The Treatment of Outliers and Influential Observations in Regression-Based Impact Evaluation

Jeremy M. Schutte and Daniel M. Violette, XENERGY Inc.

The treatment of outliers and influential observations in multivariate regression analysis is becoming a pressing issue as more utilities move to regression-based analysis in the evaluation of DSM programs. Because the treatment selected for outliers and influential observations can significantly affect the evaluation outcome, this issue has gained the attention of evaluators and regulators alike. This is a complex issue and is just beginning to be explored in current evaluation literature. This paper presents an overview of some current problems and a review of some options available to evaluators seeking to resolve these issues.

The current debate centers around two questions: How to identify outliers and influential observations, and how to treat outliers and influential observations? The first question can be answered by discussing the various definitions of outliers—observations out of range, more than a specified number of standard deviations from the mean, or with pre to post changes in usage greater than a certain limit, and standard regression diagnostics—studentized residuals, DFBETAS, DFITS, and the hat matrix. The second question is addressed in a discussion of treatment methods—removing or downweighting outliers with identified data quality problems, removing or downweighting outliers falling outside predetermined bounds, removing or downweighting observations with calculated regression diagnostics outside a predetermined range, or a combination of several methods. Lastly, we will describe the pros and cons of robust regression techniques and their ability to mitigate the leverage some observations have under OLS regression and the use of bootstrapping techniques as an alternative solutions to the removal of observations from the analysis.

Introduction

In the broadest sense, the term "outlier" describes an observation that is out of the norm. Bollen and Jackman (1990) describe outliers as "observations that are distinct from most of the data points in the sample." What defines an outlier as "distinct" in each instance is a function of the metric used and the context in which the observations are examined. In a DSM evaluation context, if you were to examine the distribution of usage levels, scatterplots of gross energy change vs. predicted savings, or number of standard deviations from the mean, the observations in a single sample classified as outliers would vary with each perspective.[1]

Another measure of an observation's "distinctness" is its influence on modeled findings. An influential data point is an observation that "has a demonstrably larger impact on the calculated values of various estimates (coefficients, standard errors, t-values, etc.) than is the case for most of the other observations" (Belsley et al. 1980, p. 11). Since not all outliers are influential observations, the influential observation can be considered a special case of outlier.

How to Identify Influential Observations?

Where the general case of outlier is defined by the values or statistics generated directly from the data—means, standard deviations, change over time—regression diagnostics are used to determine the "influence" of observations in statistical models. Again, depending upon the metric used and the context, in this case the functional form of the model, which observations are determined influential will vary. In the industry literature, Violette et al. (1991) presents three diagnostics that can be used to measure influence-the hat matrix, studentized residuals, and DFITS. This paper will discuss these three, plus partial regression plots and DFBETAS. These diagnostics can be produced by most statistical packages on the market today. The reader's attention should be directed to Violette et al. (1991), Belsley et al. (1980), and Bellman and Jackson (1990) for detailed descriptions and specific characteristics of each diagnostic method. This paper will focus on the practical aspect of their application. Often a combination of diagnostics will be used depending upon the evaluator's experience and preferences. A brief description of each diagnostic follows.

Partial Regression Plots

The partial regression plot provides an easy means for the visual inspection for leverage points in multivariate specifications. In the case of a single explanatory variable, a scatter plot of Y against X can be performed to visually inspect for the presence of outliers. In the event there is more than one explanatory variable, the use of partial regression plots can be used in a similar manner.

Consider the regression equation:

$$Y=b_0+b_1X_1+b_2X_2+\epsilon \tag{1}$$

The partial regression plot for X_1 is constructed by first regressing Y on X_2 and X_1 on X_2 and graphing the resulting two sets of residuals. A similar plot for X_2 is constructed by regressing Y on X_1 and X_2 on X_1 and graphing. The examination of these residual plots gives a visual indication of outliers.

Hat Matrix

Given the general linear regression model:

$$Y=X\beta+\epsilon \tag{2}$$

where Y is an nx1 vector of values of the dependent variable, X is an nxp matrix of the explanatory variables, ß is a pxl vector of parameters to be estimated, and e is an nx1 vector of disturbances. The ordinary least squares estimate of ßis:

$$b=(X'X)^{-1}X'Y \tag{3}$$

The fitted values of the dependent variable are therefore:

$$\hat{Y}=Xb=X(X'X)^{-1}XY=HY \tag{4}$$

H is referred to as the "hat matrix" because it transforms Y into $\hat{Y}$—the actual values into predicted values.

The diagonal elements of H, h_i, *give* the "leverage" of Y on $\hat{Y}$. The values of h_i are bounded by 1 and 1/n, and the closer to 1/n, the less the leverage of Y on $\hat{Y}$. In addition, it has been shown that the sum of the *hi's are* equal to the number of X variables, p. Therefore, observations associated with an $h_i > 2p/n$ are often considered suspect.

Studentized Residuals

Another method that can be used to determine influential observations is to compare the residuals for each observation. With the equation:

$$var(e_i) = \sigma(1-h_i) \tag{5}$$

it can be discerned that those observations with the greatest leverage, hi, have the smallest variance. Thus, it is best to standardize the residuals to account for these different variances. The studentized residual is defined as:

$$e_i^* = e_i/\sqrt{(s_i^2(1-h_i))} \tag{6}$$

where s_i^2 is the sample estimate of the disturbance value when the ith case is removed.

Studentized residuals have uniform variances and in practical applications are assumed to be distributed close to a t-distribution. Frequently, a studentized residual above 1.96 indicates an observation with high influence. (It must be noted that not all observations with high influence will have a high studentized residual. A highly leveraged observation that pulls the regression line through it will have a low studentized residual.)

DFITS

The hat matrix and studentized residuals are designed to detect observations with high leverage and high residuals, respectively. However, these measures of influence won't always be in agreement. The measure DFITS was designed to explicitly account for the leverage and residual magnitude of each observation. The formula for DFITS:

$$DFITS = \sqrt{(h_i/(1-h_i))}/(e_i/\sqrt{s_i^2(1-h_i)}) \tag{7}$$

This explicitly combines the above two methods so that the DFITS is affected by leverage points and large studentized residuals. Belsley et al. (1980) suggest that a rough rule is that observations with a $DFITS_i$ greater than $2\sqrt{(p/n)}$ should be investigated.

DFITS can be interpreted as the scaled change in the fitted value when that observation i is removed.

$$DFITS_i = \frac{\hat{y}-y_i}{\sqrt{(s_i^2 h_i)}} \tag{8}$$

DFBETAS

Where DFITS$_i$ takes into account changes in all of the regression coefficients that result when a single observation is removed, DFBETAS provides a measure of how individual coefficients change when a case is omitted.

$$DFBETAS_{ij} = (b_j - b_j(i))/\sqrt{(s^2(i)(X'X)^{-1}_{ij}} \tag{9}$$

The numerator refers to the difference between the regression coefficient for variable *j* estimated for the full sample and the regression coefficient for variable *j* with observation *i* removed. The denominator is the estimated standard error of the *jth* regression coefficient.

Large positive and negative values of DFBETAS$_{ij}$ indicate observations that lead to large changes in the jth regression coefficient. Belsley et al. (1980) suggested a size-adjusted cut-off of $2/\sqrt{n}$. An alternative is to use the higher cutoff of 1 which will identify observations that shift the regression coefficient estimate at least one standard error.

Summary of Diagnostics

Table 1 summarizes cut-off levels for each of these regression diagnostics. A range of cutoff values is presented. The low values are the widely accepted cutoff levels. Bollen and Jackman (1990) recommend also using the high cutoff levels. If the low cutoff is found to flag a high percentage of the sample and inspection reveals no apparent problems with them, then the high cutoff may then be used.

This section drew heavily upon the work of Belsley et al. (1980) and Bollen and Jackson (1990), so the reader should consult the original texts for a more detailed discussion of the issues related to the application of these diagnostic measures.

How to Treat Outliers and Influential Observations?

There is no consensus regarding the “right” diagnostic methods to use, or the specific cutoffs to use. It would be difficult if not impossible to create an absolute set of methods and criteria for outlier analysis that would lend itself equally to every situation. This section presents the steps taken in the inspection of outliers, no matter which diagnostic or set of diagnostics used and contains brief descriptions of actual evaluation applications of diagnostic methods.

Research the Observation

In the event that a given observation is flagged as an outlier, the first step is to verify that the data for that observation is not in error. In the case of energy consumption, usage levels can be verified by contacting the customer directly, recalculating bills by hand, checking that all meters affected by the DSM activity are included, etc. Similar checks can be made on program participation data, survey data, or customer information.

Table 1. Summary of Diagnostic Cutoffs

Diagnostic	**Cutoff**	
	Low	**High**
Partial Regression Plot	Visual inspection	
Hat Matrix (h_i)	$2p/n$	$3p/n$
Studentized Residuals (e_i^*)	$\alpha/2$	$\alpha/2n$
DFITS$_i$	$2\sqrt{(p/n)}$	$\sqrt{p}$
DFBETAS$_{ij}$	$2/\sqrt{n}$	1
p is the number of estimated parameters and *n* is the sample size.		

In the event that the customer data are correct, further investigation into possible reasons for the outlier classification, in this example, deviant energy usage, are warranted. An investigation of why a participant is an outlier may reveal model mispecification. As an example, a set of residential customers with abnormal energy usage may be found to have a swimming pool, and the addition of a variable representing the presence of a swimming pool to the regression model will contribute to its ability to explain energy usage for these customers. Bollen and Jackman (1990) also suggest that the presence of outliers may indicate the need to transform key variables (e.g., to lognormal or quadratic terms, etc.) to better model the activity in question based upon existing knowledge or research conducted upon the behavior being modeled.

Other Treatment

A thorny debate occurring today centers around the treatment of outliers and influential observations. There are many who would like to see *a priori* methods for identifying outliers with clearly defined methods for treating them—usually removing or downweighting them. In practice, these methods have been frequently applied with apparent success. However, as described above, the established cutoffs are “rough” and are recommended to be adjusted to fit the needs of each analysis.

Heated discussion arises in situations where different cutoffs result in significantly different model findings. In the evaluation context, the magnitude of the final realization rate or adjustment factor can have significant revenue recovery impacts in a shared savings context. Furthermore, the “stability” of the findings as defined by their sensitivity to the removal of outliers will affect the ease with which interested parties and regulatory agencies will accept evaluation results. Another related problem identified in practice (see Example A) and in the literature (Bollen and Jackman 1990) is that successive screens of outliers, where outliers are identified and removed and the process is repeated on the remaining observations, will continue to flag observations as outliers. Much like peeling an onion, the removal of each “layer” produces a different parameter estimate and reveals another layer to be peeled back until potentially there are no remaining data—only a trail of disparate parameter estimates.

Another issue discussed in the previous section is that not all outliers are bad. To mechanically cut out all outliers beyond set bounds, the evaluator must assume that the model specification is correct and the outliers do not represent information that would contribute to the explanatory ability of the model. For example, in a model with a dummy variable for the presence of air conditioning that is 1 for a handful of observations, these observations will be flagged as influential observations based upon DFITS because of their impact on that coefficient. However, they should not be removed from the model since they enhance its ability to explain the variation in customer to customer energy usage.

The debate tends to center around the issues of what screening diagnostics to use and what cutoffs to use. In regard to the first question, the evaluators in Example B use a battery of diagnostics and remove those observations exceeding the prescribed cutoff levels. As to which cutoffs to use, the standard cutoffs described can be applied while keeping in mind that, when faced with high levels of data attrition, the cutoffs may be adjusted in accordance with the judgment of the evaluator (whether the observations flagged are confounding or adding information to the analysis). However there are often cases where interested parties cannot agree upon an *a priori* diagnostic to be used nor the cutoff levels to be used. Example A is such a case. In both cases, the interested parties must have a sound understanding of the potential bias that the selected diagnostic methods will introduce into modeled findings. In Example B, the parties involved decided that the application of preset diagnostic cutoff levels was appropriate given the data and regression model, while the parties in Example A felt that such an approach would not provide an adequate treatment of the data or modeled process.

Robust Estimation Methods

Some of the potential problems arising from the presence of influential observations in multivariate statistical models arise from the manner in which ordinary least squares (OLS) estimation fits a line through data. OLS is sensitive to outliers and influential observations.

The discussion of robust regression techniques begins by asking: If the tool (OLS) is particularly sensitive to the presence of outliers, why not find another? There are alternative estimation methods that can be easily performed with most statistical packages. There are even more alternative estimation methods than outlier diagnostic methods, with even less agreement upon a preferred approach. An introduction to these methods is presented in Berk (1990). A summary of robust regression methods is not possible here. Instead, we will present some justification an evaluator may have for further researching these methods. The application of these methods may be particularly appealing to the evaluator who has reservations regarding the deletion of observations from the sample. Or, even if there is no inherent disaffection for the deletion of observation but interested parties cannot agree upon the diagnostic nor the cutoff(s) to be used, these alternative estimation methods may be of interest. Example C presents a brief description of the application of a particular robust estimation method in an evaluation. In general, these methods provide parameter estimates in a manner that mitigates the influence of influential observations. In Example C, this is achieved by performing iterative regressions on the model that weights each observation by a function inversely proportional to its residual. In this manner, for each iteration, the larger the residual in the previous iteration, the less weight that observation has in the current regression run. This is but one example of a multitude of alternative modeling approaches to OLS. The disadvantages of these methods are their high complexity, variety, and potential difficulty in explanation and acceptance. However, since they do minimize the influence of outliers, they may be applied without the deletion of sample observations.

Bootstrapping

Like the robust regression methods, bootstrapping allows the evaluator to estimate parameters without removing observations from the sample. The premise behind bootstrapping is that multiple samples from a population sample will generate parameter estimates that approach those of the population and that these bootstrapped estimates have definite statistical properties. These methods are fairly computationally demanding, and require programming in most statistical packages. Stine (1990) presents a good introduction to bootstrapping methods. The reader is referred there for descriptions of this method.

In practice, bootstrapping requires the analysis of interest to be run on B random sub-samples of the population sample of size n where n < N and calculating an "average" parameter estimate and variance. Since n is less than the total sample size, N, the effect of outliers will not be felt in all B sub-samples. Thus, the overall impact of outliers will be dampened in the bootstrapped findings. Intuitively, since you are drawing numerous random subsamples from the sample in question, the impact of outliers will not be felt in all versions of a statistical model. In this way, the bootstrapped findings may better represent the general trend for the sample. However, if the outliers are caused by problems with the underlying data, bootstrapping findings will still be affected by these data problems. These methods do require some assumptions regarding the sampling characteristics of the data. And, as with the robust regression methods, due to the complexity of this method, there is added difficulty in explaining such methods. It is worth noting that Stine (1990) further recommends using both bootstrapping and robust regression methods to best deal with outliers and influential cases.

Conclusion

As more and more evaluations are conducted using multivariate regression methods, and computing power and statistical software become widely available, evaluators and regulators alike need to become conversant with the issues surrounding regression diagnostics. The industry needs to develop a sense of what is accepted and reasonable. This isn't to say that strict protocols are required that define which methods and what cutoffs should be used. But rather, as we hope this paper conveys, emphasis should be placed on knowledge of and experience with these methods, and thorough understanding of the evaluation "context." Unfortunately for those who prefer black and white demarcations, it is this understanding of context that in many situations may guide decisions regarding what is out of the norm and "distinct" for each evaluation.

Examples

Example A. A multivariate statistical billing analysis was performed for energy savings in a fairly heterogeneous customer sector. In the process of performing the evaluation, it was found that the removal of certain outliers, as defined by their usage levels, resulted in changes in estimated realization rate. The examination of residual plots and DFITS and DFBETA diagnostics indicated the presence of influential observations. However, the realization rate would fluctuate significantly in response to small changes in the value of the selected DFITS or DFBETA cutoff. Furthermore, successive applications of these diagnostic screens presented more outliers and equally varying realization rates. The realization rate estimated did not converge in successive applications of diagnostic screens. Evaluators and interested parties could not come to agreement regarding the appropriate solution to this issue and continue to research and debate the appropriate course of action.

Example B. Evaluators applied a battery of diagnostic screens. Partial regression plots, hat matrix, studentized residuals, DFBETAS, and DFITS were generated for each observation. All observations outside of prescribed cutoff levels for each diagnostic, about 15%, were eliminated from the analysis and the model was re-run. The re-run parameters were quite close to those found with all observations present. The full-sample finding were presented and the regression minus-outliers model was filed to present an indication of model stability.

Example C. In a conditional demand model, these evaluators applied robust regression methods to the data to test the sensitivity of the model to outliers. Iterative weighted least squares models were run where observations were weighted by a function of the inverse of their residual and zero weight for extreme outliers. The results for the robust regressions were compared to the original model to test overall sensitivity to outliers.

Endnote

1. Pigg and Blasnik (1993) make a distinction between *within facility* outliers, frequently encountered in the analysis and aggregation of several months of customer billing data, and *across facility* outliers as discussed in this paper in the context of comparing different observations. A discussion of within facility outliers is essentially a discussion of how to "clean" and process customer billing data and is not dealt with here. However, it must be kept in mind that customers with within facility outliers can also become across facility outliers depending upon the methods selected for aggregation or normalization of usage data. Thompson (1993) provides an interesting application of PRISM to process and screen residential billing data. This paper assumes that an annualized and/or normalized usage has been provided as an input into a statistical analysis.

References

Belsley, D. A., Kuh, E., and Welsch, R. E. 1980. Regression diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley, New York.

Berk, R. A. 1990. "A Primer on Robust Regression." *Modern Methods of Data Analysis,* edited by Fox, J., and Long, J. S. Sage Publications.

Bollen, K. A., and Jackman, R. W. 1990. "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases." *Modern Methods of Data Analysis,* edited by Fox, J., and Long, J. S. Sage Publications.

Pigg, S., and Blasnik, M. 1993. "Dealing with Outliers in Impact Evaluations Based on Billing Data." *Proceedings from the 1993 Energy Program Evaluation Conference,* pp. 188-198.

Stine, R. 1990. "An Introduction to Bootstrap Methods - Examples and Ideas." *Modern Methods of Data Analysis,* edited by Fox, J., and Long, J. S. Sage Publications.

Thompson, Mark. 1993. "PRISM Sample Design for Increasing Data Retention." *Proceedings from the 1993 Energy Program Evaluation Conference,* pp. 188-198.

Violette, D., Ozog, M., Keneipp, M., and Stern, F. 1991. *Impact Evaluation of Demand-Side Management Programs, Volume I: A Guide to Current Practice.* EPRI.