Attrition bias and inferences regarding earnings properties; evidence from Compustat data

Peter Easton
University of Notre Dame
Mendoza College of Business
peaston@nd.edu

Martin Kapons
Tilburg University
Tilburg School of Economics and Management
m.m.kapons@tilburguniversity.edu

Peter Kelly
University of Notre Dame
Mendoza College of Business
Peter.W.Kelly.198@nd.edu

Andreas Neuhierl
Washington University in St. Louis
Olin Business School
andreas.neuhier@wustl.edu

Key words: attrition bias; missing observations; earnings forecasting; earnings persistence; mean reversion; random walk

JEL classification: G12 G14 G29 G31 G32 M40 M41

We thank Akash Chattopadhyay (discussant), Sanjay Bissessur, Jan Bouwens, Jeff Burks, Michael Calegari, Martijn Cremers, John Donovan, Michael Eames, Aytekin Ertan, Jere Francis, Stephan Hollander, Philip Joos, Yongtae Kim, Stephannie Larocque, Robin Litjens, Frank Moers, Steven Monahan, Hyungshin Park, Annelies Renders, Sandra Schafhaeutle, Harm Schutt, Christoph Sextroh, Richard Sloan, Jesse van der Geest, David Veenman, Jessie Watkins, David Windisch, and seminar participants at the 2018 AAA Financial Accounting and Reporting Section conference, Maastricht University, Tilburg University, the University of Amsterdam, the University of Santa Clara and the University of Notre Dame for helpful comments on an earlier draft.

Abstract

On average, across the years 1980 to 2018, almost 8.5 percent of firms on the Compustat Annual data set, which had earnings observations in year t-1, did not have earnings observations in year t. Because these disappearances were not random, there is attrition bias in estimates of earnings properties that require earnings observations in two consecutive years: namely, earnings persistence (used as an estimate of earnings quality), mean reversion, and accuracy of forecasts of earnings that are based on earnings of the prior year. We suggest three methods for imputing the disappeared earnings observations, which may be useful in future research on earnings properties. We show that conclusions about the properties of earnings change when we use imputed earnings to reduce the effects of attrition bias.

1. Introduction

We examine the effect of non-random data disappearances, alternatively labelled attrition bias or survivorship bias, on estimates of the persistence of earnings and associated inferences. On average, across the years 1980 to 2018, almost 8.5 percent of firms on the Compustat Annual data set, which had earnings observations in year t-1, did not have earnings observations in year t. These disappearances are non-random for a salient subset of observations; to wit, performance-related delistings. As a result we observe attrition bias that changes inferences regarding earnings properties - namely: earnings persistence (used as a measure of earnings quality); the extent to which earnings revert to the mean; regression parameters of earnings forecasting models; and, forecast accuracy from forecasts obtained from earnings forecasting models.[1]

Although attrition bias has had little attention in accounting research, its effect is widely acknowledged in other fields.[2] Research in finance has shown the implications of data disappearances on, for example, mutual fund performance, market overreaction, IPO performance, and the size effect (Brown, Goetzmann, Ibbotson and Ross, 1992; Shumway, 1997; Shumway and Warther, 1999). Research in medicine acknowledges that non-random attrition can be of great importance when evaluating clinical trials and has developed statistical methods for dealing with missing observations in randomized control trials (for example, Lachin, 2000).

[1] A number of papers acknowledge the effect of disappearance from Compustat on earnings persistence: for example, Nissim and Penman (2001), Dechow and Ge (2006), and Lawrence, Sloan and Sun (2018). Specifically, Dechow and Ge (2006) note in footnote 11 that firms that delist for performance-related reasons introduce a survivorship bias, which upwardly biases earnings persistence coefficients. Lawrence, Sloan and Sun (2018) argue that performance-related delistings are an extreme form of curtailments, which affect estimates of earnings persistence. These papers do not, as we do, conduct a formal analysis of the effects of disappearance on earnings persistence parameters and on the success (or lack thereof) of earnings forecasting models.

[2] We are aware of only one study in the accounting literature that explicitly studies the effect of data disappearances; Beaver, McNichols and Price (2007) analyse the effect of delisting returns on accounting-based trading strategies. Early studies of the time-series of earnings, which examined the effect of a form of survivorship bias on the inference that earnings follow a sub-martingale, considered a form of survivorship bias that was quite different to the bias due to disappearance from the data set, which is the focus of our study. Ball and Watts (1979) examined the effects of survivorship bias in samples that required 20 years of past earnings data for estimating the time-series properties of earnings as well as the effect of Compustat (at that time) creating a data set of firms that were of current interest to data users and then back-filling the data from past years.

Brilleman, Pachana and Dobson, 2010 note that attrition due to the death and decline in the health of participants can cause particular problems in studies of older people. The overarching concern with non-random attrition is falsely rejecting the true null hypothesis; that is, conclusions are subject to type I errors due to a bias arising from non-random data disappearances. We address this concern in the context of inferences based on time-series properties of earnings.

First, prior to our empirical analysis, we use simulated data to study the effects of attrition bias in a controlled environment and to highlight the potential effects. Our simulation assumes that earnings for firms that exist in the Compustat data in year t-1 follow a random walk to year t. We estimate the earnings persistence (EP) model, as defined in Li and Mohanram (2014), on these simulated data. This model estimates future earnings as a function of current earnings, a loss dummy, and an interaction term for negative earnings. We remove firms from these simulated data in a way that mimics the disappearance from Compustat and we find that the coefficient relating negative earnings in year t-1 to earnings in year t becomes significantly less than one. That is, even though the true data generating process is a random walk, we show that data disappearances can contribute to the estimate of differential persistence between positive and negative earnings reported in the literature. Consistent with the extant literature, we find evidence of mean reversion among loss firms, but we show that mean reversion can be attributed to data disappearances.[3] The estimated coefficients in the cross-sectional model become more biased and the forecasts perform better relative to a random walk as the fraction of

[3] We view mean reversion as the antonym of persistence. We use the estimate of the regression coefficient relating earnings of year t-1 to earnings of year t as the estimate of earnings persistence. One minus this estimate is the estimate of the degree of mean reversion.

firms that disappear increases. The bias arises because observations that disappear represent a non-random selection of non-surviving earnings observations.

Next, we undertake an empirical analysis of earnings of firms that remain in Compustat. We impute the values of missing observations in order to estimate the effects of delisting bias. We consider three conservative procedures for imputing earnings of firms that disappeared. After each type of imputation of time t earnings, the coefficient relating negative earnings of year t-1 to earnings of year t, which is the estimate of the persistence of losses, moves significantly closer to one and in many years it is not significantly different from one; in other words, the estimate of the tendency of negative earnings to revert to the mean is less than we estimate with the original Compustat data. Furthermore, after each type of imputation, negative earnings are no longer significantly less persistent than positive earnings. In fact, negative earnings are significantly more persistent than positive earnings. In other words, in this (persistence) sense the conclusion is that the reporting of losses is of a higher quality than the reporting of profits; attrition bias leads to the opposite conclusion.[4]

We also use imputed earnings for observations that disappeared in order to assess the effects of attrition bias on forecast accuracy; we compare the accuracy of regression-based forecasting models with forecasts based on a random walk. We find evidence in the Compustat data consistent with the evidence from our simulation. For the actual Compustat dataset with missing observations, regression-based forecasting models perform better in terms of forecast accuracy than the random walk model. When we use our replacement procedures for missing observations, all models lose their superior forecasting ability over a random walk model – the

[4] We discuss the use of persistence as an indicator of earnings quality (as outlined by Dechow, Ge, and Schrand (2010) later in our paper.

error differences relative to the random walk model are zero or close to zero; this is particularly so for forecasts based on negative earnings.[5]

We observe differences in the persistence of earnings over time. Positive earnings have become less persistent (that is, there is more mean-reversion) and attrition bias does not have a significant effect on the estimates of the persistence of positive earnings. On the other hand, there is no apparent change in the persistence of losses and we show that there is no evidence of mean reversion after removing the effects of attrition bias.

We observe that disappearing firms with negative earnings tend to be small, which suggests that the attrition bias may be concentrated in these smaller firms. To examine this more closely, we split the sample each year into those above the median market capitalization and those below the median market capitalization. Importantly, these small firms comprise less than three percent of the market capitalization of the entire Compustat dataset even though they represent 50 percent of the observations. We find that the attrition bias is much more evident for the small firms.[6] Hence, researchers might, for some research questions, consider focusing their analyses on large firms, which are less prone to non-random data disappearances.

The remainder of the paper proceeds as follows. We briefly describe our data in section 2. Section 3 describes the observations that disappear from the Compustat Annual file and provides evidence that the disappearances are non-random. In section 4, we discuss the effects of attrition bias on estimates of earnings persistence and mean reversion using simulations and analysis of Compustat data. Section 5 considers the effects of attrition bias on estimates of the accuracy of regression-based forecasts, again using simulations and analysis of Compustat data.

[5] This is the case for performance-related disappearances but not so for non-performance-related disappearances.
[6] The forecast error differences relative to the random walk are very small for large firms. This is the case both with and without imputing earnings data.

In section 6, we show that the effects of attrition bias are concentrated in smaller firms. Consistent with the accounting literature and the extant cross-sectional regression-based earnings forecasting models to which we refer, we study the persistence of the level of earnings. In short, our study of persistence focuses on whether firms with high (low) earnings continue to have high (low) earnings. In order to connect to the finance literature on mean reversion, particularly Fama and French (2000), we briefly, in section 7, examine the persistence of changes in earnings. We conclude with Section 8.

2. Data

Firm-level accounting data are obtained from the CRSP/Compustat merged data set. Firm-level returns and delisting codes are obtained from CRSP. We delete observations that are duplicates by permno and fiscal year. We exclude ADRs, closed-end funds, and REITs (that is, we retain observations with share codes equal to 10 or 11). We only consider firms that have U.S. dollars as their currency. Earnings are defined as income before extraordinary items minus special items.[7] We replace special items with zero if missing. The sample period is 1980 to 2018.[8]

3. Characteristics of Disappearing Firms

3.1 Annual observations in the year before disappearance

In this sub-section, we document the characteristics of firms that disappear from Compustat and we show that these firms differ from those that remain. The majority of disappearing firms experience losses in the year prior to disappearing and the firms that do not

[7] We repeat all of our analyses using earnings after special items (Compustat data ib) rather than earnings before special items (Compustat data ib-spi). All results are qualitatively very similar.
[8] Hayn (1995) and Beaver, McNichols and Price (2007) note that there are very few losses or few performance-related delistings recorded prior to 1980. We, therefore, start our analysis in 1980 in order to obtain more reliable estimates of attrition bias due to delistings.

disappear have much less negative equity-market returns in the next year compared to those that disappear; this is especially the case for performance-related delistings.[9] Further, we show that the decline in year-over-year quarterly earnings for the firms that disappear from the Compustat Annual data file but are present for part of the year on the Compustat Quarterly file is greater for loss firms that disappear than for loss firms that remain in the Annual Compustat data file.

We follow Shumway's (1997) definitions of delistings: performance-related delistings are firms with CRSP delisting codes 500 and 520 to 584 (inclusive) with missing future earnings; firms with other delisting codes with missing future earnings are non-performance-related delistings. Other delistings are all observations that do not have a CRSP delisting code but have missing future earnings.[10]

Our comparison of the characteristics of the firms that disappear from Compustat with the characteristics of the firms that remain is presented graphically in Figures 1A to 1D and is summarized in Table 1. Data for firms that disappear and those that remain in the sample are presented. Disappearing firms are also broken down into different sub-groups - those that disappear for known performance-related reasons, those that disappear for known non-performance related reasons, and those that disappear for other reasons. The most striking features of these graphs are that negative earnings firms are disproportionately represented among disappearing firms and performance-related delistings differ noticeably from other firms that disappear and from those that remain.

[9] We calculate delisting returns using CRSP delisting returns and we adjust for missing delisting returns following Shumway (1997). We split disappearing firms into performance-related disappearances and non-performance-related disappearances following the CRSP delisting codes. We follow Shumway (1997) and assume a delisting return of -0.3 (-30 percent) for firms with performance-related delistings and missing delisting returns.

[10] 191 of these firms reappeared on the Compustat dataset in the year after disappearing. Including or excluding these observations has no noticeable effect on the results of our analyses, which is expected because we find that the attrition bias is due to performance-related delistings.

Figure 1A plots the percentage of firms that disappear from Compustat for performance-related, non-performance related and other reasons. These percentages vary a great deal over time with the most performance-related delistings (5.7 percent of all available observations) occurring in 2000 (the year of the burst of the internet bubble), the most non-performance related delistings (6.9 percent) occurring in 1999 and the most "other" disappearances (2.6 percent) occurring in 1998. Across all firm-year observations (see Table 1) the percentages of firms in each of these categories that disappear from Compustat are 2.75 percent, 4.07 percent and 1.65 percent for performance-related delistings, non-performance-related delistings and "other" disappearances, respectively. Not surprisingly, firms that disappear for performance-related reasons tend to have negative earnings in the year prior to disappearance rather than positive earnings (2.38 percent compared with 0.37 percent). Conversely, firms that disappear for non-performance-related reasons tend to have positive earnings in the year prior to disappearance rather than negative earnings (2.98 percent compared with 1.09 percent).[11]

Figure 1B plots the median net income scaled by market capitalization in the year before disappearance from Compustat for each of the four sub-samples of observations. The noticeable feature of this graph is that the median earnings for performance-related delisting were negative in all years. Median scaled earnings for non-performance-related delistings were positive in all years. In the year before the delistings, the median scaled earnings of firms in each of the categories that disappear from Compustat are -0.49, 0.05 and 0.01 for performance-related delistings, non-performance-related delisting and "other" disappearances, respectively, across all firm-year observations (see Table 1). Notably, the lower earnings for observations that are

[11] This much more closely reflects the ratio of positive earnings firms to negative earnings firms in the full sample.

delisted for performance-related reasons were driven by observations with negative earnings in the year prior to delisting (median scaled earnings of -0.64).

Figure 1C plots the median returns in the year of the disappearance from Compustat. We see the worst returns for firms that are delisted for performance-related reasons (median return of -56 percent across all years). Also, there are superior returns for firms that are delisted for non-performance reasons (median return of 29 percent across all years compared with a median return across all years of five percent for firms that remain).

Figure 1D plots the percentage of earnings in the year prior to disappearance that are negative for each of the four categories. Again, the difference of firms that are delisted for performance-related reasons from all other firms is noticeable. In most years, the majority of firms delisted for performance-related reasons experience a loss in the year prior to delisting (the percentage of losses in this category is 86.55 percent). Firms that are delisted for "other" reasons (49.09 percent) have noticeably more losses than firms that remain and firms that are delisted for non-performance-related reasons. Firms that are delisted for non-performance-related reasons have slightly less losses (26.78 percent) than all firms that remain in the Compustat data (27.56 percent).

3.2 Quarterly observations in the year of disappearance

While we cannot observe annual earnings of firms that disappear in the actual year of disappearance, we can, for many of these firms, observe the quarterly earnings of disappearing firms in some of the earlier quarters within the actual year of disappearance. We expect to observe deteriorating earnings performance before delisting.

Table 2 reports the time-series means and medians of quarterly scaled earnings for observations that remain in the Compustat data, performance-related disappearances, non-

performance-related disappearances, and other disappearances. Earnings at time t denote quarterly observations for the last quarter before disappearance, except for firms remaining in the dataset. Earnings are scaled by equity market value. We split the observations by the sign of earnings in the year before disappearance. We are particularly interested in year over year changes in earnings.

Panel A reports quarterly earnings for each quarter t and change in earnings from quarter t-4 to t for all firms that remain on Compustat. The statistics in this panel are the base for comparison with statistics for observations that disappear, reported in Panels B, C, and D. Quarter t in these panels is the quarter immediately preceding disappearance.

Panel B reports the quarter-over-quarter changes in quarterly earnings before disappearance of firms that are delisted for performance-related reasons. Both, the positive earnings and the negative earnings sub-samples have negative changes in scaled earnings; mean (median) change of -0.28 (-0.10) for the positive earnings sub-sample and mean (median) change of scaled earnings of -0.73 (-0.17)) for the negative earnings sub-sample. Further, the mean (median) scaled negative earnings in the quarter before disappearance -0.88 (-0.26) is lower than the mean/median scaled earnings of any other subset of observations (see, Panels A, C and D).

Panel C reports the quarterly earnings before disappearance for non-performance-related reasons. Firms with scaled positive earnings only experience a decline for the median (-0.00), but not for the mean (1.05), which is high due to extreme outliers.[12] Scaled negative earnings experience a decline (mean of -0.04 and median of -0.00) and the resultant scaled earnings in the last quarter before disappearance is also negative (mean of -0.12 and median of -

[12] As changes of scaled earnings may be driven by changes in equity market value rather than by changes in earnings, we also analyse earnings scaled by sales, earnings scaled by total assets, and earnings scaled by equity book value; the patterns of earnings, especially for the performance-related delistings, change very little.

0.03). When we compare the performance-related delisting sample with the non-performance-related delisting sample, we see that performance-related delisting firms have higher negative changes in year over year scaled earnings than the non-performance-related delisting sample.

Panel D reports scaled earnings in the last quarter before disappearance for firms that disappear for other reasons. Although earnings changes and earnings levels do not show the strong negative pattern observed for performance-related disappearances, we still see evidence of deteriorating performance. Overall, this suggests that delistings are not random; they are tilted towards negative earnings firms with deteriorating performance.

In sum, the annual and quarterly descriptive analyses show that the firms that disappear from the Compustat data (particularly those that are delisted for performance-related reasons) differ considerably from those that remain. This suggests that estimates of persistence of earnings, especially estimates of the persistence of earnings of firms with negative earnings, may be affected by attrition bias. We examine the implications of disappearance in the next section.

4. Effects of Disappearance on Estimates of Earnings Persistence

The time series properties of earnings (earnings persistence in particular) have been extensively researched in firm-specific annual earnings time-series (for example, Watts, 1970; Watts and Leftwich, 1977) and in cross-sectional analyses (Beaver and Morse (1978); Kormendi and Lipe, 1987; Easton and Zmijewski, 1989; Dechow and Ge, 2006). Researchers are increasingly using earnings persistence parameters estimated via cross-sectional regressions to forecast annual earnings (Hou, van Dijk and Zhang, 2012; Gerakos and Gramacy, 2013; So, 2013; Li and Mohanram, 2014).[13] Dechow, Ge and Schrand (2010) list earnings persistence as

[13] See, for example, Chang, Landsman and Monahan (2013), Jones and Tuzel (2013), Lee, So and Wang (2017), and Patatoukas (2012).

one of five earnings properties that are used as proxies for earnings quality. The reason for its use in this way is "firms with more persistent earnings have a more ''sustainable'' earnings/cash flow stream that will make it a more useful input into DCF-based equity valuations." It follows that false inferences about persistence may lead to false inferences about earnings quality.

The results discussed in the previous section show that firms do not disappear at random from Compustat and, in fact, the disappearance is related to the earnings outcome. This suggests that attrition bias may lead to incorrect inferences about earnings persistence. We use simulation and imputation of the disappeared data to examine the validity of this suggestion. We show that attrition bias may, indeed, lead to false inferences about earnings persistence. In turn, attrition bias may lead to false inferences about mean reversion and earnings quality.

4.1. Simulations based on random walk earnings and deletion of poor performers

In this section, we show that disappearance from a data set, which has characteristics similar to those of the Annual Compustat data, can lead to biased coefficient estimates and the false conclusion that the earnings time-series does not follow a random walk even though the earnings in the data are generated via a random walk model. We discuss this issue in the context of a linear regression model:

$$y_i = x_i'\beta + \varepsilon_i.$$

We assume the data generating process is a random walk and that the error term is normally distributed. This implies a symmetric distribution, and, therefore, the mean and the median will be the same. Thus, to show that the median regression estimator is biased, it suffices to show that $E[\varepsilon_i|x_i] \neq 0$.

When the selection function, $s(\cdot)$, is random, that is, independent of epsilon:

$$E[\varepsilon_i|x_i, s(\cdot)] = E[\varepsilon_i|x_i] = 0.$$

Therefore, our estimator will be unbiased when the firms are removed at random. However, firms are more likely to disappear from the data when performance is poor, or when they receive a series of bad shocks. Therefore, we consider a selection function that depends on earnings (outcome) values.

Specifically, we assume a selection function that depends on the values of y_i, that is, $s(y_i)$. We assume for the sake of simplicity that we observe y_i if $y_i > c_i$, that is, if the outcome is above some threshold c_i, otherwise the observation is deleted from the data set. In other words, the selection function assumes the following functional form: $s(y_i) = 1$ if $y_i > c_i$ and $s(y_i) = 0$ otherwise. To determine if the estimator is biased, we need to check if

$$E[\varepsilon_i | x_i, s(y_i)] = 0?$$

But, here we have

$$E[\varepsilon_i | x_i, x_i'\beta + \varepsilon_i > c_i] \neq 0,$$

because the second part of the conditioning set is clearly not independent of the values of ε_i. Therefore, conditioning on the outcome will lead to biased coefficients.

4.1.1 The data generating process

The data are generated as a random walk. That is, we simulate earnings in levels by the following process:

$$E_{i,t} = E_{i,t-1} + \varepsilon_{i,t}$$

We use "actual data" for $E_{i,t-1}$ and take the panel structure of Compustat as given. We scale earnings by market equity. A firm i will enter our simulation when we first see it in Compustat. We vary the fraction of observations that are deleted from the panel in the implementation.

A critical component in this simulation is the distribution of $\varepsilon_{i,t}$. We choose, $\varepsilon_{i,t} \sim N(0, \sigma_{i,t})$. The question is what is a reasonable choice for $\sigma_{i,t}$. Because the normalization factor, market equity, is sometimes close to zero, the resultant simulated data may have vast extremes. Therefore, we use the interquartile range, which is an outlier-robust measure of dispersion, from the "actual data" to motivate our choice. The interquartile range of $\hat{e}_{i,t} = (E_{i,t} - E_{i,t-1})/E_{i,t-1}$ is approximately 0.65. Since we assume that $\hat{e}_{i,t}$ is normally distributed, the relation between the interquartile range and the standard deviation is as follows:

$$IQR = 2\Phi(0.75) \times \sigma,$$

where Φ is the cumulative distribution function of the standard normal distribution. If we apply this relation, we have $\sigma_{i,t} \approx \frac{0.65}{1.35} = 0.37$. Thus, we simulate a firm-time specific distribution, $\varepsilon_{i,t} \sim N(0, \sigma_{i,t})$ with $\sigma_{i,t} = 0.37 \times |E_{i,t-1}|$.

4.1.2 Simulation results

We generate data using the data generating process described above and then we estimate the coefficients relating earnings of t-1 to earnings of t for sub-samples with positive and negative earnings in year t-1, via the following autoregressive model (the EP model) as suggested by Li and Mohanram (2014):[14]

$$E_{i,t} = \beta_0 + \beta_1 E_{i,t-1} + \beta_2 I(E_{i,t-1} < 0) + \beta_3 I(E_{i,t-1} < 0) \times E_{i,t-1} + \varepsilon_{i,t} \quad (1)$$

where $E_{i,t}$ is earnings of firm i in year t, $I(E_{i,t-1} < 0)$ is an indicator variable equal to one if earnings in time t-1 are negative, zero otherwise. We use the market value of equity as a scaling variable to ensure that firms with high earnings do not get a disproportionate weight in the

[14] This model has also been employed as a regression-based, cross-sectional forecasting model in Li & Mohanram (2014).

estimation.[15] That is, the dependent variable $E_{i,t}$ is scaled by market value of equity at time t-1 and the independent variables are contemporaneously scaled (that is, the numerator and denominator are both at time t-1).

We vary the deletion of the most negative earnings each year between zero and ten percent in one percent increments and study the parameter estimates. We are particularly interested in the coefficient estimates $\hat{\beta}_2$ and $\hat{\beta}_3$. Our main question is: does truncation lead to biased coefficient estimates? We estimate the model for each year by median regression since median regressions produce better earnings forecasts (a detailed explanation is provided in the Appendix).[16]

Figure 2 presents the coefficient estimates with varying levels of deletion. For the complete simulated data set, the coefficient estimates are centered at the true value. That is, the estimated coefficients of the intercept, the loss dummy, and the interaction term are zero on average; the estimated coefficient on earnings is one on average. With deletion, we see that the estimates of coefficients on the intercept and the coefficients on positive earnings do not significantly differ from the estimates of the coefficients without deletion. However, we see that the average estimated coefficient for the loss dummy and the interaction term is below zero and becomes more negative the more we delete observations. This is consistent with our conjecture that non-random truncation of the sample (removing the worst-performing firms in each year)

[15] Arguments for scaling earnings by equity market capitalization are provided in Easton and Sommers (2003). We chose not to use the book value of equity because about 30 percent of performance-related delistings have negative book value and the majority of these firms report losses.

[16] We estimate the models using median regressions. There is evidence that median regressions produce better forecasts (Evans, Njorge and Yong, 2017 and Tian, Yim and Newton, 2020) and OLS regression produce unstable forecasts across firms and time (Nissim and Penman, 2001). Further, given that absolute forecast error is the primary metric to evaluate earnings forecasts, aligning the loss function of the estimation method and the out-of-sample forecasts seems natural. We also provide a theoretical and empirical motivation for this choice in the Appendix.

leads to biased estimated coefficients. To further illustrate the implications of non-random data disappearance we will revisit the simulated data in the next section.

4.2. Analyses based on actual Compustat data

We estimate annual median regressions based on model (1) using actual Compustat data. The reported coefficient estimates in Table 3 are means of the estimates of the coefficients across years.[17] The first column shows the estimated coefficients for the actual data (that is, with data disappearances). The estimate of the coefficient relating earnings of the current period to earnings of the subsequent period is statistically significant (coefficient estimate of 0.754 with a standard error of 0.042) for the sub-sample of firms with positive earnings. The estimate of the coefficient on the negative earnings interaction (labelled "Negative Earnings Dummy * Earnings") is statistically significantly negative (-0.224 with a standard error of 0.052) implying that negative earnings have a much greater tendency to mean revert (that is, they are of lower quality) than positive earnings.

4.3 Analyses with earnings of firms that disappear

4.3.1 Imputation of earnings for firms that disappear

For the analyses in this section, we, first, replace earnings of firms that disappear with their earnings of the previous year; that is, we assume earnings of these disappeared observations follow a random walk. We consider this a conservative assumption because the summary statistics presented in Section 3 show that disappearing firms are likely to have had lower earnings, on average, in the year of disappearance than in the prior year. Furthermore, if the null hypothesis in the analysis of the earnings time-series is that earnings follow a random walk, any other replacement procedure would impose a type I error a priori.

[17] Reported standard errors are Fama-MacBeth standard errors adjusted for serial correlation (one lag).

We next replace missing observations with earnings implied by earnings observations that are reported on the Quarterly Compustat dataset for the earlier quarters of the year in which they disappear from the Annual Compustat dataset. Specifically, we take the sum of the existing quarterly earnings in the year of disappearance and linearly extrapolate this sum into an annual earnings number. When, for example, a firm that disappeared had earnings data for two quarters in the year of disappearance, we take the sum of the two quarterly earnings and multiply the sum by two. When a firm that disappeared had earnings data for three quarters in the year of disappearance, we take the sum of the three quarters and multiply the sum by 4/3. If missing annual earnings observations do not have quarterly earnings observations in the delisting year, earnings are imputed using the yearly, delisting-type-specific (performance-related, non-performance-related, other) median growth rate obtained from the extrapolated quarterly earnings observations for profit and loss firms, respectively. This alternative will also be conservative if, as is likely, earnings in the actual quarter of delisting (which are unobservable) are abnormally low.

Finally, we use CRSP delisting returns to extrapolate earnings in the year before delisting to the year of delisting; the extrapolation is based on the median (trailing) earnings to price ratio for firms in the same Fama/French (48) industry.[18]

[18] Less conservative assumptions for data imputations may lead to regression coefficient estimates that are much closer to one. An example of a less conservative assumption would be recognizing the fact that earnings of loss firms that remain in the data set in year t tend to become less negative. If earnings follow a random walk on average in the negative earnings sub-sample, the probability of earnings increasing is similar to the probability of earnings decreasing; it follows that the year t earnings of firms that disappear are more likely to be more negative in year t than in year t-1. In other words, our assumption that, for disappeared firms, earnings of year t are the same as earnings of year t-1 is conservative.

We do all analyses with earnings imputed via each of these methods as well as with the average of the three imputed earnings numbers; the average may reduce the effects of idiosyncrasies in any one of the individual imputation methods.

4.3.2 Results from analyses of firms that disappear

In the second column of Table 3 Panel A, we report the estimates of the coefficient relating earnings of year t-1 to earnings of year t for the completed data set where we replace earnings of firms that disappear with the average of the earnings from each of the imputation methods. For these data, the coefficient relating earnings of the current year to positive earnings of the subsequent year is statistically significant (coefficient estimate of 0.776 with a standard error of 0.040) and it is very similar to the estimate of this coefficient in column one (0.754), which is based only on firms that remain in the Compustat sample. However, after including the imputed earnings, the estimate of the coefficient on the negative earnings interaction changes from statistically significantly negative (-0.224 with a standard error of 0.052) in column 1 to statistically significantly positive (0.101 with a standard error of 0.052) implying that negative earnings have a much *lesser* tendency to mean revert (that is, they are of higher quality) than positive earnings. That is, we reach the contrary conclusion when we control for attrition bias.

Column 3 of Table 3 Panel A reports the estimated coefficients when only performance-related delistings are replaced with the average imputed earnings. The results are similar to those reported in column 2. The coefficient relating positive earnings of the current year to earnings of the subsequent year is statistically significant (coefficient estimate of 0.751 with a standard error of 0.042). The estimate of the coefficient relating negative earnings in the current year to earnings of the subsequent year is 0.751 + 0.069 = 0.820 with a standard error of 0.034. Again, this coefficient estimate remains statistically significantly less than one. The observation that it

is higher shows that the tendency of negative earnings to mean revert is less than we would conclude with the incomplete data set.

Column 4 of Table 3 Panel A reports the estimated coefficients when only non-performance-related delistings and other delistings are replaced with average imputed earnings. The results are similar to those reported in column 1. This implies that it is not the non-performance-related delistings or "other" delistings that drive the attrition bias; non-random disappearances caused by performance-related delistings affect the estimated coefficients.

In Panel B, we replace missing annual earnings observations with annual earnings implied by random walk forecasts. We find similar results to what we find when we replace missing annual earnings with the average earnings of the three imputation methods. The estimate of the negative earnings coefficient increases from 0.530 to 0.973 after replacing missing observations. Furthermore, this effect is, primarily, driven by replaced earnings observations of performance-related delistings.

In Panels C and D, we replace missing annual earnings observations with earnings implied by extrapolating observed quarterly earnings in the quarters prior to disappearance and those implied by delisting returns and the industry earnings/price ratio. Again, we find similar results to what we find when we replace missing annual earnings with the average of earnings imputed via the three imputation methods. The estimate of the negative earnings coefficient increases from 0.530 to 0.922 and 0.905 after replacing missing observations. Furthermore, this effect seems to be mostly driven by replacing earnings observations for performance-related delistings.

To summarize, attrition bias due to the disappearance of performance-related delistings leads to a type I error. Lower estimates of earnings persistence coefficients for negative earnings

firms support the false conclusion that reporting of earnings for loss-making firms is of *lower* quality than the reporting of profits.

Since the classic paper of Beaver and Morse (1978), mean reversion in earnings has been demonstrated by forming portfolios based on price-deflated earnings and tracking the earnings over future years. In a similar fashion, each year we formed deciles of observations ranked on earnings/price ratios and we compared the median earnings/price ratio in the next year with and without imputation. Attrition bias affected the estimate of mean reversion for the decile with lowest earnings/price. For this decile, which had a median earnings/price ratio of -0.54, the median earnings/price ratio in the next year was -0.34 without imputation and -0.41 with imputation implying an attrition bias of -0.07 (untabulated). We find no evidence of attrition bias for the other deciles.

4.4. Changes in persistence over time

The plots in Figure 1 show considerable changes over time: in the percentage of firms disappearing from the Annual Compustat data; in the earnings prior to disappearance; in the returns in the year of disappearance; and, in the percentage of disappearing firm that experienced a loss in the year prior to disappearance. These changes suggest that there may be changes in the persistence of earnings over time.[19]

Figure 4 plots the annual estimates of the coefficients relating positive earnings in year t-1 to earnings in t and negative earnings in year t-1 to t, with and without imputation of disappeared earnings. We also show the 95 percent confidence interval around these estimates based on bootstrapped standard errors. The results in Figure 4A show that positive earnings have

[19] Several studies have shown changes in the types of firms included in the Compustat data over time. For example, Francis and Schipper (1999), Banker, Huang and Nataragan (2011), Eisfeldt and Papanikolaou (2013) and Srivastava (2014)) all document an increase in intangible assets. These changes also suggest that there may be changes in persistence over time.

become less persistent (that is, there is more mean-reversion); the estimates of the coefficients relating positive earnings at t-1 to earnings at t are close to one until 1985, when they become significantly less than one and remain so. Attrition bias does not have a significant effect on the estimates of the persistence of positive earnings; although the estimates of the coefficients without the imputed data are less than those with imputation, the confidence intervals around these coefficients overlap in all years. On the other hand, losses have become more persistent in later years. The 95 percent confidence interval around the blue line, which is the plot of estimated persistence based on the data without imputation of disappeared observations is always less than one suggesting that there is significant mean reversion in negative earnings. Notably, however, after removing the effects of attrition bias by imputing the earnings of the disappeared earnings data, there is a significant increase in the estimate of persistence and it is not significantly different from one in many of the later years.

5. Effects of Disappearance from Compustat on Accuracy of Regression-based Forecasting Models

In this section, we highlight the importance of the disappearance of observations from the Compustat data when estimating cross-sectional earnings forecasting models based on these data. Because we are interested in the time-series of earnings, we focus on the earnings persistence model (that is, the autoregressive model employed in the previous analyses) although inferences and conclusions are very similar for analyses based on the model in Hou, van Dijk and Zhang (2012), HVZ, and for those based on the residual income (RI) forecasting model in Li and Mohanram (2014). Specifically, the HVZ, EP and RI models produce forecasts that are more accurate than a random walk when we ignore the effect of disappearance from the data. In other words, the null hypothesis that earnings forecasts from these cross-sectional regression models do not beat a random walk is rejected. We show that this may be a type I error. When we replace

the disappeared earnings observations and analyze the completed data, we can no longer reject the null hypothesis that the regression models do not beat a random walk. We begin by demonstrating the effect of attrition bias on estimates of forecast accuracy with simulated data.

5.1 Simulating forecasts

We run a simulation to show that, in a setting in which the random walk is the true forecasting model, truncation leads to a forecast bias, thereby producing forecasts that appear to be more accurate than a random walk. That is, absent truncation, the estimated coefficients of the EP model would not lead to a rejection of the random walk as the true earnings process. When truncation is introduced, the coefficient estimates in the EP model are no longer consistent with the random walk model. The forecasts generated from the model lead to the spurious conclusion that the EP model provides a better forecast than the true model, i.e. the random walk.

For the simulated data, as described above, we obtain random walk forecasts and forecasts from the estimated EP model. We compute the mean absolute forecast error, or MAFE (the mean of the absolute value of the difference between earnings realizations and earnings forecasts). We compare the MAFE generated by the EP model and the MAFE from the RW model.

Since raw measures of MAFE are hard to interpret, we compute the following normalized quantity,

$$MAFE_s = \frac{MAFE_s^{\mathrm{EP}}}{MAFE_s^{\mathrm{RW}}}$$

where $MAFE_s^{\mathrm{EP}}$ is the mean absolute percentage error in forecasts based on the EP model and $MAFE_s^{\mathrm{RW}}$ is the mean absolute percentage error in forecasts based on the random walk model. That is, we normalize the errors by the random walk median absolute forecast error. Thus, if the relative error is greater than one, then the random walk is apparently a superior model. We also

examine the distribution (over the simulations) of $MAFE_s$ for different levels of truncation each year. The results of these analyses are summarized in Figure 3. The Figure on the left shows the relative mean absolute forecast errors for the model estimated with median regressions and shows how the forecast error increases as we increase the percentage of deleted observations.

We see a clear pattern; the EP model performs better relative to a random walk as the truncation percentage increases. We repeat the simulation for the relative mean squared error (on the right side of Figure 3) and the same pattern is observed - note that the magnitudes are comparable but because the distribution of MSE is more dispersed, the boxplots for MSE require a larger interval for the y-axis.

5.2 Pooled cross-section and time-series analysis of forecast accuracy

We present the results from the estimation of the accuracy of regression-based forecasts in Table 4, column 1.[20] Forecast accuracy is defined as the mean absolute forecast error (the mean of the absolute value of the difference between earnings realizations and earnings forecasts), and we report the error relative to the mean absolute forecast error of the random walk forecast.

To evaluate forecast accuracy comprehensively, we require realized earnings for all observations in the estimation sample and in the out-of-sample period. Hence, we examine the effects of disappearance on forecast accuracy using the four forms of data replacement discussed in the analyses of the effects of disappearance on estimates of earnings persistence. While the first row of Table 4 reports forecast accuracy results for actual Compustat data, the second row

[20] Consistent with much of the recent literature, we cluster standard errors in two-dimensions - by year and by firm. We cluster standard errors by firm because the errors may be correlated within firms due to common accounting treatment across years. We cluster standard errors by year because earnings shocks could affect particular industries, which could lead to correlated errors within years (Petersen, 2009).

reports forecast accuracy results for Compustat data after replacing all missing earnings observations with the average of the three imputed earnings and the remaining rows report the results for only imputing performance-related delistings and only imputing non-performance-related and other delistings, respectively. The regression-based forecasts based on the data without imputed earnings are significantly more accurate than a random walk for both positive earnings (difference of -0.002) and much more so for negative earnings (-0.036).[21] When we impute earnings for the observations that disappeared, the difference drops substantially to -0.002 for loss firms and this difference is no longer statistically different from zero (standard error of 0.001).[22] The results are quite similar across each of the imputation methods. As with the results from the estimation of the regression coefficients, the attrition bias is due primarily to performance related delistings; this can be seen in rows three and four of each Panel, respectively.

The changes in firm characteristics over time that are shown in Figure 1 as well as the differences in estimates of persistence and the effect of attrition bias on the estimates shown in Figure 4 suggest that there may also be changes in forecast accuracy over time. Figure 5A shows the year-by-year difference between the MAFE for forecasts based on the regression model and those based on the random walk for observations with positive earnings in year t-1 and for observations with negative earnings in year t-1. Confidence intervals (95 percent) around these MAFEs are also shown. Figure 5B is a similar graph for observations with negative earnings in year t-1. Unlike the pattern of declining estimates of the coefficient relating positive earnings at t-1 to earnings at t, there is no apparent evidence of a time-trend for differences in accuracy. The

[21] The error differences for positive earnings are -0.002 and -0.001 and for negative earnings are -0.363 and -0.373 for the RI model and the HVZ model, respectively.
[22] The error differences are -0.002 and -0.003 for the RI model and the HVZ model, respectively. Both error differences are not statistically significantly different from zero.

evidence of attrition bias in the estimate of the forecast accuracy of forecasts based on negative earnings in the pooled cross-section time series analyses (above) is also evident in the year-by-year analyses where the difference between the regression-based forecasts and random walk after imputation of delisted data is significantly less than the difference before imputation and not significantly different from zero in 28 out of 37 years.[23]

6. The Effects of Disappearance are much more Evident for Small Firms

Firms that disappear from Compustat for performance-related reasons are much smaller than firms that remain (performance-related disappearances have a mean/median market capitalization of 57.11/7.31 million compared to 2,340.49/134.45 million for firms that remain on Compustat – see Table 1). Furthermore, firms that disappear for performance-related reasons tend to have negative earnings in the year before disappearance (see Table 1) and these negative earnings tend to get worse in the year of disappearance (see Table 2). This suggests that the over-statement of mean reversion may be concentrated in these smaller firms. To examine this more closely, we split the sample each year into those above the median market capitalization and those below the median market capitalization. Importantly, firms below the median market capitalization comprise less than three percent of the market capitalization of the entire Compustat data even though they represent 50 percent of observations. We find that the over-statement of the mean reversion is much greater for the small firms and the false rejection of the null that the regression-based forecasts beat random walk forecasts occurs only for these small firms. The results for the analyses of the two sub-samples of observations are summarized in Tables 5, 6 and 7.

[23] The number of years in which the accuracy difference is not significantly different from zero when no disappearances are imputed is 15 out of 37 years.

The statistics in Table 5 highlight differences in earnings in the year before disappearance and returns in the year of disappearance between remaining firms and disappearing firms. For the sample of larger firms, the main reason for disappearance is non-performance related and most of these firms (1.78 percent) experienced positive earnings in the year before disappearance (mean/median scaled earnings of 0.07/0.06, which is quite similar to those of firms remaining on Compustat, 0.07/0.06). Delisting returns for these firms were substantially greater (mean/median returns of 35/27 percent compared with 14/10 percent for those that remain on Compustat). This is in contrast to the sample of smaller firms where the main reason for disappearance is performance-related and most of these firms (2.30 percent) experience very negative earnings in the year of disappearance (mean/median scaled earnings of -2.34/-0.65, which is much lower than those of firms remaining on Compustat, -0.43/-0.15). Delisting returns for these firms were substantially more negative (mean/median returns of -45/-57 percent compared with 16/-11 percent for those that remain on Compustat).

We report scaled earnings and the change in scaled earnings in the last quarter before disappearance from Compustat in Table 6. Again, we focus on the main reason for disappearance – non-performance related delistings of large firms with positive earnings in the year prior to delisting and performance-related delistings for small firms with negative earnings in the year prior to delisting. Median scaled quarterly earnings in the last quarter before delisting for the large firms with positive earnings in the year prior to the year of disappearance, which are delisted for non-performance-related reasons, are similar to those of firms that remain on Compustat.[24] However, mean and median scaled quarterly earnings in the last quarter before

[24] The very high mean scaled earnings (1.65) and mean change in scaled earnings (1.63) is driven by outlying observations where the scalar (market capitalization) is very low. Upon inspection of the outliers, the extreme observations are five companies of which two were acquired by other companies, two merged with other companies, and one was subject to a leveraged buyout. When using lagged market equity for these companies, the mean scaled earnings for all reported observations is 0.01 and the mean change in scaled earnings is -0.01.

delisting for small firms that disappear for performance-related reasons are much more negative (-0.86/-0.25) than scaled quarterly earnings of firms that remain on Compustat (-0.16/-0.04). Also, the decline in quarter-over-quarter earnings is much greater for these firms (-0.71/-0.16) than those that remain on Compustat (-0.03/0.00).

The differences between the earnings and returns of disappearing large firms and disappearing small firms summarized in Tables 5 and 6 suggest that the effects of disappearance on inferences about mean reversion and random walk properties of earnings may be quite different across these two sub-samples. Results of analyses of mean reversion for each of these sub-samples are summarized in Panel A of Table 7 and results of analyses of forecasts are summarized in Panel B of Table 7. We provide results for analyses where disappeared earnings observations are replaced with the average of earnings from the three imputation methods.[25]

Evidence of the overstatement of the tendency of earnings to revert toward the mean is presented in Panel A of Table 7 for both large firms and for small firms, although it is much more evident for small firms. Completing the data set by replacing missing observations increases the estimates of the coefficient relating positive earnings to next-period earnings by a small amount from 0.814 to 0.830 for larger firms and the estimate of the coefficient relating negative earnings to next-period earnings also increases by a small amount from 0.528 to 0.562 for these firms. For the smaller firms, the estimate of the coefficient relating positive earnings to next-period earnings increases from 0.682 to 0.711. The key observation is that, consistent with the estimates for the pooled large and small firms, attrition bias has a substantial effect on the estimate of the coefficient on the negative earnings interaction dummy. After removing the

[25] Results are quite similar and inferences are the same if we replace disappeared observations via any of the three imputation methods.

effects of attrition bias this estimate changes from significantly negative (-0.162 with a standard error of 0.058), implying that the reporting of losses is of *lower* quality than the reporting of profits, to significantly positive (0.183 with a standard error of 0.058), implying that the reporting of losses is of *higher* quality than the reporting of profits.

Table 7, Panel B summarizes the results for analyses of forecast errors for the two subsamples of observations. The attrition bias that was evident in the assessment of forecast accuracy, which was observed for the pooled positive and negative earnings samples, is apparently due to attrition bias in the small firm negative earnings sample (the difference in forecast accuracy with attrition bias is -0.023 and it is -0.002 without attrition bias).[26]

7. Persistence of Earnings Changes

Consistent with the accounting literature and the extant cross-sectional regression-based earnings forecasting models to which we refer, we have, in the previous analyses, studied the persistence of the level of earnings. In short, our study of persistence focused on whether firms with high (low) earnings continue to have high (low) earnings. In this section we briefly examine a related but different phenomenon; mean reversion of changes in earnings as studied by Fama and French (2000). Similar to our analyses of the time-series of earnings levels, we estimate: the persistence (and mean reversion) of earnings changes; the effect of attrition bias on these estimates; the accuracy of forecasts from cross-sectional regression forecasting models based on earnings changes; and, the effects of attrition bias on the comparison of the accuracy of these models compared with random walk forecasts.

[26] Similar attrition bias is seen in the estimates of differences between the estimates of forecast accuracy with the other regression based models (HVZ and RI); for the HVZ model, the difference in forecast accuracy with attrition bias is -0.024 and it is -0.001 without attrition bias; for the RI model, the difference in forecast accuracy with attrition bias is -0.023 and it is -0.002 without attrition bias.

We run the following regression:

$$(E_{i,t} - E_{i,t-1}) = \alpha_1 + \alpha_2(E_{i,t-1} - E_{i,t-2}) + \alpha_3 I(E_{i,t-1} - E_{i,t-2} < 0) + \alpha_4 I(E_{i,t-1} - E_{i,t-2} < 0) * (E_{i,t-1} - E_{i,t-2}) + \alpha_5 I(E_{i,t-1} < 0) + \alpha_6 I(E_{i,t-1} - E_{i,t-2} < 0) * I(E_{i,t-1} < 0) + \alpha_7 I(E_{i,t-1} < 0) * (E_{i,t-1} - E_{i,t-2}) + \alpha_7 I(E_{i,t-1} < 0) * I(E_{i,t-1} - E_{i,t-2} < 0) * (E_{i,t-1} - E_{i,t-2}) + \varepsilon_{i,t} \quad (2)$$

where $I(E_{i,t-1} - E_{i,t-2} < 0)$ is a dummy variable equal to one if the change in earnings from *t-2* to *t-1* is negative, zero otherwise. All other variables are as defined in regression (1).

The results of our analyses are summarized in Table 8. For ease of interpretation we report the coefficient estimates for the four dummy variable partitions of the data: positive earnings levels, positive earnings change; positive earnings levels, negative earnings change; negative earnings levels, positive earnings change; and, negative earnings levels, negative earnings change. The most noticeable result is the more-than-complete mean reversion (coefficient estimate of -1.140 with a standard error of 0.096) for the firms that have experienced a loss and a negative change in earnings; that is it seems that a dollar decrease in earnings is followed, on average by a $1.14 increase in earnings in the next year. Importantly, however, a considerable portion of this estimated mean reversion is due to attrition bias; after imputing earnings for the firms that disappear from the Compustat data, the estimate of mean reversion is -0.616 (with a standard error of 0.112).[27]

Although the extant regression-based forecasting literature focusses on earnings levels, we, for the sake of completeness, analyze the effects of attrition bias on the assessment of the forecast accuracy of forecasts based on regression (2). The results in Table 8, Panel B show that

[27] We also examined changes in persistence over time for each of the partitions of the data. Unlike the analyses of persistence of earnings levels, there were no clear trends. There was, however, a noticeable change in 2009 and 2010 following the global financial crisis; the extent of mean revision increased considerably for observations with positive earnings changes and profits and for firms with negative earnings changes and losses. Also, unlike other years, there was evidence of persistence of negative earnings changes for profitable firms.

attrition bias leads to the incorrect conclusion that the regression based forecasts are significantly more accurate than random walk forecasts (a difference of -0.01 percent of price with a standard error of 0.001) whereas after talking account of attrition bias, the difference is not significantly different from zero (a difference of -0.001 percent of price with a standard error of 0.001).

8. Conclusion

In this paper, we highlight the existence and importance of the fact that observations disappear from the Compustat data set. Specifically, we highlight the effect of these disappearances on estimates of earnings persistence and on the evaluation of the accuracy of regression based forecasting models, which are based on observations of earnings in two consecutive years. We show that this previously neglected consideration can lead to type I errors. For example, after removing the effects of attrition we observe that earnings of loss-making firms are more persistent than earnings of profit-making firms; before removing the attrition bias, earnings of loss making firms are estimated to be less persistent.

Our results suggest that the effects of attrition bias may be considerable in studies that examine the time-series properties of earnings and regression-based forecasts of earnings. Thus, we recommend that researchers evaluate the accuracy of their models using subsamples where the effects of disappearance from the data set is not likely to be a large concern. We also suggest methods that future researchers could use to replace missing earnings observations in order to mitigate the effects of attrition bias.

References

Ball R., and Watts, R. 1979. Some additional evidence on survival bias. *Journal of Finance.* 34(1): 197-206.

Banker, R.D., Huang, R., Natarajan, R. 2011. Equity incentives and long-term value created by SG&A expenditure. *Contemporary Accounting Research.* 28: 794–830.

Basu, S. and Markov, S. 2004. Loss function assumptions in rational expectations tests on financial analysts' earnings forecasts. *Journal of Accounting and Economics*. 38: 171-203.

Beaver, W., and Morse, W. 1978. What determines price-earnings ratios. *Financial Analysts Journal.* 34(4): 65-78.

Beaver, W., McNichols, M. and Price, R. 2007. Delisting returns and their effect on accounting-based market anomalies. *Journal of Accounting and Economics*. 43(2-3): 341-368.

Bradshaw, M. 2011. Analysts' forecasts: What do we know after decades of work? Unpublished working paper. Boston College.

Brilleman, S., Pachana, N. and Dobson, A. 2010. The impact of attrition on the representativeness of cohort studies of older people. *BMC Medical Research Methodology.* 10(71): 1-9.

Brown, S., Goetzmann, W., Ibbotson, R. and Ross, S. 1992. Survivorship bias in performance studies. *The Review of Financial Studies.* 5(4): 553-580.

Chang, W., Landsman, W., and S. Monahan. 2013. Selecting an accounting based valuation model, INSEAD working paper.

Dechow, P. and Ge, W. 2006. The persistence of earnings and cash flows and the role of special items: Implications for the Accrual Anomaly. *Review of Accounting Studies*. 11(2): 253-296.

Dechow, P. Ge, W., and Schrand, C. 2010. Understanding earnings quality: A review of proxies, their determinants and their consequences. *Journal of Accounting and Economics.* 50(2-3): 344-401.

Easton, P. and Sommers, G. 2003. Scale and scale effect in market-based accounting research. *Journal of Business Finance and Accounting.* 30 (1-2): 25-56.

Easton, P. and Zmijewski, M. 1989. Cross-sectional variation in the stock market response to earnings announcements. *Journal of Accounting and Economics*. 11: 117-141.

Eisfeldt, A. and Papanikolaou, D. 2013. Organization capital and the cross section of expected returns. *The Journal of Finance*. 68: 1365–1406.

Evans, M., Njoroge, K., and Yong, K. 2017. An examination of the statistical significance and economic relevance of profitability and earnings forecasts from models and analysts. *Contemporary Accounting Research.* 34 (3): 1453-1488.

Fama, E. and French, K. 2000. Forecasting profitability and earnings. *Journal of Business.* 73 (2): 161-175.

Francis, J. and Schipper, K. 1999. Have financial statements lost their relevance? *Journal of Accounting Research.* 37 (2): 319–352.

Freeman, R. and Tse, S. 1992. A nonlinear model of security price responses to unexpected earnings. *Journal of Accounting Research*. 30 (2): 185-209.

Gerakos, J. and Gramacy, R. 2013. Regression-Based Earnings Forecasts. Working Paper: University of Chicago.

Gu, Z. and Wu, J. 2003. Earnings skewness and analysts forecast bias. *Journal of Accounting and Economics*. 35 (1): 5-29.

Hayn, C. 1995. The information content of losses. *Journal of Accounting and Economics.* 20(2): 125-153.

Horowitz, J. 1998. Bootstrap methods for median regression models. *Econometrica*. 66 (6): 1327-1351.

Hou, K., van Dijk, M. and Zhang, Y. 2012. The implied cost of capital: A new approach. *Journal of Accounting and Economics*. 53: 504-526.

Jones, C.S. and Tuzel, S. 2013. Inventory investment and the cost of capital. *Journal of Financial Economics*. 107(3): 557-579.

Koenker, R. 2005. Quantile regression. Cambridge University Press, Cambridge, United Kingdom.

Kormendi, R. and Lipe, R. 1987. Earnings innovations, earnings persistence and stock returns. *The Journal of Business.* 60(3): 323-345.

Lachin, J.M. 2000. Statistical considerations in the intent-to-treat principle. *Controlled Clinical Trials*. 21: 167-189.

Lawrence, A., Sloan, R. and Sun, E. 2018. Why are losses less persistent than profits? Curtailment vs. conservatism. *Management Science*. 64(2): 495-981.

Lee, C., So, E., and C. Wang. 2017. Evaluating Firm-level expected return proxies. Working paper, Stanford University.

Li, K. and Mohanram, P. 2014. Evaluating cross-sectional forecasting models for implied cost of capital. *Review of Accounting Studies*. 19(3): 1152-1185.

MacKinnon, J. and White, H. 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*. 29(3): 305-325.

Nissim, D. and Penman, S. 2001. Ratio analysis and equity valuation: From research to practice. *Review of Accounting Studies.* 6: 109-154.

Patatoukas, P.N. 2012. Customer-base concentration: implications for firm performance and capital markets. *The Accounting Review.* 87(2): 363-392.

Petersen, M., 2009. Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies.* 22(1): 435-480.

Shumway, T., 1997. The delisting bias in CRSP data. *The Journal of Finance.* 52(1): 327-340.

Shumway, T. and Warther, V. 1999. The delisting bias in CRSP's Nasdaq data and its implications for the size effect. *The Journal of Finance.* 54(6): 2361-2379.

So, E. 2013. A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics.* 108(3): 615-640.

Srivastava, A. 2014. Why have measures of earnings quality changed over time? *Journal of Accounting and Economics.* 57(2): 196-217.

Tian, H., Yim, A., and Newton, D. 2020. Tail heaviness, asymmetry, and profitability forecasting by quantile regression. *Management Science.* 55(12).

Watts, R., 1970. The informational content of dividends. Working Paper. University of Chicago.

Watts, R. and Leftwich, R. 1977. The time-series of annual earnings. *Journal of Accounting Research.* 15(2): 253-271.

Figure 1: Comparison of Firm Characteristics for Delisting and Non-Delisting Observations

These figures compare the characteristics of firms that disappear from Compustat with the characteristics of firms that remain. The percentage of firms disappearing (Figure 1A) is the percentage of disappearances relative the total number of observations by year. Earnings and the percentage of loss firms are for the year prior to disappearing. Earnings (Figure 1B) are scaled by market value of equity. Annual returns (Figure 1C) are for the year of the disappearance. Percentage of loss firms (Figure D) is the percentage of firms with negative earnings. Performance-related delistings are firms with CRSP delisting codes 500 and 520 to 584 (inclusive) with missing future earnings; firms with other delisting codes with missing future earnings are non-performance-related delistings. Other delistings are all observations that do not have a CRSP delisting code but have missing future earnings.

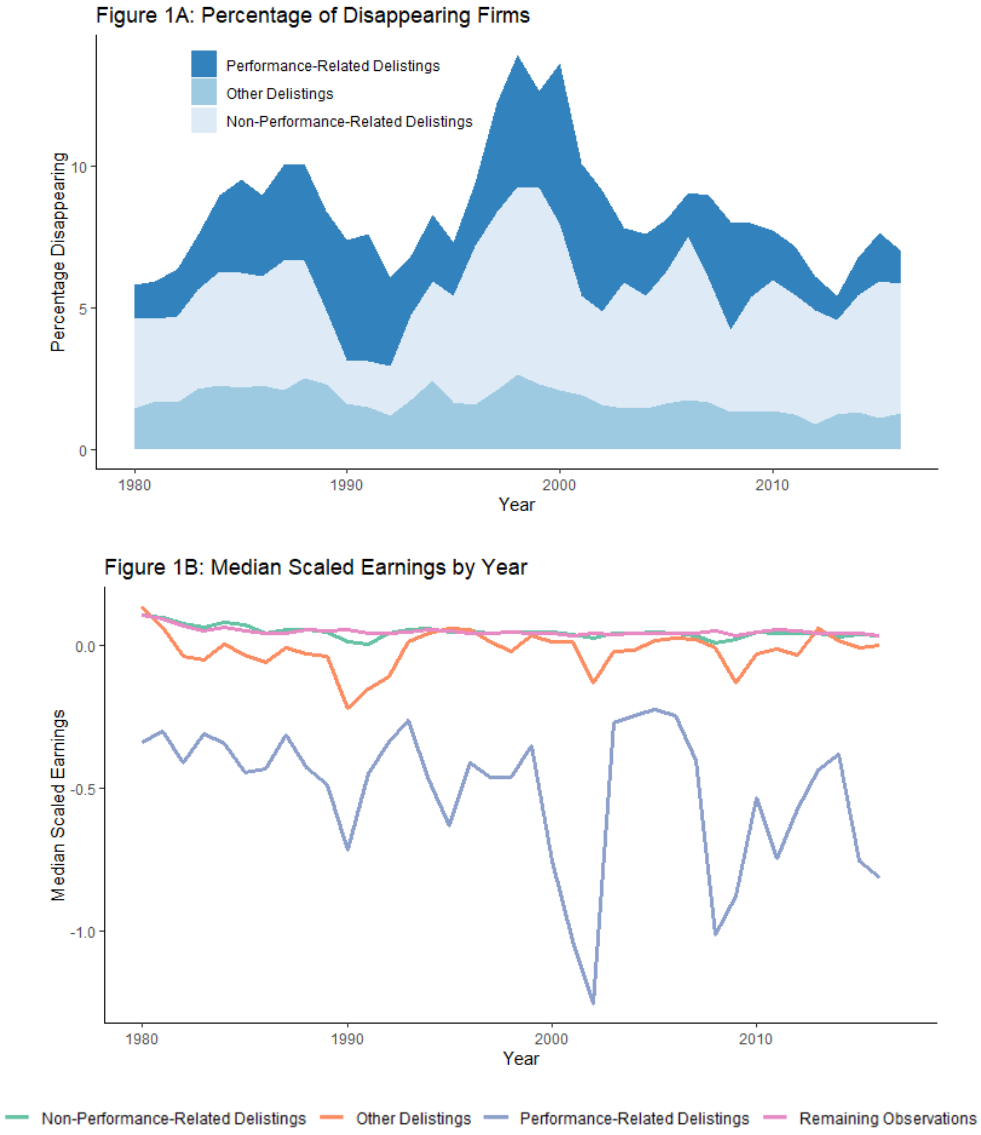


Figure 1C: Median Annual Returns by Year

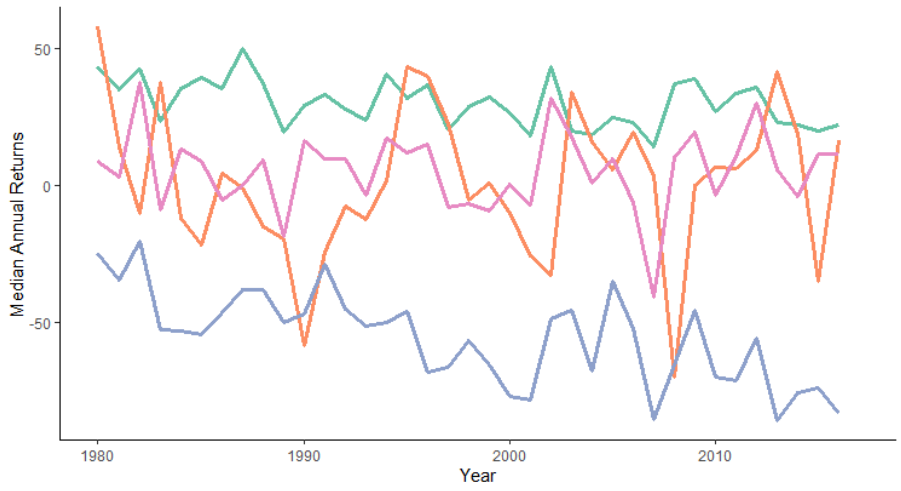


Figure 1D: Percentage of Loss Firms by Year

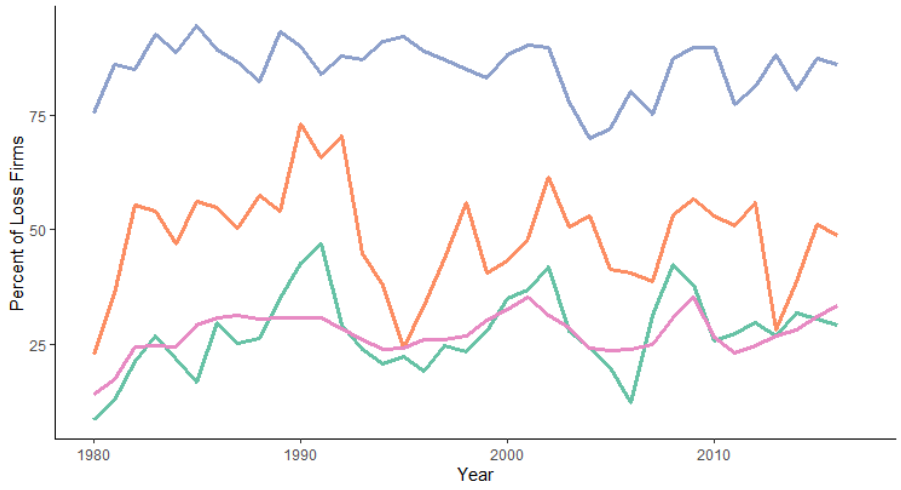


Figure 2: Coefficients from Regressions of Earnings on Lagged Earnings Using Simulated Data in which the Percentage of Deleted Observations Varies

These figures compare the estimates of coefficients from the following regression for data simulated as a random walk from the previous year's observed earnings on Compustat

$$E_{i,t} = \beta_0 + \beta_1 E_{i,t-1} + \beta_2 I(E_{i,t-1} < 0) + \beta_3 I(E_{i,t-1} < 0) \times E_{i,t-1} + \varepsilon_{i,t} \qquad (1)$$

where $E_{i,t}$ is earnings of firm i in year t, $I(E_{i,t-1} < 0)$ is an indicator variable equal to one if earnings in time t-1 are negative, zero otherwise. Coefficient estimates (mean and 95 percent confidence interval) are shown for simulated data in which we delete zero to ten percent of the simulated data (truncation fraction).

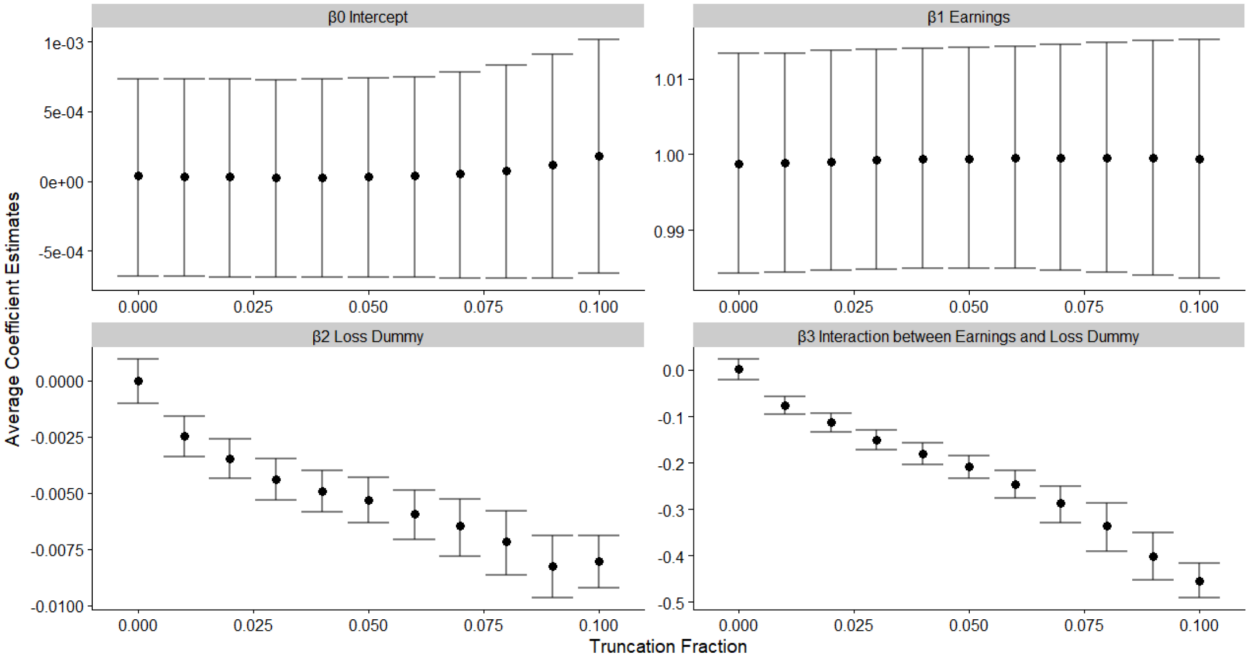


Figure 3: Forecasts of Earnings from a Regression-based Forecasting Model Using Simulated Data in which the Percentage of Deleted Observations Varies

These figures compare the forecast accuracy estimates of coefficients from the following regression for data simulated as a random walk from the previous year's observed earnings on Compustat

$$E_{i,t} = \beta_0 + \beta_1 E_{i,t-1} + \beta_2 I(E_{i,t-1} < 0) + \beta_3 I(E_{i,t-1} < 0) \times E_{i,t-1} + \varepsilon_{i,t} \qquad (1)$$

where $E_{i,t}$ is earnings of firm i in year t, $I(E_{i,t-1} < 0)$ is an indicator variable equal to one if earnings in time t-1 are negative, zero otherwise. EP model. Normalized mean absolute forecast errors are calculated as follows

$$MAFE_s = \frac{MAFE_s^{\mathrm{EP}}}{MAFE_s^{\mathrm{RW}}}$$

where $MAFE_s^{\mathrm{EP}}$ is the mean absolute percentage error in forecasts based on model (1) and $MAFE_s^{\mathrm{RW}}$ is the mean absolute percentage error in forecasts based on the random walk model. (MAFE is the mean of the absolute value of the difference between earnings realizations and earnings forecasts). The graph shows the median, the 95% confidence interval for the median, the 25th and 75th quantile, and the outlying observations which are smaller/higher than 1.5*IQR (IQR is the interquartile range).

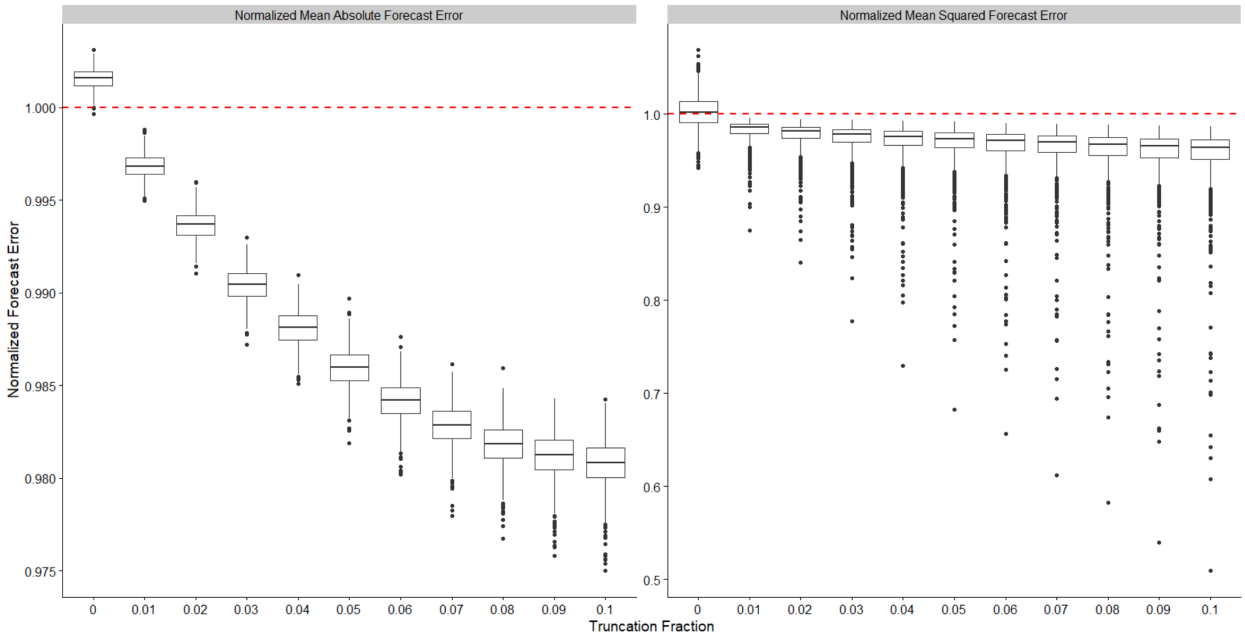


Figure 4: Time-series of Estimated Annual Slope Coefficients with and without Imputation for Missing Earnings Observations

This figure shows parameter estimates obtained from median regressions (regression (1)) for the sample with and without imputation. The negative earnings coefficient represents the sum of the positive earnings coefficient and the interaction term coefficient. Missing observations are imputed using the average of the three imputation procedures. 95 percent confidence intervals are shaded in gray. Confidence intervals are obtained from bootstrapped standard errors.

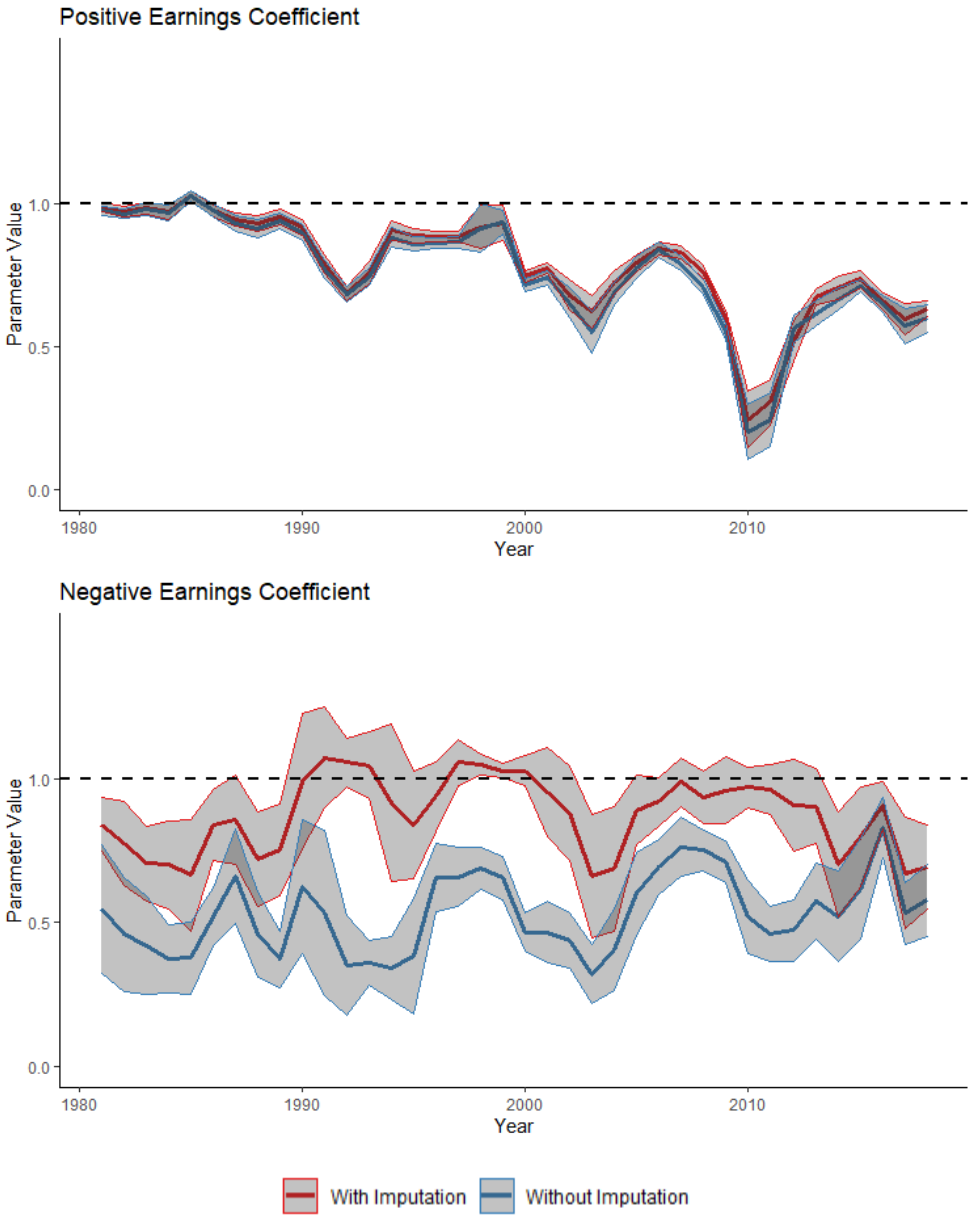
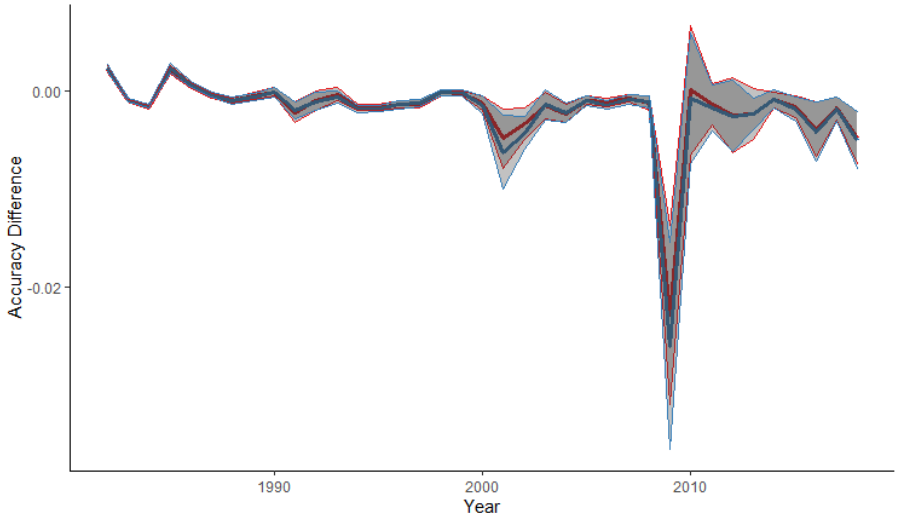


Figure 5: Time-series of Forecast Accuracy Difference between the EP Model and Random Walk

This figure shows the difference between the forecast accuracy of the EP model (regression (1)), with and without imputations, and the forecast accuracy of the random walk model overt time. Forecast accuracy is the absolute value of the difference between earnings realizations and earnings forecasts. Missing observations are imputed using the average of the three imputation procedures. 95 percent confidence intervals are shaded. Standard errors for the confidence intervals are computed as the standard deviation of the annual accuracy differences divided by the square root of the number of observations per year.

Panel A: Subsample with positive earnings at time t



Panel B: Subsample with negative earnings at time t

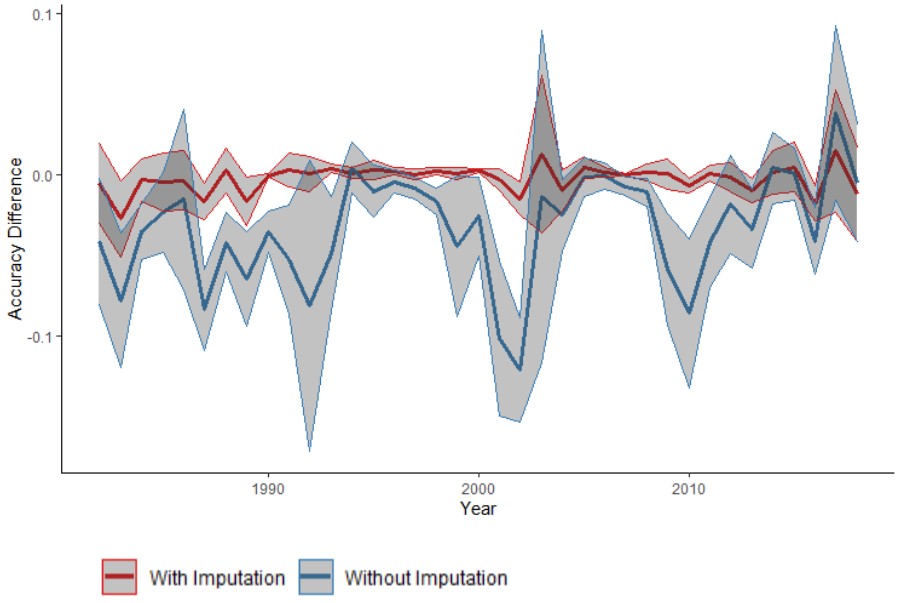


Table 1: Summary Statistics

The table compares the characteristics of firms that disappear from Compustat with the characteristics of firms that remain. Earnings (scaled by market value of equity obtained from Compustat at the fiscal year end) are reported for the year before the disappearance. Positive and negative earnings are split into separate samples in the year prior to delisting. Returns are reported for the year of the disappearance. Delistings are considered to be performance delistings when the delisting code equals 500, or falls between 520 and 584 (inclusive), non-performance delistings when the delisting code does not equal to 500, or does not fall between 520 and 584 (inclusive), and other disappearances when there are missing future annual earnings and the delisting code is missing.

Panel A: Firms remaining in Compustat (N=175,641)

	All Firms		Positive Earnings		Negative Earnings	
	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings in year before disappearance	-0.04	0.05	0.08	0.06	-0.36	-0.12
Annual returns (percent) in year other firms disappear	15	5	16	9	13	-11
Equity market value	2,340.49	134.45	3,075.81	226.87	406.87	38.81
Percentage of all observations		91.52%		66.31%		25.22%

Panel B: Firms disappearing for performance-related reasons (N=5,283)

	All Firms		Positive Earnings		Negative Earnings	
	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings in year before disappearance	-1.96	-0.49	0.25	0.07	-2.30	-0.64
Annual returns (percent) in year of disappearance	-43	-56	-21	-25	-46	-59
Equity market value	57.11	7.31	204.36	16.76	34.21	6.50
Percentage of all observations		2.75%		0.37%		2.38%

Panel C: Firms disappearing for non-performance-related reasons (N=7,812)

	All Firms		Positive Earnings		Negative Earnings	
	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings in year before disappearance	-0.10	0.05	0.08	0.06	-0.59	-0.13
Annual returns (percent) in year of disappearance	40	29	41	30	39	22
Equity market value	1,186.20	144.41	1,450.14	203.70	460.76	56.75
Percentage of all observations		4.07%		2.98%		1.09%

Panel D: Firms disappearing for other reasons (N=3,172)

	All Firms		Positive Earnings		Negative Earnings	
	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings in year before disappearance	-0.35	0.01	0.08	0.07	-0.81	-0.25
Annual returns (percent) in year of disappearance	8	0	30	28	-16	-40
Equity market value	412.42	35.24	615.07	81.04	199.27	14.94
Percentage of all observations		1.65%		0.85%		0.81%

Table 2: Scaled Quarterly Earnings for Remaining and Disappearing Firms

This table provides time-series averages of the mean and median scaled earnings for the quarters before the firm disappears from the Compustat Annual data (except for the section related to firms remaining in the sample). Time t is defined as the last quarter before the firm disappears. Positive and negative earnings are split in the year prior to delisting. Scaled earnings are earnings deflated by market value of equity (market value of equity is obtained from Compustat at the fiscal year end). Performance-related delistings are firms with CRSP delisting codes equal to 500 or between 520 and 584 (inclusive) with missing future earnings; firms with other delisting codes that are missing future earnings are non-performance-related delistings. Other delistings are all observations that do not have a CRSP delisting code but have missing future earnings.

Panel A: Firms remaining in Compustat

	Positive Earnings		Negative Earnings	
	Mean	Median	Mean	Median
Scaled Earnings$_t$	0.03	0.02	-0.13	-0.02
Scaled Earnings$_t$ – Scaled Earnings$_{t-4}$	0.00	-0.00	-0.03	0.00

Panel B: Firms disappearing for performance-related reasons

	Positive Earnings		Negative Earnings	
	Mean	Median	Mean	Median
Scaled Earnings$_t$	-0.24	-0.03	-0.88	-0.26
Scaled Earnings$_t$ – Scaled Earnings$_{t-4}$	-0.28	-0.10	-0.73	-0.17

Panel C: Firms disappearing for non-performance-related reasons

	Positive Earnings		Negative Earnings	
	Mean	Median	Mean	Median
Scaled Earnings$_t$	1.07	0.01	-0.12	-0.03
Scaled Earnings$_t$ – Scaled Earnings$_{t-4}$	1.05	-0.00	-0.04	-0.00

Panel D: Firms disappearing for other reasons

	Positive Earnings		Negative Earnings	
	Mean	Median	Mean	Median
Scaled Earnings$_t$	-0.04	0.01	-0.33	-0.08
Scaled Earnings$_t$ – Scaled Earnings$_{t-4}$	-0.06	-0.01	-0.22	-0.03

Table 3: Coefficient Estimates for Earnings Scaled by Market Equity

This table presents average coefficient estimates for the EP model (regression (1)) when estimating the model on the actual sample and when estimating the model on a modified sample that replaces missing observations due to disappearance from Compustat. In Panel A, we use the average imputed earnings of our three imputation procedures. In Panel B, we use earnings from the previous period to replace missing earnings observations. In Panel C, we replace missing annual earnings with an extrapolated value based on the firm's reported quarterly earnings in the associated year. In Panel D we impute earnings using the median earnings/price ratio of firms in the same Fama/French 48 industry and CRSP delisting returns. The Negative Earnings Coefficient is the joint effect of the respective earnings coefficient ("Earnings") and the interaction term coefficient ("Negative Earnings Dummy *Earnings"). Standard errors (in parentheses) are Fama-MacBeth standard errors adjusted for serial correlation. * indicates significance at the ten percent level, ** indicates significance at the five percent level, and *** indicates significance at the one percent level.

Panel A: Average of imputation procedures

	Actual Sample	Modified Sample – All obs. imputed	Modified Sample – Perf. del. imputed	Modified Sample – Non-perf. del. and other imputed
Intercept	0.019*** (0.002)	0.018*** (0.002)	0.019*** (0.002)	0.018*** (0.002)
Earnings	0.754*** (0.042)	0.776*** (0.040)	0.751*** (0.042)	0.779*** (0.039)
Negative Earnings Dummy	-0.047*** (0.004)	-0.017*** (0.003)	-0.021*** (0.003)	-0.041*** (0.004)
Negative Earnings Dummy *Earnings	-0.224*** (0.052)	0.101* (0.052)	0.069 (0.061)	-0.174*** (0.048)
Negative Earnings Coefficient	0.530*** (0.028)	0.877*** (0.028)	0.820*** (0.034)	0.604*** (0.029)
Average number of observations per year	9,155	9,898	9,382	9,671

Panel B: Random walk imputation

	Actual Sample	Modified Sample – All obs. imputed	Modified Sample – Perf. del. imputed	Modified Sample – Non-perf. del. and other imputed
Intercept	0.019*** (0.002)	0.014*** (0.002)	0.017*** (0.002)	0.016*** (0.002)
Earnings	0.754*** (0.042)	0.831*** (0.034)	0.788*** (0.037)	0.802*** (0.037)
Negative Earnings Dummy	-0.047*** (0.004)	-0.011*** (0.002)	-0.013*** (0.003)	-0.037*** (0.003)
Negative Earnings Dummy *Earnings	-0.224*** (0.052)	0.143*** (0.039)	0.137** (0.052)	-0.162*** (0.046)
Negative Earnings Coefficient	0.530*** (0.028)	0.973*** (0.013)	0.925*** (0.024)	0.640*** (0.028)
Average number of observations per year	9,155	10,018	9,432	9,741

Panel C: Extrapolated quarterly earnings imputation

	Actual Sample	Modified Sample – All obs. imputed	Modified Sample – Perf. del. imputed	Modified Sample – Non-perf. del. and other imputed
Intercept	0.019*** (0.002)	0.020*** (0.002)	0.021*** (0.002)	0.019*** (0.002)
Earnings	0.754*** (0.042)	0.732*** (0.045)	0.733*** (0.047)	0.752*** (0.041)
Negative Earnings Dummy	-0.047*** (0.004)	-0.013*** (0.003)	-0.016*** (0.003)	-0.041*** (0.003)
Negative Earnings Dummy *Earnings	-0.224*** (0.052)	0.190*** (0.069)	0.136* (0.072)	-0.146*** (0.050)
Negative Earnings Coefficient	0.530*** (0.028)	0.922*** (0.037)	0.869*** (0.041)	0.606*** (0.028)
Average number of observations per year	9,155	10,010	9,424	9,742

Panel D: Annual earnings imputations based on delisting returns and median industry EP ratio

	Actual Sample	Modified Sample – All obs. imputed	Modified Sample – Perf. del. imputed	Modified Sample – Non-perf. del. and other imputed
Intercept	0.019*** (0.002)	0.017*** (0.002)	0.018*** (0.002)	0.018*** (0.002)
Earnings	0.754*** (0.042)	0.802*** (0.038)	0.771*** (0.038)	0.788*** (0.040)
Negative Earnings Dummy	-0.047*** (0.004)	-0.015*** (0.003)	-0.021*** (0.003)	-0.040*** (0.004)
Negative Earnings Dummy *Earnings	-0.224*** (0.052)	0.104** (0.048)	0.080 (0.056)	-0.180*** (0.049)
Negative Earnings Coefficient	0.530*** (0.030)	0.905*** (0.024)	0.851*** (0.031)	0.608*** (0.030)
Average number of observations per year	9,155	9,899	9,383	9,671

Table 4: Forecast Accuracy

This table compares the accuracy of the regression-based EP model (regression (1) with forecasts based on a random walk. Accuracy is defined as the absolute value of the difference between realized earnings and forecasted earnings. The difference in accuracy is derived by subtracting the accuracy of the random walk model from the accuracy of the EP model. A negative value implies a better forecast accuracy performance of the EP model compared to the random walk model. In Panel A, we use the average imputed earnings of our three imputation procedures. In Panel B, we use earnings from the previous period to replace missing earnings observations. In Panel C, we replace missing annual earnings with an extrapolated value based on the firm's reported quarterly earnings in the associated year. In Panel D, we impute earnings using the median earnings/price ratio of firms in the same Fama/French 48 industry and CRSP delisting returns. Accuracy is calculated as the mean absolute deviation of the forecast from the actual divided by market value of equity as of the previous fiscal year-end. Earnings is defined as income before extraordinary items minus special items from Compustat. Standard errors are reported in parentheses. We cluster standard errors in two-dimensions - by firm and by year. * indicates significance at the ten percent level, ** indicates significance at the five percent level, and *** indicates significance at the one percent level.

Panel A: Average of imputation procedures

	All observations	Profit firms	Loss firms
Difference in Accuracy – Actual Sample	-0.012***	-0.002***	-0.036***
	(0.002)	(0.001)	(0.006)
	N = 104,171	N = 72,189	N = 31,982
Difference in Accuracy – Modified Sample, All obs. Imputed	-0.002***	-0.002***	-0.002
	(0.001)	(0.001)	(0.001)
	N = 112,328	N = 76,216	N = 36,112
Difference in Accuracy – Modified Sample, Perf. del. Imputed	-0.003***	-0.002***	-0.004*
	(0.001)	(0.001)	(0.002)
	N = 106,911	N = 72,543	N = 34,368
Difference in Accuracy – Modified Sample, Non-perf. del. and other imputed	-0.009***	-0.002***	-0.025***
	(0.002)	(0.001)	(0.005)
	N = 109,588	N = 75,862	N = 33,726

Panel B: Random walk imputation

	All observations	Profit firms	Loss firms
Difference in Accuracy – Actual Sample	-0.012***	-0.002***	-0.036***
	(0.002)	(0.001)	(0.006)
	N = 104,171	N = 72,189	N = 31,982
Difference in Accuracy – Modified Sample, All obs. Imputed	-0.000	-0.001*	0.001**
	(0.000)	(0.000)	(0.000)
	N = 113,947	N = 76,570	N = 37,377
Difference in Accuracy – Modified Sample, Perf. del. Imputed	-0.001***	-0.001**	-0.000
	(0.000)	(0.001)	(0.001)
	N = 107,585	N = 72,578	N = 35,007
Difference in Accuracy – Modified Sample, Non-perf. del. and other imputed	-0.007***	-0.001**	-0.019***
	(0.001)	(0.001)	(0.004)
	N = 110,533	N = 76,181	N = 34,352

Panel C: Extrapolated quarterly earnings imputation

	All observations	Profit firms	Loss firms
Difference in Accuracy – Actual Sample	-0.012*** (0.002) N = 104,171	-0.002*** (0.001) N = 72,189	-0.036*** (0.006) N = 31,982
Difference in Accuracy – Modified Sample, All obs. imputed	-0.001 (0.001) N = 113,960	-0.002*** (0.001) N = 76,570	0.001 (0.001) N = 37,390
Difference in Accuracy – Modified Sample, Perf. del. Imputed	-0.002** (0.001) N = 107,585	-0.002*** (0.001) N = 72,578	-0.001 (0.002) N = 35,007
Difference in Accuracy – Modified Sample, Non-perf. del. and other imputed	-0.008*** (0.002) N = 110,546	-0.002*** (0.001) N = 76,181	-0.023*** (0.005) N = 34,365

Panel D: Annual earnings imputations based on delisting returns and median industry EP ratio

	All observations	Profit firms	Loss firms
Difference in Accuracy – Actual Sample	-0.012*** (0.002) N = 104,171	-0.002*** (0.001) N = 72,189	-0.036*** (0.006) N = 31,982
Difference in Accuracy – Modified Sample, All obs. imputed	-0.001*** (0.001) N = 112,324	-0.002*** (0.001) N = 76,215	-0.001 (0.001) N = 36,109
Difference in Accuracy – Modified Sample, Perf. del. Imputed	-0.002*** (0.001) N = 106,911	-0.002*** (0.001) N = 72,543	-0.002 (0.002) N = 34,368
Difference in Accuracy – Modified Sample, Non-perf. del. and other imputed	-0.009*** (0.002) N = 109,584	-0.002*** (0.001) N = 75,861	-0.025*** (0.005) N = 33,723

Table 5: Summary Statistics – Samples with above and below Median Equity Market Value

The table compares the characteristics of firms that disappear from Compustat with the characteristics of firms that remain for samples formed each year based on being above or below the median market value of equity. Earnings (scaled by market value of equity obtained from Compustat at the fiscal year end) are reported for the year before the disappearance. Positive and negative earnings are split in the year prior to delisting. Returns are reported for the year of the disappearance. Delistings are considered to be performance delistings when the delisting code equals 500, or falls between 520 and 584 (inclusive), non-performance delistings when the delisting code does not equal to 500, or does not fall between 520 and 584 (inclusive), and other disappearances when there are missing future annual earnings and the delisting code is missing.

Panel A: Firms remaining in Compustat

	Above the median market value				Below the median market value			
	Profit Firms		Loss Firms		Profit firms		Loss Firms	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings in year before disappearance	0.07	0.06	-0.15	-0.06	0.09	0.07	-0.43	-0.15
Annual returns in year other firms disappear	0.14	0.10	0.04	-0.09	0.19	0.06	0.16	-0.11
Equity market value	4,911.05	736.76	1,473.03	372.76	75.88	33.17	54.60	20.79
Percentage of all obs.		41.14%		6.26%		25.17%		18.95%

Panel B: Firms disappearing for performance-related reasons

	Above the median market value				Below the median market value			
	Profit Firms		Loss Firms		Profit firms		Loss Firms	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings in year before disappearance	0.10	0.06	-1.14	-0.35	0.27	0.08	-2.34	-0.65
Annual returns in year other firms disappear	-0.43	-0.71	-0.85	-0.94	-0.18	-0.25	-0.45	-0.57
Equity market value	1,425.98	261.53	538.96	174.73	29.56	12.93	16.86	6.17
Percentage of all obs.		0.05%		0.08%		0.32%		2.30%

Panel C: Firms disappearing for non-performance-related reasons

	Above the median market value				Below the median market value			
	Profit Firms		Loss Firms		Profit firms		Loss Firms	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings in year before disappearance	0.07	0.06	-0.16	-0.06	0.09	0.07	-0.77	-0.19
Annual returns in year other firms disappear	0.35	0.27	0.30	0.16	0.50	0.36	0.43	0.24
Equity market value	2,370.72	574.83	1,414.91	391.40	87.01	46.16	59.45	28.97
Percentage of all obs.		1.78%		0.32%		1.20%		0.76%

Panel D: Firms disappearing for other reasons

	Above the median market value				Below the median market value			
	Profit Firms		Loss Firms		Profit firms		Loss Firms	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings in year before disappearance	0.08	0.06	-0.20	-0.06	0.09	0.07	-0.89	-0.30
Annual returns in year other firms disappear	0.30	0.29	-0.01	-0.17	0.29	0.27	-0.18	-0.43
Equity market value	1,323.78	310.72	1,370.99	283.78	61.61	29.75	31.14	10.52
Percentage of all obs.		0.37%		0.10%		0.48%		0.70%

Table 6: Scaled Quarterly Earnings for Remaining and Disappearing Firms – Samples with above and below Median Equity Market Value

This table provides time-series averages of the mean and median scaled earnings for the quarters before the firm disappears from the data (except for the section related to firms remaining in the sample) for samples formed each year based on being above or below the median market value of equity. Time t is defined the last quarter before the firm disappears. Positive and negative earnings are split in the year prior to delisting. Scaled earnings are earnings deflated by market value of equity (market value of equity is obtained from Compustat at the fiscal year end). Performance-related delistings are for firms with CRSP delisting codes 500 or between 520 to 584 (inclusive) and missing future earnings; firms with other delisting codes with missing future earnings are non-performance-related delistings. Other delistings are all observations that do not have a CRSP delisting code but have missing future earnings.

Panel A: Firms remaining in Compustat

	Above the median market value				Below the median market value			
	Profit Firms		Loss Firms		Profit firms		Loss Firms	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings$_t$	0.02	0.02	-0.03	-0.01	-0.02	0.01	-0.16	-0.04
Scaled Earnings$_t$ – Scaled Earnings$_{t-4}$	-0.00	-0.00	0.01	0.01	-0.04	-0.00	-0.03	0.00

Panel B: Firms disappearing for performance-related reasons

	Above the median market value				Below the median market value			
	Profit Firms		Loss Firms		Profit firms		Loss Firms	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings$_t$	-0.58	-0.50	-1.07	-0.60	-0.23	-0.08	-0.86	-0.25
Scaled Earnings$_t$ – Scaled Earnings$_{t-4}$	-0.60	-0.52	-0.93	-0.47	-0.27	-0.14	-0.71	-0.16

Panel C: Firms disappearing for non-performance-related reasons

	Above the median market value				Below the median market value			
	Profit Firms		Loss Firms		Profit firms		Loss Firms	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings$_t$	1.65	0.01	-0.04	-0.01	0.01	0.01	-0.14	-0.05
Scaled Earnings$_t$ – Scaled Earnings$_{t-4}$	1.63	-0.00	-0.01	0.00	-0.01	-0.01	-0.05	-0.01

Panel D: Firms disappearing for other reasons

	Above the median market value				Below the median market value			
	Profit Firms		Loss Firms		Profit firms		Loss Firms	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Scaled Earnings$_t$	-0.05	0.00	-0.21	-0.15	-0.08	0.01	-0.38	-0.12
Scaled Earnings$_t$ – Scaled Earnings$_{t-4}$	-0.07	-0.02	-0.20	-0.13	-0.11	-0.01	-0.25	-0.04

Table 7: Coefficient Estimates and Forecast Accuracy – Samples with above and below Median Equity Market Value

This table presents average coefficient estimates for the earnings persistence (EP) model (regression (1) when estimating the model on the actual sample and when estimating the model on a modified sample that replaces missing observation due to disappearance from Compustat for samples formed each year based on being above or below the median market value of equity. Missing observations are imputed using the average of the three imputation procedures (Panel A). The Negative Earnings Coefficient is the joint effect of the respective earnings coefficient ("Earnings") and the interaction term coefficient ("Negative Earnings Dummy *Earnings"). Panel B compares the accuracy of the regression-based EP model with forecasts based on a random walk. Accuracy is defined as the absolute value of the difference between realized earnings and forecasted earnings. The difference in accuracy is derived by subtracting the accuracy of the random walk model from the accuracy of the EP model. A negative value implies a better forecast accuracy performance of the EP model compared to the random walk model. In Panel A, standard errors (in parentheses) are Fama-MacBeth standard errors adjusted for serial correlation. In Panel B, standard errors (in parentheses) are clustered in two dimensions – by firm and by year. * indicates significance at the ten percent level, ** indicates significance at the five percent level, and *** indicates significance at the one percent level.

Panel A: Coefficient Estimates for Earnings Scaled by Market Equity

	Above median market value		Below median market value	
	Actual Sample	Modified Sample – All obs. imputed	Actual Sample	Modified Sample – All obs. imputed
Intercept	0.018*** (0.001)	0.017*** (0.001)	0.018*** (0.003)	0.016*** (0.003)
Earnings	0.814*** (0.029)	0.831*** (0.027)	0.682*** (0.050)	0.711*** (0.049)
Negative Earnings Dummy	-0.019*** (0.003)	-0.018*** (0.003)	-0.060*** (0.006)	-0.016*** (0.004)
Negative Earnings Dummy *Earnings	-0.285*** (0.059)	-0.268** (0.057)	-0.162*** (0.058)	0.183*** (0.058)
Negative Earnings Coefficient	0.528*** (0.052)	0.562*** (0.052)	0.520*** (0.030)	0.894*** (0.027)
Average number of observations per year	4,355	4,582	4,093	4,559

Panel B: Forecast Accuracy

Accuracy of EP – Accuracy of RW	-0.003*** (0.000)	-0.003*** (0.000)	-0.023*** (0.005)	-0.002** (0.001)
Number of observations	47,597	49,921	48,288	53,621

Table 8: Coefficient Estimates and Forecast Accuracy for Earnings Changes

This table presents average coefficient estimates for the earnings changes model (regression (2)) when estimating the model on the actual sample and when estimating the model on a modified sample that replaces missing observation due to disappearance from Compustat. Missing observations are imputed using the average of the three imputation procedures (Panel A). Panel B compares the accuracy of the regression-based EP model with forecasts based on a random walk. Accuracy is defined as the absolute value of the difference between realized earnings and forecasted earnings. The difference in accuracy is derived by subtracting the accuracy of the random walk model from the accuracy of the EP model. A negative value implies a better forecast accuracy performance of the EP model compared to the random walk model. In Panel A, standard errors (in parentheses) are Fama-MacBeth standard errors adjusted for serial correlation. In Panel B, standard errors (in parentheses) are clustered in two dimensions – by firm and by year. * indicates significance at the ten percent level, ** indicates significance at the five percent level, and *** indicates significance at the one percent level.

Panel A: Coefficient Estimates for Earnings Changes Scaled by Market Equity

	Actual Sample	Modified Sample – All obs. imputed	Modified Sample – Perf. del. obs. imputed	Modified Sample – Non-perf. del. and other obs. imputed
Positive Earnings Lvl., Positive Earnings Ch.	-0.185*** (0.045)	-0.162*** (0.043)	-0.180*** (0.045)	-0.163*** (0.042)
Positive Earnings Lvl., Negative Earnings Ch.	0.060*** (0.018)	0.060*** (0.020)	0.072*** (0.022)	0.049*** (0.017)
Negative Earnings Lvl., Positive Earnings Ch.	0.122*** (0.017)	0.039*** (0.011)	0.049*** (0.012)	0.094*** (0.017)
Negative Earnings Lvl., Negative Earnings Ch.	-1.140*** (0.096)	-0.616*** (0.112)	-0.724*** (0.116)	-0.976*** (0.092)
Average number of observations per year	8,375	9,063	8,583	8,855

Panel B: Forecast Accuracy

Accuracy Difference	-0.010*** (0.001)	-0.001 (0.001)	-0.001* (0.001)	-0.007*** (0.001)
Number of observations	150,328	162,587	154,088	158,827

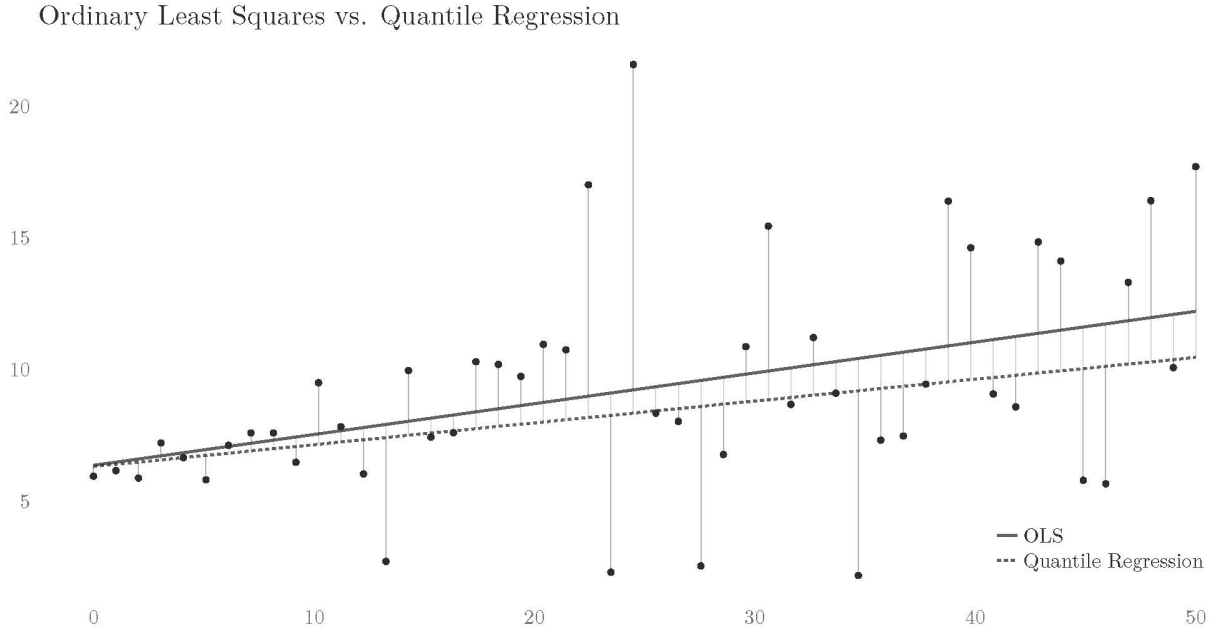
Appendix:

1. Motivation for use of Median Regressions

Before continuing with a more formal discussion of median regressions, we briefly compare them to least squares regressions to build some intuition as to why it may be the appropriate methodology for estimation earnings forecasting models. OLS regression minimizes the sum of squared deviations between an observed outcome and a within sample prediction – the prediction is modelled as a linear function of a set of observed predictors. A median regression (the special case of quantile regression), minimizes the sum of absolute deviations. Figure A1 illustrates the differences. The solid line is the least squares estimate and the dotted line is the median regression estimate.

Figure A1: Least Squares versus Median Regression estimate

This figure shows the difference between a coefficient estimated by ordinary least squares and a coefficient estimated by median regressions.



Large deviations have a stronger influence on the slope of the solid line compared to the dotted line. If both methods of regression are readily available, which one should we use? It

depends on the purpose and how the difference between the prediction and the observed outcome is evaluated, that is, which loss function is used. OLS may be better in some circumstances, whereas median regression (or other quantiles) might be more desirable in other cases.

If we are interested in making predictions and are seeking to minimize expected square loss then least squares is indeed optimal in the population. More formally, we can state that if the goal is the minimization of the following loss function,

$$\min_g E[(y - g(x))^2],$$

where y is the response, i.e. realized earnings, and x is a vector of the observed correlates, then choosing $g(x) = E[y|x]$ is the best choice. In the evaluation of earnings forecasts, however, the common practice [see Bradshaw (2011) for a summary] is to evaluate the quality of a prediction by looking at the mean absolute forecast error (MAFE). There are a number of motivations for this commonly used loss function over the mean squared error loss function. First, in terms of valuation, there is little reason to believe that the market cares more about large deviations. Freeman and Tse (1992) show that earnings response coefficients (ERCs) are non-linear. Namely, ERCs are S-shaped – concave over positive surprises and convex over negative surprises. In other words, the market seems to care proportionally less about large surprises. As such, there is little justification for a loss function that places proportionally larger weight on large surprises relative to small surprises. Further, Gu and Wu (2003) document that the market reacts positively to analyst earnings forecast errors, but reacts negatively to earnings skewness. Finally, Gu and Wu (2003) and Basu and Markov (2004) show that analysts are judged based on their mean absolute forecast error (MAFE). In summary, the standard practice, motivated by empirical evidence, is to look at averages (over firms and years) of

$$|Earnings - Predicted\ Earnings|.$$

In other words, the extant literature typically focuses on the mean (or median) absolute deviation between realized and predicted earnings. If minimizing the mean absolute forecast error is the objective, it seems natural to incorporate this into the prediction directly. Therefore, it seems prudent to use a prediction model that directly minimizes expected mean absolute deviations. More formally, we should choose g to minimize:

$$\min_g E[|y - g(x)|].$$

That is, rather than minimizing expected squared deviations, we find a $g(x)$ that minimizes expected absolute deviations. The conditional median, $med(y|x)$, minimizes expected absolute deviations. Since our goal is to estimate the linear conditional mean and quantile functions, we aim to find α's to minimize:

$$\min_{\{\alpha_0,\alpha_1,\ldots,\alpha_n\}} \sum_{i=1}^{i=N} |y_i - \alpha_0 - \alpha_1 * x_{i,1} - \alpha_2 * x_{i,2} - \cdots - \alpha_n * x_{i,n}| .$$

At first, this objective function may look somewhat clumsy because it is not differentiable everywhere, but the important property is its convexity which makes numerical solutions very efficient and computationally as simple as OLS.[28]

Since our objective is to obtain the best possible earnings forecast measured via the absolute deviation from realized earnings, it seems quite natural to make forecasts using a median regression model. Similar to OLS regression, inference for estimated parameters in the case of quantile regression is usually based on asymptotic approximation, i.e., an application of a

[28] Standard statistical software packages such as R or Stata have readily available routines for quantile regression. In R, an implementation of estimation and inference for quantile regression is available in the "rq" package. Stata provides quantile regression functionality through the "qreg" command – one could run a simple quantile regression of y on x as "qreg y x".

central limit theorem. In the case of linear quantile regression, we have the following result [Koenker (2005)]:

$$\sqrt{N}(\widehat{\alpha_\tau} - \alpha_\tau) \rightarrow^{d} N(0, V_\tau)$$

where τ is the quartile, $V_\tau = \tau(1-\tau)(E(x_i x_i' f(0|x_i)))^{-1}(E(x_i x_i'))(E(x_i x_i' f(0|x_i)))^{-1}$, and $f(e|x_i)$ is the conditional density of e_i given $x_i = x$. For the purposes of this paper, we consider the median, or when $\tau = 0.5$. In empirical work, V_τ has to be estimated. Koenker (2005) shows how to estimate this matrix directly. Alternatively, inference can also be obtained through an application of the bootstrap [Horowitz (1998)]. Overall, statistical inference for the estimated parameters of a median regression model is readily available.

2. Parameter Stability of Median Regressions and OLS

We document how median regressions perform compared to OLS regressions for cross-sectional regressions of earnings on lagged earnings. Figure A2 shows that autoregressive coefficients for positive and negative earnings are more stable when estimated by median regressions compared with coefficients estimated by OLS.

Figure A2: Coefficient Estimates for OLS and median regressions

This figure shows the difference between a parameter estimates for ordinary least squares and for median regressions. The negative earnings coefficient represents the sum of the positive earnings coefficient and the interaction term coefficient. Confidence intervals are shaded. Standard errors for median regressions are bootstrapped standard errors. Standard errors for OLS are White ("robust") standard errors (MacKinnon & White, 1985).