Influential Observations and Inference in Accounting Research

Andrew J. Leone
University Miami
a.leone@miami.edu

Miguel Minutti-Meza
University of Miami
mminutti@bus.miami.edu

Charles Wasley
University of Rochester
wasley@simon.rochester.edu

October 2013
(First Draft: April 2012)

Abstract

The existence of potentially influential or outlier observations is ubiquitous in empirical accounting research. The purpose of this paper is to summarize the various methods used by accounting researchers to identify and control for influential observations; assess the effectiveness of such methods and of alternative methods from the statistics literature; and, provide guidance for future studies so they may systematically identify and mitigate the impact of influential observations. A survey of articles published in accounting journals shows considerable variation in the way researchers deal with influential observations prior to estimating a regression model. We demonstrate that the common approach of winsorizing each variable has only a modest impact on parameter estimates compared to "doing nothing". We also demonstrate that truncation tends to bias coefficients toward zero. More generally, we show that both winsorizing and truncating do little to accommodate potential bias caused by unusual and infrequent events (i.e., data points that are not random errors). Alternatively, use of robust regression significantly reduces such bias and is preferable to winsorization or truncation when regression is used for hypothesis testing. In addition, robust regression methods are available in commonly-used statistical packages and they do not entail the *ad hoc* choice of winsorization or truncation rules, thus providing a convenient way to control for influential observations and enhance inter-study comparability.

We thank workshop participants at the University of Arizona, University of California, San Diego, University of Chicago, Florida Atlantic University, University of Maryland, Texas A&M, Pennsylvania State, Tilburg University and University of Toronto Accounting Conferences. Special thanks go to Ray Ball, Christian Leuz, Jeff McMullin, Dan Taylor, Tojmme Rusticus, and Jerry Zimmerman for helpful comments.

1. Introduction

This study examines the statistical problems related to the presence of influential observations in regressions estimated using accounting and market data. The objectives of this study 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 alternative 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. A common practice in accounting studies is to truncate or winsorize extreme values of each variable prior to estimating a regression model. This study aims to redirect researchers' attention from extreme values to overall model fit (i.e. regression residuals) when attempting to identify and treat influential observations.

In our review of the accounting literature, we find significant variation in the methods used to account for potentially influential observations, but the majority of studies either winsorize or truncate data. Winsorizing or truncating outliers, which are data points located unusually far from the mean of the sample, is a reasonable approach when extreme values are likely caused by data errors. However, when extreme values are generated from the underlying data-generating process, *ex ante* truncation or winsorization (prior to estimating regressions) can lead to either less efficient or biased estimates, depending the source of the extreme values. Suppose, for example, the "true" relation between variables x and y (x,y), both with standard normal distributions, is y=2x+error. Now consider two data points, (5,10), and (-5,10). If a researcher winsorizes observations at 2 standard deviations from the mean, then the data points would be transformed to (2,2) and (-2,2), respectively. In the first case, (5,10), winsorizing reduces efficiency in a regression of y on x, since the data point represents a "good leverage

point."[1] In the second case, winsorizing reduces but does not eliminate the influence of this "bad leverage point."[2] This observation (transformed to -2, 2), will "pull" the slope of the regression line down.

An alternative approach is to focus on *influential* observations, determined by their influence on parameter estimates from regression procedures. *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.). Methods for identifying influential observations include Belsley et al. (1980), and more recently, robust regression procedures. Returning to the simple example above, these methods would identify the data point, (10,5) as a "good leverage" point because the relation between y and x is consistent with the bulk of the data. On the other hand, (10,-5) would be identified as an influential observation and it would be dropped from the sample in estimating the relation between y and x.

By identifying influential observations, researchers can learn more about the data generating process of the dependent variable. In the case of the data point (10,-5), the large value of y might have been generated by an omitted variable. Therefore, identification of this influential observation can help the research improve model specification (e.g., add additional variables to the model). As 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* (omitted variables) affecting a

[1] A good leverage point is one where an extreme value of x, occurs in combination with an extreme value of y along the true regression line.
[2] Bad leverage points occur when an extreme value of x occurs in combination with a value of y that is far from the true regression line.

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.[3] If outliers result from measurement error, researchers can correct, discard, or adjust observations to fit between a lower and an upper bound. In the second case, unusual data points are usually not random and 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.[4]

In the OLS framework, ignoring the underlying causes of these unusual observations *influences* the model estimates and 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 statistical analyses. Moreover, researchers should attempt to mitigate the effect of these influential observations 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 influential observations.

[3] 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 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%."

[4] 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.

Collectively, the issues described above provide part of our motivation to assess the efficacy and trade-offs associated with various methods to address the inference problems related to the occurrence of influential/outlier observations in the samples encountered in accounting research. We use a combination of simulations to compare the performance of winsorization, truncation, and robust regression in three general settings: (1) when the data does not contain influential/outlier observations; (2) when the data contains influential/outlier observations occurring at random; and, (3) when the data contains influential/outlier observations resulting from events that are correlated with both the dependent variable and the explanatory variables.

We document that robust regression based on MM-estimation has the most desirable characteristics when it comes to treating influential/outlier observations when compared to other approaches commonly used in the accounting literature.[5] First, when influential/outlier observations are uncorrelated with the independent variables, MM-estimation yields parameter estimates that are unbiased and identical to those estimated under OLS. Second, MM-estimation is the most effective at mitigating bias induced by a correlation between influential/outlier observations and both the dependent variable and the explanatory variables. Third, MM-Estimation has up to 95% efficiency relative to OLS meaning that Type II errors are unlikely to be a consequence of its use. Finally, our simulation results supporting the use of robust regression methods are confirmed by a replication of a published accounting research study, namely the differential persistence of various types of accruals in Richardson et al. (2005).

In contrast to robust regression based on MM-estimation, alternatives commonly found in the accounting literature are much less effective and may even induce bias. When influential/outlier observations are randomly distributed in the data, winsorizing at the top and

[5] Huber (1964, 1973) proposes a method named M-estimation. The M in M-estimation stands for "maximum likelihood type". MM-estimation is a subsequent approach proposed by Yohai (1987). More details about this approach are provided in Section 4.

bottom 1% yields unbiased estimates. However, winsorizing is ineffective at mitigating bias induced by influential/outlier observations 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. In contrast and as noted in Kothari et al. (2005), truncation based on extreme values of the dependent variable generates a downward bias in the estimated coefficients. None of the other alternatives we evaluate generates a bias when influential/outlier observations are randomly distributed. However, when influential/outlier observations are correlated with an independent variable, truncation exhibits less bias than winsorizing (or doing nothing), but the bias is even less using robust regression.

In general, robust regression can be used in any situation where researchers use OLS. If there are no influential/outlier observations, both estimation methods will produce similar coefficients. However, in the presence of influential observations (and assuming the researcher wants to estimate a model yielding the most reliable inferences), use of robust regression is a viable alternative to winsorization or truncation. This is because it provides a well-founded statistical compromise between including all the data points and treating them equally in OLS regression versus excluding observations entirely from the analysis. The results of our study should enable accounting researchers to make informed research design choices when it comes to mitigating the influence of influential/outlier observations on the inferences they draw about their hypotheses. Robust regression based on MM-estimation yields estimates that are resistant to influential/outlier observations that are randomly distributed throughout the data, in addition to significantly reducing the bias created by influential/outlier observations that are correlated with the independent variable of interest. Moreover, MM-estimation is highly efficient relative to

OLS, and perhaps more importantly, is not subject to biases induced by winsorization or truncation.[6]

The remainder of the paper is organized as follows. Section 2 reviews the accounting literature to document the approaches used to account for influential/outlier observations; Section 3 discusses the consequences of influential/outlier observations in the context of OLS estimation; Section 4 discusses robust estimation; Section 5 outlines our simulation analyses and reports the related results; Section 6 compares the approaches to account for influential/outlier observations by replicating a published accounting study; Section 7 discusses implementation of robust regression methods in accounting settings, and Section 8 concludes.

2. Literature review: how are influential/outlier observations treated in accounting research studies?

2.1 Background

Accounting archival studies typically aim to test a theory or hypothesis about the relation between one or many causal/explanatory variables (the *x's)* and an outcome or dependent variable *y*.[7] An issue that is well-known in accounting dating back to Ball and Foster (1982) is the difficulty of drawing causal inferences in quasi-experimental settings like those typically encountered in empirical accounting capital markets research. As highlighted by Cook and Campbell (1979, p. 37, see also Ball and Foster, 1982) "Accounting for the third-variable alternative interpretations of presumed (causal) relationships is the essence of internal validity." Cook and Campbell (1979, p. 56) also note an important shortcoming of non-experimental data: "Instead of relying on randomization to rule out most internal validity threats, the investigator

[6] We stress that using robust regression is not a substitute for careful analysis of the data and thorough statistical modeling. As noted in the statistics literature, ultimately influential/outlier observations may provide interesting case studies and should be identified and their implications for inferences discussed.

[7] For example, researchers have tested hypotheses about the effect of firm characteristics such as disclosure quality, size, profitability, and leverage on dependent variables such as returns, discretionary accruals, analysts' forecast errors, management compensation, and audit fees.

has to make all the threats explicit and then rule them out one by one." With regard to drawing causal inferences in empirical accounting research settings, in their review of corporate financial accounting research, 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."

It is generally accepted that accounting and stock return data contain extreme values and an early paper by Kennedy et al. (1992) examined alternatives to adjust for outliers in regression models. Intuitively, extreme data values in accounting are caused by unusual/infrequent events (e.g., write downs, new block-buster products, etc., see Appendix A for examples) rather than primarily by data errors. Of importance is that in some cases even just a few influential/outlier observations can bias inferences. For example, Guthrie et al. (2012) demonstrate that previous results by Chhaochharia and Grinstein (2009) are driven by two observations in a sample of 865 firms.

Extreme values influence parameter estimates for even the most basic relationships. For this reason, researchers often take measures to deal with potential extreme values in their data. To illustrate, Figure 1 provides a plot of three-day cumulative abnormal returns on quarterly earnings forecast errors (CARs).[8] Figure 1A is a plot of all observations in the sample using the raw data. Figure 1B is the same data after winsorizing CARs and forecast errors at 1% and 99%. Figure 1C is the data after winsorizing CARs and forecast errors at 5% and 95%. The slope of the regressions (i.e., the ERCs) for Figures 1A, 1B, and 1C are 0.00002, 0.4088, and 1.618,

[8] The sample underlying the figures is as follows. We begin with all quarterly EPS forecasts on IBES's detail file (WRDS dataset: DET_EPSUS) from 2005-2011. We use the median of the most recent forecast by all analysts making forecasts less than 90 days before the earnings announcement to calculate earnings forecast errors as Actual EPS - Median EPS forecast, scaled by stock price on the day prior to the return accumulation period (four trading days prior to the earnings announcement). CARs are three-day abnormal returns from the day before through the day after the announcement where abnormal returns equal raw returns minus the value-weighted market return.

respectively. These plots provide initial compelling evidence that the results of even the most basic ERC regression are strongly influenced by extreme observations and, consequently, by the choice of whether or not to, and how a researcher chooses to winsorize (or truncate) the data.

2.2 How are influential/outlier observations treated in accounting studies published between 2006 and 2010? [9]

To systematically document how accounting researchers choose to account for influential/outlier observations, we reviewed 857 studies published between 2006 and 2010 in Contemporary Accounting Research, Journal of Accounting Research, Journal of Accounting and Economics, Review of Accounting Studies, and The Accounting Review. The studies examined span a variety of areas including auditing, analysts' forecasts, management compensation, earnings management, conservatism, taxes, disclosure, and the earnings-returns relation.

As shown in Table 1, Panel A, 69% (590) of the studies are archival with the remaining 31% split between 12% (101) analytical, 12% (106) experimental, and 7% (60) discussion and review studies (studies including both an analytical model and empirical tests are classified as archival). We searched the body, footnotes, and tables of each archival study for discussion of the treatment of influential observations/outliers. Only 68% of the archival studies (404 of 590) mention the presence of extreme observations/outliers or describe any procedures to deal with such observations (the percentage of studies not mentioning outliers ranges from 20% in RAST to 37% in JAR and CAR). We recognize that our classification of studies is subject to some limitations. For example, the procedures used to deal with extreme observations/outliers are not

[9] The corporate finance literature has also recognized the importance of influential/outlier observations in capital markets data. For example, in a review paper on capital structure Frank and Goyal (2005, 172) note, "The standard data sources such as Compustat have a nontrivial number of observations that seem quite anomalous. For instance, data items that by definition cannot be negative are sometimes coded as negative. Sometimes data items are coded in ways that result in the balance sheet not balancing or the cash flow identities not matching up. In some cases, a firm will have a value of some variable that is several orders of magnitude too large to be plausibly correct." Frank and Goyal (2005) also note "It is particularly common to winsorize each tail at 0.5% or 1%."

mutually exclusive; some studies use dichotomous dependent variables, while many studies use more than one regression model. Nevertheless, we believe our review serves at least two useful purposes. First, it organizes in a systematic way how the treatment of extreme observations/outliers varies across the literature. Second, it provides evidence demonstrating that there is need for greater consistency in the treatment for extreme observations/outliers in accounting research.

Table 1, Panel A shows the breakdown of archival studies using winsorization, truncation, and other procedures. The most common solutions to address the presence of influential/outlier observations are winsorization and truncation, with 88% of the studies using one of these procedures separately or combined. As seen from Panel A, winsorization is the most common procedure used to deal with influential/outlier observations with 55% of the studies (221 of 404) winsorizing at least one variable. Winsorization alters the original data by imposing an upper and lower bound on influential/outlier observations by setting them equal to a researcher-specified percentile of the distribution (with most studies using the top and bottom 1%).

Table 1, Panel B shows the breakdown of studies using winsorization of continuous variables. As seen from Panel B, of the 221 studies that apply winsorization, 151 (68%) winsorize the dependent variable, 202 (91%) winsorize at least one independent variable, and 132 (60%) winsorize a combination of the dependent and independent variables. It is worth noting that 29 studies use a general winsorization rule for most variables, but an *ad-hoc* rule for a subset of variables. Examples of the *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; winsorizing all variables except the log of total assets; and setting extreme values of effective tax rates to one and zero.

Table 1, Panel C shows the breakdown of studies that use truncation on continuous variables. Truncation is the second most common procedure with 40% of the studies (161 of 404) truncating at least one variable. While truncation also imposes an upper and lower bound on the data, it discards observations beyond researcher-specified cut-offs. The cut-offs can be 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. Of the 161 studies truncating observations, 143 (89%) truncate the dependent variable, 139 (86%) truncate at least one independent variable, and 121 (75%) truncate a combination of the dependent and independent variables. It is worth noting that 59 studies truncate observations based on a rule different than a percentile cut-off. For example, truncation of earnings scaled by total assets at plus and minus three, or truncation of firms with stock prices less than $5.

With regard to other approaches to account for influential/outlier observations we find that 13% of studies reviewed not using winsorization or truncation instead employ a diverse set of procedures including using ranks of the dependent and/or independent variables, log transformations, and some forms of robust regression (e.g., median regressions, least trimmed squares, and MM-estimation).

Overall, the procedures used to identify and adjust for influential/outlier observations in accounting settings are, for the most part, inconsistently applied across studies. Between the studies using winsorization and truncation we identified 88 studies that used an *ad-hoc* rule for a subset of variables. We also identified 38 archival studies where the procedures employed were unclearly stated. For example, some studies noted that they used the Belsley et al. (1980) regression diagnostics, including standardized residuals or Cook's D to truncate observations with large residuals, but many of these studies do not report the cut-offs employed or provide

enough detail regarding the procedures employed. Such inconsistencies make it difficult, if not impossible to replicate these studies, to make comparisons with the results of similar studies, or to reconcile conflicting results from similar studies.

2.3 Influential/outlier observations and the skeweness of stock returns

The appropriate treatment of extreme observations is a source of debate in papers examining market efficiency with respect to accounting numbers. At the center of the debate is the extreme positive skewness in stock returns and the use of those returns as the dependent variable. For example, Kraft et al. (2005) examine the association between stock returns and accruals and show that the accrual anomaly is influenced by a very small number of extreme observations. Kraft et al. (2005) argue that the decision as to how to treat influential/outlier observations should depend on the purpose of the analysis. For example, if researchers are interested in testing a theory of the market pricing of accruals, then their model should fit the bulk of the data, and not reflect the effect of a relatively small number of extreme observations. On the other hand, if the purpose is to test a trading strategy and/or to predict future returns, then the only reason to truncate influential/outlier observations is the possibility that they are data errors. While data errors in returns are rare, they do occur (see footnote 1). Beyond that, Kraft et al. (2006, p. 307) note 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 the 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."

Other studies caution against the truncation of stock returns, for example using simulation 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), while Core (2006, p. 350) states, "in general, deleting based on the robust regression techniques employed and advocated by KLW seems inappropriate… deleting extreme observations from skewed return data leads to biased estimates and can bias inferences." In addition, 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 instead conclude that the association between accruals and stock returns is robust to excluding influential/outlier observations, at least in a sub-sample excluding loss firms. Notwithstanding these studies, as shown below, the results of this study demonstrate that a combination of right skewness of stock returns and truncation based on large realizations of stock returns leads to biased inferences.

The seemingly conflicting views enumerated above have led researchers in a variety of directions when it comes to accounting for the effect of influential/outlier observations. Evidence in support of this conclusion is that our review identified 157 studies using stock returns as the dependent variable (at least in one regression model) where 53% (83) of such studies winsorized or truncated extreme stock returns while the remainder 47% (74) used the raw data. As discussed above, the objective of this study is to propose and calibrate an alternative technique to address the inference problems associated with influential/outlier observations without having to eliminate such observations from the data or having to resort to variable-by-variable winsorization or truncation rules.

In contrast to the papers just noted, which argue against the truncation of skewed stock returns, numerous studies in the analyst forecast error literature truncate forecast errors. The

distribution of forecast errors is left skewed (i.e., has large negative errors) and forecast errors are used both as dependent and independent variable in a variety of studies. Abarbanell and Lehavy (2003, p. 114) highlight the issue that "many studies implicitly limit observations in their samples to those that are less extreme by choosing ostensibly symmetric rules for eliminating them, such as winsorization or truncations of values greater than a given absolute magnitude." Abarbanell and Lehavy (2003) argue "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."

3. Influential/outlier observations and OLS estimation

3.1 A framework for examining the effect of extreme/influential/outlying observations

An OLS framework is the most widely used methodology used to assess relations between variables in accounting research studies. In a simple OLS regression we have:

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

with the expected value of *y* given *x* as:

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

Since parameters are estimated by minimizing the sum of squared errors, $\hat{\beta}$ is the mean effect of *x* on *y,* but not necessarily the "typical" effect. Since OLS parameters are based on the conditional mean of the dependent variable they suffer from the same problems as the mean itself, which is that they are susceptible to the effects of influential/outlier observations.

In theory, to address the effect of influential/outlier observations on the estimated parameters of an OLS model researchers should consider what causes the extreme values of *x* and *y*. Along these lines, the distribution of the values of *y* will depend on the researchers' sample selection procedure, the distribution of the *x's*, and additional unknown sources of

variation captured by the error term. Extreme values of *y* can be caused by a number of things including extreme values of the hypothesized *x's* that influence *y*; nonlinearities in the relation between *x* and *y*; or additional variables that also influence *y,* but are unknown to the researcher or omitted from the model. Of importance at the outset of the discussion here is that not all extreme observations have the same effect on the estimated OLS parameters. In particular, it is important to differentiate (as we do below) among four different statistical concepts/effects: univariate outliers, regression or multivariate outliers, leverage points, and influential observations.

3.2 Outliers as a correlated omitted variable problem

As discussed above, potential influential/outlier observations in either the dependent or independent variables can be the result of data errors or the result of other problems. For example, influential/outlier observations in the dependent variable can arise from skewness in the independent variables or from differences in the data generating process for a small subset of the sample. Influential/outlier values of the dependent variable caused by skewness in the independent variable, called good leverage points, are not necessarily problematic because such extreme values of *y* are generated by large values of *x*. That said, potential inference problems are caused by extreme values of *y* not explained by *x*. Such observations may be the result of a different data generating process, for example, the result of an unknown or omitted variable that frequently takes on a value of zero, but which occasionally takes on a different value and when it does so it has a major impact on *y*. In accounting settings such observations may be either one-time events or idiosyncratic firm characteristics. As an example based on accruals, INSMED, Inc., a medical technology company had total accruals scaled by total assets of 206% because it reported a gain of $123 million on the sale of technology to Merck and had total assets of only

$4 million at the start of the year. Such infrequent yet significant events also occur in stock returns. For example, OSICOM Technologies earned a buy-and-hold abnormal return of 459% between 6/1/1995 and 5/31/96, a significant amount of which is attributable to agreements it announced during the year. Specifically, 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 (in fact, OSICOM had a buy-and-hold return of 46% for the two days May 30 and 31).

Depending on their magnitude, ignoring low-frequency 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. Additionally, holding the dollar value of any event constant, the impact of the event is likely to be much more significant for smaller firms. For example, the $123 million dollar gain reported by INSMED, Inc., had a dramatic impact on its accruals (206%), but had Merck reported the same gain, its accruals would have increased by only 2%. The key point is that infrequent events like these are likely to be correlated with firm size, and firm size is often correlated with variables commonly used to test hypotheses in accounting settings.

To provide some evidence on the pervasiveness (or frequency) of the problem in actual accounting data, Figure 2 plots the relation between total accruals and total assets. The purpose of Figure 2 is to use real accounting data to illustrate the correlation between infrequent events and the behavior of variables commonly used in accounting research (e.g., accruals). Using COMPUSTAT data from 1972-2001 for all firms with stock prices in excess of $5 we construct 50 bins based on total assets. Figure 2 reports box plots of accruals for each bin. The key takeaway from Figure 2 is that (not surprisingly) extreme accruals occur more frequently in

smaller firms (see Appendix A for actual examples of extreme accruals and stock returns). The challenge for accounting researchers interested in testing hypotheses about accruals is retaining those 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 events that are not included in the model. In the following section we discuss robust regression techniques that are intended to do precisely this.

4. 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 standard errors that account for heteroscedasticity and/or error dependence. For our purposes, 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 of circumstances" (see 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. 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 and 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 to the presence of influential/outlier observations under certain conditions

The OLS method to estimate 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 squared residuals, OLS gives excessive importance to observations with very large residuals. In terms of the definitions discussed above, OLS has *unbounded influence*. In fact, even a single outlier can have a significant impact on the fit of the regression line/surface, which means the *breakdown point* of OLS is zero. In addition, extreme/ outlying observations can also be associated with non-constant error variance, violating one of the OLS assumptions, causing the OLS estimates to lose efficiency because they give equal weight to all observations. Stated differently, OLS weights influential/outlier observations equally, even when the influential/outlier observation(s) contain less information about the true relation between x and y.

4.3 Types of robust regression estimators

Robust regression methods 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, Least Median Squares LMS, and Least Trimmed Squares LTS), R-estimators, S-estimators, M-estimators, GM-estimators, and MM-estimators.[10]

[10] LTS estimation was proposed by Rousseeuw (1984); M-estimation by Huber (1964, 1973); S-estimation by Rousseeuw and Yohai (1984); and, MM-estimation by Yohai (1987). The discussion in this section is based on the

In general, L-estimators rely on minimizing a modified version of the sum of squared residuals criteria, such as the sum of the absolute values of the residuals (LAV), the median of the squares residuals (LMS), or 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. R-estimators rely on minimizing the sum of a score of the ranked residuals, but most R-estimators have low breakdown points. S-estimators take a different perspective, focusing on the minimum variance property of the OLS estimators. S-estimators minimize a measure of the dispersion of the residuals that is less sensitive to influential/outlier observations than the OLS variance. However, these estimators have very low efficiency compared to OLS.

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, which means that 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 bi-square functions. The final weights are informative and can be used to identify which observations are extreme. The general criteria to be minimized is:

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

Note that OLS can be considered a special case within this class of estimators where the square function is used to weight the residuals. MM-estimators are widely used and combine a high breakdown point with high efficiency (MM-estimation has up to 95% efficiency relative to

reviews of robust methods in Andersen (2008); Maronna, Martin and Yohai (2006); and, Fox and Weisberg (2010) (see also, Chen (2002), and Verardi and Croux (2009)).

OLS and it is usually pre-set to 85% in MM-estimation commands in STATA, SAS and R).

MM-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 estimation procedure follows these steps: (1) first pass coefficients are calculated using some form of resistant regression (usually the S-estimator with Huber or bi-square 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% from the previous iterations.[11]

5. Simulation analysis and procedures

5.1 Overview

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 coefficients if they are correlated with *x*. In this section we use Monte Carlo simulations to evaluate the relative effectiveness of methods commonly used in accounting research to account for such bias and its related effect on the inferences drawn from the analysis. As explained before, the approaches used in accounting range from "do nothing" to winsorization and truncation. A key aspect of the simulations is that we used known conditions of the data to compare the previously mentioned approaches with robust regression based on MM-estimation

[11] We are only able to identify a handful of accounting studies using robust regression (see Aboody 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 contains excerpts from these studies describing the procedure used. In most cases, the procedures were mentioned in a footnote without enough information for us to replicate them. Appendix C provides guidance to implement robust regression in commonly-used statistical packages.

to see which performs best at mitigating the impact of influential/outlier observations on coefficient estimates and overall inferences.

To perform the simulations we specify a data generating process for a dependent variable *y* consisting of a variable of interest *x,* and a variable *z* that is zero for most of the sample. Conceptually, *z* can be thought of as an infrequent event that generates extreme values of *y* when it occurs. To model the 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* (which are caused by *z*) are not properly dealt with by the underlying model.

Equation (4) describes the data 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, $v \sim N(3,1)$, and $e \sim N(0,1)$. By construction, *z* is zero approximately 98% of the time. We assume *z* is not observed by the researcher and, therefore, leave the generated value as is. The variable *z* is continuous to simulate a data generating process where extreme/ outlying observations (1) occur as a result of a correlated omitted variable; (2) exhibit variation in magnitude with respect to the bulk of the data; (3) the *x* and *y* variables have smooth distributions; (4) the distribution of *y* is skewed.[12]

For simplicity and without loss of generality, we set □□= 0, □□ 0, set out□□ 0,.0, except when we test for potential Type I (Type II) errors where we set □□whe, and □□= 0. In our tests we generate 250 samples of 2,000 observations and for each sample we estimate

[12] Although a single large influential/outlier observation can significantly bias OLS estimates, if there is a fixed number of influential/outlier observations, increasing the total number of observations would reduce the individual influence of each. We simulate a data generating process where influential/outlier observations do not occur at random, but rather occur as the result of a correlated omitted variable. The percentage of influential/outlier observations in our simulated datasets remain constant resulting from a correlated omitted variable, which means increasing the number of observations would not help in reducing the influence of such influential/outlier observations. This feature is consistent with accounting panel datasets where increasing the number of years adds more observations with large values of the *x* and *y* variables.

regressions employing the following alternative approaches to accounting for potential influential/outlier observations:

1) "Do nothing." As shown by our literature review, approximately one third of published archival papers do not report addressing the existence of influential/outlier observations.
2) Winsorize *y* and *x* at the extreme 1% of their distribution.
3) Winsorize *x,* but leave *y* as is.
4) Truncate *y* and *x* at the extreme 1% of their distribution.
5) Truncate *x,* but leave *y* as is.
6) Leave *x* and *y* variables as they are and use robust regression based on MM-estimation.

5.2 Graphical illustration of alternative identification and treatment of influential/outlier observations

Before reporting our main simulation results, we construct a sample of 4,000 observations and generate plots to illustrate the impact of alternative approaches to treat influential/outlier observations. Figure 3a presents a plot of 4,000 observations where *y* is generated from the following data generating process:

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

In this case, since *z* is correlated with *x* and *y*, if left unaccounted for (i.e., under the "do nothing" approach) the influential/outlier observations will bias the coefficient on *x*. The circles in the plot in Figure 3 are the observations where *z* is nonzero. As expected based on the data generating process underlying *y*, these "shocks" occur for roughly 2% of the sample and only when *x* is extreme. As can be seen from the plot, these observations will induce an upward bias in β. Consistent with this, the estimated β for this sample (represented by the black line) is 0.89 (from Eq. 5 the expected value of □ □should be 0.80).

To further illustrate the consequences of the approaches adopted in many accounting studies to deal with influential/outlier observations, Figure 4a uses the same data as Figure 3, but

adds lines representing the top and bottom 1% of *x* and *y*. The key takeaway from the plot is that while winsorizing or truncating will reduce the influence of *z* to a small degree, it will also reduce the influence of good leverage points. To better appreciate this point, Figure 4b plots the data after winsorizing *x* and *y*. The key feature of the Figure is that winsorizing does little to reduce the impact of *z* on *y* because, while magnitudes of the *y*'s are reduced, many are retained and are still large relative to the rest of the sample. Consistent with this, estimation of Eq. (5) on the winsorized data yields a slope coefficient of 0.87, which reduces the bias caused by *z*, but only by 0.02. Figure 4c plots the data after winsorizing *x,* but not *y*. As documented by our review of the literature (see Table 1) this is a fairly common practice in accounting, particularly when *y* is returns. Unfortunately, this approach generally makes the problem worse than doing nothing at all because it can bias the coefficient away from zero. For example, assume for a given observation that e = 0, $x = 5$, $z = 0$, $y = 4$ ($0.8*x$), and that *x* is winsorized at 2. The result is an *artificial* error of 2.4 (4 - 0.8*2) that biases the regression line upwards. More precisely, the estimated slope coefficient if only *x* is winsorized is 0.91, which is greater than the estimate of 0.89 in the "do nothing" case (see Figure 3) and also greater than the true value of β equal to 0.80.

Following the approach used in Figures 4a, 4b, and 4c to visually illustrate the effects of winsorization as a means to account for potential influential/outlier observations, Figures 5a and 5b adopt a similar approach to illustrate the effects of truncation on the data and the resulting inferences about the coefficients. To save space, we only discuss the main features and takeaways of these Figures. Figure 5a plots the same data as before but after truncating extreme values of *x* and *y* (i.e., the top and bottom 1% of their underlying distributions). This truncation rule (which is not uncommon in accounting research settings) yields a parameter estimate of

roughly 0.77, evidence of a downward bias given that the true β is 0.80. Following, Figure 5b illustrates a truncation rule applied just to extreme values of x and demonstrates that truncating only the independent variable does not reduce the bias caused by z. For example, as illustrated by the circles in this figure, many of the extreme values of y caused by z will still remain in the sample and the estimated slope coefficient is 0.90.

Turning to robust regression based on MM-estimation, Figure 6 illustrates the observations considered to be influential/outlier observations, which will be down-weighted when robust regression based on MM-estimation is used. In the Figure the diamonds represent data points with extreme values of y caused by z considered to be the influential/outlier observations; the circles symbolize data points that might be considered influential/outlier observations, but which are *not* influenced by z (i.e., when z is zero); while the triangles are data points with extreme values of y that are influenced by z (i.e., when z is not zero), but are not considered influential/outlier observations. The key takeaway is that the advantage of a robust estimation procedure (like MM estimation) is that good leverage points are retained, while other potential influential/outlier observations get down-weighted. In other words, extreme values of y caused by extreme values of x are retained which increases efficiency relative to a naive winsorization or truncation rule. As a result, robust estimation has the advantages of being both consistent and efficient relative to naïve winsorization and/or truncation rules.

5.3 Simulation Results

Baseline simulation results are reported in Table 2, Panel A where influential/outlier observations are generated by an infrequent event that is randomly distributed and independent of x. We generate values for y following the data generating process specified in Eq. (4) except that instead of extreme events z occurring only when x is in the top decile, z occurs with equal

probability (2%) across the entire x distribution. Values of *y* are generated with □ being set to 0.8 or zero and □ being set to 1.0 or zero. In the first three regression models reported in Panel A (first 3 rows of Panel A), *z* is omitted from the estimation to simulate a typical case where an infrequent event is omitted from the researcher's model. In a second set of estimated regressions (the 4th through 6th rows of Panel A) z is included in the regression, the ideal solution of course, if the researcher could in fact identify such cases. The table reports mean estimates of □ for 250 samples of 2,000 observations under the alternative treatment of influential/outlier observations. Bias is the difference between the "true" parameter value and the mean estimate of □.

The benchmark case is "do nothing" where OLS is estimated without truncation, winsorizing, or down-weighting outliers. In this case where the influential/outlier observations are randomly distributed, not surprisingly, estimates of □ are unbiased in all cases. In addition, when the influential/outlier observations are randomly distributed, the results are virtually identical when *x* and *y* are winsorized at the top and bottom 1%. This consistent with Kothari et al. (2005) who find that truncating *y* and *x* imparts a downward bias, but only when there is a correlation between *x* and *y*.

Turning to the case where *y* is generated when □ = 0, truncation does not impact □ (i.e., truncation does not bias □ to a value less than zero). The intuition for this is that the truncated observations are not influential/outlier observations generated by extreme values of *x* (because *x* is uncorrelated with *y*). This means that no good leverage points that might induce a relation between *x* and *y* are lost (see Appendix D for an analytical representation of bias caused in parameter estimates as a result of truncation).

In contrast, when *x* is correlated with *y* and the true □tis 0.8 the estimated coefficient is biased down, as expected. Specifically, as shown under the heading "truncate" in Panel A, the

estimated is biased downward by amounts ranging from 0.03 to 0.06 depending on the specification. While this is unlikely to impact significance levels in our simulations since is set to 0.8, truncation can generate Type I or Type II errors when the true value of fis closer to zero. These findings suggest truncation of influential/outlier observations should be avoided in favor of other alternatives.

As with winsorizing in the baseline case, robust regression yields unbiased estimates in all specifications. Moreover, there is no evidence that robust regression biases coefficients when influential/outlier observations occur randomly.

In contrast to Panel A, Panel B reports results where the infrequent events *z*, are correlated with *x*. In our analysis, such events only occur when *x* is in the top decile of its distribution. As a concrete example, *x* might be negatively correlated with size and *z* tends to occur only for small firms. As we shall show, these non-random influential/outlier observations are a concern for researchers attempting to draw inferences about how *x* influences *y*.

Referring to Panel B, when we set = 1.0 to generate the *y* values, we have a classic correlated omitted variables problem that will bias estimates of when the influential/outlier observations are ignored. More specifically, when influential/outlier observations are untreated/ignored (i.e., the "do nothing" case), is biased upwards by 0.10 (see the second regression model in Panel B). Even in the case were *x* is uncorrelated with *y* by construction (i.e., = 0, see the third regression model in Panel B), the mean coefficient estimate is 0.10 (i.e., biased upward). Examination of the results under winsorizing (see the "Winsorize" column) reveals that winsorization does virtually nothing to reduce the influence of the correlated influential/outlier observations (i.e. the *z's*). Moreover, the coefficient of 0.09 is almost identical to that (i.e., 0.10) in the "do nothing" case. Turning to truncation, it appears to mitigate the bias,

but this is largely an artifact of truncation's tendency to bias parameter estimates downward. The final two columns of Panel B report the robust regression results where we find that this approach leads to an 80% reduction in the bias caused by the influential/outlier observations (i.e., 0.10 versus 0.02). While the bias is not completely eliminated (i.e., = 0.02), its impact is substantially reduced. In untabulated results, we find that the bias created by the influential/outlier observations is significantly different from zero in 6% of the 250 samples ($p < 0.01$) under the robust regression alternative compared to 90% of the time when influential/outlier observations are winsorized.

To give a sense for the sensitivity of our results to parameter choices, Figures 7 and 8 demonstrate the impact of the bias for alternative values of □□ (i.e., different degrees of association between *x* and *y*). In Figure 7, the *y* values are generated with *y*=1; with alternative values of □ varying from -0.8 to +0.8; and with shocks that are randomly distributed and independent of *x*. These conditions imply that the expected bias in □ is zero. As Figure 7 illustrates, the bias is, in fact, zero for all methods except truncation. Under truncation, the bias is inversely related to the underlying value of □. This is because truncation biases the coefficient towards zero, but the impact of the truncation bias approaches zero as the true relation between *y* and *x* approaches zero. This evidence suggests that in this setting truncating on *y* and *x* should not cause the Type I errors as suggested by Kothari et al. (2006).

Figure 8 is similar to Figure 7 except that the shocks are correlated with *x*, causing a correlated omitted variables bias of approximately 0.10. As reported in Table 2, winsorizing does little to mitigate the bias while robust regression reduces the bias to 0.02. From Figure 8 we can see that the bias reported in Table 2 does not vary across parameter estimates for winsorization or robust regression. However, the Figure shows that robust regression is preferable to

winsorization despite the strength of the association between *y* and *x* and that robust regression is preferable to truncation over a large range of associations between *y* and *x* (-0.8 to +0.3). Moreover, when the association between *y* and *x* is strongly positive (+0.5 to +0.8) truncation is slightly better than robust regression (recall that the bias is inversely related to the parameter estimates when truncation is used). However, without knowledge of the true underlying association between *y* and *x*, the results suggest that robust regression is preferable to winsorization because winsorization causes a constant bias, and preferable to truncation as well because truncation causes a variable bias.

Overall, the results in Table 2 and visual analysis provided in Figures 7 and 8 demonstrate that robust regression is unaffected by random influential/outlier observations and significantly reduces bias caused by extreme events that are correlated with a researcher's variable of interest. In sum, winsorizing does little to mitigate the influence of correlated influential/outlier observations while truncation imparts a downward bias on parameter estimates.[13]

5.4 Trimming the independent variable but not the dependent variable.

Our literature review found that a common practice in accounting settings is to either winsorize or truncate the independent variable, but not the dependent variable. As illustrated in Figures 3c and 4a this practice is likely to bias coefficients away from zero. To provide some analysis of this alternative, Table 3 reports results were only the independent variable is winsorized or truncated.

[13] We highlight that our simulations intend to highlight the potential differences between approaches in conditions applicable to accounting research studies. The literature on robust regression has extensively demonstrated both analytically and empirically the validity of the robust regression estimators. For an extensive review of the robust regression literature see Andresen (2008).

In Panel A of Table 3 where extreme events are randomly distributed and uncorrelated with *x,* winsorizing *x,* but not y imparts an upward bias in □, increasing it from 0.80 to 0.82. As discussed earlier, winsorizing *x* "leverages up" its impact on *y*. In contrast, truncation of *x* appears to eliminate the downward bias caused by truncation of both *x* and *y* as reported in Table 2. More importantly, however, as reported in Panel B of Table 3, truncating *x* does nothing to mitigate the bias caused by correlated influential/outlier observations. More specifically, Panel B's results reinforce the resulting upward bias in □ □caused by winsorizing only the independent variable. In summary, the evidence reported in Table 3 suggests that truncating the independent variable but not the dependent variable may not mitigate the impact of influential/outlier observations.

6. Winsorization, truncation, and robust regression in the context of a published accounting study

In this section we replicate a published accounting study to illustrate how different approaches to deal with influential/outlier observations may yield different estimates (and potentially different inferences). The study we selected (Richardson et al. 2005, hereafter RSST) examined the differential persistence of various types of accruals on overall earnings' persistence. We selected this setting because it employs variables that are common to many studies in the accounting literature; it is widely accepted that there are influential/outlier observations in accruals components and earnings; and RSST (2005) used a winsorization rule different than the otherwise common top and bottom 1% of each variable. Our objective is not to criticize this study, but to simply illustrate how different approaches to deal with influential/outlier observations affect estimates (and potentially altering inferences).

6.1 Data, variable measurement, and earnings persistence models

We use all observations from Compustat with available data to calculate the variables of interest over the 1988 to 2001 period. Following RSST (see their Table 5) we estimate the following models:

$$ROA_{i,t+1} = \beta_0 + \beta_1 ROA_{i,t} + \varepsilon_{t+1,} \quad (6)$$

$$ROA_{i,t+1} = \beta_0 + \beta_1 ROA_{i,t} + \beta_2 TACC_{i,t} + \varepsilon_{t+1,} \quad (7)$$

$$ROA_{i,t+1} = \beta_0 + \beta_1 ROA_{i,t} + \beta_2 \Delta WC_{i,t} + \beta_3 \Delta NCO_{i,t} + \beta_4 \Delta FIN_{i,t} + \varepsilon_{t+1,} \quad (8)$$

where for each firm i and fiscal-year t, ROA = Operating income after depreciation; TACC = total accruals from the balance sheet approach (= $\Delta WC + \Delta NCO + \Delta FIN$); ΔWC = change in net working capital (WC = current operating assets – current operating liabilities); ΔNCO = change in net non-current operating assets (NCO = non-current operating assets – non-current operating liabilities); ΔFIN = change in net financial assets (FIN = financial assets – financial liabilities); with all variables scaled by average total assets.

In Eq. (6), given that earnings, scaled by average total assets, are fairly persistent, β_1 is expected to be positive and close to one. In eq. (2) $\square_1$ is expected to be positive and close to one, while $\square_2$ is expected to be negative due to reversals in accruals that reduce the persistence of earnings. Finally, in Eq. (3) $\square_1$ is expected to be positive and close to one, while $\square_2$ to $\square_4$ are expected to be negative with $\square_2 \prec \square_3 \prec \square_4$ due to reversals in accruals with different degrees of reliability.

6.2 Descriptive statistics and results

Table 4 presents descriptive statistics for the variables used in the analysis. The raw and winsorized data have a total of 65,994 firm-year observations. Panel A uses the raw data values for each variable; Panel B winsorizes each variable using a +1.0 and -1.0 cutoff (following RSST 2005); Panel C winsorizes each variable at the 1st and 99th percentiles; and Panel D truncatres

each variable at the 1st and 99th percentiles. The descriptive statistics show that all variables are characterized by influential/outlier observations. Not surprisingly, the descriptive statistics also show that the different winsorization and truncation approaches produce distributions with different means and standard deviations (i.e., different distributional properties). Of note is that since truncation is based on the top and bottom 1% of the data for each *individual* variable it results in deletion of approximately 6% of the total number of observations (this is why the size of the truncated dataset is 62,227 observations compared to 65,994 in the other three cases).[14]

Table 5 reports the results of estimating the coefficients of Eq. (6) to (8) under the alternative approaches to account for influential/outlier observations. Panel A uses the raw data (i.e., no winsorization or truncation) for each variable; Panel B uses values winsorized at +1.0 and -1.0 as done by RSST (2005); Panel C uses winsorized values based on the 1st and 99th percentiles of each variable's underlying distribution; and Panel D uses data were each variable has been truncated at its 1st and 99th percentile. In Panels A to D, estimation is based on OLS. In Panel E the raw values of each variable are used in conjunction with robust regression based on MM-estimation. In Panels A to D standard errors are clustered by firm. In Panel E robust standard errors are estimated using a bootstrap-cluster procedure (300 replications) to cluster by firm.[15]

[14] As a result, truncation is difficult to implement consistently in studies estimating different models because researchers face the choice of deleting more observations as the number of models with different variables increases, or estimating each model on a different sub-sample as a result of different truncation criteria.

[15] Clustering by firm is preferable, although qualitatively similar, to the Fama and Macbeth approach used by RSST (p. 464) to mitigate the impact of auto-correlated errors over time (see Gow et al 2010). For the robsust regression results bootstrap standard errors were estimated by repeatedly sampling from the original sample. Sampling was done by drawing firm clusters with replacement in order to account for correlation between observations within a cluster. The bootstrap estimate of the standard error is the standard deviation of the bootstrap sampling distribution. Estimating bootstrap-based clustered standard errors is consistent with the findings of Cameron et al. (2008) and the recommendations of Andersen (2008, 71). This approach can also be extended to two-way (firm and year) clustering. More details are provided in Appendix C.

Turning to the results, use of the raw data values (see Panel A) yields coefficient estimates that would seem to overestimate the persistence of *ROA* ($\gamma_1 > 1$). In addition, the results do not show any effect of accruals on earnings persistence since γ_1 to γ_4 are insignificant. On the other hand, robust regression (Panel E) and truncation (Panel D) produce similar results to RSST's (2006) findings (which are replicated in Panel B and based on RSST's winsorization at +1.0 and -1.). While truncation yields similar results in this case, as noted above, under truncation sample size is reduced by 6% in this setting, something that may be unappealing in other accounting settings where sample size may be a consideration). Turning to the commonly applied 1% and 99% winsorization rule (see Panel C) reveals that the coefficient γ_3 on the ΔNCO variable is positive and statistically significant at the 5% level ($\gamma_3 = 0.025$), but the coefficient γ_4 on the ΔFIN variable is insignificant (model, $\gamma_4 = -0.004$). Thus, under this alternative the researcher would not have been able to reject the null hypothesis as to the reliability of the ΔFIN variable.

While one might debate the reasonableness of winsorizing the data using a +1.0 and -1.0 bound based on the researchers' priors about the reasonable bounds for these variables (as done in RSST, 2005), such a debate is unnecessary and more importantly obscures the main takeaway of our analysis. Specifically, our results demonstrate that robust regression using MM-estimation yields coefficient estimates that are consistent with RSST's (2006) predictions on the one hand, while on the other hand not requiring an *ad hoc* procedure to deal with potential influential/outlier observations at the outset. Such findings support our simulation results and recommendation of using robust regression procedures as a standard practice and as a preferable alternative to ad hoc winsorization and/or truncation rules that often vary from study-to-study (approaches which serve to reduce inter-study comparability in the process).

7. Additional discussion of robust regression estimation

7.1 Are there any drawbacks of using robust regression instead of OLS?

Robust regression can be used in any situation where researchers can use OLS regression. If there are no leverage points or influential observations, both estimation methods will produce similar coefficients. However, in the presence of influential observations (the more common and likely case in accounting research), our results lead us to recommend that researchers go through the following steps:

1) Estimate the main model of interest using OLS.
2) Estimate the main model of interest using robust regression.
3) If there are major differences between the coefficient estimates of (1) and (2), identify the observations behind the differences in estimates, as well as the potential causes of the influential observations (e.g., data-errors or model misspecification).
4) The main model can be modified in a number of ways, for example:
    a. Using log, rank, or other transformations without changing the linearity assumptions or the data itself.
    b. Using a non-normal distribution that accommodates fat tails, and
    c. Using a non-linear specification.
5) Include a discussion of the impact of influential/outlier observations as part of the study.

In the presence of influential observations, assuming the researcher has aimed to estimate the best possible model, using robust regression is a viable (and non ad hoc) strategy since it is a compromise between including all the data and treating all observations equally under OLS regression versus excluding any observations entirely from the analysis. That said, robust regression is not a substitute for careful data analysis and thorough modeling. Ultimately, influential/outlier observations may provide interesting case studies and/or ways to modify or extend a theory and should always be identified and discussed.

7.2 Robust regression used as a diagnostic method to identify influential observations

Robust regression can be used not only to estimate and test hypotheses about coefficients of interest, but also to identify influential observations. Once identified, such observations can be analyzed separately from the rest of the data. For example, researchers may plot the robust standardized residuals against a measure of the influence of the explanatory variables (see e.g., Verardi and Croux 2009). In addition, the researcher can observe the weight given to each observation in the final iteration of MM-estimation, which means that observations with low or zero weights are candidates for investigation. Finally, MM-estimation is readily available in commonly used statistical packages like STATA, SAS and R. This approach is qualitatively similar to the residual diagnostics proposed by Belsley et al. (1980). However, a general problem with residual diagnostics is that they may prompt the researcher to delete influential/outlier observations (i.e. observations with large residuals, Cook's distance or DFITS) but new influential/outlier observations can appear in subsequent iterations.

7.3 Implementing robust regression

Appendix C outlines how to implement robust regression estimation in commonly used statistical packages (e.g., STATA, SAS and R). Although it is generally straightforward to estimate robust regression, we note three practical concerns in using robust regression. First, a common concern in most accounting settings is cross-sectional dependence and its effect on estimated standard errors. Currently available robust regression commands do not estimate clustered standard errors by default (either one way or two-way). However (as we do), it is possible to estimate robust standard errors using a bootstrap method by drawing clusters with replacement for each bootstrap sample.

Second, there exist a variety of robust regression methods (Andersen 2008 provides a thorough review) and within each method there are several estimation options. Our analysis focused on MM estimation, introduced by Yohai (1987), which combines high breakdown value estimation and M-estimation, and which has both the high breakdown property and a higher statistical efficiency than S estimation. Finally, while there is a potential loss of efficiency using robust regression when compared to OLS, given that accounting studies typically have very large samples, loss of efficiency is unlikely to be a concern when using robust regression. As a final thought, we note that research on robust regression techniques continues to evolve in the statistics and econometrics literatures.

8. Conclusion

In this study we examine the statistical problems related to the presence of influential observations in regressions estimated using accounting and market data. In order to provide a relevant background for our analyses, first we review the accounting literature to identify the various methods used to mitigate the impact of influential/outlier observations in accounting research. We document that the most common solutions to address the presence of influential/outlier observations are winsorization or truncation. However, there is significant variation in how these approaches are implemented (see Table 1). For example, some studies winsorize and/or truncate the independent variables, but not the dependent variables, while other studies winsorize and/or truncate both. Beyond that, some studies only winsorize or truncate a subset of variables. Perhaps more importantly, roughly a third of all studies do not discuss the treatment of influential/outlier observations at all. Such wide variation in the treatment of influential/outlier observations makes it difficult to compare results across studies.

We also evaluate alternative approaches currently used in the accounting literature to account for influential/outlier observations and compared them to robust regression based on MM-Estimation. We assess the effectiveness of these approaches in cases where influential/outlier observations occur randomly in the data, as well as where such observations are correlated with the independent variable of interest. One motivation for our analysis is that the causes of influential/outlier observations in accounting numbers and stock returns do not appear to be random. For example, extreme events that generate influential/outlier observations are likely to occur more frequently for smaller firms. Since it is widely accepted that accounting numbers contain influential/outlier values we investigated and compared how various procedures mitigate the impact of infrequent, but correlated events. The results of our simulations demonstrate that that the common (but rather ad hoc) approach of winsorizing the raw data does little to mitigate the influence of correlated influential/outlier observations. In particular, winsorizing only the independent variable biases the coefficients away from zero, increasing the probability of a Type I error. In addition, we find that the common approach of truncating the raw data tends to bias coefficients downward, except in the special (and rare) case where the independent variable is uncorrelated with the dependent variable (i.e. the "true" parameter is zero).

The advantage of robust estimation based on MM-estimation is that it offers both consistent and highly efficient estimation even in the presence of influential/outlier observations. We find that robust regression substantially reduces the bias (by 80%) induced by correlated influential/outlier observations. Beyond our simulation evidence, our conclusions about the advantages of robust estimation compared to winsorization and truncation are confirmed by the replication of a published accounting study. Specifically, our replication of Richardson et al.

(2005), focusing on the differential persistence of various accruals, demonstrates that robust regression based on MM-Estimation yielded reasonable parameter estimates without making changes to the original underlying data by making *ad hoc* assumptions about the bounds at which to winsorize or truncate influential/outlier observations.

Our findings lead us to recommend that future studies consider the use of robust regression procedures as a standard practice. Doing so will not only help to strengthen potential causality, but also enhance inter-study comparability (perhaps even aiding in reconciling conflicting results). We hope that our findings will help in redirecting researchers' attention from variable-by-variable truncation or winsorization of extreme values to model fit when attempting to identify and treat influential observations.

APPENDIX A – Examples of low frequency events causing extreme outcomes as measured by extreme buy-and-hold stock returns (BHAR) and extreme positive and negative total accruals

Extreme BHAR:

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" (two-day BHAR = 46%). 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" (two-day BHAR = 30%).

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

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

Extreme Negative Accruals (Examples from bottom 1% of distribution):

OPKO Health, Inc. 2007 (Accurals = -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, 2002 (Accruals = -990%): Expenses paid with stock ($2,129,635) and assets of only $407K.

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

Extreme Positive Accruals (Examples from top 1% of distribution):

Raytech 2001 (Accruals = 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 = 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 = 50%): Recognition of a deferred tax asset $6,700,000.

APPENDIX B – Examples of accounting studies using 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% 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% and 0.5% 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%iles. 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 – Robust regression (MM-estimation) in commonly-used statistical packages

Robust regression (MM-estimation) can be easily implemented, usually with just a few extra lines of code beyond OLS.

- In STATA using the mm_regress (Verardi and Croux 2009) and robreg (Jann 2010) commands with documentation and examples provided in Verardi and Croux (2009) and in STATA help.

- In SAS using the proc robustreg command with documentation and examples provided in the SAS Institute Paper 265-67 by Colin Chen.

- In R using the robustbase package and the command lmrob (Rousseeuw et al. 2012) with documentation and examples provided in the software package.

To produce standard error estimates for our tests we wrote a STATA ado file that computes one-way (firm) or two-way (firm and year) clustered bootstrapped standard errors for robreg. Additional datasets and examples of the code will be made available online by the authors upon request.

APPENDIX D – Impact of truncation on estimates of β

The analysis below shows 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).

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

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

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

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

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

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

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

The analysis 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 which means that truncation will bias coefficients towards zero except for cases where β_h is of a different sign than the underlying parameter value (which seems unlikely or at least rare).

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.

Chaochharia, V., and Y. Grinstein. 2009. CEO compensation and board structure. *Journal of Finance* 64: 231–261

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.

Frank, M., and V. Goyal. 2005. Trade-off and pecking order theories of debt. In *Handbook of Empirical Corporate Finance*, edited by B. E. Eckbo. Amsterdam, The Netherlands: Elsevier Science B.V.

Gow, I., G. Ormazabal, and D. Taylor .2010. Correcting for cross-sectional and time-series dependence in accounting research. *The Accounting Review* 85: 483–512.

Guthrie, K., J. Sokolowsky, and K. Wan. 2012. CEO compensation and board structure revisited. *The Journal of Finance* 57(3): 1149–1168.

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.

Jann, B. 2010. ROBREG: Stata module providing robust regression estimators. http://ideas.repec.org/c/boc/bocode/s457114.html

Kennedy, D., J. Lakonishok, and W. Shaw. 1992. Accomodating outliers and nonlinearity in decision models. *Journal of Accounting, Auditing and Finance* 7(2): 161–190.

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.

Petersen, M. 2009. Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies* 26(4): 435–480.

Richardson, S., R. Sloan, M. Soliman, and I. Tuna. 2005. Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics* 39:437–485.

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.

Rousseeuw, P., C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, M. Maechler. 2012. robustbase: basic robust statistics. R package version 0.9-7. URL http://CRAN.R-project.org/package=robustbase

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

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 1 – Plots of cumulative abnormal returns and analyst earnings forecast errors around earnings announcement dates

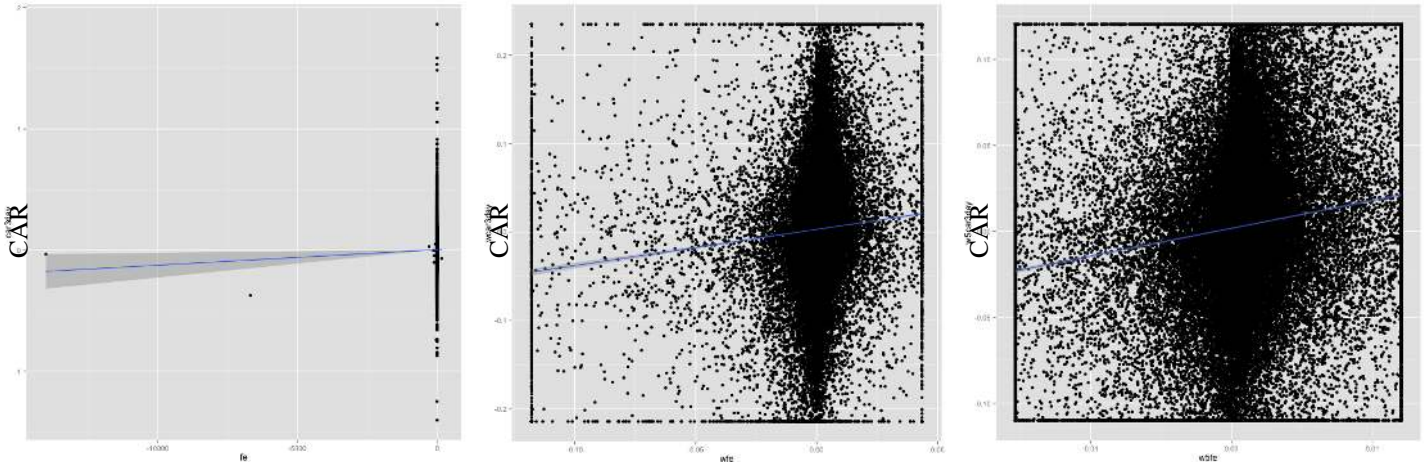


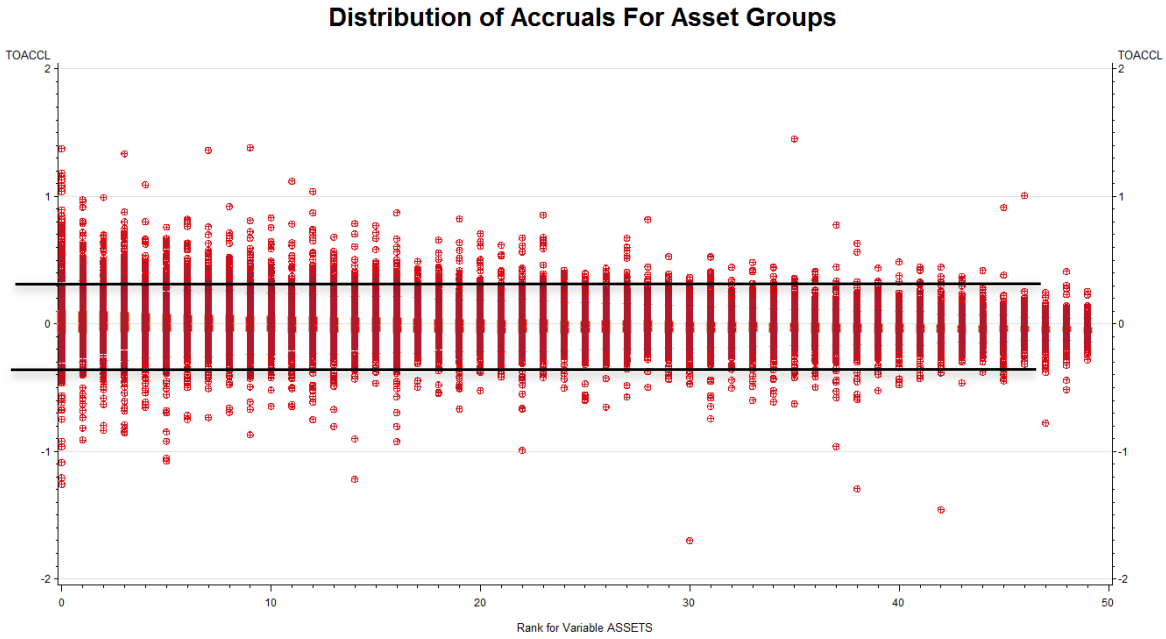
Figure 1A: Raw Data (β = ERC = 0.00002)

Figure 1B: Data winsorized at 1% and 99%. (β = ERC = 0.408)

Figure 1C: Data winsorized at 5% and 95%. (β = ERC = 1.618)

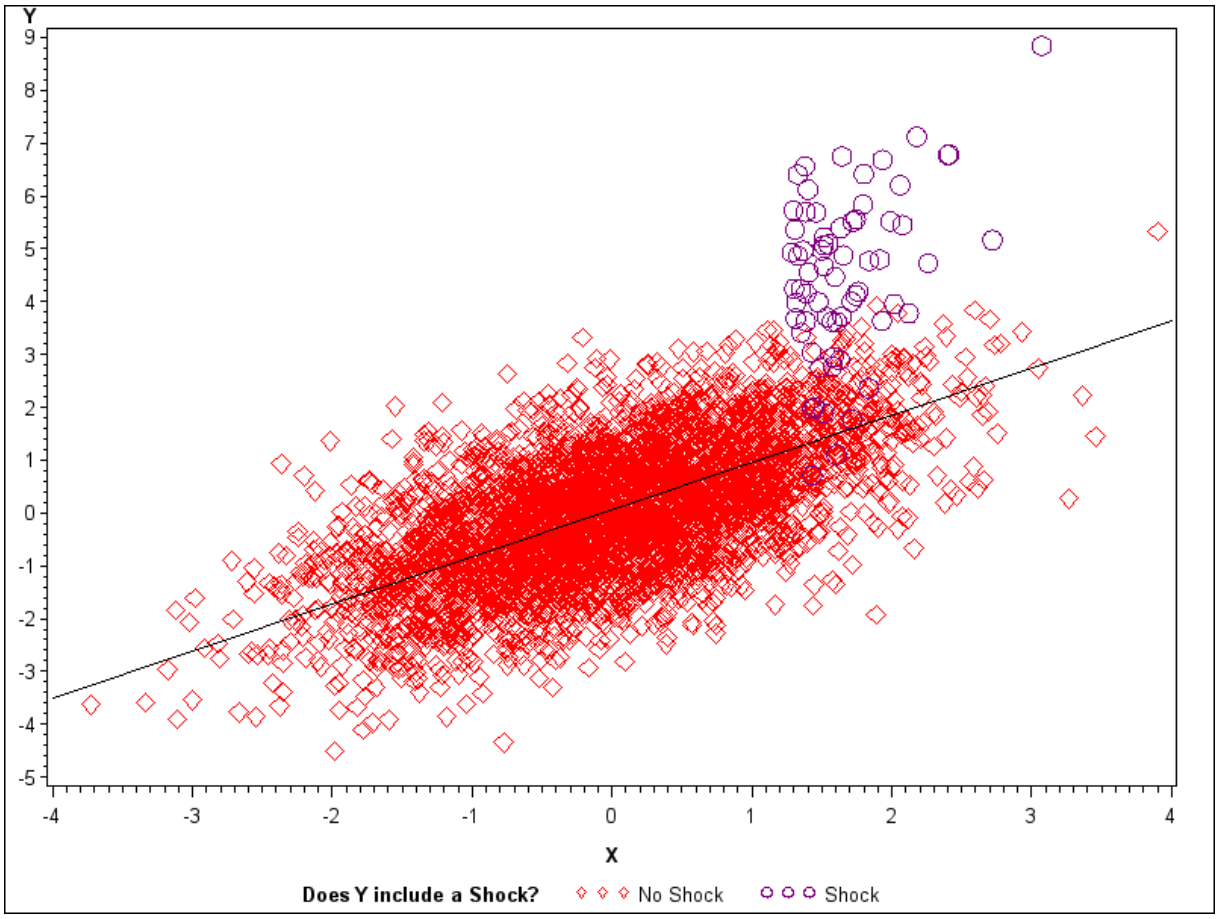
The sample underlying the figures is constructed as follows. We begin with all quarterly EPS forecasts on IBES's detail file (WRDS dataset: DET_EPSUS) from 2005-2011. We use the median of the most recent forecast by all analysts making forecasts less than 90 days before the earnings announcement to calculate the earnings forecast error as Actual EPS - Median EPS forecast, which is scaled by stock price on the day prior to the return accumulation period (four trading days prior to the earnings announcement). CARs are three-day abnormal returns from the day before through the day after the earnings announcement where abnormal returns equal raw returns minus the value-weighted market return.

Figure 2 – Accrual outliers and Firm Size



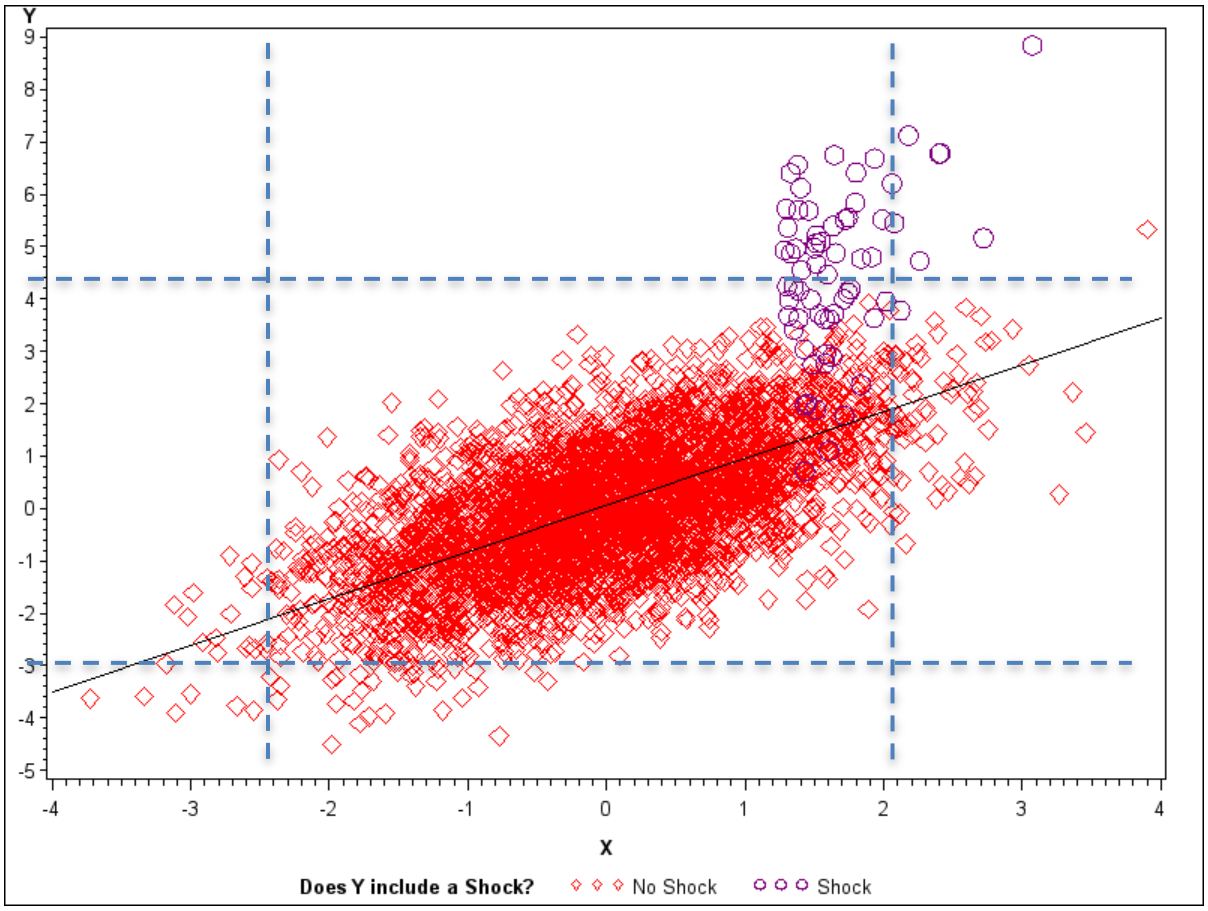
The Figure is a box plot of accruals for total assets grouped into fifty bins. Accruals and assets data are obtained from COMPUSTAT and include all firms with sufficient data to compute total accruals (balance sheet approach) from 1972-2001. The horizontal lines are the top and bottom 1% of the accruals distribution.

Figure 3 - Plot of a simulated dataset and a fitted OLS regression line, which yields an estimated slope coefficient of 0.894



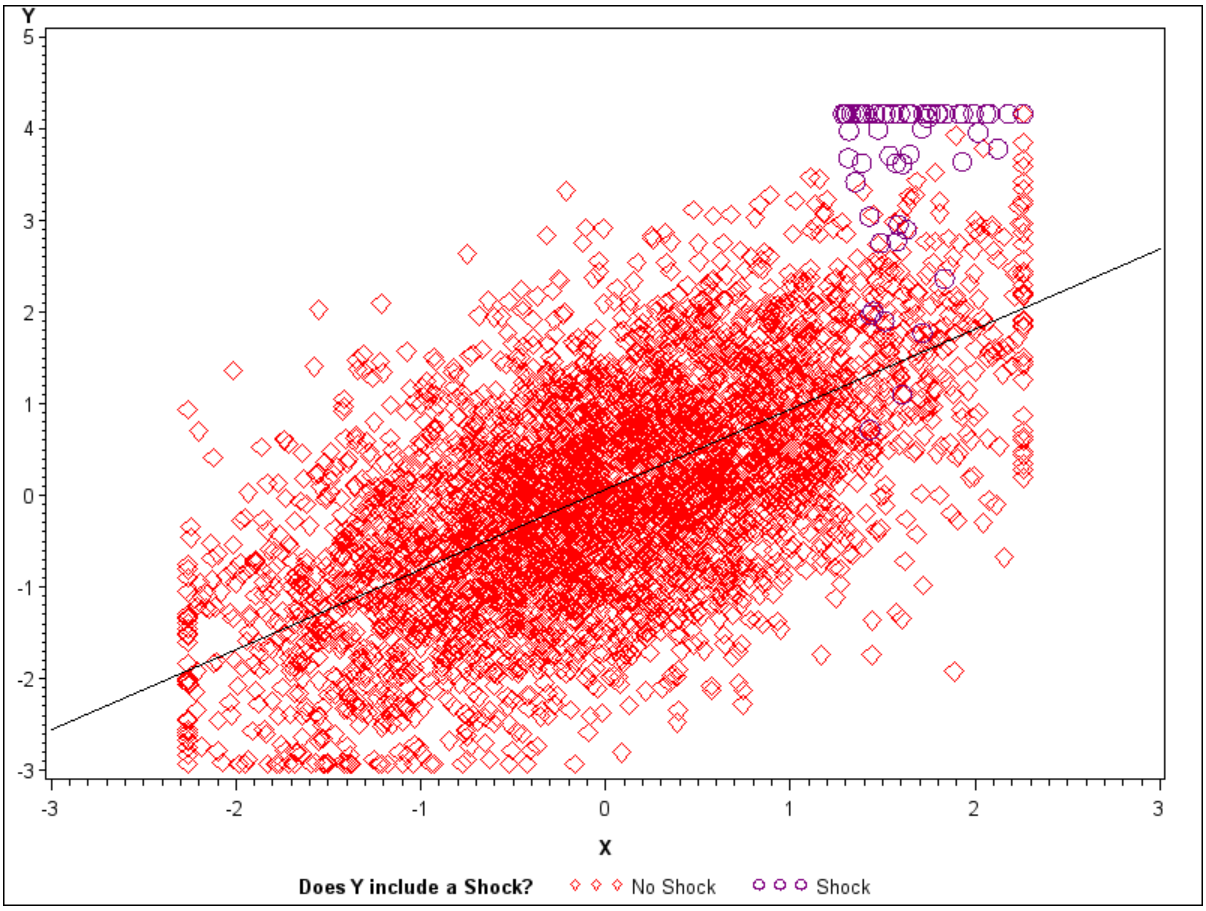
The Figure presents a plot of simulated data consisting of 4,000 values of y generated from the following data generating process: $y = bx + z + e$, where $x \sim N(0,1)$ and $e \sim N(0,1)$. An "extreme event," v, is generated from N(3,1). v is multiplied by d, an indicator variable equal to one when a random draw from a uniform distribution has a value equal to or exceeding 0.8, and x falls in the top 10% of its distribution. Thus $z = d*v$ is non-zero roughly 2% of the time. β is assigned a value of 0.8 in the simulated data. An estimate of b is generated using the following OLS model: $y = \alpha + \beta x + e$. In the Figure, the square dots represent the normally occurring observations (no shock) while the round dots represent simulated influential/outlier observations (the shock).

Figure 4a – Plot of a simulated dataset (same data as Figure 3) with lines added to signify the top and bottom 1% of the distributions of *x* and *y*



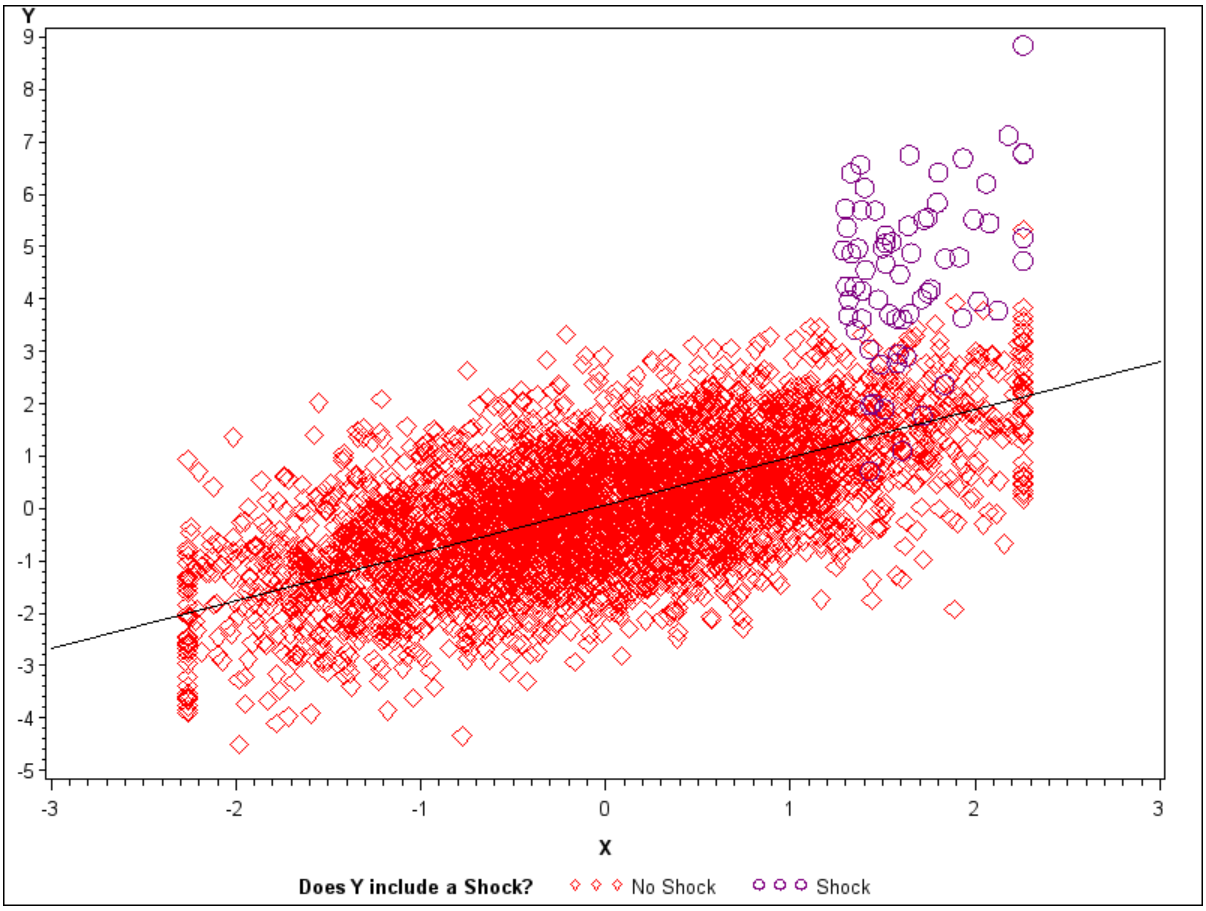
The data underlying the Figure is the same as that for Figure 3 except that lines have been added at the 1st and 99th percentiles for the simulated values of both *x* and y (dotted lines). These lines are added because values below (above) the 1st (99th) percentile are typically assumed to be extreme/ outlying observations under the commonly used truncation or winsorization rules applied in accounting research. In the Figure, the square dots represent the normally occurring observations (no shock) while the round dots represent simulated influential/outlier observations (the shock). An estimate of *b* is generated using the following OLS model: $y = \alpha + \beta x + e$. As in Figure 3, Figure 4a presents a scatter-plot of simulated data consisting of 4,000 values of y generated from the following data generating process: $y = bx + z + e$, where $x \sim N(0,1)$ and $e \sim N(0,1)$. An "extreme event," *v*, is generated from N(3,1). *v* is multiplied by *d*, an indicator variable equal to one when a random draw from a uniform distribution has a value equal to or exceeding 0.8, and *x* falls in the top 10% of its distribution. Thus $z = d*v$ is non-zero roughly 2% of the time. β is assigned a value of 0.8 in the simulated data (i.e., before any winsorization or truncation).

Figure 4b - Plot of a simulated dataset and a fitted OLS regression line after winsorizing both *x* and *y* at the top and bottom 1%, which yields an estimated slope coefficient of 0.87



The data underlying the Figure is the same as that for Figure 3 except that the data have been winsorized at the 1st and 99th percentiles (both *x* and y). This approach mirrors the commonly used winsorization rules applied in accounting research where values below (above) the 1st (99th) percentile are typically assumed to be influential/outlier observations. In the Figure, the square dots represent the normally occurring observations (no shock) while the round dots represent simulated influential/outlier observations (the shock). An estimate of b is generated using the following OLS model (on the winsorized data): $y = \alpha + \beta x + e$. As in Figure 3, Figure 4a presents a scatter-plot of simulated data consisting of 4,000 values of y generated from the following data generating process: $y = bx + z + e$, where $x \sim N(0,1)$ and $e \sim N(0,1)$. An "extreme event," v, is generated from N(3,1). v is multiplied by d, an indicator variable equal to one when a random draw from a uniform distribution has a value equal to or exceeding 0.8, and x falls in the top 10% of its distribution. Thus $z = d*v$ is non-zero roughly 2% of the time. β is assigned a value of 0.8 in the simulated data (i.e., before any winsorization).

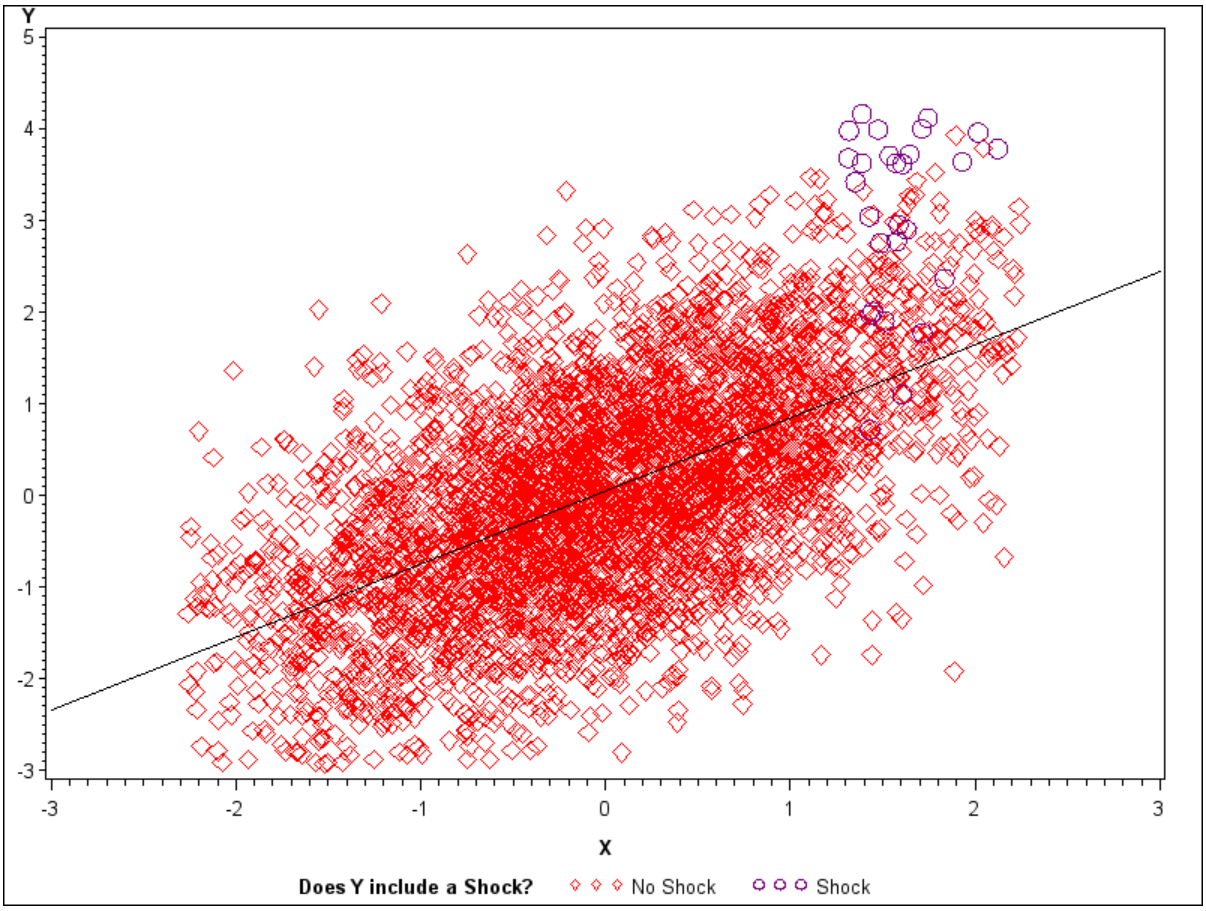
Figure 4c - Plot of a simulated dataset and a fitted OLS regression line after winsorizing only *x* at the top and bottom 1%, which yields an estimated slope coefficient of 0.91



The data underlying the Figure is the same as that for Figure 3 except that the x variable has been winsorized at the 1st and 99th percentiles. This approach mirrors the commonly used winsorization rules applied in accounting research where values below (above) the 1st (99th) percentile are typically assumed to be influential/outlier observations. In the Figure, the square dots represent the normally occurring observations (no shock) while the round dots represent simulated extreme/ outlying observations (the shock). An estimate of b is generated using the following OLS model (on the winsorized data): $y = \alpha + \beta x + e$. As in Figure 3, Figure 4a presents a scatter-plot of simulated data consisting of 4,000 values of y generated from the following data generating process: $y = bx + z + e$, where $x \sim N(0,1)$ and $e \sim N(0,1)$. An "extreme event," v, is generated from N(3,1). v is multiplied by d, an indicator variable equal to one when a random draw from a uniform distribution has a value equal to or exceeding 0.8, and x falls in the top 10% of its distribution. Thus $z = d*v$ is non-zero roughly 2% of the time. β is assigned a value of 0.8 in the simulated data (i.e., before any winsorization).

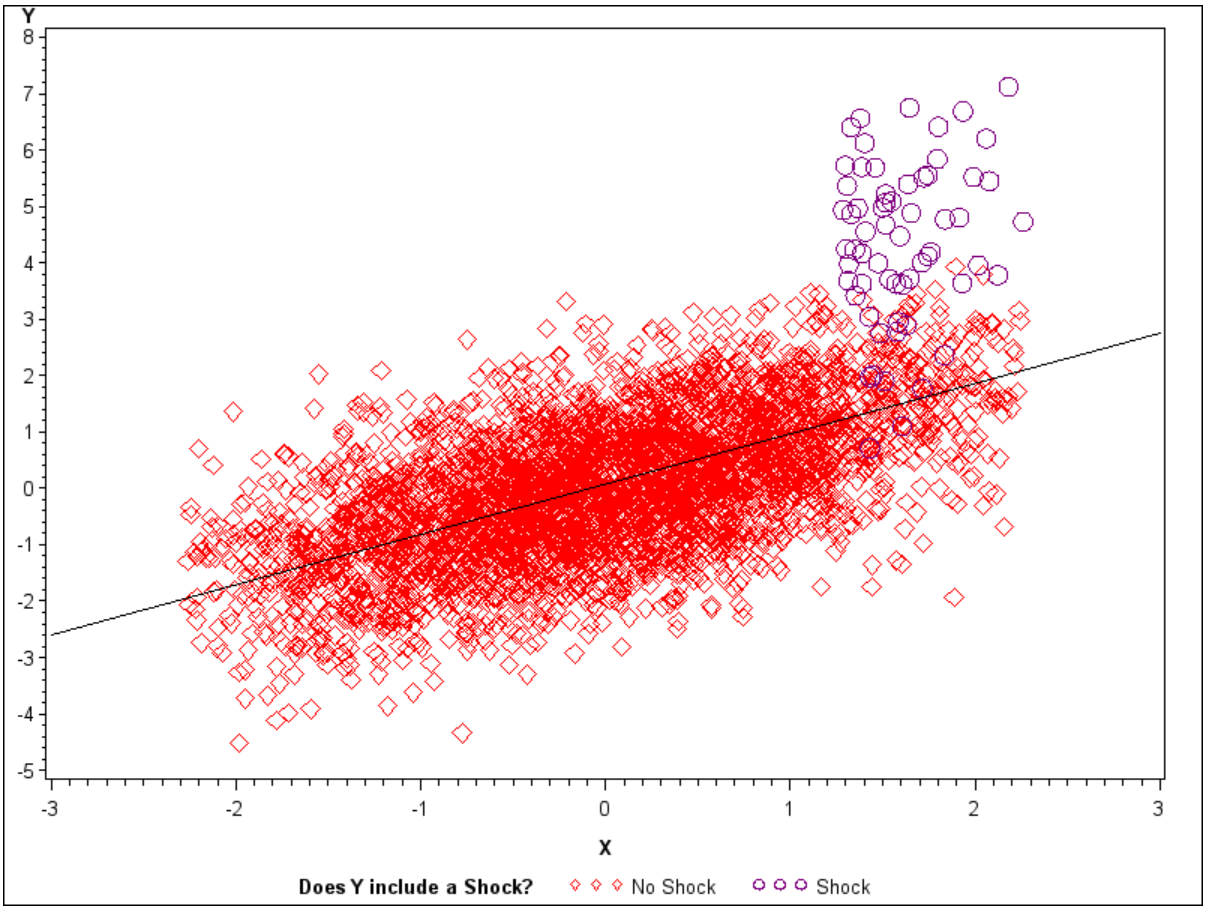
.

Figure 5a - Plot of a simulated dataset and a fitted OLS regression line after truncating both *x* and *y* at the top and bottom 1%, which yields an estimated slope coefficient of 0.77



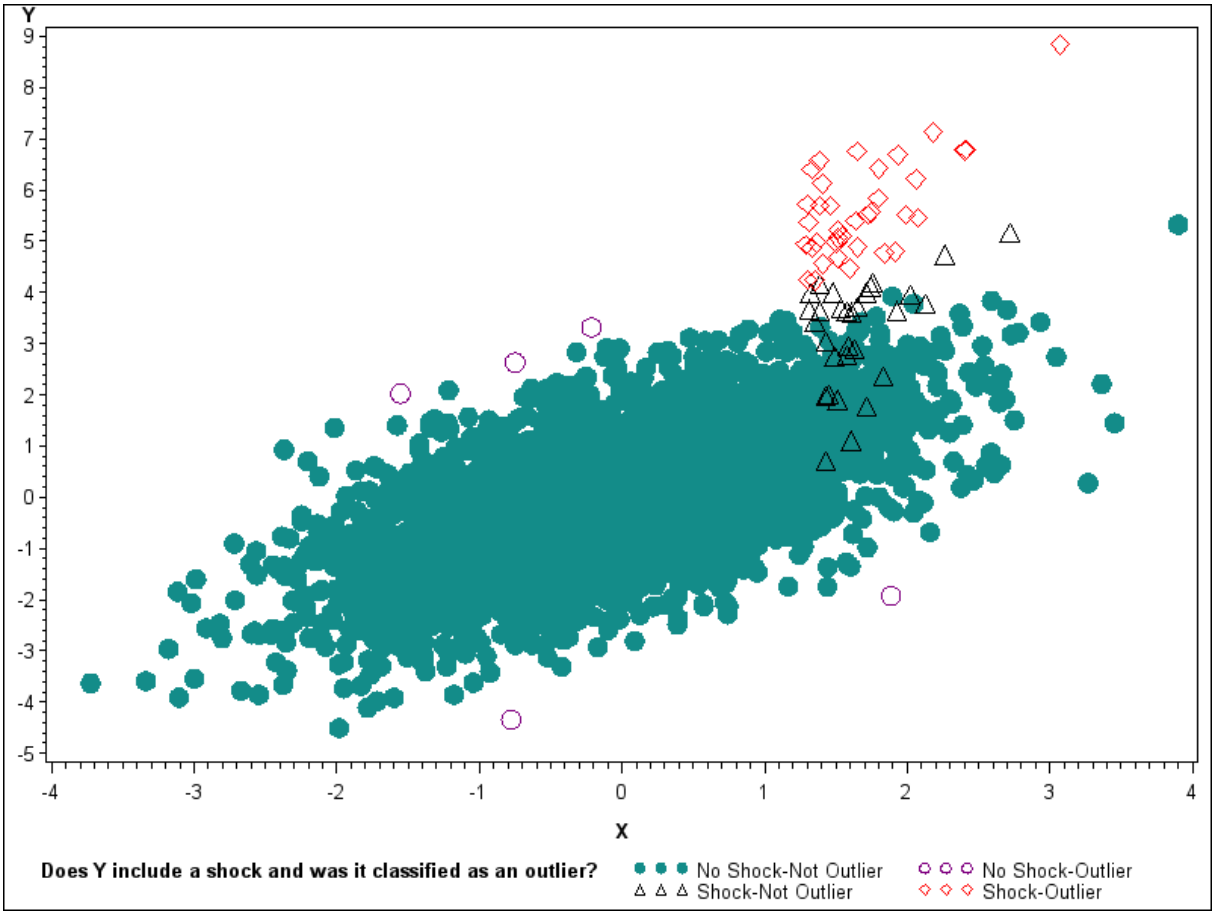
The data underlying the Figure is the same as that for Figure 3 except that the data have been truncated at the 1st and 99th percentiles (both *x* and y). This approach mirrors the commonly used truncation rules applied in accounting research where values below (above) the 1st (99th) percentile are typically assumed to be influential/outlier observations. In the Figure, the square dots represent the normally occurring observations (no shock) while the round dots represent simulated influential/outlier observations (the shock). An estimate of b is generated using the following OLS model (on the truncated data): $y = \alpha + \beta x + e$. As in Figure 3, Figure 4a presents a scatter-plot of simulated data consisting of 4,000 values of y generated from the following data generating process: $y = bx + z + e$, where $x \sim N(0,1)$ and $e \sim N(0,1)$. An "extreme event," v, is generated from N(3,1). v is multiplied by d, an indicator variable equal to one when a random draw from a uniform distribution has a value equal to or exceeding 0.8, and x falls in the top 10% of its distribution. Thus $z = d*v$ is non-zero roughly 2% of the time. β is assigned a value of 0.8 in the simulated data (i.e., before any truncation).

Figure 5b - Plot of a simulated dataset and a fitted OLS regression line after truncating *x* at the top and bottom 1%, which yields an estimated slope coefficient of 0.90.



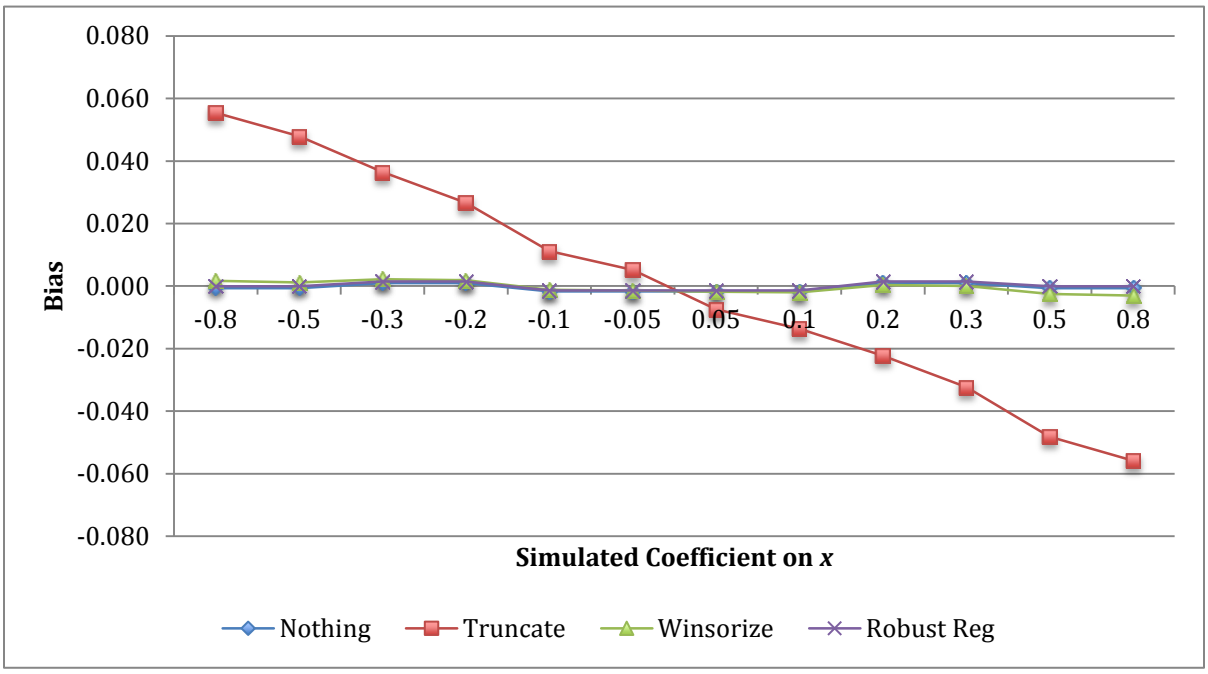
The data underlying the Figure is the same as that for Figure 3 except that the x variable has been truncated at the 1st and 99th percentiles). This approach mirrors the commonly used truncation rules applied in accounting research where values below (above) the 1st (99th) percentile are typically assumed to be influential/outlier observations. In the Figure, the square dots represent the normally occurring observations (no shock) while the round dots represent simulated extreme/ outlying observations (the shock). An estimate of b is generated using the following OLS model (on the truncated data): $y = \alpha + \beta x + e$. As in Figure 3, Figure 4a presents a scatter-plot of simulated data consisting of 4,000 values of y generated from the following data generating process: $y = bx + z + e$, where $x \sim N(0,1)$ and $e \sim N(0,1)$. An "extreme event," v, is generated from N(3,1). v is multiplied by d, an indicator variable equal to one when a random draw from a uniform distribution has a value equal to or exceeding 0.8, and x falls in the top 10% of its distribution. Thus $z = d*v$ is non-zero roughly 2% of the time. β is assigned a value of 0.8 in the simulated data (i.e., before any truncation).

Figure 6 - Plot of a simulated dataset and a fitted robust regression line based on MM-estimation, which yields an estimated slope coefficient of 0.82



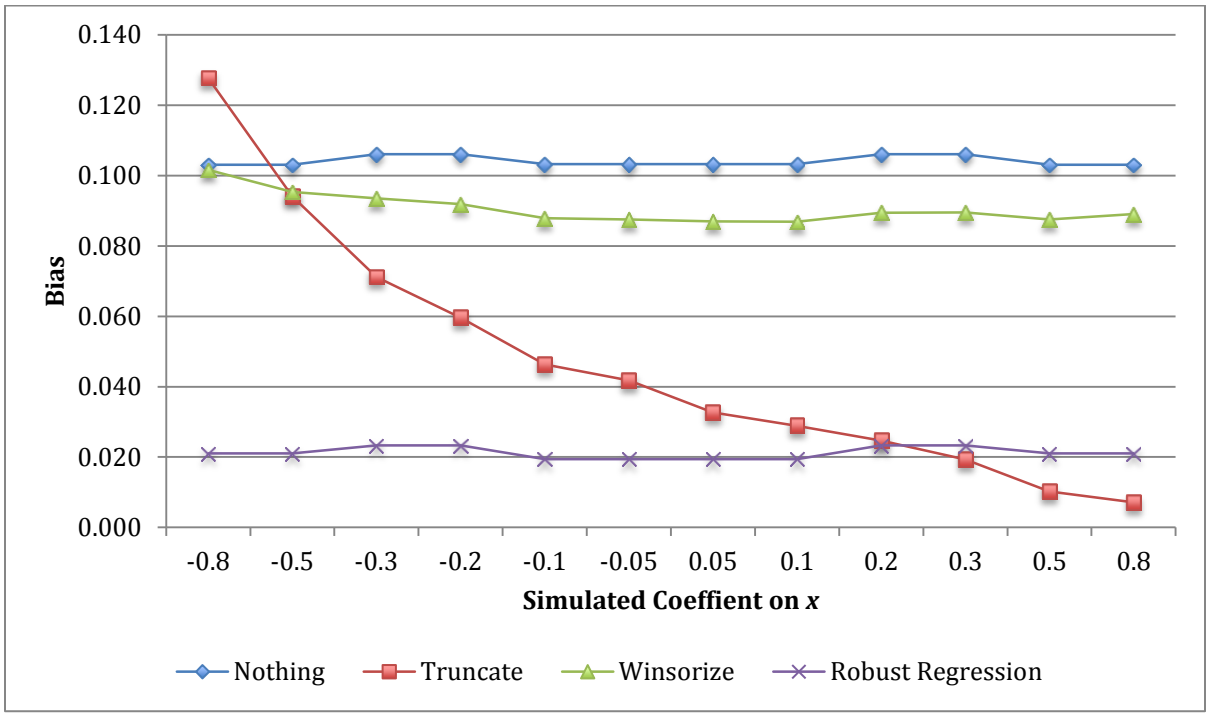
The figure illustrates how MM-estimation identifies influential/outlier observations. An estimate of *b* is generated using MM-estimation using all of the data. The full dots represent the normally occurring observations not down-weighted by robust regression (no shock and not an influential/outlier observation). The circle dots represent the normally occurring observations down-weighted by robust regression (no shock, but an influential/outlier observation). The triangle dots represent simulated influential/outlier observations not down-weighted by robust regression (shock, but not an influential/outlier observation). The square dots represent simulated influential/outlier observations down-weighted by robust regression (shock, and influential/outlier observation). The data underlying the Figure is the same as that for Figure 3. As in Figure 3, Figure 4a presents a scatter-plot of simulated data consisting of 4,000 values of y generated from the following data generating process: $y = bx + z + e$, where $x \sim N(0,1)$ and $e \sim N(0,1)$. An "extreme event," v, is generated from N(3,1). v is multiplied by d, an indicator variable equal to one when a random draw from a uniform distribution has a value equal to or exceeding 0.8, and x falls in the top 10% of its distribution. Thus $z = d*v$ is non-zero roughly 2% of the time. An estimate of *b* is generated using MM-estimation and using all of the data.

Figure 7 – Estimated β coefficients from simulations without influential/outlier observations (i.e., expected bias = 0) where true β varies from -0.8 to +0.8



The Figure plots the results of simulations showing how the estimated coefficient, β, is different from (i.e., biased) the true coefficient when there are no simulated influential/outlier observations (shocks) and where true β ranges from -0.8 to +0.8. The lines in the figure represent the average bias induced by the alternative estimation procedures. For each value of β the line represents the average β from 250 replications of sample size 2,000. y is generated from the following model: $y = bx + e$, where $x \sim N(0,1)$ and $e \sim N(0,1)$ where b is assigned values ranging from -0.8 to +0.8. Estimates of b are then generated using OLS or robust regression. The regression model is: $y = \alpha + \beta x + e$, and, therefore, estimates of b should be unbiased (i.e., if b is set to 0.8, the estimate β should also be 0.8). OLS is estimated under three alternative treatments for influential/outlier observations. In the first, "do nothing" case we do not winsorize or truncate extreme values of y and x. In the second case, "Truncate," we drop observations where either y or x falls in the top or bottom 1% of its respective distributions. In the third case, ("Winsorize") we winsorize x and y when values fall in the top or bottom 1% of their respective distribution. Finally, we use robust regression estimates based on MM-estimation.

Figure 8 - Estimated β coefficients from simulations with influential/outlier observations (i.e., expected bias = 0.10) where true β varies from -0.8 to +0.8



The Figure plots the results of simulations showing how the estimated coefficient, β, is different from (i.e., biased) the true coefficient when there are simulated influential/outlier observations (shocks) and where true β ranges from -0.8 to +0.8. The lines in the figure represent the average bias induced by the alternative estimation procedures. For each value of β the line represents the average β from 250 replications of sample size 2,000. y is generated from the following model: $y = bx + z + e$, where $x \sim N(0,1)$, $e \sim N(0,1)$, and an "extreme event" variable, v, is generated from N(3,1). The variable v is then multiplied by d, which is an indicator variable equal to one when a random draw from a uniform distribution has a value equal to or exceeding 0.8, and x falls in the top 10% of its distribution. Thus $z = d*v$ is non-zero roughly 2% of the time. β is assigned values ranging from -0.8 to 0.8. Estimates of b are then generated using either OLS or robust regression. The regression model is: $y = \alpha + \beta x + e$. Since the variables x and z are correlated by construction, the omission of z from the estimation equation causes the estimated β to be biased. By construction, the bias will be roughly 0.10. To assess the extent to which this bias can be mitigated by truncating, winsorizing, or robust regression, OLS is estimated under three alternative treatments for the influential/outlier observations. In the first, "do nothing" case we do not winsorize or truncate extreme values of y and x. In the second case, "Truncate," we drop observations where either y or x falls in the top or bottom 1% of its respective distributions. In the third case, ("Winsorize") we winsorize x and y when values fall in the top or bottom 1% of their respective distribution. Finally, we use robust regression estimates based on MM-estimation.

Table 1: How Are Influential/outlier Observations Treated in Accounting Research Settings?

The table presents a literature review of studies published between 2006 and 2010 in Contemporary Accounting Research, Journal of Accounting Research, Journal of Accounting and Economics, Review of Accounting Studies, and The Accounting Review. The body, footnotes and tables of each study was searched for discussion of the treatment of influential/outlying observations. The studies reviewed span auditing, properties of analysts' forecasts, management compensation, earnings management, conservatism, tax, disclosure, and earnings-returns associations, etc.,. Studies that include both an analytical model and empirical tests are classified as archival.

Panel A: Studies by general category

Description	*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 influential/outlier observations	404	68%
Archival studies do not addressing influential/outlier observations	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%
Subtotal (N and % are not additive)	355	88%
Archival studies using other techniques	49	12%
Total archival studies addressing influential/outlier observations	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%
Subtotal (N and % are not additive)	83	53%
Returns used raw as dependent variable	74	47%
Total archival studies with returns as dependent variable	157	100%

Panel B: Studies using winsorization

Description	*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 (N and % are not additive)	221	

Panel C: Studies using truncation

Description	*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 (N and % are not additive)	161	

Table 2: Main simulation results

The table reports simulation results where an independent variable *x*, is generated from a normal distribution with mean zero and a standard deviation of one, and with an "extreme event" variable *v*, that is generated from a normal distribution with a mean of three and a standard deviation one. In Panel A, *v* is multiplied by 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. In 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. Conversely, 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 $z = d*v$ is non-zero roughly 2 percent of the time, but in Panel B it 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, α is set to zero. Four *y* variables are generated by varying values of *b* (zero or 0.8) and γ (zero or 1.0). A total of 250 samples are generated with 2,000 observations each. 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 values of b are means of the 250 underlying regressions for that condition. Bias is the difference between the mean and "true" parameter value (zero or 0.8). All regressions are estimated using OLS, except for robust regression which is based on MM-estimation.

Panel A: Infrequent events *z* are independent of *x*

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

Panel B: Panel B: Infrequent events *z* are correlated with *x*

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

Table 3: Simulation results under winsorizing or truncating only the independent variable *x*

This table reports simulation results similar to Table 2 except that only the independent variable *x* is winsorized or truncated. Otherwise, as in table 2, the independent variable *x*, is generated from a normal distribution with mean zero and a standard deviation of one, and with an "extreme event" variable *v*, that is generated from a normal distribution with a mean of three and a standard deviation one. In Panel A, *v* is multiplied by 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. In 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. Conversely, 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 *z*= *d*v* is non-zero roughly 2 percent of the time, but in Panel B it 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, α is set to zero. Four *y* variables are generated by varying values of *b* (zero or 0.8) and γ (zero or 1.0). A total of 250 samples are generated with 2,000 observations each. 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 values of b are means of the 250 underlying regressions for that condition. Bias is the difference between the mean and "true" parameter value (zero or 0.8). All regressions are estimated using OLS, except for robust regression which is based on MM-estimation.

Panel A: Infrequent events *z* are independent of *x*

	Parameter Values		**"Do Nothing"**		**Winsorize**		**Truncate**	
Regression Model:	β	γ	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias
$Y = \alpha + \beta x + e$	.80	0	0.80	0.00	0.82	0.02	0.80	0.00
$Y = \alpha + \beta x + e$	.80	1	0.80	0.00	0.82	0.02	0.80	0.00
$Y = \alpha + \beta x + e$	0	1	0.00	0.00	0.00	0.00	0.00	0.00
$Y = \alpha + \beta x + \gamma z + e$	.80	0	0.80	0.00	0.82	0.02	0.80	0.00
$Y = \alpha + \beta x + \gamma z + e$	.80	1	0.80	0.00	0.82	0.02	0.80	0.00
$Y = \alpha + \beta x + \gamma z + e$	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 Model:	β	γ	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias	$\hat{\hat{\beta}}$	Bias
$Y = \alpha + \beta x + e$	.80	0	0.80	0.00	0.81	0.01	0.80	0.00
$Y = \alpha + \beta x + e$	.80	1	0.90	0.10	0.92	0.12	0.90	0.10
$Y = \alpha + \beta x + e$	0	1	0.10	0.10	0.11	0.11	0.10	0.10
$Y = \alpha + \beta x + \gamma z + e$	.80	0	0.80	0.00	0.81	0.01	0.80	0.00
$Y = \alpha + \beta x + \gamma z + e$	.80	1	0.80	0.00	0.81	0.01	0.80	0.00
$Y = \alpha + \beta x + \gamma z + e$	0	1	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Descriptive Statistics for RSST (2005) Earnings and Accrual Reliability Analysis

The table reports descriptive statistics for the variables used the analysis. The raw and winsorized data (Panels A-C) have a total of 65,994 firm-year observations. The truncated data (Panel D) has a total of 62,227 firm-year observations. Variable definitions are as follows, $ROA_{t,t+1}$ = Operating income after depreciation; $TACC_t$ = Total accruals using the balance sheet approach = $\Delta WC + \Delta NCO + \Delta FIN$; ΔWC_t = Change in net working capital where WC = current operating assets – current operating liabilities; ΔNCO_t = Change in net non-current operating assets where NCO = non-current operating assets – non-current operating liabilities; and ΔFIN_t = Change in net financial assets where FIN = financial assets – financial liabilities (see the text for additional discussion of the variables).

Panel A: Raw data

Variable	*Mean*	*Std. Dev.*	*Minimum*	*Lower Quartile*	*Median*	*Upper Quartile*	*Maximum*
ROA_{t+1}	-0.362	50.173	-13,000	-0.063	0.055	0.115	4
ROA_t	-0.108	4.538	-698	-0.056	0.059	0.119	4
$TACC_t$	0.120	15.887	-1,400	-0.074	0.044	0.194	2,628
ΔWC_t	0.042	7.941	-690	-0.034	0.006	0.053	1,314
ΔNCO_t	0.036	0.948	-199	-0.025	0.017	0.086	22
ΔFIN_t	0.023	7.259	-437	-0.084	-0.001	0.051	1,580

Panel B: Winsorized at +1.0 and -1.0 (following RSST, 2005)

Variable	*Mean*	*Std. Dev.*	*Minimum*	*Lower Quartile*	*Median*	*Upper Quartile*	*Maximum*
ROA_{t+1}	-0.023	0.267	-1.000	-0.063	0.055	0.115	1.000
ROA_t	-0.016	0.264	-1.000	-0.056	0.059	0.119	1.000
$TACC_t$	0.054	0.334	-1.000	-0.074	0.044	0.194	1.000
ΔWC_t	0.005	0.157	-1.000	-0.034	0.006	0.053	1.000
ΔNCO_t	0.041	0.216	-1.000	-0.025	0.017	0.086	1.000
ΔFIN_t	-0.017	0.242	-1.000	-0.084	-0.001	0.051	1.000

Panel C: Winsorized at the 1st and 99th percentiles

Variable	*Mean*	*Std. Dev.*	*Minimum*	*Lower Quartile*	*Median*	*Upper Quartile*	*Maximum*
ROA_{t+1}	-0.048	0.388	-3.051	-0.063	0.055	0.115	0.367
ROA_t	-0.036	0.359	-2.791	-0.056	0.059	0.119	0.379
$TACC_t$	0.051	0.376	-1.640	-0.074	0.044	0.194	1.389
ΔWC_t	0.005	0.138	-0.688	-0.034	0.006	0.053	0.518
ΔNCO_t	0.041	0.205	-0.757	-0.025	0.017	0.086	0.903
ΔFIN_t	-0.017	0.244	-1.002	-0.084	-0.001	0.051	1.051

Panel D: Truncated at the 1st and 99th percentiles

Variable	*Mean*	*Std. Dev.*	*Minimum*	*Lower Quartile*	*Median*	*Upper Quartile*	*Maximum*
ROA_{t+1}	-0.020	0.280	-3.039	-0.049	0.057	0.114	0.367
ROA_t	-0.009	0.260	-2.763	-0.041	0.062	0.118	0.379
$TACC_t$	0.052	0.295	-1.624	-0.068	0.043	0.183	1.387
ΔWC_t	0.006	0.112	-0.688	-0.032	0.006	0.051	0.517
ΔNCO_t	0.041	0.168	-0.757	-0.023	0.017	0.083	0.903
ΔFIN_t	-0.018	0.197	-1.001	-0.080	-0.002	0.047	1.051

Table 5: Earnings and Accruals Persistence: **Regression Results**

This table reports regression estimates of the coefficients for equations (6) – (8) in the text using the raw data and various approaches to account for influential/outlier observations. Panel A uses the values of the raw data for each variable. Panel B winsorizes each variable with a +1.0 and -1.0 cutoff (following RSST 2005). Panel C winsorizes each variable at the 1st and 99th percentiles. Panel D truncates each variable at the 1st and 99th percentiles. Estimation in Panels A-D is based on OLS. In Panel E the values of the raw data of each variable are used with robust regression based on MM-estimation. In Panels A-D standard errors are clustered. In Panel E robust standard errors are estimated using a bootstrap procedure (based on 300 replications). Variable definitions are: $ROA_{t,t+1}$ = Operating income after depreciation; $TACC_t$ = Total accruals using the balance sheet approach = $\Delta WC + \Delta NCO + \Delta FIN$; ΔWC_t = Change in net working capital where WC = current operating assets – current operating liabilities; ΔNCO_t = Change in net non-current operating assets where NCO = non-current operating assets – non-current operating liabilities; and, ΔFIN_t = Change in net financial assets where FIN = financial assets – financial liabilities (see the text for additional discussion of the variables).

Panel A: Raw data (OLS estimation)

	Model (1)	Model (2)	Model (3)
Variables	*Dependent Variable = ROA_{t+1}*		
ROA_t	1.144***	1.154***	1.178***
	(4.61)	(5.01)	(10.55)
ΔWC_t			-0.199
			(-1.20)
ΔNCO_t			-0.001
			(-0.01)
ΔFIN_t			0.187
			(1.04)
$TACC_t$		-0.037	
		(-0.62)	
Intercept	-0.238	-0.232	-0.230
	(-1.23)	(-1.22)	(-1.20)
Observations	65,994	65,994	65,994
Adj. R^2	0.011	0.011	0.011

Panel B: +1 and -1 Winsorized data and OLS estimation

	Model (1)	Model (2)	Model (3)
Variables	***Dependent Variable = ROA_{t+1}***		
ROA_t	0.822***	0.837***	0.839***
	(74.94)	(72.76)	(72.99)
ΔWC_t			-0.096***
			(-11.83)
ΔNCO_t			-0.022***
			(-3.17)
ΔFIN_t			-0.018***
			(-2.97)
$TACC_t$		-0.045***	
		(-11.40)	
Intercept	-0.010***	-0.007***	-0.009***
	(-3.35)	(-2.68)	(-3.12)
Observations	65,994	65,994	65,994
Adj. R^2	0.660	0.663	0.663

Panel C: 1 and 99% Winsorized data and OLS estimation

	Model (1)	Model (2)	Model (3)
Variables	***Dependent Variable = ROA_{t+1}***		
ROA_t	0.869***	0.877***	0.880***
	(57.49)	(58.43)	(58.54)
ΔWC_t			-0.124***
			(-5.93)
ΔNCO_t			0.025**
			(2.13)
ΔFIN_t			-0.004
			(-0.38)
$TACC_t$		-0.029***	
		(-4.17)	
Intercept	-0.017***	-0.015***	-0.017***
	(-4.98)	(-4.45)	(-4.86)
Observations	65,994	65,994	65,994
Adj. R^2	0.645	0.645	0.647

Panel D: 1 and 99% Truncated data and OLS estimation

	Model (1)	Model (2)	Model (3)
Variables	***Dependent Variable = ROA_{t+1}***		
ROA_t	0.858***	0.871***	0.874***
	(60.59)	(57.06)	(56.11)
ΔWC_t			-0.131***
			(-11.49)
ΔNCO_t			-0.028***
			(-4.61)
ΔFIN_t			-0.017**
			(-2.27)
$TACC_t$		-0.047***	
		(-10.64)	
Intercept	-0.012***	-0.010***	-0.011***
	(-4.45)	(-3.80)	(-4.10)
Observations	62,227	62,227	62,227
Adj. R^2	0.633	0.636	0.636

Panel E: Raw data and robust regression MM-estimation

	Model (1)	Model (2)	Model (3)
Variables	***Dependent Variable = ROA_{t+1}***		
ROA_t	0.849***	0.863***	0.866***
	(176.17)	(135.80)	(170.133)
ΔWC_t			-0.067***
			(-7.37)
ΔNCO_t			-0.032***
			(-7.35)
ΔFIN_t			-0.016***
			(-3.07)
$TACC_t$		-0.027***	
		(-9.65)	
Intercept	0.009***	0.010***	0.010***
	(5.31)	(6.71)	(6.62)
Observations	65,994	65,994	65,994
Adj. R^2	0.339	0.341	0.341