Appendix C: Cost Estimating Methodologies

The cost estimator must select the most appropriate cost estimating methodology (or combination of methodologies) for the data available to develop a high quality cost estimate. The three basic cost estimating methods that can be used during a NASA project's life cycle are analogy, parametric, and engineering build-up (also called "grassroots") as well as extrapolation from actuals using Earned Value Management (EVM). This appendix provides details on the following three basic cost estimating methods used during a NASA project's life cycle:

For additional information on cost estimating methodologies, refer to the GAO Cost Estimating and Assessment Guide at *http://www.gao.gov/products/GAO-09-3SP*.

Figure C-1 shows the three basic cost estimating methods that can be used during a NASA project's life cycle: analogy, parametric, and engineering build-up (also called "grassroots"), as well as extrapolation from actuals using Earned Value Management (EVM).

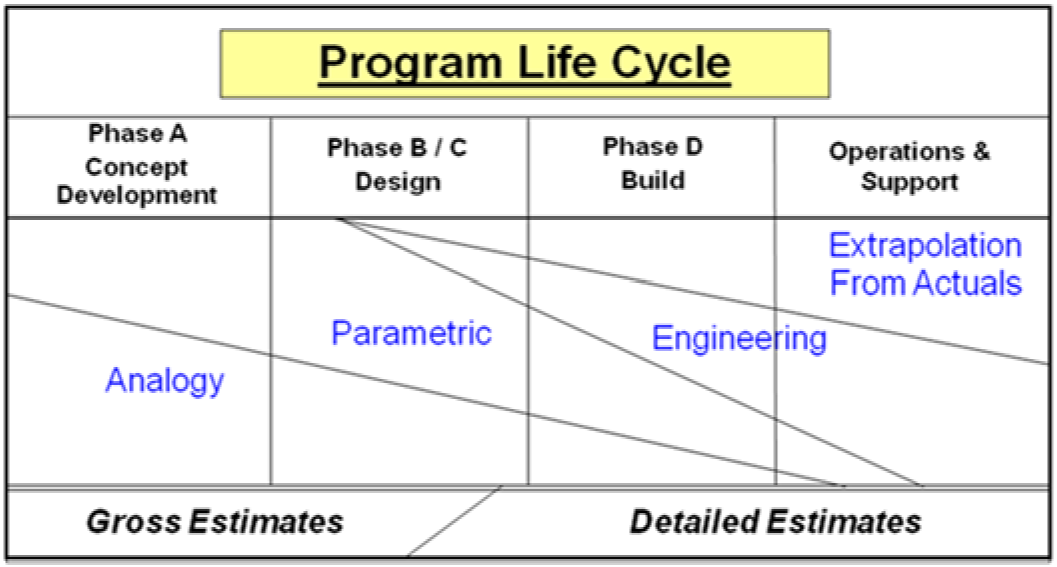

Figure C-1. Use of Cost Estimating Methodologies by Phase[1]

When choosing a methodology, the analyst must remember that cost estimating is a forecast of future costs based on the extrapolation of available historical cost and schedule data. The type of cost estimating method used will depend on the adequacy of Project/Program definition, level of detail required, availability of data, and time constraints. The analogy method finds the cost of a similar space system, adjusts for differences, and estimates the cost of the new space system. The parametric method uses a statistical relationship to relate cost to one or several technical or programmatic attributes (also known as independent variables). The engineering build-up is a detailed cost estimate developed from the bottom up by estimating the cost of every activity in a project's Work Breakdown Structure (WBS).

Table C-1 presents the strengths and weaknesses of each method and identifies some of the associated applications.

[1] Defense Acquisition University, "Integrated Defense Acquisition, Technology, and Logistics Life Cycle Management Framework chart (v5.2)," 2008, as reproduced in the International Cost Estimating and Analysis Association's "Cost Estimating Body of Knowledge Module 2."

Table C-1. Strengths, Weaknesses, and Applications of Estimating Methods

Methodology	Strengths	Weaknesses	Applications
Analogy Cost Estimating	Based on actual historical data	In some cases, relies on single historical data point	• Early in the design process • When less data are available • In rough order-of-magnitude estimate • Cross-checking • Architectural studies • Long-range planning
	Quick	Can be difficult to identify appropriate analog	
	Readily understood	Requires "normalization" to ensure accuracy	
	Accurate for minor deviations from the analog	Relies on extrapolation and/or expert judgment for "adjustment factors"	
Parametric Cost Estimating	Once developed, CERs are an excellent tool to answer many "what if" questions rapidly	Often difficult for others to understand the statistics associated with the CERs	• Design-to-cost trade studies • Cross-checking • Architectural studies • Long-range planning • Sensitivity analysis • Data-driven risk analysis • Software development
	Statistically sound predictors that provide information about the estimator's confidence of their predictive ability	Must fully describe and document the selection of raw data, adjustments to data, development of equations, statistical findings, and conclusions for validation and acceptance	
	Eliminates reliance on opinion through the use of actual observations	Collecting appropriate data and generating statistically correct CERs is typically difficult, time consuming, and expensive	
	Defensibility rests on logical correlation, thorough and disciplined research, defensible data, and scientific method	Loses predictive ability/credibility outside its relevant data range	
Engineering Build-Up	Intuitive	Costly; significant effort (time and money) required to create a build-up estimate; Susceptible to errors of omission/double counting	• Production estimating • Negotiations • Mature projects • Resource allocation
	Defensible	Not readily responsive to "what if" requirements	
	Credibility provided by visibility into the BOE for each cost element	New estimates must be "built up" for each alternative scenario	
	Severable; entire estimate is not compromised by the miscalculation of an individual cost element	Cannot provide "statistical" confidence level	
	Provides excellent insight into major cost contributors (e.g., high-dollar items).	Does not provide good insight into cost drivers (i.e., parameters that, when increased, cause significant increases in cost)	
	Reusable; easily transferable for use and insight into individual project budgets and performer schedules	Relationships/links among cost elements must be "programmed" by the analyst	

C.1. Analogy Cost Estimating

NASA missions are generally unique, but typically few of the systems are completely new systems; they build on the development efforts of their predecessors. The analogy estimating method takes advantage of this synergy by using actual costs from a similar program with adjustments to account for differences between the analogy mission and the new system. Estimators use this method in the early life cycle of a new program or system when technical definition is immature and insufficient cost data are available. Although immature, the technical definition should be established enough to make sufficient adjustments to the analogy cost data.

Cost data from an existing system that is technically representative of the new system to be estimated serve as the Basis of Estimate (BOE). Cost data are then subjectively adjusted upward or downward, depending upon whether the subject system is felt to be more or less complex than the analogous system. Clearly, subjective adjustments that compromise the validity and defensibility of the estimate should be avoided, and the rationale for these adjustments should be adequately documented. Analogy estimating may be performed at any level of the WBS. Linear extrapolations from the analog are acceptable adjustments, assuming a valid linear relationship exists.

Table C-2 shows an example of an analogy:

Table C-2. Predecessor System Versus New System Analogy

	Predecessor System	New System
Solar Array	A	B
Power	2.3 KW	3.4 KW
Solar Array Cost	$10M	?

Assuming a linear relationship between power and cost, and assuming also that power is a cost driver of solar array cost, the single-point analogy calculation can be performed as follows:

Solar Array Cost for System B = 3.4/2.3 * $10M = $14.8M

Complexity or adjustment factors can also be applied to an analogy estimate to make allowances for year of technology, inflation, and technology maturation. These adjustments can be made sequentially or separately. A complexity factor usually is used to modify a cost estimate for technical difficulty (e.g., an adjustment from an air system to a space system). A traditional complexity factor is a linear multiplier that is applied to the subsystem cost produced by a cost model. In its simplest terms, it is a measure of the complexity of the subsystem being priced compared to the single point analog data point being used.

This method relies heavily on expert opinion to scale the existing system data to approximate the new system. Relative to the analog, complexities are frequently assigned to reflect a comparison of factors such as design maturity at the point of selection and engineering or performance parameters like pointing accuracy, data rate and storage, mass, and materials. If there are a number of analogous data points, their relative characteristics may be used to inform the assignment of a complexity factor. It is imperative that the estimator and the subject matter expert (SME) work together to remove as much subjectivity from the process as possible, to document the rationale for adjustments, and to ensure that the estimate is defensible.

Complexity or adjustment factors may be applied to an analogy estimate to make allowances for things such as year of technology, inflation, and technology maturation. A complexity factor is used to modify the cost estimate as an adjustment, for example, from an aerospace flight system to a space flight system due to the known and distinct rigors of testing, materials, performance, and compliance requirements between the two systems. A traditional complexity factor is a linear multiplier that is applied to the subsystem cost produced by a cost model. In its simplest terms, it is a measure of the complexity of the subsystem being estimated compared to the composite of the cost estimating relationship (CER) database being used or compared to the single point analog data point being used.

The following steps would generally be followed to determine the complexity factor. The cost estimator (with the assistance of the design engineer) would:

- Become familiar with the historical data points that are candidates for selection as the costing analog;
- Select that data point that is most analogous to the new subsystem being designed;
- Assess the complexity of the new subsystem compared to that of the selected analog in terms of:
  - Design maturity of the new subsystem compared to the design maturity of the analog when it was developed;
  - Technology readiness of the new design compared to the technology readiness of the analog when it was developed; and
  - Specific design differences that make the new subsystem more or less complex than the analog (examples would be comparisons of pointing accuracy requirements for a guidance system, data rate and storage requirements for a computer, differences in materials for structural items, etc.).
- Make a quantitative judgment for a value of the complexity factor based on the above considerations; and
- Document the rationale for the selection of the complexity factor.

Table C-3 presents the strengths and weaknesses of the Analogy Cost Estimating Methodology and identifies some of the associated applications.

Table C-3. Strengths, Weaknesses, and Applications of Analogy Cost Estimating Methodology

Strengths	Weaknesses	Applications
Based on actual historical data	In some cases, relies on single historical data point	• Early in the design process • When less data are available • In rough order-of-magnitude estimate • Cross-checking • Architectural studies • Long-range planning
Quick	Can be difficult to identify appropriate analog	
Readily understood	Requires "normalization" to ensure accuracy	
Accurate for minor deviations from the analog	Relies on extrapolation and/or expert judgment for "adjustment factors"	

C.2. Parametric Cost Estimating[2]

Parametric cost estimates are a result of a cost estimating methodology using statistical relationships between historical costs and other program variables (e.g. system physical or performance

[2] The information in this section comes from the GAO Cost Estimating and Assessment Guide – Best Practices for Developing and Managing Capital Program Costs, GAO-09-3SP, March 2009.

characteristics, contractor output measures, or personnel loading) to develop one or more cost estimating relationships (CERs). Generally, an estimator selects parametric cost estimating when only a few key pieces of data are known, such as weight and volume. The implicit assumption in parametric cost estimating is that the same forces that affected cost in the past will affect cost in the future. For example, NASA cost estimates are frequently of space systems or software. The data that relate to these estimates are weight characteristics and design complexity, respectively. The major advantage of using a parametric methodology is that the estimate can usually be conducted quickly and be easily replicated. Figure C-2 shows the steps associated with parametric cost estimating.

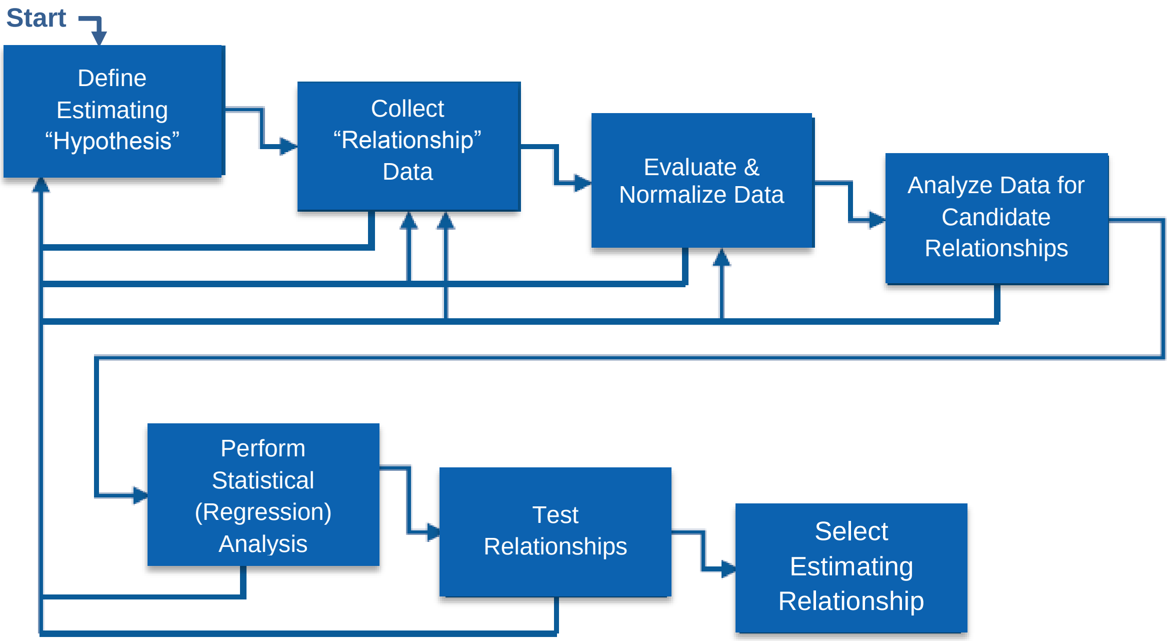
Start
Define Estimating “Hypothesis”
Collect “Relationship” Data
Evaluate & Normalize Data
Analyze Data for Candidate Relationships
Perform Statistical (Regression) Analysis
Test Relationships
Select Estimating Relationship

Figure C-2. Parametric Cost Modeling Process

In parametric estimating, a cost estimator will either use NASA-developed, commercial off-the-shelf (COTS), or generally accepted equations/models or create her own CERs. If the cost estimator chooses to develop her own CERs, there are several techniques to guide the estimator.

To develop a parametric CER, the cost estimator must determine the drivers that most influence cost. After studying the technical baseline and analyzing the data through scatter charts and other methods, the cost estimator should verify the selected cost drivers by discussing them with engineers, scientists, and/or other technical experts. The CER can then be developed with a mathematical expression, which can range from a simple rule of thumb (e.g., dollars per kg) to an equation having several parameters (e.g., cost as a function of kilowatts, source lines-of-code [SLOC], and kilograms) that drive cost.

Estimates created using a parametric approach are based on historical data and mathematical expressions relating cost as the dependent variable to selected, independent, cost-driving variables. Generally, an estimator selects parametric cost estimating when only a few key pieces of data, such as weight and volume, are known. The implicit assumption of parametric cost estimating is that the same forces that affected cost in the past will affect cost in the future. For example, NASA cost estimates are frequently of space systems or software. The data that relates to estimates of these are weight characteristics and design complexity, respectively.

The major advantage of using a parametric methodology is that the estimate can usually be conducted quickly and be easily replicated. Most estimates are developed using a variety of methods where some general principles apply.

Note that there are many cases when CERs can be created without the application of regression analysis. These CERs are typically shown as rates, factors, and ratios. Rates, factors, and ratios are often the result of simple calculations (like averages) and many times do not include statistics.

- A rate uses a parameter to predict cost, using a multiplicative relationship. Since rate is defined to be cost as a function of a parameter, the units for rate are always dollars per something. The rate most commonly used in cost estimating is the labor rate, expressed in dollars per hour. Other commonly used rates are dollars per pound and dollars per gallon.
- A factor uses the cost of another element to estimate a new cost using a multiplier. Since a factor is defined to be cost as a function of another cost, it is often expressed as a percentage. For example, travel costs may be estimated as 5 percent of program management costs.
- A ratio is a function of another parameter and is often used to estimate effort. For example, the cost to build a component could be based on the industry standard of 20 hours per subcomponent.

Parametric estimates established early in the acquisition process must be periodically examined to ensure that they are current throughout the acquisition life cycle and that the input range of data being estimated is applicable to the system. Such output should be shown in detail and well documented. If, for example, a CER is improperly applied, a serious estimating error could result. Microsoft Excel and other commercially available modeling tools are most often used for these calculations. For more information on models and tools, refer to Appendix E.

The remainder of the parametrics section will cover how a cost estimator applies regression analysis to create a CER and uses analysis of variance (ANOVA) to evaluate the quality of the CER.

Regression analysis is the primary method by which parametric cost estimating is enabled. Regression is a branch of applied statistics that attempts to quantify the relationship between variables and then describe the accuracy of that relationship by various indicators. This definition has two parts: (1) quantifying the relationship between the variables involves using a mathematical expression, and (2) describing the accuracy of the relationship requires the computation of various statistics that indicate how well the mathematical expression describes the relationship between the variables. This chapter covers mathematical expressions that describe the relationship between the variables using a linear expression with only two variables. The graphical representation of this expression is a straight line. Regression analysis is the technique applied in the parametric method of cost estimating. Some basic statistics texts also refer to regression analysis as the Least Square Best Fit (LSBF) method, also known as the method of **Ordinary Least Squares (OLS)**.

The main challenge in analyzing bivariate (two variable) and multivariate (three or more variables) data is to discover and measure the association or covariation between the variables—that is, to determine how the variables relate to one another. When the relationship between variables is sharp and precise, ordinary mathematical methods suffice. Algebraic and trigonometric relationships have been studied successfully for centuries. When the relationship is blurred or imprecise, the preference is to use statistical methods. We can measure whether the vagueness is so great that there is no useful relationship at all. If there is only a moderate amount of vagueness, we can calculate what the best prediction would be and also qualify the prediction to take into account the imprecision of the relationship.

There are two related, but distinct, aspects of the study of association between variables. The first, regression analysis, attempts to establish the nature of the relationship between variables—that is, to study the functional relationship between the variables and thereby provide a mechanism for predicting or forecasting. The second, correlation analysis, has the objective of determining the degree of the relationship between variables. In the context of this appendix, we employ regression analysis to develop an equation or CER.

If there is a relationship between any variables, there are four possible reasons.

1. The first reason has the least utility: chance. Everyone is familiar with this type of unexpected and unexplainable event. An example of a chance relationship might be a person totally unfamiliar with the game of football winning a football pool by correctly selecting all the winning teams. This type of relationship between variables is totally useless since it is unquantifiable. There is no way to predict whether or when the person would win again.
2. A second reason for relationships between variables might be a relationship to a third set of circumstances. For instance, while the sun is shining in the United States, it is nighttime in Australia. Neither event caused the other. The relationship between these two events is better explained by relating each event to another variable, the rotation of Earth with respect to the Sun. Although many relationships of this form are quantifiable, we generally desire a more direct relationship.
3. The third reason for correlation is a functional relationship, one which we represent by equations. An example would be the relationship: F = ma, where F = force, m = mass, and a = acceleration due to the force of gravity. This precise relationship seldom exists in cost estimating.
4. The last reason is a **causal** relationship. These relationships are also represented by equations, but in this case a cause-and-effect situation is inferred between the variables. It should be noted that a regression analysis does not prove cause and effect. Instead, a regression analysis presents what the cost estimator believes to be a logical cause-and-effect relationship. It's important to note that each causal relationship enables the analyst to imply that the relationship between variables is consistent. Therefore, two different types of variables will arise.
   a. There will be unknown variables called **dependent** variables designated by the symbol Y.
   b. There will be known variables called **independent** variables designated by the symbol X.
   c. The dependent variable responds to changes in the independent variable.
   d. When working with CERs, the Y variable represents some sort of cost, while the X variables represent various parameters of the system.

As noted above in #4, regression analysis is used not to confirm causality, but rather to **infer causality**. In other words, no matter the statistical significance of a regression result, causality cannot be proven. For example, assume a project designing a NASA launch system wants to know its cost based upon current system requirements. The cost estimator investigates how well these requirements correlate to cost. If certain system requirements (e.g., thrust) indicate a strong correlation to system cost, and these regressions appear logical (i.e., positive correlation), then one can **infer** that these equations have a causal relationship—a subtle yet important distinction from proving cause and effect. Although regression analysis cannot confirm causality, it does explicitly provide a way to (a) measure the strength of quantitative relationships and (b) estimate and test hypotheses regarding a model's parameters.

Prior to performing regression analysis, it is important to examine and normalize the data as follows[3]:

(1) Make inflation adjustments to a common base year.
(2) Make learning curve adjustments to a common specified unit, e.g., Cost of First Unit (CFU).
(3) Check independent variables for extrapolation.
(4) Perform a scatterplot analysis.

[3] For more details on data normalization, refer to Task 7 (Gather and Normalize Data) in section 2.2.4 of the Cost Estimating Handbook.

(5) Check for database homogeneity.
(6) Check for multicollinearity.

The first step of the actual regression analysis is to postulate what independent variable or variables (e.g., a system's weight, X) could have a significant effect on the dependent variable (e.g., a system's cost, Y). This step is commonly performed by creating a scatterplot of the (X, Y) data pairs then "eyeballing" to identify a possible trend. For a CER, the dependent variable will always be **cost** and each independent variable will be a **cost driver**. Each cost driver should be chosen only when there is correlation between it and cost and because there are sound principles for the relationship being investigated. For example, given analysts assume that the complexity (X) of a piece of computer software drives the cost of a software development project (Y), the analysts can investigate their assumption by plotting historical pairs of these dependent and independent variables (Y versus X). Plotting this historical data of cost (Y) versus weight (X) produces a scatterplot as shown in Figure C-3.

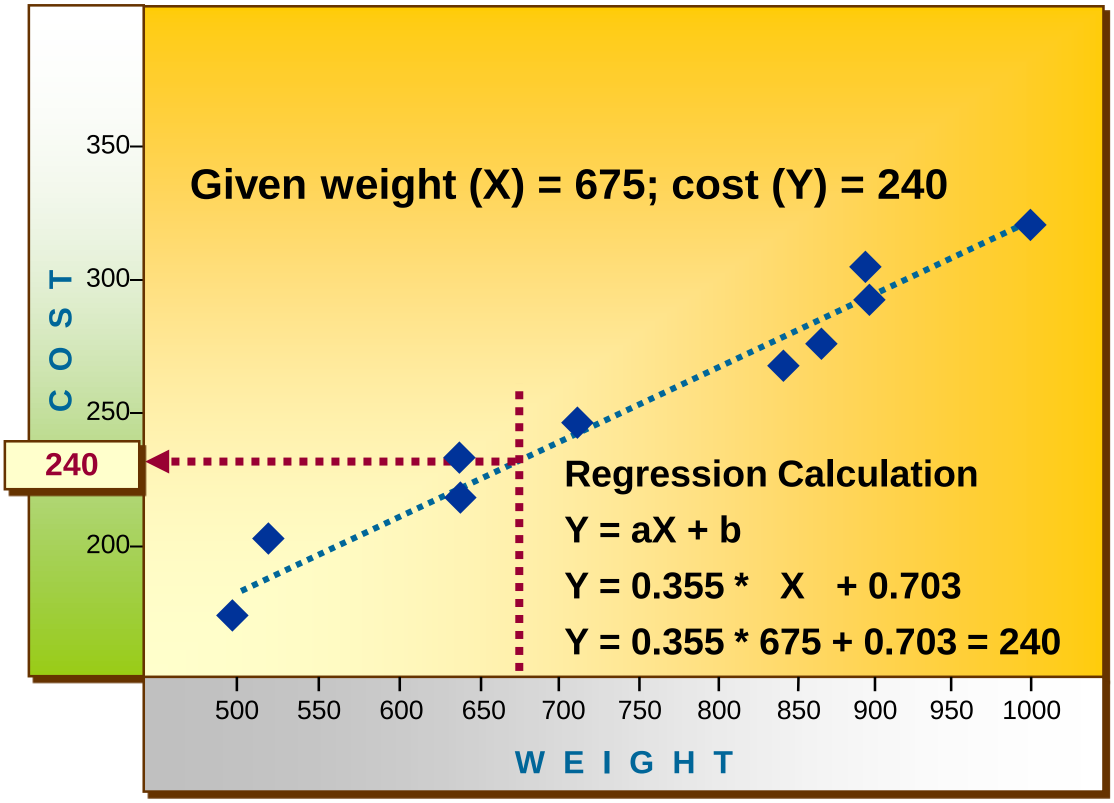

Figure C-3. Scatterplot and Regression Line of Cost (Y) Versus Weight (X)

The point of regression analysis is to "fit" a line to the data that will result in an equation that describes that line, expressed by Y = Y-intercept + (slope) (X). In Figure C-1, we assume a positive correlation, one that indicates that as weight increases, so does the cost associated with the weight. It is much less common that a CER will be developed around a negative correlation, i.e., as the independent variable increases in quantity, cost decreases. Whether the independent variable is complexity or weight or something else, there is typically a positive correlation to cost.

The next step in performing regression analysis to produce a CER is to calculate the relationship between the dependent (Y) and independent (X) variables. In other words, see if the data infer any reasonable degree of cause and effect. The data are, in most cases, assumed to follow either a linear or nonlinear pattern. For the regression line in Figure C-1, the notional CER is depicted as a linear relationship of Cost = A + B * Weight.

As noted in the beginning of this section, the most widely used regression method is the OLS method, which can be applied for modeling both linear and nonlinear CERs (when having one independent variable). Through the application of OLS, section C.2.1. provides details on how to model dependent and independent variables in linear equation form, Y = Y-intercept + (slope) (X). OLS is used again in section C.2.2. to calculate nonlinear models of the form $Y = AX^B$. In order to apply OLS in section C.2.2. (on what is thought to be a nonlinear trend), the nonlinear historical (X, Y) data is transformed using logarithms.

Table C-4 serves as a reference for describing key symbols used in regression analysis. This summary table includes not only symbols that make up a regression model but also important symbols used to assess these models.

There are several other regression methods to produce nonlinear models that bypass the need to transform the historical (X, Y) data. These methods, which were developed to address limitations associated with OLS, include:

- Minimum Unbiased Percentage Error (MUPE) Method
- Zero Percent Bias/Minimum Percent Error (ZPB/MPE) Method (also known as ZMPE Method)
- Iterative Regression Techniques

Such nonlinear regression methods are out of the scope of this handbook and, therefore, will not be covered in Appendix C. For more information on the MUPE Method, ZMPE Method, and Iterative Regression Techniques, refer to the “Regression Methods” section of Appendix A of the 2013 Joint Cost and Schedule Risk and Uncertainty Handbook (CSRUH) at *https://www.ncca.navy.mil/tools/tools.cfm*.

The remainder of Section C.2. covers the following steps in performing regression analysis and selecting the best CER:

(1) Review the literature and scatterplots to postulate cost drivers of the dependent variable.
(2) Select the independent variables(s) for each CER.
(3) Specify each model’s functional form (e.g., linear, nonlinear).
(4) Apply regression methods to produce each CER.
(5) Perform significance tests (i.e., t-test, F-test) and residual analyses.
(6) Test for multicollinearity (if multiple regression).
(7) See if equation causality seems logical (e.g., does the sign of slope coefficient make sense?).
(8) For remaining equations, down-select to the one with highest R^2 and/or lowest SE.
(9) Collect additional data and repeat steps 1–8 (if needed).
(10) Document the results.

These steps begin with how to produce and assess a simple linear regression (SLR) model.

Table C-4. Summary of Key Symbols in Regression Analysis

Symbol	Description	Definition	Evaluation

X, Y	Data Observations	Y= dependent variable
		X= independent variable
			Check and correct any errors, especially outliers in the data. If data quality is poor, it may be necessary to opt for an analysis method other than regression.

$\overline{X}$, $\overline{Y}$	Average or Mean	$\overline{X}$ = mean of actual X_i's $\overline{Y}$ = mean of actual Y_i's.	Helpful to describe the central tendency of the data being evaluated
$\hat{Y}_i$	Calculated Y	$\hat{Y}_i$ = dependent variable	Derived using SLR, $\hat{Y}_i$ is the predicted or “fitted” value associated with X_i, and will differ from Y_i whenever Y_i does not lie on the regression line.
$\hat{b}_i$ or $\hat{B}_i$	Estimated Coefficient of Each Independent Variable (i.e., each estimated regression parameter)	Value of y-intercept, each slope in a linear equation, and/or each exponent in a nonlinear equation	If t-stats are below threshold or values seem illogical, re-specify the model (e.g., with other independent variables and/or another functional form).
e_i	Error or “Residual”	Difference between an actual Y (Yi) and its respective predicted Y (Y)	Check for transcription errors. Take appropriate corrective action.
R^2	Coefficient of Determination	Measures degree of overall fit of the model to the data	The closer R^2 is to 100 percent, the better the fit.
R_a^2	R^2 Adjusted for Degrees of Freedom	R^2 formula is adjusted to account for contribution of one or more additional explanatory variables.	One indication that an explanatory variable is irrelevant is if the value R_a^2 goes down when the explanatory variable is added to the equation.
SST	Sum of Squared Total	$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$	Used to compute R^2 and R_a^2. The higher the SST, the more disperse the actual Y-data is from the mean of Y.
SSR	Sum of Squared Regression (Explained Error)	$SSR = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$	Used to compute R^2 and R_a^2. The higher the SSR, the more “different” the regression line is from the mean of Y.
SSE	Sum of Squared Errors (Unexplained Error)	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	Used to compute R2 and R_a^2. The higher the SSE, the more disperse the predicted Y-data is from actual Y-data.

C.2.1. Simple Linear Regression (SLR) Models

“**Simple**” refers to the fact that only **one independent** variable is required to predict the value of Y. In developing CERs, SLR analysis will be used most of the time. Although this may seem like an oversimplification of the problem, there are several good reasons for taking this approach. Costs should logically, and often do, vary in a linear fashion with most physical and performance characteristics. If not exactly linear, linear approximations are often adequate, especially when considering the accuracy of the data. In addition, many curvilinear and exponential functions can be transformed into a linear form, thereby lending themselves to linear analysis. And finally, our sample size is often so small that we cannot justify using any other form.

In the SLR model, a **dependent**, or explained, variable Y is related to an **independent**, or explanatory, variable X by the following expression:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X} + \mathrm{E}$$

This expression is an SLR model for a population where β_0, the Y - intercept, and β_1, the slope, are the unknown regression parameters called the population regression coefficients and E is the random error, or residual disturbance term. The dependent variable is Y and the independent variable is X.

Designating the variable as dependent or independent refers to the mathematical or functional meaning of dependence; it implies neither statistical dependence nor cause and effect. We mean only that we are regarding Y as a function of X.

It should be noted that the SLR model cited above has two distinct parts: the systematic part, $\beta_0 + \beta_1 \mathrm{X}$, and the stochastic or random part, E. This dissection shows that the model is probabilistic rather than deterministic. The stochastic nature of the regression model implies that the value of Y can never be predicted exactly as in a deterministic case. The uncertainty concerning Y is attributable to the presence of E. Since E is a random variable, it imparts randomness to Y.

In order to ensure the validity of using SLR analysis, five assumptions must be made.

1. The functional form is specified correctly. This means a linear relationship exists between X and Y, and that only one independent variable is sufficient to explain the variation in the dependent variable.
2. The independent variables are assumed to be measured without error. This is, in effect, saying that any deviation will be restricted to the dependent variable.
3. The residuals, E, are assumed to be normally distributed about the regression line. This assumption allows us to use certain statistics and tests of significance to evaluate the validity of the regression line. It is known that the difference between two random variables, regardless of their original distribution, tends to be normally distributed.
4. The residuals, E, are assumed to come from an identically and independently distributed random distribution with mean zero and constant variance. This means that the residuals cannot be predicted from a knowledge of the independent variable, X.
5. The database is homogeneous. This means that the items in the database are of the same category of product.

Using the preceding assumptions, we can derive estimators for the unknown regression parameters, β_0 and β_1, and inferences by using these estimators. It should be stressed that, in practice, one or more of the assumptions is often violated. Frequently, the independent variables are not measured without error, and the sample size is so small that assumptions about normality are invalid. However, an essential part of regression analysis is to validate these assumptions (or see if they are violated). In the section on residual analysis, we discuss some techniques used to check these assumptions.

The population regression model represents the equation of a straight line. Since sample data will virtually always be used, it is rare when all points in the population are known or identified. The model for the sample data is:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X$$

This expression is an SLR model for a sample where the ^ indicates a predicted (i.e., calculated) value of Y and calculated values of the slope $(\hat{b}_1)$ and intercept $(\hat{b}_0)$ of the regression line. Similar to the SLR model for a population, X is the independent variable. All predictors are unbiased estimators of their predicted values, meaning they are as likely to be overestimated as they are to be underestimated.

C.2.1.1. Calculating Coefficients for a Simple Linear Regression (SLR) Model

As noted in the beginning of Section C.2, basic statistics texts refer to regression analysis as the Least Square Best Fit (LSBF) method, also known as the method of **Ordinary Least Squares (OLS)**.The following two equations allow us to solve for the slope $(\hat{b}_1)$ and intercept $(\hat{b}_0)$ of the regression line.

$$\hat{b}_1 = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} \qquad \hat{b}_0 = \frac{\sum Y - \hat{b}_1 \sum X}{n}$$

Upon solving the above equations, we fully define the regression line. Therefore, for any value of the independent variable (a value of some parameter of a given system) within the range of the given independent variable, we can determine a predicted "average" value of the dependent variable (a cost). That is,

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X, \text{ where the } \wedge \text{ indicates a predicted (i.e., calculated) value.}$$

Refer to the box on the following page for an example of how to calculate coefficients for a simple linear regression model.

Regression Analysis Example

At this point, we present an example to use throughout this discussion. NASA is considering buying a newly developed storage tank that weighs approximately 2.2 tons. Data are available on 10 systems that have been procured in the past, giving the weight and first unit cost of each of these tanks. Therefore, a regression line can be computed, and an estimate of the first unit cost of the new tank can be made based on its weight. In this case, the independent variable is weight (X) in tons, and the dependent variable is first unit cost (Y) in thousands ($K). Table C-5 presents the data for the 10 systems. Before continuing, we must ensure that data normalization requirements (described in Section 2.3.2 of the main body of this handbook) have already been accomplished. Also included are calculations required to create the inputs required to compute, the regression coefficients, $\hat{b}_1$ and $\hat{b}_0$.

Table C-5. Data Table for Equation Inputs

X (tons)	Y ($K)	XY	X^2	Y^2
1.0	6	6.0	1.00	36
1.5	11	16.5	2.25	121
2.0	14	28.0	4.00	196
1.8	16	28.8	3.24	256
1.9	20	38.0	3.61	400
2.5	18	45.0	6.25	324
3.0	22	66.0	9.00	484
3.4	30	102.0	11.56	900
3.7	26	96.2	13.69	676
3.9	31	120.9	15.21	961
∑X = 24.7	**∑Y = 194**	**∑XY = 547.4**	**$\sum X^2$ = 69.81**	**$\sum Y^2$ = 4354**

We will use these formulas to determine the regression equation. We must first calculate the slope because it is required to determine the intercept.

$$\hat{b}_1 = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2} \qquad \hat{b}_0 = \frac{\sum Y - \hat{b}_1 \sum X}{n}$$

Upon substitution we have:

$$\hat{b}_1 = \frac{[\ (10)\ (547.4)] - [(24.7)(194)]}{[(10)(69.81)] - (24.7)^2} = \frac{682.2}{88.01} = 7.75 \qquad \hat{b}_0 = \frac{[194 - (7.75)(24.7)]}{10} = \frac{2.58}{10} = 0.26$$

and the regression equation is:

$$\hat{Y} \quad 7.751392 \quad 7.75\ X \qquad 0.2654062$$

The units related to the independent variable must be the same as those attached to the X values in the dataset that created the equation. If the weight for the new storage tank is provided in pounds or kilograms, we must first ensure our input is in tons before plugging it into the equation. The units related to the dependent variable are the same as those used for the Y values in the dataset. The predicted average first unit cost for the new system is:

$$\hat{Y} = 0.26 + 7.75(2.2) = \$17.310\text{K or } \$17{,}310.$$

Regression analysis coefficients are sensitive to rounding error. The analyst cannot expect these errors to somehow cancel each other. As a general rule, always carry intermediate calculations at least two decimal places further than the number of places desired in the final answer. The reason that regression analysis is especially subject to rounding error is that often it is necessary to calculate two large numbers and the difference between them. If the difference is small, then it may disappear after rounding.

The regression coefficients may or may not have a practical meaning, depending on the problem. In the previous problem, weight was used to predict storage tank cost. The value of the y-intercept,implies that a storage tank with a weight of zero tons would cost $260. This is illogical. The y-intercept is not a fixed component of cost. The intercept is a reflection of cost variation that is not captured by weight.

C.2.1.2. Extrapolation Out of the Relevant Range of the CER

Another reason for the inconsistency is that it requires us to predict a cost that is outside the relevant range. Predictions using a CER are only valid when applied within the relevant range of the equation. Relevant range refers to the range of the values of the independent variable contained in the dataset. Based on the above dataset, we can use the CER to predict first unit costs for vehicles weighing between 1.0 and 3.9 tons. A weight of zero tons is outside the range of the data. Extrapolating outside the range of the data is dangerous because we can make no inference as to the behavior of the data beyond the sample range. In this case we may say that as vehicle weight is reduced beyond a certain point, cost may increase.

Extrapolation is used when regressing time-series data because it is assumed that the relationship continues through time.

The regression slope, $\hat{b}_1$, is of both theoretical and practical importance. Theoretically, together with $\hat{b}_0$, we can determine the position of the regression line by the slope. Also, we use the value to test the significance of the total regression. The slope measures the amount of change (increase or decrease) in the mean value of Y for a one-unit increase in X. Each time the vehicle's weight increases by 1 ton, cost increases by $7.75K.

C.2.1.3. Regression Statistics of Simple Linear Models

The discussion thus far has focused on quantifying the relationship between independent and dependent variables. This constitutes only one half of the regression analysis problem. We must now compute the statistics and tests of significance, which measure the validity of the regression line.

Just because we can create a regression equation does not mean that we should use it. Regression statistics help us determine how well the CER predicts the dependent variable. They can also tell us whether a trend exists. To help determine whether or not to use a regression equation, the estimator should at least calculate then assess the regression's coefficient of determination (R2), standard error of the estimate (SE), coefficient of variation (CV), and the "strength" of regression coefficients (hypothesis tests or "t-tests"). These four areas of analyzing the regression output are covered in the next several pages.

<u>Calculating the Coefficient of Determination (R^2) for Simple Linear Regression (SLR) Equation</u>

The coefficient of determination, R^2, is the statistic that will be used more than any other as a measure of the accuracy of the regression line fit to the sample data points. R^2 can have a value between 0 and 1 (or 0 percent and 100 percent), where the higher the number, the better the "fit" of the regression line to the actual data. For example, if SLR produces a regression, Cost = 0.34 + 2.7 Mass, with an R^2 of 0.93; this implies that 93 percent of the variation in Cost can be explained by the variation in Mass—an example where Mass is a very good predictor of Cost.

In order to calculate an R^2, one first needs to understand these three terms: explained deviation, unexplained deviation, and total deviation. Refer to the regression line, $\hat{Y} = \hat{b}_0 + \hat{b}_1 X,$ in Figure C-4. As this graphic illustrates, a single deviation of an actual Y-value from the mean value of Y can be split into the two parts: unexplained and explained deviation.

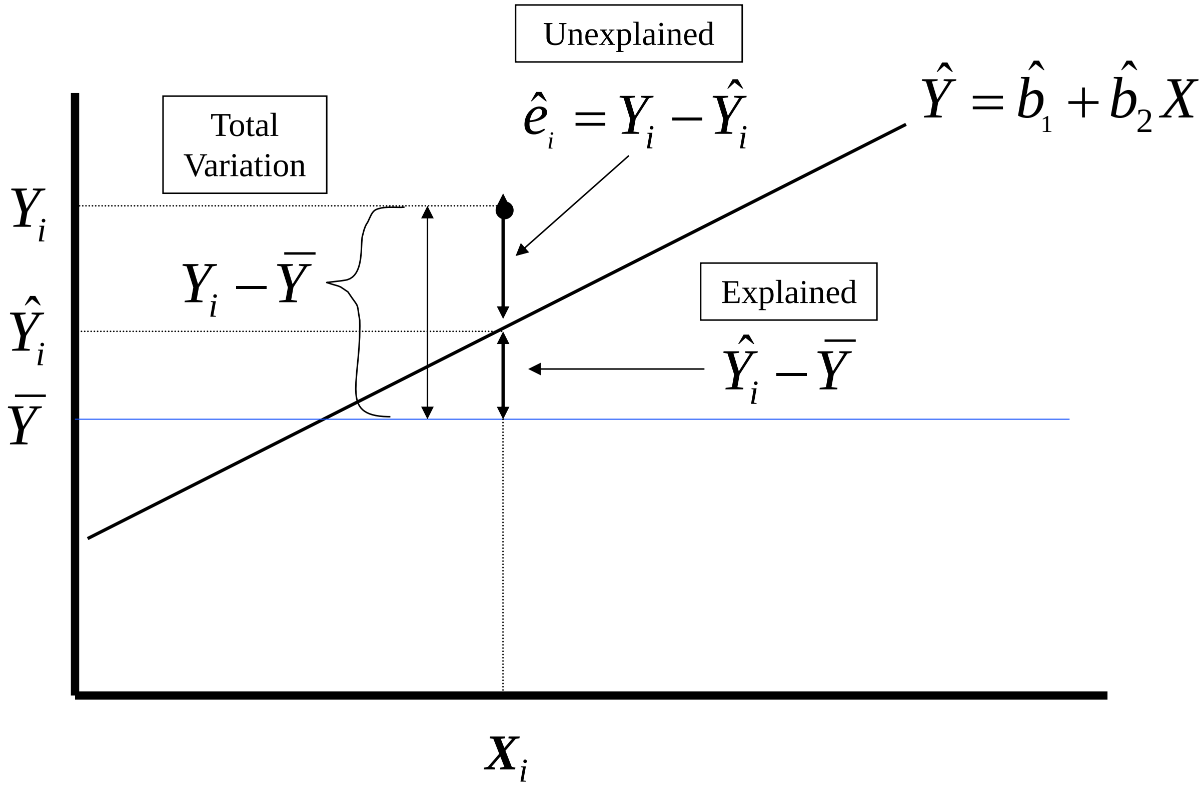
Unexplained
$\hat{e}_i = Y_i - \hat{Y}_i$
$\hat{Y} = \hat{b}_1 + \hat{b}_2 X$
Total Variation
Y_i
$Y_i - \overline{Y}$
$\hat{Y}_i$
$\overline{Y}$
Explained
$\hat{Y}_i - \overline{Y}$
X_i

Figure C-4. Illustration of Deviations About the Mean $\overline{Y}$ and Regression Line $\hat{Y}_i$.

Here are a few useful observations that are depicted in Figure C-1:

1. X_i represents the *i*th value of the independent variable.
2. Y_i is the actual observed value associated with X_i.
3. $\hat{Y}_i$ is the dependent variable. Derived using SLR, $\hat{Y}_i$ is the predicted or "fitted" value associated with X_i and will differ from Y_i whenever Y_i does not lie on the regression line.
4. $\overline{X}$ and $\overline{Y}$ are the mean values of the actual X_i's and Y_i's.
5. $\overline{Y}$ always crosses somewhere on the regression line.
6. $\overline{Y}$ serves as a reference point from which to measure each “explained deviation” and “total deviation” as a basis for the coefficient of determination, R^2.

The SLR output in Figure C-1 reveals three basic types of deviations:

1. The first is the explained deviation and is defined as the difference between the predicted value of Y and the mean value of Y. It’s a measure of how much each Y value on the **regression line** differs from the mean value of Y.

$$\text{Explained Deviation} = \left(\hat{Y}_i - \overline{Y}\right)$$

2. The second type of deviation is the unexplained deviation, which is defined as the difference between the actual value of Y and the predicted value of Y. It's a measure of how much each actual Y value differs from its respective predicted Y value on the regression line. This unexplained deviation is often referred to as an "**error**" or "residual" of a predicted value of Y.

   Unexplained Deviation = $\left(Y_i - \hat{Y}_i\right)$

3. The third type of deviation is total deviation, which is defined in one of two ways: (a) as the difference between the actual value of Y and the mean value of Y, or (b) as the total of explained deviation and unexplained deviation.

   Total Deviation = $\left(Y_i - \overline{Y}\right)$

   Total Deviation = Explained Deviation + Unexplained Deviation

   $$\left(Y_i - \overline{Y}\right) = \left(\hat{Y}_i - \overline{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

In order to calculate the coefficient of determination, R^2, each of these three types of deviation must be squared, then summed. As a result, the squaring-then-summing of each:

1. explained deviation produces the Sum of Squared Regression (SSR) where

   SSR = Sum of (Each Explained Deviation) ^{2} = $\Sigma\left(\hat{Y}_i - \overline{Y}\right)^2$

2. unexplained deviation produces the Sum of Squared Error (SSE) where

   SSE = Sum of (Each Unexplained Deviation) ^{2} = $\Sigma\left(Y_i - \hat{Y}_i\right)^2$

3. total deviation produces the Sum of Squared Total (SST) where

   SST = Sum of (Each Total Deviation) ^{2} = $\Sigma\left(Y_i - \overline{Y}\right)^2$

   or

   SST = SSR + SSE.

   $$\sum\left(Y - \overline{Y}\right)^2 = \sum\left(Y - \hat{Y}\right)^2 + \sum\left(\hat{Y} - \overline{Y}\right)^2$$

The SST formula, **SST = SSR + SSE**, serves as a good starting point to derive the R^2 formula.

Division by the term on the left yields:

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

1 = *Percent of Explained Deviation* + *Percent of Unexplained Deviation*

Notice that *the Percent of Explained Deviation* is the definition of R^2. Therefore, we can rearrange the formula to solve for *Percent of Explained Deviation,* aka R^2.

Percent of Explained Deviation = 1 – *Percent of Unexplained Deviation*

$$\frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Since SSR or R^2 is a ratio of a part to the whole, it will range in value from 0 to 1. For use in conjunction with CERs, an R^2 greater than or equal to 90 percent is desirable. The coefficient of determination can be used to compare various regression lines to determine the best fitting line (the regression line with the highest R^2).

R^2 = *Percent of Explained Deviation*

$$R^2 = \frac{SSR}{SST} = \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum(Y - \bar{Y})^2}$$

This statistic essentially measures the ratio of the explained deviation to the total deviation. Since the explained deviation represents the regression line's ability to predict, it should be as big a proportion of the total deviation as possible. In other words, given that SST is constant, the higher the SSR, the higher the value of R^2.

R^2 = 1 – *Percent of Unexplained Deviation*

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum(Y - \bar{Y})^2}$$

Taking a closer look at the previous formula, the regression line that best fits the sample data points is defined when the sum of the squares of the unexplained deviations is a minimum. Each unexplained deviation is the vertical distance from the line to the actual data point. This criterion makes the regression line as close to all the points as possible. In other words, given that SST is constant, the smaller the SSE, the higher the value of R^2.

An alternative way to calculate R^2 is shown below:

$$R^2 = \frac{\left[\sum XY - n\left(\overline{X}\right)\left(\overline{Y}\right)\right]^2}{\left[\sum X^2 - n\left(\overline{X}\right)^2\right]\left[\sum Y^2 - n\left(\overline{Y}\right)^2\right]}$$

This method bypasses the need to calculate SSR, SSE, and SST. Note, however, that SSR, SSE, and SST values are necessary for other regression statistics covered later in this section.

Referring to the earlier example in Section C.2.1. and using the same data, the coefficient of determination is calculated to be:

$$\mathrm{R}^2 = \frac{\left[547.4 - 10(2.47)(19.4)\right]^2}{\left[69.81 - 10(2.47)^2\right]\left[4354 - 10(19.4)^2\right]} = \frac{4653.97}{5196.11} = 0.895664$$

This would imply that the variations in weight explain about 90 percent of the total variation in storage tank cost. The regression line explains 90 percent of the deviation, leaving only 10 percent to chance.

Calculating the Coefficient of Correlation (R)

The coefficient of correlation is a statistic that is essentially the same measure as the coefficient of determination. It is symbolized by R and is, in fact, plus or minus the square root of the coefficient of determination. Using the coefficient of correlation will yield one piece of information not given by the coefficient of determination. For a regression line with a negative slope, R is negative. For a regression line with a positive slope, R is positive. Obviously, the range of R differs from that of R^2 since R can vary from –1 to +1. The computational formula for R is as follows:

$$R = \sqrt{R^2} or \frac{\sum XY - n\overline{X}\overline{Y}}{\sqrt{\sum X^2 - n\overline{X}^2}\sqrt{\sum Y^2 - n\overline{Y}^2}}$$

For the example problem, since the slope is positive, we know that R is positive. The coefficient of correlation is:

$$\sqrt{\mathrm{R}^2} = \sqrt{0.90} = 0.948 \cong 0.95$$

Calculating the Standard Error of the Estimate (SE)

The standard error of the estimate is analogous to the sample standard deviation. It is a measure of the deviation of the sample points from the regression line or the disturbances from the regression line. Although both R and R^2 also measure the goodness of fit of the regression line to the data points, they are only relative measures and are affected by the slope of the regression line. The standard error of the estimate is an absolute measure of the deviation and its sign is unaffected by the slope of the regression line. The standard error of the estimate is the square root of the sum of the squares of the unexplained deviations divided by the degrees of freedom and is symbolized by SE.

$$\mathrm{SE} = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{\mathrm{n - k - 1}}} = \sqrt{\frac{\mathrm{SSE}}{\mathrm{n - k - 1}}} = \sqrt{\mathrm{MSE}}$$

where:

n is the sample size,

k is the number of independent variables, and

MSE stands for “Mean Squared Error”

Since $\overset{\wedge}{\mathrm{b}}_0$ and $\overset{\wedge}{\mathrm{b}}_1$ are both determined from the same data points from which SE is calculated, there are two restrictions on the data and, thus, n - 2 degrees of freedom when calculating the standard error of the estimate. The term k refers to the number of independent variables contained in the CER. For simple linear equations, there is only one independent variable that must be used to predict Y and therefore only one slope value that must be calculated. The numerator of the SE equation $\sum\left(Y-\hat{Y}\right)^2$ is called the sum of the squared (unexplained) errors, or simply SSE. SSE divided by the degrees of freedom is called the mean squared error (MSE), which is the variance of the error term.

As previously mentioned, SE is an absolute measure of the deviation. It will, therefore, have the units that are associated with the dependent variable. The only magnitude restriction on SE is that it cannot be negative. Because of these facts, the standard error of the estimate cannot be used by itself to evaluate a regression line. It can only be used to compare different regression lines. The standard error of the estimate is also used to calculate other statistics and tests of significance. Since SE is a measure of unexplained error, it should be as small as possible. The computational formula for SE is as follows:

$$\text{SE} = \sqrt{\frac{\sum Y^2 - \hat{b}_0 \sum Y - \hat{b}_1 \sum XY}{n - 2}}$$

Considering the example problem and using the computational formula for SE, the standard error of the estimate is calculated as follows:

$$\text{SE} = \sqrt{\frac{4354\ -\ 0.26(194)\ -\ 7.75(547.4)}{10-2}}\ =\ \$2.77\text{K}$$

Calculation of the Coefficient of Variation (CV)

The coefficient of variation (CV) is a statistic that allows us to use the standard error of the estimate to evaluate a regression line. It is actually a relative standard error of the estimate since it becomes a dimensionless quantity. The formula for CV is

$$\text{CV} = \frac{\text{SE}}{\overline{\text{Y}}}$$

When developing a CER by regression analysis, the CV should be less than or equal to 20 percent, and preferably less than 10 percent. The coefficient of variation for the example problem is

$$\text{CV} = \frac{2.77}{19.4} = 0.14$$

The CV is particularly appropriate for deciding among competing CERs that seem otherwise appropriate. The one with the lowest CV is preferred, as long as both CERs are logical and have sufficient coefficients of determination and significant slopes.

Inferences about Population Regression Coefficients for SLR Models (Hypothesis Tests)

Having obtained the sample regression equation and having concluded that the regression equation may be a useful one on the basis of the standard error of the estimate and the coefficient of determination, the analyst might assume the equation to be a valid predictive device. However, the equation may still contain some error. Predictions may not be precise due to sampling error.

The significance of the sampling error associated with the regression coefficients $\hat{b}_0$ and $\hat{b}_1$ may be evaluated by testing the sample distributions of both the intercept and the slope, assuming they are approximately normal. This assumption of normality allows the analyst to perform a hypothesis test for the significance of the intercept or the slope.

From a practical standpoint, the significance of the intercept is of little importance. As previously stated, it is usually outside the range of the data and represents extrapolation. On the other hand, the significance of the value of the slope should be tested. The hypothesis test might be used to determine if the slope is significantly different from any value. In practice, a hypothesis test is performed to determine if the coefficient of the independent variable (the slope) is significantly different from zero. A slope of zero would imply that the relationship is purely chance. A slope that was not significantly different from zero would mean that knowledge of the value of X would be of no use in predicting the value of Y.

In order to be able to perform this test, another statistic, the standard error of the slope, must be computed. It is symbolized by S_1 and is computed as follows. Most statistical packages give S_1 as an output value.

$$S_1 = \frac{SE}{\sqrt{\sum X^2 - n\overline{X}^2}}$$

Steps in performing a hypothesis test on the slope coefficient are much simpler than those described in testing the mean of a population.

(1) Establish the null and alternative hypotheses. When testing the slope this step is always set up as follows:

(a) The null hypothesis is $H_0\colon b_1 = 0$.

(b) The alternative hypothesis is $H_1\colon b_1 \neq 0$.

(2) Determine the level of significance (α) desired for the test.

(3) Find $t_p = t_{(1-\alpha/2,\ n-2)}$ from the t - table. Because the alternative hypothesis is always a “not equal to,” we always have a two-tailed test.

(4) Calculate $t_c = \frac{\hat{b}_1 - b_1}{S_1} = \frac{\hat{b}_1 - 0}{S_1} = \frac{\hat{b}_1}{S_1}$.

(5) Make a decision based on the following decision rules.

(a) If $|t_c| \rangle |t_p|$, reject the null hypothesis.

Conclude that the slope coefficient **is** statistically significant.

(b) If $|t_c| \leq |t_p|$, do not reject the null hypothesis.

Conclude that the slope coefficient **is not** statistically significant.

Following the steps above, the example problem will yield the following results.

(1) $H_0 : b_1 = 0$

$H_1 : b_1 \neq 0$

(2) $\alpha = 0.05$

(3) $t_p = t_{(1-0.05/2,n-k-1)} = t_{(0.975,8)} = 2.306$

(4) To calculate t_c, we must determine $S_1 = \dfrac{2.77}{\sqrt{69.81 - 10(2.47)^2}} = 0.935362$

$$t_c = \frac{7.75}{0.93} = 8.287054$$

(5) Since $|t_c| > |t_p|$, we reject the null hypothesis.

Conclude that the slope coefficient is statistically significant.

In general, retain the sample regression equation as a predictive model if $\hat{b}_1$ is found to be significantly different from zero; otherwise the sample regression equation should be discarded.

C.2.1.4. Residual Analysis of Simple Linear Regression (SLR) Models

Once we develop a fitted regression model, we need to examine the model's appropriateness for its application. Recall the assumptions of SLR discussed earlier in this chapter (the residuals should be independent, normally distributed, random variables with mean of zero and constant variance). Graphical residual analysis allows us to see whether the errors satisfy these assumptions. The analysis is made easy by use of automated statistics packages. The plot in Figure C-5 of the residuals against the independent variable helps us to evaluate the model.

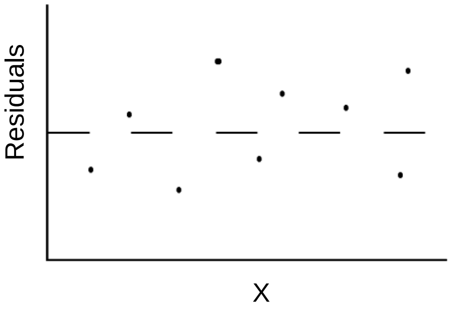

Figure C-5

The plot, a random scatter about the zero line, indicates that the errors are independent and random. Also, the points fall within an equal band above and below the zero line, indicating a mean of zero and constant error variance. This indicates that the regression function used is an appropriate model for the data.

The normal probability plot in Figure C-6 helps in studying the normality of the error terms. The plot shows that the points form a reasonably linear pattern and are close to the line of identity. The assumption of normally distributed error terms in the regression model seems reasonable.

Plotting residuals against the fitted Y values can help identify whether the error variance is constant. Also, plots of residuals against new variables not already in the model can help reveal omissions of important variables.

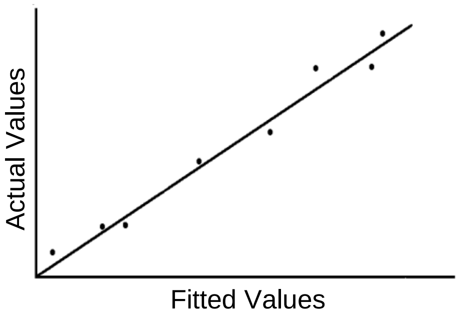

Figure C-6

C.2.1.5. Confidence and Prediction Intervals of Simple Linear Regression (SLR) Models

The confidence interval and prediction interval are two different intervals on the regression line that give different information. The confidence interval gives a region into which the true (population) regression line can be expected to fall with a probability of 1 - α. This interval actually gives boundary limits for the slope of the population line. It says nothing about the range into which a future prediction might fall. The prediction interval describes an interval into which a given percentage of the population is expected to fall. Because of this feature, prediction intervals can be used to predict the values of future dependent variables within a given range and a given confidence level.

The prediction interval exists due to the variance between the predicted value of the dependent variable, $\hat{Y}$, and the actual value, Y. The standard deviation of this distribution is

$$\text{SE}\sqrt{\frac{n+1}{n} + \frac{(X_0 - \overline{X})^2}{\sum X^2 - n\overline{X}^2}}$$

where X_o is the value of X for which we are computing the prediction interval.

With this information, the procedure for constructing a prediction interval about a regression line is outlined as follows.

(1) Determine the significance level (α).

(2) Find $t_p = t_{(1-\alpha/2, n-2)}$ from the t table.

(3) Determine the standard error of the estimate, SE.

(4) Select at least three values of X (called X_0) to evaluate. Include the mean and one value on either side of the mean.

(5) Construct the prediction interval for each value of X_0.

$$\text{PI} = \hat{Y}_0 \pm t_p \ \text{SE}\sqrt{\frac{n+1}{n} + \frac{(X_0 - \overline{X})^2}{\sum X^2 - n\overline{X}^2}}$$

For the example problem, a prediction interval will now be constructed according to the outlined procedure.

(1) Let the significance level be 0.10.

(2) Find $t_p = t_{(1-\alpha/2,\ n-2)}$ from the t table. $t_{(0.95,\ 8)} = 1.86$.

(3) Find SE = 2.77 (from previous example).

(4) For this example, the three values selected were

$$X_0 = X_{\min} = 1.00$$

$$X_0 = \overline{X} = 2.47$$

$$X_0 = X_{max} = 3.90$$

(5) Calculate the PI.

For $X_0 = X_{min} = 1.00$, $\hat{Y}_0 = 8.01$.

$$PI = 8.01 \pm (1.86)(2.77)\sqrt{\frac{11}{10} + \frac{(1.00 - 2.47)^2}{69.81 - 61.01}} = 8.01 \pm 5.98$$

$$PI = 2.03 \le Y \le 13.99$$

For $X_0 = \overline{X} = 2.47$, $\hat{Y}_0 = 19.40$.

$$PI = 19.40 \pm (1.86)(2.77)\sqrt{\frac{11}{10} + \frac{(2.47 - 2.47)^2}{69.81 - 61.01}} = 19.40 \pm 5.41$$

$$PI = 13.99 \le Y \le 24.81$$

For $X_0 = X_{max} = 3.90$, $\hat{Y}_0 = 30.49$.

$$PI = 30.49 \pm (1.86)(2.77)\sqrt{\frac{11}{10} + \frac{(3.90 - 2.47)^2}{69.81 - 61.01}} = 30.49 \pm 5.93$$

$$PI = 24.56 \le Y \le 36.42$$

Figure C-7 shows the regression line and the prediction interval for the sample problem.

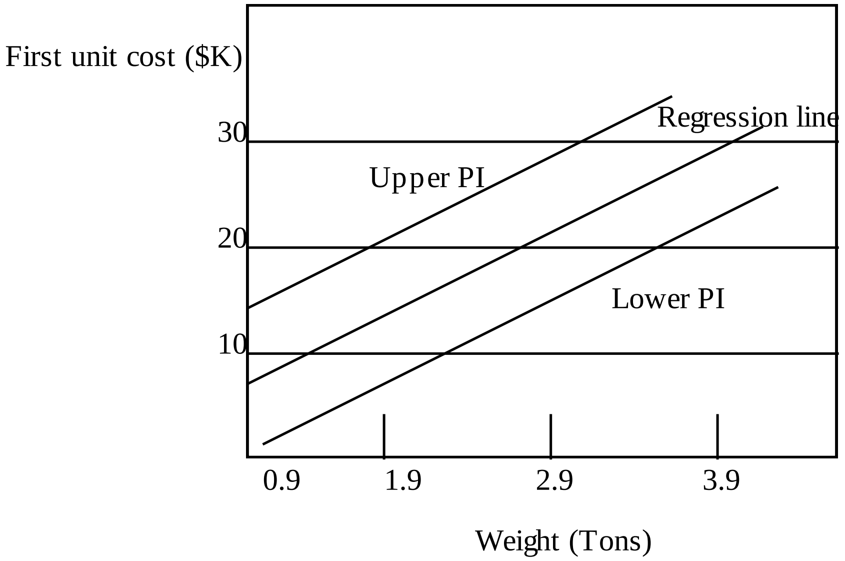

Figure C-7

The prediction interval is very similar to a confidence interval about the regression line. However, the prediction interval will always be slightly wider than the confidence interval. Prediction intervals are usually derived about a specific prediction point with respect to the Y axis, rather than for the regression line as a whole.

C.2.1.6. Summary: Simple Linear Regression (SLR) Models

Regression analysis, while a very useful device, often opens up many opportunities for misinterpretation. The most unwarranted interpretation stems from the confusion between association and causation. Regression analysis only shows the degree of association. Many people take the independent variable to be the cause and the dependent variable to be the effect. This is not a necessary condition.

The link between association and causation can be stated as follows: The presence of association does not necessarily imply causation, but true causation always implies association (i.e., correlation). Statistical evidence can only show the degree of association between variables. Whether causation exists or not depends purely on reasoning (logic). For example, there is reason to believe that an increase in the weight of a modern tank “causes” an increase in cost (a direct or positive association). There is also reason to support the premise that a decrease in the weight of an aircraft (range, velocity, and payload held constant) “causes” an increase in cost (an inverse or negative association). Showing that an increase in the inches of padding in the commander’s seat of a tactical vehicle is associated with high cost does not show cause and effect. There is probably a better, more logical predictor of total vehicle cost.

Why is it that association does not show cause and effect? One reason is that the association between variables may be pure chance, such as soil erosion in Alaska and the amount of alcohol consumed in South America. Another reason is that association between two variables may be due to influence of a third common variable. Since 1945, for example, there has been a close relationship between teachers’ salaries and liquor consumption. A plausible explanation is that both variables have been influenced by the continuous increase in national income during the same period. Another possibility is that in the real relationship we may not be able to determine which variable is the cause and which is the effect. It may be that spending more money on education increases the likelihood of a better educated populace, which gets better jobs and thus has a higher per capita income. Which one is the cause and which is the effect?

Regression analysis may lend itself to misinterpretations and spurious results. One reason for spurious results is a flaw in logic, as discussed previously. Spurious conclusions may also result from extrapolations beyond the range of the data. Regression analysis is an attempt to establish a relationship within a range. The relationship may, in fact, change radically over a different range.

Since in regression analysis we are minimizing the deviations about the mean, regression analysis is sensitive to changes in the mean. The mean may be distorted by errors in the sample data or by the presence of outlier observations. Hence, spurious results may occur.

C.2.2. Simple Nonlinear Regression Models

A linear function relating the dependent and independent variables frequently does not adequately fit the data. In this event, simple nonlinear regression models should be considered. Often it will be apparent from a scatterplot of the data that a linear model is inappropriate. The analyst should then consider fitting one or more nonlinear models to the data. Fortunately, many nonlinear models are intrinsically linear. That is, by a suitable transformation of the dependent and/or independent variables, a linear relationship between the variables might be uncovered. The SLR can then be used to fit a straight line to the transformed data. It is important to note, however, that with the modern-day computer, analysts can easily perform nonlinear regression without needing to transform the data (i.e., the computer solves for the lowest sum-squared error using an iterative method).

Logarithms. Logarithmic (log for short) transformations can often be used to “straighten out” data that do not plot linearly. The logarithm to the base b of a number X, log_bX, is the power to which b must be raised to yield X. For example, the logarithm in base 10 of 100 is 2 (i.e., log_{10}100 = 2) because $10^2 = 100$. Logarithms to the base *e* (*e* = Euler’s constant = 2.718282 . . .) are called natural logarithms and are abbreviated *ln*. The natural log of 100 is 4.6 (i.e., *ln* 100 = 4.6) because $e^{4.6} = 2.718282^{4.6} = 100$. Natural log transformations are used in this unit of instruction.

Power (Log-Log) Model. If a scatterplot of the data on ordinary graph paper does not appear linear, the analyst should plot the data on log-log paper. This task is made considerably easier by using the built-in charting capability found with most spreadsheets. If the data appear to “straighten out” on the log-log paper, then a power function in the following form may be appropriate:

$$Y = AX^B$$

Plotting X versus Y on log-log paper is equivalent to plotting *ln* X versus *ln* Y on ordinary graph paper. Consider the data set in Table C-5:

Table C-5. Power Model Data Set

X	Y	*ln* X	*ln* Y
1	100	0.0000	4.6052
2	80	0.6931	4.3820
4	64	1.3863	4.1589
8	51	2.0794	3.9318
16	41	2.7726	3.7136

As shown in Figure C-8, a scatterplot of the original data, X versus Y, shows a nonlinear relationship.

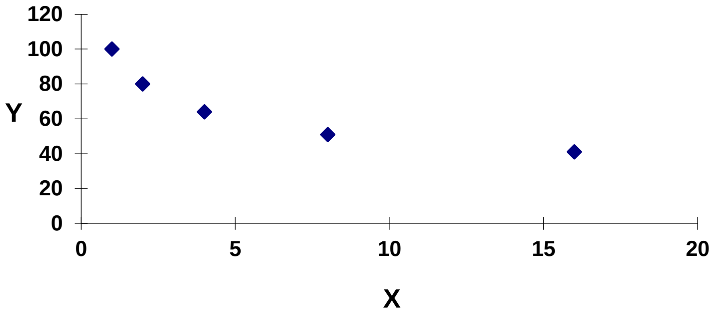

Figure C-8

As shown in Figure C-9, a scatterplot of the transformed data, *ln* X versus *ln* Y, shows an intrinsically linear relationship.

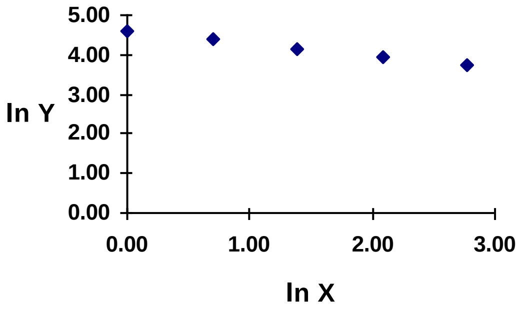

Figure C-9

Assuming that the most accurate model for such data is a power function, taking the natural log of both sides of the power function yields:

$$\ln Y = \ln A + B \ln X$$

This is the equation for a straight line in **fit space**, where the transformed variables "exist," with slope B and intercept *ln* A.

C.2.2.1. Calculating Coefficients of Simple Nonlinear Regression Models

Now that a nonlinear functional form can be expressed in linear form, the method of least squares can be applied in order to calculate predictive values of slope B, intercept *ln* A, and A (i.e., the unit space equivalent of *ln* A).

To fit a power function, first take the natural log of the sample X and Y data. Then apply least squares regression to *ln* X and *ln* Y to solve for the estimated slope ($\hat{B}$) and intercept (*ln* $\hat{A}$) using the following formulas:

$$\hat{B} = \frac{n\Sigma(\ln X \ln Y) - (\Sigma \ln X)(\Sigma \ln Y)}{n\Sigma(\ln X)^2 - (\Sigma \ln X)^2}$$

$$\ln\hat{A} = \frac{\Sigma \ln Y - \hat{B}\Sigma \ln X}{n}$$

$$\hat{A} = e^{\ln\hat{A}}$$

From these estimates for the slope and intercept, the resulting model can be developed in fit space as the equation for a line of the form:

$$\ln \hat{Y} = \ln \hat{A} + \hat{B} \ln X$$

or in **unit space**, where the original or untransformed variables "exist," as a power model of the form:

$$\hat{Y} = \hat{A}X^{\hat{B}}$$

These equations will now be applied to the data set from Table C-6 using a convenient layout to organize the necessary calculations:

Table C-6: Power Model Calculation Layout

X	Y	*ln* X	*ln* Y	*ln* X *ln* Y	(*ln* X)^2
1	100	0.0000	4.6052	0.0000	0.0000
2	80	0.6931	4.3820	3.0372	0.4804
4	64	1.3863	4.1589	5.7655	1.9218
8	51	2.0794	3.9318	8.1758	4.3239
16	41	2.7726	3.7136	10.2963	7.6873
		Σ=6.9314	Σ=20.7915	Σ=27.2748	Σ=14.4134

$$\hat{B} = \frac{5(27.2748) - (6.9314)(20.7915)}{5(14.4134) - (6.9314)^2} = -0.3222$$

$$ln\hat{A} = \frac{20.7915 - (-0.3222)(6.9314)}{5} = 4.6050$$

$$\hat{A} = e^{4.6050} = 99.9830$$

$$\hat{Y} = 99.9830X^{-0.3222}$$

C.2.2.2. Regression Statistics of Simple Nonlinear Models

Similar to the process described for an SLR equation, we proceed to compute the statistics and tests of significance that measure the validity of the nonlinear regression curve. Such statistics can determine how well the CER predicts the dependent variable. They can also tell us whether a trend exists.

Calculation of the Coefficient of Determination (R^2) for Nonlinear Models

Unlike the calculations shown for a linear equation in section C.2.1.3., the statistics for a nonlinear equation are based upon transformed X and Y data. Analysts should be cautious when interpreting statistics from regression models that involve transformations of the variables, particularly the dependent variable, Y. All spreadsheet regression tools and most statistics packages report regression statistics based on transformed data.

Recall one formula for R^2 = *Percent of Explained Deviation*

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

For both the power and exponential models, R^2 is calculated based on the natural log transformation of Y. This is not a linear transformation. For example, suppose you transform the values 100 and 10 from unit space to fit space as follows: *ln* 100 = 4.6 and *ln* 10 = 2.3. Prior to the transformation, one number is 10 times the other. After the transformation, one number is only twice the other. Therefore, when comparing a linear model with a nonlinear model, R^2 for the nonlinear model would be larger than R^2 for the linear model, and the nonlinear model would be favored over the linear. This scaling effect distorts the statistic and makes it impossible to accurately compare linear and nonlinear models using transformed R^2.

A similar problem occurs with other statistics. Recall that the goal of least squares regression is to minimize the sum of squared unexplained errors, SSE.

$$\mathrm{SSE} = \Sigma(\mathrm{Y} - \hat{\mathrm{Y}})^2$$

The units of SSE are Y^2. If the dependent variable has been transformed, as with both the power and exponential models, the SSE shown by all spreadsheet regression tools and most statistics packages is in units of (*ln* Y)^2. Therefore, SSEs between linear and nonlinear models are not directly comparable. The standard error of the estimate, SE, is in units of Y for a linear model and in units of *ln* Y for the power and exponential models, so this statistic cannot be directly compared between linear and nonlinear models.

To overcome these problems, use the following guidelines:

a) Only compare R^2 between models of the same type (i.e., linear versus linear, power versus power, exponential versus exponential), and only then if the models have the same number of independent variables. If R^2 for each equation is shown in unit space (e.g., output from a statistical computer package), then comparison of R^2 is acceptable, and guidelines (b) and (c) are not necessary.
b) Unlike R^2, unit space SEs are comparable between types of models. The model with the smallest unit space SE would be preferred. To compare models of different types, use a statistics package that reports regression statistics, such as SE, in unit space. The statistics for linear models are always unit space statistics.
c) If a computer package that reports regression statistics in unit space is not available, compute unit space SE manually.

Calculating the Standard Error of the Estimate (SE) for Simple Nonlinear Models

The definition and application of SE for a nonlinear equation is no different than that for a linear equation. Therefore, refer to section C.2.1.3. for a more detailed description of SE.

To obtain unit space statistics for a nonlinear model like the power model:

1) use the equation $\hat{\mathrm{Y}} = \hat{\mathrm{A}}\mathrm{X}^{\hat{\mathrm{B}}}$ to compute a fitted value, $\hat{\mathrm{Y}}$, for each observed X value
2) use the following equation to compute unit space SE:

$$\mathrm{SE} = \sqrt{\frac{\Sigma(\mathrm{Y} - \hat{\mathrm{Y}})^2}{\mathrm{n} - \mathrm{k} - 1}}$$

where:

n is the sample size, and

k is the number of independent variables

Continuing with the problem from earlier where the power (log-log) model was developed as $\hat{\mathrm{Y}}$ = $99.9830X^{-0.3222}$ and using a convenient layout for organizing the necessary calculations, the standard error of the estimate can be readily calculated, as shown in Table C-7.

Table C-7: Calculating the Standard Error of the Estimate

X	Y	$\hat{Y}$	$(Y-\hat{Y})^2$
1	100	99.9830	0.0003
2	80	79.9713	0.0008
4	64	63.9650	0.0012
8	51	51.1624	0.0264
16	41	40.9222	0.0061
			Σ=0.0348

$$SE = \sqrt{\frac{0.0348}{5-1-1}} = 0.1077$$

C.2.2.3. Summary: Simple Nonlinear Regression Models

This section presented the rationale for using, and the equation for developing, one type of nonlinear regression model: the power (log-log) model. An example demonstrating the development of this model was presented. Following a discussion on various statistics including R^2, SSE, and SE, the example problems were continued with manual calculations of unit space SE for each model.

C.2.3. Multiple Regression Models (Linear and Nonlinear)

Many situations require two or more independent variables to adequately describe the process or yield sufficiently precise inferences. For example, a regression model for predicting the demand for repair parts for a particular weapon system may use as independent variables the average operating hours, the average miles driven, and an environmental variable such as mean daily temperature. Regression models containing two or more independent variables are called multiple regression models. This section extends the procedures for SLR to multiple regression and discusses some special considerations when using multiple regression models.

C.2.3.1. Regression Coefficients for Multiple Regression Models (Linear and Nonlinear)

The general multiple linear regression model is:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \ldots + \hat{b}_k X_k$$

where $\hat{Y}$ = represents the estimated value of the dependent variable

X_i = ith independent variable

k = number of independent variables

$\hat{b}_i$ = estimated regression parameters

The general multiple power regression model is:

$$\hat{Y} = \hat{A} X_1^{\hat{B}_1} X_2^{\hat{B}_2} \ldots X_k^{\hat{B}_k}$$

where $\hat{Y}$ = represents the estimated value of the dependent variable

X_i = ith independent variable

k = number of independent variables

$\hat{B}_i$ = estimated regression parameters

To ensure the validity of these models, several assumptions must be checked. The assumptions are similar to those for an SLR model.

(1) The functional form is specified correctly.

(2) The values of the independent variables are known without error.

(3) The residuals or error terms, $e_i = Y_i - \widehat{Y}_i$, are independent, normally distributed multivariate random variables with mean zero and constant variance.

(4) The error terms are not correlated.

(5) The data are homogeneous.

(6) The independent variables are independent of each other.

The values of the regression parameters are to be estimated by using the sample data. This estimation is done by employing the method of least squares where we seek to minimize the sum of the squares of the error terms. The method leads to a system of normal equations that can be solved simultaneously to obtain the parameter estimates.

C.2.3.2. Regression Statistics for Multiple Regression Models

As with the nonlinear models, there are new issues that must be accounted for when using multiple regression models.

The Adjusted Coefficient of Multiple Determination: Adjusted R^2

The **coefficient of multiple determination,** R^2, is defined analogously to the coefficient of determination in SLR. It measures the proportionate reduction in the total variation from the mean Y value associated with all the independent variables in the multiple regression model.

$$R^2 = \frac{\sum(\text{Explained Deviations})^2}{\sum(\text{Total Deviations})^2} = \frac{\sum\left(\hat{Y} - \overline{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}$$

The coefficient of multiple determination is sometimes modified to recognize the number of independent variables in the model. This modification corrects for the loss of degrees of freedom resulting from adding independent variables. Fundamentally, the R^2 can either increase or remain the same with the adding of additional independent variables. This is because explained deviation can increase or remain the same while total deviation always remains the same. A modified measure is appropriate and recognizes the increase, but it also accounts for the loss of the degree of freedom for each additional independent variable. This measure is the **adjusted coefficient of multiple determination,** R_a^2.

$$R_a^2 = 1 - \frac{\sum(\text{Unexplained Deviation})^2 / (n-k-1)}{\sum(\text{Total Deviation})^2 / (n-1)} = 1 - \frac{\left(1 - R^2\right)(n-1)}{(n-k-1)}$$

where n = number of data points

k = number of independent variables

R_a^2 may decrease when another independent variable is introduced into a model since the loss of a degree of freedom in the denominator (n-k-1) can more than offset the decrease in unexplained deviation. If all CERs are of the same type (simple linear, nonlinear, multiple) with the same number of independent variables, it is not necessary to calculate R_a^2.

Calculating Standard Error of the Estimate (SE) for Multiple Regression Models (Linear and Nonlinear)

The SE formula (covered in sections C.2.1.3. and C.2.2.2.) can also be applied to multiple linear and/or multiple nonlinear models.

Recall from the previous two sections that the **standard error of the estimate,** SE, is defined as:

$$SE = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{n - k - 1}}$$

where:

- n is the sample size,
- k is the number of independent variables, and
- when k = 1, the formula reduces to the SE formula for simple regression.

The number of independent variables influences the standard error of the estimate. For example, all other things being equal, as the number of independent variables go up, so does the SE. This is because an increase in “k” leads to a decrease in the value of the denominator “n-k-1.”

For the SE formula to have a “real number” solution, sample size (n) must exceed the number of independent variables (k) by at least one. For example, if you want a regression model with 4 independent variables (k = 4), and your sample size (n) is equal to 5, the SE formula denominator will equal the square root “n-k-1” = 5-4-1 = 0. This is a case where the SE formula has no solution; you would need to at least (a) collect more actual X and Y data pairs or (b) use less than 4 independent variables.

C.2.3.3. Residual Analysis for Multiple Regression Models (Linear and Nonlinear)

Once the fitted regression model is developed, the model's appropriateness for its application needs to be examined. Recall the assumptions of multiple regression discussed previously (the residuals should be independent, normally distributed, random variables with mean of zero and constant variance). Graphical residual analysis allows us to see whether the errors satisfy these assumptions. The analysis is made easy by use of automated statistics packages. The plot, as described in Figure C-3, of the residuals against one of the independent variables helps us to evaluate the model. A plot such as this for the residuals versus each independent variable should be examined.

The plot depicted in Figure C-3 of the actual Y values versus the fitted Y values shows that the points form a reasonably linear pattern and are close to the line of identity (45 ° line). The assumption of normally distributed error terms seems reasonable. Also, plotting residuals against fitted Y values can help identify whether the error variance is constant.

Inferences about Population Regression Coefficients for Multiple Regression Models (Hypothesis Tests)

In multiple regression, not only do we test the individual slopes, we also test the overall equation. The F test is used to determine whether the regression relationship between the dependent variable and the set of independent variables is statistically significant.

(1) Formulate the hypotheses.

$$H_0: b_1 = b_2 = ... = b_k = 0$$

$$H_1: \text{Not all } b_k = 0$$

(2) Choose the desired level of significance (α).

(3) Find $F_{p(1-\alpha,\ k,\ n-k-1)}$ from the F distribution table.

(4) Calculate $F_c = \dfrac{R^2 / k}{(1-R^2)/ (n-k-1)}$

(Note: F_c is generally given in the output from a statistical package.)

(5) Make a decision using the following rules:

(a) If $F_c > F_p$, reject the null hypothesis. Conclude that the regression relationship between Y and the independent variables **is** statistically significant to the degree that such a determination can be made by F tests. Proceed to testing each slope as described in the SLR section.

(b) If $F_c \leq F_p$, fail to reject the null hypothesis. Conclude that the relationship between Y and the independent variables **is not** statistically significant. Discard the equation.

Once the existence of a regression relationship has been established, testing of individual regression coefficients can begin. Tests of the individual regression coefficients are done the same way as in simple regression; however, the t statistic must now account for n-k-1 degrees of freedom. Do a t test for each slope coefficient.

(1) Formulate the hypotheses.

$$H_0: b_i = 0$$

$$H_1: b_i \neq 0$$

(2) Choose the desired level of significance (α).

(3) Find $t_{p(1-\alpha/2,\ n-k-1)}$ from the t distribution table.

(4) Calculate t_c (normally given with the computer regression output) .

$t_p = t_{(1-\alpha/2,\, n-k-1)}$ and $t_c = \dfrac{\hat{b}_i - b_i}{S_1} = \dfrac{\hat{b}_i\ -\ 0}{S_1} = \dfrac{\hat{b}_i}{S_1}$,

where $S_1 = \dfrac{SE}{\sqrt{\sum X_i^2 - n\overline{X}_i^2}}$

(5) Make a decision using the following rules.

(a) If $|t_c| > |t_p|$, reject the null hypothesis.

Conclude that the slope **is** statistically significant.

(b) If $|t_c| \leq |t_p|$, fail to reject the null hypothesis.

Conclude that the slope **is not** statistically significant.

The test of a single coefficient is a conditional test, since it tests the significance of b_i, given the presence of the other elements of the model. If we fail to reject the null hypothesis for any slope, then we discard the equation and recompute the regression without that variable.

Checking for Multicollinearity within Multiple Regression Models.

When there are exact dependencies between the independent variables (one or more X values can be expressed in terms of another X value or linear combination of X values), then the least squares solution technique fails. This condition in which the sample observations of the independent variables are highly correlated among themselves is called **multicollinearity**. When multicollinearity is present, we encounter two major consequences:

- The estimated value of the regression coefficient, b_k (for a given independent variable), may vary greatly depending on which other independent variables are in the model. Consequently, the value of the coefficient does not really indicate the contribution of that particular variable.
- The regression coefficients tend to have extremely large standard deviations.

Once the presence of multicollinearity is identified, remedial actions can be taken.

- Obtain additional observations in an attempt to break the pattern of multicollinearity.
- Transform some of the independent variables to lessen the degree of multicollinearity.
- Omit some of the independent variables causing the multicollinearity.

To check for multicollinearity, use the correlation matrix provided by most statistical regression packages, or run linear regressions between all pairs of independent variables. If you are using nonlinear models, transform the variables before performing the linear regression. For any pair of independent variables, if $R^2 > 0.5,\ (|R| > 0.7)$, then there is a relationship between the independent variables that is likely to affect the regression parameters. The resulting tests for significance and other inferences are invalid. That pair of independent variables should not be used together in a regression model. To check for multicollinearity manually, use the following formula for **each pair** of independent variables.

$$R_{x_1x_2} = \frac{\sum X_1X_2 - n\overline{X}_1\overline{X}_2}{\sqrt{\sum X_1^2 - n\overline{X}_1^2}\,\sqrt{\sum X_2^2 - n\overline{X}_2^2}}$$

C.2.3.4. Summary: Multiple Regression Models (Linear and Nonlinear)

This section presented several multiple regression models that can be used when simple regression is not sufficient to explain relationships between independent and dependent variables. Multiple regression models use two or more independent variables in an attempt to explain the relationship with the dependent variable. Three of the most commonly used multiple regression models in cost estimating are the multiple linear, power, and exponential regression models. Multiple regression analysis should be performed to identify as many of these models as seem appropriate. A model selection procedure for determining the best equation is presented in the next section. A number of statistical tests are available for determining if a regression equation is valid. Residual analysis is used to determine if the errors have a constant variance, are independent, are normally distributed, and are random. F and t tests are utilized to determine the statistical significance of the overall regression relationship and individual regression slope coefficients. Multicollinearity is a condition where pairs of independent variables are highly correlated. Whenever multiple regression is performed, it is imperative to determine if multicollinearity exists.

Performing Regression Analysis Using Statistical Software

(1) Consider first trying to fit the (X, Y) data in the form of simple nonlinear model,

$$\hat{Y} = \hat{A}X^{\hat{B}}.$$

If the exponent is close to 1.00, then rerun the functional form as a linear regression,

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X.$$

(2) A similar procedure can also be applied to a multiple nonlinear model,

$$\hat{Y} = \hat{A}X_1^{\hat{B}_1}X_2^{\hat{B}_2} \ldots X_k^{\hat{B}_k}.$$

If any of the exponents are close to 1.00, then rerun that part of the functional form with no exponent. Note that such a process could produce a multiple regression model with both linear and nonlinear characteristics. For example:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1X_1 + X_2^{\widehat{B_2}}$$

(3) If all exponents of the multiple nonlinear model are close to 1.00, rerun the regression as a multiple linear model,

$$\hat{Y} = \hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2 + \ldots + \hat{b}_kX_k$$

(4) In all cases above, check if regression statistics and equation logic are reasonable.

C.2.4. Model Selection Process

This section provides an overview to help the analyst specify the most appropriate regression model. The following steps attempt to ensure good coverage of those principles inherent in sound analysis and model selection. The order in which the steps are performed can be varied. If no automated statistics package is available, it may be more efficient to delay some of the more laborious calculations and graphs until the number of models has been pared down by using the simpler checks. Recall from nonlinear regression that only **unit space statistics** can be used to compare models from different groups (i.e., linear, power, etc.). R_a^2 statistics from transformed models cannot be compared to R_a^2 statistics from untransformed models. If your statistics package does not provide unit space SEs for transformed models, you must calculate SEs manually using the formula described in the nonlinear regression section before comparing SEs (SEs for simple and multiple linear models are already unit space statistics).

(1) If automated with unit space statistics, choose a model with **maximum** R_a^2 or **minimum SE**.

(2) If not,

(a) Group models by type (linear, power, etc.),

(b) Choose model with highest R_a^2 from each group

(c) Calculate unit space SE for each model selected in 2(b) above and select the model with the lowest SE.

(d) Perform significance tests (if manual). If model fails, go to (b) and choose next highest R_a^2,

(e) Check for multicollinearity (if manual). If present, go to (b) and choose next highest R_a^2,

(f) Perform residual analysis (if manual). If bad, go to (b) and choose next highest R_a^2, and

(g) Perform a logic test to ensure that the relationship described in the CER makes sense. If not, go back to (b) and choose the next highest R_a^2.

Finally, if a good CER does not emerge, be aware of nonparametric techniques. There are many other methods of cost estimating besides the statistical approach.

C.2.5. Summary: Parametric Cost Estimating

The regression approach depends on adherence to key parametric assumptions, including random data selection, precise measurement of independent variables, correct specification of functional form, and normal distribution of the error terms. In practice, we must sometimes violate those assumptions, yet still produce a useful CER.

It is imperative that the analyst conduct any regression method with normalized data and then proceed in a step-wise fashion to develop and assess the quality of each CER, ultimately selecting the most preferred CER (whenever possible).

C.3. Engineering Build-Up Cost Estimating

The third type of cost estimate is an engineering build-up, also referred to as a “grassroots-level” approach, or detailed “bottom-up” estimate. The detailed engineering build-up cost estimate is developed from the bottom up by estimating the cost of every activity in a project’s WBS, summing these estimates, and adding appropriate overheads. It is used primarily when there is adequate project maturity to define the scope of work, schedule discrete activities, and determine the resources required to perform those activities. The source and structure of an engineering build-up estimate provides much more detail than estimates by analogy or parametrics. The tradeoff, however, is that producing an engineering estimate is labor intensive, slow and expensive. The strengths, weaknesses, and applications of the engineering build-up method are summarized in Table C-8.

This costing methodology involves the computation of the cost of a WBS element by estimating at the lowest level of detail (often referred to as the “work package” level) wherein the resources to accomplish the work effort are readily distinguishable and discernable. This is often referred to as the Cost Breakdown Structure (CBS) or the Cost Estimating Structure (CES). In most cases, the labor requirements are estimated separately from material requirements. Overhead factors for cost elements such as Other Direct Costs (ODCs), General and Administrative (G&A) expenses, materials burden, and fee are applied to the labor and materials costs to complete the estimate. A technical person who is very experienced in the activity typically works with the cost analyst, who prepares these engineering build-up estimates. The cost estimator's role is to review the grassroots estimate for reasonableness, completeness, and consistency with the project GR&A. It is also the cost estimator's responsibility to test, understand, and validate the knowledge base and data used to derive estimates.

Table C-8. Strengths, Weaknesses, and Applications of Engineering Build-Up Methodology

Strengths	Weaknesses	Applications
Intuitive	Costly; significant effort (time and money) required to create a build-up estimate; Susceptible to errors of omission/double counting	• Production estimating • Negotiations • Mature projects • Resource allocation
Defensible	Not readily responsive to "what if" requirements	
Credibility provided by visibility into the BOE for each cost element	New estimates must be "built up" for each alternative scenario	
Severable; the entire estimate is not compromised by the miscalculation of an individual cost element	Cannot provide "statistical" confidence level	
Provides excellent insight into major cost contributors (e.g., high-dollar items).	Does not provide good insight into cost drivers (i.e., parameters that, when increased, cause significant increases in cost)	
Reusable; easily transferable for use and insight into individual project budgets and performer schedules	Relationships/links among cost elements must be "programmed" by the analyst	

Figure C-10 illustrates a method for deriving an engineering build-up estimate. While this is a simple illustration of the engineering build-up methodology, it is important to remember to conduct other detailed activities such as documenting the Basis of Estimates and schedules and applying wage and overhead rates.

The development of the engineering build-up estimate must be preceded by and based on the scope/objectives definition and estimating structure definition, as well as the WBS, normalized data gathering, and the establishment of Ground Rules and Assumptions (GR&A) that are discussed in Section 2.2.1 of the Cost Estimating Handbook.

Cost estimating guidelines should be documented prior to the development of the engineering build-up estimate. The cost estimating guidelines ensure consistency among the estimates by providing a common set of documents to be used for the development of the estimate. The cost estimating guidelines include or identify the location of the technical baseline definition, WBS, WBS dictionary, top-level schedule, Responsibility Assignment Matrix (RAM), GR&A, and instructions as to how the cost information is to be submitted. The estimator can then follow the guidelines, further develop the WBS and the schedule, and use the appropriate source data to identify the resources required to perform each activity. The basis for the resource estimate is then recorded in the BOE document (see section C.3.3).

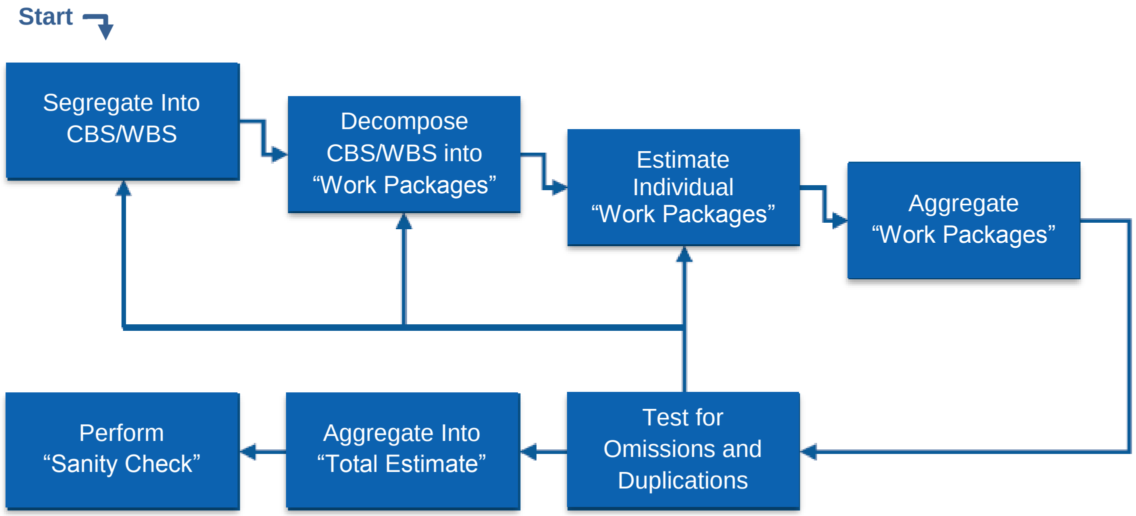
Start
Segregate Into CBS/WBS
Decompose CBS/WBS into "Work Packages"
Estimate Individual "Work Packages"
Aggregate "Work Packages"
Test for Omissions and Duplications
Aggregate Into "Total Estimate"
Perform "Sanity Check"

Figure C-10. Method for Developing Engineering Build-Up Estimate

In addition to being used for detailed cost estimates, engineering build-up cost estimates can be used for cross-checks of mission costs if equivalent costs from other approaches (analogy or parametric) are available. Engineering build-up cost estimates can also be used for what-if analyses and trade studies.

C.3.1. Estimating the Cost of the Job

Developing an engineering build-up cost estimate is a two-part process that includes defining the scope of the job and then estimating the resources required to perform the work. The scope of the project/Program must be defined by the p/P Manager. Resources are labor (hours/Full-Time Equivalents [FTEs], Work Year Equivalents [WYE], and labor categories), procurements/subcontracts, travel, General and Administrative (G&A) expenses, indirect, and Other Direct Costs (ODCs). Indirect costs are those derived from general business expense: a business expense that is not directly connected to a specific product or operation. ODCs are those costs that can be related directly and traced to the production of a product or to a particular function or service.

Often the labor requirements are estimated separately from material requirements. Whenever applicable, overhead factors for ODCs, G&A expenses, and materials burden (e.g., storage fees) are applied to labor and material costs to complete the estimate. Where the activity necessitates, a technical person who is very experienced in the activity typically works with the cost estimator, who prepares these engineering build-up estimates.

Life-Cycle Cost (LCC) includes the total of the direct, indirect, recurring, nonrecurring, and other related expenses both incurred and estimated to be incurred in the design, development, verification, production, deployment, prime mission operation, maintenance, support, and disposal of a project, including closeout, but not extended operations. The LCC of a project or system can also be defined as the total cost of ownership over the project or system's planned life cycle from Formulation (excluding Pre–Phase A) through Implementation (excluding extended operations). The LCC includes the cost of the launch vehicle.

C.3.1.1. Defining the Scope of the Job

To build up the cost, the project must be well understood and broken down into small discrete tasks or activities at the lowest level of the WBS. The discrete tasks may include design, analysis, drawings, board assembly, parts lists, materials, testing, test setup, and chamber time, among many others. The project/Program Manager must provide all the relevant information to the estimator to facilitate the estimator's understanding of the scope of work. Job definition also includes developing the schedule for the lower-level activities. These schedules are the source data used to build the project Integrated Master Schedule (IMS).

C.3.1.2. Estimating the Resources

Once the discrete task or activities have been defined and scheduled, the estimator is ready to estimate the direct, time-phased resources for each element of cost (labor, procurements/subcontracts, travel, indirect, and ODCs) required to perform the defined work activities. Table C-9 identifies the direct elements of cost to be estimated, as well as the typical resources and the types of source data that may be used. Examples of data sources are listed starting with the most important source.

Table C-9. Resources and Source Data by Cost Element

Element of Cost	Resources	Source Data
Labor	• FTEs or hours • Labor categories	• Actual labor hours/labor categories for recent relevant experience • Engineering estimate
Travel	• Reason for the trip • Number of trips, or duration • Number of people • Destinations • Airfare • Per diem	• Recent experience for specific related work outlined in the solicitation • Published Government approved rates for airfare and per diem
Materials	• A list of the materials, parts, equipment, and other items required to perform the scope of work. The list includes the vendor name, the price per unit, and the needed quantity and extended price.	• Recent vendor quotes • Catalogs • Web sites • Historical data • Engineering estimate
Subcontracts	• The cost of subcontracts required to perform the scope of work, including the vendor name, description, and dollar value.	• Recent quotes • Historical data • Engineering estimate
Indirect Cost	• Traceable cost derived from general business expense: a business expense that is not directly connected to a specific product or operation (e.g., computing, maintenance, security, etc.).	• Recent quotes • Historical data • Valid paid and dated invoicing statements
Other Direct Cost	• Traceable cost—not identified in the categories above—to specific services or service centers (e.g., labor, material, fuel, power, etc.)	• Recent quotes • Historical data • Valid paid and dated invoicing statements

The direct labor hours required to complete the work are estimated from engineering drawings and specifications, usually by an industrial engineer (IE) using company or general industry "standards". The engineers also estimate raw materials and purchase parts requirements. The remaining elements of cost, such as tooling, quality control, other direct costs, and various overhead charges including systems engineering and project management, are factored from the estimated direct labor and/or material content of the work. The actual portion of the cost estimated directly is thus a fraction of the overall cost of the system.

An IE (or similar specialist) may use a variety of techniques in estimating the direct labor and material cost of each discrete work element. For example, an IE may use an analogy to estimate one work element; a parametric CER based on an industry database of like work elements to estimate a second work element; and a set of work standards based on work activities (e.g., milling .002 inches from a 6 inch diameter rod 3 inches long) to estimate a third work element.

Uncertainty in this type of cost estimate is due to the use of multiplicative factors on the relatively small direct labor/material base that was estimated. This can result in significant error in the total system cost estimate. The uncertainty, however, can be assessed and managed. Another potential problem is that since the cost estimate is the summation of many estimates, it may be hard to maintain the documentation to support the estimate.

Because, in most cases, an engineering build-up estimate is based on standards, either company-specific or industry-wide, the contractor's cost estimate should be "attainable". By definition, standards are attainable values for specific work under given conditions. The engineering build-up estimate is thus a tool for the manufacturer to control the work on the floor (process control). The technique has its greatest value once the design has stabilized and the system is in production.

As NASA systems development programs tend to be on the leading edge of technology, much effort is spent getting the system to work, which translates into redesign and modifications. This design metamorphosis should be reflected in the engineering estimate. However, engineers may, due to the unknown aspects of the program, underestimate the number of design iterations and therefore underestimate the cost of the work element(s).

The engineering build-up cost estimate is most often used during and after Phase C (Final Design and Fabrication). This technique encourages the contractor to do his homework early on and define all the work down to the lowest level of the WBS. It is also a great process control technique at the production facility. The technique, generally accomplished by hardware manufacturers, is the most costly in time and people.

There are also situations where the engineering community provides their “professional judgment,” but only in the absence of empirical data. Experience and analysis of the environment and available data provide latitude in predicting costs for the estimator. This method of engineering judgment and expert opinion is known as the Delphi method. The cost estimator’s interview skills are important when relying on the Delphi method to capture and properly document the knowledge being shared from an engineer’s expert opinion. Delphi method usually involves getting a group of experts to converge on a value by iterating estimates using varying amounts of feedback. During this process, individuals are generally not identified to the outside and, in some experiments, not identified to each other.

C.3.2. Pricing the Estimate (Rates/Pricing)

The resources required for the job are raw information that must be priced and summarized. Pricing the estimate entails the application of rates to the resources that determine the price for the effort. The rates to be used are specific to the individual contractor or Center. It is usually recommended and preferred that approved forward pricing rates be used. In the event that forward pricing rates do not exist, the individual contractor’s or Center’s published rates should be used. In any case, the rates used for pricing and their source should be fully documented.

C.3.3. Documenting the Estimate—Basis of Estimate (BOE)

The BOE explains how the individual costs used in development of the estimate were derived, including any resources that were determined. The primary components of the BOE are task description, source data, rationale/methodology, documentation, and mathematical calculations.

The following outline expands on the structure and content of what should be included in a BOE:

1. Task Description
    a. Defines the work being performed
    b. Cross-references the WBS
    c. Addresses the specifics of:
        i. *Who* will perform the work
        ii. *What* tasks will be performed
        iii. *Where* and *when* will the work be performed
2. Source Data
    a. Identifies and describes the sources of data used
    b. Sources of data may include the following:
        i. Historical databases
        ii. Cost models
        iii. SME input
        iv. Source quotes and valid paid dated invoice statements
3. Rationale/Methodology
    a. Documents *why* an estimating technique was used
    b. Shows *how* the source data were adjusted for similarities and/or differences and also the associated assumptions and judgments used to develop the estimate
4. Mathematical Calculations
    a. Documents all calculations such as:
        i. Adjustments to actual costs
        ii. Application of complexity factors
        iii. Rates and factors used
        iv. Escalation from the historical data
        v. Time-phasing of the current work effort compared to that of the historical data

C.3.4. Summary: Engineering Build-Up Estimating

The project life-cycle phase and the maturity of project definition will impact the level of detail that can be planned in an engineering build-up cost estimate. The available level of detail will impact the number of BOEs and the level of documentation available for the estimate. A simple example of estimating using the Engineering Build-Up method is provided in the box on the following page. The estimator should be aware that an insufficient level of definition will negatively affect the accuracy of the estimate. The estimator may consider a different methodology than engineering build-up estimating, if this is the case, or seek a hybrid estimate.

Simple Example of Estimating using the Engineering Build-Up Method

Numbers and values associated with WBS, weights and CERs vary from system to system and from service to service. All numeric values shown in this example are for illustrative purposes only. For this example, mass was selected as the unit of measure (UOM). However, there are other commonly used UOMs such as length, square feet, thrust and source lines of code (SLOC).

When estimating by the engineering build-up method for this example, an analyst needs to estimate direct labor hours associated with attaching heat shielding (WBS 06.05.01) to the spacecraft. For example:

<u>DIRECT LABOR HOURS TO ASSEMBLE & ATTACH HEAT SHIELDING</u>

	Heat Shielding Mass	x	Labor CER	=	Direct Labor Hours
WBS 06.05.01	20 kilograms	x	25 hours/kg	=	500 Hours

Keep in mind that the labor to assemble and attach the heat shielding may be performed by a team of assemblers/integrators. Therefore, for example, the 500 hours of direct labor may actually be completed in a 160 hour work-month if multiple people are performing this task.

The 500 hours of direct labor can be converted to direct labor cost by applying a labor rate as follows:

<u>DIRECT LABOR COST TO ASSEMBLE & ATTACH HEAT SHIELDING</u>

	Direct Labor Hours	x	Labor Rate	=	Direct Labor Cost
WBS 06.05.01	500 hours	x	$80/hour	=	$40,000

Overhead cost must also be estimated. This is calculated by applying an overhead rate factor to the direct labor cost:

<u>OVERHEAD COST TO ASSEMBLE & ATTACH HEAT SHIELDING</u>

	Direct Labor Cost	x	Overhead Rate	=	Overhead Cost
WBS 06.05.01	$40,000	x	1.50	=	$60,000

The cost of direct labor plus overhead results in a total "burdened" labor cost of $100,000 to attach the heat shielding to the spacecraft.

The cost of the heat shielding material must also be estimated. This example assumes a material cost of $20,000 per unit kilogram:

<u>MATERIAL COST OF HEAT SHIELDING</u>

	Heat Shielding Mass	x	Material CER	=	Material Cost
WBS 06.05.01	20 kilograms	x	$20,000/kg	=	$400,000

The final step in applying the engineering build-up cost estimating method is to add the given WBS' costs of direct labor, overhead and material. For this example, this is denoted as: Cost $_{\text{WBS 06.05.01}}$ = Direct Labor Cost + Overhead Labor Cost + Material Cost. Substituting our estimates for each cost category yields Cost $_{\text{WBS 06.05.01}}$ = $40,000 + $60,000 + $400,000. Therefore, Cost $_{\text{WBS 06.05.01}}$ = $500,000. Note that Cost $_{\text{06.05.01}}$ is just one of many cost elements of the engineering estimate.

(5) Consider first trying to fit the (X, Y) data in the form of simple nonlinear model,

$$\hat{Y} = \hat{A}X^{\hat{B}}$$