Causal Inference with Corrupted Data

Measurement Error, Missing Values, Discretization, and Differential Privacy

Anish Agarwal[1], Rahul Singh[2]

[1]Columbia IEOR
[2]Harvard Economics

EEA ESEM 2023

Outline

2020 Census will have differential privacy

- (slowly) breaking news: April 22, 2022

The New York Times

The 2020 Census Suggests That People Live Underwater. There's a Reason.

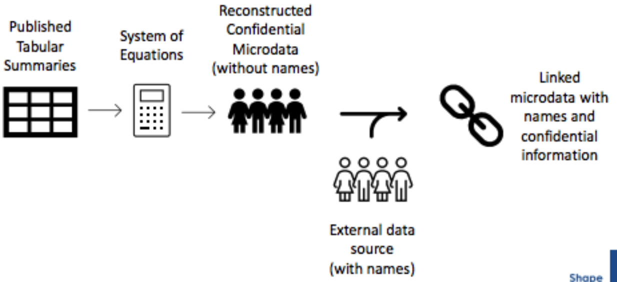
2020 Census will have differential privacy

- differential privacy achieved by injecting synthetic noise
  - "We are deploying differential privacy, the gold standard for privacy protection in computer science and cryptography, to preserve confidentiality in the 2020 Census and beyond"
  - "There are many variants of differential privacy. The one selected for the 2020 Census introduces controlled noise into the data"
- previously implemented in Apple iOS and Google Chrome

- painful trade-off between privacy and precision (Duchi et al. 2018, Abowd + Schmutte 2019, Hotz et al. 2022)

Why will the Census have privacy?

- a simulated attack on the 2010 Census

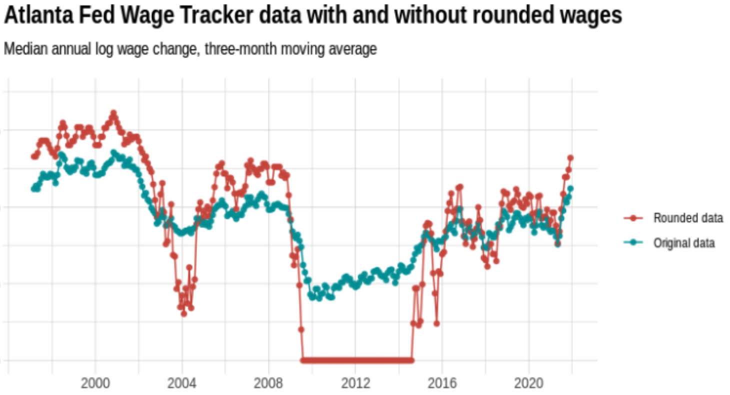


Shape your future START HERE > United States Census 2020

- what did they find?
  - "Our simulated attack demonstrated that, depending on the quality of the external data used, between 52 and 179 million respondents to the 2010 Census can be correctly re-identified from the reconstructed microdata"

Another recent announcement: discretization

- to further protect privacy, the Bureau will discretize wage data



(Zipperer 2022)

- after backlash, the policy was delayed

What are experts saying?

- Prof. Cynthia Dwork (computer science)
  - "Imagine a kind of weaponization, one where somebody decides to make a list of all the gay households across the country. I expect there will be people who would write the software to do that."
- Prof. Charles Manski (econometrics)
  - "This is not a minor technical issue. It's an inescapable tension between enhancing privacy and enhancing data usability."
- Prof. John Abowd (US Census Bureau)
  - "Until now, our discipline has ceded one of the most important debates of the information age to computer science."

What are experts saying?

- Prof. Cynthia Dwork (computer science)
  - "Imagine a kind of **weaponization**, one where somebody decides to make **a list of all the gay households across the country.** I expect there will be people who would write the software to do that."
- Prof. Charles Manski (econometrics)
  - "This is **not a minor technical issue.** It's **an inescapable tension** between enhancing privacy and enhancing data usability."
- Prof. John Abowd (US Census Bureau)
  - "Until now, our discipline has ceded one of the most important debates of the information age to computer science."

What are experts saying?

- Prof. Cynthia Dwork (computer science)
  - “Imagine a kind of **weaponization,** one where somebody decides to make **a list of all the gay households across the country.** I expect there will be people who would write the software to do that.”
- Prof. Charles Manski (econometrics)
  - “This is **not a minor technical issue.** It‘s **an inescapable tension** between enhancing privacy and enhancing data usability.”
- Prof. John Abowd (US Census Bureau)
  - “Until now, our discipline has ceded **one of the most important debates of the information age** to computer science.”

What does this mean for causal inference?

- economists are worried about these “new” data corruptions
  - differential privacy
  - discretization
- even before 2020, the Census had “old” data corruptions...
  - missing values
  - measurement error
- we propose a new end-to-end procedure
  1. data cleaning (slow rate)
  2. estimation (fast rate)
  3. inference (adjusted confidence interval)

What does this mean for causal inference?

- economists are worried about these "new" data corruptions
  - differential privacy
  - discretization

- even before 2020, the Census had "old" data corruptions...
  - missing values
  - measurement error

- we propose a new end-to-end procedure
  1. data cleaning (slow rate)
  2. estimation (fast rate)
  3. inference (adjusted confidence interval)

What does this mean for causal inference?

- economists are worried about these “new” data corruptions
  - differential privacy
  - discretization

- even before 2020, the Census had “old” data corruptions...
  - missing values
  - measurement error

- we propose a new end-to-end procedure
  1. data cleaning (slow rate)
  2. estimation (fast rate)
  3. inference (adjusted confidence interval)

Related work

- semiparametric statistics
  - asymptotic variance (Newey 1994, Robins et al. 1995, Hirano et al. 2003)
  - targeted machine learning (van der Laan + Rubin 2006, Zheng + van der Laan 2011, Luedtke + van der Laan 2016)
  - debiased machine learning (Chernozhukov et al. 2016, 2018, 2021)
- error-in-variable regression
  - auxiliary info: repeated measurement, instrument, negative control (Hausman et al. 1991, Schennach 2007, Maio et al. 2018, Deaner 2018)
  - Lasso and Dantzig: covariance of measurement error must be known (Loh + Wainwright 2012, Rosenbaum + Tsybakov 2013, Belloni et al. 2017)
  - principal component regression (Stock + Watson 2002, Agarwal et al. 2020)
- PCA for large factor models
  - identification, inference for latent factors (Bai 2003, Bai + Ng 2013)
- treatment effects with corrupted data
  - multiple imputation (Rubin 1976, Meng 1994)
  - synthetic control: ATT in panel data by factor model (Athey et al. 2021, Xiong + Pelger 2019, Agarwal et al. 2020, Feng 2020)

Motivation: Takeaway

causal inference with the 2020 US Census?

Outline

Model: Causal parameter

- $Y_i \in \mathbb{R}$ outcome
- $D_i \in \{0, 1\}$ treatment
- $X_{i,\cdot} \in \mathbb{R}^p$ covariates

- for today, we focus on ATE with i.n.i.d. data

$$\theta_0 = \frac{1}{n} \sum_{i=1}^{n} \theta_i, \quad \theta_i = \mathbb{E}[\, Y_i^{(1)} - Y_i^{(0)}]$$

- the paper considers LATE, elasticity, CATE, etc

Model: Data corruption

However, we observe $(Y_i, D_i, Z_{i,\cdot})$ rather than $(Y_i, D_i, X_{i,\cdot})$

$$Y_i = \gamma_0(D_i, X_{i,\cdot}) + \varepsilon_i$$
$$Z_{i,\cdot} = [X_{i,\cdot} + H_{i,\cdot}] \odot \pi_{i,\cdot}$$

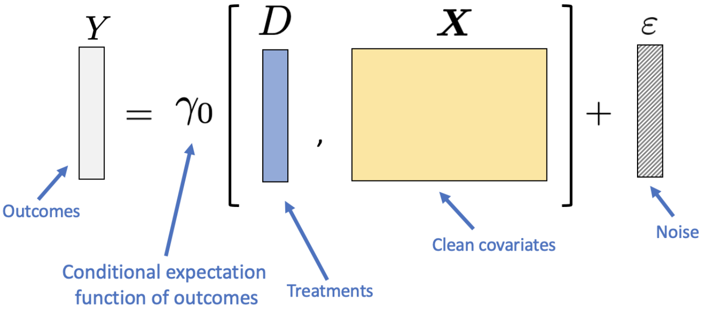
This model encompasses all four types of corruption.

Signal:	5.15	-0.50	-3.49	-2.42	1.11
Measurement error:	5.59	-0.80	-4.09	-1.84	-0.54
Missingness:	5.15	-0.50	?	-2.42	1.11
Discretization:	6	0	-2	-3	2
Privacy:	5.05	-0.37	-3.63	-3.05	1.25

Model: Data corruption

However, we observe $(Y_i, D_i, Z_{i,\cdot})$ rather than $(Y_i, D_i, X_{i,\cdot})$

$$Y_i = \gamma_0(D_i, X_{i,\cdot}) + \varepsilon_i$$
$$Z_{i,\cdot} = [X_{i,\cdot} + H_{i,\cdot}] \odot \pi_{i,\cdot}$$

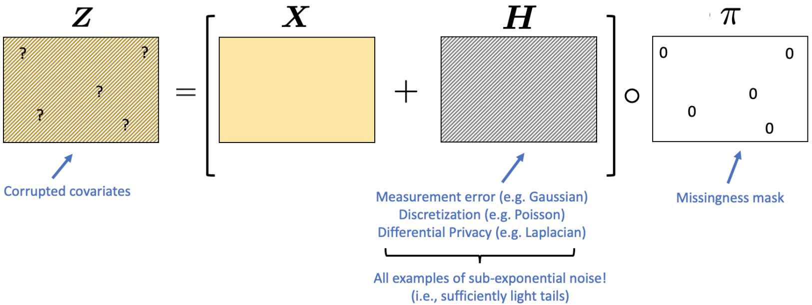


Model: Data corruption

However, we observe $(Y_i, D_i, Z_{i,\cdot})$ rather than $(Y_i, D_i, X_{i,\cdot})$

$$Y_i = \gamma_0(D_i, X_{i,\cdot}) + \varepsilon_i$$

$$Z_{i,\cdot} = [X_{i,\cdot} + H_{i,\cdot}] \odot \pi_{i,\cdot}$$

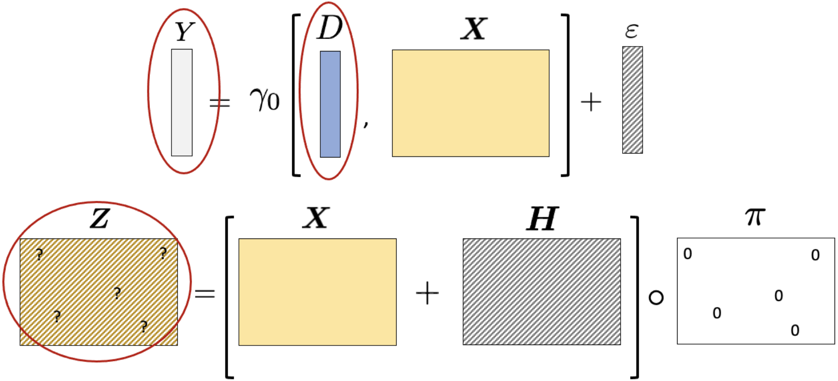


Model: Data corruption

However, we observe $(Y_i, D_i, Z_{i,\cdot})$ rather than $(Y_i, D_i, X_{i,\cdot})$

$$Y_i = \gamma_0(D_i, X_{i,\cdot}) + \varepsilon_i$$

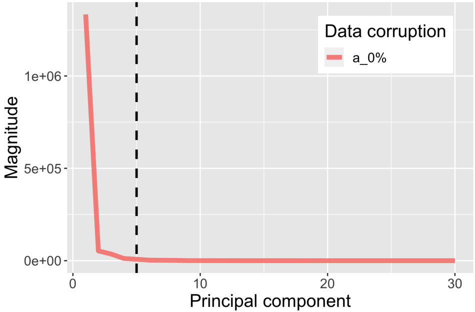
$$Z_{i,\cdot} = [X_{i,\cdot} + H_{i,\cdot}] \odot \pi_{i,\cdot}$$



Model: Key assumption

Assumption: true covariates X are approximately low rank

Why? It holds in Census data (Autor et al. 2013)



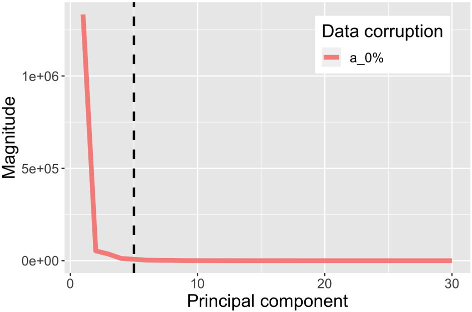
Intuition: repeated measurement model

- average disability benefits
- average medical benefits
- average unemployment benefits

Model: Key assumption

Assumption: true covariates X are approximately low rank

Why? It holds in Census data (Autor et al. 2013)



Intuition: repeated measurement model

- average disability benefits
- average medical benefits
- average unemployment benefits

Census is $\sim$ *low rank*;
has $\sim$ repeated measurements

Outline

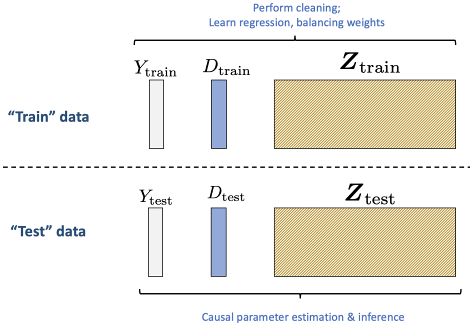
Proposal: Desired procedure

We would like a procedure that

- estimates causal parameters as if data were uncorrupted
- adjusts for data cleaning in the confidence interval
- does not require knowledge of the corruption covariance structure
- preempts the looming trade-off of privacy versus precision

Proposal: Algorithm

Using the split sample



1. data cleaning: $\hat{\mathrm{X}}$ using "train"
2. regression: $\hat{\gamma}$ using "train"
3. balancing weights: $\hat{\alpha}$ using "train"
4. causal parameter: $\hat{\theta}$ using "test"
   - implicit data cleaning of Z_{test}!

Proposal: Theory

Assume

1. each row of measurement error $H_{i,\cdot}$ is mean zero and subexponential
2. each row of missingness $\pi_{i,\cdot}$ is subexponential
3. $r \approx rank(\mathrm{X})$ and the singular values are well-balanced

Theorem (informal):

$$\hat{\mathrm{X}} \xrightarrow{p} \mathrm{X}, \quad \hat{\theta} \xrightarrow{p} \theta_0, \quad \frac{\sqrt{n}}{\sigma}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1), \quad \mathbb{P}(\theta_0 \in CI) \to 0.95$$

Interpretation

- from data cleaning to confidence interval
- $\hat{X} - X$ converges at rate slower than $n^{-1/2}$
- yet $\hat{\theta} - \theta_0$ converges at rate $n^{-1/2}$

Proposal: Theory

Assume

1. each row of measurement error $H_{i,\cdot}$ is mean zero and subexponential
2. each row of missingness $\pi_{i,\cdot}$ is subexponential
3. $r \approx rank(\mathrm{X})$ and the singular values are well-balanced

Theorem (informal):

$$\hat{\mathrm{X}} \xrightarrow{p} \mathrm{X}, \quad \hat{\theta} \xrightarrow{p} \theta_0, \quad \frac{\sqrt{n}}{\sigma}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1), \quad \mathbb{P}\left(\theta_0 \in CI\right) \to 0.95$$

Interpretation

- from data cleaning to confidence interval
- $\hat{X} - X$ converges at rate slower than $n^{-1/2}$
- yet $\hat{\theta} - \theta_0$ converges at rate $n^{-1/2}$

Proposal: Takeaway

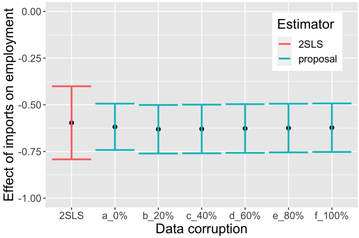
slow data cleaning,
yet *fast* causal inference

Outline

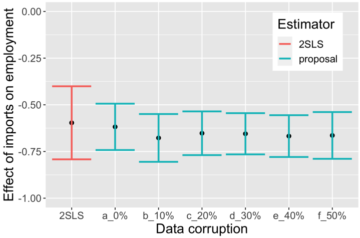
Case study: Import competition

- what is the effect of import competition on the US labor market?
- Census data at commuting zone level (Autor et al. 2013)
- can we recover the same effects with synthetic corruption?
  - differential privacy calibrated to 2020 Census levels
- causal parameter: partially linear IV

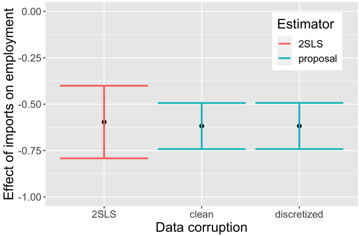
Case study: Synthetic corruption



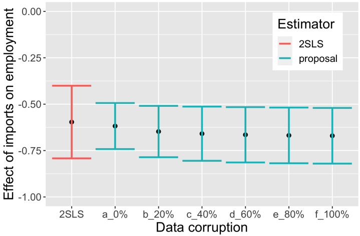
(a) Measurement error



(b) Missing values

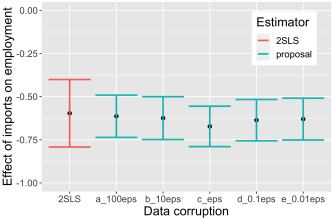


(c) Discretization



(d) Differential privacy

Case study: Calibration



Calibrated differential privacy

Results with formal differential privacy guarantee

- plausible deniability that any individual contributed data to a CZ
- parametrized by ε_{DP}, a measure of privacy loss
- calibrate Laplacian variance to ε_{DP} and variation within the CZ (Dwork et al. 2006)

Case study: Takeaway

both *privacy* and *precision*

Case study: Takeaway

hide your cake and eat it too

Conclusion

- goal: causal inference using 2020 Census
  - abstractly: learn causal parameter from corrupted data
  - concretely: overcome trade-off between privacy and precision
- we propose new data cleaning-adjusted confidence intervals
- bridge matrix completion $(\hat{\mathrm{X}} - \mathrm{X})$ with semiparametrics $(\hat{\theta} - \theta_0)$
- future work: confounded noise, sample selection bias

I would love to talk more!

- email: rahul_singh@fas.harvard.edu