ELEMENTARY STATISTICAL TECHNIQUES USED IN COST ESTIMATING RELATIONSHIPS (CER's) DEVELOPMENT

Dragoş STUPARU, Tomiţă VASILE
UNIVERSITY OF CRAIOVA

Abstract:
The Cost Estimating Relationship (CER) is a mathematical expression which describes, for predicative purposes, the cost of an item or activity as a function of one or more independent variables. The methodology can provide accurate and supportable contractor estimates, lower cost proposal processes, and more cost-effective estimating systems. The contractor community uses parametric cost models, especially during product concept definition. These estimates are used for decision making regarding bid strategies and are used as submittals to the government. It is only at the production and full scale development phase that parametrics are not commonly utilized for official proposal submissions (although contractors still frequently use parametrics to generate target costs estimates).

Key words: *parametric cost estimating, regression analysis, curve fitting, curvilinear regression*

JEL classification: *C25 - Discrete Regression and Qualitative*

1. INTRODUCTION

The origins of parametric cost estimating date back to World War II. The war caused a demand for military aircraft in numbers and models that far exceeded anything the aircraft industry had manufactured before. While there had been some rudimentary work from time to time to develop parametric techniques for predicting cost, there was no widespread use of any cost estimating technique beyond a laborious buildup of labor-hours and materials.

The Military saw a need for a stable, highly skilled cadre of analysts to help with the evaluation of such alternatives. Around 1950, the military established the Rand Corporation in Santa Monica, California, as a civil "think-tank" for independent analysis. Over the years, Rand's work represents some of the earliest and most systematic studies of cost estimating in the airplane industry.

In the mid 1950's, Rand developed the most basic tool of the cost estimating discipline, the Cost Estimating Relationship (CER), and merged the CER with the learning curve to form the foundation of parametric aerospace estimating. This estimating approach is still used today.

Defined, a parametric cost estimate is one that uses Cost Estimating Relationships (CER's) and associated mathematical algorithms (or logic) to establish cost estimates. For example, detailed cost estimates for manufacturing and test of an end item (for instance, a hardware assembly) can be developed using very precise Industrial Engineering standards and analysis. Performed in this manner, the cost estimating process is laborious and time consuming. However, if history has demonstrated that test (as the dependent variance) has normally been valued at about 25% of the manufacturing value (the independent variable), then a detailed test estimate need not be performed and can simply be computed at the 25% (CER) level. It is important, though, that any CER's used be carefully tested for validity using standard statistical approaches.

Parametric techniques are a credible cost estimating methodology that can provide accurate and supportable contractor estimates, lower cost proposal processes, and more cost-effective estimating systems.

But the Cost Estimating Relationships have weaknesses:

1. CER's are sometimes too simplistic to forecast costs. Generally, if one has detailed information, the detail may be reliably used for estimates. If available, another estimating approach may be selected rather than a CER.

2. Problems with the database may mean that a particular CER should not be used. While the analyst developing a CER should validate that CER, it is the responsibility of any user to validate the CER by reviewing the source documentation. Read what the CER is supposed to estimate, what data were used to build that CER, how old the data are, how they were normalized, etc. Never use a cost model without reviewing its source documentation.

Now that we know what a CER is, how to develop a CER, when to use a CER, and some of a CER's strengths and weaknesses, we can develop techniques for building CER's. The LSBF technique is only one mathematical transformation of the database - the linear regression model. An analyst should be mindful that little in the estimating world is linear.

2. REGRESSION ANALYSIS

The purpose of regression analysis is to improve our ability to predict the next "real world" occurrence of our dependent variable. Regression analysis may be defined as the mathematical nature of the association between two variables. The association is determined in the form of a mathematical equation. Such an equation provides the ability to predict one variable on the basis of the knowledge of the other variable. The variable whose value is to be predicted is called the **dependent variable.** The variable about which knowledge is available or can be obtained is called the **independent variable.** In other words, the dependent variable is dependent upon his value of independent variables.

The relationships between variables may be linear or curvilinear. By linear, we mean that the functional relationship can be described graphically (on a common X-Y coordinate system) by a straight line and mathematically by the common form:

$$y = a + bx$$

where

y represents the calculated value of the dependent variable,

x - the independent variable,

b - the slope of the line, the change in *y* divided by the corresponding change in x

a and *b* are constants for any value of *x* and *y*

Looking at the bi-variate regression equation — the linear relationship of two variables — we find that regression analysis can be described by an equation. The equation consists of two distinctive parts, the functional part and the random part. The equation for a bi-variate regression population is:

$$Y = A + BX + E$$

where A + BX is the functional part (a straight line) and E is the random part.

A and B are parameters of the population that exactly describe the intercept and slope of the relationship.

E represents the ran or "error" part of the equation. The random part of the equation is always there because the errors of assigning values, the errors of measurement, and errors of observation. These types of errors are always with us because of our human limitations, and the limitations associated with real world events.

Since it is practically impossible to capture data for an entire population, we normally work with a sample from that population. We denote that we are working with a sample by adjusting our equation to the form:

$$y = a + bx + e,$$

where a + bx represents the functional part of the equation and *e* represents the random part.

Our estimate of A and B in the population are represented by *a* and *b*, respectively, in the sample equation. In this sense then, *a* and *b* are statistics. That is, they are estimates of population parameters. As statistics, they are subject to sampling errors. As such, a good random sampling plan is important.

1. The values of the dependent variable are distributed by a normal distribution function about the regression line.

2. The mean value of each distribution lies on the regression line.

3. The variance of each array of the independent variable is constant.

4. The error term in any observation is independent of the error term in all other observations. When this assumption is violated, data is said **autocorrelated.** This assumption fixes the error term to be a truly random variable.

5. There are no errors in the values of the independent variables. The regression model specifies that the independent variable be a fixed number — not a random variable. For example, you wish to estimate the cost of a new bomber aircraft at mach 2, then mach 2 is the value of the independent variable. Mach 2 is a fixed number. The regression model will not handle errors in the independent variables.

6. All causation in the model is one way. This simply means that if causation is built into the model, the causation must go from he independent variable to the dependent variable. Causation, though neither statistical nor a mathematical requirements, is a highly desirable attribute when using the regression model for forecasting. Causation, of course, is what we, as cost analysts, are expected to determine. We do this in our hypothesis of a logical mathematical relationship in either building or reviewing a CER equation.

3. CURVE FITTING

There are two standard methods of curve fitting. One method has the analyst plot the data and fit a smooth curve to the data. This is known as the **graphical method.** The other method uses formulas or a "best-fit" approach where an appropriate theoretical curve is assumed and mathematical procedures are used to provide the one "best-fit" curve; this is known as the **Least Squares Best Fit (LSBF) method.**

We are going to work the simplest model to handle, the straight line, which is expressed as:

$$Y = a + bx$$

Graphical Method

To apply the graphical method, the data must first be plotted on graph paper. No attempt should be made to make the smooth curve actually pass through the data points which have been plotted; rather, the curve should pass between the data points leaving approximately an equal number of points on either side of the line. For linear data, a clear ruler or other straightedge may be used to fit the curve. The objective in fitting the curve is to "best-fit" the curve to the data points plotted; that is, each data point plotted is equally important and the curve you fit must consider each and every data point.

Although considered a rather outdated technique today, plotting the data is still always a good idea. By plotting the data, we get a picture of the relationship and can easily focus on those points which may require further investigation. Hence, as a first step, we should plot the data and note any data points which may require further investigations before developing a forecasting graphical curve or mathematical equation.

LSBF Method

The LSBF method specifies the one line which best fits the data set we are working with. The method does this by minimizing the sum of the squared deviations of the observed values of Y and calculated values of Y. For example, if the distances: ($Y_{1,}$. Y_{C1},), (Y_2 - Y_{C2}), (Y_3 - Y_{C3}), (Y_4 - Y_{C4}), etc., parallel to the Y-axis, are measured from the observed data points to the curve, then the LSBF line is the one that minimizes the following equation (see Figure l):

$$(Y_1\text{-}Yc_1)^2 + (Y_2 - Y_{C2})^2 + (Y_3 - Y_{C3})^2 + ... + (Y_n - Y_{Cn})^2$$

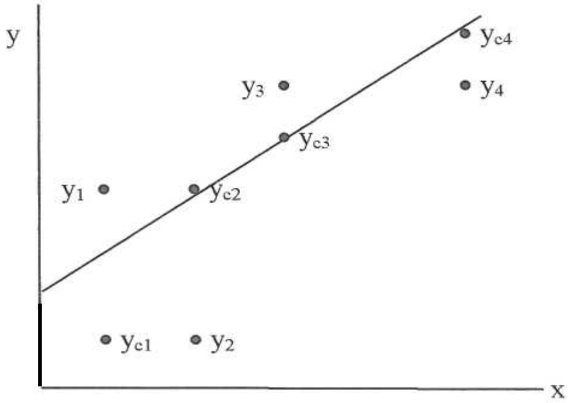

Figure 1. The LSBF Line

In other words, the sum of the deviations from the observed value of Y, and the calculated value of Y-Y_c squared, is a minimum; i.e., $(Y - Y_c)^2$ is a minimum. This same distance, (Y-Y_c) is the error term or residual. Therefore, the LSBF line is one that can be defined as:

ΣE^2 is a minimum because $\Sigma(Y - Y_c)^2 = \Sigma E^2$. For a straight line,

Y = a + bx and, with N points, we have

$(X_l, Y_1), (X_2, Y_2), (X_3, Y_3), ...(X_n,Y_n)$

The sum of the squares of the distances is a minimum if, $\Sigma Y = AN + B\Sigma X$ and $\Sigma XY = A\Sigma X + B\Sigma X^2$

These two equations are called the **normal equations** of the LSBF line. Reference to any comprehensive statistical textbook will illustrate that these two equations do meet the requirements of the properties of ordinary LSBF regression. These properties are:

1. The technique considers all points.

2. The sum of the deviations between the line and observed points is zero, that is, $\Sigma (Y - Y_c)^2 = \Sigma E^2$ = a minimum.

Similarities between these two properties and the arithmetic mean should also be observed. The arithmetic mean is the use of the values of the independent variable divided by the number of observations or $\Sigma X/n = \bar{x}$ and the sum of the "Ys" divided by the number of observations or $\Sigma y/n = \bar{y}$. Now, instead of considering the mean as a point when dealing with only one variable, we are now using a line ~ the LSBF regression line. Note that:

- The parameters, *a* and *b*, define a unique line with a Y-intercept of a and a slope of b.

- To calculate the values needed to solve for *a* and *b*, we need a spreadsheet (See Table 1). For example, use the values in Table 2.

Table 1. Table Needed to Get Sums, Squares and Cross Products

X	Y	X*Y	X^2	Y^2
X_1	Y_1	$X_1 * Y_1$	X_1^2	Y_1^2
X_2	Y_2	$X_2 * Y_2$	X_2^2	Y_2^2
X_3	Y_3	$X_3 * Y_3$	X_3^2	Y_3^2
-	-	-	-	-
-	-	-	-	-
ΣXn	**ΣYn**	**Σ (Xn * Yn)**	**ΣXn^2**	**ΣYn^2**

Table 2

X	Y	XY	X^2	Y^2
4	10	40	16	100
11	24	264	121	576
3	8	24	9	64
9	12	108	81	144
7	9	63	49	81
2	3	6	4	9
36	**66**	**505**	**280**	**974**

We can substitute the table values into the equations for *b* and *a*:
where:

$$\begin{cases} \Sigma Y = aN + b\Sigma X \\ \Sigma XY = a\Sigma X + b\Sigma X^2 \end{cases}$$

Solving for b, we get: $b = \dfrac{\Sigma XY - \bar{y}\Sigma X}{\Sigma X^2 - \bar{x}\Sigma X}$

$\bar{x}$ =ΣX/n =36/6=6, $\bar{y}$ =Σy/n=66/6=11

b = (505-11·36)/(280-6·36)

b = 1.703125

Solving for *a* yields: $a = \bar{y} - b\bar{x}$

a = 11-(1.703125) ·6

a =0.78125

Therefore, the regression equation (calculated y) is Y_c =0.78125+1.703125*x*

4. MULTIPLE REGRESSION

In simple regression analysis, a single independent variable (X) is used to estimate the dependent variable (Y), and the relationship is assumed to be linear (a straight line). This is the most common form of regression analysis used in contract pricing. However, there are more complex versions of the regression equation that can be used to consider the effects of more than one independent variable on Y. That is, multiple regression analysis assumes that the change in Y can be better explained by using more than one independent variable. For example, automobile gasoline consumption may be largely explained by the number of miles driven. However, it might be better explained if we also considered factors such as the weight of the automobile. In this case, the value of Y would be explained by two independent variables.

$$Yc = A + B_1X_1 + B_2X_2$$

where:

Yc = the calculated or estimated value for the dependent variable

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

X_1 = the first independent (explanatory) variable

B_1 = the slope of the line related to the change in X_1, the value by which change when X_1 changes by one

X_2 = the second independent variable

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

Multiple regression will not be considered in depth in this paper. Consult a statistics text to learn more about multiple regression.

5. CURVILINEAR REGRESSION

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. For example, improvement curve analysis uses a special form of curvilinear regression. As with multiple regression, consult a statistics text to learn more about curvilinear regression.

"Goodness" of fit, R and R^2

Now that we have developed the LSBF regression equations, we need to determine how good the equation is. That is, we would like to know how good a forecast we will get by using our equation. In order to answer this question, we must consider a check for the "goodness" of fit, the coefficient of correlation (R) and the related coefficient of determination (R^2). There are a number of other statistics we could check which would expand upon our knowledge of the regression equation and our assurance of its forecasting capability. Some of these will be discussed late.

Correlation Analysis

One indicator of the "goodness" of fit of a LSBF regression equation is correlation analysis. Correlation analysis considers how closely the observed points fall to the LSBF regression equation we develop. The assumption is that the more closely the observed values are to the regression equation, the better the fit; hence, the more confidence we can expect to have in the forecasting capability of our equation. It is important to note that correlation analysis refers only to the "goodness" of fit or how closely the observed values are to the regression equation. Correlation analysis tells us nothing about cause and effect, however.

Coefficient Of Determination

The coefficient of determination (R^2) represents the proportion of variation in the dependent variable that has been explained or accounted for by the regression line. The value of the coefficient of determination may vary from zero to one. A coefficient of determination of zero indicates that none of the variation in Y is explained by the regression equation; whereas a coefficient of determination of one indicates that 100 percent of the variation of Y has been explained by the regression equation. Graphically, when the R^2 is zero, the observed values appear as in Figure 2 (bottom) and when the R^2 is one, the observed values all fall right on the regression line as in Figure 2 (top).

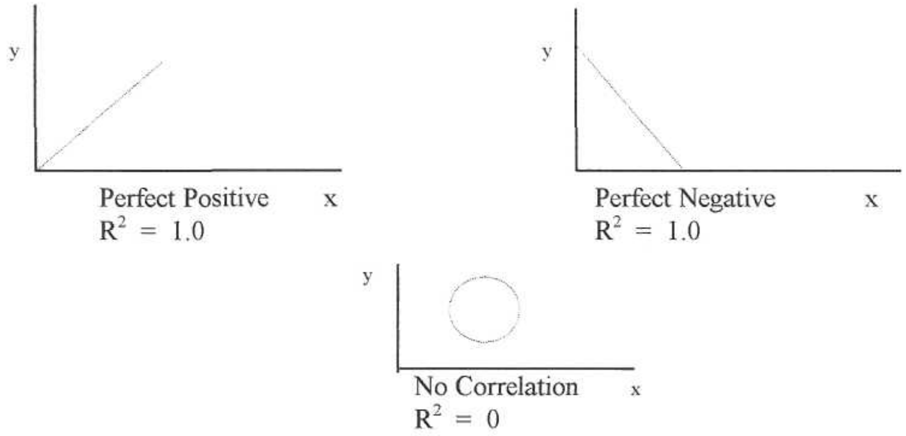

Figure 2

In order to calculate R^2 we need to use the equation:

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

R^2 tells us the proportion of total variation that is explained by the regression line. Thus R^2 is a relative measure of the "goodness" of fit of the observed data points to the regression line. For example, if we calculate R^2 using the formula above and find that $R^2 = 0.70$, this means that 70% of the total variation in the observed values of Y is explained by the observed values of X. Similarly, if $R^2 = 0.50$, then 50% of the variation in Y is explained by X. If the regression line perfectly fits all the observed data points, then all residuals will be zero, which means that $R^2 = 1.00$. In other words, a perfect straight-line fit will always yield $R^2 = 1$. As the level of fit becomes less accurate, less and less of the variation in Y is explained by Y's relation with X, which means that R^2 must decrease. The lowest value of R^2 is 0, which means that none of the variation in Y is explained by the observed values of X. Some applications require R^2 of at least 0.7 or 0.8. An $R^2 < 0.25$, which corresponds to an $R < 0.5$, would never be acceptable.

Coefficient Of Correlation

The coefficient of correlation ($R=\sqrt{R^2}$) measures both the strength and direction of the relationship between X and Y. The meaning of the coefficient of correlation is not as explicit as that of the coefficient of determination.

We can determine whether R is positive or negative by noting the sign of the scope of the line, or b. In other words, R takes the same sign as the slope; if b is positive, use the positive root of R^2 and vice versa. For example, if $R^2 = 0.81$; then $R = + 0.9$ and we determine whether R takes the positive root (+) or the negative root (-) by noting the sign of *b*. If *b* is negative, then we use the negative root of R^2 to determine R. So to calculate R we need to know the sign of the slope of the line.

It is most important to note that R does not tell us how much of the variation in Y is explained by the regression line. R is only valuable in telling us whether we have a direct or an inverse relationship and as a general indicator of the strength of the association.

6. THE LEARNING CURVE

Basic form of the "learning curve" equation is,
$y = a \cdot x^b$ or, Log y = Log a + b Log x where,

y = Cost of Unit /x (or average for x units)
a = Cost of first unit
b = Learning curve coefficient

Note that the equation Log y = Log a + b Log x is of precisely the same form as the linear equation y = a + bx. This means that the equation Log y = Log a + b Log x can be graphed as a straight line on log-log graph paper and all the regression formulae apply to this equation just as they do to the equation y = a + bx. In order to derive a learning curve from cost data (units or lots) the regression equations need to be used whether or not the calculations are performed manually or using a statistical package for your personal computer. In this sense, the learning curve equation is a special case of the LSBF technique.

Since in learning curve methodologies cost is assumed to decrease by a fixed amount each time quantity doubles, then this constant is called the learning curve "slope" or percentage (i.e., 90%). For example,

For unit #1 $Y_1 = A(1)^b = A$ (First Unit Cost) and

For unit #2 $Y_2 = A(2)^b$ = Second Unit Cost So,

Y=A·"Slope", Slope = 2^b, and, Log Slope = bLog2. Therefore, b = Log Slope/Log2.

For a 90% "Slope", b = Log 0.9/Log2 = -0.152.

If we assume that A = 1.0, then the relative cost between any units can be computed: $Y_3 = (3)^{-0152} = 0.8462$, $Y_6 = (6)^{-0152} = 0.7616$

Note that:

$Y_6 = 0.7616 = 0.9\ Y_3 = 0.9\ \ 0.8462$

Any good statistical package (for instance StatView) can perform all the calculations (and many others) shown. A quality package will let you customize your results (create presentations) save your work, and calculate all these statistics: frequency distributions, percentiles, t-tests, variance tests, Pearson correlation and covariance, regression, ANOVA, factor analysis and more. Graphics and tables such as scattergrams, line charts, pie charts, bar charts, histograms, percentiles, factors, etc., should be available to the user. Statistical analysis is greatly simplified using these tools.

REFERENCES

1. Stuparu, D., Vasile, T. – „*Matematici aplicate în economie*”, Editura Şcoala Mehedințiului, Drobeta Turnu Severin, 2002,

2. Vasile, T. – “*Metode statistice în managementul afacerilor*”, Editura Sitech, Craiova, 2008,

3. Vasilescu, N., Costescu, M., Ionaşcu, C., Babucea, G., Vasile, T., Stuparu, D. – „*Statistică*”, Editura Universitaria, Craiova, 2003,

4. Wonnacott, T., H., Wonnacott, R., J. – „*Statistique*”, Editure Economica, Paris, 1991.