# How Much Should We Trust Instrumental Variable Estimates in Political Science?
# — Practical Advice Based on Over 60 Replicated Studies
# 我们能在多大程度上相信政治科学研究中的IV估计?

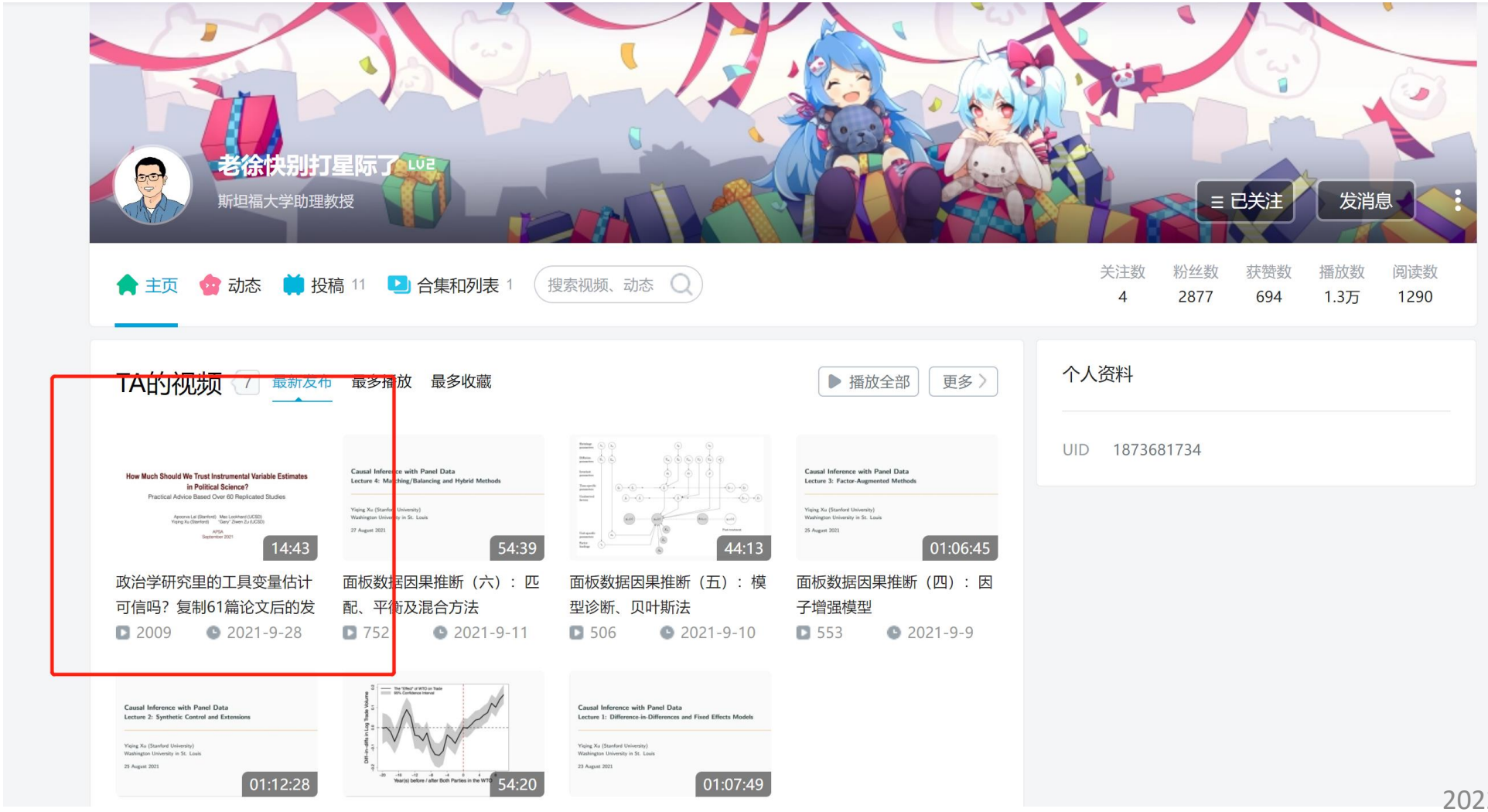
(Apporva Lal; Mac Lockhart; Yiqing Xu; Ziwen Zu)

汇报人：陈泽宇
2018级经济学本科

中国人民大學
RENMIN UNIVERSITY OF CHINA

# 目录

# 问题提出

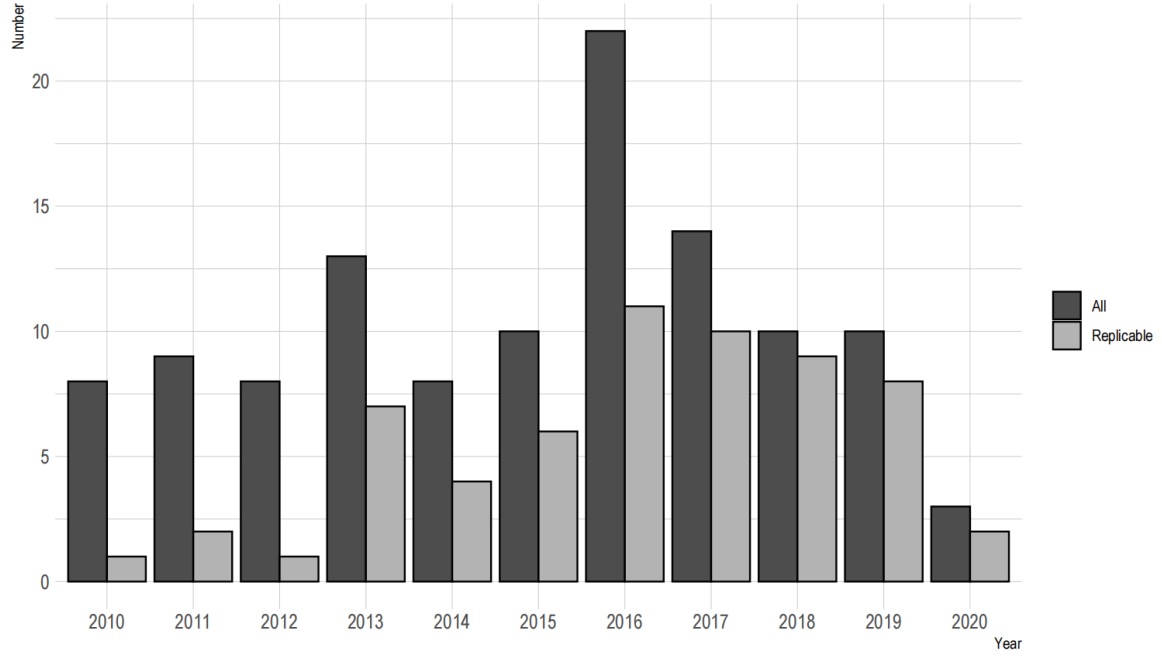

老徐快别打星际了 LV2
斯坦福大学助理教授
已关注
发消息
主页
动态
投稿 11
合集和列表 1
搜索视频、动态
关注数 4
粉丝数 2877
获赞数 694
播放数 1.3万
阅读数 1290
TA的视频 7
最新发布
最多播放
最多收藏
播放全部
更多
How Much Should We Trust Instrumental Variable Estimates in Political Science?
Practical Advice Based Over 60 Replicated Studies
14:43
政治学研究里的工具变量估计可信吗？复制61篇论文后的发
2009 2021-9-28
Causal Inference with Panel Data
Lecture 4: Matching/Balancing and Hybrid Methods
54:39
面板数据因果推断（六）：匹配、平衡及混合方法
752 2021-9-11
44:13
面板数据因果推断（五）：模型诊断、贝叶斯法
506 2021-9-10
Causal Inference with Panel Data
Lecture 3: Factor-Augmented Methods
01:06:45
面板数据因果推断（四）：因子增强模型
553 2021-9-9
Causal Inference with Panel Data
Lecture 2: Synthetic Control and Extensions
01:12:28
54:20
Causal Inference with Panel Data
Lecture 1: Difference-in-Differences and Fixed Effects Models
01:07:49
个人资料
UID 1873681734

# 问题提出

## 工具变量的普及

- 政治学领域的研究中，工具变量方法越来越普及。

FIGURE A1. PAPERS USING INSTRUMENTAL VARIABLES PUBLISHED IN THE *APSR*, *AJPS*, AND *JOP*, BY YEAR.



# 问题提出

## 质疑

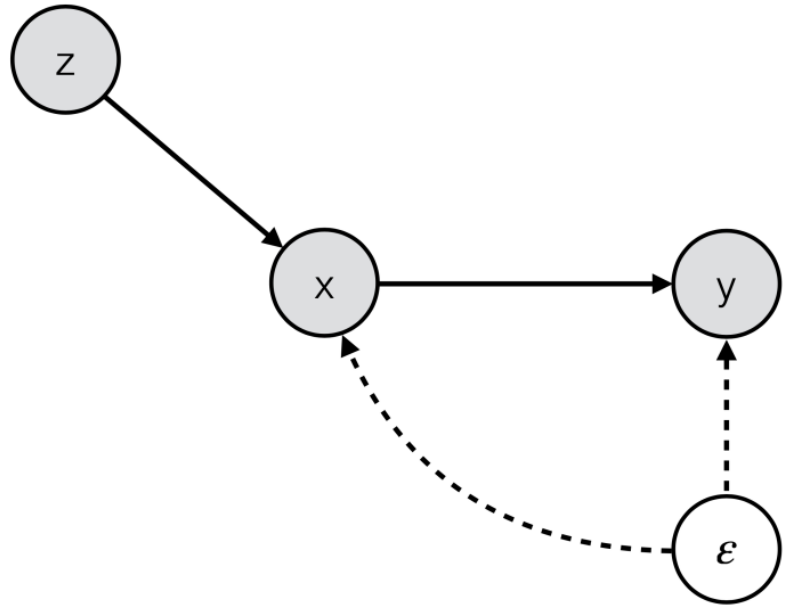
- 使用IV的研究中，IV估计得到的x对y的影响往往要大于OLS估计出来的影响。
- 在2016年NBER的政治经济学专场会议上，Alberto Alesina质疑道：“为什么在政治经济学中，IV估计值总是比OLS估计值大五倍？”
- 本文要回答的问题：
  - Is that true?
  - Why does it happen?
  - What are the implications?

# 工具变量方法的回顾

## 回归方程与两条假设

- Structural equation: $y_i = \alpha + \beta x_i + \varepsilon_i$
- First-stage equation: $x_i = \pi_0 + \pi_1 z_i + \nu_i$
- Reduced-form: $y_i = \underbrace{(\alpha_0 + \beta\pi_0)}_{\gamma_0} + \underbrace{(\beta\pi_1)}_{\gamma_1} z_i + (\beta\nu_i + \varepsilon_i)$

- 相关性：$cov(x_i, z_i) \neq 0$
- 排他性：$cov(\varepsilon_i, z_i) = 0$



# 工具变量方法的回顾

## 工具变量方法使用时的陷阱

- 相关性：$cov(x_i, z_i) \neq 0$
- 排他性：$cov(\varepsilon_i, z_i) = 0$

- (1)满足排他性、弱工具变量：$\hat{\beta}_{IV}$一致性满足，但是$\widehat{SE}$会很大，有限样本中近似正态难以满足。
  - 怎样才算是强工具变量?
  - 经验法则：F>10
  - F⩾104.7 (Lee et al., 2020)
- (2)不满足排他性、强工具变量：$\hat{\beta}_{IV}$不一致，但偏误不会太大，很可能要比$\hat{\beta}_{OLS}$的偏误要小(IV还是比OLS好)。
- (3)不满足排他性、弱工具变量：$\hat{\beta}_{IV}$不一致，而且弱工具变量会放大这种偏误，很可能比$\hat{\beta}_{OLS}$的偏误还大。

# 现有研究中的IV估计

## 考察论文的范围

- 2010年至2020年6月发表在政治学三大刊上使用IV估计作为主要识别策略的文章。
  - American Political Science Review (APSR): 38
  - American Journal of Political Science (AJPS): 31
  - Journal of Politics (JOP): 46

- 同时，还需要满足四个标准：
  - 正文中出现IV估计的结果，并用于支撑文中的主要结论；
  - 使用线性模型(即被解释变量为离散变量的研究不纳入考虑)；
  - 不考察同一个回归模型中存在多个内生变量的研究；
  - 不考虑动态面板结构下的GMM估计(Bun and Windmeijer, 2010)。

# 现有研究中的IV估计

## 考察论文的范围

- 当然，还必须要得能找到数据(Harvard Dataverse、作者网站……)：70/115=61%
  - “2016至2017年以来，在新的政策要求作者公开提供可复制的材料以后，数据的可用性有了明显的改善。尽管没有一个期刊要求第三方可以实现完全复制，但在我们看来这依然将构成一个重大的改进。”
  - 70篇研究中还有2篇数据是不完整的，导致无法复制结果。
  - 61篇可复制的研究中有3篇的复制结果与原文结果不一致，但作者予以保留。
  - 另外，61篇中有3篇文章有2个IV估计，因此总共有64个IV估计。

TABLE 1. DATA AVAILABILITY AND REPLICABILITY OF PAPERS USING IVS

| | #All Papers | Incomplete Data | Incomplete Code | Replication Error | Replicable |
|---|---|---|---|---|---|
| APSR | 38 | 23 | 0 | 2 | 13 (34%) |
| AJPS | 31 | 7 | 1 | 0 | 23 (74%) |
| JOP | 46 | 17 | 2 | 2 | 25 (54%) |
| Total | 115 | 47 | 3 | 4 | 61 (53%) |

# 现有研究中的IV估计

## 大家都用什么当IV?

- 最多的是用理论来论述IV满足排他性假设。

TABLE 2. TYPES OF IVS

| IV Type | #Papers | Percentage% |
|---|---|---|
| **Theory** | 40 | 62.5 |
| Geography/climate/weather | 10.5 | 16.4 |
| History | 10 | 15.6 |
| Treatment diffusion | 2.5 | 3.9 |
| Others | 17 | 26.6 |
| **Experiment** | 12 | 18.8 |
| **Rules & policy changes** | 5 | 7.8 |
| Change in exposure | 3 | 4.7 |
| Fuzzy RD | 2 | 3.1 |
| **Econometrics** | 7 | 10.9 |
| Interactions/"Bartik" | 5 | 7.8 |
| Lagged treatment | 1 | 1.6 |
| Empirical test | 1 | 1.6 |
| **Total** | 64 | 100.0 |

***Note:*** One paper uses both geography-based instruments and an instrument based on treatment diffusion from neighbors. We count 0.5 for each category.

# 现有研究中的IV估计

**大家都用什么当IV?**

- Theory：Geography/climate/weather类型的IV
  - Zhu (2017)：跨国公司的经济活动→地方腐败程度，IV：各省与五个经济中心的加权距离；
  - Hager and Hilbig (2019)：历史上的继承习俗→社会平等，IV：各地到莱茵河或内卡河的距离；
  - Grossman et al. (2017)：地方政府数量→公共品质量，IV：中小河流和地块的数量；
  - Henderson and Brooks (2016)：民主党的投票率→现任者国会中的点名位置，IV：选举日前是否下雨。

# 现有研究中的IV估计

**大家都用什么当IV?**

- "Rain, Rain, Go Away: 176 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable" (Mellon, 2021, SSRN)
  - "A review of 279 studies reveals 176 variables which have been linked to weather: all of which represent potential exclusion violations."

# 现有研究中的IV估计

**大家都用什么当IV?**

- Theory：Historical类型的IV
  - Vernby (2013)：选民中非公民的比例→公共支出，IV：历史上两次移民高峰时各地的移民数占当地总人口的比例；
  - Spenkuch and Tillmann (2018)：天主教徒的比例→纳粹的选票份额，IV：历史上当地的领主是否接纳新教。

# 现有研究中的IV估计

**大家都用什么当IV?**

- Theory：Treatment diffusion类型的IV
  - Dube and Naidu (2015)：美国对哥伦比亚的军事援助→国内政治冲突，IV：美国对拉丁美洲其他地区的军事援助；
  - Grossman et al. (2017)：地方政府数量→公共品质量，IV：其他国家的地方政府数量；
  - Dorsch and Maarek (2019)：是否是民主国家→基尼系数，IV：其他国家是民主国家的比例。

# 现有研究中的IV估计

## 大家都用什么当IV?

- 第二多的是用实验：工具变量通常是“个体是否被鼓励受到处理”。
- 估计得到的是compliers的局部平均处理效应(LATE)。

TABLE 2. TYPES OF IVS

| IV Type | #Papers | Percentage% |
|---|---|---|
| **Theory** | 40 | 62.5 |
| Geography/climate/weather | 10.5 | 16.4 |
| History | 10 | 15.6 |
| Treatment diffusion | 2.5 | 3.9 |
| Others | 17 | 26.6 |
| **Experiment** | 12 | 18.8 |
| **Rules & policy changes** | 5 | 7.8 |
| Change in exposure | 3 | 4.7 |
| Fuzzy RD | 2 | 3.1 |
| **Econometrics** | 7 | 10.9 |
| Interactions/“Bartik” | 5 | 7.8 |
| Lagged treatment | 1 | 1.6 |
| Empirical test | 1 | 1.6 |
| **Total** | 64 | 100.0 |

***Note:*** One paper uses both geography-based instruments and an instrument based on treatment diffusion from neighbors. We count 0.5 for each category.

# 现有研究中的IV估计

## 大家都用什么当IV?

- 第三类是用规定或政策形成的准实验。
- 但作者这里只指两类：模糊断点回归、因为出生或者某些资格影响是否受到处理。

- Dinas (2014)：投票给某一党派→自身对该党派的倾向性，IV/forcing variable：年龄(18岁之后才享有投票权)

TABLE 2. TYPES OF IVS

| IV Type | #Papers | Percentage% |
|---|---|---|
| **Theory** | 40 | 62.5 |
| Geography/climate/weather | 10.5 | 16.4 |
| History | 10 | 15.6 |
| Treatment diffusion | 2.5 | 3.9 |
| Others | 17 | 26.6 |
| **Experiment** | 12 | 18.8 |
| **Rules & policy changes** | 5 | 7.8 |
| Change in exposure | 3 | 4.7 |
| Fuzzy RD | 2 | 3.1 |
| **Econometrics** | 7 | 10.9 |
| Interactions/"Bartik" | 5 | 7.8 |
| Lagged treatment | 1 | 1.6 |
| Empirical test | 1 | 1.6 |
| **Total** | 64 | 100.0 |

***Note:*** One paper uses both geography-based instruments and an instrument based on treatment diffusion from neighbors. We count 0.5 for each category.

# 现有研究中的IV估计

## 大家都用什么当IV?

- 第四类基于计量经济学的假设或理论：滞后项、Bartik IV……
- 随着时间推移，使用随机实验作为IV的研究越来越多，Theory的比例基本不变，使用计量假设或理论的研究减少了。

- Lorentzen et al. (2014)：城市中最大企业的影响力→环境信息透明度，IV：8年前该城市最大企业的影响力。

TABLE 2. TYPES OF IVS

| IV Type | #Papers | Percentage% |
|---|---|---|
| **Theory** | 40 | 62.5 |
| Geography/climate/weather | 10.5 | 16.4 |
| History | 10 | 15.6 |
| Treatment diffusion | 2.5 | 3.9 |
| Others | 17 | 26.6 |
| **Experiment** | 12 | 18.8 |
| **Rules & policy changes** | 5 | 7.8 |
| Change in exposure | 3 | 4.7 |
| Fuzzy RD | 2 | 3.1 |
| **Econometrics** | 7 | 10.9 |
| Interactions/"Bartik" | 5 | 7.8 |
| Lagged treatment | 1 | 1.6 |
| Empirical test | 1 | 1.6 |
| **Total** | 64 | 100.0 |

***Note:*** One paper uses both geography-based instruments and an instrument based on treatment diffusion from neighbors. We count 0.5 for each category.

# 现有研究使用IV估计时出现的问题

## 复制第一阶段F统计量

- 基于四类不同的标准误计算的F统计量：
  - (1) 经典的渐进标准误；
  - (2) Huber-White稳健标准误；
  - (3) 聚类稳健标准误；
  - (4) Bootstrap标准误。

- 作者主要使用基于Bootstrap标准误得到的F统计量。
- 如果数据可以聚类，就使用cluster-bootstrapping。

# 现有研究使用IV估计时出现的问题

**为什么要用Bootstrap标准误？(附录A.3.3)**

- 因为这种标准误是最为保守的，可以减少使用弱工具变量导致错误结论的可能。
- 模拟一个数据生成过程

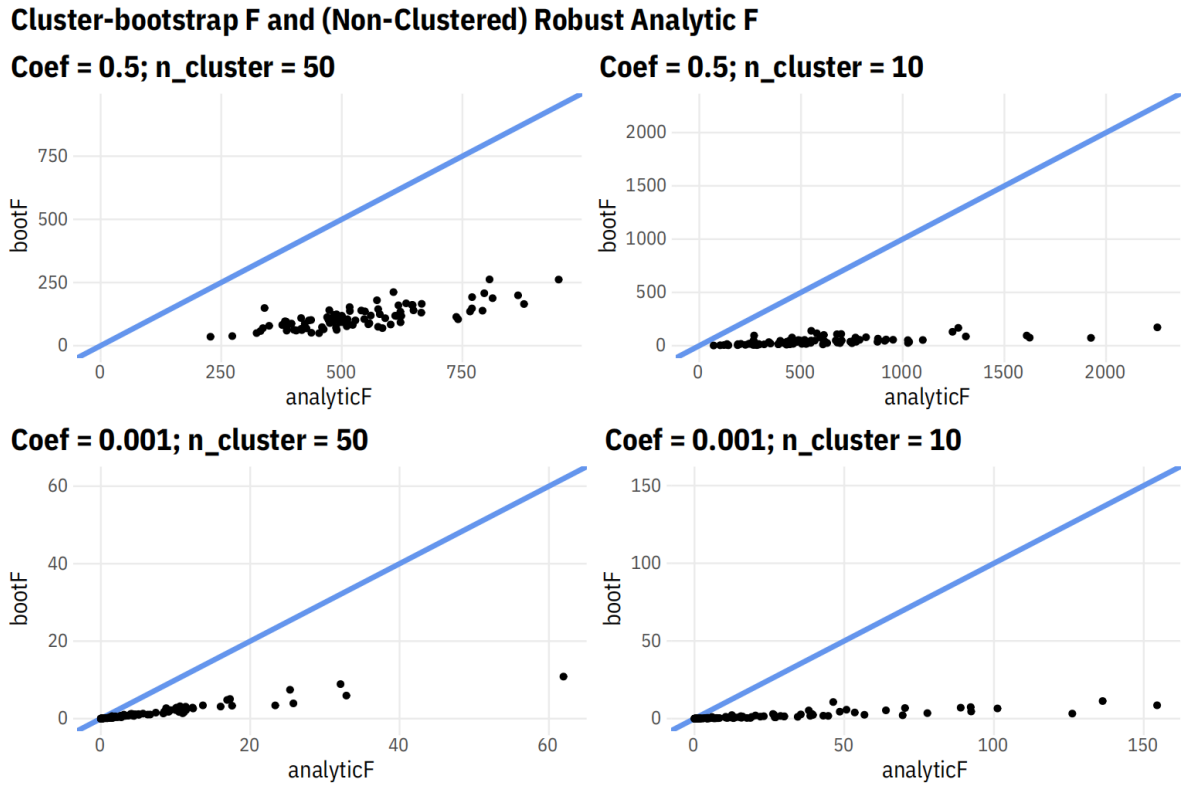
$$\begin{aligned}\text{clustered instrument and error components}\quad & \nu_j, \eta_j \sim \mathcal{N}(0, 0.5)\\ \text{instrument}\quad & z_i \sim \mathcal{N}(0,1) + \nu_j\\ \text{error}\quad & \varepsilon_i \sim \mathcal{N}(0,1) + \eta_j\\ \text{endogenous variable}\quad & x_i = \pi z_i + \varepsilon_i\end{aligned}$$

- 进行两类F统计量的比较
  - 分析型的F统计量：t-test ($H_0: \pi = 0$)。
  - 基于Bootstrap计算的F统计量：$\pi^2/\hat{\sigma}^2$，其中$\hat{\sigma}^2$是抽样估计得到的方差。

# 现有研究使用IV估计时出现的问题

## 为什么要用Bootstrap标准误?

- 未考虑聚类结构的分析型F统计量明显高估了工具变量的相关性。
- 无论是强工具变量还是弱工具变量(coef.=0.5或0.001)，也无论聚类的组数是多还是少(n_cluster=50或10)，分析型F统计量的高估都非常明显。



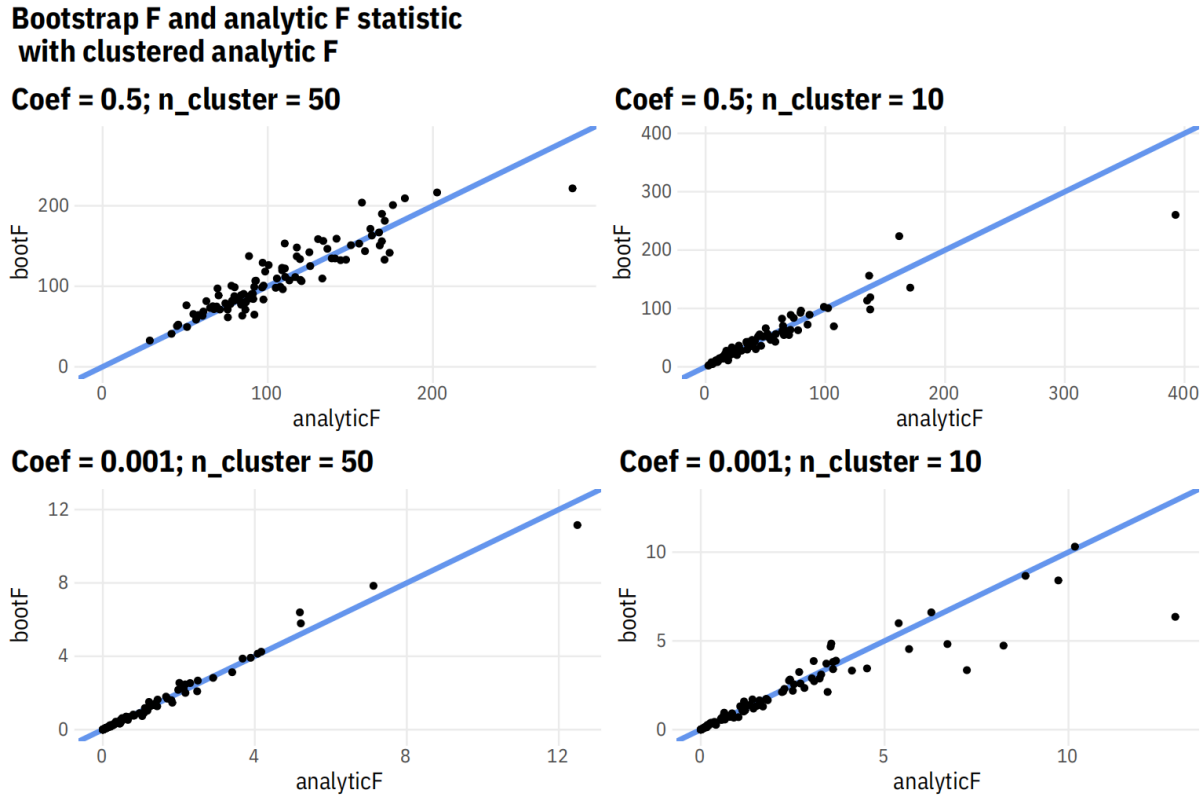

(a) Cluster-bootstrap $F$ statistic vs. Huber-white (non-clustered) $F$ statistic

# 现有研究使用IV估计时出现的问题

## 为什么要用Bootstrap标准误?

- 在正确聚类的情况下，两类F统计量基本上是等价的。
- 并且识别弱工具变量的效果都还不错。

- 但是基于Bootstrap标准误的F统计量还是要更加保守一些。




(b) Cluster-bootstrap $F$ statistic vs. cluster-robust analytic $F$ statistic

# 现有研究使用IV估计时出现的问题

**Finding 1. 第一阶段F统计量**

- 64个IV估计中，14个(22%)没有汇报第一阶段F统计量。
- 在剩下的50个IV估计中，10个(20%)使用了经典的渐进标准误(未考虑异方差和组内相关)。

# 现有研究使用IV估计时出现的问题

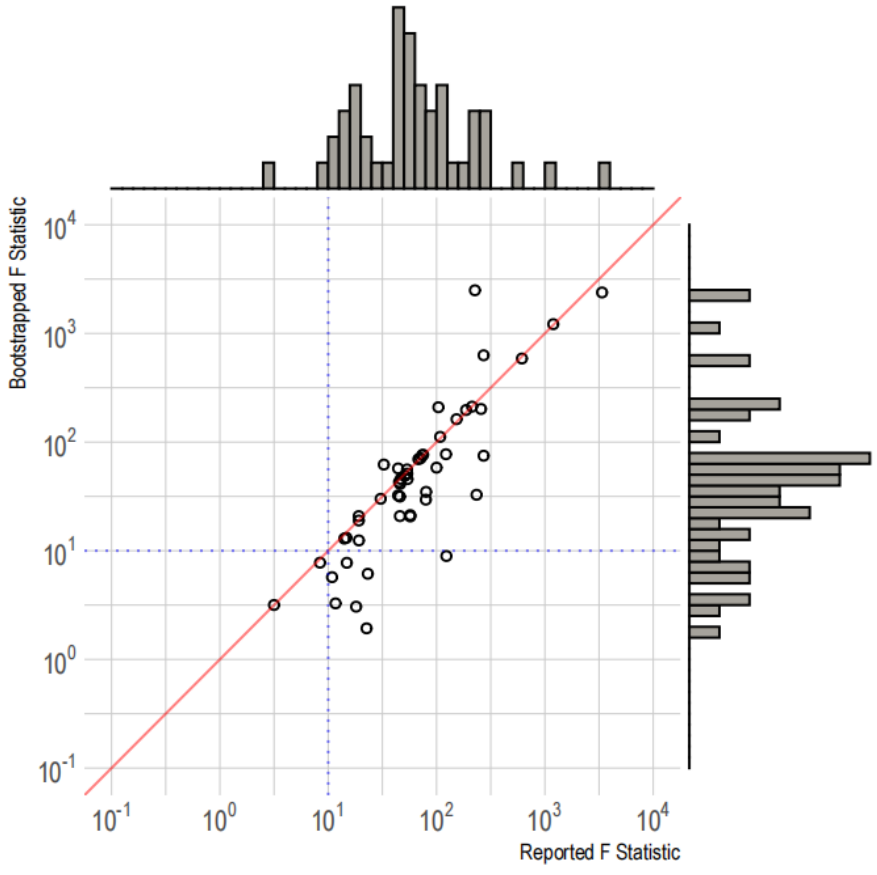
**Finding 1. 第一阶段F统计量**

- 复制的F统计量更大：15个，占30%。
- 原文汇报的F统计量更大：35个，占70%。

- 12个复制的F统计量小于10的IV中：
  - 3个没有汇报F统计量；
  - 7个汇报的F统计量大于10。

- 虽然81%的IV(51个)的复制F统计量大于10，但如果使用更加严格的要求(F≥104.7)，那么只有31%的IV(20个)符合要求。

FIGURE 2. DISTRIBUTIONS OF $F$ STATISTICS AND $z$-SCORES: REPORTED VS. REPLICATED



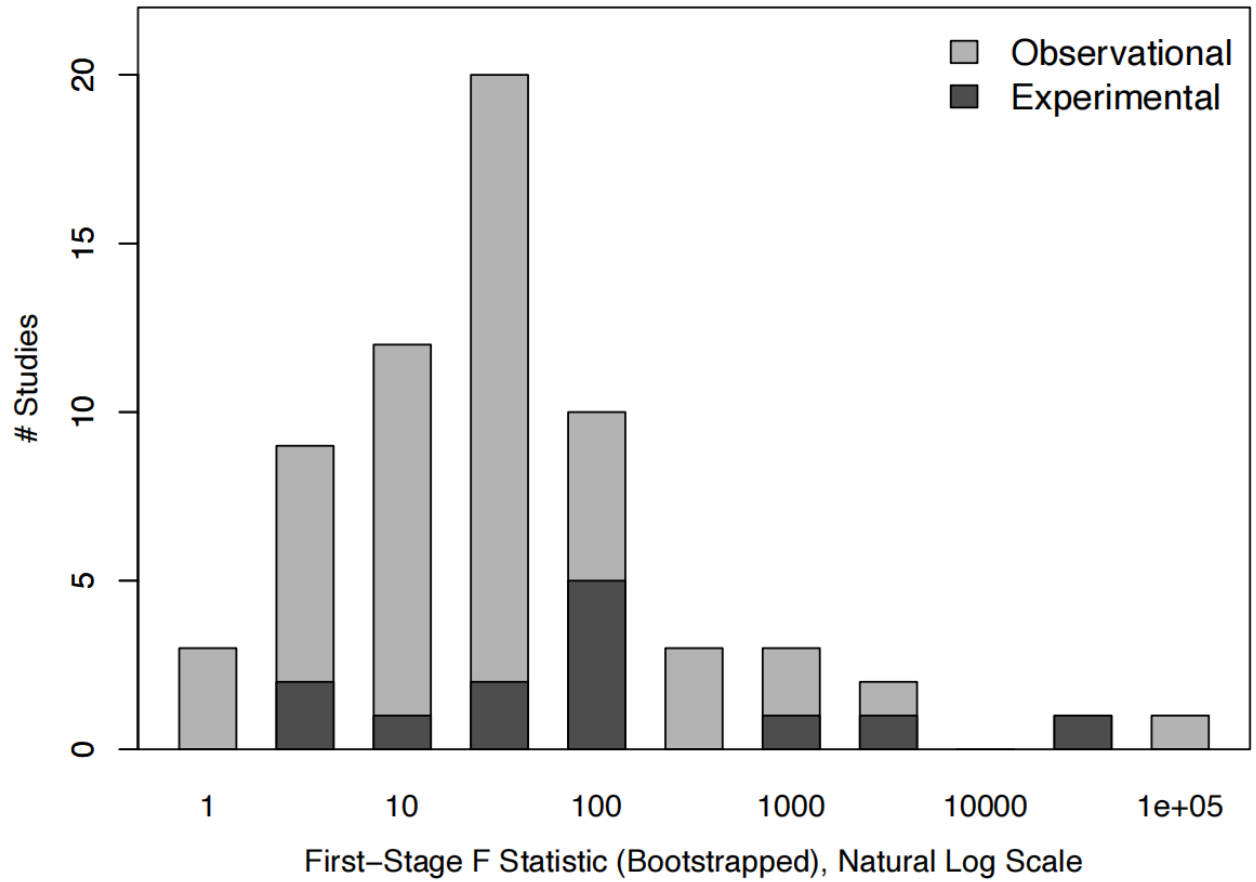

(a) First-Stage Partial $F$-Statistic

# 现有研究使用IV估计时出现的问题

**Finding 1. 第一阶段F统计量**

- 基于实验的工具变量往往更强。
- 复制的F统计量的中位数：
  - 实验：122.5
  - 非实验：41.2

FIGURE A2. HISTOGRAM OF BOOTSTRAPPED FIRST-STAGE *F* STATISTIC



# 现有研究使用IV估计时出现的问题

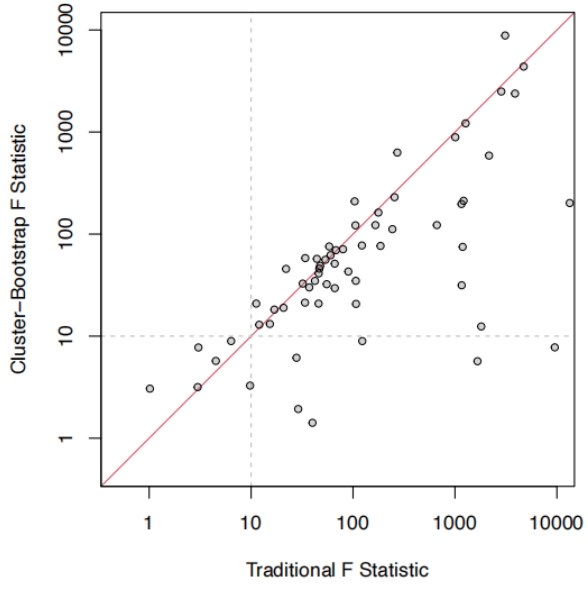
## Finding 1. 第一阶段F统计量(总结)

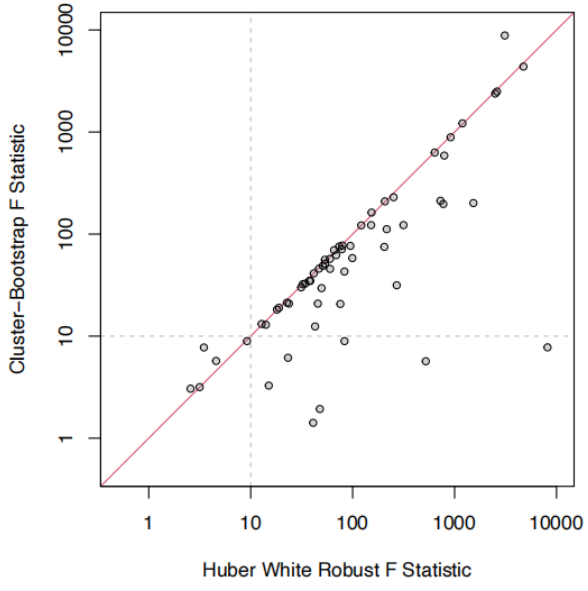
- 22%的研究者(14个)没有汇报F统计量。
  - 不过可能没有那么严重?
  - 不汇报的可能很多是基于实验的IV?

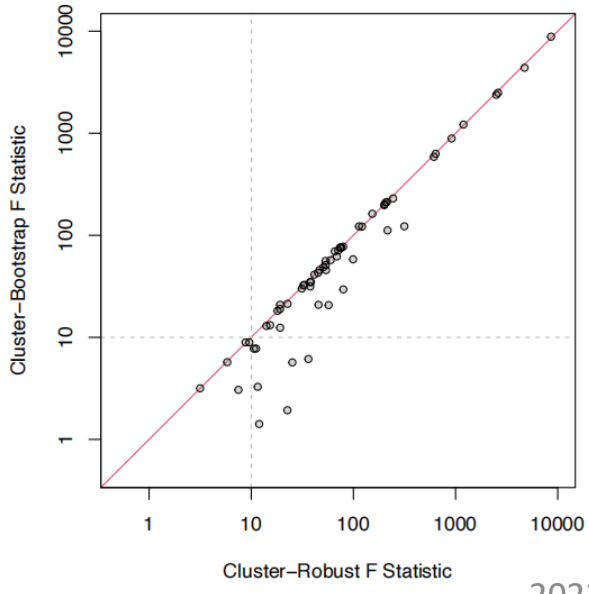
# 现有研究使用IV估计时出现的问题

## Finding 1. 第一阶段F统计量(总结)

- 相当多的研究所汇报的F统计量要大于较为保守的Bootstrap统计量。
  - 保守就好吗?
  - 应该说，很多研究不是保守不保守的问题，而是存在可以聚类的结构但是却没有考虑的问题，从而导致高估了工具变量的相关性。



# 现有研究使用IV估计时出现的问题

**复制z-score**

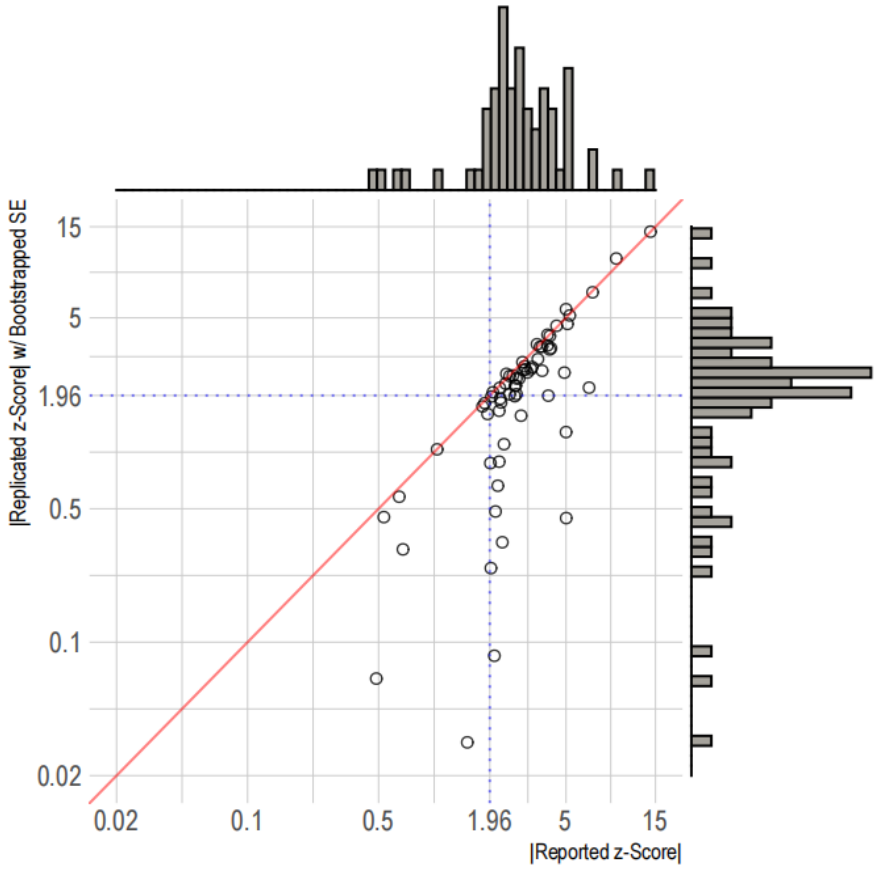
- z-score的计算：$z = \hat{\beta}_{2SLS} / \widehat{SE}(\hat{\beta}_{2SLS})$
- 复制时依然使用Bootstrap标准误进行比较。

# 现有研究使用IV估计时出现的问题

**Finding 2. 统计推断**

- 存在明显的p值操纵(p-hacking)：汇报的z-score集中在1.96附近(**)。
- 与汇报的z-score相比，基于Bootstrap标准误得到的z-score往往更小(主要是因为这些研究中使用了渐进标准误)。
- 很多原文中显著的结果其实是不显著的，设定5%的显著性水平，不显著的IV估计从9个增加到了26个(占41%)。

FIGURE 2. DISTRIBUTIONS OF $F$ STATISTICS AND $z$-SCORES: REPORTED VS. REPLICATED



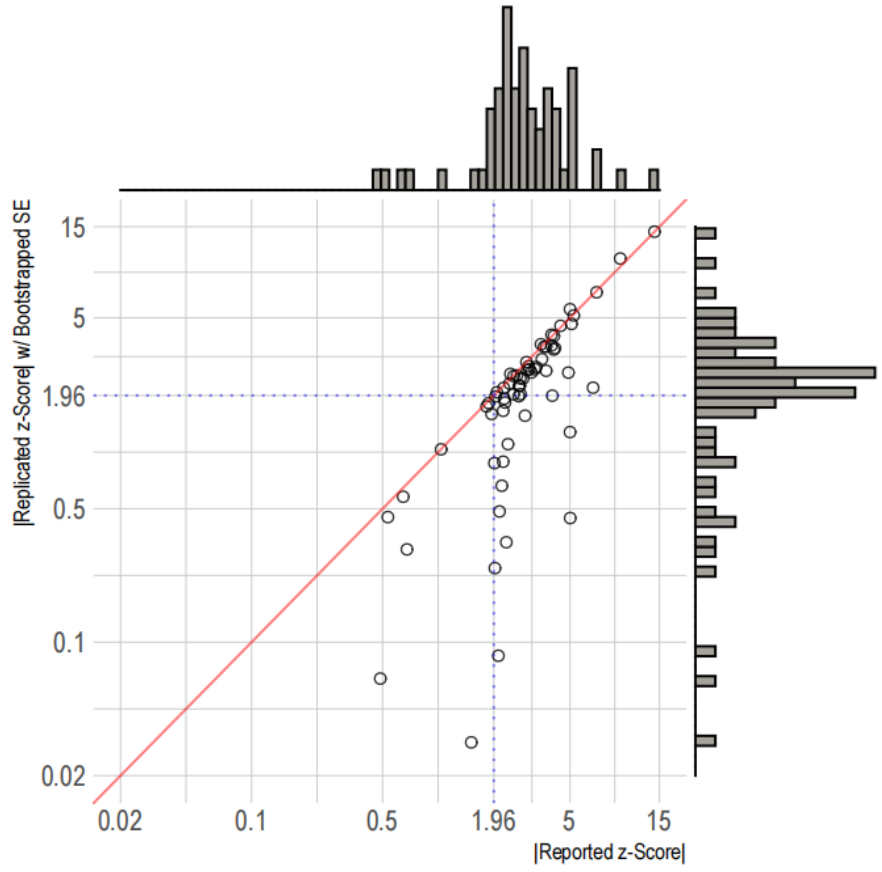

(b) $z$-Score for 2SLS Estimates

# 现有研究使用IV估计时出现的问题

**Finding 2. 统计推断**

- Lee et al. (2020)：如果仅仅以F⩾10作为强工具变量的标准，那么z-score要大于3.43才能有5%的显著性。
- 依据这一标准，只有14个IV(占21.9%)符合这个要求。

FIGURE 2. DISTRIBUTIONS OF $F$ STATISTICS AND $z$-SCORES: REPORTED VS. REPLICATED




(b) $z$-Score for 2SLS Estimates

# 现有研究使用IV估计时出现的问题

**Finding 2. 统计推断(总结)**

- 由于很多研究错误的估计了标准误(比如使用渐进标准误)，导致标准误被低估了，从而夸大了核心解释变量系数的显著性。

- 尽管有些研究的工具变量并不是很强，大多数研究也没有尝试使用一些专门为弱工具变量设计的检验或方法，比如：
  - Anderson–Rubin test (2个);
  - The conditional likelihood–ratio test (Moreira, 2003) (1个);
  - Constructs confident sets (Mikusheva and Poi, 2006) (没有)。

# 现有研究使用IV估计时出现的问题

**比较OLS和2SLS的估计系数**

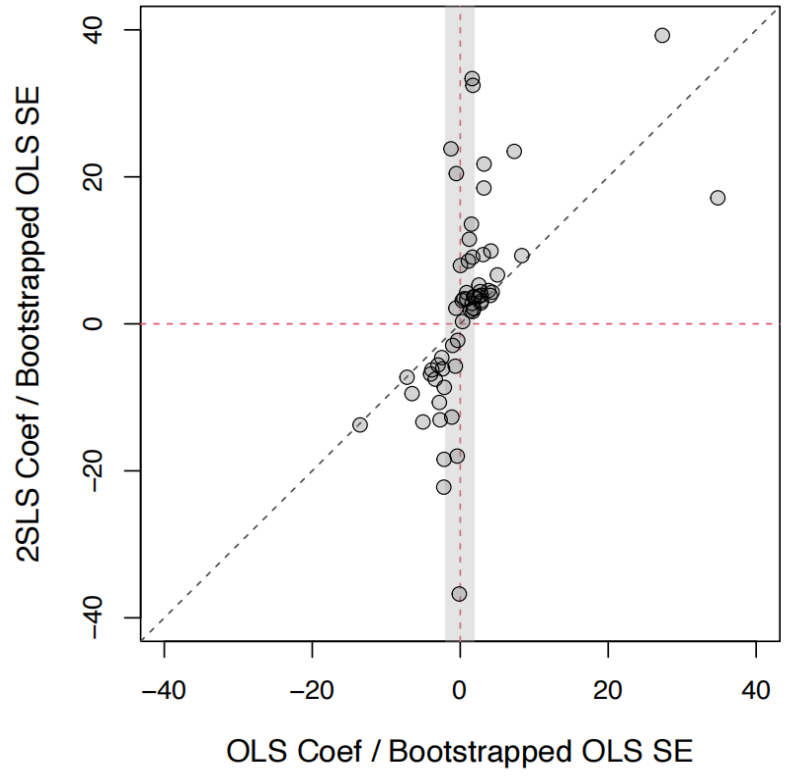
- 首先计算出OLS和2SLS的点估计值；
- 然后使用OLS中的Bootstrap标准误对两个点估计值进行标准化。

# 现有研究使用IV估计时出现的问题

## Finding 3. OLS和2SLS的点估计值差异

- OLS和2SLS估计出来的系数符号基本上是一致的(94%，60/64)。
- 但是，2SLS估计得到的x对y的影响基本都大于OLS估计出的影响(92%，59/64)。

FIGURE 3. RELATIONSHIP BETWEEN OLS AND 2SLS ESTIMATES



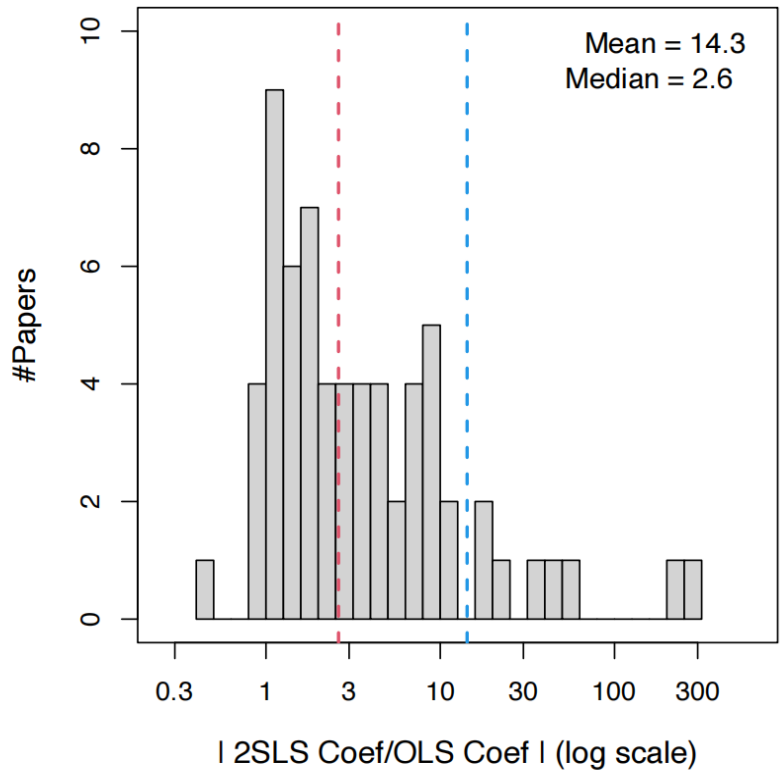

(a) Scatterplot

# 现有研究使用IV估计时出现的问题

## Finding 3. OLS和2SLS的点估计值差异

- 从绝对量上来看，2SLS估计得到的y对x的影响平均是OLS估计的14.3倍，中位数是2.6倍。

FIGURE 3. RELATIONSHIP BETWEEN OLS AND 2SLS ESTIMATES




(b) Ratio Histogram

# 现有研究使用IV估计时出现的问题

**Finding 3. OLS和2SLS的点估计值差异**

- 有可能是因为工具变量的排他性假设不满足!

Because plim $\widehat{\beta}_{2SLS} = \beta + Bias_{IV}$ and plim $\widehat{\beta}_{OLS} = \beta + Bias_{OLS}$, we have

$$\text{plim} \left| \frac{\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS}}{\widehat{\beta}_{OLS}} \right| = \left| \frac{Bias_{2SLS}}{\beta + Bias_{OLS}} \right| \leq \left| \frac{Bias_{2SLS}}{Bias_{OLS}} \right| = \left| \frac{\rho(Z,\varepsilon)}{\rho(X,\varepsilon)} \right| \cdot \frac{1}{|\rho(X,\hat{X})|}, \qquad (4.1)$$

- 如果满足排他性假设，即$\rho(Z,\varepsilon) = 0$，那么$\left|\frac{\widehat{\beta}_{2SLS}-\widehat{\beta}_{OLS}}{\widehat{\beta}_{OLS}}\right|$与$|\rho(X,\hat{X})|$不应该展现出相关性。

↑

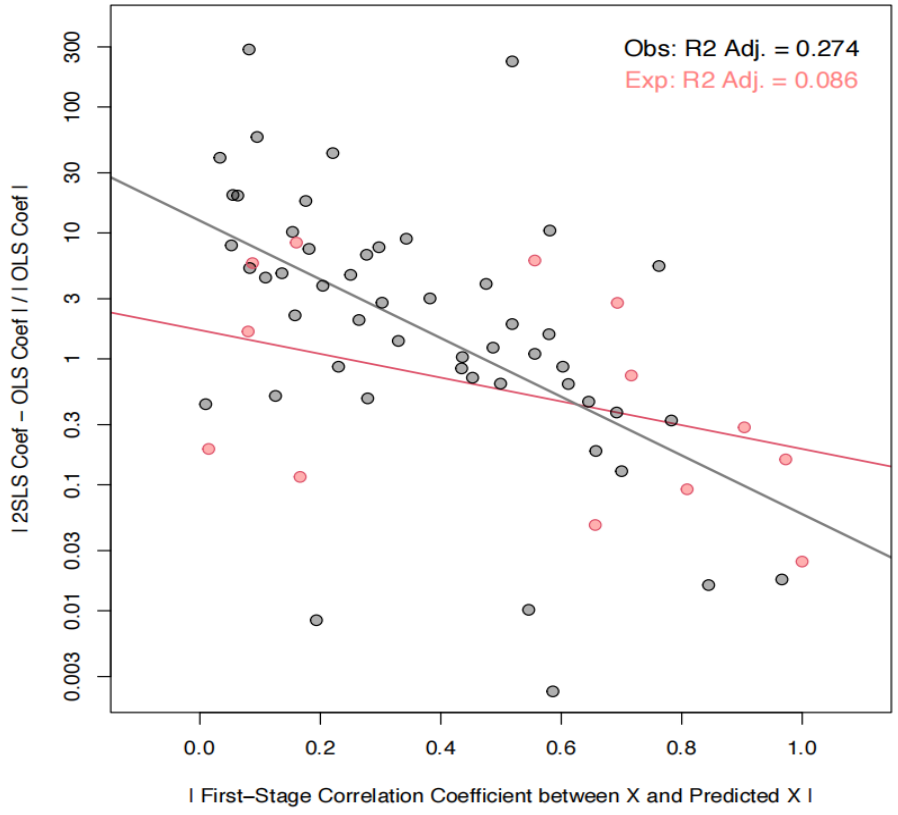
体现IV的强弱

# 现有研究使用IV估计时出现的问题

## Finding 3. OLS和2SLS的点估计值差异

- 但实际上IV的强弱却会影响偏差的大小，尤其是在观测数据中。
  - 观测数据：$R^2$=0.274；
  - 实验数据：$R^2$=0.086。

FIGURE 4. IV STRENGTH AND OLS-IV DISCREPANCY




(a) Full Sample

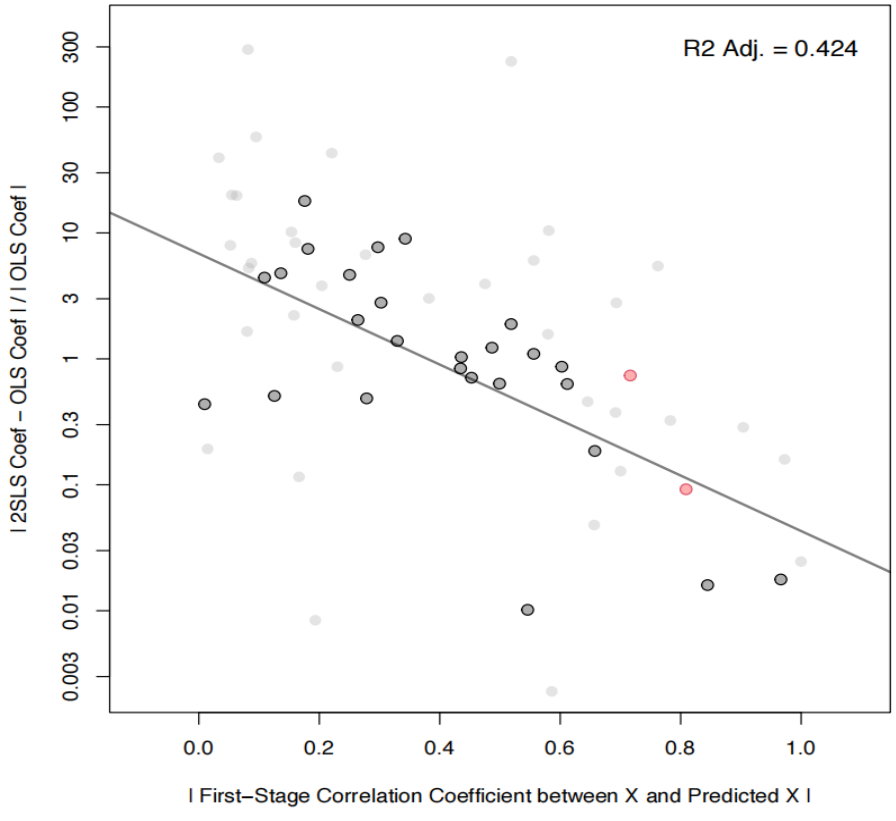
# 现有研究使用IV估计时出现的问题

## Finding 3. OLS和2SLS的点估计值差异

- 即使仅仅考虑那些汇报的OLS结果也显著，而且OLS和2SLS的估计结果符号相同的IV估计，这种负相关关系依然存在。

- 这些结果说明，2SLS估计结果普遍大于OLS估计结果(尤其是大很多的时候)是因为排他性假设不完全满足时，IV估计存在的偏误被自身的弱相关性放大了。

FIGURE 4. IV STRENGTH AND OLS-IV DISCREPANCY




(b) Subsample with Significant OLS Results

# 现有研究使用IV估计时出现的问题

**Finding 3. OLS和2SLS的点估计值差异——有没有其他可能的解释?**

- 解释一：异质处理效应
  - IV估计是对complier的LATE，有可能处理就是会对complier影响更大。
  - 19篇(占31%)研究据此解释2SLS的估计结果。

- 解释二：发表偏误
  - 弱工具变量下2SLS估计的波动会很大，只有那些往大的方向波动的实证结果可能显著，进而被保留下来。

# 现有研究使用IV估计时出现的问题

**Finding 3. OLS和2SLS的点估计值差异——解释一和二并不能完全解释！(附录A.3.2)**

- 模拟一个考虑异质处理效应的数据生成过程。

$$
\begin{aligned}
y_i &= 5 + \beta_i x_i + (u_i + b_i) \\
x_i^* &= (\kappa\pi_i) z_i + \left(0.2 v_i + \sqrt{1-(\kappa\pi_i)^2} \cdot a_i\right) \\
x_i &= 1\{x_i^* > 0\},\ z_i \overset{i.i.d.}{\sim} Bern(0, 0.5) \qquad \text{(binary-binary case)}
\end{aligned}
$$

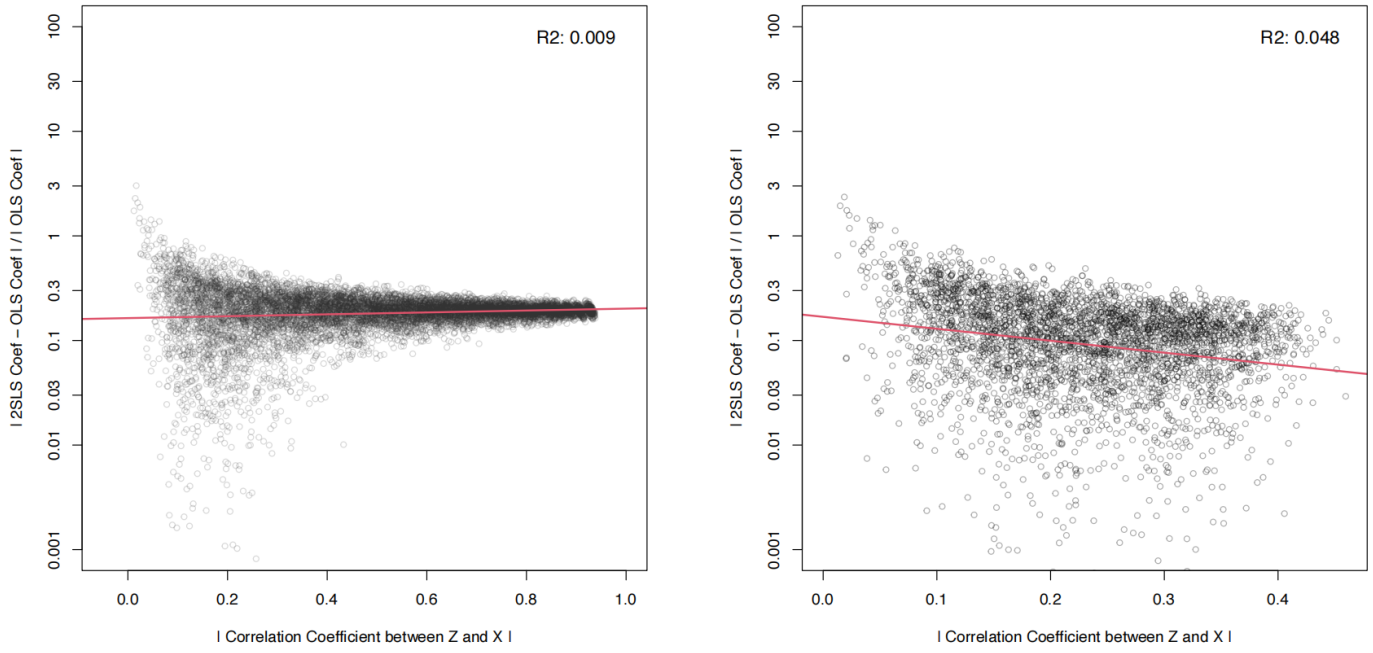
$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \overset{i.i.d.}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$; $a_i \overset{i.i.d.}{\sim} N(0,1)$, $b_i \overset{i.i.d.}{\sim} N(0,1)$ are i.i.d. errors; first stage and reduced form coefficients may be correlated, i.e., $\begin{bmatrix} \pi_i \\ \beta_i \end{bmatrix} \overset{i.i.d.}{\sim} N\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \lambda \\ \lambda & 0.5 \end{bmatrix}\right)$, in which $\sigma$ controls the amount of heterogeneity in $\beta_i$ and $\pi_i$ while $\lambda$ controls their correlation. In addition, we use $\kappa$ to control the strength of the instrument. The sample size is 1,000.

# 现有研究使用IV估计时出现的问题

**Finding 3. OLS和2SLS的点估计值差异——解释一和二并不能完全解释！(附录A.3.2)**

- (a)图：无异质处理效应($\sigma^2 = 0$)，工具变量的强度与偏差大小的负相关关系不存在。
- (b)图：存在很强的异质处理效应($\sigma^2 = 0.1, \lambda = 0.7$)，负相关关系存在，但是$R^2$只有0.048。

FIGURE A7. TREATMENT STRENGTH AND OLS-IV DISCREPANCY CONTINUOUS-CONTINUOUS CASE



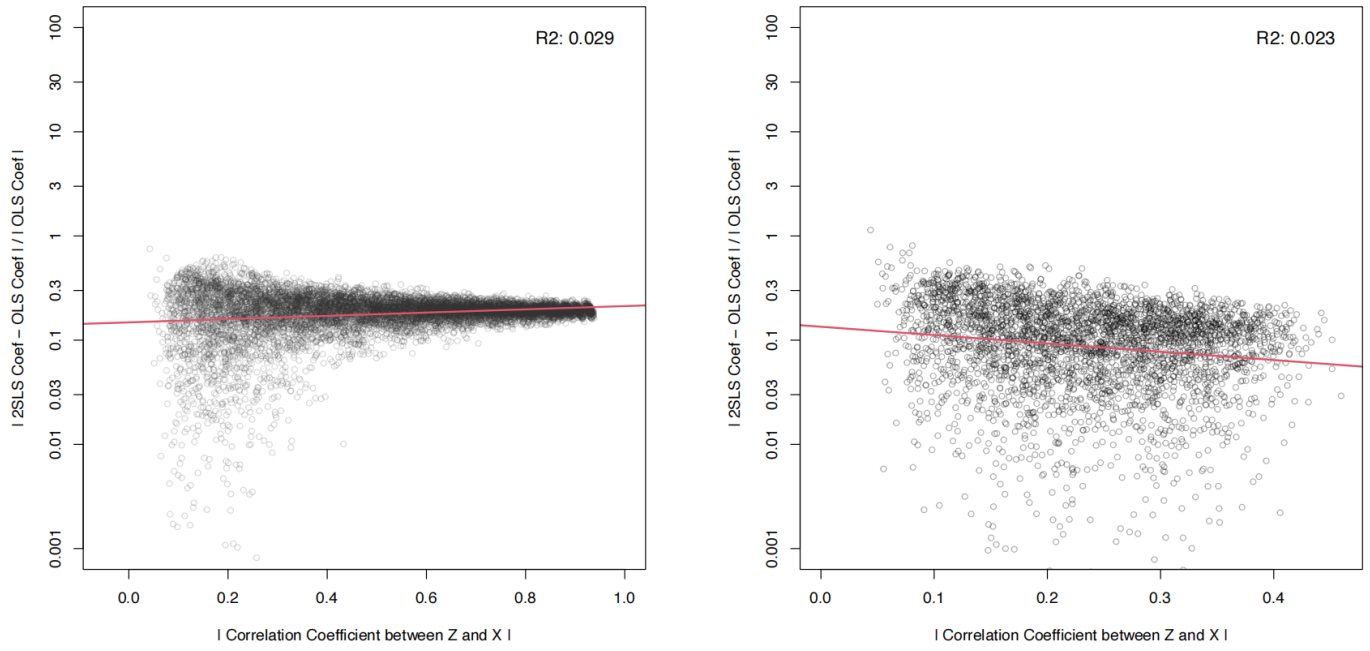

(a) Constant Effect ($\sigma^2 = 0$)　　(b) Heterogeneous Effect ($\sigma^2 = 0.1, \lambda = 0.7$)

# 现有研究使用IV估计时出现的问题

**Finding 3. OLS和2SLS的点估计值差异——解释一和二并不能完全解释！(附录A.3.2)**

- 考虑发表偏误：仅保留5%显著的结果(小尾巴被删掉了)。
- 负相关关系被削弱了。
- 因此，异质处理效应和发表偏误无法全部解释。

FIGURE A9. TREATMENT STRENGTH AND OLS-IV DISCREPANCY
CONTINUOUS-CONTINUOUS CASE: 5% SIGNIFICANCE




(a) Constant Effect ($\sigma^2 = 0$)　(b) Heterogeneous Effect ($\sigma^2 = 0.1, \lambda = 0.7$)

# 现有研究使用IV估计时出现的问题

**Finding 3. OLS和2SLS的点估计值差异——有没有其他可能的解释?**

- 解释三：存在向下的测量误差
  - 向下的测量偏误越大，工具变量就会越弱，同时2SLS和OLS估计结果的差值也会越大。
  - 但是，在作者复制的这些研究中，很少有研究说，存在测量误差是使用工具变量的主要原因。
  - 而且，其实研究中更担心的是OLS估计向上偏误，即高估x对y的影响。
  - 以及，Figure 4表明，即使考虑那些OLS显著且与2SLS同符号的研究(即OLS的内生性没那么严重，IV仅作为稳健性检验)，依然存在这种关系，这就说明不太可能是因为存在向下的测量误差。

# 现有研究使用IV估计时出现的问题

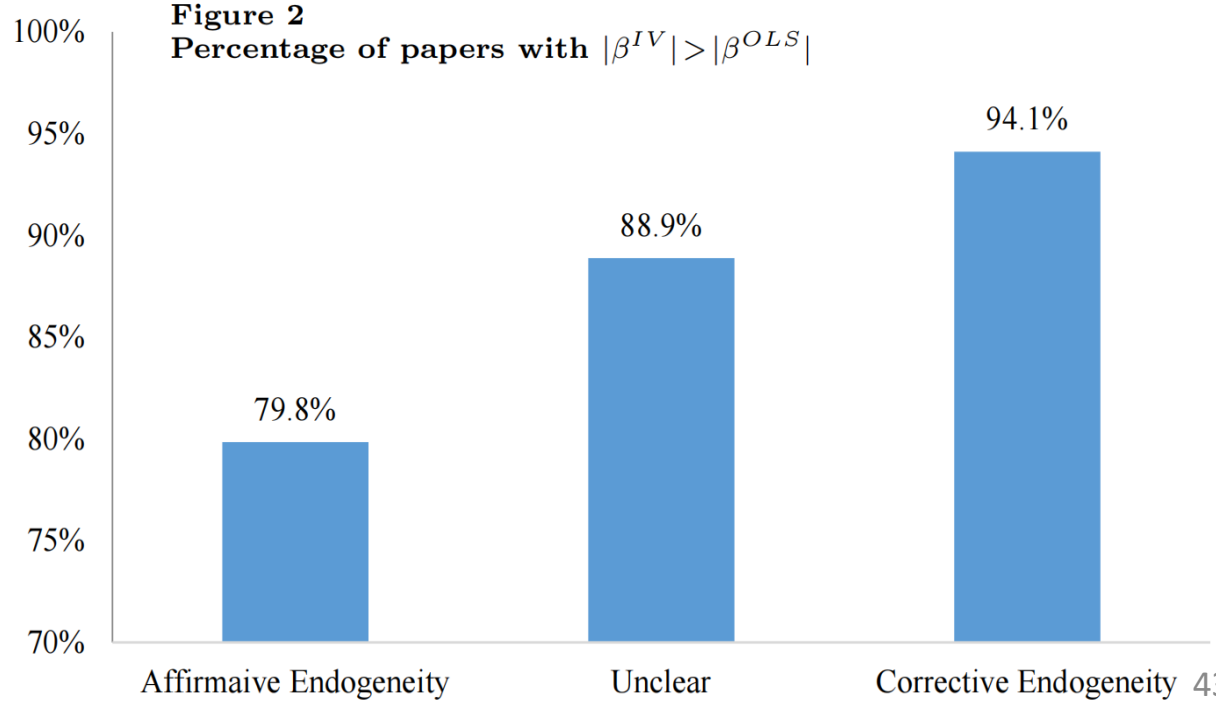
**Finding 3. OLS和2SLS的点估计值差异(总结)**

- 2SLS估计结果普遍大于OLS估计结果，是因为排他性假设不完全满足时，IV估计存在的偏误被自身的弱相关性放大了。
- 反映了研究者对于工具变量的使用不太谨慎，在IV的排他性约束普遍不完全满足的时候，却错误使用了标准误导致高估了工具变量的相关性，忽略了弱工具变量带来的潜在问题。

# 拓展：经济学研究中呢?

**"Have Instrumental Variables Brought Us Closer to the Truth" (Jiang, 2017)**

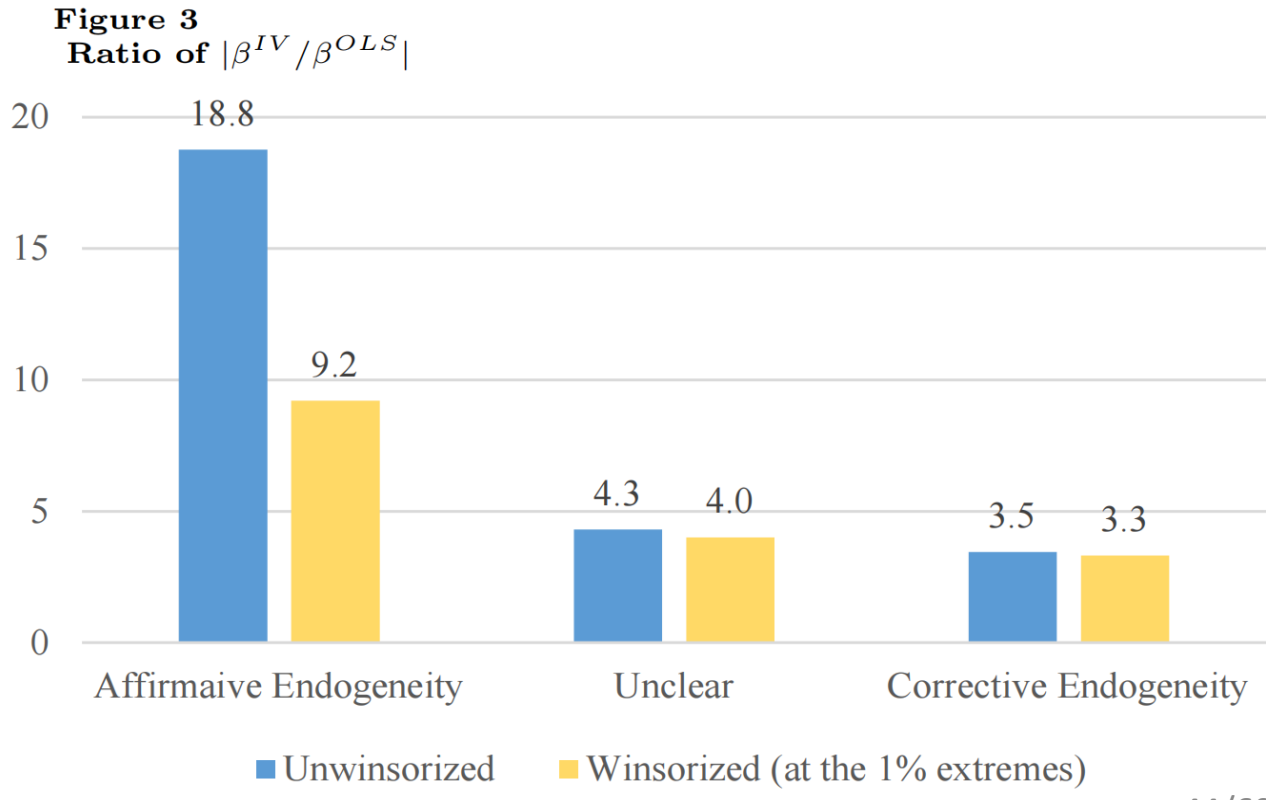
- 考察了2003年-2014年发表在金融学三大刊(JF, JFE, RFS)上的255篇使用IV的研究。
- 根据作者的分类，255篇研究中，有171篇(占67.1%)属于“支持型内生性(affirmative endogeneity)”，46篇(占18%)属于“纠正型内生性(corrective endogeneity)”，剩下38篇(占14.9%)属于无法判断的。
- 但属于“支持型内生性”的研究中依然有高达79.8%的研究是$|\beta^{IV}| > |\beta^{OLS}|$的。



# 拓展：经济学研究中呢?

"Have Instrumental Variables Brought Us Closer to the Truth" (Jiang, 2017)

- 平均的偏离程度是很大的，甚至在属于“支持型内生性”的研究中，偏离的反而最厉害。



# 拓展：经济学研究中呢?

"Have Instrumental Variables Brought Us Closer to the Truth" (Jiang, 2017)

- 解释一：可能是因为IV估计的是LATE。
- 金融学中的工具变量常常是供给侧发生的创新，而compliers往往正是那些对成本敏感的群体：在原来的条件下，成本大于收益，它们选择不受到处理；而在新的条件下，成本约束减小了，于是收益大于成本了，他们选择接受处理。
- 但是对于那些从处理中获得的收益比较小的群体而言，即使成本约束减少了，收益依然无法cover成本，于是他们不接受处理，成为never-takers。正因如此，LATE要更大。
- “如果在已发表的研究中发现的因果效应集中在LATE明显大于ATE的子样本中，那么我们对经济关系的学习，将永远不会收敛到人口中心普遍存在的情况。”

# 拓展：经济学研究中呢?

"Have Instrumental Variables Brought Us Closer to the Truth" (Jiang, 2017)

- 解释二：弱工具变量
- “工具变量越弱，对于工具变量的纯度(purity)的要求就要越高。”

$$\beta_{IV}=\frac{cov(\beta x_i+\iota z_i+\eta_i,z_i)}{cov(x_i,z_i)}=\frac{\beta cov(x_i,z_i)+\iota var(z_i)}{cov(x_i,z_i)}=\beta+\frac{\iota}{\gamma}. \quad (3)$$

- 解释三：发表偏误

# 如何更好地判断IV的有效性?

## Zero-first-stage tests (Bound and Jaeger, 2000)

- 判断IV有效性的难点在于排他性假设无法直接验证。
- 许多研究通过argue说明工具变量满足排他性，但其实我们是可以做一些检验的。

- ZFS tests是一种安慰剂检验：
  - 使用一个其他的样本进行回归，在这个样本中，工具变量并不影响个体是否受到处理(因此，在first-stage中，z对x的影响是zero，所以叫ZFS)。
  - 那么如果排他性假设满足，那么reduced-form估计得到的z对y的影响应该是0。

# 如何更好地判断IV的有效性?

## Zero-first-stage tests：一个例子

- "Long-term persistence" (Guiso et al., 2016)
- 重新考察一个著名的假说：那些在中世纪实现自治(独立)的意大利城市，在今天拥有更高的社会资本(Leonardi et al., 2001)。

- 因果关系：在中世纪是否实现自治→今天的社会资本：
  - x：在中世纪是否实现自治；
  - y：人均慈善捐赠、人均器官捐赠、儿童是否在考试中作弊；
  - z：在中世纪是否是主教所在地。

# 如何更好地判断IV的有效性?

## Zero-first-stage tests：一个例子

- 因果关系：在中世纪是否实现自治→今天的社会资本；
- IV：在中世纪是否是主教所在地。

- 历史背景
  - 公元1000年前，北意大利是罗马帝国的一部分，但之后罗马帝国走向衰弱，慢慢解体，导致在北意大利出现了一些独立的城邦。
  - 独立城邦出现的核心在于建立起了一种“誓约”：城邦的成员互相帮助，共同解决集体问题，共同抵御罗马帝国的王权。主教是誓约的担保人。

  - 南意大利则不同，在1061年至1091年间，诺曼王朝入侵了意大利南部，形成了一个强大的国家，保证了秩序和稳定。
  - 由于诺曼王朝强大的中央权力，南方地区并没有出现自由城邦。

# 如何更好地判断IV的有效性?

Zero-first-stage tests：一个例子

# 如何更好地判断IV的有效性?

## Zero-first-stage tests：一个例子

- 在南意大利，无论一个城市有没有主教，都没有自由城邦出现(never-takers)；
- 因此在南方样本中，IV不可能通过“自由城邦渠道”影响社会资本。
- 在排他性假设成立时，在reduced-form的估计中，南方样本不应该显著。

TABLE 6. Validating the instrument.

**A. Regressions of civic capital in the Center–North and in the South**

| | Center–North sample | | | South sample | | |
|---|---|---|---|---|---|---|
| | (I) Nonprofit org. | (II) Organ donation org. | (III) Cheating in mathe-matics | (IV) Nonprofit org. | (V) Organ donation org. | (VI) Cheating in mathe-matics |
| Ease of coordination | 1.61** | 0.47*** | −0.66*** | 0.18 | 0.19*** | −0.04 |
| | (0.219) | (0.047) | (0.118) | (0.137) | (0.065) | (0.309) |
| Elevation | 1.93*** | −0.25*** | 0.94** | 1.43*** | −0.04 | 0.72 |
| | (0.475) | (0.062) | (0.441) | (0.257) | (0.083) | (0.541) |
| Max difference in | 1.35*** | 0.01 | 0.26* | −0.08 | −0.05* | 0.06 |

# 如何更好地判断IV的有效性?

## 从ZFS tests到"local-to-zero (LTZ)" correction

- Van Kippersluis and Rietveld (2018)提出，可以将ZFS的思路和"plausibly exogenous"方法(Conley et al., 2012)结合起来。

$$Y = X\beta + Z\gamma + \varepsilon$$
$$X = Z\Pi + \nu$$

- Conley et al. (2012)证明，当IV不太外生， 也就是$\gamma$服从某个分布$\mathbb{F}$的时候，2SLS的估计结果满足:

$$\widehat{\boldsymbol{\beta}} \sim^{a} \mathcal{N}\left(\boldsymbol{\beta}, \mathbb{V}_{2SLS}\right) + \mathbf{A}\gamma \quad \text{where } \mathbf{A} \equiv (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \tag{5.2}$$

$$\widehat{\boldsymbol{\beta}} \sim^{a} \mathcal{N}\left(\boldsymbol{\beta} + \mathbf{A}\mu_{\gamma}, \mathbb{V}_{2SLS} + \mathbf{A}\boldsymbol{\Omega}\mathbf{A}'\right) \tag{5.3}$$

- 思路：既然已经使用ZFS tests近似将$\gamma$估计出来了，那么就可以把$\gamma$纳入2SLS的估计中，将$\hat{\gamma}_{ZFS}$作为$\mu_{\gamma}$。

# 如何更好地判断IV的有效性?

## 从ZFS tests到"local-to-zero (LTZ)" correction

- "plausibly exogenous"方法可以通过Stata中的plausexog命令(Clarke, 2014)来实现(ssc install即可)。
- 在徐轶青老师他们的这篇文章中，他们也提供了一个类似的R包"ivDiag"(accompanying this paper)(但我没找到)。

# 如何更好地判断IV的有效性?

## LTZ correction：案例(Guiso et al., 2016)

TABLE 6. Validating the instrument.

**A. Regressions of civic capital in the Center–North and in the South**

| | Center–North sample | | | South sample | | |
|---|---|---|---|---|---|---|
| | (I) Nonprofit org. | (II) Organ donation org. | (III) Cheating in mathe-matics | (IV) Nonprofit org. | (V) Organ donation org. | (VI) Cheating in mathe-matics |
| Ease of coordination | 1.61** | 0.47*** | −0.66*** | 0.18 | 0.19*** | −0.04 |
| | (0.219) | (0.047) | (0.118) | (0.137) | (0.065) | (0.309) |
| Elevation | 1.93*** | −0.25*** | 0.94** | 1.43*** | −0.04 | 0.72 |
| | (0.475) | (0.062) | (0.441) | (0.257) | (0.083) | (0.541) |
| Max difference in | 1.35*** | 0.01 | 0.26* | −0.08 | −0.05* | 0.06 |

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 |

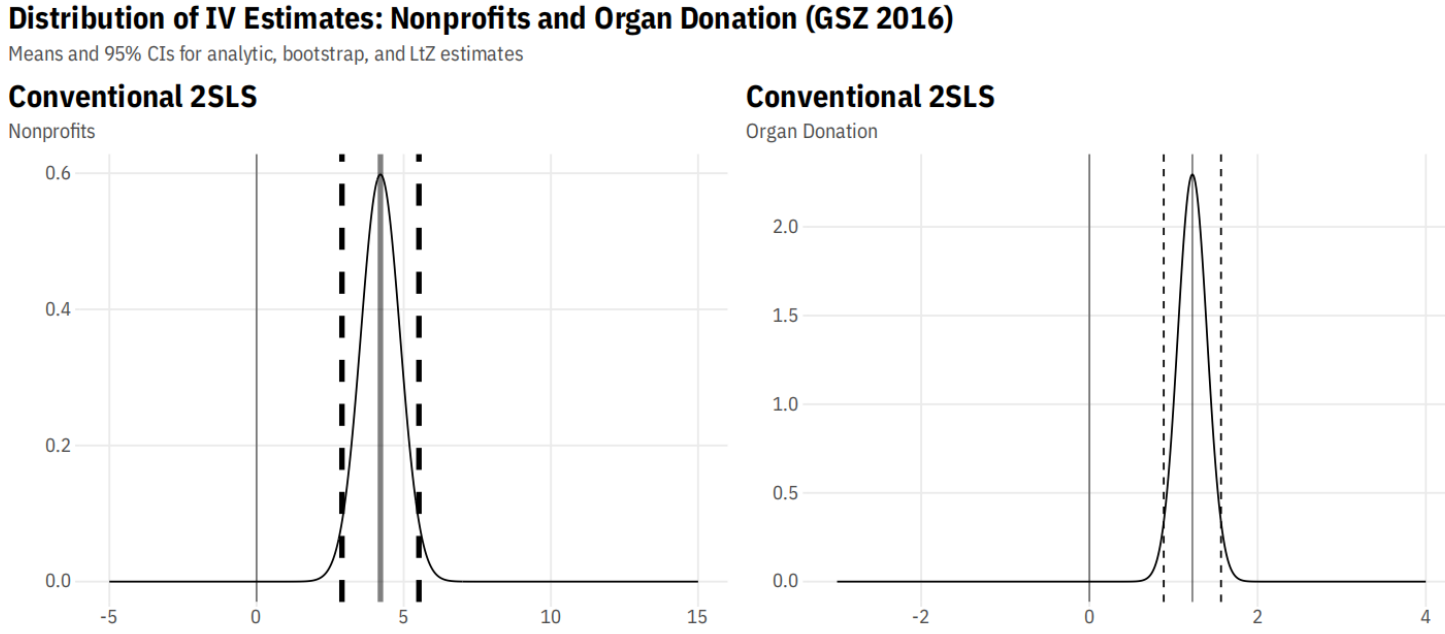
***Note:*** Bootstrapped SEs are in the parentheses. See Figure A5 in the SM for the original table.

# 如何更好地判断IV的有效性?

## LTZ correction：案例(Guiso et al., 2016)

- 经典的稳健标准误低估了估计结果的不确定性。

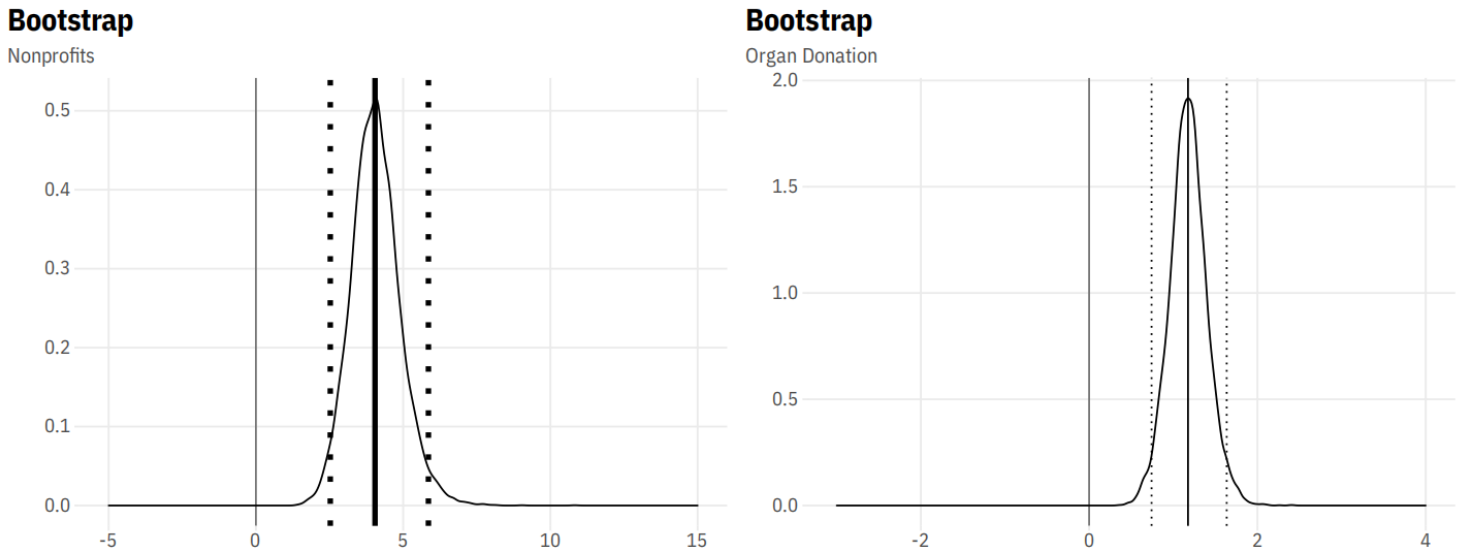
FIGURE 5. IV COEFFICIENTS FOR NON-PROFITS PER CAPITA AND ORGAN DONATION



# 如何更好地判断IV的有效性?

## LTZ correction：案例(Guiso et al., 2016)

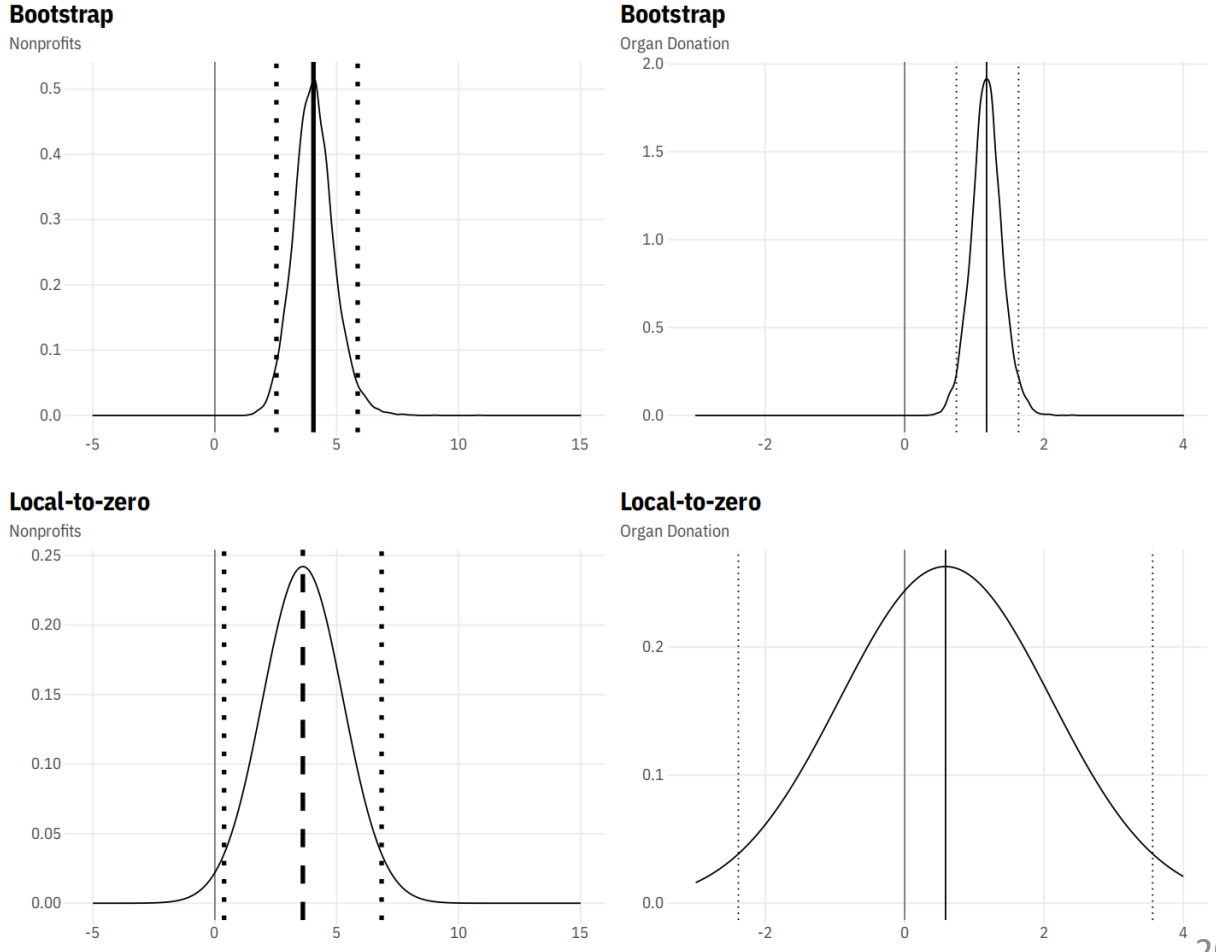
- 使用Bootstrap稳健标后，置信区间变大了，但依然显著(原文汇报的结果)。



# 如何更好地判断IV的有效性?

## LTZ correction：案例(Guiso et al., 2016)

- 使用LTZ方法之后，系数估计值有所减小，标准误进一步增大。
- 前者依然显著，但显著程度有所下降，而后者变得不再显著。



# 总结：使用IV的十个建议

## 十个建议

- 核心思想：观测数据的IV很难像实验IV那样在设计上确保外生，所以使用时要更加谨慎！

- (1)在使用OLS的时候，想清楚潜在的内生性是会使得x对y的影响被高估还是低估。如果担心的是OLS低估x对y的影响，那么可能没必要使用IV估计。
(Jiang, 2017)建议大家在使用IV之前花些篇幅对可能存在的偏误进行讨论。

- (2)在研究设计的阶段，就要想好所选择的IV能不能对x产生(准)随机的冲击。Rubin说过，在因果推断的研究中，研究设计比后续分析更重要("design trumps analysis")。

- (3)跑完第一阶段的回归之后，画图看看$x$和$\hat{x}$之间的关系(排除了协变量和固定效应之后)，瞪大眼睛目测一下IV的强度。

# 总结：使用IV的十个建议

## 十个建议

- (4)利用Bootstrap标准误来计算第一阶段的F统计量，如果数据是可以聚类的，那么应该使用聚类Bootstrap。如果第一阶段的F统计量很大(比如说F>104.7)，那么可以继续。(Jiang, 2017)工具变量的强度要公开透明，别只汇报一个包含很多控制变量的$R^2$。

- (5)类似地，也要用Bootstrap方法来计算2SLS的标准误和置信区间。同时使用一些针对弱IV的检验方法，比如说AR test。

- (6)如果预计OLS会高估x对y的影响，但是2SLS的估计结果依然比OLS高很多，那么就要小心了。

- (7)如果有充分的理由相信，处理对于compliers产生的效应要比其他群体大很多，那么请认真分析这一群体来解释这一情况。

# 总结：使用IV的十个建议

## 十个建议

- (8)因为never-takers是进行ZFS tests的合适群体，如果能够找到这么一个群体，那么请进行安慰剂检验，尝试说明IV不会对这一群体的y产生影响。
- (9)在此基础上，利用ZFS test的结果，使用LTZ correction得到IV的估计值和置信区间，并与原来的估计值和执行区间进行对比。
- (10)(Jiang, 2017)在使用IV估计出x对y的影响之后，一定不要忘记思考系数的现实意义。无论你的系数在统计上多么严谨，如果与现实不太符合，也值得再好好琢磨。

希望对大家之后使用IV有所帮助！