# Regression Adjusted Estimators in Randomized Experiments With A Diverging Number of Covariates

Lihua Lei & Peng Ding

Department of Statistics, UC Berkeley

ACIC, 2018

# WARNINGS

# WARNINGS

- No fancy design! We analyze completed randomized experiments;

BORING

# WARNINGS

- No fancy design! We analyze completed randomized experiments;
- No fancy method! We analyze ordinary least squares;

BORING $\times 2$

# WARNINGS

- No fancy design! We analyze completed randomized experiments;
- No fancy method! We analyze ordinary least squares;
- No fancy regime! We analyze $p < n$.

BORING $\times 3$

# WARNINGS

- No fancy design! We analyze completed randomized experiments;
- No fancy method! We analyze ordinary least squares;
- No fancy regime! We analyze $p < n$.

FUNDAMENTAL!

# Table of Contents

# Setup

- Unit denoted by $i$; $n$ total units;
- Potential outcomes: $(Y_i(1), Y_i(0))$ ;
- Covariate vector: $x_i \in \mathbb{R}^p$;
- Binary treatment indicator: $T_i \in \{0, 1\}$;
- SUTVA; Observed outcome: $Y_i^{\text{obs}} = Y_i(1)T_i + Y_i(0)(1 - T_i)$;

# Setup

- Unit denoted by $i$; $n$ total units;
- Potential outcomes: $(Y_i(1), Y_i(0))$ ;
- Covariate vector: $x_i \in \mathbb{R}^p$;
- Binary treatment indicator: $T_i \in \{0, 1\}$;
- SUTVA; Observed outcome: $Y_i^{\mathrm{obs}} = Y_i(1)T_i + Y_i(0)(1 - T_i)$;
- **Finite population perspective:** $(Y_i(1), Y_i(0), x_i)$ are **fixed**;
- **No assumption** on the functional relation between $(Y_i(1), Y_i(0))$ and $(T_i, x_i)$.

# Randomized Experiment

- We consider the completed randomized experiment with $n_1$ units uniformly assigned into the treatment group, i.e.

$$P\left(T_1 = t_1, \ldots, T_n = t_n\right) = \left(\begin{array}{c} n \\ n_1 \end{array}\right)^{-1},$$

for any $t_1 + \cdots + t_n = n_1$;

- $\mathcal{T}_1$ and $\mathcal{T}_0$ are the sets of treated and controlled units, with

$$n_1 \triangleq |\mathcal{T}_1|, \quad n_0 \triangleq |\mathcal{T}_0|;$$

- Inferential target: average treatment effect (ATE):

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \tau_i, \quad \text{where } \tau_i = Y_i(1) - Y_i(0).$$

# Regression Adjusted Estimator: Motivation

- Fix $\beta_1, \beta_0 \in \mathbb{R}^p$. Consider the estimator

$$\hat{\tau}(\beta_1, \beta_0) = \frac{1}{n_1} \sum_{i \in \mathcal{T}_1} (Y_i^{\text{obs}} - x_i^T \beta_1) - \frac{1}{n_0} \sum_{i \in \mathcal{T}_0} (Y_i^{\text{obs}} - x_i^T \beta_0);$$

- Assume $\sum_{i=1}^n x_i = 0$ WLOG, it is easy to show that

$$\mathbb{E}\hat{\tau} = \tau.$$

# Regression Adjusted Estimator: Motivation

- $\hat{\tau}(\beta_1, \beta_0)$ is the difference-in-means estimator with $(Y_i(1), Y_i(0))$ replaced by

$$(Y_i(1) - x_i^T \beta_1, Y_i(0) - x_i^T \beta_0).$$

- Under mild conditions,

$$\frac{\sqrt{n}(\hat{\tau}(\beta_1, \beta_0) - \tau)}{\sigma(\beta_1, \beta_0)} \xrightarrow{d} N(0, 1);$$

- Similar to difference-in-means estimators,

$$\sigma(\beta_1, \beta_0)^2 = \frac{n}{n_1} S_1^2 + \frac{n}{n_0} S_0^2 - S_\tau^2$$

  where $S_1^2, S_0^2, S_\tau^2$ are the population variances of $(Y_i(1) - x_i^T \beta_1)_{i=1}^n, (Y_i(0) - x_i^T \beta_0)_{i=1}^n$ and $(\tau_i - x_i^T (\beta_1 - \beta_0))_{i=1}^n$.

# Regression Adjusted Estimator: Motivation

- The optimal choice of $(\beta_1, \beta_0)$ is

$$\beta_1^* = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i(1) - x_i^T \beta)^2,$$

$$\beta_0^* = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i(0) - x_i^T \beta)^2;$$

- These are the population ordinary least squares (OLS) estimators by regressing $(Y_i(1))$ and $(Y_i(0))$ on $(x_i)$;
- This adjustement always gives a better asymptotic variance than the difference-in-means estimator.

$$\sigma(\beta_1^*, \beta_0^*)^2 \leq \sigma(0,0)^2.$$

# Regression Adjusted Estimator: Point Estimate

- However $(\beta_1^*, \beta_0^*)$ cannot be obtained from the observed data;
- Replace them by empirical OLS estimates,

$$\hat{\beta}_1 = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \in \mathcal{T}_1} (Y_i^{\text{obs}} - x_i^T \beta)^2,$$
$$\hat{\beta}_0 = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \in \mathcal{T}_0} (Y_i^{\text{obs}} - x_i^T \beta)^2.$$

- The regression adjusted estimator is defined as

$$\hat{\tau} \triangleq \hat{\tau}(\hat{\beta}_1, \hat{\beta}_0).$$

# Regression Adjusted Estimator: Asymptotic Normality

- Lin [2013] proves that, under strong assumptions and holding $p$ fixed, while as $n \to \infty$,

$$\frac{\sqrt{n}(\hat{\tau} - \tau)}{\sigma(\beta_1^*, \beta_0^*)} \xrightarrow{d} N(0, 1).$$

- Recall that $\sigma(\beta_1^*, \beta_0^*)$ is the asymptotic variance of difference-in-means estimators with potential outcomes

$$(e_i(1), e_i(0)) \triangleq (Y_i(1) - x_i^T \beta_1^*, Y_i(0) - x_i^T \beta_0^*),$$

  where $\beta_1^*, \beta_0^*$ are population OLS estimates in the treatment and the control group, respectively.

# Regression Adjusted Estimator: Variance Estimator

- Let $(\hat{e}_{1,i}, \hat{e}_{0,i})$ be the sample OLS residuals:

$$\hat{e}_{1,i} = Y_i^{\text{obs}} - x_i^T \hat{\beta}_1, \quad \text{if } T_i = 1,$$

$$\hat{e}_{0,i} = Y_i^{\text{obs}} - x_i^T \hat{\beta}_0, \quad \text{if } T_i = 0;$$

- Estimate $\hat{\tau}(\beta_1^*, \beta_0^*)^2$ by

$$\hat{\sigma}^2 = \frac{n}{n_1} s_{1,e}^2 + \frac{n}{n_0} s_{0,e}^2$$

where $s_{1,e}^2$ and $s_{0,e}^2$ are sample variances of $(\hat{e}_{1,i})$ and $(\hat{e}_{0,i})$;

- Under the same set of assumptions, Lin [2013] proves that

$$\lim \hat{\sigma}^2 \geq \sigma(\beta_1^*, \beta_0^*)^2.$$

# Table of Contents

# What Happens in High Dimensions? A Toy Simulation

- Consider $n$ units;
- Generate $x_i \stackrel{i.i.d.}{\sim} N(0, I_{n\times n})$ and fix them;
- Centered the columns of $X$ ($x_i \to x_i - 1/n \sum_{i=1}^{n} x_i$);
- Generate $(Y_i(1), Y_i(0)) \stackrel{i.i.d.}{\sim} N(0, I_{2\times 2})$ and fix them.

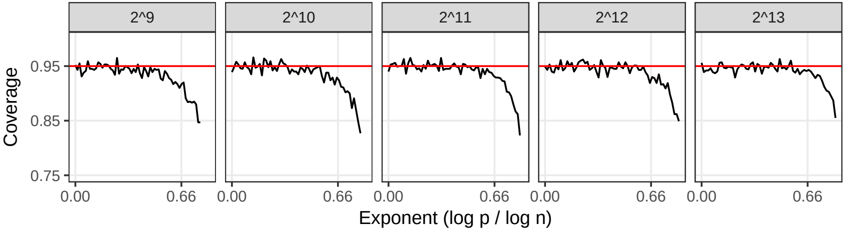
# What Happens in High Dimensions? A Toy Simulation

- Take $p = \lceil n^\gamma \rceil$ for each $\gamma \in [0, 1]$;
- Replace $X$ by the submatrix formed by the first $p$ columns;
- Randomly assign $n/2$ units into the treatment group and get the regression adjusted estimate;
- Repeat for $1000$ times to obtain $\hat{\tau}^{(1)}, \ldots, \hat{\tau}^{(1000)}$;
- Compute $95\%$ empirical coverage by

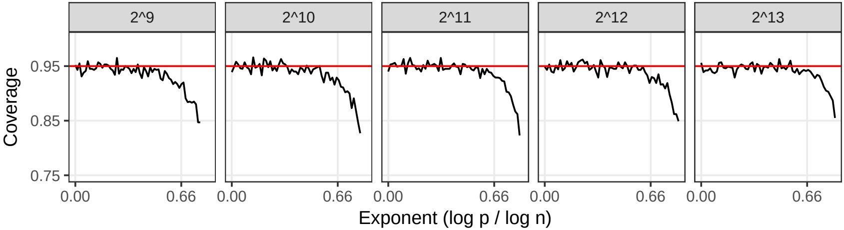
$$\frac{1}{1000} \sum_{k=1}^{1000} I(|\hat{\tau}^{(k)} - \tau| \leq 1.96\sigma(\beta_1^*, \beta_0^*)).$$

Note that $\sigma(\beta_1^*, \beta_0^*)$ is the theoretical asymptotic variance.

# What Happens in High Dimensions? A Toy Simulation



# What Happens in High Dimensions? A Toy Simulation




**Question: largest $p$ that supports the classical result?**

# Notation

- $X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n \times p}$;
- WLOG assume $X$ has centered columns; otherwise replace $X$ by its de-centered version;
- $Y(1), Y(0) \in \mathbb{R}^n$ are the vectors of potential outcomes;
- $T \in \{0, 1\}^n$ is the vector of treatment assignments, with $n_1$ units uniformly assigned $1$;
- $X, Y(1), Y(0)$ are all **fixed**; only $T$ is **random**.
- $\beta_1^*, \beta_0^* \in \mathbb{R}^p$ are the population OLS estimates;
- $e^{(1)}, e^{(0)} \in \mathbb{R}^n$ are the population OLS residual vectors

$$e^{(1)} = Y(1) - X\beta_1, \quad e^{(0)} = Y(0) - X\beta_0.$$

# Coherence/Maximum Leverage Score

- Hat matrix/Projection matrix:

$$H = X(X^T X)^{-1} X^T;$$

- Leverage scores: diagonal elements of $H$, measuring the "influence" of each observation;

- Coherence/Maximum leverage score:

$$\kappa \triangleq \max_i H_{ii} = \max_i (X(X^T X)^{-1} X^T)_{ii};$$

- It always holds that

$$\frac{p}{n} = \frac{\mathrm{tr}(H)}{n} \le \kappa \le \|H\|_{\mathrm{op}} \le 1.$$

# Large Sample Property for Regression Adjusted Estimators

## Theorem 1 (L. and Ding, 2018).

*Under extremely mild assumptions, stated in a few slides,*

1. $\hat{\tau}$ *is consistent;*
2. *The variance estimator is asymptotically conservative:*

$$\lim \frac{\hat{\sigma}^2}{\sigma(\beta_1, \beta_0)^2} \geq 1;$$

3. $\hat{\tau}$ *is asymptotically normal,*

$$\frac{\sqrt{n}(\hat{\tau} - \tau)}{\sigma(\beta_1, \beta_0)} \xrightarrow{d} N(0, 1).$$

*if we further have*

$$p\kappa \to 0,$$

# Large Sample Property for Regression Adjusted Estimators

- Note that asymptotic normality requires

$$p\kappa \to 0;$$

- In the favorable case where all leverage scores are close, i.e. $\kappa = O(p/n)$, the condition reads

$$\frac{p^2}{n} \to 0 \Longrightarrow p = o(n^{1/2});$$

- Lin [2013]'s result extends to $p = o(n^{1/2})$ with weaker assumptions;
- However, there is still a gap between $n^{1/2}$ and $n^{2/3}$;
- The result can be improved if more assumptions are imposed.

# Debiased Estimator

- Let $\hat{\beta}_1, \hat{\beta}_0$ be the sample OLS estimates

$$\hat{\beta}_1 = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \in \mathcal{T}_1} (Y_i^{\text{obs}} - x_i^T \beta)^2,$$

$$\hat{\beta}_0 = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i \in \mathcal{T}_0} (Y_i^{\text{obs}} - x_i^T \beta)^2;$$

- Let $\hat{e}_{1,i}, \hat{e}_{0,i}$ be the sample OLS residuals

$$\hat{e}_{1,i} = Y_i^{\text{obs}} - x_i^T \hat{\beta}_1, \quad \text{if } T_i = 1,$$

$$\hat{e}_{0,i} = Y_i^{\text{obs}} - x_i^T \hat{\beta}_0, \quad \text{if } T_i = 0.$$

# Debiased Estimator

- Define the bias estimator:

$$\widehat{\text{bias}} \triangleq = \frac{n_1}{n_0} \sum_{i \in \mathcal{T}_0} H_{ii} \hat{e}_{0,i} - \frac{n_0}{n_1} \sum_{i \in \mathcal{T}_0} H_{ii} \hat{e}_{1,i};$$

- The debiased regression adjusted estimator is defined as

$$\hat{\tau}^{\text{de}} \triangleq \hat{\tau} - \widehat{\text{bias}}.$$

# Large Sample Property for Debiased Estimators

**Theorem 2 (L. and Ding, 2018).**

*Under extremely mild assumptions, stated in a few slides,*

1. $\hat{\tau}^{\mathrm{de}}$ *is consistent;*
2. $\hat{\tau}^{\mathrm{de}}$ *is asymptotically normal,*

$$\frac{\sqrt{n}(\hat{\tau}^{\mathrm{de}} - \tau)}{\sigma(\beta_1, \beta_0)} \xrightarrow{d} N(0, 1).$$

*If we further have*

$$p\kappa^2 \log p \to 0.$$

# Large Sample Property for Debiased Estimators

- The asymptotic normality requires a weaker condition

$$p\kappa^2 \log p \to 0$$

- In the favorable case where $\kappa = O(p/n)$, the condition reads

$$\frac{p^3 \log p}{n^2} \to 0 \Longrightarrow p = o\left(\frac{n^{2/3}}{(\log n)^{1/3}}\right);$$

- Almost fill in the gap, up to log-factors.

# Table of Contents

# Quantities in Assumptions

- Moments of $e^{(1)}, e^{(0)}$:

$$\mathcal{E}_2 \triangleq \max\left\{\frac{1}{n}\sum_{i=1}^{n}\left(e_i^{(1)}\right)^2, \frac{1}{n}\sum_{i=1}^{n}\left(e_i^{(0)}\right)^2\right\}$$

$$\mathcal{E}_\infty \triangleq \max\left\{\max_i |e_i^{(1)}|, \max_i |e_i^{(0)}|\right\};$$

- Correlation between $e^{(1)}$ and $e^{(0)}$:

$$\rho \triangleq \frac{\sum_{i=1}^{n} e_i^{(0)} e_i^{(1)}}{\sqrt{\sum_{i=1}^{n}(e_i^{(0)})^2}\sqrt{\sum_{i=1}^{n}(e_i^{(1)})^2}}.$$

# Assumptions

**A1** $\dfrac{n_0}{n}, \dfrac{n_1}{n} \geq \pi > 0$ (only for clean results; can be removed);

**A2** $\kappa = o\left(\dfrac{1}{\log p}\right)$.

# Assumptions

**A1** $\dfrac{n_0}{n}, \dfrac{n_1}{n} \geq \pi > 0$ (only for clean results; can be removed);

**A2** $\kappa = o\left(\dfrac{1}{\log p}\right)$.

For consistency:

**A3** $\mathcal{E}_2 = O\left(\frac{n}{p}\right)$ (only for presentation; can be further weakened)

# Assumptions

**A1** $\frac{n_0}{n}, \frac{n_1}{n} \geq \pi > 0$ (only for clean results; can be removed);

**A2** $\kappa = o\left(\frac{1}{\log p}\right)$.

For consistency:

**A3** $\mathcal{E}_2 = O\left(\frac{n}{p}\right)$ (only for presentation; can be further weakened)

For variance estimation and asymptotic normality:

**A4** $\rho > -1 + \eta$ for some $\eta > 0$;

**A5** $\mathcal{E}_\infty^2 / n\mathcal{E}_2 = o(1)$. (Lindeberg-Feller type condition)

# Remarks on Assumptions

On the covariate matrix $X$

- $\kappa = o(1/\log p)$ is a very mild condition! Recall that

$$\frac{p}{n} = \frac{\mathrm{tr}(H)}{n} \leq \kappa \leq 1;$$

- No other assumption on $X$! In literature extra strong assumptions are imposed on the largest/smallest singular value, or even on the fourth moment of covariates, which excludes many realistic cases, especially when interaction terms are incorporated;

- No assumption that $n_1/n, n_0/n$ have asymptotic limits.

# Remarks on Assumptions

On population residuals $e^{(1)}$ and $e^{(0)}$:

- The second moment of $e^{(1)}, e^{(0)}$ is allowed to diverge! In literature a finite fourth moment is assumed and seems crucial in their analysis [Lin, 2013, Bloniarz et al., 2016];
- The moment condition holds for $t(2)$-like residuals;
- No assumption that the asymptotic variance converges to some limit in probability;
- Variance estimation and asymptotic normality even do not require any moment condition, except the Lindeberg-Feller condition.

# Table of Contents

# Lalonde Data

- LaLonde [1986] analyzes the impact of National Supported Work (NSW) Demonstration, a labor training program, on postintervention income levels;
- The study includes a complete randomized experiment;
- We use the dataset in NBER data archive of Dehejia and Wahba [1999], with $n = 445$ units and $n_1 = 185$ units are assigned into the program;
- The outcome is the earnings in 1978;
- The dataset has $10$ covariates: age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), nodegree (1 if no degree, 0 otherwise), RE74/RE75 (earnings in 1974/1975), u74/u75 (1 if RE74/RE75 = 0, 0 otherwise)

# Simulating Potential Outcomes on Lalonde Data

- We form $X$ by including all covariates and two-way interaction terms, and removing the ones perfectly collinear to others;
- $X$ ends up with $p = 49$ columns;
- Run OLS on treatment units and control units to get $\hat{\beta}_1$ and $\hat{\beta}_0$;
- Simulate potential outcomes by

$$Y_i^*(1) = x_i^T \hat{\beta}_1, \quad Y_i^*(0) = x_i^T \hat{\beta}_0.$$

- Perturb them by adding a noise $\epsilon^{(1)}, \epsilon^{(0)} \in \mathbb{R}^n$ and truncate them at zero,

$$Y_i(1) = \max\{0, Y_i^*(1) + e_i^{(1)}\}, \quad Y_i(0) = \max\{0, Y_i^*(0) + e_i^{(0)}\}.$$

# Simulating Potential Outcomes on Lalonde Data

- We form $X$ by including all covariates and two-way interaction terms, and removing the ones perfectly collinear to others;
- $X$ ends up with $p = 49$ columns;
- Run OLS on treatment and control units to get $\hat{\beta}_1$ and $\hat{\beta}_0$;
- Simulate potential outcomes by

  $$Y_i^*(1) = \max\{x_i^T \hat{\beta}_1 + e_i^{(1)}, 0\}, \quad Y_i^*(0) = \max\{x_i^T \hat{\beta}_0 + e_i^{(0)}, 0\},$$

  where $e_i^{(1)}, e^{(0)} \overset{i.i.d.}{\sim} N(0, I_{n\times n})$;
- Compute the true parameters, including $\tau, \beta_1, \beta_0, \sigma^2(\beta_1, \beta_0)$.

# Simulating Potential Outcomes on Lalonde Data

For each $p \in \{1, \ldots, 49\}$,

Step 1 random select $p$ columns from $X$ as the covariate matrix;

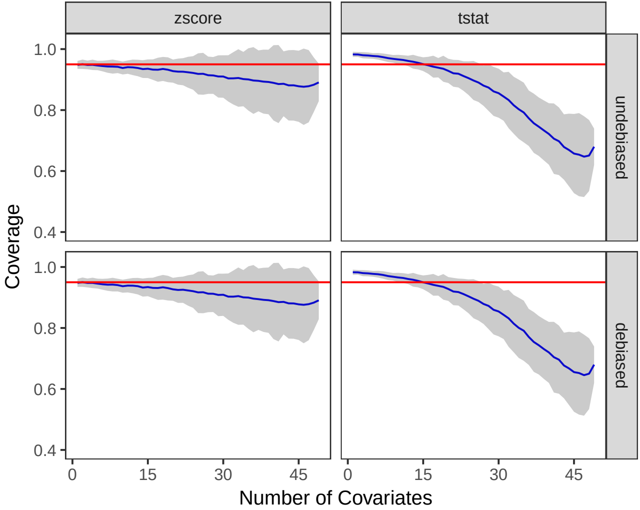
Step 2 randomly generate $1000$ assignment vectors with $n_1$ units in the treatment group;

Step 3 Obtain $1000$ estimates;

Step 4 Summarize the estimates to obtain $95\%$ coverage, bias, variance inflation ratio and p-values from Shapiro's normality test.

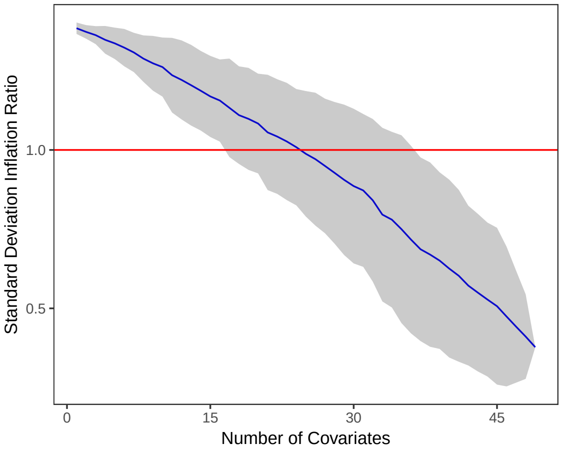
Repeat the above procedure for $50$ times and obtain the confidence intervals for the summarized measures.

# 95% Coverage

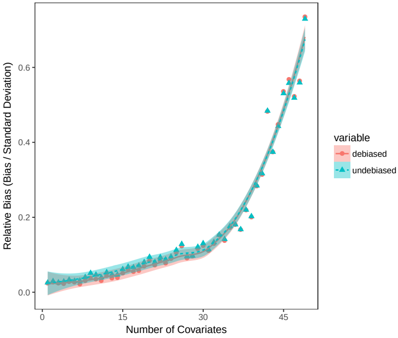


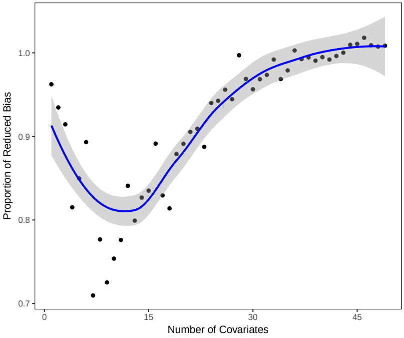
zscore
tstat
undebiased
debiased
Coverage
1.0
0.8
0.6
0.4
0
15
30
45
Number of Covariates

# Variance Inflation

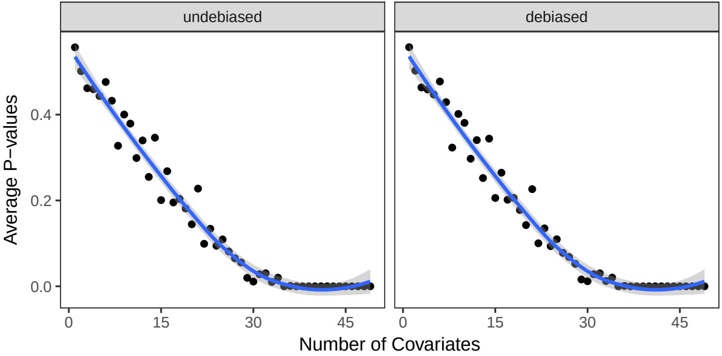


# Relative Bias



# Normality Test



# References


Adam Bloniarz, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S Sekhon, and Bin Yu. Lasso adjustments of treatment effect estimates in randomized experiments. Proceedings of the National Academy of Sciences, 113(27):7383–7390, 2016.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association, 94(448):1053–1062, 1999.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. The American economic review, pages 604–620, 1986.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. The Annals of Applied Statistics, 7(1):295–318, 2013.

Thanks!