Econometrics III: Applied Methods — NovaSBE, 2026

Problem Set 1 Solutions: Roy Model and Power Calculations

Exercise 1: Rubin Model and Roy Model

Setup: Training increases wages from w_0 to $w_1 = w_0 + \delta$ at cost c. Both $\delta > 0$ and w_0 are heterogeneous; initially $\delta \perp w_0$.

Q1. *(0.5 pts)* What is the treatment impact of a given individual i? What is the average treatment impact in the population?

Solution:

Individual treatment effect: $\tau_i = w_{1i} - w_{0i} = \delta_i$.
Average Treatment Effect (ATE):

$$\text{ATE} = \mathbb{E}[\delta_i] = \mathbb{E}[w_{1i} - w_{0i}]$$

Q2. *(0.5 pts)* Write the decision model of attending the training or not (called a Roy model).

Solution:

Individual i attends training if the benefit exceeds the cost:

$$D_i = \mathbb{1}\{w_{1i} - w_{0i} \geq c\} = \mathbb{1}\{\delta_i \geq c\}$$

Since agents know their counterfactual wages with certainty, they select into treatment based on comparing their individual gain δ_i to the common cost c.

Q3. *(1.5 pts)* 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?

Solution:

Observed wage: $w_i = D_i \cdot w_{1i} + (1 - D_i) \cdot w_{0i}$.
Average wage of treated (those with $D_i = 1$, i.e., $\delta_i \geq c$):

$$\begin{aligned}
\mathbb{E}[w_i|D_i = 1] &= \mathbb{E}[w_{1i}|\delta_i \geq c] \\
&= \mathbb{E}[w_{0i} + \delta_i|\delta_i \geq c] \\
&= \mathbb{E}[w_{0i}|\delta_i \geq c] + \mathbb{E}[\delta_i|\delta_i \geq c] \\
&= \mathbb{E}[w_{0i}] + \mathbb{E}[\delta_i|\delta_i \geq c] \quad (\text{using } \delta \perp w_0)
\end{aligned}$$

So this observed average measures the average baseline wage, plus the average treatment effect on the treated (the δ for those who select into treatment).

Average wage of untreated (those with $D_i = 0$, i.e., $\delta_i < c$):

$$\begin{aligned}
\mathbb{E}[w_i|D_i = 0] &= \mathbb{E}[w_{0i}|\delta_i < c] \\
&= \mathbb{E}[w_{0i}] \quad (\text{using } \delta \perp w_0)
\end{aligned}$$

The independence assumption $\delta \perp w_0$ is crucial: conditioning on δ (which determines selection) does not affect the distribution of baseline wages w_0. Both groups have the same expected baseline wage $\mathbb{E}[w_{0i}]$.

Q4. *(1.5 pts)* If you compare average wages of the treated and untreated, what parameter do you estimate? Interpret that parameter.

Solution:

The naive comparison estimates:

$$\mathbb{E}[w_i|D_i = 1] - \mathbb{E}[w_i|D_i = 0] = \mathbb{E}[w_{1i}|\delta_i \geq c] - \mathbb{E}[w_{0i}|\delta_i < c]$$

Add and subtract $\mathbb{E}[w_{0i}|\delta_i \geq c]$:

$$= \underbrace{\mathbb{E}[w_{1i} - w_{0i}|\delta_i \geq c]}_{\text{ATT}=\mathbb{E}[\delta_i|\delta_i \geq c]} + \underbrace{\mathbb{E}[w_{0i}|\delta_i \geq c] - \mathbb{E}[w_{0i}|\delta_i < c]}_{\text{Selection bias}}$$

Since $\delta \perp w_0$, the selection bias term equals zero:

$$\mathbb{E}[w_{0i}|\delta_i \geq c] = \mathbb{E}[w_{0i}|\delta_i < c] = \mathbb{E}[w_{0i}]$$

Therefore, the naive comparison identifies the **Average Treatment Effect on the Treated (ATT)**:

$$\boxed{\mathbb{E}[\delta_i|\delta_i \geq c]}$$

This is the average gain for those who *chose* to participate—people with above-average treatment effects (since they self-selected on $\delta_i \geq c$).

RCT Setup: Three groups: $Z = 0$ (no access), $Z = 1$ (normal cost c), $Z = 2$ (subsidized cost $c - s$).

Q5. *(1.5 pts)* Compute the value of the average wage in each of these populations.

Solution:

Let $p = \Pr(\delta_i \geq c)$ denote the share of trainees under normal cost, and $q = \Pr(\delta_i \geq c - s)$ under the subsidy.

Group $Z = 0$ (no access): Everyone is untreated.

$$\mathbb{E}[w_i|Z = 0] = \mathbb{E}[w_{0i}]$$

Group $Z = 1$ (cost c): Fraction p trains.

$$\begin{aligned}\mathbb{E}[w_i|Z = 1] &= p \cdot \mathbb{E}[w_{1i}|\delta_i \geq c] + (1 - p) \cdot \mathbb{E}[w_{0i}|\delta_i < c] \\ &= \mathbb{E}[w_{0i}] + p \cdot \mathbb{E}[\delta_i|\delta_i \geq c]\end{aligned}$$

(using $\delta \perp w_0$ to simplify)

Group $Z = 2$ (cost $c - s$): Fraction $q > p$ trains.

$$\mathbb{E}[w_i|Z = 2] = \mathbb{E}[w_{0i}] + q \cdot \mathbb{E}[\delta_i|\delta_i \geq c - s]$$

Q6. *(1 pt)* Note that you can estimate the proportion of trainees in each random group. Using this, show how you can recover the same parameter as in Q4 from the wage difference between groups $Z = 1$ and $Z = 0$.

Solution:

Let $p_1 = \Pr(D = 1|Z = 1)$ be the observed treatment rate in group 1 (and $p_0 = 0$ for group 0). This measures $p = \Pr(\delta_i \geq c)$.

The wage difference is:

$$\mathbb{E}[w_i|Z = 1] - \mathbb{E}[w_i|Z = 0] = p_1 \cdot \mathbb{E}[\delta_i|\delta_i \geq c]$$

Since we can estimate p_1 directly from the data, we recover the ATT via:

$$\boxed{\frac{\mathbb{E}[w_i|Z = 1] - \mathbb{E}[w_i|Z = 0]}{p_1} = \mathbb{E}[\delta_i|\delta_i \geq c] = \text{ATT}}$$

This is a **Wald estimator** (IV with Z as instrument for D).

Q7. *(1 pt)* Explain why the independence between δ and w_0 ensures that the naive wage comparison can estimate a treatment parameter without the experiment. Why can't we obtain ATE, though?

Solution:

Independence $\delta \perp w_0$ eliminates **selection on levels**: treated and untreated have the same baseline wage distribution. This removes the selection bias term in Q4, so naive comparison identifies ATT.

However, we **cannot obtain ATE** because there is still **selection on gains**: people with $\delta_i \geq c$ self-select into treatment. Since trainees have above-average treatment effects by construction, $\mathbb{E}[\delta_i|\delta_i \geq c] > \mathbb{E}[\delta_i]$ (unless δ is constant).

The ATE would require observing treatment effects for non-participants, which is impossible under self-selection.

Q8. *(1.5 pts)* 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.

Solution:

Define the **compliers**: individuals with $c - s \leq \delta_i < c$. These would train with subsidy but not without.

The **Local Average Treatment Effect (LATE)** for compliers:

$$\text{LATE} = \mathbb{E}[\delta_i|c - s \leq \delta_i < c]$$

From Q5:

$$\mathbb{E}[w_i|Z = 2] - \mathbb{E}[w_i|Z = 1] = q \cdot \mathbb{E}[\delta_i|\delta_i \geq c - s] - p \cdot \mathbb{E}[\delta_i|\delta_i \geq c]$$

The additional trainees in group 2 are exactly the compliers. Their contribution:

$$\mathbb{E}[w_i|Z = 2] - \mathbb{E}[w_i|Z = 1] = (q - p) \cdot \mathbb{E}[\delta_i|c - s \leq \delta_i < c]$$

Therefore:

$$\boxed{\text{LATE} = \frac{\mathbb{E}[w_i|Z=2] - \mathbb{E}[w_i|Z=1]}{q-p}}$$

where $q - p = \Pr(D = 1|Z = 2) - \Pr(D = 1|Z = 1)$ is the compliance rate.

Q9. *(0.5 pts)* What can you estimate if $s = c$?

Solution:

If $s = c$, the training is **free** for group $Z = 2$. Since $\delta > 0$ for everyone, *all* individuals in group 2 will take the training: $q = 1$.

Comparing $Z = 2$ (all trained) with $Z = 0$ (none trained):

$$\mathbb{E}[w_i|Z=2] - \mathbb{E}[w_i|Z=0] = \mathbb{E}[w_{1i}] - \mathbb{E}[w_{0i}] = \boxed{\text{ATE} = \mathbb{E}[\delta_i]}$$

With full subsidy, we eliminate selection entirely and identify the population average treatment effect.

Now assume δ and w_0 are correlated: $w_0 = a + \rho\delta + \varepsilon$ where $\varepsilon \perp \delta$.

Q10. *(1.5 pts)* 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.

Solution:

The naive comparison now includes selection bias (from Q4):

$$\begin{aligned}&\mathbb{E}[w_i|D=1] - \mathbb{E}[w_i|D=0]\\ &= \underbrace{\mathbb{E}[\delta_i|\delta_i \geq c]}_{\text{ATT}} + \underbrace{\mathbb{E}[w_{0i}|\delta_i \geq c] - \mathbb{E}[w_{0i}|\delta_i < c]}_{\text{Selection bias}}\end{aligned}$$

Since $w_{0i} = a + \rho\delta_i + \varepsilon_i$ and $\varepsilon \perp \delta$:

$$\mathbb{E}[w_{0i}|\delta_i \geq c] - \mathbb{E}[w_{0i}|\delta_i < c] = \rho\left(\mathbb{E}[\delta_i|\delta_i \geq c] - \mathbb{E}[\delta_i|\delta_i < c]\right)$$

The term in parentheses is **positive** (treated have higher δ).
Sign of bias:

- $\rho > 0$: Upward bias. High-gain individuals also have high baseline wages. The naive comparison overstates ATT.
- $\rho < 0$: Downward bias. High-gain individuals have low baseline wages (e.g., training helps the disadvantaged most). The naive comparison understates ATT.
- $\rho = 0$: No bias (back to the independent case).

Q11. *(1 pt)* Show that comparing group $Z = 1$ with group $Z = 0$ identifies the same treatment parameter as before.

Solution:

The key insight: **randomization breaks the selection bias**.

In the RCT:

$$\mathbb{E}[w_i|Z=1] - \mathbb{E}[w_i|Z=0] = \mathbb{E}[w_i|Z=1] - \mathbb{E}[w_{0i}]$$

Since Z is randomly assigned, $Z \perp (w_0, \delta)$. The treatment probability in group 1 is still $p = \Pr(\delta_i \geq c)$, and:

$$\mathbb{E}[w_i|Z=1] = \mathbb{E}[w_{0i}] + p \cdot \mathbb{E}[\delta_i|\delta_i \geq c]$$

Therefore:

$$\frac{\mathbb{E}[w_i|Z=1] - \mathbb{E}[w_i|Z=0]}{p} = \mathbb{E}[\delta_i|\delta_i \geq c] = \text{ATT}$$

The Wald/IV estimator recovers ATT regardless of the correlation between δ and w_0, because randomization ensures the control group provides a valid counterfactual for baseline outcomes.

Exercise 2: Power Calculation

Q1. *(0.5 pts)* Call y the outcome variable (log wage) and T the treatment variable. We assume the model $y = \alpha + \beta T + u$. Under randomization, what does β estimate?

Solution:

Under randomization, $T \perp (Y(0), Y(1))$, so:

$$\beta = \mathbb{E}[y|T=1] - \mathbb{E}[y|T=0] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \boxed{\text{ATE}}$$

The coefficient estimates the Average Treatment Effect of training on log wages.

Q2. *(1.5 pts)* Show that in this model, the OLS estimator of β has variance:

$$V(\hat{\beta}_{OLS}) = \frac{1}{\bar{T}(1-\bar{T})} \cdot \frac{\sigma^2}{N}$$

where $\bar{T}$ is the treatment share, $\sigma^2 = V(u)$, and N is total sample size.

Solution:

Method 1 (Difference in means): The OLS estimator equals the difference in sample means:

$$\hat{\beta} = \bar{y}_1 - \bar{y}_0$$

where $\bar{y}_1$ is the mean for treated ($N_1 = N\bar{T}$ observations) and $\bar{y}_0$ for control ($N_0 = N(1-\bar{T})$).

Variance:

$$V(\hat{\beta}) = V(\bar{y}_1) + V(\bar{y}_0) = \frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_0} = \sigma^2 \left(\frac{1}{N\bar{T}} + \frac{1}{N(1-\bar{T})} \right)$$

$$= \frac{\sigma^2}{N} \cdot \frac{1}{\bar{T}(1-\bar{T})}$$

Method 2 (OLS formula): Using $(X'X)^{-1}\sigma^2$:

$$V(\hat{\beta}) = \frac{\sigma^2}{\sum_i (T_i - \bar{T})^2} = \frac{\sigma^2}{N \cdot \bar{T}(1-\bar{T})}$$

since $\sum_i (T_i - \bar{T})^2 = N\bar{T}(1-\bar{T})$ for binary T.

Q3. *(0.5 pts)* What is the probability of allocation into treatment that ensures the most precise estimator, for a given sample size N?

Solution:

We minimize $V(\hat{\beta})$ by maximizing $\bar{T}(1-\bar{T})$.

Taking the derivative: $\frac{d}{d\bar{T}}[\bar{T}(1-\bar{T})] = 1 - 2\bar{T} = 0 \implies \bar{T} = 0.5$

Therefore, $\boxed{\bar{T} = 0.5}$ (50-50 allocation) minimizes variance.

Intuition: Equal allocation maximizes the "effective sample size" for estimating the treatment effect.

Q4. *(0.5 pts)* The dataset "power.RData" contains 3000 observations for the log wage, from a population that has not received training. Using this data, estimate σ^2.

Solution:

```
# R code:
load("assignments/data/power.RData")
sigma2 <- var(power$lnw)
sigma <- sqrt(sigma2)
```

Results:

```
sigma^2 = 0.1165
sigma   = 0.341
```

The residual variance (outcome variance in the untreated population) is approximately $\boxed{\sigma^2 \approx 0.117}$.

Q5. *(1 pt)* 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.

Solution:

The MDE formula for a two-sided test at significance α and power $1 - \kappa$:

$$\boxed{\text{MDE} = (z_{1-\alpha/2} + z_{1-\kappa}) \cdot \sqrt{\frac{\sigma^2}{N \cdot \bar{T}(1-\bar{T})}}}$$

For $\alpha = 0.05$ (5%) and power = 80% ($\kappa = 0.2$):

- $z_{0.975} = 1.96$ (critical value for significance)
- $z_{0.80} = 0.84$ (critical value for power)

So: $\text{MDE} = 2.8 \cdot \text{SE}(\hat{\beta})$
Interpretation of components:

- σ^2: Higher outcome variance → larger MDE (harder to detect effects)
- N: Larger sample → smaller MDE (more precise estimates)
- $\bar{T}(1-\bar{T})$: Maximized at 50-50 allocation; imbalanced designs increase MDE
- $z_{1-\alpha/2}$: Lower significance level (smaller α) → larger MDE
- $z_{1-\kappa}$: Higher power → larger MDE (stricter detection requirement)

Q6. *(1 pt)* What sample size is required to achieve a MDE of 20% of the standard deviation σ? Call it N^*.

Solution:

Target: MDE $= 0.2\sigma$. With 50-50 allocation ($\bar{T} = 0.5$):

$$0.2\sigma = 2.8 \cdot \sqrt{\frac{\sigma^2}{N^* \cdot 0.25}} = 2.8 \cdot \frac{2\sigma}{\sqrt{N^*}}$$

Solving for N^*:

$$\sqrt{N^*} = \frac{2.8 \times 2}{0.2} = 28 \implies \boxed{N^* = 784}$$

General formula: For MDE $= k\sigma$:

$$N^* = \frac{4 \cdot (z_{1-\alpha/2} + z_{1-\kappa})^2}{k^2} = \frac{4 \times 2.8^2}{k^2} = \frac{31.36}{k^2}$$

Q7. *(2 pts)* Each treatment costs 500 euros, and each survey costs 10 euros. Given that all individuals in the experiment must be surveyed, what is the sample size and allocation rate that minimizes cost while ensuring MDE $= 0.2\sigma$?

Solution:

Cost function: $C = 10N + 500N_T = 10N + 500N\bar{T} = N(10 + 500\bar{T})$
MDE constraint:

$$0.2\sigma = 2.8 \cdot \sqrt{\frac{\sigma^2}{N\bar{T}(1-\bar{T})}} \implies N \cdot \bar{T}(1-\bar{T}) = \frac{2.8^2}{0.04} = 196 \implies N = \frac{196}{\bar{T}(1-\bar{T})}$$

Substituting into cost:

$$C = \frac{196(10 + 500\bar{T})}{\bar{T}(1-\bar{T})}$$

Minimize over $\bar{T}$: Taking FOC and solving (or numerically):

$$\frac{dC}{d\bar{T}} = 0 \implies \bar{T}^* = \sqrt{\frac{10}{500 + 10}} \approx 0.14$$

Optimal allocation: $\boxed{\bar{T}^* \approx 14\%}$ (treat fewer people because treatment is expensive)
Required sample size:

$$N^* = \frac{196}{0.14 \times 0.86} \approx 1,628$$

Total cost: $C = 1628 \times (10 + 500 \times 0.14) \approx 1628 \times 80 =$ 130,240 euros
Compare to 50-50: $N = 784$, $C = 784 \times (10 + 250) = 203,840$ euros.

The optimized allocation reduces cost from 204k to 130k euros by using more subjects (1,628 vs 784) but treating a smaller fraction (14% vs 50%). Since treatment is 50 times more expensive than surveys, it's cost-effective to survey more people but treat fewer of them.

Q8. *(1 pt)* Keep a random sample of size $N^* = 784$. Allocate treatment at 50% probability. Set $\beta = 0.2\sigma$ and generate $y = \alpha + \beta T + u$. Run the regression. Is $\hat{\beta}$ significant at 95%? How likely was this?

Solution:

```
# R code:
load("assignments/data/power.RData")
set.seed(123)  # for reproducibility

sigma <- sd(power$lnw)
beta_true <- 0.2 * sigma
N_star <- 784

# Draw random sample
idx <- sample(1:nrow(power), N_star)
data <- power[idx, ]

# Assign treatment
data$T <- rbinom(N_star, 1, 0.5)

# Generate outcome: y = alpha + beta*T + u
# Use original lnw as the baseline (untreated outcome)
data$y <- data$lnw + beta_true * data$T

# Run regression
model <- lm(y ~ T, data = data)
summary(model)

# Check significance
p_value <- summary(model)$coefficients["T", "Pr(>|t|)"]
is_significant <- p_value < 0.05
```

Expected result: The coefficient should be significant approximately **80% of the time**—this is exactly what 80% power means!

By design, we set the true effect equal to the MDE. At the MDE, we have exactly 80% probability of rejecting the null when it is false.

Class exercise: Count significant results across students. Should be close to 80%.

Q9. *(1.5 pts)* Now assume only 60% of those offered treatment actually take it: $P(D = 1|T = 1) = 0.6$ and $P(D = 1|T = 0) = 0$. How would you estimate β? What is the new MDE?

Solution:

Estimation: With non-compliance, the naive regression of y on D is biased. Use **Instrumental Variables** with assignment T as instrument for actual treatment D.

The **Wald/IV estimator**:

$$\hat{\beta}_{IV} = \frac{\mathbb{E}[y|T=1] - \mathbb{E}[y|T=0]}{\mathbb{E}[D|T=1] - \mathbb{E}[D|T=0]} = \frac{\text{ITT}}{0.6 - 0} = \frac{\text{ITT}}{0.6}$$

This identifies the **LATE** for compliers (those who take treatment when offered).

New MDE: The IV estimator has larger variance than OLS:

$$\text{SE}(\hat{\beta}_{IV}) = \frac{\text{SE}(\hat{\beta}_{ITT})}{0.6}$$

Therefore:

$$\text{MDE}_{IV} = \frac{\text{MDE}_{ITT}}{0.6} = \frac{0.2\sigma}{0.6} = \boxed{0.333\sigma}$$

The MDE increases by a factor of $1/0.6 \approx 1.67$. Non-compliance reduces effective power.

Expected significance rate: With true $\beta = 0.2\sigma$ but MDE $= 0.333\sigma$, the true effect is only $0.2/0.333 = 0.6$ of the MDE.

The test statistic is: $z = \frac{\beta}{\text{SE}} = \frac{0.6\times\text{MDE}}{\text{SE}} = 0.6 \times 2.8 = 1.68$

Power $= \Phi(1.68 - 1.96) + \Phi(-1.68 - 1.96) \approx \Phi(-0.28) \approx 39\%$

So the coefficient should be significant only about **39%** of the time (down from 80%).

```
# R simulation code:
data$D <- ifelse(data$T == 1, rbinom(sum(data$T), 1, 0.6), 0)
data$y <- data$lnw + beta_true * data$D

library(AER)
iv_model <- ivreg(y ~ D | T, data = data)
summary(iv_model)
```