Causal Inference Crash Course Part 4: Best Practices: Outliers, Class Imbalance, Feature Selection, and Bad Control

Julian Hsu

Causal Inference Series

1) Foundations
2) Defining Some ATE/ATET Causal Models
3) ATE/ATET Inference, Asymptotic Theory, and Bootstrapping
4) **Best Practices: Outliers, Class Imbalance, Feature Selection, and Bad Control**
5) Heterogeneous Treatment Effect Models and Inference
6) Difference-in-Difference Models for Panel Data
7) Regression Discontinuity Models
8) Arguable Validation

Overview

- This presentation will outline best practices for issues around causal inference which can be applied to other ML settings.

A. Outliers;
B. Class Imbalance in Propensity Scores;
C. Feature Selection; and
D. Bad Control

- For each issue, we will discuss what the problem is, why it’s a problem, and a solution outline.

A. Outliers

Why are outliers problems?

- Generally, treatment effects estimates are about the average.
  - Average treatment effect
  - Average treatment effect on the treated
  - Conditional average treatment effect
- This is represented in their technical implementation by the statistical conditions for estimation.
- For example, the unconfoundedness assumption can be represented as $E[u_i|X_i, W_i] = 0$

Outliers skew the average

- Obviously, outlier values cause the average to take on extreme values
- This is also a theoretical problem because we have already decided our metric of interest is the average. The average by its nature is sensitive to outliers.
  - The median not so much, but we'll return to the median later.
- Outliers can be a problem if the data is is meant to be representative, but we still have low sample size.

Two types of outliers

$$Y_i = \tau W_i + g(X_i) + u_i$$

- This means that outlier values in Y_i are potentially a problem. Outlier values in X_i (without any corresponding outlier values in $\mathrm{Y_i}$) can be addressed with feature generation.
- We will discuss two types of outliers with simulations:

1. Outliers in Y_i due to random noise, or large values of $\mathrm{u_i}$;
2. Outliers in Y_i due to outlier values in X_i.

Outlier Y_i values due to large values of u_i

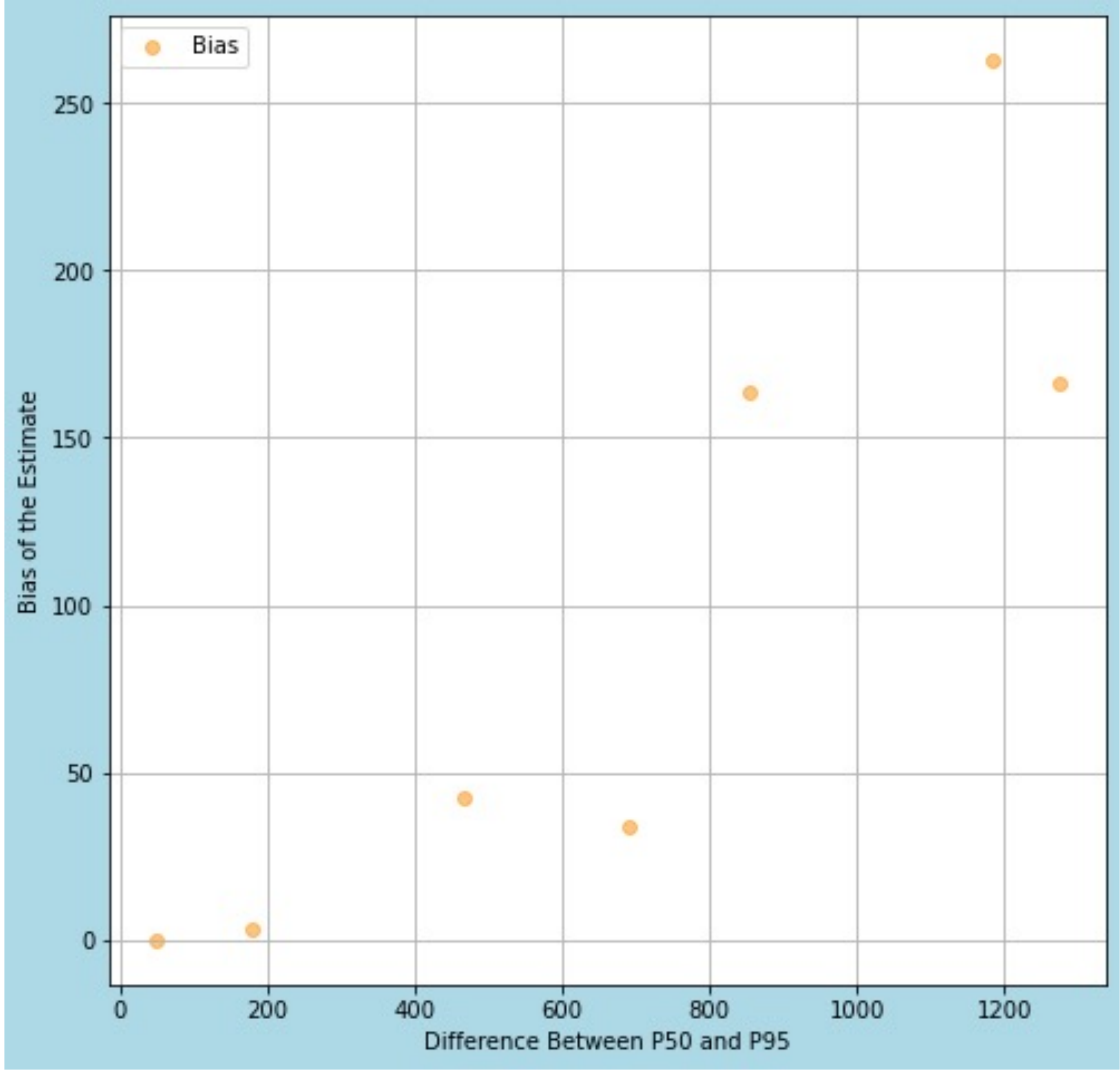
- Generally, large values of u_i are not a concern for the estimate, but can be a concern with inference. This is because u_i will be identified as random noise.
- Simulation setup:

$$Y_i = \tau W_i + g(X_{1i}, X_{2i}) + u_i$$
$$W_i = h(X_{1i}, X_{2i}, \eta_i)$$

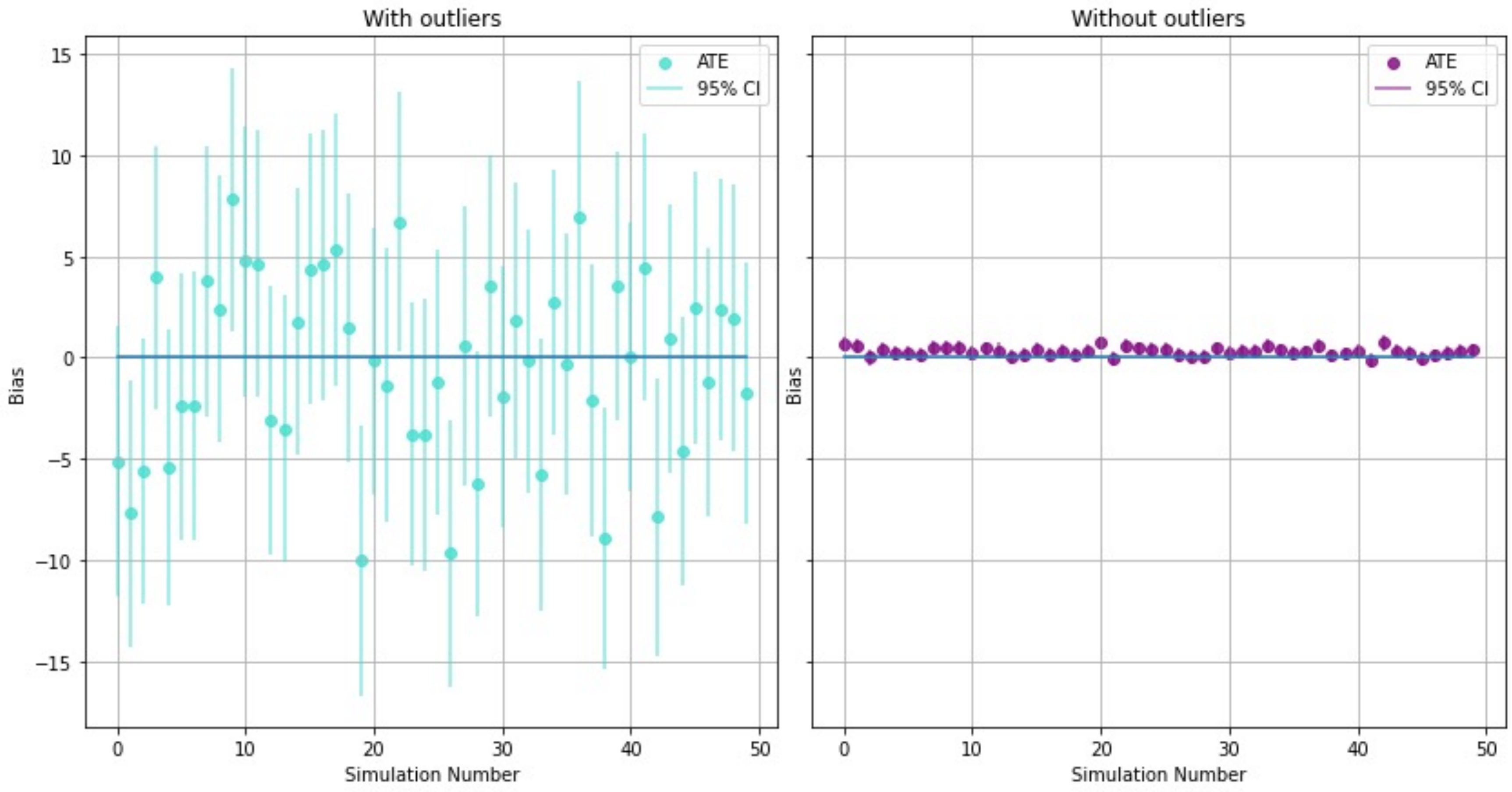
- Where u_i and η_i are random draws, and X_{1i} and X_{2i} are independent normally distributed features
- τ is the treatment effect, and the parameter of interest

Simulation evidence

- Simulation created outlier values in Y_i with large values of u_i.
  - 10% of observations have large values of u_i.
- As the distribution becomes more skewed, the bias increases



Simulation results on the bias



- With and without outliers, the estimate $\hat{\tau}$ has small bias.
- However, with outliers, the confidence intervals are much larger due the additional statistical noise.

Outlier Y_i values due to large values of X_i

- This is a much larger concern because large values of Y_i are not driven by random noise. This means that conditioning on X_i raises theoretical concerns.
- In a simulation similar to before, outlier values of X_{1i} and X_{2i} create outlier values of Y_i. The estimated ATE is **500% larger** than the true treatment effect.
- Discuss three approaches:
    1. Conditioning on generated features;
    2. Truncation;
    3. Winsorization

Condition on generated features of X_{1i} and X_{2i}

Indicator of whether $\mathrm{X_i}$ is an outlier	Indicator of whether $\mathrm{X_i}$ is an outlier interacted with $\mathrm{X_i}$	Natural log of $\mathrm{X_i} + 1$	Estimate (True Value is 50)	Standard Error
X			333.766	360.346
X	X		288.162	361.047
X		X	328.213	358.607
X	X	X	262.311	358.995

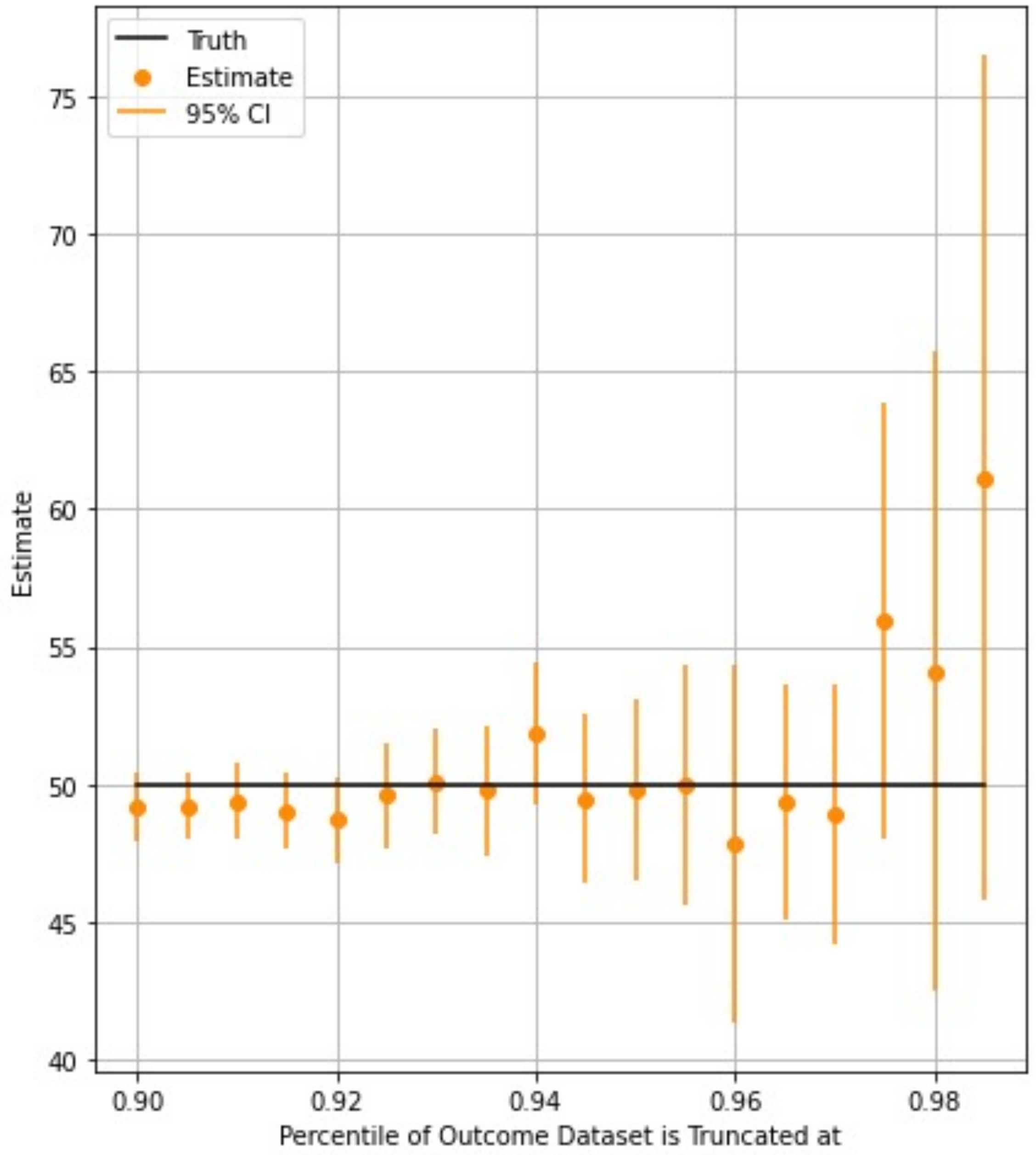
- Conditioning on additional features drives the estimate to be closer to the truth, but at best the estimate is more than 400% larger
- There is also no impact on the standard error

Truncating Values of Y_i

- Truncation is removing observations based on values of Y_i .
- However, it is unclear how much to truncate. The more data is truncated, the less <u>natural variation</u> in the data is removed.
- No principled way to determine the best truncation point.

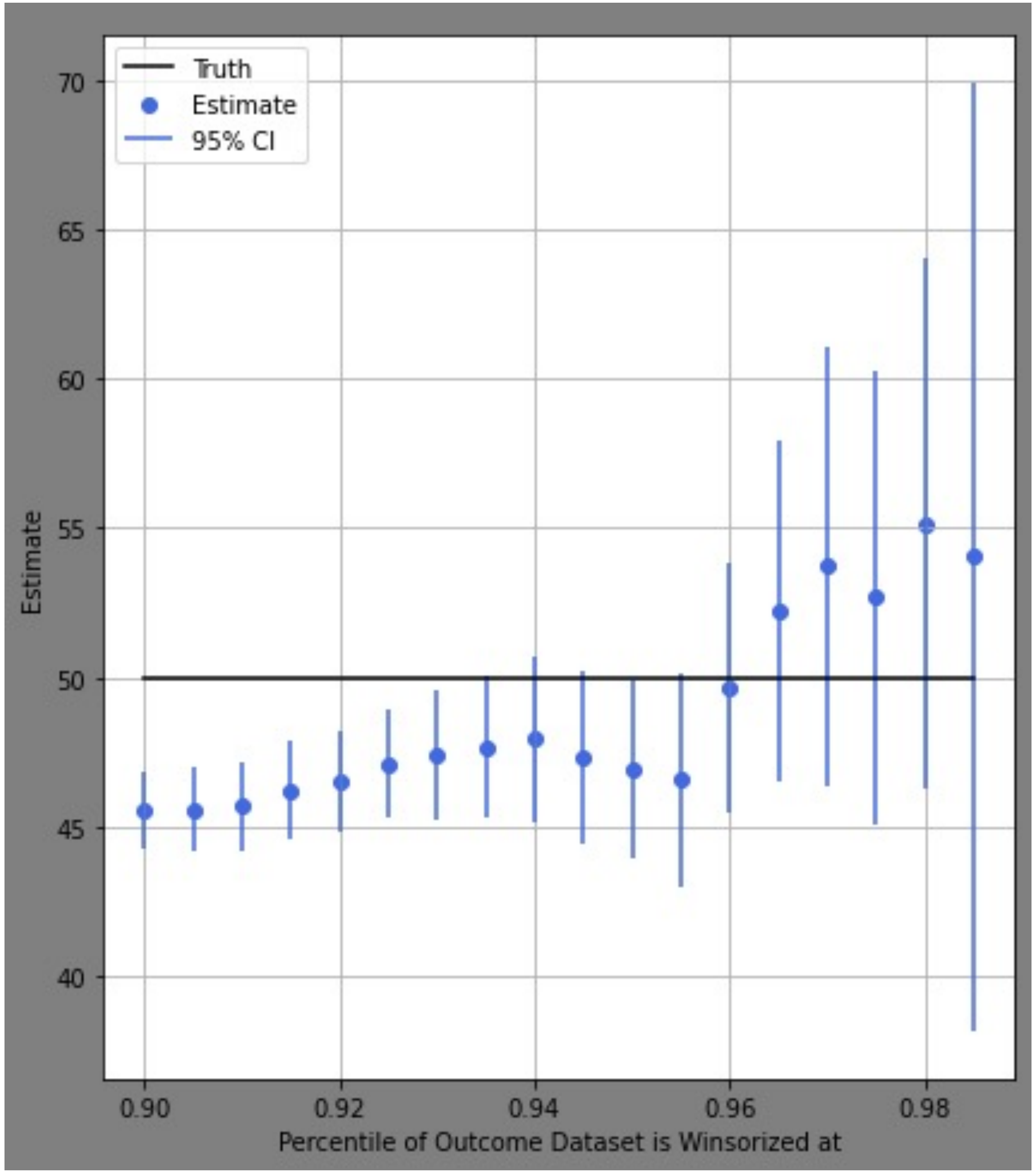
Truncation simulation evidence

- Truncated data based on the 90th , 91st, … 99th percentile in Y_i
- The less truncation, the more biased and less precise the estimate is.
- However, the idea of removing data is not palatable and will likely break down in more flexible data settings

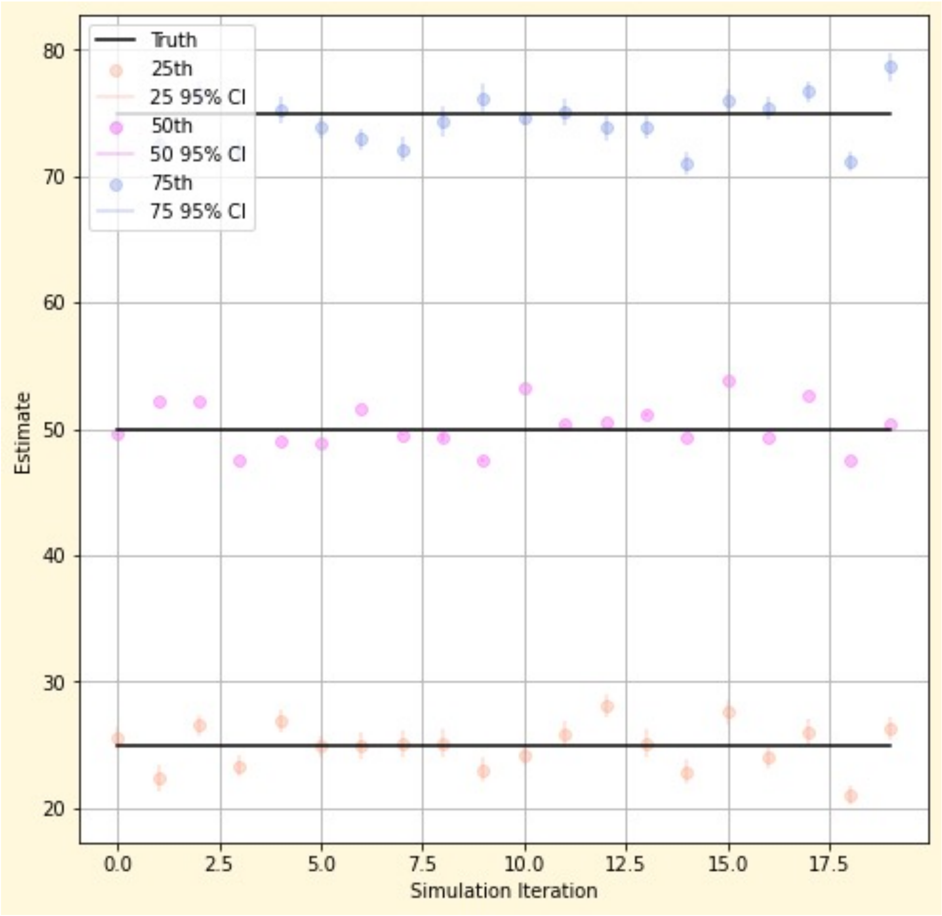


Winsorizing Values of Y_i

- Winsorizing is replacing values of Y_i with a top coded or bottom coded number
- Like truncation, it is unclear how much to winsorize.
- Simulation evidence shows that more winsorization leads to more biased estimates and more precision



WIP – Median and Quantile Treatment Effects



B. Class Imbalance in Propensity Scores

WIP

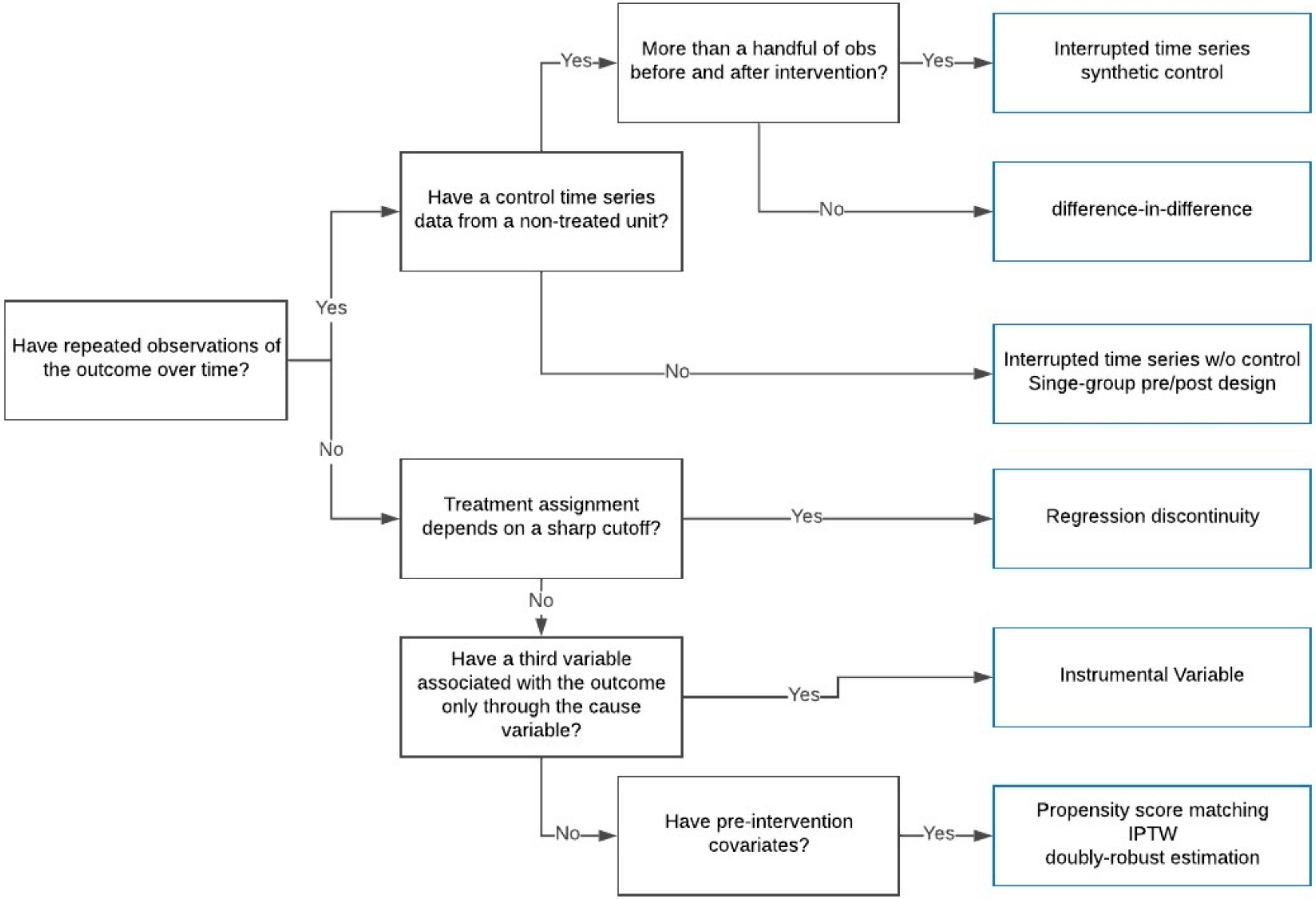
C. Feature Selection

WIP

D. Bad Control

WIP

Appendix



Source: https://eng.uber.com/causal-inference-at-uber/