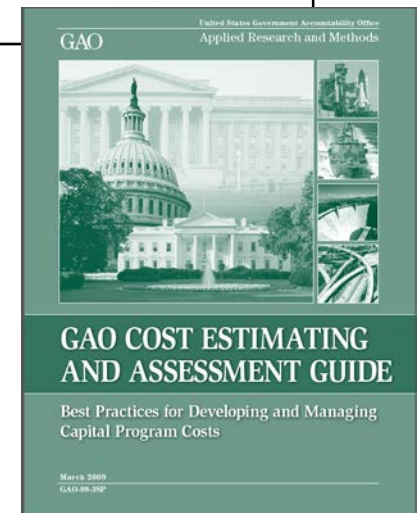
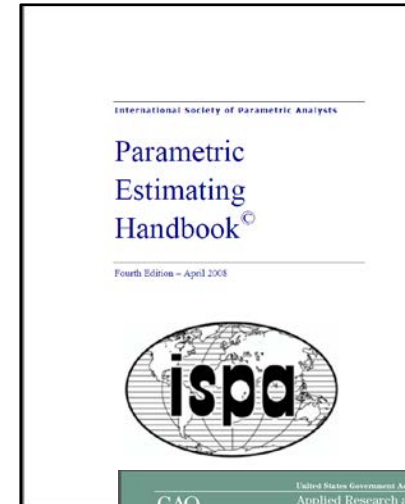




Parametric Estimating – Linear Regression

There are a variety of resources that address what are commonly referred to as parametric or regression techniques. The Parametric Estimating Handbook, the GAO Cost Estimating Guide, and various agency cost estimating and contract pricing handbooks will typically outline the steps for developing cost estimating relationships (CERs), and provide explanations of some of the more common statistics used to judge the quality of the resulting equation.

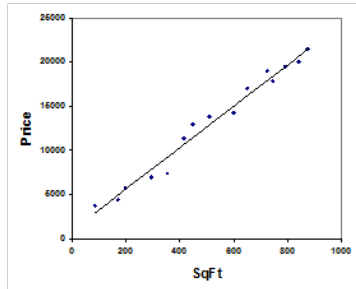
This job aid outlines such steps and statistics, beginning with an “at-a-glance” overview, and then offering somewhat more expanded discussions in the subsequent pages.



Developing Cost Estimating Relationships

Developing Linear Cost Estimating Relationships

“Cost” is used in the general sense to refer to resources such as dollars, hours, quantities of materials, etc.



$$y = mx + b \quad Y_c = A + BX \quad \hat{y}_x = b_0 + b_1x$$

1. Identification of Cost Drivers (eXplanatory variables)

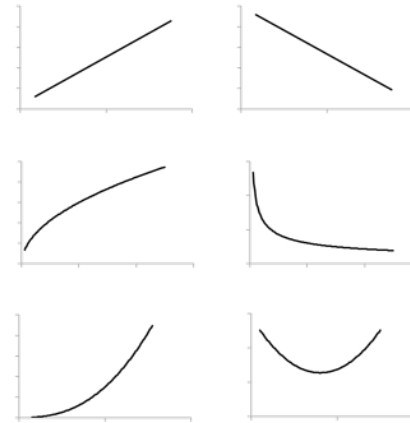
Meaningful cost drivers that capture physical, technical, performance, and other characteristics are identified thru discussions with experts, review of technical literature, industry reports, previous analyses, and personal experience.

Qualities of a good cost driver:

- Causal (direct relationship)
- Major (important characteristic)
- Significant (explains variation)
- Quantifiable (easily measured)
- Collectable (data availability)
- Predictable (known in advance with some degree of confidence)

2. Specification (what you expect)

Based on your understanding of the cost drivers, what do you expect the relationship to look like between “cost” and the cost drivers.



3. Data Collection

Gathering data on the cost and cost drivers for the same or similar items or services.

What data sources are available; what data has been used in the past.

Experts should be consulted when selecting similar items and services.

“Similar” selection criteria:

- Same form, fit, function as what is being estimated
- Same drivers as what is being estimated
- Same relationships between the variables as what is being estimated

4. Normalization

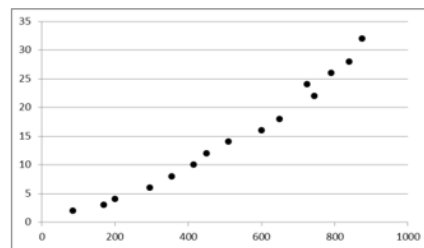
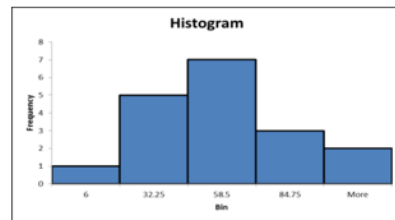
Are the previous prices, costs, hours, materials, etc. a valid basis for comparison.

You need to account for:

- Discounts due to quantity
- Learning curves
- Inflation or escalation
- Differences in content (i.e. what is or is not included in the data)
- Material differences
- Differences in complexity
- Performance differences
- Varying seller pricing strategies
- Acquisition environment
- Contract type
- Market conditions
- Changes in technology
- Are you looking at what it cost, or, what we paid

5. Graphical Analysis

You want to note trends, patterns, and unusual values in the data. Are the relationships what we expected to see.



6. Selecting a Fitting Approach and Fitting the Data

A) Linear

Linear with Intercept $\hat{Y}_x = b_0 + b_1 X$

Linear without Intercept (Factors) $\hat{Y}_x = b_1 X$

$$b_1 = \frac{\hat{Y}_x}{X}$$

B) Non-Linear

Linear ($\hat{Y}_x = b_0 + b_1 X$) with an X transformation such as: X^2 or $\frac{1}{X}$

Transform X and Y $\hat{Y}_x = b_0 (X)^{b_1}$

Higher order models

$$\hat{Y}_x = b_0 + b_1 X + b_2 X^2$$

Is the equation consistent with my specification (expectation)?

6. A) Linear Model with Intercept

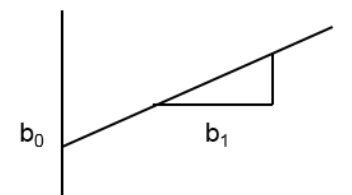
$$\hat{Y}_x = b_0 + b_1 X$$

\hat{Y}_x Predicted average value of Y for a given value of X

b_0 Y intercept

b_1 slope; change in Y given a one-unit change in X

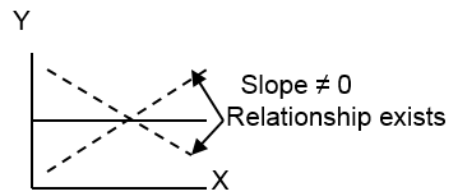
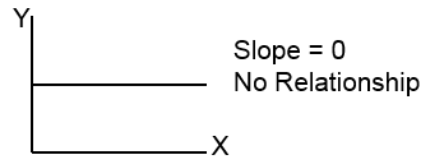
X value of the independent variable for what you are predicting



Developing Cost Estimating Relationships

7. Confidence or Significance

How confident am I that there is a statistical relationship between the X variable and the Y variable? Should I consider using this equation or not?

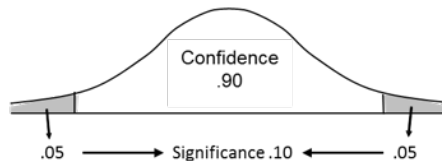


7. Confidence (cont.)

The T statistic measures how far the slope is from 0 in standard deviations. The further from 0, the more confident we are that there is a relationship between the dependent and independent variable.

$$T - \text{statistic} = \frac{\text{Slope}}{\text{Std Dev of Slope}}$$

Regression outputs typically provide the confidence level (or significance level) associated with the T-statistic.



ANOVA (Analysis of Variance)

SST – Sum of Squares Total (the total squared variation of the observations around the mean)

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

SSR – Sum of Squares Regression (the amount of the variation around the mean explained by the equation)

$$SSR = \Sigma(\hat{Y} - \bar{Y})^2$$

SSE – Sum of Squares Error (the amount of the variation around the mean not explained by equation, i.e. the unexplained variation)

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

SSE DF – Degrees of Freedom (n – p) Or (n – k – 1)

“p” is the number of estimated parameters in the equation (e.g. b_0, b_1, b_2)

“k” is the number of independent variables in the equation and the “1” represents the intercept

8. Accuracy

How accurate is the equation?

$$\text{Variance (MSE)} = \frac{SSE}{n - p}$$

MSE: Mean (average) Squared Error

$$\text{Standard Error (SE)} = \sqrt{\text{Variance}}$$

“Average” or “typical” estimating error

$$\text{Coefficient of Variation (CV)} = \frac{SE}{\bar{Y}}$$

CV x 100 can be interpreted as the “average” percent estimating error

9. Variation

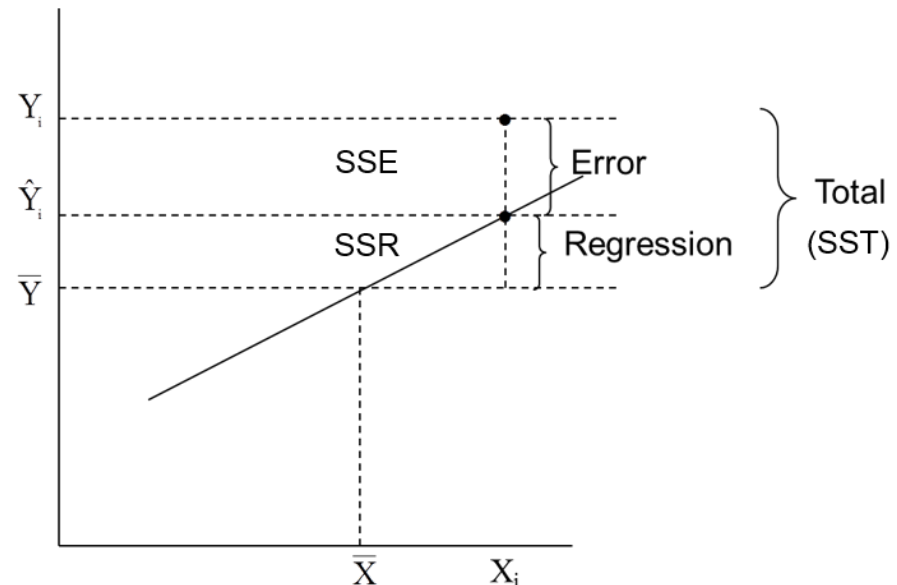
How much of the variation in the dependent variable can be explained by the variation in the independent variable?

Coefficient of Determination (R^2)

$$R^2 = \frac{SSR}{SST} \text{ or } 1 - \frac{SSE}{SST}$$

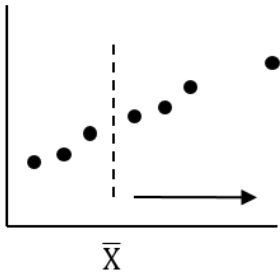
Sometimes considered a measure of the *strength* of the relationship between the variables.

R^2 is a measure of *correlation*, not *causation*, so don't just assume that an association implies causation.



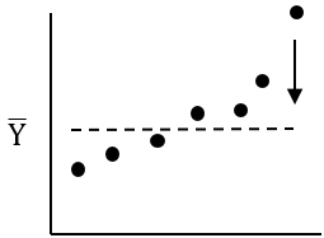
Developing Cost Estimating Relationships

10. Outliers with respect to X and Y



$$\frac{X - \bar{X}}{S_x} > \pm 2 \text{ std devs from } \emptyset \text{ is an outlier}$$

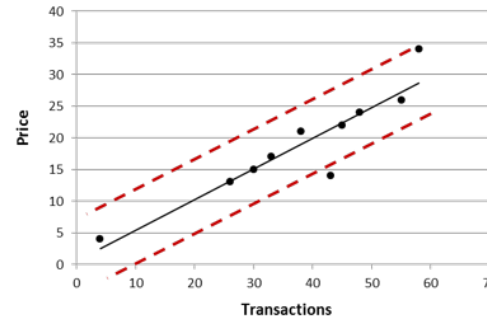
$$\text{Leverage (LV)} = \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}$$



$$\frac{Y - \bar{Y}}{S_y} > \pm 2 \text{ std devs from } \emptyset \text{ is an outlier}$$

Is it part of the population; are there classes within the data; could it be a measurement error; normalization error; data entry error; or an unusual event?

11. Outliers with respect to the predicted value (\hat{Y}) (prediction problems)



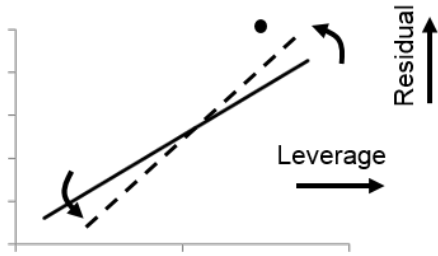
Questions:
Part of the population;
classes within the data;
measurement error;
normalization error;
data entry error; or
unusual event?
Missing a cost driver?
Wrong model form?

$$\text{Standardized Residual} = \frac{(Y - \hat{Y})}{SE} > \pm 2 \text{ std errors from } \emptyset \text{ is an outlier}$$

$$\text{Studentized Residual} = \frac{(Y - \hat{Y})}{SE\sqrt{(1 - \text{Leverage})}} > \pm 2 \text{ std errors from } \emptyset \text{ is an outlier}$$

12. Influential Observations

Is there a particular data point having significantly more influence on the slope and intercept of the equation than the other data points?

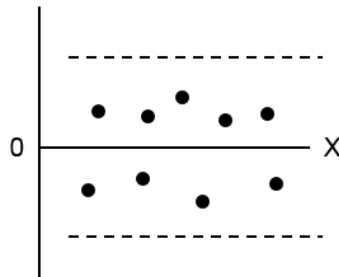


Questions:
Part of the population; classes within the data; measurement, normalization, or data entry error; or unusual event?
Missing cost driver? Wrong model?

13. Residuals

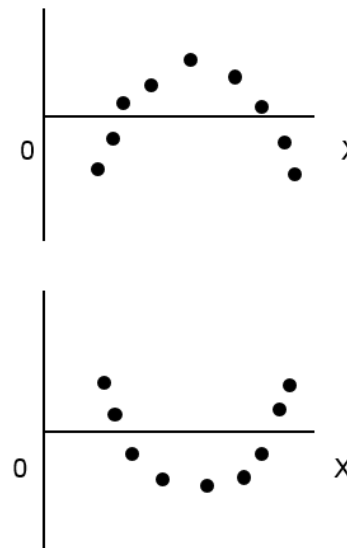
Did we properly fit the data (i.e. are the residuals randomly distributed about zero with a constant variance across the range of X values)?

random pattern; constant variance

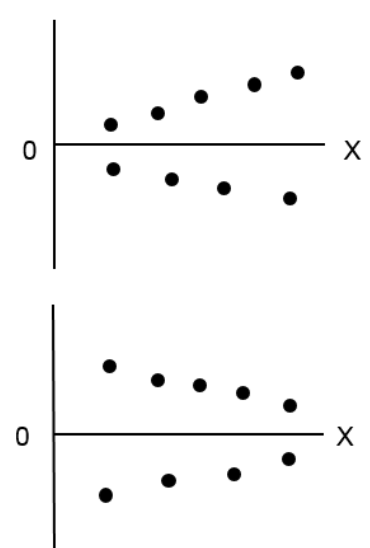


$$\text{Residual } (e_i) = Y - \hat{Y}$$

Problem: Some non-random patterns in the residuals can be indications of nonlinear data.



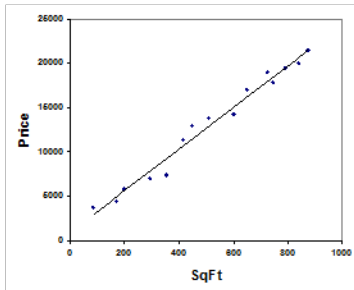
Problem: Some non-constant patterns can be indications of problems with the form of the equation.



Developing Cost Estimating Relationships

Developing Linear Cost Estimating Relationships

“Cost” is used in the general sense to refer to resources such as dollars, hours, quantities of materials, etc.



$$y = mx + b \quad Y_c = A + BX \quad \hat{y}_x = b_0 + b_1x$$

The term “cost estimating relationship” or CER is used here in the context of an equation where we predict the outcome of one variable as a function of the behavior of one or more other variables.

We refer to the predicted variable as the dependent or Y variable. Those variables that can be said to explain or drive the behavior of the predicted variable are consequently called explanatory variables or cost drivers, also known as the independent or X variables.

The terms CERs, equations, or models are often used interchangeably, although the term “model” is sometimes reserved to describe an assemblage of CERs, equations, or factors such as is often the case with respect to a software cost estimating “model”.

Identification of Cost Drivers

Is it possible for a variable to be a major cost driver, but not be a significant cost driver?

1. Identification of Cost Drivers (eXplanatory variables)

Meaningful cost drivers that capture physical, technical, performance, and other characteristics are identified thru discussions with experts, review of technical literature, industry reports, previous analyses, and personal experience.

Qualities of a good cost driver:

- Causal (direct relationship)
- Major (important characteristic)
- Significant (explains variation)
- Quantifiable (easily measured)
- Collectable (data availability)
- Predictable (known in advance with some degree of confidence)

Consider the engine in a car. The engine is both a major cost element of the car's price, and a major cost driver in that variations in the size of the engine generally have a significant impact on the cost.

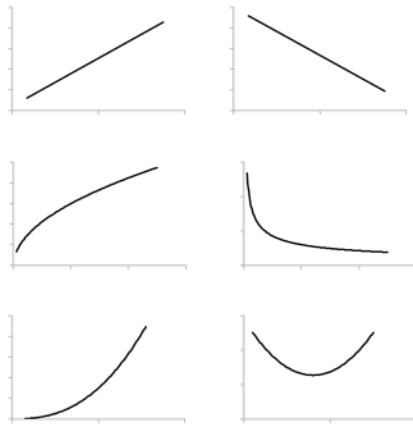
However, if you were trying to discern why the prices of a certain set of cars varied, and those cars all had engines of a similar size, then the engine size would not be "significant" in that the engine size would not discriminate between, or explain, differences in the prices.

Nevertheless, it would still be important to document the engine size as being a major cost driver in the event you were faced with estimating the price of a car with a different engine size. In that case, the engine size might then become a significant cost driver.

Specification of the Relationship

2. Specification (what you expect)

Based on your understanding of the cost drivers, what do you expect the relationship to look like between “cost” and the cost drivers.



Generally speaking, there are two approaches to fitting data. In one approach you simply rely on the data to “speak for itself”. The analyst either observes patterns in the data and makes a selection, or they run successive regressions of varying types and determine which equation “best fits” the data. This approach is best suited for situations where there is an abundance of data, or compelling patterns in the data.

However, the analyst faced with smaller data sets and less compelling data is better advised to first seek the expectations of the subject matter experts, hypothesize a relationship, and then test that hypothesis. This makes for a more sound and defensible estimating relationship.

Data Collection

3. Data Collection

Gathering data on the cost and cost drivers for the same or similar items or services.

What data sources are available; what data has been used in the past.

Experts should be consulted when selecting similar items and services.

“Similar” selection criteria:

- Same form, fit, function as what is being estimated
- Same drivers as what is being estimated
- Same relationships between the variables as what is being estimated

One challenge that analysts sometimes face is a shortage of similar items for comparison. When searching for “similar” items, the analyst should consider the level at which that similarity need exist. For example, an analyst pricing a component on a stealth ship need not unnecessarily constrain themselves to stealth ships when it is possible that the same or similar component is used on a variety of ships, and perhaps ground vehicles and aircraft as well.

Another consideration for expanding the number of similar items is to estimate at lower levels of the work breakdown structure when necessary, and when such data exists.

Familiarize yourself with the data bases that your organization and others maintain.

Normalizing the Data

4. Normalization

Are the previous prices, costs, hours, materials, etc. a valid basis for comparison.

You need to account for:

- Discounts due to quantity
- Learning curves
- Inflation or escalation
- Differences in content (i.e. what is or is not included in the data)
- Material differences
- Differences in complexity
- Performance differences
- Varying seller pricing strategies
- Acquisition environment
- Contract type
- Market conditions
- Changes in technology
- Are you looking at what it cost, or, what we paid

You probably learned at some point in a science class that when constructing experiments to study the effects of one thing upon another that it was important to hold everything else constant as much as possible. That's essentially the idea behind normalizing the data, to get a more true measure of the relationship between the Y and X variables.

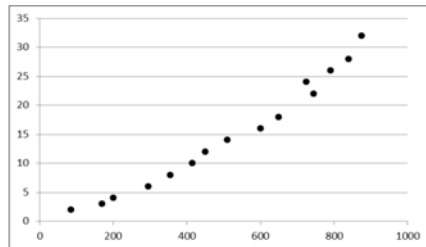
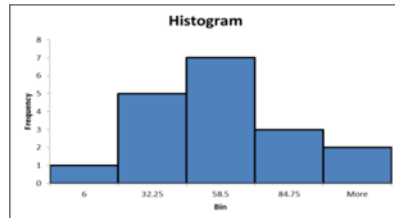
One of the other benefits of going through the process of making these adjustments in the data is that it forces us to develop a better understanding of the effects that each of these factors has on "cost", such as the changes in quantity, technology, material, labor differences, etc.

Look for checklists that your agency and others use when comparing prices, costs, hours, etc.

Graphical/Visual Analysis of the Data

5. Graphical Analysis

You want to note trends, patterns, and unusual values in the data. Are the relationships what we expected to see.



If you're counting on the R^2 , T statistic, F statistic, standard error, or coefficient of variation to tell you that you are not properly fitting the data, or that you have outliers or gaps in the data, then get ready to be disappointed, because they won't!

Histograms, number lines, scatterplots, and more. There is just no substitution for looking at the data.

Hopefully the scatterplots will be consistent with your expectations from the specification step. If not, then this is the opportunity to reengage with your subject matter experts.

Scatterplots may highlight subgroups or classes within the data; changes in the cost estimating relationship over the range of the data; and unusual values within the data set.

Selecting a Fitting Approach for the Data

6. Selecting a Fitting Approach and Fitting the Data

A) Linear

Linear with Intercept $\hat{Y}_X = b_0 + b_1 X$

Linear without Intercept
(Factors) $\hat{Y}_X = b_1 X$

$$b_1 = \frac{\hat{Y}_X}{X}$$

B) Non-Linear

Linear ($\hat{Y}_X = b_0 + b_1 X$) with an X transformation such as: X^2 or $\frac{1}{X}$

Transform X and Y $\hat{Y}_X = b_0 (X)^{b_1}$

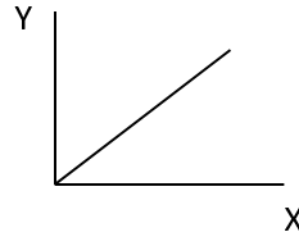
Higher order models

$$\hat{Y}_X = b_0 + b_1 X + b_2 X^2$$

Is the equation consistent with my specification (expectation)?

This step is the extension of the specification and visual analysis steps. We will concern ourselves here with linear relationships. See the Nonlinear job aid for discussions on fitting nonlinear data.

The most common linear relationship is a factor. The use of a factor presumes a direct proportional relationship between the X and Y variables.



A factor can be derived from a single data point or from a set of data points. Regression can be used to derive a factor when dealing with a set of data points by specifying a zero intercept or “constant is zero” in applications such as Excel. (See the Factors job aid for further discussion on the use of factors.)

The Linear Equation

The equation of a line has been represented by a host of different letters and characters, at our earliest ages commonly by ($y = mx + b$).

6. A) Linear Model with Intercept

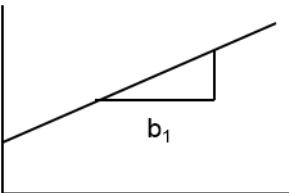
$$\hat{Y}_X = b_0 + b_1 X$$

\hat{Y}_X Predicted average value of Y for a given value of X

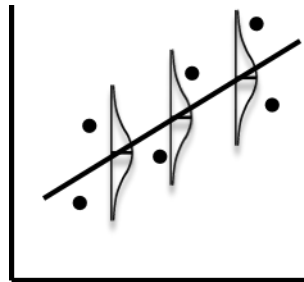
b_0 Y intercept

b_1 slope; change in Y given a one-unit change in X

X value of the independent variable for what you are predicting



One of the assumptions regarding the regression line is that there is a distribution of Y values around every point on the line. When we fit a line through the data, the resulting line represents the mean or average value of Y for every value of X.



So we need to recognize there is a range of possible Y values for a given value of X, and that the estimated Y value is essentially the average of the possible outcomes.

For example, given the equation:

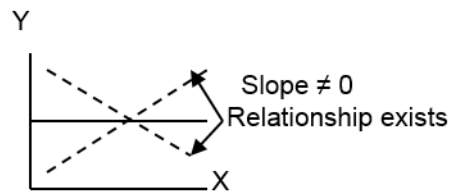
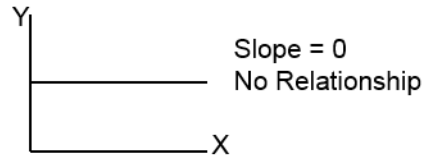
$$\begin{aligned} \text{TV Price} &= 76.00 + 13.25 (\text{Inches}) \\ &= 76.00 + 13.25 (32) \\ &= \$500 \end{aligned}$$

We would say that the average price for a TV with a diagonal of 32 inches is estimated to be \$500.

Determining the Confidence Level

7. Confidence or Significance

How confident am I that there is a statistical relationship between the X variable and the Y variable? Should I consider using this equation or not?



The T statistic measures how far the slope is from 0 in standard deviations. The further from 0, the more confident we are that there is a relationship between the dependent and independent variable.

$$T - \text{statistic} = \frac{\text{Slope}}{\text{Std Dev of Slope}}$$

Regression outputs typically provide the confidence level (or significance level) associated with the T-statistic.



The T-test is a hypothesis test. The null hypothesis is that there is no relationship between the X and Y variables, in which case the slope would equal zero.

The alternate hypothesis is that the X and Y variables are related, which would be evidenced by a positive or negative slope.

The test requires that we measure how far the slope is from zero in units of standard deviations (T calc) so that we can associate the distance with a measure of probability. Most regression applications report the T calc value and the probability associated with it.

The probability can either be stated in terms of the level of significance or as the level of confidence. Some applications report "P" or a "P value" which is the level of significance. A "P value" of 0.10 equates to a 0.90 or 90% level of confidence. Exercising some latitude with the terminology, we might say that we are 90% confident that the X and Y variables are related.

The analyst then needs to decide what constitutes an acceptable level of confidence that would warrant considering the equation.

ANOVA (Analysis of Variance)

SST – Sum of Squares Total
(the total squared variation of the observations around the mean)

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

SSR – Sum of Squares Regression
(the amount of the variation around the mean explained by the equation)

$$SSR = \Sigma(\hat{Y} - \bar{Y})^2$$

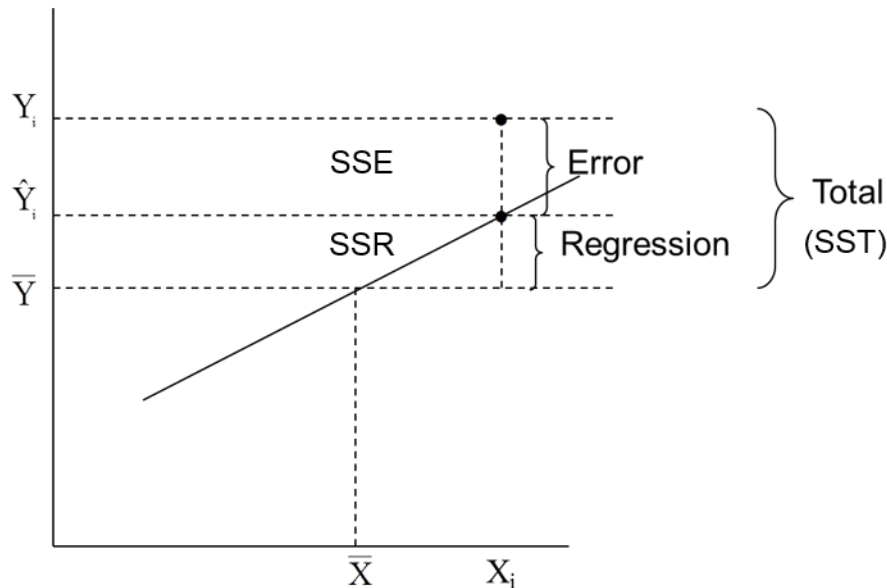
SSE – Sum of Squares Error
(the amount of the variation around the mean not explained by equation, i.e. the unexplained variation)

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

SSE DF – Degrees of Freedom ($n - p$) Or ($n - k - 1$)

“p” is the number of estimated parameters in the equation
(e.g. b_0, b_1, b_2)

“k” is the number of independent variables in the equation and the “1” represents the intercept



The ANOVA Table

The next two topics will make reference to SSE, SSR, and SST. They also refer to the Degrees of Freedom (DF) notation:

$$n - p$$

or alternatively

$$n - k - 1$$

DF is the adjustment made to sample statistics to better estimate the associated population parameters (e.g. the sample variance as an estimate of the population variance).

In the diagram, Y_i is the actual value of Y for a given X.

The Y with the hat or ^ symbol is the predicted value of Y for a given X.

The Y with the line over it (pronounced Y bar) represents the average of all the Y values.

How Accurate is the Equation?

8. Accuracy

How accurate is the equation?

$$\text{Variance (MSE)} = \frac{\text{SSE}}{n - p}$$

MSE: Mean (average) Squared Error

$$\text{Standard Error (SE)} = \sqrt{\text{Variance}}$$

“Average” or “typical” estimating error

$$\text{Coefficient of Variation (CV)} = \frac{\text{SE}}{\bar{Y}}$$

CV x 100 can be interpreted as the “average” percent estimating error

Ideally we would test the accuracy of the equation by reserving some of the data points, and then assessing how well we estimated those points based on the equation resulting from the remaining data.

Unfortunately, due to the far too common problem of small data sets we cannot afford the luxury of reserving data points. As an alternative, we measure how well we fit the data used to create the equation.

The SSE is the sum of the squared errors around the regression line. By dividing the SSE by “ $n - p$ ” or alternatively by “ $n - k - 1$ ”, we arrive at the variance of the equation, and what we might loosely call the average squared estimating error.

The square root of the variance is commonly called the standard error or standard error of the estimate. Again, we could loosely interpret this as the average estimating error.

By comparing the standard error to the average value of Y we get a relative measure of variability called the coefficient of variation, which we might think of as an average percent estimating error.

How much of the variation in Y has been explained?

9. Variation

How much of the variation in the dependent variable can be explained by the variation in the independent variable?

Coefficient of Determination (R^2)

$$R^2 = \frac{SSR}{SST} \text{ or } 1 - \frac{SSE}{SST}$$

Sometimes considered a measure of the *strength* of the relationship between the variables.

R^2 is a measure of *correlation*, not *causation*, so don't just assume that an association implies causation.

Presumably we are developing a CER to explain the variation in our Y variable for the purpose of better estimating it. It seems reasonable then to ask how much of the variation in the Y variable that we have been able to account for.

The sum of squares total, or SST, represents the total squared variation around the mean of Y. The sum of squares regression, or SSR, represents the portion of the variation in the SST that we have accounted for in the equation. You might say SSR is the variation in Y that we can associate with the variation in X.

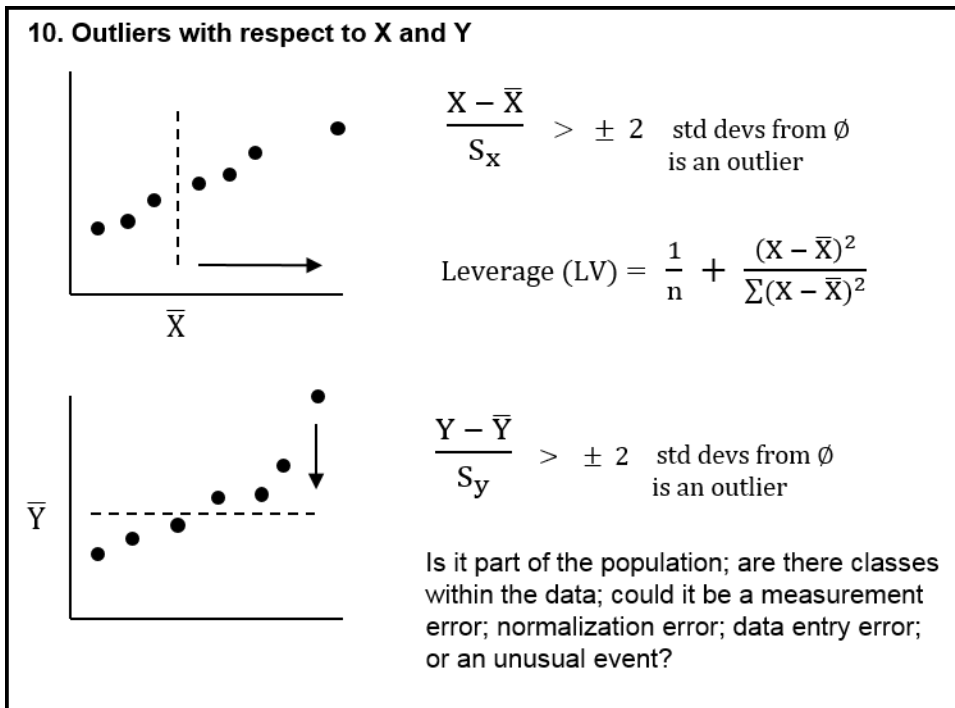
R squared is the ratio of the explained variation (SSR) to the total variation (SST). While calculated as a decimal value between 0 and 1, it's commonly expressed as a percentage between 0 and 100. If the equation bisected all of the data points, R squared would be 100%, meaning all of the variation in Y would have been explained by the variation in X.

R squared should not be used to imply causality, but rather used to support the causality supposition from the step where we identified what we believed to be causal cost drivers.

X and Y Outliers

We want to be aware of unusual values in the data for a number of reasons. It may indicate deficiencies in how we've normalized the data. It could be that there are differences in the data that either we were unaware of, or if we were aware of, we didn't know the impact those differences would have. And there is the concern of how that unusual value impacts our ability to fit the remaining data.

One test for what we call outliers, is to measure how far an X or Y value is from the mean of X or Y, in units of standard deviations. A more robust technique for the X variable is to calculate the leverage value for each X. Well-named, the higher the leverage value, the more leverage a particular X data point has on the slope and intercept of the equation. Outliers should be investigated prior to consideration for removal from the data set.



When does an X or Y become an outlier? It's subjective, but one convention is that when a value is more than plus or minus two standard deviations from the mean it is considered an outlier.

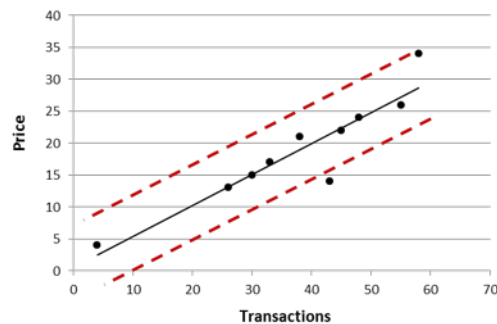
Leverage uses P (the number of coefficients in the equation) divided by "n" (the sample size). A multiplier of 2 or 3 times P/n is used as the point at which an X observation becomes an outlier.

Outliers with respect to the Predicted value of Y

Generally speaking, when we fit a line through the data there will be some variation between the actual Y values and the predicted Y values, which is to be expected. These differences are generally referred to as the “residuals”. Of concern is when particular observations vary significantly more between the actual and predicted values than what is typical for the majority of the data points.

One test for this type of outlier takes the difference (residual) between the actual value and the predicted value, and divides that difference by the standard error. As before, rules of thumb such as two or three standard errors are used to identify outliers. The shortcoming of this approach is that by dividing all the residuals by the standard error, it presumes the error is constant, which in fact it is not.

11. Outliers with respect to the predicted value (\hat{Y}) (prediction problems)



Questions:
Part of the population;
classes within the data;
measurement error;
normalization error;
data entry error; or
unusual event?
Missing a cost driver?
Wrong model form?

$$\text{Standardized Residual} = \frac{(Y - \hat{Y})}{SE} > \pm 2 \text{ std errors from } \emptyset \text{ is an outlier}$$

$$\text{Studentized Residual} = \frac{(Y - \hat{Y})}{SE\sqrt{(1 - \text{Leverage})}} > \pm 2 \text{ std errors from } \emptyset \text{ is an outlier}$$

The error between the sample equation and population equation increases as we move either direction from the center of the data (in terms of X).

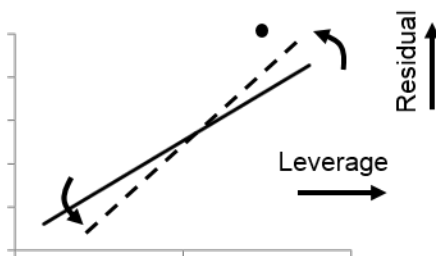
To compensate for this effect, an alternate calculation using the leverage value is employed to reflect the increased error that occurs the further a particular observation is from the center of the X data.

Influential Observations in the Data

Every data point will have some influence on the resulting equation. The concern is when a particular data point is having significantly more influence than the other data points in determining the slope and intercept of the equation.

12. Influential Observations

Is there a particular data point having significantly more influence on the slope and intercept of the equation than the other data points?



Questions:

Part of the population; classes within the data; measurement, normalization, or data entry error; or unusual event? Missing cost driver? Wrong model?

These influential data points tend to have high leverage values. Also, if we were to remove the data point from the data set and recalculate the equation, we would find that an influential observation tends to have a large residual when estimated using an equation that did not contain the data point.

The resulting effect is that an influential observation essentially pulls the line toward itself, and away from the general pattern in the data, which in turn reduces the accuracy of the equation for predicting.

One of the statistics used to identify influential observations is called Cook's Distance.

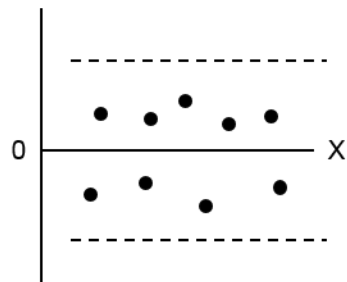
One means of dealing with an influential observation is to consider restricting the range of the data so as to not include the data point if it is not within the estimating range of interest.

Residual Analysis

13. Residuals

Did we properly fit the data (i.e. are the residuals randomly distributed about zero with a constant variance across the range of X values)?

random pattern; constant variance



$$\text{Residual } (e_j) = Y - \hat{Y}$$

Recall that a residual is the difference between the actual Y value and the predicted Y value. There is an expectation that if the data has been properly fit, the residuals (as measured on the X axis) would fall randomly about zero, and the dispersion would be fairly consistent (i.e. having a constant variance) as you look from left to right.

Residual plots are commonly used to assess the nature of the residuals, which requires us to note two things. One, because this is a visual assessment it is fairly subjective in nature. Two, because it is visual, residual plots are much more conclusive the larger the data set. Smaller data sets (again, small is subjective) may not be as compelling as to whether you have properly fit the data or not.

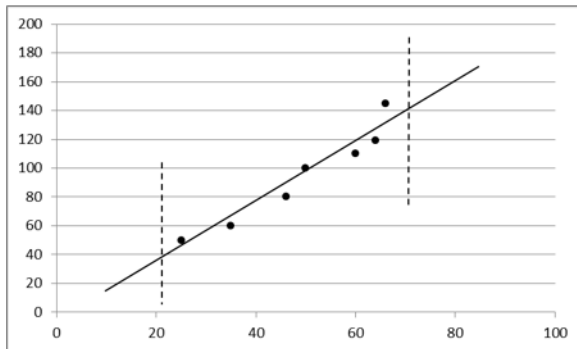
Curved patterns or V-shaped patterns in the residuals may suggest the need for a different type of model.

Estimating within the Relevant Range of the Data

It's always important to note where you are estimating with respect to the data. In other words, are you estimating near the center of the data set or at the ends of the data set? How well did the equation fit the data points in the range of the data for which you are estimating? Are you estimating outside the range of the data set, and if so, how far outside the range? Do you expect the relationship between the X and Y to continue outside the range of the existing data? Do you have expert opinion to support that?

Estimating in the “relevant” range of the CER

The range over which an estimating relationship is valid for use, roughly defined by the upper and lower bounds of the independent variable. The parameters of what is being estimating should be within the range of the data. Also, if there is correlation between the X values in the data set, then what is being estimating must exhibit the same relationships.



Should you adjust the predicted value from the equation to consider differences between the data set and that for which you are estimating, for example, differences in material, complexity, technology?

Keep in mind that adjustments will “bias” the results of the equation, and that the equation statistics such as the R squared and the standard error are no longer reflective of the adjusted number.