How Much Should We Trust Instrumental Variable Estimates in Political Science?

Practical Advice Based on 67 Replicated Studies

Apoorva Lal Mac Lockhard
Yiqing Xu Ziwen Zu

May 2023

Motivation

- Instrumental variable (IV) strategies have been widely used in the social sciences, including political science.
  - As an attempt to establish causality in the absence of experiments, RD, and longitudinal data
  - >150 papers in APSR, AJPS and JOP during the past decade (2011-2022)

- IV designs require demanding identification assumptions; results need to be interpreted with caution (Angrist, Imbens & Rubin 1996; Sovey & Green 2011)

- "How come IV estimates are always 5 times bigger than OLS estimates in political economy?" (Alberto Alesina, 2016 NBER Summer Institute)
  - Is that true? Why does it happen? What are the implications?

This Paper

- We replicate 67 papers published in the APSR, AJPS, and JOP that employ an IV design as one of the main identification strategies

- We find that
  - First-stage *F* statistic is often overestimated
  - Classical asymptotic standard errors often severely underestimate the uncertainties around the 2SLS estimates with the presence of outliers and non-i.i.d. errors (Young 2022)
  - In one-third of the replicated studies, the 2SLS estimates are 5 times bigger than the OLS estimates
  - 2SLS/OLS ratio is negatively correlated with the strength of the instrument esp. when the IVs are non-experimental

- We provide practical recommendations, including a local-to-zero test, to alleviate these issues

Roadmap

- IV Strategy: Notations & Review
- Replications
  - Data
  - Findings
  - Fixes
- Conclusion

IV Designs: Notations

- Notations: Treatment *d*; Outcome *y*; Instrument *z*

- Parameterization

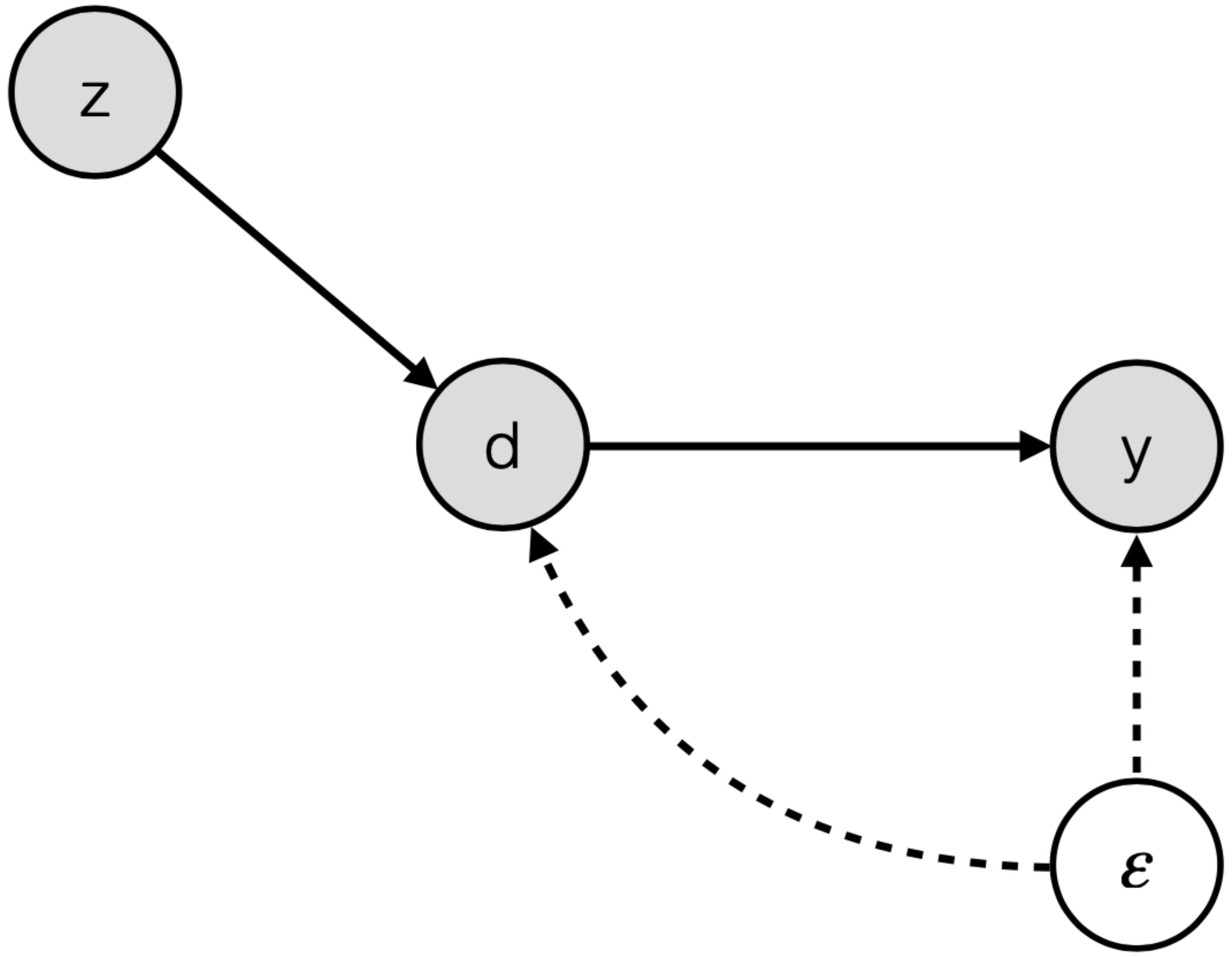
$$
\begin{aligned}
y &= \alpha + \tau d + \varepsilon \\
d &= \pi_0 + \pi' z + \nu
\end{aligned}
$$

- Assumptions
  - Relevance: $\pi \neq 0$
  - Exogeneity (unconfoundedness & exclusion restriction): $Cov(z, \varepsilon) = 0,\ \mathbb{E}[\varepsilon] = 0$

- The 2SLS estimator

$\hat{\tau}_{2SLS} = (\mathbf{d}' P_z \mathbf{d})^{-1} \mathbf{d}' P_z \mathbf{y}$ and $\hat{\tau}_{IV} = (\mathbf{z}'\mathbf{d})^{-1} \mathbf{z}'\mathbf{y}$ (if exactly identified)

(LLN on a "ratio" —> large finite sample bias)



Potential Problems in IV Estimation

- Weak instruments (Fieller 1954; Charles & Starz 1990; Staiger & Stock 1997; Angrist & Pischke 2008)
  - Under i.i.d. errors, exacerbate finite sample bias of $\hat{\tau}_{2SLS}$ (toward OLS)
  - Large variances: $\hat{\mathbb{V}}(\hat{\tau}_{2SLS}) \approx \hat{\mathbb{V}}(\hat{\tau}_{OLS})/R^2_{dz}$
  - Exacerbate finite sample bias of $\hat{\mathbb{V}}(\hat{\tau}_{2SLS})$, leading to wrong test statistics
  - Exacerbate bias from failure of the exclusion restriction (more to follow)

- Problem with the classic asymptotic SE estimator
  - Classical asymptotic variance estimator yield large finite sample biases (Young 2022)
  - Bootstrap procedures behave much better (Cameron, Gelbach, Miller 2008; Davidson & MacKinnon 2012)

- Failure of the exclusion restriction

$$\text{plim } \hat{\tau}_{2SLS} = \tau + \frac{Cov(z,\varepsilon)}{Cov(z,d)} \qquad \Rightarrow \qquad \frac{\text{plim } \hat{\tau}_{2SLS} - \tau}{\text{plim } \hat{\tau}_{OLS} - \tau} = \frac{\rho(z,\varepsilon)}{\rho(d,\varepsilon)} \frac{1}{\rho(z,d)}$$

Roadmap

- IV Designs: A Refresher
- Potential Problems in IV Estimation
- Replications
  - Data
  - Findings
  - Zero-First-Stage
- Recommendations

Data

- APSR, AJPS, and JOP: All papers using IV as one of the main identification strategies from 2011 to 2020
- Criteria
  - IV results supporting the main argument
  - Linear models
  - Exclude dynamic panels using GMM
  - Multiple endogenous variables
- For each design, selecting the most prominent IV result

	All Papers	Incomplete Data	Incomplete Code	Replication Error	Replicable
APSR	30	16	0	3	14 (42%)
AJPS	33	3	1	1	25 (83%)
JOP	51	19	3	1	28 (55%)
Total	**114**	**38**	**4**	**5**	**67 (59%)**

Types of IVs

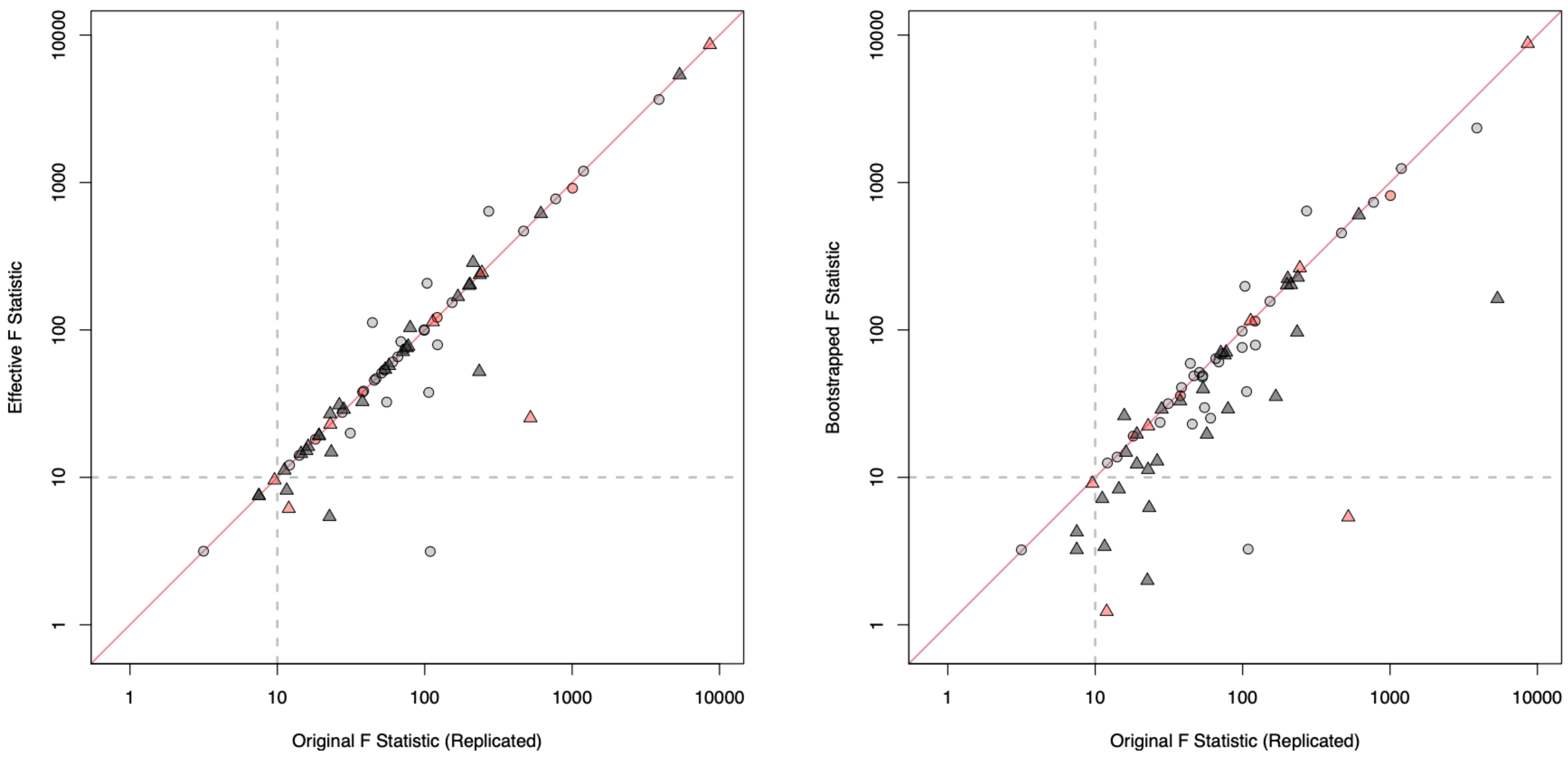
Type of IV	Number of Papers	Percentage
Theory	42	60%
Geography/Climate/Weather	13	19%
History	11	16%
Diffusion	2	3%
Others	16	23%
Experiments	12	17%
Rules (including fuzzy RD)	7	10%
Econometrics	9	13%
Total	**70** (designs)	**100%**

Procedure

- Select the main IV specification that plays a central role in supporting a main claim in the paper
- Compute the first-stage partial F statistics based on (1) classic analytic SEs, (2) Huber White heteroskedastic-robust SEs, (3) cluster- robust SEs, and (4) bootstrapped SEs, as well as (5) the effective F (Olea & Plueger 2013).
- Replicate the original IV result using the 2SLS estimator and apply four different procedures for inference
  1. Conventional t-test based on the analytic SE
  2. Bootstrap-*c* (“c” for coefficient) and bootstrap-*t* (“t” for t-statistics) (Young 2022)
  3. The Anderson-Rubin test (Anderson & Rubin 1949)
  4. The *tF* procedure, which smoothly adjusts the t–ratio critical values based on the first-stage F statistic (Lee et al. 2022)
- Calculate the ratio between 2SLS and OLS estiamtes

Finding 1: First-Stage *F* Statistics

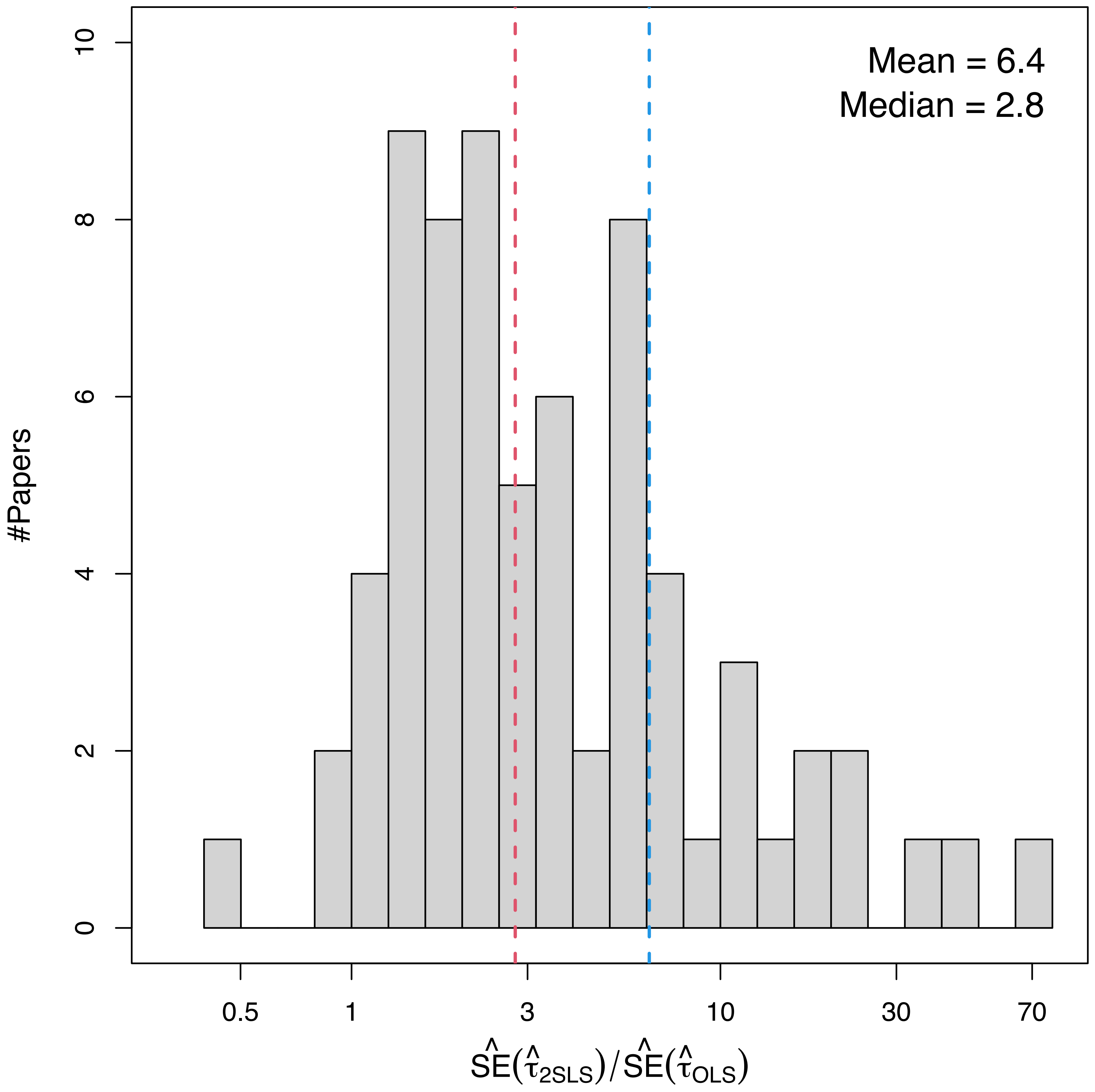
- 17% (12 out of 70) do not report first-stage *F* statistic
- Almost none applies bootstrap or the effective *F*
- 11% (8 out of 70) have effective *F* statistics under 10
- 17% (12 out of 70) have bootstrapped *F* statistics under 10

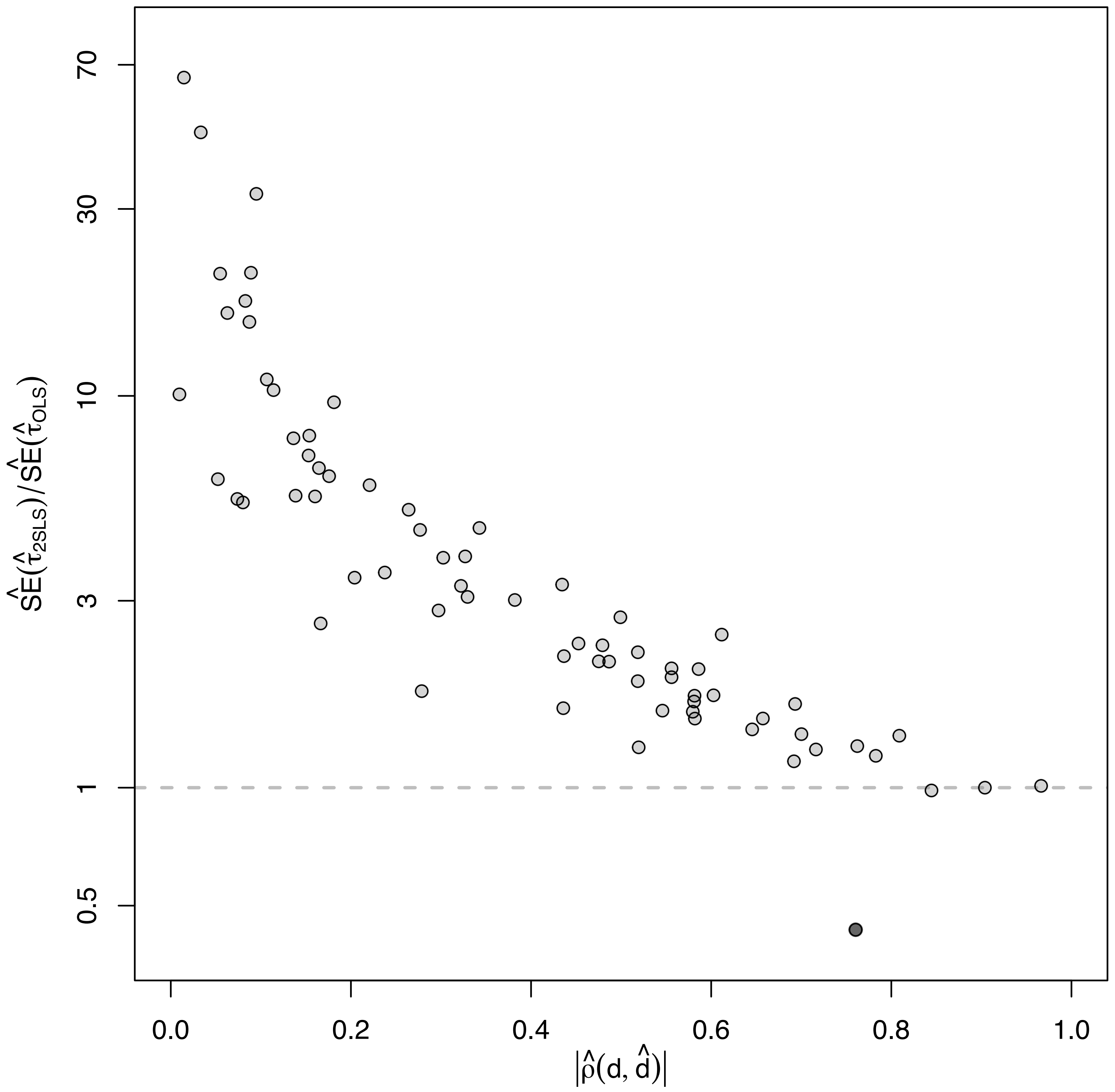


(a) Original F vs. Effective F (b) Original F vs. Bootstrapped F

Finding 2: Inference

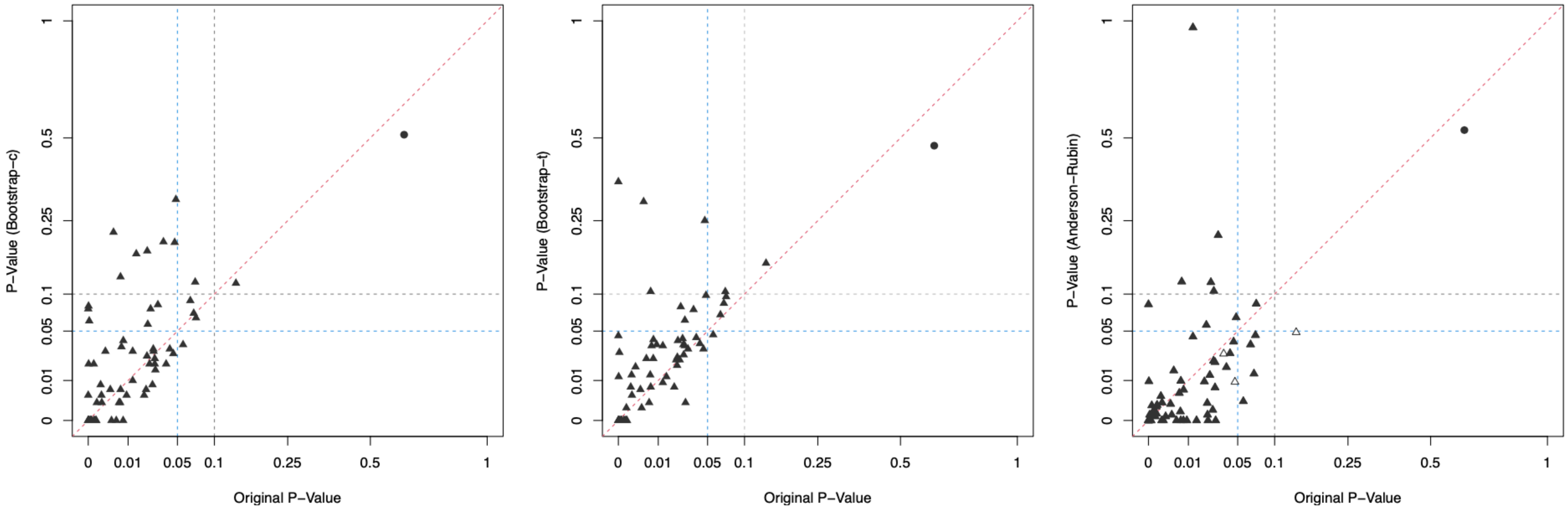
- SE estimates for the 2SLS estimates are usually much larger than those of the OLS estimates





Finding 2: Inference

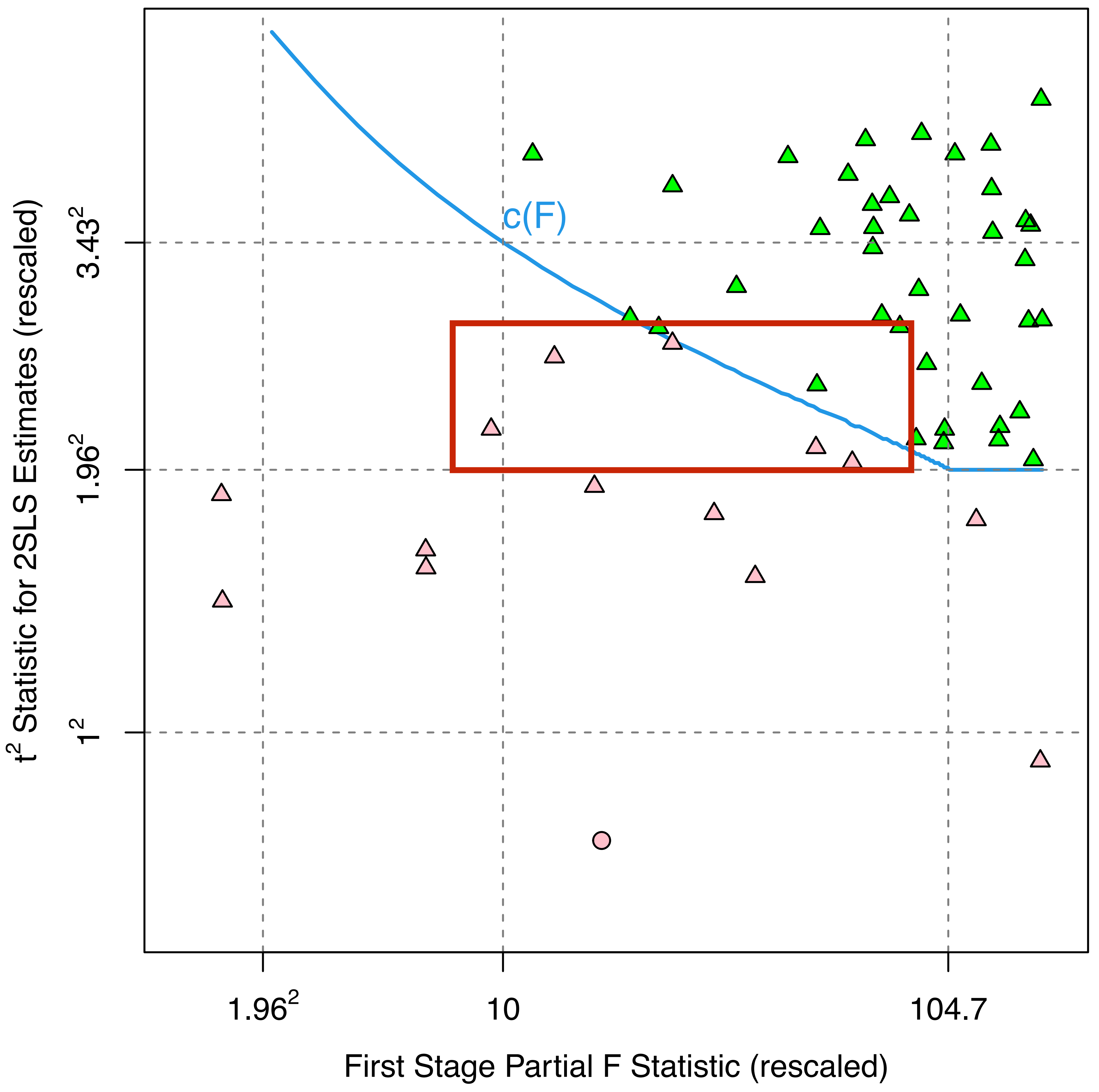
- Using the Anderson-Rubin test, 19% designs become statistically insignificant at 5%
- Using the bootstrap-*t* and bootstrap-*c* methods, 21% and 29% designs become statistically insignificant at 5%, respectively



(a) Bootstrap-c method (b) Bootstrap-t method (c) Anderson-Rubin

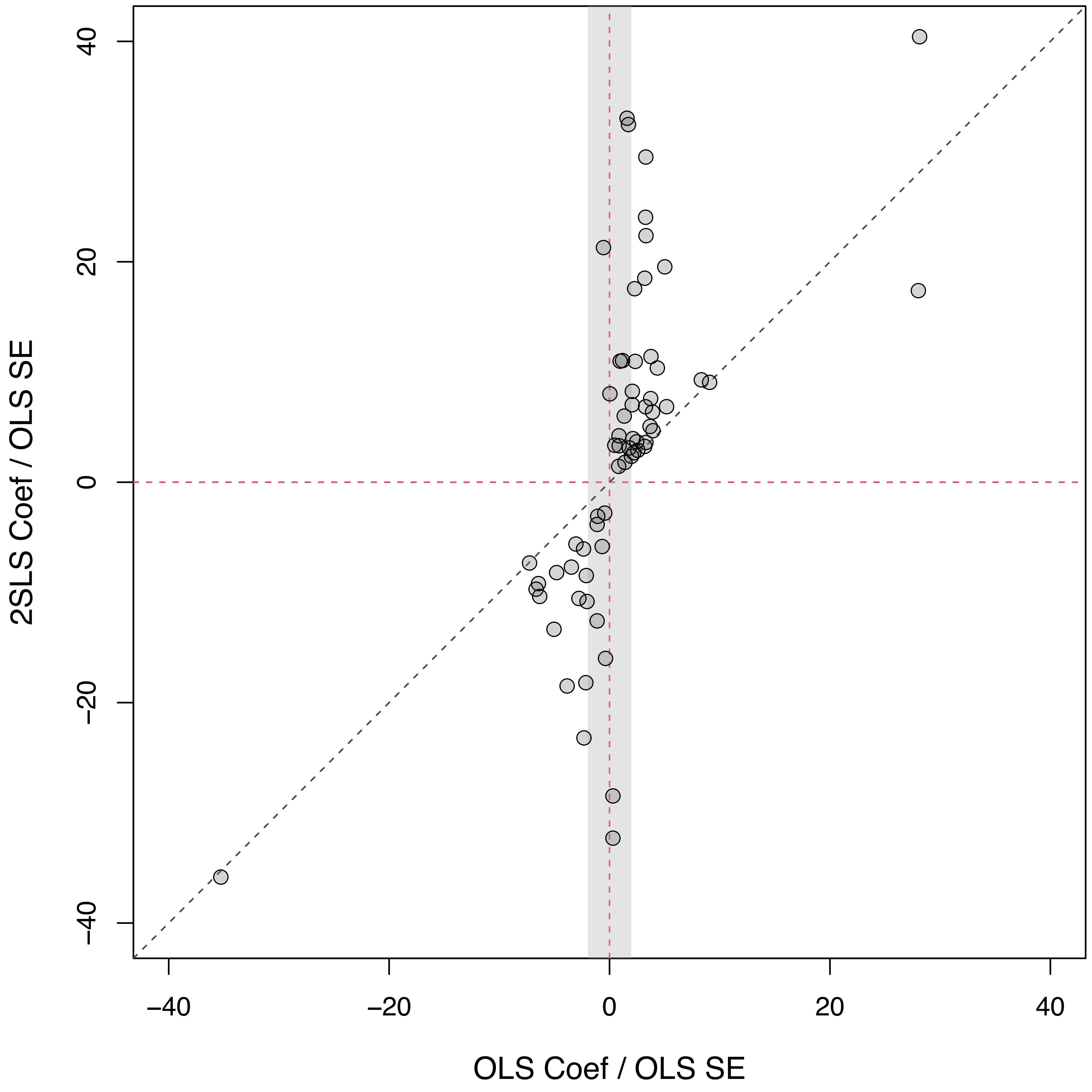
Finding 2: Inference

- For the just identified cases (one-treatment, one-instrument), we can use the *tF* procedure
- As a result, 30% (16 out of 54) designs become statistically insignificant at 5%.
- 5 studies deemed statistically significant when using the conventional fixed critical values (e.g. 1.96) for the *t*-test become statistically insignificant using the *tF* procedure



Finding 3: 2SLS vs OLS

- In most papers, 2SLS and OLS estimates are of the same signs.



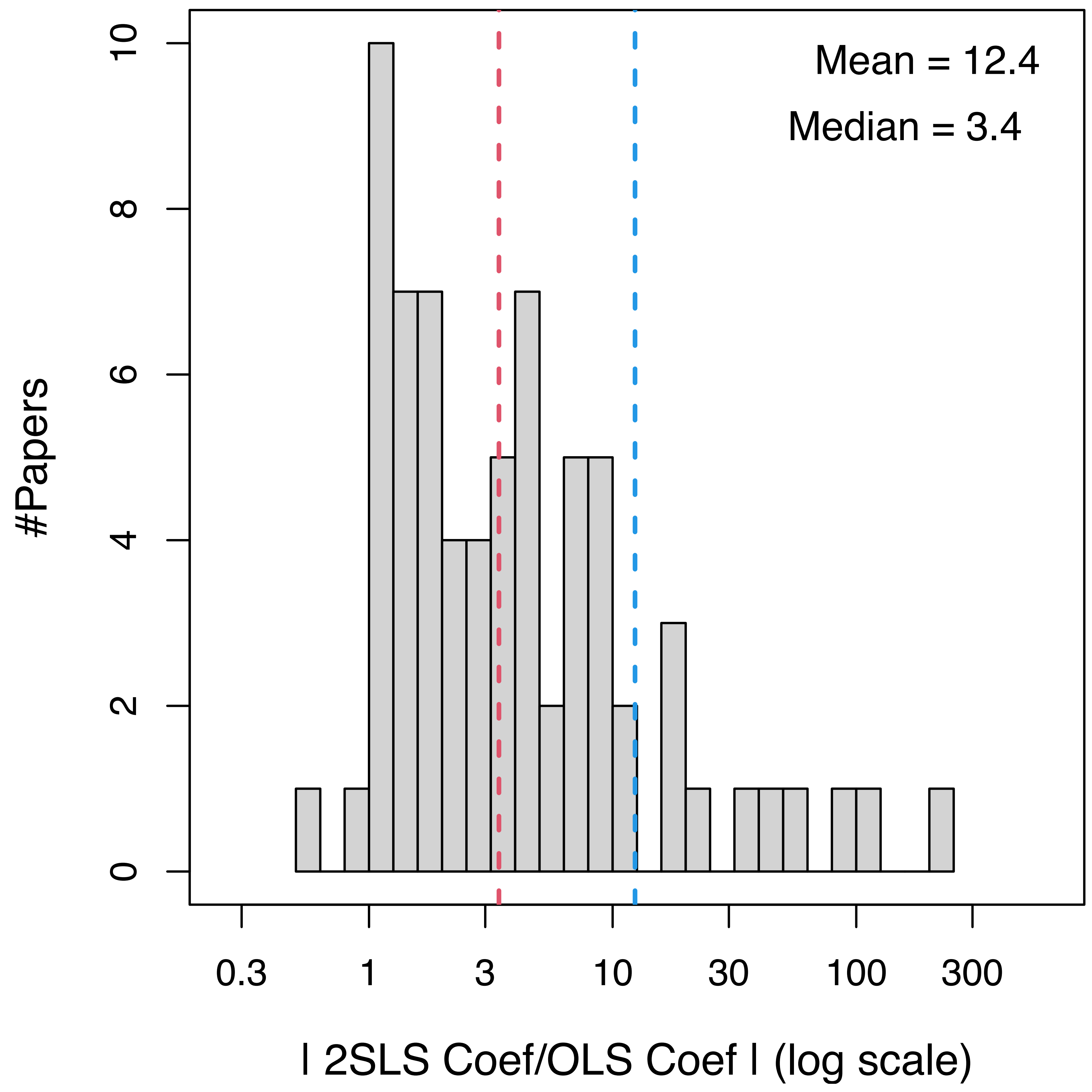
Finding 3: 2SLS vs OLS

- In most papers, 2SLS and OLS estimates are of the same signs.

- In 97% (68 out of 70) designs, the magnitudes of the 2SLS estimates are bigger than those of the OLS estimate

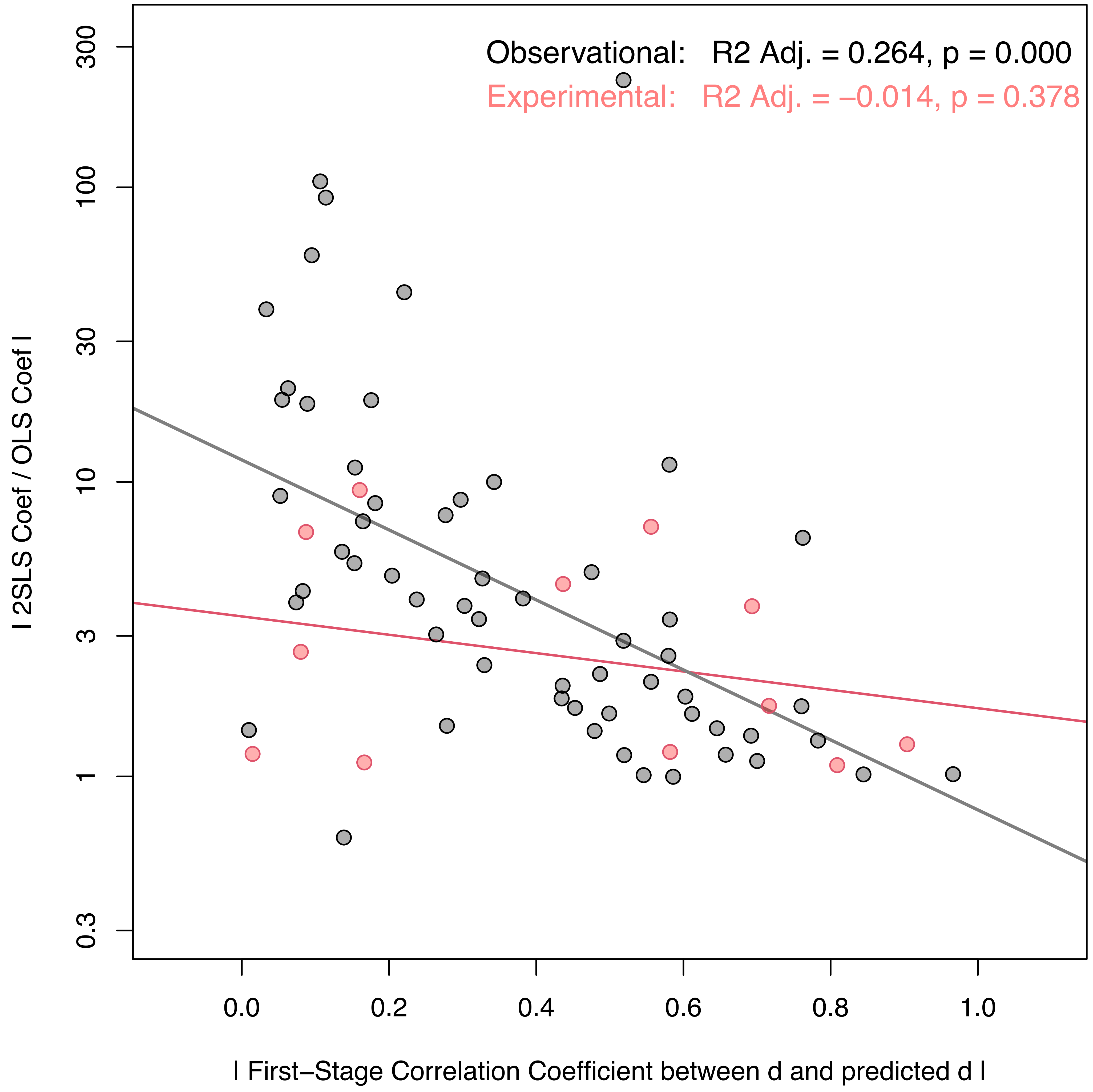
- In 34% of them, the ratio is bigger than 5.

- Excluding those that explicitly claim to expect downward biases in OLS results, the numbers are 96% and 35%.



Finding 3: 2SLS vs OLS

- A strong negative correlation between the ratio and first-stage correlation coefficient
- The relationship is robust to removing studies with statistically insignificant OLS estimates
- Possible explanations:
  1. Failure of IV exogeneity
  2. Publication bias
  3. HTE
  4. Measurement error in the treatment



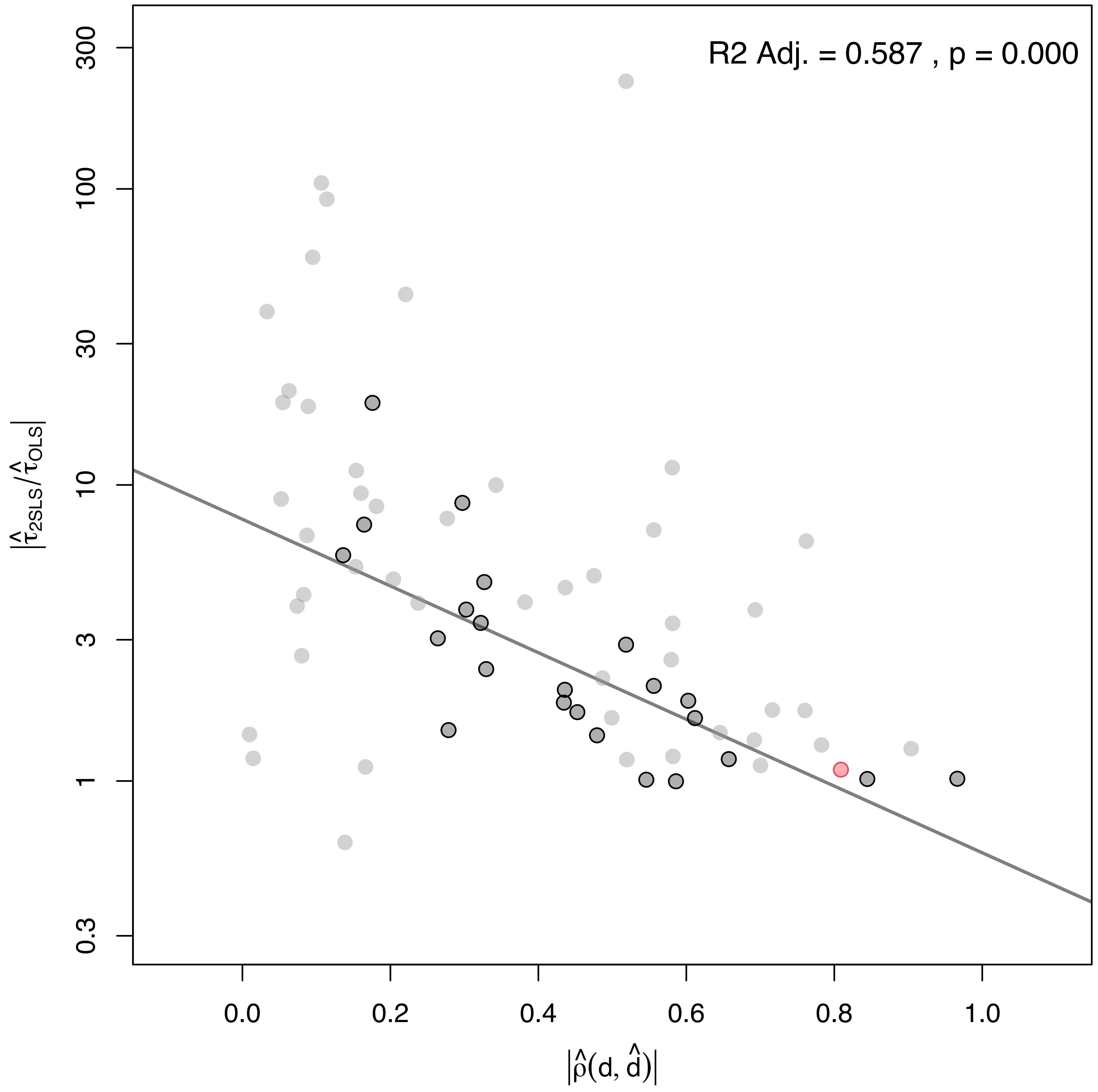
Finding 3: 2SLS vs OLS

- A strong negative correlation between the ratio and first-stage correlation coefficient

- The relationship is robust to removing studies with statistically insignificant OLS estimates

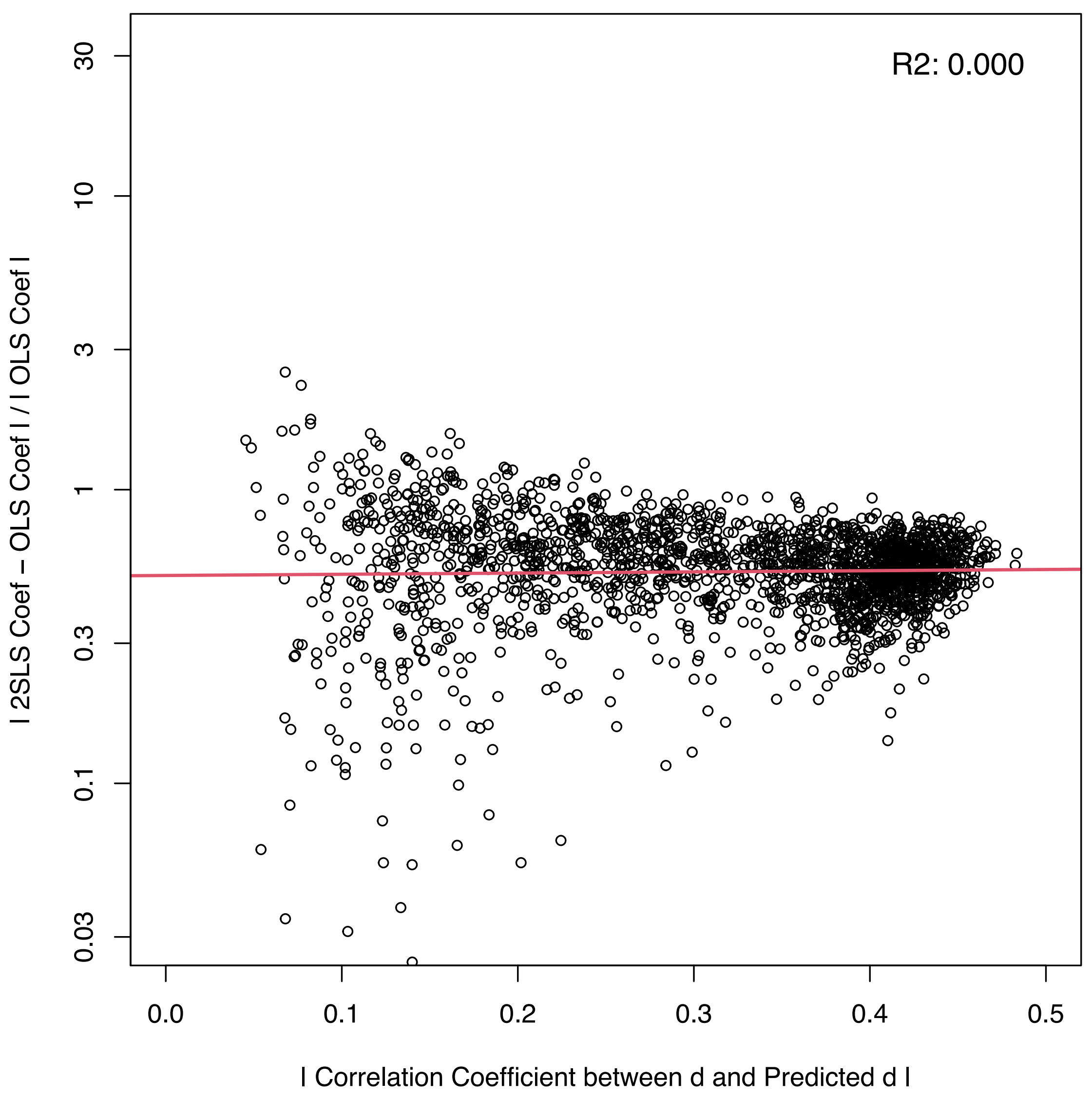
- Possible explanations:

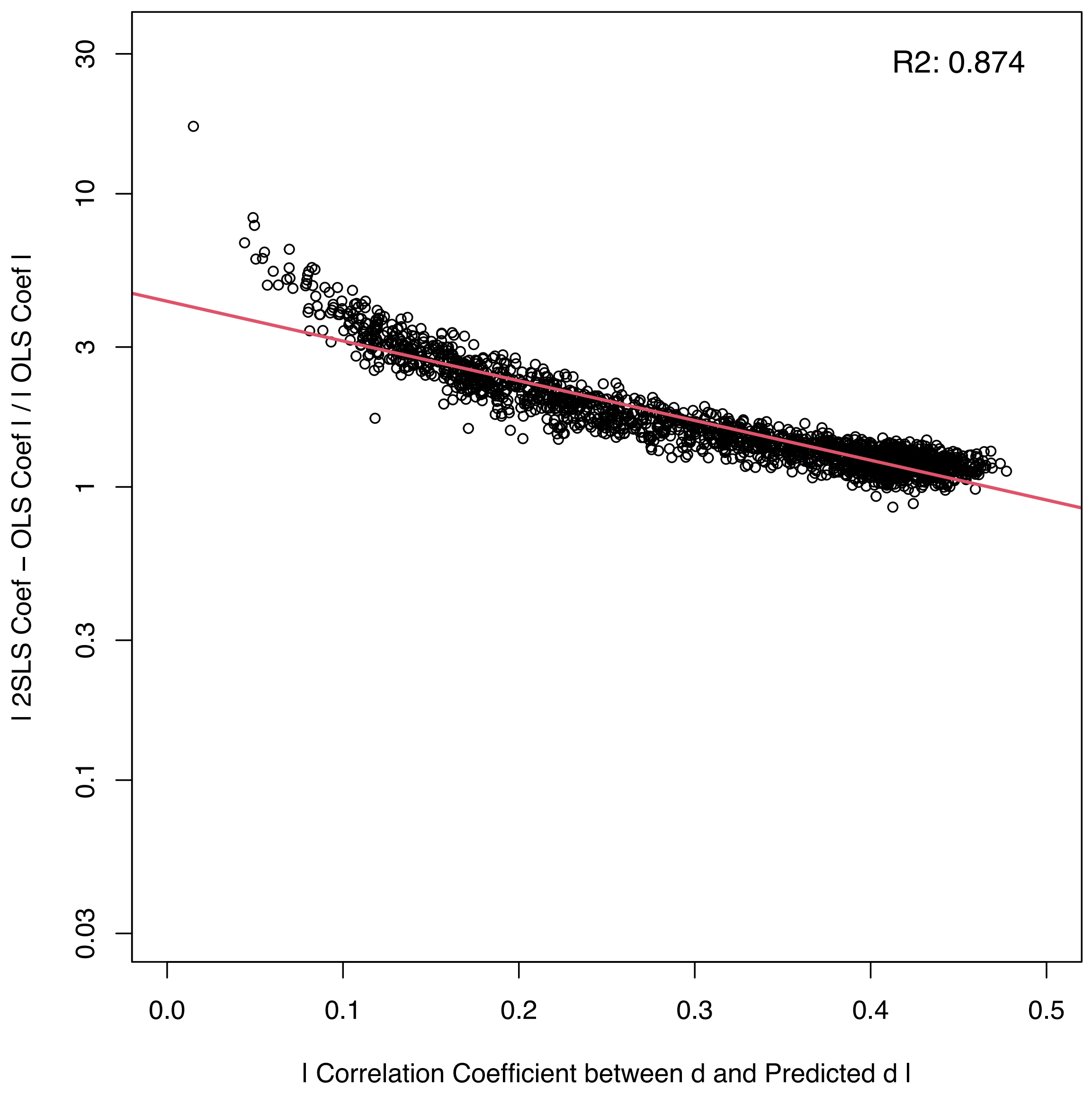
  1. Failure of IV exogeneity
  2. Publication bias
  3. HTE
  4. Measurement error in the treatment



Monte Carlo Evidence

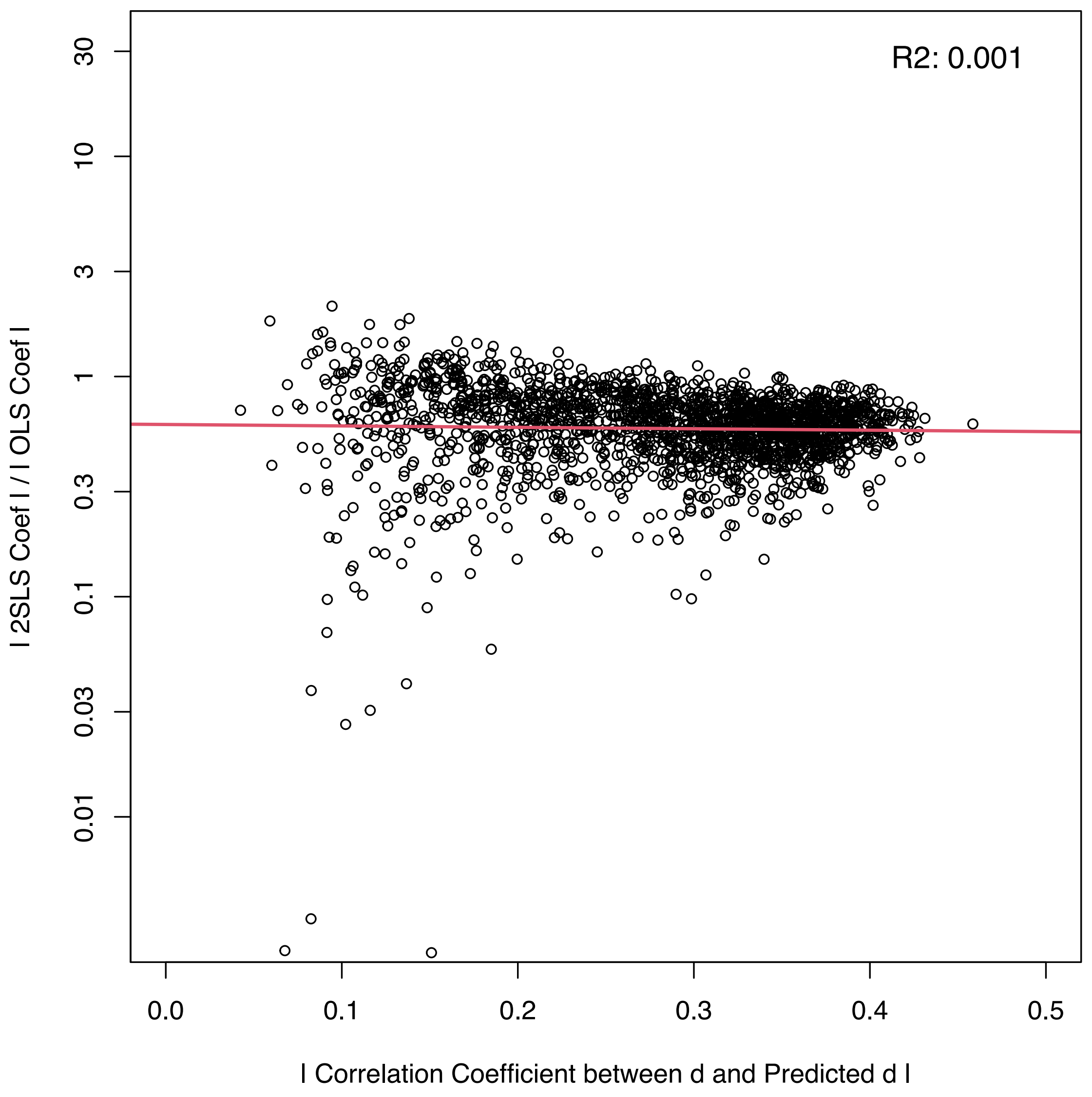
Benchmark

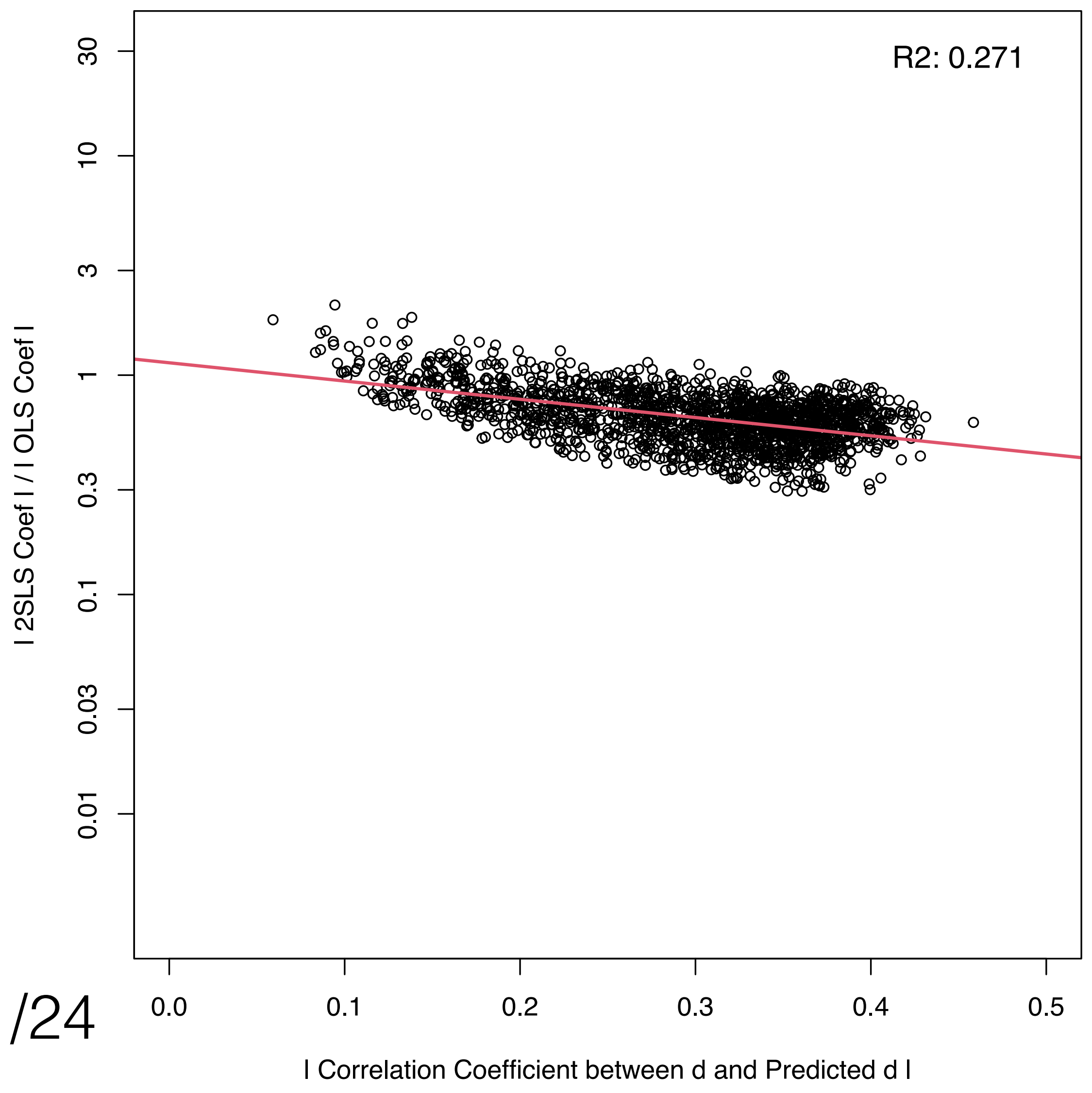




Exclusion restriction violation

HTE





HTE + Publication bias

Fixing Exclusion Restriction Failures is Difficult

- Potential solutions
  - "Design trumps analysis" (Rubin 2008)
  - "Zero-first-stage" (ZFS) test and "local-to-zero" (LTZ) correction

- ZFS test (Bound & Jaeger 2000)
  - Running first stage and reduced-form regressions in places where there *should* be no effect

- LTZ correction (Conley, Hansen & Rossi 2012; van Kippersluis and Rietveld 2018)
  - What would the 2SLS estimate be if a direct effect $d \to y$ existed?
  - We can use the coefficient from a ZFS test based on a subsample to set a prior for the direct effect

$$\hat{\tau} \sim N(\tau + A\mu_\gamma, \mathbb{V}_{2SLS} + A\Omega A')$$

Guiso, Sapienza & Zingales (2016)

- Research question: the impact of self-governing tradition on modern-day social capital

- Outcome: Social capital today
  Treatment: "Free city experience"
  Instrument: Bishop seat in the middle ages

- "Zero-first-stage" in southern Italy; expect LTZ correction has small influences

TABLE 4. REPLICATION OF GSZ (2016) TABLE 6
REDUCED FORM REGRESSIONS

	North		South (ZFS)	
Outcome Variables	Nonprofit (1)	Organ Donation (2)	Nonprofit (3)	Organ Donation (4)
Bishop (IV)	1.612	0.472	0.178	0.189
	(0.219)	(0.047)	(0.137)	(0.065)
Observations	5,357	5,535	2,175	2,178

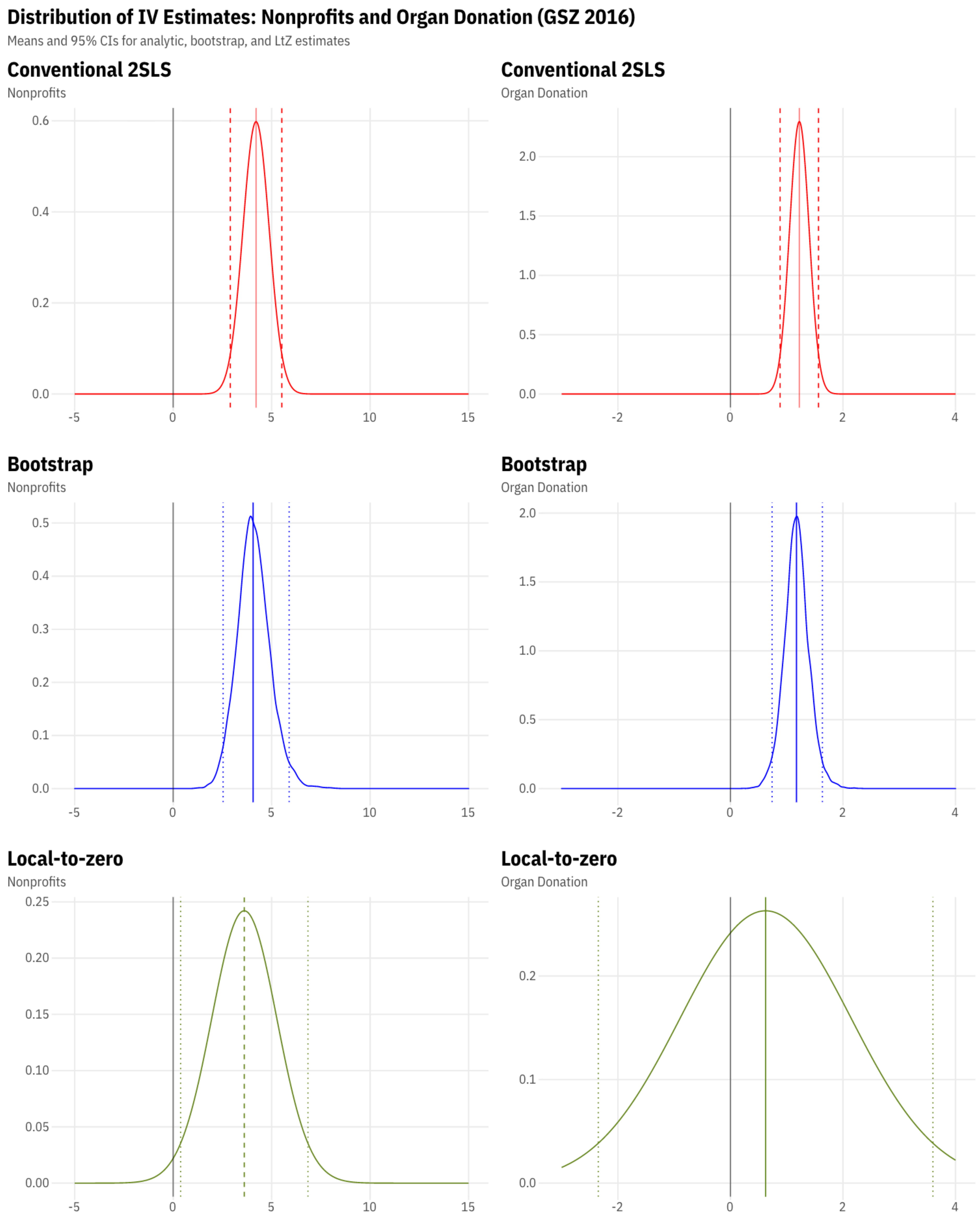
Note: Bootstrap SEs are in the parentheses.

Guiso, Sapienza & Zingales (2016)

- Research question: the impact of self-governing tradition on modern-day social capital

- Outcome: Social capital today
  Treatment: "Free city experience"
  Instrument: Bishop seat in the middle ages

- "Zero-first-stage" in southern Italy; expect LTZ correction has small influences



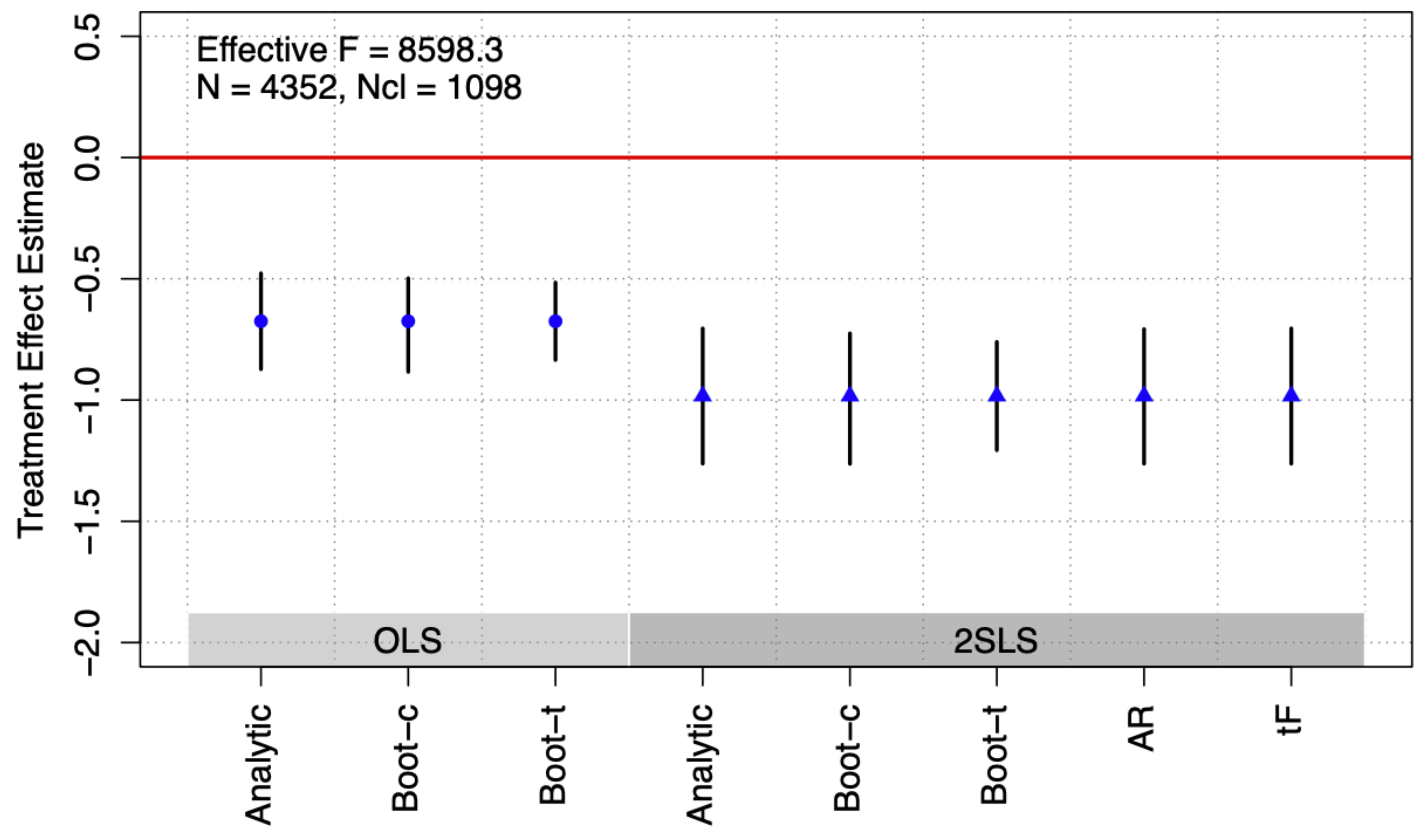
Final Thoughts

- Root cause
  - — IV estimates are much more uncertain than OLS estimates
  - — Violations of unconfoundedness or exclusion restrictions are common
  - ➡ Incentives to *p*-hack & publication biases
  - ➡ Large IV-OLS discrepancy

- IV is a design-based method; it should be used like one
  - — Be extra-cautious when IVs are not generated by experiments & rules (fuzzy RD)
  - — Finding one good IV is difficult; finding multiple good ones is super-difficult if not impossible — they should be justified individually (Angrist, Imbens & Graddy 2000; Angrist, Lavy & Schlosser 2010)
  - — If possible, characterize compliers and never-takers (ZFS) (Abadie 2003, Marbach & Hangartner 2020)

A Checklist

- Think hard about the design; commit to the direction of the selection bias
- Obtain first-stage partial *F* statistic (e.g., the effective *F*)
- Use conservative and weak-IV robust methods to conduct inference
- Ask if a large 2SLS/OLS ratio is plausible
- For observational studies, conduct a placebo test, e.g. a ZFS test, and sensitivity analysis



- R Package available at https://yiqingxu.org/packages/ivDiag/
  Thank you!