Econ 488, Second problem set

2025-01-17

- **Due Date:** 11:59pm on January 23. Submission should be via gradescope (accessible via canvas).
- Other instructions:
  - You may work on this problem set in groups of up to 3 students. One submission per group (via Gradescope), and please indicate your group's members.
  - You are welcome to practice typewriting the solutions in LaTeX or Rmarkdown to see which option you like best, or whether typewriting is a good choice for you. I have made templates to get you started (See the markdown template file).
  - Unless otherwise stated, please justify your answer. Also, try to answer concisely: please limit each response to at most two sentences.

Questions:

Question A: homogeneous treatment effects model

Suppose that for Walmart customer i, D_i is a binary variable indicating 'Sam's Club Plus' membership, and Y_i indicates spending at Walmart in the following 12 months.[1] Walmart is interested in the causal effect of Sam's Club Plus membership on customer spending at Walmart.

1. Describe in words and mathematically (using potential outcomes), the unit-level treatment effect (ITE_i) in this example.

2. Show that Y_i can be expressed as
$$Y_i = \alpha + ITE_i D_i + U_i.$$
What are α and U_i?[2] What is $E[U_i]$? Provide some examples of other determinants that may be part of U_i.

   A very common assumption is the **homogeneous treatment effects** assumption, which says that the unit-level treatment effect is the same for all units. Mathematically, for all i, $ITE_i = \beta$ for some β (which does not depend on i). Given part 2, it follows that we have the linear model
$$Y_i = \alpha + \beta D_i + U_i.$$
From previous classes, we know that the least squares regression estimator for β is
$$\hat{\beta} = \frac{\widehat{Cov}(D_i, Y_i)}{\widehat{Var}(D_i)},$$
where $\widehat{Cov}$ and $\widehat{Var}$ indicate sample covariance and variance, respectively. Suppose that one could show that
$$E[\hat{\beta}] = \frac{Cov(D_i, Y_i)}{Var(D_i)},$$
where Cov and Var are population covariance and variance, respectively.

[1] Membership in the Sam's Club Plus program implies, among other things, that customers earn cash rewards (e.g. they get $10 back for every $500 spent on qualifying purchases), enjoy free-shipping on many 13 items, and reduced 2-day shipping charges. Sam's Club Plus charges an annual fee of $100. Shoppers may use brick-and-mortar Walmart stores, or shop online at Walmart.com

[2] Notice that α does not depend on i, but the other variables may do so.

3. Compute the bias of $\hat{\beta}$ for β.[3] Provide a condition under which the bias is zero.

   In fact, since D_i is binary, the least squares estimator takes a simpler expression:

$$\hat{\beta} = \widehat{E}(Y_i|D_i = 1) - \widehat{E}(Y_i|D_i = 0),$$

   and thus

$$E[\hat{\beta}] = E(Y_i|D_i = 1) - E(Y_i|D_i = 0),$$

4. Recompute the bias of $\hat{\beta}$ for β using these new expressions, and show that it gives the same answer as the previous question. Provide a (different) condition under which the bias is zero.[4]

Question B: Regression model for the ATU

1. Continuing with the previous definition of Y_i and D_i, explain how you can choose α and U_i so that, for all i,[5]

$$Y_i = \alpha + ATU \times D_i + U_i.$$

2. What is an "exogeneity" condition for this regression model?

Question C: Conditional random assignment

Continuing with the Walmart example, suppose that Walmart's data science team designed an experiment in which D_i was randomly assigned conditional upon the length of time that i has been a Walmart customer, which is encoded as X_i. For example, it may be that they wish to provide more Plus memberships to longer-term customers. Technically, assume that the conditional random assignment (CRA) assumption holds:

$$(Y(0), Y(1)) \perp D \mid X$$

1. Based on our understanding of randomly assigned treatment meaning that D_i is (conditionally) independent of all pre-determined variables, which of the following variables do you expect to be/not be uncorrelated with treatment (within each X_i group):

   - Number of visits taken to Walmart in the previous 12 months.
   - Total spending on Amazon.com in the next 12 months.
   - Number of children in the customer's household.

2. Explain how you could use the variables (if any) that you expect to be (conditionally) uncorrelated with treatment to check whether the CRA assumption holds. If your check passes, can we conclude that the CRA assumption is true? What if the check fails?

   Recall that the conditional average treatment effect is defined as

$$CATE(x) = E[Y(1) - Y(0) \mid X = x]$$

3. If for a given customer tenure level x, $0 < \Pr(D = 1 \mid X = x) < 1$, show that the $CATE(x)$ is identified.

4. Explain which part of your argument fails if, for the given tenure level x, $\Pr(D = 1 \mid X = x) = 0$ or $\Pr(D = 1 \mid X = x) = 1$

5. Suppose that for every tenure level, x, $0 < \Pr(D = 1 \mid X = x) < 1$, build on your answer to part 3 to show that the ATE is identified.

[3] Recall the following definition of "bias". Let $\hat{\theta}_n$ be an estimator of θ. Its bias is defined as: $\text{Bias}_\theta\left(\hat{\theta}_n\right) = E\left(\hat{\theta}_n\right) - \theta$. We say the estimator is unbiased if its bias equals 0 (for all values of θ).

[4] The conditions that enable regression to be unbiased are often called "exogeneity".

[5] Recall the ATU is defined as $E[Y(1) - Y(0)|D = 0]$

Question D: Empirical question

Consider the data file "SCP_data.csv", which contains the following variables:

- `scp`: Sam's Club Plus membership (=1 if yes)
- `spend`: annual spending at Walmart in the 12 months following the `scp` membership status
- `tenure`: the (rounded down) number of years that each customer has been frequenting Walmart

There are two additional variables: `spend0` and `spend1`. These are the potential outcomes corresponding to each level of `scp`, which I could include as this is an artificial dataset I created. Of course, in a real dataset the counterfactual outcome would be missing for each unit.

We are interested in learning the ATE of `scp` on `spend`

1. First suppose the following CRA assumption

$$(spend(0), spend(1)) \perp scp \mid tenure$$

   Explain why, under this assumption, we cannot learn the ATE. (Hint: plot the `tenure` variable.) Might your answer change if the sample size was bigger?

2. Instead consider a different CRA assumption:

$$(spend(0), spend(1)) \perp scp \mid tenureD \tag{1}$$

   for

$$tenureD = \begin{cases} 1 \text{ if tenure } \leq 1 \\ 3 \text{ if tenure } > 5 \\ 2 \text{ otherwise} \end{cases}$$

   Generate this `tenureD` variable ('D' stands for 'Discrete'), and explain why we are able to learn the ATE under this assumption.

3. Under the assumption in equation (1), estimate the conditional (on $tenureD$) average treatment effects using one linear regression[6]. (Of course, use only data that you would have available to you in the real dataset: `scp`, `spend` and `tenureD`—do not use the counterfactual outcomes!)

4. Consider two estimators for the average treatment effect:

   - $\widehat{E}[\widehat{CATE}(tenureD)]$, where $\widehat{CATE}(tenureD)$ is your estimator from the previous question
   - $\widehat{E}[spend|scp = 1] - \hat{E}[spend|scp = 0]$

   Which (if any) of these estimators do you prefer? Why? Compute both of these estimators.

5. Since you have the counterfactual outcomes available to you, you can estimate the ATE directly. Compute this (usually infeasible) estimator and compare it to the values of the estimators you computed in the previous question. Is this consistent with what you would expect?

[6]You may use the 'lm()' function in R. Keep in mind that tenureD is a categorical variable.