Advanced Applied Econometrics
Final paper

[redacted]

Introduction

This paper explores econometric techniques for panel and cross-sectional data. The author's prior thesis work's [redacted] dataset is reprised, a panel of S&P 500 firms over a 10-year period. The thesis was done on the interplay between R&D, cash holdings and firm value, and as such extends on the literature of valuation of cash. As there is ample evidence that a strong relation exists between these three variables, they present a unique opportunity to apply and reflect on a variety of econometric approaches in the context of corporate finance variables.

Literature

There is extensive literature linking firms' cash holdings, R&D activity, and market value, often analyzed through the lens of agency theory and financing constraints.

Firms often maintain cash reserves to finance investments when external capital is costly or unavailable (Opler et al., 1999). This precautionary motive is especially relevant for R&D-intensive firms, as innovation projects are risky and require continuous funding to avoid costly interruptions (Brown & Petersen, 2011). Studies show that greater cash holdings can facilitate sustained R&D activity, enabling firms to seize growth opportunities and buffer against market downturns/credit market risk (Bates et al., 2018). These findings are consistent with the view that greater cash holdings enable firms to undertake more R&D. Prior research also examines the "value of cash," emphasizing that higher cash holdings can signal reduced financial constraints and enhance firm value (Opler et al., 1999; Bates et al., 2018). Cash provides flexibility for future investments and serves as a buffer against shocks, which investors may price positively.

A measure that is often interpreted as a forward-looking indicator of growth opportunities is Tobin's Q, which reflects investors' valuation of a firm's assets relative to their replacement cost (Hall, 1999). In this paper, an approximate Q is used (firm market capitalization plus its long term debt and current liabilities, divided by the book value of its total assets), which is in accordance with Chung and Pruitt (1994) who derive this as an empirical version that captures at least 96% of the variability of the original Q.

These findings motivate two hypotheses: (i) higher firm value is associated with greater R&D intensity, and (ii) higher cash holdings are associated with higher firm value. Hypothesis 1 is analysed by focusing on the cross sectional data of 2018, while the second hypothesis uses the full panel structure.

Data

The dataset and variable construction are largely based on my earlier thesis work [redacted] 2023), which used S&P 500 firm data over a 10-year period (2010–2019). The data is extracted from Orbis Global for 498 firms.

In line with Faulkender and Wang (2006), financial firms are removed to ensure consistency, as their capital structures and regulatory environments differ significantly from those of non-financial firms, making direct comparisons problematic. The final sample includes 424 firms.

Variables:

The variables used in this analysis are based on prior research in corporate finance, particularly the literature on cash holdings, firm valuation, and innovation intensity (e.g., Opler et al., 1999; Saddour, 2006; Booth & Zhou, 2013). The following section outlines the construction of key variables.

R&D intensity (RD) is calculated as R&D expenditures over turnover. **Firm value** (Tobin's Q, also proxied as growth opportunities) combines market capitalization and liabilities over total assets. **Liquidity** (CSH) is cash and cash equivalents over total assets, and **profitability** (PROF) is return on assets (net income over total assets). **Operational risk** (RISK) is the standard deviation of a firm's ROA over time. **Sales growth** (GROWTH) is the year-on-year change in turnover, **size** (SIZE) is proxied by total assets, **leverage** (LEVERAGE) is total liabilities over assets, and **payout** (PAYOUT) is dividend yield. Sales growth is different from Q in that it doesn't capture future growth expectations.

All variables are winsorized at the 1st and 99th percentiles to reduce outlier effects. Selected variables, such as Q cash and size, are log-transformed to reduce skewness. These processed variables are used in the final analysis.

Results

Tobit cross sectional model

(1) Censored regression - Tobit

As R&D intensity is heavily zero inflated (185 zero-values out of 433 observations), the decision was made to not just utilize OLS to estimate the effect of Cash, but to employ censored regression. What we therefore observe is either Y=0 or Y=Y* if Y*>0 for the model Y*=x'b + e. Employing the Tobit model with left censoring allows us to account for the zero-inflation. Accounting for the left censoring is quite important, because we have essentially a mix of probit and linear regression where the expected value of Y conditional on y > 0 has an additional term that includes sigma. This term (the expected value of the error term for the truncation of the distribution) has to be accounted for and wouldn't be if simple OLS would be utilized, resulting in sample selection bias (non-random truncation of the dependent variable).

I begin by manually coding the Tobit model's Maximum Likelihood function, as discussed in the lecture, to gain a deeper understanding of the estimation process and to incorporate an exponential function for the **heteroskedastic** variance term (since sigma appears in the Tobit log-likelihood).

Two models are subsequently specified, a heteroskedastic (where the variance is assumed to vary for each variable) and a homoscedastic model (where constant variance is assumed). A likelihood ratio (LR) test is then performed to check for heteroskedasticity, and with a Chi-square statistics of 29, the hypothesis that the error variance is constant is rejected, concluding that there is evidence of heteroskedasticity in the Tobit model. Only Q (growth opportunities) and Operational risk were statistically significant in the Tobhet: component. Therefore, I re-specified the model to include only these two variables in the heteroskedasticity equation for a more parsimonious specification.

For further testing and interpretation of marginal effects, the Intreg command is used in Stata, where likewise a het(Q, Risk) component is added to model the heteroskedasticity. The second big assumption for ML-based models like Tobit is that of **normality**. The Lagrange multiplier test (LM test) is applied in line with Verbeek (2017) and a LM statistic of 13 is obtained, so the normality assumption is violated. It is noted, however, that this test works best asymptotically, and the dataset only counts 424 observations. While the coefficient estimates may remain consistent under correct mean specification, efficiency is lost and

standard errors/p-values may be misleading. A further step would be to test semi parametric models, but this is out of the scope of this paper.

The **marginal effect interpretation** most relevant for the economic question is that of the marginal change in Y conditional on X (E(Y|X)), which I calculate through the margins command (refer to **table 1**). The AME of cash holdings (log of cash/total assets, winsorized at 1) on expected observed R&D is positive and highly significant (0.0144, SE 0.0018, $p < 0.001$), indicating that firms with greater cash resources spend more on R&D on average. Because the variable is in logs, a 10% increase in cash is associated with an increase in expected R&D intensity of about 0.0014 units, ceteris paribus.

I further look at the Probit type marginal effects, i.e., the probability of observing any R&D conditional on the covariates. It is again noted that the AME of cash holdings on the probability of observing R&D is positive and highly significant (0.1230, SE 0.0127, $p < 0.001$). A 10% increase in cash holdings is on average expected to increase the probability of observing R&D by 12.3 percentage points, ceteris paribus.

(2) Sample selection model (Heckman)

I performed a Cragg test to evaluate whether the Tobit model's assumption of a single process governing both the decision to engage in R&D and the level of R&D spending holds. The test compares the Tobit specification to a less restrictive two-part model. The Tobit model assumes that the same latent process, with identical coefficients, determines both (i) the probability of undertaking any R&D and (ii) the conditional amount of R&D given it is positive. In contrast, the two-part model allows these two stages to have different determinants. Three models are estimated, 1) a probit model for the probability of reporting any R&D; (2) a truncated regression for the amount of R&D among firms with positive R&D; and (3) a Tobit model on the full sample. The resulting p-value ($p>0.001$, $df = 9$) strongly rejects the assumption of a single process, indicating that a two-part model provides a significantly better fit for the data.

However, as the variable set of the first stage involves the same variables as that of the second stage, we may face an identification problem due to quasi-linearity of the Mills ratio, which likely results in inconsistent results (Puhani, 2000). Given the context of the economic question, no variable (exclusion restriction) is available that determines the probability of doing R&D that doesn't also determine the amount. Therefore, I do not estimate the Heckman model here and retain the Tobit (I) model as the focal specification, while acknowledging that the assumption of a single latent process is violated.

(3) OLS Comparison (Angrist and Pischke, 2009, p.94)

Angrist and Pischke (2009, p.78) argue that in the case of limited dependent variables, of which censored regression is an example, the classical Tobit framework is only appropriate when the zeros in the dependent variable are the result of censoring from a latent continuous process. When zeros are real observed outcomes rather than censored values, as is the case for R&D, the Tobit model's assumptions about the data-generating process are likely violated. They therefore recommend estimating the relationship using OLS, which treats zeros as valid data points, requires weaker assumptions, and avoids the strong distributional restrictions inherent in Tobit models. As the Cragg test rejects the standard Tobit assumption of a single latent process and the normality assumption is also violated, it is only prudent to follow Angrist and Pischke's (2009) advice to compare results with an OLS estimation. OLS treats zeros as genuine outcomes instead of censored values and under weaker assumptions yields consistent and unbiased estimates of the best linear approximation to $E[y|x]$, that are in practice close to the results obtained through marginal effects for a probit or Tobit model. This provides a robust benchmark against which to assess the results of the Tobit marginal effects.

In line with this recommendation, **Table 1** reports the marginal effects from the censored regression alongside the OLS coefficients. The estimated effects from OLS are generally similar in sign and significance to the Tobit marginal effects for $E[Y|X]$, suggesting that treating the zeros as genuine observations does not materially alter the main inferences. The magnitude furthermore doesn't diverge much from the censored regression marginal effects (1a). As OLS provides consistent estimators under weaker assumptions, and various assumptions were violated for the Tobit model (single latent process, normality), these results provide a useful point of comparison for interpreting the magnitude and direction of effects, which helps to understand the practical impact of deviations from the Tobit model's assumptions. The OLS model has an R-squared of 46%, meaning that the covariates are able to explain up to 46% of the variance in R&D. This is moderate, but not surprising given the complexity of the process that drives firms to engage in R&D. A pseudo R-squared could also be calculated for the Tobit model, but as it does not measure the proportion of variance explained and has no direct interpretation as a goodness-of-fit statistic, it is omitted from the discussion.

Table 1: Regression results

VARIABLES	(1) Censored regression	(1a) Marginal Effects E(Y|X)	(1b) Marginal Effects Pr(Y>0)	(2) OLS Coefficients
Tobin's Q	0.061***	0.042***	0.260***	0.064***
	(0.011)	(0.006)	(0.047)	(0.008)
Cash	0.027***	0.014***	0.123***	0.014***
	(0.003)	(0.002)	(0.013)	(0.002)
Operational Risk	0.356**	0.270***	1.494***	0.622***
	(0.166)	(0.089)	(0.736)	(0.161)
Growth	-0.015	-0.008	-0.069	-0.034
	(0.024)	(0.013)	(0.107)	(0.021)
Leverage	-0.108***	-0.058***	-0.492***	-0.092***
	(0.021)	(0.011)	(0.095)	(0.016)
Size	0.014***	0.008***	0.065***	0.012***
	(0.004)	(0.002)	(0.017)	(0.003)
Profitability	-0.369***	-0.196***	-1.674***	-0.341***
	(0.078)	(0.041)	(0.341)	(0.079)
Constant	-0.196**			-0.171**
	(0.096)			(0.074)
Observations	380	380	380	380

Note. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For the Tobit model, heteroskedasticity is modelled for Q and Risk through the command het(). Payout is left out of the regression.

(4) Discussion

A key econometric concern that remains is that of **endogeneity**: cash holdings may be correlated with unobserved firm characteristics that also affect R&D activity, such as managerial ability, innovation culture, or access to credit markets. The resulting **omitted variable bias** could render the estimated coefficients inconsistent. Moreover, **reverse causality** cannot be ruled out. Firms undertaking R&D projects may deliberately maintain larger cash reserves to finance uncertain innovation expenditure, creating simultaneity between cash and R&D. This reverse channel would also bias estimates in the absence of valid instruments.

A standard approach to address endogeneity is **two-stage least squares (2SLS)**. However, no suitable instrumental variables are available that are strongly correlated with cash holdings yet uncorrelated with the R&D error term. In case a valid IV regressor would be present, a Hausman test would be used to formally test whether endogeneity is present.

Overall, while the Tobit model accounts for censoring, the violation of its key assumptions (single latent process, normality) and the potential for endogeneity mean that its coefficient estimates should be interpreted with caution. The OLS specification is free of the Tobit distributional assumptions and provides a consistent estimate of the best linear approximation

to E[Y|X] under weaker conditions. Given the similarity in signs, significance, and magnitudes between the OLS and Tobit marginal effects for E[Y|X], the OLS results are concluded to me more credible, and in this context may better reflect the true underlying relationship.

Dynamic panel

To test the second hypothesis, a dynamic model is constructed. Tobin's Q is inherently persistent over time because it incorporates expectations about future profitability and investment opportunities. Ignoring this persistence can bias the estimated impact if past Q influences current Q. This decision is further supported by a correlation between Q and its lag of 0.94 and a significant coefficient in a static fixed effects model for the lagged Q. I first start by estimating an Arellano–Bond difference GMM, and further verify the results by also estimating with reduced instruments as well as through system GMM. Instrumental variables with Anderson-Hsiao estimation are left out of the analysis, as the AH-IV isn't efficient (not all information is used) and there are sufficient periods (10) available to explore GMM.

Variable operational risk is left out of the further analyses as its construction leads to a within variation of almost 0, leading to the variable being omitted from the panel regression output. DY is again omitted to preserve observations.

I estimate a two-step AB difference GMM with robust Windmeijer-corrected standard errors. Two-step GMM is used because heteroskedasticity across firms is likely as the panel spans diverse firm sizes and industries, thereby violating the homoscedasticity assumption of one-step GMM. Variable assumptions for deriving appropriate instruments are made primarily based on economic reasoning and later reflected on using specification tests:
Endogenous variables include cash, growth, leverage and profitability. Firms can adjust cash reserves in response valuation or investment opportunities. For sales growth, shocks in demand may raise both growth and valuation within the same year. Capital structure adjusts in response to current conditions and valuation (leverage). Lastly, profitability is realized within the period, the same shocks that lift Q often lift profits.
Predetermined variables include R&D intensity and Size. R&D budgets are typically set at the start of the year based on prior conditions, unlikely to react to same-year Q shocks. Size is a slow-moving state variable driven by accumulated investment
Year is presumed to be **strictly exogenous** and assumed to be uncorrelated with firm-specific error terms.

Table 2 reports estimates from four dynamic panel GMM models. The first two columns use the difference GMM estimator, while the last two apply the system GMM estimator. All models

include year fixed effects. Lag ranges include 1–9 (max) for predetermined and 2–9 for endogenous for Model (1), and 1–2 and 2–3 for Model (2). Model (3) incorporates additional moment conditions by using system GMM with lag ranges 2–4 for endogenous regressors and the maximum available for predetermined ones, while Model (4) applies the same tighter lag restrictions as Model (2)

The key econometric difference is that system-GMM augments the difference equation with an additional level equation, instrumented using lagged differences. This improves efficiency when the dependent variable is highly persistent, but requires an extra identifying assumption: that changes in the regressors are uncorrelated with firm-specific effects. For cash and size, this assumption is likely reasonable, as year-to-year fluctuations primarily reflect transitory shocks rather than fixed characteristics. For R&D, it is more debatable, since a persistent innovation culture could affect both levels and changes. If this assumption fails, system GMM becomes inconsistent.

Table 2: Panel GMM estimation overview

Tobin's Q	(1) A-B GMM	(2) A-B GMM	(3) System GMM	(4) System GMM
Tobin's Q (t-1)	0.151**	0.032	0.640***	0.587***
	(0.062)	(0.069)	(0.036)	(0.053)
Cash	0.022	0.027	0.037***	0.031*
	(0.017)	(0.029)	(0.013)	(0.018)
R&D intensity	-0.796	-3.573**	0.4*	0.460
	(0.759)	-1.722	(0.241)	(0.302)
Leverage	0.036	0.159	0.312***	0.391***
	(0.125)	(0.183)	(0.080)	(0.099)
Size	-0.240***	-0.194**	-0.143**	-0.149***
	(0.048)	(0.084)	(0.025)	(0.035)
Growth	0.008	0.032	-0.144***	-0.159**
	(0.066)	(0.080)	(0.067)	(0.080)
Profitability	1.034***	0.993**	0.782***	1.123***
	(0.288)	(0.393)	(0.782)	(0.309)
Groups	380	380	384	384
Instruments	260	99	263	157
AR(1) p-value	0.000	0.001	0.000	0.000
AR(2) p-value	0.748	0.395	0.938	0.961
Hansen p-value	0.108	0.004	0.011	0.000

Note. Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. AR(2): Arellano-Bond test for second-order autocorrelation in first differences. Hansen: test of overidentifying restrictions. All models are two-step GMM. Year dummies are omitted from table.

The results reflect these theoretical considerations. Difference GMM produces near-zero coefficients on the lagged dependent variable, consistent with weak instruments for highly persistent series. System GMM yields larger and significant lag coefficients, as expected,

illustrating the efficiency gain from additional moment conditions. Hansen tests show mixed results: acceptable in Models (1), but very low in (2), (3) and (4), indicating possible instrument validity concerns despite restricting lag ranges. AR(2) tests confirm no second-order autocorrelation for all specifications, supporting the dynamic specification. Because the Hansen J-test rejects instrument validity in most GMM specifications, the estimates may be biased and inconsistent rather than merely inefficient, so the results should be interpreted with caution.

Conclusion

This paper applied multiple econometric techniques to investigate the relationships among cash holdings, R&D intensity, and firm value, using both cross-sectional and panel data.

For the cross-sectional analysis of R&D intensity, the Tobit model was chosen to address zero-inflation. However, diagnostics challenge key Tobit assumptions: the normality test was rejected, and the Cragg test suggests that the decision to undertake R&D and the intensity level are driven by distinct processes. In line with Angrist and Pischke (2009), OLS was estimated as a robustness check, showing similar signs and magnitudes to Tobit marginal effects. Given the weaker distributional assumptions and the context, this suggests OLS provides a more reliable approximation.

For the dynamic panel analysis of Tobin's Q, the application of difference and system GMM addresses endogeneity and persistence in firm value. As expected, difference GMM produced weak lag effects, while system GMM improved efficiency by adding level equations. The lagged dependent variable remained significant in system GMM, confirming strong persistence in Tobin's Q. While the Hansen tests cast doubt on instrument validity, the stability of coefficients across specifications and the absence of AR(2) suggest the results should be interpreted as indicative rather than definitive.

Despite Hansen tests casting doubt on instrument validity in some GMM specifications, the estimator remains valuable because it addresses endogeneity and dynamic bias better than static methods. The comparison between Tobit, OLS, and GMM illustrates how estimator assumptions shape inference: while Tobit accounts for censoring, assumption violations shifted credibility toward OLS. Dynamic GMM revealed that firm value shows strong persistence over time and is influenced by feedback from its own past values, although the specification tests suggest these results should be interpreted cautiously. This paper

underscores the fact that econometric inference is heavily influenced by assumptions and adequate robustness checks should be performed.

References

Faulkender, M., & Wang, R. (2006). Corporate Financial Policy and the Value of Cash. The Journal of Finance, 61(4), 1957–1990. https://doi.org/10.1111/j.1540-6261.2006.00894.x

Chung, K. H., & Pruitt, S. W. (1994). A simple approximation of Tobin's q. *Financial management*, 70-74.

Verbeek, M. (2017). *A guide to modern econometrics*. John Wiley & Sons.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of economic surveys*, *14*(1), 53-68.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.