

9.0 PARAMETRIC ESTIMATING

9.1 Introduction

Chapters 9, 10, and 11 discuss extensively the three main estimating methodologies: parametric, analogy, and engineering, respectively. The reader was introduced to these estimating methodologies in Chapter 3, Section 3.3 in the context of the cost estimating process. This chapter provides a full and detailed treatise on parametric estimating.

How a cost estimator develops parametric estimates and evaluates their quality in both a statistical and intuitive sense is provided herein. The chapter begins with a brief overview of parametric estimating (Section 9.2), followed by a history of this type of estimating (Section 9.3). Section 9.4 discusses parametric estimating in greater detail. Section 9.5 explains how the statistical relationship between the cost to be predicted and the cost predictor or cost driver is developed. It also presents the statistical measures that allow the cost estimator to assess the quality of the parametric estimate and the likely accuracy of the estimate. The limitations of parametric estimators are discussed in Section 9.6. Section 9.7 delves into a special type of parametric estimate in wide use in estimating: the learning or cost improvement curve.

9.2 Overview of Parametric Estimating

Parametric estimating is the process of estimating cost by using mathematical equations that relate cost to one or more physical or performance characteristics of the item being estimated. A simple example of a parametric estimate is the use of square footage to estimate building costs. Square footage is a physical characteristic of a building that has been shown through statistical analyses of building trends to be one way of estimating building costs. (Rodney D. Stewart, *The Cost Estimator's Reference Manual*, page 225)

Parametric estimates are often used in the early phases of a system's life cycle. At that stage of the life cycle, basic physical or performance characteristics may be available, but detailed designs may not be. Thus, parametric approaches may be the only option. Even later in a system's life cycle, however, a parametric approach might be used, for instance for certain elements of a detailed estimate.

Parametric estimating equations are often called Cost Estimating Relationships or CERs. In the rest of this chapter, the two terms are used interchangeably. A discussion of the history of parametric estimating will shed some light upon its usefulness.

9.3 History of Parametric Estimating

Parametric estimating resulted from the need for an alternate method of estimating costs early in the development cycle. In the 1950s, the Rand Corporation first began to pursue methodically the development of statistical techniques for estimating the costs of military hardware in the early design phases. The approach worked well for estimating the cost of airframes early in the design process. This technique further evolved as learning curve theory was mated to parametric estimating. The result of joining parametric estimating and cost improvement curve methods

was to allow the estimating of design through production costs early in the development cycle. Parametric estimating was at that time a relatively radical departure from more traditional detailed estimating techniques. The estimating community, however, had discovered in CERs a useful method of producing early life cycle estimates without the time-consuming and input-intensive detailed methods previously in use. Parametric methods enjoy widespread use today. The greater availability of computers helped spur the use of CERs because of the greater ease of doing statistical analyses and handling large amounts of data on a computer. (*Cost Estimator's Reference Manual* by Rodney Stewart, page 227-228)

9.4 Cost Estimating Relationships (CERs)

A CER predicts the cost of some part of a program or of the entire program based on specific design or program characteristics. A CER may be used, for example, to predict the cost of an entire spacecraft based on its in-orbit weight. Software costs are often estimated with a CER based on how many lines of program code are written. One of the oldest relationships uses the weight and speed of an airplane to provide a prediction of the airframe's cost. Another type of CER relates the cost of one program element to another. For example, modification costs often are estimated based on the dollar size of airplane flyaway cost. Equation 9.1 presents an example of a CER, drawn from H.E. Boren and J. Dryden in *A Computer Model for Estimating Development and Procurement Costs of Aircraft*.

Equation 9.1

When using a CER, the cost is unknown, but there is some information about the size, shape, or performance of the piece of equipment to be costed or some information on the dollar size of other cost elements that enables the cost estimator to estimate the unknown cost based on the known information. When developing or using CERs, cost estimators must be aware of the data upon which it was based. Differences between the historic programs and a new program for which a cost estimate is needed may be significant and could render the CER useless, or at least require a major adjustment to the estimate or database. Assumptions and inherent limitations associated with the CER should be addressed prior to its use. Arguments for its validity should be included in the cost estimate documentation.

$ML = 0.63 \times Wt^{0.68} \times S^{1.21}$ <p>Where:</p> <p><i>ML</i> = Non-recurring manufacturing labor hours</p> <p><i>Wt</i> = Airframe unit weight in pounds</p> <p><i>S</i> = Maximum speed at best altitude in knots</p>

CERs have been developed for nearly every major commodity type and cost element and are applied to estimate costs in all phases of a system's life. CERs come in several different functional forms based upon a variety of cost drivers. The next two sections discuss the different types of CERs and their uses.

9.4.1 Types of CERs

CERs can be divided into several classes depending on: 1) the kind of costs to be estimated, 2) the cost drivers chosen to predict costs, 3) the complexity of the estimating relationship, and 4) the aggregation level of the CER. Other classifications are surely possible, but these will be addressed in this section.

CERs Based on the Kind of Costs to be Estimated

The kind of costs to be estimated can be grouped into the three phases of a program's life cycle:

- Research, Engineering and Development (RE&D)
- Production
- Operating and Support (O&S)

These distinctions are important because the kind of costs to be estimated will guide the cost estimator in the search for cost drivers to use in the estimating relationship. O&S cost estimates must consider both equipment characteristics and the support and logistic structure. When estimating maintenance costs, the reliability and maintainability of the equipment are important, but so is the level of maintenance support (e.g., field level, depot level, etc.). The level of maintenance support is a function of the established maintenance concept for that piece of equipment. In contrast, CERs in RE&D generally use equipment characteristics as primary cost drivers and usually are not based on how the equipment is to be developed. Cost estimators, who are estimating production costs, also must estimate cost/quantity relationship curve effects. Sometimes these effects are built into CERs.

CERs Classified by Type of Cost Driver

CERs also are classified by the type of cost driver. Over the years, cost estimators have discovered a variety of quantitative cost drivers to apply to CERs. The most common variable for hardware remains weight and for software, the most common variable is its size. Other system attributes, such as physical, technical, and performance characteristics, also are used. Besides weight, physical characteristics include volume, length, number of parts, and density. Examples of technical parameters (factors that produce performance) include system or subsystem power requirements and scan rate. Performance characteristics include speed, range, accuracy, reliability, etc.

Physical, technical, and performance characteristics are not the only variables that have been used to develop CERs. Cost estimators recognize that hybrid variables like hard drive speed to memory size ratio, the system environment, the system mission and function, and the technological level of the system in relation to the state-of-the-art, can all play an important role in determining costs. There is almost no end to possible quantitative cost drivers.

Cost estimators have long recognized that technology - specifically the degree of technical advance sought in a new system - can affect a system's cost dramatically. However, measuring how far the proposed system is beyond state-of-the-art can be difficult. Currently, cost

Parametric Estimating

estimators use several approaches. One approach is to use time as a proxy for technological advance. Thus, a CER may include the year development begins or the date of first flight as a proxy for the technological advance cost driver. Another approach counts the number of new designs since the first operational system was deployed. Still another approach uses a subjective measure in which the cost estimator, along with system engineers, selects a level of technical advance or system complexity. This can be represented by a continuous variable running from 0 (off the shelf, no new technology) to some number N (brand new technology, major advances in the state-of-the-art); or this variable can be represented by a binary variable, where 1 indicates a major technical advance is required, and 0 indicates no technical advance. Other approaches to quantifying technological advances are possible and should be investigated. To pursue this, the cost estimators must learn as much as possible about both how the system works and what technological improvements will be implemented to increase system performance.

Table 9.1 provides an example of possible Information Technology (IT) cost drivers that a cost estimator might consider when developing an IT CER.

Table 9.1 Potential Airframe Cost Drivers

<u>Physical</u>	
➤	Software size
➤	Number of servers
➤	Length of communications links
➤	Number of sites
➤	Number of positions
<u>Performance</u>	
➤	Processor speed
➤	Communications link speed
➤	Memory capacity
<u>Environment</u>	
➤	Levels of maintenance planned
➤	Support concept
<u>Time</u>	
➤	Date of first operational site
<u>Technological Advance</u>	
➤	Level of technical advance required

Another type of cost driver commonly used in building CERs is the use of one cost element to predict the cost of another element. For example, Engineering Change Orders (ECOs) may be estimated as a percent of the cost of the prime mission equipment. Such cost-to-cost CERs are often used to estimate portions of O&M costs and non-hardware acquisition costs. They are sometimes referred to as factors.

CERs Classified by Complexity of the Estimating Relationship

Equation 9.2

CERs can be simply two variable equations, or they can be complicated multivariate equations. J. Gibson, in *The ASD ECO Model User's Guide* presents the simple CER relating ECOs/Management Reserve (ECO/MR) during full-scale production (FSD) to total FSD costs (TFSDC) in Equation 9.2. An example of a more complex CER is presented in Equation 9.3, from B. W. Boehm and B. K. Clark's 1997 briefing *An Overview of the COCOMO 2.0 Software Cost Model*.

$$ECO/MR = 0.10 \times (TFSDC)$$

Equation 9.3

$$ESLOC = ASLOC \times ((AA+SU)/100 + 0.4xDM + 0.3xCM + 0.3xIM)$$

Where:

ESLOC = equivalent new software size of reused software

ASLOC = size of the software being adapted in Source Lines of Code (SLOC)

AA = rating of the assessment and assimilation of the adaptive software

SU = rating of the current programmers' software understanding of the adaptive software

DM = percent of design modification

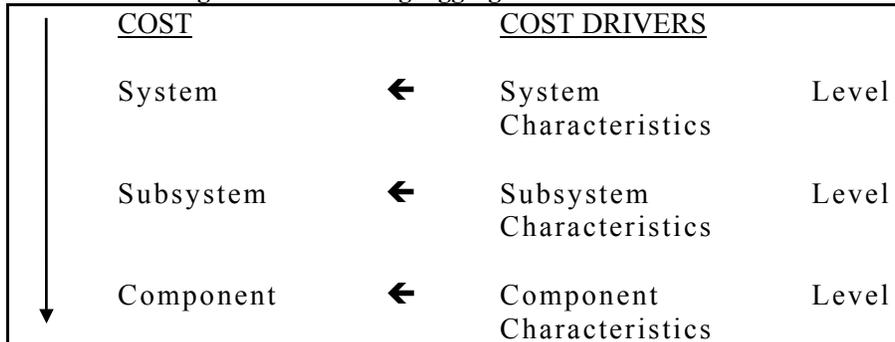
CM = percent of code modification

IM = percent of the original integration effort required for integrating the reused software

CERs Classified by Aggregation Level

CERs can also be classified in terms of the aggregation level of the estimate. For instance, CERs can be developed for the whole system, major subsystems, other major non-hardware elements (training, data, etc.) and components. The aggregation level of the costs to be estimated should be matched by the aggregation level of the cost drivers, as shown in Figure 9.1. For instance, system costs may be estimated as a function of total system weight, while a particular subsystem will be estimated by that subsystem's weight.

Figure 9.1 Matching Aggregation Levels of CERs



9.4.2 Uses of CERs

CERs are used to estimate costs at many points in the acquisition cycle when little is known about the cost to be estimated. As more cost information becomes available, more detailed methods (e.g., engineering methods) of costing become feasible. CERs are of greatest use in the early stages of a system's development. CERs can play a valuable role in estimating the cost of a design approach, especially when conceptual studies and broad configuration trade-offs are being considered.

In the source selection process, CERs can serve as checks for reasonableness on bids proposed by contractors. Many contractors use CERs to help formulate their bids.

Even after the start of the development and production phases, CERs can be used to estimate the costs of non-hardware elements. For example, they can be used to make estimates of O&S costs. This may be especially important when trying to determine downstream costs of alternative design, performance, logistic, or support choices that must be made early in the development process.

9.5 Developing CERs

As discussed earlier, a CER is a mathematical equation that relates one variable such as cost (a dependent variable) to one or more other cost drivers (independent variables). The objective of constructing the equation is to use the independent variables about which information is available or can be obtained to predict the value of the dependent variable that is unknown. A classic CER uses airframe weight, which can be estimated early in an airplane's development, to predict airframe cost, which is not known until much later in the program's life.

To make an estimate using CERs or to assess CERs developed by others, the cost estimator must have an understanding of basic statistics, including the meaning of such terms as mean, standard deviation, correlation, and so on. The reader is referred to Appendix 9B for a refresher on basic statistics.

In most of the discussion of basic statistics in Appendix 9B, the concern is with estimating characteristics of single variable probability distributions. Measures of central tendency (mean, median, and mode) are discussed, as well as two measures of dispersion (range and standard deviation). Two variable distributions are also examined. Scatter diagrams are discussed as a means of exploring the relationship between two variables. The correlation coefficient is introduced as a measure of the strength of the association between two variables. These are subjects the reader should understand before proceeding to the discussion of how to develop a CER, which is based upon statistics.

The purpose of this section is to describe the mathematical steps required to construct a CER and introduce several related statistics used to evaluate the quality of the CER. The discussion

presented here assumes the reader has read, or is otherwise familiar with, the material presented in Appendix 9B. Although the discussion in this handbook is limited to simple CERs (i.e., a single independent and a single dependent variable), the generalization to multiple, independent variables is briefly discussed. Further discussion can be found in more advanced CER texts.

The classical CER example that relates airframe weight to airplane cost is an example of a simple relationship developed from a set of two-variable data. Suppose two measurements were taken on n airframes, where X_i denotes the weight of airframe i and Y_i denotes the cost of airplane i . Then one would obtain a set of n pairs of measurements:

$$\begin{aligned} &(X_1, Y_1) \\ &(X_2, Y_2) \\ &\vdots \\ &\vdots \\ &(X_n, Y_n) \end{aligned}$$

Table 9.2 displays hypothetical cost and weight measurements for 10 airframes. This data will be used to demonstrate the techniques discussed in the remainder of the chapter

Table 9.2 Sample Airframe Cost and Weight Data

Airplane	Cost ¹	Weight ²
727	5.07	9.2
MD-95	7.67	14.8
DC-10	24.01	26.5
DC-9	20.27	18.4
767	13.0	16.4
737	4.04	12.1
MD-80	9.23	12.3
L1011	13.69	16.1
747	17.58	17.6
757	10.99	17.3
(1) Cumulative average cost of the first 100 airplanes produced, in millions of FY 1981 dollars.		
(2) Weight in thousands of pounds.		

The objective in developing a CER is to determine the relationship, if any, between X and Y (e.g., airframe weight and airplane cost). If such a relationship is found, it can be used to predict the costs of a new airplane if the cost estimator has some information on the new airplane's weight. One way to proceed is to construct a functional relationship between X and Y . This procedure is called regression analysis.

The first step in regression analysis is to hypothesize a relationship, usually involving one or more parameters, between the independent and dependent variables. This is discussed in Section 9.5.1. Once a relationship is selected, a curve fitting technique is required to determine the specific values of the parameters. The method of least squares curve fitting is discussed in

Section 9.5.2 and several simple nonlinear models are described in Section 9.5.3. Measures of “goodness of fit” and confidence intervals are presented in Sections 9.5.4 and 9.5.5, respectively. More general methods of regression are provided in Section 9.5.6. A note on computer packages to assist in constructing CERs is given in Section 9.5.7.

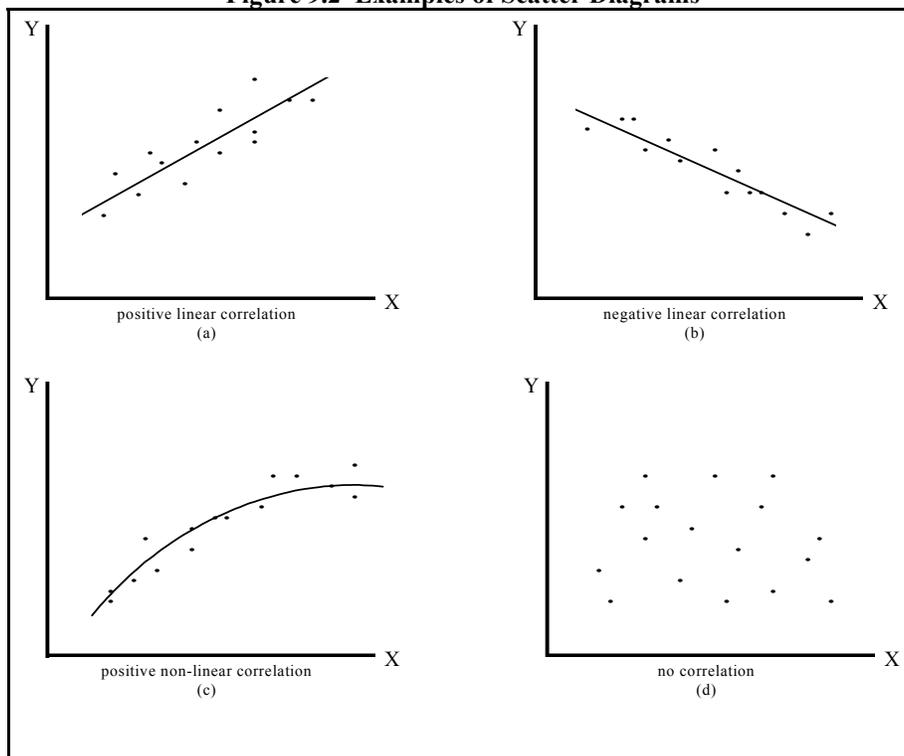
9.5.1 Hypothesizing Functional Relationships

There are essentially two approaches to hypothesizing a functional relationship between the independent and dependent variables in a regression analysis.

The first approach is to hypothesize a relationship on the basis of a priori assumptions. For example, it is reasonable to hypothesize that airframe costs increase as airframe weight increases (at least within a certain range of weight). However, it would not be plausible to assume there is a relationship between sunspots and airplane costs. The cost estimator must review what factors might cause costs to increase and measure them directly or indirectly. The weight relationship seems reasonable because the more material that the airframe comprises, the more one would expect an airframe to cost. Other relationships might be hypothesized for which there is no direct measure. For example, the airframe’s technology level could affect costs, but there is no direct measure of technology. Hence, the cost estimator may resort to an indirect measure such as time. Once the cost estimator has a list of hypothetical relationships, the cost estimator should determine what kind of relationship is expected. Is the relationship expected to be positive (as weight increases cost increases) or negative? Determining this before collecting and analyzing the data enables the cost estimator to judge the reasonableness of the estimating relationship from an intuitive sense.

The second approach is to construct and study a scatter diagram of the two variables. For example, the relationship between the X and Y variables presented in Figure 9.2 (a and b) clearly suggests a linear relationship. Figure 9.2 (c) suggests a non-linear relationship and Figure 9.2 (d) suggests that X and Y are not related at all.

Figure 9.2 Examples of Scatter Diagrams



In practice, it is best to employ both approaches. That is, after hypothesizing one or more functional relationships between the independent and dependent variables, the cost estimator should plot the data on a scatter diagram. If the scatter diagram does not confirm the hypothesized relationship, the cost estimator should rethink the a priori notions and try to explain the discrepancy. There is no simple, direct way of determining a functional relationship; the process requires good judgment and experience that are gained only through repeated use of CERs. Once the relationship has been hypothesized and the data collected and normalized, the cost estimator should use curve-fitting techniques to specify the relationship in mathematical terms.

9.5.2 Curve Fitting Techniques

Two methods for fitting a curve to a set of bivariate data are described in this section. The first method is visual inspection of the scatter diagram and drawing a suitable curve through the data points. This approach has several advantages - it is easy and quick to do, no calculations are required, and consideration can be given to outliers. The principal disadvantage of this approach is that the location and shape of the curve through the data points is based upon individual, subjective judgment.

The second approach is the least squares method. This method has the disadvantage that all data points are given equal weight. The cost estimator cannot give less weight to outliers except by excluding them from the sample altogether. However, the advantages are significant. The

Parametric Estimating

approach results in selection of a best-fitting curve according to a precise definition. Least squares avoids the subjectivity inherent in the graphical approach, and the estimated regression equation facilitates predictions (there is no need to refer to a graphical representation).

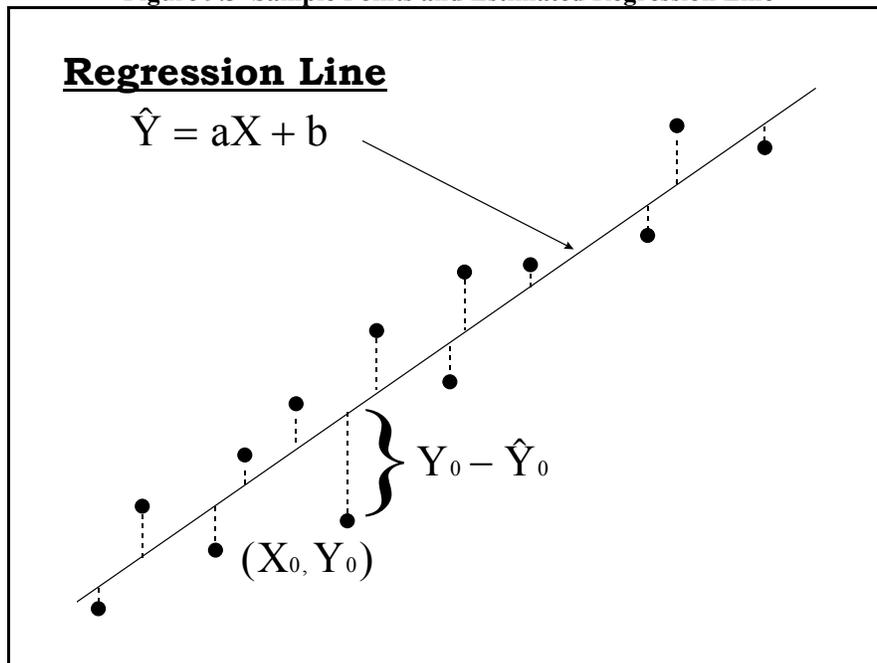
Equation 9.3

This figure depicts a scatter diagram in which an estimated regression line has been drawn through several plotted data points. The vertical distance from the estimated curve to the observed value (X_0, Y_0) is given by $\hat{Y}_0 - Y_0$. If there are n data points, similar distances can be obtained for each of the n (X, Y) pairs. The least squares curve through the plotted data points is defined to be the one that minimizes the sum of the n squared vertical distances, i.e., the curve that minimizes Equation 9.4. To illustrate the least squares method, refer to Figure 9.3.

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \dots + (Y_n - \hat{Y}_n)^2$$

Where:
 \hat{Y} = The expected value of Y which is generated by the regression equation
 Y = Observed values of Y, i.e., data points

Figure 9.3 Sample Points and Estimated Regression Line



Equation 9-4

$$\hat{Y} = aX + b$$

The least squares curve is a straight line of the form of Equation 9.5 where b is the Y-axis intercept and a is the slope of the curve. The least squares method gives rise to unique values of the two parameters b and a . Once these parameters are found, the regression line is completely specified. The formulas for estimating b and a are derived in the addendum to this chapter and displayed in Worksheet 9.1.

FAA Life Cycle Cost Estimating Handbook

The parameters of the regression line can be found by performing the computations indicated in Worksheet 9.1. For example, suppose the cost estimator had cost and weight data on the ten airframes presented in Table 9.2. A completed sample worksheet is given in Worksheet 9.2 that illustrates the computations that the cost estimator would need to make.

Worksheet 9.1 Worksheet for Computing Regression Line Parameters

X_i	Y_i	$X_i Y_i$	X_i^2	Y_i^2 *
X_1	Y_1	$X_1 Y_1$	X_1^2	Y_1^2
X_2	Y_2	$X_2 Y_2$	X_2^2	Y_2^2
.
.
.
X_n	Y_n	$X_n Y_n$	X_n^2	Y_n^2
ΣX_i	ΣY_i	$\Sigma X_i Y_i$	ΣX_i^2	ΣY_i^2
$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$				
$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$				
$a = \frac{\sum_{i=1}^n X_i Y_i - \left(\frac{\sum_{i=1}^n X_i}{n}\right)\left(\frac{\sum_{i=1}^n Y_i}{n}\right)}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}$				
$b = \bar{Y} - a\bar{X}$				
<p>* This column will be used in subsequent computations (see Section 9.4.4).</p>				

The estimated regression line and data points are plotted in Figure 9.4. Note that airframe cost (Y-axis) can be estimated by inspection of the regression curve at any given airframe weight (X-axis). Alternatively, the regression equation computes the airframe cost, given any airframe weight within the range of the data. For example, for a weight of 22,000 pounds, the regression equation developed in Worksheet 9.1 and applied in Worksheet 9.2 yields a predicted cost of \$20.19 million.

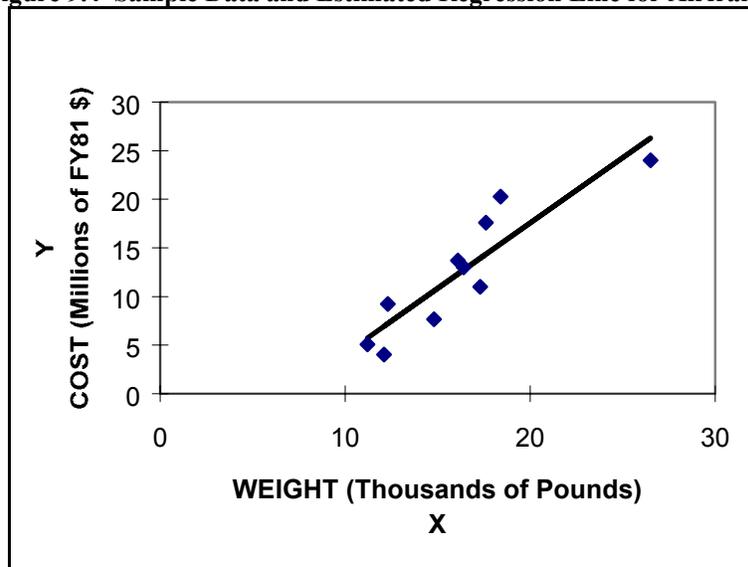
$Y = -9.29 + 1.34X = -9.29 + 1.34 \times (22000 \div 1000) = \20.19 million

Parametric Estimating

Worksheet 9.2 Completed Worksheet for Airframe Example

X_i Weight (lbs ÷ 1000)	Y_i Cost (\$M)	$X_i Y_i$ Weight × Cost	X_i^2 Weight Squared	Y_i^2 Cost Squared
11.2	5.07	56.8	125.4	25.7
14.8	7.67	113.5	219.0	58.8
26.5	24.01	636.2	702.3	576.5
18.4	20.27	373.0	338.6	410.9
16.4	13.0	213.2	269.0	169.0
12.1	4.04	48.9	146.4	16.3
12.3	9.23	113.5	151.3	85.2
16.1	13.69	220.4	259.2	187.4
17.6	17.58	309.4	309.8	309.0
17.3	10.99	190.1	299.3	120.8
$\Sigma X_i = 162.7$ $\Sigma Y_i = 125.6$ $\Sigma X_i Y_i = 2275.0$ $\Sigma X_i^2 = 2820.3$ $\Sigma Y_i^2 = 1959.6$ $\bar{X} = 16.3$ $\bar{Y} = 12.6$ $a = 1.34$ $b = -9.29$ $n = 10$				

Figure 9.4 Sample Data and Estimated Regression Line for Airframes



9.5.3 Simple Non-Linear Relationships

Although this handbook is limited to the development of CERs using simple linear regression techniques, this does not preclude consideration of certain non-linear relationships. By applying appropriate variable transformations, some non-linear relationships can be converted into equivalent linear relationships. In addition to treating simple linear relationships of the form $\hat{Y} = aX + b$, the curve fitting techniques discussed can be applied easily to the non-linear relationships listed in Table 9.3. For example, if the scatter diagram suggests that an exponential relationship might exist, then the cost estimator should first transform all the Y data values by taking their logarithms. The least squares method can then be applied to the transformed data in order to estimate the curve parameters. However, in this case, the least squares estimate of a represents the logarithm of a ($\log a$), and b represents $\log b$ in the exponential curve.

Table 9.3 Simple Non-linear Curves and Variable Transformations

Curve Type	Curve Formula	Equivalent Curve Formula	Req. X-Values	Transform Y-Values	Least Squares Estimator Of Intercept b=	Least Squares Estimator F Slope a=
Hyperbolic	$Y = \frac{1}{aX + b}$	$1/Y = aX + b$	None	1/Y	b	a
Exponential	$Y = ba^X$	$\log Y = \log b + X \log a$	None	$\log Y$	$\log b$	$\log a$
Geometric	$Y = bX^a$	$\log Y = \log b + a \log X$	$\log X$	$\log Y$	$\log b$	a

9.5.4 Determining the Goodness of Fit

In the univariate statistics discussed in Appendix 9B, the standard deviation is introduced as one measure of dispersion. All the variability in a random variable is captured in the standard deviation, regardless of the source. In regression analysis, however, the variability in the dependent variable Y is correlated with the independent variable X.

Figure 9.5 depicts a single observed data point (X_i, Y_i) , the plotted point (\bar{X}, \bar{Y}) computed from the data, and the fitted regression curve. The total deviation of Y_i from \bar{Y} is the sum of the deviation of Y_i from \hat{Y}_i and \hat{Y}_i from \bar{Y} , or mathematically as shown in Equation 9.6.

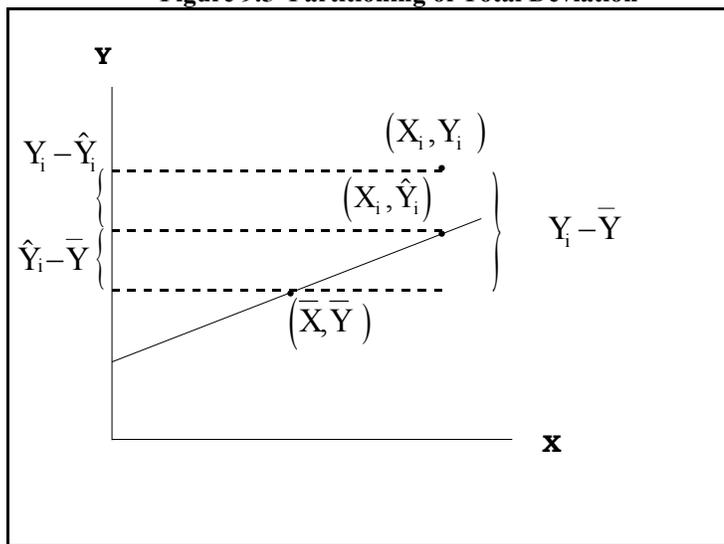
Equation 9.5

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

The second term in the right hand side of Equation 9.6 $(\hat{Y}_i - \bar{Y})$ is explained by the relationship between Y and X,

that is, by the regression of Y on X. The first term in the right hand side, $(Y_i - \hat{Y}_i)$, is due to random variation and, hence, is unexplained. By squaring both sides of Equation 9.6 and applying the summation operator over all n data points, Equation 9.7 is obtained.

Figure 9.5 Partitioning of Total Deviation



Parametric Estimating

Thus, the total variability in Y, given by the left hand side of Equation 9.7, is partitioned into a component that is attributable to the relationship between the dependent and independent variables (explained), and a component that is attributable to random variation (unexplained). The ratio of the explained portion of variability to the total variability provides a measure of the goodness-of-fit of the regression equation to the sample data. This ratio, called the coefficient of determination, is denoted by R². Hence,

Equation 9.6

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$R^2 = \frac{\text{Portion of variation due to regression}}{\text{Total variation}}$$

Equation 9.7

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Recall that $\hat{Y}_i = aX_i + b$ and $b = \bar{Y} - a\bar{X}$ (refer to Table 9.3). By substituting into the numerator of Equation 9.8 for \hat{Y}_i and \bar{Y} , Equation 9.8 can be written as Equation 9.9, given the two mathematical relationships below it.

Equation 9.8

$$R^2 = \frac{a^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}$$

Given:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}$$

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}$$

Recalling the expression for a, from Worksheet 9.1, Equation 9.9 can be written as Equation 9.10.

Equation 9.9

The coefficient of determination ranges between zero and one. Since R² is the fraction of variation explained by the regression, as R² approaches one, the “goodness of fit” increases. If all the plotted data points are close to the regression line, then R² will be close to one (R² equals one when all data points fall on the regression line). As the points become more scattered, R² will move closer to zero. Using the previous airframe example and Worksheet 9.2, R² can be computed as follows.

$$R^2 = \frac{a \left\{ \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{n} \right\}}{\sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n}}$$

Since R²=0.814 (close to one), the estimated regression line fits the data reasonably well. The fraction of variation left unexplained is,

$$1 - R^2 = 1 - 0.814 = 0.186.$$

$$R^2 = \frac{(1.34) \left[2275.0 - \frac{(162.7)(125.6)}{10} \right]}{1959.6 - \frac{(125.6)^2}{10}} = 0.814$$

The sample correlation coefficient r, discussed in Appendix 9B, is the square root of the coefficient of determination (R²). The difference between the two lies in their interpretation. In correlation, r estimates the population correlation coefficient, ρ. In regression, however, the independent variable X is assumed to be non-random. R² is simply a measure of the goodness-of-fit of the regression line.

The confidence one can place in whether a valid relationship exists depends on the computed R² value and the number of data points. Tables for using the t statistic to assess both the slope a and intercept b can be found in many college level statistics books. Although the quality of the relationship is measured by testing the confidence one can place in the a value, the intercept b should be tested to assess the CER’s usefulness in providing high confidence forecasts.

9.5.5 Estimating Confidence Regions

There are several statistical techniques for estimating confidence regions around predicted values. They vary depending on the amount of data available and the data distribution assumptions. Many textbooks, such as R. C. Owen’s *Two-Variable Linear Regression Analysis for Introductory Quantitative Analysis*, describe the use of the Standard Error of the Prediction (SEP). SEP is most applicable for cost estimating activities where data availability is limited. Even more textbooks describe the Standard Error of the Estimate (SEE). However, its use is limited to situations where more data are available and the value of the independent variable for which an estimate is desired is near the mean of the data values. The SEE will give a deceptively narrow prediction confidence interval; therefore, the SEP is a more appropriate measure for cost estimating.

In cost estimating, the typical situation involves a CER developed using a small database (less than 20 data points) and input values that are not close to the mean of the independent variables. This leads to very wide confidence limits for the predicted values of the dependent variable. Cost estimators generally will be better off trying to use a second estimating method to support

their estimates rather than attempt to prove statistically that their cost estimate has a high probability of lying within narrow bounds. Therefore, predictive confidence intervals often are not used in cost estimating.

9.5.6 Generalization of Simple Regression Analysis

Thus far the discussion has been limited to CER development using simple regression analysis: a single independent variable and a single dependent variable. For many cost applications, knowledge about a single key cost driver is all that is required to predict certain cost elements.

In other applications, however, a single independent variable may not be adequate to predict cost reliably. For example, more than one cost driver may be required to describe the manufacturing cost of a component. In these instances, it is useful to broaden simple regression techniques in order to accommodate additional cost drivers. This more general form of regression is called multiple regression because there are multiple independent variables that are used to predict the value of the single dependent variable. Thus, the functional relationship between the independent variables, denoted X_i , and the dependent variable, denoted Y , may have the following linear form if there are p independent variables:

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_p X_p$$

Where α_0 represents a constant and the α_i (for $i = 1, \dots, p$) are the coefficients of the independent variables (analogous to a in the simple regression case). The α_i can represent the relative importance, or weight, of each of the independent variables, provided the X_i are commensurable.

An important assumption of multiple regression is that the independent variables are truly statistically independent of each other (i.e., $r=0$ for all pairs of independent variables). If this is not the case, which it frequently is not, a condition called multi-collinearity is said to exist. However, some multicollinearity can be tolerated. Moderate to severe multicollinearity (values of r over 0.7) will cause problems in using the prediction equation for cost trade studies, where one wants to see how costs vary as a function of individual variables. However, multicollinearity (high r values) can be tolerated when making a single point estimate.

The computations involved in multiple regression are more difficult than those for simple regression; therefore, multiple regressions should be performed using current computer software packages. More advanced textbooks, such as N. Nie et. al.'s *SPSS: Statistical Package for the Social Sciences*, should be referenced for more detailed discussions on multiple regression.

9.5.7 A Note on Computer Applications

Today, there are many regression analysis packages available which can compute the various parameters and statistics used in regression analysis easily. Most computerized statistical packages perform simple and multiple regression, and many of them provide useful information on significance test computations and interpretations. Thus, if CERs are to be used frequently, the cost estimator should investigate how to access and use a statistical package rather than

perform the calculations by hand. In addition, many hand calculators have special functions to perform simple regression.

9.6 Limitations of CERs

Like all estimating techniques, CERs have their limitations. The cost estimator must be fully aware of these limitations to properly convey the degree of confidence one should have in the cost estimate. This section addresses the major limitations associated with using CERs.

9.6.1 Quality and Size of the Database

Credible CERs demand quality data and enough data to estimate the relationship. Quality data means actuals (e.g., actual historical costs, actual weight, speed, etc.). When the cost estimator does not work with actuals, care must be given to estimating and interpreting the CER. Of course, actuals are not always available, forcing the cost estimator to rely on cost data from contractor bids and/or other projections. If a cost estimator were to use the airplane cost CER developed in Section 9.5, the cost estimator probably would not have actual airplane weight, only an estimate of the weight. Unfortunately, actual weight is not available until the airplane is produced and even historical actuals may contain measurement errors and anomalies. Moreover, historical data is often quite time consuming to collect. These factors place limits on the quality of the data available to build CERs. As a result, the cost estimator must be sensitive to these issues.

The size of the database also places limitations on CER credibility. In general, the more data points the cost estimator has, the more confidence the cost estimator will have in the CER and its predictions. Larger values of n will usually result in smaller values of SEE and SEP. For small values of n , the size of the confidence intervals becomes unacceptably large. Thus, the cost estimator must be aware of quantity and quality of the data used to assess the quality of the CER properly. Sample sizes of 30 or more are valuable because they allow one to assume a normal distribution in situations where the Central Limit Theorem is applicable.

9.6.2 Past Costs as Predictors of Future Costs

When using a CER, the cost estimator makes the assumption that information from the past is a good predictor of the future. Therefore, CERs assume that relationships that held true in the past will remain roughly the same in the future. Put another way, one is assuming that all factors affecting costs (e.g., productivity, material type, etc.) will affect future costs in approximately the same way they affected past costs. A CER prediction further assumes that the future program will have several management and technical problems, just as the programs in the historical database.

Parametric Estimating

These assumptions may be unrealistic for two reasons. First, historical relationships between costs and cost drivers can change as technology changes. For example, the increased use of composite materials that are lighter and stronger, but which cost more than previously used metals, offer the prospect of reversing the positive airplane cost/weight relationship. Technology can thus alter the validity of CERs derived historically. Second, it is more than likely that management has learned from previous successes and failures. Managers are trying actively to ensure that a new program will not repeat past management and technical problems.

The cost estimator must consider whether technological changes (including changes in manufacturing technology) may invalidate a CER. Likewise, the cost estimator must review how management practices and acquisition strategy are likely to alter historical cost-to-cost driver relationships. Additionally, studies (Daly, Gates and Schuttling, *The Effect of Price Competition on Weapon System Acquisition Costs*; Kratz, L. A., *Dual Source Procurement: An Empirical Investigation*) show that competition during the production phase reduces unit costs; thus, if the program is to be dual-sourced, the cost estimator may have to consider the effects of competition in the cost estimate.

One way to make these adjustments would be to develop a CER using only a select portion of the data (assuming there is enough data). To develop the CER only those programs that were subject to competition would be included. More advanced regression techniques such as weighting schemes might also be used. It is important to remember that there are some built-in assumptions when using past costs as predictors of future costs, and the cost estimator must be careful when interpreting the results.

9.6.3 Cause and Effect versus Correlation

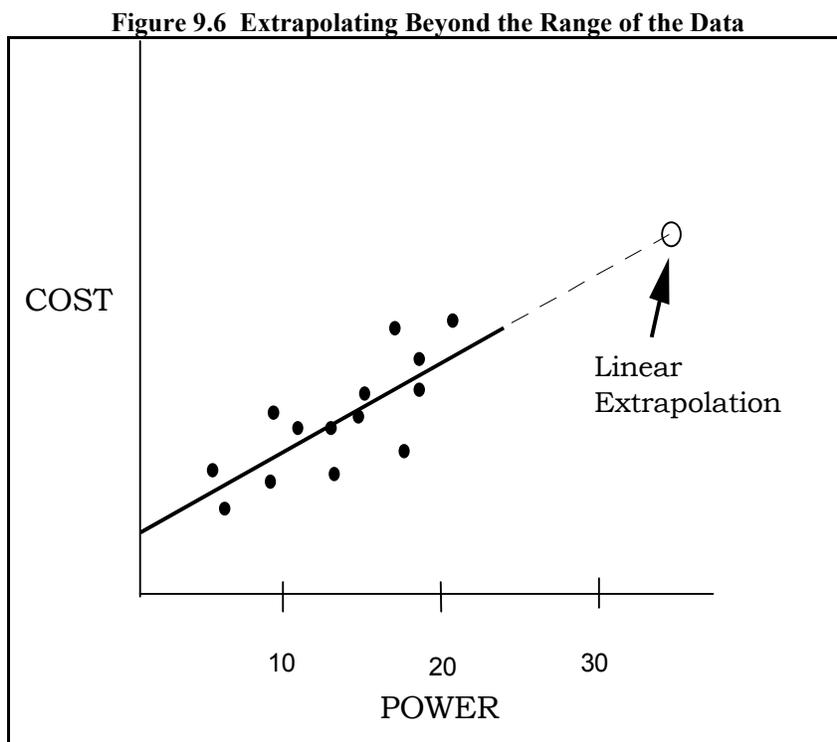
Section 9.5.4 described the computation and meaning of the coefficient of determination, R^2 . The square root of that statistical measure, r , shows the degree of association between the independent variable and the dependent variable. The higher the value of r , the closer the association between the two variables. A high r , however, does not imply there is a cause and effect relationship between the two variables. The cost estimator must provide that interpretation. When doing so, the cost estimator must think through what imputing a cause and effect relationship between the two variables really means. Thus, the cost estimator must ask this question of all potential cost drivers: How do I expect this cost driver to affect cost? One might possibly find a relationship between cost and sunspots, but what cost estimator really expects the occurrence of sunspots to drive cost?

Some relationships that may appear plausible at first glance, in actuality are implausible. For instance, if a cost estimator wanted to examine the hypothesis that a large number of air traffic controllers at an airport indicates better air space management, the cost estimator might regress the number of controllers per airport against the number of flight delays per airport. The resulting regression might show that the greater the number of controllers, the greater the number of flight delays. Does this mean more controllers result in poorer air traffic services? Not likely, instead other explanations could also account for the high R^2 . Large airports have more flights and thus more delays. Another reason might be that larger airports tend to be located in northern areas where there is more inclement weather. When thinking through this

example, the cost estimator concludes that the number of flight delays at an airport may not be the best measure of good air traffic service. The lesson here is to think through the estimating problem before performing the regression because a high correlation does not imply necessarily a cause and effect.

9.6.4 Going Outside the Range of Data Applicability

CERs are derived from a set range of data. Using the CER to extrapolate well beyond that range must be done with great care. For example, in Figure 9.6 cost estimates of power requirements for cooling a site's existing ADP between 5 and 20 kilowatts can be developed with some confidence. A cost estimate for cooling the site after receiving new ADP with power requirements of 35 kilowatts is subject to more uncertainty. Can the cost estimator be sure that the linear cost/power relationship that held for lower power requirements continues at much higher power levels? Clearly, the cost estimator should consider carefully whether such extrapolation is feasible. Some input from knowledgeable engineers could provide valuable guidance on whether to extrapolate the CER.



9.6.5 Tests of Reasonableness

When using any kind of an estimating relationship, the cost estimator should check to ensure that the relationship, cost drivers, and results of a CER are intuitively plausible. The statistics generated in a regression analysis are helpful in this regard. For example:

Parametric Estimating

- Correlation coefficient, r . This statistic should have the same algebraic sign as the regression coefficient a (the slope of the regression line). In other words, if the slope of the regression line is positive, the correlation coefficient should be positive.
- Standard Error of the Estimate (SEE) and Standard Error of the Prediction (SEP). SEE and SEP confidence bounds can be drawn around the regression line to give the cost estimator a sense of the uncertainty associated with the CER.
- Other statistics. The F-statistic and t-statistic (not discussed in this handbook) are useful in establishing the uncertainty associated with the regression coefficients b and a (α_0 and α_i). Refer to any basic statistics book for a discussion of these.

The cost estimator also must examine the relationship form carefully. The relationship between the cost drivers (independent variables) and the cost to be predicted (dependent variable) may be linear within a specified region, but curvilinear at extreme values of the independent variable. For example, component cost may be linearly related to power requirements within a certain range; however, at some threshold, costs may go up at an increasing rate. The cost estimator should try as many functional relationships as feasible.

In addition to statistical evaluation, other things can be done by the cost estimator to ensure a quality estimate and a reliable CER. For example, the estimator can:

- Make a “test” estimate for some recent system that was not included in the database and check to see if the CER’s “test” estimate is in agreement with the actual system cost.
- Perform sensitivity analysis with the CER and show that all results are logical and reasonable.
- Have independent technical experts review and endorse the selection of the cost driver variables used and the reasonableness of sensitivity analysis results.
- Show that the model produced good estimates for those systems in the database most like the new system.
- If possible, gather enough historical data points so the new system’s variable values are within the ranges of those in the database (i.e., avoid the need for data extrapolation).

Finally, the cost estimator must recognize that some cost estimating problems are not amenable to simple regression analysis and that more advanced statistical techniques need to be applied (e.g., multiple regression, multivariate techniques) and perhaps even some non-statistical techniques (e.g., expert judgment, elicitation techniques). In the final analysis, the intuition, experience, and judgment of the cost estimator are indispensable components in developing reliable cost estimates.

9.7 Introduction to Cost Improvement Curves

This section addresses the application of cost improvement curves to the cost estimating process. Cost improvement curves have been called by many names including learning curves, progress curves, cost/quantity relationships, and experience curves. Specific types (i.e., mathematical models) of cost improvement curves often have been named after the men who proposed them or companies that first used them. They include Wright, Crawford, Boeing, and Northrop curves. All of these names refer to one of two mathematical models generally agreed to describe best how costs or labor hours decrease as the quantity of an item being produced increases. These two models are described most accurately as the unit curve and cumulative (cum) average curve. The differences between the two models can be important and will be described later in Section 9.7.2. The differences are important because there are times when use of one model clearly is preferred over the other. The two models use what look like identical equations. However, because of the differences in the definition of the cost or hour term, they compute different total cost or hour values for identical first unit (T_1) and slope values.

The primary purpose of this section is to provide an introduction to basic cost improvement curve theory. While the theory is applicable equally to labor hours and costs - more exactly constant dollar costs - only costs will be addressed in the computation discussions presented in this section.

Throughout this section, the term total production costs will be used. Used herein, it means total recurring production costs; that is the total cost for activities and material requirements that are common to every production unit. Recurring costs do not include non-recurring costs, such as basic and rate tooling, which must be added in most cases to get a true total production cost.

At the outset it must be pointed out that cost improvement curve theory has been found to be a useful estimating tool in the past. However, it is based on observations, most of which do not fit either the unit or cum average curve equations exactly. No one can describe totally the cause and effect mechanisms that produce the cost decreases forecast by the theory. There are many uncertainties associated with cost estimates for future activities. While cost improvement curve analysis methods have been, and will continue to be, useful cost estimating tools, their use is also a source of estimate uncertainty. It is prudent financial management to review actual data from time to time, after the estimate has been made, to determine if cost reduction projections are being met. Section 9.7 provides brief historical, theoretical, and application information on cost improvement curves.

9.7.1 Brief History

Since the first paper on cost improvement curves in the airplane industry was published in the 1930s, much has been written on the subject. Louis E. Yelle, in *The Learning Curve: Historical Review and Comprehensive Survey*, provides over 90 references published before 1967. The most important fact derived from Yelle's research is that in the past, costs have been observed to go down in a somewhat predictable manner as the quantity increased. This has resulted in industry personnel planning and managing to assure the predicted cost reductions are achieved, and the government, as a buyer, expecting to see such reductions in the prices it pays for

systems. Much has been written on what causes costs to decrease. It is agreed widely that the decreases result from many things including - job familiarization by workmen doing repetitious jobs, general improvement in tool design and usage, production control improvements, improved materials flow, reduced scrap, design fixes and simplification, and many other factors. On the other hand, very little is known about the relative magnitude of the reductions associated with each of the many individual sources of improvement or exactly how each component of improvement can be predicted. It is important for an estimator to study the process to which learning is being ascribed before accepting the learning curve as a reasonable estimating approach. For instance, if a company has experienced 85 percent learning curves historically, but recently has automated its process significantly, it is not to be expected that the process will involve as much learning, since machines do not learn. In this section the word learning is used to describe everything being done to reduce costs. Since factors well beyond the usual definition of learning are involved, quotes will be used to indicate this special meaning of the word learning.

9.7.2 Brief Theory of Cost Improvement Curves

As already mentioned, cost improvement curve theory states that as the quantity of items produced doubles, costs decrease at a constant rate. This constant rate will depend on many factors related to the process being modeled. Equations 9.11 and 9.12 describe the learning curve concept.

Equation 9.10

In reviewing Equations 9.11 and 9.12, it is important to note that the form of the equations is the same. Both plot as straight lines when the variables are transformed into

$$Y_x = T_1 \cdot X^b$$

Where:
 Y_x = The cost required to produce the Xth unit
 T_1 = The theoretical cost of the first production unit
 X = The sequential number of the unit for which the cost is to be computed
 b = A constant reflecting the rate costs decrease from unit to unit

their logarithmic form. They differ only in the definition of the Y term. Equation 9.11 describes the basis for the unit curve. It is used to describe or model the relationship between the cost of individual units. Equation 9.12 describes the basis for the cumulative average or cum average curve. It is used to describe the relationship between the average cost of different quantities of units. The significance of the cum in cum average is that the average costs are computed for the first X units. Therefore, the total cost for X units is the product of X times the cum average cost. Unfortunately, there is no easy way to get the exact total cost of the first X units produced using the unit curve theory without a computer, although there are approximation formulas.

Equation 9.11

Both the unit and cum average cost improvement curve equations describe and model the observation that costs decrease a constant percent every time the quantity

$$\bar{Y}_x = T_1 \cdot X^b$$

Where:
 \bar{Y}_x = The average cost of the first X units
 T_1 = The theoretical cost of the first production unit
 X = The sequential number of the last unit in the quantity for which the average cost is to be computed
 b = A constant reflecting the rate costs decrease from unit to unit

doubles. This is reflected in the curves through the b value, a constant reflecting the amount of the decrease for every doubling of quantity. The b value for both curves is computed by Equation 9.13.

As an example using the unit curve, if the first unit cost 100 and the second unit cost 90, or 90 percent of unit 1, the unit curve would have a 90 percent slope, and the S value would be 0.9. The resulting b value would be the $\log 0.9 / \log 2$ or **Equation 9.12**

$-.045758 / 0.30103$, or -0.15200 . The b value is determined in the same way for the cum average curve. However, using the same first unit (T_1) value

$b = \frac{\log S}{\log 2}$
<p><i>Where:</i> $S =$ The cost/quantity slope expressed as a decimal value.</p>

and slope, one will always get lower cum total costs using the cum average curve because of the difference in how Y_x and \bar{Y}_x are defined. In the example above where the first unit cost 100 and the second 90, the total cost for the two is 190, based on use of the unit curve. Using the cum average curve the \bar{Y}_x for the same T_1 value (100), slope value of .9, and x value of 2, would yield a total cost of 2 times 90, or 180.

Since these two models of how costs decrease with quantity are clearly different, a cost estimator must always know which type of curve is required. If provided historical slope data, a cost estimator must know which curve type was assumed to derive the given slope values.

One other piece of theory is important to the applications of cost improvement curve theory. It is that when a procurement, whether new or continuing after a design change, consists of some elements being produced for the N^{th} time and other elements being produced for the M^{th} time, where N and M are not equal, the total cost of the total unit can be estimated using the sum of values computed from two cost improvement curves. The theory extends to any number of curves as long as the T_1 , slope, and quantity values for each are appropriate for the items or fraction of the total item applicable to each curve. This is often the case where two or more systems use the same engine or some other major component.

9.7.3 Importance of Cost Improvement Curves to Cost Estimating

Cost improvement curves have long been recognized in the airframe industry and widely used by industry and government cost estimators. Subsequently, cost improvement curves have been applied to almost all production cost estimates, especially where the quantity of production units involved justified planning and tooling activities greater than those used to produce prototype items. Cost improvement curves can be applicable to production quantities as small as two units if the product is not machine made.

Cost improvement curves are one of the most widely understood concepts of all cost analysis tools. Therefore, cost estimators can expect questions from various levels of management on all aspects of their use in developing a cost estimate. Where quantities exceed 100 units, a change of only a few percent in the slope value can make a large change in the total procurement cost

Parametric Estimating

value. Many managers know this and may challenge the slope values used to argue for higher or lower estimates. The cost estimator must be prepared to defend all cost improvement curve methods, assumptions, and input values used to develop an estimate.

9.8 Summary

This chapter has dealt with the subject of parametric estimating, often used interchangeably with the term CER. Parametric estimating is the process of estimating cost by using mathematical equations that relate cost to one or more physical or performance characteristics of the item being estimated. Since physical or performance characteristics of a system are known early in a system's life cycle, parametric estimating methods are particularly needed for early life cycle estimates, although they are used throughout the life cycle. The use of parametric methods has gained increasing acceptance because of the inherent advantages of the methods; they can generate complete estimates with little detail and relatively small time investment.

9A. Least Squares Formula Derivation

The derivation of the least squares formulas for estimating b (y-intercept) and a (slope) is given below.

- Step 1: Observe Equations 9A.1 and 9A.2 and note that the “F” in Equation 9A.2 must be minimized.

Equation 9A.1

$$\sum_{i=1}^n (\hat{Y}_i - Y_i)^2 = \sum_{i=1}^n (b + aX_i - Y_i)^2$$

Equation 9A.2

$$F = \sum_{i=1}^n (b + aX_i - Y_i)^2$$

- Step 2: Square the expression $(b + aX_i - Y_i)$ in Equation 9A.1 and apply the summation operator to get Equation 9A.3.

Equation 9A.3

$$F = nb^2 + 2ba \sum_{i=1}^n X_i - 2b \sum_{i=1}^n Y_i + a^2 \sum_{i=1}^n X_i^2 - 2a \sum_{i=1}^n X_i \sum_{i=1}^n Y_i + \sum_{i=1}^n Y_i^2$$

- Step 3: Take the partial derivatives of F in Equation 9A.2 with respect to b and a. Then, set these partial derivatives equal to zero to get Equations 9A.4 and 9A.5.

Equation 9A.4

$$\frac{\partial}{\partial b} = nb + a \sum_{i=1}^n X_i - \sum_{i=1}^n Y_i = 0$$

Equation 9A.5

$$\frac{\partial}{\partial a} = b \sum_{i=1}^n X_i - \sum_{i=1}^n X_i^2 - \sum_{i=1}^n X_i Y_i = 0$$

- Step 4: Multiply Equation 9A.3 by $(\sum X_i)$, and Equation 9A.4 by n. Subtract the resulting equations to generate Equation 9A.6.

Equation 9A.6

$$a \left[n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2 \right] - n \sum_{i=1}^n X_i Y_i + \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right) = 0$$

Parametric Estimating
Appendix 9A

- Step 5: Solve Equation 9A.6 for a to generate Equation 9A.7.

Equation 9A.7

$$a = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}$$

- Step 6: Solve for b in Equation 9A.5 to get Equation 9A.8.

Equation 9A.8

$$b = \frac{\sum_{i=1}^n Y_i - a \sum_{i=1}^n X_i}{n} = \bar{Y} - a\bar{X}$$

9B. Basic Statistics for Cost Estimators

Introduction

This is a basic statistics reference for cost estimators.

Probability Distributions

Very few things in life are certain. Just as the actual outcome of a good horse race seldom can be predicted with confidence, the actual cost of an airplane seldom can be predicted to the dollar. Moreover, knowing the cost of one system or cost element in the Work Breakdown Structure may not provide much insight into the cost of another system or cost element because of the differences in technology, manufacturing process, labor skill, etc. How then does the cost estimator assess the uncertainty inherent in a cost prediction?

The cost estimator will not be able to specify with certainty the cost of a given element of the total system cost. The uncertainty, however, can be captured in the form of a probability distribution (sometimes referred to as a frequency distribution) on that cost element. A probability distribution gives two basic pieces of information:

- The possible values or range of values that the cost element might assume; and
- The likelihood that each of these values will be realized.

Figure 9B.1 depicts several probability distributions. In constructing a probability distribution, the only mathematical requirements are:

- That the probability assigned to each possible value (given by the height of the curve) be non-negative; and
- That the area under the curve sum to one.

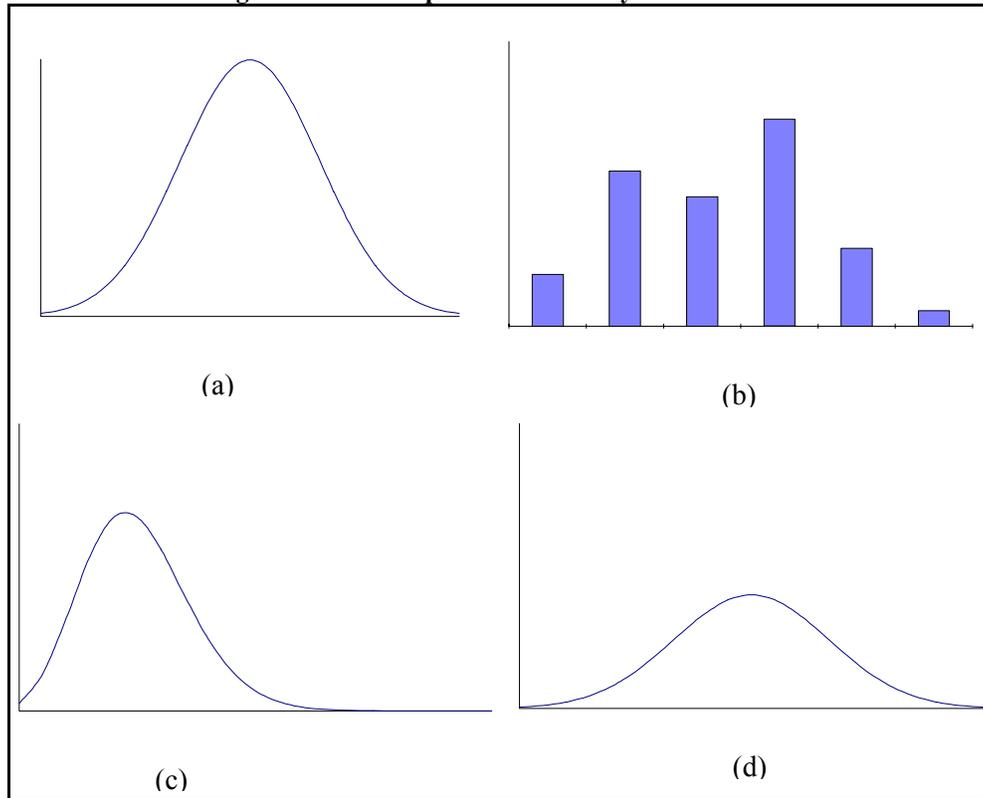
The possible shapes are limitless. The height of the curve above the X-axis represents the relative likelihood that the cost value lying immediately below it will be realized.

Graphically, probability distributions may be depicted as smooth curves or histograms. Figure 9B.1, parts (a), (c), and (d) depict smooth curve distributions, and part (b) depicts a distribution in histogram form. The difference between smooth curve and histogram forms lies largely in how the distribution is constructed. In the histogram form, the cost estimator groups data into specific intervals (e.g., cost intervals) and centers each of the histogram bars on the midpoints of the intervals. For example, if 20 of a total of 100 cost observations fell in the interval \$100,000 to \$150,000, the cost estimator would assign a probability of 0.2 (20/100) to a histogram bar centered at \$125,000. Smooth curve distributions may be generated from histograms by drawing a smooth curve through the midpoint of the top of all the histogram bars.

Parametric Estimating
Appendix 9B

Smooth distributions also may reflect certain shapes that correspond to specific analytical distribution forms. By knowing the distribution's parameters, one can simply plot the distribution. For example, part (a) of Figure 9B.1 depicts a normal, or bell-shaped, probability distribution. Normal distributions provide the basis for many statistical estimation theories.

Figure 9B.1 Examples of Probability Distributions



Another important property of probability distributions is symmetry. Symmetry must always be measured relative to some point, line, plane, or other geometric reference. The symmetry in a probability distribution is specified relative to the mean. Figure 9B.1(a) depicts a symmetrical distribution as does Figure 9B.1(d). Skewness is a property of asymmetrical distributions. Roughly speaking, a skewed distribution is one that has a long tail at one end. Figure 9B.1(c) depicts a distribution that is skewed to the left.

Most of the information contained in a distribution is reflected in its shape. Two characteristics of shape are:

- a tendency for data values to concentrate around certain values, or
- a tendency for data values to disperse.

Cost Estimating Handbook
Appendix 9B

The next two sections are devoted to discussions of measures of central tendency and measures of dispersion. When two or more variables are under consideration, other statistics become important. Since CERs seek to exploit the relationship between two or more variables (e.g., cost and weight), the last section discusses a measure of association between two variables.

When using statistical methods, a cost estimator needs to understand whether his data constitute a population or a sample. A population consists of all the data of a specified type. A sample consists of part of a population, selected at random from the entire population. Some statistical formulas vary depending on whether the data being used is a population or a sample. As a general convention, Greek letters are used for population parameters and English letters for sample parameters.

Measures of Central Tendency

When analyzing historical cost data, it is often observed that while costs may vary over some range, there is a tendency for observations to cluster around certain values. In a sense, this clustering locates the middle of the distribution. It is desirable to identify the value corresponding to the center of distribution, but this depends on how this middle value is defined. Different definitions give rise to different measures. In this section, three measures of central tendency are given - mean, median, and mode.

Mean

The most commonly used measure of central tendency is the mean or arithmetic average. The mean of a probability distribution has a geometric interpretation. It represents the middle of the distribution in the sense that it is the center of gravity. If the distribution were balanced on a fulcrum, the X value corresponding to the point of balance would be the mean value, denoted by \bar{X} .

For a given set of n values X_1, X_2, \dots, X_n (e.g., number of lines of computer code written by each of n programmers during a single hour), the mean is their sum divided by n, the number of values in the set. The mean is expressed mathematically in Equation 9B.1.

Equation 9B.1

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

where \sum is the summation operator

Parametric Estimating
Appendix 9B

CASE STUDY 9B.1. CALCULATING THE MEAN

Nine programmers picked at random were given the same programming task. After one hour, their coding sheets were collected and the following results were noted:

Programmer (i)	Lines of Code Written (Xi)
1	22
2	21
3	34
4	18
5	22
6	12
7	22
8	28
9	21

The mean number of lines of code written for this group of nine programmers is computed as:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9)$$

$$\bar{X} = \frac{1}{9}(22 + 21 + 34 + 18 + 22 + 12 + 22 + 28 + 21)$$

$$\bar{X} = \frac{200}{9}$$

$$= 22.2 \text{ lines per hour}$$

Median

Another measure of central tendency is the median or middle value of the probability distribution. The median is that value that bisects the probability distribution into two areas of equal size. The median is equivalent to the 50th percentile. This means that 50 percent of the probability lies above the median and 50 percent lies below. In other words, one is just as likely to observe values above the median as below it.

The median is frequently a more useful measure of central tendency than the mean, especially if the distribution is highly skewed. Highly skewed distributions tend to force the mean away from the median. The greater the separation between the two, the more important the choice of measure becomes. If the cost estimator has reason to believe that exceptionally high (or low)

Cost Estimating Handbook
Appendix 9B

values experienced in the past are very unlikely to repeat themselves in the future, then the median may be the better choice. On the other hand, if widely divergent values are expected to persist into the future, the mean may be more appropriate, since it implicitly gives more weight to outlying values. A good guideline is to use the more conservative estimate (i.e., the one leading to a higher cost estimate).

CASE STUDY 9B.2. CALCULATING THE MEDIAN

To calculate the median, it is first necessary to arrange the data in ascending order. Continuing our previous example, the data arrange as follows:

Programmer (i)	Lines of Code Written (Xi)
6	12
4	18
2	21
9	21
1	22
5	22
7	22
8	28
3	34

Since the median is the middle value of the frequency distribution and 22 lines of code is the middle value, then 22 is the median of this distribution.

In the example above, the number of data points was odd ($n=9$). The reader may ask how one goes about finding the median when the number of data points is even. The answer is that one averages the two middle values. For example, if a tenth data point is added corresponding to 19 lines of code, the data would display as follows (in ascending order).

Programmer (i)	Lines of Code Written (Xi)
6	12
4	18
10	19
2	21
9	21
1	22
5	22
7	22
8	28
3	34

Parametric Estimating
Appendix 9B

The median is computed as:

$$\text{Median} = \frac{X_9 + X_{10}}{2} = \frac{21 + 22}{2} = 21.5$$

Suppose now that another, especially bright programmer is added who writes 72 lines of code in one hour. The new mean for this frequency distribution of eleven data points is:

$$\bar{X} = \frac{1}{11}(12 + 18 + 19 + 21 + 21 + 22 + 22 + 22 + 28 + 34 + 72) = 26.5$$

The median is now 22. But which of these two measures of central tendency is more appropriate to use? Note that only three programmers out of eleven wrote more lines than the mean value of 26.5. The new mean value is heavily influenced by the large number of lines of code written by programmer 11. That is, programmer 11 is an outlier. Since this programmer has exceptional capability, the mean is biased and does not represent the preponderance of programmers. Therefore, the median (22) gives a better indication of the center of the frequency distribution.

Mode

The last measure of central tendency to be discussed in this chapter is the mode. The mode is simply the most frequently observed value, that is, the X-value corresponding to the highest point in the frequency distribution. The mode cannot be computed algebraically and must be determined by inspection of the frequency distribution. In the previous example, the mode is 22 since this value occurs most often.

Some distributions will have more than one mode (bimodal distribution). That is, there are two X-values around which data values tend to cluster. Other distributions may not have a mode at all if there are no repeated data values.

The normal probability distribution, which will be discussed in a later section, has the property that the mean, median, and mode all have the same value.

Measures of Dispersion

The last section was devoted to a discussion of measures of central tendency. The propensity for data values to concentrate around certain X-values. This section is devoted to just the opposite - the tendency for data values to spread. Two measures of dispersion are discussed below - range and standard deviation.

Range

The range is a simple statistic that represents the difference between the extreme values of the distribution. It is computed by taking the arithmetical difference between the largest and smallest data values.

CASE STUDY 9B.3. CALCULATING THE RANGE

Continuing our programming example, the largest value was 72 lines of code per hour; the smallest was 12 lines. The range is computed as follows:

$$\text{Range} = 72 - 12 = 60 \text{ lines of code per hour}$$

The range is of limited value as a measure of dispersion because it does not depict the shape of the distribution - merely the range of values over which observations have been taken. Moreover, the value of the range has no meaning except in relation to the magnitude of the mean (or other measure of central tendency). For example, a range of 1,000 lbs. is small in the context of comparing airframe weights but large in the context of comparing the weight of avionics boxes. The standard deviation, which is discussed next, provides an answer to this measurement problem.

Standard Deviation

The standard deviation provides a standard measure of the degree of dispersion in a probability distribution. It is defined according to the formula in Equation 9B.2 for a sample, and Equation 9B.3 for a population.

Equation 9B.2

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

Equation 9B.3

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (X_i - \mu)^2}{n}}$$

To compute the standard deviation of a sample, first compute the sum of squared deviations of the individual observations from the mean (this is the numerator under the radical). Then, divide this result by (n-1). Finally, take the square root.

CASE STUDY 9B.4. CALCULATING STANDARD DEVIATION

The standard deviation of the number of lines of code written per hour is computed by:

- Step 1: Compute the sum of the squared deviations from the mean ($\bar{X} = 26.5$) as shown in Worksheet 9B.1.

Parametric Estimating
Appendix 9B

- Step 2: Divide the result obtained in Step 1 by (n-1).

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{2588.75}{10}$$

- Step 3: Take the square root of the result obtained in Step 2.

$$\sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = \sqrt{258.875} = 16.09$$

Worksheet 9B.1. Computation of Sum of Squared Deviations

$(X_1 - \bar{X})^2$	=	$(12-26.5)^2$	=	210.25
$(X_2 - \bar{X})^2$	=	$(18-26.5)^2$	=	72.25
$(X_3 - \bar{X})^2$	=	$(19-26.5)^2$	=	56.25
$(X_4 - \bar{X})^2$	=	$(21-26.5)^2$	=	30.25
$(X_5 - \bar{X})^2$	=	$(21-26.5)^2$	=	30.25
$(X_6 - \bar{X})^2$	=	$(22-26.5)^2$	=	20.25
$(X_7 - \bar{X})^2$	=	$(22-26.5)^2$	=	20.25
$(X_8 - \bar{X})^2$	=	$(22-26.5)^2$	=	20.25
$(X_9 - \bar{X})^2$	=	$(28-26.5)^2$	=	2.25
$(X_{10} - \bar{X})^2$	=	$(34-26.5)^2$	=	56.25
$(X_{11} - \bar{X})^2$	=	$(72-26.5)^2$	=	2070.25
$\sum_{i=1}^n (X_i - \bar{X})^2$			=	2588.75

Thus the standard deviation in the example above is 16.09 lines of code per hour. Note that the standard deviation is expressed in the same units as the variable being analyzed.

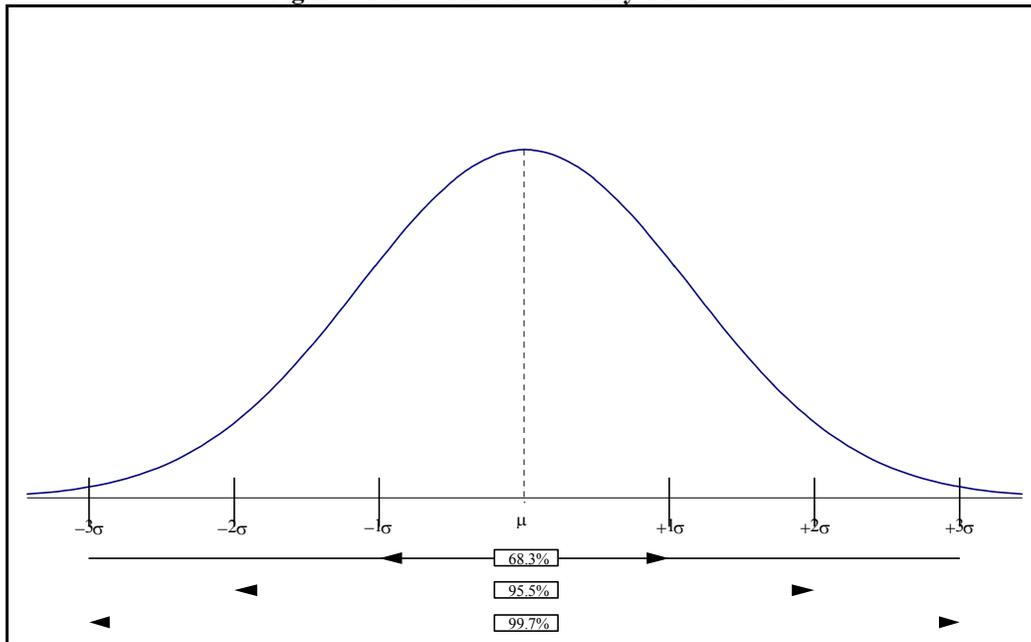
The standard deviation provides a standard measure of dispersion. Knowing the standard deviation allows one to assign probabilities that observations will occur in various intervals over the full range of the distribution. This is true regardless of the nature of the probability distribution. However, confidence limits depend on the population's distribution and on whether one is dealing with a population or a sample.

Cost Estimating Handbook
Appendix 9B

Figure 9B.2 depicts a normal probability distribution with a mean of μ and a standard deviation of σ . Many natural phenomena obey the normal probability law and, hence, have normal probability distributions. The normal distribution is symmetrical about its mean. Note in Figure 9B.2 that the mean, μ in the normal distribution is also the mode and the median. The only other parameter needed to define a normal distribution completely is the standard deviation, σ .

As stated above, the assumption that the data follow a normal distribution allows one to make some assertions about the probability that observations will fall within a specified interval. In the case of a normal distribution, 68.3 percent of the observations will fall within one standard deviation of the mean, 95.5 percent within two standard deviations of the mean; and 99.7 percent within three standard deviations. (Normal probability tables exist in most statistics texts.) One can also specify the probability that an arbitrary value of X_0 or less will be observed if μ and σ are known. This is important since frequently one would like to know whether a specific observation constitutes a likely or unlikely event.

Figure 9B.2 Normal Probability Distribution



If two normal distributions have the same mean but different standard deviations, the one with the larger standard deviation has greater dispersion. This is true in the case of normal distributions but not necessarily true in the case of asymmetrical distributions. Two distributions can have the same mean and standard deviation but widely differing shapes. Only by looking at either the entire distribution or more detailed statistics (which are beyond the scope of this handbook) can one gain a full appreciation of the uncertainty contained in the distribution. Hence, the standard deviation (as a single measure of uncertainty (or risk)) must be used with caution.

Parametric Estimating
Appendix 9B

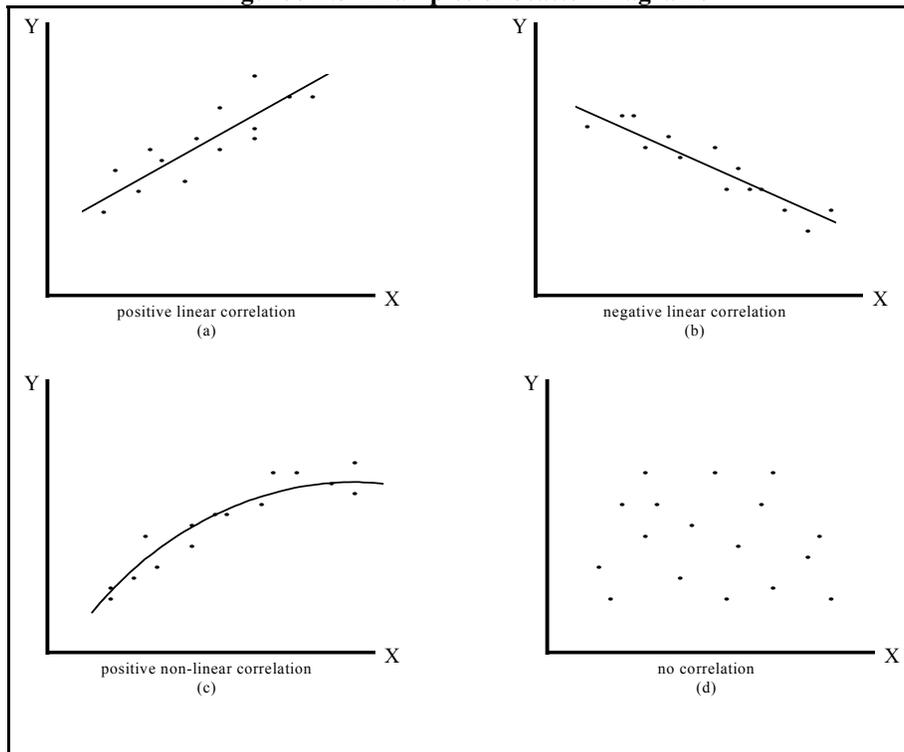
A Measure of Association

In cost estimating, cost estimators attempt to find relationships between two or more variables. Some of these relationships are deterministic (certain) in nature. For example, the relationship between programmer labor cost and labor hours is deterministic if a single labor rate is used for all programmers. If labor hours are known, then labor costs can be stated with certainty.

Other relationships are probabilistic (uncertain) in nature. For example, the relationship between lines of code written and software labor costs depends on the difficulty of the coding task and the proficiency of the programmer. If only given the number of lines of code written, one cannot assert with certainty the programming cost. However, one can measure the strength of the association between these two variables.

A good way to represent the relationship is by means of a scatter diagram. Let X_i represent the number of lines of code written by programmer i and Y_i represent the cost incurred. If there are n programmers, then there are n points (X_i, Y_i) which, when plotted, yield a scatter diagram of these two variables. Figure 9B.3 depicts four possible scatter diagrams that might result from plotting the n points. (There are other possibilities as well.)

Figure 9B.3 Examples of Scatter Diagrams



Part (a) of Figure 9B.3 depicts a situation in which Y tends to increase in proportion to X ; this situation reflects positive or direct linear correlation. Part (b) depicts negative or inverse linear correlation (i.e., as X increases, Y tends to decrease proportionately). Part (c) depicts a situation

Cost Estimating Handbook
Appendix 9B

where Y tends to increase as X does, but at a non-proportional or decreasing rate; this situation characterizes positive non-linear correlation. Finally, part (d), depicts a situation in which there is no apparent correlation between X and Y.

In essence, cost estimators need a measure that captures the strength of the association between X and Y. The correlation coefficient provides such a measure.

Correlation Coefficient (r)

The sample correlation coefficient (r), is an estimator of the population correlation coefficient (ρ). The correlation coefficient is a unitless measure of the degree of linear association between two random variables. The formula for computing the sample correlation coefficient follows.

Equation 9B.4

$$r = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sqrt{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} \cdot \sqrt{\sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}}}$$

The sample correlation coefficient can vary between -1 and +1, inclusive. If r = +1 (-1), the correlation is said to be perfectly positive (negative) which means that all sample data points lie on a straight line.

CASE STUDY 9B.4. CALCULATING THE COEFFICIENT OF CORRELATION

Suppose that in addition to the number of lines of code written by each of the nine programmers, the number of months of programming experience was also identified. The data set now appears as follows:

Programmer (i)	Lines of Code Written (X _i)	No. Months of Experience (Y _i)
1	22	16
2	21	18
3	34	22
4	18	15
5	22	33
6	12	9
7	22	40
8	28	38
9	21	30

Parametric Estimating
Appendix 9B

Worksheet 9B.2 illustrates the computations required to obtain the sum of squares and cross products. Substituting into equation 9B.4 gives the following result:

$$r = \frac{5156 - \frac{(200)(221)}{9}}{\sqrt{4742 - \frac{(200)^2}{9}} \cdot \sqrt{6403 - \frac{(221)^2}{9}}} = \frac{244.89}{(17.25)(31.24)} = 0.454$$

Thus, there is a mild positive correlation between programmer productivity and experience. This suggests that there are other variables that are just as important as experience (e.g., education, motivation, intelligence, etc.). The amount of confidence one can place in the correlation between two variables depends on the value of r and the sample size.

**Worksheet 9B.2. Computation of
Correlation Coefficient (r)**

X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
22	16	484	256	352
21	18	441	324	378
34	22	1156	484	748
18	15	324	225	270
22	33	484	1089	726
12	9	144	81	108
22	40	484	1600	880
28	38	784	1444	1064
21	30	441	900	630
200	221	4742	6403	5156
$\Sigma X_i Y_i = 5156$		$\Sigma Y_i^2 = 4742$		
$\Sigma X_i = 200$		$\Sigma Y_i^2 = 6403$		
$\Sigma Y_i = 221$		$n = 9$		

A calculated r value of 0.454 is low; therefore, the data associated with it could not be used to make high confidence estimates. The F and t statistics should be used to assess a confidence measure of the relationship between the variables. The F and t statistics are discussed in most statistical texts and can be computed by nearly all statistical computer packages.