Investment-q sensitivity under endogenous truncation *

Ilan Cooper†
BI Norwegian Business School

Daniel Kim‡
BI Norwegian Business School

Moshe Kim§
University of Haifa

Sunday 12th February, 2023

Abstract

The empirical investment literature studying the determinants of investment relies almost exclusively on truncated samples of publicly listed firms due to the lack of data on private firms. This truncation, however, is not random because listing is a choice for many firms, whereas others cannot list due to their characteristics. This endogenous truncation causes biased estimates. We correct for the bias and find that the investment-cash flow sensitivity disappears and the relation between investment and Tobin's q is much stronger. We provide an econometric framework to correct for endogenous truncation bias that can also be applied in other economic contexts.

Keywords: *Endogenous truncation bias, Tobin's q, cash flow coefficient, bias correction*
JEL codes: C24, G30, G31, G32 .

*We have benefited from the comments of the seminar participants at the Norwegian Business School (BI Oslo). We also thank Stefan Hirth (discussant), Nir Billfeld and Adam Winegar for their helpful comments.
†ilan.cooper@bi.no
‡daniel.s.kim@bi.no
§kim@econ.haifa.ac.il

1 Introduction

The q-theory of investment predicts that the marginal value of capital, namely marginal q, is a sufficient statistic for investment behavior. Hayashi (1982) shows that under linearly homogenous technologies marginal q is identical to average q (we intermittently refer to average q as q). Given that average q is, in principle, observable, studies have regressed investment on q in order to test the q-theory of investment. However, these empirical studies typically find that while q is positively and significantly related to investment, cash flows affect investment positively even when controlling for q.

The literature offers several explanations for the sensitivity of investment to cash flows. A prominent explanation for this finding, proposed first in Fazzari, Hubbard, and Petersen (1988), is that difficulty in obtaining outside financing forces firms to utilize internal funds when undertaking real investment, leading to investment sensitivity to cash flows. An alternative strand of the literature suggests that the sensitivity of investment to cash flows is a consequence of measurement errors (see Erickson and Whited (2000, 2002, 2012); Erickson, Jiang, and Whited (2014); Gilchrist and Himmelberg (1995); Abel (2018); Chalak and Kim (2020)). These papers argue that measurement errors drive a spurious positive cash flow coefficient in investment regressions. Another strand of the literature (Gomes, 2001; Cooper and Ejarque, 2003; Alti, 2003; Abel and Eberly, 2011) develops theoretical models in which marginal q is not equal to average q and analyzes the investment-cash flow sensitivity within these theoretical models.

In this paper we offer a new perspective to examining the empirical investment equations without resorting to finance constraints or measurement error, and within the classical framework of Hayashi (1982). Naturally, due to lack of data for unlisted companies the studies testing the q-theory of investment utilize truncated samples of only publicly listed firms. Ignoring privately held firms has several ramifications. First, unlisted firms are a very important and significant part of the economy and them not being included in the sample may have severe and significant ramification on results particularly if the non inclusion is endogenous. Asker, Farre-Mensa, and Ljungqvist (2015) report that in 2010 listed firms constitute only 0.06% of all firms —indeed a very small share of firms in the economy. In addition they report that in 2010 privately held firms accounted for 53% of aggregate investment, 69% of private sector employment, 59% of sales, and 49% of aggregate pre-tax profits. They also report that 86% of firms with 500 employees or more were privately held. Second, the truncation caused by not observing unlisted firms is not random. For many firms listing and not-delisting are choice variables. That is, some unlisted firms choose to become public while some public firms decide to go private. Moreover, for a listed firm staying listed is also a decision made at each point in time. Additionally, some firms lack the choice and become delisted due to bankruptcy or they cannot list to begin with. In either case, studies show that the listing likelihood depends on firm characteristics (see Pagano, Panetta, and Zingales (1998); Chemmanur and Fulghieri (1999); Chemmanur, He, and Nandy (2010); Mehran and Peristiani (2010); Djama, Martinez, and Serve (2012); Doidge, Karolyi, and Stulz (2017)). Thus, the Compustat sample of firms, largely utilized by empirical researchers, is not a random sample and is severely truncated. Consequently, the resulting endogenous truncation renders estimates obtained from the truncated sample potentially biased. In this paper we analyze the truncation bias and offer a solution to correct for that bias in the context of the q-theory of investment.

Our contribution is thus, twofold. First, we demonstrate that endogenous truncation leads to biased coefficient estimators. We subsequently derive an endogenous truncation bias correction. Our correction can be applied in other studies that employ endogenously truncated samples. Second, we apply the endogenous truncation bias correction in the context of the q-theory of investment that requires a panel structure, and find strong support for the theory's predictions in a long sample period (1971-2018).

Econometrically, truncation bias arises when the disturbance in the investment regression equation correlates with the disturbance in the listing outcome estimation. These two estimations are seemingly unrelated. However, such correlation of the disturbance terms can arise, particularly because the dependent variables in these two estimations belong to the same firms. One potential economic explanations for this correlation is a manager who desires to grow fast, possibly due to overconfidence (Malmendier and Tate, 2005, 2008; Malmendier, Tate, and Yan, 2011). One might argue that both disturbance terms reflect managerial overconfidence. Consider, for example, a private firm with an overconfident manager wishing to grow. That is, the manager overestimates the firm's expected future cash flows and capital needs. Based on the prior literature the decision to list is a function of fundamentals (such as size). That is, there is a threshold level of fundamentals that once crossed the firm will go public. For example, Doidge, Karolyi, and Stulz (2017) present a model in which there is a level of fundamentals (size in the model) that once given the manager's overconfidence, at some point in time the firm could decide to list in spite of a level of fundamentals that is below the optimal threshold. Hence, this decision reflects a positive disturbance term in the listing function.

The manager's overconfidence also implies that the firm will invest more than its q-theory's predicted investment. Assuming that the firm's investment becomes more consistent with the q-theory as the firm ages, the firm's overinvestment in the post-IPO period will be reflected as a positive disturbance in the investment regression equation. Thus, a positive correlation between the disturbance term arises if a sufficiently large number of IPO firms have managers who wish to grow fast. Interestingly, we find that indeed the investment disturbance term declines as the firm ages. Relatedly, Brau and Fawcett (2006) conduct a survey among CFOs and find that the primary motivation for going public is to facilitate acquisitions. Malmendier and Tate (2008) find that the odds of making an acquisition are 65% higher if the CEO is classified as overconfident. Jain and Kini (1994) find a significant decline in operating performance subsequent to the initial public offering (IPO), consistent with overinvestment during the post-IPO period. Brau, Couch, and Sutton (2012) analyze a large sample of IPOs and find that IPOs that acquire within a year of going public significantly underperform in terms of stock returns for 1- through 5-year holding periods following the first year, also consistent with the acquisition constituting overinvestment. A similar positive correlation between the residuals can arise in the case that managers possess superior information about the prospect of the firm. As such, the non-randomly truncated sample could yield biased slope coefficients.

In order to correct for endogenous truncation bias, one might first think of the seminal Heckman procedure (Heckman, 1979). However, the Heckman procedure cannot help to address the endogenous truncation bias because empiricists do not observe data on the unlisted truncated firms. Thus, researchers lack data to estimate a probit regression in implementing the Heckman procedure. Instead, similar to Robinson (1988), we remove the bias term by cancelling out the conditional expectation term. We embed our debiasing methodology in Wansbeek and Kapteyn (1989) to address unbalanced panel data. Naturally, the offered econometric framework is not limited to testing the q-theory and can be applied to empirical research when only a truncated data set is available (e.g. Compustat).[1]

Key features of our framework that are important to emphasize are as follows. First, the q-theory applies to both public and private firms. That is, q is a sufficient statistic for investment for both types of firms. Second, the endogenous truncation bias arises even though public and private firms are identical in terms of their true investment-q sensitivity.

In our empirical tests we employ the Tobin's q measure that Peters and Taylor (2017) made available over the sample period 1971-2018.[2] Our findings can be summarized as follows. First, when not correcting for the truncation bias in investment regressions the cash

[1] In fact, our econometric framework shares conceptual similarity to the one recently proposed in the computer science literature (e.g. Billfeld and Kim (2016, 2019)). What sets apart ours from Billfeld and Kim's is two features that are unique in the economics/finance application: panel feature and specifications driven by economic theory.

[2] Peters and Taylor (2017) show that their Tobin's q that accounts for intangibles is a superior proxy for both tangibles and intangibles investment opportunities. They made the data available via WRDS.

flow coefficient is large and highly statistically significant. However, when correcting for the truncation bias the cash flow coefficient ceases being significant both economically and statistically, whereas the q coefficient is approximately four times larger. Second, when applying both the Erickson, Jiang, and Whited (2014) correction for measurement error in q and the truncation bias correction, the q coefficient more than doubles (relative to either correcting only for the measurement error or only for the truncation bias) and is highly statistically significant, whereas the cash flow coefficient is small and statistically insignificant. This finding suggests that correcting for both measurement error and truncation bias has important implications when testing the q-theory.

The disturbance term in the listing function is not observable. Therefore a direct computation of the correlation among this term and the disturbance term in the investment equation is not possible. Thus, one cannot definitely conclude that an endogenous truncation bias exists in our sample. To address this concern we simulate the data under no correlation among the disturbance terms. Consistent with our expectations, the lack of correlation implies that the estimated coefficients on Q and cash flows are very similar when applying a bias correction and when not applying such correction. The fact that in the data there is a large impact for the correction bias suggests that the endogenous truncation bias indeed exists in the data.

We also examine the evolution of the truncation bias over time. Doidge, Karolyi, and Stulz (2017) document a clear pattern in the number of publicly listed firms in the U.S.. The number of public firms in the U.S. rose steadily from 1975 and has reached a peak in 1996. Subsequently to 1996 there has been a sharp decline in the number of publicly listed firms.[3] This decline, relative to other countries is termed by Doidge, Karolyi, and Stulz (2017) the listing gap. In order to examine the severity of the truncation bias over time we first show that investment-cash flow sensitivities moderately decline over time when not correcting for the truncation bias. This finding is consistent with Chen and Chen (2012) who show that

[3]The propensity to list follows a rather similar pattern.

the investment-cash flow sensitivity has diminished recently.

Subsequently, we present two important results. First, the truncation bias-corrected sensitivities are smaller in magnitude than uncorrected sensitivities for each of the years in our entire sample periods. Second, most interestingly, we estimate the dynamics of the truncation bias-corrected sensitivities of investment to cash flows. We define the severity of the truncation bias as the difference between the uncorrected and truncation bias-corrected cash flow sensitivities. Interestingly, we find that the severity of the truncation bias varies over time in a very similar fashion to the (inverse of) number of listed firms in the U.S.. That is, from 1975 to the mid-1990s the truncation bias severity diminishes gradually, then it reaches a trough in the mid-1990s and subsequently it rises again. This pattern mirrors the (reciprocal) pattern in the number of firms that are publicly listed in the U.S.; as the number of listed firms rises, the severity of the truncation bias decreases, reflecting the fact that there are fewer private firms. However, as the listing gap widens in the period following the mid 1990s, the severity of the truncation bias rises again.

While most papers studying the determinants of investment employ data on publicly listed firms, a notable exception is Asker, Farre-Mensa, and Ljungqvist (2015) who study the investment behavior of private firms in the short sample period of 2001 through 2011. They find that private firms' investment is more responsive to changes in investment opportunities, and conclude that short-termist pressures distort the investment decisions of public firms. Our paper differs from Asker, Farre-Mensa, and Ljungqvist (2015) along several important dimensions. First, as mentioned above, we show that endogenous truncation bias arises even if the true investment sensitivity to q is the same for both listed and delisted firms. Next, unlike us, they do not examine investment-cash flow sensitivities. Third, our sample from 1971 to 2018 spans a substantially longer period than the ten-year sample of Asker, Farre-Mensa, and Ljungqvist. Fourth, Asker, Farre-Mensa, and Ljungqvist focus on *tangible* investment whereas we use measures that capture both *tangible* and *intangible* investment.

Lastly, we emphasize the truncation bias that arises in many empirical studies in finance and economics and provide a general econometric framework to correct for that bias that can be applied in many other studies.

Our paper contributes to the recent strand of literature expressing renewed interest in testing the q-theory of investment and the investment-cash flow sensitivity. Chen and Chen (2012) show that investment-cash flow sensitivities among manufacturing firms have declined over time and especially since the mid-1990s. We find that the significance of cash flows among *all* firms (both economically and statistically) is induced by endogenous truncation bias on the 1971-2018 sample. We also find that the investment-cash flow sensitivities have declined over time even after we correct for the endogenous truncation bias.

Peters and Taylor (2017) show that the classical q-theory works better for firms and years with more intangible capital. They propose a new measure of q that accounts for intangible capital and show that it explains better both physical investment as well as intangible investment. We employ their measure as our benchmark measure as it is superior to the traditional measure of q in explaining both types of investment as well as total investment. Moreover, we also employ a traditional measure of q (asset-deflated definition as termed in Erickson and Whited (2012)) and show that our results are robust to alternative definitions. Andrei, Mann, and Moyen (2019) show that since the middle of the 1990s the relation between investment and q has become remarkably tight (however they do not examine the dynamics of the investment-cash flow sensitivity). They propose a model with innovation and learning to account for their findings. We find that when correcting for the endogenous truncation bias, the relation of investment to q is strong throughout the sample period of 1971 to 2018 and not only in recent years. Importantly, our contribution extends beyond testing the q-theory as it provides a framework for correcting for endogenous truncation bias that potentially arises in many empirical frameworks in economics.

There is ample evidence regarding the determinants of firms' decisions to become publicly listed. Pagano, Panetta, and Zingales (1998) find that the likelihood of an IPO is increasing in the company's size and the industry's market-to-book ratio. Chemmanur and Fulghieri (1999) theoretically study the determinants of the going public decision. Chemmanur, He, and Nandy (2010) and Campbell, Hilscher, and Szilagyi (2008) show that firm characteristics are determinants of its decision to go public and stay listed, respectively. Given this literature we employ the following characteristics for the determinants to publicly list: the market to book ratio, total factor productivity, size, sales growth, net income, excess equity return, leverage, stock return equity, cash, and stock price.

The rest of the paper is organized as follows. Section 2 describes a heuristic argument for the endogenous truncation bias. It details an example and provides the intuition for the endogenous truncation bias. Section 3 presents the econometric framework that includes the endogenous truncation bias correction. In Section 4 we conduct a simulation that illustrates that standard OLS regression suffers from endogenous bias, and how the econometric framework can alleviate this problem. Section 5 describes the data and variable construction. The empirical results appear in Section 6. Section 7 concludes.

2 Heuristic Argument

In order to provide intuition, in this section we present an heuristic example for the bias in investment-q sensitivities arising from endogenous truncation. We first explain the original sample. Then, we explain how firms' endogenous listing decisions can yield biased estimates.

We assume that the classical q-theory holds and accordingly construct pairs of firm characteristics: investment and q. More specifically, firms' investment can be explained by firms' q and the innovation term, ϵ_1, that is orthogonal to q.

$$\text{Investment} = \alpha + \beta \cdot q + \epsilon_1 \tag{1}$$

Here, β is the true investment-q sensitivity. Andrei, Mann, and Moyen (2019) show that both innovations and learning endogenously make q more volatile. Accordingly, we assume that q is stochastic and follows the trinomial distribution.

$$q = \begin{cases} q_L & \text{with probability } 1/3 \\ q_M & \text{with probability } 1/3 \\ q_H & \text{with probability } 1/3 \end{cases} \tag{2}$$

where $q_H > q_M > q_L$. Similarly, we assume that the innovation term ϵ_1 follows the trinomial distribution:

$$\epsilon_1 = \begin{cases} -1 & \text{with probability } 1/3 \\ 0 & \text{with probability } 1/3 \\ 1 & \text{with probability } 1/3 \end{cases} \tag{3}$$

We assume q and ϵ_1 to be independent of each other. Accordingly, there are nine different types of (investment, q) pairs and each type is equally likely. We number each type and graphically illustrate their investment and q in Figure 1 where for simplicity we assume that $\alpha = 0$. The dashed line is the best fitted line. As shown, the line drawn based on the non-truncated sample has slope that is equal to the true investment-q sensitivity: β.

Now, we illustrate the endogenous truncation bias. To this end, we first need to construct the endogenously truncated sample. We assume that the firms are listed only when

$$X + \epsilon_2 \geq 0$$

where X is firm characteristic (e.g. firm size) and ϵ_2 is an innovation term that is orthogonal to X. In other words, firms become truncated out of the sample if their characteristics do not satisfy the above condition.

We first look at a case where firm characteristic X is positively correlated with q and ϵ_1 is positively correlated with ϵ_2. One might think that the decision to become listed predates the decision of how much to invest, and therefore the correlation between investments' residual (ϵ_1) and listing's residual (ϵ_2) should be zero. However, non-zero correlation between ϵ_1 and ϵ_2 can potentially arise because firms contemporaneously decide on how much to invest and whether to stay listed (or become listed) *every* period.

To illustrate these positive correlations, we construct X and ϵ_2's distributions as follows:

$$X = \begin{cases} -1 & \text{when } q = q_L \\ 0 & \text{when } q = q_M \\ 1 & \text{when } q = q_H \end{cases}$$

$$\epsilon_2 = \begin{cases} -1 & \text{when } \epsilon_1 = -1 \\ 0 & \text{when } \epsilon_1 = 0 \\ 1 & \text{when } \epsilon_1 = 1 \end{cases}$$

We now check which type of firm gets truncated. The intuition here is as follows. When q is large, the firm is likely to become listed because its characteristics X are sufficiently high to satisfy the listing requirement. However, when q is small, firms need to have sufficiently large innovation terms, ϵ_2, in order to satisfy the listing requirement. For instance, let us discuss firm 1 and 2 in Figure 1. Because Firm 1's q is q_L and $\epsilon_1 = 1$, Firm 1 has $X = -1$ and $\epsilon_2 = 1$. This satisfies the listing requirement and thus firm 1 stays in the sample and does *not* get truncated. Now, we discuss firm 2. Because firm 2's q is q_L and $\epsilon_1 = 0$, firm 2 has $X = -1$ and $\epsilon_2 = 0$. This does not satisfy the listing requirement and thus firm 2 gets truncated.

We apply similar logic to determine whether other firms are truncated or not. In sum, firm 1, 4, 5, 7, 8 and 9 are listed as shown in the left panel of Figure 2. The solid red line is the best fitted line of the truncated sample. As shown in Panel A, the solid red line is flatter than the dashed line. This illustrates that the linear regression based on the truncated sample is *downward* biased. In Section 5, we show that the positive correlation between two disturbance terms is indeed positive in the empirical counterpart. We further show that linear regression yields *downward* biased q coefficient. For completeness, in the rest of this section, we consider other cases.

Let us now consider an alternative case where the firm characteristic X is positively correlated with q *and* ϵ_1 is negatively correlated with ϵ_2 as follows

$$X = \begin{cases} -1 & \text{when } q = q_L \\ 0 & \text{when } q = q_M \\ 1 & \text{when } q = q_H \end{cases}$$

$$\epsilon_2 = \begin{cases} -1 & \text{when } \epsilon_1 = 1 \\ 0 & \text{when } \epsilon_1 = 0 \\ 1 & \text{when } \epsilon_1 = -1 \end{cases}$$

We apply a similar logic to determine which firms are truncated. In sum, firm 3, 5, 6, 7, 8 and 9 are not truncated as graphically illustrated in Panel B of Figure 2. The fitted line of the truncated sample is steeper than the fitted line of the untruncated sample. This illustrates that the linear regression based on the truncated sample is *upward* biased.

Section 4.2 illustrates the above two cases. Moreover, the same section discusses how the biased estimate in investment-q sensitivity yields a biased estimate in investment-cash flow sensitivity. There are two remaining cases: 1) X is negatively correlated with q *and* ϵ_1 is positively correlated with ϵ_2 and 2) X is negatively correlated with q *and* ϵ_1 is negatively correlated with ϵ_2. Similar to the logic above,linear regressions based on the truncated sample still yields biased estimates.

3 Empirical Framework: Econometrics

In this section, we provide the econometrics that underlie our empirical strategy. We formally write the investment equation and the listing/not-delisting function. We next list the necessary identification assumptions. Then, we formulate the endogenous truncation bias that arises if one focuses on the endogenously truncated sample. Lastly, we show how we correct for the bias. We end the section with a related discussion.

3.1 Setup

We denote a firm i's outcome variable (e.g. the firm's investment) at time t as Y_{it}^* where:

$$Y_{it}^* = \alpha_i + \delta_t + X_{it} \cdot \beta' + \epsilon_{1,it} \tag{4}$$

where i ($i = 1, ..., H$) denotes firms and t ($t = 1, ..., T$) denotes years. X_{it} is a vector of firm characteristics, α_i is firm fixed effect, and δ_t is time fixed effect. Let N_t be the number of observed firms in year t. Let $N = \sum_t N_t$.

Our coefficient of main interest is β. We first specify the exogeneity assumption as follows:

Assumption 1. *$\forall t \in \{1, ..., T\}$ and $\forall i \in \{1, .., N\}$, $(\alpha_i, \delta_t, X_{it})$ are uncorrelated with $\epsilon_{1,it}$.*

In the absence of an endogenously truncated sample, **Assumption** 1 would have provided the sufficient conditions for the OLS regression to yield a consistent estimate for β. In the rest of this section, we first show how the OLS regression estimate suffers from the bias when the sample is endogenously truncated. Then, we show how we correct for that bias.

Let $\mathbb{1}\{\cdot\}$ be an indicator function. We denote the firm's truncation outcome as D^*_{it} where:

$$D^*_{it} = \mathbb{1}\{g(Z_{it}) + \epsilon_{2,it} \geq 0\} \tag{5}$$

Here, D^*_{it} is a binary variable and is determined based on continuous function (g) of the firm characteristics, Z_{it}. Notably, our econometric framework is valid for any continuous functional forms for g. This functional form captures both firms' listing *and* not-delisting decisions. Similar to **Assumption** 1, we specify the exogeneity assumption as follows:

Assumption 2. *$\forall t \in \{1, ..., T\}$ and $\forall i \in \{1, .., N\}$, Z_{it} are uncorrleated with $\epsilon_{2,it}$.*

Lastly, we specify exogeneity assumptions that span over Equations (4) and (5). The disturbance term in the investment equation is uncorrelated with covariates for listing/not-delisting decisions. Moreover, disturbance term in listing/not-delisting decision is uncorrelated with covariates for investment equation.

Assumption 3. *$\forall t \in \{1, ..., T\}$ and $\forall i \in \{1, .., N\}$, $(\alpha_i, \delta_t, X_{it})$ are uncorrelated with $\epsilon_{2,it}$ and Z_{it} are uncorrleated with $\epsilon_{1,it}$.*

For clarification, the above three assumptions do *not* necessarily imply that the disturbance terms $(\epsilon_{1,it}, \epsilon_{2,it})$ are jointly uncorrelated. In fact, non-zero correlation between disturbance term, $\epsilon_{1,it}$, and the other disturbance terms, $\epsilon_{2,it}$, is necessary to have non-zero endogenous truncation bias. We provide more discussion below.

3.2 Endogenous Truncation Bias

The observed investment outcome is:

$$Y_{it} = \begin{cases} Y^*_{it} & \text{if } D^*_{it} = 1 \\ \text{Unobserved} & \text{Otherwise} \end{cases}$$

Closely following Ichimura (1993), conditional on being listed and not-delisted at least up to period t, we can rewrite Equation (4) as:

$$Y_{it} = \alpha_i + \delta_t + X_{it}\beta + U_{it} + \mathbb{E}\left[\epsilon_{1it}|\alpha_i, \delta_t, X_{it}, D_{it}^* = 1\right] \tag{6}$$

where U_{it} is mean-zero disturbance term and the last term, $\mathbb{E}[\cdot]$, is the truncation bias term. The truncation bias term can be re-written as :

$$\mathbb{E}[\cdot] = \int_{u=-\infty}^{g(Z_{it})} \mathbb{E}\left[\epsilon_{1,it}|\epsilon_{2,it} = u\right] f_{\epsilon_{2,it}}(u)du$$

Because the truncation bias term is a function in terms of Z_{it}, we can rewrite Equation (6) as

$$Y_{it} = \alpha_i + \delta_t + X_{it}\beta + M(Z_{it}) + U_{it} \tag{7}$$

Here, $M(Z_{it})$ is endogenous truncation bias that we try to correct for. And $M(Z_{it})$ is non-zero only when $(\epsilon_{1,it}, \epsilon_{2,it})$ are jointly correlated.

3.3 Endogenous Truncation Bias Correction

Now, we describe how we correct for the bias. We first stack up Equation (6) and premultiply both sides by a matrix $\boldsymbol{P}$ (See Section A.1 for more details):

$$\boldsymbol{PY} = \boldsymbol{PX}\beta + \boldsymbol{PM} + \boldsymbol{PU} \tag{8}$$

This operation helps us remove both firm and year fixed effects. Here, $\boldsymbol{PM}$ is a demeaned version of the endogenous truncation bias.

Next, we start to remove the endogenous truncation bias. We first apply the conditional expectation operator on both sides:

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{PY}|\boldsymbol{Z}\right] &= \mathbb{E}\left[\boldsymbol{PX}|\boldsymbol{Z}\right]\beta + \mathbb{E}\left[\boldsymbol{PM}|\boldsymbol{Z}\right] + \mathbb{E}\left[\boldsymbol{PU}|\boldsymbol{Z}\right] \\ &= \mathbb{E}\left[\boldsymbol{PX}|\boldsymbol{Z}\right]\beta + \boldsymbol{PM} + \mathbb{E}\left[\boldsymbol{PU}|\boldsymbol{Z}\right] \end{aligned} \tag{9}$$

where the last equality holds because $\boldsymbol{M}$ is a function of $\boldsymbol{Z}$. Next, we subtract Equation (9) from Equation (8) to get (Robinson, 1988):

$$\overbrace{\boldsymbol{PY} - \mathbb{E}\left[\boldsymbol{PY}|\boldsymbol{Z}\right]}^{\boldsymbol{\Delta PY}} = \overbrace{\left(\boldsymbol{PX} - \mathbb{E}\left[\boldsymbol{PX}|\boldsymbol{Z}\right]\right)}^{\boldsymbol{\Delta PX}}\beta + \overbrace{\boldsymbol{PU} - \mathbb{E}\left[\boldsymbol{PU}|\boldsymbol{Z}\right]}^{\boldsymbol{\Delta PU}}$$

Thus, we get the regression equation:

$$\boldsymbol{\Delta PY} = \boldsymbol{\Delta PX}\boldsymbol{\beta} + \boldsymbol{\Delta PU} \tag{10}$$

We use Equation (10) to estimate $\boldsymbol{\beta}$ as follows:

$$\hat{\boldsymbol{\beta}} = \left((\boldsymbol{\Delta PX})'(\boldsymbol{\Delta PX})\right)^{-1}(\boldsymbol{\Delta PX})'\boldsymbol{\Delta PY} \tag{11}$$

As shown, the endogenous truncation term, $\boldsymbol{PM}$, does not appear in Equation (11). We would like to point out that this argument holds for any continuous functional form for $M(Z_{it})$. That is, our econometric framework works for any functional forms of $g(\cdot)$ that capture both firms' listing and not-delisting outcomes. In the present model $\boldsymbol{M}$ is canceled out and thus, β can be estimated independently of $\boldsymbol{M}$. Section A.2 addresses inference results. Please see Section A.3 for cases where $\boldsymbol{X}$ and $\boldsymbol{Z}$ consist of common covariates.

3.4 Discussion

One might wonder how this is different from the sample selection procedure that was proposed by Heckman (1979). The Heckman method can be applied when data on both truncated firms and non-truncated firms is observable, allowing probit regression estimation. Unfortunately, researchers do not easily observe data on *truncated* firms. For a similar reason, the method proposed by Malikov, Kumbhakar, and Sun (2016) is not applicable.

One might notice some similarities between our econometric framework and IV framework where truncation controls Z_{it} could be thought of as instruments. It is true that truncation controls are uncorrelated with the disturbance term ($\epsilon_{1,it}$) in the investment equation and thus share similarities to a valid instrument in IV. However, truncation decisions (Equation (5)) are functions of not only truncation controls but also of disturbance terms ($\epsilon_{2,it}$) and the disturbance term is correlated with the investment equation's disturbance term. Thus, it requires an entirely different approach to correct for the bias.

4 Monte Carlo Simulation

In this section, we use simulation to illustrate the econometric framework. In Section 4.1, we describe a data generating process that gives rise to endogenous truncation. Section 4.2 uses the generated data to illustrate that standard OLS suffers from endogenous truncation bias, the exact problem that our econometric framework attempts to address. Then, we illustrate that our econometric framework helps to alleviate this bias.

4.1 Data Generating Process

We specify the data generating process to mimic the key features, but not all, of the empirical counterpart.

4.1.1 Classical q-theory

An empirical setup that tests the classical q-theory is:

$$\frac{I_{it}}{K_{it-1}} = \alpha_i^I + \delta_t^I + \beta_q \cdot q_{i,t-1} + \epsilon_{i,t}^I \tag{12}$$

where $I_{i,t}$ is firm i's investment and $q_{i,t-1}$ is Tobin's q. Moreover, α_i^I is firm fixed effect, δ_t^I is year fixed effect, and $\epsilon_{i,t}^I$ is a disturbance term. Here, the linear specification is the exact prediction of the classical q-theory. This equation is similar to Equation (4) discussed in Section 3. Moreover, the classical q-theory implies that firm fixed effect, year fixed effect, and q are uncorrelated with the disturbance term ($\epsilon_{i,t}^I$), and thus helps to satisfy **Assumption** 1.

In order to empirically validate the classical q-theory, empiricists regress investment rate on q and CF (cash flow). The q-theory predicts that the q regression coefficient should be statistically significant whereas CF regression coefficients should not be statistically significant. In all the simulations, we set $\beta_q = 0.073$. We choose this particular number to exactly match the empirical estimate once the endogenous truncation bias is corrected for (see Table 5). We elaborate on the choice of the distribution parameters of the variables in Section 4.1.3.

4.1.2 Endogenous Truncation

Now we formulate firms' listing decision which gives rise to the endogenous truncation under certain conditions, as follows:

$$D_{i,t}^* = \mathbb{1}\left\{\alpha_x + \beta_x \cdot X_{i,t} + \epsilon_{i,t}^X \geq 0\right\} \tag{13}$$

Here, $X_{i,t}$ represents explanatory variables that determine whether the firm i is truncated or not. $\epsilon_{i,t}^X$ is innovation term that is orthogonal to $X_{i,t}$, helping to satisfy **Assumption** 2.

This functional form captures Equation (5).

We would like to re-emphasize that our econometric framework is general enough to accommodate any functional form (e.g. linear or non-linear) of the listing decision function, as well as any number of covariates. In fact, as shown in a few papers such as Chemmanur, He, and Nandy (2010), the relevant covariates are market-to-book ratios, size, sales growth, and TFP. Next, even after firms become listed, firms have to decide whether to stay listed every period. For many reasons, firms often delist. For instance, firms can go into strategic bankruptcy, and this can lead to delisting. Alternatively, firms can choose to become private (Mehran and Peristiani, 2010; Djama, Martinez, and Serve, 2012).

For an illustrative purpose, we define the joint function of listing and not-delisting to be linear functions of firm characteristic, $X_{i,t}$. This rather simple-looking specification is well motivated by the practice in the literature. Researchers (e.g. Chemmanur, He, and Nandy (2010)) model firms' listing probability as logit specification, e.g. $\frac{1}{1+\exp\left(\beta_s Size+\epsilon^s\right)}$. Subsequently, researchers use data-driven threshold $\overline{p}$ to determine which firms are listed $\left(\frac{1}{1+\exp\left(\beta_s Size+\epsilon^s\right)} \geq \overline{p}\right)$ and which firms are not listed $\left(\frac{1}{1+\exp\left(\beta_s Size+\epsilon^s\right)} < \overline{p}\right)$. We can rewrite this in a form of Equation (13) as follows:

$$D^* = \mathbb{1}\left\{\ln\left(\frac{1-\overline{p}}{\overline{p}}\right) - \beta_s Size - \epsilon^s\right\} \tag{14}$$

A similar argument can be applied to firms' delisting decision (Campbell, Hilscher, and Szilagyi, 2008). Nonetheless, our econometric framework is general enough to accommodate any function form (e.g. linear or non-linear) of any number of covariates. We label the subset of the sample as "non-truncated" when Equation (14) is true. In order to mimic the empirical fact that researchers only observe the firm-years when firms choose to list and stay listed, we focus our analysis only on the non-truncated simulated data.

In our simulations, we set $\beta_x = 0.5$. Positive β_x implies that firms are more likely to list when $X_{i,t}$ is large. This number was chosen so that simulated data feature are close to the empirical counterparts in many observable aspects including the biased and unbiased estimate of β_q. More specifically, as shown in Table 1 and 5, biased estimate of β_q is 0.017 in both simulated sample and empirical counterpart. Similarly, unbiased estimate of β_q is 0.073 in both simulated sample and empirical counterpart.

4.1.3 Panel Construction and Variables

In order to mimic the real data as closely as possible, we do the following. First, we keep the simulated data in panel structure. For each firm's time-series, we randomly remove data points. This yields an *unbalanced* panel data. It is worth pointing out that this operation would create random truncation. Contrary to the endogenous truncation bias, this random truncation would *not* generate bias because this truncation is random. Section 4.2 discusses the exact properties of the simulated sample, including the sample size.

We add firm fixed effects and year fixed effects. We construct Tobin's q q, cash flow CF, and other firm characteristics X as follows:

$$
\begin{aligned}
q_{i,t} &= \alpha_i^q + \delta_t^q + \eta_{i,t}^q \\
CF_{i,t} &= \alpha_i^c + \delta_t^c + \eta_{i,t}^c \\
X_{i,t} &= \alpha_i^x + \delta_t^x + \eta_{i,t}^x
\end{aligned}
$$

where α_i^q, α_i^c, and α_i^x are firm fixed effect and δ_t^q, δ_t^c, and δ_t^x are year fixed effect.

We simulate the random numbers as follows. To highlight the main driving force behind the endogenous truncation bias, we allow only three pairs of numbers to be correlated. First, we allow $\eta_{i,t}^q$ to be correlated with $\eta_{i,t}^c$ at correlation 0.55. We choose 0.55 to exactly match the correlation between q and cash flow in the empirical data.

Second, we allow $\eta_{i,t}^q$ to be correlated with $\eta_{i,t}^x$ at correlation 0.75. This allows q and firm characteristic X to be positively correlated. This correlation is not particularly easy to match the empirical counterpart because the simulation assumes, for tractability, that there is one covariate in X whereas there are eleven covariates in the empirical counterpart to X (see Section 5.1 for the comprehensive list). Nonetheless, this is a fairly realistic assumption because eight out of eleven covariates are positively correlated with q. Moreover, this correlation, along with other numbers, helps us to match the magnitude of the truncation bias, which is the key contribution of our paper. We remind the readers that our exercise here is purely for an illustrative purpose and we leave the precise estimation of the underlying parameters to the future study.

Moreover, we allow the investment innovation term ($\epsilon^I_{i,t}$) and truncation innovation term ($\epsilon^X_{i,t}$) to be correlated at non-zero values. We want to point out that the sign of the correlation between these two terms determines the direction of the bias. For illustrative purposes, we run two sets of simulations: when the correlation is set at 0.4 and when the correlation is set at -0.4. Lastly, $var(\eta^I_{i,t}) = var(\epsilon^X_{i,t}) = 0.08$, $var(\eta^q_{i,t}) = 0.96$, $var(\eta^c_{i,t}) = 0.01$ and $var(\eta^x_{i,t}) = 0.51$. These numbers help us to match the empirical counterparts in many observable aspects such as the biased estimate of β_q.

4.2 Simulation Results

We run 1,000 simulations. For each simulation, as discussed in Section 4.1, we construct unbalanced and truncated panel data that include only listed firms. Each row in Table 1 summarizes each simulation result when the investement innovation terms and truncation innovation terms are positively correlated. The very last row shows the average over 1,000 simulations.

The first column shows the number of firm-years in the untruncated sample. The second column shows the number of firm-years in the truncated sample, and the third column presents the number of firms in the corresponding truncated sample. As shown, on average, the number of firm-years in the untruncated and randomly unbalanced panel is 540,100. The endogenous truncation reduces the sample size to 270,049 firm-years. In other words, the panel sample size is reduced by 50% because either firms did not list or firms de-listed. In this truncated sample, there are roughly 5000 firms and each firm has 54 years worth of data on average.

The next four columns correspond to the panel OLS regression results without correcting for endogenous truncation bias. Standard errors are clustered at the firm and year levels. As a reminder, as discussed in Section 4.1, the coefficient for q should be 0.073. However, the average of the coefficients for q is 0.017 (0.001). The OLS estimate is clearly biased downward. This biased estimate is close to the empirical counterpart that is documented in Table 5. The *downward* bias is driven by the *positive* correlation between investment innovation terms and truncation innovation terms. Similarly, the OLS estimate of q is *upward* biased when investment innovation terms and truncation innovation terms are *negatively* correlated (see Table 2 for the relevant results).

Now we discuss the CF coefficient. To that end, we first denote the biased estimate of q as $\hat{\beta}_q$. Thus, $(\beta_q - \hat{\beta}_q)$ is positive where β_q is true investment-q sensitivity. We then rewrite Equation (12) as

$$\begin{aligned}\frac{I_{it}}{K_{it-1}} &= \alpha_i^I + \delta_t^I + \beta_q \cdot q_{i,t-1} + \epsilon_{i,t}^I \\ &= \hat{\beta}_q \cdot q_{i,t-1} + \underbrace{((\beta_q - \hat{\beta}_q) \cdot q_{i,t-1} + \alpha_i^I + \delta_t^I + \epsilon_{i,t}^I)}_{\text{Biased residual}}\end{aligned}$$

Because the biased residual is positively correlated with $q_{i,t-1}$ and $q_{i,t-1}$ is positively correlated with $CF_{i,t}$, the regression coefficient for CF is positive. In other words, the upward-biased estimate of CF is driven by downward-biased estimate of q *and* positive correlation between q and CF. As a reminder, as discussed in Section 4.1, the true coefficient for CF is 0. Thus, the endogenous truncation bias leads to an upward bias in the linear regression coefficient for CF.

This is illustrated in the table. The coefficient for CF is statistically significant at 0.302

(with a standard error of 0.007), which is clearly biased upward. This statistical significance is akin to what we observe empirically and is what motivated some empiricists (proposed first in Fazzari, Hubbard, and Petersen (1988)) to reject the classical q-theory.

The last four columns in Table 1 correspond to the bias-corrected estimates using our empirical framework. Similarly, standard errors are clustered at the firm and year levels. As expected, the truncation bias is correctly accounted for. The average of the coefficients for q is 0.073 (0.001), which is not statistically different from its true value at 0.073. Moreover, the coefficient for CF is 0.003 (0.380) and is not statistically significantly different from zero, consistent with the classical q-theory.

As discussed in Section 2, linear regression yields downward-biased estimate of q when a truncation control is positively correlated with q *and* truncation innovation and investment innovation are positively correlated. These results are summarized in Table 1. In contrast, when a truncation control is positively correlated with q *and* the truncation innovation and investment innovation are negatively correlated, the coefficient for q is upward-biased and the coefficient for CF is downward biased. The simulation results for the case of negative correlation of the disturbances are summarized in Table 2. Because the truncation innovation and investment innovation are unobservable, we cannot observe the correlation between the two innovation terms. Nonetheless, given that our empirical analysis reveals that the q coefficient is *downward* biased and the CF coefficient is *upward* biased (see Section 6.2), one may argue that Table 1 better reflects the empirical data.

Finally, we simulate a case where the truncation innovation and investment innovation are not correlated. As summarized in Table 3, both the uncorrected and the corrected estimate is almost exactly the true coefficient. This simulation exercise is important for two reasons. First, this confirms that the correlation between the two innovation terms is the main driver behind the endogenous truncation bias as the uncorrected coefficients are exactly what DGP specifies. Second, more importantly, this illustrates a null hypothesis where the endogenous truncation bias does not exist. More specifically, in case the null hypothesis is true, the simulation exercise shows that our framework does not, correctly, add any additional noise to it. This implies that if we observe truncation-corrected estimates that are different from the uncorrected ones, that is an evidence that the data are contaminated with truncation bias.

5 Empirical Framework: Application

In this section, we discuss how the econometric framework (discussed in Section 3) is applied to the data.

5.1 Endogenous Truncation: Listing and Not-Delisting

Public firms' data are readily available but private firms' data are hardly available. Consequently, many researchers use public firms' data (e.g. Compustat) to test economic theories. In order for firm i's data to be available at year t, the firm i should have decided to go public at year $t' \leq t$ *and* the firm i remains listed every year between year t' and t. That is, firm i makes decisions at every point in the interval $[t', t]$. Given that these decisions are endogenously determined, the empirical tests based on public firms' data likely suffer from the endogenous truncation bias. In order to address such bias, we use the framework discussed in Section 3. As discussed, we need to determine covariates that enter the firm's endogenous truncation decision, namely listing and not-delisting. For that, we turn to the related literature.

We first discuss firms' listing decision. According to the literature,[4] firms tend to go public when they possess larger growth opportunities, are larger in size, have larger sales growth, and have higher TFP. We follow Imrohoroglu and Tuzel (2014) when constructing

[4] Doidge, Karolyi, and Stulz (2004); Pagano, Panetta, and Zingales (1998); Chemmanur and Fulghieri (1999); Chemmanur, He, and Nandy (2010). See Ritter and Welch (2002) for further literature review.

the TFP measure. Next, we need to discuss how firms decide not to delist, thereby staying in the sample. Campbell, Hilscher, and Szilagyi (2008) study firms' delistings, and find that firms' delisting decision depends on firms' growth opportunities, net income, leverage, excess equity return, stock return volatility, firm size, cash, and stock price. We use the same covariates to study firms' *not*-delisting decision. Variable definitions are provided in Section 6.1.

5.2 Corporate Investment

Now, we discuss our main application: how does a firm's cash flow affect its investment and what is its investment to q sensitivity? Under the classical assumptions,[5] the classical q-theory predicts that marginal q is a sufficient statistic for investment, that is, cash flow should not affect a firm's investment after controlling for a firm's marginal q. However, researchers have not reached a consensus on its empirical validity. Erickson and Whited (2000) and Erickson and Whited (2012) apply the econometric method in Erickson and Whited (2002), which uses higher order moments to identify the equation coefficients, and cannot reject that the effect of cash flow on investment is zero, thereby corroborating the prediction of the q-theory. Almeida, Campello, and Galvao (2010) use lagged variables in a panel structure as instrumental variables to address the measurement error in Tobin's q and find that cash flow affects investment positively, contradicting the theoretical prediction in the absence of financing frictions (see also Fazzari, Hubbard, and Petersen (1988) Gilchrist and Himmelberg (1995) Love (2003) Chen and Chen (2012)). Most recently, Chalak and Kim (2020) provide an econometric framework to study multiple equations *jointly* and apply the framework to empirically study the same question.

[5]This result assumes quadratic investment adjustment costs, constant return to scale, perfect competition, and an efficient financial market (see Hayashi (1982)).

5.3 Endogenous Truncation and Investment

After we account for fixed effects, the relevant linear regression model is:

$$\frac{I_{it}}{K_{it-1}} = \alpha_0 + \beta \cdot q_{it-1} + \epsilon^1_{it} \tag{15}$$

where ϵ^1_{it} is a mean-zero disturbance term in the investment equation. Also, firms are endogenously included in the sample, following Equation (5). Then, we have

$$\begin{aligned}
& E\left[\left.\frac{I_{it}}{K_{it-1}}\right| \left\{q_{it-1}, g(Z_{it}) + \epsilon^2_{it} \geq 0\right\}\right] \\
=& E\left[\left(\alpha_0 + \beta \cdot q_{it-1} + \epsilon^1_{it}\right) \mid \left\{q_{it-1}, g(Z_{it}) + \epsilon^2_{it} \geq 0\right\}\right] \\
=& E\left[\left(\alpha_0 + \beta \cdot q_{it-1}\right) \mid \left\{q_{it-1}, g(Z_{it}) + \epsilon^2_{it} \geq 0\right\}\right] + E\left[\epsilon_1 \middle| \left\{q_{it-1}, g(Z_{it}) + \epsilon^2_{it} \geq 0\right\}\right] \\
=& E\left[\left(\alpha_0 + \beta \cdot q_{it-1}\right) | q_{it-1}\right] + E\left[\epsilon^1_{it} \middle| \left\{q_{it-1}, g(Z_{it}) + \epsilon^2_{it} \geq 0\right\}\right] \\
=& E\left[\left(\alpha_0 + \beta \cdot q_{it-1} + \epsilon^1_{it}\right) | q_{it-1}\right] + E[\epsilon^1_{it} | g(Z_{it}) + \epsilon^2_{it} \geq 0] \\
=& E\left[\left.\frac{I_{it}}{K_{it-1}}\right| q_{it-1}\right] + \underbrace{E[\epsilon^1_{it} | g(Z_{it}) + \epsilon^2_{it} \geq 0]}_{\text{Endogenous truncation bias}}
\end{aligned}$$

where the third equality holds due to **Assumption** 3, and the fourth equality holds due to **Assumption** 1.

Even though we are truly after $E\left[\left.\frac{I_{it}}{K_{it-1}}\right| q_{it-1}\right]$, because we only observe non-truncated firms' data, the aforementioned empirical studies rely on $E\left[\left.\frac{I_{it}}{K_{it-1}}\right| \{q_{it-1}, g(Z_{it}) + \epsilon^2_{it} \geq 0\}\right]$. Thus, the bias term, $E[\epsilon^1_{it} | g(Z_{it}) + \epsilon^2_{it} \geq 0]$, arises. Our econometric framework helps to correct for this bias term.

6 Findings

6.1 Data and Variable Construction

We closely follow the literature in selecting the sample and constructing the variables (see e.g. Almeida and Campello (2007); Erickson and Whited (2012); Erickson, Jiang, and Whited (2014); Peters and Taylor (2017))). We use data from Compustat on firms between 1971 to 2018. Our sample starts from year 1971 because the investment and total q measures provided by Peters and Taylor (2017) are available after year 1971. We remove financial firms (Standard Industrial Classification (SIC) code 6000 to 6999) and regulated firms (SIC code 4900 to 4999). Following the literature (e.g. Peters and Taylor (2017); Andrei, Mann, and Moyen (2019)), we drop any observations with less than $5 million in gross property, plant, and equipment. We delete firm-year observations with missing data on one of the variables used in the analysis. Lastly, we winsorize the sample at the 1% and 99% level. The final sample is an unbalanced panel of 54,899 firm-year observations from 1971 through 2018, with 1,144 firms per year on average.

Table 4 shows descriptive statistics for all the firm characteristics used in the analysis. For our main analysis (presented in Section 6.2), we use investment, q, and cash flow measures that account for both tangibles and intangibles. These are introduced by Peters and Taylor (2017) and are widely used in subsequent papers including, for example, Andrei, Mann, and Moyen (2019). We prefix the three measures with *total*. Total investment ratio is capital expenditure (Compustat item: CAPX) plus R&D (Compustat item: XRD) and 30% of SG&A (Compustat item: XSGA), divided by the lagged sum of gross PP&E (Compustat item: PPEGT) and intangible capital. Total cash flow is the sum of income before extraordinary items (Compustat item: IB) and depreciation and amortization (Compustat item: DP), divided by the lagged sum of gross PP&E (Compustat item: PPEGT) and intangible capital. The intangible capital series is downloaded from the online resources for Peters and

Taylor (2017), from which we also obtain the series of total q.

As one of our robustness checks (summarized in Section 6.5.2), we use alternative definitions for investment, q, and cash flow. To differentiate from the main analysis counterparts, we prefix alternative measures with *tangible*. We define tangible investment ratio as capital expenditure (Compustat item: CAPX), divided by lagged total assets (Compustat item: AT). We measure tangible q by market value of equity (Compustat item: $(PRCC\_F \times CSHO)$) plus total assets (Compustat item: AT) minus common equity (Compustat item. CEQ) minus deferred taxes (Compustat item TXDB) divided by total assets (Compustat item: AT). We define tangible cash flow as the sum of income before extraordinary items (Compustat item: IB) and depreciation and amortization (Compustat item: DP), divided by lagged total assets (Compustat item: AT).

Both sets of aforementioned measures are similar in magnitude to what is documented (Peters and Taylor, 2017; Erickson and Whited, 2012). Notably, the difference between total investment ratio/q and tangible investment ratio/q is attributed to the lack of intangible component in the latter. Even though total cash flow and tangible cash flow share the same numerator (Compustat item IB), they are quite different because the former's denominator accounts for intangible whereas the latter does not.

Next, we closely follow the literature to define the covariates that determine the firms' listing decision. Covariates behind the listing includes market to book ratio, sales growth, TFP, and size. We define market to book ratio as the same as tangible q above. We define sales growth as growth in net sales (Compustat item: SALE). We closely follow Imrohoroglu and Tuzel (2014) to construct panel data of TFP. We define firm size as the natural logarithm of net sales.

Lastly, we closely follow Campbell, Hilscher, and Szilagyi (2008) to construct covariates for firms' endogenous not-delisting decision. We define netincome as net income over market value of total assets (Campbell's code: NIMTAAVG), leverage as total liabilities over book value of total assets (Campbell's code: TLMTA), excess equity return as log of gross excess return over market retun (Campbell's code: EXRETAVG), stock return volatility as square root of the sum of squared firm stock returns over a 3-month period (Campbell's code: SIGMA), relative size as as log of firm's market equity over the total market valuation (Campbell's code: RSIZE), cash as annualized, stock of cash and short-term investments over the market value of total assets (Campbell's code: CASHMTA), and stock price as log of price per share (Campbell's code: PRICE).

6.2 Main Results

We estimate Equation (15) by using total investment ratio, total q, and total cash flow. Table 5 highlights our main contribution to the literature. Column (1) reproduces the result that the previous literature has documented. As shown, the CF coefficient is large, with a value of 0.486, and is statistically significant, rejecting the classical q-theory. Column (1) suffers from the endogenous truncation bias because firms' listing decision and investment decision are correlated even after accounting for controls.

Column (2) through (8) summarize the results when we correct for endogenous truncation bias. These columns illustrate bias-corrected coefficient estimates with increasing numbers of correction controls. As we add more controls, we get closer to more complete specification. Thus, **Assumption** 1, **Assumption** 2 and **Assumption** 3 become less likely to be violated. Correction controls, $C_2 - C_8$, are set as follows. C_2 and C_3 includes only the determinants for listing decision. $C_4 - C_8$ control for full determinants of listing *and* the subset of determinants for not-delisting decisions. $C_2 = \{\text{MB}\}$ and $C_3 = C_2 \cup \{\text{TFP, Size, Sales growth}\}$. $C_4 = C_3 \cup \{\text{Net income, Excess equity return}\}$, $C_5 = C_4 \cup \{\text{Relative size, Leverage}\}$, $C_6 = C_5 \cup \{\text{Stock return equity}\}$, $C_7 = C_6 \cup \{\text{Cash}\}$, and $C_8 = C_7 \cup \{\text{Stock price}\}$.

We now discuss the results. First, the CF coefficient's significance level decreases as we get closer to the complete specification (column (8)). When we get sufficiently close to the complete specification, CF coefficient is not statistically significant (column (6), (7), and (8)). Thus, we cannot reject the classical q-theory. This implies that, absent truncation bias correction, endogenous truncation bias leads to overestimating the statistical significance of CF. Moreover, these results illustrate how the endogenous truncation bias gets better corrected as we suffer less from omitted variable problems.

Second, the q coefficient stays statistically significant yet its magnitude increases as we get closer to the complete specification. Specifically, the q coefficient has more than quadrupled after correcting for the endogenous truncation bias using the most exhaustive list of correction controls. This implies that, absent truncation bias correction, the bias incorrectly lowers the economic significance of q.

6.3 Correlation Between Two Disturbance Terms

Section 6.2 shows that linear regression among truncated sample yields a *downward* biased q coefficient. In order to show that the truncation bias is the main driver behind it, as discussed and illustrated in Section 2 and 4, it is important to show that the investment disturbance term is positively correlated with listing disturbance term. Unfortunately, it is empirically challenging to provide direct evidence because both disturbance terms are unobservable and data for unlisted firms are not available. Thus, we provide two suggestive evidence.

The first suggestive evidence relates investment disturbance terms to the firm age. we estimate investment disturbance terms after regressing firm investment on Q and accounting for firm fixed effects and year fixed effects. For each firm age, we take cross-sectional average of firm disturbance terms and plot it in Figure 3. As shown, investment disturbance decreases over firm age. This implies that there seems to be over-investment (positive investment disturbance) when firms are young and under-investment (negative investment disturbance) when firms are old. This is consistent with the IPO literature (e.g. Brau and Fawcett (2006))'s finding that firms decide to go public to engage in M&A and subsequently large lump sum of investment. This shows that listing disturbance terms (large means high likelihood of IPO) are positively correlated with investment disturbance terms (large means over-investment).

Next, we use firms' pre-default stage: nearly, but not quite, default. We use firms' entering pre-default stage to proxy delisting. This assumption allows us to observe data even after the firms enter pre-default stage and estimate the disturbance terms. We closely follow Elkamhi, Ericsson, and Parsons (2012); Chen, Hackbarth, and Strebulaev (2022); Elkamhi, Kim, and Salemo (2022) to define pre-default state where operating income is smaller than interest expense.[6] Using this approximation, we first estimate investment disturbance terms after accounting for firm fixed effects and year fixed effects. This implicitly assumes that investment-q sensitivities are the same between pre-default state and non pre-default state. Next, we estimate listing disturbance terms. For this, we first construct dummy variable that is 0 if it is in pre-default stage and 1 otherwise. Then, we regress the dummy variable on the aforementioned determinants for listing decision, firm fixed effects, and year fixed effects. We label the residuals as listing disturbance terms. Finally, we estimate the correlation between the investment disturbance terms and listing disturbance term. The correlation is 0.0101 and is statistically significant at 5% level.

6.4 Evolution of Endogenous Truncation Bias

In this subsection we explore the time variation of the truncation bias severity by comparing the uncorrected cash flow coefficient to the truncation bias-corrected cash flow coefficient. That is, we explore how the severity of the truncation bias varies over time. We relate this time variation to the literature. We undertake the following steps. First, for each firm i, we estimate cash flow sensitivity using firm i's entire time-series data during its existence in the panel. We note that β_1 is fixed for each firm. Second, for each year, we average these

[6] To be more precise, we say that firms are in pre-default state if the interest-coverage ratio (Compustat item: XINT divided by OIBDP) is greater than 1.

β_1's across all firms that are present in Compustat that year. These steps are equivalent to Andrei, Mann, and Moyen (2019) and ensure that the time-series pattern is driven by changes in the composition of Compustat firms.

In both panels, the black dashed lines in Figure 4 show that when not correcting for the truncation bias cash flow sensitivities have decreased over time, confirming Chen and Chen's (2012) main result. The time-series average before 1996 is 0.68 whereas that after 1996 is 0.61. In both panels, the black solid line shows how cash flow sensitivities change once the endogenous truncation bias is corrected for. Bias-corrected cash flow sensitivities are always smaller than uncorrected cash flow sensitivities. This illustrates that the truncation bias correction makes it tougher to reject the classical q-theory. The time-series average before 1996 is 0.35 whereas that after 1996 is 0.25. This shows that investment-cash flow sensitives have declined over time even after we correct for the endogenous truncation bias.

The gap between the dashed line (the uncorrected cash flow coefficients) and the solid black line (the truncation-bias corrected cash flow coefficients) reflects the severity of the truncation bias. Interestingly, this gap mirrors the (inverse of) number of listed firms in the U.S. (as well as, to some extent, the listing propensity) as reported in Doidge, Karolyi, and Stulz (2017). As the number of listed firms and the listing propensity rise towards the mid-1990s, the truncation bias becomes less severe. Subsequently to the mid-1990s the number of U.S. listed firms and the listing propensity fall substantially, accompanied by a substantial rise in the truncation bias severity.[7]

[7]Similar to Doidge, Karolyi, and Stulz (2017), we use the number of total listed domestic companies in the U.S. from World Bank's World Development Indicators (WDI) database and panel (a) plots the number. Again, similar to Doidge, Karolyi, and Stulz (2017), we construct listing propensity and panel (b) plots the number. Listing propensity is the number of total listed domestic companies in the U.S. divided the total number of firms in the U.S. We get the total number of firms in the U.S. from Census' Business Dynamics Statistics (BDS, https://www.census.gov/programs-surveys/bds.html)

6.5 Robustness Check

6.5.1 Measurement Error

In a series of papers Erickson and Whited study how measurement error drives the rejection of the q-theory and the significance of cash flows in investment regressions.[8] They propose an econometric framework to correct for the measurement error. Erickson and Whited then demonstrate that when applying their measurement error correction cash flows cease being significant, and the q-theory has good explanatory power.

We study how our truncation bias correction fares in face of Erickson and Whited's measurement error correction. We apply the Erickson, Parham, and Whited (2016) procedure to correct for the measurement error.[9] The results are presented in Table 6. Column (1) does not implement any correction, column (2) corrects only for the measurement error, and column (3) corrects for endogenous truncation bias using C_8 correction controls.[10] Column (4) corrects for both measurement error and endogenous truncation bias.

We first discuss column (2) and (3) relative to column (1). Correcting for either only measurement error or truncation bias leads the q coefficient to quadruple. When employing the Erickson-Whited correction for measurement error, we find that the investment-cash flow coefficient declines substantially. Our finding differs from Erickson, Jiang, and Whited (2014) which finds investment-cash flow coefficient to completely disappear. The different result is attributed to different definitions: Erickson, Jiang, and Whited focus on *tangible* investment whereas we use measures that capture both *tangible* and *intangible* capitals when measuring investment as well as q. Column 3 shows that correcting for the endogenous truncation bias yields statistically insignificant CF coefficient. Now, we discuss column (4). Correction for both measurement error and truncation bias leads q coefficient to increase

[8] Erickson and Whited (2000); Erickson and Whited (2002); Erickson and Whited (2012); Erickson, Jiang, and Whited (2014)

[9] https://ideas.repec.org/c/boc/bocode/s457525.html

[10] As described in Section 6.2, C_8 includes the following correction control variables: MB, TFP, size, sales growth, net income, excess equity return, relative size, leverage, stock return equity, cash, and stock price.

by almost nine times. Moreover, q coefficient' statistical significance level has increased. When we correct for both, CF coefficient is statistically indistinguishable from zero and thus we cannot reject the classical q-theory. This underscores that truncation bias correction complements measurement error correction.

6.5.2 Alternative Variable Definitions

Measures proposed by Peters and Taylor (2017) are not the only ways to proxy investment, cash flow and q. In this subsection, we show results when we use measures that are normalized by total assets, with detailed definitions discussed in Section 6.1. Contrary to Peters and Taylor (2017)'s, these measures do not account for intangible capitals but are nonetheless widely used (e.g. Asker, Farre-Mensa, and Ljungqvist (2015)). As seen in Table 7, the results are statistically and economically similar to our main specification results that were discussed in Section 6.2.

We now discuss the results. Tangible q is the market to book ratio whereas tangible cash flow is total-asset deflated cash flow. Column (1) of Table 7 reproduces the result that the previous literature has documented. As shown, CF coefficient is statistically significant, rejecting the classical q-theory. However, we note that the regression in column (1) likely suffers from the endogenous truncation bias because firms' listing decision and investment decision are correlated even after accounting for controls. Column (2) through (8) summarize the result when we correct for endogenous truncation bias. These columns illustrate coefficient estimates for the regressions with increasing numbers of correction controls. Correction controls are set similarly to the main results. As argued before, as we add more controls, we get closer to a complete specification.

First, the CF coefficient's significance level decreases as we get closer to the complete specification (column (8)). When we get sufficiently close to the complete specification, the CF coefficient is not statistically significant (column (6), (7), and (8)). Thus, we cannot reject the classical q-theory. This implies that, absent truncation bias correction, endogenous truncation bias leads to overestimation of the statistical significance of CF. Second, the q coefficient stays statistically significant yet its magnitude generally increases as we move closer to the more complete specification. The q coefficient has increased by at least 31% after correcting for the endogenous truncation bias using the most exhaustive list of correction controls. This implies that, absent truncation bias correction, the bias incorrectly lowers the economic significance of q.

7 Conclusion

In this paper we provide an econometric framework to correct for endogenous truncation that characterizes many empirical settings in economics and finance. We focus on the endogenous truncation bias in estimates of investment-cash flow sensitivity and investment-q sensitivity. This bias occurs because existing studies of this topic almost exclusively use truncated samples of only publicly listed firms. The exclusion of privately held firms generates a bias because the listing decision is largely endogenous and depends on firm characteristics, leading to biased OLS estimates. We subsequently apply our proposed endogenous truncation correction. The results change starkly; in the sample 1971-2018 cash flows ceases to be statistically and economically significant, whereas the investment-q sensitivity rises sharply. Moreover, we show that the endogenous truncation bias declines in the period prior to 1996, when the number of listed firms rises. Since 1996 the bias rises again, when the number of listed firms falls substantially. The same econometric framework can be applied in many other empirical studies that use truncated sample data.

References

Abel, A. B. (2018). The effects of q and cash flow on investment in the presence of measurement error. *Journal of Financial Economics 128*(2), 363–377.

Abel, A. B. and J. C. Eberly (2011). How q and cash flow affect investment without frictions: An analytic explanation. *Review of Economic Studies 78*(4), 1179–1200.

Almeida, H. and M. Campello (2007). Financial constraints, asset tangibility, and corporate investment. *Review of Financial Studies 20*, 1429–1460.

Almeida, H., M. Campello, and A. Galvao (2010). Measurement error in investment equations. *Review of Financial Studies 23*(9), 3279–3328.

Alti, A. (2003). How sensitive is investment to cash flow when financing is frictionless? *Journal of Finance 58*(2), 707–722.

Andrei, D., W. Mann, and N. Moyen (2019). Why did the q theory of investment start working? *Journal of Financial Economics 133*, 251–272.

Asker, J., J. Farre-Mensa, and A. Ljungqvist (2015). Corporate investment and stock market listing: A puzzle? *Review of Financial Studies 28*(2), 342–390.

Billfeld, N. and M. Kim (2016). Truncation bias. *Available at SSRN 2786263*.

Billfeld, N. and M. Kim (2019). Semiparametric correction for endogenous truncation bias with vox populi-based participation decision. *IEEE Access 7*, 12114–12132.

Brau, J. C., R. B. Couch, and N. K. Sutton (2012). The desire to acquire and ipo long-run underperformance. *Journal of Finance 47*, 493–510.

Brau, J. C. and S. E. Fawcett (2006). Initial public offerings: An analysis of theory and practice. *Journal of Finance 61*, 399–436.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business and Economic Statistics 29*, 238–249.

Campbell, J. Y., J. Hilscher, and J. Szilagyi (2008). In search of distress risk. *Journal of Finance 63*, 2899–2939.

Chalak, K. and D. Kim (2020). Measurement error in multiple equations: Tobin's q and corporate investment, saving, and debt. *Journal of Econometrics 214*, 413–432.

Chemmanur, T. and P. Fulghieri (1999). A theory of the going-public decision. *Review of Financial Studies 12*, 249–279.

Chemmanur, T., S. He, and D. Nandy (2010). The going-public decision and the product market. *Review of Financial Studies 23*, 1855–1908.

Chen, H. and S. Chen (2012). Investment-cash flow sensitivity cannot be a good measure of financial constraints: Evidence from the time series. *Journal of Financial Economics 103*, 393–410.

Chen, Z., D. Hackbarth, and I. A. Strebulaev (2022). A unified model of distress risk puzzles. *Journal of Financial Economics 146*(2), 357–384.

Cooper, R. and J. Ejarque (2003). Financial frictions and investment: requiem in q. *Review of Economic Dynamics 6*(4), 710–728.

Djama, C., I. Martinez, and S. Serve (2012). What do we know about delistings? a survey of the literature. *Working Paper*.

Doidge, C., G. A. Karolyi, and R. M. Stulz (2004). Why are foreign firms listed in the u.s. worth more? *Journal of Financial Economics 71*, 205–238.

Doidge, C., G. A. Karolyi, and R. M. Stulz (2017). The u.s. listing gap. *Journal of Financial Economics 123*, 464–487.

Elkamhi, R., J. Ericsson, and C. A. Parsons (2012). The cost and timing of financial distress. *Journal of Financial Economics 105*, 62–81.

Elkamhi, R., D. Kim, and M. Salemo (2022). How large are pre-default costs of financial distress? estimates from a dynamic model. *Working Paper*.

Erickson, T., C. H. Jiang, and T. M. Whited (2014). Minimum distance estimation of the errors-in-variables model using linear cumulant equations. *Journal of Econometrics 183*, 211–221.

Erickson, T., R. Parham, and T. Whited (2016). Xtewreg: Stata module to estimate errors-in-variable model with mismeasured regressors. *Statistical Software Components S457525, Boston College Department of Economics, revised 02 Sep 2016*.

Erickson, T. and T. Whited (2000). Measurement error and the relationship between investment and q. *Journal of Political Economy 108*, 1027–1057.

Erickson, T. and T. Whited (2002). Two-step gmm estimation of the errors-in variables model using high-order moments. *Econometric Theory 18*, 776–799.

Erickson, T. and T. Whited (2012). Treating measurement error in tobin's q. *Review of Financial Studies 25*, 1286–1329.

Fazzari, S. M., G. R. Hubbard, and B. C. Petersen (1988). Financing constrained and corporate investment. *Brookings Papers Econ. Activity 1*, 141–195.

Gilchrist, S. and C. P. Himmelberg (1995). Evidence on the role of cash flow for investment. *Journal of Monetary Economics 36*(3), 541–572.

Gomes, J. (2001). Financing investment. *American Economic Review 91*(5), 1263–1285.

Hayashi, F. (1982). Tobin's marginal q and average q: Neoclassical interpretation. *Econometrica 50*(1), 213–224.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics 58*, 71–120.

Imrohoroglu, A. and S. Tuzel (2014). Firm-level productivity, risk, and return. *Management Science 60*, 2073–2090.

Jain, B. A. and O. Kini (1994). The post-issue operating performance of ipo firms. *Journal of Finance 49*, 1699–1726.

Love, I. (2003). Financial development and financing constraints: International evidence from the structural investment model. *Review of Financial Studies 16*(3), 765–791.

Malikov, E., S. C. Kumbhakar, and Y. Sun (2016). Varying coefficient panel data model in the presence of endogenous selectivity and fixed effects. *Journal of Econometrics 190*, 233–251.

Malmendier, U. and G. Tate (2005). Ceo overconfidence and corporate investment. *Journal of Finance 60*, 2661–2700.

Malmendier, U. and G. Tate (2008). Who makes acquisitions? ceo overconfidence and the market's reaction. *Journal of Financial Economics 89*, 20–43.

Malmendier, U., G. Tate, and J. Yan (2011). Overconfidence and early-life experiences: The effect of managerial traits on corporate financial policies. *Journal of Finance 66*, 1687–1733.

Mehran, H. and S. Peristiani (2010). Financial visibility and the decision to go private. *Review of Financial Studies 23*, 519–547.

Pagano, M., F. Panetta, and L. Zingales (1998). Why do computers go public? an empirical analysis. *Journal of Finance 53*, 27–64.

Peters, R. and L. Taylor (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics 123*, 251–272.

Ritter, J. and I. Welch (2002). A review of ipo activity, pricing, and allocations. *Journal of Finance 57*, 1795–1828.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica 56*(4), 931–954.

Wansbeek, T. and A. Kapteyn (1989). Estimation of the error-components model with incomplete panels. *Journal of Econometrics 41*, 341–361.

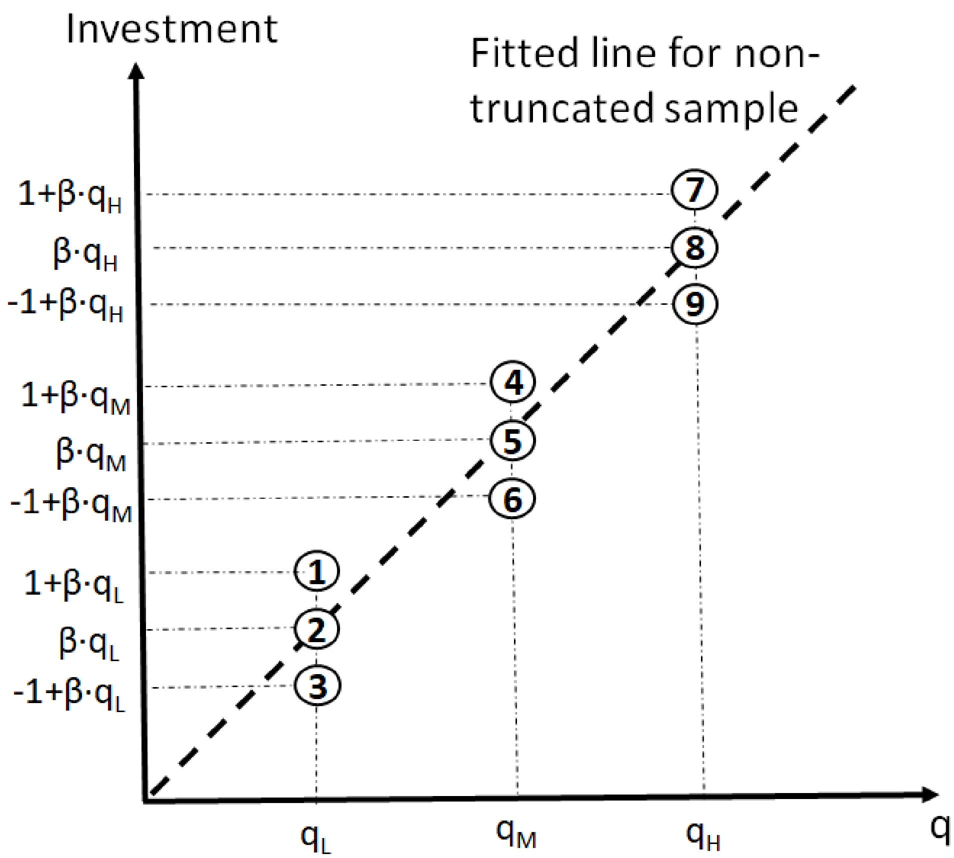
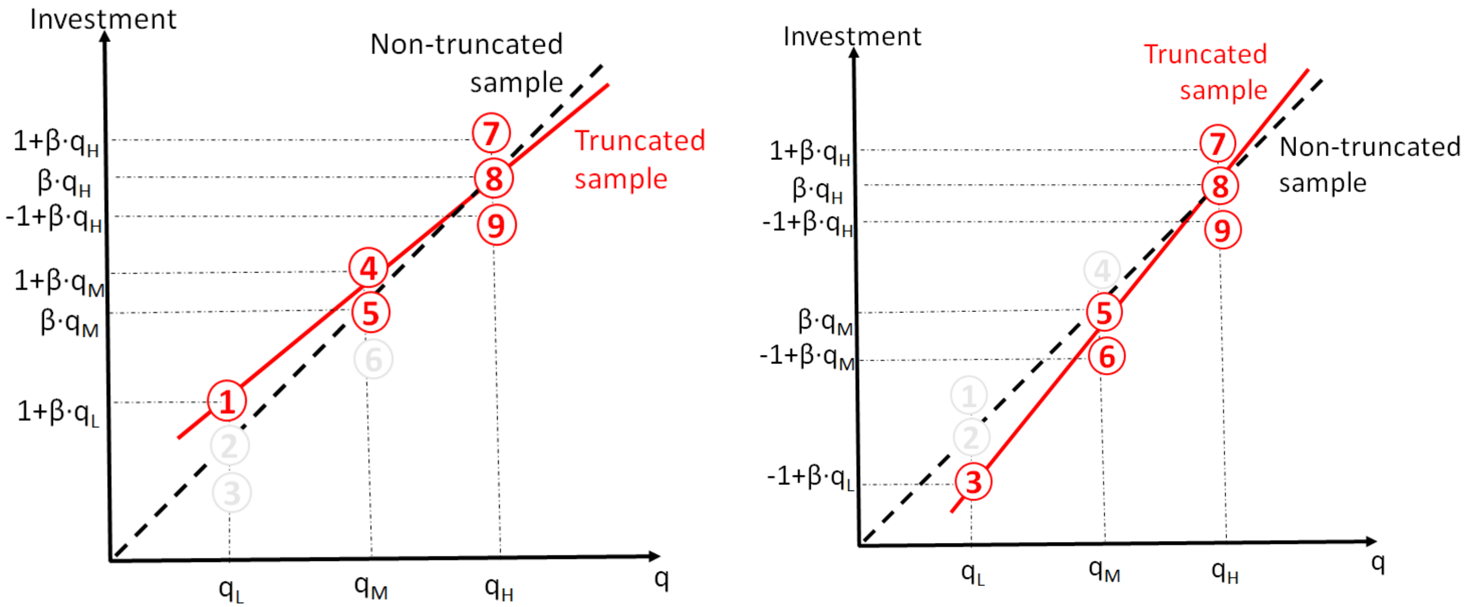


Figure 1: Un-truncated sample

This graph illustrates nine different equally likely types of (investment, q) pairs. Investment and q are constructed based on Equations (1), (2), and (3). The dashed line is the best fitted line. As shown, the line drawn based on the non-truncated sample has slope that is equal to the true investment-q sensitivity: β.



Panel A: Downward biased slope

Panel B: Upward biased slope

Figure 2: Panel A illustrates a case where firm characteristic X is positively correlated with q and ϵ_1 (investment disturbance term) and ϵ_2 (listing disturbance term) are positively correlated. When q is large, most firms will get listed because their other related characteristics are sufficiently high to satisfy the listing requirement. However, when q is small, firms need to have sufficiently large innovation terms in order to satisfy the listing requirement. Following this logic, we find that firm 1, 4, 5, 7, 8 and 9 are listed. The solid red line is the best fitted line of the truncated sample. As shown, the solid red line is flatter than the dashed line. This illustrates that the linear regression based on the truncated sample is *downward* biased. Panel B illustrates an alternative case where firm characteristic X is positively correlated with q *and* ϵ_1 and ϵ_2 are negatively correlated. Following the similar logic, firm 3, 5, 6, 7, 8 and 9 are listed. The solid red line is the best fitted line of the truncated sample. As shown, the solid red line is flatter than the dashed line. This illustrates that the linear regression based on the truncated sample is *upward* biased.

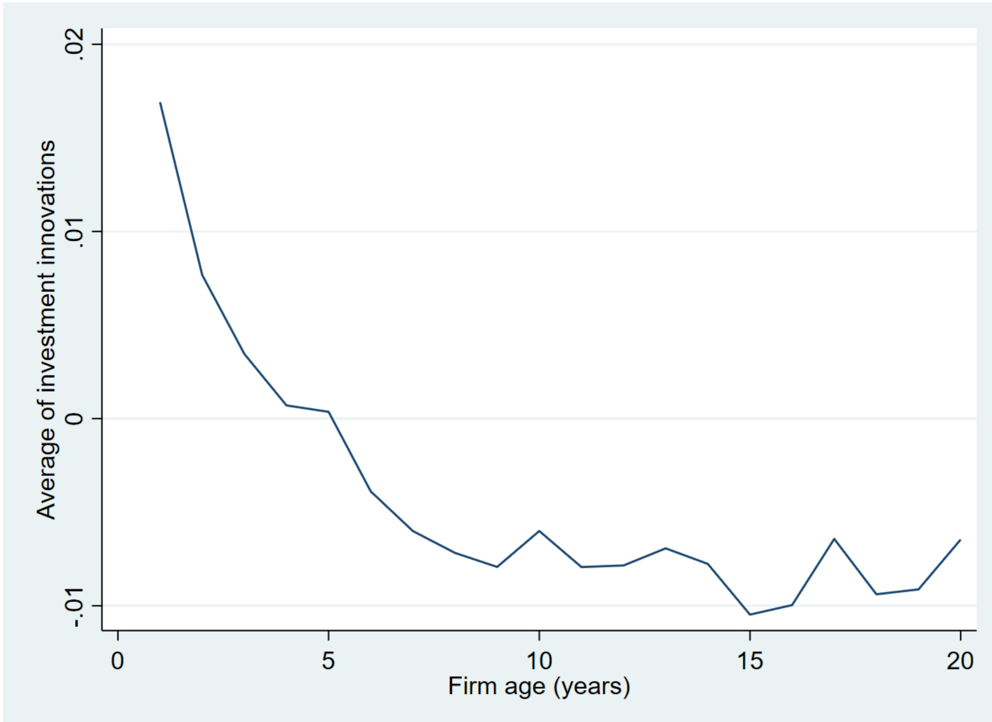
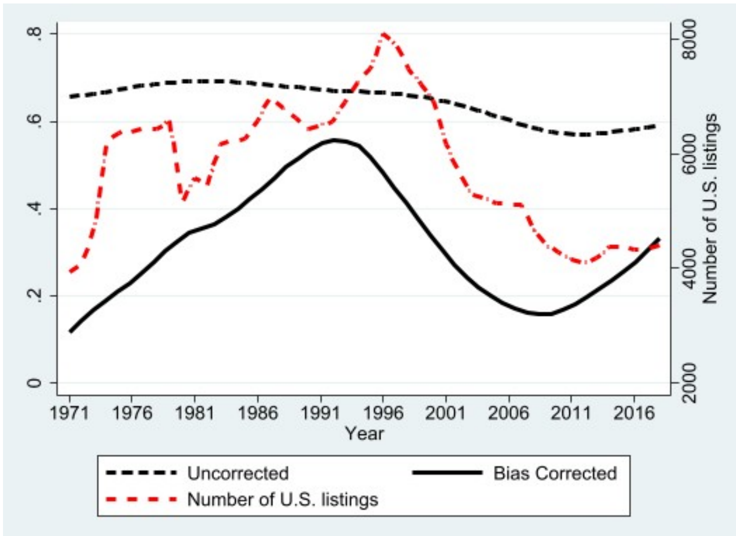
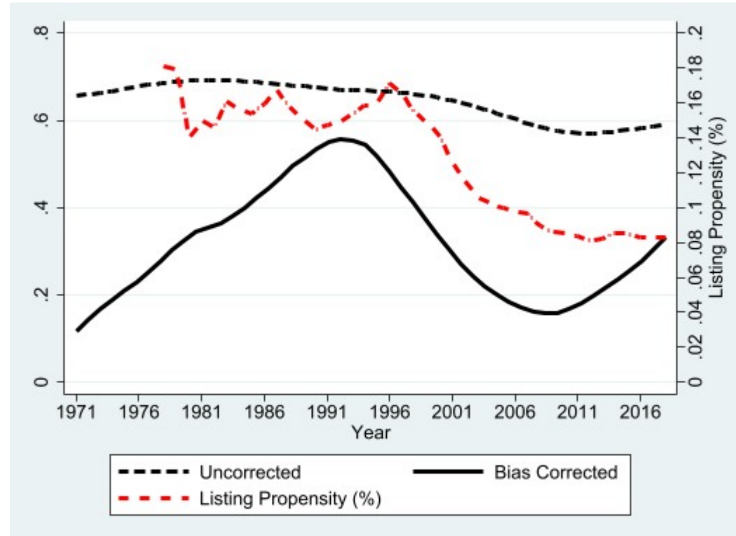


Figure 3: Investment innovation term over firm age

This graph illustrates investment disturbance terms to firm age. We estimate investment disturbance terms after regressing firm investment on Q and accounting for firm fixed effects and year fixed effects. For each firm age, we take cross-sectional average of firm disturbance terms. Investment disturbance decreases over firm age. This implies that there seems to be over-investment (positive investment disturbance) when firms are young and under-investment (negative investment disturbance) when firms are old.



Panel A: Comparison to number of U.S. listings



Panel B: Comparison to listing propensity (%)

Figure 4: Time Series of cash flow sensitivities based on the true data. In both panels, the dashed line shows how cash flow sensitivities change over time when the truncation bias is not corrected. In both panels, the solid line shows how cash flow sensitivities change over time when the truncation bias is corrected. In Panel A, the dashed-dotted line shows the number of total listed domestic companies in the U.S. In Panel B, the dashed-dotted line shows the listing propensity in percentages. Listing propensity is the number of total listed domestic companies in the U.S. divided by total the number of firms in the U.S.

Table 1: Simulation: Positive Correlation Between Investment and Truncation Innovations

This table presents results based on 1,000 simulations when investment innovation terms and truncation innovation terms are positively correlated. For each simulation, we construct unbalanced yet truncated panel data that include only listed firms. Each simulated panel follows DGP discussed in Section 4.1. The true coefficient for q is 0.073 and the true coefficeint for CF is 0. Every row, except for the last one, summarizes each simulation's results. The very last row shows the average over 1,000 simulations. The first column shows the untruncated sample size in firm-years. The second column shows the truncated sample size and the third column shows the number of firms in the corresponding truncated sample. The next 4 columns correspond to the panel OLS regression results without correcting for endogenous truncation bias. Standard errors are clustered at firm and year level. The last 4 columns correspond to the bias-corrected estimates using our empirical framework.

sim#	Untruncated sample size	Truncated sample size	Number of firms	Uncorrected				Truncation-bias Corrected			
				q		CF		q		CF	
				est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
1	542336	270580	5000	0.018	0.001	0.286	0.007	0.074	0.001	-0.249	0.327
2	543732	272238	5000	0.018	0.001	0.293	0.007	0.074	0.001	0.297	0.231
3	544600	271785	5000	0.015	0.001	0.310	0.007	0.070	0.001	-0.191	0.662
4	535465	267303	5000	0.016	0.001	0.310	0.007	0.071	0.002	0.867	0.201
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
999	541874	271147	5000	0.017	0.001	0.297	0.007	0.073	0.001	0.472	0.625
1000	539290	269827	5000	0.017	0.001	0.303	0.007	0.073	0.001	0.261	0.300
Average	540100	270049	5000	0.017	0.001	0.302	0.007	0.073	0.001	0.003	0.380
T-stats					20.958		44.284		49.354		0.008

Table 2: Simulation: Negative Correlation Between Investment and Truncation Innovations

This table presents results based on 1,000 simulations when investment innovation terms and truncation innovation terms are negatively correlated. For each simulation, we construct unbalanced yet truncated panel data that include only listed firms. Each simulated panel follows DGP discussed in Section 4.1. The true coefficient for q is 0.073 and the true coefficeint for CF is 0. Every row, except for the last one, summarizes each simulation's results. The very last row shows the average over 1,000 simulations. The first column shows the untruncated sample size in firm-years. The second column shows the truncated sample size and the third column shows the number of firms in the corresponding truncated sample. The next 4 columns correspond to the panel OLS regression results without correcting for endogenous truncation bias. Standard errors are clustered at firm and year level. The last 4 columns correspond to the bias-corrected estimates using our empirical framework.

sim#	Untruncated sample size	Truncated sample size	Number of firms	Uncorrected				Truncation-bias Corrected			
				q		CF		q		CF	
				est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
1	538021	268152	5000	0.129	0.001	-0.302	0.007	0.073	0.001	-0.133	0.319
2	530900	265209	4999	0.127	0.001	-0.285	0.007	0.072	0.001	-0.592	0.286
3	535280	267599	5000	0.128	0.001	-0.287	0.007	0.072	0.001	1.581	0.344
4	532949	267057	4999	0.129	0.001	-0.303	0.007	0.073	0.001	-0.120	0.224
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
999	534213	267160	5000	0.128	0.001	-0.302	0.007	0.071	0.001	-0.358	0.248
1000	541805	270932	5000	0.130	0.001	-0.306	0.007	0.075	0.001	0.149	0.238
mean	540001	269992	5000	0.129	0.001	-0.302	0.007	0.073	0.001	0.005	0.381
T-stats					162.630		-44.242		49.364		0.013

Table 3: Simulation: Zero Correlation Between Investment and Truncation Innovations

This table presents results based on 1,000 simulations when investment innovation terms and truncation innovation terms are not correlated. For each simulation, we construct unbalanced yet truncated panel data that include only listed firms. Each simulated panel follows DGP discussed in Section 4.1. The true coefficient for q is 0.073 and the true coefficeint for CF is 0. Every row, except for the last one, summarizes each simulation's results. The very last row shows the average over 1,000 simulations. The first column shows the untruncated sample size in firm-years. The second column shows the truncated sample size and the third column shows the number of firms in the corresponding truncated sample. The next 4 columns correspond to the panel OLS regression results without correcting for endogenous truncation bias. Standard errors are clustered at firm and year level. The last 4 columns correspond to the bias-corrected estimates using our empirical framework.

sim#	Untruncated sample size	Truncated sample size	Number of firms	Uncorrected				Truncation-bias Corrected			
				q		CF		q		CF	
				est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
1	542539	270721	5000	0.073	0.002	0.957	0.361	0.073	0.001	0.000	0.007
2	545833	272825	5000	0.074	0.002	-0.813	0.121	0.073	0.001	0.006	0.007
3	539005	269131	5000	0.071	0.002	-0.020	0.298	0.072	0.001	0.010	0.007
4	539928	270230	5000	0.074	0.002	-0.515	0.383	0.072	0.001	0.001	0.007
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
999	538653	269006	5000	0.069	0.002	-0.259	0.259	0.072	0.001	0.015	0.007
1000	543752	272151	5000	0.074	0.002	-0.410	0.546	0.072	0.001	0.001	0.007
mean	540114.8	270050.3	4999.843	0.073	0.002	0.017	0.380	0.073	0.001	0.000	0.007
T-stats					47.594		0.046		89.597		0.014

Table 4: Sample Descriptive Statistics

This table presents sample descriptive statistics for the firms in our data. Total investment ratio is capital expenditure (Compustat item: CAPX) plus R&D (Compustat item: XRD) and 30% of SG&A (Compustat item: XSGA), divided by the lagged sum of gross PP&E (Compustat item: PPEGT) and intangible capital. Total q is downloaded from Peters and Taylor (2017). Total cash flow is the sum of income before extraordinary items (Compustat item: IB) and depreciation and amortization (Compustat item: DP), divided by the lagged sum of gross PP&E (Compustat item: PPEGT) and intangible capital. The intangible capital series is downloaded from the online resources for Peters and Taylor (2017). Tangible investment ratio is capital expenditure (Compustat item: CAPX), divided by lagged total assets (Compustat item: AT). Tangible q is market value of equity (Compustat item: $(PRCC\_F \times CSHO)$) plus total assets (Compustat item: AT) minus common equity (Compustat item. CEQ) minus deferred taxes (Compustat item TXDB) divided by total assets (Compustat item: AT). Tangible cash flow is the sum of income before extraordinary items (Compustat item: IB) and depreciation and amortization (Compustat item: DP), divided by lagged total assets (Compustat item: AT). Sales growth is growth in net sales (Compustat item: SALE). TFP is downloaded from Imrohoroglu and Tuzel (2014). Firm size is natural logarithm of net sales. Not-delisting covariates closely follow Campbell, Hilscher, and Szilagyi (2008).

	N	Mean	SD	Min	Max
Total investment ratio	54,899	0.210	0.139	0.0389	0.821
Total q	54,899	1.038	1.564	-0.564	9.008
Total cash flow	54,899	0.208	0.145	-0.0652	0.840
Tangible investment ratio	54,899	0.0692	0.0645	0.00418	0.429
Tangible q	54,899	1.754	1.163	0.570	6.722
Tangible cash flow	54,899	0.106	0.0814	-0.150	0.369
Listing covariates					
Sales growth	54,899	0.0725	0.211	-2.378	8.146
TFP	54,899	0.0229	0.917	-8.205	4.319
Size	54,899	5.922	1.786	2.598	10.42
Not-delisting covariates					
Net income	54,899	0.00663	0.0135	-0.0665	0.0270
Excess equity return	54,899	-0.00390	0.0272	-0.107	0.0669
Relative size	54,899	-9.672	1.798	-13.69	-6.792
Leverage	54,899	0.392	0.241	0.0371	0.931
Stock return volatility	54,899	0.517	0.270	0.189	1.638
Cash	54,899	0.0879	0.0903	0.00243	0.394
Stock price	54,899	2.337	0.618	-3.408	2.708

Table 5: Main Result

The table summarizes results of estimating Equation (15). The analysis uses measures that account for both tangible and intangible investment. Total investment ratio is capital expenditure (Compustat item: CAPX) plus R&D (Compustat item: XRD) and 30% of SG&A (Compustat item: XSGA), divided by the lagged sum of gross PP&E (Compustat item: PPEGT) and intangible capital. Total q is downloaded from Peters and Taylor (2017). Total cash flow is the sum of income before extraordinary items (Compustat item: IB) and depreciation and amortization (Compustat item: DP), divided by the lagged sum of gross PP&E (Compustat item: PPEGT) and intangible capital. The intangible capital series is downloaded from the online resources for Peters and Taylor (2017). The sample period spans from 1971 to 2018. The analysis accounts for both firm fixed effects and year fixed effects. The reported standard errors are clustered at firm and year level. Column (1) does not correct for endogenous truncation bias whereas column (2) through column (8) correct for endogenous truncation bias. Correction controls, $C_2 - C_8$, are set as follows. C_2 and C_3 includes only the determinants for listing decision. $C_4 - C_8$ control for full determinants of listing *and* the subset of determinants for not-delisting decisions. $C_2 = \{\text{MB}\}$ and $C_3 = C_2 \cup \{\text{TFP, Size, Sales growth}\}$. $C_4 = C_3 \cup \{\text{Net income, Excess equity return}\}$, $C_5 = C_4 \cup \{\text{Relative size, Leverage}\}$, $C_6 = C_5 \cup \{\text{Stock return equity}\}$, $C_7 = C_6 \cup \{\text{Cash}\}$, and $C_8 = C_7 \cup \{\text{Stock price}\}$. Variable definitions are provided in Section 6.1.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Total q	0.0178***	0.0199***	0.0214***	0.0305***	0.0388**	0.0559*	0.0715*	0.0727*
	(0.00122)	(0.00120)	(0.00126)	(0.00701)	(0.0174)	(0.0283)	(0.0370)	(0.0375)
Total cash flow	0.486***	0.476***	0.433***	0.375***	0.316*	0.108	-0.0657	-0.0831
	(0.0203)	(0.0199)	(0.0211)	(0.0794)	(0.172)	(0.254)	(0.334)	(0.343)
Observations	54,899	54,899	54,899	54,899	54,899	54,899	54,899	54,899
Firm FE	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Trunc' Bias Correction	NO	YES	YES	YES	YES	YES	YES	YES
Correction Controls	$\varnothing$	C_2	C_3	C_4	C_5	C_6	C_7	C_8

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 6: Robustness Check: Measurement Error

The table summarizes results of estimating Equation (15) after correcting for endogenous truncation bias and/or measurement error. The analysis uses measures that account for both tangible and intangible investment. Total investment ratio is capital expenditure (Compustat item: CAPX) plus R&D (Compustat item: XRD) and 30% of SG&A (Compustat item: XSGA), divided by the lagged sum of gross PP&E (Compustat item: PPEGT) and intangible capital. Total q is downloaded from Peters and Taylor (2017). Total cash flow is the sum of income before extraordinary items (Compustat item: IB) and depreciation and amortization (Compustat item: DP), divided by the lagged sum of gross PP&E (Compustat item: PPEGT) and intangible capital. The intangible capital series is downloaded from the online resources for Peters and Taylor (2017). The sample period spans from 1971 to 2018. The analysis accounts for both firm fixed effects and year fixed effects The reported standard errors are clustered at firm and year level. Column (1) does not correct for any correction. Column (2) corrects for measurement error by following Erickson, Jiang, and Whited (2014). Column (3) corrects for truncation bias using C_8 correction controls. Column (4) corrects for both measurement error and truncation bias using C_8 correction controls. As defined in Table 5, $C_8 = \{\text{MB}, \text{TFP}, \text{Size}, \text{Sales growth}, \text{Net income}, \text{Excess equity return}, \text{Relative size}, \text{Leverage}, \text{Stock return equity}, \text{Cash}, \text{Stock price.}\}$. Variable definitions are provided in Section 6.1.

	(1)	(2)	(3)	(4)
Total q	0.0178***	0.0640***	0.0727*	0.152***
	(0.00122)	(0.00394)	(0.0375)	(0.0578)
Total cash flow	0.486***	0.206***	-0.0831	-0.164
	(0.0203)	(0.0264)	(0.343)	(0.200)
Observations	54,899	54,899	54,899	54,899
Firm FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Truncation Bias Correction	NO	NO	YES	YES
Measurement Error Correction	NO	YES	NO	YES
Correction Controls	$\varnothing$	$\varnothing$	C_8	C_8

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 7: Robustness Check: Alternative Variable Definitions

The table summarizes results of estimating Equation (15). The analysis uses measures that account for only tangible investment. Tangible investment ratio is capital expenditure (Compustat item: CAPX), divided by lagged total assets (Compustat item: AT). Tangible q is market value of equity (Compustat item: $(PRCC\_F \times CSHO)$) plus total assets (Compustat item: AT) minus common equity (Compustat item. CEQ) minus deferred taxes (Compustat item TXDB) divided by total assets (Compustat item: AT). Tangible cash flow is the sum of income before extraordinary items (Compustat item: IB) and depreciation and amortization (Compustat item: DP), divided by lagged total assets (Compustat item: AT). The sample period spans from 1971 to 2018. The analysis accounts for both firm fixed effects and year fixed effects. The reported standard errors are clustered at firm and year level. Column (1) does not correct for endogenous truncation bias whereas column (2) through column (8) correct for endogenous truncation bias. Correction controls, $C_2 - C_8$, are set as follows. C_2 and C_3 includes only the determinants for listing decision. $C_4 - C_8$ control for full determinants of listing *and* the subset of determinants for not-delisting decisions. $C_2 = \{\text{MB}\}$ and $C_3 = C_2 \cup \{\text{TFP, Size, Sales growth}\}$. $C_4 = C_3 \cup \{\text{Net income, Excess equity return}\}$, $C_5 = C_4 \cup \{\text{Relative size, Leverage}\}$, $C_6 = C_5 \cup \{\text{Stock return equity}\}$, $C_7 = C_6 \cup \{\text{Cash}\}$, and $C_8 = C_7 \cup \{\text{Stock price}\}$. Variable definitions are provided in Section 6.1.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tangible q	0.0125***	0.0140***	0.0159***	0.0174***	0.0182***	0.0156***	0.0167***	0.0164***
	(0.000815)	(0.000809)	(0.000914)	(0.00161)	(0.00216)	(0.00304)	(0.00369)	(0.00380)
Tangible cash flow	0.151***	0.143***	0.114***	0.106***	0.0577*	0.0376	0.0321	0.0329
	(0.0124)	(0.0119)	(0.0112)	(0.0316)	(0.0311)	(0.0288)	(0.0305)	(0.0308)
Observations	54,899	54,899	54,899	54,899	54,899	54,899	54,899	54,899
Firm FE	YES	YES	YES	YES	YES	YES	YES	YES
Year FE	YES	YES	YES	YES	YES	YES	YES	YES
Trunc' Bias Correction	NO	YES	YES	YES	YES	YES	YES	YES
Correction Controls	$\varnothing$	C_2	C_3	C_4	C_5	C_6	C_7	C_8

Robust standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A Math Appendix

A.1 Matrix Operator

Let d_t be the $(N_t \times H)$ matrix obtained by omitting rows of firms absent in year t from the $(H \times H)$ identity matrix. Denote $\boldsymbol{\iota}_T$ is a $T \times 1$ vector of ones and $\boldsymbol{\iota}_H$ is a $H \times 1$ vector of ones. In our present case of an unbalanced panel, matrix $\mathbb{D}$ gives the dummy-variable structure.

$$\mathbb{D} = \left(\underset{N \times H}{\mathbb{D}_1}, \underset{N \times T}{\mathbb{D}_2} \right) = \begin{bmatrix} d_1 & d_1 \boldsymbol{\iota}_H & \dots & 0 \\ : & & \ddots & : \\ d_T & 0 & \dots & d_T \boldsymbol{\iota}_H \end{bmatrix}$$

We express the equation above in vector form in order to employ the unbalanced panel demeaning operator, characterized in matrix form (Wansbeek and Kapteyn, 1989).

$$\boldsymbol{Y} = \mathbb{D}_1 \boldsymbol{\alpha} + \mathbb{D}_2 \boldsymbol{\delta} + \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{M} + \boldsymbol{U}$$

In the representation above the observations are ordered lexicographically firstly by time and secondly by firm (the index i changes more frequently).

The following operator matrices are constructed:

$$\begin{aligned} \overline{\mathbb{D}} &:= \mathbb{D}_2 - \mathbb{D}_1 \boldsymbol{\Delta}_H^{-1} \boldsymbol{\Delta}'_{HT} \\ \boldsymbol{q} &:= \boldsymbol{\Delta}_T - \boldsymbol{\Delta}_{HT} \boldsymbol{\Delta}_H^{-1} \boldsymbol{\Delta}'_{HT} = \mathbb{D}'_2 \overline{\mathbb{D}} \\ \boldsymbol{P} &:= \left(\boldsymbol{I}_N - \mathbb{D}_1 \boldsymbol{\Delta}_H^{-1} \mathbb{D}'_1 \right) - \overline{\mathbb{D}} \boldsymbol{q}^- \overline{\mathbb{D}}' \end{aligned}$$

where superscript $\cdot^-$ implies a generalized inverse. $\Delta_H := \mathbb{D}_1^T \mathbb{D}_1$ is a diagonal $(H \times H)$ matrix in which h-th element indicates the number of years the firm has been observed. $\Delta_T := \mathbb{D}_2^T \mathbb{D}_2$ is a diagonal $(T \times T)$ matrix in which t-th element indicates the number of firms in each year and $\Delta_{HT} := \mathbb{D}_2^T \mathbb{D}_1$ is the $(T \times H)$ matrix of zeros and ones indicating the firm's presence or absence in a certain year.

A.2 Inference

The two-way clustering covariance matrix is defined as a combination of one-way cluster-robust matrices, where the clusters are across firms, time and the intersections of firms and time Cameron, Gelbach, and Miller (2011):

$$\widehat{\mathbb{V}}\left[\widehat{\boldsymbol{\beta}}\right] = \overbrace{\widehat{\mathbb{V}}^{H}\left[\widehat{\boldsymbol{\beta}}\right]}^{\text{across firms}} + \overbrace{\widehat{\mathbb{V}}^{T}\left[\widehat{\boldsymbol{\beta}}\right]}^{\text{across time}} - \overbrace{\widehat{\mathbb{V}}^{H\cap T}\left[\widehat{\boldsymbol{\beta}}\right]}^{\text{across firms and time}}$$

We denote $\mathbb{S}^{\chi}$ with $\chi \in \{H, T, H\cap T\}$ an indicator matrix, such that its ij-th entry equal unity if observations i and j belong to the same cluster χ $\forall\, \chi \in \{H, T, H\cap T\}$:

$$\widehat{\mathbb{V}}^{\chi}\left[\widehat{\boldsymbol{\beta}}\right] = \left((\boldsymbol{\Delta P X})'(\boldsymbol{\Delta P X})\right)^{-1}(\boldsymbol{\Delta P X})'\left((\sqrt{c^{\chi}}\widehat{\boldsymbol{\Delta P U}})(\sqrt{c^{\chi}}\widehat{\boldsymbol{\Delta P U}})'\cdot\mathbb{S}^{\chi}\right)(\boldsymbol{\Delta P X})\left((\boldsymbol{\Delta P X})'(\boldsymbol{\Delta P X})\right)^{-1}$$

where $c^{\chi} = \frac{n_{\chi}}{n_{\chi}-1}\frac{N-1}{N-K}$ with n_{χ} specifying the number of observations in cluster χ, is a correction for small-sample in two-way clustering. The notation $\cdot$ is an element-wise product.

A.3 Special case: Common covariates

We now attend to motivate the validity of Eq. (11) in cases where $\boldsymbol{X}$ and $\boldsymbol{Z}$ consist of common covariates. Without loss of generality, suppose that $[\boldsymbol{Z}] = [\widetilde{\boldsymbol{Z}}, \boldsymbol{X}]$ such that $\widetilde{\boldsymbol{Z}}$ represents the covariates which are not included in $\boldsymbol{X}$. Let $\{\chi_1, \ldots, \chi_{NT}\}$ be NT vectors each of size $1 \times p_{\chi}$ satisfying,

$$\widehat{\chi} = \boldsymbol{P}\chi \quad \text{with} \quad \{\chi, \widehat{\chi}\} \in \left\{\left\{\boldsymbol{X}, \widehat{\boldsymbol{X}}\right\}, \left\{\boldsymbol{Y}, \widehat{\boldsymbol{Y}}\right\}, \left\{\widetilde{\boldsymbol{Z}}, \widehat{\widetilde{\boldsymbol{Z}}}\right\}, \left\{\boldsymbol{U}, \widehat{\boldsymbol{U}}\right\}\right\}.$$

Equivalently in vector notations,

$$[\widehat{\chi}_1, \ldots, \widehat{\chi}_{NT}]^{T} = \boldsymbol{P}\,[\chi_1, \ldots, \chi_{NT}]^{T} \quad \text{with} \quad \{\chi, \widehat{\chi}\} \in \left\{\{\mathrm{x}, \widehat{\mathrm{x}}\}, \{\mathrm{y}, \widehat{\mathrm{y}}\}, \left\{\widetilde{\mathrm{z}}, \widehat{\widetilde{\mathrm{z}}}\right\}, \{\mathrm{u}, \widehat{\mathrm{u}}\}\right\}.$$

This gives the following sequence of difference equations for $j = 1, \ldots, NT$,

$$\widehat{\mathrm{y}}_j - \mathbb{E}\left[\widehat{\mathrm{y}}_j | \mathrm{x}_j = x, \widetilde{\mathrm{z}}_j = z\right] = \overbrace{\left(\widehat{\mathrm{x}}_j - \mathbb{E}\left[\widehat{\mathrm{x}}_j | \mathrm{x}_j = x, \widetilde{\mathrm{z}}_j = z\right]\right)}^{\widehat{\mathrm{x}}_j - \mathbb{E}[\widehat{\mathrm{x}}_j | \mathrm{x}_j = x]}\beta + \widehat{\mathrm{u}}_j - \mathbb{E}\left[\widehat{\mathrm{u}}_j | \mathrm{x}_j = x, \widetilde{\mathrm{z}}_j = z\right]$$

It is important to note that $\widehat{\mathrm{x}}_j$ is not a deterministic function of x_j. This is so because $\widehat{\mathrm{x}}_j$ depends on the entire matrix $[\mathrm{x}_1, \ldots, \mathrm{x}_{NT}]^T$ through the equation $\widehat{\boldsymbol{X}} = \boldsymbol{P}\boldsymbol{X}$. Consequently, generally $\widehat{\mathrm{x}}_j \neq \mathbb{E}\left[\widehat{\mathrm{x}}_j | \mathrm{x}_j = x\right]$. To conclude, $\boldsymbol{P}\boldsymbol{X}$ is not absorbed when $\boldsymbol{M}$ is canceled out.