Econometrics III: Applied Methods
NovaSBE, 2026

Problem Set 1: Randomization and Controls

Exercise A: Rubin Model and Roy Model

This exercise introduces the potential outcomes framework and the Roy model of self-selection. You will see how selection bias arises from individual optimization and how randomization solves it.

We consider a labor market training program that is offered to workers. This training may increase a given individual's wage from w_0 to w_1. Attending the program has a cost c. We assume that each agent knows his or her wage outcome, with and without training, with certainty. We also assume that both counterfactual wages in the population are heterogeneous, but the cost c is common to everyone. We write

$$w_1 = w_0 + \delta \tag{1}$$

where δ and w_0 are heterogeneous in the population. We impose $\delta > 0$ and, at the beginning of this exercise, we assume that δ and w_0 are uncorrelated. We will waive this assumption later.

1. What is the treatment impact of a given individual i? What is the average treatment impact in the population?
2. Write the model of individual's choice of whether to attend the training or not (called a **Roy model**). *You can assume that no factors other than the ones mentioned affect their decision.*
3. Based on that decision rule, people sort themselves into the training program. We observe average wages of treated and untreated. What does each of those averages measure?
4. If you compare average wages of the treated and untreated, what parameter do you estimate? Interpret that parameter.

Now assume that a randomized control trial is run. Three groups are randomly formed in the population. The first group ($Z = 0$) cannot access the training at all. The second group ($Z = 1$) faces the normal rules (cost c). The third group ($Z = 2$) is offered a subsidy $s < c$ **if** they choose to attend the training.

5. Compute the value of the average wage in each of these populations.
6. Note that in the data you would be able to estimate the proportion of trainees in each random group. Using this, show how you can recover the same parameter as in question 4 from the wage difference between groups $Z = 1$ and $Z = 0$.
7. Explain why the independence between δ and w_0 ensures that the naive wage comparison can estimate a treatment parameter without the experiment. Explain, however, why we cannot obtain the ATE.
8. Show that it is possible to identify the impact of the training on the population that is induced to participate by the subsidy s (and would not participate otherwise), using group $Z = 2$ and group $Z = 1$. Define that parameter formally and explain how you can compute it.

9. What can you estimate if $s = c$?

Now assume that δ and w_0 are correlated and assume that $w_0 = a + \rho\delta + \varepsilon$ where a and ρ are fixed parameters and ε is a residual uncorrelated with δ with mean zero.

10. Show that the naive comparison of the wages of treated and untreated absent an experiment would no longer identify a treatment parameter. Discuss the sign of the bias depending on ρ.

11. Show that comparing group $Z = 1$ with group $Z = 0$ identifies the same treatment parameter as before.

Exercise B: Power Calculation

This exercise develops the theory of statistical power and minimum detectable effects (MDE). You will learn how to determine sample sizes for experiments and understand the trade-offs involved.

We want to evaluate the wage impact of providing workers with a training program. We plan to randomize the training among a number of workers from a firm. Given the cost of the training itself and the cost of data collection, we need to determine the minimum sample size that should be required.

1. Call y the outcome variable, which will be the log wage and T the treatment variable ($T = 1$ if treated and $T = 0$ otherwise). We will assume the following model for the evaluation:
$$y = \alpha + \beta T + u$$
where β is the parameter of interest. Under the hypothesis that training treatment T is randomized, what does β estimate?

2. Show that in this model, the Ordinary Least Squares estimator of β has the following variance:
$$V(\hat{\beta}_{OLS}) = \frac{1}{\overline{T}(1-\overline{T})}\frac{\sigma^2}{N}$$
where $\overline{T}$ is the empirical average of T, $\sigma^2 = V(u)$ and N is *total* sample size. (You can either use the OLS matrix formula you are familiar with, or use the fact that this is identical to comparing the average means of treated and untreated and compute the variance of this mean difference.)

3. What is the probability of allocation into treatment that ensures the most precise estimator, for a given sample size N?

4. The dataset “power.csv” contains 3000 observations for the log wage, from a population that has not received training, but may enter the program if it is started. Using this data, estimate σ^2.

5. We consider minimum detectable effects (MDE) of power 80%, for a test of significance level 5%. Recall the expression of this MDE and interpret the impact of each of its components.

6. What sample size is required to achieve a MDE of 20% of the standard deviation σ? Call it N^*.

7. Each treatment costs 500 euros, and each survey costs 10 euros. Given that all individuals in the experiment, whether treated or untreated, have to be surveyed, what is the sample size and the rate of allocation into treatment that minimizes the cost of the experiment, while ensuring a MDE of 20% of the standard deviation?

8. Forget about the previous question and keep a random sample of size N^* from your initial sample.[1]

   Allocate treatment in the remaining data, at a 50% treatment probability, using the command "`gen T=(uniform()>.5)`". Set β equal to 20% of the standard deviation σ (this is now your true value of the parameter) and generate variable y according to the model in question 1, for every individual in the sample, depending on his treatment status. Then run the regression once. Is $\hat{\beta}$ significant at the 95% level? How likely was this to happen? Check this with the class: count the share of students who had a significant coefficient.

   Good Stata programmers should write a loop to generate again and again this procedure and check the share of significant coefficients.

9. Now assume that some workers may refuse the training when it is (randomly) offered to them. Let T denote whether a worker is offered the training, and let D denote whether they actually take it. Assume that only 60% of those offered the treatment actually take it (i.e., $P(D = 1|T = 1) = 0.6$ and $P(D = 1|T = 0) = 0$). How would you estimate the parameter β?

   Given your sample size, what is the MDE now? To simulate this, among individuals with $T = 1$, randomly assign 60% to have $D = 1$ (and the rest $D = 0$). Set $D = 0$ for all individuals with $T = 0$. Generate outcome y as $y = \alpha + \beta D + u$. Estimate β: did you expect it to be significant? How often should it be significant (this requires looking into the Normal distribution table)?

[1] For instance using the following code: `gen i=uniform(); sort i; keep if _n <= N*`.

Exercise C: Omitted Variable Bias and the Logic of Controls

This exercise develops the theory of omitted variable bias and explores when adding control variables helps identification—and when it hurts.

C.1 The OVB Formula

Suppose the true data generating process is:

$$Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i$$

where $\mathbb{E}[\varepsilon_i | X_i, Z_i] = 0$. You are interested in estimating β, the causal effect of X on Y. However, you do not observe Z and instead estimate the "short" regression:

$$Y_i = \tilde{\alpha} + \tilde{\beta} X_i + u_i$$

1.1 Use the formula for the probability limit of the OLS estimator to show that $\hat{\tilde{\beta}}$ from the short regression yields:

$$\text{plim } \hat{\tilde{\beta}} = \beta + \gamma \cdot \delta$$

where δ is defined as the slope coefficient from the population regression of the omitted variable Z on the included regressor X. This is the *omitted variable bias (OVB) formula.*

1.2 Interpret each component of the OVB formula. Based on this, explain the conditions under which there would be no bias even when Z is omitted from the regression.

C.2 Application: Returns to Education

Consider estimating the returns to education using a Mincer wage equation. Let Y_i be log wages, X_i be years of education, and Z_i be "ability" (which is unobserved). Suppose the true model is:

$$\log(wage_i) = \alpha + \beta \cdot educ_i + \gamma \cdot ability_i + \varepsilon_i$$

2.1 What sign do you expect for γ? What sign do you expect for δ? Based on these predictions, what is the expected sign of the omitted variable bias when ability is omitted?

2.2 A researcher argues: "The bias from omitting ability is probably small because ability only explains a small fraction of the total variation in wages." Is this argument correct? Explain your answer.

2.3 Suppose $\gamma = 0.3$ (a one standard deviation increase in ability raises log wages by 0.3), $\delta = 0.15$ (each year of education is associated with 0.15 SD higher ability), and the true return to education is $\beta = 0.06$ (6% per year). Calculate what the short regression (omitting ability) would estimate. Comment.

C.3 OVB in Practice (R or Stata)

Use the dataset `ovb_simulation.csv`, which contains 2,000 simulated observations. The variables are `education` (years), `ability` (standardized), `ability_proxy` (a noisy measure of ability), and `wage` (log hourly wage). The true data generating process is:

$$\texttt{wage}_i = 1.5 + 0.08 \cdot \texttt{education}_i + 0.25 \cdot \texttt{ability}_i + \varepsilon_i$$

3.1 Estimate the "short" regression of `wage` on `education` only (omitting ability). Report the coefficient on education and comment.

3.2 Now estimate the "long" regression that includes `ability` as a control. Report the coefficient on education. Then verify that the OVB formula holds empirically: (i) estimate $\hat{\gamma}$, (ii) estimate $\hat{\delta}$, (iii) compute $\hat{\gamma} \times \hat{\delta}$, and confirm this equals the difference between your short and long regression coefficients.

3.3 The variable `ability_proxy` is a noisy measure of true ability (it equals ability plus measurement error). Estimate the regression of `wage` on `education` and `ability_proxy`. Is the coefficient on education closer to the true value of 0.08 than in the short regression? Is it exactly equal to the true value? Why? Discuss what this implies for empirical research where we can only observe imperfect proxies for confounders.

C.4 Multiple Omitted Variables

Now suppose there are two omitted variables. The true model is:

$$Y_i = \alpha + \beta X_i + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \varepsilon_i$$

4.1 Derive the bias in $\hat{\hat{\beta}}$ when both Z_1 and Z_2 are omitted from the regression.

4.2 Is it possible for the biases from two omitted variables to cancel each other out? Under what conditions would this happen? Is this possibility a good reason to be unconcerned about omitted variable bias in practice? Explain why or why not.

C.5 Bad Control: Mediators

Not all control variables are "good." This section explores cases where adding a control variable *introduces* bias rather than removing it.

Consider a health intervention in a developing country. A program (X) provides nutrition supplements to children. The outcome (Y) is cognitive test scores. The supplements work by improving children's health status (M), which then improves cognition. The causal chain is:

$$X \longrightarrow M \longrightarrow Y$$

where M is a "mediator" (health status). There is no direct effect of the supplements on cognition except through health.

5.1 Suppose the structural equations are:

$$\begin{aligned} M_i &= \alpha_M + \theta X_i + \nu_i \\ Y_i &= \alpha_Y + \phi M_i + \eta_i \end{aligned}$$

where $\mathbb{E}[\nu_i | X_i] = 0$ and $\mathbb{E}[\eta_i | X_i, M_i] = 0$. What is the total causal effect of X on Y? Express your answer in terms of the structural parameters.

5.2 Now suppose a researcher estimates the regression:

$$Y_i = \tilde{\alpha} + \tilde{\beta} X_i + \tilde{\phi} M_i + u_i$$

What does $\tilde{\beta}$ estimate in this regression? Should a researcher who wants to estimate the total effect of the nutrition program control for M? Why or why not?

5.3 *(R or Stata)* Use the dataset `health_intervention.csv`, which contains `treatment` (0/1), `health_status` (0-100 scale, measured after treatment), and `test_score`. Estimate (i) the regression of test score on treatment only, and (ii) the regression of test score on treatment controlling for health status. Report both coefficients on treatment and explain why they differ. Which coefficient answers the policy question "should we implement this program?"

C.6 Bad Control: Colliders

A "collider" is a variable that is *caused by* both the treatment and the outcome (or by variables related to them). Controlling for a collider can create spurious associations where none exist.

Consider the following setup. Let X represent musical talent and Y represent physical attractiveness. Suppose these are completely independent in the population: $\text{Cov}(X, Y) = 0$. Let C represent "fame as an actor," which depends on both talent and attractiveness:

$$C_i = \alpha + \lambda_1 X_i + \lambda_2 Y_i + \xi_i$$

with $\lambda_1, \lambda_2 > 0$ (both talent and attractiveness increase the probability of becoming famous).

6.1 If you estimate the regression $Y_i = a + bX_i + e_i$ in the full population, what coefficient do you expect for $\hat{b}$? Why?

6.2 Now suppose you only observe famous actors (i.e., you condition on C being above some threshold). Among this selected sample of famous actors, would you expect talent and attractiveness to be positively correlated, negatively correlated, or uncorrelated? Provide an intuitive explanation for your answer.

6.3 Show algebraically that conditioning on C (either by restricting the sample or by controlling for C in a regression) induces a negative correlation between X and Y, even though they are unconditionally independent. You may use the fact that for jointly normal variables, $\text{Cov}(X, Y|C) = \text{Cov}(X, Y) - \text{Cov}(X, C)\text{Cov}(Y, C)/\text{Var}(C)$.

6.4 *(R or Stata)* Use the dataset `collider_simulation.csv`, which contains `talent` (standardized), `attractiveness` (standardized), `fame` (continuous score), and `is_famous` (indicator for top 20% of fame). Create a scatter plot of attractiveness (y-axis) against talent (x-axis). Plot the full sample in grey, then overlay the "famous" subsample (`is_famous == 1`) in a distinct color. Add regression lines for each group. Describe what you observe.

6.5 *(R or Stata)* Regress attractiveness on talent in three ways: (a) using the full sample, (b) restricting to the famous subsample only, (c) using the full sample but controlling for fame. Report the coefficient on talent from each regression. Explain why (b) and (c) give similar results, and why both differ from (a).

C.7 Deciding What to Control For

For each of the following scenarios, determine whether the variable Z should be included as a control in the regression. Justify your answer by identifying whether Z is a confounder, a mediator, a collider, or something else.

7.1 You are estimating the effect of a job training program (X) on wages (Y). You have data on workers' pre-program education levels (Z). Workers with more education are both more likely to enroll in training and have higher wages, regardless of training.

7.2 You are estimating the effect of a new teaching method (X) on student test scores (Y). You have data on hours spent studying (Z), which is measured after the teaching method is implemented. Students exposed to the new method tend to study more hours.

7.3 You are estimating the effect of hospital quality (X) on patient mortality (Y). You have data on whether the patient was admitted to the ICU (Z). Both sicker patients (who have higher mortality risk) and patients at low-quality hospitals are more likely to be admitted to the ICU.