ECMA 31360, Pset 1: Review of OLS for Prediction/Description and for Causal Analysis

Melissa Tartari, University of Chicago

Premise: How to consume PSets and their Pedagogy

This is the first PSet you encounter in this course. All PSets have a few things in common:

- PSets start with a **reference** to the location of two files: (a) a .TEX file with the PSet's source code, and (b) a .Rmd file serving as a template for your answers. We provide the .TEX file because, at times, students want to see the LaTeX commands used to type mathematical expressions. The .Rmd file is what you will want to open in RStudio and populate with your answers. This file lists all questions in order, reports the points attached to each question, and indicates when R code is required in the answer. It also makes grading easier because submissions conform to the same format. Initially, it may be easier to write and execute your script in a separate .R file, then copy its content into code chunks within the .Rmd file. Once you are comfortable with writing and executing commands in a .Rmd file, that's all you'll need to use.
- PSets have several **parts**. Some parts test you on the theory developed in class, others have you implement operations using R, and some combine the two tasks. Some questions will expand on what we cover in class. We hope that you'll find this rewarding, as it demonstrates your ability to use what you've learned to obtain a richer set of results.
- Each part of a PSet has **learning objectives**. You'll want to read these objectives both before and after answering the questions. When reviewing them later, check whether you have attained those objectives.
- The **order** in which questions are asked typically matters, as one question often builds on the answer or the concepts learned in previous questions. Therefore, it is important to answer the questions in the order they are presented.
- We color technical terms, or **jargon**, green. Jargon is both powerful and risky. It is powerful because it conveys a lot of information in just one or two words, but it's risky because using it incorrectly leads to confusion. By highlighting jargon, we underscore that each term has a precise definition that you need to understand to answer a question correctly.
- PSets typically include questions that ask you to closely examine a **formal definition** or the **statement of a theorem**. Often, we ask you to interpret assumptions and describe them in plain English. At times, we may ask you to prove one or more theorems. Algebra and basic calculus suffice; that is, the math involved is not overly demanding. You may or may not be accustomed to paying such close attention to formal definitions and theorem statements. Theorems are like culinary recipes: the assumptions are the ingredients, and the statement is the procedure. To get the dish cooked properly, you need to carefully gather all the ingredients in the right amounts and combine them in the right way. You also want to ensure that what you get is what you want. Similarly, definitions are essential for understanding what you're working with. If a definition isn't clear, it's like using a can of food in a recipe without knowing what's inside. Learning how to break down complex ideas into their components, follow logical arguments, and communicate technical concepts clearly is valuable in any career.
- We often ask you to **explain something in plain English**. At times, we specify your audience (e.g., a fellow classmate, a layperson on the bus, or a business stakeholder). This exercise encourages you to (a) step back from the details of the answer and its derivations to distill the core message or learning, and (b) practice communicating technical concepts clearly to a non-technical audience—a highly valuable skill.
- We provide **guidance** in two ways: (a) hints, and (b) "programming guidance." Pay attention to both. We use programming guidance as a teaching tool to expand your R knowledge and proficiency. Typically hints and guidance are placed at the end of the PSet in a separate section because it encourages you to think about the question for a while before consulting the guidance. How far can you get in answering the question without looking at the guidance?
- Each question carries one or more **points**. We do not assign half points.
- **Heads up**: PSets have a long "shelf life" because (a) future PSets build on previous PSets, and (b) midterm and final exams test you on the material you learn both in class and by solving PSets. Thus, to do well in exams we recommend that you make sure to solidify the learnings from each PSet and promptly and regularly look at the suggested solutions.

References: You can access the .TEX version of this file at `Canvas/Files/PSets/PSet1/PSet1-clean.tex` and the .Rmd version of the answer template file at `Canvas/Files/PSets/PSet1/PSet1_Solutions_Template.Rmd`.

Part 1: Review of OLS for Prediction and Description (40 points)

Learning Objectives: You review the OLS estimator in a prediction context: you have a linear-in-parameters Conditional Expectation Function (CEF), $\mathbb{E}[Y_i \mid X_i] = \beta_0 + \beta_1 X_i$, and show that the OLS estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is unbiased and consistent. This part of the PSet leverages knowledge you acquired in previous courses.

Q1. (22p) Prove **Claim 1**.

Claim 1 (Properties of the OLS Estimator when the CEF is linear-in-parameters) *Consider an IID population where (X_i, Y_i) are such that $Var[X_i] > 0$ and the CEF has the form $E[Y_i|X_i] = \beta_0 + \beta_1 X_i$, i.e., it is linear-in-parameters. Let $\{(Y_i, X_i)|i = 1, \ldots, n\}$ be a sample of size n drawn randomly from the population. Let $\{(y_i, x_i)|i = 1, \ldots, n\}$ denote a particular realization of the random sample, and assume that x_i is not constant. Regress y_i on a constant and x_i and denote the OLS values of the intercept and slope coefficient by $(\hat{\beta}_0, \hat{\beta}_1)$. As a function of the random sample, $(\hat{\beta}_0, \hat{\beta}_1)$ are unbiased and consistent estimators of the CEF's parameters.*

Q2. (2p) Prove **Claim 2**.

Claim 2 (Linearity-in-parameters of the CEF for binary variables) *Let Y and X be two RVs such that X takes values in $\{0, 1\}$. Then, there exist two scalars (β_0, β_1) such that $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$, i.e., such that the CEF is linear-in-parameters.*

Q3. (4p) You (let's say your name is Ty) are employed as a data scientist by Amazon. Your team has access to a random sample of 1,000,000 Amazon customers, some are *Prime* members, some are not. (Amazon Prime is a subscription plan which, among other things, entitles a customer to free shipping on many items sold on Amazon.com.) The metrics included in the data are a customer's total spend for the month of December 2024, denoted Y_i, and their *Prime* status, denoted D_i with $D_i = 1$ if customer i is a *Prime* subscriber, and 0 otherwise. Your stakeholder is interested in quantifying the **causal impact of *Prime* membership on customer spend**, i.e., the impact of *Prime* membership holding all other determinants of spend fixed. They regard this piece of information as central to informing an outstanding decision: whether to increase *Prime* advertisement expenditures. Alyson, your manager and previously a data scientist, sends you the email below. You ponder this request and decide to push back. Compose a response to your manager including the justification for your position.

Email from Manager: Alyson Booth

Subject: Advertising Spend Analysis

Hi Ty,
Claim 2 says that the CEF is always linear-in-parameters when X_i is binary. **Claim 1** says that when the CEF is linear-in-parameters, the OLS procedure yields good estimators (i.e., unbiased and consistent) of the CEF's parameters under the minimal added condition that X_i varies in the (population and) sample. Therefore, when X_i is binary, the OLS estimators of the slope coefficient is an unbiased and consistent estimator of the causal effect of X_i on the outcome variable Y_i. That's great for us: all you have to do is to tap into our historical customer-level data, regress customer spend on a constant and the *Prime* indicator, and use the OLS estimate of the slope coefficient to answer our stakeholder's ask. Please index on bias for action and email me the answer weaved into a business-friendly paragraph within the next 15 minutes.

Thanks, Alyson
It is always Day 1!

Q4. (4p) Alyson is less than thrilled about your push-back: they want to provide an answer to their stakeholder ASAP. They jump onto Slack and fire you a short message: "Ty: Our answer does not need to be perfect, it suffices that it is directionally correct. We must deliver results!" They also quickly reach out to Nashant, the business analyst in your team, and ask them to compute the difference in average spend between *Prime* and non-Prime members, which turns out to be \$52 per month per customer. They email you back, see below. Compose a response to your manager including the justification for your position.

Email from Manager: Alyson Booth

Subject: Advertising Spend Analysis

Hi Ty,
I appreciate your insisting on the highest standards and questioning the validity of the regression-based approach. But I am confident that you can stand behind a model-free answer that simply states a fact, namely that *Prime* status increases spend on average by $52 per month. The data truly does speak for itself! I need you to approve this answer and use it in your email to the stakeholder (put me in Cc).

Thanks, Alyson
It is always Day 1!

Q5. (2p) Prove **Claim 3**.

Claim 3 (Properties of a linear-in-parameters CEF) *Let Y and X be two RVs. Denote by $\mathcal{X}$ the support of the distribution of X. If $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$ for same scalar (β_0, β_1) (i.e., if the CEF is linear-in-parameters), then there exists a RV ε such that $Y = \beta_0 + \beta_1 X + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X = x] = \mathbb{E}[\varepsilon] = 0 \ \forall x \in \mathcal{X}$ (i.e., ε is mean-independent of X). The converse is also true: if $Y = \beta_0 + \beta_1 X + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X = x] = E[\varepsilon] = 0 \ \forall x \in \mathcal{X}$ then $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$.*

Q6. (6p) The issue is escalated to Jay, the Sr. product manager in your team. They review the email exchanges so far and write back to your and Alyson, see below. Compose a response to Jay. Note: Your numerical example (4p) should be as simple as possible and have actual numbers, so that a business audience may understand the issues that you want to explain. You should also have a more technical section (which is optional for the business audience) addressed to a data science audience, containing the technical detail that complement your numerical example (2p).

Email from Sr. Product Manager: Jay Tran

Subject: Advertising Spend Analysis

Hi Ty,
+ Alyson for visibility

- I took econometrics in college. That was a few years back but I remember that sample size matters. In our case the sample size is so large that we should trust that Nashant's estimate is close to the truth. Should we not?
- A simple and fully worked-out example that illustrates — using actual numbers — your concerns would go a long way. Can you write that up?

Thanks, Jay
We delight our customers!

Part 2: Review of OLS for Causal Inference (46 points)

Notation: Starting with this part of the PSet we no longer use upper case and small case letters to distinguish between random variables and realizations of random variables. The context determines what is what.

Learning Objectives: You review the OLS estimator in a causal context. Comparing this context to that considered in Part 1 yields a lot of learning. You do not need the course material unless explicitly stated because you use the traditional (i.e., pre Rubin Causal Model) approach to causal analysis (from econometrics courses).

Q7. (15p) Walmart Inc. is an American retail corporation that operates a chain of hypermarkets. Walmart introduced *Sam's Club Plus* in February 2018. Membership in *Sam's Club Plus* earns customers cash rewards (e.g., they get $10 back for every $500 spent on qualifying purchases), 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. 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, and zero 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:

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

Note that ρ is the same for all customers, hence the name of the model. 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: u_i is not included in the data for any i.[1] 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) (1p) Provide two additional examples of determinants of spend that may be part of u_i.

(b) (1p) 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 other determinants the same.

(c) (1p) Is $E[u_i] = 0$ an assumption or a normalization? Show it.

(d) (1p) 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?

(e) (2p) You know that expression (2) gives the closed-form of the OLS estimators $(\widehat{\alpha}, \widehat{\rho})$. Use these expressions to describe $(\widehat{\alpha}, \widehat{\rho})$ in plain English.

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

(f) (1p) Without further assumptions: Are $(\widehat{\alpha}, \widehat{\rho})$ unbiased/consistent estimators of (α, ρ)? Explain (no proof).

(g) (1p) 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.

(h) (1p) Show that ρ is identified if ZCMA holds.

(i) (1p) Can you weaken ZCMA and still achieve identification of ρ? Yes/no and justify.

(j) (2p) Assume that ZCMA holds. Is estimator $\widehat{\rho}$ unbiased? Is it consistent?

(k) (1p) Assume that $u_i \perp D_i$ in place of ZCMA. Does your answer to **Q7j** change? If it does change, 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.

Q8. (13p) 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) (9p) 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). We mentioned in class that RCTs are subject to imitations/challenges, see `CAUS_intro.pdf`. With reference to the Walmart RCT, describe situations that what would produce substitution bias and, separately, contamination bias.

(c) (2p) 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?

Q9. (14p) The setting is as in **Q7** 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, hence the name of the model. You have access to a RS of existing Walmart customers. As in **Q7**, 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.

[1] 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*.

(a) (2p) Interpret ρ_i.

(b) (5p) 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) (1p) Verify that you can rewrite model (3) as follows:

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

(d) (4p) Consider the OLS estimator of the slope parameter in model (4). In **Q7** 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. (2p) What are the substantive benefits of this result? That is, what do you learn about the causal effect of the treatment?

ii. (2p) 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?

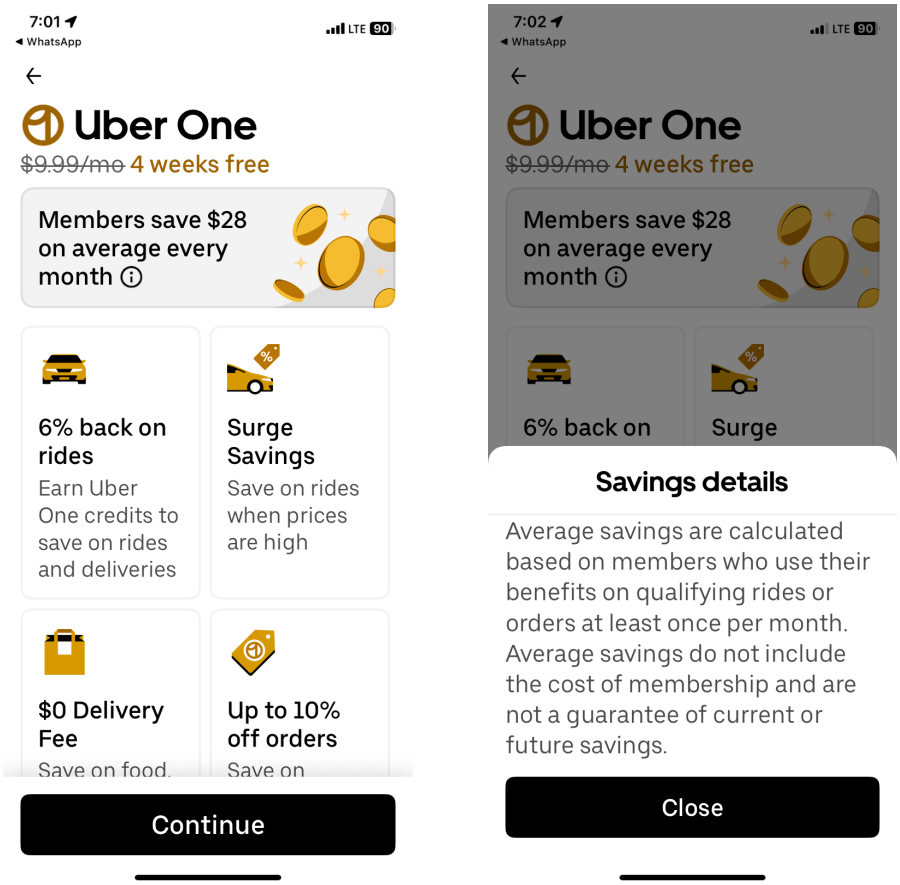
(e) (2p) Under which additional condition(s) if any does $\widehat{\rho_1}$ allow us to learn the average causal effect of treatment for the entire population?

Q10. (4p) Take stock. You've 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've considered one estimator: the OLS estimator of the intercept and slope coefficients in a linear regression of customer spend (y_i) on a constant term and the indicator of *Sam's Club Plus* membership status (D_i). You've spelled out the assumptions that suffice for the OLS estimator of the slope coefficient to be unbiased for ρ in model (1), and 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 a paragraph (3 to 4 sentences).

Part 3: Causality vs. Description — The Case of Uber One (10 bonus points)

Learning objective: Practice distinguishing answers to causal questions from answers to descriptive questions.

Q11. (10p) In 2021 Uber launched *Uber One*, a single membership that bundles benefits for Uber rides and Uber Eats. The program advertises, among other things, discounts and $0 delivery fees on eligible orders. The other day I received a prompt in the Uber app about this program. The **left** screenshot is the main promo screen. Tapping the information icon (i) next to "Members save $28 on average every month" opens the **right** screenshot, a pop-up titled "Savings details."



(a) (5p) Treat the claim "Members save $28 on average every month" as an *answer looking for a question*. What is the question that produces this answer? Is that question descriptive or causal? Briefly justify.

(b) (3p) Based on the "Savings details" pop-up (the **right** figure), write pseudocode (high-level steps) for how Uber likely computes the "$28 per month per member" figure.

(c) (2p) Now think as a rider deciding whether to join Uber One. What question do *you* want answered? Is it descriptive or causal? Explain in one or two sentences.

Part 4: A First Look at the Data from the NSW Experiment (14 points)

Learning Objectives: Starting with this PSet you use experimental data with the ultimate objective of estimating the effect of the offer of employment *cum* training on earnings. You are asked to mimic the steps of empirical analysis: 1) describe control and treated groups (this PSet); 2) test balance in predetermined characteristics to ascertain whether randomization was carried out successfully; 3) estimate impact using the difference in means estimator and regression adjusted difference in means estimator; 4) understand the conceptual difference b/w the impact of the offer of training and that of undergoing training.

Q12. (14p) The National Supported Work (NSW) demonstration project was a transitional, subsidized work experience program for people with long-standing employment problems. NSW was implemented in 1976-77 as a Randomized Control Trial (RCT). Eligible applicants were randomly assigned by the "flip of a coin" either to treatment or control. Treated individuals were offered participation in the NSW employment *cum* training program for a period of 6 to 18 months, controls were not. Those randomly selected to join the program participated in various type of work, such as restaurant and construction work. Information on pre-intervention characteristics was obtained from initial surveys and the Social Security Administration records. Both the treated and the control groups participated in follow-up surveys at specific intervals. The outcome we focus on are earnings measured in 1978, i.e., about one year post-intervention. You work with Dehejia and Wahba (1999, 2002)'s extract of the NSW original data.[2] In this extract, the treated sample consists of 185 males, and the control sample of 260 males. Here are the questions:

(a) (1p) What is the treatment in this experiment?

(b) (1p) Load and combine the NSW data from `nswre74_control.csv` and `nswre74_treated.csv` into one dataframe.

(c) (1p) Verify the counts of treated and control units.

(d) (4p) Complete Table 1, i.e., fill the blank spaces in columns numbered (4) and (5). Note: Variables whose counter is 1 through 10 are called pre-determined variables, i.e., they capture characteristics determined at or before treatment assignment; some of these variables are background characteristics (e.g., `edu`), others capture a subject's pre-RCT labor market experience (e.g., `u75`). `re78` is the observed outcome variable. `treat` is the indicator of treatment status.

Variable Counter	Variable Name	Variable Definition	Sample Average	
			Treated	Control
(1)	(2)	(3)	(4)	(5)
1	`age`	Age (in years)		
2	`edu`	Education (in years)		
3	`nodegree`	=1 if education < 12, zero otherwise		
4	`black`	=1 if Black, zero otherwise		
5	`hisp`	=1 if Hispanic, zero otherwise		
6	`married`	=1 if married, zero otherwise		
7	`u74`	=1 if unemployed in '74, zero otherwise		
8	`u75`	=1 if unemployed in '75, zero otherwise		
9	`re74`	Real earnings in '74 (in '82 $)		
10	`re75`	Real earnings in '75 (in '82 $)		
11	`re78`	Real earnings in '78 (in '82 $)		
12	`treat`	=1 if received offer of employment *cum* training, zero otherwise	1	0
Sample Size			185	260

Table 1: Sample Averages of 10 (ten) pre-determined Variables and the Outcome Variable in the NSW Data, by Group (**Treated** includes individuals assigned to treatment and **Control** includes individuals assigned not to be treated).

[2] Dehejia and Wahba (1999) Causal Effects in Nonexperimental Studies: reevaluating the Evaluation of Training Programs, *JASA*, pp. 1053-1062 and Dehejia and Wahba (2002) Propensity-score Matching Methods for Nonexperimental Causal Studies, *ReStat*, pp. 151-161. The original data (not used in the pset) is available at the ICPSR page.

(e) (5p) Proper random assignment of treatment *balances* all the subjects' characteristics, including all determinants of the outcome (but for treatment status), both observed and unobserved. Thus, an implication of proper randomization is that there should be no systematic differences (i.e., no "imbalance") between control and treatment groups in terms of their observed predetermined variables (OPVs).[3] You always want to check this implication in your data, because what you find informs how you setup estimation. Here you carry out the check as follows: test the hypothesis that the two groups have the same means for all OPVs, variable by variable. Specifically, test that each variable's mean is the same in the control and treated groups by running 10 (ten) simple linear regressions (SLR) specifications. Use a $\alpha = 5\%$ significance level and look at the relevant 10 (ten) T test statistics, comment on your findings.

(f) (2p) Write 1 to 3 sentences to provide a clear and crisp description of the sample to a lay person (imagine describing the sample to a friend over dinner).

Guidance

Q1. **Hint**: Points assigned as follows: (1p) for OLS minimization problem and FOCs; (1p) for derivation of the closed-form solution; (3p) for proving unbiasedness of $\hat{\beta}_1$; (3p) for proving unbiasedness of $\hat{\beta}_0$; (10p) for proving consistency of $\hat{\beta}_1$; (4p) for proving consistency of $\hat{\beta}_0$. Points are assigned for the proofs only in so far as the correct assumptions or properties are invoked, that is, correct derivations that lack an explicit justification don't get points. You want to use the definitions of unbiased estimator from PSet0 as well as other results reviewed in PSet0 such as the Law of Iterated Expectations (LIE), the Law of Large Numbers (LLN), the Slutsky's Theorem (SLT), and the Continuous Mapping Theorem (CMT). **Note**: Many estimators developed in the course can be had *via* linear regression, thus familiarity with OLS is important.

Q2. **Hint**: Find the specific values of (β_0, β_1) that verify the statement.

Q3. **Note**: All references to Amazon.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. The Leadership Principles cited in the email exchanges are real.

Q7. **Note**: 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.

(a) **Hint**: In the background there is a consumer demand model, i.e., you think of model (1) as a consumer expenditure function from microeconomics. **Note**: It is common to use D_i to denote a binary 0/1 variable, it is mnemonic for dummy variable to mean that it stands in for a qualitative characteristic (in this case for *Prime status*).

(b) None

(c) **Hint**: 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 re-parametrize the model so that the objects in the reformulated model are learnable.

(d) None

(e) **Note**: Makes sure that you know how to derive these closed-form expressions as we will use them repeatedly during the quarter.

(f) None

(g) None

(h) **Hint**: Express ρ *exclusively* as a function of population data moments (PDM), that is, features of the population distribution of (y_i, D_i).

(i) **Hint**: Mean independence implies zero correlation but the other way around need not obtain. However, you can verify that given two RVs X and D such that D is binary 0/1 you have $Cov(X, D) = p(1-p)\Big(E[X|D=1] - E[X|D=0]\Big)$ where $p \equiv \Pr(D=1)$. What does this imply?

(j) **Hint**: You may use proofs previously developed in this PSet.

[3] OPVs are often called "covariates," the name stems from the traditional approach to causal inference which uses OPVs as regression covariates, i.e., as right-end-side variables in a regression equation.

(k) **Note**: The symbol "$\perp$" signifies statistical independence.

Q8. **Hint for Q8b**: Feel free to use the prompt "What are substitution bias and contamination bias in RCTs?" on your preferred chat bot or web-browser.

Q12. Here is the guidance:

(a) None.

(b) **Programming Guidance**: To load data you may use `utils::read.delim( )` where `utils` is the package and `utils::read.delim( )` is one of its functions; another option is `read.csv( )`. There are other packages/functions that accomplish this task, feel free to use whichever you prefer. To stack dataframes `df1` and `df2` into a combined dataframe `df` you may use `df <- rbind(df1,df2)` (mnemonic for row bind) available in base R; see Example 4.

(c) **Programming Guidance**: To count how many units are included in each sample you may use `dplyr::tally(dplyr::group_by(df, treat))` which employs the `dplyr` package and the function `group_by( )` to group by the treatment column `treat` and then `dplyr::tally( )` to produce the counts.

(d) **Programming Guidance**: To summarize multiple columns (e.g., compute averages) you may use `dplyr::summarise_all( )` in piped format, i.e. `dplyr::group_by(df, treat) %>% dplyr::summarise_all(list(mean))`, or without piping `dplyr::summarise_all(dplyr::group_by(df, treat), list(mean))`. Pipes let you take the output of one function and send it directly to the next, which is useful when you need to apply multiple transformation to the same data set. Pipes in R look like %>% and are made available via the `magrittr` package installed as part of `dplyr`.

(e) **Hint**: Each OPV is the dependent variable in its regression equation. All 10 SLR models have the same covariates. **Programming Guidance**: To run a SLR use `stats::lm( )` where `stats` is a package and `lm( )` is a function that estimates linear-in-parameter models. It is convenient to first create the formula for the model then pass it to `lm ()`. Example: Declare `formula <- stats::formula(paste(age, ''~treat''))` to create the formula for a SLR that has `age` as the dependent variable regressed on a constant and the treatment indicator (printing `formula` to standard output (stout) yields `age` $\sim$ `treat` because the constant is left implicit and present by default). Then pass the formula to `lm( )` by typing `lm_model <- lm(formula = formula, data = df)`. To retrieve regression output use `summary(lm_model)`, where `summary( )` is available in base R. Example: Retrieve regression coefficients with `summary(lm_model)$coefficients`. To repeat for each OPV, employ a `for` loop or e.g., `lapply( )`, see here for an example.