Financial Econometrics – Cross Section and Panel Data

Miscellaneous Advice & Thoughts on Empirical Work

Andreas Fuster
Swiss Finance Institute @ EPFL

SFI Léman PhD program – 2025/6

s:fi

1. Things we did not cover
2. Log(y+1)
3. Outliers and influential observations
4. Researcher degrees of freedom
5. AI AI AI
6. Miscellaneous resources

Methods we did not cover (but that you may encounter / may want to use)

- Matching methods. Idea: if don't have exogenous variation in "treatment", at least make treated and control groups as similar as possible along observable dimensions.
  - propensity score matching (PSM) – used primarily for binary treatments
  - coarsened exact matching (CEM)
  - matching based on "distance" of X variables
  - see Roberts-Whited section 6; Verbeek section 7.2.4
- Downsides/limitations:
  - still not a "real" identification strategy – but can be combined with others (as seen in papers we discussed)
  - worry that worsen balance on unobservables
  - generally feels a bit ad-hoc – which variables used for matching, how many matched observations per treated unit, etc.

Methods we did not cover (but that you may encounter / may want to use)

s:fi

- Models for binary or discrete outcomes (logit, probit, tobit, ordered/multinomial logit, count data models [Poisson etc.])
- Selection models (Heckman)
- Duration/survival models (e.g. Cox model)
  - all of these are at least briefly discussed in Verbeek Ch. 6
- Bunching methods (see Kleven 2016, Annual Review of Econ)
- Local projections – useful to trace out dynamic effects (impulse responses) of a "shock"
  - primarily used in macro, where they have almost replaced VARs, but also seen in finance papers
  - see https://sites.google.com/site/oscarjorda/home/local-projections (including review article forthcoming in JEL)

Log(y + 1)

- Very common in corporate finance type settings: outcome variable that has many zeros & some positive values – e.g.
  - number of patents filed by a firm *f* in year *t*
  - total loan $ made by a bank *b* to a firm *f* in year *t*
- The positive values are often very skewed, so would like to use log tranformation – but can't do that with y = 0
  - also would like to keep semi-elasticity interpretation of coeffs
- So researchers very commonly use log(y + c) for c > 0 (often c = 1)
- Seems innocuous ("0 and 1 are close")…

Log(y + 1)

s:fi

- Finally somebody took a careful look at this issue:

Journal of Financial Economics
Volume 146, Issue 2, November 2022, Pages 529-551

Count (and count-like) data in finance

Jonathan B. Cohn [a] ✉, Zack Liu [b] ✉, Malcolm I. Wardlaw [c] ✉

Abstract

This paper assesses different econometric approaches to working with count-based outcome variables and other outcomes with similar distributions, which are increasingly common in corporate finance applications. We demonstrate that the common practice of estimating linear regressions of the log of 1 plus the outcome produces estimates with no natural interpretation that can have the wrong sign in expectation. In contrast, a simple fixed-effects Poisson model produces consistent and reasonably efficient estimates under more general conditions than commonly assumed. We also show through replication of existing papers that economic conclusions can be highly sensitive to the regression model employed.

(see also Chen & Roth, Logs with Zeros? Some Problems and Solutions, *QJE* 2024)

Log(y + 1)

- Clearly written paper, delivers 12 takeaways. E.g.
  - "Log1plus regression coefficients are not interpretable as semi-elasticities of the outcome variable, nor can any economically meaningful relationship between the outcome variable and a covariate be recovered from a log1plus regression coefficient."
  - "Log1plus regression is almost certain to suffer from two forms of bias that make even the sign of a relationship difficult to infer from log1plus regression coefficients."
  - these results also hold when using the inverse hyperbolic sine (IHS) transformation, a common alternative to log(1+y)

- Authors strongly advocate use of **Poisson regression** models instead, which can also be estimated with high-dimensional fixed effects
  - ppmlhdfe in Stata; glmhdfe in R

Outliers and influential observations (cf. Verbeek 2021, Section 4.1)

- Many papers in finance have to deal with the potential importance of "outliers" (unusual data points) for their results
- Common approaches: univariate trimming or winsorizing
  - e.g. winsorizing at 1st and 99th percentile of continuous variables
- This is ad-hoc but widely accepted. Would recommend at least studying robustness to different choices (e.g. winsorizing vs. trimming; different cut-offs)
  - also, consider not just sensitivity to individual obs. but entire units – e.g. if use state-level variation, drop one state at a time to assess robustness
- Adams et al. (2019) make the point that univariate trimming/winsorizing may not eliminate points with high leverage / influence on the results
- And recommend use of alternative "MM"-estimators that are more outlier-robust & can also handle many fixed effects
  - not commonly used yet to my knowledge

Researcher degrees of freedom

- Highly recommended (slightly depressing) reading: Mitton (RFS 2022)

I document large variation in empirical methodology in corporate finance regressions in top finance journals. Although methodological variation allows for customization of empirical tests to fit specific theories, it can also enable excessive reporting of statistically significant results. For example, given discretion over 10 routine methodological decisions, a researcher could report that over 70% of randomly generated variables are statistically significant determinants of leverage at the 5% level. The methodological decisions that affect statistical significance the most are dependent variable selection, variable transformation, and outlier treatment. I discuss remedies that can mitigate the negative effects of methodological variation. (*JEL* C18, C52, G30)

- He focuses on typical corporate finance regressions (leverage, profitability etc.) but things would certainly be similar e.g. in financial intermediation papers

Table 4
Current practice in empirical corporate finance: Other methodological decisions

	Profitability	Value	Leverage	Investment	Payout	Cash	ALL	ALL (2016–18)
A. Industry inclusion								
All	60%	52%	38%	39%	46%	30%	46%	34%
All except financial and utility	16%	28%	33%	34%	37%	46%	30%	34%
All except financial	13%	15%	17%	16%	14%	15%	15%	14%
Manufacturing only	3%	2%	6%	7%	1%	6%	4%	7%
Other	8%	4%	6%	5%	2%	3%	5%	11%
B. Key explanatory variable form								
Continuous—not logged	54%	59%	54%	56%	58%	51%	55%	51%
Dummy—naturally occurring	23%	20%	26%	25%	24%	32%	25%	30%
Dummy—created from continuous	15%	15%	13%	14%	16%	14%	14%	13%
Continuous—logged	8%	6%	7%	5%	2%	4%	6%	5%
C. Lag on explanatory variable								
Contemporaneous	58%	63%	65%	61%	67%	66%	62%	61%
Lagged	32%	19%	24%	30%	23%	16%	26%	30%
Both	4%	3%	3%	5%	3%	5%	4%	2%
Unclear	7%	15%	7%	4%	7%	13%	8%	7%
D. Outlier treatment								
Winsorize	49%	38%	48%	50%	49%	56%	48%	62%
Retain	43%	49%	43%	40%	47%	36%	43%	34%
Trim	8%	13%	9%	9%	4%	8%	9%	4%
E. Outlier cutoffs								
1st/99tth	74%	65%	75%	78%	82%	82%	75%	79%
5th/95th	9%	15%	5%	7%	6%	7%	8%	9%
0.5th/99.5th	5%	4%	8%	6%	2%	7%	6%	5%
2.5th/97.5th	3%	1%	2%	3%	4%	4%	3%	1%
2nd/98th	3%	2%	1%	0%	2%	0%	1%	1%
3rd/97th	1%	0%	1%	1%	0%	0%	1%	1%
10th/90th	1%	1%	0%	1%	0%	0%	1%	0%
Other/Not specified	4%	12%	8%	4%	4%	0%	6%	4%
F. Dependent variable form								
Continuous—not logged	97%	75%	91%	90%	73%	82%	87%	89%
Continuous—logged	2%	25%	8%	9%	4%	18%	10%	10%
Dummy	0%	0%	1%	2%	22%	0%	3%	0%
G. Denominator on flow/stock dependent variables								
End of year	70%	89%	40%	50%	76%	55%	62%	59%
Beginning of year	21%	11%	56%	49%	24%	41%	33%	37%
Averaged	9%	0%	4%	1%	0%	5%	4%	3%
H. Industry dummy definition								
2-digit SIC	24%	23%	29%	23%	19%	29%	25%	17%
Fama-French	12%	19%	21%	16%	14%	12%	16%	20%
3-digit SIC	8%	5%	13%	12%	12%	10%	10%	12%
1-digit SIC	3%	4%	4%	3%	7%	2%	4%	2%
4-digit SIC	2%	5%	5%	2%	5%	2%	4%	4%
NAICS	1%	0%	4%	2%	2%	0%	2%	2%
Other	8%	8%	6%	7%	7%	14%	8%	10%
Not specified	41%	38%	18%	34%	33%	31%	32%	34%

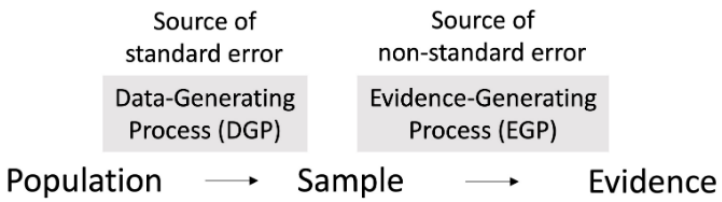
- Proposed remedies:
  - robustness checks – but: double-edged sword (can get false negatives, plus researchers often have freedom in what they report)
  - **specification checks**: graphically show distribution of results across all possible permutations of choices
  - more emphasis on economic significance

Recent example from asset pricing: "Non-standard errors" (A. Menkveld et al. JF 2024)

Abstract

In statistics, samples are drawn from a population in a data-generating process (DGP). Standard errors measure the uncertainty in sample estimates of population parameters. In science, evidence is generated to test hypotheses in an evidence-generating process (EGP). We claim that EGP variation across researchers adds uncertainty: *non-standard errors*. To study them, we let 164 teams test six hypotheses on the same sample. We find that non-standard errors are sizeable, on par with standard errors. Their size (i) co-varies only weakly with team merits, reproducibility, or peer rating, (ii) declines significantly after peer-feedback, and (iii) is underestimated by participants.

Academic research recognizes randomness in data samples by computing standard errors (SEs) for parameter estimates. It, however, does *not* recognize the randomness that is in the research process itself. We believe that such randomness is the cause of, what we will call, non-standard errors (NSEs).



- 364 authors from 34 countries and 207 institutions (incl. HEC, EPFL, Geneva)
- Data: 720 million trade records of EuroStoxx 50 index futures, spanning 17 years
- 5 hypotheses to be tested (e.g. "Market efficiency has not changed over time")
- Massive dispersion in point estimates (NSE > avg. SE), esp. if hypothesis not very concretely articulated
- Hard to predict with characteristics
- Participants underestimate dispersion

Pre-analysis plans – may be a good idea to "tie one's hands" in some applications

s:fi

- Example: Brown et al. (JME 2022), "The Convenience of Electronic Payments and Consumer Cash Demand", https://ssrn.com/abstract=3582388 / https://doi.org/10.1016/j.jmoneco.2022.06.001

> Our analysis follows a pre-analysis plan which has been registered and time-stamped at https://osf.io/scvbq/ before data delivery. In this plan we have pre-specified the hypotheses, the data cleaning and sample selection, the definition of outcome and explanatory variables, the econometric specification and statistical inference (Olken 2015). The use of a pre-analysis plan intends to eliminate biases arising from model selection as well as from the selective reporting of findings and should thus strengthen the credibility of results, in particular for proprietary data (Casey et al. 2012; Coffmann and Niederle, 2015). While pre-analysis plans are common in randomized control trials, they are much less frequent in studies using observational data (Burlig 2018). We are unaware of other papers in monetary economics which employ proprietary, observational data and are based on a pre-analysis plan.

How will AI change the research process?

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5060022,
https://github.com/velikov-mihail/AI-Powered-Scholarship

Mihail Velikov · 2nd
Assistant Professor of Finance at Penn State University
1d · Edited ·

+ Follow ···

An academic paper has excellent empirical evidence & hypotheses that perfectly match the patterns in the data.

One catch: AI wrote the hypotheses after seeing the results.

Should this matter?

New paper w/ Robert Novy-Marx on AI-Powered (Finance) Scholarship:
https://lnkd.in/eDfw6QCW

To assess this question we:

1️⃣ Mined 30K+ potential stock return predictors
2️⃣ Validated 96 robust signals using our "Assaying Anomalies" protocol
3️⃣ Used LLMs to generate 3 versions of complete papers with different hypotheses for each signal

The generated papers & code are available at:
https://lnkd.in/eCR-_utW

The papers are remarkably coherent - they include creative names for the signals, contain custom introductions providing different hypotheses for the observed predictability patterns, and incorporate citations to existing (and, on occasion, imagined) literature.

Key implication: When AI can rapidly produce plausible hypotheses for any empirical finding at unprecedented scale, how do we ensure quality control in academic research? In the paper we raise further questions about research integrity and evaluation that reflect the realities of AI-enabled research production and give some initial thoughts on ways to address those.

AI-Powered (Finance) Scholarship*

Robert Novy-Marx† Mihail Velikov‡

December 16, 2024

Abstract

This paper describes a process for automatically generating academic finance papers using large language models (LLMs). It demonstrates the process' efficacy by producing hundreds of complete papers on stock return predictability, a topic particularly well-suited for our illustration. We first mine over 30,000 potential stock return predictor signals from accounting data, and apply the Novy-Marx and Velikov (2024) "Assaying Anomalies" protocol to generate standardized "template reports" for 96 signals that pass the protocol's rigorous criteria. Each report details a signal's performance predicting stock returns using a wide array of tests and benchmarks it to more than 200 other known anomalies. Finally, we use state-of-the-art LLMs to generate three distinct complete versions of academic papers for each signal. The different versions include creative names for the signals, contain custom introductions providing different theoretical justifications for the observed predictability patterns, and incorporate citations to existing (and, on occasion, imagined) literature supporting their respective claims. This experiment illustrates AI's potential for enhancing financial research efficiency, but also serves as a cautionary tale, illustrating how it can be abused to industrialize HARKing (Hypothesizing After Results are Known).

How will AI change the research process?

s:fi

Andrew Hall · 2nd
Prof @ Stanford GSB, Hoover | Studying technology, politics, an...
2d ·

Connect

Here's proof that Claude Code can now write an entire empirical polisci paper.

Over the weekend, I had Claude Code fully replicate and extend an old paper of mine estimating the effect of universal vote-by-mail on turnout and election outcome...essentially in one shot.

After careful prompting, Claude Code:

(1) Downloaded the old paper's repo and replicated the past results, translating our old Stata Code into Python

(2) Crawled the web to get updated official election data and census data

(3) Ran new analyses extending the results through 2024

(4) Created new tables and figures

(5) Performed a lit review

(6) Wrote a wholly new paper

(7) Pushed the whole thing to a new github repo

The whole thing took about an hour.

This is an insane paradigm shift in how empirical work is done. It's obviously going to massively reshape academic research, but the implications go much broader---we can now direct AI to carry out useful quantitative research on our behalf. People and organizations should be able to learn far more far faster, if they're ready to make the jump.

Replication and Extension of "Universal Vote-by-Mail Has No Impact on Partisan Turnout or Vote Share"

Claude Code

Andrew B. Hall*

January 3, 2026

Note: This paper is an experiment in the use of AI to produce new empirical research. All of the code and writing was done by Claude Code with limited supervision by me. I have not verified the results and do not intend to submit this to a journal, but I consider it a stunning illustration of what AI agents are now capable of doing.

Abstract

We replicate and extend Thompson et al. (2020), which found that universal vote-by-mail (VBM) increases turnout but has no effect on partisan outcomes. Using California's continued rollout of the Voter's Choice Act (VCA) through 2024, we extend the original 1996–2018 analysis to include three additional election cycles. Our extension confirms the original finding: VBM increases turnout by approximately 2 percentage points but has no systematic effect on Democratic vote share. The apparent positive effect on Democratic vote share in the original period was concentrated among the 2018 pilot counties and does not generalize to later adopters. Population-weighted estimates

Miscellaneous resources

- How to make nice Beamer slides:
  - https://github.com/paulgp/beamer-tips
  - https://github.com/kylebutts/templates/tree/master/latex-slides
- More on coding style & organization (mostly Stata):
  - https://github.com/skhiggins/Stata_guide
  - https://github.com/michaelstepner/healthinequality-code/blob/main/code/readme.md
  - https://julianreif.com/guide/
- Stata – R – Python (etc.) “translations”:
  - https://lost-stats.github.io/
  - https://stata2r.github.io/
  - or ask ChatGPT / Claude / …

Miscellaneous resources

s:fi

- "Idea generation":
  - AFA 2023 panel: https://youtu.be/UkPoAktls14?si=7YMr0NeqAdLDVXT9
- Writing:
  - John Cochrane's writing tips for PhD students: https://static1.squarespace.com/static/5e6033a4ea02d801f37e15bb/t/5eda74919c44fa5f87452697/1591374993570/phd_paper_writing.pdf
  - Jesse Shapiro (very short): https://scholar.harvard.edu/files/shapiro/files/foursteps.pdf
  - McCloskey, "Economical Writing" (book – some say it is life changing, but I haven't read it): https://www.amazon.com/Economical-Writing-Deirdre-McCloskey/dp/1577660633
  - More links to more things: https://sites.google.com/site/amandayagan/writingadvice

Miscellaneous resources

- Conferences – check programs to see what people are working on. Best ones in my opinion:
  - NBER – corporate finance, asset pricing, behavioral finance (spring/fall) and Summer Institute – see https://www.nber.org/conferences (often streamed on YouTube, you should watch, esp. Summer Institute !)
  - AFA (and AEA/ASSA): https://afajof.org/annual-meeting/ https://www.aeaweb.org/conference/
  - Western Finance Association (WFA) – https://westernfinance.org/
  - European Finance Association (EFA) – https://european-finance.org/r/annual-meeting
  - SFS Cavalcade – http://sfs.org/financecavalcades/sfs-cavalcade-north-america/
  - FIRS – https://firsociety.org/conference/