Influential Observations and Inference in Accounting Research

Andrew J. Leone

University Miami

Miguel Minutti-Meza

University of Miami

Charles Wasley

University of Rochester

Very Preliminary

April 2012

I. Introduction

This study examines the statistical problems related to the presence of influential observations or outliers in financial data. The objectives are: (1) to summarize the various methods used to identify and adjust for influential observations in the accounting literature; (2) to assess the effectiveness of such methods in commonly used research designs; (3) to assess the effectiveness of new methods proposed by the statistics literature; and, (4) to provide guidance for future studies that need to identify and correct for the presence of influential observations.

Although a precise definition is hard to come by, *influential observations* are generally considered to be data points that have a large impact on the calculated values of various estimates (e.g, mean, regression coefficients, standard errors, etc.), while *outliers* are data points that are located unusually far from the mean of a sample. Influential observations or outliers are not necessarily problematic, but researchers should be aware of their potential effect on statistical inferences.[1] Belsley et al. (1980, p. 3) remark:

> "Unusual or influential data points, of course, are not necessarily bad data points; they may contain some of the most interesting sample information. They may also, however, be in error or result from circumstances different from those common to the remaining data."

This remark highlights an important issue for accounting researchers. Unusual data points may result either from erroneous data or from unusual events affecting a subset of the data. In the first case, unusual data points occur randomly in a dataset and are caused by measurement error such as coding errors or wrong data-collection procedures.[2] If outliers

[1] Throughout this study we refer to outliers, unusual, extreme, or influential observations alternatively; however, in Section 3 we provide criteria to identify the types of observations that may affect statistical inference.

[2] Kraft et al. (2006, p. 307) provide an example of a data error in CRSP: "…consider the case of Smith Corona, which filed for bankruptcy in 1996 and delisted in May of 1996 when the stock closed at $0.375. In February of 1997, the firm emerged from bankruptcy, and as part of the reorganization, the common stock at the time of the

result from measurement error, researchers can correct, discard, or adjust them to fit between a lower and an upper bound. In the second case, unusual data points likely indicate areas where a certain theory is not valid or the presence of infrequent events that generate extreme outcomes. For instance, a company deciding to sell a business taking large restructuring and asset impairment charges would have unusually low total accruals, or an exploration company reporting a large unexpected oil discovery would have unusually large stock returns.[3] In the OLS framework, ignoring the underlying causes of these unusual observations is a form of model misspecification and a potential correlated omitted variables problem. This problem may cause wrong statistical inferences if the unusual events are correlated with both the dependent variable and the variable of interest. In this case, researchers should be very cautious in generalizing the results of their initial statistical analyses. Moreover, researchers should attempt to mitigate the effect of these outliers by modifying their model to include additional variables that capture the effect of unusual events, or by implementing econometric methods that are robust to the presence of outliers.

Our literature review shows that the procedures used to identify and adjust for outlier observations are scattered and inconsistent throughout the accounting literature. In the large sample of archival studies we reviewed, most studies use linear regression but only 68 percent mention that outliers are addressed. The most common solutions to address the presence of outliers are winsorization and truncation. We found that 55 percent of those

bankruptcy was canceled and shareholders of record received one warrant to purchase shares in the new company for every 20 shares previously held. The warrants had an estimated value of $0.10 or one half a cent per original share. When the new shares began trading in February 1997, CRSP used the new trading price of $3.12 as the firm's delisting price. As a result, the calculated delisting return on CRSP is over 700%, when it actually should be closer to −100%. More importantly, using CRSP's delisting return produces a BHAR of 2,177% when it should be −135%."

[3] Kraft et al. (2006, p. 333) provides examples of these two events, in 1996 Tyler Technologies Inc. had negative 72 percent total accruals, scaled by total assets; and in 1998 Triton Energy Ltd. had 203 percent buy and hold abnormal returns.

studies winsorize outliers, but only 33 percent winsorize both the dependent and independent variables; while 40 percent truncate outliers, but only 30 percent truncate both the dependent and independent variables. In addition, we found that 22 percent of studies addressing the problem of outliers applied a general winsorization or truncation cut-off for most variables and an *ad-hoc* cut-off for a subset of variables.

Winsorizing or truncating data alters legitimately occurring extreme observations that can contain information that improves estimation efficiency. Furthermore, accounting academics have diverging opinions about the best way to deal with outliers in studies examining market efficiency (e.g., Kothari et al. 2005; Kraft et al. 2006; and Teoh and Zhang 2011). In particular, these studies have highlighted potential problems resulting from deleting outliers based on extreme values of the dependent variable (referred to as *ex-post* truncation of stock returns). Despite these concerns, roughly half of the studies in our literature review winsorize or truncate stock returns used as dependent variable.

Taken together, these issues motivate us to assess the effectiveness of various methods to address the problem of outliers. We use simulations to compare the incidence of Type I and Type II errors after winsorizing, truncating or using robust regression when the data has (1) no outliers, (2) random outliers, and (3) outliers resulting from unusual events that are correlated with both the dependent variable and the variable of interest. Winsorizing and truncating rely on modifying the properties of the data, while the various forms of robust regression rely on a transformation of the OLS framework to maximize the sum of weighted errors. Robust regression takes into account deviations from normality in the error terms in estimating the parameters of the OLS model, thus increasing the accuracy of the inference.

Recent research in statistics is advancing fast in this area and there are a number of new robust methods available to address the problem of outliers.

We find that robust regression based on MM-estimation has the most desirable characteristics in treating outliers relative to other approaches found in the literature. First, when outliers are uncorrelated with the independent variables, it yields parameter estimates that are unbiased and identical to those estimated under OLS. Second, it is most effective at mitigating bias induced by a correlation between outliers and the independent variables. Third, MM-Estimation has 95 percent efficiency relative to OLS meaning that Type II errors are unlikely to be a consequence of the estimation procedure.

In contrast to robust regression based on MM-estimation, alternatives commonly found in the accounting literature are much less effective and some methods actually induce bias. When outliers are random, winsorizing at the top and bottom 1 percent yields unbiased estimates, but winsorizing is ineffective at mitigating bias induced by outliers that are correlated with the variable of interest. In addition, the practice of winsorizing independent variables but not the dependent variable biases the coefficients away from zero.

As reported in Kothari et al. (2005), simple truncation at the extremes generates a downward bias on coefficients. None of the other alternatives we evaluate generate a bias when outliers are randomly distributed. When outliers are correlated with an independent variable, truncation exhibits less bias than winsorizing (or doing nothing) but not to the extent that robust regression does.

In summary, our study provides guidance to accounting researchers considering alternative approaches to mitigating the influence of outliers. Robust regression based on MM-estimation yields estimates that are resistant to random outliers and significantly reduces the bias created by outliers that are correlated with the independent variable of interest. Furthermore, this procedure is highly efficient, exhibiting 95 percent efficiency relative to OLS, and not subject to biases induced by winsorization or truncation.

II. Literature review

2.1 Review of studies published between 2006 and 2010

We conducted a large literature review, including 857 studies published between 2006 and 2010 in the following journals: Contemporary Accounting Research, Journal of Accounting Research, Journal of Accounting and Economics, Review of Accounting Studies, and The Accounting Review. The studies in our review span a variety of areas, such as auditing, properties of analysts' forecasts, management compensation, earnings management, conservatism, tax, disclosure, and earnings-returns associations. As illustrated in Table 1, Panel A, 69 percent (590) of the studies are archival. The remaining 31 percent are split between 12 percent (101) analytical, 12 percent (106) experimental, and 7 percent (60) discussion and review studies.[4]

We searched the body, footnotes and tables of each archival study for any discussion of the treatment of influential observations or outliers. Surprisingly, only 68 percent of the archival studies (404 of 590) mention the presence of outliers in the data or describe any procedures to deal with influential observations.[5] The most common solutions to address the presence of outliers are winsorization and truncation, 88 percent of studies use one of these procedures separately or combined. Table 1, Panel A, shows the breakdown of archival studies using winsorization, truncation and other procedures.

[4] Studies that include both an analytical model and archival empirical tests are classified as archival.

[5] This percentage of studies not mentioning outliers ranges from 20% (RAST) to 37% (JAR and CAR).

Table 1, Panel B, shows the breakdown of studies that use winsorization to deal with outliers in continuous variables. Winsorization is the most common procedure used to deal with outliers, 55 percent of the studies dealing with outliers (221 of 404) winsorize at least one variable. This strategy modifies the original dataset, imposing and upper and lower bound on outliers by setting extreme data points to be equal to a specified percentile of the distribution of values for each variable, most studies in our sample use the top 99 and the bottom 1 percent. From the 221 studies that apply winsorization, 151 winsorize the dependent variable, 202 winsorize at least one dependent variable, and 132 winsorize a combination of dependent and independent variables. It is particularly concerning that we find that 29 studies use a general winsorization cut-off for most variables and an *ad-hoc* cut-off for a subset of variables. Examples of these *ad-hoc* cut-offs are: setting extreme values of discretionary accruals as a percentage of total assets to be equal to plus and minus two, winsorize all variables except the logarithm of total assets, and setting extreme values of effective tax rate to be equal to one and minus one.

Table 1, Panel C, shows the breakdown of studies that use truncation to deal with outliers in continuous variables. Truncation is the second most common procedure used to deal with outliers, 40 percent of the studies dealing with outliers (161 of 404) truncate at least one variable. This procedure also imposes an upper and lower bound to the data, but it discards observations beyond the specified cut-offs. The cut-offs can be set based on percentiles, similar to winsorization, or based on the influence of each observation on the OLS fit, for example, eliminating observations with large residuals. The latter procedures rely on estimating a regression model and eliminating those observations with large residuals. From the 161 studies that truncate extreme observations, 143 truncate the dependent variable, 139 truncate at least one dependent variable, and 121 truncate a combination of dependent and independent variables. In our sample, there are 59 studies that truncate observations based on a rule different than a percentile cut-off. For instance, truncate earnings scaled by total assets at plus and minus three, or truncate firms with stock prices less than five.

We find that the 13 percent of studies that do not use winsorization or truncating employ diverse procedures, including: rank dependent and/or independent variables, transform variables using logarithms, and use various forms of robust regression (e.g., median regression, least trimmed squares, and MM-estimation).

Overall, we find that the procedures used to identify and adjust for outlier observations are scattered and inconsistent throughout the accounting literature. Between the studies that use winsorization and truncation, we identified 88 studies that use an *ad-hoc* rule for a subset of variables. Furthermore, we identified 38 studies in our sample of archival studies in which the procedures employed were unclear. For example, some studies note that they use the procedures from Belsley et al. (1980) but they do not describe exactly how were they performed; or other studies note that they use standardized residuals or Cook's D to truncate observations with large residuals, but do not include the cut-offs employed. Such inconsistencies make it nearly impossible to replicate these studies.

Finally, we note that our classification of studies in separate categories is subject to some limitations, given that (1) the procedures used to deal with outliers are not mutually exclusive, (2) some studies have dichotomous dependent variables, (3) most studies have more than one regression model. Nevertheless, we believe that our review shows that there is need for greater consistency in the treatment for outliers in accounting research studies.

2.2 Outliers and skeweness of stock returns

The capital markets literature has examined the distributional properties of accounting and stock returns data and the impact of those properties on the OLS assumptions.[6] The proper treatment of outliers has become a source of debate in several papers examining market efficiency. A controversial issue is the presence of extreme positive stock returns and the use of returns as dependent variable.

A study by Kraft et al. (2005) examines the association between stock returns and accruals and show that a previously documented association in the accrual anomaly literature is influenced by a small number of extreme observations. This study argues that the decision to truncate outliers should depend on the purpose of the analysis. On one side, if researchers aim to test a general theory, their statistical model should fit the bulk of the data, instead of reflecting the effect of relatively few extreme observations. On the other side, if the purpose of the analysis is to evaluate the profitability of a trading strategy and predict future returns, then the only reason to truncate extreme observations is the possibility of data errors.[7]

In contrast, other studies caution against the truncation of stock returns. Core (2006, p. 350) concludes that, "in general, deleting based on the robust regression techniques

[6] These assumptions include: (1) the dependent variable can be calculated as a linear function of a specific set of independent variables, plus a disturbance term; (2) the expected value of the disturbance term is zero; (3) the disturbance terms all have the same variance and are not correlated with one another; (4) the observations on the independent variables can be considered fixed in repeated samples; and, (5) the number of observations is greater than the number of independent variables, and there is no exact linear relationship between the independent variables.

[7] Errors in returns are rare but they are present in the data as noted previously in footnote 4. In addition, Kraft et al. (2006, p. 307) highlight that if an error impacts delisting returns and "researchers suspect that the frequency of delisting is correlated with their partitioning variable (e.g., accruals/ performance) then it will be worthwhile to report the sensitivity of reported results to extreme performing firms. If the results are robust then it is simple to rule out data errors as the source of the significant results. On the other hand, if the results change, the researcher can investigate the affected observations to verify whether there are errors in the returns calculation. Any errors can then either be corrected or deleted and the analysis re-estimated."

employed and advocated by KLW seems inappropriate… deleting extreme observations from skewed return data leads to biased estimates and can bias inferences.” Using simulations, Kothari et al. (2005) show that data truncation can induce a spurious negative relation between future returns and *ex-ante* information variables (e.g., analyst forecasts and abnormal accruals). The results of this study suggest that a combination of right skewness of stock returns and truncation based on large realizations of stock returns may bias the inferences of capital market studies using actual data. Finally, Teoh and Zhang (2011) argue that the results of Kraft et al. (2006) are attributable to non-random deletion of firms with unusually high stock returns, and also document that the association between accruals and stock returns is robust to excluding outliers for a sub-sample excluding loss firms.[8]

These seemingly conflicting views have led researchers in a variety of directions. In our literature review we identified 157 studies that use stock returns as a dependent variable at least in one regression model. From these studies, 53 percent (83) winsorize or truncate extreme stock returns, while the remainder 47 percent (74) use the raw data. This study proposes an alternative technique to address the problem of extreme returns without eliminating these observations.

III. Outliers, influential observations and OLS estimation

Most archival studies in accounting research aim to test a hypothesis about the relation between one or many causal or independent variables *x*, and an outcome or

[8] In contrast to these papers, arguing against the truncation of skewed stock returns, several studies in the analyst forecast error literature truncate large forecast errors. The distribution of forecast errors has large negative errors and is left skewed. Abarbanell and Lehavy (2003, p. 114) highlight this issue "many studies implicitly limit observations in their samples to those that are less extreme by choosing ostensibly symmetric rules for eliminating them, such as winterization or truncations of values greater than a given absolute magnitude", and argue that "such rules inherently mitigate the statistical impact of the tail asymmetry and arbitrarily transform the distribution, frequently without a theoretical or institutional reason for doing so."

dependent variable y.[9] For instance researchers are interested in the impact of individual firm characteristics such as disclosure quality, size, profitability, and leverage on dependent variables such as stock returns, discretionary accruals, analysts' forecast errors, management compensation, and audit fees.[10] The linear regression or OLS framework is the most widely used quantitative tool to assess the association between variables in accounting research studies. In a simple OLS regression:

$$y = \alpha + \beta x + \epsilon \qquad (1)$$

and the expected value of y given x is:

$$E[y|x] = \hat{\alpha} + \hat{\beta}x \qquad (2)$$

Since the parameter estimates are estimated by minimizing the sum of squared errors, $\hat{\beta}$ is the mean effect of x on y and not necessarily the "typical" effect. Given that OLS parameters are based on the conditional mean of the dependent variable, they suffer from the same problems from the mean itself and are altered by extreme observations.

In order to address the effect of extreme observations in the estimated parameters of an OLS model, researchers should consider what causes extreme values of x and y. The distribution of the values of y will depend on the researchers' sample selection procedure, the distribution of the x variables, and additional unknown sources of variation, represented by an error term. Extreme values of y can be caused by a number of reasons, including: (1) extreme values of the hypothesized x variables that influence y; (2) nonlinearities in the

[9] Ball and Foster (1982, p. 165) write: "Because the laboratory environment is unavailable, the solution cannot be to "purify" the data from a theoretical perspective. The researcher must attempt to reduce the level of anomaly implied by the imperfect construct-data correspondence, but also will have to decide how much anomaly is tolerable."

[10] A problem in determining causal relationships in accounting research is the lack of experimental data. As highlighted by Cook and Cambell (1979, p. 37): "Accounting for the third-variable alternative interpretations of presumed (causal) relationships is the essence of internal validity. " Furthermore (1979, p. 56), "with quasi-experimental groups, the situation is quite different. Instead of relying on randomization to rule out most internal validity threads, the investigator has to make all the threats explicit and then rule them out one by one."

relation between *x* and *y*; or (3) additional *x* variables, unknown to the researcher or omitted from the model, that also influence *y*. It turns out that not all extreme observations have the same effect on the estimated parameters of an OLS model. It is important to differentiate four different concepts, univariate outlier, regression or multivariate outlier, leverage points, and influential observations.

3.1 Univariate and regression outliers

In general, a univariate outlier is an observation that stands away from the rest of the distribution of a particular variable. This is the type of outlier addressed by winsorization or truncation methods. Nevertheless, even when it is informative to detect these cases, because extreme observations in either *x* or *y* can mask the underlying relation between *x* and *y*, they are not necessarily problematic for estimating the OLS parameters. In contrast, a *regression outlier* (also called *multivariate*) is an observation that stands apart from the bulk of the data, considering the *x* and *y* variables simultaneously. A regression outlier is also referred to as a *vertical outlier* when it is within the normal range for *x* but it has an unusually high or low value of *y*.

Consider the examples in Figure 1a and 1b. Figure 1a plots a number of data points (*x*,*y*) and a fitted regression line, which has a positive slope, suggesting that *y* is increasing in *x*. Suppose, however, that the *y* observation is either miscoded or is heavily influenced by another variable not included in the model, but the *x* observation for that data point is within the "usual" range compared to other data points. Figure 1b, shows that the outlier will "pull" the regression line down affecting the slope. Various types of outliers have different implications for other parameters of interest besides the slope coefficients. In particular, the presence of vertical outliers biases the intercept and inflates the standard errors.

3.2 Leverage points and influential observations

Leverage points relate to extreme values of the *x* variables that may influence the slope of the regression line. The influence of a leverage point depends of its position compared to the rest of the data. We show the effect of leverage points using two different examples. First in Figure 1c, a data point is distant from the other observations but it lies within or very close to the fitted line. This is considered a "good" leverage point in the sense that improves the overall fit of the model. In contrast, in Figure 1d, a data point is out of the "usual" *x* and *y* range compared to other observations and it is away from the fitted regression line. This is considered a "bad" leverage point, influencing the slope of the regression line. The low value of *y* combined with high value of *x* pull the regression line down. This type of "bad" leverage point is considered an influential observation, if excluded from the analysis, it would change the regression estimates substantially. In general, the observations that have the most influence on all the regression parameters are those that (1) stand apart from the bulk of the data, and (2) have high leverage (unusual *x* values).[11]

3.3 Outliers as a correlated omitted variable problem

As previously discussed, potential outliers in either the dependent or independent variables can be the result of errors or by other problems. Outliers in the dependent variable

[11] An alternative way to characterize the different types of outliers and their implications is as follows: (1) vertical outliers have usual *x* values but unusual *y* values, influencing the intercept and inflating the standard errors; (2) good leverage points have unusual *x* and *y* values, but are on the regression line, improving the estimation by reducing the standard errors; and, (3) bad leverage points have unusual *x* and *y* values, generally away from the bulk of the data, influencing the intercept and slope coefficients and inflating the standard errors.

can arise from skewness in the independent variables or from differences in the data generating process for a small subset of the sample.

As suggested in the previous Section, extreme values of the dependent variable caused by skewness in the independent variable, called good leverage points, are not necessarily problematic because extreme values of *y* are generated by large values of *x*. Nevertheless, potential inference problems are caused by extreme values of *y* not explained by *x*. These are cases may have a different data generating process, for example resulting from an unknown or omitted variable that frequently takes on a value of zero, but occasionally takes on a different value and has a major impact on *y*.

In the context of accruals, for example, INSMED, Inc., a medical technology company, had total accruals scaled by total assets of 206 percent because it reported a gain of $123 million on the sale of technology to Merck. The company had total assets of only $4 million at the start of the year. Such infrequent yet significant events also occur in the context of stock returns. For example, OSICOM Technologies earned buy-and-hold abnormal returns of 459% between 6/1/1995 and 5/31/96. A significant amount of the returns is attributable to agreements it announced during the year. On May 31, 1996, it issued a press release that it became the sole supplier of video equipment for GTE on a $259 million army contract. OSICOM had BHAR of 46 percent in two days (May 30 and May 31).

Depending on their magnitude, ignoring these low frequency but extreme events will generate large standard errors and bias the coefficient estimates. To the extent that these high-impact events are correlated with the independent variables of interest, they lead to the standard correlated omitted variables problem. Holding the dollar value of any event constant, the impact of the event is likely to appear much more significant for smaller firms.

For example, the $123 million dollar gain reported by INSMED, Inc., had a dramatic impact on accruals (206 percent), but had Merck reported the same gain, accruals would have increased by only 2 percent. Consequently, these infrequent events are likely to be related to the many variables in accounting research that are correlated with size.[12]

To illustrate the potential correlation between infrequent events and accounting variables using real data, we report the relation between total accruals and total assets in Figure 2. Using Compustat data between 1972-2001 for all firms with stock prices in excess of $5, we construct 50 bins based on total assets and report box plots of accruals for each bin. Not surprisingly, extreme accruals occur more frequently in smaller firms.

The challenge for accounting researchers is to retain extreme values of y that are "caused" by the independent variables of interest while, at the same time, limiting the influence of extreme values of y caused by infrequently occurring events that are not included in the model. In the following Section, we discuss robust regression techniques that are intended to do precisely this.

IV. Robust regression

4.1 Definition of robust estimator and properties of robust estimators

The term *robust* has many different connotations in the statistics literature. For example, the term *robust standard error* refers to a modification of the standard errors to account for heteroscedasticity or error dependence. For the purposes of this paper, we will refer to robust estimators, and particularly to robust regression, as a class of estimators that satisfy two conditions: "(1) if a small change is made to the data, it will not cause a substantial change in the estimate, and (2) the estimate is highly efficient under a wide range

[12] See Appendix A for additional examples of extreme outcomes for stock returns and accruals.

of circumstances" (Andersen 2008, p. 3). The first condition for a robust estimator is its *resistance* to the presence of unusual observations. A resistant estimator provides a valid estimate for the bulk of the data. The second condition for a robust estimator is its *efficiency*. An efficient estimator has high precision even when the distributional assumptions necessary for the estimator are not strictly met. Finally, an estimator is efficient if its variance is small, resulting in small standard errors.

The literature on robust estimators has focused on two additional properties: *breakdown point* and *bounded influence*. The breakdown point is an overall measure of the resistance of an estimator. It is the smallest fraction of the data that a given estimator can tolerate without producing an inaccurate result. When an estimator "breaks down" it fails to represent the pattern in the bulk of the data. The bounded influence property refers to the influence of each individual observation y_i on the properties of a given estimator. Or, in other words, the marginal change in an estimate by the inclusion of the additional observation y_i.

4.2 Why OLS is not robust under certain conditions

The OLS method to estimate the regression parameters is not robust because its objective function, based on the minimization of the sum of squares of the residuals, increases indefinitely with the size of the residuals. By considering the sum of the square residuals, OLS gives excessive importance to observations with very large residuals. In terms of the definitions above, OLS has *unbounded influence*. Moreover, even a single outlier can have a significant impact on the fit of the regression surface and the *breakdown point* of OLS is zero.

Outliers can also be associated with non-constant error variance, violating one of the assumptions of the OLS model and causing the OLS estimates to lose efficiency because they give equal weight to all observations. OLS weights outliers equally, even when the outliers contain less information about the true relation between *x* and *y*.

4.3 Types of robust regression estimators

The various types of robust regression estimate the parameters of a linear regression model while dealing with deviations from the OLS assumptions. There are a number of robust regression techniques, including: L-estimators (Least Absolute Values LAV, Lest Median Squares LMS, and Least Trimmed Squares LTS), R-estimators, S-estimators, M-estimators, GM-estimators, and MM-estimators.[13]

In general, the L-estimators rely on minimizing a modified version of the sum of square residuals criteria, such as the sum of the absolute values of the residuals (LAV), the median of the squares residuals (LMS), and the sum of truncated or trimmed squares residuals after estimating the regular OLS regression (LTS). Although these methods are relatively easy to compute and have bounded influence, they are generally inefficient, performing badly in small samples. Similarly, the R-estimators rely on minimizing the sum of a score of the ranked residuals, but most R-estimators have low breakdown points.

The S-estimators take a different perspective, focusing on the minimum variance property of the OLS estimators. The S-estimators minimize a measure of the dispersion of the residuals that is less sensitive to outliers than the OLS variance; however, these estimators also have very low efficiency compared to OLS.

[13] LTS estimation was proposed by Rousseeuw (1984); M-estimation was proposed by Huber (1964, 1973); S-estimation was proposed by Rousseeuw and Yohai (1984); and, MM-estimation was proposed by Yohai (1987). The discussion in this section is based on the reviews of robust methods in Andersen (2008); Maronna, Martin and Yohai (2006); and, Fox and Weisberg (2010). Additional guidance on the estimation of robust methods is available in Chen (2002), and Verardi and Croux (2009).

The M-estimators, GM-estimators and MM-estimators are based on minimizing a function of the residuals. This class of estimators minimizes the sum of a function w_{i} of the scaled residuals (scaling residuals by an estimate of their standard deviation) using weighted least squares. The weight function w_{i} is non-decreasing for positive values and less increasing than the square function, thus errors that are far from zero receive progressively less weight than errors that are closer to zero. The most commonly used weight functions are the Huber and bisquare functions. The final weights are informative and can be used to identify which observations are outliers. The general criteria to be minimized is as follows:

$$\sum_{i=1}^{n} w_i\left(\frac{e_i}{\hat{\sigma}_e}\right)x_i = 0 \qquad (3)$$

The OLS can be considered a special case within this class of estimators, where the square function is used to weight the residuals. The MM-estimators are widely used and combine a high breakdown point with high efficiency (95 compared to OLS). These estimators are computed using an iterative procedure, because the residuals cannot be found until the model is fitted, and the parameter estimates cannot be found without the residuals. The estimator procedure follows these steps: (1) first pass coefficients are calculated using some form of resistant regression (usually the S-estimator with Huber or bisquare weights); (2) the first pass coefficients are used to estimate residuals and the scale parameter; (3) a weight function is applied to the scaled residuals; (4) a second pass estimate of coefficients is obtained using weighted least squares; and, (5) the new coefficients are used for a new iteration (keeping constant the measure of the scale of the residuals). The solution is considered to have converged when the change in estimates is no more than 0.1 percent from the previous iterations.[14]

V. Simulations

As discussed in Section 3.3, extreme values of y from infrequent but extreme events, possibly arising from a data-generating process that is different from the bulk of the sample, can bias coefficient estimates if these events are correlated with *x*. In this section we conduct simulations to evaluate the relative effectiveness of methods commonly used in accounting research, "do nothing", winsorization, and truncation, and compare them to robust regression based on MM-estimation at mitigating the impact of outliers.

For our simulations, we construct a data generating process consisting of a primary variable of interest *x,* and a variable *z*, that is zero for most of the sample. The variable *z* can be thought of as an infrequent event that generates extreme values of y whenever it occurs. To illustrate the potential impact of these infrequent events, *z* can occur only for high values of x, which induces a correlated omitted variables problem if *z* is ignored and extreme values of y (caused by *z*) are not dealt with properly. Equation (4) describes the generating process for y:

$$Y = \alpha + \beta x + \gamma z + e \qquad (4)$$

where

$x \sim N(0,1)$,
$z = d * v,$
$d = 1$ if x is in the top decile of its distribution and a random draw from a uniform distribution exceeds 0.8,
$v \sim N(3,1)$,

[14] We identified a small number of paper that use various forms of robust regression, for example: Abbody et al. (2010), Bell et al. (2008), Chen et al. (2008), Choi et al. (2009), Dyreng and Bradley (2009), Kimbrough (2007); and, Ortiz-Molina (2007). Appendix B shows excerpts from these studies were robust regression procedures are mentioned. Unfortunately, we note that in most cases these procedures are just mentioned in a footnote and there is insufficient information to replicate them independently.

$e \sim N(0,1)$.

By construction, z will be zero approximately 98 percent of the time. For our primary analysis, we generate 250 samples of 2,000 observations. For simplicity, we set $\alpha = 0$, $\beta = 0.8$ or zero, and $\gamma = 1$ or zero. We set $\beta = 0$ ($\gamma = 0$) when we test for potential Type I (Type II) errors. For each sample we generate, we estimate regressions employing the following alternative approaches:

1. Do nothing. As described in our literature review of recent accounting research, approximately a third of published archival papers do not report addressing the existence of outliers in the data;
2. Winsorize y and x at the extreme 1 percent of the distributions; [15]
3. Winsorize x but leave y as is;
4. Truncate y and x at the extreme 1 percent of the distributions;
5. Truncate x but leave y as is; and,
6. Leave x and y variables as they are, but use robust regression based on MM-estimation.

5.1 Graphical illustrations of alternative identification and treatment of outliers.

Before reporting our main simulation results, we construct a sample of 4,000 observations and generate scatter plots to illustrate the impact of alternative approaches to treat outliers. Figure 3a is a scatterplot of 4,000 observations where y is generated from the following process:

$$y_i = 0.8x_i + z_i + e_i \qquad (5)$$

[15] We assume that z is not observed by the researcher and, therefore, leave the generated value as is.

Since *z* is correlated with *x* and *y*, outliers will bias the coefficient on *x* if left untreated. The circles in Figure 3a represent observations where *z* is nonzero. As expected from the construction of *y*, these "shocks" occur for roughly 2 percent of the sample and only when *x* is extreme. These observations will obviously induce an upward bias on the β coefficient. The estimated coefficient from this sample (represented by the black line) is 0.89.

Figure 4a, reports the same data used for Figure 3 but overlays lines representing cutoffs for the top and bottom 1 percent of *x* and *y*. As the plot suggests, winsorizing or truncating will reduce the influence of *z* to a small degree but it will also reduce the influence of good leverage points. Figure 4b is a scatterplot of the sample after winsorizing *x* and *y*. This plot illustrates that winsorizing does little to reduce the impact of *z* on *y* because, although magnitudes of the *y*'s are reduced, they are still retained and are still large relative to the rest of the sample. Estimation of equation (5) with the winsorized data yields a slope coefficient of 0.87, reducing the bias caused by *z* by only 0.02.

Figure 4c is a scatterplot of the data after winsorizing *x* but not *y*. As reported in Table 1 this is a fairly common practice in accounting research, particularly when the dependent variable is returns. This approach generally makes problems worse than doing nothing at all as it can bias the coefficient away from zero. For example, assume that for a given observation, e = 0, *x* = 5, *z* = 0, and y = 4 (0.8**x*). For example, if *x* is winsorized to 2, an artificial error of 2.4 (4 - 0.8*2) will bias the regression line upwards. The estimated slope coefficient after winsorizing *x* only is 0.91 compared to 0.89 when nothing winsorized and 0.8, the true value of the β coefficient.

Figure 5a depicts a scatterplot after truncating extreme values of *x* and *y* (top and bottom 1 percent). Truncation leads to a parameter estimate of roughly 0.77, which suggests some downward bias. Figure 4b illustrates a scatterplot after truncating only extreme values of *x*. As suggested by the plot, truncating only the independent variable does not reduce the bias caused by *z*. As illustrated by the circles, many of the extreme values of y caused by z still remain in the sample.

Figure 6 is a scatterplot to illustrate observations considered to be outliers, which will be down-weighted significantly using robust regression based on MM-estimation. The diamonds represent data points with extreme values of *y* caused by *z* and considered outliers. The circles symbolize data points considered outliers but that are not influenced by *z* (when *z* is zero). The triangles are data points with extreme values of *y* that are influenced by z (when z is not zero) but are not considered outliers. The advantage of this robust estimation procedure is that good leverage points are retained. That is, extreme values of *y* caused by extreme values of *x* are retained which increases efficiency relative to a simple truncation rule. Consequently, robust estimation has the advantages of being both consistent and efficient relative to winsorizing and truncating.

5.2 Simulation Results

Base simulation results are reported in Table 2, Panel A. In this panel, the outliers generated by an infrequent event are randomly distributed and independent of x. We generate values of y as described in Equation (4) except that instead of extreme events *z* occurring only when *x* is in the top decile, *z* occurs with equal probability (2 percent) across the x distribution. Values of *y* are generated with β and γ being set to 0.8 or zero, and 1.0 or zero, respectively. In the first four regressions, *z* is omitted from the regression estimation to simulate a typical case where an infrequent event is omitted from the regression.

The second four regressions are estimated with z included in the regression, which is the ideal solution if the research could identify these cases. The table reports mean estimates of β for 250 samples of 2,000 observations, with various treatments for outliers. Bias is the difference between the "true" parameter value and the mean estimate of β. Significance levels for bias are also reported in the table ($p < 0.01$***, $p < 0.05$**, and $p < 0.1$=*). Since standard errors are very small, small magnitudes of bias will be significant.

The benchmark case is "do nothing", where OLS is estimated without trimming, winsorizing or down-weighting outliers. Not surprisingly, since the outliers are randomly distributed, estimates of β are unbiased in all cases. Results are virtually identical when *x* and *y* are winsorized at the top and bottom 1 percent. As reported in Kothari et al. (2005), truncating *y* and *x* imparts a downward bias on bias but only when there is a correlation between *x* and *y*. If *y* is generated when β= 0, truncation does not impact β (bias it to a value below zero).[16] The intuition for this is that truncated outliers are not outliers generated by extreme values of *x* (since *x* is uncorrelated with *y*), which means no good leverage points that might induce a relation between *x* and *y* are lost.

In contrast, when *x* is correlated with *y* (β = 0.8 in the *y* generating process), the coefficient is biased down as expected. β is downward biased by amounts ranging from 0.03 to 0.06, depending on the specification. Although this will not likely impact significance levels in our simulations since β is set to 0.8, truncation can generate Type I or Type II errors

[16] See Appendix C for an analytical representation of bias caused in parameter estimates as a result of truncation.

if the true value of β is closer to zero. These findings suggest truncation of outliers should be avoided in favor of other alternatives.

As with winsorizing and our base case, robust regression results reveal unbiased estimates for all specifications. There is no evidence that robust regression biases coefficients when outliers are randomly occurring. In contrast to Panel A, Panel B reports results for the case when the infrequent events *z*, are correlated with *x*. Such events can only occur when x is in the top decile of its distribution. As an example, *x* might be negatively correlated with size and z tends to occur only for small firms. These non-random outliers are a concern for researchers attempting to draw inferences about how *x* influences *y*.

As in Panel A, *z* is omitted from the regressions in the first two rows of Panel B and serve as benchmarks for rows three and four, where γ = 1.0. As expected, when γ = 0, β estimates are about 0.8 (unbiased) across the board except for the case of truncation, where, again, there is a downward bias in the parameter estimate. In contrast, when γ = 1.0, we have a classic correlated omitted variables problem that will bias estimates of β if these outliers are ignored. When outliers are untreated (Do nothing case), β is biased upwards by 0.10 and, as row four shows generates a Type I error. Even though *x* is uncorrelated with *y* by construction (β= 0), the mean coefficient estimate is 0.10. Column (2) shows that winsorizing at 1 coefficient does virtually nothing to reduce the influence of the correlated outliers *z*. The coefficient is 0.09 and almost identical to the do nothing case. In Column 4, truncation appears to mitigate the bias but this is largely an artifact of a tendency for truncation to bias parameter estimates down.

Column 4 of Panel B reports the robust regression results, which show an 80 percent reduction in the bias caused by the outliers (0.10 versus 0.02). The bias is not completely eliminated but the impact is reduced substantially. In untabulated results, the bias created by the outliers is significantly different from zero in only 6 percent of the 250 samples ($p < 0.01$) compared to 90 percent of the time when outliers are winsorized. Overall, the simulation results in Table 2 show that robust regression is unaffected by random outliers and significantly reduce bias caused by extreme events that are correlated with a researchers variable of interest. Winsorizing does little to mitigate the influence of correlated outliers and truncation imparts a downward bias on parameter estimates.

5.3 Trimming the independent variable but not the dependent variable.

As our literature review suggests, it is a surprisingly common practice for researchers to either winsorize or truncate the independent variable but leave the dependent variable as it is. As illustrated by example and in Figures 3c and 4a this practice is likely to bias coefficients away from zero. In Table 3 we winsorize or truncate only the independent variable *x* and the base case is reported for comparison purposes. In Panel A, extreme events are randomly distributed and uncorrelated with *x*.

As Panel A shows, winsorizing *x* but not y imparts an upward bias on β, increasing it from 0.80 to 0.82. As discussed earlier, winsorizing only *x* "leverages" up the impact of *x* on *y*. In contrast, truncation of only the *x* variable appears to eliminate the downward bias caused by truncation of both *x* and *y* reported in Table 2. However, as reported in Table 3, Panel B, truncating only *x* does nothing to mitigate the bias caused by correlated outliers.

Panel B also reinforces the upward bias caused by winsorizing only the independent variable. In summary, based on evidence reported in Table 3, trimming the independent variables but not the dependent variable is not advisable.

VI. Conclusion

In this study we review the accounting literature and quantify the various methods used to mitigate the impact of influential observations and find significant variation. The most common solutions to address the presence of outliers are winsorization or truncation but even within these approaches implementation varies. For example, some studies trim (winsorize or truncate) the independent variables but not the dependent variables. Moreover, some studies winsorize only a subset of all variables. Almost a third of all studies do not report a treatment of any kind. This wide variation in the treatment of outliers makes comparability across studies and assessment of causality problematic.

The purpose of this study is to evaluate alternative approaches currently in the accounting literature and compare them to more formal treatment of outliers through robust regression procedures. We evaluate these approaches in cases where outliers occur randomly as well as cases where outliers are correlated with the independent variable of interest. The causes of outliers in the context of accounting information and stock returns do not appear to be random. For example, extreme events that generate extreme outcomes likely occur more frequently in smaller firms. Therefore, we are interested in how various procedures mitigate the impact of infrequent but correlated events (outliers).

We find that winsorizing does little to mitigate the influence of correlated outliers and winsorizing only the independent variable biases the coefficients away from zero, increasing the probability of a Type I error. In contrast, consistent with Kothari, et al. (2006), truncation tends to bias coefficients down, unless the independent variable is uncorrelated with the dependent variable (i.e. the "true" parameter is zero). We compare these approaches to a robust estimation procedure based on MM-estimation that is both consistent and highly efficient even in the presence of outliers. Robust regression substantially reduces bias induced by correlated outliers (80 percent reduction in bias). Based on our findings, accounting researchers should strongly consider the use of robust regression procedures as a standard practice and alternative to winsorization and truncation.

References

Abarbanell, J., and R. Lehavy. 2003. Biased forecasts or biased earnings? The role of reported earnings in explaining apparent bias and over/underreaction in analysts' earnings forecasts. *Journal of Accounting and Economics* 36: 105–146.

Aboody, D., N. Jonson, and R. Kasznik. 2010. Employee stock options and future firm performance: Evidence from option repricings. *Journal of Accounting and Economics* 50: 74–92.

Andersen, R. 2008. Modern methods for robust regression. Thousand Oaks, CA: Sage.

Ball, R. and G. Foster. 1982. Corporate financial reporting: A methodological review of empirical research. *Journal of Accounting Research* 20(Supplement): 161–234.

Bell, T., R. Doogar, and I. Solomon. 2008. Audit labor usage and fees under business risk auditing. *Journal of Accounting Research* 46(4): 729–760.

Belsley, D., E. Kuh, and R. Welsch. 1980. Regression diagnostics: Identifying influential data and sources of collinearity. New York, NY: John Wiley.

Chen, W, C. Liu, and S. Ryan. 2008. Characteristics of securitizations that determine issuers' retention of the risks of the securitized assets. *The Accounting Review* 83(5): 1181–1215.

Choi, J., J. Kim, X. Liu, and D. Simunic. 2009. Cross-Listing audit fee premiums: Theory and evidence. *The Accounting Review* 84(5): 1429–1463.

Cook, T., and D. Campbell. 1979. Quasi-experimentation: Design and analysis for field settings. Chicago, IL: Rand McNally.

Core, J. 2006. Discussion of an analysis of the theories and explanations offered for the mispricing of accruals and accrual components. *Journal of Accounting Research* 44(2): 341–350.

Chen, C., 2002. Robust regression and outlier detection with the ROBUSTREG procedure. SUGI Paper No.265-27. Cary, NC: The SAS Institute.

Dyreng, S., and B. Lindsey. 2009. Using financial accounting data to examine the effect of foreign operations located in tax havens and other countries on U.S. multinational firms' tax rates. *Journal of Accounting Research* 47: 1283–1316.

Fox, J. and S. Weisberg. 2011 An R companion to robust regression. Thousand Oaks, CA: Sage.

Kennedy, P. 2003. A guide to econometrics. Cambridge: The MIT Press.

Kimbrough. M. 2007. The influences of financial statement recognition and analyst coverage on the market's valuation of R&D capital. *The Accounting Review* 82(5): 1195–1225.

Kothari, S. 2001. Capital market research in accounting. *Journal of accounting and economic*s 31(1–3): 105–231.

Kothari, S., J. Sabino, and T. Zach. 2005. Implications of survival and data trimming for tests of market efficiency. *Journal of Accounting and Economics* 39 (1): 129–161.

Kraft, A., A. Leone, and C. Wasley. 2006. An analysis of the theories and explanations offered for the mispricing of accruals and accrual components. *Journal of Accounting Research* 44 (2): 297–339.

Maronna, R., D. martin, and V. Yohai. 2006. Robust statistics theory and methods. Hoboken, NJ: John Wiley.

Ortiz-Molina, H. 2007. Executive compensation and capital structure: The effects of convertible debt and straight debt on CEO pay. *Journal of Accounting and Economics* 43: 69–93.

Rousseeuw, P. 1984. Least median of squares regression. *Journal of the American Statistical Association* 79(4): 871–880.

Rousseeuw, P. , and V. Yohai. 1984. Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis*, edited by J. Franke, W. Härdle, and R. Martin. Lecture Notes in Statistics 26, New York, NY: Springer-Verlag.

Teoh, S., and Y. Zhang. 2011. Data truncation bias, loss firms, and accounting anomalies. *The Accounting Review* 86(4): 1445–1475.

Huber, P. 1964. Robust estimation location parameters. *Annals of Mathematical Statistics* 35(1): 73–101.

Huber, P. 1973. Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics* 1: 799–821.

Verardi, V and Croux, C. 2009. Robust regression in Stata. The Stata Journal 9(3): 439–453.

Yohai V. 1987. High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics* 15: 642–656.

Figure 1a- Normally distributed data

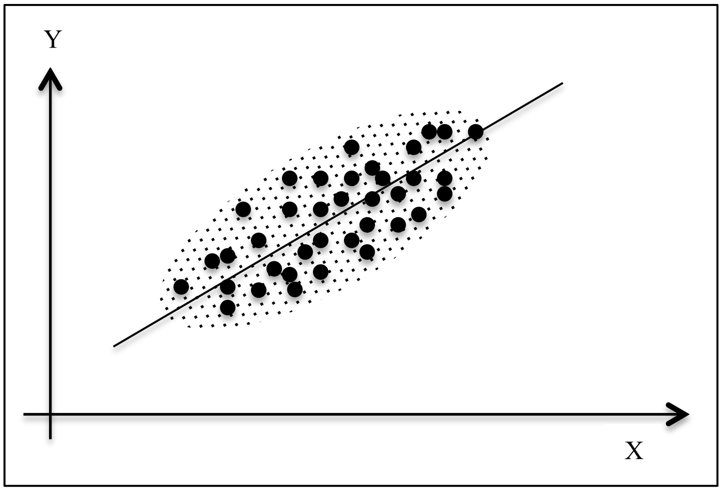

Figure 1b- Vertical outlier

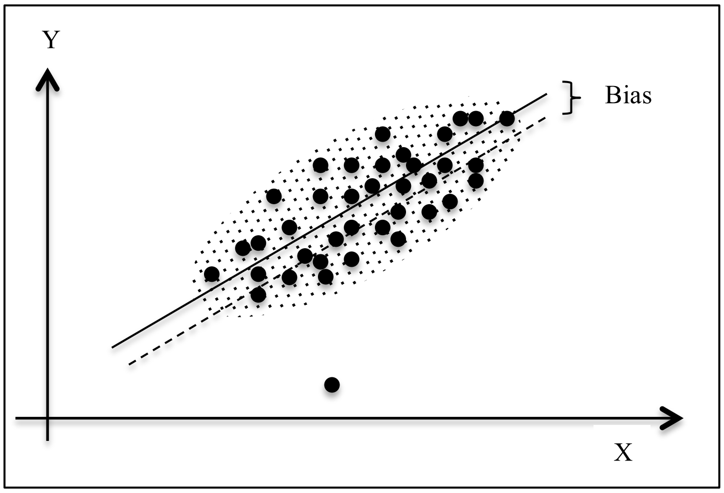

Figure 1c – Good Leverage Point

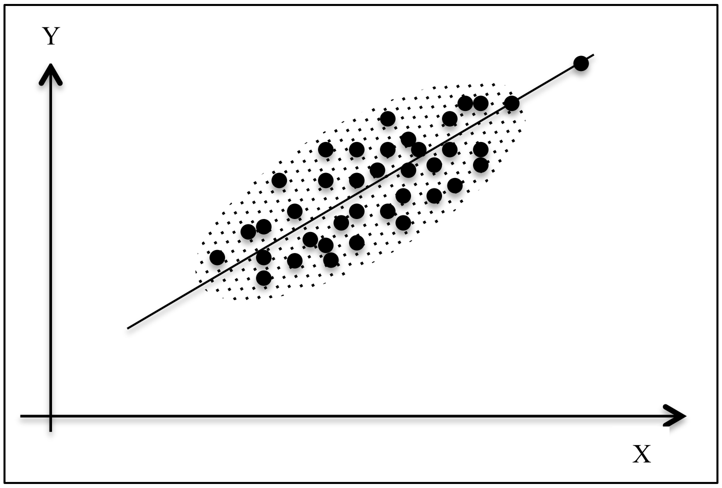

Figure 1d – Bad leverage point

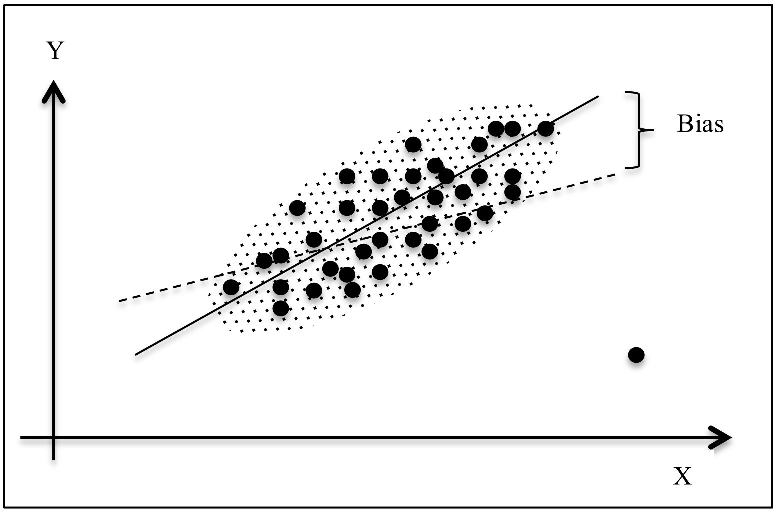

Figure 2 – Accrual outliers and Firm Size.

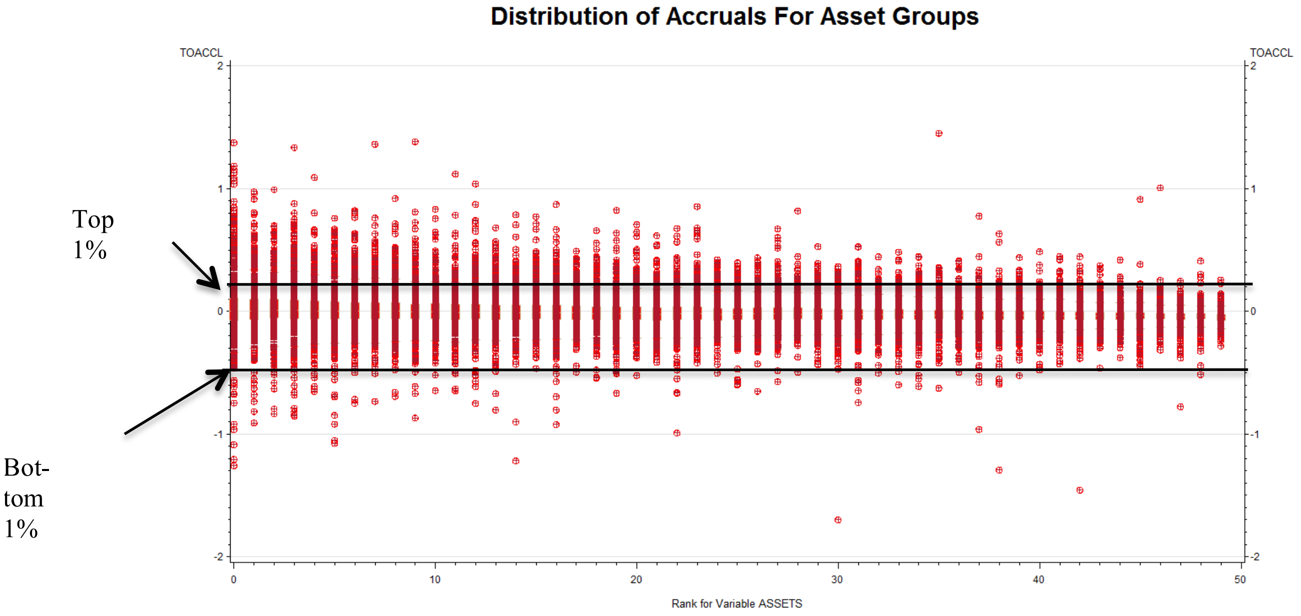

Figure 2 is a box plot of accruals for total assets grouped into fifty bins. Accruals and assets data are obtained from Compustat and include and include all firms with sufficient data to compute total accruals (balance sheet) between 1972-2001.

Figure 3- Scatterplot of (*x,y*) data points in a simulated dataset and a fitted OLS regression line. The estimated slope coefficient is 0.894

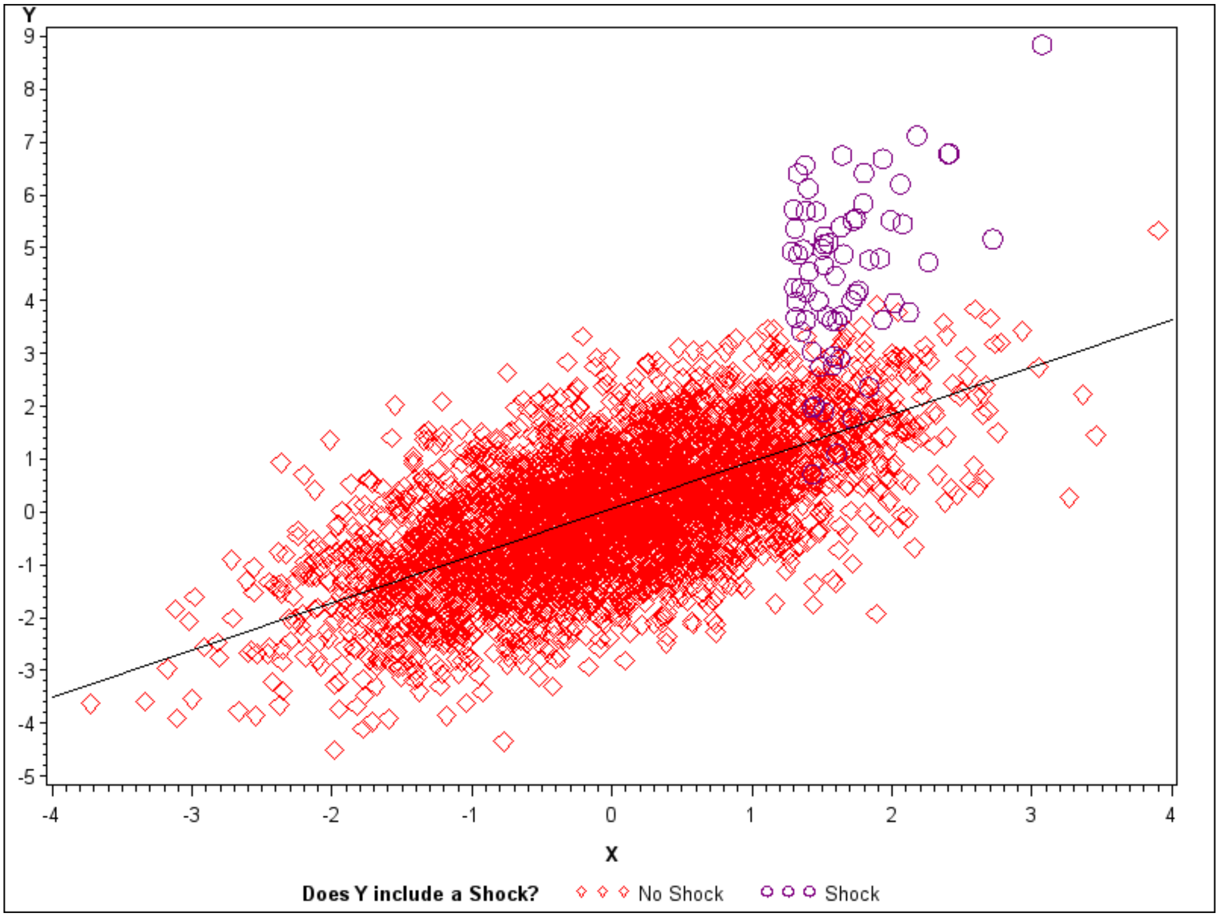

Figure 4a – Scatterplot with lines representing top and bottom 1 percent of *x* and *y* distributions.

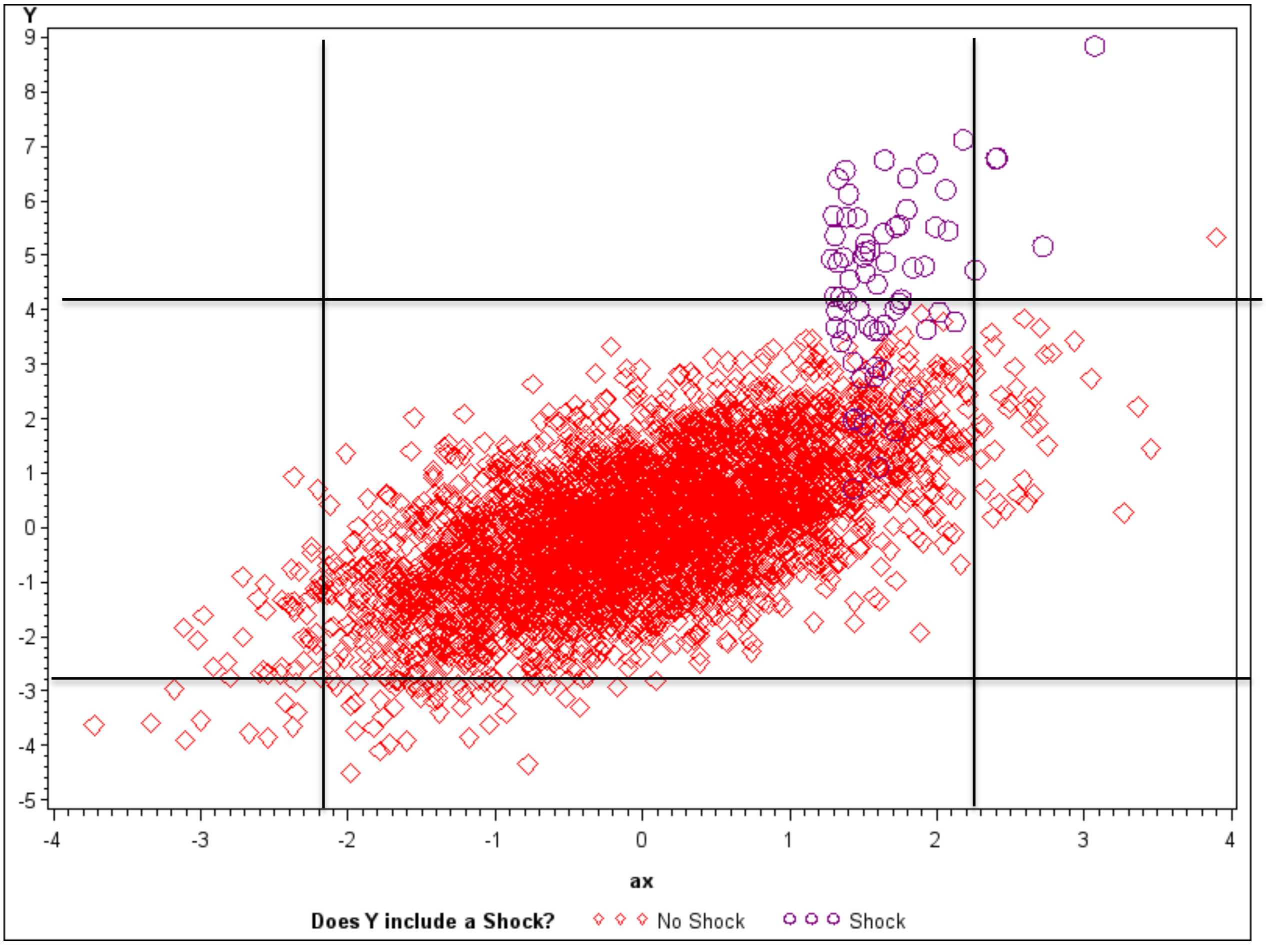

Figure 4b - Scatterplot of (*x,y*) data points in a simulated dataset and a fitted OLS regression line after winsorizing *x* and *y* at the top and bottom 1 percent. The estimated slope coefficient is 0.87.

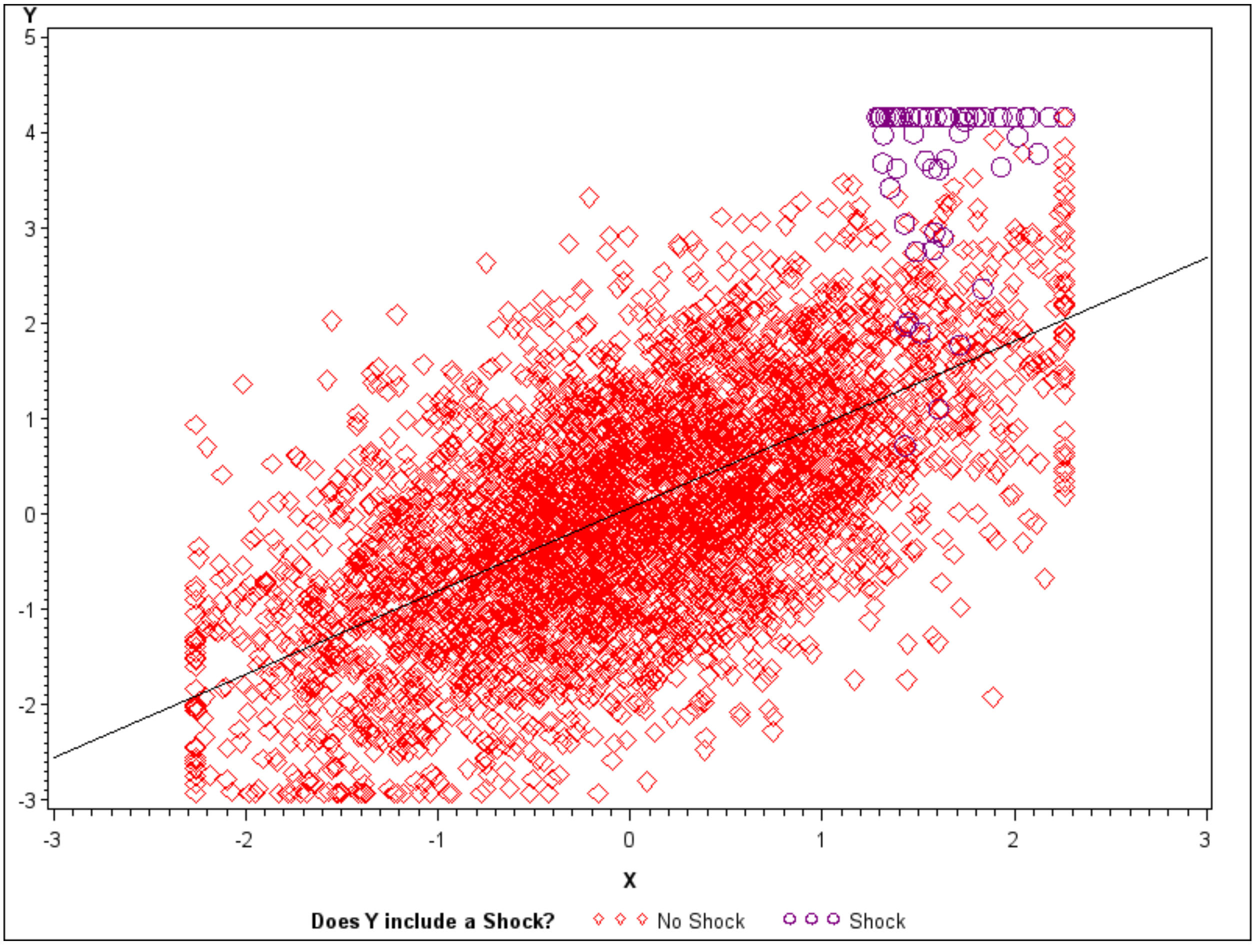

Figure 4c - Scatterplot of (*x,y*) data points in a simulated dataset and a fitted OLS regression line after winsorizing only *x* at the top and bottom 1 percent. The estimated slope coefficient is 0.91.

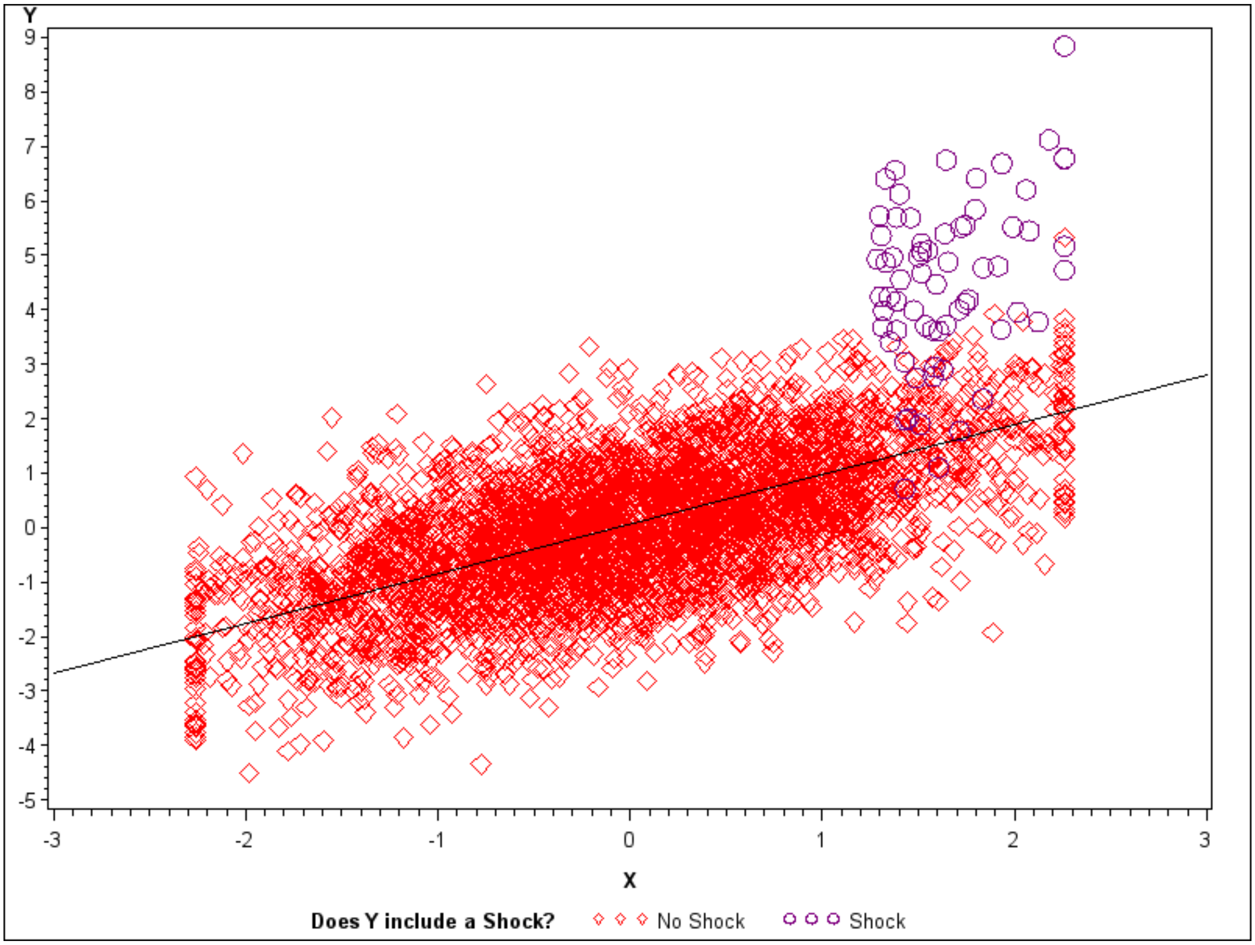

Figure 5a - Scatterplot of (*x*,*y*) data points in a simulated dataset and a fitted OLS regression line after truncating *x* and *y* at the top and bottom 1 percent. The estimated slope coefficient is 0.80.

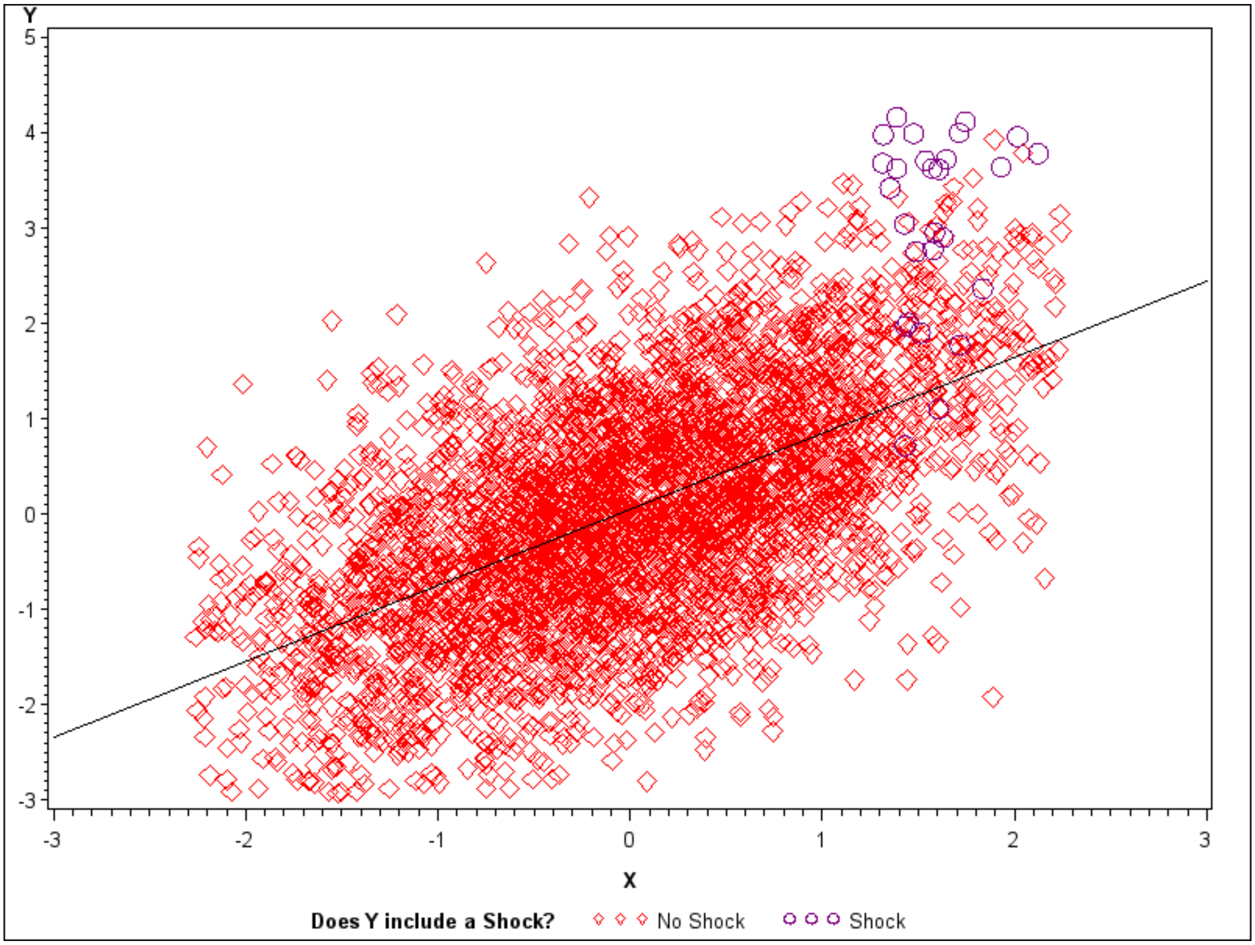

Figure 5b - Scatterplot of (*x,y*) data points in a simulated dataset and a fitted OLS regression line after truncating only *x* at the top and bottom 1 percent. The estimated slope coefficient is .80.

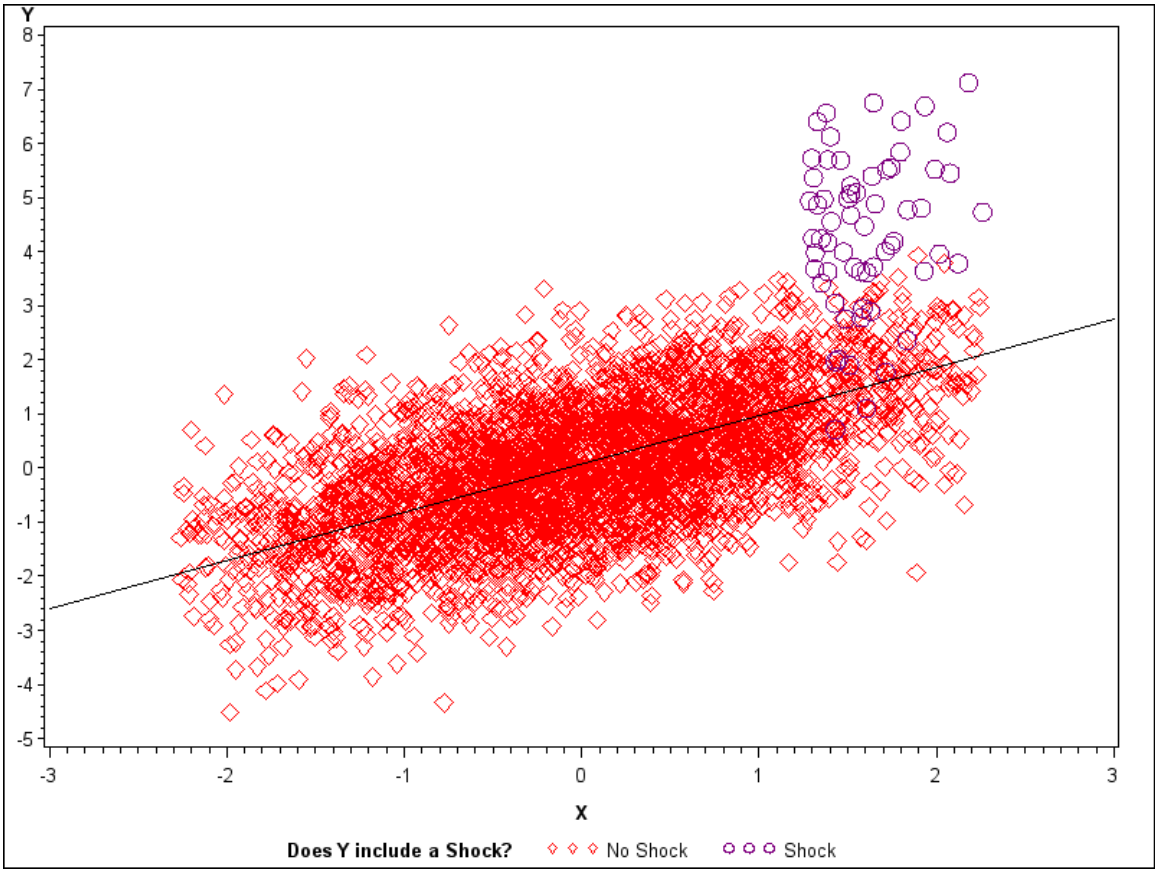

Figure 6 - Scatterplot of (*x,y*) data points in a simulated dataset and a fitted robust regression line based on MM-estimation. The estimated slope coefficient is 0.82.

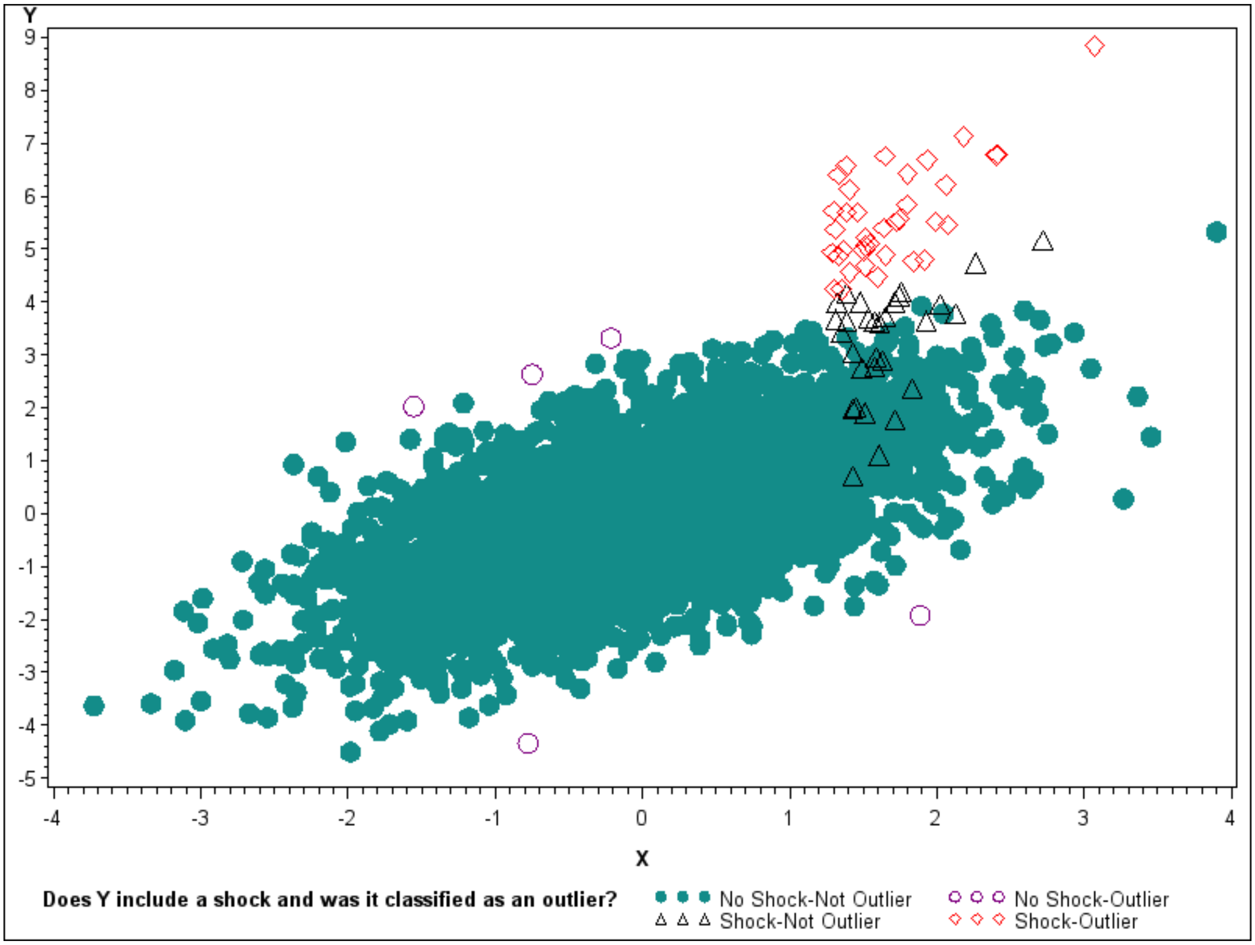

Table 1 –Literature Review
Panel A –Breakdown of studies by general category

	Number of studies		*Percentage of total*	
Archival	590		69%	
Analytical	101		12%	
Experimental	106		12%	
Discussion and reviews	60		7%	
Total number of studies		857		100%
Archival studies addressing outliers	404		68%	
Archival studies do not addressing outliers	186		32%	
Total archival studies		590		100%
Archival studies using winsorization	221		55%	
Archival studies using truncation	161		40%	
Archival studies using both winsorization and truncation	27		7%	
		355		88%
Archival studies using other techniques		49		12%
Total archival studies addressing outliers		404		100%
Archival studies with returns as dependent variable	157		27%	
Archival studies with other dependent variables	433		73%	
Total archival studies		590		100%
Returns winsorized as dependent variable	45		29%	
Returns truncated as dependent variable	43		27%	
Returns winsorized and truncated as dependent variable	5		3%	
		83		53%
Returns used raw as dependent variable		74		47%
Total archival studies with returns as dependent variable		157		100%

Panel B –Breakdown of studies using winsorization

	Number of studies	*Percentage of total*
Independent variables	202	91%
Dependent variables	151	68%
Both independent and dependent variables	132	60%
Archival studies using winsorization	221	100%

Panel C –Breakdown of studies using truncation

	Number of studies	*Percentage of total*
Independent variables	139	86%
Dependent variables	143	89%
Both independent and dependent variables	121	75%
Archival studies using truncation	161	100%

This table presents the result of our literature review. We reviewed all studies published between 2006 and 2010 in the following journals: Contemporary Accounting Research, Journal of Accounting Research, Journal of Accounting and Economics, Review of Accounting Studies, and The Accounting Review. We searched the body, footnotes and tables of each archival study for any discussion of the treatment of influential observations or outliers. The studies in our review span a variety of areas, such as auditing, properties of analysts' forecasts, management compensation, earnings management, conservatism, tax, disclosure, and earnings-returns associations. Studies that include both an analytical model and archival empirical tests are classified as archival.

Table 2 –Main Simulation Results

Panel A – Infrequent events (z) are independent of x

	Parameter Values		Do Nothing		Winsorize		Truncate		Robust Regression	
Regression	β	γ	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias
$\boldsymbol{Y = \alpha + \beta x + e}$	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	.80	0	0.80	0.00	0.80	0.00	0.75	-0.06 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	.80	1	0.80	0.00	0.80	0.00	0.75	-0.05 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	.80	0	0.80	0.00	0.80	0.00	0.74	-0.06 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	.80	1	0.80	0.00	0.80	0.00	0.77	-0.03 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Panel B – Infrequent events (z) are correlated with x

	Parameter Values		Do Nothing		Winsorize		Truncate		Robust Regression	
Regression	β	γ	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias
$\boldsymbol{Y = \alpha + \beta x + e}$	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	.80	0	0.80	0.00	0.79	-0.01 *	0.74	-0.06 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	.80	1	0.90	0.10 ***	0.89	0.09 ***	0.81	0.01 ***	0.82	0.02 ***
$\boldsymbol{Y = \alpha + \beta x + e}$	0	1	0.10	0.10 ***	0.09	0.09 ***	0.04	0.04 ***	0.02	0.02 ***
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	.80	0	0.80	0.00	0.79	-0.01 *	0.74	-0.06 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	.80	1	0.80	0.00	0.80	0.00	0.77	-0.03 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

This table reports simulation results constructed as follows. The independent variable x, is generated from a normal distribution with mean zero and standard deviation of one. An “extreme event” variable v, is generated from a normal distribution with a mean of three and standard deviation one. For Panel A, v is multiplied by a the variable d, which is one when a random draw from a uniform distribution has a valued equal to or exceeding 0.98, which means $z= d*v$, is non-zero roughly 2 percent of the time. For Panel B, a relation between x and z is induced differently. The variable d is assigned a value of zero whenever the corresponding x falls below the top decile of its distribution. If x is in the top decile of its distribution, d is assigned a value of one whenever a random draw from a uniform distribution exceeds 0.8. This implies that, again, $z= d*v$ is non-zero roughly 2 percent of the time but is correlated with x. The dependent variable y, is generated by applying the following data generating process:

$$Y = \alpha + \beta x + \gamma z + e,$$

where e is drawn from a standard normal distribution. In all cases, a is set to zero. Four y variables are generated by varying values of b (zero or 0.8) and g (zero or 1.0). A total of 250 samples are generated where n= 2,000. For each sample, x and y variables are winsorized or truncated at the top and bottom 1 percent of the sample for results reported in “Winsorize” and “Truncate,” Columns. Reported estimates of b are means from the estimation of 250 regressions under each condition. Bias is the difference between the mean and “true” parameter value (zero or 0.8). Reported significance levels are compute from t-statistics. All regressions are estimated using OLS, except for robust regression results based on MM-estimation. ***, **, * symbolize significance at p < 0.01, p < 0.05 and p < 0.10.

Table 3 – Simulation Results Winsorizing ot Truncating Only *x*

Panel A – Infrequent Events *z* are Independent of *x*

	Parameter Values		Do Nothing		Winsorize		Truncate	
Regression	β	γ	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias
$\boldsymbol{Y = \alpha + \beta x + e}$	0	0	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	.80	0	0.80	0.00	0.82	0.02 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	.80	1	0.80	0.00	0.82	0.02 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	0	1	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	0	0	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	.80	0	0.80	0.00	0.82	0.02 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	.80	1	0.80	0.00	0.82	0.02 ***	0.80	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	0	1	0.00	0.00	0.00	0.00	0.00	0.00

Panel B – Infrequent Events *z* are Correlated with *x*

	Parameter Values		Do Nothing		Winsorize		Truncate	
Regression	β	γ	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias
$\boldsymbol{Y = \alpha + \beta x + e}$	0	0	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + e}$	.80	0	0.80	0.00	0.81	0.01 ***	0.74	-0.06 ***
$\boldsymbol{Y = \alpha + \beta x + e}$	.80	1	0.90	0.10 ***	0.92	0.12 ***	0.81	0.01 ***
$\boldsymbol{Y = \alpha + \beta x + e}$	0	1	0.10	0.10 ***	0.11	0.11 ***	0.04	0.04 ***
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	0	0	0.00	0.00	0.00	0.00	0.00	0.00
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	.80	0	0.80	0.00	0.81	0.01 ***	0.74	-0.06 ***
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	.80	1	0.80	0.00	0.81	0.01 ***	0.77	-0.03 ***
$\boldsymbol{Y = \alpha + \beta x + \gamma z + e}$	0	1	0.00	0.00	0.00	0.00	0.00	0.00

This table reports simulation results constructed as reported in Table 2. In this table, however, data trimming (winsorizing and truncating is performed on *x* only and *y* is left a is. ***, **, * symbolize significance at p < 0.01, p < 0.05 and p < 0.10.

APPENDIX A – Examples of low frequency events causing extreme outcomes

Extreme Stock Returns

OSICOM Technologies 6/1/1995-5/31/96 (BHAR 459%)
5/31/96- Press release OSICOM UNIT NAMED SOLE SUPPLIER OF VIDEO EQUIPMENT FOR GTE $259 MILLION ARMY CONTRACT.” (46% two-day BHAR).

1/19/1996- Press Release “1/19/96 - ROCKWELL TO SELL ITS NETWORK SYSTEMS BUSINESS TO OSICOM” (One-day BHAR-47%).

4Kids Entertainment (A licensing Company) -5/1/1998-4/30/1999
5/17/1998 “Pokemon poised to be pop culture's next big phenomena Digicritters move out of Game Boys, into film, TV, toys and more.” (2 day BHAR 30%).

TEKELEC (468% BHAR) 5/2/1994-5/1/1995
Sept 19,1994 – announce distribution agreement with AT&T – up 25%.

Jones Medical Industries – 5/1/1995-4/30-1996 (BHAR 669%)
3/18/1996 – Announce a marketing rights deal (up 24% in 3 days).

Extreme Negative Accruals (Examples from bottom 1%)

OPKO Health, Inc. 2007 – (-1,000%)
Large write off of In-Process R&D ($243 Million on $40 million of assets in 2007 and assets of only $116k in 2006).

CARDINAL COMMUNICATIONS INC -990%, 2002.
Expenses paid with stock ($2,129,635) and assets of only $407K.

JDS UNIPHASE CORP (-290%, 2001)
Write down of good will $50 million on assets of $12 million.

Extreme Positive Accruals

Raytech (2001 accruals of 2200%)
Company emerged from bankruptcy in 2001, and adopted fresh-start accounting. Accruals actually relate to the short-year January –April 2001 and extra-ordinary gain of $6 million on assets of 300k.

INSMED (2009 accruals of 206%)
Gain on sale of an asset amounting to $127 million, with total assets of increasing from $4 million to $127million. This was the sale of intangible technology to Merck.

SOMANETICS CORPORATION (2004 accruals of 50%)
Recognition of a deferred tax asset $6,700,000.

APPENDIX B – Examples of studies that use a form of robust regression

Aboody et al. (2010)

We estimate all equations using a robust regression technique, pooling data across years. The procedure begins by calculating Cook's D statistic and excluding observations with D>1. Then, the regression is re-estimated, weights for each observation are calculated based on absolute residuals – Huber weights and biweights – and the estimation is repeated iteratively using the weighted observations until convergence in the maximum change in weights is achieved.

Bell et al. (2008)

To reduce the effects of outliers on estimated effects, we employ bounded influence ordinary least squares (OLS) (unreported results using seemingly unrelated regressions yield qualitatively similar conclusions). Estimating each model using robust regressions (excluding observations with leverage greater than one and smoothly downweighting outliers) does not materially alter the results. We report the percentiles and medians in addition to the mean values of the ratios since the mean is susceptible to the influence of outliers.

Chen et al. (2008)

To mitigate the effect of outliers, we winsorize observations in the outside l percent of each tail of each variable in Equation (1), excepting the lower tail of the variables that are bounded below by zero and have some zero observations. Our results are substantially unaffected if we do not winsorize outliers of the dependent variable, if we delete rather than winsorize outliers, and if we make the winsorization/deletion rule more or less stringent within commonly applied levels as long as extreme outliers are pulled in or deleted (e.g., 2.5 percent and 0.5 percent winsorization rules yield similar results). To reduce the potential effect of influential observations, we estimate Equation (1) using least absolute deviation (LAD) estimation, which reduces the weight placed on large model residuals compared to least-squares estimation. The results of these approaches are reported in Table 5.

Choi et al. (2009)

Unless otherwise specified, all of the regressions are estimated after removing outliers that have a Cook's (1977) distance value greater than 4/(sample size). As a result, the actual sample size is slightly smaller than 17,837 and varies across the regressions. Finally, we perform a median quantile regression and a robust regression to minimize the influence of extreme observations without removing them from regression analyses.

Dyreng and Bradley (2009)

We use robust regression to control for outliers in all tables. Because robust regression iteratively

assigns weights to observations to mitigate the influence of outliers, some observations effectively receive a weight of 0, and are not included in the regression. We report the number of observations with nonzero weights in the N for each regression, which accounts for the slightly varying N from table to table. In a prior version of the paper, we used OLS and truncated all variables at the 1st and 99th percentiles. We use robust regression in this version because we view the procedure as less subjective.
Kimbrough (2007) In addition, to mitigate the impact of outliers, I report estimates based on **Huber M** estimation, which is a robust estimation method that, instead of minimizing the sum of squared residuals, minimizes the sum of less rapidly increasing functions of the regression residuals.
Ortiz-Molina (2007) As it is typical in the executive compensation literature (and evident in Table 1), the right skewness of the data and the presence of large outliers require a robust estimation method. Following previous research, I use median regression (MR; also known as least absolute deviation regression) throughout the analysis.

APPENDIX C – Impact of truncation on estimates of β.

In the analysis below, we show the impact of truncation on estimates of β in the case of one independent variable, x. β_n, β_T, and β_h represent coefficients for the entire sample (i=1 to n), the sample after truncation (i=1 to T), and the truncated sample (i=T+1 to n).

$$\beta_n = \frac{\mathrm{Cov}_n\,(x,\,y)}{\mathrm{Var}_n\,(x)} = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{T} x_i\, y_i + \sum_{i=T+1}^{n} x_i\, y_i}{\sum_{i=1}^{T} x_i^2 + \sum_{i=T+1}^{n} x_i^2}$$

$$= \frac{\sum_{i=1}^{T} x_i\, y_i}{\sum_{i=1}^{T} x_i^2} - \frac{\left(\sum_{i=1}^{T} x_i\, y_i\right)\left(\sum_{i=T+1}^{n} x_i^2\right)}{\left(\sum_{i=1}^{T} x_i^2\right)\left(\sum_{i=1}^{n} x_i^2\right)} + \frac{\sum_{i=T+1}^{n} x_i\, y_i}{\sum_{i=T+1}^{n} x_i^2} - \frac{\left(\sum_{i=T+1}^{n} x_i\, y_i\right)\left(\sum_{i=1}^{T} x_i^2\right)}{\left(\sum_{i=T+1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} x_i^2\right)}$$

$$= \beta_T - \beta_T \left(\frac{\left(\sum_{i=T+1}^{n} x_i^2\right)}{\left(\sum_{i=1}^{n} x_i^2\right)}\right) + \beta_h - \beta_h \frac{\left(\sum_{i=1}^{T} x_i^2\right)}{\left(\sum_{i=1}^{n} x_i^2\right)}$$

$$= \beta_T \left(1 - \left(\frac{\left(\sum_{i=T+1}^{n} x_i^2\right)}{\left(\sum_{i=1}^{n} x_i^2\right)}\right) + \beta_h \left(1 - \frac{\left(\sum_{i=1}^{T} x_i^2\right)}{\left(\sum_{i=1}^{n} x_i^2\right)}\right)\right)$$

$$\beta_n = \beta_T \frac{\left(\sum_{i=1}^{T} x_i^2\right)}{\left(\sum_{i=1}^{n} x_i^2\right)} + \beta_h \frac{\left(\sum_{i=T+1}^{n} x_i^2\right)}{\left(\sum_{i=1}^{n} x_i^2\right)}$$

$$\beta_T = \beta_n \frac{\left(\sum_{i=1}^{n} x_i^2\right)}{\left(\sum_{i=1}^{T} x_i^2\right)} - \beta_h \frac{\left(\sum_{i=T+1}^{n} x_i^2\right)}{\left(\sum_{i=1}^{T} x_i^2\right)}$$

The analysis above suggests that bias caused by truncation is impacted by the beta coefficient on the truncated observations. In most cases, the direction of the β_n and β_h will be the same meaning that truncation will bias coefficients towards zero except for cases where β_h is of a different sign than the underlying parameter value.