PS Week 7, Econometrics 3

The TA will randomly pick two questions and grade them.

Dealing with non-response and sample selection in randomized experiments: bounds under a monotonicity assumption (inspired from: "Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. The Review of Economic Studies, 76(3), 1071-1102.")

Microcredit is usually perceived as a way to ensure that households who do not have access to the regular credit market can get loans to start, expand, or just maintain their businesses. Assume you want to measure whether microcredit indeed reaches this goal or not. On that purpose, you run a randomized experiment. Within a sample of n households, you randomly allocate loans to some households and not to others. Let's assume that compliance is perfect: only households allocated to the treatment group receive a loan, and all of them do so. One year later you send by mail a questionnaire to all households included in the randomization asking them whether someone in the household runs a business or not. Their answer to that question is coded as a dummy variable Y_i for whether someone in household i runs his own business or not. D_i is a dummy for whether household i was assigned to the treatment or to the control group. $Y_i(1)$ denotes whether someone in household i will run his own business one year after the lottery if household i receives the loan, and $Y_i(0)$ denotes whether someone in household i will run his own business one year after the lottery if household i does not receive the loan.

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

For now, let's assume that all households participating in the experiment sent back their questionnaire.

a) Which very simple regression can you run to measure the effect of receiving a loan on the probability that someone in the household runs his own business?

b) Give a formula for the population coefficient in that regression which will measure the effect of the treatment.

c) Show that because treatment has been randomly allocated, this coefficient is equal to the average treatment effect.

2) Why non-response is a problem.

Now assume that some households did not send back their questionnaires. This could be because they no longer live at the same place so they did not receive the questionnaire, or because they forgot / did not want to send it back. To model this, let us introduce a dummy variable R_i equal to 1 if household i sends back its questionnaire.

a) For which households do we observe Y_i?

b) As a result, among the 4 following quantities, which are those we can estimate from the sample: $E(Y_i|D_i = 1, R_i = 1)$, $E(Y_i|D_i = 1, R_i = 0)$, $E(Y_i|D_i = 0, R_i = 1)$, $E(Y_i|D_i = 0, R_i = 0)$?

c) Show that

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

and

$$E(Y_i|D_i = 0) = P(R_i = 1|D_i = 0)E(Y_i|D_i = 0, R_i = 1)+P(R_i = 0|D_i = 0)E(Y_i|D_i = 0, R_i = 0).$$

d) Can we estimate the ATE when we have non-respondents?

3) How can we deal with non-response

In this question, we assume that $P(R_i = 1|D_i = 1) \neq P(R_i = 1|D_i = 0)$. Without loss of generality, assume that $P(R_i = 1|D_i = 1) > P(R_i = 1|D_i = 0)$: the probability to respond is strictly higher if a household is assigned to the treatment group than if she is assigned to the control group. Since treatment is randomly allocated, the difference in response rates across the treatment and control groups cannot come from the fact that these groups are different. It can only come from the fact that treatment has an effect on response. For instance, control group households may resent the fact they did not receive a micro-credit, so they may be unwilling to respond the survey. To account for this, let $R_i(1)$ and $R_i(0)$ respectively denote household i's potential responses to the survey in a world in which she is assigned / not assigned to the treatment group. $R_i(0)$ is equal to 1 (resp. 0) if household i will respond (resp. not respond) to the survey if she is not assigned to the treatment group. $R_i(1)$ is equal to 1 (resp. 0) if household i will respond (resp. not respond) to the survey if she is assigned to the treatment group. Observed response is $R_i = D_iR_i(1) + (1 - D_i)R_i(0)$. As treatment is randomly allocated, it is by construction independent of all potential variables: $D_i \perp\!\!\!\perp (R_i(0), R_i(1), Y_i(0), Y_i(1))$.

Following the IV framework, always respondents are households with $R_i(1) = R_i(0) = 1$. Irrespective of whether they are assigned to the treatment or to the control group, they will always send back the questionnaire. Never respondents are households with $R_i(1) = R_i(0) = 0$. Irrespective of whether they are assigned to the treatment or to the control group, they will never send back the questionnaire. Response compliers are households with $R_i(0) = 0$ and $R_i(1) = 1$: they will send back the questionnaires if and only they are assigned to the treatment group. Finally, response defiers are households with $R_i(0) = 1$ and $R_i(1) = 0$: they will send back the questionnaire if and only if they are assigned to the control group. The main assumption in Lee (2009) is that there are no response defiers: $R_i(1) \geq R_i(0)$.

a) Under the no response defiers assumption, who are households with $\{R_i = 1, D_i = 0\}$: always respondents, response compliers, or never respondents? Subjects with $\{R_i = 1, D_i = 1\}$ include two of the three aforementioned subpopulations. Which are those two subpopulations?

b) Find one reason why the no response defiers assumption might be violated in the micro credit example. Overall, is this still a relatively credible assumption?

c) Show that under the no response defiers assumption, $P(R_i = 1|D_i = 1) - P(R_i = 1|D_i = 0) = P(R_i(0) = 0, R_i(1) = 1)$.

d) Show that under the no response defiers assumption,

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

$$\begin{aligned} E(Y_i|R_i = 1, D_i = 1) &= E(Y_i(1)|R_i(0) = R_i(1) = 1)\frac{P(R_i = 1|D_i = 0)}{P(R_i = 1|D_i = 1)} \\ &+ E(Y_i(1)|R_i(0) = 0, R_i(1) = 1)\left(1 - \frac{P(R_i = 1|D_i = 0)}{P(R_i = 1|D_i = 1)}\right), \end{aligned}$$

and finally conclude that

$$B_-^{L1} \leq E(Y_i(1) - Y_i(0)|R_i(0) = R_i(1) = 1) \leq B_+^{L1},$$

with

$$\begin{aligned} B_-^{L1} &= E(Y_i|R_i = 1, D_i = 1)\frac{P(R_i = 1|D_i = 1)}{P(R_i = 1|D_i = 0)} - \frac{P(R_i = 1|D_i = 1) - P(R_i = 1|D_i = 0)}{P(R_i = 1|D_i = 0)} \\ &- E(Y_i|R_i = 1, D_i = 0), \end{aligned}$$

and

$$B_+^{L1} = E(Y_i|R_i = 1, D_i = 1)\frac{P(R_i = 1|D_i = 1)}{P(R_i = 1|D_i = 0)} - E(Y_i|R_i = 1, D_i = 0).$$

Hint: you need to use the fact that in this exercise, the outcome Y_i is assumed to be binary.

e) What is the length of $[B_-^{L1}, B_+^{L1}]$? Compute the length of that interval when $P(R_i = 1|D_i = 0) = 0.6$ and $P(R_i = 1|D_i = 1) = 0.63$.

f) Find two variables Y_i^{-L} and Y_i^{+L}, such that you can estimate B_-^{L1} and B_+^{L1} trough two regressions of Y_i^{-L} and Y_i^{+L} on D_i among the sample of respondents. Hint: Y_i^{-L} and Y_i^{+L} are functions of Y_i, $(D_j)_{1\leq j\leq n}$, and $(R_j)_{1\leq j\leq n}$: you can use the entire sample of $(D_j)_{1\leq j\leq n}$ and $(R_j)_{1\leq j\leq n}$ but only Y_i to construct Y_i^{-L} and Y_i^{+L}. You do not need to prove that the population coefficient of D_i in these two regressions will indeed be equal B_-^{L1} and B_+^{L1}, you just need to find Y_i^{-L} and Y_i^{+L}.

g) Will the confidence intervals for B_-^{L1} and B_+^{L1} arising from those two regressions be correct?

h) The previous bounds only work with a binary Y_i. In this question and in the next, let's assume that Y_i is continuous. Let $p = \frac{P(R_i=1|D_i=0)}{P(R_i=1|D_i=1)}$, and let $G(y)$ denote the cdf of Y_i conditional on $R_i = 1$ and $D_i = 1$ ($G(y) = F_{Y_i|R_i=1,D_i=1}(y)$). Explain intuitively (no need to prove this) why

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

Hint: you can use the following fact: $G^{-1}(1-p)$ is the value of Y_i for the household at the $1-p^{th}$ percentile of the distribution of Y_i among households with $R_i = 1, D_i = 1$. Similarly, $G^{-1}(p)$ is the value of Y_i for the household at the p^{th} percentile of the distribution of Y_i among households with $R_i = 1, D_i = 1$.

i) Use the previous question to show that

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

with

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

and

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

Conclusion

B_-^{L2} and B_+^{L2} are the bounds proposed in Lee (2009). Those bounds can be used to deal with non-reponse in randomized experiments, under a monotonicity assumption on households' response behavior. More generally, those bounds can be used whenever we have sample selection. Assume you are interested in the effect of a training program on participants' labor market participation, and on their wages. You only observe wages for households who work (in the example above, you only observed the outcome for households answering your questionnaire). If the treatment has on effect on labor market participation, the samples for which you observe wages may not be comparable in your treatment and control groups (in the example above, the treatment had an effect on whether households answered the questionnaire), so comparing the wages of households that work in your treatment and in your control group may not be a valid comparison (in the example above, comparing the outcome for households responding / not responding to the questionnaire may not be a valid comparison). The structure of the problem is exactly the same as that of the non-response problem in the micro-credit example, except that now, R_i is an indicator equal to 1 if person i works, and to 0 otherwise. Thus, you can use the Lee bounds to get bounds on the average effect of the treatment on wages, among households that will always work irrespective of whether they get the training or not.