Liming Lin
Professor Clement de Chaisemartinx
Econometrics III
Problem Set 5
Oct. 12th, 2025

1 What we would do in a world without non response.

a

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

b

$$\beta_1 = \frac{Cov(D_i, Y_i)}{Var(D_i)}$$

c

The Average Treatment Effect (ATE) is defined as:

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

The OLS coefficient β_1 is:

$$\beta_1 = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

Because of random assignment, the treatment status D_i is independent of potential outcomes $(Y_i(1), Y_i(0))$. Therefore:

$$E[Y_i \mid D_i = 1] = E[Y_i(1) \mid D_i = 1] = E[Y_i(1)]$$

$$E[Y_i \mid D_i = 0] = E[Y_i(0) \mid D_i = 0] = E[Y_i(0)]$$

Substituting these into the formula for β_1 shows that:

$$\beta_1 = E[Y_i(1)] - E[Y_i(0)] = ATE$$

2 Why non-response is a problem.

a

We only observe the outcome Y_i for the households that returned the questionnaire.

b

$$E(Y_i|D_i = 1, R_i = 1) and E(Y_i|D_i = 0, R_i = 1)$$

c

The expectation of Y_i conditional on $D_i = 1$ can be expressed as the weighted average of the expectation of Y_i conditional on $D_i = 1$ and the two possible events for R_i (responding or not responding).
For the treated group ($D_i = 1$):

$$E(Y_i \mid D_i = 1) = E[E(Y_i \mid D_i = 1, R_i) \mid D_i = 1]$$

$$= E(Y_i \mid D_i = 1, R_i = 1)P(R_i = 1 \mid D_i = 1) + E(Y_i \mid D_i = 1, R_i = 0)P(R_i = 0 \mid D_i = 1)$$

The same logic applies to the control group ($D_i = 0$):

$$E(Y_i \mid D_i = 0) = E[E(Y_i \mid D_i = 0, R_i) \mid D_i = 0]$$

$$= E(Y_i \mid D_i = 0, R_i = 1)P(R_i = 1 \mid D_i = 0) + E(Y_i \mid D_i = 0, R_i = 0)P(R_i = 0 \mid D_i = 0)$$

d

No, in general, we cannot estimate the ATE. The coefficient estimated on the sub-sample of responders is:

$$\hat{\beta}_1 = E[Y_i \mid D_i = 1, R_i = 1] - E[Y_i \mid D_i = 0, R_i = 1]$$

The true ATE is:

$$ATE = E[Y_i(1)] - E[Y_i(0)]$$

Because of non-response, we can no longer rely on randomization to ensure that the sample of responders is representative of the full sample. Therefore, there is no reason to assume that $\hat{\beta}_1 = ATE$.

3 How can we deal with non-response.

a

Under no defiers assumption, for $R_i = 1, D_i = 0$ we have the always respondents.
For $R_i = 1, D_i = 1$ we have the always respondents and the response compliers.

b

The "no response defiers" assumption $R_i(1) \geq R_i(0)$ could be violated in several scenarios:

- Shame associated with failure: A household that receives a loan but whose business fails may feel embarrassed and choose not to respond, whereas they would have responded if they had not received the loan and thus not experienced that specific failure.
- Desire to hide success: A household whose business becomes very successful after receiving the loan might wish to avoid attention from tax authorities or others in the community, causing them not to respond.

Overall, these cases are extremely rare.

c

For simplicity, we define Probability of always respondents
$P(\text{Always Respondent}) = P(R_i(0) = R_i(1) = 1)$
Probability of response compliers $P(\text{Response Complier}) = P(R_i(0) = 0, R_i(1) = 1)$
By the law of total probability, and using the definitions of the groups:

$$P(R_i = 1 \mid D_i = 1) = P(\text{Always Respondent}) + P(\text{Response Complier})$$
$$P(R_i = 1 \mid D_i = 0) = P(\text{Always Respondent}) + P(\text{Response Defier})$$

Under the no response defiers assumption, $P(\text{Response Defier}) = 0$.
So the second equation becomes:

$$P(R_i = 1 \mid D_i = 0) = P(\text{Always Respondent})$$

Subtracting the second equation from the first:

$$P(R_i = 1 \mid D_i = 1) - P(R_i = 1 \mid D_i = 0) = P(\text{Response Complier})$$

By definition, a Response Complier is someone for whom $R_i(0) = 0$ and $R_i(1) = 1$.
Therefore:

$$P(R_i = 1 \mid D_i = 1) - P(R_i = 1 \mid D_i = 0) = P(R_i(0) = 0, R_i(1) = 1)$$

d

First Equation

In the group $\{R_i = 1, D_i = 0\}$. These are the responders in the control group, which means their potential outcome for responding to control is $R_i(0) = 1$. This group is composed of Always Respondents $(R_i(0) = 1, R_i(1) = 1)$ and Response Defiers $(R_i(0) = 1, R_i(1) = 0)$.
Under the no response defiers assumption, the group of Response Defiers is empty. Therefore, the group $\{R_i = 1, D_i = 0\}$ consists solely of Always Respondents.
For anyone in the control group $(D_i = 0)$, their observed outcome is their potential outcome without treatment: $Y_i = Y_i(0)$.
Therefore, the expected outcome for this group is the expected potential outcome without treatment, conditional on being an Always Respondent:

$$E(Y_i \mid R_i = 1, D_i = 0) = E(Y_i(0) \mid \text{Always Respondent}) = E(Y_i(0) \mid R_i(0) = 1, R_i(1) = 1)$$

Second Equation

The group $\{R_i = 1, D_i = 1\}$ is a mix of Always Respondents (AR) and Response Compliers (RC). Its expected outcome is a weighted average:

$$E(Y_i \mid R_i = 1, D_i = 1) = E(Y_i(1) \mid \text{AR})P(\text{AR} \mid R_i = 1, D_i = 1) + E(Y_i(1) \mid \text{RC})P(\text{RC} \mid R_i = 1, D_i = 1)$$

Using Bayes' theorem, we find the weights. The probability of being an AR given that you are a treated responder is:

$$P(\text{AR} \mid R_i = 1, D_i = 1) = \frac{P(R_i = 1 \mid D_i = 1, \text{AR})P(\text{AR})}{P(R_i = 1 \mid D_i = 1)} = \frac{1 \cdot P(\text{AR})}{P(R_i = 1 \mid D_i = 1)} = \frac{P(R_i = 1 \mid D_i = 0)}{P(R_i = 1 \mid D_i = 1)}$$

The probability of being a RC is the remainder:

$$P(\text{RC} \mid R_i = 1, D_i = 1) = 1 - P(\text{AR} \mid R_i = 1, D_i = 1) = 1 - \frac{P(R_i = 1 \mid D_i = 0)}{P(R_i = 1 \mid D_i = 1)}$$

Substituting these weights back in gives the final formula:

$$E(Y_i \mid R_i = 1, D_i = 1) = E(Y_i(1) \mid \text{AR})\frac{P(R_i = 1 \mid D_i = 0)}{P(R_i = 1 \mid D_i = 1)} + E(Y_i(1) \mid \text{RC}) \left(1 - \frac{P(R_i = 1 \mid D_i = 0)}{P(R_i = 1 \mid D_i = 1)}\right)$$

Lower and Upper Bounds

The ATE for Always Respondents is $ATE_{AR} = E(Y_i(1)|\text{AR}) - E(Y_i(0)|\text{AR})$.
From the first part of 3d, we know $E(Y_i(0)|\text{AR}) = E(Y_i|R_i = 1, D_i = 0)$.
From the second part, we can isolate $E(Y_i(1)|\text{AR})$ by rearranging
and we also simplify the notation for probabilities:

$$P_1 = P(R_i = 1|D_i = 1)$$
$$P_0 = P(R_i = 1|D_i = 0)$$
$$E(Y_i(1)|\text{AR}) = E(Y_i|R_i = 1, D_i = 1)\frac{P_1}{P_0} - E(Y_i(1)|\text{RC})\frac{P_1 - P_0}{P_0}$$

where $P_1 = P(R_i = 1|D_i = 1)$ and $P_0 = P(R_i = 1|D_i = 0)$.
To find the lower bound for ATE_{AR}, we set the unknown term $E(Y_i(1)|\text{RC})$
to its highest possible value, 1:

$$B_-^{L1} = \left(E(Y_i|R_i = 1, D_i = 1)\frac{P_1}{P_0} - \frac{P_1 - P_0}{P_0}\right) - E(Y_i|R_i = 1, D_i = 0)$$

To find the upper bound for ATE_{AR}, we set $E(Y_i(1)|\text{RC})$ to its lowest possible value, 0:

$$B_+^{L1} = \left(E(Y_i|R_i = 1, D_i = 1)\frac{P_1}{P_0}\right) - E(Y_i|R_i = 1, D_i = 0)$$

e

The length of the interval is the upper bound minus the lower bound:

$$\text{Length} = B_+^{L1} - B_-^{L1}$$
$$= \left(E(Y_i \mid R_i = 1, D_i = 1)\frac{P_1}{P_0} - E(Y_i \mid R_i = 1, D_i = 0)\right)$$

$$-\left(E(Y_i \mid R_i = 1, D_i = 1)\frac{P_1}{P_0} - \frac{P_1 - P_0}{P_0} - E(Y_i \mid R_i = 1, D_i = 0)\right)$$
$$= \frac{P_1 - P_0}{P_0}$$

Plugging in the given values:

$$\text{Length} = \frac{0.63 - 0.6}{0.6} = \frac{0.03}{0.6} = 0.05$$

f

To estimate the upper bound B_+^{L1}, we define a new outcome variable Y_i^{+L} such that:

$$Y_i^{+L} = \begin{cases} Y_i & \text{if } D_i = 0, \\ Y_i \frac{P_1}{P_0} & \text{if } D_i = 1. \end{cases}$$

A regression of Y_i^{+L} on D_i for the sample of responders will yield a coefficient equal to B_+^{L1}.
To estimate the lower bound B_-^{L1}, we define a second new outcome variable Y_i^{-L} such that:

$$Y_i^{-L} = \begin{cases} Y_i & \text{if } D_i = 0, \\ Y_i \frac{P_1}{P_0} - \frac{P_1 - P_0}{P_0} & \text{if } D_i = 1. \end{cases}$$

A regression of Y_i^{-L} on D_i for the sample of responders will yield a coefficient equal to B_-^{L1}.

g

No, the standard confidence intervals will not be correct. The variables Y_i^{-L} and Y_i^{+L} are constructed using the probabilities $P(R_i = 1|D_i = 1)$ and $P(R_i = 1|D_i = 0)$, which are unknown population parameters.
In practice, these probabilities must be estimated from the sample data. This introduces an additional source of sampling variation that is not accounted for by the standard OLS variance formula.

h

The group of treated responders $(D_i = 1, R_i = 1)$ is a mix of Always Respondents (AR) and Response Compliers (RC). We want to find the bounds for the average treated outcome for just the ARs, $E(Y_i(1)|\text{AR})$.
Let $p = P(\text{AR}|R_i = 1, D_i = 1)$ be the proportion of Always Respondents within this mixed group.
The Lower Bound:
To find the lowest possible average outcome for the ARs, we assume they are the fraction ‘p‘ of individuals with the lowest outcomes in the observed group. The left side of the inequality represents this: we calculate the average outcome for the bottom ‘p‘ quantile of the treated responders.

$E(Y_i|R_i = 1, D_i = 1, Y_i \leq G^{-1}(p))$
The Upper Bound (Optimistic Scenario):
To find the highest possible average outcome for the ARs, we assume they are the fraction ‘p‘ of individuals with the highest outcomes. The right side of the inequality represents this: we calculate the average outcome for the top ‘p‘ quantile (i.e., above the $1 - p$ quantile) of the treated responders.
$E(Y_i|R_i = 1, D_i = 1, Y_i \geq G^{-1}(1 - p))$
The true average outcome for the Always Respondents, $E(Y_i(1)|\text{AR})$, must lie between these two extreme scenarios.

i

The Average Treatment Effect for Always Respondents is given by:

$$ATE_{AR} = E(Y_i(1)|\text{AR}) - E(Y_i(0)|\text{AR})$$

We know that $E(Y_i(0)|\text{AR}) = E(Y_i|R_i = 1, D_i = 0)$. For $E(Y_i(1)|\text{AR})$, we use the bounds derived in 3h.
Lower Bound:

$$B_{-}^{L2} = E(Y_i|R_i = 1, D_i = 1, Y_i \leq G^{-1}(p)) - E(Y_i|R_i = 1, D_i = 0)$$

Upper Bound:

$$B_{+}^{L2} = E(Y_i|R_i = 1, D_i = 1, Y_i \geq G^{-1}(1 - p)) - E(Y_i|R_i = 1, D_i = 0)$$