Regression-Based Case Selection in Multi-Method Research

Jason Seawright

j-seawright@northwestern.edu

Jan. 19, 2024

Discovery vs. Confirmation

Case Selection

Case Selection

1. Study the entire population.

Case Selection

1. Study the entire population.
2. Take a random sample.

Case Selection

1. Study the entire population.
2. Take a random sample.
3. Follow some rule for deliberate case selection.

Choosing Cases

- Case-selection rules:

Choosing Cases

- Case-selection rules:
  - Random sampling

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases
  - Deviant cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases
  - Deviant cases
  - Influential cases

Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases
  - Deviant cases
  - Influential cases
  - Most-similar cases

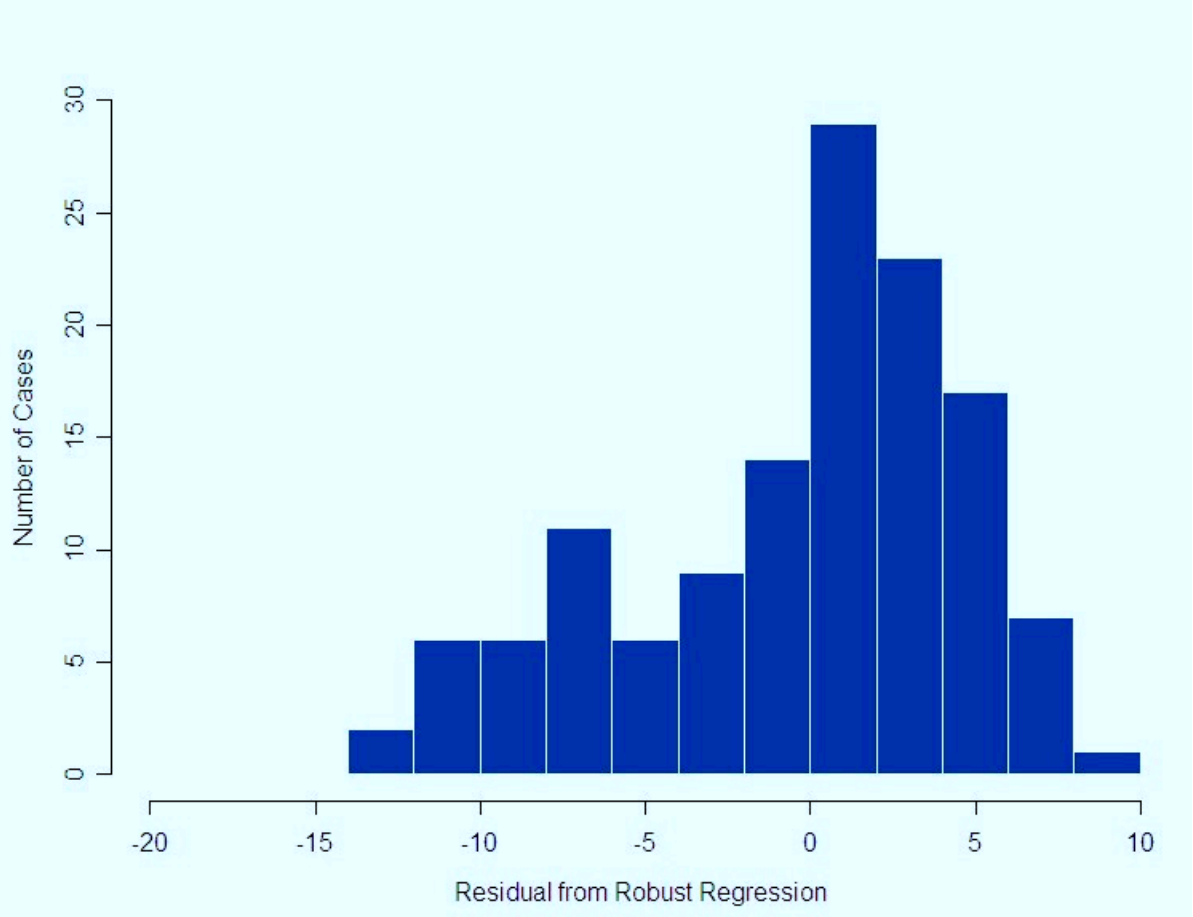
Choosing Cases

- Case-selection rules:
  - Random sampling
  - Typical cases
  - Diverse cases
  - Extreme cases
  - Deviant cases
  - Influential cases
  - Most-similar cases
  - Contrast cases

Typical Cases

$$\text{Typicality}_i = -\text{abs}[y_i - \text{E}(y_i|x_{1,i}, x_{2,i}, \ldots, x_{k,i})] \quad (1)$$

Typical Cases



Extreme Cases

$$\text{Extremity}_i = |\frac{x_i - \bar{x}}{s}| \tag{2}$$

Deviant Cases

$$\text{Deviantness}_i = -\text{Typicality}_i \tag{3}$$

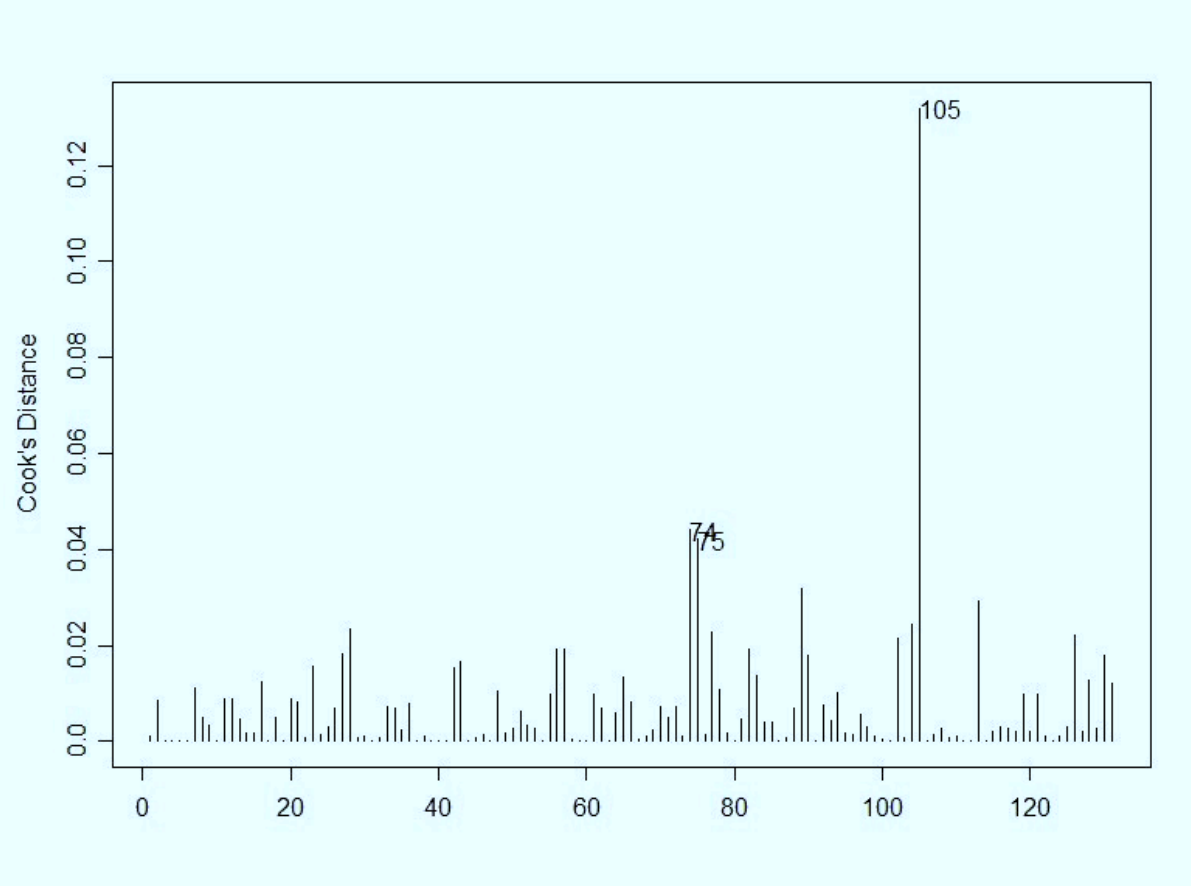
Influential Cases

- Cook's distance is a statistical measure of how much the overall regression result would change if a given case is deleted.

Influential Cases

- Cook's distance is a statistical measure of how much the overall regression result would change if a given case is deleted.
- A Cook's distance score of 1 or more usually is regarded as representing substantial influence.

Influential Cases



Most-Similar Cases

- Matching techniques are an automated way of finding most similar cases.

Contrast Cases

Measurement Error in Y

$$Y_i^* = Y_i + \delta_{Y,i}$$

Measurement Error in Y

$$Y_i^* = Y_i + \delta_{Y,i}$$

Random Sampling

Measurement Error in Y

Typical/Deviant Cases:

$$e_i = Y_i - \mathbb{H}_{i,.}Y + \delta_{Y,i}$$

Measurement Error in Y

Influential Cases

Measurement Error in Y

Extreme Cases:

$$Y_i^* = Y_i + \delta_{Y,i}$$

Measurement Error in Y

Most-Similar Cases

Measurement Error in Y

Most-Similar Cases

Contrast Cases

Measurement Error in X

$$X_i^* = X_i + \delta_{X,i}$$

Measurement Error in X

$$X_i^* = X_i + \delta_{X,i}$$

Random Sampling

Measurement Error in X

Typical/Deviant Cases:

$$e_i = Y_i - X_i\hat{\beta}^* - \delta_{X,i}\hat{\beta}^*$$

Measurement Error in X

Influential Cases

Measurement Error in X

Extreme Cases:

$$X_i^* = X_i + \delta_{X,i}$$

Measurement Error in X

Most-Similar Cases

Measurement Error in X

Most-Similar Cases

Contrast Cases

Omitted Variables

$$e_i = d_i + \gamma \tilde{Z}_i, \text{ where } \tilde{Z}_i = Z_I - \mathrm{E}(Z_i|X_i)$$

Omitted Variables

$e_i = d_i + \gamma \tilde{Z}_i$, where $\tilde{Z}_i = Z_I - \mathrm{E}(Z_i|X_i)$

Random Sampling

Omitted Variables

Typical/Deviant Cases:

$e_i = d_i + \gamma \tilde{Z}_i$

Omitted Variables

Influential Cases

Omitted Variables

Extreme Cases:

Omitted Variables

Extreme Cases:

For confounders, extreme on X may be a good strategy.

Omitted Variables

Extreme Cases:

For confounders, extreme on X may be a good strategy.

Extreme on Y maximizes:

$\hat{Y}_i + d_i + \gamma \tilde{Z}_i$

Omitted Variables

Most-Similar Cases

Omitted Variables

Most-Similar Cases

Contrast Cases

Pathway Variables

$$W_i = \nu + \mu X_i + \omega_i$$
$$Y_i = \alpha + \tau W_i + \sigma_i$$

Pathway Variables

$$W_i = \nu + \mu X_i + \omega_i$$
$$Y_i = \alpha + \tau W_i + \sigma_i$$

Random Sampling

Pathway Variables

Typical/Deviant Cases:

$e_i = \tau\omega_i + \sigma_i$

Pathway Variables

Influential Cases

Pathway Variables

Extreme Cases:

Pathway Variables

Extreme Cases:

$$W_i = \nu + \mu X_i + \omega_i$$

Pathway Variables

Extreme Cases:

$$W_i = \nu + \mu X_i + \omega_i$$

Extreme on Y maximizes:

$$Y_i = \alpha + \tau W_i + \sigma_i$$

Pathway Variables

Most-Similar Cases

Pathway Variables

Most-Similar Cases

Contrast Cases

Summary: Analytic Arguments

	Deviant	Influential	Ext. X	Ext. Y
Error in Y	Good	Mixed	Poor	Good
Error in X	Mixed	Mixed	Good	Poor
Confound	Mixed	Mixed	Good	Good
Pathway	Good	Mixed	Good	Mixed

Monte Carlo for Case Selection

Simulate case selection for the same problem 10,000 times.

Monte Carlo for Case Selection

Simulate case selection for the same problem 10,000 times.

- Analysis of presidential vote shares and the economy in Latin America, 1980-2000.

Monte Carlo for Case Selection

Simulate case selection for the same problem 10,000 times.

- Analysis of presidential vote shares and the economy in Latin America, 1980-2000.
- Add measurement error, omitted variables, etc.

Monte Carlo for Case Selection

Simulate case selection for the same problem 10,000 times.

- Analysis of presidential vote shares and the economy in Latin America, 1980-2000.
- Add measurement error, omitted variables, etc.
- 2 SD Rule

Simulation Results

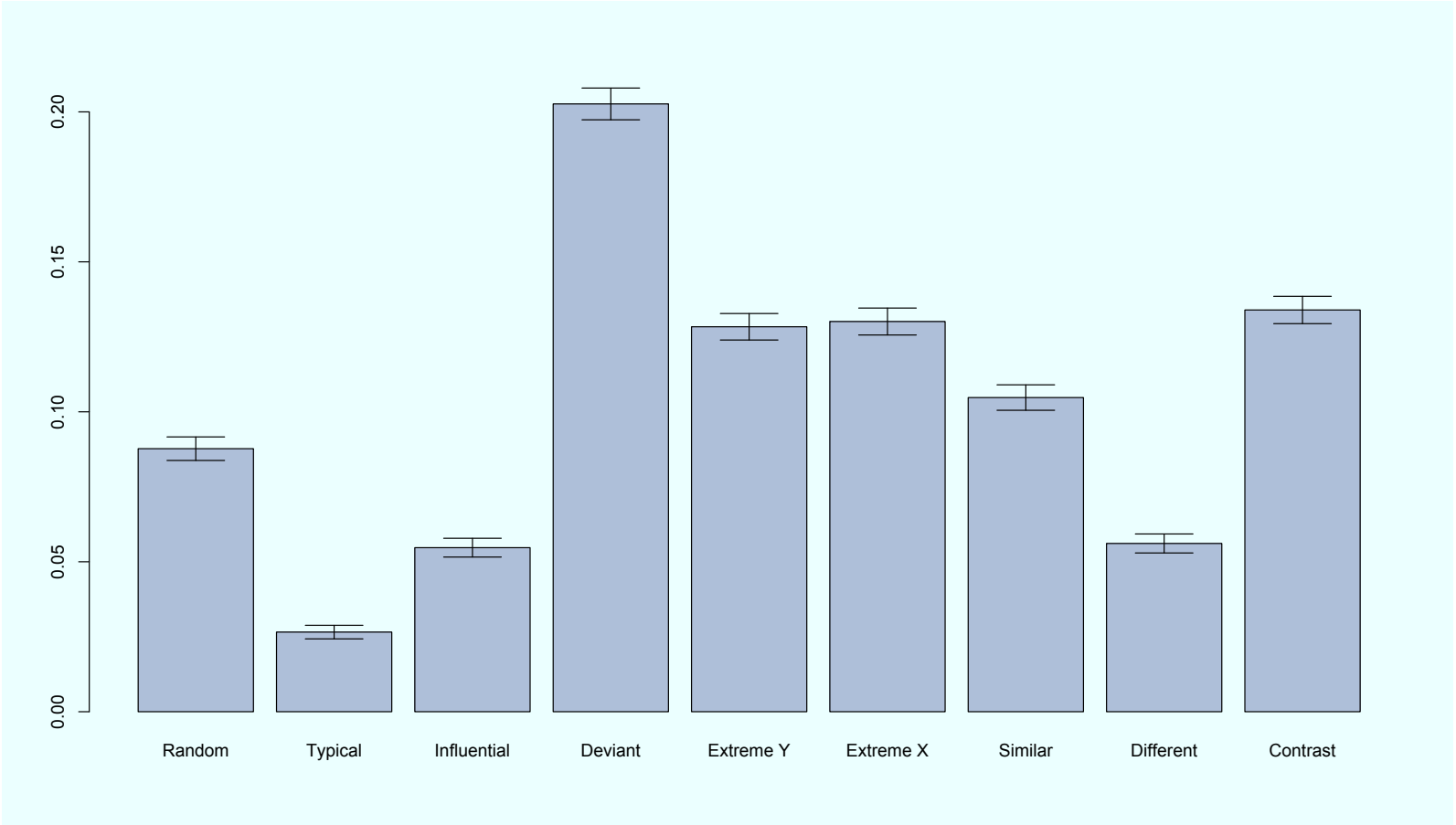


Figure: Case Selection for Finding Confounder.

Simulation Results

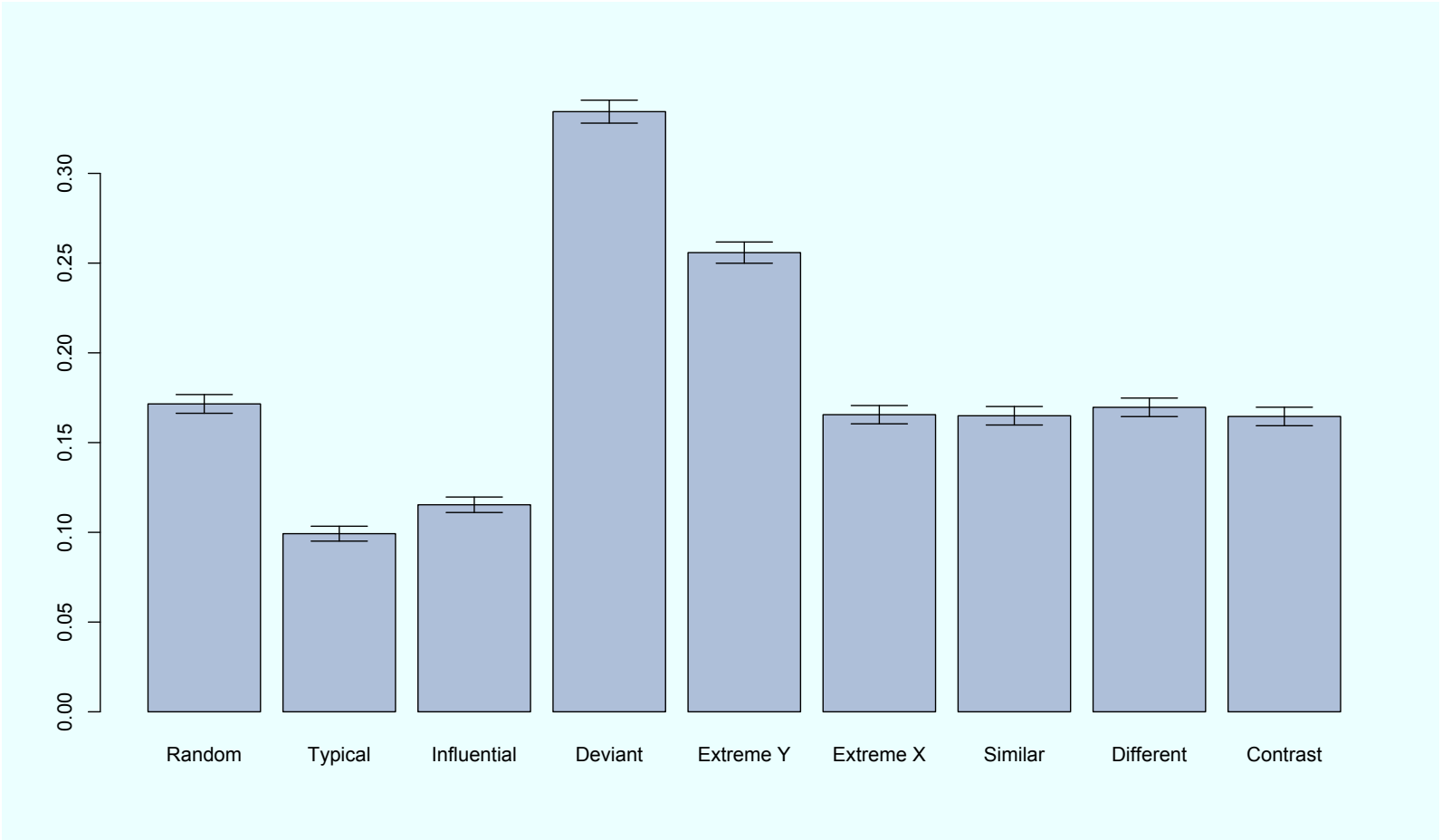


Figure: Case Selection for Other Causes.

Simulation Results

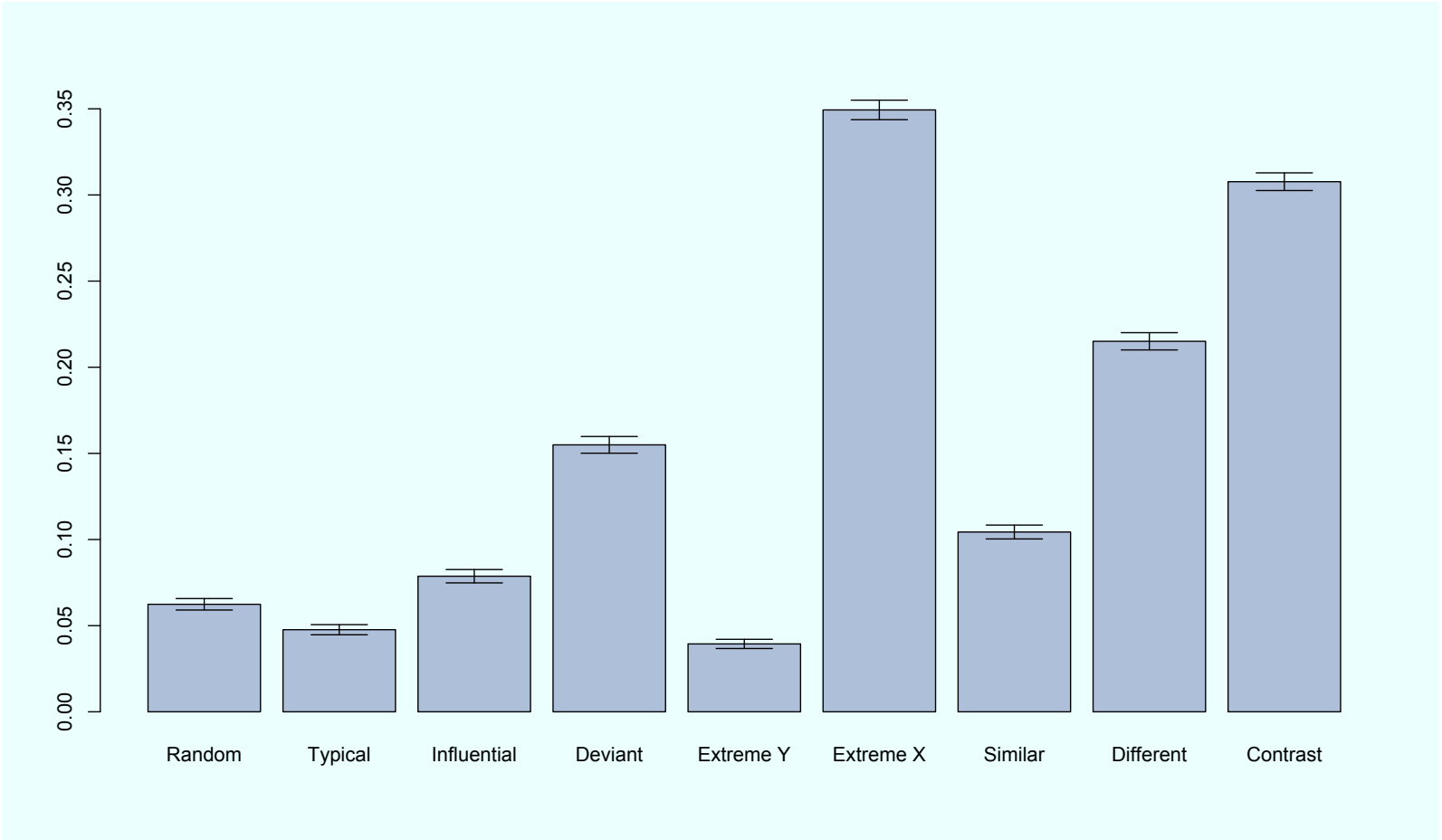


Figure: Case Selection for Exploring Mechanisms.

Simulation Results

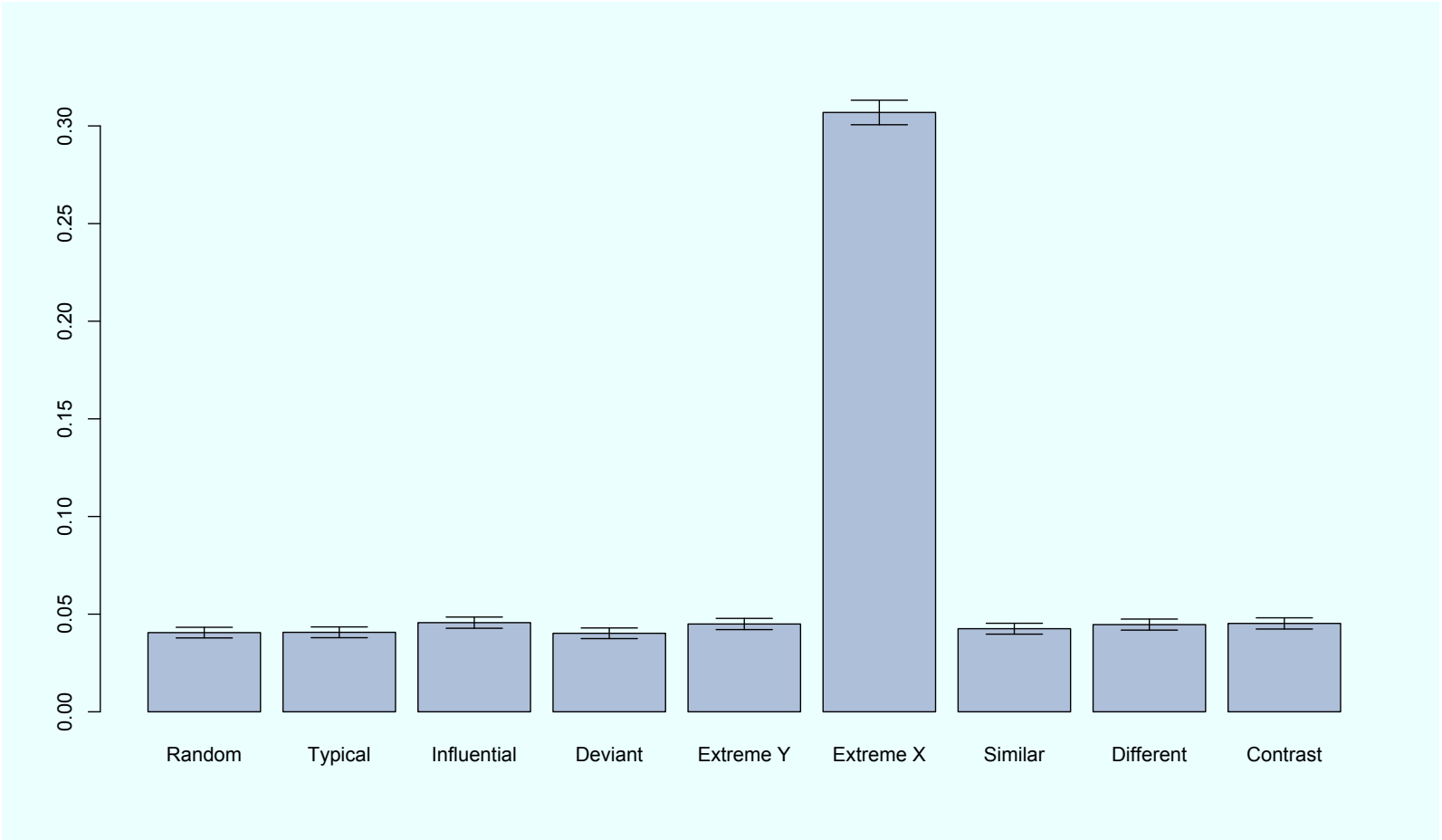


Figure: Case Selection for Error in X.

Simulation Results

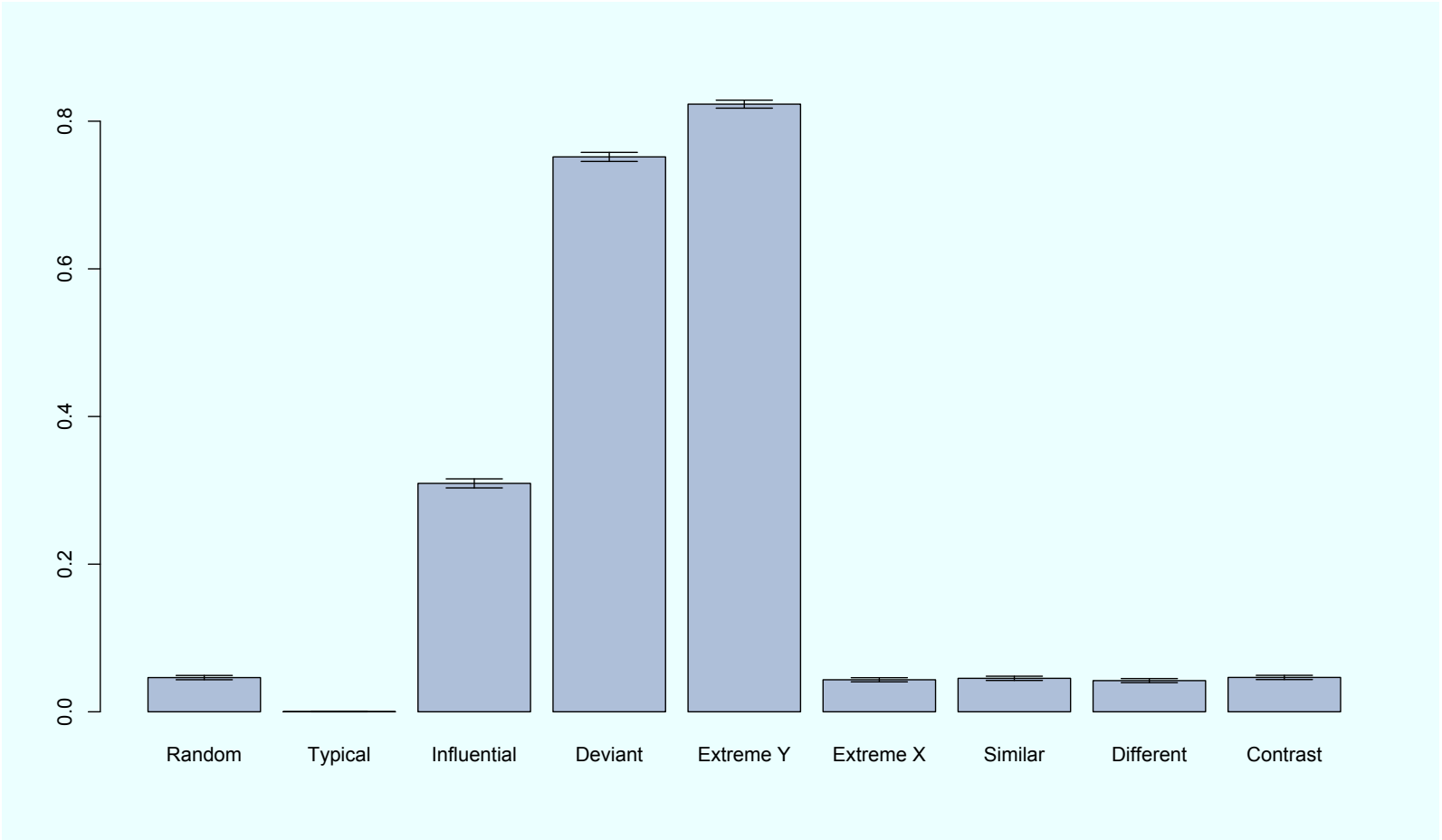


Figure: Case Selection for Error in Y.

Simulation Results

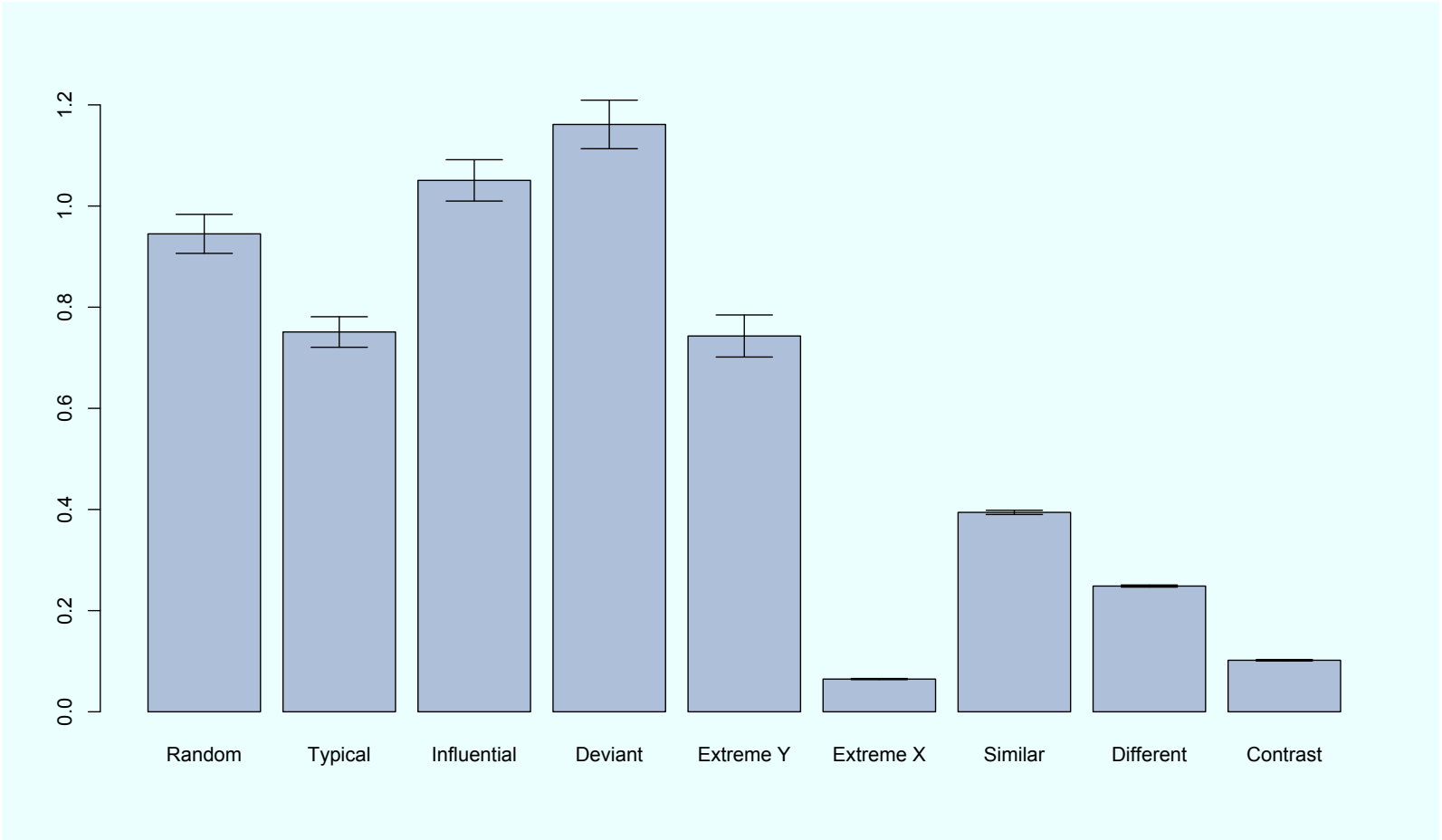


Figure: Case Selection for Estimating Overall Slope.

Propensity-Adjusted Extreme Cases

Table 3

	(1) Employment Laws
Base Density	0.0384***
	(0.00718)
Change in Density	-0.0385***
	(0.0113)
MFG	0.583
	(0.608)
logMFGWage	-0.000745
	(0.0343)
logEMPLCHG	0.104
	(0.0744)
Unemployment	-0.0670**
	(0.0324)
TotalEnactments	-1.03e-05**
	(4.47e-06)
RTW	-0.166**
	(0.0796)
Constant	0.417
	(0.558)

Figure 2: Propensity-Adjusted Extreme Values on Change in Union Density

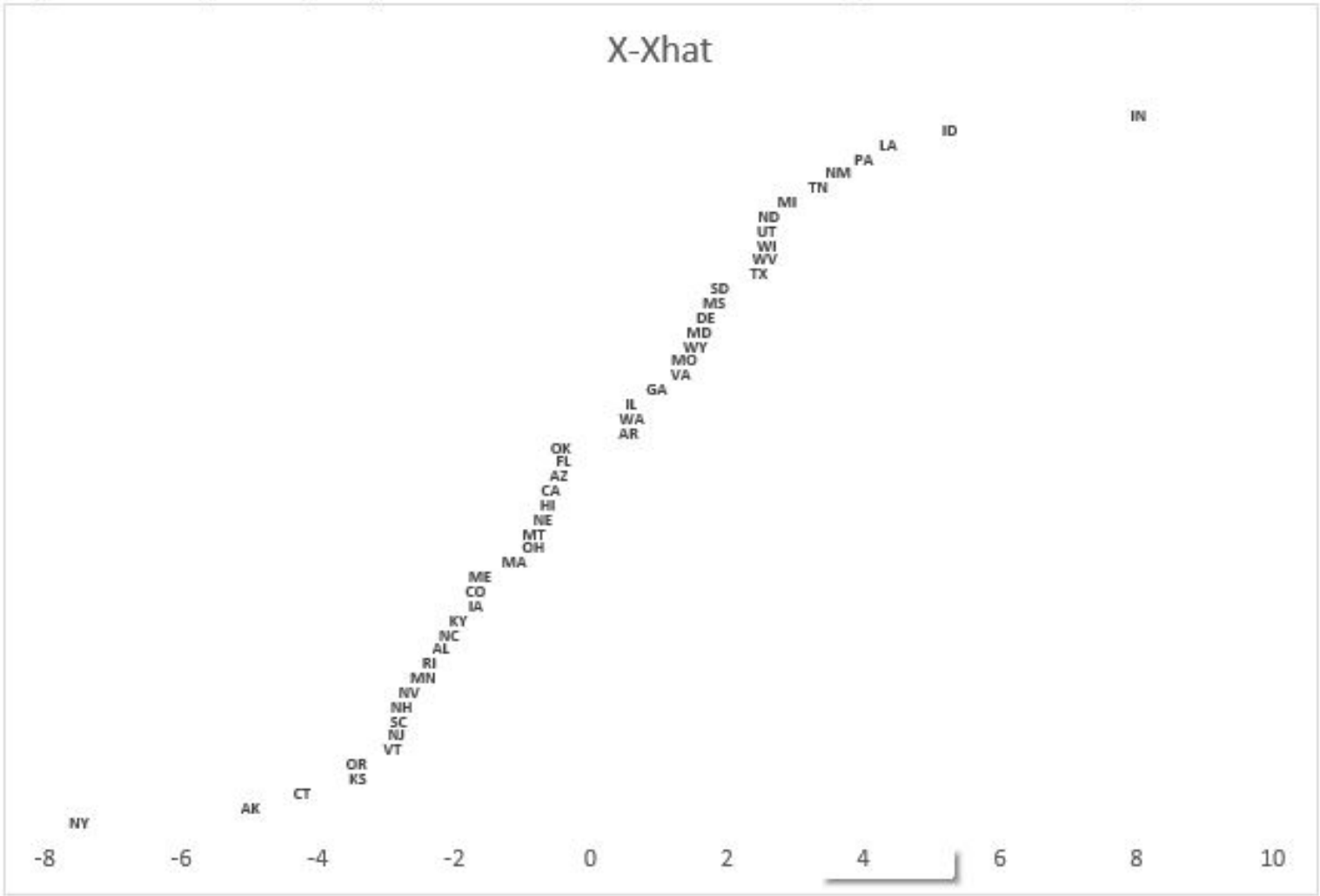


Figure 3

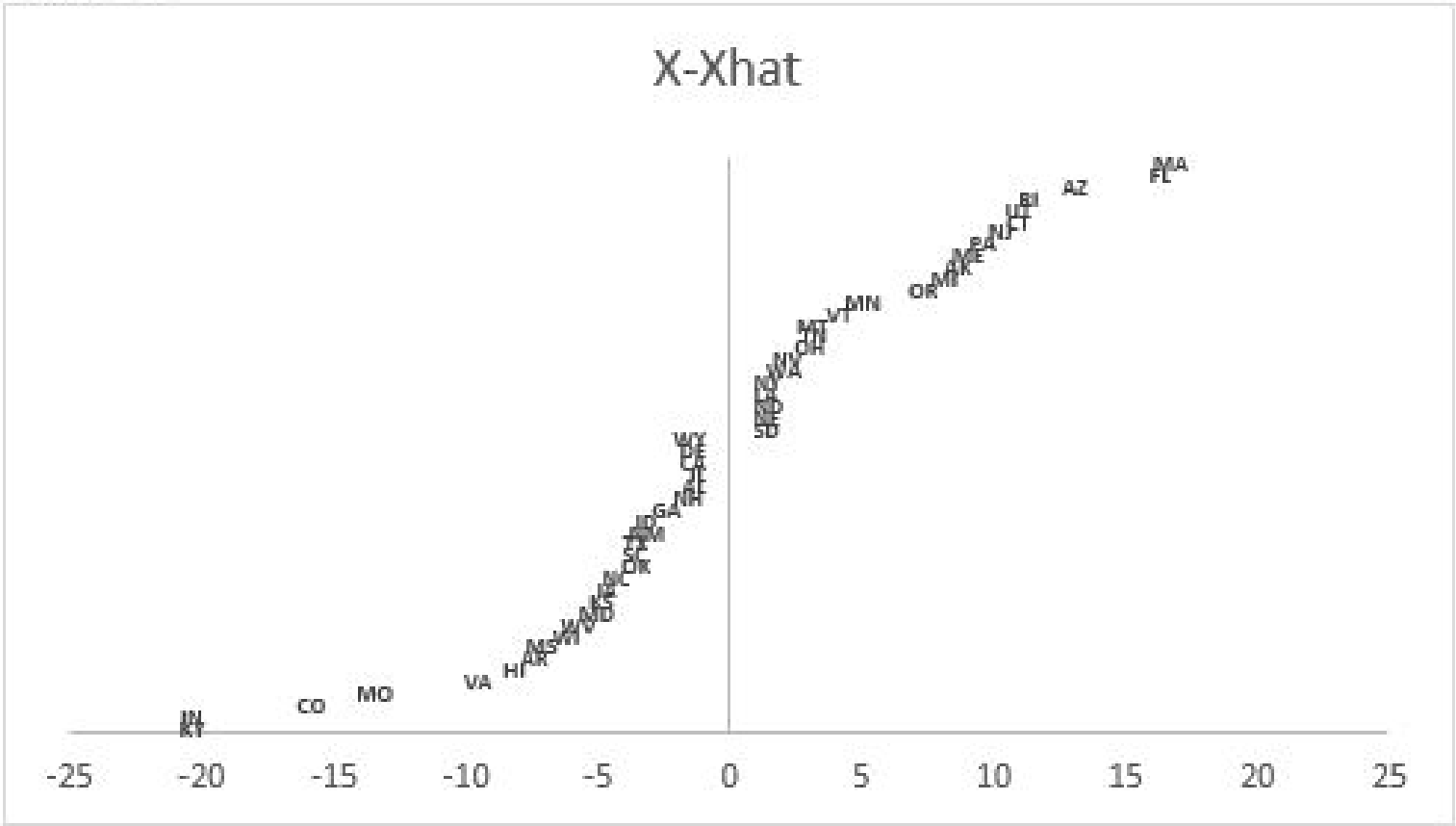
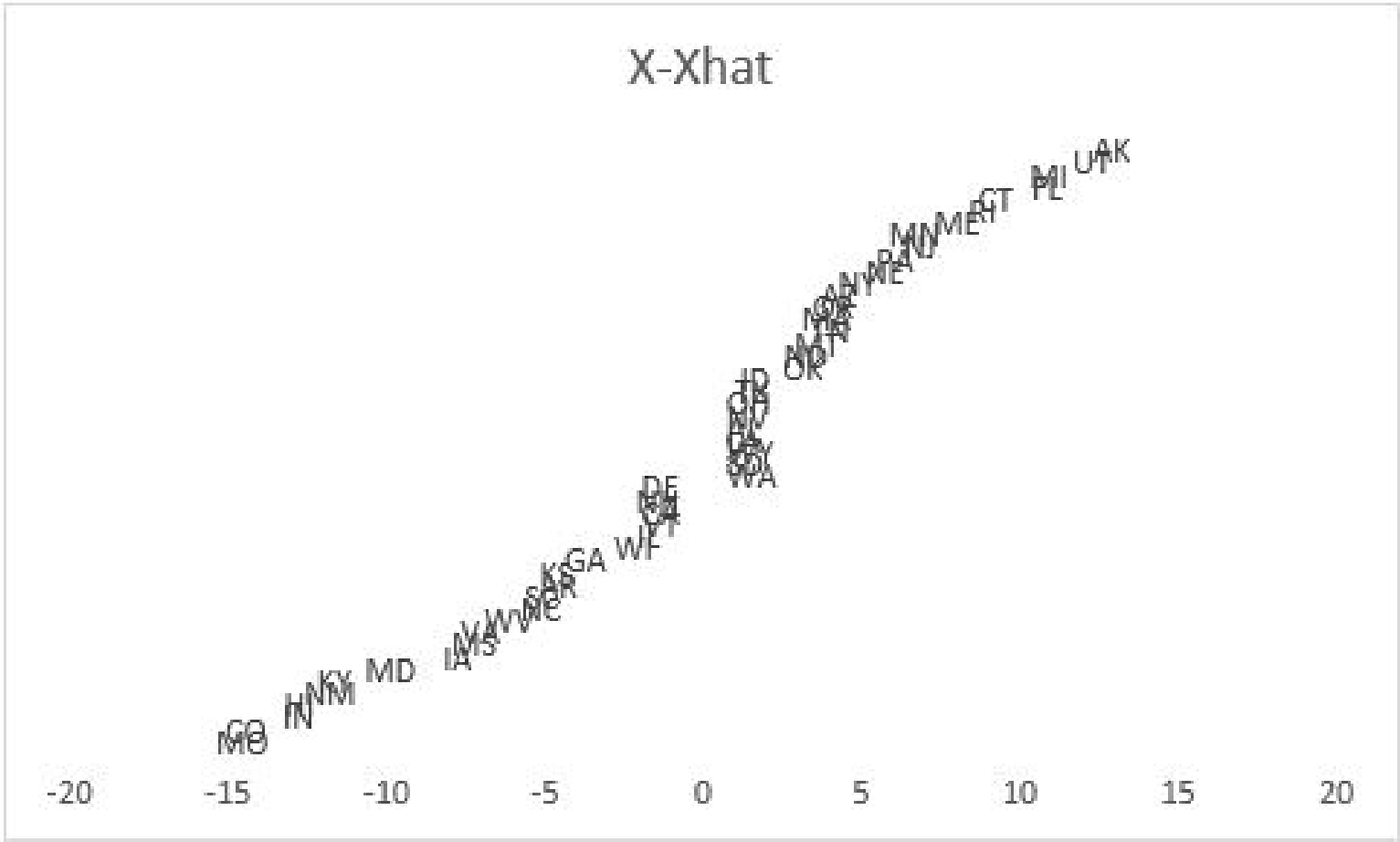


Figure 4



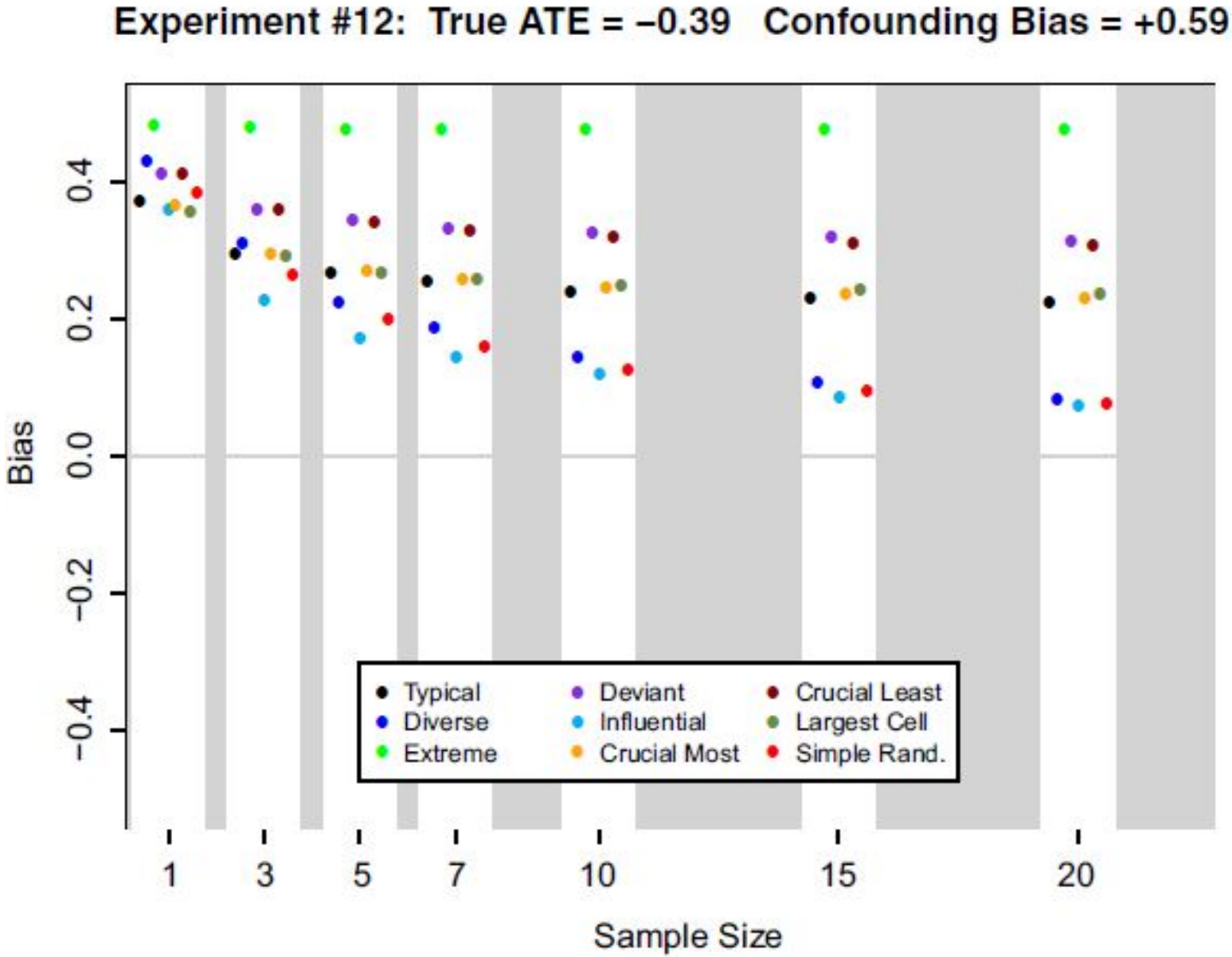
Confirmation

`http://www.dagitty.net/dags.html`

Herron and Quinn's Case Selection

Table 2. Possible Patterns of Potential Outcomes and Coarsest General Confounding Variable.

$Y_i(X_i = 0)$	$Y_i(X_i = 1)$	Z_i	
0	0	0	Never Succeed
0	1	1	Helped
1	0	2	Hurt
1	1	3	Always Succeed


Experiment #12: True ATE = −0.39 Confounding Bias = +0.59
Bias
0.4
0.2
0.0
−0.2
−0.4
Typical
Diverse
Extreme
Deviant
Influential
Crucial Most
Crucial Least
Largest Cell
Simple Rand.
1
3
5
7
10
15
20
Sample Size

Comparative Political Studies
46(10) 1301–1327

DOI: 10.1177/0010414012472469
cps.sagepub.com
SAGE

The Enduring Indispensability of the Controlled Comparison

Dan Slater[1] and Daniel Ziblatt[2]

Abstract

Do controlled comparisons still have a place in comparative politics? Long criticized by quantitatively oriented methodologists, this canonical approach has increasingly been critiqued by qualitative methodologists who recommend greater focus on within-case analysis and the confinement of causal explanations to particular cases. Such advice accords with a welcome shift from a combative "tale of two cultures" toward mutual respect for research combining qualitative and quantitative methods in the simultaneous pursuit of internal and external validity. This article argues that controlled comparisons remain indispensable amid this "multimethod turn," explicating how they too can generate both internal and external validity when their practitioners (a) craft arguments with general variables or mechanisms, (b) seek out representative variation, and (c) select cases that maximize control over alternative explanations. When controlled comparisons meet these standards, they continue to illuminate the world's great convergences and divergences across nation-states in a manner that no other methods can surpass.