

# 15. QUANTITATIVE ANALYSIS TECHNIQUES

## Table of Contents

<b>15.1 OVERVIEW.....</b>	<b>15-2</b>
<b>15.2 BASIC STATISTICS DESCRIBING CENTRAL TENDENCY.....</b>	<b>15-2</b>
15.2.1 The Arithmetic Mean.....	15-2
15.2.2 The Median.....	15-3
15.2.3 The Mode.....	15-3
<b>15.3 SAMPLING.....</b>	<b>15-4</b>
15.3.1 Determining if Sampling is Appropriate.....	15-4
15.3.2 Selecting a Sample.....	15-4
15.3.3 Analyzing the Sample and Applying Findings.....	15-7
<b>15.4 REGRESSION ANALYSIS.....</b>	<b>15-8</b>
15.4.1 Regression Analysis Concepts.....	15-9
15.4.2 Evaluating the Performance of the Regression Equation.....	15-13
15.4.3 Advanced Topics in Regression Analysis.....	15-20
15.4.4 Uses of Regression Analysis.....	15-22
<b>15.5 LEARNING (IMPROVEMENT) CURVES.....</b>	<b>15-35</b>
15.5.1 Uses of Learning (Improvement) Curves.....	15-36
15.5.2 Developing and Analyzing Improvement Curves ...	15-36
15.5.3 Note on Computer Models.....	15-42
<b>15.6 SUMMARY.....</b>	<b>15-42</b>

## 15.1 OVERVIEW

This chapter discusses the quantitative analysis techniques used in pricing analysis. Quantitative analysis entails using numerical, measurable data to perform analysis of a subject. For the analyst, the subject may be a price list, contract cost data, or a cost proposal.

## 15.2 BASIC STATISTICS DESCRIBING CENTRAL TENDENCY

The analyst's job involves collecting, evaluating, and reporting on cost and pricing data. Cost and pricing data are quantitative in nature and can be considered statistical information. As such, the data can be analyzed using descriptive statistics. One group of descriptive statistics provide information pertaining to the central tendency of data (the tendency of data to group around a central point). Due to this clustering, it is possible to develop values that are descriptive of the entire data group. An analyst can use these statistics to determine whether a price quote, labor rate, or other data are unusually above or below the market average. The most common descriptive statistics used to measure central tendency are: the arithmetic mean, the median, and the mode. Additional descriptive statistics will be explained in subsequent sections, as they become relevant.

### 15.2.1 The Arithmetic Mean

The **arithmetic mean** (or average) is the most common measure of central tendency. It is calculated by summing all of the numerical data and dividing the sum by the total number of data involved.

#### EXAMPLE: CALCULATING THE MEAN

- The prices of color monitors are:

\$260 \$279 \$255 \$265 \$259 \$270

- The mean is calculated as follows:

$$[(260 + 279 + 255 + 265 + 259 + 270) \div 6] = [1,588 \div 6] = 264.67$$

### 15.2.2 The Median

The **median** is the middle value in an ordered sequence of data. To find the median, the analyst must arrange the raw data according to value. This is called an **ordered array**. Once the data have been arranged in this fashion, the analyst can determine the median point. The median point is determined by the positioning point formula:

$$(n+1) \div 2$$

Where:  $n$  = the number of observations

The resulting number produces the location of the median value within an ordered array.

#### EXAMPLE: CALCULATING THE MEDIAN

Using the color monitor prices, the median can be determined as follows:

- 1.) Place the raw data into an ordered array:

\$255 \$259 \$260 \$265 \$270 \$279

- 2.) Utilize the positioning point formula:

$$(n+1) \div 2 = (6+1) \div 2 = 3.5$$

The median will be at location 3.5, which is between the prices \$260 and \$265. Since the median is not an actual observation in this example, it is necessary to take the average of the third and fourth observations as the median. Therefore, the median equals \$262.5  $[(260 + 265) \div 2]$ .

The median in the above example was not part of the observed data. However, this is not always the case. If the sample size of observed data had been an odd number, then the median would have been an observed number.

In some cases, the median may be more accurate than the arithmetic mean. For example, the hourly wage rates for the general population may include an unusually high wage earner that skews the mean. The median would not be subject to the same distortion.

### 15.2.3 The Mode

Several observations in a set of data may share the same value. The value in a group of data that appears with the greatest frequency is the **mode**. The mode is not a measurement; rather, it is a statistic used to describe a group of collected data. Not every set of data will have a mode, e.g., the set of color monitor prices in the previous example.

### 15.3 Sampling

The analyst often needs to evaluate hundreds of elements in a cost proposal. Evaluating each individual element can be too difficult and too time consuming to be practical. Fortunately, sampling can alleviate the task of reviewing each individual element in a cost proposal.

A **population** of data consists of all observations for a given subject. **Sampling** is the process of selecting units from a population to represent the population. Sampling allows the analyst to examine cost and pricing information without scrutinizing every piece of data within a cost proposal. Used correctly, sampling can expedite and simplify the analysis of a cost proposal without sacrificing accuracy or validity. Sampling consists of the following steps:

**APPLICATION:**

In cost and price analysis, sampling is most often used to analyze material costs. Sampling can also be used to analyze other cost elements such as direct labor. For example, if an analyst must review an estimate of proposed labor hours, the analyst can construct a population of work breakdown structure (WBS) numbers containing direct labor costs. The sampling techniques discussed in this section can then be used to select a sample from the population of WBS numbers.

- 1.) Determining if Sampling is Appropriate,
- 2.) Selecting the Sample, and
- 3.) Analyzing the Sample and Applying Findings.

#### 15.3.1 Determining if Sampling is Appropriate

Determining if sampling is appropriate depends on the size of the population and the end-user of the information. Sampling is not necessary if there is a minimal amount of data to process. The quantity of data may be small enough for complete analysis. Also, the end-user of the information (such as the contracting officer) should be considered prior to taking a sample. The end-user may not want sample information. The end-user may have a policy of checking every individual unit. Using a sample group to evaluate data wastes time if the end-user will not accept the results.

#### 15.3.2 Selecting a Sample

There are many different methods for selecting a sample. Among these methods are simple random sampling, stratified sampling, cluster sampling, systematic random sampling, and convenience sampling. Of these methods, simple random and stratified sampling are the most valuable to the analyst because they provide the most accurate and effective way to process cost and pricing data. These terms are defined in Table 15-1.



Table 15-1. Sampling Terms and Definitions

Term	Definition
Simple Random Sampling	Sampling by selecting units at random from the entire population.
Stratified Sampling	Stratified sampling divides the population units into homogenous groups (strata) and draws a simple random sample from each group.
Cluster Sampling	This approach obtains the sample group by choosing a unit at random and then selecting a group of units adjacent to the first unit that equals the desired sample group size.
Convenience Sampling	Selects the sample group in the easiest possible manner. If 20 units need to be sampled, the analyst would sample the first twenty units in a population.
Systematic Random Sampling	Selects the first unit at random and all additional units are selected at a predetermined interval from the first unit.

### Simple Random Sampling

Simple random sampling is the best method for selecting an accurate and objective sample. Randomness in sampling is a desired quality because it ensures that the sample accurately represents the population. Since each unit has an equal chance of being selected, simple random sampling is the only method that is free of bias and determines statistical confidence in the sample results. When using this method, the analyst should consider two factors: how many items must be sampled and how the random samples will be selected.

Determining the sample size should take into consideration the characteristics of the sample group. The sample group should be large enough to represent the whole population in every detail. There are many methods that can be used to select an accurate sample size. These methods, however, require extensive calculations, and the additional accuracy gained is not great enough, typically, to warrant their use. (Additional methods for selecting a sample size can be obtained from most statistics textbooks.)

A sample size of at least 30 observations can serve as a general rule for most cost and pricing sampling applications. This general rule has been developed through experience because 30 is a large enough sample to notice any pricing patterns and small enough to complete analysis within a short amount of time. The use of 30 observations as a sample size is not applicable in all situations. The total size of the population should be taken into consideration.

Thirty observations from a population of thirty is not a sample, and thirty observations from a population of one thousand may be inadequate. Another consideration is how much time the analyst has to complete the analysis. Larger sample sizes require more time to be spent analyzing the sample. Finally, if the results of analysis are inconclusive or inconsistent, then a larger sample size should be used.

To ensure the objectivity of this method, each item must have an equal opportunity to be selected and must only be selected once for analysis. Each item within the group is, therefore, assigned a sequential number. A table of random numbers or a list of computer generated random numbers is used to identify the items (by number) to be included in the sample group. For example, if the first three random numbers are 22, 64, and 5, the corresponding items, with the numbers 22, 64, and 5, from the sequential list will be analyzed. A table of random numbers can be found in most statistics textbooks. Computer generated random numbers can be obtained using most spreadsheet applications. Regardless of which method the analyst uses, it is good practice to document the method used and to include the list of randomly generated numbers as part of any sampling analysis documentation.

### Stratified Sampling

Stratified sampling is often used to analyze a bill of materials where there are a few high dollar items and several small dollar items. The first step in stratified sampling is the division of the population into homogenous strata. The second step is identifying the 100% review stratum, from which every unit in the group will be analyzed. Typically, the 100% review stratum includes those items that comprise a substantial portion of the total proposed cost (usually 80% or 90%). The analyst should be able to evaluate a significant portion of the proposed cost by analyzing relatively few units. The 80/20 rule is applicable under these circumstances. The 80/20 rule states that approximately 80% of a population's costs can be attributed to approximately 20% of the items.

Choosing the 100% review stratum does not always involve selecting those items that comprise a substantial portion of the proposed cost. Other criteria that can be used to select the 100% review stratum include:

- Costs of an unusual or sensitive nature,
- Costs in areas where the contractor's procedures or internal controls are known to be ineffective or where discrepancies have been discovered in the past,
- Costs not subject to previous reviews, and

- Costs in areas where incorrect estimates can have a substantial impact on overall cost.

Regardless of the criteria used to select the 100% review stratum, the stratum should be clearly distinguishable from the rest of the population.

The third step in stratified sampling is determining how the remaining strata will be sampled. Depending upon the situation, the remaining strata can be sampled using various sampling techniques. Simple random sampling is the method used most often to select the units to be analyzed in the remaining strata.

### 15.3.3 Analyzing the Sample and Applying Findings

Once a sample has been selected, the analyst should analyze the sample group and apply the findings to the remaining population. The process for analyzing sample data is the same as if the entire population was analyzed. Any test or question that can be asked of the population can also be asked of the sample group. The analyst can examine the prices, material quantities, labor hour estimates, or any other quantifiable information within the sample group for reasonableness and allowability. Cost and price analysis techniques for determining the reasonableness and allowability of specific cost elements are discussed in Parts III and IV of this handbook.

After a sample is analyzed, findings from the sample must be applied to the entire population to develop a recommended cost position. In cases where stratified sampling is used, the findings of the 100% review stratum apply only to the reviewed items, not the entire population. For the remaining population, findings gathered from the analysis of a randomly selected (or otherwise) sample are used. The analyst is specifically interested in determining the average difference between the proposed and evaluated price (or quantity) of units within the sample group. A total recommended cost position is developed through combining the results of the two samples. Case Study 15-1 provides an example of how the two sampling methods can be used in analysis of material costs.



**CASE STUDY 15-1. UTILIZING SAMPLING TO EVALUATE MATERIAL PRICES****Background:**

An analyst must analyze proposed material costs of \$10 million. The Consolidated Material List (CML) submitted in the contractor's proposal consists of over 1,000 different part numbers and is, therefore, too large to analyze each item individually.

**Approach:**

The analyst chooses to use stratified random sampling. The analyst selects the 100% review stratum by selecting the high dollar items that comprise 80% of the total material cost (\$8 million). The analyst examines the price for each item within this stratum and develops a recommendation for each item. When summed, the analyst's recommendations for items within the 100% review stratum total \$7.5 million.

For the remaining items within the CML (valued at \$2 million), the analyst chooses to use simple random sampling. After collecting and analyzing the random sample, the analyst determines that prices are overstated by an average of 20%. The analyst applies this finding to this stratum, reducing the proposed value of \$2 million by 20%. Therefore, the analyst's recommended position for the randomly sampled stratum is \$1.6 million.

**Conclusion:**

As a result of the sampling analysis, the analyst recommends a position for material of \$9.1 million (\$7.5 + \$1.6 million).

**15.4 REGRESSION ANALYSIS**

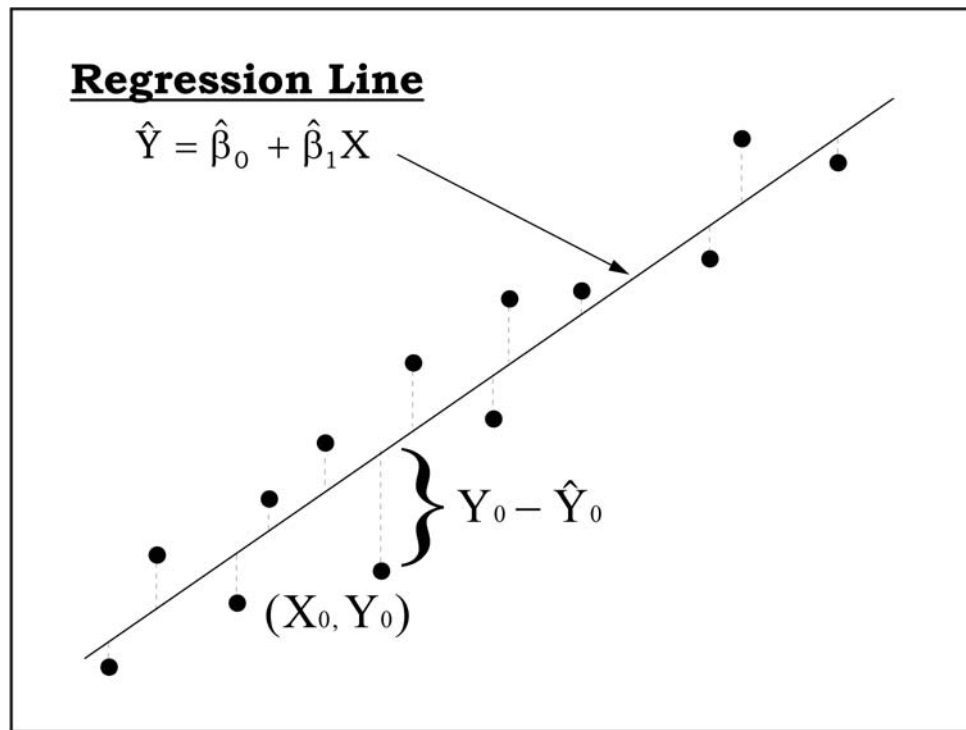
Regression analysis allows the analyst to support hypotheses regarding specific relationships between two variables. Regression analysis is a statistical tool that identifies and quantifies the effect an independent variable has on a dependent variable. The quantification of this effect results in the estimated coefficients of the independent variable(s).



### 15.4.1 Regression Analysis Concepts

The most popular technique for estimating the coefficients is the **least squares method**. To illustrate the least squares method, refer to Figure 15-1. This figure depicts a scatter diagram in which an estimated regression line has been drawn through several plotted data points. The vertical distance (residual) between the observed value ( $X_0, Y_0$ ) and the estimated curve is given by  $Y_0 - \hat{Y}_0$ . 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 the one that minimizes the sum of the  $n$  squared vertical distances. The equation of the line shown below uses  $\hat{\beta}_0$  and  $\hat{\beta}_1$  to represent the slope and Y-intercept. The same equation is often written using  $b$  and  $a$  in place of  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , respectively, to represent these variables, such as  $\hat{Y} = aX + b$

Figure 15-1. Least Squares Regression Line



The simplest equation that describes the least squares regression theory is a two-variable straight line equation or a **bivariate linear regression**. The equation takes two forms: theoretical and estimated. The theoretical equation includes a value ( $e$ ) which represents the residual difference between the theoretical and estimated values of the dependent variable. The lower the residual, the closer the estimated curve is to the actual (theoretical) curve.

Since the true or theoretical equation will never be observed, the following sections will address the estimated equation. The equations are shown below.

1.) Theoretical:  $Y = \beta_0 + \beta_1 X + e$       or      2.) Estimated:  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Where:  $Y$  = the actual (or observed) value of the dependent variable

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

$x$  = the independent variable

$\beta_0$  = the  $y$  intercept

$\beta_1$  = the slope of the line

$\hat{\beta}_0$  = the estimated value of the  $Y$  intercept; the value of  $\hat{Y}$  when  $X = 0$

$\hat{\beta}_1$  = the estimated value of the slope of the line; the change in  $Y$  divided by the change in  $X$

$e$  = the residual term ( $Y - \hat{Y}$ )

Before being overwhelmed with these formulas, it should be stated that most spreadsheet applications will calculate the slope and  $y$  intercept based on the known values of  $x$  and  $y$ . The formulas for calculating the slope (Equation 15-1) and  $y$  intercept (Equation 15-2) are as follows:

**NOTE:**

To simplify notation, the theoretical beta,  $\beta$ , will be used to represent the estimated beta,  $\hat{\beta}$ , from this point forward.

Equation 15-1. Slope

$$\beta_1 = \frac{\sum_{i=1}^n xy - n\bar{x}\bar{y}}{\sum_{i=1}^n x^2 - n\bar{x}^2}$$

Equation 15-2.  $Y$  Intercept

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Where:

$\bar{x}$  = the mean value of  $x$  observations

$\bar{y}$  = the mean value of  $y$  observations

$n$  = the number of observations in the sample

$\beta_0$  = the  $y$  intercept

$\beta_1$  = the slope of the line

Constructing a worksheet similar to Table 15-2 helps to develop an understanding of the mathematics of regression analysis. Given that most popular spreadsheet programs perform regression analysis, manual calculations are not necessary. However, Table 15-2 is followed by Case Study 15-2a that shows how a regression line can be manually calculated. Familiarity with the manual calculations will provide insight when interpreting regression output generated by spreadsheet software.

**Table 15-2. Regression Analysis Worksheet**

$X_i$	$Y_i$	$X_i Y_i$	$X_i^2$	$Y_i^2$
$X_1$	$Y_1$	$X_1 * Y_1$	$X_1 * X_1$	$Y_1 * Y_1$
$X_2$	$Y_2$	$X_2 * Y_2$	$X_2 * X_2$	$Y_2 * Y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_n$	$Y_n$	$X_n * Y_n$	$X_n * X_n$	$Y_n * Y_n$
$\Sigma X_i$	$\Sigma Y_i$	$\Sigma (X_i * Y_i)$	$\Sigma (X_i^2)$	$\Sigma (Y_i^2)$
$\bar{X} = \frac{\Sigma X_i}{n}$ <p>Mean of X Values</p>		$\beta_1 = \frac{\sum_{i=1}^n xy - n\bar{x}\bar{y}}{\sum_{i=1}^n x^2 - n\bar{x}^2}$ <p>Slope of the Line</p>		
$\bar{Y} = \frac{\Sigma Y_i}{n}$ <p>Mean of Y Values</p>		$\beta_0 = \bar{Y} - \beta_1 \bar{X}$ <p>Y intercept</p>		

## CASE STUDY 15-2a. REGRESSION ANALYSIS OF HOURLY EARNINGS

**Background:**

An analyst wants to determine the relationship between years of experience possessed by workers and the hourly wage rate.

**Data:**

The analyst has conducted a sample survey and determined the average labor rate for each year of experience. Using this information to perform a regression analysis, the analyst can determine the relationship between years of experience (the independent variable) and hourly earnings (the dependent variable).

Years of Experience	Hourly Earnings
1	12.98
2	13.94
3	17.07
4	18.27
5	21.63
6	22.84
7	24.04
8	25.84
9	26.08
10	28.61

**Steps to Define the Regression Equation (Line):**

- 1.) Set up the regression worksheet (as shown in Table 15-2).

$X_i$	$Y_i$	$X_i Y_i$	$X_i^2$	$Y_i^2$
1	12.98	12.98	1.00	168.48
2	13.94	27.88	4.00	194.32
3	17.07	51.21	9.00	291.38
4	18.27	73.08	16.00	333.79
5	21.63	108.15	25.00	467.86
6	22.84	137.04	36.00	521.67
7	24.04	168.28	49.00	577.92
8	25.84	206.72	64.00	667.71
9	26.08	234.72	81.00	680.17
10	28.61	286.10	100.00	818.53
55	211.30	1306.16	385.00	4722.37

$$\bar{X} = 5.5$$

$$\bar{Y} = 21.13$$

$$n = 10$$



2.) Obtain the  $\beta_0$  and  $\beta_1$  values using the least squares method.

$$\beta_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{1306.16 - (10)(5.5)(21.13)}{385 - 10(5.5)^2} = \frac{144.01}{82.5} = 1.746$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 21.13 - (1.746)(5.5) = 11.529$$

3.) Develop the formula for the linear regression equation.

$$\hat{Y} = 11.529 + 1.746X$$

4.) Determine the estimated hourly earnings of a worker with 11 years of experience.

$$\hat{Y} = 11.529 + 1.746(11) = 30.735 \text{ or } \$30.74$$

### 15.4.2 Evaluating the Performance of the Regression Equation

When evaluating the performance of a regression equation, an analyst usually considers the following indicators: variance, overall fit, significance, and correlation.

#### Measuring Variation using the Standard Error of Estimate

Although the least squares method produces a line that fits a group of data points with a minimum amount of variation, it is not a perfect predictor. The regression line serves only as an approximate predictor of Y for values of X. The **standard error of estimate (SEE)** is a statistical measure used to determine the variability of the actual Y values from predicted Y values. (**Predicted values** of Y are the values of Y as estimated by the regression equation.) The SEE is measured in units of the dependent variable, Y. As can be expected, a low SEE is generally preferred to a higher SEE.

The formula used to calculate SEE is shown below in Equation 15-3. Application of the formula is described in Case Study 15-2b.

Equation 15-3. SEE

$$\text{SEE} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}}$$

Where:

- $Y_i$  = the actual value of  $Y$  given  $X_i$
- $\hat{Y}_i$  = the predicted value of  $Y$  given  $X_i$
- $n$  = the number of observations

**CASE STUDY 15-2b. CALCULATING STANDARD ERROR OF THE ESTIMATE**

This continuation of Case Study 15-2 uses the results of Part a to calculate the standard error of the estimate (Equation 15-3). The first step is to calculate the predicted values of  $Y$  given  $X$ , remembering that  $Y = 11.529 + 1.746X$ .

YEAR	ACTUAL RATE	PREDICTED RATE	$(Y_i - \hat{Y}_i)^2$
1	12.98	13.28	0.09
2	13.94	15.02	1.17
3	17.07	16.77	0.09
4	18.27	18.51	0.06
5	21.63	20.26	1.88
6	22.84	22.01	0.70
7	24.04	23.75	0.08
8	25.84	25.5	0.12
9	26.08	27.24	1.35
10	28.61	28.99	0.14
		Sum =	5.68

Using the information provided in the table above, the standard error of estimate can be determined.

$$\text{SEE} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2}} = \sqrt{\frac{5.68}{8}} = 0.8426 \text{ or } \$0.84$$

This means any hourly rate generated by the formula will differ by approximately \$0.84 from the actual hourly rate.

**Measuring Overall Fit**

A correctly written equation may not model (or fit) the sample data well. Consequently, the output of the regression will be meaningless. Statistics such as the coefficient of determination ( $R^2$ ) and adjusted  $R^2$  assess the adequacy of a regression equation's "goodness of fit".

*Coefficient of Determination ( $R^2$ )*

The **coefficient of determination** indicates what percentage of the variation in the dependent variable is attributable to the independent variable. The coefficient of determination ranges between zero and one. If all the plotted data points are close to the regression line,  $R^2$  will be close to one.  $R^2$  equals one when all data points fall on the regression line. As the points become more scattered,  $R^2$  will move closer to zero.

In Case Study 15-2c, the coefficient is .978 between years of experience and hourly rate. This means that there is a strong relationship between the two variables, where 97.8% of the variation in hourly rates is attributable to the regression function. .022 ( $1 - R^2$  or  $1 - 0.978$ ), of the variation in the hourly rate can be attributed to factors not included in this regression equation.

**CASE STUDY 15-2c. CALCULATING  $R^2$** 

Utilizing the data from Part b of this case study, the  $R^2$  value can be computed as follows:

YEAR	ACTUAL RATE	PREDICTED RATE	$(Y_i - \hat{Y}_i)^2$	$(Y_i - \bar{Y})^2$
1	12.98	13.28	0.09	66.42
2	13.94	15.02	1.17	51.70
3	17.07	16.77	0.09	16.48
4	18.27	18.51	0.06	8.18
5	21.63	20.26	1.88	0.25
6	22.84	22.01	0.70	2.92
7	24.04	23.75	0.08	8.47
8	25.84	25.5	0.12	22.18
9	26.08	27.24	1.35	24.50
10	28.61	28.99	0.14	55.95
Sum =			5.68	257.06

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - (5.68/257.06) = 0.978$$

Even though  $R^2$  is the most common measure of fit, it has one major weakness. When using more than one independent variable (multivariate regression), the addition of independent variables will **NEVER** decrease the  $R^2$  value. Assume an analyst is building an equation with cost of software programs as the dependent variable and lines of code and program language as the

independent variables. To improve  $R^2$ , the analyst decides to add another independent variable. The analyst could add any variable (even the shoe size of the programmer) and the  $R^2$  will increase!

*Adjusted Coefficient of Determination (  $\bar{R}^2$  )*

$\bar{R}^2$  adjusts  $R^2$  for the number of independent variables included in a regression equation.  $\bar{R}^2$  will either decrease, increase, or stay the same with the addition of an independent variable, depending on whether the improvement of fit outweighs the loss of one **degree of freedom**. Degrees of freedom are the number of observations ( $n$ ) minus the number of independent variables ( $k$ ) minus one. If the loss of the degrees of freedom is greater than the improvement of fit, the  $\bar{R}^2$  value will decrease.

**CASE STUDY 15-2d. CALCULATING**

Utilizing the data from Part c, the  $\bar{R}^2$  value can be computed as follows:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / (n - k - 1)}{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (n - 1)} = 1 - \left( \frac{5.68 / 8}{257.06 / 9} \right) = 0.975$$

Where:

$k$  = the number of independent variables  
 $n$  = the number of observations

Measuring Significance Using the t-test

As the previous section implied, not all independent variables will have a significant impact on the dependent variable. Insignificant variables should not be included in the regression equation. The t-test is a method of determining significance. There are three elements to the t-test: t-statistic, critical t-value (from statistics tables), and the decision rule. The steps for conducting a t-test are below.

**Step 1. Calculate the t-statistic.** Most spreadsheet applications calculate the t-statistic for the independent variable to be tested. The formula for a t-statistic for the  $k$ th independent variable is shown in Equation 15-4.



**Equation 15-4. T-Statistic**

$$t_k = \frac{(\hat{\beta}_k - \beta_{HO})}{SE(\hat{\beta}_k)}$$

$$SE(\hat{\beta}_k) = \frac{SEE}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Where:

- $t_k$  = *t*-statistic for the  $k^{th}$  independent variable  
 $\beta_{HO}$  = the border value (usually zero) of the null hypothesis from the decision rule  
 $SE(\hat{\beta}_k)$  = the standard error of the  $k^{th}$  coefficient  
 $SEE$  = the standard error of the estimate (described earlier)

**Step 2. Develop the hypothesis test.** The goal of the hypothesis test is to reject the null hypothesis ( $H_0$ ). By doing so, the analyst will prove that the alternative hypothesis ( $H_A$ ) is true. Therefore, the alternative hypothesis states what the analyst assumes to be true. Two hypothesis tests are possible: one-sided and two-sided. One-sided tests indicate the sign of a coefficient. Two-sided tests indicate whether the estimated coefficient is significantly different from zero or some other value.

**Step 3. Find the critical value.** This can be one of two numbers, depending upon whether a one-sided or two-sided test is being conducted. To locate the critical t-value:

- Choose a level of significance. The most common level of significance is 95%. A significance level of 95% means that 95% of the variation of the dependent variable is explained by the regression equation.
- Determine the degrees of freedom ( $n-k-1$ ). If the regression has one independent variable ( $k=1$ ) and 10 observations ( $n=10$ ), there are 8 degrees of freedom.
- Find the corresponding critical t-value. Table 15-3 is an excerpt from a table of critical t-values. The critical t-value is 1.860 for a one-sided test with a 95% level of significance and 8 degrees of freedom.

Table 15-3. Critical T-Values

Degrees of Freedom	One-sided Two-sided	10% 20%	5% 10%	2.5% 5%	1% 2%	0.5% 1%
1		3.078	6.314	12.706	31.821	63.657
.	.	.	.	.	.	.
.	.	.	.	.	.	.
8		1.397	1.860	2.306	2.896	3.355

**Step 4. Test the hypothesis.** The decision rule is: reject the null hypothesis ( $H_0$ ) if the t-value of  $k^{\text{th}}$  variable is greater than the critical t-value and has the sign implied by the  $H_A$ . **If the null hypothesis can be rejected, then the  $k^{\text{th}}$  variable is significant.**

### Measuring the Correlation

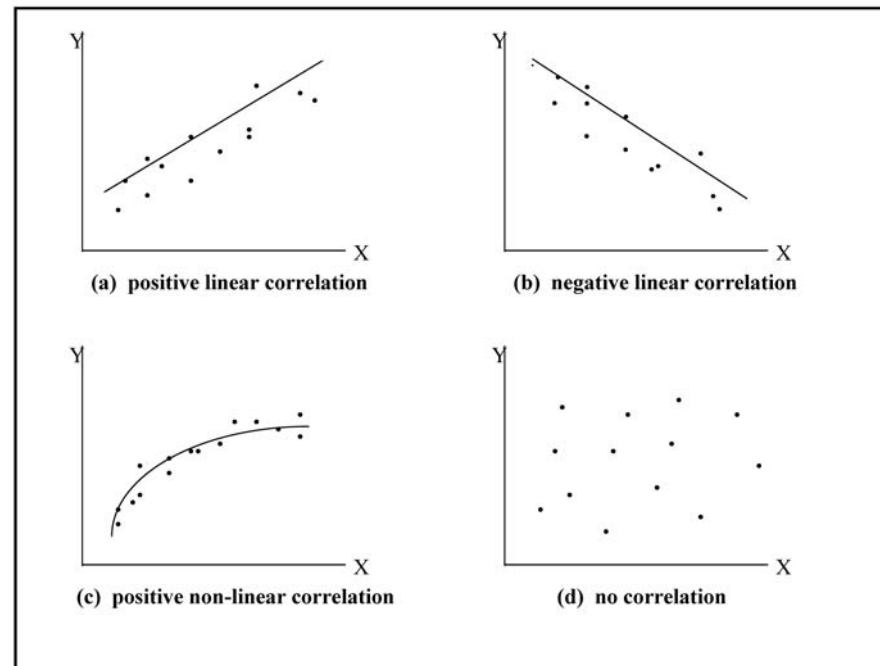
Correlation indicates the impact a change in the independent variable will have on the dependent variable. The measure of this correlation is the **coefficient of correlation**, or the **r-value**. Two variables can be positively or negatively correlated or have no correlation at all. The r-value ranges from (-1) to 1. If two variables are truly independent of each other, the coefficient of correlation would equal zero. If the variables are perfectly positively correlated ( $r = 1$ ), any change in the independent variable results in an equal change in the dependent variable. If the variables are perfectly negatively correlated ( $r = -1$ ), any change in the independent variable results in an equal and opposite change in the dependent variable.

The equation for calculating the coefficient of determination (Equation 15-5) is:

**Equation 15-5. Coefficient of Determination**

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

Where:  $X$  and  $Y$  are the variables for which correlation is being determined

**Figure 15-2. Correlation**

In regression analysis, correlation can be described by the function's slope. If the slope is positive, the correlation is positive. If the slope is negative, the correlation is negative. The value of the correlation coefficient, however, is not the same as the slope. Figure 15-2 shows possible correlation scenarios. When doing multivariate regressions, correlation between independent variables is undesirable and must be avoided. Otherwise, the results of the regression analysis will be inaccurate. Case Study 15-2e explains the calculation of  $R^2$ .

CASE STUDY 15-2e. CALCULATING  $r$ 

The data analyzed in Part a showed that for the simple linear regression fit of the data, the slope is 1.746. Therefore, the hourly wage rate is positively correlated to the number of years of experience. The two formulas above can be used to double check the previous work.

YEAR ( $X_i$ )	ACTUAL RATE ( $Y_i$ )	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
1	12.98	-4.5	-8.15	36.68
2	13.94	-3.5	-7.19	25.17
3	17.07	-2.5	-4.06	10.15
4	18.27	-1.5	-2.86	4.29
5	21.63	-0.5	0.50	-0.25
6	22.84	0.5	1.71	0.86
7	24.04	1.5	2.91	4.37
8	25.84	2.5	4.71	11.78
9	26.08	3.5	4.95	17.33
10	28.61	4.5	7.48	33.66
			Sum =	144.04

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{144.04}{\sqrt{82.5} \sqrt{257}} = 0.989$$

## 15.4.3 Advanced Topics in Regression Analysis

Multivariate Regression Analysis

Simple regression analysis uses 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. However, to explain some relationships more than one independent variable is required. For example, the manufacturing supervisor's hours may depend on both assembly hours and quality assurance hours. This type of regression analysis is referred to as multivariate regression analysis.

The functional relationship between the independent variables ( $X_i$ ) and the dependent variable ( $Y$ ) may have the following linear form if there are "p" independent variables:  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$

Where:  $\beta_0$  = a constant

$\beta_i$  = the coefficients of the independent variables (for  $i = 1, \dots, p$ ). The  $\beta_i$  can represent the relative importance or weight of each of the independent variables, provided the  $X_i$  are commensurable.



The concepts and computations involved in multivariate regression are more difficult than those for simple regression and, therefore, should be performed using current computer software packages. Statistics textbooks should be referenced for detailed discussions on multivariate regression.

### Simple Non-Linear Relationships

Not all relationships are linear (i.e., the relationship can be graphically represented by a straight line). Applying appropriate variable transformations, some non-linear relationships can be converted into equivalent linear relationships. In so doing, the curve fitting techniques discussed in sections 15.4.1 and 15.4.2 can be applied to the non-linear relationships listed in Table 15-4. For example, if the scatter diagram suggests that an exponential relationship might exist, the analyst 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**, and **b** represents logarithm **b** in the exponential curve.

**Table 15-4. 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 Of Slope A=
Hyperbolic	$Y = \frac{1}{aX + b}$	$1/Y = aX + b$	None	$1/Y$	<b>b</b>	<b>a</b>
Exponential	$Y = ba^X$	$\log Y = \log b + X \log a$	None	$\log Y$	$\log b$	$\log a$
Geometric	$Y = bX^a$	$\log Y = \log b + a \log X$	$\log X$	$\log Y$	$\log b$	<b>a</b>

### Utilizing Computer Applications to Perform Regression Analysis

Most computer spreadsheet and statistical packages perform simple and multiple regression, and many of them provide useful information on significance test computations and interpretations. Thus, the analyst should investigate how to access and use a statistical package rather than perform the calculations by hand. In addition, many calculators have special functions to perform simple regression. Figure 15-3 depicts regression output from Microsoft Excel. The statistics mentioned in this section are outlined with bold lines. The statistics were generated using the earlier example of a regression with hourly wages as the dependent variable and year as the independent variable.

Figure 15-3. Spreadsheet Output

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.989							
R Square	0.978							
Adjusted R	0.975							
Std. Error	0.843							
Observ.	10							
ANOVA								
	df	SS	MS	F	Signif. F			
Regression	1.00	251.380	251.380	354.017	0.000			
Residual	8.00	5.681	0.710					
Total	9.00	257.061						
	Coeff.	Std. Error	t Stat	P-value	Lower	Upper	Lower	Upper
Intercept	11.529	0.576	20.028	0.00	10.202	12.857	10.202	12.857
X Variable	1.746	0.093	18.815	0.00	1.532	1.960	1.532	1.960

#### 15.4.4 Uses of Regression Analysis

There are many uses of regression analysis. Two uses, forecasting and investigating a cost estimating relationship (CER), are examined in the next two sections.

##### Forecasting

Occasionally, the analyst needs to evaluate a cost proposal that spans the life of a multi-year contract. The contractor will have adjusted proposed costs for inflation and other economic impacts anticipated during the period of performance. To conduct a complete evaluation, the analyst needs to estimate how the costs should be escalated over the period of performance and compare this estimate to what the contractor has proposed. The process of predicting the impact of business and economic conditions on contract costs is **forecasting**. The analyst may develop forecasts or obtain them through outside sources.

##### *Developing a Forecast Model*

Forecasts of changes in economic conditions and how these changes alter price conditions are necessary for good contract evaluation. Forecasts of the costs of materials and labor rates are crucial to completely evaluate and negotiate long-term contracts. A complete forecast takes into account all known information about historical trends and any economic predictions that are available and relevant. Subsequently, a forecast model processes known information and predicts future costs. Two broad classes of models exist: econometric and time series analysis models. Since econometric models are quite complex and rarely used in cost and price analysis, time series analysis models will be discussed in this section. The following paragraphs cover two time series models: trend analysis and the Autoregressive Integrated Moving Average (ARIMA) model.

### *Trend Analysis*

One of the most convenient methods of developing a forecast model is trend analysis. Trend analysis considers past data and generates a least squares regression line to predict future index numbers. The steps are given below:

- Collect data, e.g., the Producer Price Index (PPI).

YEAR	INDEX NUMBER
1988	100
1989	106
1990	111
1991	118
1992	121
1993	127

- Use the formula from section 15.4.1 to develop a regression line for the index numbers.
- Plot out the regression line on a graph for all possible points.
- Use the line to estimate future index numbers.

It is important to note that the accuracy of the forecast depends upon the value of past data for predicting the future. Trend analysis is not advisable for long-term forecasts (3-5 years or more), since most series do not follow a trend such as this for a long period of time. Case Study 15-3 shows the creation of a trend analysis forecast.

## CASE STUDY 15-3. TREND ANALYSIS

Given the PPI Index numbers from above, a model can be constructed to forecast index numbers for the following three years.

$X_i$	$Y_i$	$X_i Y_i$	$X_i^2$
1	100	100	1
2	106	212	4
3	111	333	9
4	118	472	16
5	121	605	25
6	127	762	36
Sum =		2484	91

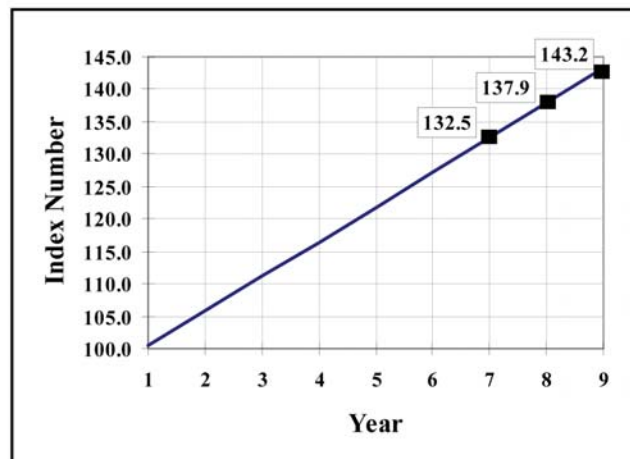
Means of Variables:

$$\bar{X} = 3.5 \quad \bar{Y} = 113.83 \quad n = 6$$

$$\beta_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{2484 - 2,390.43}{91 - 73.5} = \frac{93.57}{17.5} = 5.347$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} = 113.83 - 5.347(3.5) = 95.116$$

Using the data above, the analyst can develop a regression line with the equation:  $Y = 95.116 + 5.347X$ . This regression formula can calculate estimates of index numbers for years 7 through 9.



If all historical trends remain unchanged over the next three years, the index numbers for years 7 through 9 should equal 132.5, 137.9, and 143.2, respectively.



### *The ARIMA model*

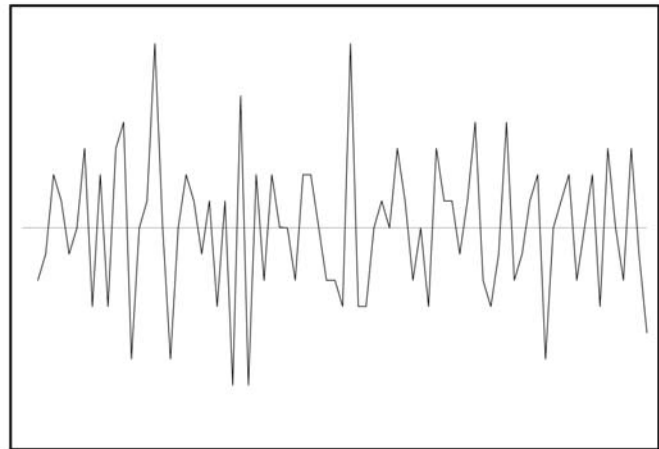
A more complex model to forecast annual time series data is the **Autoregressive Integrated Moving Average (ARIMA) model**. The ARIMA model is a good model for long-term predictions. The model uses the theory that future trends correlate highly to the trends immediately preceding them. The weight placed on data decreases as the distance between the time period of the observed trend and the time period of the forecast increases. The level of ARIMA's sophistication dictates that computer software, such as the SAS<sup>®</sup> System, be used to generate a forecast.

The ARIMA model can predict and account for economic trends. Trends can affect the economy in dramatic ways that are not represented well by a straight line model. The ARIMA model takes into account trends and smoothes out data so that forecasts can be provided with greater accuracy.

ARIMA models can predict with a high-degree of accuracy either short-term or long-term economic trends, but not both. This occurs because the correlation and error factors used to smooth out trends are based on the length of the predicted time span.

ARIMA analysis begins by transforming the data series (Y) to ensure that it is stationary (the mean and the amount of fluctuation around the mean are constant). Figure 15-4 graphically depicts a stationary data series. Economic time series data are trending (i.e., nonstationary). George Box and Gwilym Jenkins, who developed ARIMA, stated

**Figure 15-4. Stationary Data Series**



that economic time series data can be made stationary by differencing. (Differencing is simply creating a new data series by subtracting the  $n-1$  observation or data point from the  $n$ th data point). Differencing creates a new data series ( $Y^*$ ), which becomes the input for the Box-Jenkins (ARIMA) analysis. Usually only one or two differencing operations are required.

Equation 15-6 shows the general formula for ARIMA forecasting.

Equation 15-6. ARIMA

$$Y_t^* = \sum_{t=1}^p \phi_p Y_{t-p}^* + \sum_{t=1}^q \theta_q \varepsilon_{t-q}$$

Where:

- $Y_t^*$  = the forecasted value of  $Y$
- $p$  = the number of historical units of  $Y$
- $\phi$  = the correlation of  $Y$
- $\theta$  = the correlation of the error component
- $\varepsilon$  = the error component
- $q$  = the number of error terms

$Y_t^* = Y_t - Y_{t-1}$  which is the forecast value for the differenced series

- **Stage 1: Identification and model selection.** There are a variety of models for ARIMA. Each model weighs the correlation and error components differently. For example, the analyst may feel that trends of the last two time periods will continue into the future. The analyst should select a model where the trends of the last two time periods heavily influence the future. **Autocorrelation** plots of data observations are useful guidelines or tools for selecting models.
- **Stage 2: Estimation of Parameters.** Using the least squares method discussed earlier in this section, the computer software being used will determine the parameters of the model.
- **Stage 3: Diagnostic Checks.** The model must be tested to determine its accuracy. The analyst may need to spend time determining which model provides the best fit. Stages 1 and 2 may need to be repeated until an accurate model is found.

**Autocorrelation** occurs when observations are correlated with those of earlier time periods.

The statistics and measures of performance discussed in sections 15.4.1 and 15.4.2 can be used as diagnostics here. In addition, the **Chi Square statistic** can be used as a measure of the model's adequacy. The computer software being used usually generates the Chi Square statistic for a model. A model is adequate if the Chi Square statistic for the given model is less than the critical value. Stated differently, a model is adequate if the probability value generated by the software used is .05 or greater.

The mathematics and concepts involved in refining an ARIMA model are complex. A more complete explanation of the ARIMA model can be found in most forecasting textbooks.

### *Outside Sources of Forecasts*

Unfortunately, the analyst does not always have the time or resources needed to produce a complete forecast. There are many sources of forecast information available for an analyst's use. When selecting an outside source of a forecast, an analyst must consider who the source is, where the data were collected, how the data are reported, and if the forecast is applicable to the needs of the analyst. Four excellent sources of forecasts are:

- **Bureau of Economic Analysis Publications.** The Bureau of Economic Analysis, a part of the Department of Commerce, publishes the *Survey of Current Business*. The Survey publishes indices on a range of industries and reports information on trends and economic forecasts for each industry.
- **Council of Economic Advisors (CEA) and Office of Management and Budget (OMB).** The CEA publishes, as part of the *Economic Report of the President*, an annual forecast of Gross Domestic Product and other economic indicators in February of each year. The OMB revises this forecast and publishes it as the *Mid-session Review* in August of each year.
- **The Federal Reserve.** The Federal Reserve publishes the *Federal Reserve Bulletin*. The bulletin has economic indexes and collects and analyzes data on business, commodities, construction, labor, manufacturing, and trade. Each of the Federal Reserve District banks publishes information on its district's economic indicators and the outlook for that region of the country.
- **IHS Global Insight.** Global Insight provides economic forecasts, data, and analysis for countries and industries.

Other federal agencies and state governments also publish forecasts according to their mission. Private firms, private associations, industry and trade publications, and newspapers or business magazines also provide forecast data, usually at a fairly high price. An analyst, using any outside source of forecasts, should ensure that the forecast is relevant to the current situation and was produced by a credible source. Also, the analyst should reference the outside source of the forecast in any report generated which includes the forecast.

### Cost Estimating Relationships

A cost estimating relationship (CER) uses a mathematical expression relating cost as the dependent variable to one or more independent cost driving variables. Statistical techniques, using multiple historical data points, are the preferred way to develop CERs.

A CER predicts the cost of some part of a program based on specific design or program characteristics. When using a CER, the cost is unknown, but some information is known about the size, shape, or performance of the item to be costed, or the dollar size of other cost elements. The analyst is able to estimate the unknown cost based on the known information.

#### *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, and 3.) the complexity of the estimating relationship. Generally, CERs follow a cost-to-cost or parametric (cost-to-noncost) relationship.

**Cost-to-Cost Relationships** use one cost element to predict the cost of another element (e.g., using total production costs to determine the cost of quality assurance). Cost-to-cost CERs are often used to estimate portions of Operations & Support (O&S) costs and non-hardware acquisition costs.

**Parametric (Cost-to-Noncost) Relationships** use a specific characteristic (other than cost) to predict the cost of another element (e.g., using the weight of an item to estimate manufacturing costs). Parametric relationships are classified by the type of cost driver, or system attributes, such as physical, technical, and performance characteristics. Table 15-5 provides examples of parametric cost drivers.

**Table 15-5. Sample Parametric Cost Drivers**

Dependent Variable	Independent Variable
Automobile Price	Horsepower, Mileage, Fuel Efficiency, Make or Model, Year, Vehicle Condition
Cost of a House	Square Footage, Construction Materials, Total Surface Area
Clothing Cost	Fabric Quantity, Fabric Quality
Manufacturing Equipment Cost	Weight, Horsepower, Processing Speed, Processing Quality
Aircraft	Speed, Power, Wingspan, Load Capacity, Empty Weight
Computer System (Development Cost)	Lines of Code, Number of Users, Processing Speed, Number of System Functions
Shipping Costs	Weight, Distance, Speed



### *Uses of CERs*

CERs are used to estimate costs any time during the acquisition cycle when little is known about the cost to be estimated. As more cost information becomes available, more detailed 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. Even in the early stages of the acquisition process, there is a need to know how much a system will cost.

In the source selection process, CERs serve as checks for reasonableness on bids proposed by contractors, and contractors will often use CERs to 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. This may be especially important when trying to determine future costs of alternative design, performance, logistic, or support choices that must be made early in the development process.

### *Developing CERs*

When constructing the equation, the analyst uses the independent variables (X) about which information is known to predict the value of the dependent variable (Y), which is unknown. The objective in developing a CER is to determine the relationship, if any, between X and Y (e.g., lines of code and software cost). If such a relationship is found, it can be used to predict the cost of a software program when the analyst has information on the lines of code required. A functional relationship between X and Y can be constructed through regression analysis.

There are six steps to developing a CER. To make an estimate using CERs or to assess CERs developed by others, the analyst must have an understanding of these six steps.

**Step 1: Target Cost Drivers.** When targeting the type of cost driver to use, the analyst must decide whether to use a cost-to-cost or a parametric (cost-to-noncost) relationship. If a cost-to-cost relationship is used, the analyst must determine what cost element can predict the cost of another element. If a parametric relationship is used, the analyst must determine the type of cost driver, or system attribute, such as physical, technical, and performance characteristics. **Physical characteristics** include volume, length, number of parts, and density. **Technical parameters** (factors that *produce* performance) include system or subsystem power requirements and engine thrust.

**Performance characteristics** include speed, range, accuracy, and reliability. CERs also need to be classified in terms of the aggregate level of the estimate. CERs can be developed for the whole system, major subsystems, other major non-hardware elements (training, data, etc.) and components. The aggregate level of the cost drivers should match the aggregate level of the costs to be estimated, as shown in Figure 15-5. For instance, system costs may be estimated as a function of total system weight, while a subsystem will be estimated by that subsystem's weight.

Figure 15-5. Matching Aggregation Levels of CERs

COST		COST DRIVERS	
	SYSTEM	←	SYSTEM LEVEL CHARACTERISTICS
	SUBSYSTEM	←	SUBSYSTEM LEVEL CHARACTERISTICS
	COMPONENT	←	COMPONENT LEVEL CHARACTERISTICS

**Step 2: Hypothesize 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 assumptions made before reviewing the data (a priori). 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 aircraft costs. The analyst must review what factors might cause costs to increase and measure them directly or indirectly. The weight relationship is an example of a direct measure. 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 analyst may resort to an indirect measure, such as time. Once the analyst has a list of hypothetical relationships, the analyst 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 analyst to judge the reasonableness of the estimating relationship based on intuition.

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 earlier in Figure 15-2 (a and b) suggests a linear relationship. Figure

15-2 (c) suggests a non-linear relationship and Figure 15-2 (d) suggests that X and Y are not related at all.

In practice, it is best to employ both approaches. After hypothesizing one or more functional relationships between the independent and dependent variables, the analyst should plot the data on a scatter diagram. If the scatter diagram does not confirm the hypothesized relationship, the analyst should rethink 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, which are only gained through repeated use of CERs.

**Step 3: Collect and Normalize Data.** The strength of a CER depends largely on the availability, timeliness, and accuracy of data. The analyst should collect data from all available, credible sources and should normalize it to adjust for extraneous factors which could influence the validity of the data.

Sources of data include cost studies, agency cost libraries, contractor cost databases, current contract information, contractor cost proposals, outside organizations/agencies, program personnel, and technical interface organizations. These sources should be exhausted when collecting data to use in developing CERs.

Many factors influence the validity of cost data. Normalization is the adjustment of actual cost data to enable its application on a uniform basis. Generally, data must be normalized as a result of economic changes, technological changes, or differences in work content or cost accounting structures.

- **Economic changes.** Inflation and other economic variables influence cost data; and, therefore, data needs to be normalized to reflect a common economic period or base year. Index analysis serves to eliminate or minimize the economic impacts of inflation.

**Index analysis** involves the use of index numbers to determine inflation/deflation rates. Index numbers indicate the percentage change in a price relative to a base year. To perform index analysis, collect data on the base price, select an appropriate index, and use the index to determine the rate of inflation/deflation. For detailed information on selecting and using indexes see Chapter 6, "Gathering and Evaluating Data for Price Analysis".

- **Technological changes.** Changes in technology often render historical data obsolete, unless adjustments are made to account for past inefficiencies or recent improvements in development or manufacturing processes.

- **Differences in work content or cost accounting structure.** Historical data for individual cost elements derived from different programs often do not represent an identical work effort by the contractor. Also, data for individual cost elements vary from contractor to contractor as a result of different cost accounting practices and structures.

**Step 4: Utilize Curve Fitting Techniques.** There are two methods that the analyst can use to fit a curve to a set of data. 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; no calculations are required; and consideration can be given to outliers. The principle disadvantage of this approach is that the location and shape of the curve through the data points are based upon subjective judgment.

The second approach is the least squares method, discussed in section 15.4.1. This method has the weakness that all data points are given equal weight. The analyst cannot give less weight to outliers except by excluding them from the sample. However, the advantages are significant. The 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).

**Step 5: Determine Goodness of Fit and Confidence Regions.** In cost estimating, the typical situation involves a CER that is 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. The analyst is generally better off using a second estimating method to support an estimate, rather than attempting to prove statistically that the cost estimate has a high probability of lying within narrow bounds.

Should it become necessary to validate a CER using statistical techniques, several methods using a regression line exist for estimating confidence regions around a predicted dependent variable value. Techniques available to the analyst will vary depending on the amount of data available and the data distribution assumptions. Statistics textbooks or advanced cost estimating handbooks should be consulted for more information on this topic.

**Step 6: Understand the Applicability and Limitations of the CER.** Like all estimating techniques, CERs have their limitations. The analyst must be fully aware of these limitations to properly convey the degree of confidence one should have in the cost estimate.



- **Quality and Size of the Database.** Credible CERs demand quality data and enough data to estimate the relationship. To meet the requirement of quality data, the use of actuals (actual historical costs, actual weight, speed etc.) is often required. When the analyst does not work with actuals, care must be given to estimating and interpreting the CER. Of course, actuals are not always available, forcing the analyst to rely on cost data from contractor bids and/or other projections.

The size of the database also limits CER credibility. The more data points the analyst has, the more confidence the analyst can have in the CER and its predictions.

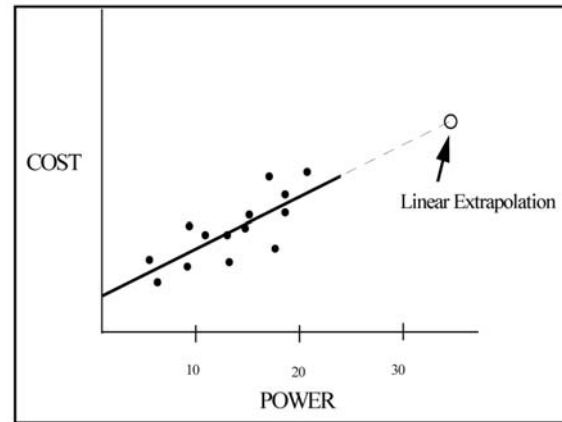
- **Past Costs as Predictors of Future Costs.** The analyst may assume that all factors affecting historical costs (productivity, material type etc.) will affect future costs in approximately the same way. This assumption may be unrealistic for two reasons. First, historical relationships between costs and cost drivers change as technology changes. Second, there has most likely been some learning from past management successes and failures. Managers strive to ensure that the new program will not repeat past management and technical problems.

The analyst must consider whether technological changes (including changes in manufacturing technology) may invalidate a CER. Likewise, the analyst must review how management practices and acquisition strategy are likely to alter historical cost to cost driver relationships. Additionally, studies show that competition during the production phase reduces unit costs. If the program is to be dual-sourced, the analyst may need to consider the effects of competition.

- **Cause and Effect versus Correlation.** A high correlation between the dependent and independent variables does not imply a cause and effect relationship. The analyst must provide that interpretation. When doing so, the analyst must think through what imputing a cause and effect relationship between the two variables means. The analyst should ask this question of all potential cost drivers: How do I expect this cost driver to affect cost? One might find a relationship between cost and sunspots, but what analyst expects the occurrence of sunspots to drive cost?
- **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 may not be accurate. The relationship between the cost drivers (independent variables) and the cost to be predicted (dependent variable) may be constant within a specified region, but may change at extreme values of the independent variable. For example, in Figure 15-6, cost estimates

**Figure 15-6. Extrapolating Beyond the Range of the Data Applicability**



of power requirements between 5 and 20 KVA can be developed with some confidence. A cost estimate for an avionics suite with power requirements of 35 KVA is more uncertain. Can the analyst be sure that the linear cost/power relationship that held for lower power requirements will continue at much higher power levels? 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. Clearly, the analyst should carefully consider whether such an extrapolation is feasible. Some inputs from knowledgeable engineers can provide valuable guidance on whether to extrapolate the CER.

- **Utilizing Tests of Reasonableness.** When using any kind of estimating relationship, the analyst should 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.

In addition to statistical evaluation, the analyst can do the following things to ensure a quality estimate and a reliable CER.

- Make a “test” estimate for a recent system, not included in the database, and check to see if the CER’s “test” estimate agrees 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 accurate 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 range of those in the database (i.e., avoid the need for data extrapolation).

Finally, simple regression analysis cannot solve all problems. Sometimes more advanced statistical techniques (e.g., multiple regression, multivariate techniques) or some non-statistical techniques (e.g., expert judgment, elicitation techniques) need to be applied.

### 15.5 LEARNING (IMPROVEMENT) CURVES

Improvement curve theory is used to estimate recurring resource requirements in operations performed repetitively. The theory is based upon the idea that as a task is performed repeatedly, the time required to complete the task will decrease. Improvement curve theory considers worker improvement, increased efficiency, and other factors that change as workers gain experience. The factors listed below should be analyzed when determining why improvements have occurred and should be considered when developing an improvement curve.

- **Improved tooling and tool use.** When a task is initiated or the production of a good starts, the workers may not be familiar with the tools or the use of the tools in the manner required. As the workers gain experience, their speed with the tools increases. Also, this factor considers that tools not available or considered during the planning of the task may be developed or introduced after production has begun.
- **Increased task familiarity among workers.** As workers become more experienced in a task, they become more efficient, and the number of simple mistakes decreases.
- **Improved production procedures.** Workers who are motivated to increase production will seek and find ways to become more efficient, thus, reducing the time to complete each task.
- **Improved productivity.** Workers and management constantly consider better ways to produce the product or perform the task without degradation of the final product.
- **Improved work flow organization.** As workers become more

familiar and experienced, they discover and eliminate tasks that are needless to production.

- **Improved engineering support.** Before production begins, engineers establish a process for line workers to follow. If a new process is developed, engineers may not know how efficiently the process will flow. Once there are actual line workers involved, engineers can identify production problems and work to solve them.

### 15.5.1 Uses of Learning (Improvement) Curves

Improvement curves are often used for pricing material and estimating labor hours. Improvement theory, however, cannot be used as an estimating tool in every situation. For an improvement curve to be used, the following factors must exist:

- **Production of complex items.** The greater the number of tasks involved in a process, the greater the chance for improvement over time.
- **No significant technological initiatives.** New technology can radically and unpredictably change the production time of an item.
- **Continuous pressure to improve.** A noticeable improvement curve does not occur naturally. Management must actively pursue improvement in the production cycle for dramatic, **measurable improvements to occur.**
- **A high degree of manual labor.** Automated labor performs at a steady pace regardless of experience. As manual labor gains experience, workers find easier ways to perform tasks, which increases efficiency and accelerates production.
- **Uninterrupted production.** Production must be continuous and not subject to any breaks, such as a labor strike or a stop work order. Work stoppages force workers to relearn tasks, negating a large degree of prior improvement.

### 15.5.2 Developing and Analyzing Improvement Curves

Improvement curve theory applies to “total production costs”. This represents only the 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 such nonrecurring costs as basic and rate tooling, which must be added in most cases to get a true total production cost.



Learning curves are usually based on labor hours or dollars-per-unit. When developing an improvement curve, it is usually preferable to use labor hours rather than dollars-per-unit. Dollars are subject to the effects of inflation or deflation, and results based on a dollars-per-unit analysis may be skewed due to economic fluctuations. If the analyst uses dollars, data must be normalized to offset economic impacts.

Learning curves are referred to by many names, including cost improvement curves, progress curves, cost/quantity relationships, and experience curves. Specific types (i.e., mathematical models) of cost improvement curves have often been named after the people who proposed them or the companies that first used them. They include the Wright, Crawford, Boeing, and Northrop curves. These names refer to one of two mathematical models generally agreed to best describe how costs or labor hours decrease as the quantity of an item being produced increases. The two models are most accurately described as the unit improvement curve and the cumulative average improvement curve.

Improvement curve theory is a useful estimating tool. However, it is based on observations, most of which do not exactly fit either the unit or cumulative average curve equations. It is prudent to review actual data to determine if actual improvement is in line with estimated projections.

**NOTE:**

When analyzing a curve created by a contractor, the analyst should question the basis for determining the rate of learning. In most cases, the contractor can provide the historical data necessary for the analyst to verify the rate of learning being used. The analyst should make sure that the contractor did not understate the rate of learning. Increases in the rate of learning may significantly decrease price.

### Unit Improvement Theory

The basis of the unit improvement theory is that as the total volume of units produced doubles, the cost per unit decreases by some constant percentage. This constant percentage is the **rate of learning**. The rate of learning is used to calculate the slope of the unit improvement curve and is usually based on historical data. The unit improvement curve, commonly referred to as the Crawford or Boeing improvement curve, is expressed mathematically in the following equation:

$$Y_x = T_1 * X^b$$

Where:

- $Y_x$  = the cost required to produce the  $X$ th 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

The observation that costs decrease by a constant percentage every time the quantity doubles is reflected in the improvement curve through the **b** value, which is computed using the following equation:

$$b = \frac{\log S}{\log 2}$$

Where:  $S$  = The cost/quantity slope expressed as a decimal value. The slope is calculated by subtracting the rate of learning from 100%.

The  $X^b$  term can also be obtained from improvement curve tables. These tables, published in improvement curve textbooks and in publications such as the DCAA Contract Audit Manual, typically provide unit and cumulative values for the  $X^b$  term for a wide range of curve slopes.

When plotted on log-log paper, the unit improvement curve plots as a straight line. A straight line on log-log paper indicates that the rate of change between two variables is constant. When plotted on standard graph paper with rectangular coordinates, the unit improvement curve plots as a hyperbolic, curved line. Case Study 15-4 shows how to calculate unit cost using the unit curve.

## CASE STUDY 15-4. CALCULATING UNIT COST USING THE UNIT CURVE

**Background:**

Information provided by the contractor shows that the first unit in the run requires 10,000 labor hours to produce and there is a 20% rate of learning. An analyst would like to determine the amount of labor hours necessary to produce the 24th unit in a production run using an unit improvement curve.

**Approach:**

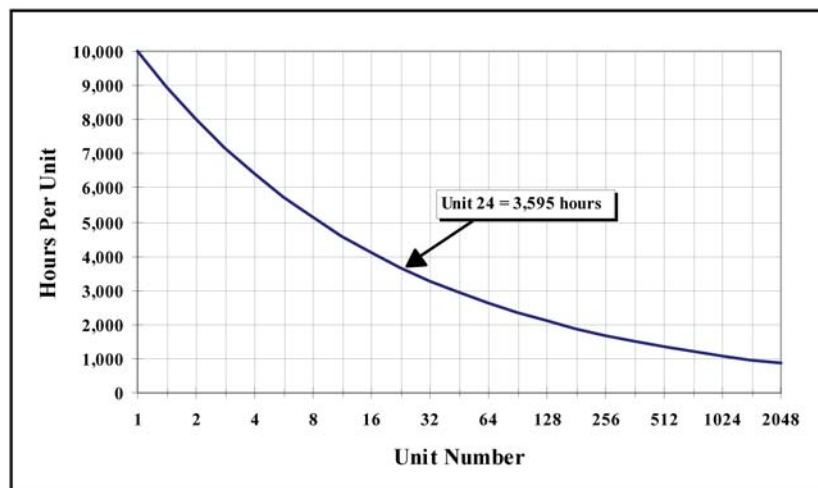
Using the equation for the unit improvement curve:  $Y_x = T_1 * X^b$

$$\text{Where: } T_1 = 10,000 \quad S = (1 - 0.20) = 0.80 \quad b = \frac{\log S}{\log 2} = \frac{\log 0.8}{\log 2} = -0.3219$$

The analyst computes the unit improvement curve to be:

$$Y_x = 10,000 * X^{-0.3219}$$

This equation can be graphed by the analyst to determine the cost of the 24th unit. The graph below shows how a graph of this curve would look.



Or, the analyst may wish to calculate the value of the 24th unit through substitution:

$$Y_x = 10,000 * X^{-0.3219}$$

$$Y_x = 10,000 * 24^{-0.3219}$$

$$Y_x = 3,595 \text{ hours}$$

Cumulative Average Improvement Theory

Cumulative average improvement theory states that as the total volume of units produced doubles, the average cost per unit decreases by some constant percentage. Similar to unit improvement theory, this constant percentage

is the rate of learning and is used to calculate the slope of the cumulative average improvement curve. The cumulative average improvement curve, also known as the Wright or Northrop curve, is expressed through the following equation:

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

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.  
This constant is calculated in the same manner as in the unit improvement equation.

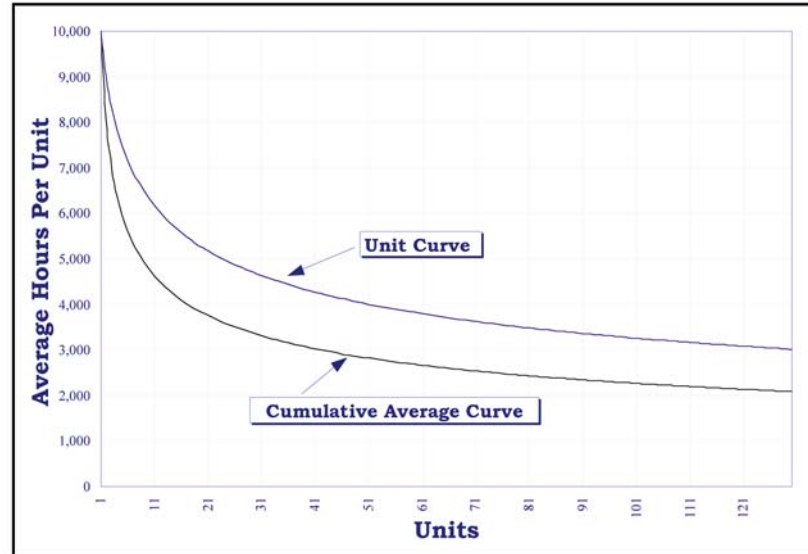
Reviewing the cumulative average improvement equation, it is important to note the similarities with the unit improvement curve equation. The form of the equations is the same. Both plot as straight lines on log-log paper and as hyperbolic lines on standard graph paper. Both calculate the constant ( $b$ ) in the same manner. The two equations differ only in the definition of the  $Y$  term. Unit curve theory describes or models the relationship between the cost of individual units. The cumulative average curve, however, describes the relationship between the average cost of different quantities of units.

The difference between the two curves is significant. Under the unit improvement theory, the cost of each unit is calculated separately using the equation from the previous section. The individual unit costs are summed to arrive at the total cost of  $X$  units. Under the cumulative average theory, the equation is used to calculate an average unit cost. The average unit cost is multiplied by  $X$  to arrive at the total cost of  $X$  units. This is better illustrated using a simple example with two data points ( $X = 2$ ), where the first unit takes 100 hours to produce ( $T_1 = 100$ ) and units thereafter follow a 20% rate of learning (yielding a slope of 0.8). Utilizing the unit improvement curve, the cost for data points 1 and 2 are 100 and 80, respectively. The total cost for the two data points, therefore, is 180 ( $100 + 80$ ). Using the cumulative average curve, the average cost ( $\bar{Y}_x$ ) for the two points is 80. The total cost is, therefore, 160 ( $2 \times 80$ ). Using the same first unit ( $T_1$ ) value and constant ( $b$ ), the analyst will always get lower total cost using the cumulative average curve because of the difference between  $Y_x$  and  $\bar{Y}_x$ . For purposes of comparing the two curves, Figure 15-7 shows both curves plotted using the average cost per unit. As demonstrated in the graph, the cumulative average curve will experience a much greater reduction in cost during the first few units of production; and it will,



therefore, have a steeper slope than the unit curve through the first stages of production.

**Figure 15-7. 80% Cumulative Average Curve with Corresponding Unit Curve**



### Selecting the Appropriate Theory

Since improvement curves are one of the most widely used and understood concepts of all cost analysis tools, analysts can expect questions regarding all aspects of a cost improvement curve used to develop an evaluated cost position. Where quantities exceed 100 units, a change of only a few percentage points in the slope value can make a large change in the total procurement cost. Contractors know this and may challenge the slope value or curve type used by the analyst in order to argue for higher or lower estimates. An analyst must be able to defend all cost improvement curve methods, assumptions, and input values used to develop an estimate.

The unit improvement curve is the predominant method used by both the Government and contractors. The cumulative average curve is usually used in situations where an end item is being produced for the first time or where design problems are not completely resolved. Generally, the following criteria can aid the analyst in determining which theory best suits the situation at hand.

1. If the estimate is for work by a specific company, use the theory applied in the plant where most of the work will be accomplished.
2. If the contractor has not been selected, look for the theory used by most companies competing for the contract.

3. If historical slope data are to be used, use the theory associated with the best historical slope data available.
4. If past data for the program is available to support an estimate for further production, use the theory that provides the best fit to the data.
5. If analysis of special curve situations is involved, methods of analysis may be more available for unit rather than cumulative average theory. (Information pertaining to special curve situations and is available in improvement curve textbooks.)
6. When totally unable to make a choice, make the estimate using both theories to see how significant the difference is relative to other areas of uncertainty, including slope uncertainty.

### 15.5.3 Note on Computer Models

Computer models exist that are designed specifically for improvement curve analysis, such as E-Z Quant. These models are accurate, fast, and easy to use. Modeling programs differ on the mathematical approach to producing curve information. When comparing the analyses generated by two or more programs, differences may occur. It may be necessary to reprocess the information using a common model.

## 15.6 SUMMARY

This chapter is not an all-encompassing, encyclopedic study of quantitative analysis. Rather, this chapter provides an understanding of the quantitative analysis techniques that are most helpful when analyzing cost and pricing data.

Analysts should understand descriptive statistics, sampling, regression analysis, and learning curves in order to thoroughly process cost and price information and provide qualified recommendations. Descriptive statistics allow the analyst to understand and describe the characteristics of a set of data. Sampling is useful because it expedites the analysis of large quantities of data without sacrificing accuracy. Regression analysis is the basis for examining and developing forecasts and cost estimating relationships. Learning curves are central to analyzing and creating cost estimates. Specific examples where these techniques are useful are discussed in the material, direct labor, and indirect cost chapters within this Handbook.

See GAO Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Capital Program Costs. GAO-09-3SP, Mar 2, 2009.

Website address: <http://www.gao.gov/products/GAO-09-3SP>