ECMA 31360, Pset 2: Using OLS to Estimate Treatment Effects (100p)

Melissa Tartari, University of Chicago

Objective of the PSet: In PSet1 you considered the OLS estimator in a prediction context: you had a linear-in-parameters CEF, $E[y_i|D_i] = \alpha + \beta D_i$ with D_i being a 0/1 variable, and showed that the OLS estimator of β is unbiased for $E[y_i|D_i = 1] - E[y_i|D_i = 0]$. Here you consider the OLS estimator in a causal context: you start with the outcome equation[1] $y_i = \alpha + \beta D_i + u_i$, where D_i and u_i are determinants of y_i (observed and unobserved respectively), and study the circumstances under which the OLS estimator of β enables inference about the causal effect of D_i on y_i. Comparing the two contexts yields a lot of learning. To answer PSet2's questions you do not need the course material unless explicitly stated because Pset2 you use the traditional (i.e., pre Rubin Causal Model) approach to causal analysis (from introductory econometrics courses).

Background for the PSet: Walmart Inc. is an American retail corporation that operates a chain of hypermarkets. Walmart introduced *Sam's Club Plus* (link) in February 2018. Membership in *Sam's Club Plus* implies that customers earn cash rewards (e.g., they get $10 back for every $500 spent on qualifying purchases), enjoy free-shipping on many 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. The questions below focus on the causal impact of *Sam's Club Plus* membership on online spend.[2]

[1] Think of this equation as stemming from some theoretical model which we have left unspecified.

[2] All references to Walmart.com and its customers in this problem set are entirely fictitious and are used solely for the purpose of illustrating statistical concepts and their application in various contexts.

Q1. (50p) Let y_i be customer i's spend at Walmart.com in a given month. Let $D_i = 1$ if customer i is a *Sam's Club Plus* member, $= 0$ otherwise. Assume membership status does not vary during the month. y_i is determined by the customer's membership status (D_i) and other determinants (u_i) according to the homogeneous treatment effects model:

$$y_i = \alpha + \rho D_i + u_i. \tag{1}$$

For example, u_i may include household income and size. As (y_i, D_i, u_i) vary across customers, we think of them as RVs. Let $E[u_i] = 0$, where the expectation is taken with respect to the distribution of u_i in the population of Walmart.com customers. You have data on a sample of size n of customers: $\{(y_i, D_i) | i = 1, \ldots, n\}$. Note that u_i is not included in the data for any of the sample customers.[3] Some of the sample customers are *Sam's Club Plus* members, some are not. As your data contains a mix of both types of customers we say that you have "observational variation in the cause or treatment". (α, ρ) are unknown parameters. Let $\overline{y}^0$ (respectively, $\overline{y}^1$) denote the sample average of y_i across sample customers with $D_i = 0$ (respectively, with $D_i = 1$).

(a) (2p) Provide additional examples of determinants of spend that may be part of u_i. **Hint**: In the background there is a consumer demand model, i.e., you think of model (1) as a consumer expenditure function from microeconomics.

Solution: One determinant of spend that may be part of u_i is the customer's age. Younger customers may spend more than older customers, for example. Another determinant of spend that may be part of u_i is the customer's location - customers in urban areas may spend more than customers in rural areas due to the convenience of online shopping, shipping times/rates, etc.

(b) (2p) Show that ρ is the causal impact of *Sam's Club Plus* membership on a customer's spend, in the sense that, ρ is the difference in a customer's spend with and without membership holding all else the same.

Solution: Given the model above, we can analyze the two cases for D_i.
When $D_i = 1$, the equation becomes:

$$y_i = \alpha + \rho + u_i$$

And when $D_i = 0$, the equation becomes:

$$y_i = \alpha + u_i$$

Now, let's consider the difference between the spend of a customer who is a Sam's Club Plus member ($D_i = 1$) and one who is not ($D_i = 0$). Holding all else the same, we can take the difference of the two above equations:

$$\overline{y}^1 - \overline{y}^0 = (\alpha + \rho + u_i) - (\alpha + u_i) = \rho$$

This demonstrates that ρ is the difference in a customer's spend with and without Sam's Club Plus membership, while holding all other factors constant - the unobserved term u_i cancels out in the difference. Thus, ρ in this model represents the causal impact of a Sam's Club Plus membership on a customer's spend at Walmart.com.

[3] That is, you observe the customer's spend and their membership status but not their household income and size, nor any of the other determinants of how much they spend on Walmart.com. This is the reason why we use the letter u, it is mnemonic for *unobserved*.

(c) (2p) Is $E[u_i] = 0$ an assumption or a normalization? Show it.[4]

Solution: $E[u_i] = 0$ is a normalization. Supposed $E[u_i] \neq 0$. Then we can reparametrize as follows:

$$y_i = \alpha + \rho D_i + u_i - E[u_i] + E[u_i]$$
$$y_i = (\alpha + E[u_i]) + \rho D_i + (u_i - E[u_i])$$
$$y_i = \tilde{\alpha} + \rho D_i + \tilde{u}_i$$

and we have arrived at the same model as before, but with a different unobserved term.

In general, if we find that for a given α and ρ, $E[u_i] \neq 0$, we can absorb the non-zero mean of the error term into the constant term α, thus allowing us to correctly assume that $E[\tilde{u}_i] = 0$. Thus, $E[u_i] = 0$ is a normalization.

(d) (2p) Let $(\widehat{\alpha}, \widehat{\rho})$ denote the Ordinary Least Squares (henceforth OLS) estimator of parameters (α, ρ) in model (1). Do you need to make any assumption on u_i to compute $(\widehat{\alpha}, \widehat{\rho})$ in a particular sample?

Solution: To compute the Ordinary Least Squares (OLS) estimators $(\widehat{\alpha}, \widehat{\rho})$ for the parameters (α, ρ) in the model $y_i = \alpha + \rho D_i + u_i$, you don't need to make any specific assumptions about the distribution or properties of the error term u_i.
The OLS estimators for α and ρ are obtained by minimizing the sum of squared differences between the observed values of y_i and the values predicted by the model for a given sample:

$$\widehat{\alpha}, \widehat{\rho} = \arg\min_{\alpha,\rho} \sum_{i=1}^{n} (y_i - \alpha - \rho D_i)^2$$

This estimation method does not require any assumptions specifically about the distribution or properties of u_i.

However, while OLS doesn't require explicit assumptions about u_i for estimation, certain assumptions under the classical linear regression are needed to establish the properties of the estimators (e.g., unbiasedness, consistency, efficiency). These assumptions include things like the error term having a mean of zero conditional on the predictors, homoscedasticity (constant variance of errors), and no correlation between the errors and the independent variables. Violations of these assumptions can affect the reliability of the OLS estimators in terms of their statistical properties.

(e) (10p) Verify expression (2). **Hint**: Leverage the derivations you did for PSet1, do not use linear algebra. **Note**: We leverage this result repeatedly, make sure to understand it both from an intuitive standpoint and algebraically.

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\rho} \end{bmatrix} = \begin{bmatrix} \overline{y}^0 \\ \overline{y}^1 - \overline{y}^0 \end{bmatrix}. \tag{2}$$

(f) (2p) Use expression (2) to describe in plain English estimator $(\widehat{\alpha}, \widehat{\rho})$.

(g) (2p) **Without further assumptions**: Are $(\widehat{\alpha}, \widehat{\rho})$ unbiased/consistent estimators of (α, ρ)? Explain (no proof).

(h) (4p) You have a random sample (RS) of Walmart.com customers. In econometrics the assumption that $E[u_i|D_i = 1] = E[u_i|D_i = 0]$ is called the "zero conditional mean assumption" (ZCMA) because, once we make the normalization $E[u_i] = 0$, it writes as $E[u_i|D_i = 1] = E[u_i|D_i = 0] = E[u_i] = 0$. Describe the ZCMA in plain English.

[4] An assumption imposes a restriction on the objects/items present in your model, the restriction may or may not hold. For example, if a claim is stated subject to an assumption then your proof will use the assumption to arrive at the result, which means that the result may not obtain had you dropped the assumption. A normalization is when you recognize that two (or more) objects/items in a model are not separately identified, i.e., there is no way to learn about each of them separately, e.g., you may only learn their sum or product, or some other function of the two (or more) objects. If you recognize such a situation in your model you reparametrize the model so that the objects in the reformulated model are learnable.

(i) (8p) Show that ρ is identified if ZCMA holds. **Hint**: Express ρ *exclusively* as a function of population data moments (PDM), that is, features of the population distribution of (y_i, D_i).

(j) (12p) Assume that ZCMA holds. Is estimator $\widehat{\rho}$ unbiased (6p)? Is it consistent (6p)? Prove it. **Hint**: Let $D = \{D_1, \ldots, D_n\}$. Show $E[\widehat{\rho}] = \rho \; \forall \rho$ and $\widehat{\rho} \xrightarrow{p} \rho \; \forall \rho$ starting with expression (2).

(k) (2p) In place of maintaining ZCMA, assume that $u_i \perp D_i$, where the symbol "$\perp$" signifies statistical independence. Does your answer to **Q1j** change? If it does, how?

(l) (2p) In light of your answers to the previous questions: Do you expect estimator $\widehat{\rho}$ to be unbiased/consistent when constructed using a sample of actual Walmart.com customers chosen at random? Explain.

Q2. (20p) Consider the time **before** Walmart introduced *Sam's Club Plus*. Sales leadership have come up with the idea of a *Sam's Club Plus* membership that offers cash back on all orders. Scientists want to design and carry out a randomized control trial (RCT) to estimate how different consumer spend would be on average with a *Sam's Club Plus* membership. The estimate would help stakeholders decide whether to roll-out a *Sam's Club Plus* program, and how to price it.

(a) (2p) Suggest two reasons why a customer's spend at Walmart.com may differ with versus without a *Sam's Club Plus* membership, all else the same.

(b) (12p) The Walmart scientists carried out an RCT: they randomly assigned (RA) *Sam's Club Plus* membership status to $10,000$ existing customers (at no charge) (treated group) and left the rest of the customers *as is* i.e., without the membership (control group). Provide a discussion of this RCT's possible limitations/challenges by following step by step the list of limitations given in `CAUS_intro.pdf`, that is, implementation hurdles, lack of generalizability, etc.

(c) (6p) The RCT is carried out. The post-experiment analysis sample includes the $10,000$ customers in the treatment group, and $10,000$ customers chosen randomly from the control group. For each sample customer, the data only records the customer's membership status and how much they spent at Walmart.com during the first month following the date of RA, i.e., $\{(y_i, D_i) \,|i = 1, \ldots, n = 20,000\}$. Can the scientists use the experimental variation in this data to estimate ρ in model (1) by OLS? Explain. How shall they interpret the resulting OLS estimate?

Q3. (20p) The setting is as in **Q1** but for the following: y_i is determined by a customer's *Sam's Club Plus*-membership status (D_i) and other unobserved determinants (v_i) according to the heterogeneous treatment effects model:

$$y_i = \alpha + \rho_i D_i + v_i. \tag{3}$$

Note that ρ_i varies across customers. You have access to a RS of existing Walmart customers. As in **Q1**, the data is the collection $\{(y_i, D_i) \,|i = 1, \ldots, n\}$. Note that ρ_i is **not** included in your data, nor is v_i.

(a) (2p) Interpret ρ_i.

(b) (4p) Think of each ρ_i as a draw from a distribution. Let $\rho \equiv E[\rho_i]$, $\rho_1 \equiv E[\rho_i | D_i = 1]$, and $\rho_0 = E[\rho_i | D_i = 0]$. Describe in plain English these three objects. Interpret in plain English the assumption $\rho = \rho_1 = \rho_0$. Speculate about why this assumption is called "no selection on gains" (specialize your answer to the situation being considered).

(c) (2p) Verify that you can rewrite model (3) as a simple linear regression model:

$$y_i = \alpha + \rho_1 D_i + u_i \text{ with } u_i \equiv v_i + (\rho_i - E[\rho_i | D_i = 1]) D_i. \tag{4}$$

(d) (6p) Consider the OLS estimator of the slope parameter in model (4). In **Q1** you established that $\widehat{\rho_1} = \overline{y}^1 - \overline{y}^0$ and $\widehat{\rho_1}$ is unbiased for ρ_1 under the ZCMA $E[u_i | D_i = 1] = E[u_i | D_i = 0]$.

i. What are the substantive benefits of this result? That is, what do you learn about the causal effect of the treatment?

ii. What does the ZCMA $E[u_i | D_i = 1] = E[u_i | D_i = 0]$ imply for the relationship between the unobserved (v_i) and observed (D_i) determinants of the outcome?

(e) (6p) Under which additional condition does $\widehat{\rho}_1$ allow us to infer the average causal effect of treatment for the entire population, rather than only for the sub-population of *Sam's Club Plus* members?

Q4. (10p) Take stock. You worked with two models: (1) $y_i = \alpha + \rho D_i + u_i$, see expression (1); and (2) $y_i = \alpha + \rho_i D_i + v_i$, see expression (3). You considered one estimator: the OLS estimator of the intercept and slope coefficients in a linear regression of customer spend on a constant term and an indicator of *Sam's Club Plus* membership status. You derived the assumptions that suffice for the OLS estimator of the slope coefficient to be unbiased for ρ in model (1), and to be unbiased for the mean of ρ_i in the sub-population of *Sam's Club Plus* members or in the entire population in model (3). What did you learn about interpreting the OLS estimator of the slope coefficient in a causal context? Write 3 to 5 sentences.