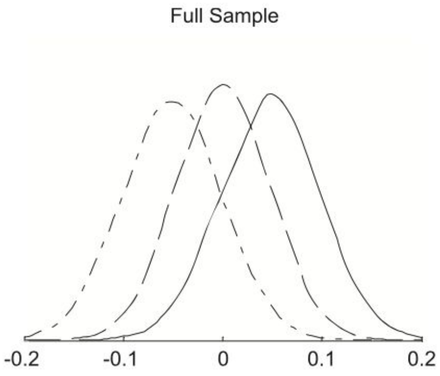
Endogenous Stratification in Randomized Experiment

Gaojia Xu, Guanqi Zeng

1. Introduction

- **Endogenous Stratification**
  - **In-sample data**
  - **Predicted outcome without treatment**
  - **Stratify by intervals of pred outcome**
  - **ATE for each stratus**
- **Bias with predicted pattern!**



Project Outline

- Demonstration & comparison of the bias problem using STAR data
  - Full sample
  - Leave-one-out
  - Sample splitting
  - 10 Fold
- Test the bias problem using STAR-based simulation
- Explore the bias problem & compare the properties of the estimators by computer-generated simulation
  - Sample size (200, 1000, 5000)
  - Number of covariates (10, 20, 40)

2.1 Method - Algorithms

- **Full sample ATE estimation**

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \boldsymbol{x}_i (1 - w_i) \boldsymbol{x}_i' \right)^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i (1 - w_i) y_i$$

- **Leave-one-out ATE estimation**

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \left(\sum_{j \neq i} \boldsymbol{x}_j (1 - w_j) \boldsymbol{x}_j' \right)^{-1} \sum_{j \neq i} \boldsymbol{x}_j (1 - w_j) y_j$$

- **Sample splitting ATE estimation**

$$\widehat{\boldsymbol{\beta}}_m = \left(\sum_{i=1}^{N} \boldsymbol{x}_i (1 - w_i)(1 - v_{im}) \boldsymbol{x}_i' \right)^{-1} \times \sum_{i=1}^{N} \boldsymbol{x}_i (1 - w_i)(1 - v_{im}) y_i.$$

2.1 Method - Algorithms

- **Full sample ATE estimation**

$$\hat{\tau}_k = \frac{\sum_{i=1}^N y_i I_{[w_i=1, c_{k-1} < x_i' \hat{\beta} \le c_k]}}{\sum_{i=1}^N I_{[w_i=1, c_{k-1} < x_i' \hat{\beta} \le c_k]}} - \frac{\sum_{i=1}^N y_i I_{[w_i=0, c_{k-1} < x_i' \hat{\beta} \le c_k]}}{\sum_{i=1}^N I_{[w_i=0, c_{k-1} < x_i' \hat{\beta} \le c_k]}}$$

- **Leave-one-out ATE estimation**

$$\hat{\tau}_k^{LOO} = \frac{\sum_{i=1}^N y_i I_{[w_i=1, c_{k-1} < x_i' \hat{\beta} \le c_k]}}{\sum_{i=1}^N I_{[w_i=1, c_{k-1} < x_i' \hat{\beta} \le c_k]}} - \frac{\sum_{i=1}^N y_i I_{[w_i=0, c_{k-1} < x_i' \hat{\beta}_{(-i)} \le c_k]}}{\sum_{i=1}^N I_{[w_i=0, c_{k-1} < x_i' \hat{\beta}_{(-i)} \le c_k]}}$$

- **Sample splitting ATE estimation**

$$\hat{\tau}_{km}^{SS} = \frac{\sum_{i=1}^N y_i I_{[w_i=1, c_{k-1} < x_i' \hat{\beta}_m \le c_k]}}{\sum_{i=1}^N I_{[w_i=1, c_{k-1} < x_i' \hat{\beta}_m \le c_k]}} - \frac{\sum_{i=1}^N y_i I_{[w_i=0, v_{im}=1, c_{k-1} < x_i' \hat{\beta}_m \le c_k]}}{\sum_{i=1}^N I_{[w_i=0, v_{im}=1, c_{k-1} < x_i' \hat{\beta}_m \le c_k]}} \qquad \hat{\tau}_k^{RSS} = \frac{1}{M} \sum_{m=1}^M \hat{\tau}_{km}^{SS}$$

2.1 Method - Algorithm

- **10 Fold ATE Estimation**

$$\hat{\boldsymbol{\beta}}_{f=j} = (\sum_{i=1}^{N} \boldsymbol{x}_i(1-w_i)I(i \notin fold_j)\boldsymbol{x}_i^{'})^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i(1-w_i)I(i \notin fold_j)y_i$$

$$\hat{\tau_k}^{10fold} = \frac{\sum_{i=1}^{N} y_i I_{[w_i=1, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} \le c_k]}}{\sum_{i=1}^{N} I_{[w_i=1, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} \le c_k]}} - \frac{\sum_{i=1}^{N} y_i I_{[w_i=0, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_{(10fold)} \le c_k]}}{\sum_{i=1}^{N} I_{[w_i=0, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_{(10fold)} \le c_k]}}$$

2.2 Method - Data sets & Simulation

- **STAR**
  - **Outcome Variables: Math test score**
  - **Treatment: Class type** (small class vs. regular-sized class)
  - **Other covariates:** African-American, female, eligibility for the free lunch program, and school attended
- **Simulation Data Sets**
  - **STAR-based simulation**
  - **Computer-generated simulation**

2.2 Method - Simulation: Computer-generated Simulation

$$y_i = 1 + \sum_{l=1}^{40} z_{li} + v_i$$

z_{li} has independent standard normal distributions

v_i has independent normal distribution with var=60

3. Results and Conclusion - STAR estimation

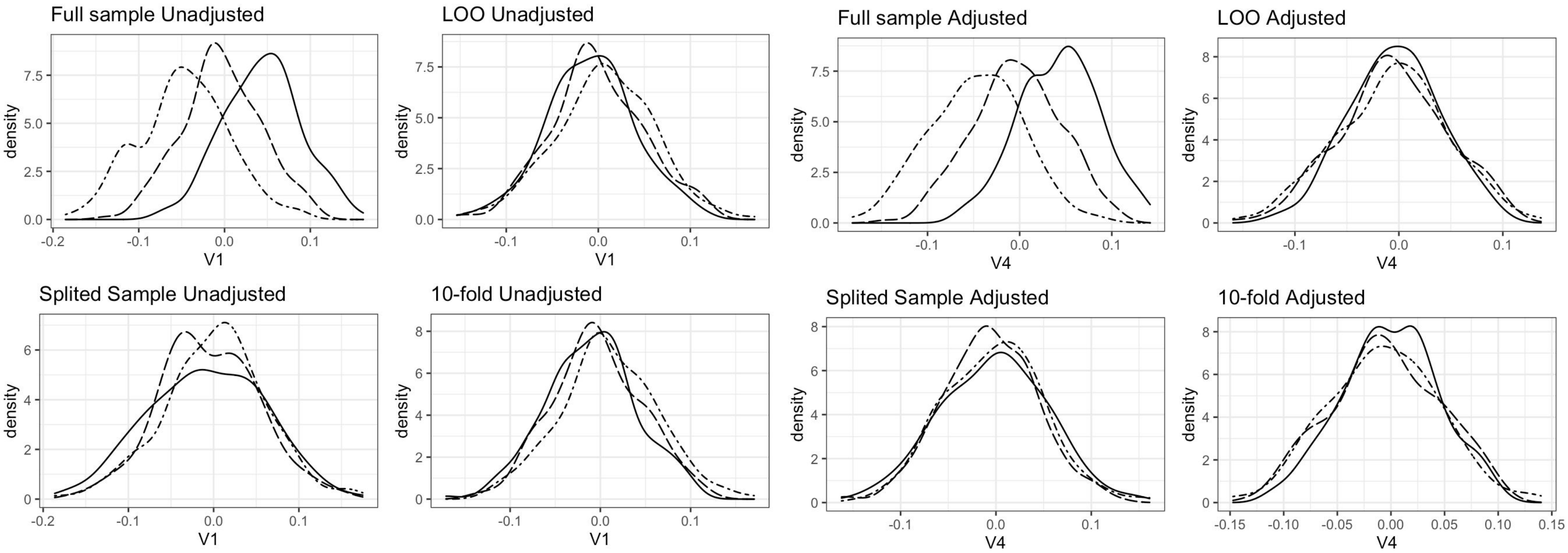
STAR Default Estimation Results

	Unadjusted	Adjusted
$\hat{\tau}$	0.1659	0.1892
se $\hat{\tau}$	0.0334	0.0295

STAR Estimation Results

	Unadj low	Unadj medium	Unadj high	Adj low	Adj medium	Adj high
$\hat{\tau}_k$	0.3705	0.2688	-0.1330	0.3908	0.3023	-0.1242
se $\hat{\tau}_k$	0.0593	0.0618	0.0636	0.0562	0.0621	0.0583
$\hat{\tau}_k^{LOO}$	0.3277	0.2499	-0.0486	0.3440	0.2730	-0.0660
se $\hat{\tau}_k^{LOO}$	0.0603	0.0659	0.0661	0.0565	0.0648	0.0611
$\hat{\tau}_k^{RSS}$	0.1829	0.2083	0.2203	0.1737	0.1933	0.2754
se $\hat{\tau}_k^{RSS}$	0.0446	0.0429	0.0427	0.0389	0.0363	0.0365
$\hat{\tau}_k^{10fold}$	0.3226	0.2809	-0.0440	0.3201	0.2789	-0.0586
se $\hat{\tau}_k^{10fold}$	0.0611	0.0634	0.0649	0.0569	0.0613	0.0604

3. Results and Conclusion - STAR-based simulation



3. Results and Conclusion - STAR-based simulation

Bias in STAR Simulations

	Unadjusted			Adjusted		
	Low	Medium	High	Low	Medium	High
$\hat{\tau}_k$	0.04	0.00	-0.05	0.04	0.00	-0.05
$\hat{\tau}_k^{LOO}$	-0.01	0.00	0.01	0.00	-0.01	-0.01
$\hat{\tau}_k^{RSS}$	-0.01	-0.01	0.00	0.00	-0.01	0.00
$\hat{\tau}_k^{10fold}$	-0.01	0.00	0.01	0.00	-0.01	-0.01

Coverage in STAR Simulations

	Unadjusted			Adjusted		
	Low	Medium	High	Low	Medium	High
$\hat{\tau}_k$	0.86	0.94	0.80	0.88	0.94	0.80
$\hat{\tau}_k^{LOO}$	0.96	0.94	0.93	0.98	0.96	0.94
$\hat{\tau}_k^{RSS}$	0.90	0.92	0.93	0.95	0.98	0.96
$\hat{\tau}_k^{10fold}$	0.94	0.96	0.93	0.98	0.96	0.94

RMSE in STAR Simulations

	Unadjusted			Adjusted		
	Low	Medium	High	Low	Medium	High
$\hat{\tau}_k$	0.06	0.05	0.07	0.06	0.05	0.07
$\hat{\tau}_k^{LOO}$	0.05	0.05	0.05	0.04	0.05	0.05
$\hat{\tau}_k^{RSS}$	0.07	0.06	0.06	0.06	0.05	0.05
$\hat{\tau}_k^{10fold}$	0.05	0.05	0.05	0.04	0.05	0.05

3. Results and Conclusion - Computer-generated simulation

Bias in Simulations Using Artificial Data

	K = 10						K = 20						K = 40					
	Unadjusted			Adjusted			Unadjusted			Adjusted			Unadjusted			Adjusted		
	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
N=200 $\hat{\tau}_k$	2.54	-0.24	-2.38	2.52	-0.30	-2.32	3.37	0.12	-2.92	3.19	-0.15	-2.80	4.72	-0.08	-4.08	4.34	0.01	-3.83
N=200 $\hat{\tau}_k^{LOO}$	-0.17	-0.01	0.11	0.13	-0.08	-0.02	-0.01	-0.05	0.55	0.31	0.08	-0.10	0.47	0.25	-0.20	1.24	0.15	-1.10
N=200 $\hat{\tau}_k^{RSS}$	-0.01	-0.08	0.13	-0.08	-0.03	0.16	0.32	0.21	0.43	0.25	0.07	0.15	0.14	0.16	0.34	0.19	-0.10	0.11
N=200 $\hat{\tau}_k^{10fold}$	-0.12	-0.19	0.29	0.38	-0.22	-0.16	-0.21	0.15	0.64	0.38	-0.11	-0.15	0.09	0.28	0.05	1.40	0.52	-1.42
N=1000 $\hat{\tau}_k$	0.34	-0.22	-0.43	0.36	-0.22	-0.45	0.78	-0.11	-0.73	0.76	-0.17	-0.76	0.83	0.02	-0.94	0.83	0.05	-0.88
N=1000 $\hat{\tau}_k^{LOO}$	-0.30	-0.21	0.22	-0.27	-0.21	0.19	0.01	-0.04	-0.02	-0.02	-0.10	-0.06	0.01	0.00	-0.14	0.04	0.04	-0.10
N=1000 $\hat{\tau}_k^{RSS}$	0.06	-0.08	-0.20	0.07	-0.10	-0.19	0.04	0.10	-0.15	0.09	0.02	-0.20	-0.17	-0.13	0.04	-0.09	-0.09	-0.05
N=1000 $\hat{\tau}_k^{10fold}$	-0.35	-0.16	0.20	-0.27	-0.14	0.13	-0.04	-0.06	0.03	0.02	-0.12	-0.08	-0.04	0.00	-0.05	0.04	0.01	-0.08
N=5000 $\hat{\tau}_k$	0.06	-0.01	-0.06	0.06	-0.02	-0.05	0.11	-0.05	-0.11	0.11	-0.06	-0.12	0.21	-0.01	-0.26	0.22	0.00	-0.23
N=5000 $\hat{\tau}_k^{LOO}$	-0.06	-0.01	0.06	-0.06	-0.02	0.07	-0.05	-0.05	0.04	-0.05	-0.05	0.03	0.05	-0.01	-0.09	0.06	-0.01	-0.06
N=5000 $\hat{\tau}_k^{RSS}$	0.01	-0.03	-0.04	0.03	-0.03	-0.02	-0.02	-0.08	0.02	-0.04	-0.11	0.02	-0.05	-0.07	0.00	-0.01	-0.07	-0.04
N=5000 $\hat{\tau}_k^{10fold}$	-0.08	0.01	0.06	-0.07	0.00	0.06	-0.05	-0.07	0.07	-0.04	-0.07	0.05	0.05	-0.03	-0.08	0.06	-0.01	-0.06

4. Discussion and Limitations

- Performance of 10 Fold ATE estimation
  - Compared with LOO & Sample Splitting
  - Sample size
  - Number of covariates
- Data Generation Process for Simulation
- Sample Splitting & Bootstrap

Reference

- Abadie, Chingos, and West (2018)

Thank you for your time~