

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

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



The objective of this two-part essay is to provide a brief overview of the proper procedures to employ when using linear regression to summarize data and project trends. In the January issue of the Consultants Update, Part 1 of this article outlined the first three steps of a five-step process. In Part 2 of the article, the last two steps are described. As a reminder, the five-step process presented is:

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

Check the model (Regression Diagnostics)

Once the tentative statistical model has been developed, it is necessary to test the model through regression diagnostics. The purpose of this step is to confirm that the assumptions that were made are within acceptable range. Note that until the regression diagnostics have been performed, the analysis has not been completed. Regression diagnostics generally include *Residuals Analysis*, *Outlier Assessment*, and *Collinearity Assessment and Remedial Methods*.

Residuals Analysis

This is the primary means of examining the acceptability of the regression model. The key concept is that, if the model is reasonably describing the relationship, the residuals should form a random pattern centered around 0. From Part 1, recall that a residual is the difference between the actual data point value and the value that is predicted by the linear regression equation (i.e. the fitted line). In order to properly perform this analysis, it is necessary to account for or correct scale dependencies and the magnitude of e_i changes throughout the dataset. The primary methods to examine and evaluate residual values are (1) Studentized Residuals, (2) Jackknife Residuals, (3) Normal Probability Plot of the residuals. Again, the purpose of these exercises is to look for non-random patterns in the residuals, which would indicate non-linear relationships or outliers among the variables. In a sense, these tests are analogous to performing a second linear regression analysis on the residuals of each predictor variable...only this time it is desirable that no trend is detected.

Most statistical programs will provide an option to compute and list these values. Plotting the studentized or jackknife residuals is a convenient way of viewing the computed values for trends that might be difficult to observe in tabular format. Likewise, preparing a Normal Probability plot of the raw residuals is always useful, although it may be a little more difficult to evaluate the existence of a trend than when using the first two methods.

Outlier Assessment

Outliers are points that have much larger absolute values (larger or smaller) than the others in the data set...in other words they don't appear to "fit." The issue of outliers should be dealt with cautiously so as not to omit valid data points. They should only be deleted from the analysis if the data points are determined to be collected inappropriately, measured erroneously or processed improperly (e.g., malfunctioning data collection equipment, erroneous data entry, etc.) If the data points are scientifically or methodologically valid, they should not be deleted because they may indicate other factors or variability at work that would be suppressed if the points were deleted.

Outliers can be roughly identified by data plots, but are more specifically identified by looking at the studentized or jackknife residuals. Any residual value that is seen to be more than three standard deviations from the mean of residuals should be examined as a potential outlier. This is especially true if the outlier is

associated with one of the extreme values of X. This is a high leverage value and can significantly influence the regression model. In addition to the residual methods mentioned above, many computer packages compute *Cook's distance*. This is useful because it computes how much the regression coefficients are changed, by deleting the particular observation in question. A Cook's $d > 1.0$ should be investigated.

Again, deleting an outlier will typically make the mathematical correlation of the variables look better, but will not improve the estimation of the actual underlying relationship if the outlier is a valid, but extreme point. Therefore, every outlier must be examined in detail, and a valid justification for eliminating it must be stated before it can be removed.

Collinearity Assessment and Remedial Methods

Typically, as the number of independent variables increases, the risk of collinearity (also known as multiple collinearity or multicollinearity) occurs. Collinearity effects arise when the independent variables are not truly independent of each other. In other words, being collinear, they are related to each other. If this issue is not recognized and resolved, the result tends to be large variations in the linear coefficients which means that resulting equation is unreliable. Also, the estimated standard deviations of the coefficients become quite large, so the confidence intervals of the coefficients are not helpful.

Statisticians have developed a number of ways of detecting this occurrence, but the most common test to identify it is by computing the Variance Inflation Factor (VIF). Without going into the mathematics of the test, when the VIF for a variable is greater than 10 it is generally considered a red flag warranting further investigation. The VIF is normally a computation option available on most statistical software packages. In most cases, by deleting or inserting predictor variables by trial and error and observing the VIF, the standard error and parameter confidence interval allows the analyst to arrive at the optimum combination of variables.

If problems persist after the regression diagnostics have been completed, several alternatives are available to recover. The following measures should be considered.

1. Examine outlier effects. Are data points valid? Should they be eliminated?
2. Abandon the regression model and develop a more appropriate model (e.g., Non-linear).
3. Transformation. This should be used with caution remembering that any regression coefficient has properties related to the transformed observations, not the original ones.

Document and interpret the model

As mentioned in Part 1 of this article, it is not uncommon for most of the attention in a statistical analysis to be focused on fitting the regression model to the data (Step 3). While this is without question an important step, hopefully, one of the take-away messages of this article is that the other four steps can be just as important. Typically, documentation and interpretation of the model (Step 5) tends to get the least amount of attention, and yet it ultimately can be the most important. No matter how suitably the math was performed in fitting the model, if an improper interpretation of the results follows, then the whole effort was unproductive. Also, if there is not proper documentation of the assumptions and tradeoffs used in developing the fitted equation, future use of the results may be applied inappropriately.

First, when documenting the analysis, make sure that you note whether the data is classified as experimental (independent variables under the control of the experimenter) or observational (neither response nor predictor variables are controlled by the experimenter). A statistical test that leads to the conclusion that $\beta_1 \neq 0$ does not necessarily establish a cause and effect relationship between the predictor and response variables. With *non-experimental data* (observational), both the X and Y variables may be simultaneously influenced by other variables not in the regression model, which can lead to erroneous inferences. On the other hand, with *controlled experimental data* there is often good evidence for a cause-effect relationship. Recall also that in either case the response variable will always be continuous, while the predictor variables can be either continuous or categorical (discrete).

One of the attractive features of the linear regression model is that the interpretation of the results is straightforward and easy to understand. Once the fitted equation has been developed, each predictor variable will have a coefficient. This coefficient may be interpreted in the following way: “*by increasing*

the predictor (X) variable by 1 unit, the response variable (Y) will increase (or decrease if negative) by the value of the coefficient, when all other variables remain constant.” This interpretation is applied to each predictor variable in turn, so it is possible to easily see which variables have the greatest effect on the response. This observation can be of great value when trying to prioritize actions based on the analysis.

Two general cautions should be noted with regard to using regression as a prediction tool. One is related to temporal extrapolations, and the other is related to data extrapolations. More specifically, regression analysis is frequently used to make inferences about the future. It is important to remember that the validity of the regression application depends upon whether basic causal conditions in the period ahead will be similar to those in existence during the period upon which the regression analysis is based. Secondly, caution is warranted when inferences are made involving predictor values that lie outside the range of observation. In this case, residual/error checking cannot be done to evaluate model validity. Thus, these extrapolations have an additional inherent error risk.

In sum, linear regression is a useful tool for organizing and developing linear relationships in a wide range of data types. Computer software makes the evaluation of even large volumes of data convenient and accurate. Most people find the results of this analysis easy to communicate and understand. However, this article and the previous one attempt to identify some of the major assumptions associated with the technique, along with some of the pitfalls to avoid when developing the model. Other subtleties and analytical features exist, but are beyond the scope of this article. The references below can be consulted for a more complete presentation of these issues.

References:

Neter, J., M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. (1996). **Applied Linear Statistical Models** (4th Ed.). Richard D. Irwin, Inc. Chicago, Ill.

Kleinbaum, D.G., L.L. Kupper, K.E. Muller, and A. Nizam. (1998). **Applied Regression Analysis and Multivariate Methods** (3rd Ed.). Duxbury Press. Pacific Grove, CA.

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.

© 2009 Chastain-Skillman, Inc. This article is taken from the 2nd quarter 2009 issue of Consultant's Update, a publication of Chastain-Skillman, Inc.