

Lab 6: Simulation Challenges

2024-05-01

Split into 3 groups

Each group will work together to simulate one of 3 scenarios over 30 minutes, and then 1 member from each group will informally present their simulation for about 5 minutes.

Group 1: Difference in Differences and Parallel Trends

The **parallel trends** assumption in the difference-in-differences model says that absent treatment, the treatment group *would have* followed the same trend that the control group did.

Consider the following scenario: three states (A, B, and C) are considering implementing near identical investment policies. It is passed in states A and B, not in C. However, in state B, the day before the policy was to be implemented, it was halted by a Federal Judge. Additionally, state C is hit with a devastating natural disaster, plummeting them into a recession.

Suppose GDP (Y) in each state i at time t is determined by the following data generating process:

$$Y_{i,t} = \alpha + \gamma_i + \beta \cdot t + \delta \cdot \mathbb{I}(\text{policy}_{i,t}) + \eta \cdot t \cdot \mathbb{I}(i = C, t \geq 0) + \varepsilon_{i,t}$$

Where:

- $\alpha = 10$
- Fixed effects: $\gamma_A = 2, \gamma_B = -1, \gamma_C = 0.5$
- Time trend: $\beta = 0.5$
- $\delta = 1$, and $\mathbb{I}(\text{policy}_{i,t}) = 1$ if the policy is implemented, 0 otherwise
- $\eta = -0.7$
- $\varepsilon_{i,t} \sim N(0, 0.3)$

You, as an economist interested in the causal effect of the policy's implementation.

1.1 Create data Write a function that generates data for states A, B, and C over periods -10 to 10, where the policy is enacted at time $t = 0$.

Graph an example dataset, showing each state's output (Y) overtime on the same graph, depicting when the policy was implemented.

1.2 Simulate with C as Control Run 1000 iterations of a simulation where you run a difference in differences regression with state A as the treatment group and C as the control. Then graph the estimates of δ in a density plot.

1.3 Simulate with B as Control Repeat **1.2**, this time with B as the control.

1.4 BONUS! Run 1.2/1.3 again, this time using both B and C as control groups.

Group 2: Instrumental Variable

Consider an agriculture market where equilibrium is determined by the two following equations for Supply and Demand:

$$\begin{aligned}\text{Supply: } q_t &= \gamma \cdot p_t + \eta \cdot w_t + \nu_t \\ \text{Demand: } q_t &= \delta \cdot p_t + \varepsilon_t\end{aligned}$$

Where $\gamma, \eta > 0$, and $\delta < 0$. p_t and q_t are de-meaned measures of price and quantity of the good, w_t represents a de-meaned measure of weather, where higher levels of w_t increase crop yields.

2.1 Create Data To generate the data, first solve for the market clearing price p_t . The exogenous variables are drawn i.i.d. (each period) from the following distributions:

- $w_t \sim U(-3, 3)$
- $\nu_t \sim N(0, 1)$
- $\varepsilon \sim N(0, 2)$

Generate the data for 100 periods $t = 1$ to $t = 100$ where:

$$\gamma = 0.5, \quad \eta = 1.5, \quad \delta = -1$$

Graph and example data set with p_t against quantity demanded (q_t).

2.2 Estimating Demand Write a function that simulates 1000 iterations of this data set, running a regression estimating δ . Collect the estimates and graph on a density plot.

2.3 IV Now repeat **2.2** using w_t as an instrument for price.

2.4 Invalid Instrument? Generate the data again, this time splitting observations into two groups of equal size, and make the following correction:

- In odd periods: $\varepsilon_t \sim N(0, 2)$
- In even periods: $\varepsilon_t \sim N(0, 2) + 0.2 \cdot w_t$

That is, there is correlation between the demand disturbances and the weather (bad weather makes consumers discouraged).

Repeat **2.2** and **2.3** again with this data generating process

Group 3: Regression Discontinuity

Consider the following scenario:

You want to estimate the effect of college on earnings. A state college only accepts SAT math scores above 400, you have access to a high school's record of 1000 students' SAT scores (SAT_i) and annual income many years later (Y_i). (Assume everyone who scored over 400 went to college).

The data generating process for income is determined by the following equation:

$$Y_i = \alpha + \delta \cdot D_i + \beta_1 \cdot SAT_i \cdot (1 - D_i) + \beta_2 \cdot SAT_i \cdot D_i + \varepsilon_i$$

$$\text{Where : } D_i = \begin{cases} 1 & \text{if } SAT_i \geq 400 \\ 0 & \text{if } SAT_i < 400 \end{cases}$$

Assume $\beta_1 < \beta_2$, $\varepsilon_i \sim N(0, \sigma_\varepsilon)$.

Additionally, suppose SAT scores are distributed according to $SAT_i \sim N(500, 120)$ (pretty close to reality)

Which variable determines the causal effect of going to college on income?

3.1 Create Data: Let:

- $\alpha = 10,000$
- $\delta = 500$
- $\beta_1 = 2$
- $\beta_2 = 3$
- $\sigma_\varepsilon = 300$

Create a function that generates a dataset of 1,000 students according to this scenario.

Plot a sample generation, labeling the cut off and whether an observation went to college or not.

3.2 Regression with Same Slope Write a function that simulates generating the data and running a *Regression Discontinuity Design* linear regression that assumes the slope of the line is **the same on either side** of the cut off.

Run this simulation 1,000 times, collecting the estimates of δ , and plot the density of these estimates.

Are these estimates biased, unbiased, or ambiguous?

3.3 Regression with Different Slopes Repeat **3.2**, this time **allowing for differences** in the slope of the regression line on either side of the cut off.

Are these estimates biased, unbiased, or ambiguous?