5.0 - Chapter Introduction

In this chapter, you will learn to use regression analysis in developing cost estimating relationships and other analyses based on a straight-line relationship even when the data points do not fall on a straight line.

Line-of-Best-Fit. The straight-line is one of the most commonly used and most valuable tools in both price and cost analysis. It is primarily used to develop cost estimating relationships and to project economic trends. Unfortunately, in contract pricing the data points that are used in analysis do not usually fall exactly on a straight line. Much of the variation in a dependent variable may be explained by a linear relationship with an independent variable, but there are usually random variations that cannot be explained by the line. The goal in establishing a line-of-best-fit is to develop a predictive relationship that minimizes the random variations. This can be done visually with a graph and a ruler, but the visual line-of-best-fit is an inexact technique and has limited value in cost or price analysis. Regression analysis is commonly used to analyze more complex relationships and provide more accurate results.

This chapter will focus on simple regression (2-variable linear regression); in which a single independent variable (X) is used to predict the value of a single dependent variable (Y). The dependent variable will normally be either price or cost (e.g., dollars or labor hours), the independent variable will be a measure related to the product (supply or service) being acquired. It may be a physical characteristic of the product, a performance characteristic of the product, or an element of cost to provide the product.

In some situations, you may need regression analysis tools that are more powerful than simple regression. Multiple regression (multivariate linear regression) and curvilinear regression are variations of simple regression that you may find useful. The general characteristics of both will be addressed later in the chapter.

5.1 - Identifying Situations For Use

Cost Estimating Relationship Development and Analysis. Regression analysis is one of the techniques most commonly used to establish cost estimating relationships (CERs) between independent variables and cost or price. If you can use regression analysis to quantify a CER, you can then use that CER to develop and analyze estimates of product cost or price.

Indirect Cost Rate Analysis (FAR 31.203). Indirect costs are costs that are not directly identified with a single final cost objective (e.g., contract item), but identified with two or more final cost objectives or an intermediate cost objective. In addition, minor direct costs may be treated as indirect costs if the treatment is consistently applied to all final cost objectives and the allocation produces substantially the same results as treating the cost as a direct cost.

Because indirect costs are not directly identified with a single final cost objective, they must be accumulated into logical cost pools and allocated to final cost objectives using indirect cost rates (e.g., overhead and general and administrative expense rates). The base used to allocate indirect costs should be selected to permit allocation of the cost pool on the basis of the benefits accruing to the various cost objectives.

Regression analysis is commonly used to quantify the relationship between the indirect cost rate base and pool over time. If you can quantify the relationship, you can then use that relationship to develop or analyze indirect cost rate estimates.

Time-Series Analysis. You can use regression analysis to analyze trends that appear to be related to time. It is particularly useful when you can identify and adjust for other factors that affect costs or prices (e.g., quantity changes) to isolate the effect of inflation/deflation for analysis. The most common applications of this type are forecasting future wage rates, material costs, and product prices.

In time-series analysis, cost or price data are collected over time for analysis. An estimating equation is developed using time as the independent variable. The time periods are normally weeks, months, quarters, or years. Each time period is assigned a number (e.g., the first month is 1, the fourth month is 4, etc.). All time periods during the analysis must be considered, whether or not data were collected during that period.

Time does not cause costs or prices to change. Changes are caused by a variety of economic factors. Do not use time-series analysis when you can identify and effectively measure the factors that are driving costs or prices. If you can identify and measure one or more key factors, you should be able to develop a better predictive model than by simply analyzing cost or price changes over time. However, if you cannot practically identify or measure such factors, you can often make useful predictions by using regression analysis to analyze cost or price trends over time.

Just remember that regression analysis will not automatically identify changes in a trend (i.e., it cannot predict a period of price deflation when the available data trace a trend of increasing prices). As a result, regression analysis is particularly useful in short-term analysis. The further you predict into the future, the greater the risk.

5.2 - Developing And Using A Simple Regression Equation

Simple Regression Model. The simple regression model is based on the equation for a straight line:

$Y_c = A + BX$

Where:

Y_c = The calculated or estimated value for the dependent (response) variable

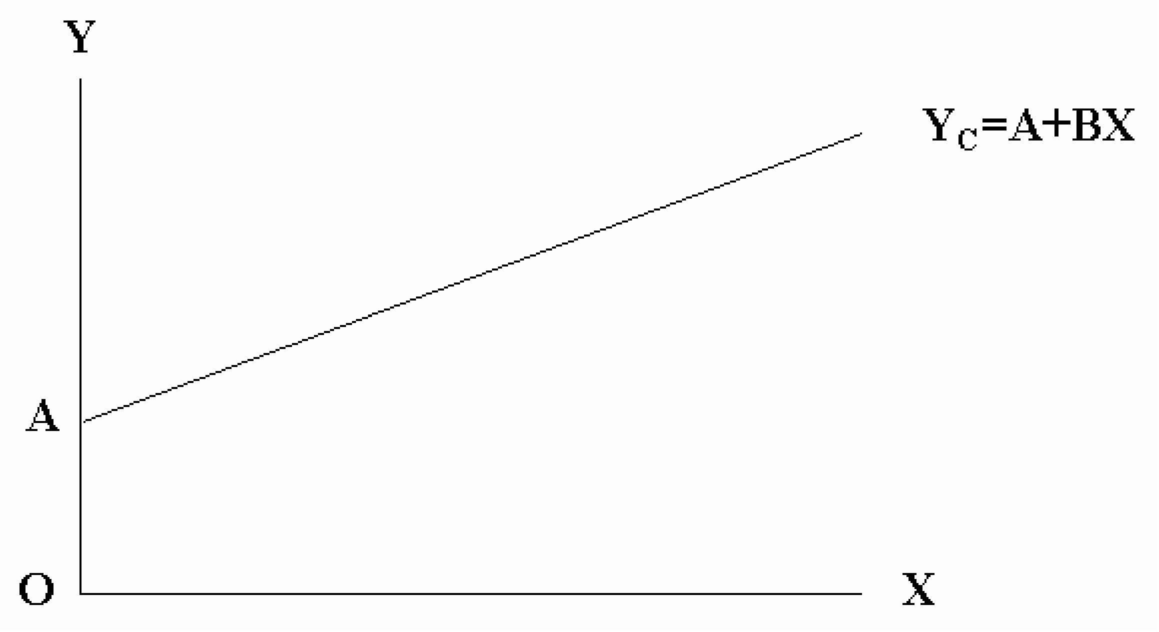
A = The Y intercept, the theoretical value of Y when X = 0

X = The independent (explanatory) variable

B = The slope of the line, the change in Y divided by the change in X, the value by which Y changes when X changes by one.

For a given data set, A and B are constants. They do not change as the value of the independent variable changes. Yc is a function of X. Specifically, the functional relationship between Yc and X is that Yc is equal to A plus the product of B times X.

The following figure graphically depicts the regression line:



Steps for Developing a 2-Variable Linear Regression Equation. To develop a regression equation for a particular set of data, use the following 5-step least-squares-best-fit (LSBF) process:

Step 1. Collect the historical data required for analysis. Identify the X and Y values for each observation.

X = Independent variable

Y = Dependent variable

Step 2. Put the data in tabular form.

Step 3. Compute $\bar{X}$and $\bar{Y}$.

$$\bar{X} = \frac{\Sigma X}{n} \qquad \bar{Y} = \frac{\Sigma Y}{n}$$

Where:

$\overline{X}$= Sample mean for observations the independent variable

$\overline{Y}$= Sample mean for observations the dependent variable

= Summation of all the variables that follow the symbol (e.g., X represents the sum of all X values)

X = Observation value for the independent variable

Y = Observation value for the dependent variable

n = Total number of observations in the sample

Step 4. Compute the slope (B) and the Y intercept (A).

$$B = \frac{\Sigma XY - n\,\overline{X}\,\overline{Y}}{\Sigma X^2 - n\overline{X}^2}$$

$$A = \overline{Y} - B\overline{X}$$

Step 5. Formulate the estimating equation.

$$Y_C = A + BX$$

2-Variable Linear Regression Equation Development Example. Assume a relationship between a firm's direct labor hours and manufacturing overhead cost based on the use of direct labor hours as the allocation base for manufacturing overhead. Develop an estimating equation using direct labor hours as the independent variable and manufacturing overhead cost as the dependent variable. Estimate the indirect cost pool assuming that 2,100 manufacturing direct labor hours will be needed to meet 19X8 production requirements.

Step 1. Collect the Historical Data Required for Analysis.

Historical Data		
Year	**Manufacturing Direct Labor Hours**	**Manufacturing Overhead**
19X2	1,200	$ 73,000
19X3	1,500	$ 97,000

19X4	2,300	$128,000
19X5	2,700	$155,000
19X6	3,300	$175,000
19X7	3,400	$218,000
19X8	2,100 (Est)	

Step 2. Put The Data In Tabular Form.

X = Manufacturing direct labor hours in hundreds of hours (00s)

Y = Manufacturing overhead in thousands of dollars ($000s)

	Tabular Presentation				
	X	**Y**	**XY**	**X^2**	**Y^2**
	12	73	876	144	5,329
	15	97	1,455	225	9,409
	23	128	2,944	529	16,384
	27	155	4,185	729	24,025
	33	175	5,775	1,089	30,625
	34	218	7,412	1,156	47,524
Column Totals	144	846	22,647	3,872	133,296

Step 3. Compute $\bar{X}$and $\bar{Y}$.

$$\begin{aligned}\bar{X} &= \frac{\Sigma X}{n} \\ &= \frac{144}{6} \\ &= 24\end{aligned}$$

$$\begin{aligned}\bar{Y} &= \frac{\Sigma Y}{n} \\ &= \frac{846}{6} \\ &= 141\end{aligned}$$

Step 4. Compute the slope (B) and the intercept (A).

$$
\begin{aligned}
B &= \frac{\Sigma XY - n\overline{X}\overline{Y}}{\Sigma X^2 - n\overline{X}^2} \\
&= \frac{22{,}647 - 6(24)(141)}{3{,}872 - 6(24)^2} \\
&= \frac{22{,}647 - 20{,}304}{3{,}872 - 3{,}456} \\
&= \frac{2{,}343}{416} \\
&= 5.6322
\end{aligned}
$$

$$
\begin{aligned}
A &= \overline{Y} - B\overline{X} \\
&= 141 - 5.6322(24) \\
&= 141 - 135.1728 \\
&= 5.8272
\end{aligned}
$$

Step 5. Formulate the estimating equation. Substitute the calculated values for A and B into the equation:

$$Y_C = A + BX$$

$$Y_C = 5.8272 + 5.6322X$$

Where:

Yc = Manufacturing overhead ($000's)

X = Manufacturing direct labor hours (00's)

Example of Estimate Using Simple Regression Equation. Estimate manufacturing overhead given an estimate for manufacturing direct labor hours of 2,100:

$$
\begin{aligned}
Y_C &= 5.8272 + 5.622X \\
&= 5.8272 + 5.6322(21) \\
&= 5.8272 + 118.2762 \\
&= 124.1034 \text{ thousand dollars}
\end{aligned}
$$

Rounded to the nearest dollar, the estimate would be $124,103.

5.3 Analyzing Variation In The Regression Model

Assumptions of the Regression Model. The assumptions listed below enable us to calculate unbiased estimators of the population) and to use these in predicting values and regression function coefficients (of Y given X). You should be aware of the fact that violation of one or more of these assumptions reduces the efficiency of the model, but a detailed discussion of this topic is beyond the purview of this text. Assume that all these assumptions have been met.

- For each value of X there is an array of possible Y values which is normally distributed about the regression line.
- The mean of the distribution of possible Y values is on the regression line. That is, the expected value of the error term is zero.
- The standard deviation of the distribution of possible Y values is constant regardless of the value of X (this is called "homoscedasticity").
- The error terms are statistically independent of each other. That is, there is no serial correlation.
- The error term is statistically independent of X.

Note: These assumptions are very important, in that they enable us to construct predictions around our point estimate.

Variation in the Regression Model. Recall that the purpose of regression analysis is to predict the value of a dependent variable given the value of the independent variable. The LSBF technique yields the best single line to fit the data, but you also want some method of determining how good this estimating equation is. In order to do this, you must first partition the variation.

- **Total Variation.** The sum of squares total (SST) is a measure of the total variation of Y. SST is the sum of the squared differences between the observed values of Y and the mean of Y.

$$SST = \Sigma(Y_i - \overline{Y})^2$$

Where:

SST = Sum of squared differences

Y_i = Observed value i

$\overline{Y}$ = Mean value of Y

While the above formula provides a clear picture of the meaning of SST, you can use the following formula to speed SST calculation:

$$SST = \Sigma Y^2 - \overline{Y}\Sigma Y$$

Total variation can be partitioned into two variations categories: explained and unexplained. This can be expressed as

SST = SSR + SSE

- **Explained Variation.** The sum of squares regression (SSR) is a measure of variation of Y that is explained by the regression equation. SSR is the sum of the squared differences between the calculated value of Y (Yc) and the mean of Y ($\overline{Y}$).

$$SSR = \Sigma(Y_C - \overline{Y})^2$$

You can use the following formula to speed SSR calculation:

$$SSR = B(\Sigma XY - \overline{X}\Sigma Y)$$

- **Unexplained Variation.** The sum of squares error (SSE) is a measure of the variation of Y that is not explained by the regression equations. SSE is the sum of the squared differences between the observed values of Y and the calculated value of Y. This is the random variation of the observations around the regression line.

$$SSE = \Sigma(Y_i - Y_C)^2$$

You can use the following formula to speed SSE calculation:

$$SSE = \Sigma Y^2 - A\Sigma Y - B\Sigma XY$$

Analysis of Variance. Variance is equal to variation divided by degrees of freedom (df). In regression analysis, df is a statistical concept that is used to adjust for sample bias in estimating the population mean.

- **Mean Square Regression (MSR).**

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{\mathrm{df}}$$

For 2-variable linear regression, the value of df for calculating MSR is always one (1). As a result, in 2-variable linear regression, you can simplify the equation for MSR to read:

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{1} \quad \text{or}$$
$$\mathrm{MSR} = \mathrm{SSR}$$

- **Mean Square Error (MSE).**

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{\mathrm{df}}$$

In 2-variable linear regression, df for calculating MSE is always n - 2. As a result, in simple regression, you can simplify the equation for MSE to read:

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2}$$

- **Analysis of Variance Table.** The terms used to analyze variation/variance in the regression model are commonly summarized in an Analysis of Variance (ANOVA) table.

ANOVA Table			
Source	**Sum of Squares**	**df**	**Mean Square****
Regression	SSR	1	MSR
Error	SSE	n-2	MSE
Total	SST	n-1	
**Mean Square = Sum of Squares/df			

Constructing an ANOVA Table for the Manufacturing Overhead Example. Before you can calculate variance and variation, you must use the observations to calculate the statistics in the table below. Since we already calculated these statistics to develop the regression equation to estimate manufacturing overhead, we will begin our calculations with the values in the table below:

Statistic	Value
?	144
?Y	846
?XY	22,647
?X^2	3,872
?Y^2	133,296
$\bar{X}$	24
$\bar{Y}$	141
A	5.8272
B	5.6322
n	6

Step 1. Calculate SST.

$$\begin{aligned} SST &= \Sigma Y^2 - \bar{Y}\Sigma Y \\ &= 133{,}296 - 141(846) \\ &= 133{,}296 - 119{,}286 \\ &= 14{,}010 \end{aligned}$$

Step 2. Calculate SSR.

$$\begin{aligned} SSR &= B(\Sigma XY - \bar{X}\Sigma Y) \\ &= 5.6322\,[22{,}647 - 24(846)] \\ &= 5.622\,[22{,}647 - 20{,}304] \\ &= 5.6322\,[2{,}343] \\ &= 13{,}196.24 \text{ (rounded to 13,196 for this example)} \end{aligned}$$

Step 3. Calculate SSE.

$$\begin{aligned} SSE &= \Sigma Y^2 - A\Sigma Y - B\Sigma XY \\ &= 133{,}296 - 5.8272(846) - 5.6322(22{,}647) \\ &= 133{,}296 - 4929.81 - 127{,}552.43 \\ &= 813.76 \text{ (rounded to 814 for this example)} \end{aligned}$$

Step 4. Calculate MSR.

$$\begin{aligned} MSR &= SSR \\ &= 13{,}196 \end{aligned}$$

Step 5. Calculate MSE.

$$
\begin{aligned}
\text{MSE} &= \frac{\text{SSE}}{n - 2} \\
&= \frac{814}{6 - 2} \\
&= \frac{814}{4} \\
&= 203.5 \text{ (rounded to 204 for this example)}
\end{aligned}
$$

Step 6. Combine the calculated values into an ANOVA table.

ANOVA Table

Source	Sum of Squares	df	Mean Square**
Regression	13,196	1	13,196
Error	814	4	204
Total	14,010	5	

**Mean Square = Sum of Squares/df

Step 7. Check SST. Assure that value for SST is equal to SSR plus SSE.

$$
\begin{aligned}
&\text{SST} = \text{SSR} + \text{SSE} \\
&14{,}010 = 13{,}196 + 814 \\
&14{,}010 = 14{,}010
\end{aligned}
$$

5.4 - Measuring How Well The Regression Equation Fits The Data

Statistics Used to Measure Goodness of Fit. How well does the equation fit the data used in developing the equation? Three statistics are commonly used to determine the "goodness of fit" of the regression equation:

- Coefficient of determination;
- Standard error of the estimate; and
- T-test for significance of the regression equation.

Calculating the Coefficient of Determination. Most computer software designed to fit a line using regression analysis will also provide the coefficient of determination for that line. The coefficient of determination (r^2)

measures the strength of the association between independent and dependent variables (X and Y).

The range of r^2 is between zero and one.

$0 < r^2 < 1$

An r^2 of zero indicates that there is no relationship between X and Y. An r^2 of one indicates that there is a perfect relationship between X and Y. As r^2 gets closer to 1, the better the regression line fits the data set.

In fact, r^2 is the ratio of explained variation (SSR) to total variation (SST). An r^2 of .90 indicates that 90 percent of the variation in Y has been explained by its relationship with X; that is, 90 percent of the variation in Y has been explained by the regression line.

$$r^2 = \frac{SSR}{SST}$$

For the manufacturing overhead example:

$$r^2 = \frac{13{,}196}{14{,}010}$$
$$= .94$$

This means that approximately 94% of the variation in manufacturing overhead (Y) can be explained by its relationship with manufacturing direct labor hours (X).

Standard Error of the Estimate. The standard error of the estimate (SEE) is a measure of the accuracy of the estimating (regression) equation. The SEE indicates the variability of the observed (actual) points around the regression line (predicted points). That is, it measures the extent to which the observed values (Yi) differ from their calculated values (Yc). Given the first two assumptions required for use of the regression model (for each value of X there is an array of possible Y values which is normally distributed about the regression line and the mean of this distribution (Yc) is on the regression line), the SEE is interpreted in a way similar to the way in which the standard deviation is interpreted. That is, given a value for X, we would generally expect the following intervals (based on the Empirical Rule):

- Yc = 1 SEE to contain approximately 68 percent of the total observations (Yi)
- Yc = 2 SEE to contain approximately 95 percent of the total observations (Yi)
- Yc = 3 SEE to contain approximately 99 percent of the total observations (Yi)

The SEE is equal to the square root of the MSE.

$$\text{SEE} = \sqrt{\text{MSE}}$$

For the manufacturing overhead example:

$$\begin{aligned}\text{SEE} &= \sqrt{204} \\ &= 14.28\end{aligned}$$

Steps for Conducting the T-test for the Significance of the Regression Equation. The regression line is derived from a sample. Because of sampling error, it is possible to get a regression relationship with a rather high r^2 (say > 80 percent) when there is no real relationship between X and Y. That is, when there is no statistical significance. This phenomenon will occur only when you have very small sample data sets. You can test the significance of the regression equation by applying the T-test. Applying the T-test is a 4-step process:

Step 1. Determine the significance level ( ).

= 1 - confidence level

The selection of the significance level is a management decision; that is, management decides the level of risk associated with an estimate which it will accept. In the absence of any other guidance, use a significance level of .10.

Step 2. Calculate T. Use the values of MSR and MSE from the ANOVA table:

$$T = \sqrt{\frac{\text{MSR}}{\text{MSE}}}$$

Step 3. Determine the table value of t. From a t Table, select the t value for the appropriate degrees of freedom (df). In 2-variable linear regression:

$$df = n - 2$$

Step 4. Compare T to the t Table value. Decision rules:

If T > t, use the regression equation for prediction purposes. It is likely that the relationship is significant.

If T < t, do not use the regression equation for prediction purposes. It is likely that the relationship is not significant.

If T = t, a highly unlikely situation, you are theoretically indifferent and may elect to use or not use the regression equation for prediction purposes.

Conducting the T-test for the Significance of the Regression Equation for the Manufacturing Overhead Example.

To demonstrate use of the T-test, we will apply the 4-step procedure to the manufacturing overhead example:

Step 1. Determine the significance level ( ). Assume that we have been told to use   = .05.

Step 2. Calculate T.

$$\begin{aligned} T &= \sqrt{\frac{MSR}{MSE}} \\ &= \sqrt{\frac{13{,}196}{204}} \\ &= \sqrt{64.69} \\ &= 8.043 \end{aligned}$$

Step 3. Determine the table value of t. The partial table below is an excerpt of a t table.

$$\begin{aligned} df &= n - 2 \\ &= 6 - 2 \\ &= 4 \end{aligned}$$

Partial t Table	
df	**t**
- - - - - - - - - - -	

2	4.303
3	3.182
4	2.776
5	2.571
6	2.447
- - - - - - - - - - - - -	

Reading from the table, the appropriate value is 2.776.

Step 4. Compare T to the t Table value. Since T (8.043) > t (2.776), use the regression equation for prediction purposes. It is likely that the relationship is significant.

Note: There is not normally a conflict in the decision indicated by the T-test and the magnitude of r^2. If r^2 is high, T is normally > t. A conflict could occur only in a situation where there are very few data points. In those rare instances where there is a conflict, you should accept the decision indicated by the T-test. It is a better indicator than r^2 because it takes into account the sample size (n) through the degrees of freedom (df).

5.5 - Calculating And Using A Prediction Interval

Formulating the Prediction Interval. You can develop a regression equation and use it to calculate a point estimate for Y given any value of X. However, a point estimate alone does not provide enough information for sound negotiations. You need to be able to establish a range of values which you are confident contains the true value of the cost or price which you are trying to predict. In regression analysis, this range is known as the prediction interval.

For a regression equation based on a small sample, you should develop a prediction interval, using the following equation:

$$Y_C \pm t(SEE)\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma X^2 - n\bar{X}^2}}$$

Note: The prediction interval will be smallest when X = $\bar{X}$. When $X = \bar{X}$, the final term under the radical sign becomes zero. The greater the difference between X and $\bar{X}$, the larger the final term under the radical sign and the larger the prediction interval.

Constructing a Prediction Interval for the Manufacturing Overhead Example. Assume that we want to construct a 95 percent prediction interval for the manufacturing overhead estimate at 2,100 manufacturing direct labor hours. Earlier in the chapter, we calculated Y_C and the other statistics in the following table:

Statistic	**Value**
Yc	124.1034
t (Use n - 2 df)	2.776
SEE	14.27
$\bar{X}$	24
ΣX^2	3,872

Using the table data, you would calculate the prediction interval as follows:

$$Y_C \pm t(SEE)\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma X^2 - n\bar{X}^2}}$$

$$124.1034 \pm 2.776(14.27)\sqrt{1 + \frac{1}{6} + \frac{(21 - 24)^2}{3{,}872 - 6(24)^2}}$$

$$124.1034 \pm 39.6135\sqrt{1 + .1667 + \frac{(-3)^2}{3{,}872 - 3{,}456}}$$

$$124.1034 \pm 39.6135\sqrt{1.1667 + \frac{9}{416}}$$

$$124.1034 \pm 39.6135\sqrt{1.1667 + .0216}$$

$$124.1034 \pm 39.6135\sqrt{1.1883}$$

$$124.1034 \pm 39.6135\,(1.0901)$$

$$124.1034 \pm 43.1827$$

When X = 21 the prediction interval is: $80.9207 \leq Y \leq 167.2861$.

Prediction Statement: We would be 95 percent confident that the actual manufacturing overhead will be between $80,921 and $167,286 at 2,100 manufacturing direct labor hours.

5.6 - Identifying The Need For Advanced Regression Analysis

Other Forms of Regression. In 2-variable regression analysis, you use a single independent variable (X) to estimate the dependent variable (Y), and the relationship is assumed to form a straight line. This is the most common form of regression analysis used in contract pricing. However, when you need more than one independent variable to estimate cost or price, you should consider multiple regression (or multivariate linear regression). When you expect that a trend line will be a curve instead of a straight line, you should consider curvilinear regression.

A detailed presentation on how to use multiple regression or curvilinear regression is beyond the scope of this text. However, you should have a general understanding of when and how these techniques can be applied to contract pricing. When you identify a situation that seems to call for the use of one of these techniques, consult an expert for the actual analysis. You can obtain more details on the actual use of these techniques from advanced forecasting texts.

Multiple Regression Situation. Multiple regression analysis assumes that the change in Y can be better explained by using more than one independent variable. For example, suppose that the Region Audit Manager (RAM) wants to determine the relationship between main-frame computer hours, field-audit hours expended in audit analysis, and the cost reduction recommendations sustained during contract negotiations.

Computer Hours	Field Audit Hours	Sustained Reduction
1.4	45	$290,000
1.1	37	$240,000
1.4	44	$270,000
1.1	45	$250,000
1.3	40	$260,000
1.5	46	$280,000
1.5	47	$300,000

It is beyond the purpose of this text to demonstrate how a multivariate equation would be developed using this data. However, we will describe the elements of the multivariate equation and the results of a regression analysis.

Three-Variable Linear Equation. Multiple regression can involve any number of independent variables. To solve the audit example above, we would use a three-variable linear equation -- two independent variables and one dependent variable.

$Y_C = A + B_1X_1 + B_2X_2$

Where:

Yc = The calculated or estimated value for the dependent (response) variable

A = The Y intercept, the value of Y when $X_1 = 0 \text{ and } X_2 = 0$

X_2 = The first independent (explanatory) variable

B_2 = The slope of the line related to the change in X_1, the value by which Y changes when X_1 changes by one.

X_2 = The second independent (explanatory) variable

B_2 = The slope of the line related to the change in X_2, the value by which Y changes when X_2 changes by one.

Results of Audit Data Three-Variable Linear Regression AnalysisI. Using the above data on audit analysis and negotiated reductions, an analyst identified the following three variables:

X_2 = Computer Hours

X_2 = Field Audit Hours

Y = Cost Reductions Sustained

The results of analysts analysis are shown in the following table:

Regression Results		
Predictor Variable	**Equation**	**r^2**
Computer Hours	$Y = A + BX_1$	.82
Field Audit Hours	$Y = A + B\ X_2$	.60
Comp Hrs and Field Audit Hrs	$Y = A + B_1X_1 + B_2X_2$	.88

You can see from the r^2 values in the above table that computer hours explains more of the variation in cost reduction recommendations sustained than is explained by field audit hours. If you had to select one independent variable, you would likely select Computer Hours. However, the combination of the two independent variables in multiple regression explains more of the variation in cost reduction recommendations sustained than the use of computer hours alone. The combination produces a stronger estimating tool.

Curvilinear Regression Analysis. In some cases, the relationship between the independent variable(s) may not be linear. Instead, a graph of the relationship on ordinary graph paper would depict a curve. You cannot directly fit a line to a curve using regression analysis. However, if you can identify a quantitative function that transforms a graph of the data to a linear relationship, you can then use regression analysis to calculate a line of best fit for the transformed data.

Common Transformation Functions	Examples
Reciprocal	$\frac{1}{X}$
Square Root	$\sqrt{X}$
Log-Log	logX
Power	X^2

For example, improvement curve analysis (presented later in this text) uses a special form of curvilinear regression. While it can be used in price analysis and material cost analysis, the primary use of the improvement curve is to estimate labor hours. The curve assumes that less cost is required to produce each unit as the total units produced increases. In other words, the firm becomes more efficient as the total units produced increases.

There are many improvement curve formulations but one of the most frequently used is:

$Y = AX^B$

Where:

Y = Unit cost (in hours or dollars of the Xth unit)

X = Unit number

A = Theoretical cost of the first unit

B = Constant value related to the rate of efficiency improvement

Obviously, this equation does not describe a straight line. However, using the logarithmic values of X and Y (log-log transformation), we can transform this curvilinear relationship into a linear relationship for regression analysis. The result will be an equation in the form:

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

Where:

logY = The logarithmic value of Y

logA = The logarithmic value of A

logX = The logarithmic value of X

We can then use the linear equation to estimate the logarithmic value of Y, and from that Y.

5.7 - Identifying Issues And Concerns

Questions to Consider in AnalysisI. As you perform price/cost analysis, consider the issues and concerns identified in this section, whenever you use regression analysis.

- ***Does the r^2 value indicate a strong relationship between the independent variable and the dependent variable?***

The value of r^2 indicates the percentage of variation in the dependent variable that is explained by the independent variable. Obviously, you would prefer an r^2 of .96 over an r^2 of .10, but there is no magic cutoff for r^2 that indicates that an equation is or is not acceptable for estimating purposes. However, as the r^2 becomes smaller, you should consider your reliance on any prediction accordingly.

- ***Does the T-test for significance indicate that the relationship is statistically significant?***

Remember that with a small data set, you can get a relatively high r^2 when there is no statistical significance in the relationship. The T-test provides a baseline to determine the significance of the relationship.

- ***Have you considered the prediction interval as well as the point estimate?***

Many estimators believe that the point estimate produced by the regression equation is the only estimate with which they need to be concerned. The point estimate is only the most likely estimate. It is part of a range of reasonable estimates represented by the prediction interval. The prediction interval is particularly useful in examining risk related to the estimate. A wide interval represents more risk than a narrow interval. This can be quite valuable in making decisions such as contract type selection. The prediction interval can also be useful in establishing positions for negotiation. The point estimate could be your objective, the lower limit of the interval your minimum position, and the upper limit your maximum position.

- ***Are you within the relevant range of data?***

The size of the prediction interval increases as the distance from $\bar{X}$ increases. You should put the greatest reliance on forecasts made within the relevant range of existing data. For example, 12 is within the relevant range when you know the value of Y for several values of X around 12 (e.g., 10, 11, 14, and 19).

- ***Are time series forecasts reasonable given other available information?***

Time series forecasts are all outside the relevant range of known data. The further you estimate into the future, the greater the risk. It is easy to extend a line several years into the future, but remember that conditions change. For example, the low inflation rates of the 1960s did not predict the hyper-inflation of the 1970s. Similarly, inflation rates of the 1970s did not predict inflation rates of the 1980s and 90s.

- ***Is there a run of points in the data?***

A run consisting of a long series of points which are all above or all below the regression line may occur when historical data are arranged chronologically or in order of increasing values of the independent variable. The existence of such runs may be a symptom of one or more of the following problems:

  - Some factor not considered in the regression analysis is influencing the regression equation (consider multivariate regression);
  - The equation being used in the analysis does not truly represent the underlying relationship between the variables;
  - The data do not satisfy the assumption of independence; or
  - The true relationship may be curvilinear instead of linear (consider curvilinear regression).

- ***Have you graphed the data to identify possible outliers or trends that cannot be detected through the mathematics of fitting a straight line?***

When you use 2-variable linear regression, you will fit a straight line through the data. However, the value of the relationship identified may be affected by one or more outliers that should not really be considered in your analysis. These can be easily identified through the use of a graph. Remember though, you cannot discard a data point simply because it does not fit on the line. The graph will help you identify an outlier, but you cannot discard it unless there is a valid reason (e.g., different methods were used for that item).

A graph can also permit you to identify situations where a single simple regression equation is not the best predictor. The graph may reveal that there is more than one trend affecting the data (e.g., the first several data points could indicate an upward trend, the latter data points a downward trend). It could also reveal the true relationship is a curve and not a straight line.

- ***Have you analyzed the differences between the actual and predicted values?***

Like the graph, this analysis will provide you information useful in identifying outliers (e.g., there may be one very large variance affecting the relationship). However, the outlier may not be as easy to identify as with a graph because the line will be pulled toward the outlier.

- ***Are you comparing apples with apples?***

Regression analysis, like any technique based on historical data, assumes that the past is a good predictor of the future. For example, you might establish a strong relationship between production labor hours and quality assurance labor hours. However, if either production methods or quality assurance methods change substantially (e.g., automation) the relationship may no longer be of any value.

- ***How current are the data used to develop the estimating equation?***

The more recent the data, the more valuable the analysis. Many things may have changed since the out-of-date data were collected.

- ***Would another independent variable provide a better estimating tool?***

Another equation may produce a better estimating tool. As stated above, you would likely prefer an equation with an r^2 of .96 over one with an r^2 of .10.

- ***Does the cost merit a more detailed cost analysis?***

If the cost is high and the r^2 is low, it may merit a more detailed analysis. For example, if you had a relatively low r^2 for a production labor effort, it may be worth considering the use of work measurement techniques in your analysis.