

# IS THERE A TREND IN THAT DATA? (PART 1 OF 2)

By James R. Chastain, Jr., PhD, PE, MPH



In our digital age we are awash in data. This is especially true in the engineering, environmental and public health fields. Whether the data results from routine compliance monitoring or special purpose studies, it is frequently desirable to infer whether a trend has developed...and if so what caused the trend. One of the most common statistical tools used for this purpose is linear regression. Linear regression can, in fact, be a very useful tool for organizing and evaluating data, however, the proliferation of personal computers and computational packages (ex. Excel<sup>®</sup>) have made it easy for inexperienced users to misapply this procedure. The purpose of this article is to discuss the basic concepts of linear regression and its application to trend estimation. At the outset it should be mentioned that this topic is fairly robust and this article sketches only a high level summary. One of my textbooks on the topic is 1396 pages long, so it's unreasonable to expect to cover all the subtleties in these few pages. Still, it is hoped that this overview will be helpful to the casual user and possibly spur an interest for further study.

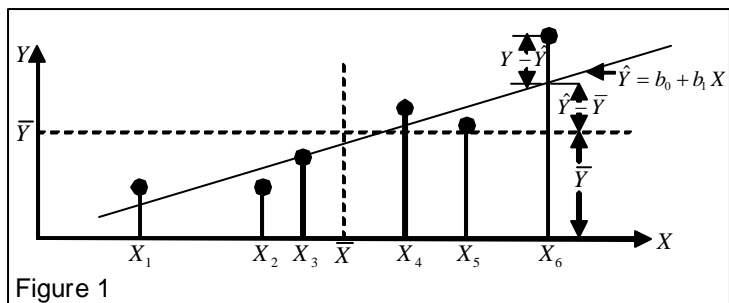
As a preliminary note, most technical professionals typically think of regression models in terms of developing a model that will allow prediction of future values. This is one use of regression, but it is important to be aware of other uses that may be of equal or greater importance. They are:

- Characterize a relationship between variables
- Control for the effects of other variables (i.e. what is the contribution of this particular variable when the other factors are held constant?)
- Determine the importance/priority among the different variables (i.e. which are the most important?)
- Assess the effects of interaction among the variables.

So an informed analysis of the data should begin with a firm concept of how the data should be collected and what the intended use of the regression equation will be. Understanding that regression analysis can provide numerous insights into the data structure and its effects can prove very helpful later in data interpretation.

## The Big Picture

While the technical details of regression analysis can become somewhat complex, the core concepts underlying linear regression are straightforward. To illustrate this a little more clearly, observe Figure 1. Fundamentally, the mathematical techniques are developed to determine the level of deviation from the grand mean (average) of all the Y-values of the data. This is noted as  $\bar{Y}$  on the figure. The individual data points are indicated by Y, and the computed y-value from the linear regression equation is  $\hat{Y}$ .



The greater the deviation from the grand mean, the more likely that there is a correlation between the predictor (independent; X-values) variables and the response (dependent; Y-values) variables. Thus, the interpretation is that unless there is a significant difference between the computed trend line and the grand mean ( $\hat{Y} - \bar{Y}$ ) no trend exists. Why is that? Because it indicates that no matter what the value of X, there is no difference in Y, thus there is no trend. It will probably be necessary to concentrate on the figure for a few minutes, but once the philosophical basis for the analysis is clear, comprehending regression output becomes more intelligible.

The mathematical techniques for determining the level of deviation from grand mean are based on (1) the concept of statistical variation (related to standard deviation) and (2) the correlation of sum of squares associated with each variable through the use of the Fisher distribution. To develop these concepts mathematically, certain assumptions are made in order to develop the computational procedure. For example, as you might expect, linear regression assumes that the relationship between the predictor and response variables are linear (as opposed to exponential, logarithmic, logistic, etc.). Mathematically, the model will be expressed in the following form for computing a

straight line that most students learned in high school:

$$y = a + bx$$

**Equation 1**

Because the most practical situations involve more than one predictor variable, the equation is generalized to the form:

$$\mu_{Y|X_1, X_2, \dots, X_p} = a + b_1x_1 + b_2x_2 + \dots + b_px_p + e$$

**Equation 2**

A very important aspect of this equation should be noted. In statistical notation,  $\mu$ , is generally used to denote the mean of values. Consequently,  $\mu_{Y|X_1, X_2, \dots, X_p}$  is to be read as saying that the regression line is the mean of y-values given the x-values of variables  $X_1, X_2, \dots, X_p$ . Therefore, understand that  $\mu$  will not estimate the range of individual values; it is an estimate of the mean. Misunderstanding this fact contributes to frequent misinterpretation of linear regression.

It might be helpful at the outset to sketch the general process to be used when developing a regression model (mathematical equation) to describe the data. Textbooks refer to this as “fitting” a model to the data. The objective is to use the fewest number of variables in the fitted model to adequately describe the response variable function. This is called the **parsimony principle**. The following five-step process applies to any regression technique (linear, exponential, logistic, etc.):

1. Perform an exploratory data analysis (get to know your data)
2. Decide which regression model to use (in this article: linear regression)
3. Fit the model to the data
4. Check the model
5. Document and interpret the model

The novice typically takes a data set, performs Step 3 by itself and then begins to draw conclusions from the resulting model. This can easily lead to poor decision-making.

### **Get to know your data**

The first step should be to perform a preliminary Exploratory Data Analysis (EDA). The objective of this effort is to get a feel for the raw data set and to look for obvious problems such as data entry errors, obvious outliers, etc. Examples of selected EDA might include: scatter-plot of data, list 10 largest/10 smallest values, mean/median, distribution tests/plots (dot plot, normal data plot, stem-leaf plot), etc. Since this article is considering linear data relationships, one purpose of examining the raw data is to get a sense of whether the data has a linear trend to it or not. It is also important to look for potential correlations between predictor variables. This may be a signal of collinearity problems, or non-linear relationships, which could reduce the validity of the model. There really is no substitute for having a “feel” for the data, including data collection procedures, plausible range of values, proper (and possible) units of measurement, data type, etc. Again, the purpose is to do as much preliminary data checking as possible prior to actually performing the regression analysis. This will save wasted effort down the road.

### **Decide what regression model to use**

If there was not a clear understanding of the likely data relationships prior to data collection, the next step is to use the EDA insights to select promising statistical models. This requires a knowledge of the underlying assumptions for the alternative statistical models as well as how to interpret and apply the results. For the purpose of this article, linear regression has been stipulated.

### **Fit the model to the data**

A formal presentation of this important step is quite involved. However, for the purpose of this article it will be assumed that a computer with a statistical software package is available which will handle the computational details. Each program has a specific data entry process so we won't consider that either. However, the computer analyses typically proceed in similar fashion.

The development of the equation for the line used to summarize the data trend (the regression line) is based upon an analysis of the sum of squares. The objective is to develop an equation for the line that results in the smallest sum of squares in deviation from the computed line. Referring to Figure 1 this is accomplished by minimizing the sum of

squares between the computed line and all of the data points ( $Y - \hat{Y}$ ) and the sum of squares between the computed line and the grand mean ( $\hat{Y} - \bar{Y}$ ). Each variable that is being considered in the analysis will contribute to the total sum of squares.

In many cases the analyst doesn't know which variables to include in the analysis and which to eliminate. Most computer packages give three options for trying to find the best combination of variables. These are usually classified as (1) forward selection, (2) backward elimination and (3) stepwise regression. There is no universally accepted "best" process, but many people use the stepwise process because it allows the computer to estimate the optimal solution. It is not a bad idea to try several selection options to see if the same result is reached.

There are a number of criteria to know which variables are important. For multi-variable regression the most common is the  $R^2$  value which is also called the **coefficient of determination**. This is analogous to the correlation coefficient,  $r^2$ , in simple two variable regression. Both parameters have a value between 0.0 and 1.0. The closer  $R^2$  is to 1.0 the more effectively the predictor and response variables are related. The adjusted  $R^2$ ,  $R^2_{adj}$ , is actually preferable when there are multiple variables, and especially if the number of data points is different with each variable.

Another common way of assessing the priority of a variable is to look at the "p-value". (The use and misuse of p-values has a long literary history, but it is beyond the scope of this article to elaborate on this topic.) By convention a variable is considered "significant" if it has a p-value of 0.05 or less. The smaller the value, the more "significant" it is deemed to be. This is an arbitrary number though and must be considered in the context of the overall analysis. One would find the value for each variable in the regression table in the computer printout. It is usually placed at the far right end of the regression analysis table.

There are numerous other techniques for fitting (selecting) the best model to the data, but these are common criteria. Other considerations might include such things as whether the line has an intercept or is fixed at the origin. The reason for this is that some situations occur where intercept values equaling zero don't have a physical meaning. For example, if one was developing a regression analysis of property damage versus wind speed, and there was an intercept value of \$50,000, the uncensored interpretation would be \$50,000 damage occurring at a wind speed of zero. As a general rule it is better to allow the intercept to float where the regression line is computed and restrict the use of the regression equation at low values of the predictor variable rather than force the line through the origin. This is an example of why one should not develop or use a regression analysis without understanding what the underlying assumptions were when the model was developed.

Up to this point no mention of the error value,  $e$  in the equation 2, has been made. Unless there is a perfect correlation in the data, the Sum of Squares analysis mentioned above will not balance perfectly. The part of the variance from the fitted regression line is called the Error Sum of Squares (SSE) and this is used to estimate how accurately the regression line estimates the actual data. This is a key component in the computation of  $R^2$  discussed above, as well as a number of the residual analysis procedures mentioned in the next section.

*Dr. Jim Chastain is the CEO and President of Chastain-Skillman, Inc. He has a Bachelor of Science in Civil Engineering (honors) and Master of Engineering from the University of Florida. He also has a Master of Public Health and Ph.D. from the University of South Florida. He is a registered Professional Engineer with over 30 years of experience and is a Diplomate of the American Academy of Environmental Engineers. He can be reached at (863) 646-1402 or jrchastain@chastainskillman.com.*