [CER prediction error may](#) be significantly lower ([Step 5: Characterize Uncertainty](#)).

²¹ Cost driver, in this handbook, refers to performance, technical, programmatic and schedule characteristics that are shown through statistical analysis, supported by operational or engineering expert judgment to influence cost. Cost contributors are the individual WBS/CES element costs that contribute to the total project cost.

The item's value, the availability of cost driver data, the need to perform What-If analysis, and the ability to reduce the error in the estimate drive the selection of cost estimating method.

2.3 Choosing Between Analogy, Straight Average or a CER

2.3.1 Assess Number of Observations (n)

The first step in data analysis is to assess the available data. Part of that assessment is evaluating the dataset to understand how many complete normalized observations are available for analysis, traditionally denoted with a lowercase n . Here, “complete” means observations with values for both final (or nearly final) actual cost (y) and all independent variables or cost drivers (x_i) of interest. Often there will be “holes” in the data set, thus, n may vary depending on the subset of cost driver variables (Section [2.7 Identify Potential Variable Sets](#)). In the Electronics example introduced in **Table 6** with nine observations, only seven of them have *Aperture* and *Cost* data, thus $n=7$.

At least five observations are required to assess the significance of a linear relationship with one independent variable or to validate normality (Section [4.2.1.4 Normality of Errors](#)). The following cut-off values are rules-of-thumb rough guidelines. The important principle is tailoring the estimating approach to the quantity and quality (completeness, reliability) of available data.

If $n = 1$, proceed to Section [2.3.2 Analogy Estimate](#).

If $n = 2, 3$, or 4 , it is feasible to use a parametric estimate, but an analogy may be more appropriate. In both cases, the alternate method should be used as a cross check. Proceed to Section [2.3.3 Estimating with Very Small Data Sets](#). (There are other options outside the scope of the handbook, such as Bayesian techniques²²).

If $n \geq 5$, develop a parametric CER. Proceed to Section [2.4 Univariate Data Analysis](#).

If the hypothesis is for functional forms containing more than one independent variable, five observations will not be sufficient. A degree of freedom is lost for every coefficient estimated in the hypothesized functional form and for every constraint placed on the regression method (see [3.3.5.5 Zero Percentage Bias Minimum Percentage Error \(ZMPE\)](#)). A common way to think of degrees of freedom is the number of independent pieces of information available to estimate another piece of information. For instance, the mean of the numbers 4, 6, 8 is 6. You can replace any two of these numbers with some other number. However, the third number is now fixed if the mean is to remain 6. Similarly, in regression analysis, subtract the number of unknowns (coefficients in the model) from the number of observations to obtain the degrees of freedom. If the CER functional form has two independent variables, then the number of unknowns will be three (the intercept plus two coefficients). Instead of $n \geq 5$, a more precise guide is $n - k \geq 3$ where k is the number of unknowns in the CER.

²² Smart, C. (14 April 2014) “Bayesian Parametrics: How to Develop a CER with Limited Data and Even Without Data”, Missile Defense Agency, Best Paper Overall ICEAA Workshop June 2014

Why is three (3) the minimum acceptable number of degrees of freedom for regression? In fact, the general literature on regression analysis suggests that a minimum of 10 observations are required for every independent variable. In cost analysis, that is a bridge too far. A simple reason to settle on three is that for less than two degrees of freedom, the standard deviation of the t-distribution is undefined (the area in the tails is too great). Since we use t-distributions for [regression validation](#), three is the absolute minimum, but always strive for more.

If the number of independent variables, k , is known, the minimum number of observations to collect is

$$n - k \geq 3.$$

2.3.2 Analogy Estimate

An Analogy Estimate uses a single historical data point, adjusted using factors for differences between the comparable historical system and the new system (e.g., a factor to account for complexity differences between two similar systems). The scalability or adjustment of the analogy should be discussed with at least one subject matter expert (SME) associated with the program. Remember to document SME rationale, basis of expertise, and contact information.

In addition to scaling a new system, use an analogy to develop cost-on-cost factors to apply to the new estimated system. For example, if training associated with a system accounts for approximately X.X% of the Average Unit Cost (AUC), then training for the new system is approximately X.X% of its AUC.

The assessment of analogy uncertainty may be subjective given the lack of data. The Joint Agency Cost Schedule Risk and Uncertainty Handbook (JA CSRUH)²³ provides a detailed discussion on how to assign uncertainty subjectively.

Less than five observations or the lack of potential cost driver data leads to an Analogy or Straight Average cost estimating method.

2.3.3 Estimating with Very Small Data Sets

With two, three, or four observations, a simple linear CER with a slope and intercept term has 0, 1, or 2 degrees of freedom, respectively, which is generally not sufficient for satisfactory statistical results.²⁴ The following methods are valid however; analysts must be prepared to defend the rationale for their chosen method:

- **Analogy:** pick the most representative data point to use as the basis for an [Analogy Estimate](#) (Section [2.3.2](#)) and consider the others in the adjustment of the Analogy.
- **Traditional Parametric:** be aware of the limitations of many of the statistical techniques to follow. Consider fixing a parameter coefficient as described in [3.4.3 Pseudo-Exact Prior](#)

²³ <http://cade.osd.mil/tools/other-cost-tools>

²⁴ This holds true under the traditional statistical paradigm where degrees of freedom are essential for hypothesis testing. Additional advanced approaches exist for handling small data samples beyond the scope of this guide, including the “Bayesian” approach. See footnote [22](#).

[Information on Parameter Values](#)). This technique will protect degrees of freedom, albeit biasing the results.

- **Bayesian Parametrics:** a technique to develop estimates from limited data. See footnote [22](#).
- See Appendix [A.3.3.1 Small Data Sets](#) for more information on the treatment and consequences of small data sets.

If proceeding with a traditional parametric estimate, first perform univariate data analysis on both the dependent and independent variables. Understanding the mean, median, range, and other statistics helps to understand the limitations of the dataset.

2.3.4 Straight Average

Estimating cost using a straight average of the normalized cost observations yields a single result. While the straight average method is considerably easier to implement, this method will generally result in a larger variance than a CER.

2.3.4.1 Arithmetic Mean

The straight average, or arithmetic mean²⁵, of a data set is the sum of the costs divided by the number of sample observations. The sample mean is an estimate of the population mean (also called the true mean).

$$\text{Sample Mean} = \bar{x} = \frac{1}{n} \sum_{i=1}^{i=n} x_i$$

2.3.4.2 Confidence and Prediction Intervals About the Arithmetic Mean

The confidence interval (CI) estimates plausible values for the population mean for a given probability level.

$$\text{Confidence Interval} = \bar{x} \pm t_{1-\frac{\alpha}{2}, n-1} \hat{\sigma} \sqrt{\frac{1}{n}}$$

$$\text{Sample Standard Deviation} = \hat{\sigma} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

Where \bar{x} is the sample mean, t is the value for the t-distribution with $n - 1$ degrees of freedom giving a probability (area) of $\alpha/2$ in the right-hand tail, and n is the number of sample observations. The sample standard deviation divided by the square root of the number of observations is the definition of the standard error.

²⁵ The arithmetic mean is not always the correct approach to establish the mean of the collected data. For additional detail and guidance, see Hu,S. (June 2010) “Simple Mean, Weighted Mean, or Geometric Mean?”, ISPA/SCEA

A different statistic, the prediction interval (PI), is required to address the probability range for a specific project future cost. The PI bounds the estimate of a future data point, x_0 , with a specified level of probability. The PI expression is:

$$\text{Prediction Interval} = \bar{x} - E < x_0 < \bar{x} + E$$

E is the half-width of the PI. Traditionally, the normal distribution is used when there are 30 or more observations. The t-distribution is used when there are fewer than 30 observations. However, this rule-of-thumb is outdated since modern software, including MS Excel, can easily provide the t-distribution for any number of observations. Some agencies use the t-distribution for up to 100 observations. In cost estimating, the t-distribution is usually used due to the small number of observations. The PI formula is:

$$\text{Prediction Interval} = \bar{x} \pm t_{1-\frac{\alpha}{2}, n-1} \hat{\sigma} \sqrt{1 + \frac{1}{n}}$$

For example (summarized in **Table 7**):

- [Table 6 Notional Electronics Data Set](#) consists of nine cost observations, $n=9$
- Sample mean, $\bar{x} = \$461\text{K}$
- Sample standard deviation, $\sigma = \$203.5\text{K}$
- The CI and PI standard errors are \$67.83K and \$214.50K, respectively
- The CI and PI bounds are derived from a t-distribution with 8 degrees of freedom
- A 95% prediction interval²⁶ (α value of 0.05), MS Excel's inverse t-distribution function T.INV(0.975, 8) yields 2.306. The 95% PI upper bound is $\$461.11 + 2.306 * \$214.50 = \$955.76$

Table 7: Assessing the Accuracy of the Electronics Univariate Analysis

Statistical Output	FY16\$K
Mean	\$461.11
Std Dev	\$203.50
Confidence Interval Standard Error of the Mean	\$67.83
Prediction Interval Standard Error of the Mean	\$214.50
t distribution at 97.5%, 8 degrees of freedom	2.3060
97.5% Confidence Interval Bound	\$617.53
97.5% Prediction Interval Bound	\$955.76

The 95% CI (error of the sample mean represents the population mean) and the PI (error of using the sample mean as a future estimate) for the electronics data are illustrated in **Figure 7**. In this example, there is a 95% probability that the actual price of the new project will be fall between -\$33.54K and \$955.76K. In this example, the lower tail of the PI extends below zero. Modeling this distribution in the

²⁶ Individual agencies may have different requirements or guidance pertaining to the significance level to be used for prediction intervals. Follow local guidance when selecting a significance level.

cost model means the simulation will draw negative cost numbers for some trials. To avoid this situation, the analyst can truncate the distribution at zero (or some other positive threshold established by the data or a SME). However, truncation changes the shape and bounds of the distribution. It eliminates some variance, but may increase the mean. The decision to truncate is an arbitrary one governed by local policy rather than mathematics. There is no consensus on whether to truncate or not. Seek specific guidance from your agency.

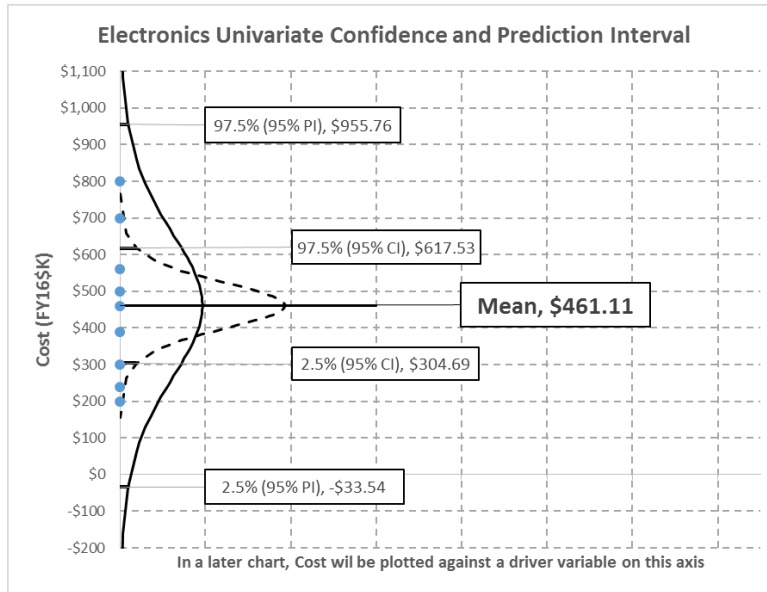


Figure 7: Confidence and Prediction Interval for the Straight Average of the Electronics Cost Data

See [Figure 74: Compare OLS CER PI to a Straight Average PI](#) for an illustration of how much smaller the prediction interval can be using regression to develop a CER.

If basing the estimate on a straight average, then the path through this guide is complete. The PI just discussed accounts for the uncertainty of the estimate (the mean). Proceed to [Step 6: Document CER](#). Otherwise, proceed with CER development by going to [2.4 Univariate Data Analysis](#).

2.4 Univariate Data Analysis

Univariate analysis is statistical analysis of a single variable. It provides a way to develop a greater understanding of each variable before generating scatter plots to uncover relationships.

2.4.1 Significant Digits

The normalized project costs introduced in [Table 6 Notional Electronics Data Set](#) are only available to the closest \$1K because that is consistent with how the data was collected (see [Figure 4: Consolidate Raw Cost Data into a Summary Table](#)). When collecting data, maintain the precision of the data. More significant digits in calculations reduce the impacts of rounding, which may have large implications when

dealing with large numbers, and long lists of numbers that sum²⁷. The final results should be reported using the precision of the original data sources.

A common theme in most references is that “All calculations should be completed prior to any rounding to avoid introducing additional error into the analytical result.”

2.4.2 Descriptive Statistics

Descriptive statistics summarize the features of the collected observations. Descriptive statistics are sample statistics, not population statistics. Even a dataset representing all programs ever executed would represent a sample from the population of all possible programs. Typical descriptive statistics include measures of central tendency, measures of dispersion, kurtosis, and skewness. These measures form the basis of virtually every quantitative analysis of data.

2.4.2.1 Measures of Central Tendency

Measures of central tendency include mean, median, and mode. See [A.3.1.2 Measures of Central Tendency](#) for details. Measures of central tendency are single statistics used to represent the dataset as a whole. In the absence of potential cost drivers, they can be used as the point estimate as described in [2.3.4 Straight Average](#).

2.4.2.2 Measures of Dispersion

Measures of dispersion characterize the variability in the dataset (i.e., how similar or dissimilar the observations are). Typical measures of dispersion are variance, standard deviation, coefficient of variation (CV), and range. The units of variance are the square of the source data units. The units of standard deviation, on the other hand, are the same units as the source data. The CV is the unitless measure of spread calculated by dividing the standard deviation by the mean. See [A.3.1.3 Measures of Dispersion](#) for details.

The variability in the dataset can help identify potential questions to ask. For instance, are all the projects acquiring similar electronics, or are there some very different projects in the dataset, thus explaining why the costs may be different? Do the most expensive projects align with a particular contract type (for instance, FFP or T&M)? Are the intended operating environments different?

In general, a wide range of observations makes a CER more useful. Analyzing measures of dispersion helps identify data elements that may need more attention. A single data point that is far away from the rest of the data (three standard deviations is a reasonable test) may not belong in the dataset and may have an undue influence on the resulting CER (Section [4.3.1 Influential Points](#)). This data point may also be the only accurate data point in the set. Do not eliminate data without a sound reason and if eliminated, document the reason.

²⁷ There are many sources of guidance on the rules for significant figures. One source is: B. Michener, C. Scarlata, and B. Hames, “Rounding and Significant Figures” U.S. Department of Energy Technical Report NREL/TP-510-42626 January 2008.

2.4.2.3 Descriptive Statistics Summary

Table 8 illustrates one way to document most of the descriptive statistics identified in this section. Some observations on these statistics include:

- **Mode:** the most common value occurring in the dataset; **Table 8** does not include the mode because none of the data is repeated.
- **Mean vs Median:** the location of the mean relative to the median provides insight into the skew of the data. The formal assessment is the skewness²⁸ statistic.
- **Range:** establishes the complete set of all possible resulting values of the data, but provides no insight into the dispersion of data between the lowest and highest value.
- **Coefficient of Variation (CV):** provides a unit less measure to compare dispersion across different variables. While the standard deviation is a measure of the data’s dispersion, it cannot be directly compared to other data sets without context. To provide context, divide the standard deviation by the mean to calculate the CV. The lower the CV, the less dispersion in the data.

Table 8: Descriptive Statistics Summary

Observation	Collected, Validated and Normalized Data			Calculated Data	
	Cost (FY16\$K)	Power (kW)	Aperture (cm ²)	Cost per Unit Power (\$K/kW)	Intensity (kW/cm ²)
Number	9	9	7	9	7
Minimum	\$200.0	5.00	8.00	\$27.778	0.625
25 percentile	\$300.0	7.00	8.45	\$32.000	0.892
Median	\$460.0	12.00	9.00	\$38.333	1.333
Mean	\$461.1	13.44	8.90	\$36.768	1.496
75 percentile	\$560.0	18.00	9.35	\$40.000	2.079
Maximum	\$800.0	25.00	9.70	\$46.154	2.577
Range	\$600.0	20.00	1.70	\$18.376	1.952
Quartile Range	\$260.0	11.00	0.90	\$8.000	1.187
Quartile/Range	0.433	0.550	0.53	0.435	0.608
Std Deviation (S)	\$203.5	7.29	0.64	\$5.985	0.772
CV	0.441	0.542	0.072	0.163	0.516

2.4.3 Generate a Histogram

A histogram is a graphical representation of the data dispersion and is a good way to visualize how the data are dispersed (its shape). Determining how to group the data (interval or bin size) is an important part of generating a histogram. A larger bin size reduces noise in the data attributed to small data set sampling. A smaller bin size provides more precision when there is enough data to warrant it. Cost analysis histograms (supporting descriptive statistics) tend to use equal interval bins. Appendix [A.3.1.1.2 Histogram](#) contains more information on bin selections.

²⁸ There are circumstances rarely found in cost analysis where the location of the mean relative to the median does not accurately establish the skew direction.

Figure 8 shows the histogram for the Electronics cost data in **Table 6** for a variety of bin size selections. The y-axis is the frequency associated with the particular cost range. For example, in the 3-bin histogram, four projects cost less than or equal to \$400K, three projects cost greater than \$400K but less than or equal to \$600K and so forth. Note how the distribution shape changes with the number of bins.

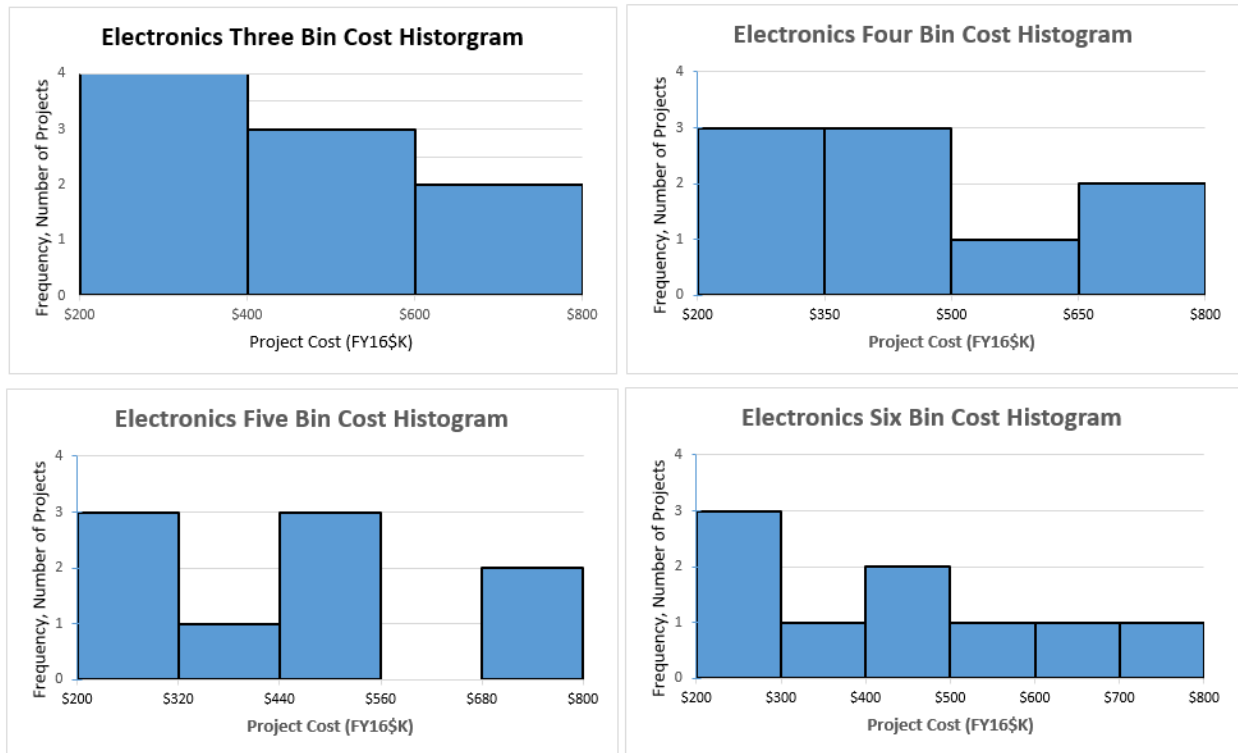


Figure 8: Histogram of Electronics Cost Data

The histogram modal bin as shown in **Figure 8** is the leftmost bin for the 3-bin and 6-bin charts. This is the mode of this particular histogram of the data, not the mode of the dataset. Note that two of the histograms have more than one modal bin.

Histograms are not always useful for small data sets. With just a few data observations, it is very difficult to gather insights into the “true shape” of the data. Creating several histograms as done in **Figure 8** can be useful in understanding where these points lie. Fortunately, as more data are collected, a clearer picture often evolves. For more information, see [A.3.1.1 Statistical Graphics](#).

Perform univariate analysis on the dependent and independent variables to understand the scope of the data, possible limitations and the presence of potential outliers

2.5 Measure Correlation between Dependent and Independent Variables

Information collected as described in section [1.3 Cost Estimate Purpose and Scope](#) of the data collection process provides the basis for identifying potential programmatic, performance and technical parameters that may be cost drivers. The objective is to enhance our understanding of potential response-predictor correlations, and not yet draw definitive conclusions. If this correlation measurement does not identify potential cost driver variables, return to [Step 1: Purpose, Scope, Collect, Validate, & Normalize](#).

Additionally, a good practice is to measure the correlation between multiple candidate drivers and avoid correlated drivers in the same CER or consider techniques as described in [3.3.6 Ridge Regression](#).

The degree that a potential driver variable correlates with the dependent variable helps to identify viable candidates for regression.

2.5.1 Correlation Types

The two correlation coefficients²⁹ that measure the relationship between collected data are:

- **Pearson product-moment (PPM) correlation coefficient**(r): a measure of the linear relationship between two sets of variables³⁰. The PPM correlation formula shown below is used in the MS Excel CORREL() function:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2(y - \bar{y})^2}}$$

where \bar{x} and \bar{y} are the arithmetic means of the respective variables. Use [Table 45: Pearson Product Moment Critical Values](#) to determine if the correlation is statistically significant. See Appendix [A.3.2.2.1 Pearson's r](#) for more detail on this coefficient.

- **Spearman's Rho**: the correlation coefficient uses the same formula as PPM on the ranks of the data instead of the raw values themselves. This method has properties similar to Pearson's r , with values ranging from -1 ("perfect" negative correlation) to +1 ("perfect" positive correlation), and is indifferent to the underlying form of the data. The Spearman coefficient provides the strength and direction of the monotonic³¹ relationship between two variables. [Table 46: Spearman's Rho Critical Values](#) contains a reference table of critical values, though approximations also exist.³² See Appendix [A.3.2.3.1 Spearman's Rho](#) for more detail on this coefficient.

When using the critical value tables, [Appendix D Correlation Critical Value Tables](#), conduct a two-sided test to determine if the correlation is significant at all. If there is an *a priori* reason to expect a positive or negative correlation, use a one-sided test. If the measured correlation is greater than the critical value, then correlation is statistically significant. See Appendix [A.2.1.4 Correlation](#) for more information regarding the mathematical definition of correlation and references for more information regarding its diagnosis and implications.

²⁹ Kendall's Tau (τ) is another measure the ordinal association between two variables, but rarely used in cost analysis. See [A.3.2.3.2 Kendall's Tau](#) for more details

³⁰ P. Garvey (1999), "Do not use Rank Correlation in Cost Risk Analysis", 32nd Department of Defense Cost Analysis Symposium is one of many references to state that PPM is the only appropriate measure of correlation for cost risk analysis

³¹ Monotonic means that as the value of one variable increases, the other always increases or always decreases.

³² Specifically, the p^{th} percentile of Spearman's Rho is approximately $z_p/\sqrt{n-1}$, where the numerator z_p is the corresponding percentile of the standard normal distribution.

Table 9 contains the PPM and Spearman correlation values for the Electronics data shown in Table 6. Both Power and Aperture are highly correlated with cost. While the values in both matrices are similar, a recommended best practice is to compare PPM and Spearman correlation coefficients. Large differences between them may indicate the relationship between the two variables is not linear.

Table 9: Normalized and Composite Variable Correlation Matrix

CORRELATION				
PPM	Cost	Power	Aperture	Intensity
Cost	1.000			
Power	0.981	1.000		
Aperture	0.939	0.943	1.000	
Intensity	0.996	0.999	0.937	1.000

Spearman	Cost	Power	Aperture	Intensity
Cost	1.000			
Power	0.983	1.000		
Aperture	0.970	0.943	1.000	
Intensity	0.995	0.996	0.964	1.000

After reviewing, the PPM and Spearman tables shown above, Power and Aperture are highly correlated. The highlighted composite variable, Intensity (created by dividing Power by Aperture area), has the highest correlation with cost.

All the values in the tables exceed the critical values in [Table 45](#) (0.582) and [Table 46](#) (0.600) for a one-tailed test at a significance level of $\alpha = 0.05$. The test provides sufficient evidence to conclude the observed positive correlations between Cost and Power, Aperture, and Intensity are statistically significant. See Appendix [A.3.2.1 Hypothesis Testing](#) for a general discussion of hypothesis testing and significance levels.

Univariate analysis can be misleading for multivariate CERs and should not be the sole rational used to screen out variables. When primary drivers are understood, identifying the second and third variables to improve the CER becomes a challenge.

*Correlation does not imply causation.
SME input is valuable to ensure relationships make sense.*

2.5.2 Identify Redundant Variables and Potential Multicollinearity

Power and Aperture are examples of two highly correlated variables indicating they have a strong relationship with each other, essentially conveying redundant information. This phenomenon, called multicollinearity, can lead to inaccurate coefficients because of undue influence of the effect of correlated independent variables. These types of variables may convey the same information from a CER perspective. One variable cannot change without the other changing as well. In the case of our notional example, a CER based on both Power and Aperture may lead to inaccurate cost sensitivity because the

variables are strongly correlated. Section [4.3.2 Multicollinearity](#) provides a more in depth discussion on the topic and analytical techniques to mitigate the effects.

2.6 Scatter Plot of the Most Promising Cost Drivers

A scatter plot is a graph of observations displaying the relationship between two variables. An ordered pair (x, y) represents each point, with the dependent variable shown on the y -axis and the independent variable on the x -axis. Many choose to build scatter plots before univariate or correlation analysis to visualize potential cost drivers. When there are only a few drivers to investigate, this is a reasonable first step. However, the descriptive statistics and correlation assessments are very easy to perform and yield useful information that will help to interpret the scatter plots. Performing the analysis and building the scatter plots are important steps in the process. The sequence is not as important.

[Table 9: Normalized and Composite Variable Correlation Matrix](#) highlights the need to develop scatter plots of cost and all independent variables. **Figure 9** is a scatter plot illustrating the relationship between Cost and Power. As Power increases, Cost also increases. This is consistent with engineering judgment obtained in Section [1.3.3](#) that illustrated a strong positive correlation. In the event Cost decreased as Power increased, recommend confirming this non-intuitive relationship with SME input. Visually observing a trend is not enough. The analyst must be able to explain and defend the relationship.

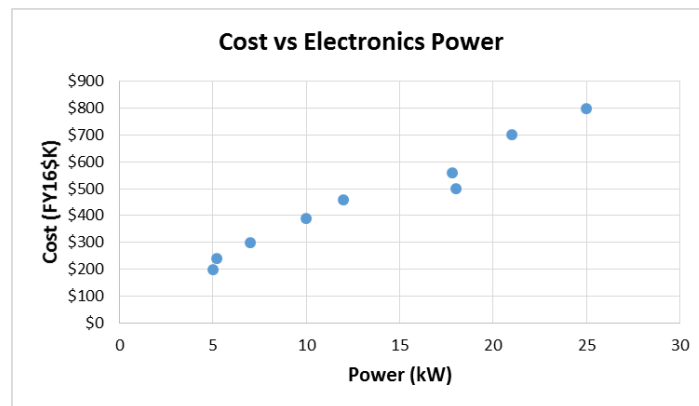


Figure 9: Scatter Plot of Cost vs. Power

Figure 10 shows a scatterplot of Cost versus Aperture. Again, there is a positive correlation, meaning as the Aperture size increases, so does Cost.

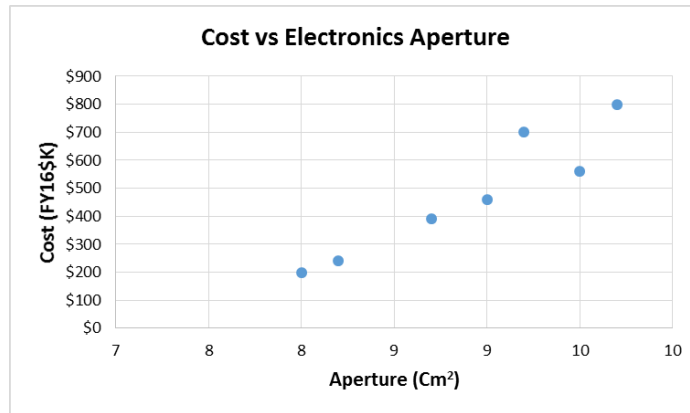


Figure 10: Scatter plot of Cost vs Aperture

Figure 11 illustrates an example of negative correlation between the factor Cost Per Kilowatt (\$/kW) versus Cost. As the factor increases, Cost decreases.

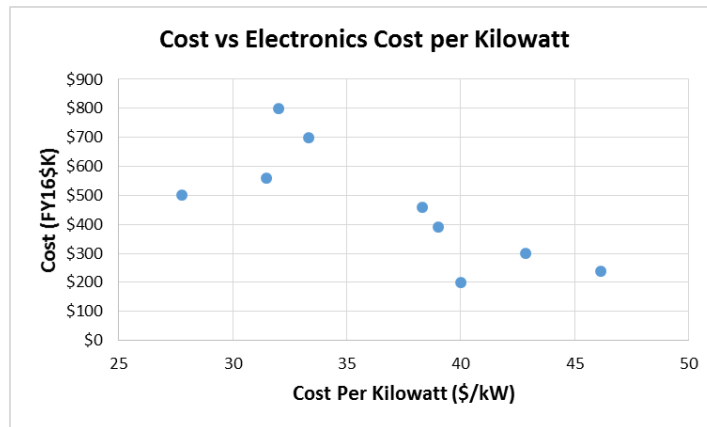


Figure 11: Scatter Plot of Cost vs Cost per Kilowatt

Figure 12 shows the relationship between Power and Aperture. Both of these variables are highly correlated with Cost. As shown in the chart and [2.5.1 Correlation Types](#), they are also highly related to each other and illustrate multicollinearity. The high correlation suggests that Power is related to Aperture. The potential negative effects of multicollinearity may be a reason to avoid using Power and Aperture in the same CER. Understanding this type of relationship is an important step to understanding the dataset. Section [4.3.2 Multicollinearity](#) provides a more in depth discussion on the topic.

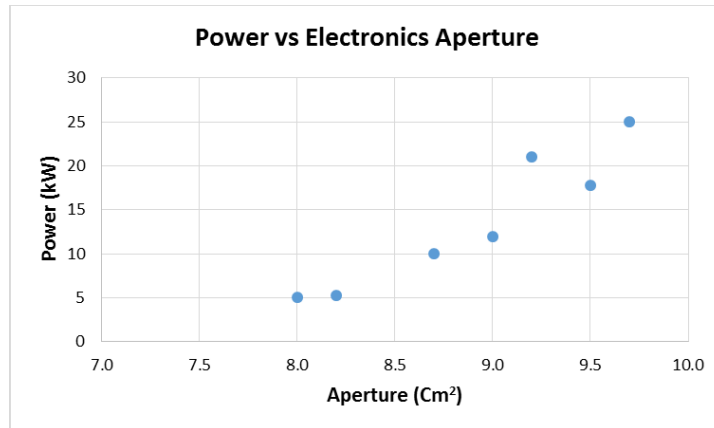


Figure 12: Demonstration of Multicollinearity between related variables

For additional details on scatter plots, see Appendix [A.3.1.1.1 Scatter Plot](#).

2.7 Identify Potential Variable Sets

Working in conjunction with the programmatic and technical experts, use the analysis described above to identify the best variable sets to develop candidate CERs. A CER requires a single dependent variable³³ (usually cost or effort) and one or more independent variables (cost drivers). The focus should be to choose quality programmatic, technical, or performance parameters that are good predictors of the dependent variable.

See Appendix [A.3.1 Descriptive Statistics](#) for more insight on using statistics to help choose the best potential variable sets and interpret their statistical measures.

2.8 Hypothesize Functional Form and Error Term

The following rules apply when trying to determine the best functional form for the data available.

- **The rules of physics must stand:** When dealing with weights, power, and other physical parameters, the equation must logically sound.
- **Keep it simple:** The relationship should be as simple as possible. Adding independent variables that do not significantly improve the statistics complicates the CER, potentially making it too difficult to use, and may needlessly consume degrees of freedom. Do not ignore engineering judgment. Variables historically known to drive cost should be given careful consideration.
- **Stratifying the data, if necessary:** Using dummy variables is a technique to derive relationships affected by categorical differences. A dummy variable can test for a statistically significant multiplier or additive cost that defines how categorical differences influence cost. See [3.2.3 Dummy Variables](#) for details, particularly when combining disparate datasets.

³³ Models with more than one dependent variable are termed multivariate multiple regression. These models are not addressed in this handbook.

- **Use statistics cautiously:** Just because the regression produces a high coefficient of determination, R^2 ([4.5.1.1 R-squared](#)), does not mean the trend line represents a useful relationship. There are other fit and predictive statistics to consider. See [4.0 Step 4: Validate CER](#) for details.

Establishing a sound hypothesis early in the process will guide the steps that follow. Understanding historical cost drivers has a positive influence on data collection and CER development.

Functional form exploration is driven by the research performed while [collecting the data](#), particularly [guidance from SMEs](#), and the results of the [correlation analysis](#). Exploring CER functional forms is an iterative process. Modeling the data often involves testing and re-evaluating functional forms and variable sets many times prior to selecting the “best” equation. Physics and engineering principles (not hypotheses) should supersede statistics in the evaluation process. This guide presents regression³⁴ equation forms in the following order:

- Linear: $y = \beta_0 + \beta_1 x$
- Power: $y = \beta_0 x^{\beta_1}$
- Exponential: $y = \beta_0 e^{\beta_1 x}$
- Logarithmic: $y = \beta_0 + \beta_1 \ln x$
- More General Functional Forms

Where,

y = estimated value

β_0 = y-intercept (vertical calibration)

x = independent variable (driver variable)

β_1 = coefficient of x

Considerations that influence the choice of [Regression Methodologies](#) are as follows:

- **Transform to linear:** Power and exponential forms can be transformed³⁵ to test for a linear relationship. The logarithmic form is already in a linear functional form.
- **Select error term type:** Selection of the error term type is independent of the functional form.

Figure 13 is a flow chart describing a systematic process to explore functional form and error term combinations. The first step is to review the plots created under [2.6 Scatter Plot of the Most Promising Cost Drivers](#) and determine by visual inspection if the data appears to be one of the following patterns:

- Straight line
- Concave up, concave down, where the cost increases/decreases with the independent variable.
- Concave down, where the cost decreases with the independent variable. A cost improvement curve (where cost decreases with quantity) is a common example.

³⁴ In cases where functional form cannot be solved using regression techniques, other techniques may be required.

³⁵ See Appendix [A.1.1 Basic Operations](#) for additional information on working with exponents and logarithms.

- Steep concave up, concave down

Other nonlinear forms are possible, but the above represents the most common. All of these examples can be modeled with a functional form that is either linear or can be transformed to linear (e.g., power, exponential, logarithmic).

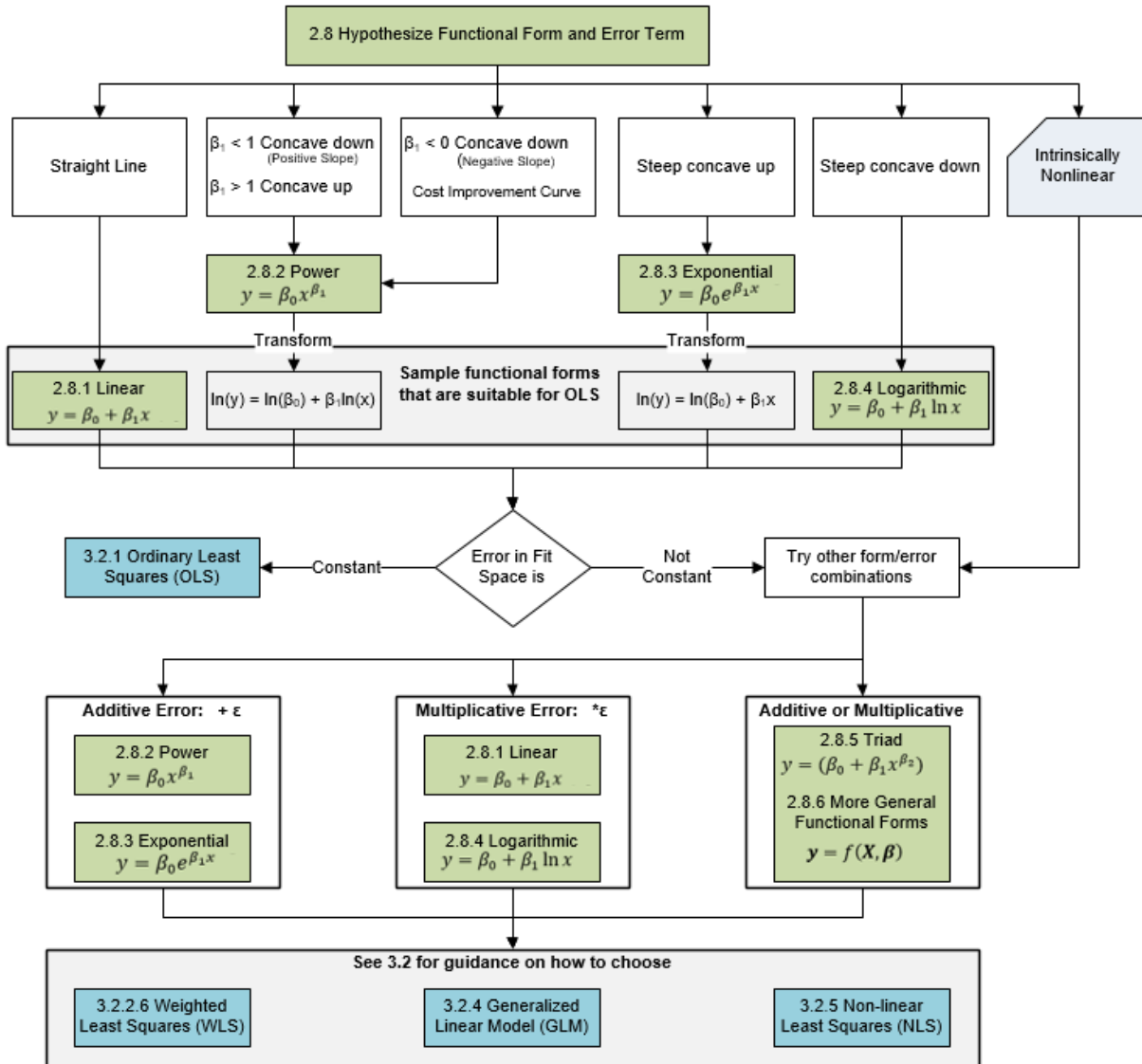


Figure 13: Selecting a Functional Form and Error Term

More complicated functional forms are possible, such as adding a non-zero y -intercept to the power form to create the triad form (Section [2.8.5 Triad Functional Form](#)), or combining multiple integer-power terms to create a polynomial. These often require more advanced regression methods and are discussed in the following sections.

Using [3.3.1 Ordinary Least Squares \(OLS\)](#) as a diagnostic tool helps to understand the data, even with small datasets. A rigorous assessment of the linear model will determine if a linear functional form and an additive error term are appropriate. If not, other methods may need to be considered as described in [Step 3: Generate CER](#).

If there is strong evidence the hypothesized relationship is intrinsically nonlinear or the error term is non-constant, then functional form/error methods other than linear/additive deserve attention, such as [3.3.2.2 Weighted Least Squares \(WLS\)](#) or [3.3.5 Non-linear Least Squares \(NLS\)](#).

Table 10 supports the following functional form discussion. The cost data in the Linear column is from **Table 6** and has been sorted on Power (kW) to visualize the data range. The other cost columns in **Table 10** contain different results from each functional form described below:

- **Power** $0 < \beta_1 < 1$: follows a concave down pattern. The β_1 represents the exponent in the power form equation.
- **Power** $\beta_1 > 1$: follows a concave up pattern
- **Exponential**: follows a steeper concave up pattern
- **Logarithmic**: follow a steeper concave down pattern

Table 10: Notional Data to Demonstrate Functional Forms

Observations	Cost \$K FY2016					Power (kW)
	Linear	Power $0 < \beta_1 < 1$	Power $\beta_1 > 1$	Exponential	Log arithmic	
Observation 1	\$200	200.0	100.0	200.0	200.0	5.0
Observation 2	\$240	215.0	41.0	116.0	250.0	5.2
Observation 3	\$300	329.0	82.0	152.0	330.0	7.0
Observation 4	\$390	472.0	154.0	191.0	475.0	10.0
Observation 5	\$460	448.0	242.0	254.0	550.0	12.0
Observation 6	\$560	575.0	497.0	415.0	750.0	17.8
Observation 7	\$500	625.0	474.0	395.0	700.0	18.0
Observation 8	\$700	706.0	655.0	533.0	775.0	21.0
Observation 9	\$800	800.0	800.0	800.0	800.00	25.0

The following charts illustrate the concave down and up patterns and functional forms they fit.

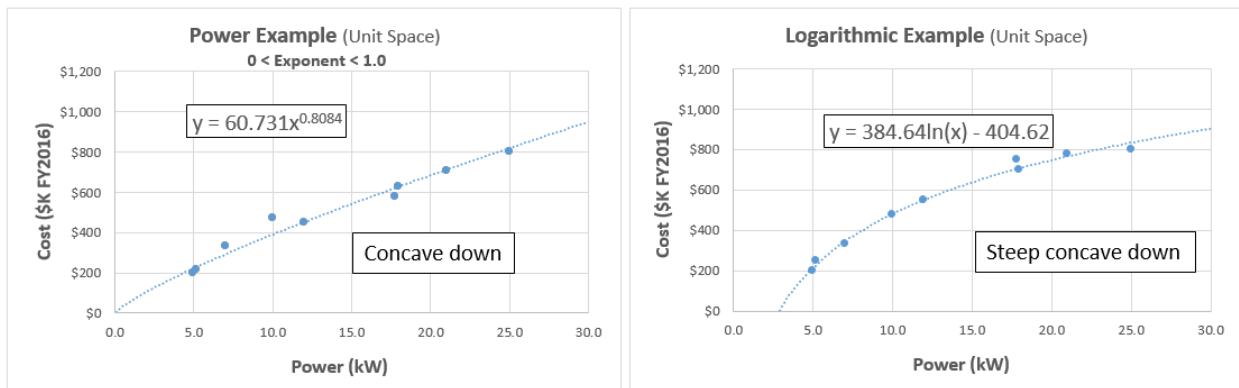


Figure 14: Concave Down Patterns

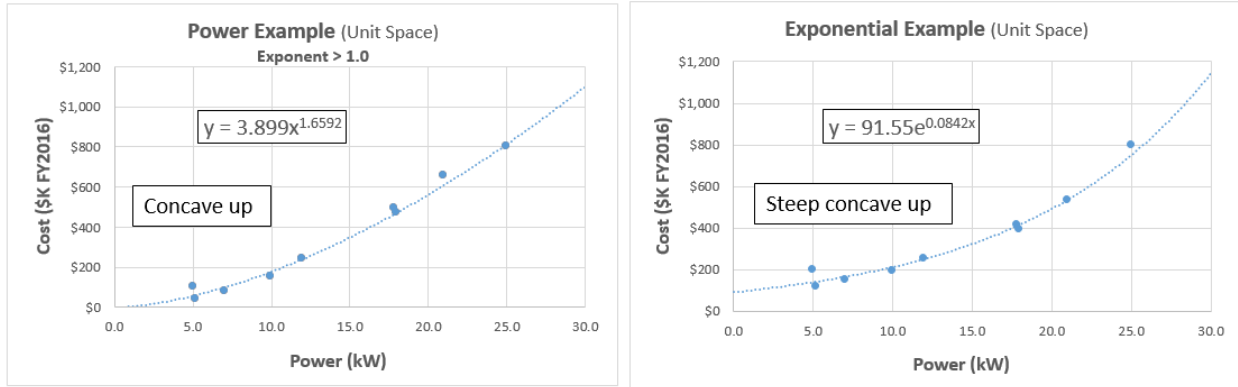


Figure 15: Concave Up Patterns

2.8.1 Linear Functional Form

The simplest functional form is a linear equation with a single explanatory variable used to predict the dependent variable. The equation is in the form:

$$y = \beta_0 + \beta_1 x$$

Where,

- y = estimated cost
- β_0 = y-intercept (vertical calibration)
- β_1 = coefficient of x (constant change in cost per constant in x)
- x = independent variable (cost driver)

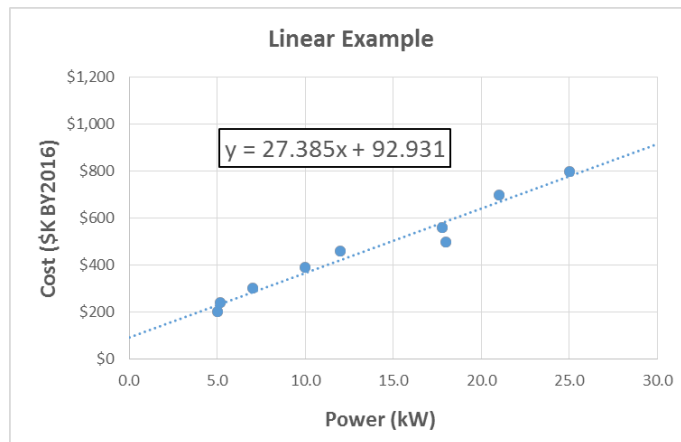


Figure 16: Linear Functional Form Example

Mathematically, the y-intercept is the value the equation yields for an x -value of zero. Unless the hypothesized model has a specific role for the y-intercept, avoid ascribing meaning to it. The y-intercept consumes another degree of freedom but provides a vertical adjustment allowing the prediction line (CER) to better fit the data. Be sure the point being estimated lies within the relevant data range. When the linear CER intercept is excluded or has a value of zero, the coefficient β_1 is called a factor (rather than a slope). Section [4.4.2.1 Intercept Term](#) discusses the intercept and its validation in more detail.

The coefficient β_1 is the crucial component of a simple linear CER, representing the slope of the line. For every unit increase (constant change) in x , there is a β_1 -unit increase (constant change) in y . This same relationship holds regardless of the value of x . A positive slope ($\beta_1 > 0$) indicates positive correlation between cost (y) and the cost driver (x) (e.g., system power). Cases where cost increases as the cost driver gets smaller (e.g., microchip size) indicate the line will have a negative slope ($\beta_1 < 0$).

When data does not demonstrate a linear relationship, transforming the dependent and/or independent variable(s) can allow for the use of linear analysis techniques to analyze non-linear data. If a scatter plot forms a discernible curve, then transformations often cause a linear pattern to emerge. This transformation is demonstrated for the Power and Exponential functional forms. The focus is currently on a single independent variable (x).

This simple linear model extends to the multiple variable cases: $y = \beta_0 + \beta_1x_1 + \beta_2x_2$ and so on. Each cost driver (x_i) has an independent additive relationship with cost (y).

2.8.2 Power Functional Form

Power Models take the form:

$$y = \beta_0x^{\beta_1}$$

Where,

y = estimated cost

β_0 = coefficient of x (multiplicative scaling)

β_1 = exponent of x (response of cost, percent change per percent change in x)

x = independent variable (cost driver)

The scatter plots of power functions appear curved, with the shape dependent upon the exponent β_1 . While a fixed percent increase in x always yields a fixed percent increase in y , regardless of position on the curve, the rate of change (slope) does change for different values of x as illustrated in **Figure 17**. For an exponent between zero and one, cost increases at a decreasing rate. For an exponent greater than one, cost increases at an increasing rate (as the cost driver increases). Even when the curved pattern is quite subtle, the regression statistics may demonstrate this is a better choice than linear.

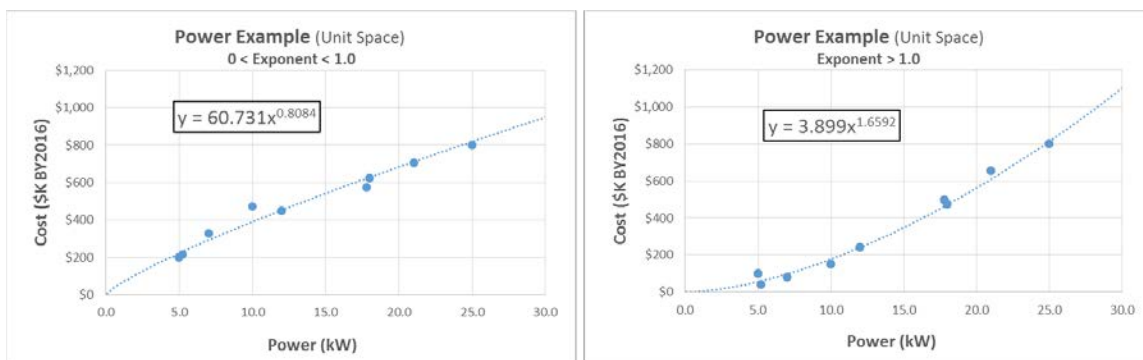


Figure 17: Power Functional Form Examples

The nonlinear Power equation may transform to a linear form as follows:

Initial Power Equation: $y = \beta_0 x^{\beta_1}$

Logarithmic Transformation of the Power Equation: $\ln(y) = \ln(\beta_0) + \beta_1 \ln(x)$

Logarithms turn multiplication into addition (products to sums), and exponentiation into multiplication. Instead of plotting x on the horizontal axis, plot $\ln(x)$. Likewise, plot $\ln(y)$ on the vertical axis, and if a power relationship is the true underlying form, expect a linear pattern to appear. In so-called “log space”, $\ln(\beta_0)$ is the y -intercept and β_1 is the slope of the transformed equation. **Figure 18** demonstrates the power equation in both unit space and the log transformed fit space.

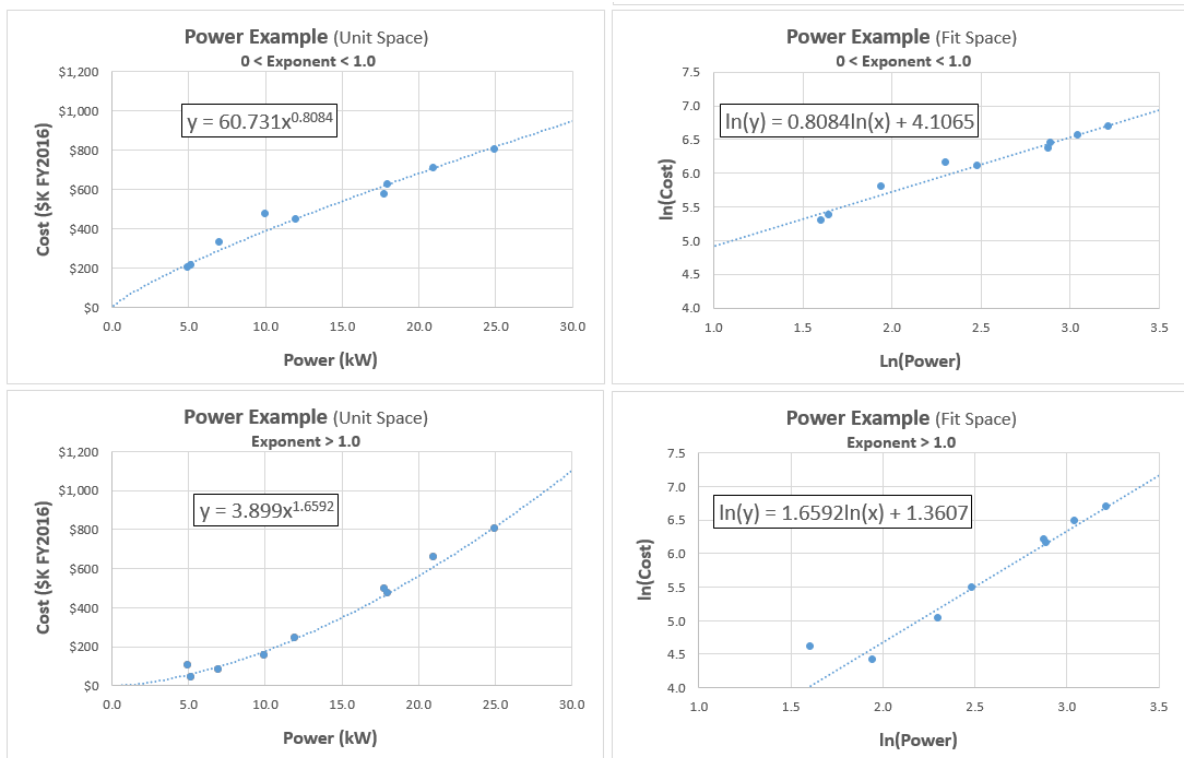


Figure 18: Power Equation in Unit and Fit Space

The case when the exponent is negative results in a negative correlation. A common case of this in cost estimating is CIC analysis, where: $-1 < \beta_1 < 0$.³⁶ A CIC is shown in **Figure 19**. The purpose of this section is not to discuss CIC theory, though it is important to note that CICs use a power relationship to relate effort to the unit number.

³⁶ In CIC applications, common notation is $y = ax^b$ for the power function where the independent variable is quantity. In the case of lot data, the independent variable is the Lot Plot Point (also called the Lot Midpoint).

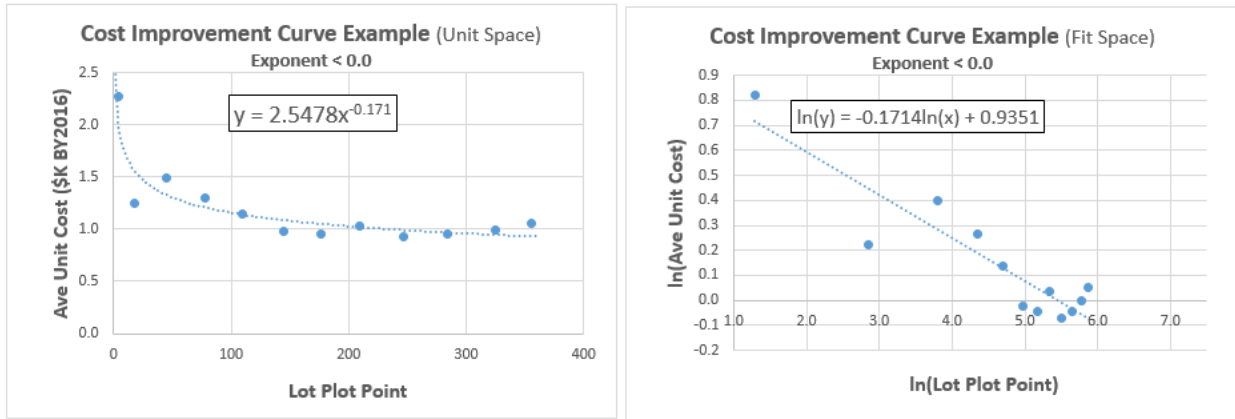


Figure 19: Example Cost Improvement Curve

2.8.3 Exponential Functional Form

Exponential Models take one of three equivalent forms:

$$\begin{aligned}
 y &= \beta_0 e^{\beta_1 x} \\
 &= \beta_0 k^x \\
 &= \beta_0 (1 + r)^x
 \end{aligned}$$

Where:

- y = estimated cost
- β_0 = coefficient (multiplicative scaling)
- $e^{\beta_1} = k = (1 + r)$ = base of the exponential function
- r = rate (percentage change per constant change in x)
- $e = 2.71828 \dots$ (base of the natural logarithm)
- x = independent variable (cost driver)

The three forms of the exponential equation are mathematically equivalent. The first is generally preferred when performing regression (which enables solving for β_1 directly). The second is simplest and thus easiest for algebraic manipulations. The third is most intuitive and relevant for non-CER cost applications, where r represents an inflation or interest rate.

Note the Power Model has a fixed exponent with x in the base. Now x is in the exponent with a fixed base. Exponential models are common in the physical and biological sciences, characterizing exponential growth (e.g., populations) and decay (e.g., radioactivity), but in cost estimating efforts they are mostly relegated to capturing economic effects over time.

The exponential equation form also transforms to a linear equation.

Initial Exponential Equation: $y = \beta_0 e^{\beta_1 x}$

Logarithmic Transformation of the Exponential Equation: $\ln(y) = \ln(\beta_0) + \beta_1 x$

As with Power, plot $\ln(y)$ on the vertical axis, but this time simply plot x on the horizontal axis. Again, expect a linear pattern to appear; assuming an exponential relationship governs the underlying data. On this “semi-log” plot, $\ln(\beta_0)$ is the y-intercept and β_1 is the slope of the transformed equation. **Figure 20** illustrates the Exponential Model.

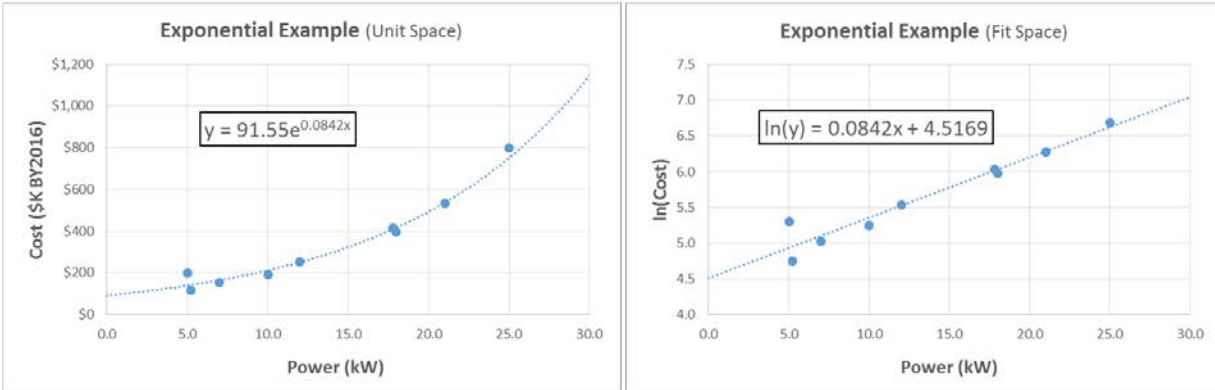


Figure 20: Exponential Functional Form in Unit and Fit Space

2.8.4 Logarithmic Functional Form

Logarithmic Models take the form:

$$y = \beta_0 + \beta_1 \ln x$$

Where,

y = estimated cost

β_0 = y-intercept (vertical calibration)

β_1 = coefficient of $\ln(x)$ (constant change per percentage change in x)

x = independent variable (cost driver)

This is the rarest of the main four functional forms in cost estimating. Similar to a Power function with $0 < \beta_1 < 1$, cost increases at a decreasing rate, but with extreme flattening as it takes an ever larger increase in x to yield the same amount increase in y .

Unlike Power and Exponential, the Logarithmic equation is already in a linear form. However, to see the linear pattern, plot the data in semi-log space to perform the fit. **Figure 21** illustrates the logarithmic form.

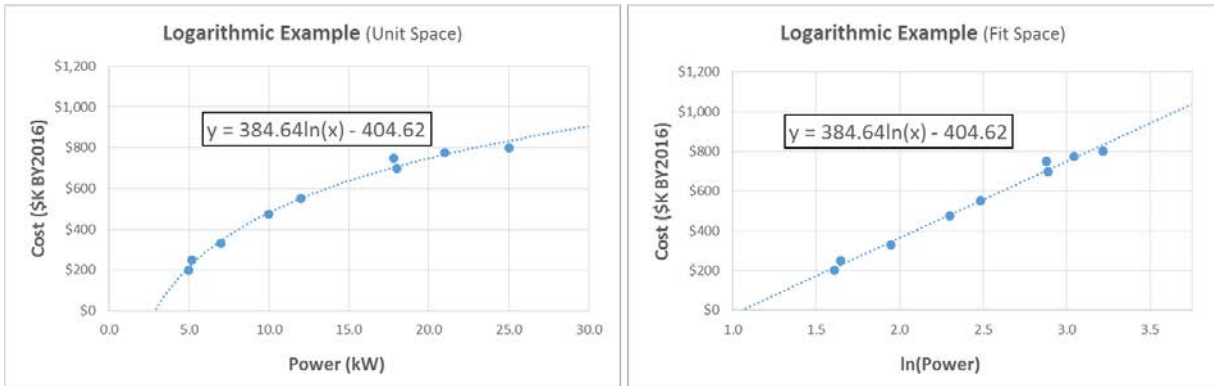


Figure 21: Logarithmic Functional Form in Unit and Fit Space

2.8.5 Triad Functional Form

[Ordinary Least Squares \(OLS\)](#) regression handles (at least initially) the four preceding functional forms and consistent combinations thereof. However, combining additive and multiplicative relationships produces more general functional forms not readily handled by OLS. The simplest and most common of these is the triad CER, named after its three parameters: a non-zero y -intercept, β_0 ; a multiplicative coefficient, β_1 ; and an exponent, β_2 .

$$y = \beta_0 + \beta_1 x^{\beta_2}$$

Where,

y = estimated cost

β_0 = y -intercept (vertical calibration)

β_1 = coefficient of x (multiplicative scaling)

β_2 = exponent of x

x = independent variable (cost driver)

Use of this functional form requires [Non-linear Least Squares \(NLS\)](#), though in some cases, the artful use of an additive transformation can preserve the use of OLS (e.g., subtracting off the vertical calibration to scale back down to zero).

2.8.6 More General Functional Forms

Combining additive and multiplicative relationships produces more general functional forms not readily handled by OLS. These forms can take on nearly any shape and therefore are very flexible. As a result, they can also be dangerous due to the temptation to “data mine” arbitrary forms with little to no practical meaning. Use these forms when the hypothesized relationships require such a non-linear form.

2.8.7 Caution When Regressing Transformed Data

CER regression fits are performed in the transformed or fit space. Use the fit space to check model assumptions (Section [4.2 Model Assumptions](#)), run model diagnostics (Section [4.3 Model Diagnostics](#)), and to check for statistical significance (Section [4.4 Model Significance](#)). However, ascribing meaning to the coefficient statistics requires transforming back to, or calculating in, unit space.

While transforming the data provides a convenient method for analyzing certain non-linear relationships, doing so often has less than desirable implications on a model. The resulting equations are biased ([Step 5: Characterize Uncertainty](#)) and rarely generate the model with the lowest error in unit space. Alternatives to performing OLS on transformed data are provided in [3.3.4 Generalized Linear Model \(GLM\)](#) or [3.3.5 Non-linear Least Squares \(NLS\)](#).

2.8.8 Error Terms

The regression process requires the functional form to include an error term, ε . The error term accounts for the variation between the fitted functional relationship and an observed, actual value. The regression process finds the equation minimizing the error between the actual and predicted dependent values. The manner in which this is defined and accomplished is the subject of [Step 3: Generate CER](#). In general, there are two distinct choices for an error term:

- **Additive:** This error term is most often associated with the linear model; however, this term can be associated with any functional form. An additive error term is appropriate if the difference between the observed values and the fitted CER is constant over the range of the data.
- **Multiplicative:** This error term is often associated with the nonlinear growth models; however, this term can be associated with any functional form. A multiplicative error term may be appropriate if:
 - The errors are proportional to the dependent variable value.
 - The range of the dependent value spans an order of magnitude or more.

The choice of the error term should be made independent from the functional form selection. A good place to start is with the OLS method, which assumes an additive error term on a linear model.

3.0 STEP 3: GENERATE CER

The goal of Step 3 is to conduct regression analysis to establish the coefficients of the proposed equation and calculate associated statistics, using the normalized data. This guide strives to provide examples when possible. While MS Excel can implement some of the techniques³⁷, the results for even the most simplistic methods, such as [Ordinary Least Squares \(OLS\)](#) as performed by the LINEST() function and Data Analysis ToolPak Regression macro, are not sufficient for [Step 4: Validate CER](#). Additionally, more advanced non-linear methods require the use of problematic numerical techniques when relying on MS Excel's built in Solver functionality. In those cases, tools such as CO\$TAT, R, or SAS JMP are more effective and return more detailed statistical results.

After the analyst completes [Step 1: Purpose, Scope, Collect, Validate, & Normalize](#) and [Step 2: Analyze Normalized Data](#) including hypothesizing the most appropriate form of the equation(s) ([2.8 Hypothesize Functional Form](#)), CER generation involves the following steps:

1. Estimate model coefficients using regression and other techniques as appropriate.
2. Assess and validate the appropriateness of the fit model ([Step 4: Validate CER](#)). If unsatisfactory, return to item 1, above.
3. Use the CER to estimate cost and quantify the error using prediction intervals.

Regression methods described in this handbook include:

- [3.3.1 Ordinary Least Squares \(OLS\)](#)
- [3.3.2 Generalized Least Squares \(GLS\)](#)
 - [3.3.2.2 Weighted Least Squares \(WLS\)](#)
- [3.3.3 Transformable Linear and the Log-Linear Model](#)
- [3.3.4 Generalized Linear Model \(GLM\)](#)
- [3.3.5 Non-linear Least Squares \(NLS\)](#)
 - [3.3.5.4 Minimum Unbiased Percentage Error \(MUPE\)](#)
 - [3.3.5.5 Zero Percentage Bias Minimum Percentage Error \(ZMPE\)](#)

Figure 22 shows one possible sequence of steps to select a regression method to generate a CER.

³⁷ Simple regression forms have closed form solutions that can be solved in matrix algebra in Excel, as well as many excellent statistics packages.

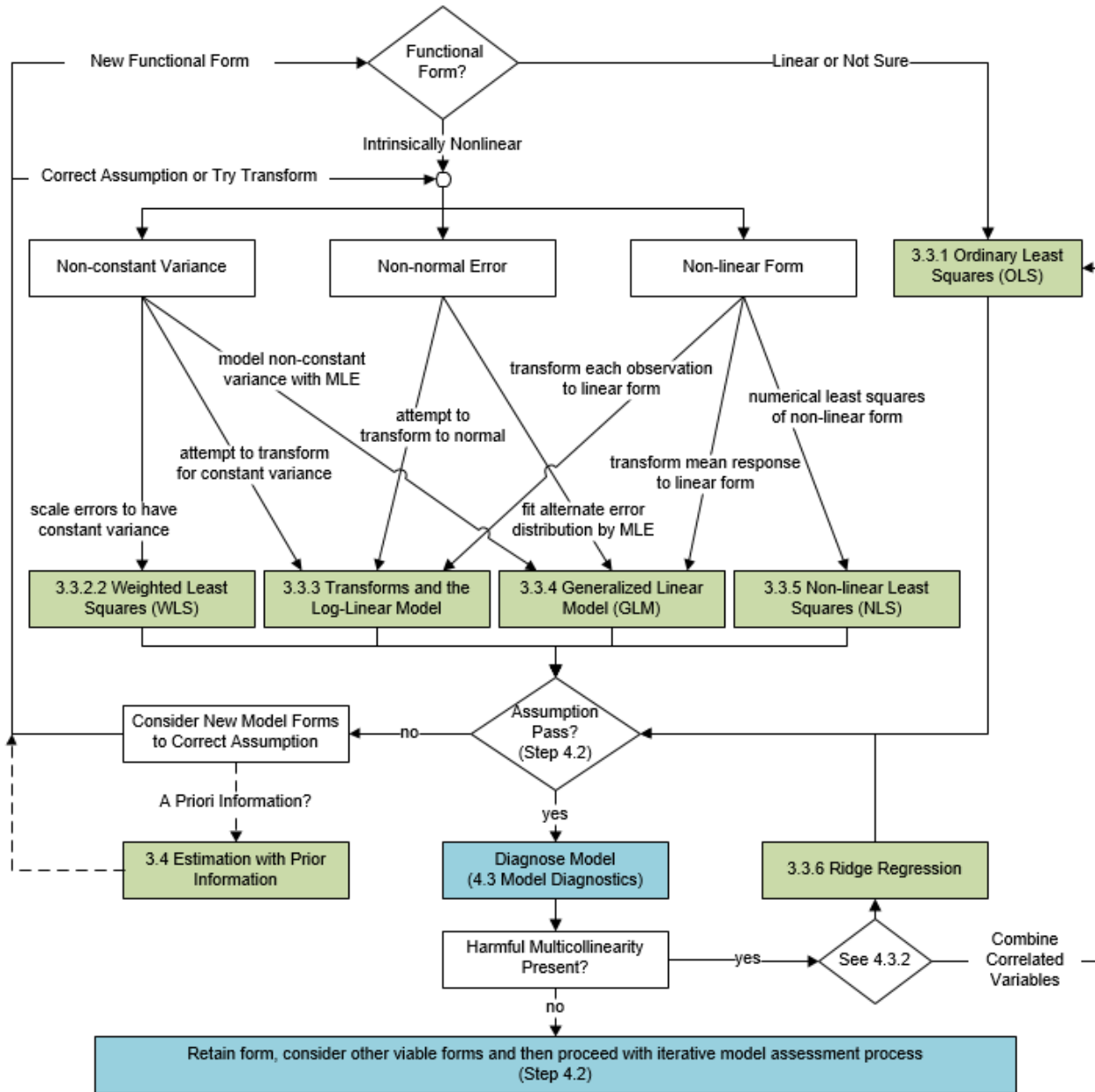


Figure 22: Step 3: Generate CER

3.1 A Guide to Regression Methodology Selection

The flow chart in **Figure 22** identifies the DoD recommended approach to determine the best regression method for a particular dataset. The sections below summarize properties and ratings of regression methodologies from a perspective of the science. The statistical perspective is often the best choice. However, there may be valid reasons to override the statistical results due to underlying physics or other considerations.

3.1.1 Using OLS as a Regression Method Baseline

A fundamental regression methodology is the linear model with a constant, normally distributed error, known as OLS. OLS remains the most popular and well-known regression method in cost analysis. For

this reason, OLS is the starting point. There are well defined tests on the OLS results to help guide the search for a more suitable method when OLS fails to perform.

There are four basic mathematical assumptions underpinning the use of the OLS model:

- (1) **Independence of errors:** Errors, ε_i , are not related
- (2) **Homoscedasticity:** Errors, ε_i , have constant variance, σ^2 , across the data range
- (3) **Normality:** The errors, ε_i , are normally distributed
- (4) **Linearity:** The relationship is defined by a constant slope in fit space

The strategy is to fit OLS and use resulting information to accept the model or diagnose inherent problems. Alternative methodologies may mitigate these problems. **Table 11** identifies the regression methodologies to use when a statistical assumption is violated using OLS. A green check indicates the corresponding regression method is an appropriate option.

Table 11: Summary of Regression Methodologies

	OLS	WLS	Log-linear	GLM	NLS
Statistical Assumptions					
(1) Independence	<i>time series methodology</i>				
(2) Homoscedasticity	✗	✓	✓	✓	✗
(3) Normality	✗	✗	✓	✓	✗
(4) Linearity	✗	✗	✓	✓	✓

Violation of the first assumption ([4.2.1.2 Independence of Errors](#)) suggests a time series model, which is beyond the scope of the handbook. In cases where there is a “x”, the analyst is recommended not to use the identified methodology. The following sections describe how methodologies marked with a check mark may overcome the assumption violations.

3.1.2 Choosing Where to Go When One or More OLS Statistical Assumptions is Violated

Once valid methodologies have been established to remedy problems with the OLS model, there are practical and mathematical properties that help guide the selection of a specific method. **Table 12** summarizes these metrics and brings to light an apparent hierarchy of methodologies. However, when OLS fails on one or more of the underlying assumptions, alternatives must be considered:

- If the only problem with OLS is that the data do not have a constant variance (e.g., multiplicative error), then [WLS](#) is a good choice. [MUPE](#) is a special case of the WLS method.
- If there is a problem with the normality assumption and/or linearity as well, Log-linear and [GLM](#) models can be used as a remedy.
- NLS is the most flexible, but least preferred method due to unfavorable statistical properties. It should be used when subject matter expertise drives towards a model form that cannot be fit by any other methodology. Use NLS when the hypothesis dictates that this method is the best choice.

Table 12 summarizes the following:

- **Engineering Assumptions:** This group of analytical assumptions should be discussed with the appropriate subject matter expert prior to model selection.

- **Mathematical Properties:** indicates which results are unbiased and generates solutions with minimum error. This section also highlights properties that have exact statistics and uncertainty metrics rather than estimated or asymptotic ones.
- **Summary Scorecard³⁸:** is a subjective illustration to summarize the ease of use, the ease of interpretation, and the statistical properties associated with each methodology. Preferences should be for models with simple interpretations and strong statistical properties.

Table 12: Summary of Methodology Properties

	OLS	WLS	Log-linear	GLM	NLS	
Engineering Assumptions						
"Multiplicative" Error Term	✗	✓	✓	✓	✓	✓ Yes
Non-linear Functional Form	✗	✗	✓	✓	✓	✗ No
Mathematical Properties						
Unbiased Coefficients	✓	✓	✗	!	✗	! Not Always
Minimum Unit Space Error	✓	✓	✗	!	!	
Exact Statistics	✓	✓	✓	✓	✗	
Exact Uncertainty Estimates	✓	✓	✓	✓	✗	
Summary Scorecard						
Ease of Application	○	◐	◐	●	○	○ Average
Ease of Interpretation	○	◐	○	◐	◐	
Statistical Properties	○	◐	○	◐	●	● Worst
Summary	○	◐	○	◐	●	

3.2 Select Variable Set

3.2.1 Value of Prior Information

Small sample sizes are the norm in defense cost analysis given limited acquisition programs and outcomes. Collection of information is further constrained by policy, sensitivity of data, cost, and long-lead times. Given this backdrop of limited degrees of freedom, any information regarding the values of regression parameters, or their relationships to one another, can be valuable in increasing the precision of estimates.

The first step, in using prior (but imperfect) knowledge is assessing the fidelity of the information and its potential contribution to the regression equation. Information hopefully represents at least an expected value of a parameter, culled from a distribution with a suitably small CV. Ideally, the information fills a gap in the knowledge of a parameter whose importance in the regression equation is well established.

The value of *a priori* information decreases with sample size. This is because the variances of unconstrained estimators generally decrease when adding more observations to the mix. Unfortunately, the DoD cost analysis environment may not provide sample sizes that can be demonstrated to meet the OLS assumptions defined earlier in this text. Prior information is useful and can be valuable.

³⁸ Add detail regarding where the assessments come from.

3.2.2 Overview

Regression analysis can only provide the answer to the specific question asked. In the context of CER Development, the question is generally: What are the best-fit coefficients for an equation relating cost to a certain set of independent variables using a certain functional form? Each variable set results in a different CER. While this process flow is worded in the singular (“CER”), analysis often requires iteration through many different variable sets.

3.2.3 Dummy Variables

One reason for using a dummy variable is to potentially differentiate between different types of impacts within each categorical group when adequate data are available. When inadequate data exists, data may be combined using dummy variables.

The range and usefulness of the classical regression model $y = X\beta + \varepsilon$ is occasionally expanded through the inclusion of dummy (also known as categorical, binary, qualitative, or Boolean) variables that may represent:

- Temporal effects (e.g., fleet OPTEMPO during wartime versus peacetime)
- Spatial effects (e.g., manufacturing labor rates today at different defense firms)
- Qualitative variables (e.g., nuclear versus non-nuclear propulsion)
- Broad grouping of quantitative variables (e.g., compensation for active and reserve component military members)

The simplest case to consider is the binary situation where the categorical variable takes on one of two values. One example of this is the inclusion or exclusion of a characteristic from data observations. For example, a major cost driver on a sea system may be due to nuclear power. A dummy variable with a value of 1 when nuclear powered and 0 when not nuclear powered could be introduced. The following equation relates ship end cost (y), ship displacement (x_1), and type of propulsion (x_2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Where,

$$x_2 = nuclear = \begin{cases} 1 & \text{if nuclear powered} \\ 0 & \text{otherwise} \end{cases}$$

Taking conditional expectations with respect to nuclear and non-nuclear propulsion (fixing at specific values of x_2) gives,

$$\begin{aligned} E(Y|X_2 = 0) &= \beta_0 + \beta_1 x_1 \\ E(Y|X_2 = 1) &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned}$$

This means that β_2 measures the cost impact of using nuclear power. As the left graphic in **Figure 23** shows, this is equivalent to a shift in the regression line holding the slope constant. The y-intercept changes position, either up or down, depending upon the sign of β_2 .

In rare cases, values of a dummy variable may impact only the y-intercept and in other cases may impact the y-intercept and the slope. Visual inspection of the data may indicate the need for the following method of dummy variable use. The following equation demonstrates the latter case for the same ship propulsion:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Taking conditional expectations yields,

$$E(Y|X_2 = 0) = \beta_0 + \beta_1 x_1$$

$$E(Y|X_2 = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1$$

And so β_3 measures a deviation in slope, shown by the right graphic in **Figure 23**.

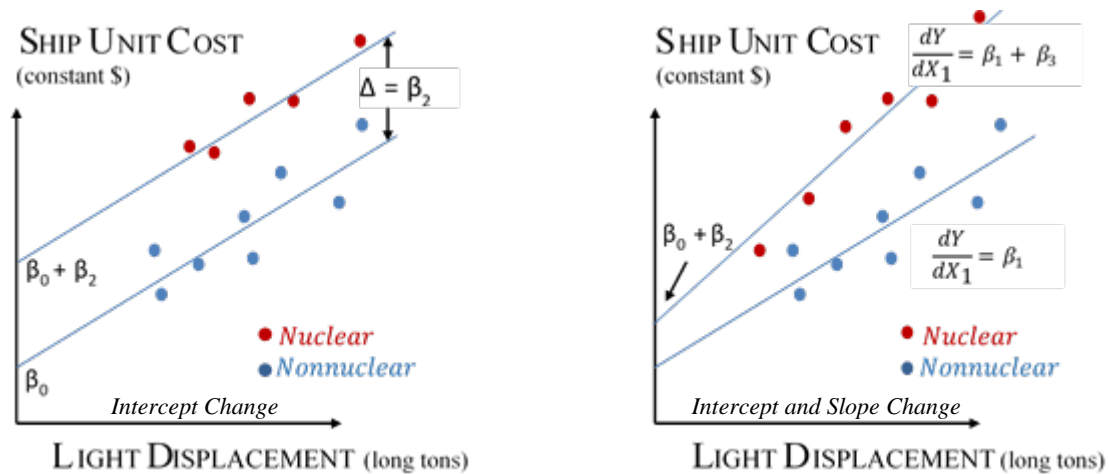


Figure 23: Dummy Variables Linear Example

When combining groups (or categories) evaluate the individual relationships (graphically and/or mathematically) first, before pooling them together using dummy variables. To be more specific, analysts should analyze separate regression equations (e.g., $y = \beta_0 + \beta_1 x_1 + \varepsilon$ for each group) before choosing a reduced model (e.g., $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$). Also, use either the Chow test or t-test to determine whether a reduced model is appropriate.³⁹

Dummy variables also apply to non-linear forms such as the Power Function (Section 2.8.2). Analogous to the linear case, the influence may be on the constant term, the exponent, or the two together. For a CIC example, under the pressure of competitive procurement, a previously sole-source firm may shift downward its CIC, steepen it, or both.⁴⁰ The dummy variable in this case is the presence or absence of a competitive procurement. **Figure 24** illustrates this example.

³⁹ Hu, S. and A. Smith, "Using Dummy Variables in CER Development," 2014 ICEAA Annual Conference, Denver, CO, 10-13 June 2014.

⁴⁰ "Analysis of Competitive Procurement of Selected Navy Weapon Systems;" Naval Center for Cost Analysis, 1990, Dr. Brian Flynn, et al.

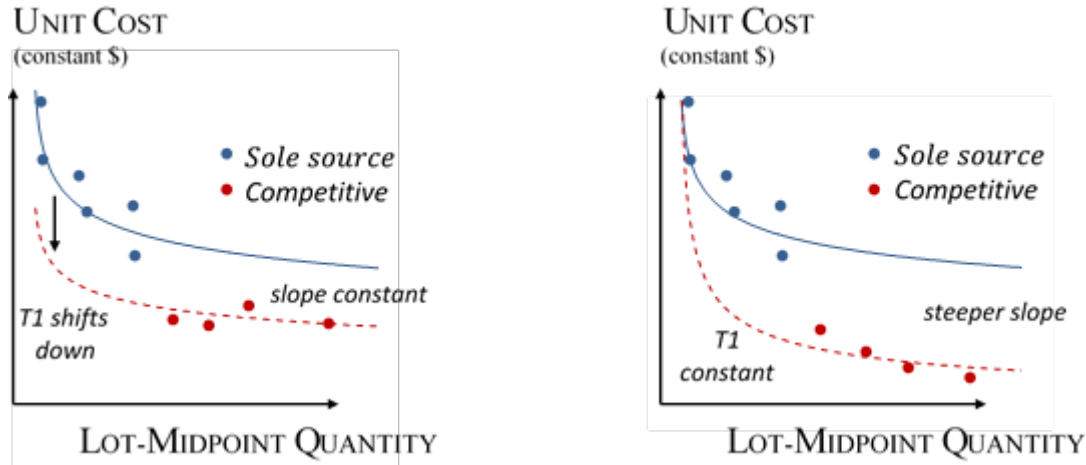


Figure 24: Dummy Variables CIC Example

A categorical variable with a levels requires $a - 1$ dummy variables. In a previous example, the two levels were $\{nuclear\ powered, other\}$ and therefore one dummy variable was required. Now, suppose a categorical variable is used to model class of ship and three categories are required: Carrier, Submarine, and Other. This has $a = 3$ levels and would require $a - 1 = 2$ dummy variables:

$$x_i = carrier = \begin{cases} 1 & \text{if carrier} \\ 0 & \text{otherwise} \end{cases}$$

$$x_j = submarine = \begin{cases} 1 & \text{if submarine} \\ 0 & \text{otherwise} \end{cases}$$

An observation would either have $x_i = 1$ and $x_j = 0$ to designate a carrier, $x_i = 0$ and $x_j = 1$ to designate a submarine, or $x_i = 0$ and $x_j = 0$ to designate other. As a result, categorical variables with multiple levels can quickly consume degrees of freedom. Further, there is sometimes an unfortunate tendency to try to explain away outliers through the use of dummy variables. Therefore, use dummy variables sparingly and judiciously.

3.3 Regression Methodologies Detail

Multiple competing components drive the development of CERs including statistical significance, logical interpretation of the equation, and availability of quality data. Evaluating cost and deriving CERs often involves testing multiple methodologies and selecting the most appropriate model (Section 4.6).

As previously discussed, potential relationships are first hypothesized using visual analysis such as scatter plots (Section 2.6). Regression analysis then develops the “best” possible CERs, or regression models, by correlating (not causation) dependent and independent variables.

The first step in identifying a potential regression methodology is to specify the model being fit. There is a dependent variable responding to a set of predictor, or independent, variables and their associated coefficients, plus variation or error. While the focus of the following sections will be on practical implementation, there are some core notations to introduce upfront. Below is a notational definition, which will carry through the remainder of this handbook.

The following is the generic regression problem:

$$y_i = f(x_{1i}, x_{2i}, \dots, x_{ki}; \beta_0, \beta_1, \beta_2, \dots, \beta_k) + \varepsilon_i \text{ for } i = 1, \dots, n$$

Where,

y_i = the i^{th} observation of the response (dependent) variable

x_{ji} = the i^{th} value of the j^{th} predictor (independent) variable for $j = 1, \dots, k$

β_j = the j^{th} coefficient for $j = 1, \dots, k$

β_0 = the intercept

ε_i = the random error associated with the i^{th} observation

f = function describing the relationship between the predictors and the response

n = number of observations

k = number of independent variables

p = number of estimated parameters

or in matrix notation (further defined in Section [3.3.1.3](#)),

$$\mathbf{y} = f(\mathbf{X}; \boldsymbol{\beta}) + \boldsymbol{\varepsilon}$$

Where,

\mathbf{y} = the $n \times 1$ vector of responses

\mathbf{X} = the $n \times (k + 1)$ matrix of predictors and the constant term

$\boldsymbol{\beta}$ = the $(k + 1) \times 1$ vector of coefficients

$\boldsymbol{\varepsilon}$ = the $n \times 1$ vector of errors

The classic model estimates independent variables and an intercept term (*i. e.*, $p = k + 1$). However, models may not include an intercept term (*i. e.*, $p = k$). The Triad Functional Form has two more parameters than independent variables (*i. e.*, $p = k + 2$). All of these representations require more observations than parameters ($n > p$).

The regression model estimates each β_j value and the ε_i 's are unknown. Each ε_i is a random value reflecting the probabilistic nature of the relationship between \mathbf{X} and \mathbf{y} . Estimated and predicted values are notated with hats, e.g., $\hat{\beta}_j$ and $\hat{\varepsilon}_i$.

The selection of a regression methodology largely depends on the explicit functional form of the relationship between \mathbf{X} and \mathbf{y} , and on the assumed structure of the error term, $\boldsymbol{\varepsilon}$. The most common strategy to fit a regression model is that of least squares. This method solves for estimates of the coefficients, $\boldsymbol{\beta}$, which minimize a function of the Residual Sum of Squares (RSS), also called Sum of Squared Errors (SSE), given a set of model assumptions. The violation of one or more underlying model assumptions in cost analysis drives or motivates the use of different regression methodologies. The remainder of this section will examine different methods of least squares for implementation in the case of different functional forms and error terms, with the exception of [3.3.4 Generalized Linear Model \(GLM\)](#), which is a method of [A.4.7.2 Maximum Likelihood Estimation \(MLE\)](#).

3.3.1 Ordinary Least Squares (OLS)

The most common regression model is OLS, which is discussed in two sections. Section [3.3.1.1 Simple Linear Regression \(SLR\)](#) introduces the model in the single independent variable. Section [3.3.1.3 Multiple Linear Regression \(MLR\)](#) takes the principles from the single independent variable case and extends them to the multiple independent variables. Section [3.3.1.3](#) will also introduce the matrix notation for more complicated functional forms. Once introduced, the remainder of this handbook will use matrix notation. Finally, OLS coefficient estimates can be solved by least squares and are the same as those solved by MLE, when model assumptions hold.⁴¹

3.3.1.1 Simple Linear Regression (SLR)

Closed-form formulas (see Appendix [A.4.1](#)) exist to solve for both the coefficient estimates and for all of the statistical metrics of interest. The Simple Linear Regression (SLR) model is the simplest functional form. Use this method when there is a single independent variable predicting a single dependent variable in a linear relationship. SLR provides a reasonable starting place to first understand the importance of a prospective predictor variable and to serve as a catalyst for developing alternative, more-sophisticated CERs. As a result, when use of a single cost driver is hypothesized, and in the absence of a strong assumption about the data, the SLR model serves as an analytical starting place. ([Step 4: Validate CER](#)). This model has many convenient properties.

SLR includes an additive error term with the assumption that the errors are independently and normally distributed. The normality assumption provides a systematic framework for conducting inference and determining significance of the results. The parameters have practical, meaningful interpretations. The intercept coefficient, β_0 , is the predicted value of the response when the independent variable is zero. Only under certain circumstances does this mathematical interpretation make sense in the context of a statistical model with prediction properties. The slope coefficient, β_1 , is interpreted as the change in the response for each positive unit increase in the independent variable.

Below is the statistical formulation of the classical, normal linear regression model, or SLR model. The first part of the statement expresses the response variable, y , is equal to an intercept parameter, β_0 , plus a slope parameter, β_1 , multiplied by the independent variable, x , plus random error, ε . The second part of the statement indicates the errors are normally distributed, independent from each other, with a mean of zero, and the same constant variance of σ^2 .

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2)$$

This statement explicitly captures the assumptions made when conducting SLR. These assumptions, which form the basis of OLS in general, are as follows:

- (1) *Independence* of errors. Each error, ε_i , is not related to any other error term.
- (2) *Homoscedasticity*. Each error, ε_i , has a constant variance, σ^2 , across the data range.

⁴¹ Further, since maximum likelihood estimators are asymptotically unbiased, consistent, and asymptotically efficient, these same properties carry over for least squares estimates in the classical normal regression model.

- (3) *Normality*. The errors are normally distributed with a mean of zero.
- (4) *Linearity*. The relationship exhibits a constant slope over the data range.
- (5) *Error term*. The error is not proportional to the independent variables

These assumptions are key to using the OLS methodology. Under assumptions (1), (2), and (4), the Gauss-Markov theorem states the coefficient estimates solved by OLS are the Best Linear Unbiased Estimators (BLUE) of the true parameter values, with the lowest variance (see Section [5.1 Adjust Point Estimate](#)). To fit this regression model by method of least squares, find values for β_0 and β_1 which minimize the sum of squared errors,⁴² represented by the following objective function:

$$\arg \min_{\beta_0; \beta_1} \sum_{i=1}^n \varepsilon_i^2 = \arg \min_{\beta_0; \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

Application

SLR is a very convenient model because the problem has a closed-form solution. Formulas exist for the estimated values of the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ as well as for the estimated variance of the error term, $\hat{\sigma}^2$. Applying these formulas to the data set produces regression results and relevant outputs.

The formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

3.3.1.2 Simple OLS Example

Consider the Electronics sample data in **Table 13** for one independent variable *Power (kW)* and dependent variable *Cost (\$M)*.

⁴² The function “arg min” represents the argument of the minimum. This returns the argument, or parameter values, which result in the minimum value of the equation. The “min” function returns the minimum value, which in this case is the sum of squared errors.

Table 13: Simple Linear Model Data Example

Observation	Cost (FY16\$K)	Power (kW)
Project 1	\$390	10.00
Project 2	\$200	5.00
Project 3	\$240	5.20
Project 4	\$300	7.00
Project 5	\$460	12.00
Project 6	\$560	17.80
Project 7	\$700	21.00
Project 8	\$800	25.00
Project 9	\$500	18.00

COSTAT is used to fit the SLR model and return the standard outputs, displayed in **Figure 25**.

I. Model Form and Equation Table

Model Form:	Unweighted Linear model
Number of Observations Used:	9
Equation in Unit Space:	Cost = 92.93 + 27.39 * Power

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	92.9309	30.9235		3.0052	0.0198	0.9802
Power	27.3853	2.0480	0.9810	13.3716	0.0000	1.0000

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
42.2261	96.23%	95.69%	0.9810

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	1	318807.5318	318807.5318	178.7998	0.0000	1.0000
Residual (Error)	7	12481.2970	1783.0424			
Total	8	331288.8889				

Figure 25: Simple Linear Regression Output

These results may vary in appearance by software package, but all should contain the same basic information. In **Figure 25**, the first table provides basic information about the model, including the regression equation. Next is a table of coefficients showing the estimated values for the regression equation, standard errors, and t-tests for significance testing. In addition, there are several regression statistics including the Standard Error and R-squared. The Analysis of Variance (ANOVA) table provides a standard view for significance testing and provides other key diagnostic values specific to the regression model. These values are discussed in [4.5 Model Quality](#).

Figure 26 shows a scatter plot with *Power (kW)* on the *x*-axis and *Cost (\$M)* on the *y*-axis and the fit regression line going through the data.

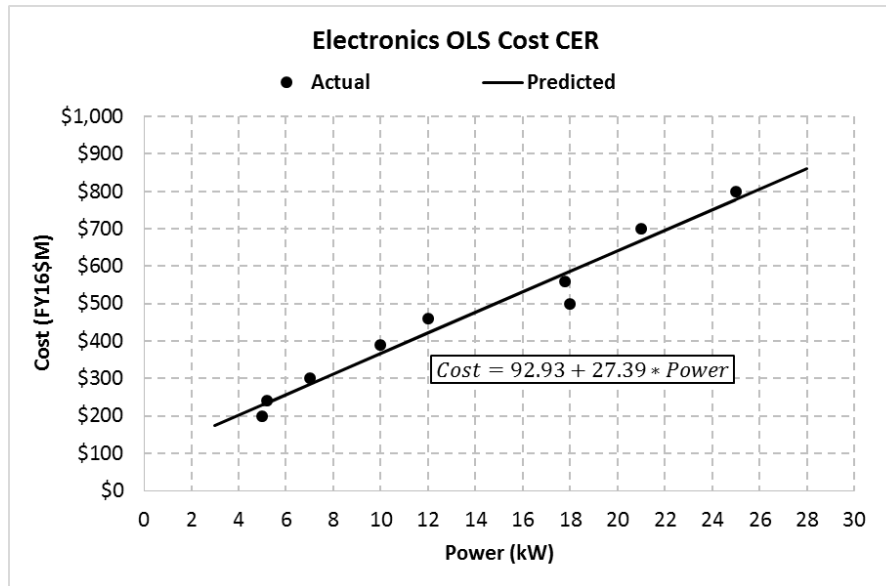


Figure 26: Simple Linear Regression Model Scatter Plot

Accepting the CER requires additional analysis. [Step 4: Validate CER](#) contains a much more in-depth model validation discussion.

3.3.1.3 Multiple Linear Regression (MLR)

The Multiple Linear Regression (MLR) model is another simple functional form and a direct extension of the simple linear model. This method is recommended when multiple independent variables may predict a single dependent variable in a linear relationship. Much like SLR, the true, underlying relationship between a dependent variable and multiple explanatory variables is always unknown. MLR provides a reasonable starting place to understand the importance of a prospective predictor variable as well as serves as a catalyst for developing alternative CERs. When multiple cost drivers are hypothesized, the MLR model is recommended as an analytical starting place ([Step 4: Validate CER](#)).

Similar to SLR, the MLR model expresses a linear functional form and assumes a normal additive error term. The difference is that the MLR model may include more than one independent variable.⁴³ The model has the same convenient properties as SLR along with a closed-form solution (see Appendix [A.4.1](#)) that allows analysts to solve for coefficient estimates and all other statistical metrics. The assumption of normally distributed error terms allows analysts to determine significance of the regression results. The MLR model includes an intercept coefficient, β_0 , that reflects the predicted value of the response when all of the independent variables are zero. However, only under certain circumstances does this mathematical interpretation make sense in the context of a statistical model with prediction properties. Each slope

⁴³ It is important to note the MLR model refers to multiple regression, not multivariate regression. The term multivariate refers to an analysis of multiple response variables, not predictors. A common example of a multivariate analysis in Cost Analysis is the joint modeling of Cost and Schedule.

coefficient, $\beta_1, \beta_2, \dots, \beta_k$, can be interpreted as the change in the response for each positive unit increase of the respective predictor when all other predictors are held constant.

While working with standard algebra is possible for the MLR model, linear algebra and matrix notation provide a standard framework with computational advantages for more complex models. Bold font denotes matrices and vectors, with matrices represented by capital letters and vectors by lowercase letters. The transpose of a matrix or vector is notated with a prime, for example \mathbf{X}' would be the transpose of \mathbf{X} . Similarly, the inverse of \mathbf{X} is notated as \mathbf{X}^{-1} .

Below is the statistical formulation of the MLR model in matrix form. This generalization is valid for any number of independent variables ($k \geq 1$) and any number of observations ($n > k + 1$), which makes it very convenient to work with. The first part of the statement expresses the response variable, or vector, \mathbf{y} , is equal to the matrix of independent variables, \mathbf{X} , multiplied by the coefficient variables, or vector, $\boldsymbol{\beta}$, plus some random error, $\boldsymbol{\varepsilon}$. The second part of the statement indicates the assumption that the error term is normally distributed with a mean of zero, illustrates constant variance of σ^2 , and has a covariance of zero.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Where,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}_{n \times (k+1)} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1) \times 1} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}$$

\mathbf{I} is the $n \times n$ identity matrix (a matrix of all zeroes except for ones in the top-left to bottom-right diagonal positions). The leading column of ones in the \mathbf{X} matrix corresponds with the intercept coefficient, β_0 .

This statement explicitly captures the assumptions made when conducting MLR. These are the same OLS assumptions introduced with SLR:

- (1) *Independence* of errors. Each error, ε_i , is not related to any other error term.
- (2) *Homoscedasticity*. Each error, ε_i , has a constant variance, σ^2 , across the data range.
- (3) *Normality*. The errors are normally distributed with a mean of zero.
- (4) *Linearity*. The relationship exhibits a constant slope over the data range.
- (5) *Error term*. The error is not proportional to the independent variables

Similar to the SLR model, assumptions (1), (2), and (4), support the Gauss-Markov theorem that states coefficient estimates solved by OLS are the Best Linear Unbiased Estimators (BLUE) of the true parameter values. To fit this regression model by method of least squares, find values for the coefficient vector $\boldsymbol{\beta}$ that minimize the sum of squared errors. The MLR method includes an objective function similar to the SLR method and is noted below:

$$\arg \min_{\boldsymbol{\beta}} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

$$= \arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta)$$

MLR is a very convenient model because the problem has a closed-form solution. A formula exists for the vector of coefficient estimates, $\hat{\beta}$, as well as for the estimated variance of the error term, $\hat{\sigma}^2$. Applying these formulas to a given data set produces relevant regression results and statistical outputs.

The formula for $\hat{\beta}$ is:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

This formula can be evaluated via matrix algebra in MS Excel.

3.3.1.4 MLR Example

Consider the same data set as introduced in Section [3.3.1.1](#) but now with a second independent variable, *Aperture*, illustrated in **Table 14**.

Table 14: Multiple Linear Model Data Example

Observation	Cost (FY16\$K)	Power (kW)	Aperture (cm ²)
Project 1	\$390	10.00	8.70
Project 2	\$200	5.00	8.00
Project 3	\$240	5.20	8.20
Project 4	\$300	7.00	
Project 5	\$460	12.00	9.00
Project 6	\$560	17.80	9.50
Project 7	\$700	21.00	9.20
Project 8	\$800	25.00	9.70
Project 9	\$500	18.00	

The matrix math used to apply the MLR model described in the previous section can also be applied using MS Excel, as illustrated in **Figure 27**. This figure provides analysts with the MS Excel formulas reflected in columns F, G, and H, that can be used to apply the MLR model to the datasets reflected in columns B, C, and D.

CER Development Handbook

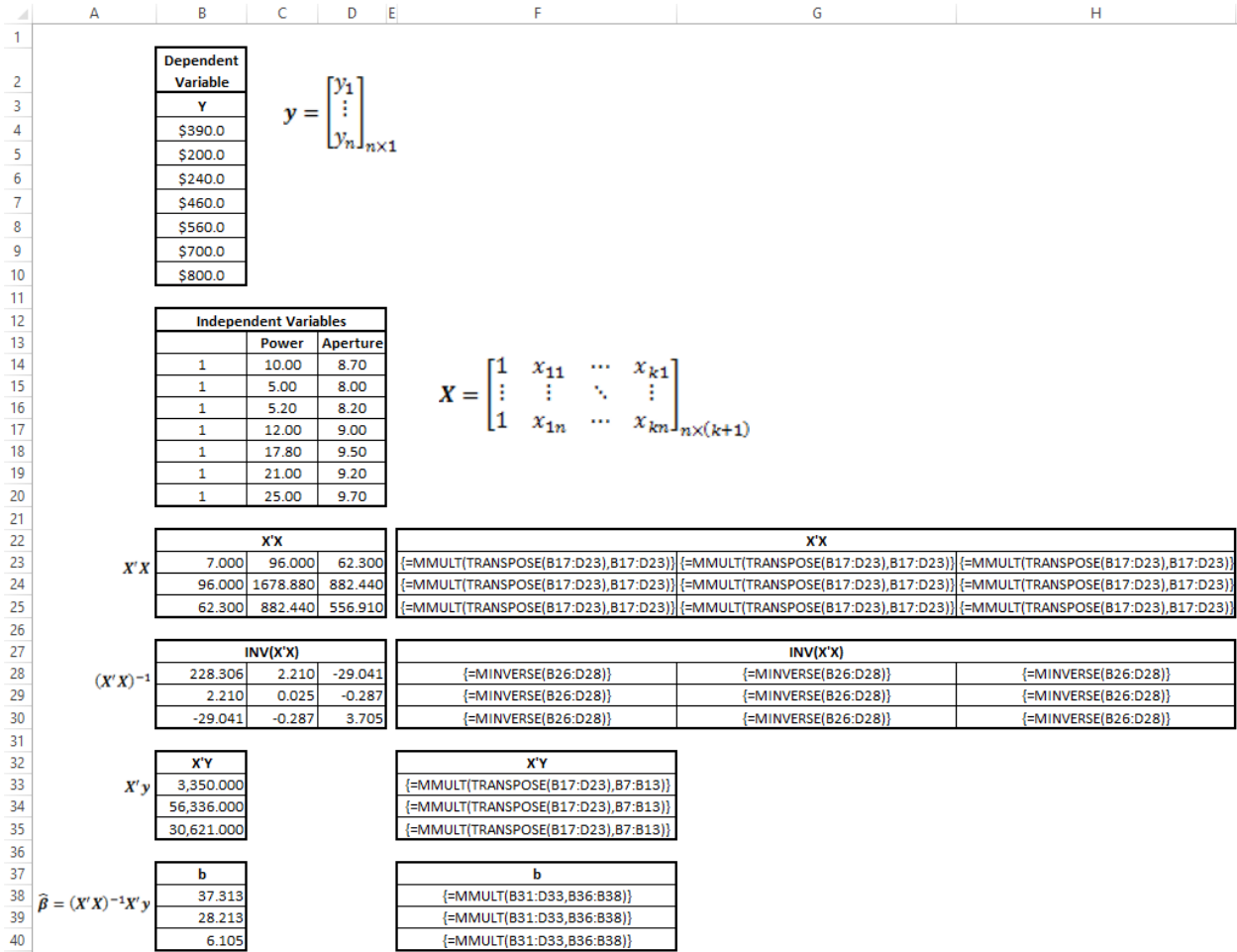


Figure 27: Multiple Linear Regression Matrix Math

Figure 28 and **Figure 29** illustrate another method of fitting the MLR model and generating relevant regression model outputs using CO\$TAT.

CER Development Handbook

I. Model Form and Equation Table

Model Form:	Unweighted Linear model
Number of Observations Used:	7
Equation in Unit Space:	Cost = 37.31 + 28.21 * Power + 6.105 * Aper

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	37.3129	449.4459		0.0830	0.9378	0.0622
Power	28.2134	4.6985	0.9777	6.0047	0.0039	0.9961
Aper	6.1047	57.2542	0.0174	0.1066	0.9202	0.0798

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
29.7453	98.83%	98.24%	0.9941

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	2	298146.5837	149073.2919	168.4858	0.0001	0.9999
Residual (Error)	4	3539.1306	884.7826			
Total	6	301685.7143				

Further Analysis of Variance

(SS explained by each variable when entered in the order given)

Due To	DF	Sequential SS
Regression	2	298146.5837
Power	1	298136.5250
Aper	1	10.0588

Pairwise Correlation Matrix

Variables	Cost	Power	Aper
Cost	1.0000	0.9941	0.9394
Power	0.9941	1.0000	0.9431
Aper	0.9394	0.9431	1.0000

Figure 28: Multiple Linear Regression Output

These results may vary in appearance by software package, but all should contain the same basic information. In **Figure 28**, the first table provides basic information about the model, including the regression equation. Next is a table of coefficients showing the estimated values for the regression equation, standard errors, and t-tests for significance testing. This table is followed by several regression statistics including Standard Error and R-squared. The next section includes the Analysis of Variance (ANOVA) table, which provides key diagnostic values relative to the regression model. The format of the results is similar to SLR, but now includes an additional independent variable.

When the MLR model is used, two additional tables are provided; the first identifies how much of the variation is explained by each variable and the second provides the correlation between each of the variables. This example includes two independent variables that reflect strong correlation, highlighted in red. This correlation table provides analysts with key indicators of potential multicollinearity impacts. Section [3.3.6 Ridge Regression](#) provides additional detail regarding methods of addressing models potentially affected by multicollinearity.

Figure 29 shows a scatter plot with the Actual Cost on the x -axis and Predicted Cost on the y -axis, with the line $Predicted = Actual$ going through the data.

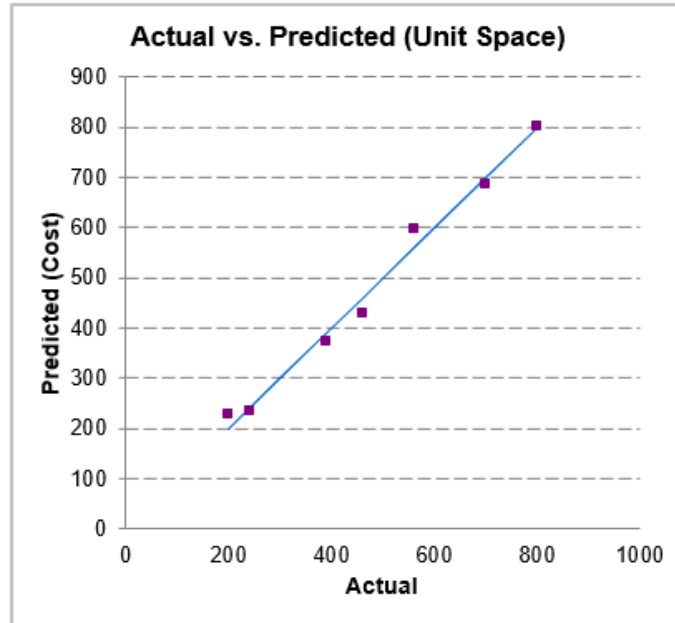


Figure 29: Multiple Linear Regression Model Predicted vs. Actual Plot

While this example illustrates a positive y-intercept value, models that output a negative y-intercept are not uncommon. Section [4.4.2.1 Intercept Term](#) provides additional detail regarding y-intercept validation methods.

Accepting the CER requires additional analysis. [Step 4: Validate CER](#) contains a much more in-depth model validation discussion.

3.3.2 Generalized Least Squares (GLS)

The OLS linear regression model relies on a set of fairly restrictive assumptions concerning the behavior of the error term. An alternative model, known as Generalized Least Squares⁴⁴, or GLS, is considerably less restrictive in this respect. GLS retains all of the assumptions of the OLS model (Section [3.3.1 Ordinary Least Squares \(OLS\)](#)) except for:

- (1) Independence of errors (Section [4.2.1.2](#))
- (2) Homoscedasticity (Section [4.2.1.3](#))

GLS allows for the possibilities of serial correlation and heteroscedasticity (non-constant variance). The model is called “generalized” because it includes other models as special cases.

⁴⁴ “On Least Squares and Linear Combinations of Observations,” Aitken, A. C.; Proceedings of the Royal Statistical Society of Edinburgh,” Vol. 55, 1935, pages 42-48.

Section [3.3.2.1](#) introduces the GLS model and Section [3.3.2.2](#) discusses the Weighted Least Squares (WLS) model. WLS is a special case of the GLS model that only allows for a violation of the homoscedasticity (2) assumption.

3.3.2.1 Generalized Least Squares (GLS)

The Generalized Least Squares (GLS) model takes on the same linear form for $f(\mathbf{X}; \boldsymbol{\beta})$ as [Multiple Linear Regression \(MLR\)](#) (referred to as OLS). GLS expresses a linear functional form and a normal additive error term. However, the assumption that the errors are independently and identically distributed as normal is no longer required. Normality of the error term is still an assumption, but now the errors may be correlated with each other and/or have non-constant variance.

Time Series analysis is a common use of the GLS model as well as other scenarios where the homoscedasticity assumption fails. Closed-form solutions ([Appendix A.4.2](#)) exist for both the coefficient estimates and for other relevant statistical metrics. The normality assumption provides a systematic framework to determine significance of the results.

Below is the statistical formulation of the GLS model. The first part of the statement expresses that the response variable, or vector, \mathbf{y} , is equal to the matrix of independent variables, \mathbf{X} , multiplied by the coefficient variables, or vector, $\boldsymbol{\beta}$, plus random error, $\boldsymbol{\varepsilon}$. The second part of the statement indicates the assumption that the error term is normally distributed with a mean of zero, and with covariance matrix $\boldsymbol{\Sigma}$.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

This statement captures the assumptions made when conducting GLS and reflects similar assumptions introduced with OLS. The covariance matrix, $\boldsymbol{\Sigma}$, is used to standardize the residual errors, removing their correlation and scaling to the same variance. The GLS model includes the following assumptions:

- (1) *Independence* of errors. Each error, ε_i , is not related to any other error term.
- (2) *Homoscedasticity*. Each error, ε_i , has a constant variance, σ^2 , across the data range.
- (3) *Normality*. The errors are normally distributed with a mean of zero.
- (4) *Linearity*. The relationship exhibits a constant slope over the data range.
- (5) *Error term*. The error is not proportional to the independent variables

Similar to OLS, under assumptions (1), (2), and (4), the Gauss-Markov theorem states that the coefficient estimates solved by GLS are the Best Linear Unbiased Estimators (BLUE) of the true parameter values. However, this assertion is only true when $\boldsymbol{\Sigma}$ is known. The GLS model has the added complexity that $\boldsymbol{\Sigma}$ is almost always unknown.

When the errors have correlation, an estimate of the covariance structure is required. Additionally, when the errors have non-constant variance, or heteroscedasticity, an estimate of the errors is required. Once $\boldsymbol{\Sigma}$ is estimated, under assumptions (1), (2), and (4), the Gauss-Markov theorem states that the coefficient

estimates solved by GLS are the Estimated Best Linear Unbiased Estimators (EBLUE) of the true parameter values.

To fit this regression model by method of least squares, find values for the coefficient vector β , which minimize a similar objective, function to that of OLS, but now on the standardized, or Mahalanobis⁴⁵ distance. Note that when the correlation matrix is (or is proportional to) the identity matrix, I (or $\sigma^2 I$), then the objective function reduces to that of OLS.

$$\arg \min_{\beta} \boldsymbol{\varepsilon}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon} = \arg \min_{\beta} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

GLS is a convenient model because the problem has a closed-form solution. A formula exists for the coefficient estimates, β , as well as for the estimated variance of the standardized error term, $\hat{\sigma}^2$. However, there is now the added complexity of having to estimate the covariance matrix prior to applying these formulas. Any selected covariance matrix will provide a valid (unbiased and consistent) estimator, but not necessarily with minimum variance. Once $\boldsymbol{\Sigma}$ is estimated, formulas can be directly applied to the data set.

Several statistical software packages include the capability to automatically produce relevant regression results and diagnostics. As previously mentioned, the most common use of GLS is to correct for heteroscedasticity impacts or analyze Time Series datasets. While GLS represents one method of Time Series data analysis, the topic is beyond the scope of this handbook.

The formula for $\hat{\beta}$ is:

$$\hat{\beta} = (\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{y}$$

3.3.2.2 Weighted Least Squares (WLS)

The Weighted Least Squares (WLS) model is a special case of GLS used to address heteroscedasticity impacts. While WLS does not allow for correlation of the errors, it does allow for non-constant variance. The violation of the OLS homoscedasticity assumption often motivates the use of WLS and is further discussed in Section [4.2.1.3](#). Closed-form solutions (see Appendix [A.4.2](#)) exist to solve for coefficient estimates as well as for other statistical metrics of interest. The normality assumption provides a systematic framework to determine significance of the results.

One common use of WLS is the linear model with multiplicative error. This example is a special case of WLS, where the model has non-constant variance that is proportional in magnitude to the response

⁴⁵ The Mahalanobis distance accounts for distance and direction. For example, the number of standard deviations a point is away from the center of mass of an ellipsoid.

variable. Despite the fact that the error appears multiplicative, another form of the WLS model includes an additive error term.⁴⁶

Below is the statistical formulation of the WLS model. The first part of the statement expresses that the response variable, or vector, \mathbf{y} , is equal to the matrix of independent variables, \mathbf{X} , multiplied by the coefficient variables, or vector, $\boldsymbol{\beta}$, plus error, $\boldsymbol{\varepsilon}$. The second part of the statement indicates the assumption that the error term is normally distributed with a mean of zero, and with covariance matrix, $\boldsymbol{\Sigma}$.

In WLS, $\boldsymbol{\Sigma}$ has the restriction that the diagonal entries⁴⁷ must all be greater than zero, and the off-diagonals must all be equal to zero. The notation $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$ allows for a more practical meaning of these diagonal entries. Once inverted, these diagonal entries are referred to as the vector of weights (inclusive of σ^2), \mathbf{w} , and the covariance matrix can be notated as $\boldsymbol{\Sigma}^{-1} = \mathbf{W} = \langle \mathbf{w} \rangle$. The optimal weights are the reciprocals of each error's variance. However, these are rarely known values and require an intermediary estimation step in order to fit the WLS model.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{W}^{-1}) \text{ and } \mathbf{W} = \langle \mathbf{w} \rangle$$

The formula above captures the assumptions made when using the WLS model. These are similar assumptions introduced with OLS. The weight matrix, \mathbf{W} , is used to standardize the residual errors scaling them to have the same variance. This topic is further discussed in Section [4.2.2 Weighted Least Squares \(WLS\)](#). The WLS model is based on the following assumptions:

- (1) *Independence* of errors. Each error, ε_i , is not related to any other error term.
- (2) *Homoscedasticity*. Each error, ε_i , has a constant variance, σ^2 , across the data range.
- (3) *Normality*. The errors are normally distributed with a mean of zero.
- (4) *Linearity*. The relationship exhibits a constant slope over the data range.
- (5) *Error term*. The error is not proportional to the independent variables

Similar to OLS, under assumptions (1), (2), and (4), the Gauss-Markov theorem states that the coefficient estimates solved by WLS are the Best Linear Unbiased Estimators (BLUE) of the true parameter values. However, this assertion is only true when \mathbf{W} is known. The WLS model has the added complexity that the weight vector, \mathbf{W} , is rarely known and an estimate of the errors is required.

Once \mathbf{W} is estimated, under assumptions (1), (2), and (4), the Gauss-Markov theorem states that the coefficient estimates solved by WLS are the Estimated Best Linear Unbiased Estimators (EBLUE) of the true parameter values. To fit this regression model by method of least squares, find values for the

⁴⁶ This is a deviation from the multiplicative error representation of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}\boldsymbol{\varepsilon}$ seen in some publications. MUPE and ZMPE are two specific methods for solving this problem. The reference in Appendix [0](#) ZMPE contains key information on both methods and recommends use of the MUPE method. In the linear model, MUPE is equivalent to an iterative application of WLS, aptly named Iteratively Reweighted Least Squares (IRLS). IRLS is covered later in this section, as well as in Section [3.3.5 Non-linear Least Squares \(NLS\)](#) for the non-linear application.

⁴⁷ The term “diagonals” refers to the diagonal entries in a square matrix from the top-left to the bottom-right diagonal positions. These are the positions of the ones in the identity matrix.

coefficient vector $\boldsymbol{\beta}$, which minimize a similar objective function to that of OLS. Note that when all the weights are the same, the objective function reduces to that of OLS.

$$\arg \min_{\boldsymbol{\beta}} \boldsymbol{\varepsilon}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon} = \arg \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

WLS has the added complexity that the vector of weights, \mathbf{w} , must be estimated prior to applying the objective formula. Statistical software packages can be used to automatically produce the relevant regression results and diagnostics, and in some cases even estimate the weight vector automatically. There are many strategies for estimating \mathbf{w} , but no perfect solution (refer to Section [4.5 Model Quality](#) for additional information). The following are several common strategies for estimating the weight vector:

Method 1: Run the OLS model and use the squared reciprocal of the residuals as the weights.

$$\hat{\boldsymbol{\beta}}_{OLS} = \text{coefficients derived from OLS}$$

$$\begin{aligned} \mathbf{e}_{OLS} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} \end{aligned}$$

$$\mathbf{w} = \frac{1}{\mathbf{e}_{OLS}^2}$$

Method 2: Use the squared reciprocal of an independent variable.

$$\mathbf{w} = \frac{1}{\mathbf{x}_j^2}$$

Method 3: Use the squared reciprocal of the response variable.

$$\mathbf{w} = \frac{1}{\mathbf{y}^2}$$

Method 4: This method attempts to minimize the relative residual sum of square error (i.e., percentage error), relative to the actual values⁴⁸ by changing \mathbf{w} .

$$PE = \frac{\mathbf{y} - \hat{\mathbf{y}}}{\mathbf{y}}$$

Method 5: Run the MUPE algorithm in CO\$TAT. In the linear model setting, MUPE produces a set of weights and returns statistical results under the WLS framework. MUPE is utilizing a non-linear methodology to minimize the relative residual sum of squares (i.e., percentage error (%Error)), *relative to the predicted values*,

⁴⁸ Tofallis, Chris (2008) "Least Squares Percentage Regression," *Journal of Modern Applied Statistical Methods*: Vol. 7: Iss. 2, Article 18.

$$PE = \frac{y - \hat{y}}{\hat{y}}$$

MUPE is iteratively reweighting the model. Starting with $w = \mathbf{1}$ (i.e., the OLS estimator), MUPE iteratively fits a WLS model using the reciprocal of the squared predicted values of the previous iteration as the current weighting vector. This repeats until the coefficient estimates converge within a user specified tolerance limit and is known as iteratively reweighted least squares (IRLS).⁴⁹ The weight vector at a given iteration step γ is expressed as,

$$w_{\gamma} = \frac{1}{\hat{y}_{\gamma-1}^2}$$

After estimating the weights, WLS becomes a very convenient model because the problem has a closed-form solution for the linear model. A formula exists for the coefficient estimates, $\hat{\beta}$, and for the estimated variance of the error term, $\hat{\sigma}^2$. Directly applying these formulas to the data set, or utilizing statistical software, produces the regression results and relevant outputs.

The formula for $\hat{\beta}$ is:

$$\hat{\beta} = (X'WX)^{-1}X'Wy$$

Consider the electronics data set with the independent variable *Power* (kW) and dependent variable *Cost* (\$K), shown in **Table 15** along with the weights for all four methods. Iteratively weighted methods change weight values at each iteration. Thus, the weights presented for Method 5, MUPE, are from the final iteration (i.e., the final set of weights used in the actual model).

Table 15: Weighted Linear Squares Data Example

Observation	Cost (FY16\$K)	Power (kW)	Weight Methods 1, 2, 3, and 5			
			1/OLS Err ²	1/Pwr ²	1/Cost ²	MUPE
Project 1	\$390	10.00	0.0018553	0.0100000	0.0000066	0.0000075
Project 2	\$200	5.00	0.0011218	0.0400000	0.0000250	0.0000195
Project 3	\$240	5.20	0.0459405	0.0369822	0.0000174	0.0000186
Project 4	\$300	7.00	0.0042319	0.0204082	0.0000111	0.0000125
Project 5	\$460	12.00	0.0006766	0.0069444	0.0000047	0.0000056
Project 6	\$560	17.80	0.0024054	0.0031562	0.0000032	0.0000029
Project 7	\$700	21.00	0.0009779	0.0022676	0.0000020	0.0000022
Project 8	\$800	25.00	0.0019865	0.0016000	0.0000016	0.0000016
Project 9	\$500	18.00	0.0001356	0.0030864	0.0000040	0.0000029

⁴⁹ This is a common definition of iteratively reweighted least squares (IRLS). However, IRLS can also be conducted under different weighting methodologies, such as using Method 1 at each iteration.

Figure 30 shows actual values compared to the predicted value of each model noted in Table 15. Data appearing to have non-constant variance in a multiplicative pattern suggest the applicability of a weighted least squares approach. That is not the case in the example shown in Figure 30.

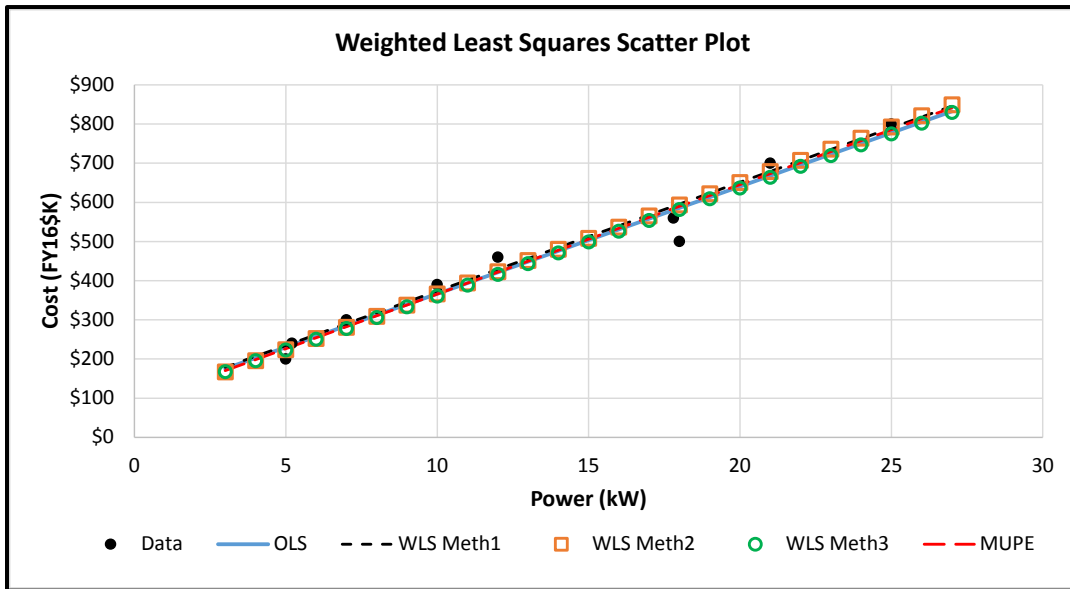


Figure 30: Weighted Least Squares Scatter Plot

This example illustrates results that are very difficult to discern between the different weighting methods shown in Table 15. This is expected. The different sets of weights in WLS often result in very similar point estimates. However, they have different variances and therefore different uncertainties around the estimate (discussed in Section [5.0: Step 5: Characterize Uncertainty](#)).

The statistical output for each method will be similar to OLS outputs, as shown in the examples of Section [3.3.1 Ordinary Least Squares \(OLS\)](#). For example, Figure 31 illustrates a statistical output for Method 3 using COSTAT.

I. Model Form and Equation Table

Model Form:	Weighted Linear model
Number of Observations Used:	9
Equation in Unit Space:	Cost = 85.13 + 27.58 * Power

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	85.1306	21.3774		3.9823	0.0053	0.9947
Power	27.5785	2.1962	0.9785	12.5576	0.0000	1.0000

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.0975	95.75%	95.14%	0.9785

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	1	1.4988	1.4988	157.6927	0.0000	1.0000
Residual (Error)	7	0.0665	0.0095			
Total	8	1.5653				

Figure 31: Weighted Least Squares Regression Output

Table 16 provides a summary of selected outputs and model fit statistics, such as intercept, slope, and R^2 Adj. SE and MAD⁵⁰ illustrate that all methods return a similar level of accuracy.

Table 16: Example WLS Methodology Comparison

Name	Equation	Intercept p-value	Slope p-value	In Unit Space		
				R ² Adj	SE	MAD
OLS	92.93 + 27.39 * Power	1.98%	0.00%	95.69%	42.23	7.17%
Method 1	95.74 + 27.78 * Power	0.00%	0.00%	95.47%	43.32	6.71%
Method 2	81.8 + 28.43 * Power	0.38%	0.00%	95.51%	43.12	7.17%
Method 3	85.13 + 27.58 * Power	0.53%	0.00%	95.60%	42.66	7.56%
Method 4 (MUPE)	86.65 + 27.95 * Power	0.40%	0.00%	95.61%	42.48	6.89%

Recalling that OLS minimizes the SSE in unit space, the R^2 value in unit space for WLS cannot be better than the OLS estimator. MAD is defined as the mean absolute deviation, which is then converted to a percentage by dividing by the number of observations, n .

Accepting the CER requires additional analysis. [Step 4: Validate CER](#) discusses the process of using statistical outputs to select the most appropriate model.

⁵⁰ MAD = Mean Absolute Deviation. Measures the average percentage by which the regression overestimates or underestimates the observed actual value.

3.3.3 Transformable Linear and the Log-Linear Model

3.3.3.1 Overview

The use of a variable transformation can remedy certain assumption violations of the [Ordinary Least Squares \(OLS\)](#) model. CERs violating homoscedasticity (Section [4.2.1.3](#)), normality of errors (Section [4.2.1.4](#)), or linearity (Section [4.2.1.5 Linearity](#)) can sometimes be corrected by applying a transformation. **Figure 32** displays just a few of the many possible transformations for which the resulting equation is linear in the parameters, and therefore can be estimated using [Ordinary Least Squares \(OLS\)](#). However, for the purposes of this section, the term ‘transform’ will refer to the Log-Linear model.

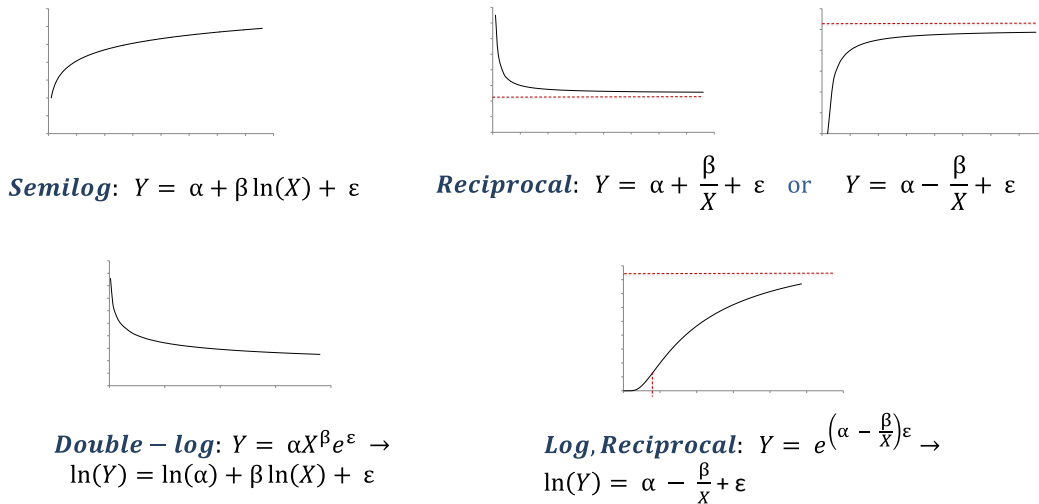


Figure 32: Common Linear Transformations

– Terminology –

The “Double-log” transformation in **Figure 32** is often referred to as *Log Ordinary Least Squares (LOLS) regression*.

Transformations are simple to implement, but they produce a regression optimized in the transformed (or fit) space, not in the unit space⁵¹. Transformations can lead to unexpected results including biased coefficient estimators and models without minimal variance in unit space. Additionally, transformations remain useful for variables illustrating non-linear trends.

The Log-Linear model is any functional form of a linear model in the transformed space after taking the logarithm, usually base e , of both sides of the equation. Since log transforms turn multiplication into addition, and the normal distribution into the lognormal distribution, a lognormal model assumes a

⁵¹ [Appendix B](#) Maximum likelihood estimation for Regression of Log Normal error (MRLN) Summary provides additional background on the popularity behind Log Ordinary Least Squares (LOLS) regression and motivation for the log-normal error term.

multiplicative lognormal error term in unit space. This translates into an additive normal error term in the transformed space, making the model a candidate for OLS.

When using Log-Linear transformation, model selection based on minimal variance in unit space is unlikely. Further, some of the coefficients when transformed back into unit space from the transform space are biased (See [Step 5: Characterize Uncertainty](#) for a brief discussion on bias).

While there are many transformations to make the model linear, the focus of this section will be on the two most common models: the [2.8.2 Power Functional Form](#) and [2.8.3 Exponential Functional Form](#). The concepts introduced for these two cases translate directly to other related transformable forms, such as those with multiple independent variables.

Below is the mathematical formulation for the [Exponential Model](#). The error term is now multiplied by, not added to, the model equation.

$$\mathbf{y} = \beta_0 e^{x\beta_1} e^\varepsilon \text{ where } \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

By applying the natural log transform, the equation transforms to,

$$\ln(\mathbf{y}) = \ln(\beta_0) + \beta_1 \mathbf{x} + \varepsilon \text{ where } \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Substitutions give an identical form to OLS,

$$\mathbf{y}^* = \beta_0^* + \beta_1 \mathbf{x} + \varepsilon \text{ where } \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Where,

$$\begin{aligned} \mathbf{y}^* &= \ln(\mathbf{y}) \\ \beta_0^* &= \ln(\beta_0) \end{aligned}$$

A similar transformation can be applied to the [Power Model](#).

$$\mathbf{y} = \beta_0 \mathbf{x}^{\beta_1} e^\varepsilon \text{ where } \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

By applying the natural log transform to each side of the equation, the equation transforms to,

$$\ln(\mathbf{y}) = \ln(\beta_0) + \beta_1 \ln(\mathbf{x}) + \varepsilon \text{ where } \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Substitution again provides an identical form to OLS,

$$\mathbf{y}^* = \beta_0^* + \beta_1 \mathbf{x}^* + \varepsilon \text{ where } \varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Where,

$$\begin{aligned} \mathbf{y}^* &= \ln(\mathbf{y}) \\ \mathbf{x}^* &= \ln(\mathbf{x}) \\ \beta_0^* &= \ln(\beta_0) \end{aligned}$$

The assumptions for the Log-Linear model in the transform space are the same as the OLS assumptions (Section [5.2.1](#)). The constant variance assumption in the transform space translates back to a constant CV

in unit space. The normal error in the transform space, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, implies $e^\varepsilon \sim LN(\mathbf{0}, \sigma^2 \mathbf{I})$. This lognormal distribution has a CV of $\sqrt{e^{\sigma^2} - 1}$. Since the variance of the assumed additive normal error term in log space, σ^2 , is constant, so is the CV in unit space throughout the data range.

3.3.3.2 Applying the LOLS Model

The Log-Linear model is a very convenient model because the problem has a closed-form solution. The OLS formulas are used for the log space coefficient estimates, $\hat{\boldsymbol{\beta}}^*$, and to estimate the variance of the error term, $\hat{\sigma}^2$. Applying these formulas to the data set, or utilizing statistical software, will produce the relevant regression outputs.

Unfortunately, the OLS estimator of β_0^* fit in log space yields a biased estimate of the parameter β_0 when transformed back into unit space. Methods exist to quantify and adjust for this bias by applying a factor. Two methods are the Goldberger factor and the PING factor (Appendix [A.4.3.1 Mean Shift](#))⁵². However, both methods are approximations and neither performs particularly well outside of the data range (extrapolation). With modern computing and software applications, analysts are recommended to fit the model in unit space rather than transform the coefficients and attempt to apply an adjustment factor.

Hu, 2005, studied this recommendation and resulted to using the MUPE (see [3.3.5.4 Minimum Unbiased Percentage Error \(MUPE\)](#)) methodology instead of the Log-Linear model. In addition, a maximum likelihood (MLE) approach applied to the log-linear model is possible. This methodology is discussed in [Appendix B Maximum likelihood estimation for Regression of Log Normal error \(MRLN\) Summary](#).

A further generalization is the recommendation of treating non-linear models as just that, non-linear models, rather than transforming for the mathematical convenience of OLS. Some non-linear methods are discussed briefly in Section [3.3.4 Generalized Linear Model \(GLM\)](#) and in more detail in Section [3.3.5 Non-linear Least Squares \(NLS\)](#).

3.3.3.3 LOLS Example

Consider the sample data in **Table 17** for one independent variable (*Intensity (kWperCm²)*) and the dependent variable *Cost (FY16\$K)*.

⁵² Hu, 2005 demonstrates improved performance of the PING factor over the Goldberger factor outside of the data range (i.e., when extrapolating).

Table 17: Power Model Data Example

Observation	Cost (FY16\$K)	Power (kW)	Aperture (cm ²)	Intensity (kWperCm ²)
Project 1	\$390	10.00	8.70	1.1494
Project 2	\$200	5.00	8.00	0.6250
Project 3	\$240	5.20	8.20	0.6341
Project 4	\$300	7.00		
Project 5	\$460	12.00	9.00	1.3333
Project 6	\$560	17.80	9.50	1.8737
Project 7	\$700	21.00	9.20	2.2826
Project 8	\$800	25.00	9.70	2.5773
Project 9	\$500	18.00		

After viewing a scatter plot of the data, COSTAT is used to fit the Power Model, $y = \beta_0 x^{\beta_1} \cdot e$, by conducting OLS regression on the data that is transformed for both $y = Cost$ (\$K) and $x = Intensity$ (kWperCm²). **Figure 33** and **Figure 34** display the regression outputs.

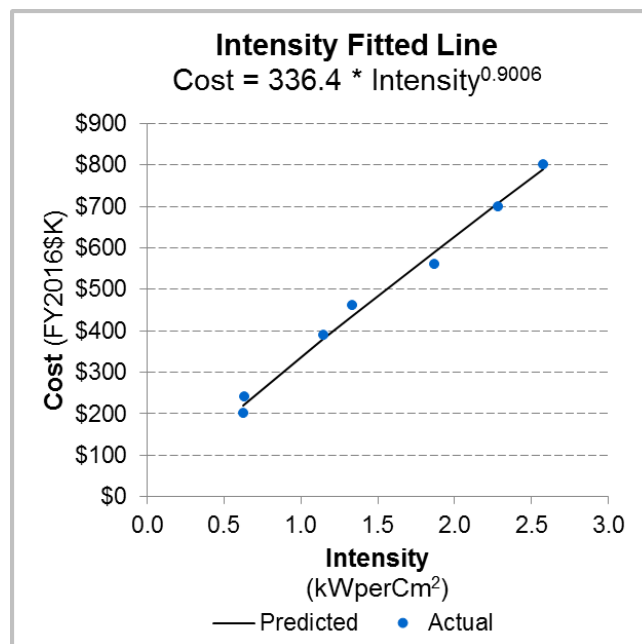


Figure 33: Log-Linear Regression Model Scatter Plot

I. Model Form and Equation Table

Model Form:	Unweighted Log-Linear model
Number of Observations Used:	7
Equation in Unit Space:	Cost = 336.4 * Intensity ^{0.9006}

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value
Intercept	5.8183	0.0278		209.0295	0.0000
Intensity	0.9006	0.0466	0.9934	19.3463	0.0000

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.0656	98.68%	98.42%	0.9934

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value
Regression	1	1.6112	1.6112	374.2811	0.0000
Residual (Error)	5	0.0215	0.0043		
Total	6	1.6327			

Figure 34: Log-Linear Regression Output

These results may vary in appearance by software package, but all should contain the same basic information. **Figure 33** shows a scatter plot with *Intensity* on the *x*-axis and *Cost* on the *y*-axis, with the fit regression curve going through the data. The results displayed in **Figure 34** are identical to those of OLS, but now represent the variables in the transformed space. The first table of coefficients is a standard output showing the estimated values for the regression equation, standard errors, and t-tests for significance. This is the model in the transform space, with both *Cost* and *Intensity* transformed. Recall that the model is solving for $\beta'_0 = \log(\beta_0)$. Thus, the regression equation in the transform space is,

$$Cost' = 5.8183 + 0.9006 \cdot Intensity'$$

Therefore, the Power Model in unit space becomes,

$$\begin{aligned} Cost &= e^{5.8183} \cdot Intensity^{0.9006} \\ &= 336.4 \cdot Intensity^{0.9006} \end{aligned}$$

Figure 34 also shows several regression statistics that are common to OLS such as the Standard Error, R-squared, and the F-statistic. The final table is the Analysis of Variance (ANOVA) table, which is a standard view for significance and provides key diagnostic values relative to the regression model. Some statistical packages also provide results in unit space.

Accepting the CER requires additional analysis. [Step 4: Validate CER](#) discusses the process of using statistical outputs to select the most appropriate model.

3.3.4 Generalized Linear Model (GLM)

Methods of least squares, such as [3.3.1 Ordinary Least Squares \(OLS\)](#) and [3.3.2 Generalized Least Squares \(GLS\)](#) can be fairly restrictive as both methods depend on a normal error distribution and operate

by minimizing the model's sum of squared error term. A [Generalized Linear Model \(GLM\)](#) is a generalization of the standard linear model allowing for non-normal error distributions (see Section [4.2.1.4 Normality of Errors](#)), and limited non-linear function forms (see Section [4.2.1.5 Linearity](#)), such as [2.8.2 Power Functional Form](#) and [2.8.3 Exponential Functional Form](#).

GLM expresses a response whose mean is a function of a linear predictor, and an additive error term that follows a distribution belonging to the exponential family (Appendix [A.2.2.3](#)). The model has many convenient properties analogous to those of OLS, but with added complexities. To accommodate non-normal error distributions, GLM utilizes [Maximum Likelihood Estimation \(MLE\)](#) (Appendix [A.4.7.2](#)). The error distribution assumption provides a systematic framework to determine significance of the results.

However, the coefficient estimates typically do not have a closed-form solution. Solving for the coefficients by maximizing the likelihood function of the model requires an algorithm that is included in many statistical software packages. Under certain conditions, statistical inference properties of the GLM are preferable to those of both the [3.3.3 Transformable Linear and the Log-Linear Model](#) and [3.3.5 Non-linear Least Squares \(NLS\)](#) forms.

There are several applications of GLM used to solve for specialized regression models. Binary response variables are predicted using logistic regression. Count data are often modeled using Poisson regression. While beyond the scope of this guide, both are common enough to warrant awareness. Of particular interest to CER construction, GLM provides the flexibility to directly fit a lognormal error term (or approximation of) and power and exponential models without having to first transform the data.⁵³

The GLM is an advanced tool in the CER toolbox and is further described in Appendix [A.4.4 Generalized Linear Model \(GLM\)](#).

While not a GLM, other methods do use the MLE approach. Specifically, a MLE approach directly applicable to the log-linear model (without the need for an approximation) is possible. This methodology is discussed in [Appendix B Maximum likelihood estimation for Regression of Log Normal error \(MRLN\) Summary](#).

3.3.5 Non-linear Least Squares (NLS)

3.3.5.1 Overview

The models presented in earlier sections all rely on their own respective sets of fairly restrictive assumptions concerning the behavior of the error term and the functional form of the model. There are many scenarios where these assumptions are either violated by the data or simply not realistic from an engineering / subject matter expert perspective. When these assumptions are violated, Non-linear Least Squares (NLS) can be used as a last resort methodology.

⁵³ In practice, gamma regression is a more common approach. The log-normal distribution can be approximated by certain parameterizations of the gamma distribution. Many properties are shared, but mathematically the gamma distribution is more convenient to work with.

NLS allows for any functional form of the model (Section [4.2.5](#)), including “additive” or “multiplicative” assumptions on the error term. A common example of a NLS form is the triad model, introduced in Section [2.8.5](#). This form arises in the “fixed-cost” CIC application, where the cost data are assumed to include fixed costs, which do not decrease over time. This form provides the analyst with flexibility and can be used to fit nearly any functional form to the data. However, NLS does not provide the desired statistical properties similar to other regression methods. As a result, use NLS as a last resort methodology.

– *Terminology* –

Non-linear Least Squares (NLS) is common terminology for the regression methodology of using numerical methods to minimize some error term (objective function) for any desired functional form and error term combination. This has also been called “General Error Regression Methodologies” (GERM.). GERM is essentially an NLS method. “ZMPE,” a special case of GERM includes a constraint (sum of percent errors is zero), is discussed in [3.3.5.5 Zero Percentage Bias Minimum Percentage Error \(ZMPE\)](#).

Non-linear models remain an area of heavy research in the statistics community and many, if not all, of the convenient properties of the preceding methods are lost. Nearly any functional form is possible and methods for handling correlated and non-constant errors of many different distributions exist. Recall the normal distribution of errors assumption is required only for inference when utilizing a least squares method.

Below is the general statistical formulation of the NLS model. The first part of the statement expresses the dependent variable, or vector, \mathbf{y} , is equal to a function of the independent variables, \mathbf{X} , and the coefficient variables, or vector, $\boldsymbol{\beta}$, plus random error, $\boldsymbol{\varepsilon}$. The second part of the statement indicates the error term is assumed to be normally distributed with a mean of zero, and with covariance matrix, $\boldsymbol{\Sigma}$. One example of a common non-linear form is referred to as the “triad model” (introduced in Section [2.8.5](#)), where $f(\mathbf{X}; \boldsymbol{\beta}) = (\beta_0 + \beta_1 x^{\beta_2}) + \boldsymbol{\varepsilon}$.

For the purposes of this guide, $\boldsymbol{\Sigma}$ has the same restriction as introduced in Section [3.3.2.2 Weighted Least Squares \(WLS\)](#), that the diagonal entries must all be greater than zero, and the off-diagonals must all be equal to zero. The reciprocal of these diagonal entries are frequently referred to as the vector of weights, \mathbf{w} , and the covariance matrix can be notated as $\boldsymbol{\Sigma} = \langle \mathbf{w}^{-1} \rangle$. In the context of the non-linear model, a common approach is the method of Iteratively Reweighted Least Squares (Appendix [A.4.7.1](#)), which iteratively selects and adjust these weights automatically when fitting the NLS model. This process is nearly identical to that covered under WLS Method 4, using the previous iterations predicted values to weight the residuals of the current iteration. This is the basis of the MUPE algorithm ([3.3.5.4 Minimum Unbiased Percentage Error \(MUPE\)](#)). However, weighting of the residuals is not required in order to run this method (i.e., $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$), thus minimizing standard “additive error.”

$$\mathbf{y} = f(\mathbf{X}; \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \text{ and } \boldsymbol{\Sigma} = \langle \mathbf{w}^{-1} \rangle$$

This statement explicitly captures the assumptions made when conducting NLS. These are the same four assumptions introduced with OLS, but with significant modifications. The covariance matrix, $\boldsymbol{\Sigma}$, is used to

standardize the residual errors, scaling the errors to have the same variance. Similar assumptions to OLS apply to these standardized data:

- (1) *Independence* of errors. Each error, ε_i , is not related to any other error term.
- (2) *Homoscedasticity*. Each error, ε_i , has a constant variance, σ^2 , across the data range.
- (3) *Normality*. The errors are normally distributed with a mean of zero.
- (4) *Functional Form*. The statement of the functional form is an assumption on the type of appropriate modeling function for \mathbf{y} .

Under the assumptions, the results of NLS are asymptotic, meaning that the coefficients are optimal in a large sample. In order to fit this regression model using the least squares method, values for the coefficient vector β use the equation below.

$$\arg \min_{\beta} \boldsymbol{\varepsilon}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon} = \arg \min_{\beta} (\mathbf{y} - f(\mathbf{X}; \beta))' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - f(\mathbf{X}; \beta))$$

3.3.5.2 Application of NLS

NLS does not have a closed-form solution. Numerical optimization methods are required to generate coefficient estimates. These methods are generally reliable, though the concept of a local versus global solution now requires consideration. Typically, non-linear optimization methods involve search methods or variations on methods of steepest descent. These methods commonly start with an initial “guess” for the optimal solution of the function and use the derivative (or approximations thereof) to iterate closer to the solution. Unfortunately, this raises the possibility that the method provides a locally optimal solution when a globally optimal solution is more desirable. As a result, the solution provided by these methods has the potential to be highly influenced by both the starting value(s) of the algorithm and by the optimization function.

Consider the plot of the function in Figure 35. Note how the three different initial conditions (circles) lead to three different minimal values (diamonds). In this case, the red initial values (two circles on the right) arrive at local minimum values while the green initial value (circle on the left) arrives at the global minimum value. The functional forms used in NLS models have the potential to lead to similar situations. In these cases, different initialized values for β could result in different final estimates for the model coefficients, $\hat{\beta}$.

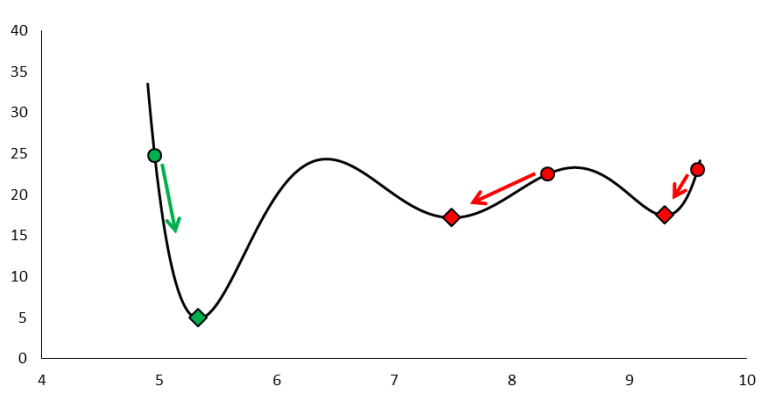


Figure 35: Local versus Global Solution

In practice, generating coefficients involves minimizing the error term for the regression. Unconstrained optimization involves the minimization of functions with no restrictions on the value of the independent variables. In most cases of regression, unconstrained non-linear optimization techniques are the most relevant. However, there are special cases where constraints upon the independent variables are required (Section [3.4 Estimation with Prior Information](#)). The independent variables that minimize the errors are the parameters of the regression function (β), and not the independent variables of the regression equation (X).

Most numeric methods of continuous non-linear optimization involve iterative methods. In these methods, each step improves upon the solution from the previous step, bringing the final solution closer to the optimum and termination of the algorithm. In the case of minimizing a sum of squared errors term, the Levenberg-Marquardt (sometimes referred to as Modified Marquardt) algorithm is often the method of choice. There are cases when a more robust but less efficient downhill simplex (Nelder-Mead) method may be preferred.

One strategy for complex functions would be to use the Levenberg-Marquardt method first. If the software package has difficulty converging using the Levenberg-Marquardt method, recommend trying the downhill simplex method. Many statistical software packages⁵⁴ allow for a custom model specification, and automatically run the algorithm to produce the regression results and relevant diagnostics.

3.3.5.3 NLS Example

Consider the same sample data introduced in Section [3.3.3 Transformable Linear and the Log-Linear Model](#) in [Table 17](#) for one independent variable *Intensity* ($kWperCm^2$) and dependent variable *Cost* (\$M). After viewing a scatter plot of the data, a Triad Model, $y = \beta_0 + \beta_1 x^{\beta_2} + \epsilon$, is fit to the data. Using CO\$TAT, the NLS regression analysis is conducted using the Levenberg-Marquardt algorithm by default (Downhill Simplex and Gauss Newton are also available) and standard outputs are returned, displayed in **Figure 36** and **Figure 37**. In this example, the additive error is minimized. If there are concerns that the variance may be non-constant, other methods may be used (e.g., IRLS, MUPE, WLS, etc.).

⁵⁴ See Appendix [A.4.5.1](#) for use of the CO\$TAT software package and Appendix [A.4.5.2](#) for use of Excel for NLS modeling. ACEIT makes use of language and methodologies similar to the NLS argument presented here, but under more cost analysis unique language (Section [0 General Error Regression Models \(GERM\)](#)). Use of Excel solver has many cautions and is a methodology for the expert user.

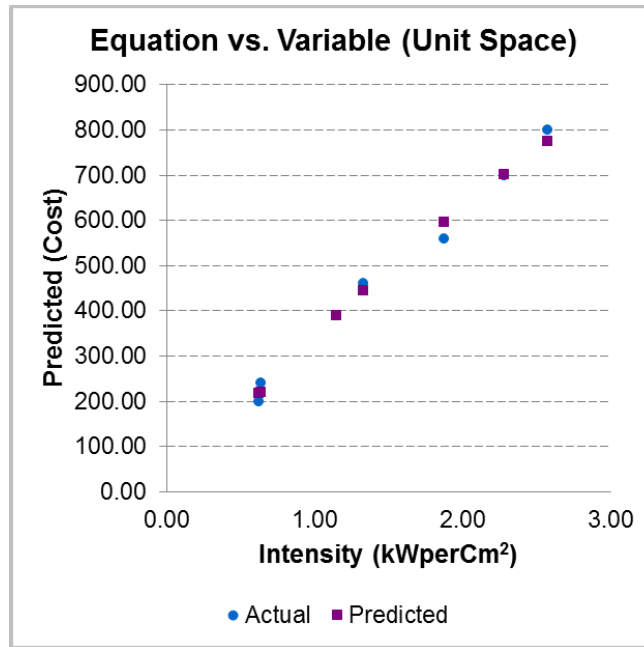


Figure 36: NLS Regression Model Scatter Plot

I. Equation Form & Error Term

Model Form:	Unweighted Non-Linear Model
Non-Linear Equation:	Cost = (-90.05) + 433.6 * Intensity ^ 0.73
Error Term:	Multiplicative
Minimization Method:	Modified Marquardt

II. Fit Measures

Coefficient Statistics Summary

Variable/Term	Coefficient Estimate	Approximate Std Error	Approximate Lower 95% Confidence	Approximate Upper 95% Confidence
a	-90.0538	200.9392	-648.2629	468.1553
b	433.5642	214.6412	-162.7092	1029.8375
c	0.7300	0.3109	-0.1335	1.5936

Least Squares Minimization Summary Statistics

Source	DF	Sum of Squares	Mean SQ = SS/DF
Residual (Error)	4	0.0200	0.0050
Total (Corrected)	6	1.6327	

Goodness-of-Fit Statistics

Std. Error (SE)	Approx. R-Squared	Approx. R-Squared
0.0707	98.77%	98.16%

Figure 37: NLS Regression Output

Figure 36 shows a common visual plotting *Intensity (kWperCm²)* on the *x*-axis and *Cost (\$K)* on the *y*-axis, with the fit regression points superimposed on the data. The results displayed in **Figure 37** are now distinctly different to those of OLS. The first table of coefficients depicts the estimated values for the regression equation and the approximate standard errors, and t-tests for significance.

Next, there are several common regression statistics such as the approximate R-squared and the Standard Error (SE). Many metrics relevant in the linear model are no longer applicable in the non-linear setting such as an ANOVA table. Statistical results produced from the NLS model are used less frequently for comparison.

However, in situations where a CER can only be fit via NLS or engineering judgment, NLS may be preferred. Be sure to consider other options when OLS fails.

Accepting the CER requires additional analysis. [Step 4: Validate CER](#) discusses the process of using statistical outputs to select the most appropriate model.

3.3.5.4 Minimum Unbiased Percentage Error (MUPE)

– Terminology –

MUPE as a whole is a broad methodology that uses iteratively reweighted least squares (IRLS) to minimize an objective function where the variance of the dependent variable is not a constant. This concept directly applies to non-linear models.

The MUPE methodology is a specific case of NLS, which minimizes a multiplicative error model. This approach is applied as a modifier to the Levenberg-Marquardt algorithm. The model weights are derived using the process described in Section [3.3.2.2 Weighted Least Squares \(WLS\)](#). All properties, assumptions, formulas, equations, and caveats discussed in Section [3.3.5](#) continue to be applicable.

MUPE is utilizing a non-linear methodology to minimize the relative residual sum of squares (i.e., percentage error (Percent Error [PE])), defined relative to the predicted values,

$$PE = \frac{y - \hat{y}}{\hat{y}}$$

Starting with $w=1$ (i.e., the additive NLS estimator), MUPE iteratively fits a NLS model using the squared predicted values of the previous iteration as the current weighting vector. This repeats until the coefficient estimates converge within a specified tolerance limit, known as iteratively reweighted least squares (IRLS). The weight vector at a given iteration step γ is expressed as,

$$w_y = \frac{1}{\hat{y}_{r-1}^2}$$

3.3.5.5 Zero Percentage Bias Minimum Percentage Error (ZMPE)

The ZMPE methodology is a constrained minimization process that can be applied to a linear or non-linear functional form. ZMPE minimizes a multiplicative error model and can be broken down into two components: the Zero Percentage Bias component and the Minimum Percentage Error component. Much like MUPE, ZMPE is utilizing a non-linear methodology to minimize the residual sum of squares relative to the predicted values,

- Zero Percentage Bias in this context is defined as the constraint that average (or equivalently the sum) of the relative residuals (i.e., the average percentage errors) must equal zero.

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{f(x_i) - y_i}{f(x_i)} \right) = 0$$

- Minimum Percentage Error in this context is defined as the standard percent error of the estimate (SPE) expressed as a percentage error (i.e., multiplicative) relative to the predicted value

$$SPE = \sqrt{\frac{1}{n - m} \sum_{i=1}^n \left(\frac{f(x_i) - y_i}{f(x_i)} \right)^2}$$

– Terminology –

The term bias is one with a strict statistical definition, covered in Section 5.0. Bias is often a desirable statistical property and it is common to hear a methodology referred to as being “biased” or “unbiased.” ZMPE uses a non-standard definition of bias. ZMPE defines bias as the sum of the weighted residuals (i.e., the residuals in fit space).

No distribution assumptions are made on the CER error term when fitting the ZMPE CER. While this may sound appealing, the lack of assumptions results in a void when attempting to validate the CER. This objective function can be presented as a special case of the one presented for the general NLS.

$$\arg \min_{\beta} \boldsymbol{\varepsilon}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\varepsilon} = \arg \min_{\beta} (f(\mathbf{X}; \boldsymbol{\beta}) - \mathbf{y})' \boldsymbol{\Sigma}^{-1} (f(\mathbf{X}; \boldsymbol{\beta}) - \mathbf{y}) \text{ such that } \sum_{i=1}^n (\varepsilon_i^*) = 0$$

Where,

$$\begin{aligned} \varepsilon_i^* &= \frac{f(x_i, \boldsymbol{\beta}) - y_i}{f(x_i, \boldsymbol{\beta})} \\ \boldsymbol{\Sigma}^{-1} &= \mathbf{W} \\ &= \left\langle \frac{1}{f(x_i, \boldsymbol{\beta})^2} \right\rangle \text{ the diagonal matrix of the reciprocal of the squared predicted values} \end{aligned}$$

This objective function appears very similar to that of NLS with a weighted residual (i.e., MUPE). However, there are two distinct differences. First, the weighting matrix, \mathbf{W} , is dependent on the predicted value of each observation, rather than the predicted value of a prior step in the numerical algorithm. Second, there is a constraint that the weighted residuals (i.e., percentage errors) must sum to zero. These two alterations have significant effects on the optimization routine used to solve the objective function. NLS methodologies derive approximate statistical results based on results from the final iteration of the optimization algorithm. This is no longer possible when using ZMPE. The following are key items to note when considering using the ZMPE methodology:

- The relative residual calculation is reversed to represent the predicted value minus the observed value. This reverses the meaning of a positive or negative residual from all of the other methodologies.
- Section 5.0 provides a definition of bias, which is different from the ZMPE definition.

- There is no consistent statistical output or methodology to validate the model.
- Cost analysts need to quantify the uncertainty about the point estimate. Other regression methods do this by calculating a prediction interval. ZMPE provides no universally accepted methodology to define the location of the point estimate in the error distribution nor the shape of the error distribution.
- In the linear case, the Zero Percentage Bias constraint is satisfied by using a WLS (Section [3.3.2.2](#)) methodology with an intercept term.
- The Zero Percentage Bias constraint reduces the degrees of freedom by one.
- The Zero Percentage Bias constraint specification defines a ZMPE CER. Without this constraint, ZMPE becomes an MPE CER.

ZMPE is a popular methodology in cost analysis. Research and papers that attempt to resolve the observations above are available in the literature. For more information on ZMPE its associated research, see Appendix [0 General Error Regression Models \(GERM\)](#).

3.3.6 Ridge Regression

3.3.6.1 Overview

Ridge regression is an extension of Section [3.3.1 Ordinary Least Squares \(OLS\)](#) for use when high multicollinearity (Section [4.3.2 Multicollinearity](#)) is present. Multicollinearity occurs when there is a near linear relationship between two or more independent variables of the regression equation. Intuitively, the nearly linear relationship between two or more of the independent variables makes it difficult for the regression to discern the true parameters, thus resulting in very large variances around the coefficient estimates. Even a very small change in the data set could result in a large change in the values of the coefficients.

If suspected, the best course of action to correct for multicollinearity is to collect more data and see if the additional data resolves the problem. For example, CIC data sets with monotonically increasing unit numbers and rising production rates are often highly collinear. In this instance, additional data collection over the complete production run may capture stable and falling production rates, resolving the multicollinearity problem. Another course of action is to remove one or more of the collinear variables from the model. In cases where these options are impractical or ineffective in dealing with multicollinearity, the method of Ridge Regression is an option.

Ridge Regression, or simply Ridge⁵⁵, usually takes on the same form for $f(\mathbf{X}; \boldsymbol{\beta})$ as OLS. Just like OLS, Ridge expresses a linear functional form and a normal additive error term; that is, the assumption that the errors are independently and identically distributed as normal. Closed-form formulas (Appendix [A.4.6](#)) exist to solve for both the coefficient estimates and for all of the statistical metrics of interest. Again, the normality assumption provides a systematic framework for inference.

⁵⁵ It is also common to see Ridge referred to as L_2 Regularization.

While many of the properties are identical to those of OLS, Ridge applies an additional constraint to the model. Ridge is derived by issuing a restriction on the length of the coefficient vector. This restriction is applied on the sum of the squared coefficients, not on the individual coefficients, and should not be confused with Section [4.2.7 Restricted Least Squares \(RLS\)](#).

This method derives estimates of the regression parameters by decreasing the variance of the parameters. However, this decreased variance in the parameter estimates introduces bias (see [Step 5: Characterize Uncertainty](#) for more on bias). This section will discuss Ridge in the context of Section [3.3.1 Ordinary Least Squares \(OLS\)](#), however, the concept applies to Section [3.3.2 Generalized Least Squares \(GLS\)](#) and Section [3.3.3 Transformable Linear and the Log-Linear Models](#).

Ridge is highly dependent on the scale of the data. As a result, all predictors must be centered around zero and scaled to have the same standard deviation. Calculate the centered and scaled predictors, \mathbf{X}^* , as follows,

$$x_j^* = \frac{x_j - \bar{x}_j}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

Below is the statistical formulation for the Ridge model in matrix form. The first part of the statement expresses that the response variable, or vector, \mathbf{y} , is equal to the matrix of scaled predictors, \mathbf{X}^* , multiplied by the coefficient variables, or vector, $\boldsymbol{\beta}$, plus some random error, $\boldsymbol{\varepsilon}$. The second part of the statement indicates the assumption that the error term is normally distributed with a mean of zero, the same constant variance of σ^2 , and a covariance of zero.

The third part to the statement indicates that the sum of squares of the coefficients is restricted to be less than some unknown constant, c . This constant translates into (but is not equivalent to) what is known as the ridge parameter in the analysis. This guide denotes the ridge parameter as λ (lowercase lambda), but it is not uncommon to see it denoted as κ (lowercase kappa), or k . This additional constraint to the problem introduces bias into the coefficient estimates. As the value of c decreases, λ increases, and the amount of bias introduced into the estimate of $\boldsymbol{\beta}$ increases, but with a corresponding reduction in variance.

$$\mathbf{y} = \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \text{ such that } \boldsymbol{\beta}' \boldsymbol{\beta} < c$$

This statement explicitly captures the assumptions made when conducting Ridge. These are the same assumptions introduced with OLS:

- (1) *Independence* of errors. Each error, ε_i , is not related to any other error term.
- (2) *Homoscedasticity*. Each error, ε_i , has a constant variance, σ^2 , across the data range.
- (3) *Normality*. The errors are normally distributed with a mean of zero.
- (4) *Linearity*. The relationship exhibits a constant slope over the data range.
- (5) *Error term*. The error is not proportional to the independent variables

In OLS, assumptions (1), (2), and (4), the Gauss-Markov theorem states the coefficient estimates are the Best Linear Unbiased Estimators (BLUE) of the true parameter values, with the lowest variance.

However, with Ridge regression the estimator is no longer unbiased, and can be shown that for any $\lambda > 0$, the Ridge estimator has a smaller variance than the BLUE estimator derived from OLS. To fit this

regression model by method of least squares, find values for the coefficient vector β , which minimize the penalized SSE. The resulting vector of coefficients is the centered and scaled predictors, X^* . Using the method of Lagrange multipliers (Appendix [A.4.7.3](#)) to enforce the constraint on β , the penalized objective function is as follows. A normal application is to define β with the smallest possible values while resulting with acceptable fit parameters⁵⁶.

$$\arg \min_{\beta} \epsilon' \epsilon = \arg \min_{\beta} ((y - X^* \beta)'(y - X^* \beta) + \lambda \beta' \beta)$$

3.3.6.2 Applying Ridge Regression

The first step in Ridge regression is to center and scale the independent variables. Since the variables are transformed to fit the model, they must also be transformed back to unit space after fitting the model to use for analysis. However, most software packages perform both of these steps automatically without the user ever being aware. Ridge is a convenient model because the problem has a closed-form solution. Once a value for the ridge parameter, λ , is selected, formulas exist for the estimated values of the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ as well as the estimated variance of the error term, $\hat{\sigma}^2$.

The formula for $\hat{\beta}$ is:

$$\hat{\beta} = (X'X + \lambda I)^{-1} X'y$$

In practice, the ridge parameter is varied and coefficient parameters are estimated for each of those values. A value of λ is selected where the parameter estimates start to stabilize (i.e., the relative change in the estimate is small as λ increases). This analysis approach of varying the value of λ and observing the convergence of the coefficient estimates creates a ridge trace (or perturbation) plot. Note as λ increases so does the standard error of the estimate, despite variance around the coefficients decreasing. A normal application is to select the smallest value of λ required for the parameter estimates to stabilize. Current practice is to limit the λ value at or below .3.

The Ridge estimator is known to be biased. The equation for the bias is,

$$bias(\hat{\beta}) = -\lambda(X^* X^* + \lambda I)^{-1} \beta$$

Where,

- $bias(\hat{\beta})$ = the $(k + 1) \times 1$ vector of coefficient bias values
- X^* = the $n \times (k + 1)$ matrix of centered and scaled predictors (and the intercept)
- I = the $(k + 1) \times (k + 1)$ identity matrix
- β = the $(k + 1) \times 1$ vector of true parameter values (unknown)
- λ = the ridge parameter (scalar)

⁵⁶ Acceptable fit parameters are typically determined by the specific cost estimating organization.

Note the bias equation contains the true parameter, β . Since β is unknown, the bias cannot be calculated in practice.

3.3.6.3 Ridge Regression Example

The data set shown in [Table 14](#) is used in this example. Power and Aperture Area (Aper) are highly correlated. Section [4.3.2](#) uses an expanded example dataset to provide more details on diagnosing multicollinearity.

As a first attempt to gain insight into the relationship between the predictors and the response, CO\$TAT is used to fit the linear regression analysis and return standard outputs. **Table 18** displays the table of coefficients from the OLS analysis. The Beta Value column represents $\hat{\beta}^*$, the coefficients associated with the centered and scaled data.

Table 18: OLS Results Before Applying Ridge Regression

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	37.3129	449.4459		0.0830	0.9378	0.0622
Power	28.2134	4.6985	0.9777	6.0047	0.0039	0.9961
Aper	6.1047	57.2542	0.0174	0.1066	0.9202	0.0798

Request the Ridge Statistics and Trace Plot from the Report Styles menu. In addition to the OLS outputs (demonstrated in Section [3.3.1.3](#)), the software returns a Ridge trace table and trace (perturbation) plot, displayed in **Figure 38** and **Figure 39**.

IV. Ridge Perturbation Parameter (RPP) & Related Statistics

RPP by Non-iterative Procedure	0.0061
RPP by Iterative Procedure	0.0068
Von Neumann Test for Autocorrelation	3.6234
Durbin-Watson Statistic	3.1057
Determinate of (X'X)	0.1106
Measure of Ill conditioning	9.0400

Ridge Trace Table

Ridge Parameters	BETA (1)	BETA (2)	SSE
0.00	0.9777	0.0174	3539.1306
0.02	0.8478	0.1371	4104.5678
0.04	0.7695	0.2055	5005.9975
0.06	0.7164	0.2489	5884.9395
0.08	0.6775	0.2782	6696.4412
0.10	0.6474	0.2990	7450.5198
0.12	0.6231	0.3141	8164.3838
0.14	0.6029	0.3253	8852.9625
0.16	0.5857	0.3337	9527.6302
0.18	0.5707	0.3400	10196.6826
0.20	0.5575	0.3447	10866.0511
0.22	0.5457	0.3482	11539.9189
0.24	0.5349	0.3508	12221.1902
0.26	0.5251	0.3526	12911.8338
0.28	0.5160	0.3537	13613.1334
0.30	0.5076	0.3544	14325.8684

Figure 38: Ridge Specific Regression Output

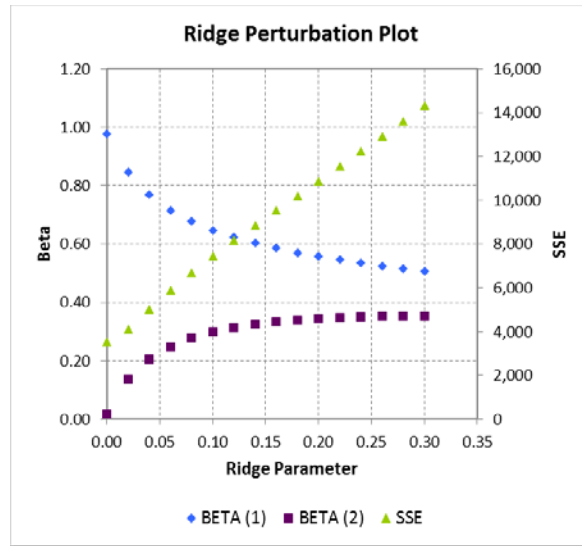


Figure 39: Ridge Trace (Perturbation) Plot

These results may vary in appearance by software package, but all should contain the same basic information. **Figure 38** shows a table of statistics which can be valuable to the identification of high multicollinearity. The Trace Table shows the Ridge parameter, λ , on the left, and the coefficient estimates and model standard error. The first row with the ridge parameter set to zero, $\lambda = 0$, is the OLS regression model. Note the coefficient values match those in the Beta Value column of [Table 18](#). This is because the Trace Table displays the centered and scaled variety of the coefficients. The recommended practice is to choose the ridge parameter, λ , that stabilizes corresponding Beta Values while minimizing SSE.

Figure 39, represents the Trace Table in graphical form. When selecting a λ , look for the spot where the coefficients start to “level out.” COSTAT tests the range of λ 's from 0 to 0.30 and in this example, the leveling happens around $\lambda = 0.12$.

After selecting a Ridge parameter, run the linear model again. The model is now run with the specific $\lambda = 0.12$, and the coefficient table is displayed in **Table 19**.

Table 19: Before and After Ridge Example Coefficients with Ridge Parameter = 0.12

Before Ridge Regression						
Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	37.3129	449.4459		0.0830	0.9378	0.0622
Power	28.2134	4.6985	0.9777	6.0047	0.0039	0.9961
Aper	6.1047	57.2542	0.0174	0.1066	0.9202	0.0798

After Ridge Regression						
Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	-750.9813	682.6385		-1.1001	0.3330	0.6670
Power	17.9806	1.6667	0.6231	10.7879	0.0004	0.9996
Aper	110.4452	20.3101	0.3141	5.4380	0.0055	0.9945

Compared to the results in [Table 18](#), the coefficients differ greatly and their standard errors are much smaller. This becomes critical in Section [4.4.2 Validate Variable Set](#) and in [Step 5: Characterize](#)

Uncertainty. In this example, the coefficients significantly varied between OLS and Ridge. There were also significant changes in the coefficient p-values. For example, *Aper* went from being not statistically significant to being statistically significant. These results were consistent with the provided SME inputs. More on this topic is covered in Section [4.3.2 Multicollinearity](#).

Accepting the CER requires additional analysis. Refer to [Step 4: Validate CER](#) for more information regarding CER validation.

3.4 Estimation with Prior Information

3.4.1 Types of Prior Information

Prior information about parameter values or relationships in a CER arises in various ways. Sometimes theoretical reasoning or practical experience suggests constraints on parameter space. CIC exponents, for example, are generally negative. In other cases, estimates obtained from previous or complementary empirical studies typically give information about current parameters. For example, visibility into Economic Order Quantities (EOQs)⁵⁷ for class standard equipment on a surface combatant program, culled from shipyard/supplier contracts, might provide bottom-up information on rate effects in a CIC for material. In all these cases, prior information exists *outside of the sample* under consideration.

Prior knowledge represents information about the population regression equation. In practical applications in defense cost analysis, three basic types of information are generally available:

- (1) Exact knowledge of parameter *relationships* (Section [3.4.2](#))
- (2) Pseudo-exact knowledge of parameter *values* (Section [3.4.3](#))
- (3) Inexact knowledge of parameter *values* (Section [3.4.4](#))

3.4.2 Exact Prior Information on Parameter Relationships

The knowledge of exact prior information on parameter relationships is a well-studied problem in the fields of statistics and econometrics. In this scenario, *a priori* knowledge provides information on how parameters must behave in relation to each other and are solved by applying constraint equations to the chosen regression methodology (Section [3.3](#)). The remainder of this section examines this problem in the context of [3.3.1 Ordinary Least Squares \(OLS\)](#).

The constraint equations specification requires two components. The first component is an $a \times p$ matrix \mathbf{R} with a column for each parameter and a row for each of the a constraints enforced on the model. The second component is an $a \times 1$ vector \mathbf{q} of equality constraints. To illustrate, consider the model with $p = 4$,

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \boldsymbol{\varepsilon}$$

⁵⁷ EOQ the number of units that a company should add to inventory with each order to minimize the total costs of inventory—such as holding costs, order costs, and shortage costs.

Due to some prior information on parameter relationships, $\beta_1 + \beta_2 = 1$. This translates to,

$$\mathbf{R} = [0 \quad 1 \quad 1 \quad 0] \text{ (i.e., a } 1 \times 4 \text{ matrix)}$$

The first column relates to β_0 , the second column relates to β_1 , and so on. In this case, the vector \mathbf{q} simply is $\mathbf{q} = [1]$, since there is only one equation, which is being set equal to 1. Thus,

$$\begin{aligned} \mathbf{R}\boldsymbol{\beta} &= \mathbf{q} \text{ (i.e., a } 1 \times 1 \text{ vector)} \\ \Rightarrow 0\beta_0 + 1\beta_1 + 1\beta_2 + 0\beta_3 &= 1 \\ \Rightarrow \beta_1 + \beta_2 &= 1 \end{aligned}$$

Additional restrictions may be added using the same logic by simply adding another row to \mathbf{R} and to \mathbf{q} . Specifying a row with all zeroes except for a single one fixes that parameter to an exact value. The equality constraint in \mathbf{q} for that row would simply be the value to be fixed. Section [3.4.3](#) studies this specific case in much greater detail. Further, hypothesis tests do exist to test the hypothesis of these restrictions using a penalization of the F-test. This test is shown in Appendix [A.4.9.1 Restricted Least Squares](#).

The method of Lagrange multipliers (Appendix [A.4.7.3](#)) is used to solve the least squares problem once the restrictions are specified. Generically, this translates to an objective function of,

$$\arg \min_{\boldsymbol{\beta}} ((\mathbf{y} - f(\mathbf{X}; \boldsymbol{\beta}))'(\mathbf{y} - f(\mathbf{X}; \boldsymbol{\beta})) + \lambda(\mathbf{R}\boldsymbol{\beta} - \mathbf{q}))$$

In the OLS case, a closed form solution to the problem exists. If $\widehat{\boldsymbol{\beta}}_{OLS}$ is the unrestricted OLS solution, then the RLS solution can be expressed as,

$$\widehat{\boldsymbol{\beta}}_{RLS} = \widehat{\boldsymbol{\beta}}_{OLS} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\widehat{\boldsymbol{\beta}}_{OLS} - \mathbf{q})$$

This estimator provides the best linear unbiased estimator *under the restrictions*. If the restrictions are misguided or otherwise incorrect, then bias is introduced into the estimator. The references in Appendix [A.4.9.1 Restricted Least Squares](#) provide specific examples of this estimator as well as goodness of fit tests for the RLS solution.

3.4.3 Pseudo-Exact Prior Information on Parameter Values

The knowledge of precise, high value prior information on the value of regression parameters is rare. However, in these scenarios, fixing these parameters to precise values has the effect of removing the parameters from the regression analysis while still allowing them to influence the response. Recall under the classical framework, regression parameters are fixed, unknown values, estimated by the regression. Fixing a parameter value changes it to a fixed, known value. This is a rather strong statistical statement and should only be considered when extremely high confidence on high value prior information is possessed.

In this case, the model results in more degrees of freedom because fewer parameters require estimation. As a result, fixing a parameter increases the number of degrees of freedom in the error term, which intuitively may lead to the conclusion of increased CER precision, but in reality may have the inverse effect depending on the quality of the prior information.

There are two ways to apply this methodology to a data set. First, exact information on a parameter value is a special case of exact information on a parameter relationship. All that changes is the formulation of the constraint equations. All values in the matrix \mathbf{R} will be zero except at the location(s) of the parameter(s) to be fixed. The resulting application is the same application introduced in Section [3.4.2](#).

Alternatively, implementation of this methodology is possible by a transformation or normalization of the regression equation. From this point, the regression methodology (Section [3.3](#)) most relevant to the CER functional form can be utilized, diagnosed, and assessed ([Step 4: Validate CER](#)). Finally, reverse the normalization to produce the final CER. While these two approaches may seem very different, both provide the exact same results.

Fixing a parameter value is a very easy technique to implement. Consider a partition of the [3.3.1 Ordinary Least Squares \(OLS\)](#) model where the data, \mathbf{X} , and the respective coefficient parameters, $\boldsymbol{\beta}$, are broken into two groups. The first group, \mathbf{X}_1 , consists of all the independent variables to be fit by OLS regression. This set of parameters is denoted as $\boldsymbol{\beta}_1$. The second group, \mathbf{X}_2 , consists of all the independent variables whose parameters, denoted by $\boldsymbol{\beta}_2$, are fixed ahead of time.

$$\mathbf{X} = (\mathbf{X}_1 | \mathbf{X}_2) \text{ and } \boldsymbol{\beta}' = (\boldsymbol{\beta}'_1 | \boldsymbol{\beta}'_2)$$

Under this definition, the regression equation becomes,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \end{aligned}$$

In this case, the claim is made that the coefficient vector $\boldsymbol{\beta}_2$ is known ahead of time to be some vector $\boldsymbol{\gamma}$. Note that on the continuous scale, the probability that $\boldsymbol{\beta}_2 = \boldsymbol{\gamma}$ is zero, so a reformulation yields,

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \text{ for fixed } \boldsymbol{\gamma} \neq \boldsymbol{\beta}_2$$

And,

$$\mathbf{y}^* = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \text{ where } \mathbf{y}^* = \mathbf{y} - \mathbf{X}_2\boldsymbol{\gamma}$$

This form represents a normalization of the response for the prior information related to \mathbf{X}_2 . The OLS methodology can now be applied to the new response \mathbf{y}^* with independent variable \mathbf{X}_1 .

To evaluate the validity of such an approach, it is important to consider potential bias and variance implications on the estimator for $\boldsymbol{\beta}_1$. Taking the expected value of the estimator gives the following result:

$$\text{bias}(\widehat{\boldsymbol{\beta}}_1) = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2(\boldsymbol{\beta}_2 - \boldsymbol{\gamma}) \quad (1)$$

Similarly, by OLS, for the model fit by fixing coefficients the variance estimator for σ^2 is,

$$\hat{\sigma}^2 = \frac{\mathbf{y}^{*'}(\mathbf{I} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1)\mathbf{y}^*}{n - p_1} \text{ where } p_1 = \# \text{ parameters in } \boldsymbol{\beta}_1$$

Again, taking the expected value of the estimator gives the following result:

$$E(\hat{\sigma}^2) = \sigma^2 + \frac{(\boldsymbol{\beta}'_2 - \boldsymbol{\gamma}')\mathbf{X}'_2(\mathbf{I} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1)\mathbf{X}_2(\boldsymbol{\beta}_2 - \boldsymbol{\gamma})}{n - p_1} \geq \sigma^2 \quad (2)$$

For both expressions (1) and (2), above, the bias is zero when either $\boldsymbol{\gamma} = \boldsymbol{\beta}_2$ or $\mathbf{X}_1 \perp \mathbf{X}_2$, that is, when the parameter was fixed to its true theoretical value or if the independent variables in \mathbf{X}_1 have zero correlation with the independent variables in \mathbf{X}_2 ; both occurring with probability zero outside of a controlled setting (i.e., a designed experiment). Also observe that the expected value for $\hat{\sigma}^2$ is σ^2 plus an always-positive quadratic form. From these expressions, it can be concluded that fixing variables based on pseudo-exact prior information *always* results in biased coefficients, and *always* results in an inflated estimate of the model variance, $\hat{\sigma}^2$.

Consider the formulations for both the MSE (Section [4.4.1](#) and Section [5.0](#)) and the Margin of Error (MOE) of a Prediction Interval (Section [5.3](#)):

$$MSE = \frac{SSE}{df_{error}}$$

$$MSE(\hat{\mathbf{y}}) = Var(\hat{\mathbf{y}}) + bias^2(\hat{\mathbf{y}})$$

$$MOE = Critical\ Value \cdot Standard\ Error$$

$$= t_{1-\frac{\alpha}{2}}(df_{error}) \cdot \sqrt{MSE} \sqrt{1 + (\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$$

The OLS model fit with all the independent variables is the one with the smallest SSE. However, in the MSE formulation, there is the df_{error} term in the denominator. Thus, if the increase in the degrees of freedom is enough to counteract the increase in SSE, the pseudo-exact prior model may have a lower MSE. A reduction in MSE occurs when the reduction in variance exceeds any increase due to bias.

The prediction interval critical value comes from the t-distribution with the degrees of freedom for the error of the model. More degrees of freedom result in lower values of the critical value, and therefore narrower prediction intervals. When moving from 1 to 2 degrees of freedom with a small α , this can have a huge impact on the MOE. However, the change from even 3 to 4 degrees of freedom may not be enough to outweigh any increase in the MSE. Any reduction in uncertainty due to a narrower prediction interval must be balanced with the unknown impact of a biased point estimate.

Fixing variable coefficients will bias all coefficients in the model and will inflate variance. This may lead to CERs that are perceived to be data driven. Only when fixing the coefficient to its exact theoretical value will the remaining coefficients be unbiased with minimum variance. The resulting model will be biased with the only question being by how much. This question is difficult to answer, and can only be hypothesized by considering the rationale used when fixing a coefficient. In most cases, the preferred methodology would be fitting the full model with fewer degrees of freedom. In the case where $n < p$, utilize an advanced regression methodology capable of solving the problem, such as LASSO (Appendix [A.4.9.5](#)).

Example

To illustrate this concept, consider the following model with $n = 6$ and four independent variables, or $p = 5$ parameters:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \varepsilon$$

In order to assess the impacts of fixing a coefficient, a data set is simulated (provided in [Appendix F](#)) and the full model is fit. **Table 20** shows the results.

Table 20: Fixed Coefficient Example – Full Model

Parameter	True Value	Estimate	Bias	Standard Error
β_0	1000	1410.16	0.00	724.84
β_1	8	3.82	0.00	4.78
β_2	-12	-11.71	0.00	9.64
β_3	3	2.97	0.00	0.77
β_4	25	27.92	0.00	68.29

Fitting this model results in only $n - p = 6 - 5 = 1$ degree of freedom. Suppose prior knowledge indicates that $\beta_3 = 3$. The normalized equation becomes,

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \varepsilon$$

$$y - 3x_3 = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_4x_4 + \varepsilon$$

$$y^* = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_4x_4 + \varepsilon \text{ where } y^* = y - 3x_3$$

The response is thus normalized for x_3 and the model now has 2 degrees of freedom. **Table 21** shows the results of the regression on the new, normalized response. The estimator remains unbiased and appears very similar to that of the full model. This is expected, since the value was fixed to the true theoretical value, and very close to the value of the OLS estimate, 2.97. In this scenario, fixing a coefficient results in unbiased coefficients for the remaining variables, with smaller standard errors, and with one additional degree of freedom. In this perfect example, the pseudo-exact prior information was very valuable and resulted in a superior model.

Table 21: Fixed Coefficient Example – Fixed $\beta_3 = 3$

Parameter	True Value	Estimate	Bias	Standard Error
β_0	1000	1406.31	0.00	506.5
β_1	8	3.84	0.00	3.4
β_2	-12	-11.75	0.00	6.8
β_3	3	-	-	-
β_4	25	26.64	0.00	40.1

Now suppose the prior information suggests that $\beta_3 = 5$. Just as before, **Table 22** shows the results of the regression on the new, normalized response. Now the estimator is biased. In fact, β_4 is no longer close to the true value, and is actually of the wrong sign. Note that the full assessment ([Step 4: Validate CER](#)) of the CER has not been conducted, as β_4 appears to be insignificant and a strong candidate for removal

from the model. In addition to being biased, the standard errors of the model are roughly twice as large compared to the original full model in [Table 20](#). Recalling the critical value in the prediction interval, increasing from 1 to 2 degrees of freedom for the error term at the $\alpha = 0.05$ significance level, lowers the critical value from 12.71 to 4.30. This methodology may generate results indicating the fixed coefficient proved valuable. While parameter standard errors doubled, the critical value in the prediction interval margin of error reduced by two thirds.

Table 22: Fixed Coefficient Example – Fixed $\beta_3 = 5$

Parameter	True Value	Estimate	Bias	Standard Error
β_0	1000	1109.65	296.66	1428.2
β_1	8	4.92	-1.08	9.5
β_2	-12	-14.83	3.08	19.1
β_3	3	-	-	-
β_4	25	-72.31	98.95	113.2

Figure 40 carried out this exercise across a range of values from $\beta_3 = -1$ to $\beta_3 = 7$. For each fixed value, a point estimate with the upper 95% prediction interval was generated at the true parameter values. The horizontal red line represents the predicted value using the full model, with the dashed red line representing the respective upper bound with 1 degree of freedom. The black line represents the predicted value with β_3 fixed according to the value represented by the x -axis. The black dashed line is the respective upper bound with 2 degrees of freedom.

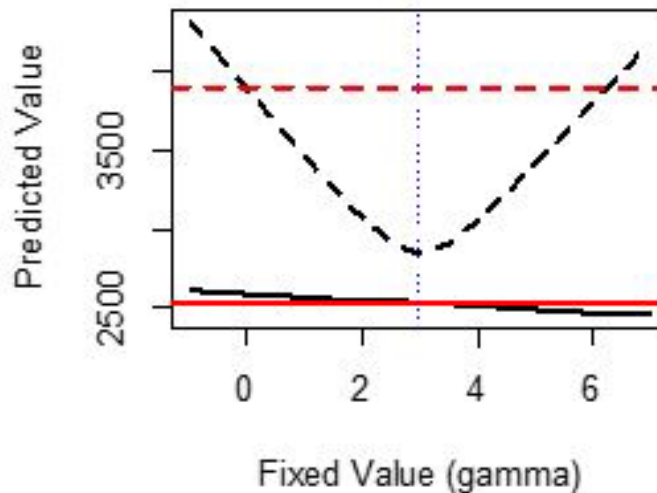


Figure 40: Fixed Coefficient Example Prediction Interval Comparison

Both results produce nearly identical point estimates. However, the point estimate with the fixed parameter is biased and therefore may be consistently inaccurate when predicting out of sample data (i.e., data not used to fit the original model). In addition, the margin of error is narrower when β_3 is fixed between roughly 0 and 6 and wider otherwise. This demonstration is just one example dataset generated with a specific error distribution, and other problems may behave drastically different. However, the takeaway is fixing a coefficient is not a preferred methodology and is highly dependent on the quality of the prior information.

3.4.4 Inexact Prior Information on Parameter Values

The exact value of a regression parameter is unknown. In the prior scenarios, parameters are restricted in the least squares objective function, forcing them to fall within a predetermined range. The parameters still maintain the property of being fixed, unknown values, and therefore do not result in any changes to the model's degrees of freedom. Inexact prior information on parameter values restricts the feasible region for regression coefficient values. When this information is good, precision of the estimators may be increased, but otherwise may force the model into drawing conclusions not supported by the data.

Restrictions are applied by using a methodology called Inequality Constrained Least Squares (ICLS). ICLS is used more as a modifier to an existing methodology (Section [3.3](#)) rather than a separate or unique methodology. Restrictions are placed on a range of values, which the parameters can assume. An example of this is limiting a cost improvement curve to values less than 100%. Parameters are then calculated and the restricted model is assessed for loss of fit compared to the unrestricted model. If the inexact prior information is deemed acceptable, the CER can then be diagnosed and assessed ([Step 4: Validate CER](#)) as normal.

Constraining a parameter value can be a difficult technique to implement. Even in the case of OLS, a closed form solution does not exist for the parameters. As a result, algorithms are used to minimize the error term under the set of constraints. The consequences and implications of such restrictions can be difficult to assess and understand. For the purposes of this guide, many of the concerns and discussions covered in Section [3.3.5 Non-linear Least Squares \(NLS\)](#) are relevant to ICLS.

Accepting the CER requires additional analysis. [Step 4: Validate CER](#) contains a much more in-depth model validation discussion.

4.0 STEP 4: VALIDATE CER

After developing the regression equation, the next step is to evaluate the CER with a critical eye focused on the model's strengths, weaknesses, and limitations. A cost estimator must provide an answer, even when data are limited or analogies tenuous. CER validation is examining the model's inputs, statistical outputs, and all other relevant metrics to support a comprehensive assessment. The "best" statistical and most compelling CER possible is desired, but decision makers should be informed of the strengths and weaknesses of the estimates, and how to best improve them going forward.

Examination of visual and numerical analyses helps assess and validate key model properties. The following are key steps before accepting a CER for use in a cost estimate:

- Understand the data relationships and coefficient estimates
- Determine consistency with engineering and physical principles
- Assess and validate the statistical model assumptions
- Identify and review high influence points such as leverage points and potential outliers
- Assess the impacts of multicollinearity
- Determine the significance of the model and independent variables
- Quantify metrics of best fit and prediction strength
- Compare and contrast multiple competing CERs to identify the "best" model

If validation of the CER is successful, proceed to [Step 5: Characterize Uncertainty](#). Even then, do so with multiple CERs for a single dependent variable, or acknowledge that the proposed solution is only one of many possible solutions. If the validation process identifies model deficiencies, circle back to [Step 3: Generate CER](#). In many cases, multiple iterations back to the previous step will be required before arriving at the "best" CER.

Sometimes the "best" CER for a given application may not be the best performing CER from a statistical perspective. As more data are collected, reassess the analysis and return to [Step 2: Analyze Normalized Data, Measure Correlation, and Hypothesize Functional Form](#). After understanding the new data, proceed to [Step 3: Generate CER](#) and generate a new CER. Cost analysis is an iterative process.

Figure 41 treats the many interrelated challenges of developing a CER in serial form. There can be issues with the underlying data set (e.g., potential outliers); choice of regression methodology (e.g., OLS vs. GLS); choice of functional form (e.g., linear vs. non-linear); and choice of variable set (e.g., omitting an important cost driver). The first step in validating a CER is to graph the CER with the associated data. [Figure 41](#) graphically outlines the process of diagnosing a model. Due to the complexity of the step, [4.2 Model Assumptions](#) has a separate, more detailed flowchart, found in [Figure 45](#).

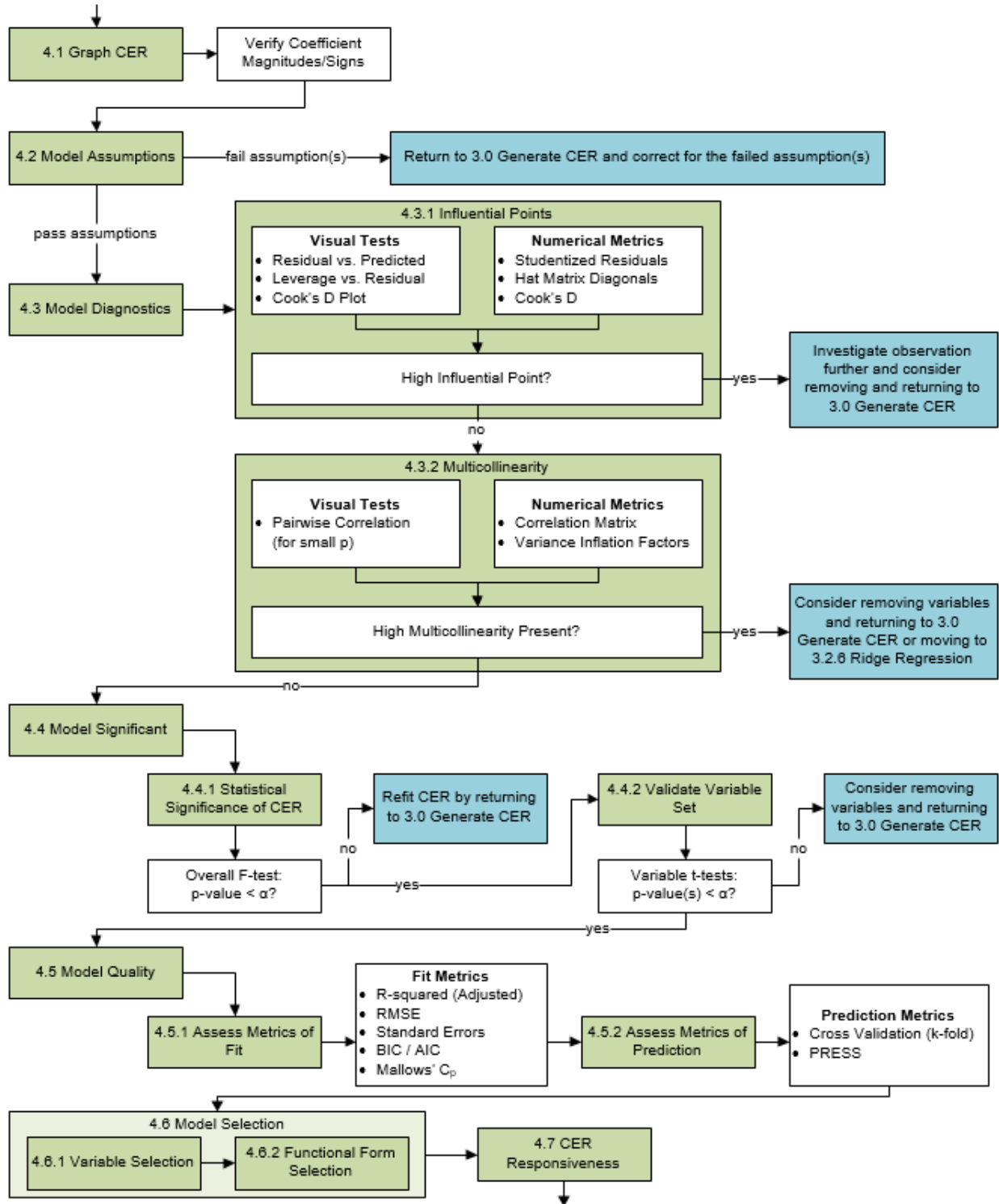


Figure 41: Step 4: Validate CER

4.1 Graph CER

Section [2.6 Scatter Plot of the Most Promising Cost Drivers](#) demonstrates the utility of a scatter plot used to identify relationships among data, particularly between cost and other independent variables. In conventional plots, the dependent variable is illustrated on the y -axis and the independent variable(s) are illustrated on the x -axis. To illustrate the importance of visualization, consider **Figure 42**. This famous dataset, known as Anscombe's Quartet⁵⁸, contains four distinct sets of points: (x_1, y_1) , (x_2, y_2) , (x_3, y_3) , and (x_4, y_4) . All four data sets share numerous statistical properties. In fact, each dataset has the same sample mean and variance for both x and y , the same correlation, or R^2 , and the same linear regression coefficients. By looking solely at these summary statistics, the datasets appear to be identical. However, the scatter plots clearly show the drastic differences in the trends of the data.

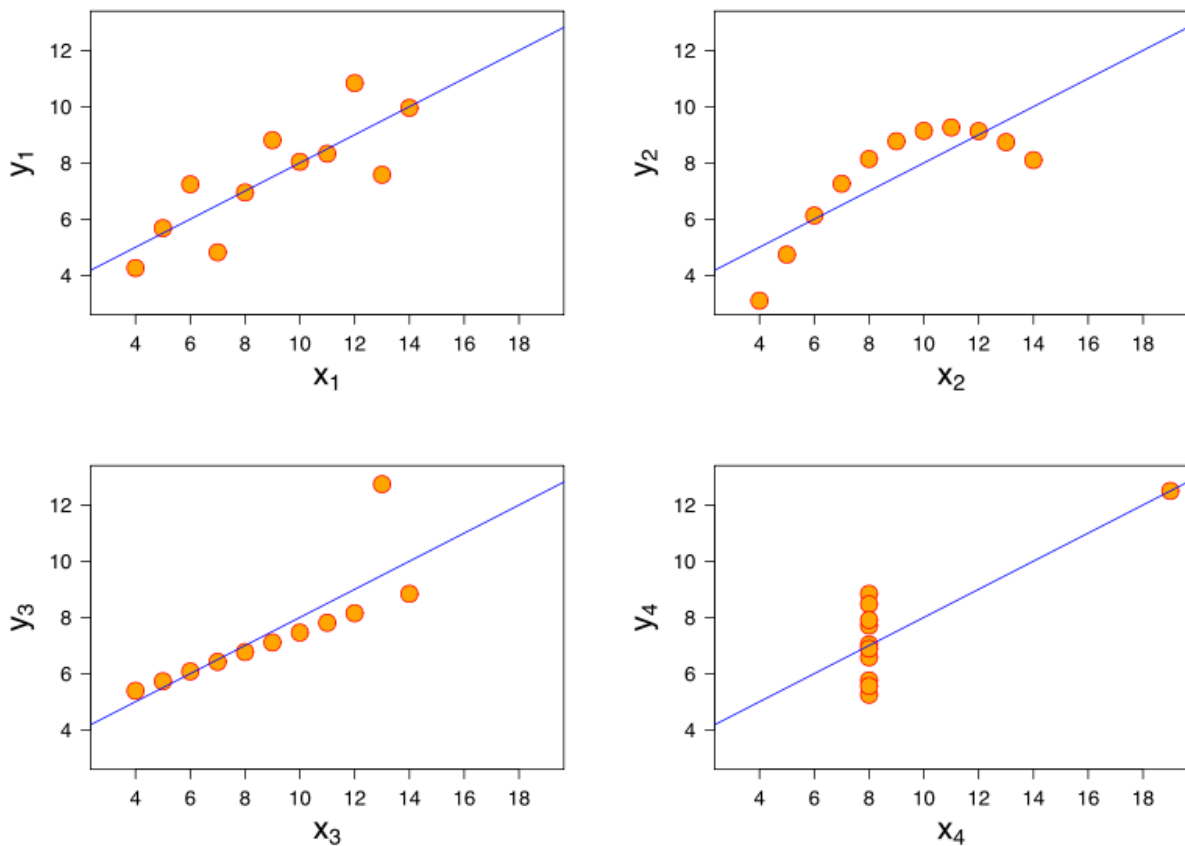


Figure 42: Anscombe's Quartet

At this stage, the graphical analysis shifts from hypothesizing a relationship to validating that relationship and understanding the observations in the context of the developed CER. Are there potential outliers

⁵⁸ Anscombe, Francis J. (1973) Graphs in statistical analysis. *American Statistician*, 27, 17–21.

(Section [4.3.1 Influential Points](#))? Is the data distributed in a reasonable way or is there a disproportionate weighting of the observations to one range of the CER versus another?

Visual analysis is just the first step in validating the CER. Utilize statistical metrics to evaluate the influence of each observation, validate the choice of regression methodology and functional form, and evaluate the statistical merits of the equation.

4.1.1 Visualizing the Simple Regression (Single Predictor) CER

The most common way to visualize a CER with a single predictor is to create a scatter plot (see [2.6 Scatter Plot of the Most Promising Cost Drivers](#)) as demonstrated using the OLS example from Section [3.3.1.3](#) in **Figure 43**. The line shown on the graph represents the CER regression equation:

$$Cost = 92.93 + 27.39 * Power$$

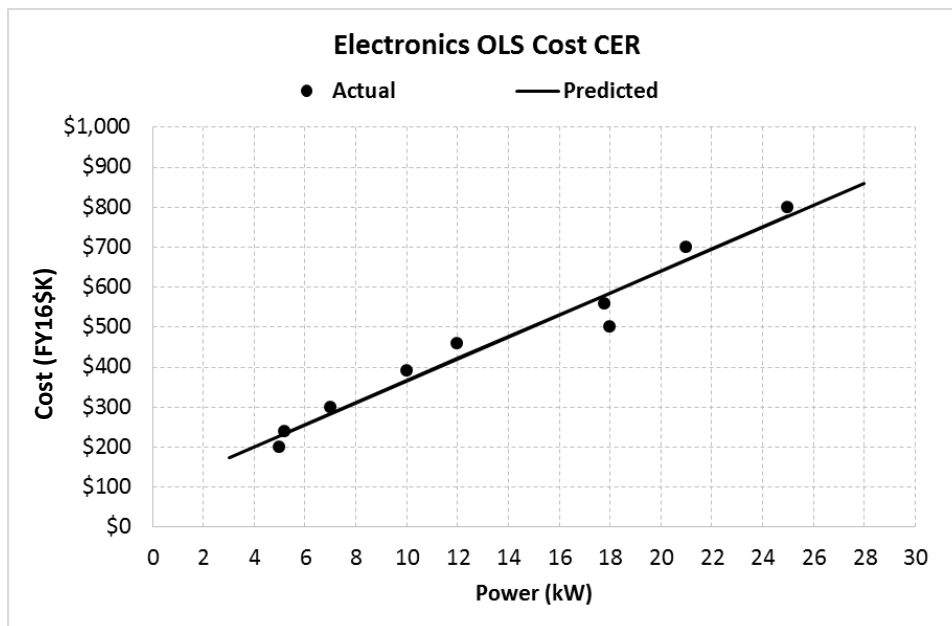


Figure 43: Graphical View of Simple CER

4.1.2 Visualizing the Multiple Regression (Single Predictor) CER

If there are two or more independent variables in the CER, the relationship can be represented by the following general equation form, where β_0 is the intercept, each β_i is the coefficient associated with the independent variable x_i ($i = 1, \dots, k$), and ϵ is the error associated with the CER.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

If there are two independent variables, the data has a 3-dimensional representation. For a linear functional form, the CER is a plane that slices through the heart of the data set, as shown in **Figure 44**. For non-linear functional forms, the CER is a curved surface.

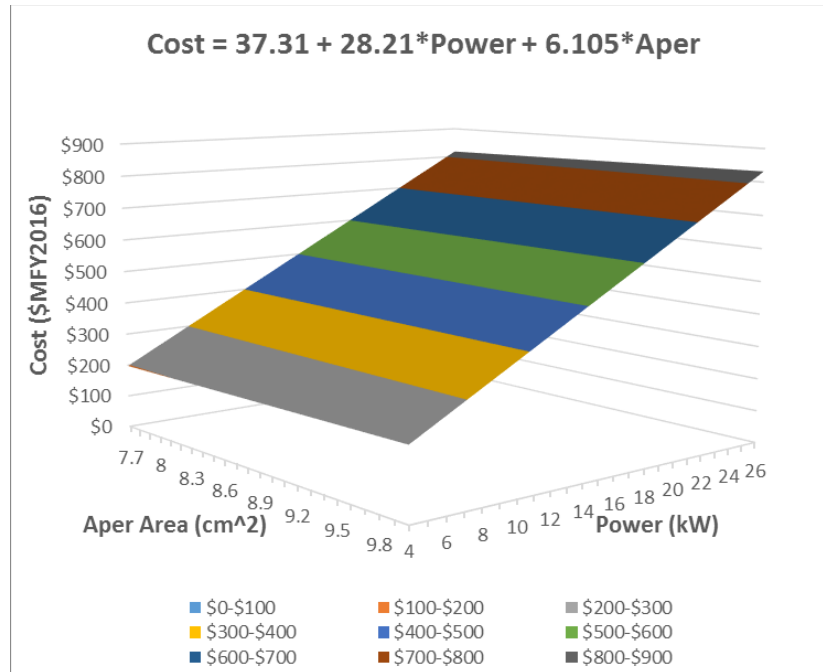


Figure 44: 3-D Visualization of Data

However as the number of independent variables grows beyond two, the model is working in higher dimensions that are not easily evaluated visually. A substitute graphic illustrating the relationship between the actual observations, versus the predicted observations may provide visual insight to a multiple variable relationship. See Section [4.5.1.8 Predicted versus Actuals Plot](#).

4.2 Model Assumptions

All data analytics methodologies come with a set of underlying assumptions. Regression analysis, and CER development, is no different. The assumptions are required for the mathematics to be able to guarantee a property, such as minimum variance or unbiasedness of the model. Distributional assumptions, such as normality, are crucial to be able to make inferences and characterize uncertainty around developed estimates.

A common regression assumption is that the errors are independently and identically distributed; that is, uncorrelated with each other and all coming from the same distribution with the same variance. Assumed is some type of functional form, be it linear or a specific non-linear model, in order to fit a regression equation. Finally, a distributional assumption, typically normality, enables a framework for statistical inference and quantifications of uncertainty. [Step 3: Generate CER](#) introduced various regression techniques, adjusting the four standard OLS assumptions noted in Section [3.1.1 Using OLS as a Regression Method Baseline](#).

There are three additional assumptions to the four standard OLS assumptions when multiple independent variables exist. Validation for these three assumptions is conceptual and less conducive to direct assessment than the other four.

- (1) The models discussed all assume that all *independent variables are non-stochastic*. This assumption states that the parameters are fixed and to be estimated, as opposed to being random

and to be predicted. In other words, the random error associated with the model all comes from variation in the response variable, y , and not from an independent variable, x . Violation or deviation from this assumption is an advanced topic beyond the scope of this guide. One way to deal with it is to utilize a Mixed or Random Effects model, briefly introduced in Appendix [A.4.9.3](#). This assumption does not suggest that uncertain inputs for predictions are inappropriate, and in general should not be of concern.

- (2) *No perfect multicollinearity*, or linear combinations, exists between the independent variables. This occurs when two (or more) predictors are exact multiples or combinations of each other. A simple example is length being given in both meters and feet. In a slightly less obvious case, suppose that a ship's crew drives the Purchased Services. Desired are the effects of both Officer and Enlisted personnel. Because Officers + Enlisted = Complement, all three cannot be used in the regression equation. This guide treats this issue as a data normalization step, covered in Section [2.5.2 Identify Redundant Variables and Potential Multicollinearity](#).
- (3) *No correlation exists* between the error term of the CER and the error term of another CER in the greater system estimate. Under the assumptions of the classical normal linear regression model, the least squares estimators of the regression coefficients are unbiased and with minimum variance. These properties flow directly from the premise that the specification of the model represents all there is to know about the regression equation. However, the integrity of these properties might be compromised if other pieces of information are in fact available. An example would be knowledge that the error term in the regression equation under consideration could be correlated with the error term in some other regression equation. By estimating each equation separately and independently, important information about the mutual correlation is discarded. While the resulting OLS estimates of the parameters remain valid, they are no longer with minimum variance (i.e., efficient). For efficient estimation, the technique of "Seemingly Unrelated Regression" is required. The technique, a form of Generalized Least Squares, is beyond the scope of this guide. This issue is not to be a concern, but rather a concept to be aware of.

Regardless of the regression methodology used to fit the CER, it is critical that the OLS assumptions are well understood and appreciated. The OLS model is often the default for the first look at the data and other methodologies are remedies for assumption violations with OLS. Even when not hypothesized to be appropriate, recalling Section [2.8 Hypothesize Functional Form](#), OLS is conducted in conjunction with the hypothesized model functional form. As a result, treatment of the OLS is at a higher level of rigor than for the other regression methodologies. The remaining methods often use the same techniques and principles as OLS, simply performed on transformed or standardized data, or with analogous metrics.

Visual analysis of specialized plots is the primary methodology for assessing the assumptions. Formal tests do exist, but are generally secondary to their visual counterparts in this context. Some tests are more popular than others, and their use depends heavily on the field of study. Even within DoD, different organizations may have different guidelines and requirements for the use of a specific formal test.

The following sections discuss the explicit assumptions by each of the methods introduced in [Step 3: Generate CER](#), how to assess them both visually and analytically, and how to remedy potential violations. While presented in a numerical order, it is important to consider and assess all four assumptions before proceeding to any next step, be it accepting or rejecting an assumption. The assumptions are all interrelated, and an apparent violation of one may be caused by a more severe violation of another.

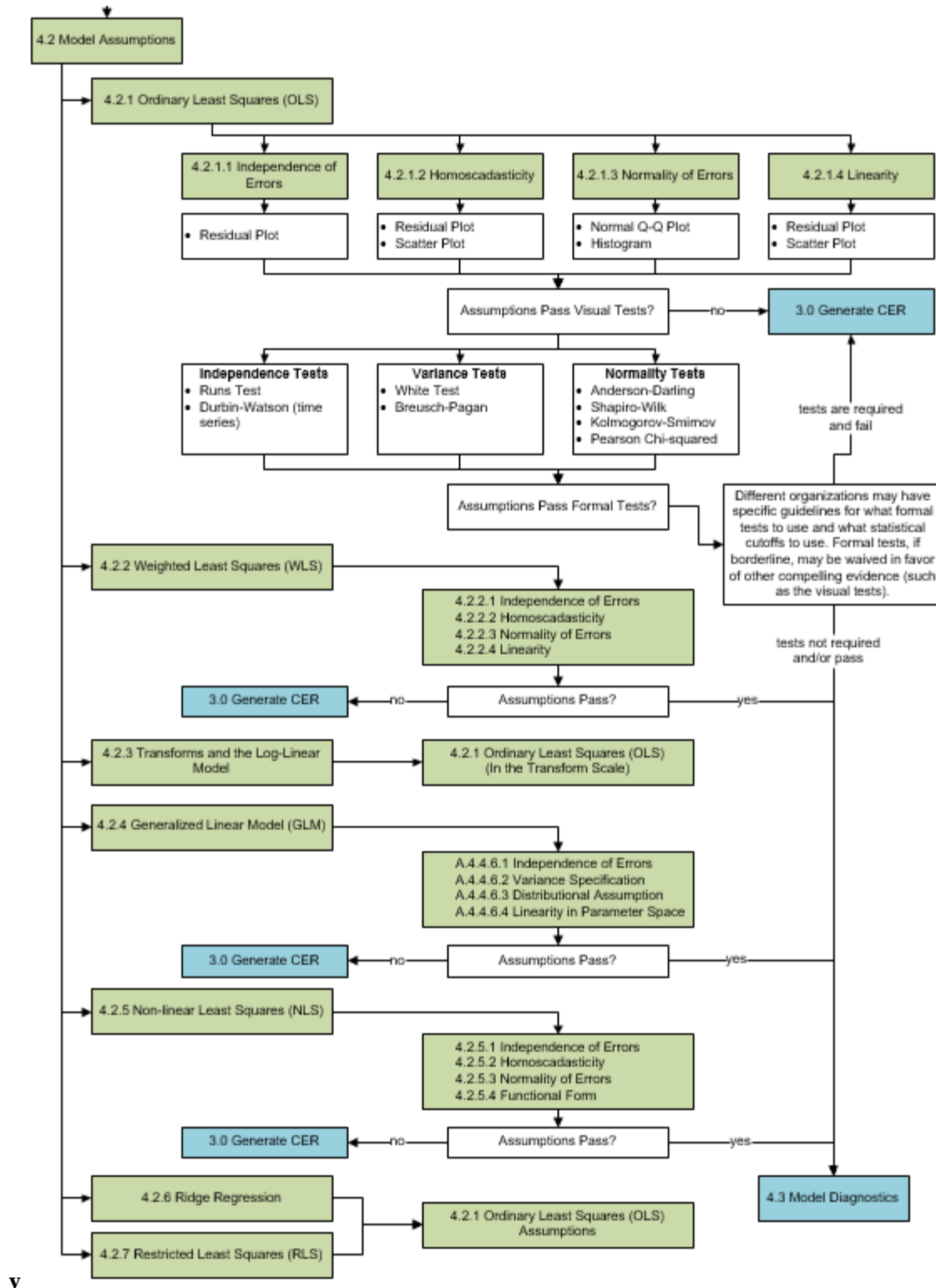


Figure 45: Step 4.2 Model Assumptions

4.2.1 Ordinary Least Squares (OLS)

Section 3.3 introduced the Ordinary Least Squares Model (OLS) in two forms: [Simple Linear Regression \(SLR\)](#) and [Multiple Linear Regression \(MLR\)](#). Both models have the same set of assumptions and use the same validation methodologies. As a result, the remainder of this section will treat OLS in the MLR case. The difference between the evaluations of the two methods is generation of a [scatter plot](#) is possible for SLR, while MLR must rely on other methods to visualize the data.

The model statement explicitly captures the assumptions of the analysis. Recalling Section 3.3.1.3, the OLS model is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

4.2.1.1 Residuals

– Terminology –

The term residual can be ambiguous. This section discusses three common and uniquely defined residuals: raw residual, internally studentized residual, and externally studentized residual.

Another common term is the “standardized residual.” The most commonly used definitions are the internally studentized residual and externally studentized residual. Some software such as MS Excel uses neither.

For the purposes of this handbook, the “standardized residual” is the internally studentized residual.

The residual error is the difference between the actual value and the predicted value. This is the raw residual and for OLS is,

$$\begin{aligned} \mathbf{e}_{raw} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

This is the simplest residual and is useful to examine how far each predicted point is from the actual value. However, this is not the correct residual to use in the sense of a residual analysis. While the errors, $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ are assumed to have constant variance, the residuals, \mathbf{e}_{raw} , do not. As a result, the residuals need to be standardized. Dividing each raw residual by the corresponding variance results in the internally studentized residual,

$$e_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Where,

$$\begin{aligned} \hat{\sigma} &= \sqrt{MSE} \\ h_{ii} &= i^{th} \text{ diagonal entry of the hat matrix, } \mathbf{H} \\ \mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{aligned}$$

The internally studentized residual follows an approximate standard t-distribution, with mean of zero and variance of one, and is acceptable to use for residual analyses. A relatively minor modification results in

the residual following a t-distribution, instead of an approximate t-distribution. This is done by calculating the standard error, $\hat{\sigma}$, for each residual based on the data set without that observation. This calculation of the residual of interest uses an independently derived error. This is the externally studentized (or deleted) residual,

$$e_{i,-1} = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{i,-1} \sqrt{1 - h_{ii}}}$$

Where,

$$\begin{aligned} \hat{\sigma}_{i,-1} &= \sqrt{MSE} \text{ (calculated without data point } i) \\ h_{ii} &= i^{\text{th}} \text{ diagonal entry of the hat matrix, } \mathbf{H} \\ \mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{aligned}$$

Standardizing the residual is not as simple as dividing by the estimate of the error, $\hat{\sigma}$. The term standardized is ambiguous and not every statistical package makes it obvious which methodology is being used. The internally studentized residual is acceptable to use and is returned by most software packages, including CO\$TAT. MS Excel returns a standardized residual but is defined incorrectly.⁵⁹ Many dedicated statistical packages provide the option to return the externally studentized residual as well, including SAS and R, referred to as ‘rstudent’, and Minitab, referred to as the ‘deleted residual’.

An analysis of the residuals is not valid unless done using either the internally or externally studentized residual. When available, is the analyst is recommended to use the externally studentized residual.

4.2.1.2 Independence of Errors

The statement of $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ captures the assumption that each error is distributed independently. Observing a value for some y_j has no impact on the expected value of some other response, y_k . A common violation of this assumption is the correlation of sequential responses in a time series application. Again, this is beyond of the scope of the material in this guide.

Recalling Section [3.3.1 Ordinary Least Squares \(OLS\)](#) and the Gauss-Markov theorem, which explain when the independence of errors assumption fails, the OLS model no longer guarantees the Best Linear Unbiased Estimate of the coefficients. As a result, the variances around the coefficients are inflated. However, the estimator is still unbiased.

4.2.1.2.1 Standardized Residuals Visual Tests

– Terminology –

When using statistical software be careful to understand is being portrayed.

⁵⁹ Excel has changed formulas for the standardized residual from Office 95 to 97 and may do so again in future. This information is current through Office 2013.

A residual plot is the most common way to assess the independence of the errors. A residual plot (i.e., a residuals versus predicted values plot) is a [scatter plot](#) with the standardized residuals on the y-axis, and the predicted value of the response, \hat{y} , on the x-axis. An ideal result displays random scatter around zero, with no apparent pattern. This type of plot is a standard output from most statistical tools and is easy to construct given the standardized residuals. Due to the standardization of the residuals, all residual plots, regardless of the data, appear on roughly the same scale.

Figure 46 displays two example residual plots. In Plot A, the independence assumption is not of a concern since the figure displays random scatter. Plot B suggests a pattern, which could indicate lagging correlation between residuals, common in a time series application. This pattern may also arise in CIC analysis with production breaks.

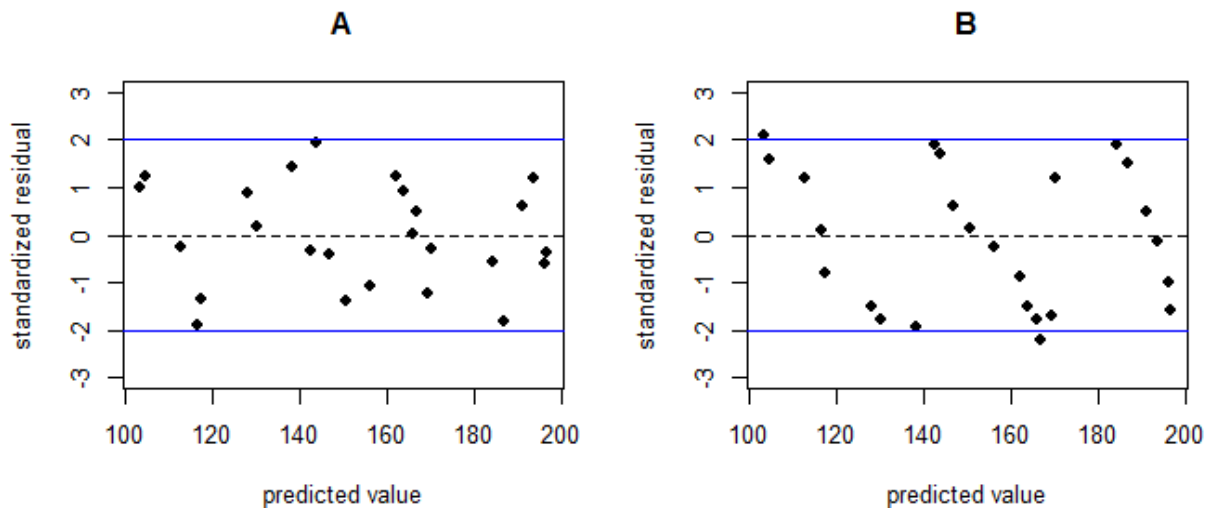


Figure 46: Independence of Errors Residual Plots

If distinct patterns are present, consider potential sources of the problem. Outside of the time series case, an apparent problem with the residual versus predicted values plot often stems from a violation of the linearity assumption, covered in more detail in [Section 4.2.1.5 Linearity](#).

Alternatively, if an independent variable has some type of sequential interpretation (e.g., time or unit number), plot this value as an independent variable on the x -axis to assess potential correlation which may have an impact on the residuals. To reiterate, this type of application is where independence of errors is most likely to be problematic, and is referred to as autocorrelation or serial correlation, which is beyond the scope of this guide.

4.2.1.2.2 Formal Tests for Independence

As previously noted, the visual tests are the main assessment of assumptions. A rejection based on an assumption by a formal test would supersede a conclusion drawn from the graphical analysis is unlikely. However, the tests can still be useful and their use is more prevalent in some fields than in others. Additionally, different organizations within DoD may have unique requirements and/or guidelines on which tests to use and at what significance level. In the case of independence of errors, there are two common tests, which are popular to test autocorrelation and constant variance. These tests are the Durbin-Watson test and the Runs test.

The Durbin-Watson (DW) test is the most widely used test for autocorrelation of the residuals, which test the correlation between a given residual and the one preceding. A p-value for the test less than the pre-specified α results in a rejection of the independence of errors assumption.

In some cases, software applications return only the test statistic, D , and not a p-value. In these cases, use a table of critical values⁶⁰ to determine the outcome of the test. The DW table requires both the sample size, n , and the number of independent variables, k . The tables then returns two critical values: an upper bound (d_U) and a lower bound (d_L) which are used as follows:

- If $D > d_U$ then fail to reject the null hypothesis and accept the independence assumption
- If $D < d_L$ then reject the null hypotheses and reject the independence assumption
- If $d_L < D < d_U$ then the test is inconclusive

The Runs test is a nonparametric test to determine if the sequence of data are random, which test whether the positive and negative elements appear at random. The test looks at the sign (+ or -) of each residual and attempts to detect if there is a pattern to their occurrences. A p-value for the test less than the pre-specified α results in a rejection of the independence of errors assumption.

4.2.1.3 Homoscedasticity

The mathematical phrase $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ captures the assumptions of identically distributed error terms with constant variance. A violation of this assumption is fairly common to cost analysis: as the response, \mathbf{y} , gets larger, so does the error.

Recalling Section [3.3.1 Ordinary Least Squares \(OLS\)](#) and the Gauss-Markov theorem, which explain, when the homoscedasticity assumption fails, the OLS model no longer guarantees the Best Linear Unbiased Estimate of the coefficients. As a result, the variances around the coefficients are inflated. However, the estimator is still unbiased.

4.2.1.3.1 Homoscedasticity Visual Tests

The most common way to assess homoscedasticity is to generate scatter plots and a residual versus predicted plot. For SLR, examine a scatter plot for the single predictor for any type of non-constant scatter. **Figure 47** shows two scatter plot examples. Plot A indicates random scatter about the regression line, indicating homoscedasticity. Plot B shows a larger variance for larger values of \mathbf{y} , which suggests multiplicative error (i.e., heteroscedasticity). Similarly, in the MLR setting, create a scatter plot for each independent variable. Look for the same type of observations at the individual level as with the SLR plot. While informative, assessing homoscedasticity based solely on scatter plots is insufficient.

⁶⁰ https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf

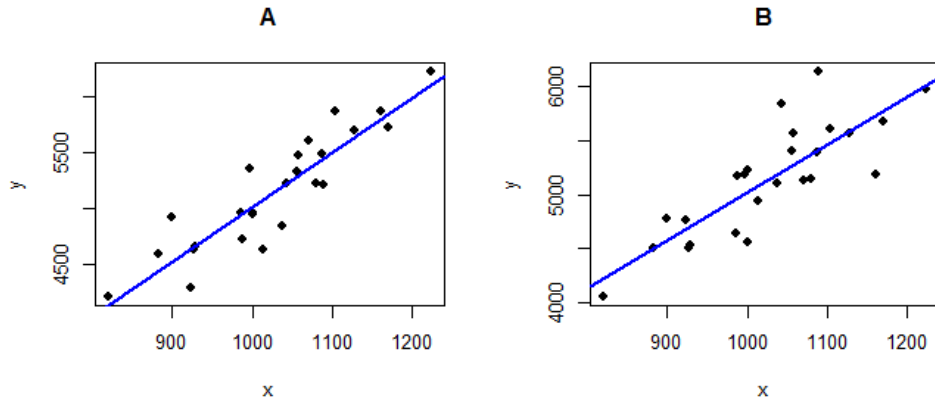


Figure 47: Scatter Plots For Assessing the Homoscedasticity Assumption

While the scatter plot(s) provide a good starting point, a residual versus predicted value plot, as defined previously in Section [4.2.1.2](#), provides a better understanding of the behavior of the errors. An ideal result displays random scatter around zero, with no apparent pattern.

Figure 48 shows four example residual plots. Case A shows random scatter, suggesting homoscedasticity. In Case B multiplicative error appears present, indicated by the cone-shaped nature of the residuals (i.e., heteroscedasticity). Cases C and D also indicate heteroscedasticity and a violation of the non-constant variance assumption. The residuals follow a curved pattern. As noted with the independence assumption, apparent problems with the residual versus predicted values plots often stem from a violation of the linearity assumption. These two plots suggest this possibility, covered in more detail in Section [4.2.1.5 Linearity](#).

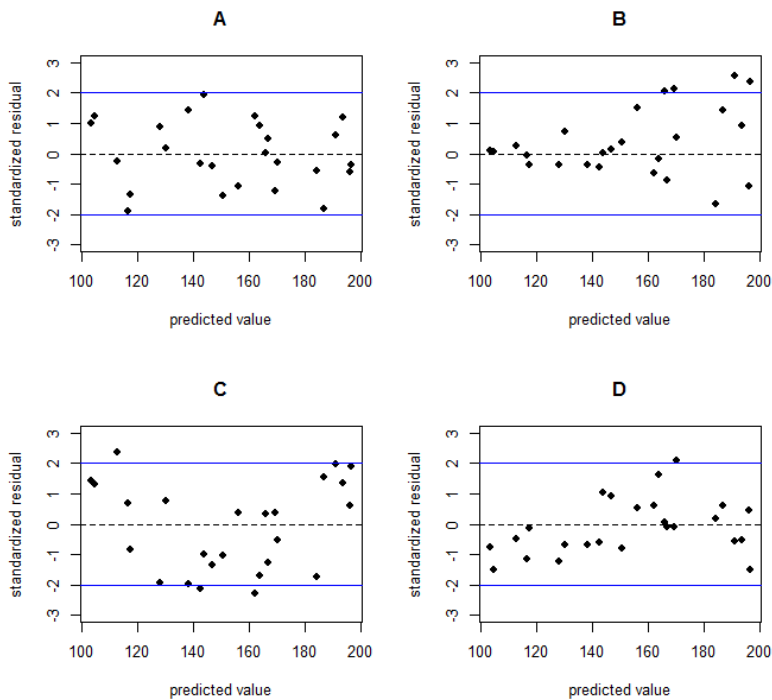


Figure 48: Residual Plots For Assessing the Homoscedasticity Assumption

4.2.1.3.2 Formal Tests

Visual tests are the main assessment of assumptions. Formal tests are still useful for validating the visual assessments. In the case of homoscedasticity, two common tests are the White test and Breusch-Pagan.

The White test is a very popular test in economics for heteroscedasticity of the errors. The test takes the approach of conducting a regression on the squared residuals of the model, based on the original predictors and the comprehensive set of second order combinations (i.e., the squared predictors and all of their combinations). The R^2 of the resulting regression is part of the resulting test statistic, testing the null hypothesis that the errors are constant. Thus, a p-value for the test less than the pre-specified α results in a rejection of the assumption of homoscedasticity of errors. The R statistics package has the White test.

The Breusch-Pagan (BP) test is an alternative to the White test, again taking a strategy of employing a regression on the residuals. This test relies on the normality assumption by use of the F-test. If the linear regression line drawn through the residuals has statistically significant parameters, then the BP test rejects the assumption of heteroscedasticity of errors. The BP test statistic produces a p-value from the Chi-squared distribution, where a p-value less than the pre-specified α results in a rejection of the assumption of homoscedasticity of errors. The R statistics package has the BP test.

4.2.1.4 Normality of Errors

The statement of $\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ captures the assumption that each error is distributed according to the normal distribution. With real world data, the normality assumption is often optimistic. Normally distributed error values are uncommon and an assessment of normality can be difficult with relatively small sample sizes, as is common with CER construction. The diagnoses and assessment of the normality assumption is very important.

Recalling Section [3.3.1 Ordinary Least Squares \(OLS\)](#) and the Gauss-Markov theorem, when the normality of errors assumption fails, the OLS model still provides the Best Linear Unbiased Estimate of the coefficients, as long as the other three assumptions hold. Normality is simply a construct for inference, albeit very useful and often essential. As a result, failure of normality still results in an unbiased estimator with minimum variance, but severely limits the ability to conduct inference including outlier detection, significance testing, and risk analysis.

4.2.1.4.1 Normality of Errors Visual Tests

A common way to assess normality when you have many observations is to generate a histogram of the residuals. A histogram may help determine if the distribution is symmetrical and bell-shaped. Another graphical option is to examine a Normal Quantile-Quantile (Q-Q)⁶¹ plot. For more information regarding histogram generation, see Section [2.4.3](#). Once constructed, compare the histogram to the shape of the normal distribution. **Figure 49** shows an example of a best normal distribution fit overlaid directly on top of the histogram.

⁶¹ An alternate display of the Quantile-Quantile is the Probability-Probability (P-P) plot. Construction of both is very similar and both convey the same information. CO\$TAT makes use of the P-P plot.

The selection of the bin size has a huge influence on the shape of the histogram. There is more than one accepted approach to estimating bins and as a result, histograms can end up being very subjective. The histogram provides a good starting place to understanding the behavior of the residuals, but is insufficient in developing a conclusive result.

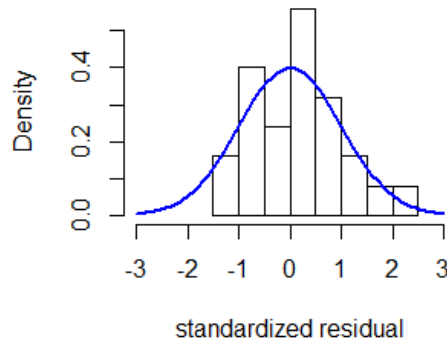


Figure 49: Histogram of Standard Residuals

The Q-Q plot is the preferred graphic to examine the behavior of data relative to any distribution of interest. A normal Q-Q plot is constructed by plotting the theoretical quantiles of the normal distributions on the x -axis and the observed standardized residual quantiles on the y -axis. Practically speaking, the plot can be thought of as a predicted versus actuals plot for the residuals. If the data are exactly normal, then the plotted standardized residuals fall in a perfectly straight line. Deviations from a line indicate deviations from the normal distribution.

Figure 50 shows four example Q-Q plots. Case A illustrates an example of normality. While not a perfect fit to a line, this is a strong case to accept the normality assumption. Cases B and C indicate some other type of behavior. Perhaps the tails are longer, or the distribution is skewed. Accepting normality in either of these two cases is unlikely and an alternate model should be considered. Case D is not ideal, but sufficient to accept normality. There seems to be behavior in the tails, which are non-normal, but nothing substantial. While an alternate model may be considered, plots such as Case D are often accepted.

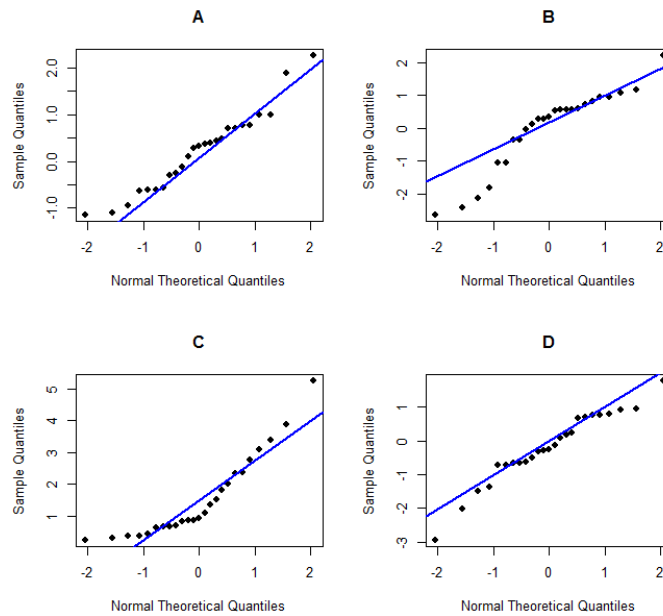


Figure 50: Normal Q-Q Plot Examples

Assessing a normal Q-Q plot takes practice. With only a few observations, a perfect linear fit is uncommon. Unless the deviation is severe with a distinct trend, the normality assumption is often accepted. Much like with the independence of the errors assumption, apparent problems with the Q-Q plot can also stem from a violation of the linearity assumption, covered in more detail in Section [4.2.1.5 Linearity](#).

4.2.1.4.2 Formal Tests for Normality

Visual tests are the main assessment of assumptions. The following paragraph describes several Formal tests that can be useful for validating the assumption of normality. The Anderson-Darling (AD) test, the Shapiro-Wilk (SW) test, the Kolmogorov-Smirnov (KS) test, and the Pearson Chi-squared (Chi-squared) test. .

The Kolmogorov-Smirnov (KS) tests for normality by examining the cumulative distribution function (CDF) of the normal distribution compared to the empirical distribution function (EDF) of the residuals. If the data are normally distributed, then the two curves lay on top of each other. The KS test statistic is the largest deviation between the two curves and tests the null hypothesis that the data does follow the distribution of interest (e.g., normal). P-values that are less than the pre-specified α results in a rejection of the assumption of normality of errors.

The Anderson-Darling (AD) test is a refinement of the KS test, placing more weights in the tails of the distribution. The AD statistic tests the same null hypothesis that the data follows the distribution of interest (i.e., normal). P-values less than the pre-specified α results in a rejection of the assumption of normality of errors.

The Shapiro-Wilk (SW) test is specifically for normality and tests the assumption that the data are normally distributed. The test statistic calculates weighted deviations of the sample data from the normal distribution and performs very well in comparison to the other normality tests. The SW statistic tests the

same null hypothesis that the data does follow the distribution of interest (i.e., normal). P-values less than the pre-specified α result in a rejection of the assumption of normality of errors.

The [Pearson Chi-squared](#) test compares discrete sections, or bins of equal probability, of data to their theoretical expectation. This test is best visualized as comparing the heights of each histogram bar to the superimposed normal density curve, even though the bin selection methodology differs from a histogram. As a result, the Chi-squared test is sensitive to the selected bin widths. There is no universal methodology for these selections, but a common approach is to use the Mann-Wald method, divided by 2. Appendix [A.3.1.1.2 Histogram](#) contains references with more information on this topic. The Chi-squared statistic tests the same null hypothesis that the data does follow the distribution of interest (i.e., normal). P-values less than the pre-specified α result in a rejection of the normality of errors assumption.

4.2.1.5 Linearity

The statement of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ captures the assumption of linearity. This assumption states that the response, \mathbf{y} , is a linear function of the predictors (or transformations of the predictors), \mathbf{X} , and the coefficient parameters, $\boldsymbol{\beta}$. When violated, the entire analysis is not considered to be best practice and statistical metrics, regardless of apparent quality, are questionable. In some cases, a linear approximation may still be appropriate or desired for simplicity, but often a non-linear form or data transformation is required to correct the problem.

Recalling Step [3.3.1 Ordinary Least Squares \(OLS\)](#) and the Gauss-Markov theorem, when the linearity assumption fails, the OLS model still provides the Best Linear Unbiased Estimate of the coefficients. However, the issue is now that the linear estimator, regardless of other properties, is questionable. As a result, the variances around the coefficients are still minimal and the estimator will still be unbiased relative to other linear estimators, but the CER is not an optimal representation of the system modeled.

4.2.1.5.1 Linearity Visual Tests

To assess linearity generate a scatter plot, a residual versus predicted plot, and a predicted versus actuals plot. For SLR, examine a scatter plot for the single predictor for a linear trend. Curvature in the scatter plot indicates a non-linear trend.

While the scatter plot provides a good starting point, a residual versus predicted value plot, as defined previously in Section [4.2.1.2](#), often provides a better understanding of the model. An ideal result displays random scatter around zero, with no apparent patterns. The exception is in the case of non-constant error where the residuals can indicate that the linear trend is still appropriate, but heteroscedastic. **Figure 51** displays two of the cases from [Figure 48](#) in Section [4.2.1.3](#).

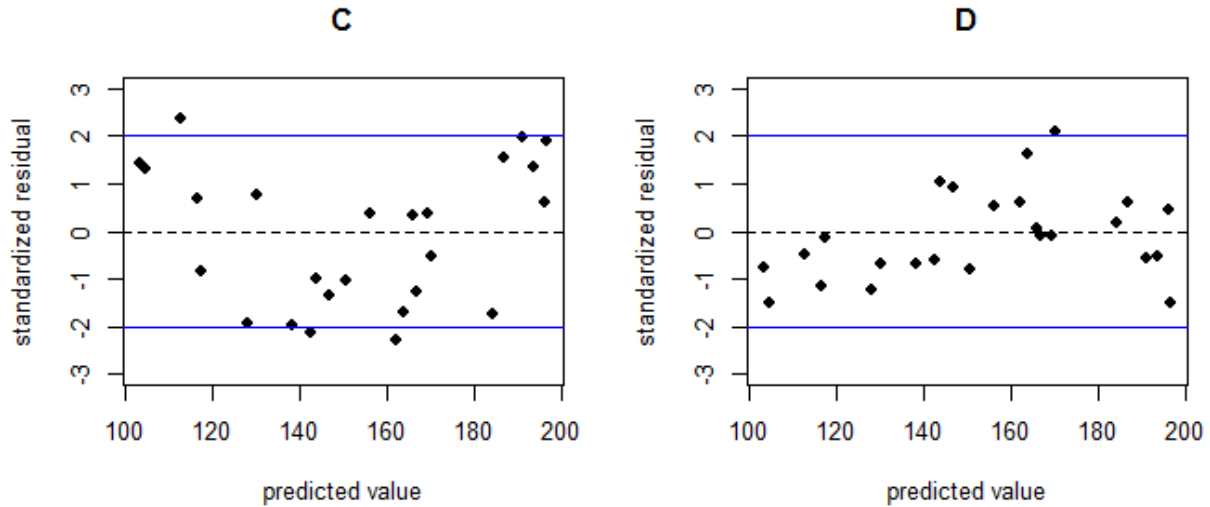


Figure 51: Non-linear Residual Plots

Plot C under-fits the model for lower values of y , over-fits in the middle of the predicted y values, and under-fits for larger values of y . Plot D indicates a similar behavior in the opposite direction. **Figure 52** further illustrates potential nonlinear impacts.

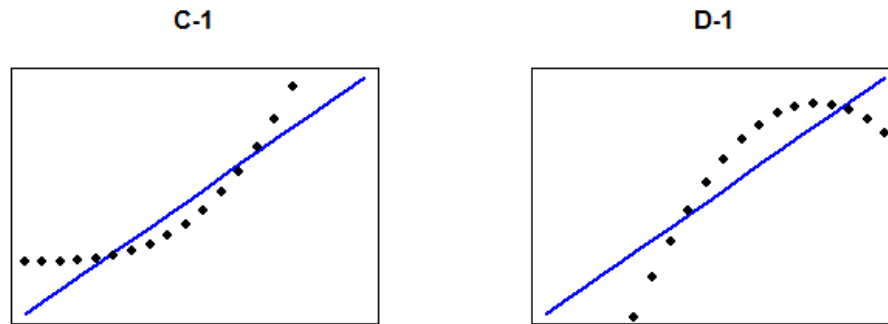


Figure 52: Non-linear Behavior of Residuals

The predicted versus actuals plot provides similar information as the residual plot. Again, the desired outcome is random scatter of the predicted points about the line $y = x$. If there is an apparent pattern such as a clustering, then the linear model may be inappropriate. **Figure 53** is an example of two predicted versus actuals plots. Plot A demonstrates random scatter supporting the acceptance of the linearity assumption. Plot B demonstrates a violation of linearity.

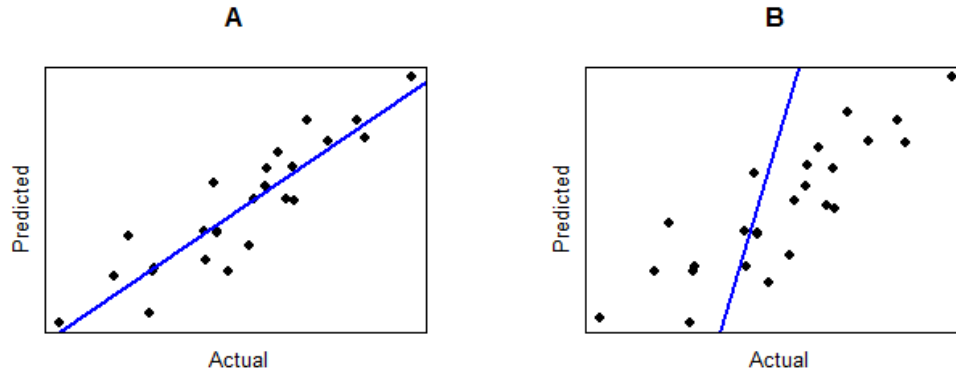


Figure 53: Linearity – Predicted versus Actuals Example

4.2.1.5.2 Formal Tests

Formal tests designed to assess linearity are uncommon. A visual analysis is considered best practice to assess the linearity assumption.

While this guide presents the assumptions sequentially, assess all four assumptions (Sections [4.2.1.2](#) through [4.2.1.5 Linearity](#)) before making a decision regarding model selection. If all four assumptions are acceptable, the next step is [4.3 Model Diagnostics](#).

4.2.1.6 Residuals Example

[Table 10](#): Notional Data to Demonstrate Functional Forms contained the originally normalized cost and power data plus adjusted cost values to demonstrate the power, exponential and logarithmic functional forms. **Figure 54** shows the residual plots from a linear fit performed on the nonlinear data developed in those data sets. Even with only nine observations, clear patterns are evident indicating that the residuals are not independent of the dependent variable.

- The “Linear on Power Data” in the upper left illustrates a non-constant variance. This is pattern is discussed in detail in [4.2.1.3 Homoscedasticity](#).
- The remaining charts show a concave down or concave up pattern. See [4.2.1.4 Normality of Errors](#) for a discussion on why these patterns are evidence that a linear fit is likely not the best fit.

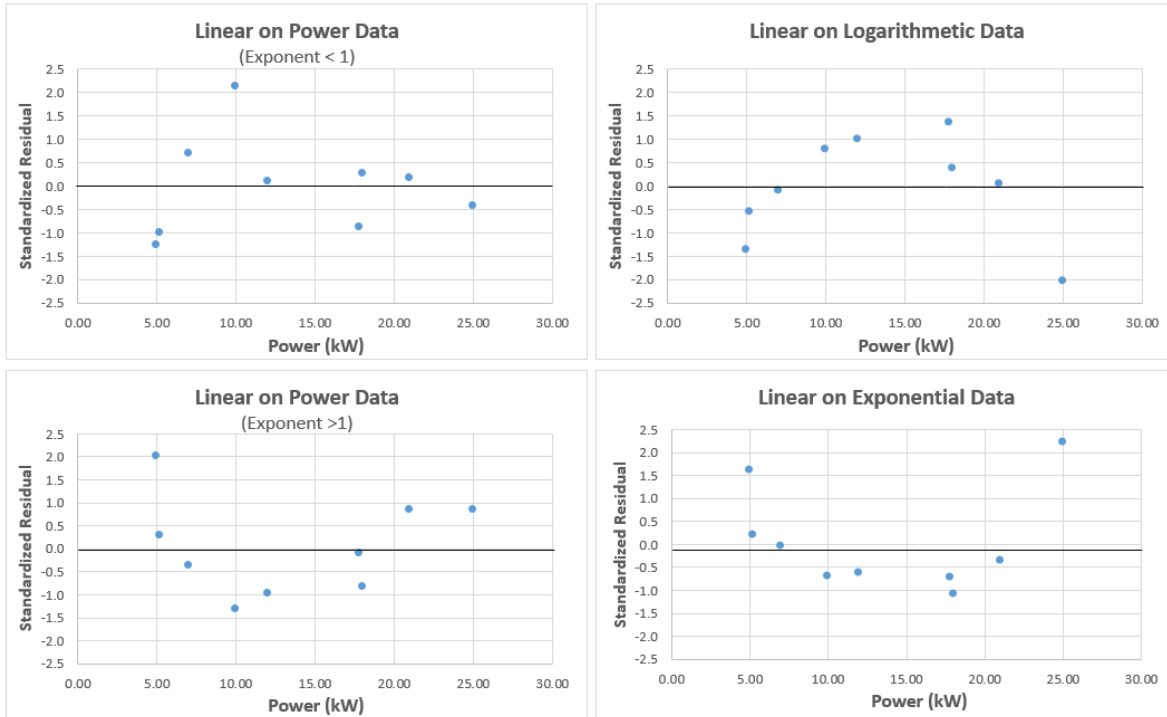


Figure 54: Residual Plots For Linear Fits on Nonlinear Data

Figure 55 is the residual plot for the linear fit demonstrated in [2.8.1 Linear Functional Form](#). The pattern suggests residuals may not be related to the dependent variable when using the linear form. An ideal result displays random scatter around zero, with no apparent pattern. **Figure 55** may not be enough evidence to reject the CER, but should be sufficient to motivate further exploration.

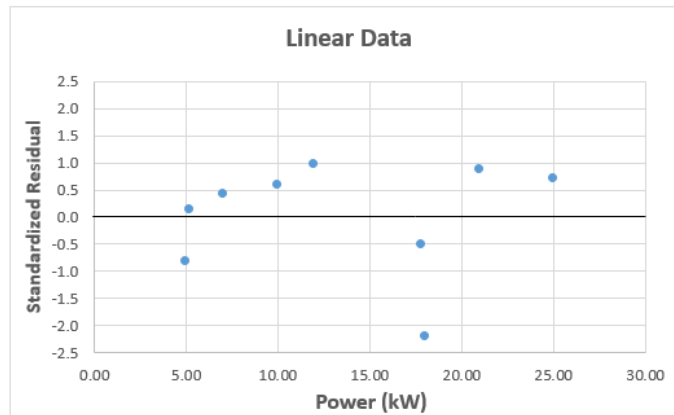


Figure 55: Residual Plot For the Example Linear Fit

Recall the OLS example introduced in Section [3.3.1.3](#) with the response variable, *Cost*, and predictors, *Power* and *Aperture*. When the example was run, COSTAT produced a Standardized Residual plot, displayed in **Figure 56**. The residuals appear to have random scatter about zero. There is no evidence of a pattern, indicating the independence of errors assumption may be accepted.

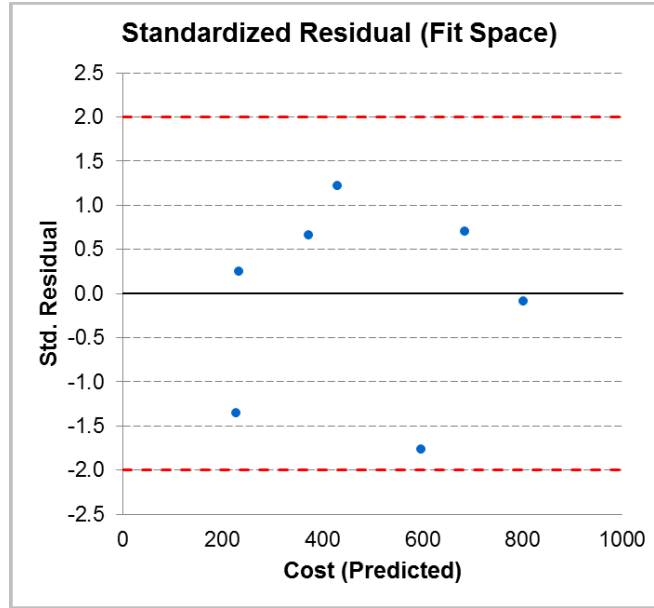


Figure 56: Standardized (Internally Studentized) Residual Plot

The following section includes examples of formal tests that can be performed to verify the visual observations:

- **Independence of Errors:**

- Section [3.3.6 Ridge Regression](#) demonstrated a method to address correlation of independent variables. **Figure 57** is the CO\$TAT Ridge Statistics output from that example and contains the Durbin-Watson test statistic, $D = 3.1057$.

IV. Ridge Perturbation Parameter (RPP) & Related Statistics

RPP by Non-iterative Procedure	0.0061
RPP by Iterative Procedure	0.0068
Von Neumann Test for Autocorrelation	3.6234
Durbin-Watson Statistic	3.1057
Determinate of (X'X)	0.1106
Measure of Ill conditioning	9.0400

Figure 57: Durbin-Watson Test Statistic

A table of critical values for $\alpha = 0.05$, $n = 7$, and $k = 2$ returns $d_U = 1.896$ and $d_L = 0.467$. Since $D = 3.1057 > d_U = 1.896$, the test fails to reject the independence of errors assumption, which aligns with the conclusion drawn from **Figure 56**.

- The remedy for a failed independence of errors assumption often involves employing a time series methodology. These methods account for correlation. Section [3.3.2 Generalized Least Squares \(GLS\)](#) introduces a framework, which can support specification of a correlation matrix, a topic covered in detail by many statistical and econometric resources.

- **Homoscedasticity:**

- The residuals in **Figure 56** appear to have random scatter about zero. There is no evidence of a pattern, and this example suggests no problem with the constant variance assumption. This is the same plot used to validate the independence of errors assumption. With practice, assessment of the independence of errors and constant variance assumptions can be done simultaneously.
- Many statistical packages provide both the White test and the Breusch-Pagan test. To illustrate the interpretation of one of these tests, consider the result for the BP test, returned by R:

```
studentized Breusch-Pagan test

data:  lm(Cost ~ Power + Aper)
BP = 2.6952, df = 2, p-value = 0.2599
```

- In this example, $p\text{-value} = 0.2599 > 0.05 = \alpha$. Indicating the BP test fails to reject the null hypothesis that the errors are homoscedastic. This result agrees with the visual analysis from **Figure 56**.
- Weighted Least Squares (WLS) regression is a common approach to remedy a failure of the homoscedasticity assumption. WLS weights the residuals of each data point separately when minimizing the least squares, adjusting them to be homoscedastic in this scaled space. Section [3.3.2.2 Weighted Least Squares \(WLS\)](#) covers this topic, providing four common methodologies for the weighted estimation. Method 1 is a general approach that can model many different types of weights. Methods 2-4 are particularly useful in multiplicative error cases, such as Case B in [Figure 48](#).

- **Normality:**

- A visual test for normality is illustrated in **Figure 58**. On the left are the standardized residuals from a [linear fit](#) of the electronics data. While the statistics of the fit are very good, the histogram suggests the normality assumption is not valid. However, the [log linear functional form](#) with independent variables *Intensity* ($kWperCm^2$) and FFP are preferred. . Note the standardized residuals are assessed in fit space.

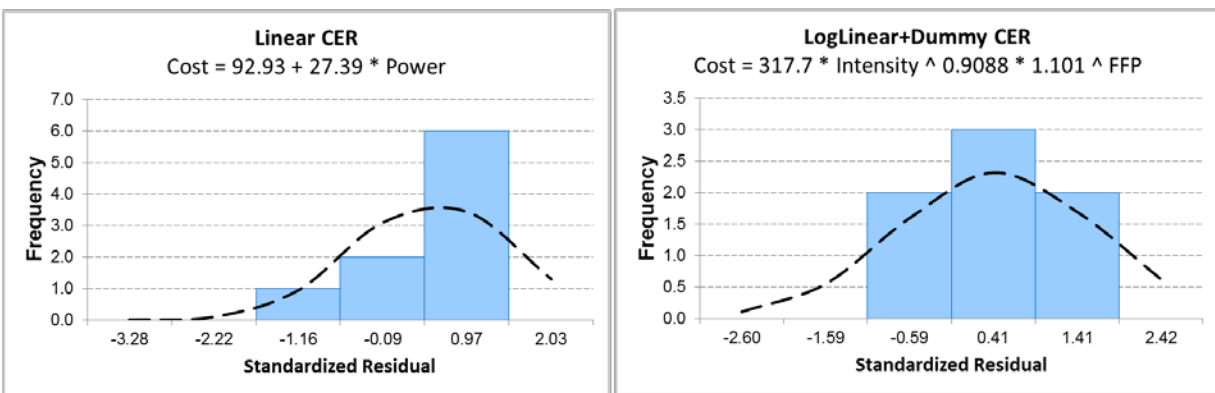


Figure 58: Histogram of Standardized Residuals

- As part of the output, CO\$TAT produced a P-P plot, displayed in **Figure 59**. CO\$TAT uses a P-P plot rather than a Q-Q plot, but the interpretation and information is the same.

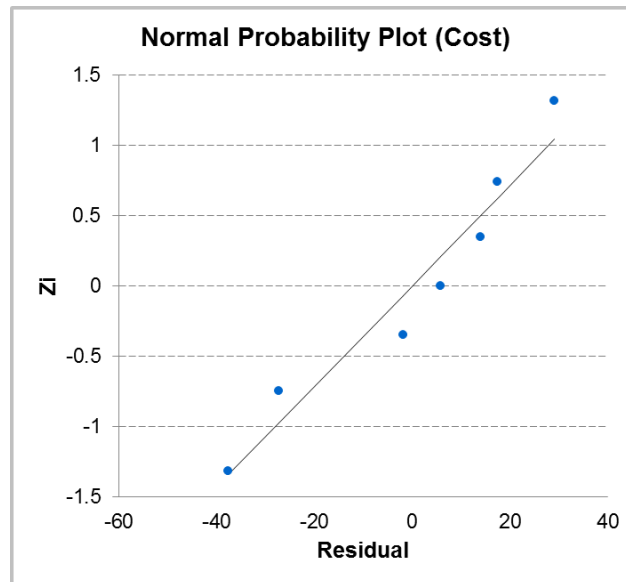


Figure 59: Normal Probability-Probability Plot from CO\$TAT

- Many statistical packages provide all four of the normality tests. To illustrate the interpretation of these tests, consider the results for the Shapiro-Wilk and Kolmogorov-Smirnov tests, returned by R:

```
Shapiro-Wilk normality test

data:  rstudent(lm(Cost ~ Power + Aper))
W = 0.8759, p-value = 0.2088

One-sample Kolmogorov-Smirnov test

data:  rstudent(lm(Cost ~ Power + Aper))
D = 0.2293, p-value = 0.7823
alternative hypothesis: two-sided
```

- In this example, both of these tests return a p-value far greater than $\alpha = 0.05$. Both the SW and KS tests fail to reject the null hypothesis in favor of the alternative —, which assumes the errors are normally distributed. This result agrees with the visual analysis from **Figure 59**.
- It is common practice to accept (with documentation) a minor violation of the normality assumption. However, there are several approaches to remedy the violation if the deviation is too severe. When the tails are too long or skewed, a transformation of the x or y variables often remedies the problem. Thus, proceeding to Section [3.3.3](#) and fitting a Log-Linear Model may be appropriate. With a skewed distribution, another option is to assume a different distribution for the errors. If selected from the exponential family, such as the log-normal distribution, the [Generalized Linear Model \(GLM\)](#) can be utilized to fit the model. Accepting the normality assumption when the evidence suggests

otherwise is common practice in cost analysis. In extreme cases, use a transform to attempt to correct the problem, or fit a GLM with an alternate distributional assumption.

- **Linearity:**
 - Recall the OLS example introduced in Section [3.3.1.3](#) with response variable *Cost* and predictors *Power* and *Aperture*. As part of the output, CO\$TAT produced a Standardized Residual plot, displayed in [Figure 56](#), and an Actual vs. Predicted plot, displayed in **Figure 60**. The predicted versus actuals plot in **Figure 60** shows no evidence of a pattern and indicates no problem regarding the linearity assumption. This plot can also help validate the independence of errors and constant variance assumptions. **Figure 60** supports accepting the linearity assumption.

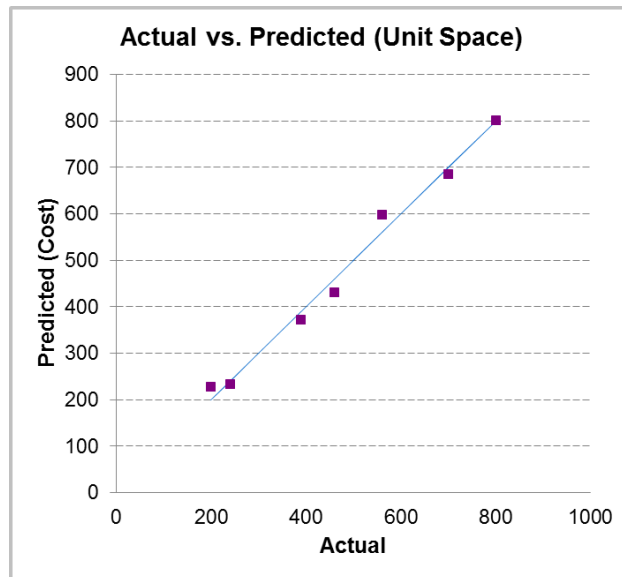


Figure 60: OLS Actual vs. Predicted Plot

- If the pattern does not support linearity, the analyst may choose to transform the variables and fit the Log-Linear model or use an alternative model form.
- Violation of the linearity assumption is the most critical assumption to address. These assumptions indicate the linearity form may not be optimal for this example. Without satisfying the linearity assumptions, the supporting statistics become meaningless when they are not satisfied.

While this guide presents the assumptions sequentially, assess all four assumptions (Sections [4.2.1.2](#) through [4.2.1.5 Linearity](#)) before making a decision about any single assumption. Only after assessing each of them, make a decision on which to address. If all four assumptions are acceptable, the next step is to proceed to [4.3 Model Diagnostics](#).

4.2.2 Weighted Least Squares (WLS)

Section [5.2.2](#) introduced the [Generalized Least Squares \(GLS\)](#) Model in two forms: the generic case and [Weighted Least Squares \(WLS\)](#). Typical use of the GLS model is in time series applications and is

beyond the scope of this guide. WLS application is similar to OLS application but introduces a weight term. Recalling Section [3.3.2.2](#), the WLS model is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{W}^{-1}) \text{ and } \mathbf{W} = \langle \mathbf{w} \rangle$$

4.2.2.1 Residuals

The residual error is the difference between the actual value and the predicted value. This is the raw residual, and for OLS is,

$$\begin{aligned} e_{ols} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

However, in WLS, the x and y variables are transformed into a scaled space with constant variance. Therefore, the residuals become,

$$\mathbf{e}_{wls} = \mathbf{W}^{\frac{1}{2}}\mathbf{y} - \mathbf{W}^{\frac{1}{2}}\mathbf{X}\hat{\boldsymbol{\beta}}$$

With this logic, the same discussion on standardizing residuals applies to WLS as with OLS (Section [4.2.1](#)). Similarly to OLS, the internally studentized residual is,

$$e_i = \frac{e_{wls,i}}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

Where,

$$\begin{aligned} \hat{\sigma} &= \sqrt{MSE} \\ h_{ii} &= i^{th} \text{ diagonal entry of the hat matrix, } \mathbf{H} \\ \mathbf{H} &= \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime} \\ \mathbf{X}^* &= \mathbf{W}^{\frac{1}{2}}\mathbf{X} \end{aligned}$$

The internally studentized residual follows an approximate standard t-distribution, with a mean of zero and variance of one. Similar to OLS, the following modification results in the externally studentized (or deleted) residual,

$$e_{i,-1} = \frac{e_{wls,i}}{\hat{\sigma}_{i,-1}\sqrt{1 - h_{ii}}}$$

Where,

$$\begin{aligned} \hat{\sigma}_{i,-1} &= \sqrt{MSE} \text{ (calculated without data point } i) \\ h_{ii} &= i^{th} \text{ diagonal entry of the hat matrix, } \mathbf{H} \\ \mathbf{H} &= \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime} \\ \mathbf{X}^* &= \mathbf{W}^{\frac{1}{2}}\mathbf{X} \end{aligned}$$

The internally studentized residual is acceptable to use and can be summarized by most software packages, including COSTAT. Many other statistical software packages provide the option to return the

externally studentized residual, including SAS and R, referred to as ‘rstudent’, and Minitab, referred to as the ‘deleted residual’. When available use the externally studentized residual. Plots of the studentized residual provide a means of visually assessing fit characteristics.

4.2.2.2 Independence of Errors

The statement of $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{W}^{-1})$ and $\mathbf{W} = \langle \mathbf{w} \rangle$ captures the assumption that each error is distributed independently with their respective variances captured by \mathbf{w} . Perform the same analysis as covered in Section [4.2.1.2](#), but now on the WLS standardized residuals.

Recalling Step [3.3.2.2](#) and the Gauss-Markov theorem, when the independence of errors assumption fails, the WLS model no longer provides the Best Linear Unbiased Estimate of the coefficients. As a result, the variances around the coefficients may be inflated. However, the estimator is still unbiased.

4.2.2.3 Independence of Errors Example

Recall the WLS example introduced in Section [3.3.2.2](#) with response variable Cost and predictor Power. This example demonstrates the results produced by WLS, compared to the OLS fit. When the regression was run, COSTAT produced a Standardized Residual plot, displayed in **Figure 61** and **Figure 62**, for OLS and WLS, respectively.

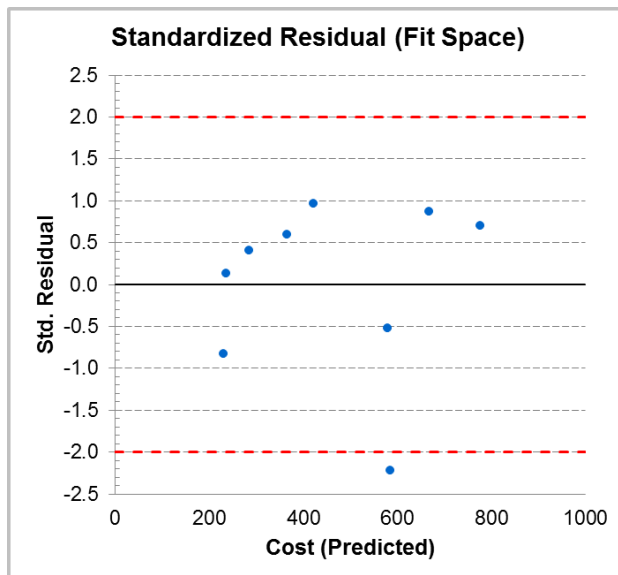


Figure 61: OLS Standardized Residual Plot

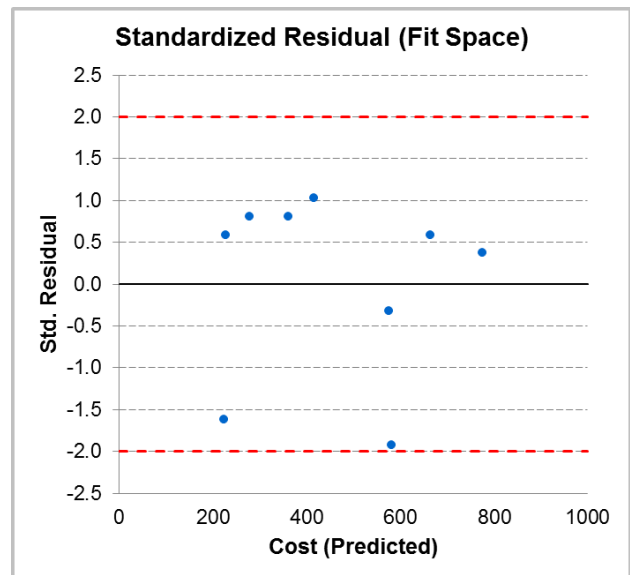


Figure 62: WLS Standardized Residual Plot

Figure 61 displays random scatter. No pattern is obvious, suggesting a violation of the independence of errors assumption. **Figure 62** appears to be very similar and leads to the same conclusion. The similarities in the plots suggest the use of WLS is not resulting in significant changes to the residuals. The OLS model is preferable in this example because the WLS model does not significantly improve residual performance.

While this guide presents the assumptions sequentially, assess all four assumptions (Sections [4.2.2.2](#) through [6.2.2.4](#)) before making a decision about any single assumption. Only after assessing each of them, make a decision on which to address. If all four assumptions are acceptable, the next step is to proceed to run [4.3 Model Diagnostics](#).

A failed independence of errors assumption often involves employing a time series methodology—, which is beyond the scope of this guide. These methods account for correlation between the residuals, or time periods, and attempts to build these relationships directly into the model. Section [3.3.2 Generalized Least Squares \(GLS\)](#) introduces a framework to address correlation impacts. The analyst is recommended to proceed with their analysis even in the presence of minor independence concerns.

4.2.2.4 Homoscedasticity

The statement of $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{W}^{-1})$ and $\mathbf{W} = \langle \mathbf{w} \rangle$ captures the assumption that the error term has a variance proportional to \mathbf{w} . Normalize the residuals for \mathbf{W} and then validate them in the same way as with OLS.

Perform the same analysis as covered in Section [4.2.1.3](#), but now with $\mathbf{W}^{\frac{1}{2}}\mathbf{X}$ and $\mathbf{W}^{\frac{1}{2}}\mathbf{y}$ instead of \mathbf{x} and \mathbf{y} for scatter plots, and on the WLS standardized residuals.

Recalling Step [3.3.2.2](#) and the Gauss-Markov theorem, when the independence of errors assumption fails, the WLS model no longer provides the Best Linear Unbiased Estimate of the coefficients. As a result, the variances around the coefficients may be inflated. However, the estimator is still unbiased.

4.2.2.5 Homoscedasticity Example

Recall the WLS example introduced in Section [3.3.2.2](#) with response variable *Cost* and predictor *Power*. This example demonstrates the results produced by WLS, compared to the OLS fit. When the regression was run, CO\$TAT produced a Standardized Residual plot, displayed in [Figure 61](#) and [Figure 62](#), for OLS and WLS.

In this example, [Figure 61](#) does not indicate a clear violation of homoscedasticity. The model was refit using WLS and [Figure 62](#) illustrates the weighted residuals from this analysis. The resulting standardized residual plot also highlights homoscedastic behavior of random scatter and no pattern.

If the WLS model did not address the heteroscedasticity concern, then fit a new WLS model with a different set of weights. Section [3.3.2.2 Weighted Least Squares \(WLS\)](#) suggests a few methods for selecting weights.

4.2.2.6 Normality of Errors

The statement of $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{W}^{-1})$ and $\mathbf{W} = \langle \mathbf{w} \rangle$ captures the assumption that each error is distributed according to the normal distribution. Perform the same analysis as covered in Section [4.2.1.4 Normality of Errors](#), but now on the WLS standardized residuals.

Recalling Step [3.3.2.2](#) and the Gauss-Markov theorem, when the normality of errors assumption fails the WLS model still provides the Best Linear Unbiased Estimate of the coefficients, as long as the other three assumptions hold. Normality is a simply construct for making model selection decisions, albeit very useful and often essential. As a result, failure of normality still results in an unbiased estimator with

minimum variance, but severely limits the ability to conduct inference including outlier detection, significance testing, and risk analysis.

4.2.2.7 Normality of Errors Example

Recall the WLS example introduced in Section [3.3.2.2](#) with response variable *Cost* and predictor *Power*. This example demonstrates the results produced by Method 3, compared to the OLS fit. When the regression was run, CO\$TAT produced a Probability Plot (P-P), displayed in **Figure 63** and **Figure 64**, for OLS and WLS.

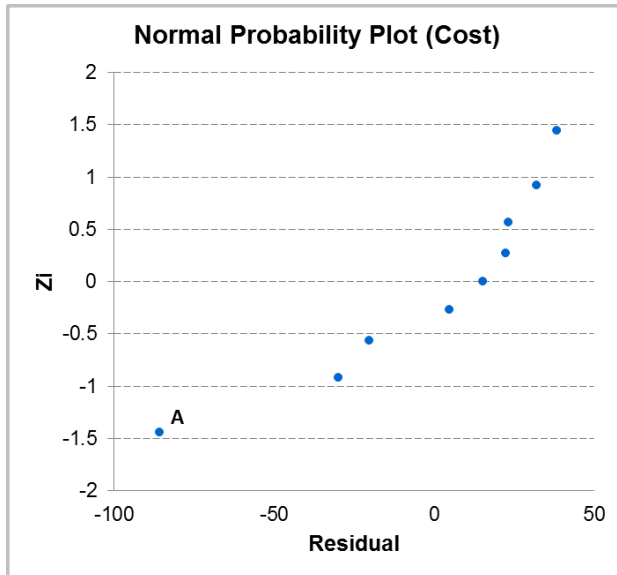


Figure 63: OLS P-P Plot

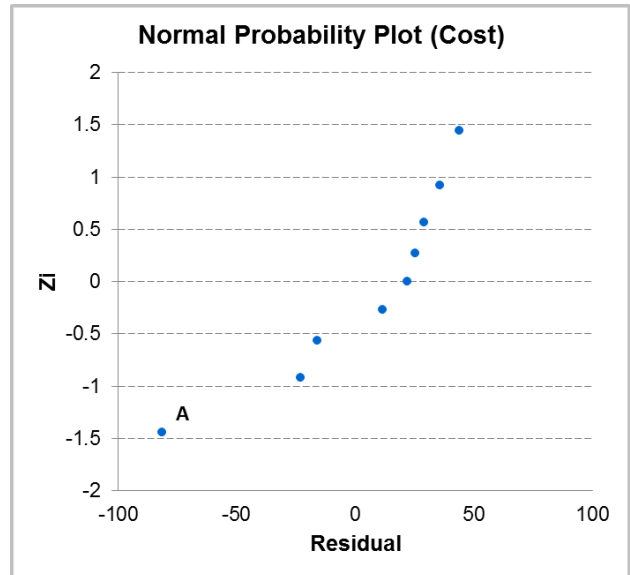


Figure 64: WLS P-P Plot

In this example both, **Figure 63** and **Figure 64** suggest normality. Ignoring the point labeled “A”, the remaining residuals follow a moderately straight line with no apparent patterns. Point A represents a deviation from the other data points and may violate normality assumptions. However, this is common when working with real-world data, and does not warrant rejection of the assumption. The presence of non-constant variance (or multiplicative error) in the dataset did not affect the ability to accept the normality assumption.

It is common practice to accept (with documentation) a minor violation of the normality assumption. However, there are several approaches to address a violation where the deviation is too severe. When the tails are too long or skewed, a transformation on x or y often remedies the problem—indicating that Section [3.3.3](#) and fitting a Log-Linear Model may be appropriate. With a skewed distribution, another option is to assume a different distribution for the errors.

If selected from the exponential family, such as the log-normal distribution, the [Generalized Linear Model \(GLM\)](#) can be utilized to fit the model. In summary, accepting the normality assumption, with cautions, is common practice. However, extreme cases sometimes call for the use a transform to mitigate any concern.

4.2.2.8 Linearity

The assessment of the linearity assumption is the same as with OLS. The [scatter plot](#) and predicted versus actual plots will be identical. The residual plot using the WLS residuals may make the results more clear, but the OLS residuals should provide the same conclusion. Section [4.2.1.5 Linearity](#) to discuss this topic in greater detail.

4.2.3 Transforms and the Log-Linear Model

Model transformations by themselves are not a standalone regression methodology. Transformations involve taking the predictor and/or response variable(s) and applying a transformation. In the case of the Log-Linear model, a transformation is done on \mathbf{y} , and sometimes \mathbf{x} . After transforming the selected variables, run the OLS model on the transformed variable set,

$$\begin{aligned}\mathbf{y}^* &= \ln(\mathbf{y}) \\ \mathbf{X}^* &= \ln(\mathbf{x})\end{aligned}$$

Now, validate the OLS assumptions using the process detailed in Section [4.2.1](#), only in the transformed space, $\{\mathbf{X}^*, \mathbf{y}^*\}$.

4.2.4 Generalized Linear Model (GLM)

Section [3.3.4](#) introduced the [Generalized Linear Model \(GLM\)](#), with more details provided in Appendix [A.4.4](#). The assumptions for GLM are different than OLS because GLM is a method of maximum likelihood, not of least squares. GLM application is an advanced topic covered in more detail under section [A.4.4.5 Validate CER \(Assumptions\)](#).

4.2.5 Non-linear Least Squares (NLS)

Section [3.3.5](#) introduced the [Non-linear Least Squares \(NLS\)](#) methodology. The assumptions for NLS are less strict than for other models, which provides greater flexibility when choosing a functional form to fit the data. The lack of well-defined assumptions unfortunately translates into less desirable verification processes and interpretations of the model. The same principles from OLS extend to NLS where assumption violations may influence the model. Recalling Section [3.3.5](#), the NLS model is,

$$\mathbf{y} = f(\mathbf{X}; \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \text{ and } \boldsymbol{\Sigma} = \langle \mathbf{w}^{-1} \rangle$$

Theorems do not govern the impacts of the violation of this assumption (e.g., bias, variance inflation) in the way the Gauss-Markov theorem governs OLS. Rarely are the NLS assumptions examined with the same rigor as the other models discussed in this handbook. Since NLS is used when other models are not feasible, the focus shifts to concepts covered in the remaining sections of Step 4 (e.g., minimizing model error).

4.2.5.1 Residuals

The residual error is the difference between the actual value and the predicted value. This is the raw residual and for NLS is,

$$\mathbf{e}_{nls} = \mathbf{y} - f(\mathbf{X}; \hat{\boldsymbol{\beta}})$$

Similar to OLS, these are not the correct residuals to use; standardization is required. This process is more complex for NLS, being dependent on the final iteration of the numerical algorithm. Refer to the residual standardization process discussed in Section [4.2.1](#).

4.2.5.2 Independence of Errors

The statement of $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \langle \mathbf{w}^{-1} \rangle$ captures the assumption that each error is distributed independently and is validated using the same OLS methods, but now on the NLS residuals. Perform the same analysis as covered in Section [4.2.1.2](#), but now on the NLS standardized residuals.

4.2.5.3 Homoscedasticity

The statement of $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \langle \mathbf{w}^{-1} \rangle$ captures the assumption that the error term has a variance proportional to \mathbf{w} . Perform the same analysis as covered in Section [4.2.1.3](#), but now with the NLS standardized residuals.

4.2.5.4 Normality of Errors

The statement of $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{W}^{-1})$ and $\mathbf{W} = \langle \mathbf{w} \rangle$ captures the assumption that each error is distributed according to the normal distribution. Perform the same analysis as covered in Section [4.2.1.4](#), but now on the NLS standardized residuals.

4.2.5.5 Functional Form

Assess the functional form in the same way as with OLS, but with the NLS residuals. With a single predictor, examine a scatter plot to determine if the curve is a reasonable fit. Examine the residual plot and predicted versus actuals plot for the same types of information as introduced in Section [4.2.1.5](#) [Linearity](#) with OLS.

4.2.6 Ridge Regression

Ridge Regression solves for the estimated coefficient vector, $\hat{\boldsymbol{\beta}}$, under a size restriction. No alterations are made to \mathbf{x} , \mathbf{y} , or to any of the predicted value calculations. As a result, calculate the residuals using the same OLS method. Validate the OLS assumptions using the process detailed in Section [4.2.1](#).

4.2.7 Restricted Least Squares (RLS)

Restricted Least Squares (Section [3.4 Estimation with Prior Information](#)) solves for the estimated coefficient vector, $\hat{\boldsymbol{\beta}}$, under a set of restrictions. No alterations are made to \mathbf{x} , \mathbf{y} , or to any of the predicted value calculations. As a result, calculate the residuals using the same OLS method. Validate the OLS assumptions using the process detailed in Section [4.2.1](#).

4.3 Model Diagnostics

After validating the model assumptions, the next step is to conduct model diagnostics on the CER. There may be outlier or leverage points with high influence on the model skewing the results. Additionally, independent variables may be highly correlated to each other, creating problems with the model statistics examined in Section [4.4 Model Significance](#). This is the multicollinearity problem described in Section [4.3.2](#).

4.3.1 Influential Points

High influence points are observations with larger impacts on the model relative to the other points. One way, which an observation can have high influence, is in the predictor space. This is referred to as a leverage point, and has a value far away from the rest of the data, creating a “lever” effect on the model. For example, if using weight to predict cost and one system weighs substantially more than the others, the mathematics cause the much heavier system to affect the model coefficients more than the other points.

– Terminology –

High influence points (HIPS) come in two forms: in the predictor space (x) or in the response space (y).

Leverage points excessively influence the model in the predictor space.

Outlier points excessively influence the model in the response space.

Using the same example as before, now suppose all the systems weigh about the same, but one has a substantially greater cost. This observation may be an outlier, skewing the model away from the actual trend of the data.

It is possible that an observation is extreme in both the predictor and the response space. This scenario results in the observation of interest having a large impact on the model, heavily skewing the model away from the actual trend of the data.

Assess leverage points and outliers simultaneously to understand the full impact of the observations on the model. The remainder of this section discusses visual tests and numerical metrics to assess both of these types of high influence points.

Each of these metrics can be found in **Table 23** which shows the Outlier Analysis Table from COSTAT for the multivariable example problem introduced with OLS in Section 3.3.1.3.

Table 23: Multivariable OLS Example Outlier Analysis Table

Obs #	Cost	Predicted Y Value	Residual	Std. Dev. Pred Y	Std. Residual	Leverage Value	Cook's Distance	Flags
1	390.0000	372.5577	17.4423	13.6074	0.6594	0.2093	0.0384	
2	200.0000	227.2173	-27.2173	21.8781	-1.3506	0.5410	0.7166	
3	240.0000	234.0809	5.9191	17.5772	0.2467	0.3492	0.0109	
4	300.0000							
5	460.0000	430.8160	29.1840	17.6362	1.2184	0.3515	0.2683	
6	560.0000	597.5062	-37.5062	20.7648	-1.7610	0.4873	0.9826	D
7	700.0000	685.9578	14.0422	22.0057	0.7016	0.5473	0.1984	
8	800.0000	801.8639	-1.8639	21.3334	-0.0899	0.5144	0.0029	
9	500.0000							

SE = 29.7453, Mean = 478.5714, Coef. of Var. = 6.22% in Fit Space

D denotes an observation with an unusual influence on the fitted regression equation.

Figures 4 and 9 represent observations that lack all selected independent variables.

4.3.1.1 Standardized Residuals

The most commonly used metric to assess outliers is the residual. Section 4.2.1 introduced the different types of residual – raw, internally studentized, and externally studentized. Examine the standardized residuals (externally studentized, if available, internally studentized otherwise) against a rule-of-thumb to

flag observations as potential outliers. The column marked “Std. Residual” in **Table 23** is the internally standardized residual returned from CO\$TAT

A common rule-of-thumb for assessing standardized residual values is ± 2 , which can be interpreted as a rough 95% level of confidence, depending on the number of degrees of freedom. For smaller datasets (i.e., $n < 10$), ± 3 may be more representative.

Note that these points are marked only as *potential* outliers because with 100 observations, expect about five to be flagged as outliers when assuming a 95% level of confidence.

To illustrate the importance of standardizing the residuals, consider the scatter plot and regression line in **Figure 65**. The red dashed line in the plot on the left signifies the true model from which the data were generated. The blue dashed line in the plot on the right is the linear model fit through the data.

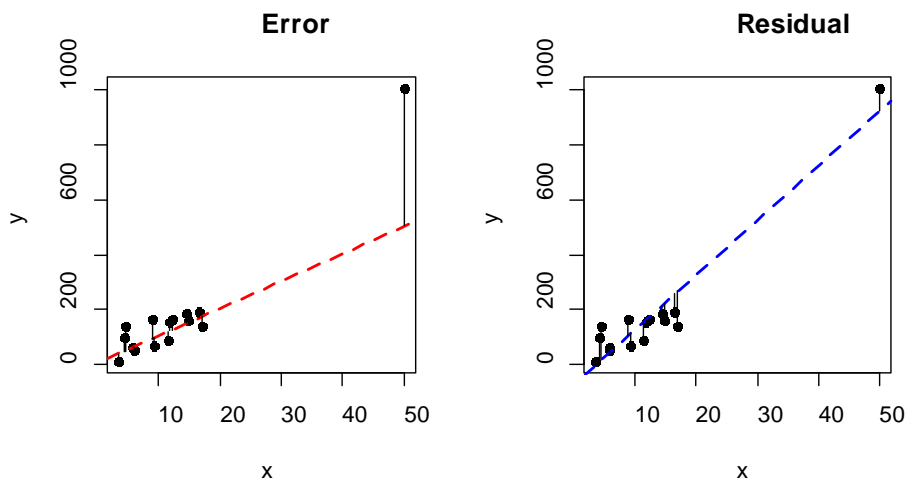


Figure 65: Error versus Residual

The data point in the top right is manually adjusted to be an outlier, while the remainder of the data follow a linear trend. **Figure 66** presents a comparison of the residuals.

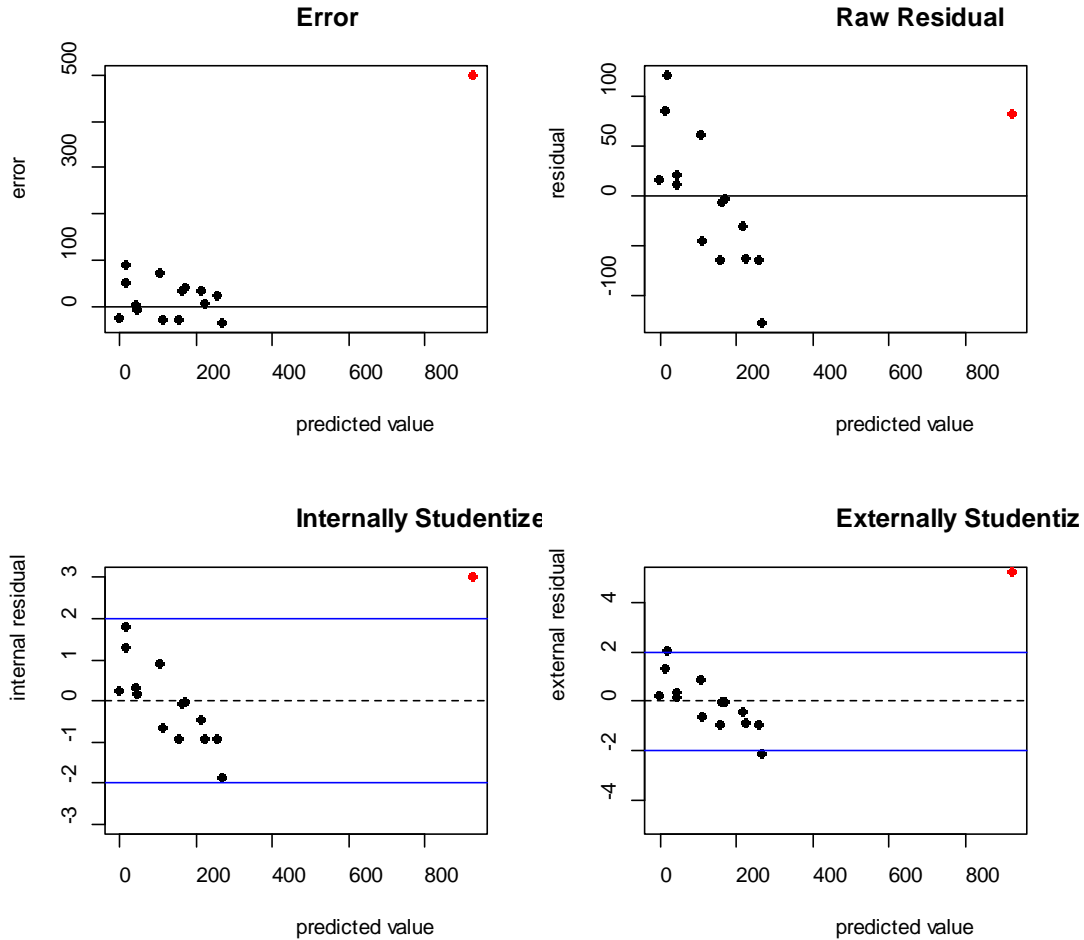


Figure 66: Comparison of Residual Types

The top left plot shows the error associated with the aforementioned outlier point. This is the actual error. In this case, the error is known since the data were simulated, but in real world situations, this is never the case. The top right plot shows the raw residuals, and the extreme point does not stand out. In fact, there are four more points with a greater residual in magnitude. The lower left plot shows the internally studentized residuals and the extreme point is correctly flagged as a potential outlier. The lower right plot shows the externally studentized residuals with the outlier value well above the rule-of-thumb cutoff of ± 2 .

4.3.1.2 Leverage Value

The most commonly used metric to assess leverage is the diagonal value of the hat, or projection, matrix. This metric is also simply referred to as the leverage value. The hat matrix is named as such because it “puts the hat on \mathbf{y} ”, meaning $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. It is defined as,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The hat matrix is a function of only the predictor values, and its diagonal entries, notated h_{ii} , represent the leverage each observation has on the model. The column marked “Leverage” in **Table 23** is the hat

diagonal leverage value returned from CO\$TAT. Leverage values and leverage plots can be generated using other analytical tools (e.g., SAS JMP, R, MS Excel)

A general rule-of-thumb is that leverage values greater than $\frac{2p}{n}$ or $\frac{3p}{n}$ (where $p = k + 1$, or the number of parameters in the model) may be of concern for having high leverage on the model.

4.3.1.3 Cook’s Distance

Cook’s Distance (Cook’s D)⁶² is a metric to assess overall influence of a point on the model, that is, in both the \mathbf{x} and \mathbf{y} directions. Cook’s D calculates a standardized distance between the coefficients, and the coefficients calculated in the absence of the observation of interest. The column marked “Cook’s Distance” in **Table 23** is Cook’s D value returned from CO\$TAT using the following formulation,

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})'(X'X)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i})}{p\hat{\sigma}^2}$$

Where,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{-i} &= \text{Coefficient vector calculated without observation } i \\ \hat{\sigma}^2 &= \text{Mean Squared Error (MSE)} \end{aligned}$$

An equivalent derivation method is,

$$D_i = \frac{1}{p\hat{\sigma}^2} \sum_{j=1}^n (\hat{y}_j - \hat{y}_{j,-i})^2$$

Where,

$$\begin{aligned} \hat{y}_{j,-i} &= \text{Predicted value calculated without observation } i \\ \hat{\sigma}^2 &= \text{Mean Squared Error (MSE)} \end{aligned}$$

A good rule-of-thumb is that a data point merits serious investigation if Cook’s Distance (D_i) exceeds the 50th percentile of the $F(p, n - p)$ distribution; the F -distribution with p and $n - p$ degrees of freedom where n is the number of observations (including the potential outlier) and p is the number of coefficient parameters in the regression model (including the y-intercept).

At the 50th percentile, $F(p, n - p) \approx 1$, the rule-of-thumb can be expressed as $D_i > 1$.

4.3.1.4 Leave-One-Out Metrics

Cross validation methods provide additional numeric ways to identify and assess influential points. In particular, “leave-one-out” regression can provide insights into how a single observation influences the

⁶² Additional metrics exist beyond Cook’s D. While not covered in this guide, DFFITS and DFBETAS are two other popular diagnostic metrics to assess influence.

overall model. Leave-one-out performs the regression n times, once for each observation with that specific observation omitted. This concept is discussed in Section [4.5.2.2](#). In the OLS case, several metrics can be calculated for assessing influential points instead of having to go through the actual mechanics of performing the regression n times.

The externally studentized residual is actually a leave-one-out metric. The residual and the externally studentized concept is discussed in prior sections, including Section [4.3.1.1](#). Recall that this residual is calculated using a σ^2 calculated under a leave-one-out method.

DFFITS is a metric similar to the externally studentized residual and compares how far off an observation is predicted from its observed value. This metric is largely redundant of information derived from Cook's D . The following is a reasonable rule-of-thumb for when DFFITS is considered to be "large", signaling an influential point:

$$DFFITS > 2 \sqrt{\frac{\bar{p}}{n}}$$

DFBETAS is a metric that compares how much each regression coefficient changes based on the removal of each observation. If there are k independent variables, then k different values of DFBETAS are calculated per observation. Again, this information can be redundant to that of Cook's D . The following is a reasonable rule-of-thumb for when DFBETAS is considered to be "large", signaling an influential point:

$$DFBETAS_i > 2 \sqrt{\frac{1}{n}}$$

Overall, leave-one-out methods for identifying outliers can be useful. A sound strategy is to consider both the studentized residual and Cook's Distance. DFFITS and DFBETAS can be examined if desired, especially if utilizing a statistical package where they are calculated automatically.

4.3.1.5 Visual Tests

Identify high influence points visually by plotting the previously discussed metrics. Visualization of data are discussed earlier in this text and examples provided below.

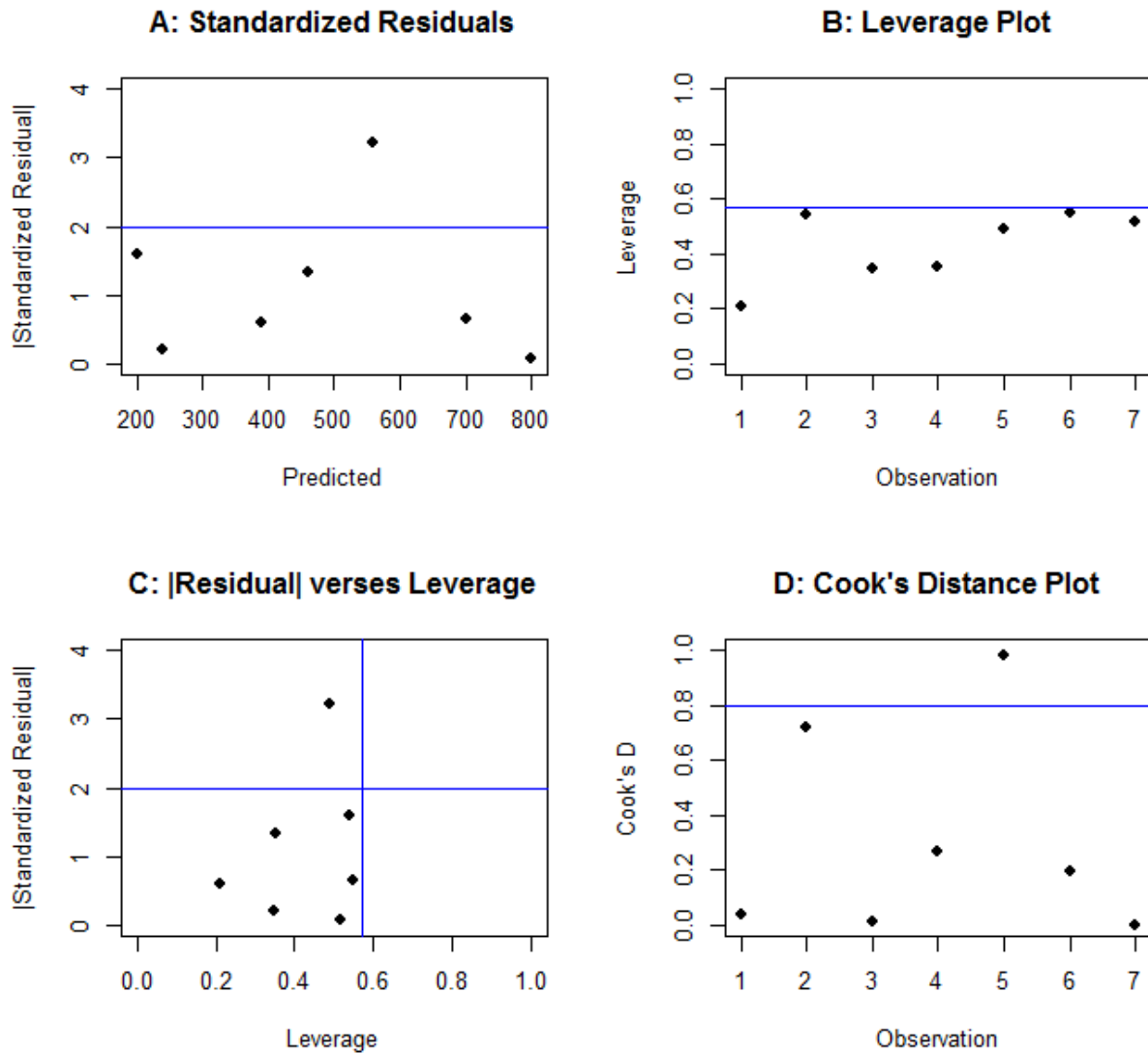


Figure 67: Diagnostic Plots for Section 3.3.1.2 OLS Example

Plot A: Standardized Residuals

To assess outliers, examine the same residuals versus predicted plots used to validate assumptions. Since the interest is now in the magnitude, it is convenient to plot the absolute value of the residuals. Observations standing out from bulk of the data and are above the rule-of-thumb cutoff line of 2 are of concern. In this example, one data point falls above the rule-of-thumb cutoff. The value has a standardized residual slightly greater than 3, which warrants a closer look.

Plot B: Leverage Plot

To assess leverage, create a [scatter plot](#) with the observation number on the x -axis, the leverage value (h_{ii}) on the y -axis, and the rule-of-thumb line drawn as a frame of reference. In this example, no points stick out as being extreme from the others. While some are right around the cutoff line, none appear to “stick out” enough to warrant additional attention.

Plot C: |Residual| versus Leverage

To assess both outliers and leverage points at the same time, create a scatter plot with the leverage value on the x -axis and the absolute value of the standardized residual on the y -axis. Add the respective rule-of-thumbs dividing the plot into four quadrants. The lower left quadrant contains observations with little concern for either being an outlier or a leverage point. The bottom right contains points with high leverage, but that are not potential outliers. The top left contains points without high leverage, but that may be potential outliers. Finally, the top right quadrant contains the points of concern; high leverage potential outliers. Again, the rule-of-thumbs provide a good starting place, but points sticking out from the rest are those of the highest concern. In this example, no points appear to be of major concern. No points fall above the cutoffs. The results agree with the conclusions drawn in Plot B and Plot C.

Plot D: Cook's Distance

To assess points as either outliers and/or leverage points at the same time, create Cook's D plot, with the observation number on the x -axis and the Cook's D statistic on the y -axis. Add the rule-of-thumb cutoff to the plot. Investigate those points crossing the rule-of-thumb threshold. In this example, Project 5 slightly stands out by being above the cutoff produced at the 50th percentile of the F-distribution. The Cook's Distance statistic is in agreement with the previous diagnostics and is not bounded at 1 (as may be suggested by the plot).

The leave-one-out metrics can be plotted similarly to Cook's Distance. Simply plot DFFITS and/or DFBETAS with their rule of thumb. Look out for values above the rule of thumb line, or ones that stand out from the remainder of the data.

4.3.1.6 Extension to Other Model Forms

The discussion of influence points focuses on the OLS model. However, these principles translate directly to more complex functional forms. The interpretations are nearly identical, with some mathematical subtleties in the background. Significant differences, where they exist, are called out and discussed. For the purposes of this guide, it is sufficient to understand that while the metrics may be calculated differently, statistical software packages output the material in a very similar fashion.

Statistical packages output leverage values and standardized residuals for all functional forms. For some, such as [Generalized Least Squares \(GLS\)](#), [Transformable Linear and the Log-Linear Model](#), and [Ridge Regression](#), the calculations are exact and based on formulas very similar to – or even transformed to be equal to – those of OLS. For others, such as the [Generalized Linear Model \(GLM\)](#) and [Non-linear Least Squares \(NLS\)](#), the results are asymptotical and based on the convergence properties of normal theory (Appendix [A.3.3.1.2 Small Data Sets – Asymptotic Results](#)).

The GLM often has better properties than NLS due to characteristics of [Maximum Likelihood Estimation \(MLE\)](#) (Appendix [A.4.7.2](#)). While these asymptotic metrics are different than OLS, their analyses and interpretations can be conducted under the same framework.

Additionally, leave-one-out regression, a type of cross validation, can calculate these metrics easily with modern computing power. This is discussed more thoroughly in Section [4.5.2.1](#).

The assessment of influence points may uncover several observations requiring further investigation. The statistics do not decide whether or not to remove a data point. Points with high leverage on the model that are not outliers are rarely of concern. Encountering observations with standardized residuals greater than two is expected. Simply having a large standardized residual does not dictate that the point should be removed. In fact, removing potential outlier points almost always results in more points being flagged as potential outliers, iteratively removing data until very little is left.

Small sample data sets further magnify the problem. Many CERs have relatively few observations. An observation, which appears to be an outlier, may actually be a very useful piece of information. Collecting more data would perhaps yield similar results and the observation would no longer be an outlier. Data are an extremely valuable commodity, and thus only remove points after a thorough examination.

After examining potential outliers in more detail, the next step is to assess multicollinearity.

4.3.2 Multicollinearity

Multicollinearity is the condition where multiple independent variables are highly correlated with each other, as discussed in Appendix A.2.1.4. This correlation makes it difficult for the model to distinguish between the predictors, resulting in high degrees of uncertainty, or variance, around the estimators. When correlation between two predictors occurs, the existence causes singularity of the covariance matrix. In fact, when perfect (or near perfect) correlation exists, calculation of the regression coefficients and respective statistics is impossible. **Figure 68** shows weak correlation between two predictors, x_1 and x_2 , on the left, and strong correlation on the right.

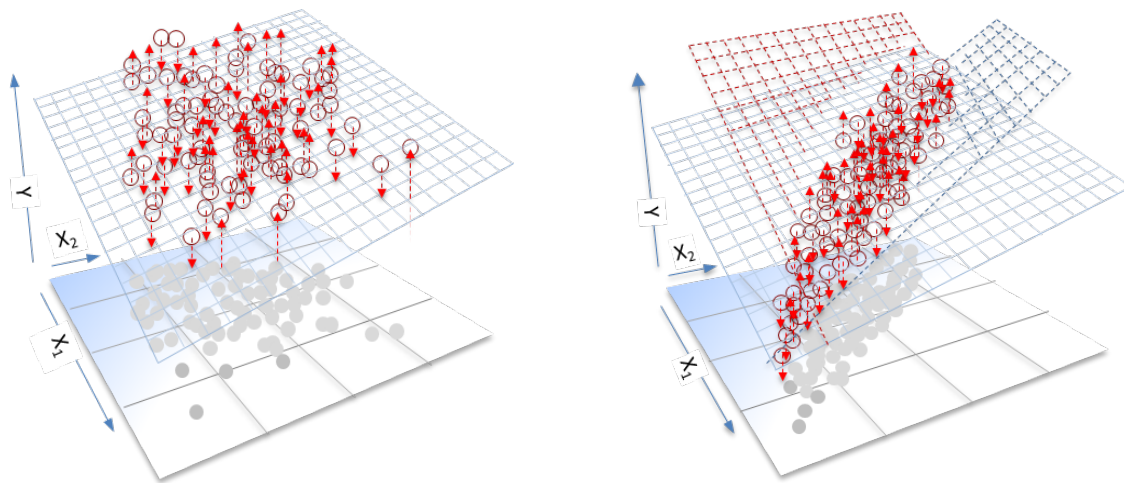


Figure 68: Weak and Strong Correlation Between Predictors

The presence of multicollinearity can significantly impact model prediction accuracy:

- (1) **Coefficients have inflated standard errors.** As a result, the model may have no significant predictors, despite being statistically significant as a whole. Section [6.4](#) covers model and variable significance.
- (2) **Coefficient estimates may not be logical.** Predictor coefficients may have the wrong sign (e.g., negative when they should be positive) and be extremely large in magnitude. This scenario could be due to choice of model form or interrelated data.

- (3) **Prediction properties are poor.** Multicollinearity causes a hidden extrapolation problem, resulting in poor prediction and excessively wide confidence intervals ([Step 5: Characterize Uncertainty](#)).

While multicollinearity can be examined at a high level by use of visuals, the most reliable diagnostics are the numerical metrics. In the case of severe multicollinearity, many software packages will return an error referring to singularity of the covariance matrix. Encountering this error signals that multicollinearity may be a severe problem. The most effective way to diagnose multicollinearity is by rule-of-thumb cutoffs on statistical metrics. The following sections discuss two of the most common metrics: the correlation matrix and variance inflation factors (VIFs).⁶³

4.3.2.1 Numerical Metrics

The first way to identify multicollinearity is to look at the correlation matrix. The correlation matrix contains the pairwise correlations between each predictor.

The correlation matrix with more than two independent variables is insufficient for identifying multicollinearity problems. A generic rule-of-thumb is that a pairwise correlation above 0.8 indicates potential severe problems with multicollinearity. However, research shows that correlations much lower can cause severe problems and therefore other metrics must be used in addition to the correlation matrix.

Table 24 presents example data to illustrate multicollinearity impacts with four independent variables.

Table 25 shows the resulting correlation matrix for this data set. In this example, x_1 , x_2 , x_3 , and x_4 all appear to be highly correlated (over 80%).

Table 24: Multicollinearity Data Example

Observation	x_1	x_2	x_3	x_4	y
1	465	9,264	2389	564	127,477
2	419	9,017	2210	517	127,092
3	537	10,673	2803	666	152,717
4	515	10,305	2456	615	146,827
5	536	10,186	2434	586	147,403
6	530	10,290	2650	644	149,954
7	466	10,271	2405	585	156,673
8	522	10,437	2689	636	141,290
9	494	9,762	2428	571	145,770

⁶³ This is not a comprehensive list of metrics. The use of eigenvalues via the spectral decomposition is another popular way to assess multicollinearity. This metric is the Condition Number and is provided by many statistical packages.

10	430	8,492	2053	536	133,256
11	475	9,584	2446	568	118,372
12	511	9,594	2638	609	125,825
13	495	9,294	2500	630	119,035
14	601	12,013	2871	685	171,826
15	642	12,736	3217	789	168,742

Table 25: Multicollinearity Data Example Correlation Matrix

Variables	y	x1	x2	x3	x4
y	1.0000	0.7294	0.8500	0.6232	0.6647
x1	0.7294	1.0000	0.9340	0.9317	0.9362
x2	0.8500	0.9340	1.0000	0.9064	0.9002
x3	0.6232	0.9317	0.9064	1.0000	0.9588
x4	0.6647	0.9362	0.9002	0.9588	1.0000

Table 25 would indicate there are several independent variables that are highly correlated. The Variance Inflation Factor (VIF) is another useful tool to identify multicollinearity. When there are more than two independent variables, VIF helps identify the correlated independent variables with the most impact on model variability. VIFs are calculated as the diagonal entries of the inverse of the correlation matrix. The VIF provides a summary of the impact on the variance, for each individual predictor.

A generic rule-of-thumb for VIFs is values greater than 5 may be cause for concern, and values greater than 10 indicate one or more independent variables should be removed from the model.

Inverting the correlation matrix presented in **Table 25**, the following diagonal entries are obtained: 12.9, 8.5, 14.6, and 15.0, for x_1 , x_2 , x_3 , and x_4 . Going back to the Ridge example, the VIF's can be verified by CO\$TAT in the Multicollinearity Analysis output, shown in **Table 26**.

Table 26: Multicollinearity Data Example Analysis

Indep Variables	Indiv R-Sqr (%)	F-Stats	Prob Related to Other Vars	Indiv R-Sqr/Model R-Sqr	VIF	Flags
x1	92.28%	43.8469	1.0000	1.0879	12.9583	X
x2	88.24%	27.5011	1.0000	1.0402	8.5003	X
x3	93.13%	49.7426	1.0000	1.0979	14.5662	X
x4	93.33%	51.3053	1.0000	1.1002	14.9924	X

Variables x_1 , x_3 , and x_4 exceed the rule-of-thumb cutoff.⁶⁴ Variable x_2 warrants further review. Removing x_1 , x_3 , or x_4 from the model may reduce the VIFs for the remaining predictors below the rule-

⁶⁴ The rule-of-thumb presented here is not the methodology used by CO\$TAT to flag highly collinear variables. For more information, see the CO\$TAT help file.

of-thumb thresholds. A best practice is to remove or change only one variable at a time and reexamine the VIF results.

4.3.2.2 Visual Tests

Visual assessment of multicollinearity is uncommon. Pairwise [scatter plots](#) can be created, known as a scatter plot matrix, plotting each predictor against each other predictor. **Figure 69** displays an example of a scatter plot matrix.

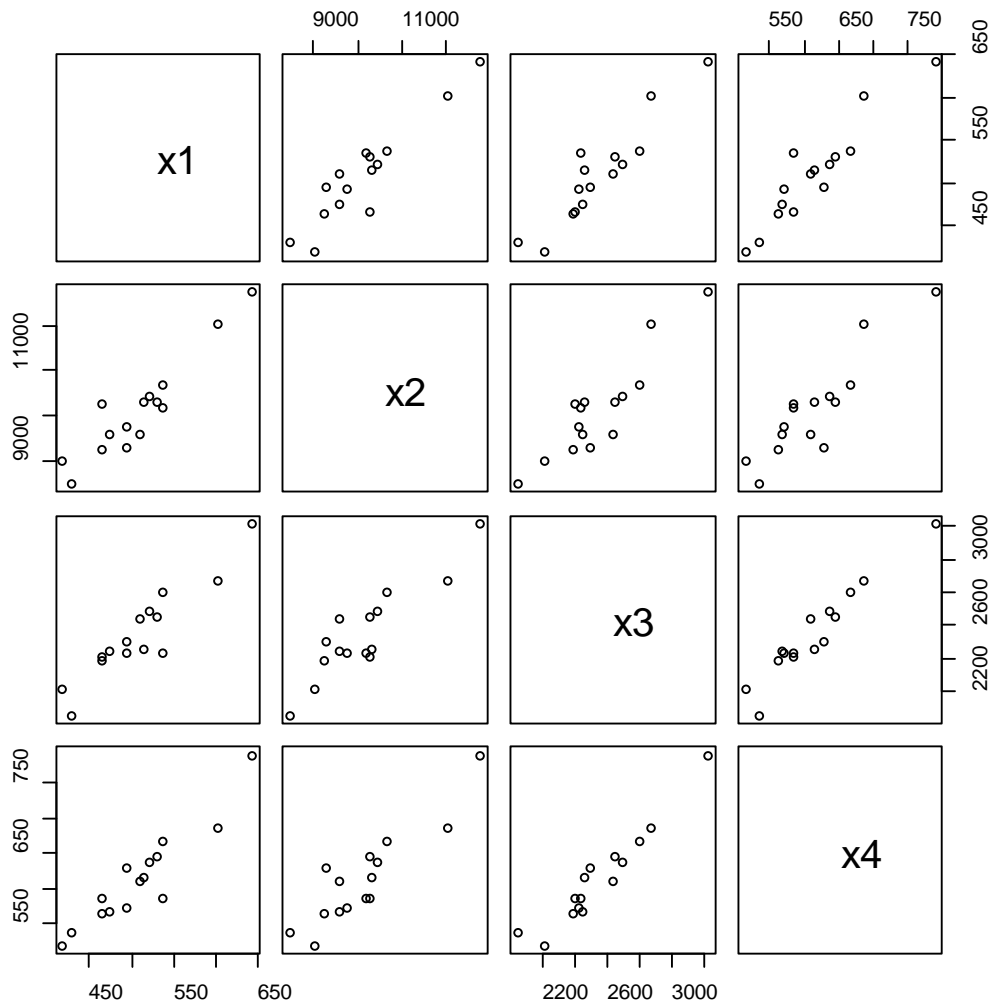


Figure 69: Scatter Plot Matrix

4.3.2.3 Extension to Other Model Forms

The discussion of multicollinearity focuses on the OLS model. However, these principles translate directly to more complex functional forms. The interpretations are nearly identical, with some mathematical subtleties in the background. For the purposes of this guide, it is sufficient to understand that while the metrics may be calculated differently, statistical software packages output the material in a very similar fashion.

Most statistical packages output correlation matrices and VIFs. Analyze these results using the same OLS framework.

4.3.2.4 Addressing Multicollinearity

When dealing with high multicollinearity, the first option is to eliminate one or more of the highly correlated variables. There are a number of ways to choose which variable to remove. The first option is to remove the variable(s) with the highest p-value, correlation coefficient, or VIF. Then, generate and validate a new CER. If multicollinearity is still present, either add the variable back into the model and remove a different variable(s), or simply remove an alternate variable(s), and repeat the fit and validation process.

Second, centering and scaling the data as described in Section [3.3.6 Ridge Regression](#) can help numerically stabilize the model. Attempt to refit the model using this transformation.

If possible, combining collinear variables into a single variable is a sound way to deal with multicollinearity. Doing so maintains the set of relevant cost drivers from the original CER, and can possibly maintain the functional form derived from prior engineering knowledge. Unfortunately, this approach is usually only possible in the simplest of CERs. Section [3.3.6 Ridge Regression](#) offers an alternative to the problem of multicollinearity. Another viable, related solution is the use of Principal Component Regression, introduced in Appendix [A.4.9.2](#) as an advanced topic.

If there are no significant multicollinearity impacts, the next step is assessing the significance of the model.

4.4 Model Significance

Many issues in CER development arise from choosing the “wrong” set of independent variables. Even when individual variables are reasonable choices, a certain combination of those variables may prove problematic. Enumerated below are four potential issues related to the choice of independent variables. The first three relate to individual variables, and the last one to the set of variables as a whole.

- (1) **Omission of a relevant variable:** This is the case where the equation excludes an important, relevant predictor. The cost driver exploration phase may have overlooked this variable altogether, or it simply may not have been included for this particular regression run. Omission of a “true” cost driver results in biased coefficient estimates for the remaining variables in the “incomplete” model. Coefficients that do not pass a “sanity check” (e.g., negative when expected to be positive, and vice versa) may be an indication of an omitted variable. Or, the model simply has a low R-squared value, indicating that a large portion of the variation in cost remains unexplained. If possible, the remedy is to include the variable or a reasonable proxy (e.g., system weight accounting for system complexity), and may require revisiting [Step 1: Purpose, Scope, Collect, Validate, & Normalize](#) and re-engaging with engineers to discuss what might be missing.
- (2) **Misspecification of a variable:** The previous case is the incorrect exclusion of a variable. Another case, the variable belongs, but its precise definition or expression is causing problems, resulting in biased and inconsistent coefficient estimates. As with many of these cases, unreasonable coefficient values can be a symptom, but the most common indication is an

insignificant variable (i.e., failed t-test) where it is expected to be significant. Specific instances and remedies are discussed below:

- a. **Incorrect dimension:** In shipbuilding, linear feet of cable may be a reasonable cost driver, but linear feet of pipe may be a problem because it fails to take into account significant differences in the diameter of various segments of pipe. Therefore, the customary cost driver is “square feet of pipe.”
 - b. **Improper use of proxy:** Perhaps out of necessity (or expediency), a proxy variable is used and is not performing well in the regression. For example, the model may use an SME-based assessment of interface complexity, but it turns out to be better to collect actual interface counts from Software Resource Data Reports (SRDRs).
 - c. **Failure to use proxy:** This is the converse of the previous case. A variable believed to be the true cost driver is not performing well in the regression, which leads the analyst to assess other variables that might be a reasonable proxy. This is why weight-based CERs are so common.
- (3) **Inclusion of an irrelevant variable:** This is the converse of case (1), above. An extra predictor that does not really “belong” is inappropriately included, resulting in inflated coefficient variances and artificially high R^2 and F-statistic values. Like case (2), failing the t-test is the most common indication. Unreasonable coefficient values can also be a symptom. Because adding an extra variable always increases R^2 , adjusted R^2 can be compared for the models with and without the variable in question. If adjusted R^2 is higher without the variable and the variable is not statistically significant, then it should be excluded.
- (4) **Strong multicollinearity amongst the independent variables:** The case of significant correlation amongst the independent variables is not uncommon and creates problems when assessing model significance. This is the multicollinearity issue, discussed in Section [4.3.2](#).

The focus is on independent variables, but the choice of dependent variable can also be problematic. It is not always as simple as taking $y = Cost$. For example, volatilities in cost are often due to varying labor rates. In these cases, try $y = Labor\ Hours$ instead and convert the result to dollars using contractor-specific rates. As another example, consider a CER constructed using a quotient (e.g., dollars per pound) as the dependent variable. It generally works better to use the denominator of that quotient as an additional independent variable. If the assumption is cost relates to weight, use weight as an independent variable. Problems with trying to estimate the quotient include:

- it entails an estimating step that may be confused as data normalization (e.g., dividing all the y-variable costs by their respective weights)
- it fails to capture the relationship between weight and any other independent variables (e.g., multicollinearity)
- it locks the model into a linear relationship with the denominator variable when it may in reality be non-linear.

While it may be appropriate to continue use of the quotient, do so only after considering the aforementioned issues. Also keep in mind the same parametric techniques used to develop CERs can be useful in a wide range of applications including estimating schedule, risk, and even technical parameters.

4.4.1 Statistical Significance of CER

An analysis of variance (ANOVA) determines the overall significance of the model. **Table 27** demonstrates a sample ANOVA table produced by CO\$TAT for the example problem introduced with OLS in Section [3.3.1.3](#). The ANOVA table contains three distinct rows: Regression (or Model), Residual (or Error), and Total (or Corrected Total). An ANOVA table has the following columns: Degrees of Freedom (DF), Sum of Squares (SS), Mean Squares (MS), F-statistic, and the F-statistic p-value. The example below also contains a column “Prob Not Zero”, which is simply $1 - P$ -value.

Table 27: ANOVA Table – Section [3.3.1.3](#) Example

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	2	298146.5837	149073.2919	168.4858	0.0001	0.9999
Residual (Error)	4	3539.1306	884.7826			
Total	6	301685.7143				

As observed in **Table 27**, not every row has information in every column. There are no total MS, and only the Regression has an F-statistic. The values in the table have the following relationships:

$$\begin{aligned}
 DF_{total} &= DF_{regression} + DF_{residual} \\
 &= n - 1 \\
 SS_{total} &= SS_{regression} + SS_{residual} \\
 MS &= \frac{SS}{DF} \\
 F &= \frac{MS_{regression}}{MS_{residual}}
 \end{aligned}$$

The remainder of this handbook uses the following notational conventions:

$$\begin{aligned}
 SSR &= SS_{regression} = SS_{model} \\
 SSE &= SS_{residual} = SS_{error} \\
 SST &= SS_{total}
 \end{aligned}$$

And similarly,

$$\begin{aligned}
 MSR &= \frac{SSR}{df_{regression}} \\
 MSE &= \frac{SSE}{df_{error}}
 \end{aligned}$$

The F-statistic, and its corresponding p-value, is a measure of the overall model significance. Appendix [A.3.2.1 Hypothesis Testing](#) provides more information on hypothesis testing and significance levels.

The F-statistic tests whether or not all of the independent variable coefficients are statistically different than zero. Predetermine a significance level, or α , ahead of time. If the p-value is less than α , the hypothesis that all the variable coefficients are equal to zero is rejected in favor of the alternative that they are not all zero, and the model as a whole is deemed significant. A traditional value to use is $\alpha = 0.05$. In

disciplines with small samples, a lower level may be justifiable. It is important to specify the level ahead of time and stay consistent to maintain a valid model.

The F-test makes no statement on which variables are significant, only whether or not the overall model is significant. In this example, since $p\text{-value} = 0.0001 < \alpha = 0.05$, the conclusion is that there is sufficient evidence to suggest that all the coefficient estimates are not zero and that the overall model is significant.

4.4.1.1 Extension to Other Model Forms

The discussion of model significance focuses on the OLS model. However, these principles translate directly to more complex functional forms. The interpretations are nearly identical, with some mathematical subtleties. For the purposes of this guide, it is sufficient to understand that while the metrics may be calculated differently, statistical software packages output the material in a very similar fashion, regardless of method.

Always test for statistical significance in the fit space, not in the unit space. Statistical packages output an ANOVA table for all functional forms solved using the method of least squares (not including the [Generalized Linear Model \(GLM\)](#)). For linear forms such as [Generalized Least Squares \(GLS\)](#), [Transformable Linear and the Log-Linear Model](#), and [Ridge Regression](#), the calculations are exact and based on formulas very similar to – or even transformed to be equal to – those of OLS. However, the concept of sums of squares is one that applies strictly to the linear model.

Non-linear functions have deviances rather than the typical regression fit statistics. Further, the F-test is dependent on the normality assumption. [Non-linear Least Squares \(NLS\)](#) can produce results that are asymptotical and based on the convergence properties of normal theory (Appendix [A.3.3.1.2 Small Data Sets – Asymptotic Results](#)). While these asymptotic metrics are different from OLS, their analysis is identical but with less concrete interpretations.

The [Generalized Linear Model \(GLM\)](#) typically does not output such a table. The results rely on likelihood ratios, which follow the Chi-squared distribution. Asymptotic results, similar to NLS, can also be produced by relying on Wald inference. Despite these differences, there is an overall model test, often called out in the output by the software package. Compare this p-value against a significance level in the same way as the OLS F-test.

4.4.1.2 Addressing CER Significance

If the model as a whole is statistically significant, the next step is to proceed to Section [4.4.2 Validate Variable Set](#). If the F-test is not rejected ($p\text{-value} > \alpha$), then the overall model is not statistically significant. In this case, explore a new model form and/or consider a new set of predictors. The data may not support the hypothesized CER, as the predictors may not be cost drivers.

4.4.2 Validate Variable Set

The review of the results should also validate the individual variables used in the analysis. **Table 28** displays the standard output, in this case from COSTAT, for the OLS example introduced in Section [3.3.1.3 Multiple Linear Regression \(MLR\)](#). This table always has a row for each parameter in the model: the independent variables and the intercept term. The table always contains the following columns: Coefficient, Std Dev of the Coef (Standard Error), t-statistic, t-statistic p-value.

The example below also contains the fields “Beta Value” and “Prob Not Zero.” The “Beta Value” is the scaled and centered value of the coefficient (as in Section [3.3.6 Ridge Regression](#)). The “Prob Not Zero” is simply $1 - p$ -value. Not all statistical packages provide these two outputs and both can be ignored in this step.

Table 28: Coefficients Table – Multiple Linear Regression Example

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	37.3129	449.4459		0.0830	0.9378	0.0622
Power	28.2134	4.6985	0.9777	6.0047	0.0039	0.9961
Aper	6.1047	57.2542	0.0174	0.1066	0.9202	0.0798

The t-statistic is the coefficient estimate divided by its standard deviation and follows a Student’s t-distribution, with $n - p$ degrees of freedom. The t-statistics, and their corresponding p-values, measure the significance of the individual predictors in the model. Appendix [A.3.2.1 Hypothesis Testing](#) provides more information on hypothesis testing and significance levels.

The t-statistic tests whether an individual independent variable coefficient is statistically different than zero. Predetermine a significance level, or α , ahead of time. If the p-value is less than α , the hypothesis that the coefficient is equal to zero is rejected in favor of the alternative that it is not zero. A p-value greater than α suggests that the variable is not statistically significant and is a candidate for removal from the model. A traditional value to use is $\alpha = 0.05$. In disciplines with small samples, a lower level may be justifiable. It is important to specify the level ahead of time and stay consistent to maintain a valid model.

Due to correlations that may cause unexpected results when removing more than one variable, only remove one variable at a time. Removal of one variable may cause another previously insignificant variable to now become significant.

Each estimated coefficient should be statistically significant. If an insignificant variable is believed to be a “good variable”, consider the quality of the data sample or proceed back to Section [2.8 Hypothesize Functional Form](#) to alter the way the variable affects the dependent variable in the equation form. In this example, the p-value for *Aper* = .9202 > $\alpha = 0.05$ indicating the variable is not significant. The p-value for *Power* = 0.0039 < $\alpha = 0.05$ indicating *Power* is significant in the example model. These results also indicate *Aper* is a candidate for removal from the model, while *Power* should be kept in the model.

4.4.2.1 Intercept Term

In general, the intercept term is not tested for significance. Its inclusion (or exclusion) from the model is governed by physical properties of the system and not from the statistics, highlighting the difference between a mathematical model and a statistical/econometric regression equation⁶⁵. Mathematically, it is the CER result when all of the predictors are zero. While it is tempting to assign meaning to the intercept,

⁶⁵ Rao, Potluri, and Roger LeRoy Miller. Applied Econometrics. Belmont, CA: Wadsworth Pub., 1971. Print.

practically it is rarely interpretable. Additionally, negative intercepts are not of concern so long as the CER result is meaningful (generally the case when predicting away from the origin).

There are cases where it may make sense to test the intercept for significance. A zero-intercept model with one independent variable is known as a Factor CER. Factor CERs are common in modeling below-the-line WBS elements (e.g., training, testing or data as a factor of production costs).

- With small sample sizes, every degree of freedom is vital. With only 3 or 4 observations, the simple linear regression model has only 1 or 2 degrees of freedom. Removing the intercept from the model conserves an additional degree of freedom. If the intercept is close to zero and statistically insignificant, it may be advantageous to remove the intercept from the CER.
- When predicting values close to zero, the intercept gains more importance. When the data are far away from zero, a negative value is not of concern. The intercept is simply another degree of freedom helping “adjust” the model.
- When predicting small values close to zero, a negative intercept may produce unacceptable results. In these cases, if the negative intercept is statistically insignificant, it may be advisable to remove it from the model. If the intercept is negative and highly significant, there may be other misspecifications in the model, such as missing predictors, inconsistent normalization, or poor quality of data.
- When the intercept is negative, a reasonable sanity check is to test data over the range of interest, particularly on the side that results in smaller predicted values. For example, test values on the extreme low end of what may be used in the CER. If uncertain of these values, the minimum values for the sample dataset can serve as a good proxy. If these values result in negative or unrealistic outputs, then the negative intercept may need to be removed.

In general, include the intercept in the model and test significance, unless there is logic for a zero intercept. If the regressed intercept is positive and just marginally significant, then leave it in the model. If the intercept is near zero in value and statistically insignificant, and the sample size is small, then it may be appropriate to delete the intercept from the model to conserve degrees of freedom.

In some regressions, there may be a statistically significant negative intercept. If the data are far away from the origin, then it is okay to have a moderately negative intercept. However, if the data are very close to the origin and the predicted (negative) intercept is statistically very significant, then check the validity of the model to see if there are any misspecifications in the model form or any omitted explanatory driver variables. Additionally, the no-intercept equation might not necessarily go through the center of the dataset but will go through the origin.

4.4.2.2 Extension to Other Model Forms

The discussion of variable validation focused on the OLS model. However, these principles translate directly to more complex functional forms. The interpretations are nearly identical, with some mathematical subtleties. For the purposes of this guide, it is sufficient to understand that while the metrics may be calculated differently, statistical software packages output the material in a very similar fashion.

Statistical packages output a table of coefficients for all functional forms. For linear forms such as [Generalized Least Squares \(GLS\)](#), [Transformable Linear and the Log-Linear Model](#), and Ridge Regression, the calculations are exact and based on formulas very similar to – or even transformed to be

equal to – those of OLS. The [Generalized Linear Model \(GLM\)](#) uses likelihood ratio tests to produce tests analogous to the t-tests, but following the Chi-squared distribution. [Non-linear Least Squares \(NLS\)](#) produces asymptotic results based on the convergence properties of normal theory (Appendix [A.3.3.1.2 Small Data Sets – Asymptotic Results](#)). Presentation of these metrics is in a nearly identical format as the coefficients table for OLS. While these asymptotic metrics are different than OLS, their analysis is identical but with less concrete interpretations.

4.4.2.3 Addressing Driver Significance

If all the variables are significant the next step is to proceed to Section [4.5 Model Quality](#). If the t-tests are not rejected (*e. g.*, $p\text{-value} > \alpha$), then the respective variable is not statistically significant. In this case, explore a new model form and/or consider a new set of variables.

4.5 Model Quality

After validating the model assumptions, assessing high influence points and multicollinearity impacts, as well as confirming model and prediction coefficient significance, assess the model for fit and prediction quality. It is often instinct to choose the model that best fits the data. However, when constructing CERs, the real interest is usually prediction. A strong fitting model is great, but with poor prediction properties, that model is not optimal. The following sections examine several common metrics to assess both fit and prediction of the model.

4.5.1 Assess Metrics of Fit

This section describes the common metrics of fit, their use in assessing the OLS CER, and how they translate to other regression methodologies. A summary of some key SLR values and statistics for the Electronics data are provided in **Table 29**.

Table 29: Simple Linear Regression Formula Summary

Parameter	Formula	Result
Slope	$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$	27.3853
Intercept	$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$	92.9309
Standard Error of the Estimate	$S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$	\$42.2261
Standard Error of the Intercept	$S_y = S_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum(x_i - \bar{x})^2}}$	\$30.9235
Standard Error of the Slope	$S_b = \frac{S_e}{S_x \sqrt{n - 1}}$	\$2.0480
Standard Error Confidence Interval (CI)	$S_{yc} = S_e \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$	Power = 13.44 \$14.08
		Power = 26 \$29.31
Standard Error Prediction Interval (PI)	$S_{yp} = S_e \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$	Power = 13.44 \$44.51
		Power = 26 \$53.57

4.5.1.1 R-squared

The coefficient of determination (R^2) is the most commonly used metric and represents the percent of total variation in the response variable explained by the model. The formula is,

$$R^2 = 1 - \frac{SSE}{SST}$$

The R^2 metric is bounded by 0 and 1. A value of zero suggests no correlation, or that taking the mean of the response is the optimal fit to the data. A value of one suggests a perfect linear fit to the data.

Therefore, models with high R^2 values are desired. However, there are many cautions to consider with

this metric.⁶⁶ As more predictors are added to the model, R^2 cannot get smaller; it can only get larger. As a result, when using multiple predictors in the model, use the adjusted R^2 instead,

$$R_{adj}^2 = 1 - \frac{SSE/DF_{error}}{SST/DF_{total}}$$

This adjustment applies a penalty for adding variables to the equation. An additional predictor must be sufficiently valuable to outweigh this penalty in order for the R_{adj}^2 to go up.

While high R_{adj}^2 values are desired, there is no clear cutoff for what constitutes an acceptable threshold.

Different fields of study have different standards on how well a model should fit, and models with low R_{adj}^2 values can still prove useful when no other alternative exists. Further, the R_{adj}^2 value is highly dependent on scale and cannot be compared across models with different forms. As a result, R_{adj}^2 is most useful for comparing models built on the same response, y , with different predictors, x_i , in a linear setting. In this sense, the R_{adj}^2 can be used as a selection criteria to determine which subset of predictors produces the best fitting model.

Never use R^2 or R_{adj}^2 as the sole indicator of model quality or acceptability.

4.5.1.2 Extension of R^2 to Other Model Forms

The R^2 statistic is difficult to interpret when translated to other forms. As previously mentioned, it is highly dependent on scale. The definition of sums of squares is one unique to the linear model. They do not exist in the non-linear setting. The model has deviances instead, which are conceptually similar. In these settings, the R^2 can be calculated in unit space as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

And,

$$R_{adj}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / DF_{error}}{\sum_{i=1}^n (y_i - \bar{y})^2 / DF_{total}}$$

The following are notes on R^2 for the other regression methods focused on in this guide:

⁶⁶ Kvalseth, Tarald O. "Cautionary Note about R^2 ." *The American Statistician* 39.4 (1985): 279.

(1) Weighted Least Squares (WLS)

The R^2 produced for WLS by software is in the weighted space. As a result, it is often very high (i.e., closer to 1). This metric is not useful to assess how well the model fits. It is not comparable to OLS, and not even to other WLS models with different weighting schemes. To derive a useful metric, the R^2 must be calculated in unit space. This metric can be used to compare between the OLS model and the WLS models. However, the OLS model will always have the highest unit space R^2 . Further, the WLS model with the highest R^2 is not always optimal. Other metrics, such as the coefficient standard errors (Section [4.5.1.6 Coefficient Standard Errors](#)) should be considered when choosing the optimal weights, \mathbf{w} .

(2) Transformable Linear and the Log-Linear Model

The R^2 produced by the Log-Linear model is in the log transformed space. As a result, it is a measure on how well the transformed data fits. This transformed model may not be the optimal fit, and may even result in a negative R^2 in unit space. In these cases, the regression is a worse predictor than taking the average of the data, which has $R^2 = 0$. Further, the unit space R^2 is not comparable to the OLS R^2 .

(3) Generalized Linear Model (GLM) and Non-linear Least Squares (NLS)

The R^2 is primarily a concept for linear regression and is not comparable back and forth between different functional forms.

4.5.1.3 Root Mean Squared Error (RMSE)

The RMSE is the estimate of the standard deviation of the error term, σ , where $\hat{\sigma}^2 = MSE$ and $\hat{\sigma} = RMSE$. When examining RMSE, lower values are highly desired. RMSE is in the formula to calculate coefficient ($\hat{\beta}$) standard errors, lower values can lead to higher chances of statistically significant predictors, as well as narrower confidence intervals. Another name for the RMSE is the Standard Error of the Estimate (SEE).

Much like R^2 , it is important to distinguish between the fit space and unit space calculation of RMSE. Calculate the RMSE in unit space as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{DF_{error}}}$$

4.5.1.4 Standard Percent Error (SPE)

The standard error of estimate for regression methods using a multiplicative error term has a different formulation than that derived for OLS. This formulation is referred to as the standard percent error (SPE), and can be derived using the following formulation:

$$SPE = \sqrt{\frac{\sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{\hat{y}_i}\right)^2}{GDF}}$$

Where,

y_i is the actual observation in unit space

\hat{y}_i is its predicted value for the i th data point (for $i = 1, \dots, n$)

n is the total number of observations

GDF = $n - p$ for MUPE; GDF = $n - p - 1$ for regression methods that include a constraint, e.g., ZMPE (except for factor CERS)

p is the total number of estimated parameters

SPE is used to measure the multiplicative error for the model's overall estimation of error. It is used to define the uncertainty of CER results for methods such as MUPE and ZMPE.

4.5.1.5 Extension of RMSE to Other Model Forms

Root Mean Squared Error is applicable to all of the functional forms. Its calculation is the same for linear and non-linear functional forms fit by least squares methodologies. One difference between methodologies is that the GLM model uses maximum likelihood to estimate model parameters. MLE estimates the dispersion parameter (e.g., $\hat{\sigma}^2$ for normal error) for the model rather than using the least squares MSE calculation. RMSE can be used to compare models of different forms (such as [Ordinary Least Squares \(OLS\)](#), [Transformable Linear and the Log-Linear Model](#), and [Non-linear Least Squares \(NLS\)](#)), but this should be done using the same set of independent variables each time.

4.5.1.6 Coefficient Standard Errors

[Table 28: Coefficients Table – Multiple Linear Regression Example](#) contains the coefficient standard errors. Under the normality and linearity assumptions, not only is the response, \mathbf{y} , normally distributed, but the parameter coefficients, β_j , are as well.

Therefore, the smaller the coefficient standard errors, the smaller the error term of the normal distribution surrounding each coefficient. This is an important concept because confidence and prediction intervals (discussed under [Step 5: Characterize Uncertainty](#)) are a function of the covariance matrix. The $(k + 1) \times (k + 1)$ covariance matrix for $\hat{\beta}$ is,

$$\begin{aligned} \text{Var}(\hat{\beta}) &= \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1} \\ &= \text{RMSE} \cdot (\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

The variance, or the coefficient standard error, surrounding each $\hat{\beta}_j$ is the j^{th} diagonal of $\text{Var}(\hat{\beta})$. The smaller the RMSE, the more desirable the model. The equation above shows how the variance of the coefficients is a multiple of the RMSE. The second half of the equation displays the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix. The correlation matrix is defined as $\mathbf{X}^{*'}\mathbf{X}^*$, where \mathbf{X}^* is the \mathbf{X} matrix scaled to have variance of one and is unitized. The inverse of this matrix, $(\mathbf{X}^{*'}\mathbf{X}^*)^{-1}$, is how the VIFs were calculated. The $(\mathbf{X}'\mathbf{X})^{-1}$ matrix is similar, but in the scale of the raw data rather than unitized. Large VIFs result in large diagonals that are in $(\mathbf{X}'\mathbf{X})^{-1}$ relation to the size of the coefficient estimate.

4.5.1.7 Extension of Coefficient SE to Other Model Forms

Models with small coefficient standard errors are preferred. Comparison across model forms is possible, but may not be the best metric to use when trying to decide between multiple functional forms, such as linear versus non-linear. Examining the standard errors is very useful when employing [Weighted Least Squares \(WLS\)](#). In this setting, the selected set of weights should be the one, which has minimum error

around the coefficient estimates. Provided the R^2 and RMSE are not significantly worse than the other weighting options, this can be a good selection strategy for deciding between sets of weights.

4.5.1.8 Predicted versus Actuals Plot

Figure 70 shows the actual observed costs and the predicted costs for the Section [3.3.1.3](#) OLS Example. It is the visual display of the dependent variable actual observations versus the dependent variable estimated values. A perfect CER has all observations on the 45-degree line through the origin ($y = x$, or Predicted = Actual).

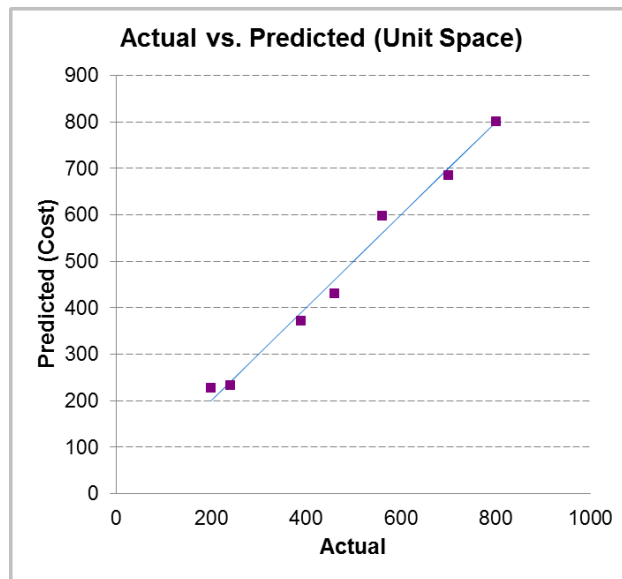


Figure 70: Predicted versus Actual Plot

Figure 70 illustrates a strong regression fit as indicated by the proximity of the observed points to the predicted line, highlighted in blue. This plot also enables a visual check for the regression model's accuracy when compared to the actual response variable values. This graph provides an optimal visualization technique for the results of the regression, which is particularly important in the case of two or more independent variables, where the regression itself may be difficult to visualize.

4.5.1.9 Bayesian and Akaike Information Criterion

Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are two likelihood-based metrics. These metrics are available from most statistical packages and are relative metrics. AIC and BIC by themselves provide no information on how well the model fits. However, they can be used as a comparison between different model forms and aid in the selection of a functional form, provided that the dependent variable, y , has the same numerical value for each model (e.g., not comparing y to $\log y$). These metrics are not OLS specific and provide a good way to compare different CER forms. In terms of model selection, lower AIC and BIC values are preferred.

AIC is often used due to its practical, probabilistic interpretation. The relative probability of $model_i$ minimizing the loss is,

$$e^{\frac{\min AIC - AIC_i}{2}}$$

For example, suppose there are three models with the following AIC values: 51, 53, and 59. **Table 30** can be produced using the above formula:

Table 30: AIC Relative Probability Example

Model	AIC	Relative Probability
1	51	$e^{\frac{51-51}{2}} = 1.0000$
2	53	$e^{\frac{51-53}{2}} = 0.3679$
3	59	$e^{\frac{51-59}{2}} = 0.0183$

In this example, Model 2 has a 36.79% probability of being “optimal” relative to Model 1, and Model 3 has only a 1.83% probability.

The Joint Cost and Schedule Risk and Uncertainty Handbook (CSRUH) contains more information on AIC and BIC.

4.5.1.10 Extension of BIC and AIC to Other Model Forms

Both BIC and AIC are applicable to all functional forms. Use them to compare models of different forms (such as [Ordinary Least Squares \(OLS\)](#), [Transformable Linear and the Log-Linear Model](#), and [Non-linear Least Squares \(NLS\)](#)).

4.5.1.11 Mallows' C_p

The Mallows' C_p statistic is an estimate of the mean squared prediction error for the OLS model. The metric accounts for both variance and bias within the estimator, and adjusts for the impact of adding more than necessary predictors to the model. Much like AIC and BIC, Mallows' C_p is not used on its own but can be very useful when comparing multiple models to each other.

In linear regression with a normally distributed error term, Mallows' C_p is equivalent to AIC. The formula for Mallows' C_p is,

$$C_p = \frac{SSE}{MSE} - n + 2p$$

Similar to Root Mean Squared Error (RMSE), models with smaller values of C_p are preferred.

4.5.1.12 Extension of Mallows' C_p to Other Model Forms

Mallows' C_p is only used for assessing linear regression models (discussed in further detail under Section [3.3.1 Ordinary Least Squares \(OLS\)](#), Section [3.3.2.2 Weighted Least Squares \(WLS\)](#)). For other model forms such as [Non-linear Least Squares \(NLS\)](#), use alternative but similar metrics, such as AIC.

4.5.2 Assess Metrics of Prediction

A popular method to assess the prediction ability of a model is Cross Validation and the PRESS statistic. The following sections discuss these concepts and their use in assessing prediction ability of a CER in the context of OLS, and then how that translates to other regression methodologies.

Cross validation is a methodology in which a subset of the observations is held out in a “test” dataset, while the model is fit using the remaining “training” dataset. Cross validation can serve multiple purposes. A primary objective is to prevent model over-fitting. Additionally, it can be useful in identifying highly influential points on the model.

4.5.2.1 Cross Validation (k-Fold)

One way to conduct cross validation is called *k*-fold cross validation. The variable *k* refers to the number of subsets formed from partitioning the data set. For example, 10-fold cross validation splits the data into 10 roughly equal subsets. First, fit the model using nine of these subsets. Then, test the model using the remaining subset and record a measure of prediction or accuracy, usually Mean Squared Error. Repeat this process, predicting each of the 10 subsets based on the remaining nine. Two popular selections for *k* are 5 and 10, since they represent 80/20 and 90/10 splits on the data, respectively.

This process creates 10 different MSE statistics, all which should be similar to each other. If not, that is an indication that one of the subsets performed different when predicting the response.

4.5.2.2 Cross Validation (Leave-One-Out)

A more extreme example of cross validation is leave-one-out. In this scenario, use all but one data point to fit the model, and calculate model metrics for the single “left out” point. Repeat this process for all observations in the model. This methodology is discussed in Section [4.3.1.4 Leave-One-Out Metrics](#). In general, a good way to use this information is to examine which data observations have leave-one-out results greatly different from the rest. This can signal observations that significantly impact the model’s performance and may help identify potential leverage points and outliers.

Cross Validation is convenient due to its applicability to all of the functional forms. Its calculation is the same for linear and non-linear functional forms. Use the resulting metric from Cross Validation to compare models of different forms (such as [Ordinary Least Squares \(OLS\)](#), [Transformable Linear and the Log-Linear Model](#), and [Non-linear Least Squares \(NLS\)](#)), and across different subsets of predictors.

4.5.2.3 Predicted Residual Sum of Squares (PRESS)

The PRESS statistic is an aggregated metric from Leave-One-Out cross validation. The sum of the squared errors for each predicted point is the Predicted Residual Sum of Squares (PRESS). PRESS is a popular statistic used to assess prediction ability of the model. Its interpretation is the same as SSE, but provides a better view on how well the model predicts. A further convenience is that PRESS can be calculated with a relatively simple formula. There is no need to run all the *n* regressions, making it convenient from a computational standpoint. Many statistical packages including SAS JMP, Minitab, and R return the PRESS statistic. Smaller values indicate a lower prediction error and are desirable.

A simple way to calculate PRESS for OLS is to first calculate the PRESS residual, $e_{i,PRESS}$ for each observation.

$$e_{i,PRESS} = \frac{y_i - \hat{y}_i}{1 - h_{ii}}$$

The value h_{ii} is the leverage value of observation *i*, or the *i*th diagonal entry of the predicted matrix, *H*. Note the similarities with the internally and externally studentized residuals.

Once calculated, PRESS is simply the sum of squares of the individual PRESS residuals.

$$PRESS = \sum_{i=1}^n e_{i,PRESS}^2$$

The Outlier Analysis Table from CO\$TAT in [Table 23](#) again provides the necessary information to calculate PRESS for the example problem introduced with OLS in Section [3.3.1.2](#). In order to calculate the PRESS residual, the following terms are required:

$$\begin{aligned} y_i &= \text{Cost} \\ \hat{y}_i &= \text{Predicted Y Value} \\ h_{ii} &= \text{Leverage} \end{aligned}$$

For example, the first observation can be derived using the following statement:

$$\begin{aligned} e_{1,PRESS} &= \frac{390 - 372.5577}{1 - 0.2093} \\ &= 22.0593 \end{aligned}$$

While the CO\$TAT result shows all 9 observations from the original dataset, keep in mind that Observations 4 and 9 are not actually used in this CER since they have no data for Aperture (see data in [Table 6](#)). In this example, $n = 7$ due to two observations that are excluded. After carrying out this calculation for each observation, the resulting value is⁶⁷:

$$\begin{aligned} PRESS &= \sum_{i=1}^7 e_{i,PRESS}^2 \\ &= 22.0593^2 + \dots + (-3.8383)^2 \\ &= 12,439.1171 \end{aligned}$$

PRESS is convenient due to its applicability to all of the functional forms. Its calculation is the same for linear and non-linear functional forms. Use PRESS to compare models of different forms (such as [Ordinary Least Squares \(OLS\)](#), [Transformable Linear and the Log-Linear Model](#), and [Non-linear Least Squares \(NLS\)](#)), and across different subsets of predictors.

4.5.2.4 Predicted R²

PRESS is an absolute measure. Using this measure alone, analysts do not really know how robust the regression model is. The Predicted R² statistic is defined to be the fraction of the variation in the dependent variable explained by the “leave-one-out” model, which puts PRESS in perspective:

⁶⁷ Allen, David M. “The Relationship between Variable Selection and Data Augmentation and a Method for Prediction.” *Technometrics*, vol. 16, no. 1, 1974, pp. 125–127. www.jstor.org/stable/1267500

$$\text{Predicted } R^2 = 1 - \text{PRESS}/\text{SST}$$

SST is the total sum of squares about the dependent variable. The Predicted R^2 values cannot be compared across model forms because SST is different for the OLS and LOLS model.

The Predicted R^2 statistic is commonly used to determine how well the model predicts for new observations, but it would be unusual for it to give a different answer than PRESS. If Predicted R^2 is substantially lower than R_{adj}^2 that is an indication that the model is inflated by over-fitting. Predicted R^2 can be more useful than R_{adj}^2 in measuring the predictive power of the model because it is calculated using observations not included in the model. PRESS and/or Predicted R^2 are recommended calculations for assessing any regression equation.

4.6 Model Selection

The preceding sections of [Step 4: Validate CER](#) introduce the basic toolset for validating and assessing a regression model. After the iterative process of checking assumptions, running model diagnostics, checking statistical significance, and assessing model quality, a statistically sound CER is hopefully achieved. However, it is possible that multiple CERs appear valid and the question becomes: which one should be used? This leads to the topic of Model Selection, which is further broken out into the following two questions:

- (1) Within a singular functional form, what variable set should be included in the model?
- (2) Given two or more valid models of different functional form, which one is better?

The remaining parts of this section study both of these cases using the previously introduced tools and concepts. Section [4.6.1 Variable Selection](#) addresses question (1) and Section [4.6.2 Functional Form Selection](#) addresses question (2).

Keep in mind that the statistics and numerical metrics are there to guide the model selection. Make sure to compare each potential model with the original hypothesis as developed in Section [1.3.4](#). In an ideal world, the data and statistical methods will support the hypothesized CER. When this is the case, the process of model selection is easy. When this is not the case, it is important to consider and try to understand why the hypothesis and statistical results are in conflict. It could be that the data are simply poor and not an accurate representation of reality.

The solution may be a subtle change in the formulation or addressing a mathematical nuance. Perhaps multicollinearity is creating noise or maybe the selected regression method is not the best way to fit the hypothesized CER. Another possibility is that the data are now providing new insights that may cause an update to the original hypothesis. Be sure to consider all the information gathered along the way towards constructing each CER and make sure that candidate CERs are not simply arbitrary mathematical models with seemingly “good” statistics.

While many statistical tools are available to inform the model decision making process, the selected model must be logical, comprehensible, and justifiable. When considering multiple models with similar fit statistics, the model with the most logical and simplistic form is preferred.

4.6.1 Variable Selection

Variable selection within a model is a heavily studied area of statistics. Suppose a given regression methodology and model functional form are to be used to model a response y based on k potential predictors, x_1, \dots, x_k . What subset of these predictors produces the “optimal” model? While the term optimal is subjective, there are standard metrics and practices to help reach an agreeable, defensible solution. This section introduces several common approaches to variable selection when creating a CER.

There are many factors to consider when determining if a predictor belongs in the model or not. The predictor must make sense when considering the physics of the chosen variables. Correlation does not imply causation, it takes thought and understanding beyond what the statistics can provide to draw this line. [1.3 Cost Estimate Purpose and Scope](#) discusses this concept in more detail.

Correlation does not imply causation.

It is also important that the subset of independent variables create a model, which satisfies all assumptions and is statistically valid. This means that the model satisfies the criteria put forth in Sections [4.2 Model Assumptions](#), [4.3 Model Diagnostics](#), and [4.4 Model Significance](#). Note that the results from Section [4.4 Model Significance](#) are binary: statistically significant, or not statistically significant.

The statistical tests, such as the overall model F-test (Section [4.4.1](#)), only suggest if the model is statistically significant at the predetermined significance (α) level. Do not compare p-values across models. Comparing p-values between models and concluding that one model is more statistically significant than the other is an incorrect and a false interpretation. Additionally, adjusting the α level after the fact is not an acceptable practice and invalidates the analysis.

Finally, the subset of independent variables should create a model that performs well for both metrics of fit and prediction, as covered in Section [4.5 Model Quality](#). It is not wise to select one criteria, such as R_{adj}^2 , and blindly choose the variable subset with the highest/lowest value. Be sure to examine multiple criteria and select a model that performs highly across all relevant criteria.

4.6.1.1 Common Selection Strategies

There are several traditional, iterative approaches utilized to search for a model with some optimal criteria. The simplest method adds and/or removes independent variables based on significance tests (Section [4.4.2 Validate Variable Set](#)).

Forward selection begins with no variables in the model. At each step, the variable with the lowest (closest to zero) statistically *significant* p-value is entered into the model. The process continues until there are no more significant predictors to add.

Backward selection begins with all variables in the model. At each step, the variable with the highest (closest to one) statistically *insignificant* p-value is removed from the model. The process continues until all predictors in the model are significant.

While not common in cost analysis, backwards selection is a reasonable starting approach when there are many potential independent variables. It can be used to reduce many (e.g., 30 or more) predictors down to a more reasonable amount (e.g., 10).

Stepwise selection⁶⁸ begins as either forward selection or backwards selection (though starting from backwards is more common). With stepwise selection, the model can either add or remove a variable at each step. The process continues until all predictors in the model are significant and all predictors not in the model are insignificant. Stepwise selection is usually the default between the choice of forward, backward, and stepwise. The following two paragraphs describe how this selection process works.

– *Terminology* –

In regression analysis there is an unknown, true theoretical regression equation made up of both an actual underlying relationship and random noise. This is shown by the generic form $y = f(X; \beta) + \varepsilon$.

Over-fitting is when the model is describing noise (ε) and occurs when the model is more complex than necessary (e.g., has too many predictors).

For forward, backward, and stepwise selection, a p-value is selected at a predetermined significance (α) level as a criteria for which a variable will enter/exit the model. This p-value is usually set higher than the traditional $\alpha = 0.05$, usually around $\alpha = 0.10$ or $\alpha = 0.15$ to enter the model, and as high as $\alpha = 0.20$ to exit the model.

In addition to using the F-statistic as the entry/exit criteria, all of the above methodologies can be modified to add/remove variables based on other fit/prediction criteria (Section [4.5 Model Quality](#)). A popular, and more technically sound, approach to forward, backward, and stepwise selection is to use Mallows' C_p as the entry/exit criteria into the model which helps hedge against overfitting concerns.

All-subsets selection is the “brute force” approach of examining all possible subsets of the model. Construct a model for each possible combination of variables and sort them by some selection criteria. While computationally intensive, modern algorithms and improved computing speed now provide a more practical approach for as many as $p = 30$ independent variables. Use this method to create a short list of candidate models for more rigorous manual examination.

Do not use R^2 or R^2_{Adj} as the selection criteria under an all-subsets approach. Doing so will almost certainly result in overfitting (i.e., selecting a CER with too many independent variables). Be sure to consider multiple metrics simultaneously.

Although popular, none of the aforementioned selection methodologies are without their flaws. Recalling Section [3.3.1 Ordinary Least Squares \(OLS\)](#), the OLS model produces the Best Linear Unbiased Estimator (BLUE). However, this does not account for the selection procedure. Running many models in a forward, backward, or stepwise selection procedure introduces bias and inflates the variance of the result. The resulting statistics from the selected model, including t-statistics and coefficient standard errors, are invalid since they do not account for the selection procedure. The test distributions no longer

⁶⁸ CO\$TAT includes an “analyst in the loop” feature called Stepwise Analysis that helps identify the next step to improving a cost estimating relationship. This is not the same as Stepwise Regression. More details on this feature is found in the CO\$TAT help file.

hold, and the errors are understated. Additionally, there is no guarantee that the procedure will find the model with the “optimal” metric of interest (e.g., the highest R^2).

The all-subsets approach should generally be preferred over stepwise regression. The results provide more details and the ability to choose the “best” model based upon multiple sources of information on a set of candidate models, including subject matter expert knowledge of the variable set. For example, suppose the all-subsets approach is used and two models, Model A and Model B, rank very favorably across R^2_{adj} , Root Mean Squared Error (RMSE), AIC, BIC, PRESS, and C_p .

4.6.1.2 Conclusion

While the processes, tools, and methodologies discussed in Section 4.0 are powerful, it is important for analysts to understand the mathematical relationships and outputs specific to a given a model. **Table 31** summarizes the metrics introduced in Section [4.5 Model Quality](#) and their use in variable selection. A detailed analytical comparison of variable subsets examines R^2_{adj} , RMSE (or MSE), PRESS, AIC (or BIC), and C_p .

Every analysis does not need to use every metric, but using multiple metrics and selecting a logical model, which performs favorably across all of them, is a sound strategy for performing variable selection. Note that at this stage all candidate models have passed the statistical significance requirements for both the F and t statistics (Section [4.4 Model Significance](#)). Do not use these statistics to compare between models.

Table 31: Variable Selection Metrics

Section	Metric	Variable Selection Application	Excel and CO\$TAT ⁶⁹
4.5.1.1	R^2 and R^2_{adj}	As variables are added to the model, R^2 always increases. Do not use R^2 for variable selection. The R^2_{adj} statistic adjusts for the number of predictors in the model and can decrease if an insignificant predictor is added. The R^2_{adj} statistic can be useful for comparing between different variable sets within a functional form, but should not be used as the sole metric. Higher values for R^2_{adj} suggest a better model fit.	Yes
4.5.1.3	Root Mean Squared Error (RMSE)	RMSE (or just MSE) is an estimate of the model standard deviation (or variance). It can be used to compare between different variable sets within a functional form, with lower	Yes

⁶⁹ Indicates if the metric is readily accessible from Excel and CO\$TAT 7.4. Minitab, SAS, SAS JMP, STATA, and R support all these metrics.

		values suggesting a better model. The RMSE should not be used as the sole selection metric.	
4.5.1.6	Coefficient Standard Errors	Coefficient standard errors can be difficult to compare across multiple variable sets. There are multiple values to consider, so while they can be valuable, they are not traditionally used in model selection. Lower standard error values suggest less variation around the estimates and are preferred.	Yes
4.5.1.8	Predicted versus Actuals Plot	Predicted versus Actuals Plots are useful to assess a single model, but are not an ideal tool for comparing across models with multiple variable sets. They provide no insight to the inclusion of insignificant variables and are not traditionally used in the variable selection process.	Yes
4.5.1.9	BIC and AIC	BIC and AIC are likelihood-based metrics for quality of the model. Both are used to compare between different variable sets within a functional form, with lower values suggesting a better model. AIC is one of the more popular metrics to use when selecting a model. While it is best to compare multiple metrics to select the optimal model, if a sole metric were to be used, it most commonly would be AIC.	No
4.5.1.11	Mallows' C_p	Mallows' C_p is a traditional statistic for comparing between different variable sets within a linear model. Lower values suggest a better model.	No ⁷⁰
4.5.2.1	Cross Validation and PRESS	Cross Validation statistics and the PRESS statistic are comparable across different variable sets within a functional form. It is common to use results from the Cross Validation to compare models. It is most common to report the PRESS statistic to compare between variable subsets. Like RMSE, lower PRESS values suggest a better model.	No

A common strategy to compare models is to create a table of candidate models, such as **Table 32**. To create the table, first run the all-subsets regression. Sort the results on each of the metrics of interest, in this case R_{adj}^2 , RMSE, AIC, C_p , and PRESS. Select the top ten performing models for each for further comparison. There will be much overlap of candidate models, so select models that rank in the top ten for all of the metrics as the final candidates. This will usually result in around five to eight models, but it varies depending on the problem. Finally, compare the short list of models. By examining the metrics, a few models typically stand out from the rest as being more statistically favorable. From there, apply subject matter expertise to make the final selection. Models with fewer predictors or containing predictors deemed more important by logic should be heavily favored.

⁷⁰ Mallows' C_p is not readily provided by Excel nor CO\$TAT, but can easily be calculated using the formula provided in Section [4.5.1.11](#).

Table 32 displays an arbitrary table for the sole purpose of illustrating the concept. From the statistics, it is clear that models 1, 2, and 3 perform better across all metrics compared to models 4 and 5. Models 1 and 2 are simpler than model 3, which is preferred (i.e., Models 1 and 2 include one less variable).

There will always be judgment involved in model selection, but the creation of a short list of candidate models can greatly assist in the process of selecting a variable set.

Table 32: Variable Selection Table

Model	Predictors	R^2_{adj} (rank)	RMSE (rank)	AIC (rank)	C_p (rank)	PRESS (rank)
1	x1 x2 x4 x6 x7	0.8264 (1)	29.3 (2)	103.9 (2)	5.26 (2)	6648 (3)
2	x1 x2 x5 x7 x8	0.8205 (2)	29.7 (3)	103.2 (1)	4.35 (1)	6596 (2)
3	x1 x2 x4 x6 x7 x8	0.8186 (3)	28.4 (1)	106.3 (3)	6.92 (3)	6528 (1)
4	x1 x3 x4 x5 x6	0.8172 (4)	32.6 (5)	110.9 (4)	7.04 (4)	6692 (4)
5	x1 x2 x4 x5	0.8103 (5)	30.1 (4)	112.6 (5)	7.89 (5)	6795 (5)

4.6.2 Functional Form Selection

Selecting between multiple valid models of different functional forms can be a difficult task. It is not common to have multiple forms all satisfying their respective sets of assumptions. However, the problem does arise and it is often very difficult to select which form is the “correct” one. In the context of functional form selection, a valid model is one, which satisfies requirements set forth in [Section 4.2 Model Assumptions](#) and in [Section 4.4 Model Significance](#). Conclusions drawn in [Section 4.3 Model Diagnostics](#) result in actions affecting all candidate models, such as removing a point or removing a collinear variable. Again, it is worth noting that the [Section 4.4 Model Significance](#) cannot be compared between models.

While advanced modern selection algorithms exist, there are no traditional automatic model selection methodologies used to choose between candidate models. The following are several rules-of-thumb to follow when deciding between multiple functional forms.

- (1) **Does the functional form make sense?** With computer regression software, it is easy to develop a model with many variables in complex relationships that fits the data very well. Such a model may predict future observations poorly due to overfitting, loss of degrees of freedom, etc.
- (2) **How well do the assumptions fit?** As covered in [Section 6.2](#), validation of model assumptions can be subjective. There can be different degrees of how well the assumptions fit. For example, a linear model may have a residual plot that is not ideal, but not enough to completely reject the linearity assumption ([Section 4.2.1.5 Linearity](#)). However, the residual plot may look much better and fully satisfy the assumption when using a non-linear form. This may be reason enough to prefer the non-linear form, since this form generates improved fit statistics than compared to the linear form.
- (3) **How do the models compare across the relevant metrics?** Not all of the metrics presented in [Section 4.5 Model Quality](#) are valid for comparison across functional forms. However, there are some statistics that can be used for comparison purposes (e.g., RMSE, SSE, etc.).

Table 33 summarizes the metrics introduced in [Section 4.5 Model Quality](#) and their use in functional form selection. An analytical comparison of functional forms examines far fewer metrics than comparing

different variable subsets. Every analysis does not need to use every metric, but using multiple metrics and selecting a logical model, is a sound strategy for performing variable selection.

It is possible and there are cases where there is simply no model that provides a reasonable result. When the possibilities have been exhausted, there are times when a simple average (i.e., the analogy approach) may be the best choice moving forward (Section [2.3.4](#)).

Table 33: Functional Form Selection Metrics

Section	Metric	Functional Form Selection Application	Excel and CO\$TAT ⁶⁷
4.5.1.1	R^2 and R_{adj}^2	Neither R^2 nor R_{adj}^2 should be used to compare between models of different functional forms.	Yes
4.5.1.3	Root Mean Squared Error (RMSE)	RMSE (or just MSE) is an estimate of the model variance. It can be used to compare between functional forms, however each form must contain the same variable set. Additionally, the unit space metric should be used. Lower values suggest a better model, and the RMSE should not be used as the sole selection metric.	Yes
4.5.1.6	Coefficient Standard Errors	Coefficient standard errors can be difficult to compare across many models. Since there are usually several valid functional forms to choose from, comparison of standard errors is often more practical for functional form selection as opposed to variable selection. Lower standard error values suggest less variation around the estimates and are preferred.	Yes
4.5.1.8	Predicted versus Actuals Plot	Predicted versus Actuals Plots are useful to visualize the fit of a model. If one functional form predicts the in-sample data more accurately than another form, that may indicate the more appropriate model.	Yes
4.5.1.9	BIC and AIC	BIC and AIC are likelihood based metrics accounting for fit of the model. They are both used to compare different functional forms, with lower values suggesting a better model. However, all models should use the same dependent variable (e.g., one model cannot use y while another uses $\log y$) AIC is one of the more popular metrics to use when selecting a model. While it is best to compare multiple metrics to select the optimal model, if a sole metric were to be used, it would most commonly be AIC.	No
4.5.1.11	Mallows' C_p	Mallows' C_p is a traditional statistic for comparing between variable sets within a linear model. Since it only applies to the linear model, it is not applicable for comparing between functional forms	No
4.5.2.1	Cross Validation and PRESS	Cross Validation statistics and the PRESS statistic are comparable across different functional forms. It is common to use results from the Cross Validation to compare models. The use of k-fold Cross Validation is a common way to compare across functional forms, especially with complex models.	No

In summary, RMSE, Predicted versus Actuals plots, AIC (or BIC), Coefficient Standard Errors, and Cross Validation (including PRESS) can be used as statistical metrics to select between multiple model forms.

Create a table, much like [Table 32](#), to facilitate the comparisons. However, statistical comparisons are only valid if (1) the model makes sense and (2) the assumptions are satisfied.

4.7 CER Responsiveness

Part of the identification of cost drivers is understanding the variable sets believed to be important to the equation (Section [2.7:Identify Potential Variable Sets](#)). A perfect data set is almost never possible. More often than not, analogies are not perfect, and the data are more dissimilar than they are similar from a system-level perspective. For example, when trying to evaluate the cost for a propulsion system, the same engine type may be in a variety of different systems or platforms with very little in common other than propulsion requirements. Treating the data to make it more similar by using categorical variables, or excluding observations in particular relationships is often necessary. That does not mean the data are bad, they just may only be relevant to a small subset of the CERs developed to populate a WBS for a given system.

Therefore, the analysis must be qualified. Clearly articulate the range or system properties associated with the analysis to qualify its limitations. Assessing and documenting a CER's responsiveness over a particular scope of variable parameters will allow an understanding of the CER's limitations and strengths.

Sensitivity Analysis, like most of the previous discussion, is an iterative process conducted while evaluating the CER. To conduct a sensitivity sanity check, evaluate the set of independent variables and determine a reasonable range over which they should vary. At a minimum, evaluate the model over the range of data for which the CER is based, but it can be desirable to extrapolate beyond the range of the data. If the estimated parameter value for the new system being estimated falls outside the historical range, the evaluated sensitivity range must account for adequate uncertainty.

Once these ranges are determined, vary each independent variable across its range, holding the other variables constant at their planned values. Depict the resulting variation in cost in a tornado chart.

Consider referring back to Section [2.4 Univariate Data Analysis](#) to see the influence of a single variable and its impact on the CER.

CER inputs are rarely are precise. For example, weights are usually engineering estimates and SLOC counts are usually assessments from software engineers. The total uncertainty in the estimate includes not only the statistical uncertainty in the CER itself, but the uncertainty in the independent variables that are inputs to the CER.

5.0 STEP 5: CHARACTERIZE UNCERTAINTY

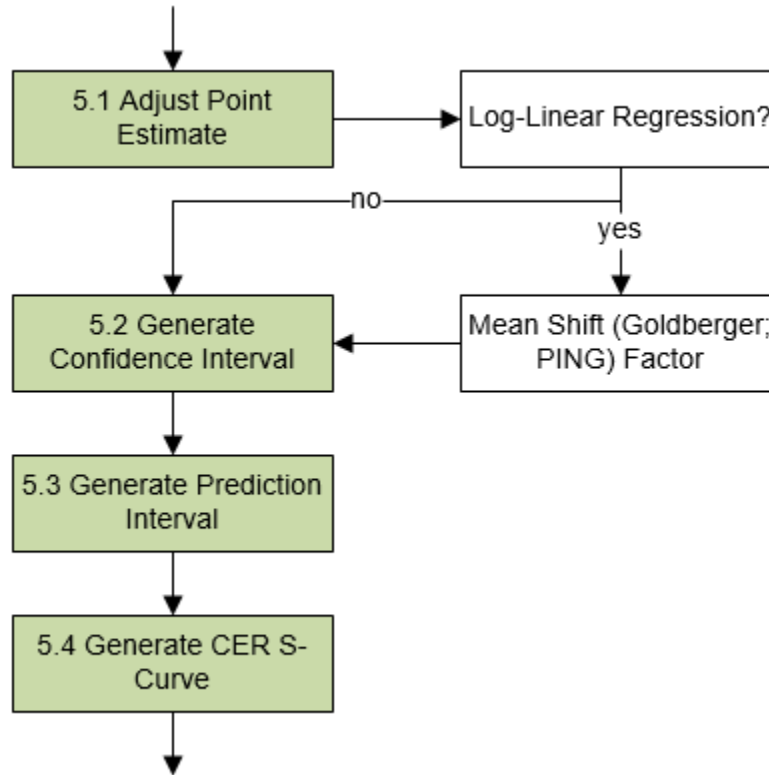


Figure 71: Step 5: Characterize Uncertainty

Risk and Uncertainty Analysis are vital parts of cost estimating, usually conducted at the Cost Model level. Understanding the risk and uncertainty associated with a CER is crucial to accurate implementation. The Mean Squared Error (MSE) specifies the overall error in the model. The focus is typically on prediction, and there is a high interest in the MSE of the predicted values, \hat{y} . This value can be derived using the following statement,

$$MSE(\hat{y}) = Var(\hat{y}) + bias^2(\hat{y})$$

The total error in the prediction is the variance of the prediction, plus the squared bias of the prediction. When possible, it is preferred to use a methodology where $bias^2(\hat{y}) = 0$. However, it may be advantageous to introduce bias if it reduces the variance enough, and therefore the overall MSE. This is what [3.3.6 Ridge Regression](#) does, demonstrated by **Figure 72**.

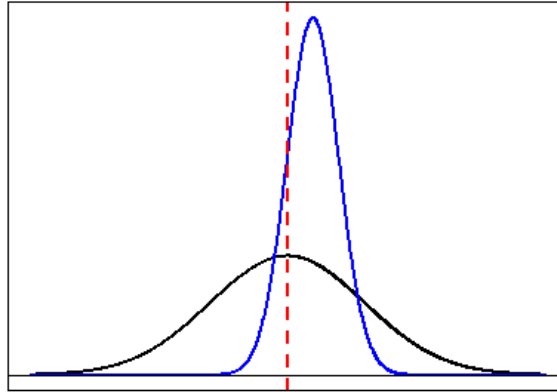


Figure 72: Bias versus Variance

The wide density in black has zero bias with the true parameter value running through its peak, as designated by the dashed red line. The narrower density in blue does not estimate at the true value, but has less variation.

– *Terminology* –

The term bias has a very strict and well-defined statistical meaning. Bias is a theoretical mathematical calculation. Outside of a simulated environment, bias cannot be measured. In certain cases, it is possible to estimate it, but its calculation is dependent on the unknown, true value of the parameter(s) of interest.

Bias for an estimator $\hat{\theta}$ for estimating an unknown parameter θ is defined as:

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

In other words, bias is how far off from the true parameter value the estimated parameter actually is. In the case of regression models, an unbiased methodology has the following properties:

$$E(\hat{\beta}) = \beta$$

$$E(\hat{\sigma}^2) = \sigma^2$$

Or, the estimated coefficient parameters and the estimated variance mathematically achieve their true theoretical values.

When quantifying uncertainty in the model, it is often impossible to separate variance and bias out from the MSE. Calculations can be derived, but are dependent on the true parameter values β and σ^2 , which are unknown. However, it is still important to keep the concept in mind when generating confidence and prediction intervals.

A resource for this analysis is the Joint Agency Cost and Schedule Risk and Uncertainty Handbook (JA CSRUH), 16 September 2014. In particular, see Section 2.4.2 Uncertainty of Parametric CERs (p. 16 ff.).

5.1 Adjust Point Estimate

5.1.1 Overview

The bias of an estimator is the difference between the expected value and the true value of the parameter being estimated. An estimator with zero bias is called unbiased. OLS Regression, under the correct assumptions, produces unbiased CERs. The estimate produced by the CER is a “best guess” as to the true average cost for a given element and corresponding input parameters. Otherwise the estimator is said to be biased. Mathematically the expected value of the estimated coefficients, $\hat{\beta}$, is equal to the true parameter values, β . The [3.3.1 Ordinary Least Squares \(OLS\)](#), [3.3.2 Generalized Least Squares \(GLS\)](#), and [3.3.4 Generalized Linear Model \(GLM\)](#) models result in unbiased coefficient estimates in unit space. Section [3.3.5 Non-linear Least Squares \(NLS\)](#) methods may result in unbiased estimators depending on the functional form, $f(\mathbf{X}; \beta)$, with one such example being the Power and Exponential Models (Section [2.8.2](#) and [2.8.3](#)). Section [3.3.6 Ridge Regression](#) intentionally introduces bias to combat the multicollinearity problem.

It is important to note the language that the aforementioned methods are unbiased *under the correct assumptions*. For example, one assumption that is not always stated is the independent variables in the CER are the “true” independent variables. Consider the following model,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

This form is the unknown theoretical model, with two predictors x_1 and x_2 . Suppose the following model is fit to test a given hypothesis,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

The model is under-fit, leaving out the “true” predictor x_2 . In this scenario, the coefficient estimate $\hat{\beta}_1$, and \hat{y} , are biased. There are other ways in which the resulting model can end up being biased, but corrections and estimations of the bias are possible only when it is known to be present, and even then is a challenge. As a result, there is no need to adjust the point estimate for any of the methodologies presented, with the exception of Log-Linear Regression.

5.1.2 Adjusting the Log-Linear Regression Result

One of the most popular methodologies with known bias is Log-Linear Regression (Section [3.3.3](#)). Since the OLS Regression equation is unbiased, that means that for Log-Linear Regression, the fit equation is unbiased in log space. Unfortunately, the transformation back to unit space creates a bias such that Log-Linear models systematically underestimate the response. One way to remember this is that the additive normal error (assumed by OLS) in log space becomes a multiplicative lognormal error in unit space. The former is the “related normal” distribution for that lognormal distribution. This transformation preserves the median, not the mean. A lognormal distribution shifts the mean to the right of the median, and shifts the mode to the left.

Log-Linear CERs estimate a median cost, not a mean. One solution for estimating at the mean is to apply the PING factor or the Goldberger factor (Appendix [A.4.3.1 Mean Shift](#))⁷¹. This shift is also inherent in the Confidence and Prediction intervals. Alternatively, it may be easier to fit the non-linear Power or Exponential model directly in its non-linear form, utilizing either the [3.3.4 Generalized Linear Model \(GLM\)](#), [3.3.5 Non-linear Least Squares \(NLS\)](#) or [Appendix B Maximum likelihood estimation for Regression of Log Normal error \(MRLN\) Summary](#).

5.2 Generate Confidence Interval

– Terminology –

A confidence interval shows upper and lower bounds for a predicted mean response. In other words, a CER confidence interval may convey error around the average cost of a future system, were it to be produced many times.

5.2.1 Overview

The Confidence Interval (CI) about the regression equation captures only the uncertainty in the regression equation itself. Since a mean prediction for cost is desired, the statement can be made that the true mean response, \bar{y} , will fall within the CI $(1 - \alpha) \cdot 100\%$ of the time, where α is the significance level. The lower the significance level value, the higher the confidence level, and the wider the CI will have to be. One mnemonic is that the CI accounts for both the “bounce” and “wiggle” of the regression line, the former referring to the vertical uncertainty of the y-intercept, the latter to the diagonal uncertainty of the slope.

The generic formulation of the confidence interval is,

$$PE - CV \cdot SE < \bar{y} < PE + CV \cdot SE$$

Or in interval notation,

$$(PE - CV \cdot SE, PE + CV \cdot SE)$$

Where,

PE = Point Estimate

CV = Critical Value

SE = Standard Error

Common notation is to combine $CV \cdot SE$ into a single term called the Margin of Error (MOE). Suppose a new set of values for the independent variables is to be predicted, notated as,

⁷¹ Note that the mean may be more influenced by the input values in a highly non-linear CER than the PING factor and in these cases, it may be simpler to note the CER is producing the median and model uncertainty accordingly. This would simplify documentation and use of the model.

$$\mathbf{x}'_0 = (1, x_{0,1}, x_{0,2}, \dots, x_{0,k})$$

And the point estimate for this new data point is $f(\mathbf{x}'_0; \hat{\boldsymbol{\beta}})$, or simply $\mathbf{x}'_0 \hat{\boldsymbol{\beta}}$ for OLS. The critical value relates the confidence level back to the parametric distribution assumed by the model, which is usually normality. Because the variance is an estimate, the normality assumption corresponds to the critical value coming from a t-distribution with $n - p$ degrees of freedom, at the $1 - \frac{\alpha}{2}$ level, denoted as $t_{1-\frac{\alpha}{2}}(n - p)$ and is usually around 2, depending on the number of degrees of freedom. The standard error (SE) is calculated off of the covariance matrix of the coefficient parameter estimates, at \mathbf{x}'_0 :

$$SE = \sqrt{\hat{\sigma}^2(\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$$

Where,

$$\hat{\sigma}^2 = MSE$$

Putting it all together, a $(1 - \alpha) \cdot 100\%$ Confidence Interval for OLS at the point \mathbf{x}'_0 is,

$$\left((\mathbf{x}'_0 \hat{\boldsymbol{\beta}}) - t_{1-\frac{\alpha}{2}}(n - p) \cdot \sqrt{\hat{\sigma}^2(\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}, (\mathbf{x}'_0 \hat{\boldsymbol{\beta}}) + t_{1-\frac{\alpha}{2}}(n - p) \cdot \sqrt{\hat{\sigma}^2(\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \right)$$

In the single variable case with the new point x_0 , this simplifies down to,

$$\left((\beta_0 + \beta_1 x_0) - t_{1-\frac{\alpha}{2}}(n - p) \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i x_i^2 - n\bar{x}^2}}, (\beta_0 + \beta_1 x_0) + t_{1-\frac{\alpha}{2}}(n - p) \cdot \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_i x_i^2 - n\bar{x}^2}} \right)$$

An alternate form of the scaling factor ($\hat{\sigma}$ times the radical above), is:

$$\sqrt{\left(\frac{\hat{\sigma}}{\sqrt{n}}\right)^2 + [SE(\beta_1) \cdot (x_0 - \bar{x})]^2}$$

This more clearly illustrates the “bounce” and the “wobble” as the two components, combined in a Pythagorean Theorem manner (square root of sum of squares). The former is the uncertainty in the y-intercept, which is the Root Mean Squared Error (RMSE) shrunk by a factor of the square root of n . The latter is the standard error of the slope, β_1 , times the distance away from the mean value of the input parameter.

The interpretation of the confidence interval can be a bit tricky. The statement for a 95% CI would be that there is 95% confidence that the true mean response at the independent variable input \mathbf{x}_0 lies within the confidence interval. Running an infinite number of analyses on random samples of the same underlying true population dataset, 95% of the resulting CIs would capture the true values of the parameters.

5.2.2 Extension to Other Model Forms

The general form of the confidence interval extends to other regression methodologies. The form is always $PE \pm CV \cdot SE$. The point estimate (PE) is easy to calculate by entering the new data and the coefficient estimates into the functional form. The critical value (CV) is from the assumed distribution. The standard error (SE) can become tricky for complex model forms, involving complex matrix calculus. Fortunately, software is able to output CI at requested locations for \mathbf{X} . A summary of the general methodologies follows:

[Generalized Least Squares \(GLS\)](#) – The exact SE can be calculated once the weights, \mathbf{w} , are estimated. Calculate the MSE as with OLS, and the covariance matrix as $MSE \cdot (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, compared to $MSE \cdot (\mathbf{X}'\mathbf{X})^{-1}$ for OLS.

[Transformable Linear and the Log-Linear Model](#) – Produce the confidence interval as a whole in the log transformed space under the same methodology as OLS (but after applying the appropriate adjustment as discussed in Section [5.1](#)). Once produced, transform the entire interval back into unit space in the same fashion as with the point estimate. Note that this procedure results in the CI no longer being symmetric.

[Generalized Linear Model \(GLM\)](#) – A common way to generate a confidence interval for the generalized linear model is to make use of likelihood ratios and the corresponding chi-squared distribution. This is more complex, but many statistical packages are able to produce them automatically. Alternatively, the final iteration of the numerical algorithm can return approximate standard errors for use in confidence interval calculations. This appeals to asymptotic normal properties and works well for large sample sizes but not so well for smaller samples, as are common in cost analysis. Then, generate the CI under the same formulations as OLS.

[Non-linear Least Squares \(NLS\)](#) – The second method for GLM is often used to generate the confidence intervals for NLS. This is one of the reasons why a GLM model, when possible, is preferred to NLS: it can make use of the likelihood ratio tests and intervals. Alternatively, increasingly powerful computers have made the bootstrap approach popular (See Appendix [A.4.7.4](#) and [A.4.8.2](#)).

5.3 Generate Prediction Interval

– Terminology –

A prediction interval shows upper and lower bounds for a single predicted response. In other words, the CER prediction interval defines the potential error around the predicted cost of the future system.

5.3.1 Overview

The Prediction Interval (PI) is a direct extension on the CI. When making an estimate, the result of interest is typically not the mean response, but rather the predicted response for a single circumstance. As a result, the error for predicting a single point rather than a mean response is required. The following demonstrates the general concept of the difference between the two types of intervals:

$$CI: \bar{y}_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}}$$

$$PI: y_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}} + \varepsilon_0$$

The variance for ε_0 must be added into the interval for prediction, but not for the confidence interval. Since $Var(\varepsilon_i) = \hat{\sigma}^2$, for OLS, all the calculations remain the same, except now the standard error (SE) is defined as,

$$SE = \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2(x_0'(X'X)^{-1}x_0)}$$

$$= \hat{\sigma} \sqrt{1 + (x_0'(X'X)^{-1}x_0)}$$

Due to this, the PI is always wider than the CI, since there is more uncertainty involved in estimating a single observation than a mean. From here, all the calculations and extensions to other models remain identical to those introduced with the CI.

5.3.2 Example

The single predictor case makes it easy to visualize both confidence and prediction intervals. **Figure 73** shows a [scatter plot](#) with both a 95% confidence interval and prediction interval for the example in Section [3.3.1.1 Simple Linear Regression \(SLR\)](#). In this example, the confidence interval is fairly tight about the predicted line, but the low sample size and moderately sized variance cause for a much wider prediction interval. Both the confidence interval and the prediction interval are narrowest at the mean point (\bar{x}, \bar{y}) .

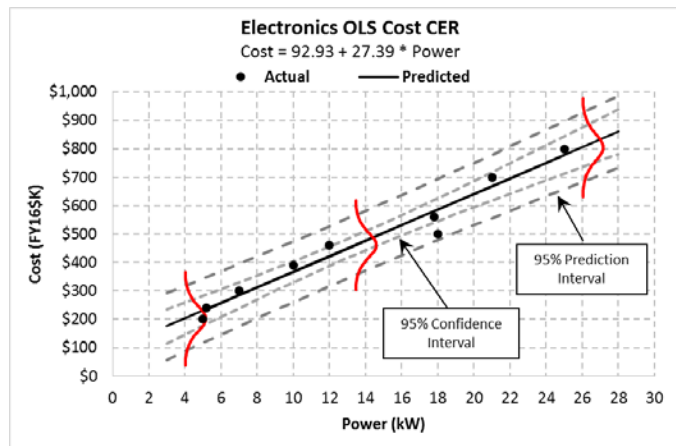


Figure 73: OLS Example 95% Confidence/Prediction Intervals

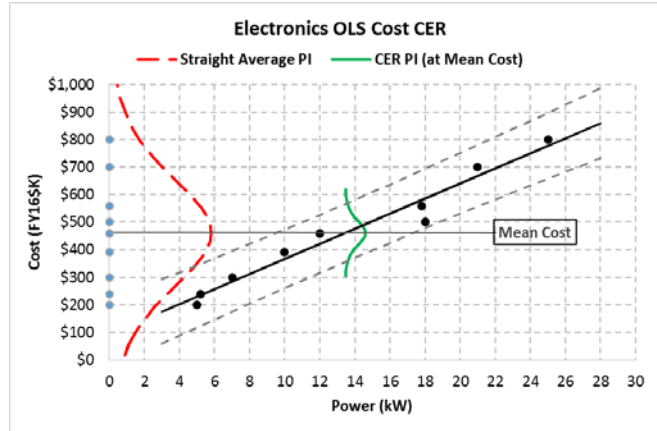


Figure 74: Compare OLS CER PI to a Straight Average PI

Figure 74 demonstrates how much smaller the OLS CER prediction interval is compared to the straight average PI shown in [Figure 7: Confidence and Prediction Interval for the Straight Average of the Electronics Cost Data](#).

Recall the OLS example from Section [3.3.1.3](#) using *Power* and *Aperture* to model *Cost*. Using CO\$TAT, a prediction interval is requested at the point $x'_0 = (15, 9)$, or at *Power* = 15 and *Aperture* = 9). **Figure 75** displays the prediction interval results.

VI. Prediction Intervals

Estimate Inputs

Input	Example
Power	15.0000
Aper	9.0000
Confidence Level (%)	90.00%

Prediction Results

Result	Example
Lower Bound	447.5255
Estimate	515.4563
Upper Bound	583.3871
Delta(%)	
Lower Bound	13.1788
Upper Bound	13.1788
RI\$(%) Multiplier	
Lower Bound	86.8212
Upper Bound	113.1788

Figure 75: OLS Example Prediction Interval Output

The top table simply displays the inputs for both predictors, as well as the stated confidence interval. The bottom table returns the lower bound, the point estimate, and the upper bound. The Delta(%) is the percentage difference the lower and upper bound lie from the estimate. For example,

$$\begin{aligned} \frac{\text{Estimate} - \text{Lower Bound}}{\text{Estimate}} &= \frac{515.4563 - 447.5255}{515.4563} \\ &= 0.131788 \\ &= 13.1788\% \end{aligned}$$

The RISK(%) Multiplier is simply a percentage factor applied to the estimate to produce the respective lower and upper bound. Note how the range of possible answers increases as you move towards the upper and lower bounds of the independent variable.

Based on this example, the conclusion would be that there is 90% confidence that the true predicted value for cost when *Power* = 15 and *Aperture* = 9 is between \$447.5K and \$583.4K.

5.4 Generate CER S-Curve and Histogram

To illustrate the full range of uncertainty around a predicted cost, it may be useful to construct an “S-curve” and/or “Histogram” of the estimate. The S-curve is a visual of the cumulative distribution function (CDF) and the histogram is a visual of the probability density function (PDF).

– *Terminology* –

The Cumulative Distribution Function (CDF) and Probability Density Function (PDF) are the formal statistical terms. The CDF is referred to as an “S-curve” and the PDF as a “histogram.” Note that the PDF will not necessarily actually be “bell shaped” in nature, but it is often referred to as such all the same.

Figure 76 and **Figure 77** are an example of the CDF and PDF. These plots are for the OLS Example prediction interval calculated in the preceding Section [5.3](#).

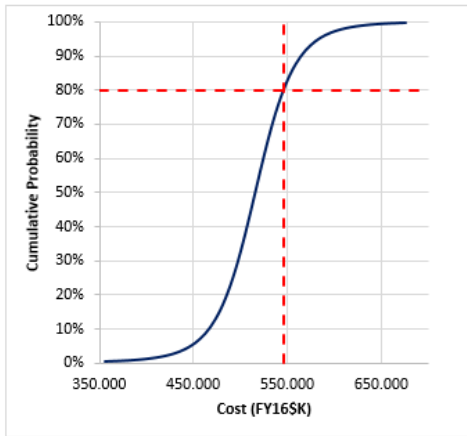


Figure 76: OLS Example Prediction Interval CDF

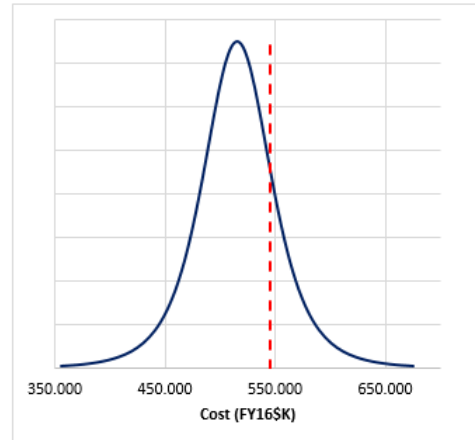


Figure 77: OLS Example Prediction Interval PDF

The S-curve provides the cumulative distribution of the predicted value (cost) evaluated at the requested point (*Power* = 15 and *Aperture* = 9). This interprets as a confidence interval from zero to the value on the curve. Construct the curve in the same way as a prediction interval, but vary the confidence level from 0 to 1. Therefore, the point estimate and the standard error remain the same as with the prediction interval, but with a varied critical value across the full range of probabilities,

$$CV = t_{\gamma}(n - p) \text{ for } \gamma = \{0.00, 0.01, 0.02, \dots, 1.00\}$$

In **Figure 76**, the intersection of the red lines on the CDF indicates the 80% estimate cost. Continuing the example, the interpretation would be that there is 80% confidence that the true predicted value for cost when *Power* = 15 and *Aperture* = 9 is less than \$545.4K FY16 (marked by the red vertical line).

After a CER has been developed, diagnosed, and verified, and after quantifying Risk and Uncertainty by creating Confidence and Prediction Intervals, the next step is to document the CERs. [Step 6: Document CER](#) discusses the steps to properly document the work up to now and produce a final, defensible product.

6.0 STEP 6: DOCUMENT CER

Regardless of how much work it takes to develop a CER, the final product will not stand without proper documentation. Effective documentation ensures the result can be traced back to the source data. Documentation also explains why the source data were selected and how this information was normalized. A well-documented CER enables an analyst to use and defend the CER in an informed manner. Documentation also provides other analysts with adequate information necessary to replicate the CER and to update, should additional data become available. Documentation may vary depending on the intended purpose of the CER, but the key components listed below and illustrated in **Figure 78** are beneficial regardless of use:

1. Scope / Purpose of the Recommended Cost Estimating Relationship
2. Data
 - a. Sources
 - b. Raw Data
 - c. Normalized Data
3. CER Development
 - a. Identify Cost Drivers
 - b. Document Functional Form (Algebraic Equation) and Coefficient Values
 - c. Document Statistics (Goodness of Fit) and Model Selection Methodology
 - d. Characterize CER uncertainty

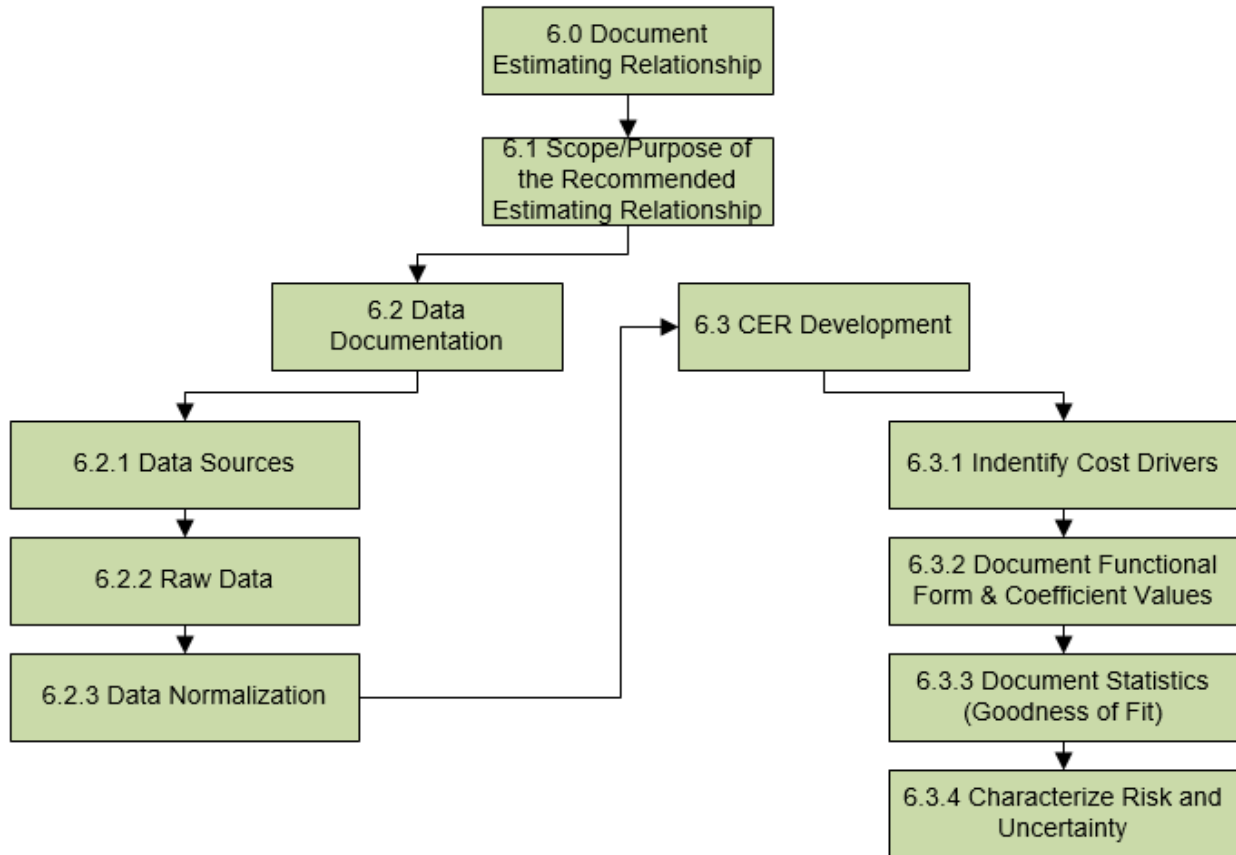


Figure 78: Documenting the Estimating Relationship

The following sections provide guidance on how to document a CER using the Electronics dataset as an example.

References for documenting a cost estimate are: GAO Cost Estimating and Assessment Guide, GAO-09-3SP, Government Accountability Office (GAO), March 2009; Department of the Navy NCCAINST 4451.1B, *Cost Estimating Documentation Policy*, 28 September 2012; and NAVSEA Cost Estimating Handbook (CEH), Naval Sea Systems Command, 2005.

6.1 Scope/Purpose of the Recommended Cost Estimating Relationship

Define the cost element estimated by the CER. This enables the cost analyst using the CER to understand the scope of the estimate generated from the CER.

When developing a parametric relationship, the more contextual data that is available, the easier to determine whether an old system is truly representative of the new system. The System Description includes a generalized description of the system type, as outlined by the system Work Breakdown Structure.

Section [1.3.3](#) provides direction on developing an influence diagram illustrating the interrelationships of the variables thought to impact cost. In regards to CER documentation, start by capturing the objective for the CER. A good practice is including a qualitative description of the dependent variable for the CER

(what is being estimated). This enables the cost analyst using the CER to understand the scope of the estimate generated by the CER. Providing specific definitions for each component of the WBS structure helps to clearly explain what information is included and/or excluded for a particular item. For example:

Electronics for the Air Vehicle (AV): The AV is the airborne platform for the UAV System. The AV serves as the carry vehicle for mission payloads or data link relay. The AV includes all of the hardware and equipment aboard the aircraft to enable flight. These components include: Airframe, Propulsion, Navigation and Guidance, Communications, Air Vehicle Central Computer, Flight Termination System, and Integration and Assembly.

A CER was developed for the production of UAV guidance electronics. Since multiple units will be purchased, first unit cost and related information was collected or derived for each historical program.

The recommended CER is:

$$T_1 = 317.7 * Intensity ^ 0.9088 * 1.101 ^ FFP$$

Where,

T_1 = Theoretical first unit cost FY2016\$K inclusive of overhead, not including fee

Intensity = The power/aperture area in kilowatts per centimeter squared

FFP = 1 if the contract strategy is Firm Fixed Price, 0 if Time and Materials

The scope/purpose of the CER drives the data collection and analysis. Section 6.2 addresses documentation for the data (both dependent and independent variables) included in the analysis.

6.2 Data Documentation

Documenting the data is critical to establishing credibility and traceability of the CER. Documentation for the CER should include the data sources, the raw data, methods utilized for normalizing data, and the resulting normalized data. Data Sources should include the identification of technical, programmatic and cost data sources. For details, see [Step 1: Purpose, Scope, Collect, Validate, & Normalize](#).

To document the data utilized to develop the CER, provide tables with the raw data, normalized data, and a quantitative summary of the data. The remainder of this section addresses the quantitative summary of data sources and the database development and analysis.

6.2.1 Data Sources

Documenting data sources is important to provide for authenticity and accuracy. Documentation should include information associated with the specified systems and information about obtaining the data.

Identification of data sources should include:

- Performing organization (e.g., contractor name or government field activity)
- Data source provider and pedigree (e.g., engineer or technical document)
- Contract number (if applicable)
- Period of performance
- Program name
- Report format (e.g., raw accounting data or contractor cost reports)

- Percent complete of the contract (if using a cost report)
- Source of the data and dates collected

Documentation requires a means to catalog, distinguish, and analyze the information. Consider developing a database, or data table, to document individual cost elements. Data should be stored in a format that allows access to all stakeholders.

6.2.2 Raw Data

Catalog and save all raw data in the original format. Then input the raw cost, technical, and programmatic data into a spreadsheet or database tool. Document the available cost data and pedigree, including data source, life cycle phase, cost interpretation, and quantity. For example, the data for Project 1 in **Table 34** should include a supporting narrative such as: Project 1 cost is the Average Unit Procurement Cost (AUPC) FY2004\$K for the Project 1 contractor (use the project and contractor name) from the SAR report dated December, 2014, contract number 1234-C-08-789, with Period of Performance (PoP) spanning November 1, 2003 – September 30, 2014.

Table 34: Raw Cost Data and Notes

Observation	Price (\$K)	Fee Included	Type	Dollar Type	Year	Source	Comments
Project 1	\$377.5	\$65.0	Unit Cost, 90% Complete	Constant	2004	SAR - 2014	Fee provided by PMO
Project 2	\$190.1	\$32.9	Unit 10 EAC, 95% Complete	Constant	2011	PMO	PMO data in constant dollars (BY16)
Project 3	\$234.6		Unit CostC, 100% Complete	Constant	2014	SAR - 2015	
Project 4	\$343.4	\$79.0	EAC Total, 99% Complete	Budget	2007	CSDR (1921)	as of 31 May 2007
Project 5	\$521.7	\$108.0	FFP Values	Budget	2008	Contract	
Project 6	\$544.4		EAC, 98% Complete	Budget	2013	IPMR	as of 30 Sep 2013
Project 7	\$782.4	\$198.0	EAC, 95% Complete	Budget	2005	CPR	as of 30 April 2005
Project 8	\$944.8	\$192.0	FFP Value	Budget	2011	Contract	
Project 9	\$479.0		T&M Ceiling Value	Budget	2012	Contract	
Project 10	\$1,000.0	Unknown	Estimate	Constant	2007	Foreign	Source in US\$, unknown Foreign to US adjustment
Project 11	\$861.0	\$145.0	Actual	Constant	2015	PMO	Not a comparable system

Table 35: Technical and Programmatic data

Observation	Power (kW)	Aperture (cm^2)	Unit Number	Learning Slope	Service	Contract Type
Project 1	10.00	8.70	1		Air Force	FFP
Project 2	5.00	8.00	10	0.95	Air Force	T&M
Project 3	5.20	8.20	1		Air Force	FFP
Project 4	7.00		1		Air Force	T&M
Project 5	12.00	9.00	1		Air Force	FFP
Project 6	17.80	9.50	1		Air Force	T&M
Project 7	21.00	9.20	1		Air Force	T&M
Project 8	25.00	9.70	1		Air Force	FFP
Project 9	18.00		1		Air Force	T&M
Project 10	6.20	8.20	1		N/A	FFP
Project 11	13.00	9.25	1		Army	FFP

After compiling the raw data, this information needs to be normalized to enable consistent analysis between different systems.

6.2.3 Data Normalization

Section [1.6](#) describes normalization methods. This section of the CER documentation should capture the key data and steps utilized in the normalization of the raw data. For data normalization, be sure to retain copies of the original raw data and normalized data. If escalation indices are used, be sure to collect the data source and publication date of the selected indices. Also, be sure to document the number of significant digits used in the normalization process.

Project 2 in [Table 34](#) provided raw cost data for Unit 10. The goal is to collect or derive the first unit cost (T_1). Choosing to estimate the T_1 means the actual or theoretical T_1 of the source project is a good analogy to ours. Consequently, use the source slope to estimate T_1 for Project 2. **Table 36** shows the results of applying a unit CIC with a 95% slope to derive the T_1 .

Table 36: Deriving First Unit Cost

Project 2		
Unit 10 Cost	\$157.2	Price - Fee
T1	\$186.4	Given 95% Slope, unit learning

The next step is to convert each cost to FY2016. **Table 37** summarizes this process. However, the source of the FY to FY raw escalation indices is required. In addition, to develop the weighted escalation indices needed to convert TY to FY, the source of the outlay profile must also be identified. Table 37 lacks adequate information for complete documentation. Additional documentation fields include data source, outlay profiles, and appropriation (if applicable).

Table 37: Converting Raw Cost to a Base Year 2016 Cost

Observation	First Unit Cost	Dollar Type	Year	BYtoBY	TYtoBY	Normalized Cost
Project 1	\$312.5	Constant	2004	1.248132		\$390.0
Project 2	\$186.4	Constant	2011	1.073029		\$200.0
Project 3	\$234.6	Constant	2014	1.023132		\$240.0
Project 4	\$264.4	Budget	2007		1.134794	\$300.0
Project 5	\$413.7	Budget	2008		1.111997	\$460.0
Project 6	\$544.4	Budget	2013		1.028587	\$560.0
Project 7	\$584.4	Budget	2005		1.197909	\$700.0
Project 8	\$752.8	Budget	2011		1.062648	\$800.0
Project 9	\$479.0	Budget	2012		1.043737	\$500.0

Table 38 summarizes the normalized data that will be used to derive the CER. All data are in consistent units and context.

Table 38: Key Normalized Electronics Data

Observations	Cost (FY16\$K)	Power (kW)	Aperture (cm ²)	Intensity (kW/cm ²)	FFP (1) or T&M (0)
Project 1	390	10.0	8.7	1.1494	1
Project 2	200	5.0	8.0	0.6250	0
Project 3	240	5.2	8.2	0.6340	1
Project 4	300	7.0			0
Project 5	460	12.0	9.0	1.3330	1
Project 6	560	17.8	9.5	1.8740	0
Project 7	700	21.0	9.2	2.2830	0
Project 8	800	25.0	9.7	2.5770	1
Project 9	500	18.0			0

The most effective way to ensure an analysis is repeatable and updateable is to provide the raw and normalized data tables along with all the adjustment factors in the documentation.

Once the estimate's scope and the system description have been documented, the resulting documentation illustrates how the estimate was derived using the selected CER.

6.3 CER Development

[Step 2: Analyze Normalized](#) Data provides additional information regarding how to utilize collected data and to determine the appropriate regression methodology. In terms of documentation, ensure that the following elements are included:

- (1) Identify Cost Drivers
- (2) Document Regression Method Selection and CER Functional Forms
- (3) Document the Selected CER
- (4) Characterize CER Uncertainty

6.3.1 Identify Cost Drivers

[Step 2: Analyze Normalized](#) Data provides more information on the steps to identify the relevant cost drivers. Section [2.5](#) through [2.8](#) discussed cost driver exploration and selection. Documentation should include a discussion of all technical, programmatic, and cost parameters considered to ensure a complete understanding of the totality of the effort and analysis involved in the CER development.

Including an influence diagram ([Figure 3: Simplified Influence Diagram Example](#)) in the CER documentation is a good way to visualize relationships between data and highlight potential cost drivers.

Include a description clearly stating how correlation was examined in the cost driver assessment process. For example, **Table 39** contains PPM correlation coefficients and **Table 40** provides the Spearman Rank correlation coefficients for the Electronics data.

Table 39: Pearson Product Moment Correlation

	Cost	Power	Area	Intensity	FFP
Cost	1.000				
Power	0.981	1.000			
Area	0.939	0.943	1.000		
Intensity	0.996	0.999	0.937	1.000	
FFP	0.053	0.053	0.000	-0.118	1.000

Table 40: Spearman Rank Correlation

	Cost	Power	Area	Intensity	FFP
Cost	1.000				
Power	0.983	1.000			
Area	0.970	0.943	1.000		
Intensity	0.995	0.996	0.964	1.000	
FFP	0.000	0.000	0.000	0.000	1.000

In addition to providing the tables, discuss which values led to the selected parameters highlighted in the following steps. Be sure to include supporting visual representations, graphs, and tables as these artifacts are critical components of a well-documented cost estimate. **Table 39**, **Table 40**, and **Figure 79** provide examples of these visual illustrations.

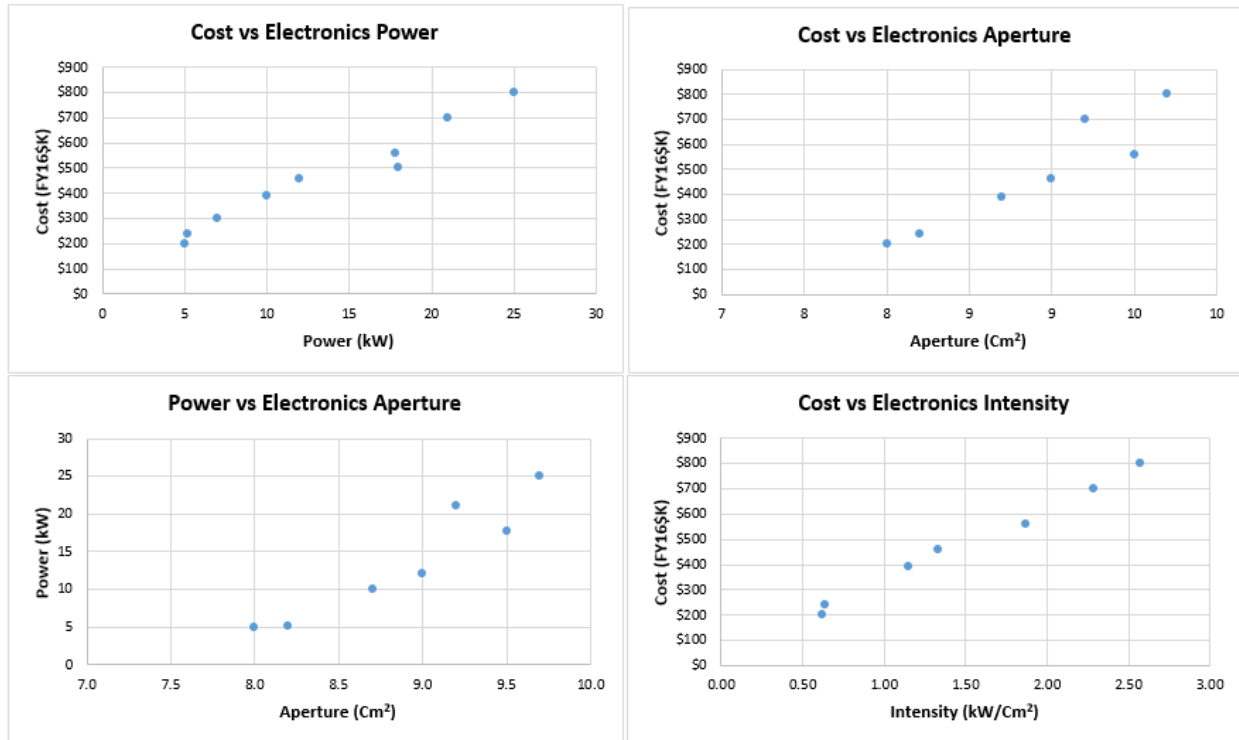


Figure 79: Electronics Data Scatter Plots

6.3.2 Document Regression Method Selection and CER Functional Forms

[Step 3: Generate CER](#) provides additional information on how to choose a regression method and CER functional form. All CERs evaluated in the process of determining the best CER should be documented in this section. An abbreviated example of a supporting rationale is:

“For the Electronics data, the scatter plots and pairwise analysis indicate the potential for linear relationships. Subject matter expert advice or historical CERs on similar systems confirm this selection or point to other starting points. [Ordinary Least Squares \(OLS\)](#) was selected due to the presence of strong linear correlation. Dummy variable and log linear OLS are also explored.”

Section [2.8](#) discusses functional forms and Section [4.4](#) identifies the relevant statistics to determine if the CER is statistically significant. Documentation should include the criteria used to evaluate the CERs, the following model-fit metrics provide an example (Note: this example provides notional values as each organization may reference different predictive metric benchmarks for each predictive metric):

- p-value \leq 5% for t and F
- $R^2_{adj} > 60\%$

Table 41 summarizes the statistical significance for the ten functional forms that were explored. Several failed the criteria despite having excellent R^2_{Adj} . Although the intercept p-value fails for functional form 4 and 5, they are still labeled as “Passed” consistent with guidance in [4.4 Model Significance](#).

Table 41: Summary Results: Fit Statistics

	Name	Status	Equation	DF	STATISTICAL SIGNIFICANCE				
					F	p-value Intercept	p-value b1	p-value b2	p-value b3
1	Linear Pwr	Passed	Cost = 92.93 + 27.39 * Power	7	0.0%	2.0%	0.000		
2	Linear Pwr + Aper	Failed C	Cost = 37.31 + 28.21 * Power + 6.105 * Aper	4	0.0%	93.8%	0.004	0.920	
3	Linear Pwr+Aper+Dum	Failed C	Cost = 212 + 30.4 * Power + (-19) * Aper + 32.95 *	3	0.0%	62.6%	0.005	0.731	0.200
4	Linear Intensity	Passed	Cost = 46.03 + 289 * Intensity	5	0.0%	7.2%	0.000		
5	Linear Intensity + Dum	Passed	Cost = 21.3 + 292 * Intensity + 35.65 * FFP	4	0.0%	14.4%	0.000	0.015	
6	LogLinear Pwr	Passed	Cost = 64.59 * Power ^ 0.7649	7	0.0%	0.0%	0.000		
7	LogLinear Pwr + Aper	Failed C	Cost = 64.13 * Power ^ 0.8071 * Aper ^ (-0.03651)	4	0.0%	28.3%	0.015	0.984	
8	LogLinear Pwr+Aper+Dum	Failed C	Cost = 587.1 * Power ^ 0.9441 * Aper ^ (-1.231) *	3	0.0%	6.3%	0.005	0.366	0.070
9	LogLinear Intensity	Passed	Cost = 336.4 * Intensity ^ 0.9007	5	0.0%	0.0%	0.000		
10	LogLinear Intensity + Dum	Passed	Cost = 317.7 * Intensity ^ 0.9088 * 1.101 ^ FFP	4	0.0%	0.0%	0.000	0.031	

Section [4.5 Model Quality](#) identifies the statistics used to assess the predictive power of the CERs. **Table 42** summarizes some predictive statistics for the six statistically significant functional forms. There are several other predictive metrics available, as described in [4.5.2 Assess Metrics of Prediction](#). Each organization may reference different predictive metric benchmarks for each predictive metric.

Table 42: Summary Results: Predictive Statistics

	Name	Status	Equation	R ² Adj (%)	PREDICTIVE STATISTICS				
					MAD(%) (unit space)	CV(%) (unit space)	PRESS	Pred R ²	SE (unit space)
1	Linear Pwr	Passed	Cost = 92.93 + 27.39 * Power	95.7%	7.2%	6.6%	19101	94.2%	42.2
4	Linear Intensity	Passed	Cost = 46.03 + 289 * Intensity	98.9%	4.9%	3.6%	4844	98.4%	23.1
5	Linear Intensity + Dum	Passed	Cost = 21.3 + 292 * Intensity + 35.65 * FFP	99.7%	1.5%	1.6%	1775	99.4%	11.4
6	LogLinear Pwr	Passed	Cost = 64.59 * Power ^ 0.7649	96.5%	7.0%	6.8%	0.0938	94.7%	43.9
9	LogLinear Intensity	Passed	Cost = 336.4 * Intensity ^ 0.9007	98.4%	4.7%	3.6%	0.0533	96.7%	22.7
10	LogLinear Intensity + Dum	Passed	Cost = 317.7 * Intensity ^ 0.9088 * 1.101 ^ FFP	99.5%	2.6%	2.5%	0.0270	98.3%	20.5

Equation 5 has the best predictive statistics, but is not selected. **Figure 80** plots the standardized residual for candidate CERs 5 and 10. CER 10 appears to conform to the “normally” distributed assumption better than CER 5 (see [Section 4.2.1.4](#)). **CER 10 is selected.** Additionally, CER 10 has a multiplicative error term, which is generally preferred in cost analysis because the error scales with the estimate⁷².

⁷² See JA CSRUH, para 2.4.2.1

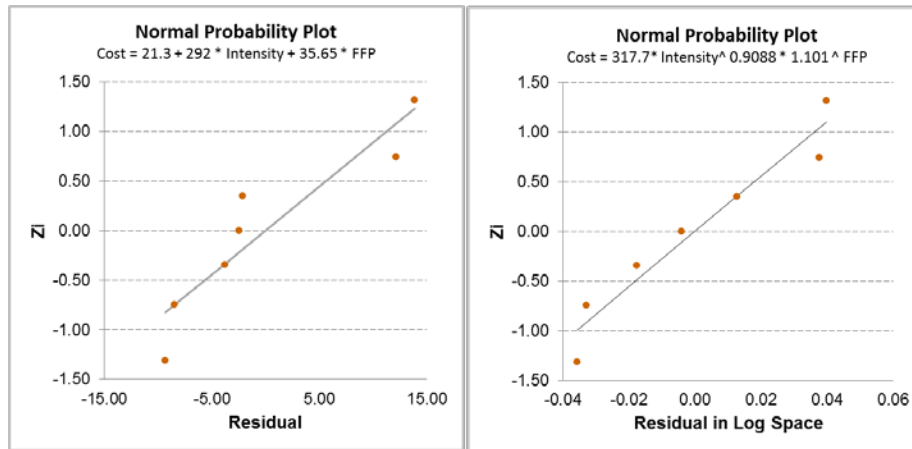


Figure 80: Electronics Data Normal Probability Plot

6.3.3 Document the Selected CER

[Step 4: Validate CER](#) provides a detailed explanation of the statistics associated with the CER. The following is a list of key information that to be included:

- When the regression was performed and by whom
- Sample size, degrees of freedom, regression method, and the actual regression equation
- If any observations were manually excluded, they should be identified with an explanation for their exclusion (there needs to be a sound reason to exclude any data point)
- The [Coefficient Summary Table](#) reports the value of each coefficient developed by the regression method, its standard error, t-statistic, p-value and the potential range of the coefficient (Note: not all regression methods will yield a p-value).
- The Coefficient of Determination R^2 and R^2_{adj} . (Note: not all regression methods will yield a R^2 and R^2_{adj} value).
- The [Analysis of Variance \(ANOVA\) Table](#) provides the sum of squares (regression, residual/error, and total), mean squared error (MSE), and overall model F-statistic and p-value.
- A correlation matrix of the independent variables used in the CER (and/or variance inflation factors (VIFs)) to assess the presence of multicollinearity
- An outlier table in order to identify dependent and independent values that are outliers and if they have a significant influence on the regression result. Typical elements of an outlier analysis table include: residuals, standard residuals (indicates observations with an unusual response), leverage (test statistic for an extreme value of the independent variable), Cook's D (indicates the observation influencing the fitted regression)
- Summary of predictive statistics evaluated in unit space
- **Table 43** identifies a few recommended charts to include as part of the CER documentation
- Prediction Intervals: Often, the regression software can generate the estimate prediction interval.

Table 43: Summary of Graphs to Include in Documentation

Graph	Significance to Analyst Understanding
Residual Plot	<ul style="list-style-type: none"> • Reasonableness of the model assumptions. “Poor” residual plots may indicate an inappropriate model in terms of functional form and/or cost drivers as well as inflated uncertainty in the result. • Presence of potential outlier and influential points on the analysis.
Normal Probability Plot	<ul style="list-style-type: none"> • Reasonableness of the normality assumption. “Poor” normality plots may indicate a non-normal model, which may have impacts on functional form and/or cost drivers as well as understated uncertainty in the result.
Predicted versus Actual Plot	<ul style="list-style-type: none"> • Reasonableness of the model assumptions. “Poor” plots may indicate an inappropriate model in terms of functional form and/or cost drivers.
Leverage Plot	<ul style="list-style-type: none"> • Presence of observations in the dataset “far” away from the center of the data, which may have large impacts on the model.

Figure 81 and **Figure 82** are examples of required documentation. The CER is generated from the data in [Table 38](#).

LogLinear Analysis for Dataset CO\$TAT Electronics Data, Loglinear, Intensity * Dummy

Monday, 18 July 2016, 1:15 PM

I. Model Form and Equation Table

Model Form:	Unweighted Log-Linear model
Number of Observations Used:	7
Equation in Unit Space:	Cost = 317.7 * Intensity ^{0.9088} * 1.101 ^{FFP}

II. Fit Measures (in Fit Space)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value
Intercept	5.7612	0.0238		242.4076	0.0000
Intensity	0.9088	0.0272	1.0024	33.4250	0.0000
EXP_FFP	0.0961	0.0293	0.0985	3.2835	0.0304

Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.0382	99.64%	99.46%	0.9982

Analysis of Variance

Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value
Regression	2	1.6269	0.8135	558.6391	0.0000
Residual (Error)	4	0.0058	0.0015		
Total	6	1.6327			

Further Analysis of Variance

(SS explained by each variable when entered in the order given)

Due To	DF	SS
Regression	2	1.6269
Intensity	1	1.6112
FFP	1	0.0157

Pairwise Correlation Matrix

Variables	Cost	Intensity	FFP
Cost	1.0000	0.9934	0.0067
Intensity	0.9934	1.0000	-0.0915
FFP	0.0067	-0.0915	1.0000

Multicollinearity Analysis

Indep Variables	Indiv R-Sqr (%)	F-Stats	Prob Related to Other Vars	Indiv R-Sqr/Model R-Sqr	VIF	Flags
Intensity	0.84%	0.0423	0.1548	0.0084	1.0085	
FFP	0.84%	0.0423	0.1548	0.0084	1.0085	

Figure 81: Documenting Fit Statistics and ANOVA

CER Development Handbook

Outlier Analysis Summary

Observations exhibiting unusual values	
Dependent Variable	
Independent Variable	
Observations influencing coefficients	#2

Outlier Analysis Table

Obs #	Log of Cost	Predicted Y Value	Residual	Std. Dev. Pred Y	Std. Residual	Leverage	Cook's Distance	Flags
1	5.9661	5.9837	-0.0176	0.0193	-0.5331	0.2541	0.0323	
2	5.2983	5.3340	-0.0357	0.0310	-1.5940	0.6565	1.6190	D
3	5.4806	5.4430	0.0376	0.0267	1.3770	0.4883	0.6032	
4	5.7038							
5	6.1312	6.1184	0.0128	0.0192	0.3884	0.2517	0.0169	
6	6.3279	6.3320	-0.0040	0.0235	-0.1336	0.3791	0.0036	
7	6.5511	6.5114	0.0397	0.0259	1.4134	0.4591	0.5651	
8	6.6846	6.7175	-0.0329	0.0273	-1.2308	0.5112	0.5281	
9	6.2146							

SE = 0.0382, Mean = 6.0629, Coef. of Var. = 0.63% in Fit Space

D denotes an observation with an unusual influence on the fitted regression equation.

III. Predictive Measures (in Unit Space)

Percentage Error Table

Obs #	Name	Actuals	Predicted	Residuals	% Errors	Flags
1	Project 1	390.0000	396.9196	-6.9196	1.7742	
2	Project 2	200.0000	207.2658	-7.2658	3.6329	@
3	Project 3	240.0000	231.1378	8.8622	-3.6926	
4	Project 4	300.0000				
5	Project 5	460.0000	454.1346	5.8654	-1.2751	
6	Project 6	560.0000	562.2569	-2.2569	0.4030	
7	Project 7	700.0000	672.7500	27.2500	-3.8929	
8	Project 8	800.0000	826.7338	-26.7338	3.3417	
9	Project 9	500.0000				
Avg	(Arith)	478.5714	478.7426	-0.1712	0.04%	
Avg	(Absolute)			12.1648	2.57%	

@ Refer to outlier analysis table.

Summary of Predictive Measures

Average Actual (Avg Act)	478.5714
Standard Error (SE)	20.4693
Root Mean Square (RMS) of % Errors	2.88%
Mean Absolute Deviation (Mad) of % Errors	2.57%
Coef of Variation based on Std Error (SE/Avg Act)	4.28%
Coef of Variation based on MAD Res (MAD Res/Avg Act)	2.54%
Pearson's Correlation Coefficient between Act & Pred	99.73%
Adjusted R-Squared in Unit Space	99.17%

Figure 82: Documenting Outlier Analysis and Predictive Measures

Figure 83 provides an example of several charts that should be included as part of the final documentation.

V. Charts

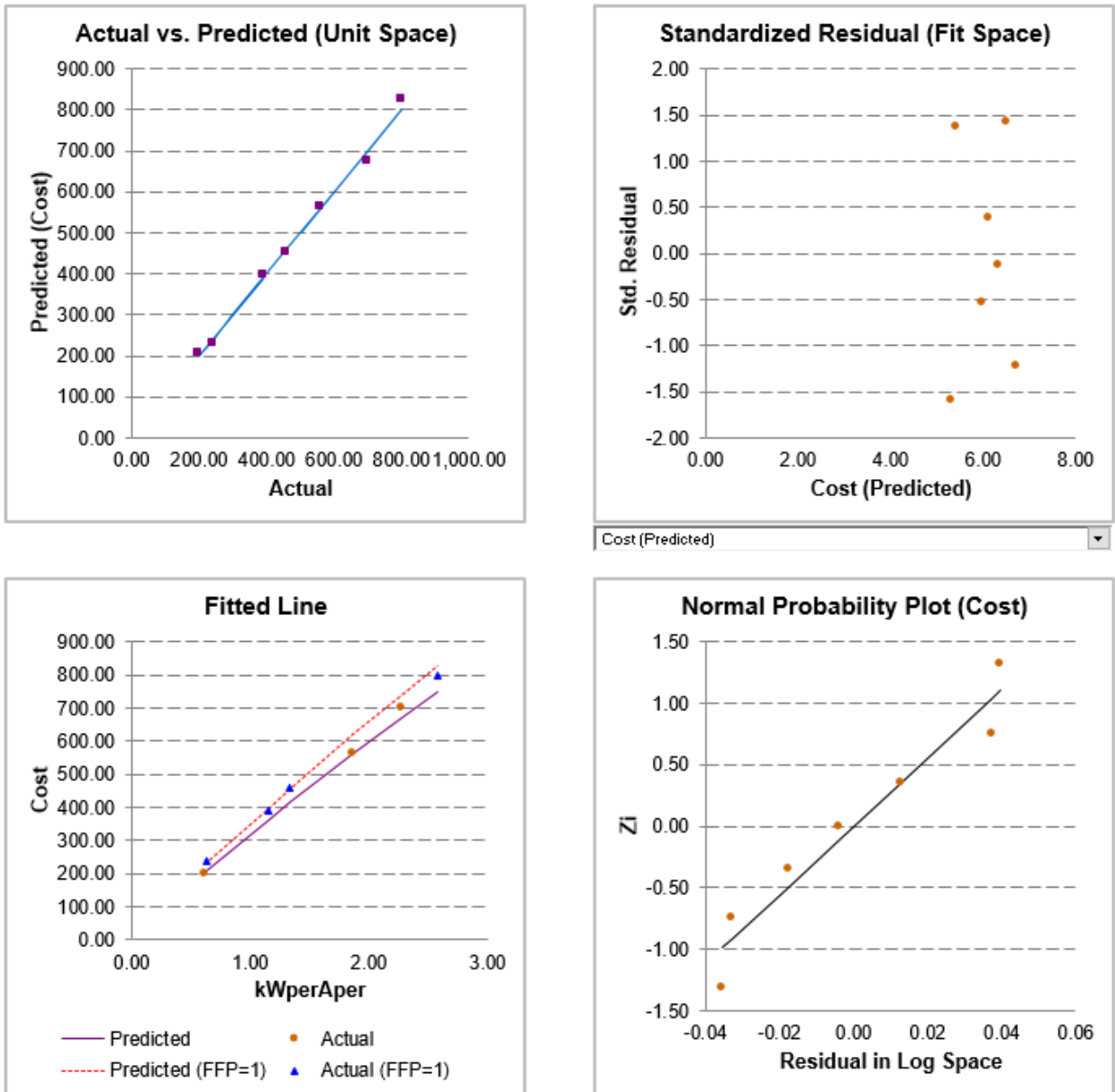


Figure 83: Representative Charts to Document the CER

6.3.4 Characterize CER Uncertainty

[Step 5: Characterize Uncertainty](#) provides guidance on how to model CER uncertainty. How this process is executed can be heavily influenced by the regression method selected. Regardless of the method selected, the documentation should report the confidence interval and the prediction interval (see [5.2](#) and [5.3](#) for additional information).

For generic methods of defining the CER confidence and prediction intervals, see the Joint Agency Cost Schedule Risk and Uncertainty Handbook.

If the point estimate values for the new project are known, the prediction interval can be calculated for OLS and estimated for MUPE and NLS. **Figure 84**, identifies the prediction intervals for an FFP contract based on a low, nominal and high value for Intensity (kWperCm²) at the 90% confidence level. In this case, the CER result is known to be the median of a lognormal distribution. Using @RISK, Crystal Ball or ACEIT, a normal or lognormal distribution can be modeled from two known points. The CER result (median) is one point. The other can be either the low (5th percentile) or high (95th percentile) value documented in **Figure 84**.

IV. Prediction Intervals

Estimate Inputs

Input	Low	Nominal	High
Intensity	0.6250	1.5000	2.5000
FFP	1.0000	1.0000	1.0000
Confidence Level (%)	90.00%	90.00%	90.00%

Prediction Results

Result	Low	Nominal	High
Lower Bound	206.5070	461.2974	728.1603
Estimate	228.1539	505.5589	804.2527
Upper Bound	252.0698	554.0672	888.2966
Delta(%)			
Lower Bound	9.4878	8.7550	9.4612
Upper Bound	10.4824	9.5950	10.4499

Figure 84: Documenting Representative Prediction Intervals

While the aforementioned information is critical, when in doubt include as much information as possible (even if in appendices) as this information may be useful in later analysis. Ensure the documentation provides enough information to effectively identify and explain sources of uncertainty.

Appendices

APPENDIX A GENERAL THEORY.....	195
A.1 Arithmetic	196
A.1.1 Basic Operations	196
A.1.1.1 Exponentiation	196
A.1.1.2 Logarithm	196
A.1.2 Weights	196
A.1.3 Linear Algebra	197
A.2 Probability	197
A.2.1 Foundations of Probability.....	197
A.2.1.1 Discrete Distributions.....	197
A.2.1.2 Continuous Distributions	198
A.2.1.3 Percentiles	199
A.2.1.4 Correlation	199
A.2.1.5 Covariance	199
A.2.2 Probability Distributions.....	200
A.2.2.1 Discrete Distributions.....	200
A.2.2.2 Continuous Distributions	200
A.2.2.3 Exponential Family Distributions	200
A.3 Statistics	200
A.3.1 Descriptive Statistics.....	201
A.3.1.1 Statistical Graphics.....	201
A.3.1.2 Measures of Central Tendency.....	205
A.3.1.3 Measures of Dispersion.....	209
A.3.1.4 Outlier Analysis	211
A.3.2 Inferential Statistics.....	212
A.3.2.1 Hypothesis Testing.....	212
A.3.2.2 Parametric Statistics/Tests	214
A.3.2.3 Non-parametric Statistics/Tests.....	218
A.3.3 Data Analysis Challenges	221
A.3.3.1 Small Data Sets	221
A.3.3.2 Missing Data	224
A.3.3.3 Extreme Observations.....	224
A.3.4 Data Mining	224
A.4 Regression Analysis	225
A.4.1 Ordinary Least Squares (OLS).....	225
A.4.2 Generalized Least Squares (GLS).....	225
A.4.3 Log-Linear Regression.....	225
A.4.3.1 Mean Shift.....	225
A.4.3.2 Unbiased.....	226
A.4.4 Generalized Linear Model (GLM).....	227
A.4.4.1 Generate GLM CER.....	227
A.4.4.2 GLM Model	228
A.4.4.3 GLM Application.....	230

A.4.4.4 GLM Example.....	230
A.4.4.5 Validate CER (Assumptions).....	232
A.4.4.6 Residuals	232
A.4.5 Non-linear Least Squares (NLS).....	233
A.4.5.1 CO\$TAT Application.....	233
A.4.5.2 Excel Application.....	234
A.4.6 Ridge Regression	235
A.4.7 Mathematical/Numerical Techniques	235
A.4.7.1 Iteratively Reweighted Least Squares (IRLS).....	235
A.4.7.2 Maximum Likelihood Estimation (MLE)	235
A.4.7.3 Lagrange Multipliers.....	236
A.4.7.4 Bootstrap	236
A.4.8 Minimum-Unbiased-Percentage-Error (MUPE).....	236
A.4.9 General Error Regression Models (GERM).....	236
A.4.9.1 Zero-Percentage Bias (ZPB) Minimum Percentage Error (ZMPE)	236
A.4.9.2 GERM Significance	237
A.4.9.3 GERM Uncertainty (Bootstrapping)	237
A.4.10 Advanced Regression Methodologies.....	237
A.4.10.1 Restricted Least Squares	237
A.4.10.2 Principal Component Analysis.....	238
A.4.10.3 Mixed Models	238
A.4.10.4 General Estimating Equations.....	238
A.4.10.5 LASSO and the Elastic Net.....	239
A.5 Influence Diagram.....	239
APPENDIX B MAXIMUM LIKELIHOOD ESTIMATION FOR REGRESSION OF LOG NORMAL ERROR (MRLN) SUMMARY.....	241
APPENDIX C CER DEVELOPMENT CHECKLIST	244
APPENDIX D CORRELATION CRITICAL VALUE TABLES	245
APPENDIX E REFERENCES	247
APPENDIX F DATA SETS	248
F.1 Electronics Example	248
F.2 Cost Improvement Curve Example.....	248
F.3 Power Density Example.....	249
F.4 Pseudo-Exact Prior Information Example	249

APPENDIX

APPENDIX A GENERAL THEORY

This Appendix is intended to provide supplementary content to complement the CER Development Guide, including the most important analytical techniques from mathematics, applied mathematics, probability and statistics, and operations research commonly used in cost estimating and risk analysis. While the Flow Chart steps, discussed in the main body of this documentation, mention many of these techniques and present examples, refer to the Appendix sections below for detailed derivations and computations.

Where time and space considerations have precluded a thorough discussion, links to some external references are provided to help understand and recreate the needed analytical steps to implement these techniques. Motivated analysts may want to search further, as extensive statistical references are available through other venues (e.g., the internet). Most of the resources identified in this appendix provide reference to additional related material. This appendix provides links to resources, (color-coded) in the following categories:

Statistical Textbooks: Many of the mathematical and statistical techniques are not unique to cost estimating, and discussed at length in various college-level textbooks.

CEBoK[®]: The Cost Estimating Body of Knowledge is a desktop reference, certification study guide, and training curriculum provided by the International Cost Estimating and Analysis Association (ICEAA). It is a seminal reference divided into sixteen modules, and the most recent version is v1.2 (2013). A companion guide is the Parametric Estimating Handbook (PEH), produced by the legacy International Society of Parametric Analysts (ISPA), now part of ICEAA. As of this writing, PEH content is in the process of merging with CEBoK[®], and references provide direct links to version 4 of the PEH where appropriate.

Professional papers: Often the most thorough references for cost-specific application of analytical techniques are papers presented in forums such as the annual ICEAA conference or the Department of Defense and Department of Navy Cost Analysis Symposia (DoDCAS/DONCAS), or published in the Journal of Cost Analysis and Parametrics (JCAP).

DoD Instructions, Handbooks, etc.: DoD issued instructions are useful resources to trace and comply with current DoD policy and procedures.

Course Materials: For instance Defense Acquisition University (DAU) course BCF 204, Intermediate Cost Analysis has extensive course materials, available at no cost, that serve as a basic reference for cost analysts.

A.1 Arithmetic

A.1.1 Basic Operations

A.1.1.1 Exponentiation

Exponentiation is repeated multiplication. The notation 5^3 is equivalent to $5 \times 5 \times 5$. In that case, 3 is the exponent (or power) while 5 is the base. Define negative exponents via $b^{-n} = \frac{1}{b^n}$, and fractional exponents by $b^{\frac{m}{n}} = \sqrt[n]{b^m}$. In the context of cost estimating, this operation is most applicable in the case of exponential and power functional forms.

A few useful algebraic properties of exponents are as follows:

$$\begin{aligned} e^x \cdot e^y &= e^{x+y} \\ \frac{e^x}{e^y} &= e^{x-y} \\ (e^x)^y &= e^{xy} \end{aligned}$$

A.1.1.2 Logarithm

Logarithms are the inverse of Exponentiation. That is, for a base b the log of a number x is written as $\log_b(x)$ and is defined as the number y such that $b^y = x$. That is, the logarithm answers the question “To what exponent should I raise the base to get the given number?” As a result, the first property of logarithms is $\log_b(b^y) = y$. Here, logarithms “undo” exponentiation, just as division “undoes” multiplication. In the context of cost estimating, this operation is most applicable in the transformations of independent variables in linearizing a [Power Functional Form](#) and (less often) the logarithmic functional form of CERs.

A few useful algebraic properties of logarithms are as follows:

$$\begin{aligned} \log_b(x) + \log_b(y) &= \log_b(xy) \\ \log_b(x) - \log_b(y) &= \log_b\left(\frac{x}{y}\right) \\ x \cdot \log_b(y) &= \log_b(y^x) \end{aligned}$$

Hass, Joel, George B. Thomas, Jr. and Maurice D. Weir, [University Calculus](#), page 376-8.

A.1.2 Weights

In this context weights, or weightings, are methods used to give certain elements greater influence than other elements on the final result. Examples include weights used to correct for differing sample sizes when aggregating averages, and criteria weighting in decision analysis or analysis of alternative evaluations.

Ascher, Uri M. and Chen Grief, [A First Course in Numerical Methods](#), page 373.

A.1.3 Linear Algebra

Vector and matrix representations are useful tools to simplify the mathematics used in applied statistics and econometrics.

Shores, Thomas S., Applied Linear Algebra and Matrix Analysis.

A.2 Probability

In its most general terms, probability is a measure of how likely an event (or set of events) is to occur. It is a broad area of mathematics that involves making conclusions about occurrences based on an understanding of the underlying process driving those occurrences. Conversely, statistics examines a set of occurrences and makes more general conclusions about the underlying process dictating those occurrences. These two areas, probability and statistics, go hand-in-hand.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 7-81.

A.2.1 Foundations of Probability

Probability is an area of mathematics focused on the underlying principles of random variables, which in turn help to serve as a foundation for the field of statistics. This is an extremely broad field with wide-reaching practical impact that includes results such as the central limit theorem, the law of large numbers, and distributions of non-deterministic events.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 7-81.

A.2.1.1 Discrete Distributions

A discrete distribution defines the probability of occurrence of events from a discrete random variable. A discrete random variable is one that can only take on discrete values. Examples include flipping a coin, rolling die, and whether or not a schedule slip occurs. For the coin flip there are only two possibilities, for the rolling die there are six, and the schedule slip will either occur or it will not. This is in contrast to continuous distributions that describe the occurrence of events from a set of continuous possibilities, such as the height of a child on their 5th birthday, or the duration of schedule slip experienced by a program. In these cases, the possibilities are from a continuous set (e.g., the child could be 12.3 in. tall, 38.6 in. tall, some other value between these heights, or virtually any other number).

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 42-49, 59.

A.2.1.1.1 *Probability Mass Function (pmf)*

A probability mass function, f_X , defines the probability that a discrete random variable X will take on a particular value. For a random variable X the probability that X will take on the value a is written as $f_X(a)$. For example, if a flip of the coin is a random variable Y then the probability that Y will be heads is $f_Y(\text{heads}) = \frac{1}{2}$.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 21, 62.

A.2.1.1.2 *Cumulative Distribution Function (cdf), Discrete*

The cumulative distribution for a random variable X is a function, $F_X(a)$, that calculates the probability of the random variable attaining a value less than or equal to a . For a discrete random variable, that means that the cdf is simply the sum of the pmf evaluated at values less than or equal to a . For example, the probability that the roll of a die will be less than or equal to 4 is the probability that the roll is a 1, plus the probability that the roll is a 2, plus the probability that the roll is a 3, plus the probability that the roll is a 4. More succinctly, $F_X(4) = f_x(1) + f_x(2) + f_x(3) + f_x(4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3}$.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 21, 23, 62.

A.2.1.2 Continuous Distributions

A continuous distribution defines the probability of occurrence for events from a continuous random variable. Continuous random variables are those random variables that can take on values from anywhere within a range of values. Examples include the weight of a ship, or the total time for schedule competition. This is in contrast to discrete distributions that describe the occurrence of events from a set of discrete possibilities such as flipping a coin. The value of a coin flip is either heads or tails; a coin flip resulting in a value between heads and tails is impossible.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 49-62.

A.2.1.2.1 *Probability Density Function (pdf)*

The probability density function, f_X , of a continuous random variable X describes the likelihood for X to take on a given value. Unlike in probability mass functions for discrete variables, for a continuous distribution the probability of attaining a single variable is zero despite the fact that $f_X(a)$ may not be equal to zero. This is a property of the fact that continuous distributions can attain an unaccountably infinite number of values and thus the probability of any single event is effectively zero. However, for continuous distributions the cumulative distribution function is useful in determining the probability that the random variable takes on a value within some range.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 22-23, 62.

A.2.1.2.2 *Cumulative Distribution Function (cdf), Continuous*

The cumulative distribution for a random variable X is a function, $F_X(a)$, that calculates the probability of the random variable attaining a value less than or equal to a . For a continuous random variable the cdf is an integral that starts at $-\infty$ and ends at a , $F_X(a) = \int_{-\infty}^a f_X(x)dx$. The cdf is also useful in determining the probability of the random variable attaining a value between a and b . Such a probability is simply $F_X(b) - F_X(a) = \int_a^b f_X(x)dx$.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 21, 23, 62.

A.2.1.3 Percentiles

The n^{th} percentile of a random variable or distribution is the value under which n percent of the possible values occur. For example, the 35th percentile of a dataset is the smallest value x such that 35% of the observations in the set are less than or equal to x .

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 29-30.

A.2.1.4 Correlation

In the broad sense, correlation refers to any type of statistical dependency between data sets. A common example of correlation is the measure of the linear relationship between two random variables. For random variables X and Y with means μ_X and μ_Y , and standard deviations σ_X and σ_Y , the equation for the PPM correlation coefficient between X and Y is,

$$\text{corr}(X, Y) = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

The correlation always takes on a value between -1 and 1, inclusive.

A positive PPM correlation indicates that relatively high values of one random variable signify a likelihood of relatively high values of the other. Conversely, relatively low values of one random variable indicate relatively low values of the other. In the case where correlation is negative, high values in one variable indicate a likelihood of low values in the other variable. It is important to remember that PPM correlation measures the magnitude and strength of linear relationships. A PPM correlation of zero, implying no linear relationship, can exist between two random variables with a strong non-linear relationship.

In cost estimating, well-known correlations include positive relationships between cost and weight of a vehicle, as well as costs between program management and overall program cost, excluding program management.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 38-39.

A.2.1.5 Covariance

Covariance is a measure of how much two random variables change together. If the variables tend to show similar behavior (i.e., greater values in one variable imply greater values in the other variable), then those random variables are said to be positively correlated. Conversely if they show opposite behaviors (i.e., greater values in one variable imply lesser values in the other) then the variables are said to be negatively correlated. For random variables X and Y with means μ_X and μ_Y , the equation for the covariance of X and Y is $\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$, and is sometimes notated as $\sigma_{X,Y}$. The covariance differs from correlation in that it is not bounded between -1 and 1. This makes comparisons of

covariance between different pairs of distributions difficult if the magnitude of those distributions differ greatly.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 36-38.

A.2.2 Probability Distributions

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, page 62.

A.2.2.1 Discrete Distributions

- Uniform Distribution (Discrete)
- Bernoulli Distribution
- Binomial Distribution

A.2.2.2 Continuous Distributions

- Uniform Distribution (Continuous)
- Normal Distribution
- Log-normal Distribution
- Student's t Distribution
- Exponential Distribution
- F Distribution
- Chi-Squared Distribution

A.2.2.3 Exponential Family Distributions

The exponential family of probability distributions are a convenient set of distributions which can take on the following generic form:

$$f(x|\boldsymbol{\theta}) = h(x) \cdot c(\boldsymbol{\theta}) \cdot e^{\sum_{i=1}^k w_i(\boldsymbol{\theta}) \cdot \tau_i(x)}$$

Common distributions belonging to the exponential family include:

- | | |
|---------------|------------|
| • Exponential | • Poisson |
| • Normal | • Binomial |
| • Chi-squared | • Beta |
| • Gamma | |

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 24, 50-51, 59, 63.

A.3 Statistics

In its most general terms, statistics is the study of a set of observed data and drawing conclusions about the underlying processes driving those observations. It is a broad area of applied mathematics closely related to probability. The two major subfields of statistics are descriptive statistics and inferential statistics. Descriptive statistics provide quantifiable information summarizing the primary features of

observed data. Inferential statistics is the field of drawing conclusions about observed data and the underlying sources of that data. In a way, descriptive statistics can be thought of as the “reporting” side of data collection and analysis. Inferential statistics can be thought of as the “analysis” side of data collection and analysis.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate.

A.3.1 Descriptive Statistics

Descriptive statistics is a major branch of statistics that focuses on providing quantifiable information by summarizing the primary features of a set (or sets) of observed data. In practice this involves visualizations of data (e.g., histograms, scatter plots), and quantification of data set properties (e.g., sample mean, sample variance, min, max, interquartile range).

In cost estimating this topic is most often encountered in the analysis of historical data. For example, a succinct description of vehicle production times might be best described via the mean and variance for production times, and a scatter plot of production times over time.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, page 2.

A.3.1.1 Statistical Graphics

Statistical graphics is a broad term including all data visualization methods commonly used in descriptive statistics reporting. The goal of these methods is to convey relevant information about the data sets in a clear and concise way. The most common types of plots and graphs are included in the subsections below.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate.

A.3.1.1.1 *Scatter Plot*

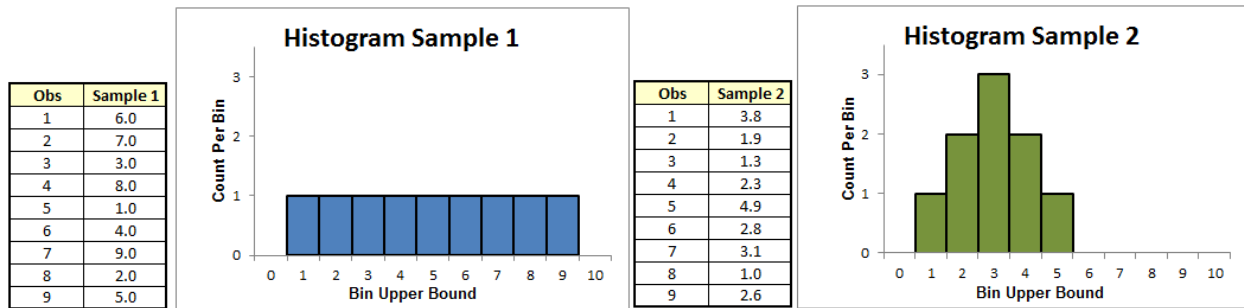
A scatter plot is a graphical representation of values for two variables for a set of data. These plots are easy to generate using virtually any statistical software package. The plots help an analyst develop an intuitive understanding of the data set’s properties and the possible relationships between variables within the data set. For this reason it is highly recommended to generate scatter plots between pairs of variables within the data sets at the start of any data analysis effort.

The data are displayed as a collection of points in the Cartesian plane. The location of each point is determined by the values of the variables being plotted. For instance, a scatter plot of weight vs. cost for ships would have weight on the horizontal axis and cost on the vertical axis. A point of the scatter plot is placed on the graph representing each ship in the data set with the vertical position of the point determined by the cost of the ship and the horizontal position determined by the weight of the ship.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 132-134, 149, 289, 348-349.

A.3.1.1.2 Histogram

A histogram is a graphical representation of the distribution of the data for one variable of the data set. The standard format is a column chart with the horizontal axis corresponding to one variable from the data set, and the vertical axis corresponding to a count of occurrences. The horizontal axis is broken into ranges, or bins, and the height of the column for each range is determined by the number of observations with values in that range for the given variable.



A critical part of the construction of any histogram is determining how many intervals (bins) the data should be grouped into. For the purposes of visualizing the data, the choice of bin size should be one that best illustrates the overall distribution of the data set. Almost universally, the width of each bin is the same for each bin, but the determination of an appropriate bin size is not always clear.

There is no standard approach, but there are several well-known options. Excel takes the square root of the sample size. The Mann-Wald method is used by @RISK, Crystal Ball and CO\$TAT to determine the number of bins for the Chi-squared goodness of fit test. Mann-Wald-divided-by-two is recommended for use as a first approximation of bin count for histograms and as the basis for the chi-squared goodness-of-fit test for samples with less than thirty observations.

For analytical purposes such as distribution fitting and/or hypothesis tests (e.g., Chi-squared test), a more systematic approach is required. In these cases the bin widths are of equal probability, not interval. The Joint Agency Cost and Schedule Risk and Uncertainty Handbook provides a more in-depth discussion on common selection methodologies for these purposes.

Table 44 contains methods to determine histogram bin width⁷³.

⁷³ From JA CSRUH Table A-14., page A-43.

Table 44: Methods to Determine Histogram Bin Width

Mann-Wald/2 and 5% significant level version	$2 \left(\frac{2n^2}{(\Phi^{-1}(\alpha))^2} \right)^{0.2}$ $5\% \text{ siglevel} = 1.88n^{0.4}$ $\text{ROUND}(4*(2*\text{ObsCount}^2/(\text{NORMSINV}(\text{ChiSigLvl}))^2)^{0.2},0)$
Sturges (performs poorly for n<30)	$1 + \log_2(n) \cong 1 + 3.322 \log_{10}(n) \cong 1 + 1.443 \ln(n)$
Scott's Choice	$\frac{\sqrt[3]{n}(\text{max} - \text{min})}{3.5s}$ $\text{ROUNDUP}((n^{1/3}*\text{SampleRange})/(3.5*\text{Stdev}(\text{Sample})),0)$
Freedman-Diaconis	$\frac{\sqrt[3]{n}(\text{max} - \text{min})}{2IQR}$ $\text{ROUNDUP}((n^{1/3}*\text{SampleRange})/(2*\text{SampleInnerQuad}),0)$
Square Root choice	\sqrt{n}

A.3.1.1.3 Empirical CDF

An empirical cumulative distribution function (CDF) plot is similar in principle to a cumulative distribution function of a probabilistic distribution. The plot is a two-dimensional plot with the horizontal axis providing values from the range of the data set, and the vertical axis listing the percentiles from 0 to 100. The line plot describes, for any value in the range of the data set, what percent of the data set is below that value.

An S-curve, such as one derived from risk analysis, is by far the most common instance of an empirical CDF in cost estimating.

van der Vaart, A.W., Asymptotic Statistics, page 265.

A.3.1.1.4 Bar and Column Charts

A bar chart is a plot of bars corresponding to different observations of the data set. The relative size of the bars indicates the magnitude of the variable for each element from the data set represented in the chart. Commonly, a bar chart is one where the bars are drawn horizontally, with “longer” bars corresponding to those observations with larger values. Alternatively, a column chart draws the bars vertically, with “taller” bars, in this case columns, corresponding to observations with larger values.

Histograms and S-curves (Empirical CDFs) are column charts commonly encountered in cost estimating.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 109, 148.

A.3.1.1.5 *Q-Q Plot*

A Q-Q plot, or “quantile-quantile plot” is used to compare two distributions. In descriptive statistics this commonly involves comparing observed data to a theoretical distribution the data are hypothesized to fit. In essence, this plot is simply a scatter plot with percentiles of each distributions on the vertical and horizontal axes. If the data sets are of the same size, or if one is derived from a theoretical distribution, then the plot is simply a plot of points where the horizontal location of the point is determined by the percentile of one data set and the vertical location is determined by the other data set. The case with data sets of different size is more complex and requires a level of interpolation between the points of the smaller data set.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 271, 289.

A.3.1.1.6 *Box Plot*

Box plots are variations of bar charts that display the major quartiles of a data set or distribution. They are also known as box-and-whisker plots. Rather than a full column or bar however, the chart consists of a rectangle for each category. The rectangle contains a line indicating the median (50th percentile) of the distribution. The short ends of the rectangle indicate the first and third quartiles (25th and 75th percentile respectively). In some variations, “whiskers” beyond the edges of the rectangle are applied to indicate the 2nd and 98th percentiles. Other variations on this idea exist, but the format described here is the most common.

Box plots are discussed on slides 57 and 58 in the Related and Advanced Topics section of CEBoK® Module 6 Basic Data Analysis Principles, with a strong emphasis on visual display of information. They are presented as an alternative to histograms, highlighting the quartiles (i.e., 25th, 50th, and 75th percentiles) of the data and any potential outliers.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 121-122, 148.

A.3.1.1.7 *Stem-and-Leaf Plot*

A stem-and-leaf plot is similar to a histogram, and like a histogram conveys the distribution of points within a data set. The format consists of a vertical line with all but the last digit of each data point on the left hand side of the line and the final digit of each data point on the right hand side of the line. Much like the binning required for generating a histogram, some rounding may be required to create a stem-and-leaf plot. These plots maintain a level of order to the data and at least two significant digits of the data values. For these reasons, stem-and-leaf plots are sometimes preferred over histograms in non-parametric data analysis.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 120-121, 148.

A.3.1.2 Measures of Central Tendency

In statistics, the central tendency of a data set is a value that is “typical” of the distribution or a variable of the data set. There are a number of different ways to interpret the concept of “typical,” and as a result there are multiple measures of central tendency.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 111-113.

A.3.1.2.1 Mean

The mean, commonly referred to as the arithmetic mean, average, or the expected value, is a measure of central tendency that is most commonly referred to when describing the “center” of a data set or distribution. If a random variable X is discrete, the mean of X is denoted as $E[X]$ and calculated as $E[X] = \left(\frac{1}{n}\right) \sum_i x_i f_X(x_i)$ where x_i is a possible value of X , and $f_X(x_i)$ is the probability mass function of the random variable X . If X is a continuous random variable, then $E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$ where f_X is the probability density function of X .

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 111-113.

Arithmetic Mean

The arithmetic mean is the best known mean, so much so that in most cases when someone says the word “mean” or “average” they are almost always referring to the arithmetic mean.

Arithmetic Mean, Unweighted

As the arithmetic mean is the most common version of the mean, the unweighted version is the most common version of the arithmetic mean. It is simply the sum of data set values divided by the number of values. For example, the arithmetic mean of the numbers (5, 6, 8, 9, 56) is $(5 + 6 + 8 + 9 + 56)/5 = 16.8$. In general, for a set of observations $x_i, i = 1, \dots, n$ the equation for the unweighted arithmetic mean is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

In cost estimating this calculation is useful in those cases where a single average is sufficiently representative of a group. For example, calculating the average labor rate across all individuals.

Arithmetic Mean, Weighted

In the unweighted arithmetic mean, all points have an equal amount of influence on the final average. However, in the weighted arithmetic mean extra “weight” (w_i) is given to some observations, and as a result, those points have a greater effect on the final mean. The weighted arithmetic mean for a set of observations x_i , and weights w_i , for $i = 1, \dots, n$ is $\bar{x} = \frac{\sum_i w_i x_i}{\sum_i w_i}$. Notice that if all of the weights are 1 (i.e., $w_i = 1$ for all i), then the weighted mean simply reduces to the unweighted mean. Often the sum of the

weights is constrained to 1, so that a data point is given more or less weight than in the unweighted mean depending on whether $w_i > 1/n$ or $w_i < 1/n$.

In cost estimating, the weighted arithmetic mean is useful in calculating an average for a set using only averages from distinct subsets within the group. For example, if a program had 40 FTEs from contractor A at an average labor rate of \$60/hr, and 12 FTEs from contractor B at an average labor rate of \$78/hr, the average across all of the contractors can be found by weighting each of the average labor rates by the number of contractors. That is, the weights are 40 and 12, and the average labor rate across all contractors is,

$$\frac{40 \left(\frac{\$60}{\text{hr}} \right) + 12 \left(\frac{\$78}{\text{hr}} \right)}{40 + 12} = \frac{\$3336}{52} \approx \frac{\$64}{\text{hr}}$$

Note how this is more representative of the average labor rate for the workforce between the two contractors as opposed to simply taking the average of \$60/hr and \$78/hr.

Moriarity, D.J., Basic Statistics Review – Part One, California Polytechnic Institute, pages 5-6.

Geometric Mean

The geometric mean is useful for averaging sets of positive numbers usually interpreted according to their product. For example, a growth factor is a compounding value that is almost always multiplied by another such factor. Thus, a geometric mean is a more representative average of a set of growth rates than the arithmetic mean would be. There are two variations of the geometric mean: the unweighted geometric mean and the weighted geometric mean. Usually, the term “geometric mean” is referring to the unweighted geometric mean.

Kalder, Robin S., Ed.D., “Geometric Mean – What Does it Mean?”, Department of Mathematical Sciences, Central Connecticut State University, New Britain, CT, June 2012. Note: extensive references.

Geometric Mean, Unweighted

For a set of numbers x_i ($i = 1, \dots, n$), the geometric mean is calculated as the n^{th} root of the product $\sqrt[n]{\prod x_i}$. This calculation is commonly used in cost estimating when compounding the average growth rate across periods of equal length. In the case where the periods are of unequal length, a weighted geometric mean should be used.

Kalder, Robin S., Ed.D., “Geometric Mean – What Does it Mean?”, Department of Mathematical Sciences, Central Connecticut State University, New Britain, CT, June 2012. Note: extensive references.

Geometric Mean, Weighted

Like the unweighted geometric mean, the weighted geometric mean is best applied in the case when the data set exclusively involves values that are multiplicative in practice. In the unweighted case all values have equal influence over the final mean, but in the weighted case a set of weights (w_i) are applied to the data. As a result, some observations will have a greater influence over the final mean than others. The weighted geometric mean for a set of observations x_i , and weights w_i , for $i = 1, \dots, n$ is

$$\bar{x} = \sqrt[W']{\prod x_i^{w_i}}$$

where $W' = \sum_i w_i$. One application of the weighted geometric mean is finding the average growth rate between a set of growth rates across periods of different lengths. For example, a weight of $w_i = 6$ would be applied to growth rates across a six-month period, while a weight of $w_i = 12$ would be applied to growth rates across a year.

http://en.wikipedia.org/wiki/Weighted_geometric_mean

Harmonic Mean

The harmonic mean is an average that is most appropriate when an average of rates is desired. As an example in cost estimating, the harmonic mean is commonly applied to inflation outlay rates or expenditure profiles. There are two variations of the harmonic mean: the unweighted harmonic mean and the weighted harmonic mean. The unweighted harmonic mean is commonly referred to as simply the “harmonic mean.”

Wilson, Jim, “The Haarmonic Mean,” Mathematics Education EMAT 4600/6600, The University of Georgia. Available at <http://jwilson.coe.edu/EMT725/HM/HM.html>.

Harmonic Mean, Unweighted

For a set of number x_i ($i = 1, \dots, n$), the unweighted harmonic mean is calculated as

$$\left(\frac{1}{n} \cdot \sum_i \frac{1}{x_i}\right)^{-1}$$

The relationship between the harmonic mean and arithmetic mean should be clear, and the harmonic mean can be remembered as “the reciprocal of the arithmetic mean of the reciprocals.” This average is most appropriate in those cases where the data set is an average of values defined in relation to some unit such as speed (miles/hour), or in the case of inflation, inflation indices representing a rate of change from one period to the next.

van Belle, Gerald, Lloyd D. Fisher, Patrick J. Heagerty, Thomas Lumley, Biostatistics: A Methodology For the Health Sciences, Wiley & Sons, October 2004. See section 10.5.3, page 396.

Venderschel, David, PhD, “Why is harmonic mean used for speeds, not arithmetic mean?”, Rice University, October 2017. Available at <https://www.quora.com/Why-is-harmonic-mean-used-for-speeds-not-arithmetic-mean>.

Harmonic Mean, Weighted

Like the unweighted harmonic mean, the weighted harmonic mean is best applied in the case when the data set exclusively involves values that are defined in relation to a unit. In the unweighted case all values have equal influence over the final mean, but in the weighted case a set of weights (w_i) are applied to the

data. As a result, some observations will have a greater influence over the final mean than others. The weighted harmonic mean for a set of observations, x_i , and weights w_i , for $i = 1, \dots, n$ is $\frac{\sum_i w_i}{\sum_i \frac{w_i}{x_i}}$.

One application of the weighted harmonic mean is the averaging of inflation indices across multiple time periods of different length. In that case, each data point should be weighted by the length of the corresponding time period.

Agrawal, Pankaj, Richard Borgman, John M. Clark and Robert Strong, "Using the Price-to-Earnings Harmonic Mean to Improve Firm Valuation Estimates," *Journal of Financial Education*, Vol 36, No 3/4, Fall/Winter 2010. Abstract from these University of Maine and University of Missouri authors available at <https://www.jstor.org/stable/41948650>.

Root Mean Square (RMS)

RMS is a measure of the magnitude in varying values. It is calculated as the square root of the arithmetic mean of the square values of a data set. In other words, for a data set x_i for $i = 1, \dots, n$ the RMS is

$x_{rms} = \sqrt{\left(\frac{1}{n}\right) \sum_i x_i^2}$. In cost estimating, this calculation of a mean is most common when performing regression and calculating the root-mean-square error of the predicted values.

Weisstein, Eric W., "Root-Mean-Square," from Mathworld, A Wolfram Web Resource. Available at <http://mathworld.wolfram.com/Root-Mean-Quare.html>.

Generalized Mean

The generalized mean is defined such that all the other means described in this section are a special case of the generalized mean. Specifically, for observations x_i for $i = 1, \dots, n$, and some exponent $p \neq 0$ the generalized mean is defined as

$$M_p = \left(\frac{1}{n} \sum_i x_i^p \right)^{\frac{1}{p}}$$

When $p = 0$, $M_p = M_0 = \sqrt[n]{\prod x_i}$, the geometric mean. Similarly, $p = 1$ gives the arithmetic mean and $p = -1$ the harmonic mean. A weighted version of the generalized mean also exists.

Sheldon, Neil, "The Generalized Mean," *Teaching Statistics*, Vol 26, Issue 1, February 2004.

Cumulative Average

The cumulative average is also known as the cumulative moving average and is most relevant in time series data. For example, consider taking a daily reading of plant growth over the course of the summer. A moving average of that data might consist of the average of the 5 most recent days as opposed to an average of all the measurements until this point.

In cost estimating, a common application of cumulative average is in the CUMAV method for CIC analysis.

Hyndman, Rob J., “Moving averages,” November 2009. Available at <https://robjhyndman.com/papers/movingaverage.pdf>.

A.3.1.2.2 *Median*

The median is a measure of central tendency that is defined by the value that partitions the data set or distribution into two sets of equal size. For probability distributions, the median m is the value such that there is a 0.5 probability the random variable will take on a value less than m . An equivalent definition of the median of a random variable X is the value m such that $F_X(m) = 0.5$ where F_X is the cumulative distribution function of the random variable X .

For data sets, one must be careful in the calculation of the median. If there is an odd set of observations then the median is straightforward. For example, the data set 1, 2, 3, 4, 5 has a median of 3 as the set 1,2 is the same size as the set 4,5. If there is an even set of observations however, such as the case 1, 2, 3, 4, 5, 6, then the median may be unclear. Clearly the median must be equal to some number between 3 and 4 in order to partition the data set into two equal subsets of size 3. There are an infinite number of values between 3 and 4, but convention holds that the average of the two “center-most” numbers be reported as the median. Thus, for the data set 1, 2, 3, 4, 5, 6, the median is 3.5.

The definition and calculation and significance of the median are discussed on slides 14 and 15 in the Unit III section of CEBoK® Module 6 Basic Data Analysis Principles.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 111-113.

A.3.1.2.3 *Mode*

The mode is simply the most frequent value of the data set. For example, the data set 1, 2, 3, 4, 4, 4, 5, has a mode of 4. For distributions with pdf or pmf f_X the mode is the value m such that $f_X(m) \geq f_X(x)$ for x equal to all possible values of the random variable X . It should be noted that the mode of a data set or distribution is not necessarily unique. The pdf of a uniform distribution has constant value and so every value attainable by the random variable is the “mode.” Similarly, the data set 1, 1, 1, 6, 6, 6, has both 1 and 6 as modes. In these cases, the mode is not well suited for describing a “typical” value of the data set or distribution, and a different measure of central tendency may be more useful.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 111-113.

A.3.1.3 Measures of Dispersion

In statistics, dispersion describes the amount of variation in a data set or among possible values for a random variable. One can also think of this as the amount of “spread” within the data or distribution. The most common measures are the standard deviation, variance, and CV described below (though other measures do exist). These are important in describing data or distributions since a measure of central tendency gives only a one-dimensional view of a data set or distribution. For example, the data set 1, 1, 1, 1, 1, 1 has the same mean, median and mode as the data set -10, -10, 1, 1, 1, 10, 10.

Measures of dispersion are of particular interest as they quantify (in part) the amount of uncertainty inherent in estimates.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 113-117.

A.3.1.3.1 *Standard Deviation*

The standard deviation of a data set is a measure of the amount of dispersion. Similarly, for a distribution, the standard error is a measure of the amount of spread between elements of a sample from the distribution. In either case, the standard deviation is calculated as the square root of the variance of the data, and is expressed in the same units as the data or random variable for which it is calculated.

It is important to recognize the difference between the standard deviation of a population or a probability distribution, and the standard deviation of a sample data set from the population. In the case of a data set consisting of all members of a population, or of a probability distribution, full knowledge of the random variable is known. As a result, the calculation of the mean is not an estimate and can be used with complete confidence in its value. In the case of a data set sampled from the population, the mean is an *estimate* of the population mean. Commonly the letter σ is used to denote the population standard deviation which can be calculated from the population variance. The letter s is used to denote the sample standard deviation and can be calculated from the sample variance.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 115-116.

A.3.1.3.2 *Variance*

The variance of a data set is a measure of dispersion. Similarly, for a distribution, the variance is a measure of the amount of spread between elements of a sample from the distribution. The calculation of variance for a random variable X is $E[(X - E[X])^2]$.

When calculating the variance of a data set one must consider whether the data set is a population of data or a sample of data from the population. If the data set x_i ($i = 1, \dots, n$) is the population of the data, then the calculation for the population variance is $\sigma^2 = \left(\frac{1}{n}\right) \sum_{i=1}^n (x_i - \bar{x})^2$, where \bar{x} is the arithmetic mean of the data set. For a sample data set x_i , $i = 1, \dots, n$ the sample variance is calculated as $s^2 = \left(\frac{1}{n-1}\right) \sum_{i=1}^n (x_i - \bar{x})^2$. Note that in the calculation for s^2 the summation is divided by $n - 1$ rather than n . This is because in a sample, the mean is only an estimate of the “true” mean, and as a result, one degree of freedom is lost in estimating the sample variance.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 115-116.

A.3.1.3.3 *Coefficient of Variation (CV)*

The coefficient of variation is a measure of the amount of spread within a distribution or data set. It is calculated as the ratio of the standard deviation divided by the mean, $cv = \frac{\sigma}{\mu}$. The measure CV has an advantage over the variance and standard deviation as it is a unitless measure.

The variance is in squared units of a data set, and the standard deviation is in the same units of the data set. This does not allow for a comparison of standard deviation or variance between data sets due to the magnitude associated with these units. For example, the standard deviation in weight of 100 cars might be on the order of hundreds of pounds. Measuring that same standard deviation in tons would result in a much smaller measure of the standard deviation. However, the CV for both of these measurements is the same as the differences in magnitude would be “divided out” by the magnitude differences in the mean.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, page 116.

A.3.1.3.4 *Range*

The range of a data set or random variable is an easy to compute measure of dispersion. It is simply the smallest value in the data set (or attainable by the random variable), subtracted from the largest value. The range is also a somewhat primitive measure of dispersion as it ignores any measure of likelihood within the data. For example, the data set -10, 9, 9, 9, 9, 9, 10 has the same range as -10, -9, -9, -9, -9, 10. The range is still a useful measure for “well behaved data” and, like all other elements of descriptive statistics can be employed if it provides information representative of the data set.

A.3.1.4 Outlier Analysis

Outliers are simply portions of a data set that are “distant” from other elements in the data set. When considering a data set an examination of possible outliers is important. These outlying observations may have an undo effect on the analysis. For example, an outlier may inflate or deflate the calculation of the mean to the point where it is no longer representative of the data set. Additionally, the outlier may be the result of an error in data collection. Identifying that as an issue may help in identifying other issues with the data set that need to be addressed.

There is no concrete definition of what constitutes an outlier, but there are a number of mechanisms for detecting observations that could be outliers. The simplest of which is a visual examination of a scatter plot for every variable and data point of the data set. Points on the scatter plot far removed from other points in the plot deserve, at the least, close consideration. More advanced statistical tests include Chauvenet's criterion, Grubbs' test, Peirce's criterion, Dixon's Q test and Mahalanobis distance.

Once an outlier has been identified the first course of action should always be to examine and attempt to identify the cause of the outlier. Upon closer inspection one might find that the outlier is due to a result in data reporting or collection. Alternatively it could be that the outlier is simply due to chance. Remember, there is always a chance of recording an observation in the sample that is far from the population mean. Outliers are complicated even more when dealing with small data sets. It may be difficult to determine if a data point is representative of the larger population of data by comparing it to only a few other observations from the sample.

There is no universally accepted definition for what constitutes an outlier and a cost estimator must use their own best judgment as to how these data elements should be handled. After identifying the cause of the outlier, one might decide to remove the points as it may unduly influence the calculations of the analysis. Alternatively, an analyst may decide that, while the values of the observation seem extreme, the

data point is a valid observation from the population of data and needs to be included. In either case, a discussion of outliers should be included in any documentation associated with the analysis.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 112, 148, 386, 414, 440.

A.3.2 Inferential Statistics

Inferential statistics is the branch of statistics focused on drawing conclusions from data associated with random variables. Descriptive statistics can be thought of as the “reporting” side of data collection and analysis. Inferential statistics can be thought of as the “analysis” side of data collection and analysis. A sound inferential analysis of data should have its foundation in an understanding of the data. As a result before an inferential statistics process is started, a set of descriptive statistics should be generated and examined for the raw data.

Inferential statistics includes any method used to make conclusions beyond what is included in the data. This includes any analysis involving the relationships between variables in a data set, and forecasts about future observations. In cost estimating, inferential statistics examples include confidence tests for sample statistics, such as the level of certainty with which an analyst can calculate the average value of historical costs given a sample of those costs.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 196-236.

A.3.2.1 Hypothesis Testing

A hypothesis test is the formal statistical procedure used to test a specific scientific hypothesis about a population based on a data sample. A hypothesis can range from a claimed measurement in a formally designed experiment to an observed trend in a survey. In data analysis and CER construction, many of the hypotheses revolve around trends of the data (e.g., positive or negative slope), significance of a variable or model, or validation of some assumption (e.g., normality, constant variance, etc...).

Hypothesis tests are expressed as a null hypothesis, denoted as H_0 . The null is the status quo argument, or hypothesis being tested by the scientific experiment. The null hypothesis is compared against a test statistic. If the value of the test statistic is unlikely to have occurred by random chance under some pre-specified probability threshold, then the null is rejected in favor of the alternative hypothesis, denoted H_1 or H_A . Otherwise, the test fails to reject and the null hypothesis is retained. The conclusions drawn by hypothesis tests are conveyed using very careful language. A hypothesis test can never prove or accept a null hypothesis; it can only reject or fail to reject the null.

In general, hypothesis tests can be broken out into two major groups: parametric and non-parametric. Parametric tests make a strict assumption on the distribution of the test statistic. If the assumption is true, the tests can be very powerful and support relatively low sample sizes. Non-parametric tests do not assume a distribution for the test statistic. While this makes the tests less restrictive, they are often less powerful and require larger sample sizes in order to reject the null hypothesis.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 196, 208-227.

A.3.2.1.1 *Type I and Type II Error*

Statistically, there are two types of errors that can occur when conducting a hypothesis test.

A Type I error is the rejection of the null hypothesis in favor of the alternative, when in reality the null hypothesis is true. For example, consider the test of a cost driver, *weight*. In a CER, the hypothesis test for significance of *weight* is expressed as,

$$H_0: \text{weight} = 0$$

$$H_1: \text{weight} \neq 0$$

If the sample data were to suggest the rejection of H_0 in favor of the conclusion that *weight* is a significant driver (i.e., not equal to zero), but in reality *weight* has no impact on the theoretical model of the population, then a Type I error has been committed. Thus, an insignificant variable is incorrectly included in the CER model.

A Type II error is the failure to reject the null hypothesis in favor of the alternative, when in reality the alternative hypothesis is true. Consider the same example as above regarding the cost driver *weight*. If now the sample data were unable to reject H_0 in favor of the conclusion that *weight* is a significant driver when it actually is, then a Type II error has been committed. Thus, a significant variable is incorrectly excluded from the CER model.

In general, an error of Type I is considered to be more severe. The rate of a Type I error is denoted by α and is controlled by the scientist by setting α at a predetermined level. This rate is theoretical, and the actual error rate may actually be (in cases, substantially) higher or lower.

A Type II error is considered to be less severe and is not directly controlled. The rate of a Type II error is denoted by β and can be tedious to calculate even in simple applications. This error rate is directly related to the concept of power. A test's power is its ability to correctly reject a null hypothesis and is expressed as $1 - \beta$. Besides changing the hypothesis test or one of the hypotheses, the scientist can only control the power of the model by changing the sample size, n .

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 211, 228.

A.3.2.1.2 *Significance Levels*

When conducting a hypothesis test, a threshold value for the test statistic is specified beforehand. This is often done probabilistically on the p-value of the test statistic and is referred to as the α level. If the p-value of the test is below the declared significance level of the test, then the null hypothesis is rejected in favor of the alternative. In a CER, this has a direct impact on whether a model is declared statistically significant or not and whether or not a cost driver is included in the model.

Selecting a significance level is an important part of CER construction. Selecting a cutoff too low (close to zero) can result in no predictors and no model being deemed statistically significant. The result would

be no model, which is often not an option. Raising the cutoff increases the likelihood of rejecting a null hypothesis, making it much easier to obtain statistically significant results. However, too high of an α will result in an unacceptable Type I error rate (Appendix [A.3.2.1.1](#)) and false conclusions on models and the value of certain cost drivers.

Traditional values for α in regression models include 0.05, and 0.10. There are many different philosophies and no agreed upon value exists. The most critical rule is that whatever value is selected, it must be done prior to viewing the data and the test results. Selecting a significance value after the fact invalidates the analysis.

Different fields of study, and different organizations within DoD, may have different guidelines and requirements on significance levels. In CER construction, datasets are often small in sample size. As a result, many models may be unable to achieve low (close to zero) significant levels. Even in these scenarios, a model is required. In this case, raising the significance level may be done, as long as cautions are well understood.

A.3.2.1.3 *Controlling Type I Error*

In an analysis, a significance level, α , is selected for individual hypothesis tests. When constructing a CER model, it is often the case that multiple variables are being tested under multiple simultaneous hypothesis tests. As a result, the actual Type I error rate is being compounded to a much higher rate. For example, this problem arises when creating confidence intervals for each parameter in the model. Suppose there are 20 predictors in a model, each with an α level of 0.05. Since the Type I error rate is 1 in 20, it is expected that about one of these parameters has a confidence interval without the true parameter value being contained in it (with no way of knowing).

There are many methodologies to adjust the overall error rate for multiple comparisons. A few of the most common ones include:

- Fisher's Least Significance Difference (LSD)
- Bonferonni's Correction
- Tukey's Method
- Scheffé's Method

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 227, 468, 469, 477-478, 491.

A.3.2.2 Parametric Statistics/Tests

Parametric statistics is a branch of inferential statistics that analyzes data based on an assumption of the underlying distributions from which the data has been sampled. For example, one might assume that annual maintenance costs for aircraft are normally distributed. Parametric statistics include a number of tests that can be applied to a sample of these maintenance costs provided the assumption regarding their normality is true.

In the case where such an assumption cannot be made, non-parametric statistical methods provide an alternative set of analyses. However, if one can safely make certain assumptions about the underlying

processes of the data set, parametric statistical methods are much more straightforward and simpler than non-parametric methods.

van der Vaart, A.W., Asymptotic Statistics.

A.3.2.2.1 *Pearson's r*

Pearson's r is analogous to correlation in probability. It is a measure of the linear correlation between two random variables, and is defined as the covariance between the variables divided by the product of the standard deviation for each variable. Like correlation, this measure is bounded by -1 and 1, with 1 indicating a strong positive relationship between the variables, and -1 indicating a strong negative relationship between the variables.

Pearson's r , like correlation, is a measure of the linear relationship between variables. There are many other possible relationships besides linear relationships (see functional forms in Section 2.8) for which this measure is not appropriate in evaluating.

van der Vaart, A.W., Asymptotic Statistics, page 242.

A.3.2.2.2 *t-test*

The t-test is a statistical hypothesis most commonly used to determine whether two sets of data are significantly different from one another (e.g., the mean of population X is larger than the mean of population Y), or if a value estimated from a set is significantly different from some constant value (e.g., the population mean is greater than 14). The primary assumptions of the test statistic is that the variables are normally distributed, and that the sample data being tested is an independent sample from the population.

The t-test is extremely prevalent in statistics in that there are a number of variations to account for cases with sets of equal sizes, sets of unequal size, sets with equal variance, sets with unequal variance, and any combination of these. Additionally, the "one-tailed" test should be used when inequality is part of the hypothesis (e.g., the mean of X is greater than the mean of Y), and the "two-tailed" test should be used when the hypothesis involves equality (e.g., the mean of X is equal to the mean of Y). Analysts should be careful to apply the appropriate calculations when employing the t-test.

In cost estimating the t-test is commonly encountered when considering the statistical significances within estimated parameters of a regression. There the hypothesis is usually that the coefficient is not equal to zero and a two-tailed test is conducted.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 210-211.

A.3.2.2.3 *F-test*

The F-test is a statistical test most often used in comparing statistical models. In cost estimating the F-statistic is most commonly encountered in an ANOVA table or when validating the statistical significance of a regression model.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, page 287.

A.3.2.2.4 *Confidence Interval (CI)*

A confidence interval is a range set around a population estimate that helps indicate the reliability of the estimate. For example, values are sampled from a population and used to estimate the mean of the population. Since this is only an estimate, it is not absolutely certain that the population mean is the same. However, if the sample is assumed to be a representative (random) sample and then an interval can be defined around the sample mean such that, to some degree of confidence, the population mean lies within that interval. The more certain the actual value falls within the confidence interval, the larger the interval must be. Thus, any confidence interval consists of not only a range, but a level of certainty associated with the calculation of that range.

There are a variety of ways to calculate a confidence interval and all are dependent upon the random variable being estimated, and the assumptions underlying the population and sample. The most common however include confidence intervals around an estimate of a population mean, and intervals around the estimates for parameters of a regression model.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 197, 203-208.

A.3.2.2.5 *Prediction Interval (PI)*

A prediction interval is a range that estimates, with a certain level of confidence, where a future observation will fall based on previous observations. Without prior knowledge of the population, there is no way to be absolutely certain of an interval for the next observation. For this reason, every prediction interval is calculated via an assumed level of confidence for the prediction. Higher levels of confidence are associated with wider prediction intervals.

There are a variety of ways to calculate a prediction interval and all are dependent upon the underlying assumptions regarding the underlying population from which the next observation is assumed to come. As prediction intervals are forecasts regarding events that have yet to occur, a prediction interval for a given confidence level is always larger than the confidence interval for the same level of confidence.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 257-258, 362, 386, 412.

A.3.2.2.6 *Durbin-Watson Test*

The Durbin-Watson test is the most widely used test to test for autocorrelation of the residuals. It tests the null hypothesis that the correlation between a given residual and the one preceding it is zero. Thus, a p-value for the test less than the pre-specified α would result in a rejection of the assumption of independence of errors. The test is most useful in time series application and is very popular in the field of Economics.

Montgomery, D.C., E.A. Peck, and G.C. Vining, Introduction to Linear Regression Analysis (Third Edition), Wiley & Sons, 2001.

The White test is a popular test in economics for heteroscedasticity of the errors. The test takes the approach of conducting a regression on the squared residuals of the model based on the original predictors and the comprehensive set of second-order combinations. That is, the squared predictors and all of their combinations. The R^2 of the resulting regression is part of the resulting test statistic, testing the null hypothesis that the errors are constant. Thus, a p-value for the test less than the pre-specified α would result in a rejection of the assumption of homoscedasticity of errors.

White, H., “A heteroscedasticity Consistent Covariance Matrix Estimator and a Direct Test of Heteroscedasticity,” *Econometrica*, Vol 48, pp. 817-818, 1980.

A.3.2.2.7 *Breusch-Pagan Test*

The Breusch-Pagan (BP) test is an alternative to the White test, again taking a strategy of employing a regression on the residuals. This test relies on the normality assumption by use of the F-test. If the linear regression line drawn through the residuals has all of its parameters statistically significant, then the BP test rejects the assumption of homoscedasticity of errors.

Pardoe, Dr. Iain, Dr. Laura Simon, and Dr. Derek Young, “STAT 501/Regression Methods,” Eberly College of Science, Pennsylvania State University, 2018. See section 7. Available at <https://onlinescourses.science.psu.edu/stat501>.

A.3.2.2.8 *Anderson-Darling Test*

The Anderson-Darling test is a statistical test of whether a given sample of data are drawn from a given probability distribution. In its basic form, the test assumes that there are no parameters to be estimated in the distribution being tested, in which case the test and its set of critical values is distribution-free. However, the test is most often used in contexts where a family of distributions is being tested, in which case the parameters of that family need to be estimated and account must be taken of this in adjusting either the test-statistic or its critical values. The test can be applied to a variety of probability distributions, but its most common application is in normality testing. When applied to testing if a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting most departures from normality. Because this is the most common application in cost estimating, it is placed under Parametric Statistics, but in the former case it would fall under the following Non-parametric Statistics section, like the Chi-square and Kolmogorov-Smirnov (KS) tests, which also involve the empirical distribution function (EDF).

Natrella, Mary, Carroll Croarkin, and many others, [NIST/SEMATECH e-Handbook of Statistical Methods](#), National Institute of Standards and Technology (NIST) and SEMATECH consortium, updated October 2013. Available at <http://www.itl.nist.gov/div898/handbook/index.htm>. See section 1.3.5.14.

A.3.2.2.9 *Shapiro-Wilk*

The Shapiro-Wilk test is specifically for normality and tests the assumption that the data are from a normal distribution. The test statistic calculates weighted deviations of the sample data from the normal distribution. The test has been proven to perform very well in comparison to the other normality tests and is often favored by analysts. The SW statistic tests the same null hypothesis that the data does follow the

distribution of interest (i.e., normal). Thus, a p-value for the test less than the pre-specified α would result in a rejection of the assumption of normality of errors.

Natrella, Mary, Carroll Croarkin, and many others, NIST/SEMATECH e-Handbook of Statistical Methods, National Institute of Standards and Technology (NIST) and SEMATECH consortium, updated October 2013. Available at <http://www.itl.nist.gov/div898/handbook/index.htm>. See section 7.2.1.3.

A.3.2.2.10 *Cook's Distance*

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, page 456.

A.3.2.2.11 *Akaike Information Criterion (AIC)*

Joint Cost and Schedule Risk and Uncertainty Handbook (CSRUH), 16 September 2014, Appendix A.9.6, <https://www.ncca.navy.mil/tools/csruh/index.cfm>

Akaike, H., "Akaike's Information Criterion," In: Lovric M. (eds), *International Encyclopedia of Statistical Science*, Springer-Verlag, Berlin Heidelberg, 2011.

Hu, Shuhua, "Akaike Information Criterion," Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, February 2012. Available at http://www4.ncsu.edu/~shu3/Presentation/AIC_2012.pdf.

A.3.2.2.12 *Bayesian Information Criterion (BIC)*

Joint Cost and Schedule Risk and Uncertainty Handbook (CSRUH), 16 September 2014, Appendix A.9.7, <https://www.ncca.navy.mil/tools/csruh/index.cfm>

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 646-650.

A.3.2.3 Non-parametric Statistics/Tests

Non-parametric statistics is a branch of inferential statistics focused on the analysis of data without assuming an underlying distribution from which the data has been sampled. For example, if a certain type of data are collected for the first time, there may be no historical basis for assuming a theoretical distribution for the data.

Parametric statistical methods are based on assumptions that non-parametric methods avoid. For this reason, non-parametric methods are generally simpler and more robust in that those assumptions need not be tested nor valid for the results to hold. The downside however, is that for the same level of confidence, a non-parametric method usually needs more observations than the parametric methods.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate.

A.3.2.3.1 *Spearman's Rho*

Spearman's rank correlation coefficient, also known as Spearman's Rho (ρ) is a test that measures the level of statistical dependence between two random variables⁷⁴. It is similar to correlation in that it is bounded between -1 and 1. In the cases where there are no repeated observations, a value of $\rho = 1$ implies perfect positive dependence and $\rho = -1$ implies perfect inverse dependence. With data sets containing repeated values, the calculations and interpretations differ slightly. Unlike Pearson's r , Spearman's Rho does not assume a strictly linear relationship. Instead, it only assumes that the relationship between the two variables can be described as a monotonic function (either never decreasing for positive dependence between variables, or never increasing for negative dependence between variables).

Conover, W.J., Practical Nonparametric Statistics (Third Edition), Wiley & Sons, 1999. Section 5.4 "Measures of Rank Correlation." See pp. 314-319 for a specific discussion of Spearman's Rho

http://www.unesco.org/webworld/idams/advguide/Chapt4_2.htm

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 586-589, 604.

A.3.2.3.2 *Kendall's Tau*

The Kendall rank correlation coefficient, or "Kendall's Tau," measures the correlation of rank between two variables within a data set. The observed values for the variables are ranked independently of each other, and those rankings are compared to a ranking of the variables as they were collected in the original data. Like correlation, the resultant calculation is bounded between -1 and 1, with a value of 1 or -1 indicating a perfect relationship, or perfect inverse relationship respectively.

Conover, W.J., Practical Nonparametric Statistics (3rd ed.). Wiley (1999). Section 5.4 "Measures of Rank Correlation." See pp. 319-323 for a specific discussion of Kendall's Tau.

http://www.unesco.org/webworld/idams/advguide/Chapt4_2.htm

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 589-592, 604.

A.3.2.3.3 *Runs Test*

The Runs test is a nonparametric test to determine if the sequence of data are random. The test looks at the sign (+ or -) of each residual and attempts to detect if there is a pattern to their occurrences. It tests the null hypothesis that the positive and negative elements appear at random. Thus, a p-value for the test less

⁷⁴ For a specific discussion of Spearman's Rho, see: Conover, W.J., Practical Nonparametric Statistics (3rd ed.). Wiley (1999). Section 5.4 "Measures of Rank Correlation." See pp. 314-319
http://www.unesco.org/webworld/idams/advguide/Chapt4_2.htm

than the pre-specified α would result in a rejection of the assumption of independence of errors. The test is most useful in time series application.

Natrella, Mary, Carroll Croarkin, and many others, NIST/SEMATECH e-Handbook of Statistical Methods, National Institute of Standards and Technology (NIST) and SEMATECH consortium, updated October 2013. Available at <http://www.itl.nist.gov/div898/handbook/index.htm>. See section 1.3.5.13.

A.3.2.3.4 *Wilcoxon Two-Sample Test*

The Wilcoxon Two-sample test, also called the Mann-Whitney U test, is a non-parametric test for whether two populations are statistically significantly different from one another. The test is especially useful when only ordinal information about the data set is available. It differs from student's t-test in that it does not assume a normal distribution for the underlying populations and can be applied more widely (such as to ordinal data).

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 575-579, 596, 603.

A.3.2.3.5 *Kruskal-Wallis Test*

The Kruskal-Wallis test is used for determining whether two samples come from the same distribution, and can also be used in testing the independence of more than two variables. It is a non-parametric equivalent of an ANOVA test (specifically the "one-way" ANOVA). Unlike the ANOVA test, it does not assume that the samples are from a normal distribution.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 581-582, 603-604.

A.3.2.3.6 *Wilcoxon Test*

The Wilcoxon test is used to compare two related samples to determine if the population means calculated from those samples are statistically significantly different. It is a non-parametric version of the paired t-test. It differs from the paired t-test in that it does not assume the populations are normally distributed, and it can be applied to ordinal data.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 568-573.

A.3.2.3.7 *Friedman's Test*

Friedman's test is a non-parametric test used to detect differences in evaluation across multiple measurements. Commonly this is used to detect differences between "blocks" of a designed experiment.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 583, 584-585, 604.

A.3.2.3.8 *Chi-square test*

Pearson's chi-squared test, frequently referred to as simply "Chi-squared test," tests for the goodness of fit between observed data and some theoretical distribution, or if observations from a sample are statistically independent.

In cost estimating, the Chi-Squared test is most frequently used in validating the assumptions of the regression model, such as normality of the residuals.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, page 318.

A.3.2.3.9 *Kolmogorov-Smirnov (KS)*

The Kolmogorov-Smirnov test, or "KS test," is a test to determine if two continuous probability distributions are statistically significantly different. The one-sample test compares observed data to a theoretical distribution, such as comparing a set of sample values to the normal distribution. The two-sample test is used in comparing the distributions of two sets of observed data.

In cost estimating it is commonly used to compare the residuals of a regression analysis to the normal distribution in a test of the normality assumption.

van der Vaart, A.W., Asymptotic Statistics, page 290.

A.3.3 Data Analysis Challenges

There are a number of challenges commonly encountered by cost estimators. These include small data sets, data sets with missing or incomplete information, and the presence and influence of extreme observations. In all of these cases, the best course of action is to first attempt to remedy the problem if possible. That includes collecting more data, finding the information missing from the collected data set, and determining the cause of the unusual observations, respectively. As it is not always possible to correct such errors, it is important to understand the implications and proceed with the analysis under caution.

A.3.3.1 Small Data Sets

Small data sets make it difficult to form conclusions about the population from which the data were sampled. Additionally, these conclusions are difficult to test due to the effect of small sample sizes on test statistics. There are some benefits to small data sets. First, with small data sets all of the data are easily comprehended. All relevant information about a data set with five points can be easily conveyed with scatter plots and a few descriptive statistics.

Cost analysts are often expected to "make do" with small data sets. As a possible remedy, the analyst should consider combining the current data set with a second one to produce a larger data set. In doing so, the analyst must consider whether the two data sets are similar enough with respect to the goals of the analysis. For example, the analyst might combine a small data set of components from one submarine class with a similar data set from a different submarine class. If the element of cost being estimated is similar enough between the submarine classes, then combining them makes sense. Otherwise, the additional data may be more harmful than helpful by skewing the results.

What constitutes a “small” data set is very application specific. Some fields of study are used to years of repeated observations, and a sample size of $n = 100$ may be considered miniscule. In others, an event may be so rare in occurrence that $n = 2$ is considered quite good. However, despite what may be a “good” sample based on scarcity of data, small sample sizes can have large statistical implications. With small samples, variances tend to be large and thus produce wide confidence and prediction intervals. Larger variances also make it more difficult to determine statistical significance of results.

Collins, Justin, Jordan Brown, Christine Schammel, PhD, Kevin Hutson, PhD, and W. Jeffery Edenfield, MD, “Meaningful Analysis of Small Data Sets: A Clinician’s Guide,” Greenville Health System (GHS) Proc., June 2017; 2(1); pages 16-19. Authors from GHS Institute for Translational Oncology Research, Pathology Consultants, and Furman University Department of Mathematics. Available at <http://hsc.ghs.org/wp-content/uploads/2016/11/GHS-Proc-Finding-Meaning-In-Small-Data-Sets.pdf>.

Seibert, Carl F., and Sarcy Clay Siebert, Data Analysis with Small Samples and Non-Normal Data: Nonparametrics and Other Strategies, Oxford University Press, 2018.

A.3.3.1.1 *Small Data Sets – Degrees of Freedom*

Closely tied to the sample size is the number of degrees of freedom for the model (i.e., DF_{error}). With 2, 3, or 4 observations, a simple linear CER with a slope and intercept term has 0, 1, or 2 degrees of freedom, respectively. This may be insufficient in conducting statistical inference.

To “save” a degree of freedom, different approaches have been taken. The simplest solution is to drop a parameter from the model. For example, dropping the intercept term from the equation results in the Factor CER, and one additional degree of freedom. When multiple predictors are being used, one (or more) may be dropped from the equation. Despite desires to retain these parameters in the model, the statistics may simply make it infeasible.

Another approach is to “fix” a coefficient to a precise value. This practice is discussed in detail in Section [3.4.3 Pseudo-Exact Prior Information on Parameter Values](#).

The LASSO method is a modern regression methodology capable of fitting models with more parameters than observations (Appendix [A.4.9.5 LASSO and the Elastic Net](#)).

As the number of parameters approaches the sample size (i.e., low degrees of freedom), the model can suffer from overfitting. This results in a great model fit, but with poor prediction abilities.

Smith, Martha K., “Common Mistake Mistakes in Using Statistics: Spotting and Avoiding Them,” University of Texas at Austin, June 2014. Available at <https://www.ma.utexas.edu/users/mks/statmistakes/overfitting.htm>.

Draper, Norman R. and Smith, Harry, Applied Regression Analysis (Third Edition), John Wiley & Sons, Inc., 1998.

A.3.3.1.2 *Small Data Sets – Asymptotic Results*

Further complicating the problem is the case of asymptotic results, such as with Section [3.3.5 Non-linear Least Squares \(NLS\)](#). These models make use of properties that are accurate with large sample sizes. Defining a large sample size is difficult; it depends on many factors such as the variance of the sample and the complexity of the model. Some problems may behave well with $n = 30$, while others may require larger samples such as $n = 100$. However, almost surely results with $n < 10$, as is common in cost analysis, will be insufficiently small.

In many cases, use of these methods is for lack of better options. The small sample size is just a fact of life, and the analyst may proceed to use the analysis under caution, understanding that prediction intervals and statistical significance may be inaccurate. In particular, the interval may not cover the true value at the assumed significance level (α), and the F-test and t-tests may have Type I error rates (see Appendix [A.3.2.1.1 Type I and Type II Error](#)) different from the assumed significance level (α).

A.3.3.1.3 *Small Data Sets – Bootstrap*

The bootstrap method (Appendix [A.4.7.4 Bootstrap](#), [A.4.8.2 GERM Uncertainty \(Bootstrapping\)](#)) is an alternative to relying on asymptotic statistics. Implementation requires iterative resampling from the data set sample and deriving the appropriate statistics from each sample. Studies have shown strong performance for these methodologies.

http://ocw.mit.edu/courses/sloan-school-of-management/15-450-analytics-of-finance-fall-2010/lecture-notes/MIT15_450F10_lec09.pdf

Book, Stephen A., “Prediction Bounds for General-Error-Regression Cost-Estimating Relationships,” *Journal of Cost Analysis and Parametrics (JCAP)*, Vol 5 No 1, 2012.

Dunlop, Dorothy D. and Ajit C. Tamhane, [Statistics and Data Analysis: From Elementary to Intermediate](#), pages 597-601.

A.3.3.1.4 *Small Data Sets – Bayesian Analysis*

Another way to work with small sample sizes is to use Bayesian statistical methods. Bayesian methods do not make use of degrees of freedom and are thus not constrained by their limitations. Under the classical, or “frequentist,” framework, the regression coefficients are unknown, *fixed* values. The Bayesian paradigm views these coefficients as unknown, *random* values. As a result, these analyses rely on prior distributions on the parameters, either informative (e.g., $\beta_1 > 0$), or uninformative (e.g., $-\infty < \beta_1 < \infty$), weighted against the data. Bayesian analysis comprises advanced methods that are difficult to implement from a statistical perspective. However, their results are based on probabilities and can be very intuitive to understand.

Smart, Christian, PhD., “Bayesian Parametrics: How to Develop a CER with Limited Data and Even Without Data”, ICEAA Professional Development & Training Workshop, Denver, 9-13 June 2014.

Dunlop, Dorothy D. and Ajit C. Tamhane, [Statistics and Data Analysis: From Elementary to Intermediate](#), pages 646-650.

A.3.3.2 Missing Data

In statistics, missing data refers to an observation for which there is no data recorded for one or more of the variables. There are a number of technical approaches to the topic of missing data. However, in cost estimating, the term missing data are usually meant more generally to describe data that has not been recorded, is not available to the analyst, or has been provided but is known to be inaccurate.

In some cases, cost analysis provides a mechanism for dealing with missing data. For example, in CIC analysis, calculations on lot production costs can provide insight when unit level data are not available.

Soley-Bori, Marina, “Dealing with missing data: Key assumptions and methods for applied analysis,” Technical Report No. 4, Boston University School of Public Health, Department of Health Policy & Management, May 2013. Available at <https://pdfs.semanticscholar.org/5691/e4052ddc076059184c9d055c30211ba815b1.pdf>.

There is a wealth of information regarding the identification of extreme observations (outliers) and assessing their impact on the analysis. For the cost analyst usually faced with small data sets, the evaluation of whether to include or remove an extreme data point is not always clear. Identifying an extreme point does not imply that the point should be removed. It only implies that the point is in need of further investigation. If it is found that the data point was recorded improperly, or that there is some fundamental difference between the point and others in the data set (other than the value of the data point), then the removal of that point may be appropriate. However, if by all other measure, the point is a valid element of the data set then it should be included. Extreme observations can be beneficial in data analysis. Specifically, they can be used to form bounds necessary for risk and sensitivity analysis.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 112, 148, 386, 414, 440.

A.3.4 Data Mining

Data mining is the process of identifying patterns in large data sets. It contrasts traditional statistical analysis where the process is to formulate a hypothesis and collect data separately, and then evaluate the hypothesis by examining the data. In data mining, the data are “mined” for conclusions with the intention of extracting data in an understandable way. It should be noted that “data mining” is a bit of a “buzzword” that encompasses numerous methods previously developed and more general terms such as “large scale data analysis” are more relevant.

Hand, Mannila, and Smyth, Principles of Data Mining, Massachusetts Institute of Technology Press, Cambridge, MA, 2001. ISBN 026208290X.

Berry and Linoff, Mastering Data Mining, Wiley & Sons, 2000. ISBN 0471331236.

Delmater and Hancock, Data Mining Explained, Digital Press, 2001. ISBN 1555582311.

A.4 Regression Analysis

A.4.1 Ordinary Least Squares (OLS)

The coefficients for OLS can be calculated from a formula. For the single predictor case,

$$\hat{\beta}_1 = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

And,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$$

In matrix notation, useful for the multiple predictor (as well as the single predictor) case,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

And,

$$\hat{\sigma}^2 = \frac{\mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}}{n - (k + 1)}$$

A.4.2 Generalized Least Squares (GLS)

The coefficients for GLS can be solved for with a formula. In matrix notation, useful for the multiple predictor (as well as the single predictor) case,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$$

And,

$$\hat{\sigma}^2 = \frac{\mathbf{y}'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1})\mathbf{y}}{n - (k + 1)}$$

A.4.3 Log-Linear Regression

A.4.3.1 Mean Shift

A.4.3.1.1 Goldberger Factor

In log space the standard log-linear regression equation with multiplicative error is $\mathbf{y} = \beta_0 \mathbf{x}^{\beta_1} \cdot e^{\boldsymbol{\varepsilon}}$, and after transformation to unit scale the equation is $\ln \mathbf{y} = \ln \beta_0 + \beta_1 \ln \mathbf{x} + \boldsymbol{\varepsilon}$. Then using the substitutions $\mathbf{y}^* = \ln \mathbf{y}$, $\mathbf{x}^* = \ln \mathbf{x}$, and $\boldsymbol{\beta}_0^* = \ln \beta_0$ yields a standard OLS equation $\mathbf{y}^* = \boldsymbol{\beta}_0^* + \beta_1 \mathbf{x}^* + \boldsymbol{\varepsilon}$. Thus, the values of $\boldsymbol{\beta}_0^*$ and β_1 can be estimated via ordinary least squares and then transformed to derive the values for β_0 , \mathbf{x} , and \mathbf{y} in the log space equation. For example, $\boldsymbol{\beta}_0^* = \ln \beta_0$ implies that $\beta_0 = e^{\boldsymbol{\beta}_0^*}$. Unfortunately, transforming the OLS estimate of log space yields a biased estimate of the parameter β_0 .

In (Goldberger 1968) the author demonstrated that the process above for a log-linear equation with multiplicative error term yields a biased estimate for the parameter β_0 . Fortunately, the other parameter estimate β_1 is still unbiased. Goldberger provides an equation for estimating the term a that is unbiased (as well as being minimum-variance). This equation, is shown below and its derivation can be found in (Goldberger 1968). The unbiased estimator of β_0 is equal to $e^{\beta_0^*} F$ where:

$$F = \sum_{j=0}^{\infty} \frac{f_j (cw)^j}{j!},$$

And,

$$c = 0.5m \text{ where } m \text{ is the unscaled variance of the estimate for } \beta_0$$

$$w = s^2 \text{ where } s \text{ is the residual variance from least squares estimation}$$

$$f_j = \frac{(0.5v)^j \Gamma(0.5v)}{\Gamma(0.5v + j)} \text{ where } v = n - k - 1 \text{ degree of freedom and } \Gamma \text{ is the Gamma function}$$

$$\Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx$$

It should be noted that this is a theoretical calculation for the value of F . The denominator $j!$ grows rapidly and so the equation for F quickly converges in j . In practice, the summation of terms is usually only taken out to $j = 4$ or 5 .

Goldberger, Arthur S., "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica*, Vol 35, 1968, pp. 464-472.

A.4.3.1.2 PING Factor

Hu, Dr. Shu-Ping, "The Impact of Using Log-Error CERs Outside the Data Range and PING Factor," 5th Joint Annual ISPA/SCEA Conference, Broomfield, CO, 14-17 June 2005.

A.4.3.2 Unbiased

The Gauss-Markov Theorem states that in the classical linear regression model, the least squares estimator is the minimum variance linear unbiased estimator of the coefficient. An unbiased estimator has an expected value for its sample distribution that reflects the value of the population parameter. If many samples of size n are drawn from the overall population, the resulting coefficients would generally reflect the true value of the coefficient based on the full population. By contrast, a biased estimator does not reflect the true value of the coefficient based on the full population.

The property of unbiasedness holds true in the numerical scale in which OLS regression is done. If a log-linear transform is used to linearize a functional form, bias is introduced by the act of conversion to unit space, as described in Section 5.1, though the median of the distribution remains unbiased in this conversion. [Restricted Least Squares](#) regression results in unbiased coefficients as long as the restriction holds true in the population (which is usually not the case in practice). In the case of ridge regression, the estimated coefficient is biased by the mechanical perturbation introduced with this process.

OLS also has properties that hold in very large sample sizes. A large sample property of OLS is that it is consistent. A consistent estimator has a sampling distribution that converges to the true population parameter as n approaches infinity.

Note "unbiased" refers to a property of the mean of the sampling distribution for the estimator, whereas consistency refers to the convergence of the sample distribution to the true population parameter as n approaches infinity.

Omitting key driving variables in the hypothesized model may result in both biased and inconsistent estimators of the remaining variables. This is why a good, well thought out hypothesis tied to the underlying process being modeled is more important than randomly fitting models solely on the basis of the best statistical fit.

A.4.4 Generalized Linear Model (GLM)

A.4.4.1 Generate GLM CER

Methods of least squares, such as [3.3.1 Ordinary Least Squares \(OLS\)](#) and [3.3.2 Generalized Least Squares \(GLS\)](#) can be fairly restrictive in the sense that they depend on a symmetrical error distribution and operate by minimizing the sum of squared errors of the model. A [Generalized Linear Model \(GLM\)](#) is a generalization of the standard linear model allowing for non-normal error distributions, such as lognormal (see Section [4.2.1.4 Normality of Errors](#)), and limited non-linear function forms (see Section [4.2.1.5 Linearity](#)), such as the [2.8.2 Power Functional Form](#) and [2.8.3 Exponential Functional Form](#). Of particular interest to CER construction, GLM provides the flexibility to directly fit a lognormal error term (or approximation of) and power and exponential models without having to first transform the data.

[Appendix B Maximum likelihood estimation for Regression of Log Normal error \(MRLN\) Summary](#) provides an overview on a variation of this method.

GLM expresses a response whose mean is a function of a linear predictor, and an additive error term that follows a distribution belonging to the exponential family (Appendix [A.2.2.3](#)). The model has many convenient properties analogous to those of OLS, but with added complexities. To accommodate non-normal error distributions, GLM utilizes [Maximum Likelihood Estimation \(MLE\)](#) (Appendix [A.4.7.2](#)). The error distribution assumption provides a systematic framework to conduct inference and to determine significance of the results. The parameters may have practical, meaningful values, depending on the specific form of the model. However, the coefficient estimates typically do not have a closed-form solution. Solving for the coefficients by maximizing the likelihood function of the model requires an algorithm and many software packages are able to accommodate the GLM without a problem. Additionally, under certain conditions, statistical inference properties of the GLM are preferable to those of both the [3.3.3 Log-Linear model](#) and more generalized [3.3.5 Non-linear Least Squares \(NLS\)](#) forms.

There are several common specific applications of GLM used to solve specialized regression problems. Binary response variables are predicted using logistic regression. Count data are often modeled using Poisson regression. While beyond the scope of this guide, both are common enough to be aware.

van der Vaart, A.W., [Asymptotic Statistics](#), page 234.

A.4.4.2 GLM Model

Below is the general statistical formulation of the GLM model. This generalization is valid for any error distribution that is a member of the exponential family, which includes the normal and lognormal distributions. In addition, the selection of the link function is flexible. The link function, g , is the relationship between the mean, or expected value, of the response (denoted $E[\mathbf{y}]$) to the linear predictor. Many different forms of the GLM exist, but most applicable to cost may be those with either normal or lognormal error, and a log link function. The first part of the statement expresses that the response variable, or vector, \mathbf{y} , is equal to a function of the matrix of linear predictors, \mathbf{X} , multiplied by the coefficient variables, or vector, $\boldsymbol{\beta}$, plus some random error, $\boldsymbol{\varepsilon}$. The second part of the statement indicates the assumption that the errors are all independently and identically distributed according to some exponential family distribution specified by some set of parameters, $\boldsymbol{\theta}$.

Since a function of the response is linear with respect to the predictors, $g(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and,

$$\mathbf{y} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim \text{Exp Family}(\boldsymbol{\theta})$$

The GLM supports specification of the variance as well, done as some function $h(x)$ of the mean,

$$\begin{aligned} E[\mathbf{y}] &= g^{-1}(\mathbf{X}\boldsymbol{\beta}) \\ \text{Var}[\mathbf{y}] &= h(E[\mathbf{y}]) \\ &= h(g^{-1}(\mathbf{X}\boldsymbol{\beta})) \end{aligned}$$

Unlike the link function for the mean, the model does not require a specification of the variance. It can be taken to be the identity, $h(x) = \mathbf{I}$, which is an assumption of constant variance. However, the formulation as a function of the mean does lend itself well towards fitting multiplicative type error.

Suppose,

$$g(x) = x \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

These conditions satisfy the requirement that the error is distributed according to the exponential family and specifies the identity as the link function. The GLM model now becomes,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

This special case looks identical to OLS and has a closed-form solution. In fact, these coefficient estimates are the same as those generated by OLS, and thus the least squares and maximum likelihood estimates of the coefficients of the linear model with the normality assumption are equivalent.

Now suppose,

$$g(x) = \ln(x) \text{ and } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

Thus, $g^{-1}(x) = e^x$, and again, the error satisfies the requirement of GLM. The GLM model now is,

$$\mathbf{y} = e^{\mathbf{X}\boldsymbol{\beta}} + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

This is an example of a non-linear form fit by a GLM. GLM is transforming the mean of the response in order to fit the estimate. The same model fit with Log-Linear regression transforms each individual response in order to fit the estimate. As a result, GLM fits directly in the unit space, and no messy transformations or factors are required to interpret the results. While the form does not appear to resemble either the [Power Functional Form](#) or [Exponential Functional Form](#) introduced in Section 2.8, some algebraic manipulation can show that the forms are in fact equivalent. Consider the simple case with one independent variable,

$$\begin{aligned} E[y] &= e^{X\beta} \\ &= e^{\beta_0 + \beta_1 x} \\ &= e^{\beta_0} \cdot e^{\beta_1 x} \\ &= \beta'_0 \cdot e^{\beta_1 x} \end{aligned}$$

$$\text{And, } \beta'_0 = e^{\beta_0}$$

This form represents the exponential model. Now consider the same simple case but with a log transform applied to x ,

$$\begin{aligned} E[y] &= e^{X\beta} \\ &= e^{\beta_0 + \beta_1 \ln(x)} \\ &= e^{\beta_0} \cdot e^{\beta_1 \ln(x)} \\ &= e^{\beta_0} \cdot e^{\ln(x^{\beta_1})} \\ &= \beta'_0 \cdot x^{\beta_1} \end{aligned}$$

$$\text{And, } \beta'_0 = e^{\beta_0}$$

This form represents the power model.

The statement of the model explicitly states the assumptions made when conducting GLM. These are analogous to the four OLS assumptions:

- (1) *Independence of errors.* The errors, $\boldsymbol{\varepsilon}$, follow an exponential family distribution specified such that no covariance exists between the errors of each observation.
- (2) *Variance Specification.* The errors, $\boldsymbol{\varepsilon}$, come from the function specified by $h(g^{-1}(\mathbf{X}\boldsymbol{\beta}))$.
- (3) *Distributional Assumption.* The error term follows the specified exponential family distribution.
- (4) *Linearity in the Parameter Space.* The relationship between the predictors, \mathbf{X} , and the parameters, $\boldsymbol{\beta}$, is linear with the correct link function, $g^{-1}(\mathbf{X}\boldsymbol{\beta})$.

Under these assumptions, the results of GLM are asymptotic ([Appendix A.3.3.1.2 Small Data Sets – Asymptotic Results](#)), meaning that the coefficients are optimal in large samples. To fit this regression model by method of maximum likelihood, find values for the coefficient vector $\boldsymbol{\beta}$ that maximize the objective function. The objective function is dependent on the distributional assumption, and is essentially the likelihood function of the specified distribution's parameters, $\boldsymbol{\theta}$, dependent on the known data, \mathbf{X} and \mathbf{y} . Its generic expression is,

$$\arg \max_{\beta} L(\theta | X, y)$$

A.4.4.3 GLM Application

GLM is a less convenient model because the problem rarely has a closed-form solution. The Iteratively Reweighted Least Squares (IRLS) (Appendix [A.4.7.1](#)) technique was developed specifically to solve this problem, and does so efficiently. Standard approaches, such as the popular Gauss-Newton and Levenberg-Marquardt algorithms (more details provided in Section [3.3.5](#)) are able to implement IRLS. Many statistical software packages have GLM built in and automatically produce the regression results and relevant diagnostics. These results are often displayed in a way analogous to the outputs of OLS. To run the analysis, it is required to enter the data along with the desired link function and distributional assumption.

A.4.4.4 GLM Example

Consider sample data with one independent variable *Weight* and dependent variable *Cost*. After viewing a scatter Plot of the data, **Figure 85**, a Power Model, $y = \beta_0 x^{\beta_1} + \epsilon$, is fit to the data by utilizing the GLM. This is done by log transforming $x = \text{Weight}$ and then fitting the GLM with the log link function, $g(x) = \log(x)$ and Normal (or Gaussian) error. **Figure 86** displays the GLM regression analysis outputs on the transformed data.

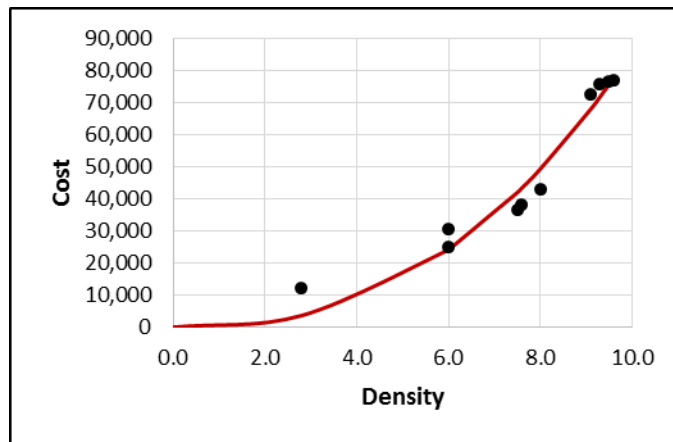


Figure 85: GLM Regression Model Scatter Plot

Coefficients

	Estimate	Std. Error	t-value	Pr(> t)
(Intercept)	5.624	0.711	7.91	0.0000
log(Density)	2.492	0.326	7.65	0.0001

Residual deviance: 333569671 on 8 degrees of freedom

R-squared GLM:
0.9379

Adjusted R-squared GLM:
0.9302

Analysis of Deviance Table

	Df	Deviance
Model	1	5.04E+09
Residuals	8	3.34E+08
Null (Total)	9	5.38E+09

Figure 86: GLM Regression Output

The statistical package R generated these results and the format may vary in appearance by software package, though all should contain the same basic information.⁷⁵ **Figure 85** is a common visual plotting of *Density* on the *x*-axis and *Cost* on the *y*-axis, with the fit regression line going through the data. The appearance of results displayed in **Figure 86** are now distinctly different to those returned by OLS, seen in [Figure 25](#). The first table of coefficients is a standard output showing the estimated values for the regression equation, standard errors, and t-tests for significance. This is the model in the transform space, with only *Density* transformed. Recall that the model is solving for $E[y] = e^{x\beta}$ with *x* being log transformed. Thus, the regression equation is,

$$Cost = e^{5.624 + 2.492 \cdot \ln(Density)}$$

Algebraic manipulation of this form yields the power model,

$$\begin{aligned} Cost &= e^{5.624} \cdot Density^{2.492} \\ &= 277.051 \cdot Density^{2.492} \end{aligned}$$

Next, there are several regression statistics that are common such as the R-squared GLM and the residual deviances. GLM does not have sums of squares, but rather now has an analogous metric called deviances. Many metrics relevant in the linear model are no longer applicable in the non-linear setting. However, the

⁷⁵ R Core Team, “R: A language and environment for statistical computing. R Foundation for Statistical Computing,” Vienna, Austria, 2013. URL <http://www.R-project.org>.

results are analogous in appearance and the approach for much of the analysis is the same as with OLS. The final table is the Analysis of Deviance table as shown in **Figure 86**, which is a standard view for significance and key diagnostic values in a regression model. Again, this table has significant deviation from OLS.

A.4.4.5 Validate CER (Assumptions)

Section [3.3.4](#) introduced the [Generalized Linear Model \(GLM\)](#), with more details covered in Appendix [A.4.4 Generalized Linear Model \(GLM\)](#). The assumptions and validation of the GLM are different than OLS because GLM is a method of maximum likelihood, not of least squares. However, the same basic principles are required from the data. The model statement explicitly states the assumptions of the analysis. Recalling Section [3.3.4](#), the GLM model is,

$$E[\mathbf{y}] = g^{-1}(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\varepsilon} \text{ where } \boldsymbol{\varepsilon} \sim \text{Exp Family}(\boldsymbol{\theta})$$

$$\text{Var}[\mathbf{y}] = h(g^{-1}(\mathbf{X}\boldsymbol{\beta}))$$

A.4.4.6 Residuals

The residual error is the difference between the actual value and the predicted value. This is the raw residual, and for the GLM is,

$$\mathbf{e}_{glm} = \mathbf{y} - g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}})$$

Similarly to OLS, these are not the correct residuals to use; standardization is required. This process is more complex for GLM, being dependent on the final iteration of the numerical algorithm. As a result, it is best to be aware of the standardization process as defined for OLS, but use the standardized residuals generated by the statistical software.

A.4.4.6.1 Independence of Errors

The assessment of the independence of errors assumption is largely the same as with OLS, but with the residual plots constructed on the GLM residuals. See Section [4.2.1.2 Independence of Errors](#).

A.4.4.6.2 Variance Specification

The assumption for GLM is not necessarily that the errors have constant variance, but that specification of the variance function, $h(g^{-1}(\mathbf{X}\boldsymbol{\beta}))$ is correct. Verify this much in the same way as the homoscedasticity assumption; by examining the residual plot on the GLM residuals. See Section [4.2.1.3 Homoscedasticity](#). If violated, consider a new functional form for h .

A.4.4.6.3 Distributional Assumption

The assumption for the GLM model for any exponential family distribution, not necessarily normality. The software generates a similar Q-Q plot as before, but using the assumed distribution. Validate the plot in the same way as introduced in Section [4.2.1.4 Normality of Errors](#).

A.4.4.6.4 Linearity in the Parameter Space

This step verifies both the linearity in the parameter space and the link function $g^{-1}(\mathbf{X}\boldsymbol{\beta})$. Look for evidence that the model specification is correct. Much like OLS, use a scatter plot in the single variable

case; and a residual plot and predicted versus actual plot in the multiple predictor case. Look for the same symptoms as with OLS in Section [4.2.1.5 Linearity](#).

A.4.5 Non-linear Least Squares (NLS)

van der Vaart, A.W., [Asymptotic Statistics](#), page 57.

A.4.5.1 CO\$TAT Application

Generating the coefficients for NLS models can be accomplished via the non-linear analysis methods available within CO\$TAT (part of the ACEIT software suite). Using CO\$TAT, virtually any functional form can be defined and the software will calculate estimates of the parameters for that functional form. In addition to defining the functional form, options involving the optimization procedure used to find the parameters can be set. The following discussion on CO\$TAT application heavily leverages methodologies found in [A.4.8 Minimum-Unbiased-Percentage-Error \(MUPE\)](#), but is applicable to fitting a NLS model in the more general sense as discussed in Section [3.3.5 Non-linear Least Squares \(NLS\)](#).

When using the Non Linear analyses built into CO\$TAT to generate the parameters a number of option settings must be specified. Most of these options include details pertinent to the optimization routine used to generate the coefficient, but still others involve a description of the functional form for which the parameters are to be derived. A thorough explanation of these options and the necessary inputs from the analyst can be found in the CO\$TAT help file. However, a brief discussion of some of the options is included here.

Functional Form

When entering the functional form, initial values for the parameters to be generated need to be included. As the function being minimized is non-linear, there is a potential for the generated parameters to be highly influenced by the choice of these initial values. With that in mind, the parameters should be initialized via a best guess and not simply the default value (usually 1). Values of a “best guess” for each parameter can be derived from inspection of scatter plots, similar CERs, or even input from subject matter experts.

Error Metric

In addition to the functional form, the error term for the model needs to be specified. The options for error term are Additive, Multiplicative, and Minimum Unbiased Percentage Error (MUPE). Further information on these error terms can be found in Book and Young’s 1997 paper “General-Error Regression for Deriving Cost-Estimating Relationships.” The choice of error term determines which error metric the algorithm tries to minimize. Choosing an additive error term generates coefficients that minimize the error function $\sum_{i=1, \dots, n} (\varepsilon_i)^2$.

A multiplicative error term is defined by $\varepsilon_i = y_i / f(x_i; \boldsymbol{\beta})$. Thus, $f(x_i; \boldsymbol{\beta})$ close to y_i yields a residual ε_i close to 1 rather than 0 in the case of additive errors. Choosing a multiplicative error term generates coefficients that minimize the squared deviation of the errors from 1. In other words, the error function minimized when the Multiplicative error is chosen is $\sum_{i=1, \dots, n} (\varepsilon_i - 1)^2$. This method is known as minimum percentage error (MPE) regression. Book and Young’s paper notes that the estimates

minimizing this error function are positively biased. For this reason a third, unbiased, error function is available in CO\$STAT.

Due to the bias inherit in estimating parameters by minimizing the error function $\sum_{i=1, \dots, n} (\varepsilon_i - 1)^2$, the method of minimum unbiased percentage error (MUPE) was developed. This method is a special case of iteratively reweighted least squares regression, and more information on the topic can be found in Appendix [A.4.7.1 Iteratively Reweighted Least Squares \(IRLS\)](#) and Appendix [A.4.8 Minimum-Unbiased-Percentage-Error \(MUPE\)](#).

Optimization Method

There are a number of settings for the minimization algorithm. The available algorithms are Modified Marquardt (more formally known as the “Levenberg and Marquardt” method), Downhill Simplex, and the Gauss-Newton method. Unfortunately there is no single “best” method and none of these methods can guarantee a globally minimal solution to minimizing the chosen error metric. However, the Marquardt method is the default and it provides a good balance between computational efficiency (speed) and robustness (i.e., it is capable of finding global minimums despite poor parameter initialization).

After selecting the optimization method, settings such as the maximum number of iterations, the convergence tolerance, and differential delta for approximation of derivatives need to be chosen. For convergence tolerance, a smaller value will calculate the minimum value to a higher degree of precision (i.e., more decimal places), however, it may take longer to do so. Similarly, a higher number of iterations will allow the algorithm to work towards convergence longer (which in turn provides more accuracy) but will require more time to do so. Lastly, the delta for approximating derivatives directs the algorithm towards a minimum more precisely, but does so with an increased probability of overlooking a globally minimal solution. In any case, the defaults for these settings are a good starting point and should only be manipulated if (a) the algorithm is not converging at all, or (b) the solutions to which the algorithms converged are unreasonable.

The “best” algorithm to use and the “best” settings to employ are highly dependent upon not only the functional form and the number of parameters to be estimated, but also on the data set being analyzed. With this in mind, it is recommended that all three algorithms be used to generate parameters in the case that one method outperforms the others. Along those same lines, a variety of runs within each method using different initializations for each parameter will help reduce the chance of failing to identify a globally minimal solution. Lastly, if the algorithm is not converging in the allotted iterations, increases to the number of iterations, or decreases to the convergence tolerance might be in order. Regardless, the analyst should attempt multiple varied approaches for arriving at parameter estimates that minimize the specified error metric.

A.4.5.2 Excel Application

As an alternative to CO\$STAT, the SOLVER function in Excel can be used to generate parameters for NLS. Using SOLVER to minimize the error metric can be a cumbersome task as it requires not only building the model in the Excel worksheet, but also selecting the parameters that will be used by the SOLVER routines. In statistical packages (including CO\$STAT) the generation of statistics for NLS models is fully automated. Using SOLVER requires the analyst to perform all required calculations themselves.

While Excel provides many relevant functions and operations that aid in these calculations (e.g., “T.TEST”), building these calculations into an Excel worksheet are tedious and require a strong understanding of matrix algebra and statistics. For example, standard error estimates can be calculated by using the Jacobian (or Fisher’s Information) matrix at the last iteration of the algorithm. This information is lost with SOLVER and would require expert knowledge of the mathematics in order to duplicate. In general, a knowledge of statistics far greater than that provided in this guide would be required.

Despite these drawbacks, most analysts are comfortable working in Excel. For this reason, SOLVER remains a viable option for those adventurous analysts seeking more control and transparency into parameter generation. However, its use should be cautioned, if not discouraged, since it is very easy to solve for parameter coefficients without fully understanding what is being done and without any of the proper precautions and [4.3 Model Diagnostics](#).

A.4.6 Ridge Regression

In matrix notation, useful for the multiple predictor (as well as the single predictor) case,

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'y$$

And,

$$\hat{\sigma}^2 = \frac{y'(I - X(X'X + \lambda I)^{-1}X')y}{n - (k + 1)}$$

Hoerl, Arthur E., and Robert W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics*, Vol 12, No 1, February 1970. Authors from University of Delaware and I.E. du Pont de Nemours & Co. Published by American Statistical Association and American Society for Quality. Available at <http://www.jstor.org/stable/1267351> or <http://math.arizona.edu/math574m/Read/Ridge.pdf>.

A.4.7 Mathematical/Numerical Techniques

A.4.7.1 Iteratively Reweighted Least Squares (IRLS)

Iteratively Reweighted Least Squares is a mechanism for calculating parameter estimates for general error regression models. In each iteration step, the parameter estimates are updated based on minimizing the weighted squared error

IRLS is a general technique, of which MUPE is a specific application.

Daubechies, Ingrid, Ronald Devore, Massimo Fornasier, and C. Sinan Güntürk, “Iteratively Re-weighted Least Squares Minimization for Sparse Recovery.” Available at <https://arxiv.org/pdf/0807.0575.pdf>.

A.4.7.2 Maximum Likelihood Estimation (MLE)

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 615-630.

A.4.7.3 The Method of Lagrange Multipliers

Suppose that $f(x, y, z)$ and $g(x, y, z)$ are differentiable and $\nabla g \neq \mathbf{0}$ when $g(x, y, z) = 0$. To find the local maximum and minimum values of f subject to the constraint $g(x, y, z) = 0$ (if these exist), find the values of x, y, z and λ that simultaneously satisfy the equations

$$\nabla f = \lambda \nabla g \text{ and } g(x, y, z) = 0.$$

For functions of two independent variables, the condition is similar, but without the variable z .

A.4.7.4 Bootstrap

Hass, Joel, George Thomas, Jr., and Maurice D. Weir, University Calculus, page 769.

Dunlop, Dorothy D. and Ajit C. Tamhane, Statistics and Data Analysis: From Elementary to Intermediate, pages 597-601.

A.4.8 Minimum-Unbiased-Percentage-Error (MUPE)

Some key references on MUPE and other GERM techniques include:

Hu, Dr. Shu-Ping, “The Minimum-Unbiased-Percentage-Error (MUPE) Method in CER Development,” 3rd Joint Annual ISPA/SCEA International Conference, Vienna, VA, 12-15 June 2001.

Hu, Dr. Shu-Ping Hu, and Alfred Smith, “Why ZMPE When You Can MUPE?,” IPISA / SCEA Annual Conference and Training Workshop, New Orleans, LA, 12-15 June, 2007.

General Error Regression Models (GERM)

General Error Regression is a broad term used to describe any regression methodology involving functional forms that cannot be transformed into a linear functional form or one or more independent variables. In practice, estimating the coefficients of these models often requires the use of an optimization routine such as SOLVER in MS Excel.

It is important to remember that the assumptions of general regression models are unlikely to have the same assumptions as OLS. Additionally, an alternative set of tests for statistical significance of the model or model parameters may need to be employed.

Book, Stephen A., and Lao. N., “Minimum-Percentage-Error Regression under Zero-Bias Constraints,” *Proceedings of the Fourth Annual U.S. Army Conference on Applied Statistics, 21-23 October 1998*, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.

Book, Stephen A., and Philip H. Young, “General-Error Regression for Deriving Cost-Estimating Relationships,” *The Journal of Cost Analysis*, Vol 14, 1997 pp.1-28.

Book, Stephen A., “Modern Techniques for Multiplicative-Error Regression,” IPISA / SCEA Annual Conference and Training Workshop, New Orleans, LA, 12-15 June, 2007.

Zero-Percentage Bias (ZPB) Minimum Percentage Error (ZMPE)

While advocating MUPE over ZMPE, this paper has key information on the implementation and evaluation of both methods:

Hu, Dr. Shu-Ping, and Alfred Smith, “Why ZMPE When You Can MUPE?,” IPISA / SCEA Annual Conference and Training Workshop, New Orleans, LA, 12-15 June , 2007.

A.4.8.1 GERM Significance

Anderson, Tim, “A Distribution-Free Measure of the Significance of CER Regression Fit Parameters Established Using General Error Regression Methods,” IPISA / SCEA Professional Development and Training Workshop, St. Louis, MO, 2009.

A.4.8.2 GERM Uncertainty (Bootstrapping)

Book, Stephen A., “Prediction Bounds for General-Error-Regression Cost-Estimating Relationships,” *Journal of Cost Analysis and Parametrics (JCAP)*, Vol 5 No 1, 2012.

A.4.9 Advanced Regression Methodologies

A.4.9.1 Restricted Least Squares

Most often the exact relationship among the parameters, e.g., $\beta_1 + \beta_2 = 1$ or $\beta_1 + \beta_2 = 1.25$ is unknown. Instead, the desire is to verify certain prior information on the parameters. The interest is in testing a general hypothesis before running the restricted least squares. In OLS regression, the null hypothesis for the F-test is that none of the regression coefficients are statistically significant (i.e., all are equal to zero). This is shown compactly as:

$$H_0: \boldsymbol{\beta} = \mathbf{0}$$

For this more general problem of testing relationships, the null hypothesis generalizes as follows:

$$H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{a}$$

where C is used to denote a given matrix of r constraints, $\boldsymbol{\beta}$ is a vector of coefficients, and a is a vector specified by the hypothesis. The test statistic is then given by,

$$F = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}}_1 - \mathbf{a})' [\mathbf{C}(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{C}']^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_1 - \mathbf{a})/r}{SSE/n - k - 1}$$

It can be shown that the above equation follows an F distribution with r and $n - k - 1$ degrees of freedom when H_0 is true. Note that \mathbf{X}_1 is the centered design matrix (i.e., each independent variable has its mean subtracted from each observation), k is the number of independent variables, and n is the sample size.

See the white paper “A Priori Knowledge in Linear Regression Analysis,” which draws on the below sources, for more details.

Wooldridge, J. M., *Introductory Econometrics: A Modern Approach* (Third Edition), Thomson South-Western, 2006.

Draper, Norman R. and Smith, Harry, Applied Regression Analysis (Third Edition), John Wiley & Sons, Inc., 1998, Section 9.5 “Restricted Least Squares”

Jaggi, S. and Sivaramane, N., “Restrictions in Regression Model,” Indian Agriculture Statistics Research Institute, India.

A.4.9.2 Principal Component Analysis

Principal Component Analysis (PCA) is a regression method commonly used in the presence of high multicollinearity. PCA makes use of the eigenvalue decomposition of the $\mathbf{X}'\mathbf{X}$ matrix to perform regression. Eigenvectors are defined as being orthogonal to one another. That is, there is no correlation between the eigenvectors, or principal components. Thus, PCA regresses on the linearly independent principal components. The regression may be analyzed and variable selection can be performed without inflated variance due to multicollinearity. The model can then be transformed back into the original variable space, or the full regression may be run from scratch using the down-selected variable set from the PCA.

PCA as an estimator is highly related to [Ridge Regression](#). In fact, Ridge Regression smoothly shrinks the estimates, while PCA proceeds more discretely in steps. The topic has been studied in detail and many statistical packages are capable of running PCA.

Wold, Svante, Kim Esbensen, and Paul Geladi, “Principal Component Analysis,” *Chemometrics and Intelligent Laboratory Systems*, Elsevier Science Publishers B.V., Amsterdam, Vol 2, pages 37-52, 1987.

Abdi, Herve and Lynne J. Williams, “Principal Component Analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, 2010. Available at <https://pdfs.semanticscholar.org/53b9/966a0333c9c9198cdf03efc073e991647c12.pdf>

A.4.9.3 Mixed Models

When a regression is performed, it is assumed that all independent variables are fixed and therefore the model has a single random error term. This is what is known as a fixed effects model. In certain situations, this assumption is inappropriate. The full dataset may be a hierarchical structure of different populations, defined by a categorical variable. Each of these population subsets of the data are in need of its own error term. Mixed Models allow for a second (or multiple) error terms to be integrated into the model. This can produce superior results, as only the overall error (SSE) is used for inferential purposes. A model with only random variables and no fixed variables is known as a random effects model.

Setlman, Howard J., MD and PhD, *Experimental Design and Analysis*, Carnegie Mellon University, Department of Statistics & Data Science, Pittsburgh, PA, September 2015. See Chapter 15. Available at http://www.stat.cmu.edu/_hseltman/309/Book/Book.pdf

A.4.9.4 General Estimating Equations

General Estimating Equations (GEE) are a generalization of the [Generalized Linear Model \(GLM\)](#) that allows for correlation between responses. Therefore, GEE is to GLM as Generalized Least Squares (GLS) is to OLS.

van der Vaart, A.W., *Asymptotic Statistics*, page 401.

A.4.9.5 LASSO and the Elastic Net

The Least Absolute Shrinkage and Selection Optimizer (LASSO) is a shrinkage estimator similar in concept to Ridge Regression. To explain the concept, the idea of a norm is introduced. Simply stated, a norm is a functional representation of size for a vector. The following are two common measures:

$$\begin{aligned} L_1 &= \|\mathbf{x}\|_1 \\ &= \sum_i |x_i| \end{aligned} \qquad \begin{aligned} L_2 &= \|\mathbf{x}\|_2 \\ &= \sum_i x_i^2 \end{aligned}$$

Ridge issues the constraint on the L_2 norm of the coefficients in the model, $\|\boldsymbol{\beta}\|_2 < c$. For this reason, Ridge is also commonly referred to as L_2 regularization. The result is the objective function,

$$\arg \min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_2)$$

LASSO issues the constraint on the L_1 norm of the coefficients in the model, $\|\boldsymbol{\beta}\|_1 < c$. For this reason, LASSO is also commonly referred to as L_1 regularization. The result is the objective function,

$$\arg \min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_1)$$

The absolute value function creates a point on the objective function where no derivative exists. As a result, LASSO has no closed-form solution, but there are efficient algorithms to solve the problem. LASSO has a convenient property that as the restriction parameter λ becomes larger (and therefore c becomes smaller), coefficient estimates move exactly to zero. Ridge regression simply shrinks the parameters, while LASSO performs automatic variable selection as well.

While appealing, in the presence of severe multicollinearity Ridge regression outperforms LASSO. Another hybrid type model exists which shares properties of both the Ridge estimator and the LASSO estimator, called the elastic net. The elastic net has an objective function defined as,

$$\arg \min_{\boldsymbol{\beta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda_1 \|\boldsymbol{\beta}\|_2 + \lambda_2 \|\boldsymbol{\beta}\|_1)$$

The elastic net maintains properties of both Ridge regression and LASSO. The estimator performs well in the presence of multicollinearity, and also sends coefficients to be exactly equal to zero.

Elastic net is popular in the setting where $n \ll p$, that is, where there are many more parameters of interest than observations.

Tibshirani, Robert, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58.1 (1996): 267-88, JSTOR Web, 07 Jan. 2015.

A.5 Influence Diagram

An influence diagram is a graphical depiction of a web of variables and their interrelationships. An arrow from one variable to another, labeled with either a plus sign or a minus sign, indicates positive or negative correlation, respectively.

Figure 87 illustrates a number of technical, schedule, and cost variables associated with Software Maintenance. There is a mixture of normalization steps (where factors color-coded in green may be applied) and hypothesized cost-driving relationships, indicated with a single-headed arrow and a plus or minus sign (positive or negative correlation, respectively). This latter component typifies an influence diagram, with additional discussion in Section [1.3 Cost Estimate Purpose and Scope](#).

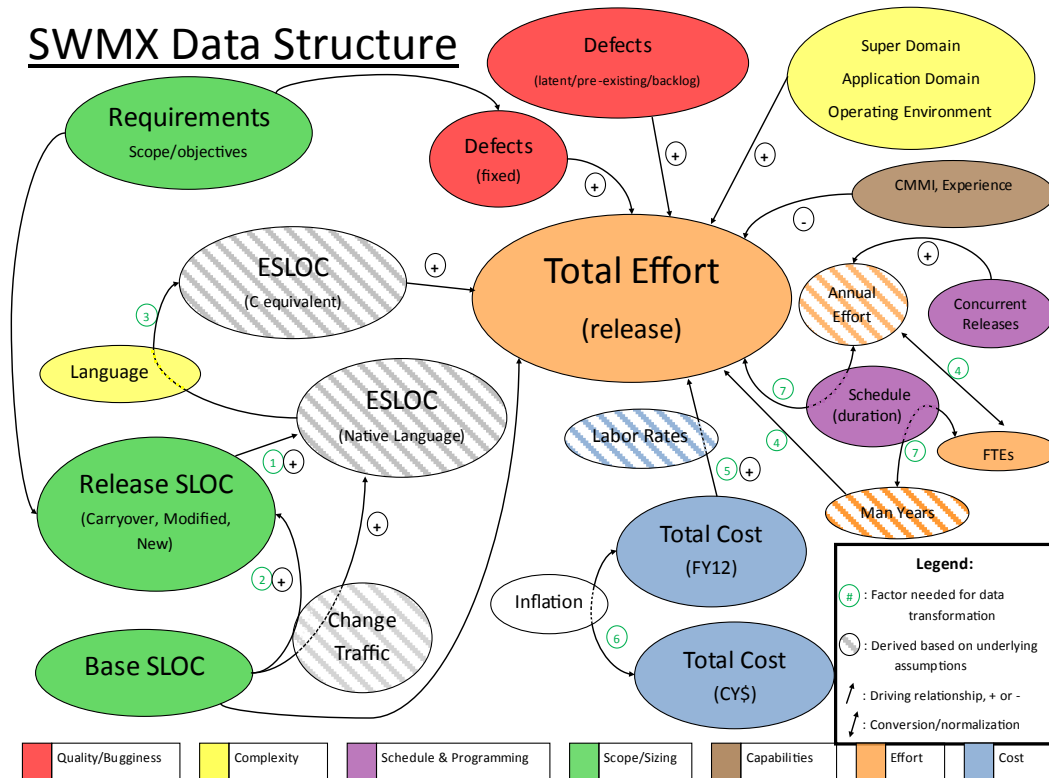


Figure 87: Example of Data Normalization and Hypothesized Relationships Between Variables

Clemen, Robert T., and Terence Reily, *Making Hard Decisions with Decision Tools* (Third Edition), published by South-Western and Cengage Learning, Mason, OH, 2014. See Chapter 3. ISBN 978-0-538-79757-3.

APPENDIX B MAXIMUM LIKELIHOOD ESTIMATION FOR REGRESSION OF LOG NORMAL ERROR (MRLN) SUMMARY

Nonlinear equations are commonly used in CERs for government projects⁷⁶. Three primary methods commonly used in the cost estimating community for calculating nonlinear CERs from historical data are log-transformed ordinary least squares (LOLS), iteratively reweighted least squares/minimum unbiased percentage error (IRLS/MUPE), and minimum percent error with zero percent bias (ZMPE). Of these three, LOLS is the oldest, primarily because it is not computationally intensive. Indeed, the parameters can be calculated using a hand calculator.

The pedigree and simplicity of LOLS have led to the perception that this method is antiquated, and should be replaced by more modern, computationally intensive methods such as IRLS/MUPE or ZMPE. In log-transforming the data we are estimating “log-dollars.” The transformed estimate is unbiased in log-space, but is biased once we transform the equation back to unit space. LOLS is estimating the median of a lognormal which is less than the mean, so LOLS is a biased estimator of the mean. The bias is low, so if we are trying to estimate the mean, LOLS will underestimate that value. The ZMPE method was developed as an alternative to LOLS.

There is strong evidence for why CER residuals should be lognormally distributed, both theoretical and empirical. Changes in costs over time are proportional to prior costs. This makes sense. Cost is more likely to increase than decrease over time, as evidenced by numerous studies on cost growth that show that over 80% of government projects experience cost growth, and on average increase by over 50%. Thus when we talk about cost changes, we almost always mean cost increases. Cost increases often do not result in funding increases in the short term due to funding constraints. Thus cost increases will result in longer schedules. Longer schedules imply a longer period in which the personnel devoted to a project will charge to that particular project. Larger projects have more personnel assigned to a project, meaning that increases in cost will result in a proportional increase in cost. What we have described is a multiplicative Central Limit Theorem for cost risk meaning that cost risk is approximately lognormally distributed.

The lognormal is widely used to model risk in other industries, such as health care and property insurance.

Each of the methods in wide use today – LOLS, ZMPE, and IRLS/MUPE – have a connection to maximum likelihood estimation. For the case of lognormally distributed residuals, LOLS is an optimal method for estimating the median.

Since there is strong evidence that CER residuals are lognormally distributed, and there are concerns with LOLS, we propose the use of maximum likelihood on the non-transformed equation to circumvent these issues. This method, which we term maximum likelihood estimation for regression of lognormal error

⁷⁶ This appendix is a summary of the following paper: Smart, Christian, PhD, “Cutting the Gordian Knot: Maximum Likelihood Estimation of Untransformed Lognormal Error”, Director, Cost Estimating and Analysis Missile Defense Agency.

(“MRLN”), is a direct and simple approach. It avoids the trouble of having to transform the data, and then retransform one of the coefficients, and eliminates the concerns about LOLS.

For MRLN, we are using the power equation, with cost as a function of one or more parameters:

$$Y = \beta_0 X_1^{\beta_1} \dots X_p^{\beta_p}$$

The mean of a lognormal density function is

$$e^{\mu + \frac{\theta}{2}}$$

To estimate the mean directly, for the i^{th} observation, we set

$$e^{\mu_i + \frac{\theta}{2}} = \beta_0 X_{i1}^{\beta_1} \dots X_{ip}^{\beta_p}$$

Taking log transformation of both sides of the above equation, we find

$$\mu_i + \frac{\theta}{2} = \ln \beta_0 + \beta_1 \ln X_{i1} + \dots + \beta_p \ln X_{ip}$$

Therefore,

$$\mu_i = \ln \beta_0 + \beta_1 \ln X_{i1} + \dots + \beta_p \ln X_{ip} - \frac{\theta}{2} = \ln \beta_0 + \sum_{j=1}^p \beta_j \ln X_{ij} - \frac{\theta}{2}$$

Recall that the likelihood for a lognormal is given by

$$L(\mu, \theta) = \prod_{i=1}^n \frac{1}{y_i \sqrt{2\pi\theta}} e^{-\frac{(\ln y_i - \mu_i)^2}{2\theta}}$$

For n observations, the log-likelihood is thus (ignoring constants)

$$l(\mu, \theta) = -\frac{1}{2\theta} \sum_{i=1}^n (\ln y_i - \mu_i)^2 - \sum_{i=1}^n \ln y_i - \frac{n}{2} \ln \theta$$

We substitute for μ to obtain

$$l(\beta_0, \beta_1, \dots, \beta_p, \theta) = -\frac{1}{2\theta} \sum_{i=1}^n \left(\ln y_i - \ln \beta_0 - \sum_{j=1}^p \beta_j \ln X_{ij} + \frac{\theta}{2} \right)^2 - \sum_{i=1}^n \ln y_i - \frac{n}{2} \ln \theta$$

Ignoring constants and rearranging we obtain

$$(\beta_0, \beta_1, \dots, \beta_p, \theta) = -\frac{n}{2} \ln \theta - \frac{1}{2\theta} \sum_{i=1}^n \left(\ln y_i - \ln \beta_0 - \sum_{j=1}^p \beta_j \ln X_{ij} + \frac{\theta}{2} \right)^2$$

Taking partial derivatives with respect to the parameters, we obtain

$$\frac{\partial l}{\partial \theta} = -\frac{n}{2\theta} - \frac{n}{8} + \frac{\sum_{i=1}^n (\ln y_i - \ln \beta_0 - \sum_{j=1}^p \beta_j \ln X_{ij})^2}{2\theta^2}$$

$$\frac{\partial l}{\partial \beta_0} = -\frac{\sum_{i=1}^n (\ln y_i - \ln \beta_0 - \sum_{j=1}^p \beta_j \ln X_{ij})}{\beta_0 \theta}$$

For $k = 1, \dots, p$,

$$\frac{\partial l}{\partial \beta_k} = -\frac{\sum_{i=1}^n \ln X_{ik} (\ln y_i - \ln \beta_0 - \sum_{j=1}^p \beta_j \ln X_{ij})}{\theta}$$

There won't typically be a closed form solution for the roots of these equations (unlike OLS), so we will need a numerical iterative routine to solve, such as the Newton-Raphson algorithm. The Newton-Raphson method was published in Joseph Raphson's *Analysis Aequationum Universalis* in 1690. While tedious, the tools to calculate nonlinear least squares have existed before the development of the least squares method by Carl Gauss in the early 19th century.

We can utilize Excel's solver routine to minimize the negative of the log likelihood. We are maximizing a negative value, so instead we minimize the negative of this log-likelihood. That is we minimize:

$$-l(\beta_0, \beta_1, \dots, \beta_p, \theta) = \frac{n}{2} \ln \theta + \frac{1}{2\theta} \sum_{i=1}^n \left(\ln y_i - \ln \beta_0 - \sum_{j=1}^p \beta_j \ln X_{ij} + \frac{\theta}{2} \right)^2$$

This is a single number, so we can use Solver to minimize this value. We have to allow the variance term and the parameter coefficients to vary in order to find this minimum value. One potential set of starting values is to use the average sample value of Y for β_0 , set the other β parameters to equal θ , and set $\theta = 1$.

APPENDIX C CER DEVELOPMENT CHECKLIST

The inspiration for this checklist came from the GAO Cost Estimating and Assessment Guide, GAO-09-3SP, March 2009 and the FAA Guide to Conducting Business Case Cost Evaluations 08 June 2016.

Step	Item	Complete
1	Purpose of the estimate	
	Customer, scope, level of detail identified	
2	Ground Rules & Assumptions	
	Base year and life cycle identified	
	Program schedule developed	
	Other program assumptions documented	
3	Cost estimating plan	
	Team members and roles & responsibilities defined	
	Deliverables & dates specified	
	Resource requirements identified	
	Cost estimating checklist developed	
4	Program definition	
	Technical baseline defined	
	Life cycle support strategy developed	
	Acquisition strategy defined	
5	Cost element structure	
	Work Breakdown Structure (WBS)/Cost Element Structure (CES) identified	
	WBS/CES dictionary developed	
	Operational and engineering insight obtained	
	Estimating method options & data sources by WBS element defined	
6	Data collection and analysis	
	Data collected and validated	
	Data organized and normalized for statistical analysis	
	Univariate analysis on dependent and independent variables completed	
	Analysis to find candidate cost drivers consistent with SME advice completed	
7	Cost estimating relationship (CER) developed and validated	
8	CER cost risk & uncertainty defined	
9	CER documented	

APPENDIX D CORRELATION CRITICAL VALUE TABLES**Table 45: Pearson Product Moment Critical Values**

n	Two-Tailed Probabilities			
	0.1000	0.0500	0.0100	0.001
	One-Tailed Probabilities			
	0.0500	0.0250	0.0125	0.005
5	0.687	0.805	0.878	0.959
6	0.608	0.729	0.811	0.917
7	0.551	0.669	0.754	0.875
8	0.507	0.621	0.707	0.834
9	0.472	0.582	0.666	0.798
10	0.443	0.549	0.632	0.765
11	0.419	0.521	0.602	0.735
12	0.398	0.497	0.576	0.708
13	0.380	0.476	0.553	0.684
14	0.365	0.458	0.532	0.661
15	0.351	0.441	0.514	0.641
16	0.338	0.426	0.497	0.623
17	0.327	0.412	0.482	0.606
18	0.317	0.400	0.468	0.590
19	0.308	0.389	0.456	0.575
20	0.299	0.378	0.444	0.561
21	0.291	0.369	0.433	0.549
22	0.284	0.360	0.423	0.537
23	0.277	0.352	0.413	0.526
24	0.271	0.344	0.404	0.515
25	0.265	0.337	0.396	0.505
26	0.260	0.330	0.388	0.496
27	0.255	0.323	0.381	0.487
28	0.250	0.317	0.374	0.479
29	0.245	0.311	0.367	0.471
30	0.241	0.306	0.361	0.463
40	0.207	0.264	0.312	0.403

Table 46: Spearman’s Rho Critical Values

Two-Tailed Probabilities				
	0.1000	0.0500	0.0100	0.001
One-Tailed Probabilities				
n	0.0500	0.0250	0.0125	0.005
5	0.800	0.900	1.000	
6	0.657	0.829	0.886	1.000
7	0.571	0.714	0.786	0.929
8	0.524	0.643	0.738	0.881
9	0.483	0.600	0.700	0.833
10	0.455	0.564	0.648	0.794
11	0.427	0.536	0.618	0.755
12	0.406	0.503	0.587	0.727
13	0.385	0.484	0.560	0.703
14	0.367	0.464	0.538	0.679
15	0.354	0.446	0.521	0.654
16	0.341	0.429	0.503	0.635
17	0.328	0.414	0.488	0.618
18	0.317	0.401	0.472	0.600
19	0.309	0.391	0.460	0.584
20	0.299	0.380	0.447	0.570
21	0.292	0.370	0.436	0.556
22	0.284	0.361	0.425	0.544
23	0.278	0.353	0.416	0.532
24	0.271	0.344	0.407	0.521
25	0.265	0.337	0.398	0.511
26	0.259	0.331	0.390	0.501
27	0.255	0.324	0.383	0.492
28	0.250	0.318	0.375	0.483
29	0.245	0.312	0.368	0.475
30	0.240	0.306	0.362	0.467
40	0.207	0.264	0.313	0.405

APPENDIX E PARTIAL REFERENCES

1. Gallant (1975) "Nonlinear Regression", *The American Statistician*, 29:2, 73-81
2. Tukey, John W. (1977) "Exploratory Data Analysis", Reading, MA: Addison-Wesley Pub., Print.
3. R. Dennis Cook (1979) "Influential Observations in Linear Regression", *Journal of the American Statistical Association*, 74:365, 169-174.
4. Kvalseth, Tarald O. (1985) "Cautionary Note about R 2." *The American Statistician* 39.4: 279. Web.
5. Willett, John B., and Judith D. Singer. (1988) "Another Cautionary Note About: Its Use in Weighted Least-Squares Regression Analysis." *The American Statistician* 42.3: 236-38. Web.
6. Norman R. Draper, Harry Smith (1998), "Applied Regression Analysis, 3rd Edition", Wiley New York
7. Greene, William H. (2003) "Econometric Analysis", 3rd ed. Upper Saddle River, NJ: Prentice Hall, Print.
8. Goldberg, Andrew S., Touw, Anduin (March 2003) "Statistical Methods for Learning Curves and Cost Analysis", CIM D0006870.A3/1 Rev.,
9. Operating and Support Cost-Estimating Guide, OSD CAIG, October 2007.
10. B. Michener, C. Scarlata, and B. Hames, (January 2008) "Rounding and Significant Figures" U.S. Department of Energy Technical Report NREL/TP-510-42626
11. Rencher, Alvin C. (2008) "Linear Models in Statistics", 2nd ed. New York: Wiley, Print.
12. Government Accountability Office (GAO) Cost Estimating and Assessment Guide (GAO-09-3SP), March 2009
13. Weapon Systems Acquisition Reform Act, Public Law 111-23— 22 May 2009
14. Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. (2009) "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", New York: Springer, Print.
15. DoD Architecture Framework Version 2.02, DoD Deputy Chief Information Officer, August 2010 <http://dodcio.defense.gov/dodaf20.aspx>.
16. Work Breakdown Structures for Defense Materiel Items (MIL-STD-881C), Department of Defense Standard Practice, 03 October 2011
17. Responsibilities and functions, relationships, and authorities of the DCAPE, Department of Defense Directive 5105.84, Director of Cost Assessment and Program Evaluation (DCAPE), 11 May 2012
18. Thompson, Wayne and Patel, Tapan (2013) "Data Mining from A to Z: Better Insights, New Opportunities", SAS Whitepaper
19. Defense Acquisition Guidebook, 16 September 2013
20. "1.3.5. Quantitative Techniques" *Engineering Statistics Handbook*. National Institute of Standards and Technology, n.d. Web. 24 September 2014.
21. Joint Agency Cost and Schedule Risk and Uncertainty Handbook (CSRUH), 16 September 2014, <https://www.ncca.navy.mil/tools/csruh/index.cfm>.

APPENDIX F DATA SETS**F.1 Electronics Example**

Observation	Cost (FY16\$K)	Power (kW)	Cost per Unit Power (\$M/kW)	Aperture (cm ²)	Power per Unit Aperture (kW/cm ²)	FFP (1) or T&M (0)
1	390	10.0	39.0000	8.70	1.149	1
2	200	5.0	40.0000	8.00	0.625	0
3	240	5.2	46.1538	8.20	0.634	1
4	300	7.0	42.8571			0
5	460	12.0	38.3333	9.00	1.333	1
6	560	17.8	31.4607	9.50	1.874	0
7	700	21.0	33.3333	9.20	2.283	0
8	800	25.0	32.0000	9.70	2.577	1
9	500	18.0	27.7778			0

F.2 Cost Improvement Curve Example

Year	Collected, Validated and Normalized Data			Calculated From Normalized Data		
	Lot Total Cost FY2016\$K	Lot QTY	Low Rate Initial Production	Ave Unit Price FY2016\$K	First Unit	Last Unit
	LotTotCost	Qty	LRIP	AUP	First	Last
2004	18.182	8	1	2.273	1	8
2005	24.975	20	1	1.249	9	28
2006	52.003	35	1	1.486	29	63
2007	37.751	29	1	1.302	64	92
2008	40.240	35	1	1.150	93	127
2009	34.302	35	0	0.980	128	162
2010	27.763	29	0	0.957	163	191
2011	37.289	36	0	1.036	192	227
2012	35.329	38	0	0.930	228	265
2013	36.291	38	0	0.955	266	303
2014	42.899	43	0	0.998	304	346
2015	18.955	18	0	1.053	347	364

F.3 Power Density Example

Observation	Density	Cost (\$K)
1	2.8	11,966.0
2	9.5	76,510.2
3	9.3	75,640.6
4	9.6	72,360.3
5	8.0	42,867.6
6	9.1	76,831.0
7	7.6	38,128.9
8	7.5	36,729.7
9	6.0	24,990.0
10	6.0	30,684.5

F.4 Pseudo-Exact Prior Information Example

Observation	x_1	x_2	x_3	x_4	y
1	98.29	39.64	367.77	4.46	2,483.19
2	110.31	45.78	443.74	5.14	2,764.18
3	98.64	48.34	410.80	3.86	2,554.84
4	98.60	41.15	333.71	4.43	2,477.89
5	93.86	48.94	365.18	5.25	2,410.59
6	81.91	41.48	504.68	5.99	2,915.43

APPENDIX G ACRONYMS

G.1 General

Aper	Aperture
AV	Air Vehicle
BLUE	Best Linear Unbiased Estimators
BLS	Bureau of Labor and Statistics
CEBoK©	Cost Estimating Body of Knowledge
CER	Cost Estimating Relationship
CES	Cost Element Structure
cm	Centimeter
DAU	Defense Acquisition University
DF or df	Degrees of Freedom
DoDCAS	Department of Defense Cost Analysis Symposium
EDF	Empirical Distribution Function
EQQs	Economic Order Quantities
Fig.	Figure
hp	Horsepower
ICEAA	International Cost Estimating and Analysis Association
IMP/IMS	Integrated Master Plan / Schedule
JA CSRUH	Joint Agency Cost Schedule Risk and Uncertainty Handbook
JCAP	Journal of Cost Analysis and Parametrics
kW	Kilowatt
lb	Pounds
LN	Natural Logarithm
Log	Logarithm
PM	Project Management
Pmf	Probability Mass Function
PoP	Period of Performance
SME	Subject Matter Expert
SS	Sum of Squares
WBS	Work Breakdown Structure
ZPB	Zero-Percentage Bias

G.2 Multicollinearity

PPM	Pearson Product-Moment Correlation
VIFs	Variance Inflation Factors (test for multicollinearity)

G.3 Cost Estimating and Regression Methods

AUC	Average Unit Cost
AUPC	Average Unit Procurement Cost
CIC	Cost Improvement Curve
GERM	General Error Regression Methods
GLM	Generalized Linear Model
GLS	Generalized Least Squares
IRLS	Iteratively Reweighted Least Squares
LC	Learning Curve
LOLS	Log Ordinary Least Squares
MPE	Minimum Percentage Error
MRLN	Maximum Likelihood Estimation for Regression of Log Normal error
MUPE	Minimum Unbiased Percentage Error
NLS	Non-linear Least Squares
OLS	Ordinary Least Squares
SLR	Simple Linear Regression
MLR	Multiple Linear Regression
WLS	Weighted Least Squares
ZMPE	Zero Percentage Bias Minimum Percentage Error

G.4 Advanced Regression Methods

GEE	General Estimating Equations
ICLS	Inequality Constrained Least Squares
LASSO	Least Absolute Shrinkage and Selection Optimizer
PCA	Principal Component Analysis

G.5 Influence Points

Cook's D	Cook's Distance
DFFITS	Difference in Fit Statistic
DFBETAS	A variation on DFFITS
HIPS	High Influence Points

G.6 Regression Statistics

%Error	Percentage Error
ANOVA	Analysis of Variance
CDF	Cumulative Density Function

CI	Confidence Interval
MLE	Maximum Likelihood
MOE	Margin of Error
MS	Mean Squares
MSE	Mean Squared Error
PDF	Probability Density Function
PI	Prediction Interval
RMSE	Root Mean Squared Error
RSS	Residual Sum of Squares
SSE	Sum of Squared Errors
SST	total sum of squares

G.7 Assumption Tests

AD	Anderson-Darling
BP	Breusch-Pagan (test for heteroscedasticity of the errors)
Chi-squared	Pearson Chi-squared
DW	Durbin-Watson (test for independence of errors)
KS	Kolmogorov-Smirnov
P-P	Probabilty-Probabilty (plot to check for normality of the errors)
Q-Q	Quantile-Quantile (plot to check for normality of the errors)
SW	Shapiro-Wilk

G.8 Fit/Predictive Statistics

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
Cp	Mallows statistic (estimate of the mean squared prediction error for the OLS model)
CV	Coefficient of Variation (standard deviation/mean)
MAD	Mean Absolute Deviation
PRESS	Predicted Residual Sum of Squares
R_{adj}^2	Coefficient of Determination Adjusted for Degrees of Freedom
SE	Standard Error
SEE	Standard Error of the Estimate
SPE	Standard Percent Error

G.9 DoD Terminology

AS	Acquisition Plan (AP) / Acquisition Strategy
BY	Base Year

CER Development Handbook

CADE	Cost Assessment Data Enterprise
CADRe	Cost Analysis Data Requirement (NASA)
CARD	Cost Analysis Requirements Description
CAPE	Cost Assessment and Program Evaluation
CCDR	Contractor Cost Data Report
CCE	DoD Component Cost Estimate
CCP	DoD Component cost position
CDD	Capability Development Document
CP	Constant Prices
CSDR	Cost and Software Data Reports (CSDR = CCDR + SRDR)
CWIPT	Cost Working Integrated Product Team
CY	Constant Year
DoD	Department of Defense
EVM	Earned Value Management
FFP	Firm Fixed Price
FY	Fiscal Year
GFE	Government Furnished Equipment
IAT&C	Integration Assembly Test and Checkout
ICD	Initial Capabilities Document
ICE	Independent Cost Estimate
ILSP	Integrated Logistics Support Plan
MIL	Military
O&S	Operating and Support
ONCE	One NASA Cost Engineering
POE	Program Office Estimate
SAR	Selected Acquisition Report (total program cost, schedule, and performance)
SEPM	System Engineering and Project Management
SRDR	Software Resource Data Report
SRU	Shop Replaceable Units
STD	Standard
T&M	Time and Materials
T1	Theoretical First Unit Cost
TEMP	Test and Evaluation Management Plan
TRA	Technical Readiness Assessment
TRA	Technical Risk Assessment
TY	Then Year

G.10 Tools

ACEIT	Automated Cost Estimating Integrated Tools
CO\$TAT	Cost Analysis Statistics Package (part of the ACEIT suite of tools)
JMP	Statistics package by the JMP business unit of SAS Institute
MiniTab	Minitab
R	The R Project for Statistical Computing
STATA	Data Analysis and Statistical Software