Publication Bias and the Cross-Section of Stock Returns

Andrew Y. Chen[1] **Tom Zimmermann**[2]

[1]Federal Reserve Board

[2]Quantco, Inc

AFA: 2018

Disclaimer: The views expressed herein are those of the author and do not necessarily reflect the position of the Board of Governors of the Federal Reserve or the Federal Reserve System

“The Lord of the p-value”

"The Lord of the p-value"

"The Lord of the p-value"

“The Lord of the p-value”

"The Lord of the p-value"

The Cross-Sectional Asset Pricing Lit

"The Lord of the p-value"

p-hacking

- data-mining, data-snooping
- suspicion and ambition
- collective re-use of data

The Cross-Sectional Asset Pricing Lit

"The Lord of the p-value"

p-hacking

- data-mining, data-snooping
- suspicion and ambition
- collective re-use of data

Journal Review

- robustness tests
- theoretical motivations
- supporting results
- a scientific, ethical culture

The Cross-Sectional Asset Pricing Lit

"The Lord of the p-value"

p-hacking

- data-mining, data-snooping
- suspicion and ambition
- collective re-use of data

Journal Review

- robustness tests
- theoretical motivations
- supporting results
- a scientific, ethical culture

The Cross-Sectional Asset Pricing Lit

Our Question: Which Side is Winning?

This Paper: A Focused, Structured Estimate of Who's Winning

This Paper: A Focused, Structured Estimate of Who's Winning

(1) Focus: replications of 172 **published cross-sectional predictors**

- Excludes non-predictive and aggregate factors in Harvey, Liu, Zhu 2016
- Excludes un-published predictors in Chordia, Goyal, Saretto 2017

This Paper: A Focused, Structured Estimate of Who's Winning

(1) Focus: replications of 172 **published cross-sectional predictors**

- Excludes non-predictive and aggregate factors in Harvey, Liu, Zhu 2016
- Excludes un-published predictors in Chordia, Goyal, Saretto 2017

(2) Structure: estimated model of biased publication

- Allows for **p-hacking** effects *and* **journal review**
- Unlike Hou, Xue, Zhang's 2017 informal approach

This Paper: A Focused, Structured Estimate of Who's Winning

(1) Focus: replications of 172 **published cross-sectional predictors**

- Excludes non-predictive and aggregate factors in Harvey, Liu, Zhu 2016
- Excludes un-published predictors in Chordia, Goyal, Saretto 2017

(2) Structure: estimated model of biased publication

- Allows for **p-hacking** effects *and* **journal review**
- Unlike Hou, Xue, Zhang's 2017 informal approach

Result:

- **Journal review** dominates. **Nearly all predictors were real!!**
  - Consistent w/ McLean-Pontiff 2016, Jacobs-Müller 2016, Yan-Zheng 2017

This Paper: A Focused, Structured Estimate of Who's Winning

Nearly all predictors were real!!

This Paper: A Focused, Structured Estimate of Who's Winning

Nearly all predictors were real!!

How can this be true??

This Paper: A Focused, Structured Estimate of Who's Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors,
    many in-sample returns will be large by pure chance

This Paper: A Focused, Structured Estimate of Who's Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors,
    many in-sample returns will be large by pure chance

  $\Rightarrow$ many predictors were fairy tales

This Paper: A Focused, Structured Estimate of Who's Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors,
    many in-sample returns will be large by pure chance

  $\Rightarrow$ many predictors were fairy tales
- Our **more structured** logic (James-Stein 1961, Efron-Morris 1973)

This Paper: A Focused, Structured Estimate of Who's Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors,
    many in-sample returns will be large by pure chance
  - $\Rightarrow$ many predictors were fairy tales
- Our **more structured** logic (James-Stein 1961, Efron-Morris 1973)
  - 172 predictors **tell us about the nature of the publication process**

This Paper: A Focused, Structured Estimate of Who's Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors,
    many in-sample returns will be large by pure chance
  - ⇒ many predictors were fairy tales
- Our **more structured** logic (James-Stein 1961, Efron-Morris 1973)
  - 172 predictors **tell us about the nature of the publication process**
  - They tell us that **journal review** dominates **p-hacking**

This Paper: A Focused, Structured Estimate of Who's Winning

Nearly all predictors were real!!

How can this be true??

- Standard logic (Bonferroni, Benjamini-Hochberg 1995)
  - After looking at 172+ predictors, **many in-sample returns will be large by pure chance**

  $\Rightarrow$ many predictors were fairy tales

- Our **more structured** logic (James-Stein 1961, Efron-Morris 1973)
  - 172 predictors **tell us about the nature of the publication process**
  - They tell us that **journal review** dominates **p-hacking**

  $\Rightarrow$ nearly all predictors were real.

Replications of 172 Published Predictors

Data: Replications of 172 Published Predictors

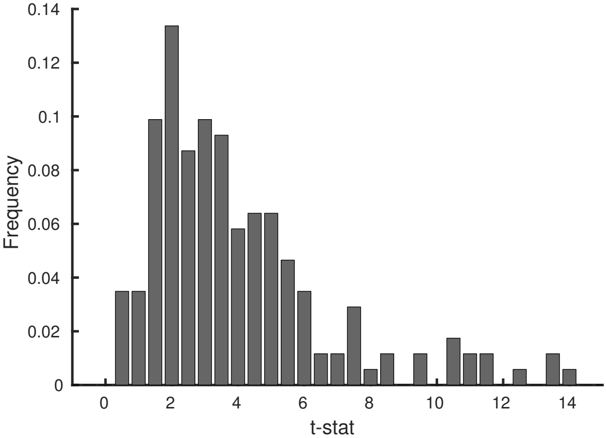
(1) Replicate McLean and Pontiff's (2016) 97 published cross-sectional predictors

(2) Replicate 75 additional variables that were

- **shown to predict cross-sectional returns**
- **published in "top-tier" journals**

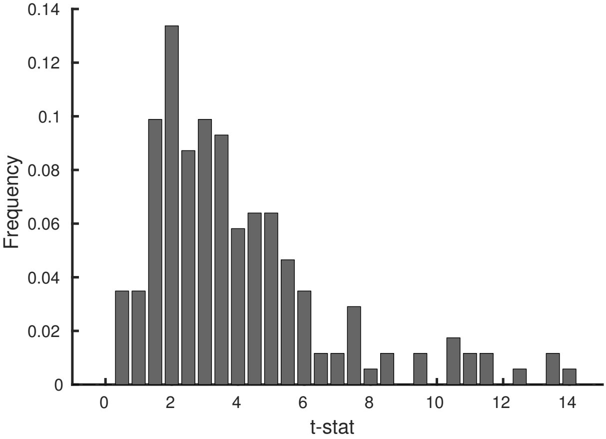
Data available at sites.google.com/site/chenandrewy/

Distribution of Replicated t-stats



- Sharp left shoulder $\Rightarrow$ strongly suggestive of **p-hacking**
- But what explains the long right tail?

Distribution of Replicated t-stats



- Sharp left shoulder ⇒ strongly suggestive of **p-hacking**
- But what explains the long right tail? ⇒ **need model**

Model and Estimation

A Statistical Model of Publication 1/2

Motivating Story:

1. Anything that might be published is submitted to journals
   - Allows for **p-hacking**
2. Only portfolios with "narratives" are considered for publication
   - Allows for **journal review**: robustness tests, supporting results, ...
3. Only narratives with high t-stats are published
   - Another **p-hacking** effect

A Statistical Model of Publication 1/2

Motivating Story:

1. Anything that might be published is submitted to journals
   - Allows for **p-hacking**
2. Only portfolios with "narratives" are considered for publication
   - Allows for **journal review**: robustness tests, supporting results, ...
3. Only narratives with high t-stats are published
   - Another **p-hacking** effect

$\Rightarrow$ statistical model of publication similar to Harvey, Liu, and Zhu's (2016) model with correlations

A Statistical Model of Publication 2/2

Key equations

- If portfolio i has a narrative,

  true return $\mu_i \sim$ scaled student's t with σ_μ, ν_μ

- **dispersion of true returns** σ_μ measures power of **journal review**
  - **large** $\sigma_\mu \Rightarrow$ **narratives find variation in true returns**

A Statistical Model of Publication 2/2

Key equations

- If portfolio i has a narrative,

  true return $\mu_i \sim$ scaled student's t with σ_μ, ν_μ

- **dispersion of true returns** σ_μ measures power of **journal review**
  - **large** $\sigma_\mu \Rightarrow$ **narratives find variation in true returns**

- In-sample returns are noisy and biased signals of μ_i

$$r_i = \mu_i + \epsilon_i$$

Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors

Maximum Likelihood Estimation

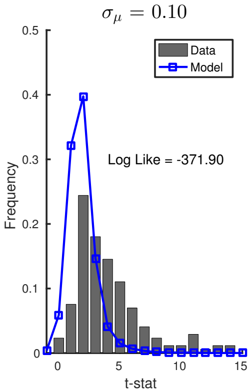
- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of σ_μ comes from dispersion of t-stats

Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of σ_μ comes from dispersion of t-stats

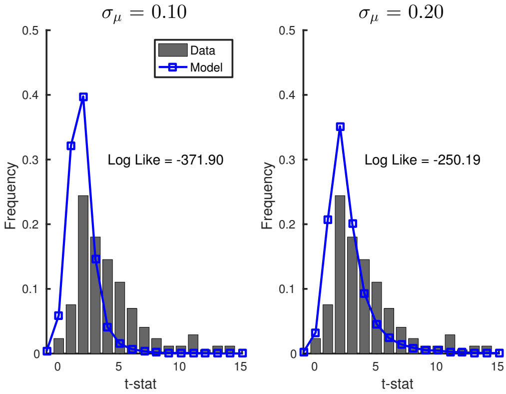
Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of σ_μ comes from dispersion of t-stats



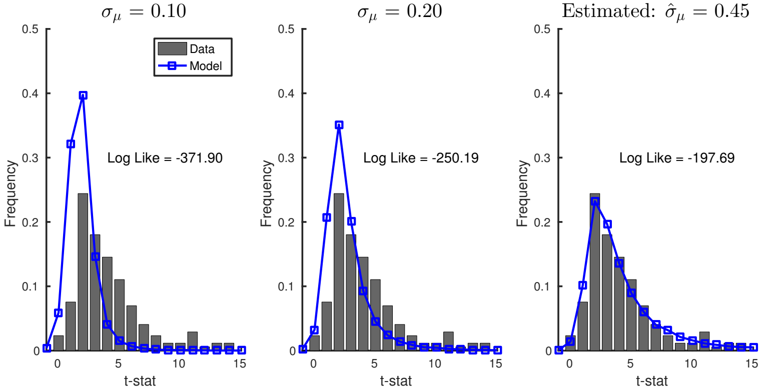
Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of σ_μ comes from dispersion of t-stats



Maximum Likelihood Estimation

- Choose 7 parameters to maximize likelihood of replicated data
  - 172 in-sample returns and standard errors
- Identification of σ_μ comes from dispersion of t-stats



Bias Adjustment and Shrinkage

- We focus on **Shrinkage** defined by

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$

  - 100% Shrinkage $\Rightarrow$ **p-hacking** dominates, bias-adjusted return $= 0$
  - 0% Shrinkage $\Rightarrow$ **journal review** works, bias-adjusted $=$ in-sample

Bias Adjustment and Shrinkage

- We focus on **Shrinkage** defined by

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$

  - 100% Shrinkage $\Rightarrow$ **p-hacking** dominates, bias-adjusted return $= 0$
  - 0% Shrinkage $\Rightarrow$ **journal review** works, bias-adjusted $=$ in-sample

- **Bayesian logic gives a shrinkage formula**
  (Dawid 1994, Senn 2008, Efron 2011, 2012)

$$\text{Shrinkage}_i \approx \frac{[\text{Standard Error}]_i^2}{\hat{\sigma}_\mu^2 + [\text{Standard Error}]_i^2}$$

$\hat{\sigma}_\mu^2$= Estimated Dispersion of True Returns

Results

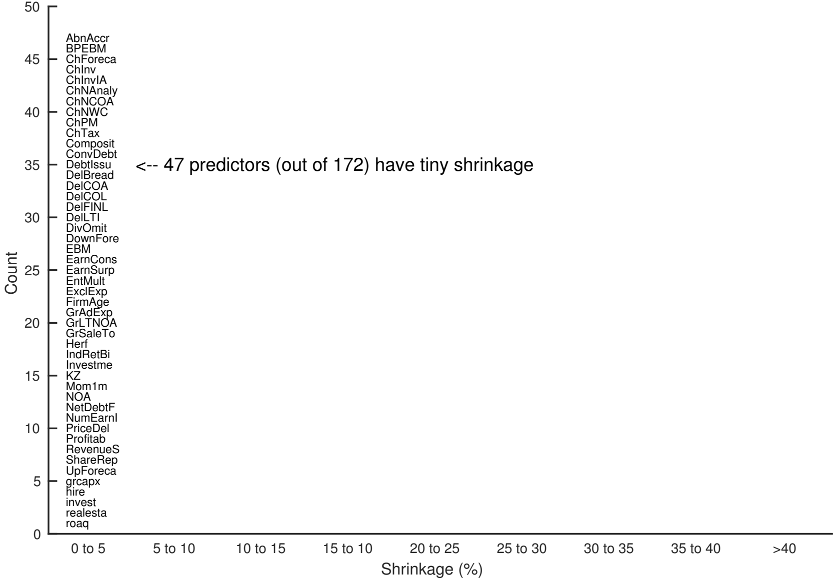
Main Result 1/2: Bias Adjustments are Modest

Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$

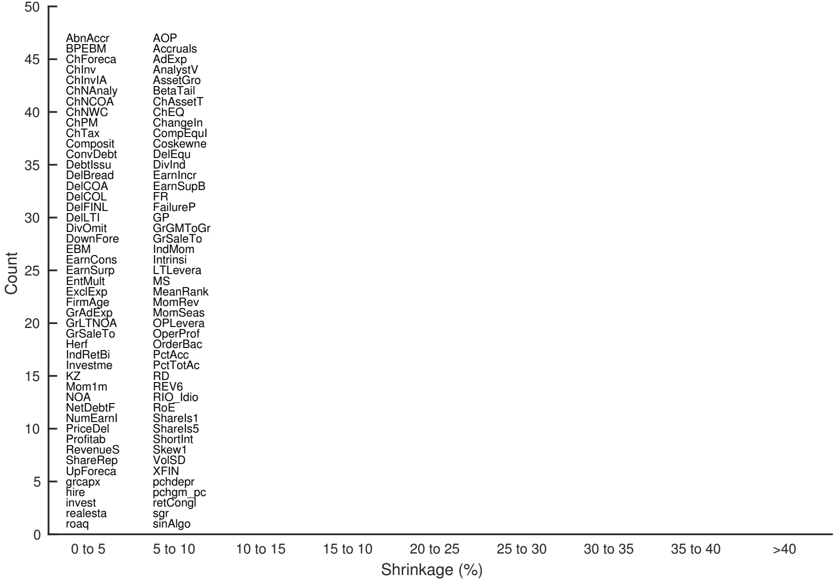
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



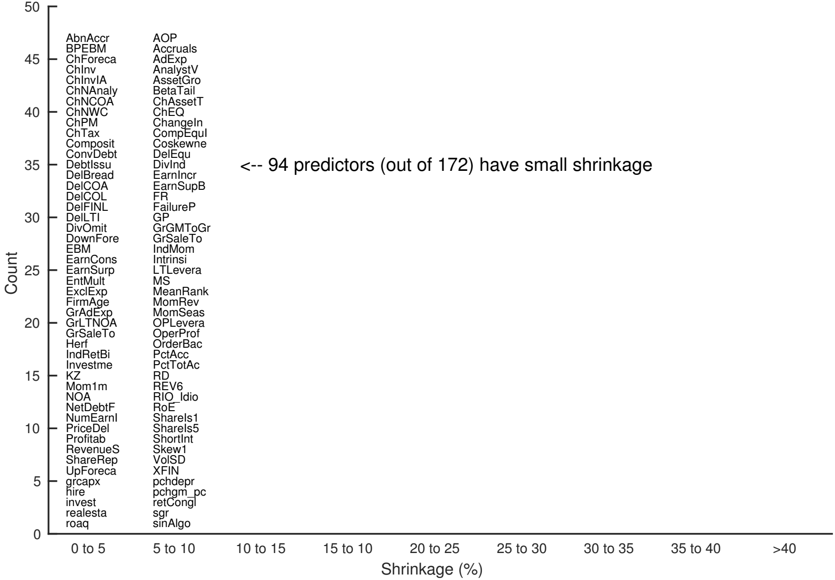
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



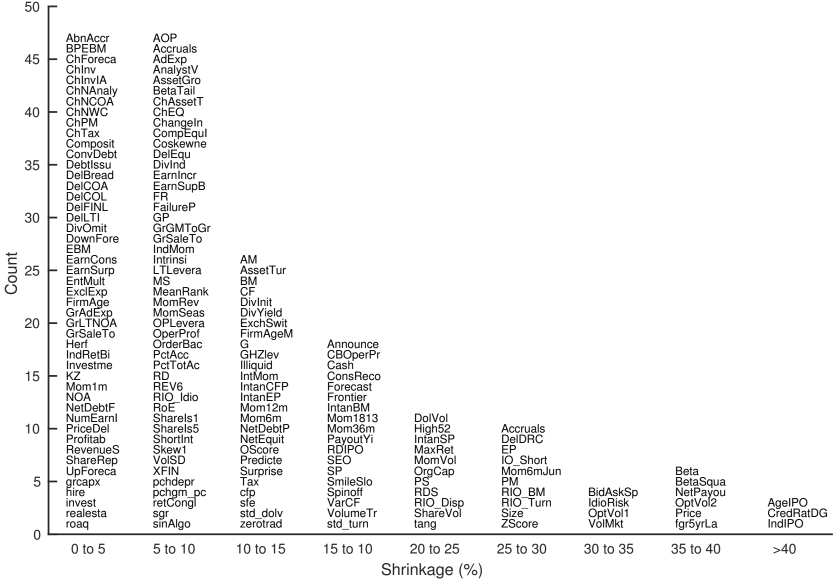
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



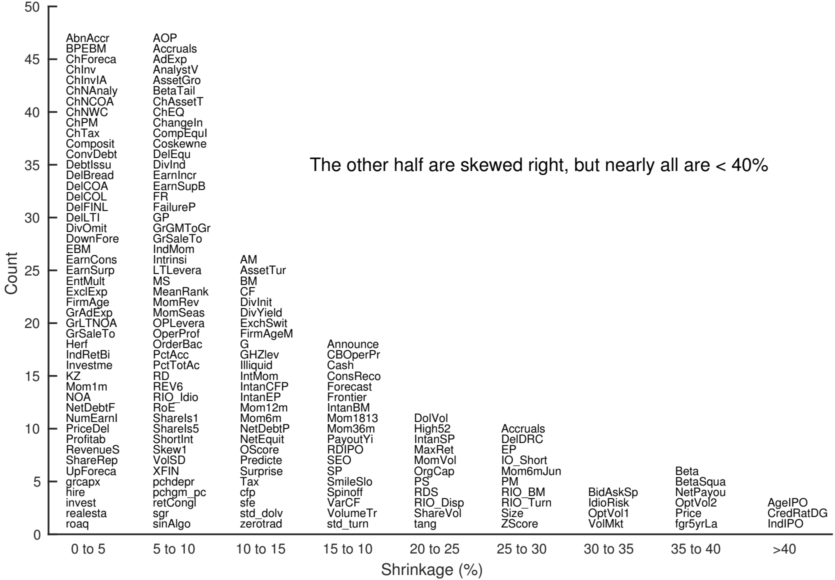
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



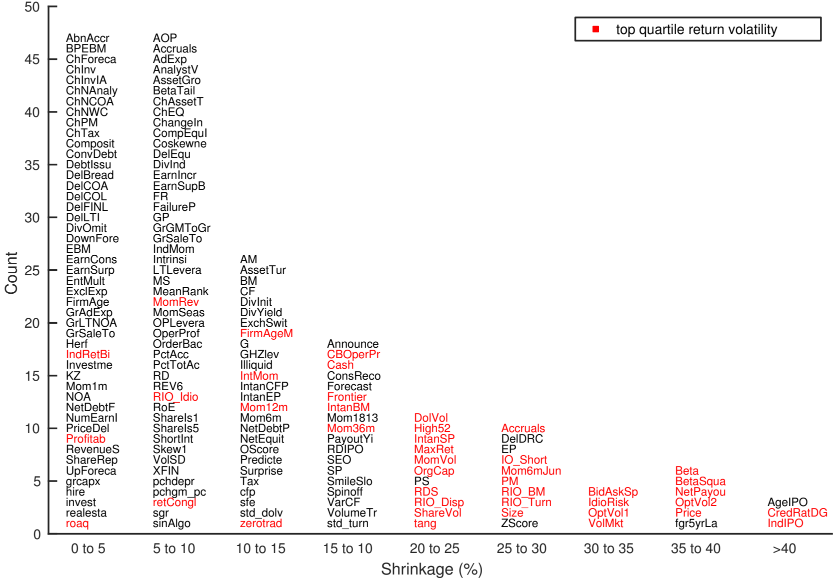
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



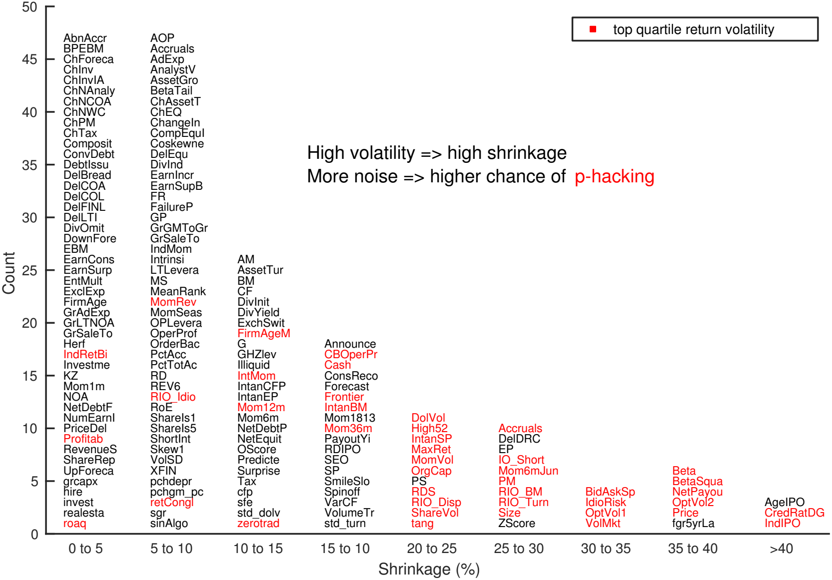
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



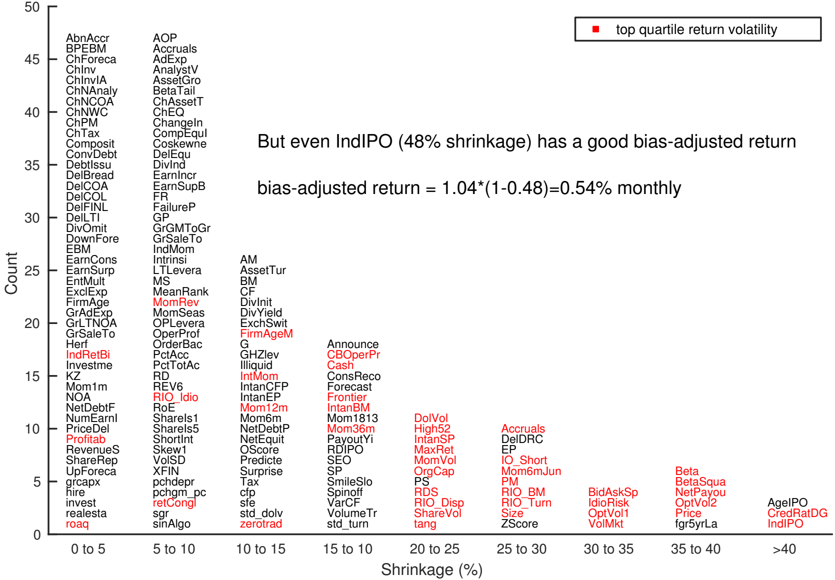
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



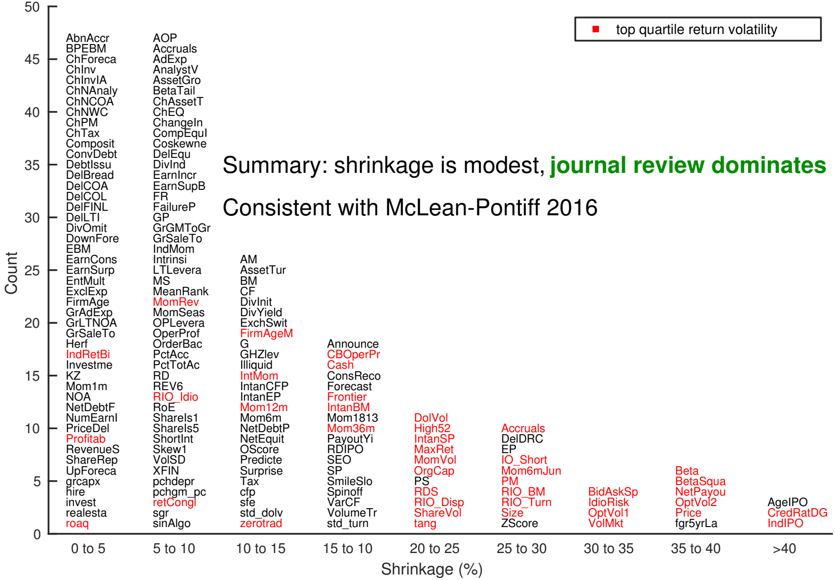
Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



Main Result 1/2: Bias Adjustments are Modest

$$[\text{Bias-Adjusted Return}]_i = (1 - \text{Shrinkage}_i)[\text{In-Sample Return}]_i$$



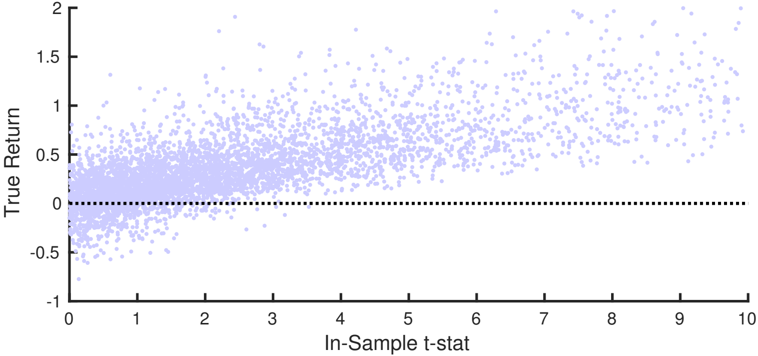
Summary: shrinkage is modest, **journal review dominates**

Consistent with McLean-Pontiff 2016

Main Result 2/2: Nearly All Anomalies were Real

Main Result 2/2: Nearly All Anomalies were Real

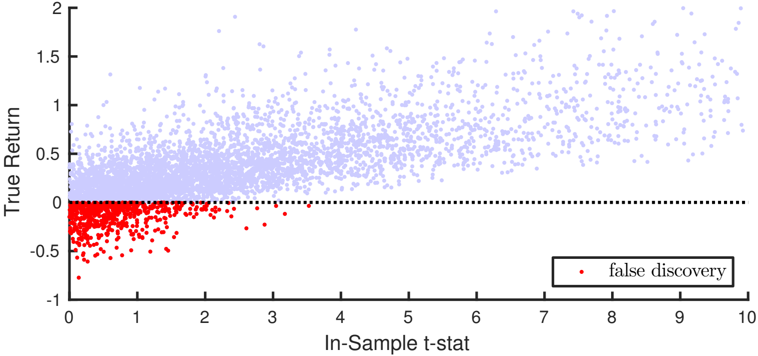
We can estimate the false discovery rate (FDR) (à la HLZ 2016)



- Simulate true returns and t-stats using estimated parameters

Main Result 2/2: Nearly All Anomalies were Real

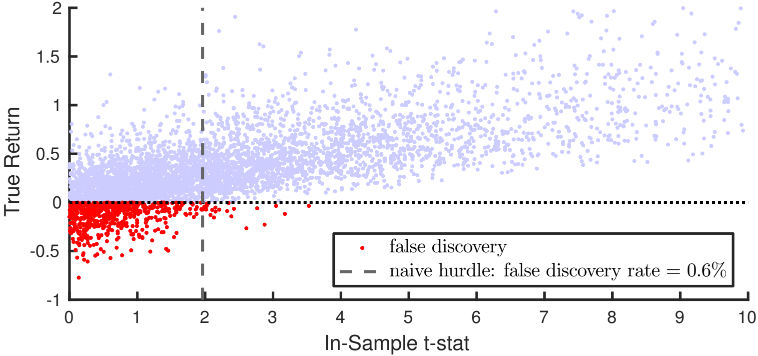
We can estimate the false discovery rate (FDR) (à la HLZ 2016)



- Define false discoveries: true returns ≤ 0 (equivalent to HLZ)

Main Result 2/2: Nearly All Anomalies were Real

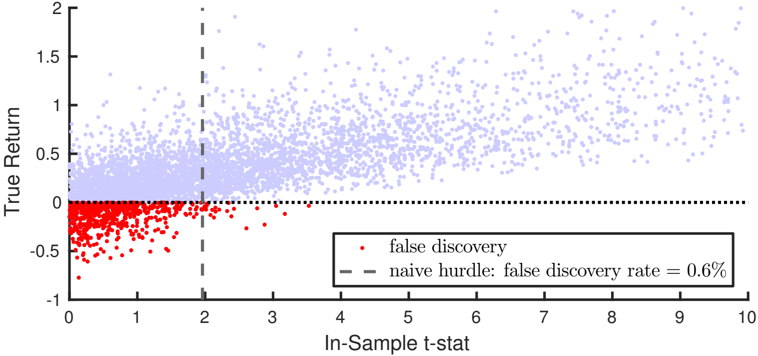
We can estimate the false discovery rate (FDR) (à la HLZ 2016)



- Calculate false discovery rate (FDR) for a given t-stat hurdle
- **Naive hurdle (1.96) implies a tiny FDR of 0.6%**

Main Result 2/2: Nearly All Anomalies were Real

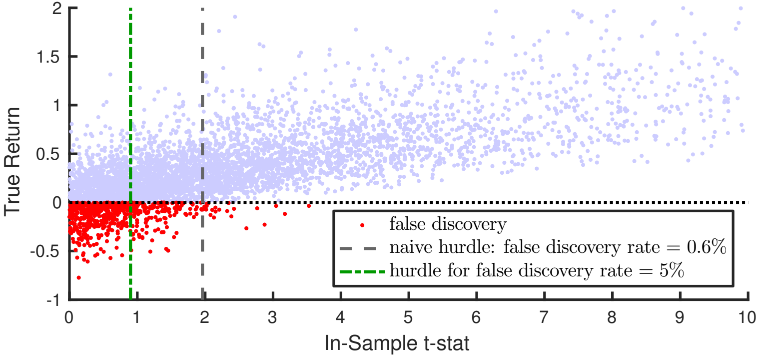
We can estimate the false discovery rate (FDR) (à la HLZ 2016)



- Calculate false discovery rate (FDR) for a given t-stat hurdle
- **Naive hurdle (1.96) implies a tiny FDR of 0.6%**
- **Nearly all anomalies were real** (in-sample)

Main Result 2/2: Nearly All Anomalies were Real

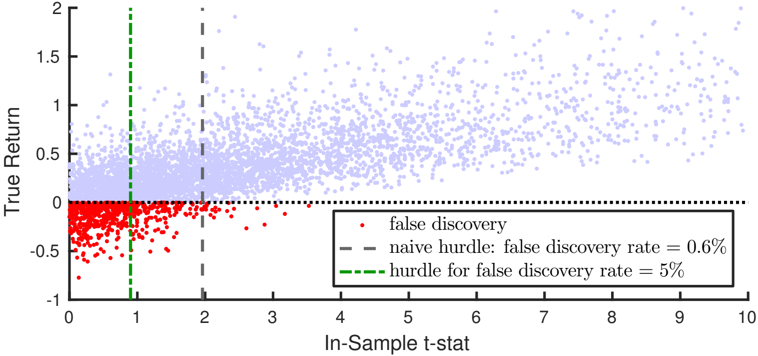
We can estimate the false discovery rate (FDR) (à la HLZ 2016)



- Can calculate hurdles for other FDRs

Main Result 2/2: Nearly All Anomalies were Real

We can estimate the false discovery rate (FDR) (à la HLZ 2016)



- Can calculate hurdles for other FDRs
- Standard t-stat hurdles can actually be **lowered!!!**

Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

How can this be true???

Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

How can this be true???

- Standard multiple-testing logic (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, **the null will be rejected by pure chance**

  $\Rightarrow$ t-stat hurdles should be raised

Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

How can this be true???

- Standard multiple-testing logic (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, **the null will be rejected by pure chance**
  - ⇒ t-stat hurdles should be raised
- Our **more structured logic** (James-Stein 1961, Efron-Morris 1973)
  - **172 tests tell us about the nature of the publication process**

Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

How can this be true???

- Standard multiple-testing logic (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, **the null will be rejected by pure chance**

  $\Rightarrow$ t-stat hurdles should be raised
- Our **more structured logic** (James-Stein 1961, Efron-Morris 1973)
  - **172 tests tell us about the nature of the publication process**
  - The publication process produces **dispersed true returns**

Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

How can this be true???

- Standard multiple-testing logic (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, **the null will be rejected by pure chance**
  - ⇒ t-stat hurdles should be raised
- Our **more structured logic** (James-Stein 1961, Efron-Morris 1973)
  - **172 tests tell us about the nature of the publication process**
  - The publication process produces **dispersed true returns**
  - ⇒ t-stats are informative about true returns

Main Result 2/2: Nearly All Anomalies were Real

Standard t-stat hurdles can actually be **lowered!!!**

How can this be true???

- Standard multiple-testing logic (Bonferroni, Benjamini-Hochberg 1995)
  - After running 172+ tests, **the null will be rejected by pure chance**

  $\Rightarrow$ t-stat hurdles should be raised
- Our **more structured logic** (James-Stein 1961, Efron-Morris 1973)
  - **172 tests tell us about the nature of the publication process**
  - The publication process produces **dispersed true returns**

  $\Rightarrow$ t-stats are informative about true returns

  $\Rightarrow$ t-stat hurdles can be lowered!

Main Result 2/2: Nearly All Anomalies were Real

- Other multiple testing studies find most results are false
  - Harvey, Liu, Zhu (2016); Chordia, Goyal, Saretto (2017)
- Difference: focus on **cross-sectional predictors** in **top-tier journals**

Main Result 2/2: Nearly All Anomalies were Real

- Other multiple testing studies find most results are false
  - Harvey, Liu, Zhu (2016); Chordia, Goyal, Saretto (2017)
- Difference: focus on **cross-sectional predictors** in **top-tier journals**

	Variable Counts		
	Harvey-Liu-Zhu	Chordia-Goyal-Saretto	Our Paper
Aggregate Risk Factor	113	0	0
X-Sectional Predictor	202	2,100,000	172
X-Sectional & Top Tier Pub	146	$<$500	151
Total	315	2,100,000	172

Main Result 2/2: Nearly All Anomalies were Real

- Other multiple testing studies find most results are false
  - Harvey, Liu, Zhu (2016); Chordia, Goyal, Saretto (2017)
- Difference: focus on **cross-sectional predictors** in **top-tier journals**

	Variable Counts		
	Harvey-Liu-Zhu	Chordia-Goyal-Saretto	Our Paper
Aggregate Risk Factor	113	0	0
X-Sectional Predictor	202	2,100,000	172
X-Sectional & Top Tier Pub	146	<500	151
Total	315	2,100,000	172

- **Suggests p-hacking much worse among aggregate risk factors and outside top journals**

Conclusion

Conclusion

- **A structured, focused estimation finds**
  - **Journal review** has triumphed over **p-hacking***

    *in top-tier pubs predicting cross-sectional stock returns, **for now**
    Consistent w/ McLean-Pontiff 2016, Jacobs-Müller 2016, Yan-Zheng 2017

Conclusion

- **A structured, focused estimation finds**
  - **Journal review** has triumphed over **p-hacking***

    *in top-tier pubs predicting cross-sectional stock returns, **for now**
    Consistent w/ McLean-Pontiff 2016, Jacobs-Müller 2016, Yan-Zheng 2017
- Suggests a **complete accounting for the typical anomaly return**
  - **13% publication bias (this paper)**
  - 35% mispricing that can be traded away (McLean and Pontiff 2016)
  - 52% trading costs (Chen and Velikov 2017)