PS5024-NonR_ProbSet_Questions.rmd

Everyone!!

Problem Set 1.

1.1. Unbiased Estimator.

1.1.a. Consider a population y_i of size N, where $i = 1, \ldots, N$. Suppose we draw from the population a random sample of size n, and define an indicator variable z_i, where $z_i = 1$ if unit i was sampled and $z_i = 0$ otherwise. Notice that z_i is the only random variable in this scenario.

The population mean is $\mu_y = \frac{1}{N}\sum_{i=1}^{N} y_i$. Consider the estimator $\tilde{y}_i = \frac{1}{n}\sum_{i=1}^{N} z_i y_i$. Intuitively, what does $\tilde{y}_i$ represent? Show that it is an unbiased estimator of μ_y. Would this still be true if we did not have a simple random sample (e.g., if we were more likely to sample higher values of y_i than lower values)?

- $\tilde{y}_i$ represents the sample mean, an estimate of the population mean.
- Note that z_i is a random variable with two possible outcomes: $z_i \in \{0, 1\}$. This represents a Bernoulli probability distribution, which has an expected value of p, the probability of success $\Pr(z_i = 1)$. In this case, we know that $z_i = 1$ exactly n times where n is the number of units in the population y_i that were randomly sampled. Furthermore, $z_i = 0$ exactly $N - n$ times. Then the probability that $z_i = 1$ is: $\Pr(z_i = 1) = \frac{n}{N}$
  - Thus, the Bernoulli distribution has the following expected value: $\mathbb{E}[z_i] = \frac{n}{N}$
- An estimator is unbiased if $\mathbb{E}(\tilde{y}_i - \mu_y) = 0 \Rightarrow \mathbb{E}(\tilde{y}_i) = \mu_y$. Let's prove this:

$$
\begin{aligned}
\mathbb{E}(\tilde{y}_i) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{N}(z_i y_i)\right] && \text{(By Definition of } \tilde{y}_i\text{)} \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{N}(z_i y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{N}\left[\mathbb{E}(z_i y_i)\right] && \text{(By linearity of expectation)} \\
&= \frac{1}{n}\sum_{i=1}^{N}\left[y_i \mathbb{E}(z_i)\right] && (y_i \text{ is not a random variable}) \\
&= \frac{1}{n}\sum_{i=1}^{N}\left[y_i \cdot \frac{n}{N}\right] && \text{(By 1st moment of Bernoulli distribution.)} \\
&= \frac{1}{n}\cdot\frac{n}{N}\sum_{i=1}^{N}[y_i] \\
&= \frac{1}{N}\sum_{i=1}^{N}[y_i] \\
&= \mu_y && \text{(As desired.)}
\end{aligned}
$$

- This would not still be true if we lacked a simple random sample. If we were to over-sample higher values of y_i than lower values, then $\mathbb{E}(\tilde{y}_i) > \mu_y$.

1.2. Potential Outcome Notation.

This problem should help you get familiar with the potential outcomes notation. Try to keep your answers brief and your language precise. Throughout the problem, assume that the Stable Unit Treatment Value Assumption (SUTVA) holds.

1.2.a. Explain the notation $Y_i(0)$.

- $Y_i(0)$ is observation i's potential outcome with no treatment (i.e., under the control condition).

1.2.b. Contrast the meaning of $Y_i(0)$ with the meaning of Y_i.

- Y_i is the variable of interest. It has two potential outcomes based on if it receives treatment or not (see equation below). $Y_i(0)$ is the potential outcome for Y_i when it doesn't treatment (i.e., under the control condition).

$$
Y_i(d) = \begin{cases} Y_i(1) & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_i(0) & \text{Potential outcome for unit } i \text{ without treatment} \end{cases}
$$

1.2.c. Contrast the meaning of $Y_i(0)$ with the meaning of $Y_i(1)$. Is it ever possible to observe both at the same time? Why?

- $Y_i(0)$ is the potential outcome of variable Y for unit i under control (i.e., no treatment).
- $Y_i(1)$ is the potential outcome of variable Y for unit i under treatment.
- It is not possible to observe both potential outcomes at the same time because we cannot observe both $Y_i(1)$ and $Y_i(0)$ for the same i at the same time. This is the fundamental problem of causal inference.

1.2.d. Explain the notation $\mathbb{E}[Y_i(0) \mid D_i = 1]$, where D_i is a binary variable that gives the treatment status for subject $i, 1$ if treated, 0 if control.

- $\mathbb{E}[Y_i(0) \mid D_i = 1]$ is the expected value of the potential outcome for Y_i when unit i does not receive treatment, conditional on unit i receives treatment.

1.2.e. Contrast the meaning of $\mathbb{E}[Y_i(0)]$ with the meaning of $\mathbb{E}[Y_i \mid D_i = 0]$.

- $\mathbb{E}[Y_i(0)]$ is Y_i's potential outcome when Y_i doesn't receive treatment (i.e., under the control).
- $\mathbb{E}[Y_i \mid D_i = 0]$ is the expected value of *both* potential outcome for Y_i when unit i does not receive treatment, conditional on unit i receives treatment.
- **Difference**: the former quantity ($\mathbb{E}[Y_i(0)]$) includes only one potential outcome, while the latter quantity ($\mathbb{E}[Y_i \mid D_i = 0]$) includes *both* potential outcomes for Y_i conditional on unit i not receiving treatment. [More elaboration??]

1.2.f. Contrast the meaning of $\mathbb{E}[Y_i(0) \mid D_i = 1]$ with the meaning of $\mathbb{E}[Y_i(0) \mid D_i = 0]$.

- $\mathbb{E}[Y_i(0) \mid D_i = 1]$ is the expected value of the potential outcome for Y_i when unit i is not treated, conditional on unit i receiving treatment.
- $\mathbb{E}[Y_i(0) \mid D_i = 0]$ is similar except it's conditional on unit i *not* receiving treatment.

1.2.g. Which of the following expectations (that you explained in parts (d) through (f)) can be identified from observed information? Do not make any additional assumptions about the distributions of Y or D, except that there is at least one observation with $D_i = 1$, and at least one with $D_i = 0$ in the observed data.

$$\mathbb{E}[Y_i(0) \mid D_i = 1]$$
$$\mathbb{E}[Y_i(0)]$$
$$\mathbb{E}[Y_i \mid D_i = 0]$$
$$\mathbb{E}[Y_i(0) \mid D_i = 0]$$

- Let's address each of these in turn:
- $\mathbb{E}[Y_i(0) \mid D_i = 1]$ is **not observable** because it's not an actualizable quantity.
  - Note that this quantity is the potential outcome for Y_i when unit i doesn't receive treatment, conditional on unit i receiving treatment.
- $\mathbb{E}[Y_i(0)]$ is generally **not observable** in experimental settings because it requires that all units receive no treatment (i.e., the control condition), which would be a terrible experiment!
  - Note that this quantity is the expected value of the potential outcome for Y_i when unit i does not receive treatment.

- $\mathbb{E}[Y_i \mid D_i = 0]$ is **not observable** because it requires us to know both potential outcomes ($Y_i(0)$ and $Y_i(1)$) conditional on unit i not receiving treatment, but we cannot observe $Y_i(1)$ when unit i doesn't receive treatment.
  - Note that this quantity is the expected value of *both* potential outcome for Y_i when unit i does not receive treatment, conditional on unit i receives treatment.
- $\mathbb{E}[Y_i(0) \mid D_i = 0]$: is the only **observable** quantity. It's the only one where the potential outcome (i.e., $Y_i(0)$) and treatment status (i.e., $D_i = 0$) line up.
  - Note that this quantity is the expected value of the potential outcome for Y_i when unit i does not receive treatment, conditional on unit i not receiving treatment.

1.3. BONUS QUESTION: Define SUTVA, both formally and substantively.

- **Stable Unit Treatment Value Assumption (SUTVA)**: an assumption that consists of two conditions/assumptions.
  - **No Interference** between units: potential outcomes for a unit must not be affected by treatment for any other units.
    - Violations: spill-over effects, contagion, dilution.
  - **Same Version** of the treatment. (i.e., treatment stability, consistency.
    - Violations: variable levels of treatment, technical errors
- **SUTVA**, formal definition:

$$Y_{(D_1, D_2, \ldots, D_N)i} = Y_{\left(D'_1, D'_2, \ldots, D'_N\right)i} \quad \text{if} \quad D_i = D'_i$$

Problem Set 2.

2.2. OLS Estimator.

2.2.a. (***) We can describe an observed outcome variable y as a function of the explanatory variables X in the following form: $\mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}}$, where y is a column vector of length n, X is an $n \times k$ matrix, $\hat{\beta}$ is a column vector of length k, and $\hat{\mathbf{u}}$ is a column vector of length n. Given this setup, consider the following quantity (where ' indicates transpose):

$$(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})$$

What does this quantity represent in the terminology of linear regression theory? Using matrix notation, solve for $\hat{\beta}$ such that this quantity is minimized. What conditions are required such that your solution for $\hat{\beta}$ is well-defined?

- This quantity represents the **sum of squared residuals**—or **Residual Sum of Squares (RSS)**—in linear regression theory.
  - Specifically, it's how much of the variance in the outcome variable is unexplained by the linear regression model.
  - Generally, it's a measure of the discrepancy between the (observed) values of the outcome variable and the (predicted) values of the linear regression model.
- Let's solve for $\hat{\beta}$ such that this quantity is minimized in the proof below. We can do this in three easy steps:
  - First, simplify the residual sum of squares expression.
  - Second, take the first derivative of with respect to $\hat{\beta}$.
  - Third, set this first derivative equal to zero and solve for $\hat{\beta}$.

$$\begin{aligned}
\text{RSS} &= (\mathbf{y} - \mathbf{X}\hat{\beta})^\top(\mathbf{y} - \mathbf{X}\hat{\beta}) && \text{(By definition of RSS)} \\
&= \left(\mathbf{y}^\top - \hat{\beta}^\top\mathbf{X}^\top\right)(\mathbf{y} - \mathbf{X}\hat{\beta}) && \text{(Distribute transpose)} \\
&= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\hat{\beta} - \hat{\beta}^\top\mathbf{X}^\top\mathbf{y} + \hat{\beta}^\top\mathbf{X}^\top\mathbf{x}\hat{\beta} && \text{(Multiply out RHS)} \\
&= \mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\hat{\beta} + \hat{\beta}^\top\mathbf{X}^\top\mathbf{x}\hat{\beta} && \text{(Combine like terms)} \\
\frac{\partial \text{RSS}}{\partial \hat{\beta}} &= \frac{\partial}{\partial \hat{\beta}}\left[\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top\mathbf{X}\hat{\beta} + \hat{\beta}^\top\mathbf{X}^\top\mathbf{X}\hat{\beta}\right] && \text{(Take derivative w.r.t } \hat{\beta}) \\
&= \mathbf{X}^\top\mathbf{y} - \left(\mathbf{X}^\top\mathbf{X}\right)\hat{\beta} && \text{(Perform derivative)} \\
0 &= \mathbf{X}^\top\mathbf{y} - \left(\mathbf{X}^\top\mathbf{X}\right)\hat{\beta} && \text{(Set equal to zero)} \\
\left(\mathbf{X}^\top\mathbf{X}\right)\hat{\beta} &= \mathbf{X}^\top\mathbf{y} && \text{(Isolate } \hat{\beta}) \\
\hat{\beta} &= \left(\mathbf{x}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y} && \text{(Multiply by inverse)} \\
& && \text{(As desired)}
\end{aligned}$$

- The following condition is required for the above solution (i.e., the $\hat{\beta}$ that minimizes RSS) to be well-defined:
  - **No Perfect Collinearity**: none of the covariates are linearly dependent on other covariates (or combinations of other covariates). Linear dependence would make would make $\left(\mathbf{x}^\top\mathbf{X}\right)^{-1}$ impossible to calculate.

2.2.b. (***) Show that the conditional expectation function $\mathbb{E}[\mathbf{y} \mid \mathbf{X}]$ is equivalent to $\mathbf{X}\beta$ for a linear regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ under the zero conditional mean assumption. (Make sure to state what that assumption is using appropriate notation.) In precise language, explain what the conditional expectation function is.

- **Zero Conditional Mean Assumption** (ZCMA): we assume that, on average, the error term will have a mean of zero, even when conditioned on all covariate values.
  - Formal expression:

$$\mathbb{E}(u|\mathbf{X}) = \mathbb{E}(u) = 0$$

- Show that $\mathbb{E}[\mathbf{y} \mid \mathbf{X}] = \mathbf{X}\beta$ under the zero conditional mean assumption (ZCMA):

$$\begin{aligned}
\mathbb{E}[y \mid \mathbf{X}] &= \mathbb{E}(\mathbf{X}\beta + u \mid \mathbf{X}) && \text{(By given linear regression equation)} \\
&= \mathbb{E}(\mathbf{X}\beta \mid \mathbf{X}) + \mathbb{E}(u \mid \mathbf{X}) && \text{(Distribute expectation)} \\
&= \mathbf{X}\beta + \mathbb{E}(u \mid \mathbf{X}) && \text{(By rules of expectation and conditional expectation)} \\
&= \mathbf{X}\beta + 0 && \text{(By zero conditional mean assumption, } u \perp\!\!\!\perp \mathbf{X}\text{)} \\
&= \mathbf{X}\beta && \text{(As desired)}
\end{aligned}$$

- I now explain what the conditional mean function is (in plain language):
  - By design, the error term u is 0 in expectation.
  - We further assume heteroskedasticity–that the error term is uncorrelated with the matrix $\mathbf{X}$.
  - Then, conditioning on $\mathbf{X}$ changes nothing about the expectation of u (adds no information since they are uncorrelated) and it remains 0.

2.2.c. Consider the following data and corresponding graph (figure 1—2.2.c graph) which plots the population regression function of Y on X (represented as A in the graph), which we assume to be linear such that $\mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$. First, express the quantities corresponding to B, C and D in terms of the conditional expectations of X and Y. Second, express B, C and D in the graph in terms of β_0 and β_1. Finally, use the provided data to estimate B, C, D, β_0 and β_1.

X	Y
0	1
0	3
1	5
1	7

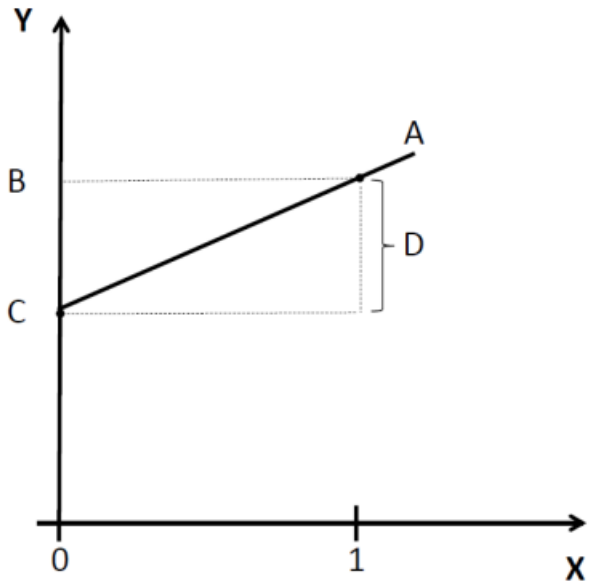

Figure 1: 2.2.c. Graph

- First, express the quantities corresponding to B, C and D in terms of the conditional expectations of X and Y.
  - $B = \mathbb{E}[Y|X = 1]$.
  - $C = \mathbb{E}[Y|X = 0]$.
  - $D = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0]$.
- Second, express B, C and D in the graph in terms of β_0 and β_1.
  - $B = \beta_0 + \beta_1$.
  - $C = \beta_0$.
  - $D = B - C = \beta_1$.
- Finally, use the provided data to estimate B, C, D, β_0 and β_1.
  - $B = \mathbb{E}[Y|X = 1] = \frac{5+7}{2} = 6$.
  - $C = \mathbb{E}[Y|X = 0] = \frac{1+3}{2} = 2$.
  - $D = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0] = 6 - 2 = 4$.
  - $\beta_0 = C = 2$.
  - $\beta_1 = D = 4$.

2.2.d. (***) Now we will derive the variance covariance matrix for $\hat{\beta}$. Assume that $\mathbb{E}[\mathbf{u}\mathbf{u}' \mid \mathbf{X}] = \sigma^2\mathbf{I}$, where $\mathbf{I}$ denotes the $n \times n$ identity matrix. Given your result for $\hat{\beta}$ from (a) and given that we can write $\hat{\beta}$ as $\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$, derive the variance of the OLS estimator $\hat{\beta}$ using matrix notation. (Hint: Remember that the variance of a random variable is defined as the expected value of the squared deviation from the mean.)

$$
\begin{aligned}
\mathbb{V}(\hat{\beta}) &= \mathbb{E}\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)^{\top} \mid \mathbf{X}\right] && \text{(By matrix definition of variance)} \\
&= \mathbb{E}\left\{\left[\beta + \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{u} - \beta\right]\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{u} - \beta\right]^{\top} \mid \mathbf{X}\right\} && \text{(By given identity for } \hat{\beta}) \\
&= \mathbb{E}\left\{\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{u}\right]\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{u}\right]^{\top} \mid \mathbf{X}\right\} && \text{(The } \beta\text{'s cancel out)} \\
&= \mathbb{E}\left\{\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{u}\right]\left[\mathbf{u}^{\top}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\right] \mid \mathbf{X}\right\} && \text{(Distribute transpose)} \\
&= \mathbb{E}\left[\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{u}\mathbf{u}^{\top}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1} \mid \mathbf{X}\right] && \text{(Remove parentheses)} \\
&= \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbb{E}\left[\mathbf{u}\mathbf{u}^{\top} \mid \mathbf{X}\right]\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1} && \text{(Pull out constants w.r.t. } \mathbf{X}) \\
&= \left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1} && \left(\text{By } \mathbb{E}\left[\mathbf{u}\mathbf{u}^{\top} \mid \mathbf{X}\right] = \sigma^2\mathbf{I}\right) \\
&= \sigma^2\mathbf{I}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{X}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1} && \left(\text{Pull out constant } \sigma^2\right) \\
&= \sigma^2\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1} && \text{(Inverses cancel)} \\
& && \text{(As desired)}
\end{aligned}
$$

2.3. Ghana's 2016 Election.

You and some colleagues are conducting an intervention in Ghana's 2016 election. Your goal is to assess the effect of deploying a new biometric voting machine on the incidence of electoral fraud at polling stations. Though Ghana is made up of 275 constituencies, due to the political context you are only allowed to perform your experiment within one constituency. Unconcerned, you and your team randomly select eight polling stations in the constituency, and among the eight, randomly assign half to receive the new voting machines and the other to serve as a control group. D_i gives the resulting treatment status for each polling station i, for $i \in \{1, \ldots, N\}$, where $D_i \in \{0, 1\}$ and $N = 8$. The outcome of interest is the percentage of votes in a polling station attributable to fraud, Y_i.

2.3.a. Assume that both parts of SUTVA hold. Calculate, explaining your answer:

2.3.a.i. For each polling station i, the number of potential outcomes that can be defined.

- There are only two potential outcomes we can define:
  - $Y_i(1)$: the value of Y for unit i under treatment.
  - $Y_i(0)$: the value of Y for unit i under control (i.e., no treatment).

2.3.a.ii. For each polling station i, the number of unit treatment effects that can be defined.

- For each polling station i, the **Average Treatment Effect (ATE)** is the only definable unit treatment effect:

$$\tau = Y_i(1) - Y_i(0)$$

2.3.a.iii. For the sample of polling stations, the number of (unconditional) average treatment effect estimands that can be defined.

- For the sample of polling stations, there is only **one** definable (unconditional) average treatment effect estimands:

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right] \\ &= \frac{1}{8}\sum_{i=1}^{8}\left[Y_i(1) - Y_i(0)\right]\end{aligned}$$

[**Check Answer**]

2.3.b. A Ghanaian political insider gets wind of your study, and gives you some information. She says that because all of the polling stations in your study are within one small constituency, there will be interference between units. She explains how "interference" might occur in this case: In Ghana, the political operatives in each polling station who are responsible for committing fraud will move elsewhere if their efforts are frustrated. Their range of movement is predetermined by the geographic influence of the party bosses. Local conditions are such that the operatives in each polling station can only move to one, and only one, other polling station, as shown in Figure 2 (2.3.b graph).

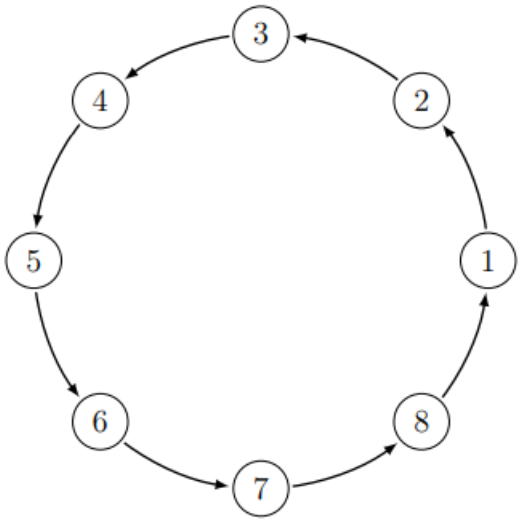

Figure 2: 2.3.b. Graph: Operatives may move to only one other polling station, as indicated by the arrows (for example, from the second to third polling station, but not the reverse).

Given this structure, again answer questions (i) - (iii).

2.3.b.i. For each polling station i, the number of potential outcomes that can be defined.

- Because we now account for the treatment administered in the original experiment and also for the potential "intervention" of operatives, the number of potential outcomes has increased. Let unit $i-1$ denote the previous polling station.
- For each polling station i, there are now **four** potential outcomes that can be defined:

- $\{Y_i(1), Y_{i-1}(1)\}$.
- $\{Y_i(0), Y_{i-1}(1)\}$.
- $\{Y_i(1), Y_{i-1}(0)\}$.
- $\{Y_i(0), Y_{i-1}(0)\}$.

2.3.b.ii. For each polling station i, the number of unit treatment effects that can be defined.

- For each polling station i, there are now **six** definable unit treatment effects:
  - $\tau_{i1} = Y_i(1,1) - Y_i(0,0)$.
  - $\tau_{i2} = Y_i(1,1) - Y_i(1,0)$.
  - $\tau_{i3} = Y_i(1,1) - Y_i(0,1)$.
  - $\tau_{i4} = Y_i(1,0) - Y_i(0,0)$.
  - $\tau_{i5} = Y_i(1,0) - Y_i(0,1)$.
  - $\tau_{i6} = Y_i(0,1) - Y_i(0,0)$.

2.3.b.iii. For the sample of polling stations, the number of (unconditional) average treatment effect estimands that can be defined.

- For the sample of polling stations, there are now **six** definable (unconditional) average treatment effect (ATE) estimands:
  - $\tau_{ATE,1} = \mathbb{E}\left[Y_i(1,1) - Y_i(0,0)\right]$.
  - $\tau_{ATE,2} = \mathbb{E}\left[Y_i(1,1) - Y_i(1,0)\right]$.
  - $\tau_{ATE,3} = \mathbb{E}\left[Y_i(1,1) - Y_i(0,1)\right]$.
  - $\tau_{ATE,4} = \mathbb{E}\left[Y_i(1,0) - Y_i(0,0)\right]$.
  - $\tau_{ATE,5} = \mathbb{E}\left[Y_i(1,0) - Y_i(0,1)\right]$.
  - $\tau_{ATE,6} = \mathbb{E}\left[Y_i(0,1) - Y_i(0,0)\right]$.

2.3.c. Your funder is keen to spend their budget, and demands that you field the intervention as described originally: $N = 8$, and for $D_i = \{0,1\}, \sum_{i=1}^{N} D_i = 4$. You set aside your concerns, and observe the data in Table 1 on the percentage of votes in each polling station attributable to fraud.

Write out an appropriate estimator for the average treatment effect (ATE) given SUTVA. Drawing on your insights so far and your knowledge of causal inference, under what conditions will this estimator be unbiased? Apply it to the data and compute an estimate of the effect of the new biometric voting machines on the incidence of fraud.

Table 1: Observed Data: Treatment and Ballot Stuffing

Unit	D_i	Y_i
1	1	4%
2	1	2%
3	0	8%
4	1	9%
5	0	12%
6	0	13%
7	0	4%
8	1	1%

- First, write out an appropriate estimator for the average treatment effect (ATE) given SUTVA. This is the difference in means estimator, which calculates the difference between the average outcome for the treatment group and the average outcome for the control group:

$$
\begin{aligned}
\hat{\tau}_{ATE} &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\
&= \frac{1}{4}\sum_{i=1}^{4}[Y_i(1) \mid D_i = 1] - \frac{1}{4}\sum_{i=1}^{4}[Y_i(0) \mid D_i = 0] \\
&= \frac{0.04 + 0.02 + 0.09 + 0.01}{4} - \frac{0.08 + 0.12 + 0.13 + 0.04}{4} \\
&= -0.0525
\end{aligned}
$$

- Thus, the difference in mean estimator suggests a 5.25% reduction in fraud.
- [CHECK ANSWER] This estimator will be unbiased under the following conditions:
  - SUTVA (non-interference and treatment equivalence).
  - Random assignment to treatment/control groups. (i.e., equal probability of receiving treatment as not receiving treatment).

[**Again, CHECK ANSWER**]

2.3.d. Given that you know the structure of the interference network, you believe you may be able to rescue something from the study. Using Figure 1 , translate Table 1 into a table formatted like Table 2 , where the latter columns give $Y_i(D_i, D_{i-1})$ for the combinations of possible values for D_i and D_{i-1}. (And for $i = 1, Y_1(D_1, D_8)$, per Figure 1). Where you can, fill the cells of your table with values from Table 1 ; leave blank the cells representing unobserved potential outcomes.

Table 2: Observed and Unobserved Potential Outcomes

Unit	D_i	D_{i-1}	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$
1						
⋮						
8						

- Here's my table:

Unit	D_i	D_{i-1}	$Y_i(1,1)$	$Y_i(1,0)$	$Y_i(0,1)$	$Y_i(0,0)$
1	1	1	4%			
2	1	1	2%			
3	0	1			8%	
4	1	0		9%		
5	0	1			12%	
6	0	0				13%
7	0	0				4%
8	1	0		1%		

2.3.e. Formally express each of the estimands described below using potential outcomes; propose appropriate estimators for each; and finally estimate them with the help of your new table.

2.3.e.i. The ATE, conditional on a neighbor taking treatment.

- **Estimand**:

$$\mathbb{E}[Y_i(1,1) - Y_i(0,1)] = \frac{1}{8}\sum_{i=1}^{8}[Y_i(1,1) - Y_i(0,1)]$$

- **Estimator**: $\mathbb{E}[Y_i|D_i = 1, D_{i-1} = 1] - \mathbb{E}[Y_i|D_i = 0, D_{i-1} = 1]$
- **Estimate**: $\frac{0.04+0.02}{2} - \frac{0.08+0.12}{2} = -0.07$
- Therefore, the ATE conditional on a neighbor receiving treatment is -7%.

2.3.e.ii. The ATE, conditional on a neighbor taking control.

- **Estimand**:

$$\mathbb{E}[Y_i(1,0) - Y_i(0,0)] = \frac{1}{8}\sum_{i=1}^{8}[Y_i(1,0) - Y_i(0,0)]$$

- **Estimator**: $\mathbb{E}[Y_i|D_i = 1, D_{i-1} = 0] - \mathbb{E}[Y_i|D_i = 0, D_{i-1} = 0]$.
- **Estimate**: $\frac{0.09+0.01}{2} - \frac{0.13+0.04}{2} = -0.035$.
  - Therefore, the ATE conditional on a neighbor taking control is -3.5%.

2.3.e.iii. The magnitude of effect modification due to assignment of treatment to a neighboring unit.

- **Estimand**:

$$\mathbb{E}[Y_i(1,1) - Y_i(0,1)] - \mathbb{E}[Y_i(1,0) - Y_i(0,0)] = \frac{1}{8}\sum_{i=1}^{8}([Y_i(1,1) - Y_i(0,1)] - [Y_i(1,0) - Y_i(0,0)])$$

- **Estimator**: $(\mathbb{E}[Y_i|D_i = 1, D_{i-1} = 1] - \mathbb{E}[Y_i|D_i = 0, D_{i-1} = 1]) - (\mathbb{E}[Y_i|D_i = 1, D_{i-1} = 0] - \mathbb{E}[Y_i|D_i = 0, D_{i-1} = 0])$.
- **Estimate**: $\frac{0.04+0.02}{2} - \frac{0.08+0.12}{2} - \left(\frac{0.09+0.01}{2} - \frac{0.13+0.04}{2}\right) = -0.07 - (-0.035) = -0.035$.
  - Therefore, the magnitude of effect modification due to assignment of treatment to a neighboring unit is -3.5%.

Problem Set 3.

3.2.f. Your friend comes back to you and asks if your result from part e) [IRRELEVANT TO THIS QUESTION] is generalizable, or just an artifact of the parameters you chose for your simulation. Respond to your friend with reference to the following formula for the sampling variance of our ATE estimate $\hat{\tau}$ under clustered random assignment, where n_1 is the number of units in treatment, n_0 is the number of units in control, M is the number of clusters, $\bar{y}_{0j}$ is average untreated potential outcome in cluster j, and $\bar{y}_{1j}$ is the average treated potential outcome in cluster j. Discuss at least two ways to reduce sampling variance in a clustered design.

$$\text{Var}(\hat{\tau}) = \frac{1}{M-1}\left(\frac{n_0}{n_1}\text{Var}\left(\bar{y}_{1j}\right) + \frac{n_1}{n_0}\text{Var}\left(\bar{y}_{0j}\right)\right)$$

- If we increase the number of clusters sampled M, the fraction $\frac{1}{M-1}$ indicates that the variation of the estimated average treatment effect decreases. However, if random assignment to treatment and control groups is left the same (i.e., the proportion of observations in the treatment and control groups remains the same as the total number of units increases), then adding new clusters will not effect the variance of the treatment mean and control mean. This indicates that increasing the number of clusters decreases variance in the estimator regardless of the values of the given parameters in this simulation.
- Two ways we can reduce sampling variance in clustered designs:
  - First, sample more clusters (based on the argument provided above).
  - Second, decrease the difference in variance between each mean (i.e., try to get as close to $\text{Var}\left(\bar{y}_{1j}\right) = \text{Var}\left(\bar{y}_{0j}\right)$ as possible). This means including in your sample more clusters that are similar to each other in the covariates. By doing this, you make it easier to attribute difference in the means to the treatment effect and not random noise.

3.3. ATE and the Difference-in-Means Variance.

Consider a field experiment that compares treatments A and B (You can think of B as the absence of a treatment, i.e. the "treatment" received by the control group). Suppose there are N subjects, indexed by $i = 1, \ldots, N$. Let x_i be the response of subject i to treatment A; likewise, y_i is the response to receiving treatment B (the control). For each i, either x_i or y_i can be observed, but not both. Let S be a random subset of $\{1, \ldots, N\}$, with n elements; this group gets treatment A, so x_i is observed for i in S. Let T be a random subset of $\{1, \ldots, N\}$, with m elements, disjoint from S. This group gets treatment B, so y_i is observed for i in T.

We estimate population means $\bar{x}$ and $\bar{y}$ by the sample means:

$$\bar{X} = \frac{1}{n}\sum_{i}^{n} x_i \quad \bar{Y} = \frac{1}{m}\sum_{i}^{m} y_i$$

Using simple sampling without replacement formulas:

$$\text{var}(\bar{X}) = \frac{N-n}{N-1}\frac{\sigma^2}{n} \quad \text{var}(\bar{Y}) = \frac{N-m}{N-1}\frac{\tau^2}{m}$$
$$\text{cov}(\bar{X}, \bar{Y}) = -\frac{1}{N-1}\text{cov}(x, y)$$

3.3.a. What is the average treatment effect parameter? Write it using the above notation and also explain what it is in words.

- The average treatment effect parameter is the difference in the respective population means, which are themselves estimated by the difference in sample means [Check Answer]:

$$\begin{aligned}\tau_{ATE} &= \bar{X} - \bar{Y} && \text{(Population difference in means)}\\ &= \frac{1}{n}\sum_{i}^{n} x_i - \frac{1}{m}\sum_{i}^{m} y_i && \text{(Sample difference in means)}\end{aligned}$$

[Check Answer]

3.3.b. What is the variance of the average treatment effect estimator, i.e. $\text{var}(\bar{X} - \bar{Y})$, using the above notation?

- The variance of the difference between random variables/estimators $\bar{X}$ and $\bar{Y}$ is their variances plus their covariance [Check Answer]:

$$\begin{aligned}\mathbb{V}(\bar{X} - \bar{Y}) &= \mathbb{V}(\bar{X}) + \mathbb{V}(-\bar{Y}) + 2 \cdot \text{cov}(\bar{X}, -\bar{Y})\\ &= \mathbb{V}(\bar{X}) + \mathbb{V}(\bar{Y}) - 2 \cdot \text{cov}(\bar{X}, \bar{Y}) && \text{(Pull out constant } -1)\\ &= \frac{N-n}{N-1}\frac{\sigma^2}{n} + \frac{N-m}{N-1}\frac{\tau^2}{m} - 2 \cdot \left[-\frac{1}{N-1}\text{cov}(x,y)\right] && \text{(Given)}\\ &= \frac{N-n}{N-1}\frac{\sigma^2}{n} + \frac{N-m}{N-1}\frac{\tau^2}{m} + \frac{2}{N-1}\text{cov}(x,y) && \text{(Simplify)}\\ &= \frac{1}{N-1}\left[\frac{(N-n)\cdot\sigma^2}{n} + \frac{(N-m)\cdot\tau^2}{m} + 2 \cdot \text{cov}(x,y)\right] && \text{(Pull out common denominator)}\end{aligned}$$

[Check Answer]

3.3.c. The usual two sample difference-in-means variance (without replacement) found in sampling textbooks is:

$$\frac{N}{N-1}\left(\frac{\sigma^2}{n} + \frac{\tau^2}{m}\right)$$

What is the difference, if any, between the usual two sample difference-in-means variance and the "true" variance expression you derived in part b? Can we observe this difference in the data we have?

- The **covariance term** $2 \cdot \text{cov}(x, y)$ is the biggest difference between the two *sample* difference in means variance and the *true* variance. The covariance term is present in the *true* variance but not the sample variance. This is because we cannot observe this covariance term, as we cannot simultaneously observe the actual potential outcomes under treatment and control.

[Check Answer]

Problem Set 5.

5.1. Curse of Dimensionality.

The curse of dimensionality makes it difficult to work in a situation where there are many pre-treatment covariates to condition on. Suppose we have covariates X_k for $k = 1, \ldots, P$, where P is the number of pre-treatment covariates (i.e., the dimensionality of the covariate space). Then x_i is a vector of covariate values for observation $i, x_i = [x_{i1}, \ldots, x_{iP}]^T$.

5.1.a. Write an expression that gives the Euclidean distance between observations i and j in terms of their covariates x_i and x_j, respectively.

- Let's first consider smaller-dimensional spaces before developing an expression for n-dimensions:
  - For $P = 1$, the Euclidean distance is: $d(x_i, x_j) = \sqrt{(x_i - x_j)^2} = \sqrt{(x_{i1} - x_{j1})^2}$.
  - For $P = 2$, the Euclidean distance is: $d(x_i, x_j) = \sqrt{(x_i - x_j)^2} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$.
- Then the expression of the Euclidean distance for any value of P is:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{P} (x_{ik} - x_{jk})^2}$$

Problem Set 6.

6.1. Fictitious Observational Study.

The table below summarizes outcomes from a fictitious observational study assessing whether receiving assistance from one's elected representative after making a request for assistance increases political self-efficacy. Data come from a survey of 800 constituents who made requests to their representatives. The survey measured political self-efficacy using a thermometer-style scale ranging from 0 to 100 . We'll denote the outcome Y_i, for constituents $i = 1, \ldots, N$, and the treatment $D_i \in \{0, 1\}$. Our quantity of interest is the average effect of receiving a representative's assistance on a constituent's self-efficacy, $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$, the ATE.

D_i	$\mathbb{E}[Y_i]$
1	73
0	56

The survey also measured constituent income, coded in our data as a discrete covariate $X_i \in \{1, 2, 3\}$, indicating low, middle, and high income. In the sample, 50% of constituents are low-income, 30% are middle-income, and 20% are high-income. Frequencies by income level are displayed in the following table.

Freq.	$X_i = 1$	$X_i = 2$	$X_i = 3$	Total
$D_i = 1$	212	163	125	500
$D_i = 0$	188	77	35	300
Total	400	240	160	800

Throughout this problem, assume that the observed data represent the entire population of interest and the variables are measured without error, so that we can ignore sampling variability. (In real applied settings, you would want to incorporate sampling uncertainty and report confidence intervals, etc., as part of your result.)

6.1.a. Write an expression for τ using the expressions in the following table. Then create a similar table that replaces the expressions in each cell with actual values, using only the observed data from this study. Leave the original expression in any cell you cannot populate with a value from the observed data.

d	$\Pr(D_i = d)$	$\mathbb{E}[Y_i(1) \mid D_i = d]$	$\mathbb{E}[Y_i(0) \mid D_i = d]$
1	$\Pr(D_i = 1)$	$\mathbb{E}[Y_i(1) \mid D_i = 1]$	$\mathbb{E}[Y_i(0) \mid D_i = 1]$
0	$\Pr(D_i = 0)$	$\mathbb{E}[Y_i(1) \mid D_i = 0]$	$\mathbb{E}[Y_i(0) \mid D_i = 0]$

- The expression for τ—where $\pi = \Pr(D_i = 1)$ and $1 - \pi = \Pr(D_i = 0)$ —follows:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_{i1} - Y_{i0}] \\
&= \pi \cdot \text{ATT} + [1 - \pi] \cdot \text{ATU} \\
&= \pi \{\mathbb{E}[Y_{i1} \mid D = 1] - \mathbb{E}[Y_{i0} \mid D = 1]\} + [1 - \pi] \{\mathbb{E}[Y_{i1} \mid D = 0] - \mathbb{E}[Y_{i0} \mid D = 0]\} \\
&= \pi \cdot \mathbb{E}[Y_{i1} \mid D = 1] - \pi \cdot \mathbb{E}[Y_{i0} \mid D = 1] + [1 - \pi] \cdot \mathbb{E}[Y_{i1} \mid D = 0] - [1 - \pi] \cdot \mathbb{E}[Y_{i0} \mid D = 0] \\
&= \pi \cdot \mathbb{E}[Y_{i1} \mid D = 1] + [1 - \pi] \cdot \mathbb{E}[Y_{i1} \mid D = 0] - \pi \cdot \mathbb{E}[Y_{i0} \mid D = 1] - [1 - \pi] \cdot \mathbb{E}[Y_{i0} \mid D = 0] \\
&= \Pr(D_i = 1) \cdot \mathbb{E}[Y_{i1} \mid D = 1] + \Pr(D_i = 0) \cdot \mathbb{E}[Y_{i1} \mid D = 0] - \\
&\quad [\Pr(D_i = 1) \cdot \mathbb{E}(Y_{i0} \mid D = 1) + \Pr(D_i = 0) \cdot \mathbb{E}(Y_{i0} \mid D = 0)]
\end{aligned}
$$

- Now create new table with derived from empirical data:

d	$\Pr(D_i = d)$	$\mathbb{E}[Y_i(1) \mid D_i = d]$	$\mathbb{E}[Y_i(0) \mid D_i = d]$
1	$\frac{5}{8}$	73	$\mathbb{E}[Y_i(0) \mid D_i = 1]$
0	$\frac{3}{8}$	$\mathbb{E}[Y_i(1) \mid D_i = 0]$	56

6.1.b. Suppose you have some reason to be concerned about the identifying assumption of conditional ignorability. Calculate sharp bounds for τ without making any assumptions at all. Explain in simple, concise language what these bounds represent. You may find it helpful to use the table you produced in part (a) and fill in the values that produce the upper and lower no-assumptions sharp bounds.

- Calculate sharp bounds for τ without making assumptions. The outcome variable *political self-efficacy* is measured using a thermometer-style scale ranging from 0 to 100.

6.1.b.i. Sharp Lower Bound: Worst Possible Outcome.

- Let's assume that the treated units would have the best possible outcome $\bar{Y} = 100$ if left untreated and the control units would have the worst possible outcome $\underline{Y} = 0$ if treated. Thus, we can reproduce the table with these pessimistic assumptions:

d	$\Pr(D_i = d)$	$\mathbb{E}[Y_i(1) \mid D_i = d]$	$\mathbb{E}[Y_i(0) \mid D_i = d]$
1	$\frac{5}{8}$	73	$\overline{Y} = 100$
0	$\frac{3}{8}$	$\underline{Y} = 0$	56

- Now let's calculate the sharp lower bound τ with this information:

$$\begin{aligned}\underline{\tau} &= \frac{5}{8} \cdot 73 + \frac{3}{8} \cdot 0 - \frac{5}{8} \cdot 100 - \frac{3}{8} \cdot 56 \\ &= -\frac{303}{8} \\ &= -37.875\end{aligned}$$

6.1.b.ii. Sharp Upper Bound: Best Possible Outcome

- Now, let's assume that the treated units would have the worst possible outcome $\bar{Y} = 100$ if left untreated and the control units would have the best possible outcome $\underline{Y} = 0$ if treated. Thus, we can reproduce the table with these optimistic assumptions:

d	$\Pr(D_i = d)$	$\mathbb{E}[Y_i(1) \mid D_i = d]$	$\mathbb{E}[Y_i(0) \mid D_i = d]$
1	$\frac{5}{8}$	73	$\underline{Y} = 0$
0	$\frac{3}{8}$	$\overline{Y} = 100$	56

- Now let's calculate the sharp upper bound τ with this information:

$$\begin{aligned}\overline{\tau} &= \frac{5}{8} \cdot 73 + \frac{3}{8} \cdot 100 - \frac{5}{8} \cdot 0 - \frac{3}{8} \cdot 56 \\ &= \frac{497}{8} \\ &= 62.125\end{aligned}$$

- Then the final interval for τ is:

$$\begin{aligned}\tau &\in [\underline{\tau}, \overline{\tau}] \\ &\in [-37.875, 62.125]\end{aligned}$$

- This interval is far too broad for it to be plausible or practical.

6.1.c. One possible threat to identification is that constituents with higher self-efficacy are more likely to follow up on their requests, and in turn more likely to receive assistance from their representatives (assume follow-up requests are not captured in the survey). If this were the case, treated constituents would tend to have greater self-efficacy than untreated units, whether or not they received treatment. State this assumption formally and use it to calculate new bounds.

- **Monotone Treatment Selection (MTS) Assumption**: In this case, the MTS assumption states that the surveyed individuals that received treatment (receiving a representative's assistance) already have higher self-efficacy.
  - [NEEDS BETTER DEFINITION/EXPLANATION OF MTS ASSUMPTION]
  - Formal expression:

$$\begin{aligned}\mathbb{E}\left[Y_{0i} \mid D_i = 0\right] &\leq \mathbb{E}\left[Y_{0i} \mid D_i = 1\right] \\ \mathbb{E}\left[Y_{1i} \mid D_i = 0\right] &\leq \mathbb{E}\left[Y_{1i} \mid D_i = 1\right]\end{aligned}$$

- If we use the MTS assumption to calculate a new upper bound, we get the following:

$$\begin{aligned}\tau &= \frac{5}{8} \cdot E\left[Y_{i1} \mid D = 1\right] + \frac{3}{8} \cdot E\left[Y_{i1} \mid D = 0\right] - \frac{5}{8} \cdot E\left[Y_{i0} \mid D = 1\right] - \frac{3}{8} \cdot E\left[Y_{i0} \mid D = 0\right] \\ &\leq \frac{5}{8} \cdot E\left[Y_{i1} \mid D = 1\right] + \frac{3}{8} \cdot E\left[Y_{i1} \mid D = 1\right] - \frac{5}{8} \cdot E\left[Y_{i0} \mid D = 1\right] - \frac{3}{8} \cdot E\left[Y_{i0} \mid D = 1\right] \\ &\leq E\left[Y_{i1} \mid D = 1\right] - E\left[Y_{i0} \mid D = 1\right] \\ &\leq 73 - 56 \\ &\leq 17\end{aligned}$$

- This is the new bound given the MTS assumption:

$$\begin{aligned}\tau &\in [\underline{\tau}, \tau_{MTS}] \\ &\in [-37.875, 17]\end{aligned}$$

- This is a tighter bound, but it still crosses zero. Therefore, it's not clear that the ATE is different than zero.

6.1.d. Now consider that elected representatives have limited time and resources and must make strategic decisions about which requests to prioritize. Representatives want to maximize the probability of winning re-election, and they know there are greater payoffs to responding to requests from high income people, who more likely to both turn out and make large campaign donations. Assume representatives can accurately guess a constituent's income level based on factors like the constituent's address and the type of request. As a result of this incentive structure, the probability of response to requests from high-income constituents is unrelated the number of times they follow up (and therefore unrelated to pre-existing levels of self-efficacy). For low- and middle-income constituents, however, positive self-selection may still hold. Express this assumption formally. Would the resulting bounds be more or less credible than those under your assumption in (c)? Explain.

- If self-efficacy is unrelated to a representative's response for high-income residents but still related to a representative's response for low- and middle-income residents, then we'd expect there to be no difference in the outcomes of treated vs. untreated high-income residents, but the same relationship as identified in problem (1c) for low- and middle-income residents.
- Let's express this formally:
  - For low income residents:

$$\mathbb{E}\left[Y_{0i} \mid D_i = 0 \cap X_i \in \{1,2\}\right] \leq \mathbb{E}\left[Y_{0i} \mid D_i = 1 \cap X_i \in \{1,2\}\right]$$
$$\mathbb{E}\left[Y_{1i} \mid D_i = 0 \cap X_i \in \{1,2\}\right] \leq \mathbb{E}\left[Y_{1i} \mid D_i = 1 \cap X_i \in \{1,2\}\right]$$

- But for high income resident, there is no difference:

$$\mathbb{E}\left[Y_{0i} \mid D_i = 0 \cap X_i = 3\right] = \mathbb{E}\left[Y_{0i} \mid D_i = 1 \cap X_i = 3\right]$$
$$\mathbb{E}\left[Y_{1i} \mid D_i = 0 \cap X_i = 3\right] = \mathbb{E}\left[Y_{1i} \mid D_i = 1 \cap X_i = 3\right]$$

- I would think that the bounds we'd fine in this case would be *more* credible because we're making the MTS assumption on a smaller segment of the survey sample.

6.1.e. Use the assumption described in (d) and the following table of conditional expectations to calculate new bounds.

D_i	$\mathbb{E}[Y_i]$	$\mathbb{E}[Y_i \mid X_i = 1]$	$\mathbb{E}[Y_i \mid X_i = 2]$	$\mathbb{E}[Y_i \mid X_i = 3]$
1	73	69	73	83
0	56	51	55	70

- [**CHECK ANSWER**]
- Okay let's do the calculations (where π isn't a set quantity. Rather, it's a placeholder for that segment of the survey population's probability/proportion of the sample) [**CHECK ANSWER**]:

$$
\begin{aligned}
\tau &= \pi \cdot E\left[Y_{i1} \mid D=1\right] + \left[1-\pi\right] \cdot E\left[Y_{i1} \mid D=0\right] - \pi \cdot E\left[Y_{i0} \mid D=1\right] - \left[1-\pi\right] \cdot E\left[Y_{i0} \mid D=0\right] \\
&= \pi \cdot E\left[Y_{i1} \mid D=1 \cap X_i=1\right] + \left[1-\pi\right] \cdot E\left[Y_{i1} \mid D=0 \cap X_i=1\right] \\
&\quad - \pi \cdot E\left[Y_{i0} \mid D=1 \cap X_i=1\right] - \left[1-\pi\right] \cdot E\left[Y_{i0} \mid D=0 \cap X_i=1\right] \\
&\quad + \pi \cdot E\left[Y_{i1} \mid D=1 \cap X_i=2\right] + \left[1-\pi\right] \cdot E\left[Y_{i1} \mid D=0 \cap X_i=2\right] \\
&\quad - \pi \cdot E\left[Y_{i0} \mid D=1 \cap X_i=2\right] - \left[1-\pi\right] \cdot E\left[Y_{i0} \mid D=0 \cap X_i=2\right] \\
&\quad + \pi \cdot E\left[Y_{i1} \mid D=1 \cap X_i=3\right] + \left[1-\pi\right] \cdot E\left[Y_{i1} \mid D=0 \cap X_i=3\right] \\
&\quad - \pi \cdot E\left[Y_{i0} \mid D=1 \cap X_i=3\right] - \left[1-\pi\right] \cdot E\left[Y_{i0} \mid D=0 \cap X_i=3\right] \\
&\leq \pi \cdot E\left[Y_{i1} \mid D=1 \cap X_i=1\right] + \left[1-\pi\right] \cdot E\left[Y_{i1} \mid D=1 \cap X_i=1\right] \\
&\quad - \pi \cdot E\left[Y_{i0} \mid D=1 \cap X_i=1\right] - \left[1-\pi\right] \cdot E\left[Y_{i0} \mid D=1 \cap X_i=1\right] \\
&\quad + \pi \cdot E\left[Y_{i1} \mid D=1 \cap X_i=2\right] + \left[1-\pi\right] \cdot E\left[Y_{i1} \mid D=1 \cap X_i=2\right] \\
&\quad - \pi \cdot E\left[Y_{i0} \mid D=1 \cap X_i=2\right] - \left[1-\pi\right] \cdot E\left[Y_{i0} \mid D=1 \cap X_i=2\right] \\
&\quad + \pi \cdot E\left[Y_{i1} \mid D=1 \cap X_i=3\right] + \left[1-\pi\right] \cdot E\left[Y_{i1} \mid D=1 \cap X_i=3\right] \\
&\quad - \pi \cdot E\left[Y_{i0} \mid D=1 \cap X_i=3\right] - \left[1-\pi\right] \cdot E\left[Y_{i0} \mid D=1 \cap X_i=3\right] \\
&\leq \quad \frac{1}{2} \cdot \left\{E\left[Y_i \mid D=1 \cap X_i=1\right] - E\left[Y_i \mid D=0 \cap X_i=1\right]\right\} \\
&\quad + \frac{3}{10} \cdot \left\{E\left[Y_i \mid D=1 \cap X_i=2\right] - E\left[Y_i \mid D=0 \cap X_i=2\right]\right\} + \frac{1}{5} \cdot 0 \\
&\leq \quad \frac{1}{2} \cdot (69-51) + \frac{3}{10} \cdot (73-55) \\
&\leq \quad \frac{72}{5} \\
&\leq \quad 14.4
\end{aligned}
$$

- Then the new bound is [**CHECK ANSWER**]:

$$\tau \in [-37.875, 14.4]$$

[**CHECK ANSWER**]

Problem Set 7.

7.1. The IV Estimator.

7.1.a. Suppose that we are interested in the effect of a potentially endogenous causal variable X_i on an outcome variable of interest Y_i. Assume that we have another variable Z_i, which is binary and is an instrumental variable for X_i. Show that the IV estimator for the effect of X_i on Y_i

$$\hat{\beta}_{IV} = \frac{\text{cov}\left(Z_i, Y_i\right)}{\text{cov}\left(Z_i, X_i\right)}$$

Can be written as

$$\frac{\left(\bar{Y}_1 - \bar{Y}_0\right)}{\left(\bar{X}_1 - \bar{X}_0\right)},$$

where $\text{cov}(\cdot)$ is the sample covariance; $\bar{Y}_0$ and $\bar{X}_0$ are the sample averages of Y_i and X_i over the part of the sample with $Z_i = 0$; and $\bar{Y}_1$ and $\bar{X}_1$ are the sample averages of Y_i and X_i over the part of the sample with $Z_i = 1$.

- i.e., we need to show that:

$$\hat{\beta}_{\text{IV}} = \frac{\text{cov}\left(Z_i, Y_i\right)}{\text{cov}\left(Z_i, X_i\right)} \Leftrightarrow \frac{\left(\bar{y}_1 - \bar{y}_0\right)}{\left(\bar{x}_1 - \bar{x}_0\right)}$$

- I first show that $\text{cov}\left(Z_i, Y_i\right) \Leftrightarrow \left(\bar{Y}_1 - \bar{Y}_0\right)$.
  - If I can show that this is true for the numerator, then it's also true for the denominator.
- First, Let's establish several preliminary conclusions and definitions that will be used in the final proof. All of these will be for Y_i (the numerator) but also hold true if we replace it with X_i (the denominator)
  - **Preliminary 1**: By the given definitions of $\bar{Y}_1, \bar{Y}_0, \bar{X}_1$, and $\bar{X}_0$ and that Z_i is a binary variable:

$$\bar{Y}_1 - \bar{Y}_0 = \mathbb{E}\left[Y_i \mid Z_i = 1\right] - \mathbb{E}\left[Y_i \mid Z_i = 0\right] \qquad (P1)$$

- **Preliminary 2**: By the definition of covariance:

$$\text{cov}\left(Z_i, Y_i\right) = \mathbb{E}\left[Z_i Y_i\right] - \mathbb{E}\left[Z_i\right]\mathbb{E}\left[Y_i\right] \qquad (P2)$$

- **Preliminary 3**: By definition of expected value for binary variable z_i:

$$\begin{aligned}\mathbb{E}\left[z_i\right] &= 1 \cdot \text{P}\left(z_i = 1\right) + 0 \cdot \text{P}\left(z_i = 0\right) \\ &= \text{P}\left(z_i = 1\right) \qquad (P3)\end{aligned}$$

- **Preliminary 4**: By the law of total expectation & discrete $Z_i \in \{0, 1\}$:

$$\begin{aligned}\mathbb{E}[Y_i] &= \sum_{Z_i=0}^{1} \{\mathbb{E}[Y_i \mid Z_i = z_i] \cdot \mathrm{P}(Z_i = z_i)\} \\ &= \mathbb{E}[Y_i \mid Z_i = 0] \cdot \mathrm{P}(Z_i = 0) + \mathbb{E}[Y_i \mid Z_i = 1] \cdot \mathrm{P}(Z_i = 1) \qquad (P4)\end{aligned}$$

- **Preliminary 5**: Note that $\mathbb{E}[Y_i|Z_i = 1] = \sum_{Y_i} y_i \frac{\mathrm{P}(Y_i, Z_i=1)}{\mathrm{P}(Z_i=1)}$. Then, by the law of total expectation & discrete $Z_i \in \{0, 1\}$:

$$\begin{aligned}\mathbb{E}[Y_i \cdot Z_i] &= \sum_{Y_i}\sum_{Z_i} y_i z_i \mathrm{P}(Y_i, Z_i) \\ &= \sum_{Y_i} y_i \cdot 0 \cdot \mathrm{P}(Y_i, Z_i = 0) + \sum_{Y_i} y_i \cdot 1 \cdot \mathrm{P}(Y_i, Z = 1) \\ &= \sum_{Y_i} y_i \cdot \mathrm{P}(Y_i, Z_i = 1) \\ &= \sum_{Y_i} y_i \cdot \mathrm{P}(Y_i, Z_i = 1) \cdot \frac{\mathrm{P}(Z_i = 1)}{\mathrm{P}(Z_i = 1)} \\ &= \mathbb{E}[Y_i \mid Z_i = 1] \cdot \mathrm{P}(Z_i = 1) \qquad (P5)\end{aligned}$$

- Given these preliminary conclusions and definitions, we can finally show that $\mathrm{cov}(Z_i, Y_i) = \left(\bar{Y}_1 - \bar{Y}_0\right)$:

$$\begin{aligned}&\mathrm{Cov}(Z_i, Y_i) = \mathbb{E}[Z_i \cdot Y_i] - \mathbb{E}[Z_i]\mathbb{E}[Y_i] && \text{(By } P2) \\ &= \mathbb{E}[Y_i \mid Z_i = 1] \cdot \mathrm{P}(Z_i = 1) - \mathrm{P}(Z_i = 1) \cdot \{\mathbb{E}[Y_i \mid Z_i = 0] \cdot \mathrm{P}(Z_i = 0) + \mathbb{E}[Y_i \mid Z_i = 1] \cdot \mathrm{P}((Z_i = 1)\} && \text{(By } P3, P4) \\ &= \mathrm{P}((Z_i = 1)\{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0] \cdot \mathrm{P}(Z_i = 0) - \mathbb{E}[Y_i \mid Z_i = 1] \cdot \mathrm{P}(Z_i = 1)\} \\ &= \mathrm{P}(Z_i = 1)\{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0][1 - \mathrm{P}(Z_i = 1)] - \mathbb{E}[Y_i \mid Z_i = 1] \cdot \mathrm{P}(Z_i = 1)\} \\ &= \mathrm{P}(Z_i = 1)\{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0] + \mathbb{E}[Y_i \mid Z_i = 0] \cdot \mathrm{P}(Z_i = 1) - \mathbb{E}[Y_i \mid Z_i = 1] \cdot \mathrm{P}(Z_i = 1)\} \\ &= \mathrm{P}(Z_i = 1)\{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0] - \mathrm{P}(Z_i = 1)(\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0])\} \\ &= \mathrm{P}(Z_i = 1)\left\{\bar{Y}_1 - \bar{Y}_0 - \mathrm{P}(Z_i = 1)\left(\bar{Y}_1 - \bar{Y}_0\right)\right\} \\ &= \mathrm{P}(Z_i = 1)\left\{[1 - \mathrm{P}(Z_i = 1)]\left(\bar{Y}_1 - \bar{Y}_0\right)\right\} \\ &= \mathrm{P}(Z_i = 1) \cdot \mathrm{P}(Z_i = 0) \cdot \left(\bar{Y}_1 - \bar{Y}_0\right)\end{aligned}$$

- Again, this also holds for the denominator.
- Thus, we get the following outcome:

$$\begin{aligned}\hat{\beta}_{IV} &= \frac{\mathrm{cov}(Z_i, Y_i)}{\mathrm{cov}(Z_i, X_i)} \\ &= \frac{\mathrm{P}(Z_i = 1) \cdot \mathrm{P}(Z_i = 0) \cdot \left(\bar{Y}_1 - \bar{Y}_0\right)}{\mathrm{P}(Z_i = 1) \cdot \mathrm{P}(Z_i = 0) \cdot \left(\bar{X}_1 - \bar{X}_0\right)} \\ &= \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0} && \text{(As desired)}\end{aligned}$$

7.1.b. Let $\mathbf{X} = [1, X_1, X_2, \cdots, X_k, D]$ and $\mathbf{Z} = [1, X_1, X_2, \cdots, X_k, Z]$. The matrix X contains the covariates (including a vector of 1s) and your treatment vector D, and Z is a matrix of the same covariates and the instrument for the treatment variable in place of the actual treatment. Y is a vector of observed outcomes. We can construct the following system of linear equations, with error terms u_2 and u_1 respectively:

$$Y = \mathbf{X}\beta + u_2$$
$$D = \mathbf{Z}\pi + u_1$$

with coefficient vectors $\beta = [\beta_0, \beta_1, \cdots, \beta_k, \beta_D]$ and $\pi = [\pi_0, \pi_1, \cdots, \pi_k, \pi_Z]$.

The IV estimator can be obtained by:

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'Y$$

7.1.b.i. What are the conditions under which the treatment effect estimate $\hat{\beta}_{IV}$ is consistent?

- There are three conditions under which $\hat{\beta}_{IV}$ is consistent:
  - 1. **Exogeneity**: The instrument Z must be as good as random conditional on the other covariates.
    - In other words, Z must be uncorrelated with the error term u_1:

$$\text{Cov}[u_1, Z] = 0$$

- 2. **Exclusion Restriction**: The instrument Z must have no direct effect on Y, the outcome variable.
  - Instead, Z should only affect Y via the treatment variable D:

$$\text{Cov}[u_2, D] = 0$$

- 3. **Relevance**: The instrument Z must be significantly correlated with the treatment variable D.
  - In other words, Z must have a (reasonably strong) effect on D.
- On a final (somewhat unrelated) note, please keep in mind that The 2SLS estimator, in contrast, is consistent, but biased.
  - This means that the 2SLS estimator only promises to be close the causal effect of interest in large samples.
    - In small samples, the 2SLS estimator can differ systematically from the population estimand.
  - The 2SLS estimator is most biased when the instruments are weak, meaning the correlation with endogenous regressors is low, and when there are many over-identifying restrictions.

7.1.b.ii. Now let's obtain the Two-stage Least Squares estimator. We can do that following these next steps.

- **(a) Run the first stage regression:** $D = \mathbf{Z}\pi + u_1 \Rightarrow \hat{\pi} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'D$
- **(b) Get fitted values:** $\hat{D} = \mathbf{Z}\hat{\pi}$
- **(c) Regress Y on $\hat{\mathbf{X}} = \left[1, X_1, X_2, \cdots, X_k, \hat{D}\right] : Y = \hat{\mathbf{X}}\beta_{2SLS} + u_3$**

Show formally that $\hat{\beta}_{2SLS} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}'Y = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'Y = \hat{\beta}_{IV}$. Comment on the steps along the way to reach your conclusion. If you need any additional assumptions, please state them.

- **Shorter Proof**: let's adopt the following preliminary assumptions/simplifications:
  - Let's make the following assumptions:
    - **P1**: $\hat{\mathbf{X}} = \mathbf{P}\mathbf{X}$ where $\mathbf{P} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top$.
    - **P2**: $\mathbf{P}$ is idempotent ($\mathbf{P}\mathbf{P} = \mathbf{P}$) and symmetric ($\mathbf{P} = \mathbf{P}^\top$).
  - Then the proof goes as follows:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \left(\hat{\mathbf{X}}^\top\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}^\top Y && \text{(Given)} \\
&= ((\mathbf{P}\mathbf{X})^\top\mathbf{P}\mathbf{X})^{-1}(\mathbf{P}\mathbf{X})^\top Y && \text{(By P1)} \\
&= (\mathbf{X}^\top\mathbf{P}^\top\mathbf{P}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{P}^\top Y && \text{(Distribute transpose)} \\
&= (\mathbf{X}^\top\mathbf{P}\mathbf{P}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{P}Y && \text{(By P2: symmetry)} \\
&= (\mathbf{X}^\top\mathbf{P}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{P}Y && \text{(By P2: idempotency)} \\
&= \left\{\mathbf{X}^\top\mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\mathbf{X}\right\}^{-1}\mathbf{X}^\top\mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top Y && \text{(By P1)} \\
&= \left(\mathbf{Z}^\top\mathbf{X}\right)^{-1}\left[\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\right]^{-1}\left(\mathbf{X}^\top\mathbf{Z}\right)^{-1}\mathbf{X}^\top\mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top Y && \text{(Distribute inverse)} \\
&= \left(\mathbf{Z}^\top\mathbf{X}\right)^{-1}\left(\mathbf{Z}^\top\mathbf{Z}\right)\left(\mathbf{X}^\top\mathbf{Z}\right)^{-1}\mathbf{X}^\top\mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top Y && \text{(Inverses cancel)} \\
&= \left(\mathbf{Z}^\top\mathbf{X}\right)^{-1}\mathbf{Z}^\top\mathbf{Z}\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\mathbf{Z}^\top Y && (\mathbf{X}^\top\mathbf{Z} \text{ cancels with inverse}) \\
&= \left(\mathbf{Z}^\top\mathbf{X}\right)^{-1}\mathbf{Z}^\top Y && (\mathbf{Z}^\top\mathbf{Z} \text{ cancels with inverse}) \\
&= \hat{\beta}_{\text{IV}} && \text{(As desired)}
\end{aligned}
$$

- **Comprehensive Proof**:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \left(\hat{\mathbf{X}}^\top \hat{\mathbf{X}}\right)^{-1} \hat{\mathbf{X}}^\top Y \\
&= \left(\begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_k \\ \hat{D} \end{bmatrix} \begin{bmatrix} 1, X_1, \ldots, X_k, \hat{D} \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_k \\ \hat{D} \end{bmatrix} Y \\
&= \left(\begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_k \\ \mathbf{Z}\hat{\pi} \end{bmatrix} [1, X_1, \ldots, X_k, \mathbf{Z}\hat{\pi}] \right)^{-1} \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_k \\ \mathbf{Z}\hat{\pi} \end{bmatrix} Y \quad \begin{pmatrix} \text{since} \\ \hat{D} = \mathbf{Z}\hat{\pi} \end{pmatrix} \\
&= \left(\begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_k \\ \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top D \end{bmatrix} \left[1, X_1, \ldots, X_n, \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} D\right] \right)^{-1} \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_k \\ \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top D \end{bmatrix} Y \begin{pmatrix} \text{By} \\ \text{definition} \\ \text{of } \hat{\pi} \end{pmatrix} \\
&= \left\{ \left(\mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top [1, X_1, \ldots, X_k, D]\right)^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top [1, X_1, \ldots, X_k, D] \right\}^{-1} \left\{ \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top [1, X_1, \ldots, X_k, D] \right\}^\top Y \\
&= \left\{ \left(\mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X}\right)^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X} \right\}^{-1} \left\{ \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X} \right\}^\top Y \\
&= \left\{ \mathbf{X}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X} \right\}^{-1} \mathbf{X}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top Y \\
&= \left\{ \mathbf{X}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{I}_n \mathbf{Z}^\top \mathbf{X} \right\}^{-1} \mathbf{X}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top Y \\
&= \left\{ \mathbf{X}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X} \right\}^{-1} \mathbf{X}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top Y \\
&= \left[\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X} \right]^{-1} \left(\mathbf{X}^\top \mathbf{Z}\right)^{-1} \mathbf{X}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top Y \\
&= \left[\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X} \right]^{-1} \mathbf{I}_n \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top Y \\
&= \left[\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top \mathbf{X} \right]^{-1} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top Y \\
&= \left(\mathbf{Z}^\top \mathbf{X}\right)^{-1} \left[\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \right]^{-1} \left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top Y \\
&= \left(\mathbf{Z}^\top \mathbf{X}\right)^{-1} \mathbf{Z}^\top \mathbf{Z}\left(\mathbf{Z}^\top \mathbf{Z}\right)^{-1} \mathbf{Z}^\top Y \\
&= \left(\mathbf{Z}^\top \mathbf{X}\right)^{-1} \mathbf{I}_n \mathbf{Z}^\top Y \\
&= \left(\mathbf{Z}^\top \mathbf{X}\right)^{-1} \mathbf{Z}^\top Y \\
&= \hat{\beta}_{\text{IV}} \quad \text{(As desired)}
\end{aligned}
$$

7.3. The Colonial Origins of Comparative Development.

In this problem we will assess one of the most famous social science articles using instrumental variables, Acemoglu, Johnson, & Robinson's 2001 paper "The Colonial Origins of Comparative

Development: An Empirical Investigation" (henceforth AJR).[1]

First, we will begin with a stylized characterization of the study. Assume that AJR use the following variables for any country i that was previously colonized:

- **Instrument $Z_i \in \{0, 1\}$: Mortality in the 17th, 18th, and early 19th centuries, 0 if low mortality, 1 if high.**
- **Treatment $D_i \in \{0, 1\}$: Modern property rights institutions, 0 if weak, 1 if strong.**
- **Outcome Y_i : Modern $\log$ GDP per capita.**

In our stylized characterization, assume that AJR use instrumental variables to estimate the effect of D_i on Y_i by instrumenting for D_i with Z_i. They find that having strong modern property rights institutions causes higher GDP per capita. (Note: As we will see in a minute, AJR include various specifications in which they also control for some pre-treatment covariates, but for now we will focus on the "simplest" empirical strategy.)

7.3.b. Name the five assumptions underpinning instrumental variables as a strategy for identifying the effect of an endogenous treatment on the outcome for compliers. Write out each assumption formally in terms of Z_i, D_i, and Y_i. In your own words, interpret each assumption with regard to the specific setup of A.JR's study. Finally, discuss the plausibility of each assumption. (Hint: It may be useful to refer to your DAG from (a) in interpreting and assessing some assumptions.)

- (1) **Stable Unit Treatment Value Assumption (SUTVA)**: for $D_i(z)$ and $Y_i(z, d)$. This means that the potential outcomes for each unit i are unrelated to the treatment status of other units $-i$. In substantive terms, there's no interference between units and there's no different versions of the treatment.
  - Formal Expression: $D_i(z) = D_i'(z) \Rightarrow Y_{(D_0, D_1, z)i} = Y_{(D_0', D_1', z)i}$ (i.e., stratified on Z_i).
  - Specific Interpretation: in this context, SUTVA means two things.
    - First, once we've conditioned on country i's mortality in the 17th, 18th, and 19th centuries (Z_i), then country i's modern log GDP per capita (Y_i) should be unrelated to the strength of other countries' property rights institutions (D_i).
    - Second, after conditioning on mortality in the 17th, 18th, and 19th centuries (Z_i) for all countries, treatment with each type of modern property rights institutions (D_i) should have roughly the same effect on every country.
  - Assumption Plausibility: It isn't clear to me that the SUTVA assumption is satisfied in this context. Settler mortality isn't the only factor that might explain differences in the effect of the treatment (property rights institutions) on a country's economic outcomes. Geographic trade openness (via factors like accessible rivers, ocean access, and proximity to wealthy trading partners) likely plays a significant role in how a country's property protections impact their economic success. Countries with greater geographic trade openness likely experience greater economic benefits from strong property rights institutions than countries with weaker geographic trade openness (who likely struggle to increase trade and investment regardless of the strength of their property rights institutions).
- (2) **Randomization of Encouragement**.
  - Formal Expression: $\{\forall y_i \in Y_i, D_{i0}, D_{i1}\} \perp\!\!\!\perp Z_i$.

[1] Thi paper has over 8,000 citations, and is heavily debated in a range of disciplines including economics, political science, and history. This problem set question is highly stylized, and if you are really interested in the substantive and methodological details of the paper we encourage you to read the paper and surrounding debates carefully. Also remember, it is always easier to criticise something than to build it yourself!

- Specific Interpretation: In this context, randomization of the encouragement means that values of the instrumental variable (high/low settler mortality) is randomly distributed across countries—or, more plausibly, that settler mortality was distributed exogenously (i.e., not under the colonizing countries' control).
- Assumption Plausibility: This assumption seems more likely to be plausible. Before the advent of proto-modern medicine, exogenous geographic factors were likely the dominant determinants of settler mortality rates.

- (3) **Exclusion Restriction**: The instrument Z_i affects the outcome Y_i only through the treatment D_i.

  - Formal Expression: $\forall d \in \{0,1\} : Y_{i1}(d) = Y_{i0}(d)$.
  - Specific Interpretation: In this context, the exclusion restriction entails that settler mortality only affects modern-day GDP per capita via its effect on property rights.
  - Assumption Plausibility: I'm not confident that the exclusion restriction is satisfied in this case. It's plausible that high settler mortality affects present-day GDP per capita through socio-cultural pathways. For example, settler mortality is likely to be higher in hot, humid climates. It's also going to be more painstakingly difficult and dangerous to carry out any outdoor manual labor in such conditions, and thus manual laborers in these countries are likely to be less productive than those in temperate climates. This added friction on economic growth over centuries will have an impact on modern GDP per capita that doesn't happen because of property rights protections.

- (4) **Monotonicity**: There are no defiers.

  - Formal Expression: $\forall i : D_{i1} \geq D_{i0}$
  - Specific Interpretation: In this case, we assume that there are no countries in which high settler mortality produced strong property rights and no countries where low settler mortality produced weak property rights.
  - Assumption Plausibility: This seems implausible if high-mortality countries had prior histories of strong property rights institutions before European colonization. In these cases, the pre-existing (native) property rights institutions may still have an effect in the modern day such that traditional norms/institutions lead to strong property rights in a region where European settlers faced a high mortality rate. This country would constitute a defier.

- (5) **Relevance**: there's nonzero average encouragement effect (i.e., the encouragement actually works).

  - Formal Expression: $\mathbb{E}\left[D_i(1) - D_i(0)\right] \neq 0$
  - Specific Interpretation: This means that the instrumental variable must have a strong effect on the treatment variable. In other words, low settler mortality lead to stronger property rights institutions (via colonization) and high settler mortality led to weaker property rights institutions.
  - Assumption Plausibility: This seems to be implausible given some of the scenarios I outlined above that would violate other assumptions. For example, what if there are high-mortality countries that had prior histories of strong property rights institutions before European colonization? Again, in this case, we would not expect settler mortality to strongly predict property right institutions.

Problem Set 8.

8.2. RDD Assumptions.

8.2.a. Write down and explain the identification assumption required for the type of RDD you implemented in Problem 1.

- Given $d \in \{0,1\} \cap x, c \in X_i : \mathbb{E}[Y_i(d)|X_i = x]$ remains continuous as $x \to c$ from both sides of $X_i = c$. i.e., unit i's characteristics do not abruptly change at the cutoff point c such that units that fall below the cutoff in reality would have the same treatment effect as those that fall above the outcome in the alternative world where they actually fall above the cutoff. In this context, that would mean that winning the election (i.e., $> 50\%$) would have the same effect on those candidates that barely lost the last election(i.e., that just barely fall below the threshold) are not different from incumbents that just barely fall above the threshold by winning the last election. It's the continuity of average potential outcomes assumption.

8.2.b. Contrast this assumption with the assumptions required for identification under selection on observables. Why do we require this new assumption?

- To do selection on observables, we assume that the treatment is as-if random after stratifying on the control variables. If we think about like a DAG, then the control variables close all the theoretically problematic back door paths between the treatment and outcome variables so that the observed relationship corresponds to the actual effect of X on Y. Controlling for these control variables makes the treatment status "as-if" random. We also assume for selection on observables that there's common support. These two selection on observables assumptions are used when treatment is not random. However, the "continuity of average potential outcomes" assumption assumes that the treatment is "as-if" random if its close enough to the cutoff point and there's no difference in the treated vs. untreated characteristics around the cutoff point. We require it to ensure there's no sorting (i.e., units just below the cutoff can self-select into being just above the cutoff).

Problem Set 9.

9.1. Card and Krueger.

Card and Krueger (1994) estimate the impact of changing the minimum wage on teenage employment. Conventional economic wisdom holds that raising the minimum wage reduces employment, especially among teenagers, who often earn the minimum wage. This is the main argument against raising it. Empirical analysis has failed to find evidence of an employment response to the minimum wage, however. In 1992, New Jersey's minimum wage increased from $4.25 to $5.05, while the minimum wage in Pennsylvania remained at $4.25. The authors use data on employment at fast-food establishments in New Jersey and Pennsylvania before and after the increase in the minimum wage to estimate the impact of the increase in minimum wage on teenage employment.

Download `card_krueger.csv` from the course website. It contains the following variables, with one row corresponding to one fast-food restaurant:

- **`state`: New Jersey or Pennsylvania**
- **`chain`: the fast-food chain to which the restaurant belongs**
- **`wage.pre`: starting wage in February 1992, in dollars per hour**
- **`wage.post`: starting wage in November 1992, in dollars per hour**
- **`emp.pre`: employment in February 1992, in number of full-time equivalent (FTE) employees**
- **`emp.post`: employment in November 1992, in number of full-time equivalent (FTE) employees**
- **`closed`: whether the store was closed in November 1992**

##$ 9.1.a. Assume that the fast-food restaurants surveyed by Card and Krueger represent a random sample from a larger population of all fast-food restaurants in New Jersey and Pennsylvania. Consider the quantities in Table 1, denoting the mean level of full-time equivalent (FTE) employment for restaurants by state and time. Define the article's difference-in-differences estimator in these terms. In your own words, what does the estimator represent?

	February	November
New Jersey	α	β
Pennsylvania	γ	δ

Table 1: Average population FTE employment by state and time

- First, calculate state-level first difference:
  - New Jersey first difference $D_{NJ} = \beta - \alpha$
  - Pennsylvania first difference: $D_{PE} = \delta - \gamma$
- Second, calculate difference in differences:

$$\begin{aligned} \text{DiD} &= D_{NJ} - D_{PE} \\ &= \beta - \alpha - (\delta - \gamma) \\ &= \beta + \gamma - \alpha - \delta \end{aligned}$$

- This difference-in-differences estimator is the difference in November and February's total teenage employment in New Jersey minus the difference in November and February's total teenage employment in Pennsylvania. New Jersey comes first because it receives the treatment (a minimum wage increase) while Pennsylvania is the control group as it did not experience a minimum wage increase.

9.1.b. Consider potential mean FTE employment among treated and untreated restaurants, pre-treatment and post-treatment. Denote treatment group membership with D_i for $d = \{0, 1\}$, index treatment status with $z = \{0, 1\}$ and index the period with $t = \{0, 1\}$. (For a unit in the treatment group, $D_i = 1$ in all periods and $z = D_i$ post-treatment.) Replace each element of Table 1 using expected potential outcomes $Y_{it}(z)$ conditional on $D_i = d$. Then, define the differences-in-differences estimator and the causal quantity of interest, the ATT, in these same terms. What assumption is necessary to identify the ATT using the differences-in-differences estimator?

- First, replace each element of table 1 using the expected potential outcomes conditional on treatment assignment:
  - $\alpha = \mathbb{E}[Y_{i0}(0)|D_i = 1]$
  - $\beta = \mathbb{E}[Y_{i1}(1)|D_i = 1]$
  - $\gamma = \mathbb{E}[Y_{i0}(0)|D_i = 0]$
  - $\delta = \mathbb{E}[Y_{i1}(0)|D_i = 0]$
- Second, calculate the difference in differences estimator for the causal quantity of interest:

$$\begin{aligned}\text{DiD} &= \beta + \gamma - \alpha - \delta \\ &= \mathbb{E}[Y_{i1}(1)|D_i = 1] + \mathbb{E}[Y_{i0}(0)|D_i = 0] - \mathbb{E}[Y_{i0}(0)|D_i = 1] - \mathbb{E}[Y_{i1}(0)|D_i = 0] \\ &= \tau_{ATT}\end{aligned}$$

- The **Parallel Trends Assumption** is the necessary assumption to identify the ATT using the difference-in-differences estimator.
  - This can be formally represented as:

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 0]$$

- In substantive terms, this is saying that New Jersey would've continued on the same trend in the counterfactual scenario where the minimum wage increase didn't happen.

9.2. Malesky et al.

Malesky et al. (2014) study the effect of recentralizing public service provision in Vietnam. The study investigates the effect of the piloted removal of elected local councils from 99 districts across the country in 2009 using a difference-in-differences design with repeated cross-sectional data. The authors find that recentralization improved public service delivery in the transportation, healthcare, and communications sectors. In this problem we will replicate and probe the validity of some of their key results.

First, download the two datasets `maleskyetal.dta` and `maleskyetal_placebo.dta` from the course website. Both datasets include the following variables for each commune i:

The main specifications used in the paper implement the following model:

$$Y_{it} = \alpha + T_t\beta + D_i\gamma + T_tD_i\theta + X_{it}^\top\delta + \epsilon_{it}$$

Where i indicates a commune and t a year. T_t is an indicator variable for the treatment period, 1 after 2009, 0 before. D_i is an indicator variable for the treatment group, 1 if in the treatment group, 0 if not. X_{it} are covariates (`lnarea`, `lnpopden`, `city`, and dummies for each value of `reg8`). The error terms in the model (ϵ_{it}) are correlated within districts (`tinh`) but not across districts.

9.2.a. Given the setup outlined above, create a three-by-three table (as in Slide 17 in the DID slides) that shows the meaning of each of $\alpha, \beta, \delta, \gamma$ and θ. Which parameter gives the difference-in-differences, the causal quantity of interest in Malesky et al's study? Explain in your own words what each "difference" represents in the differencein-differences.

	After ($T_i = 1$)	Before ($T_i = 0$)	After - Before
Treated $G_i = 1$	$\hat{\alpha} + \hat{\beta} + \hat{\gamma} + \hat{\theta} + \hat{\delta}$	$\hat{\alpha} + \hat{\beta} + \hat{\delta}$	$\hat{\gamma} + \hat{\theta}$
Control $G_i = 0$	$\hat{\alpha} + \hat{\gamma} + \hat{\delta}$	$\hat{\alpha} + \hat{\delta}$	$\hat{\gamma}$
Treated - Control	$\hat{\beta} + \hat{\theta}$	$\hat{\beta}$	$\hat{\theta}$

- As in slide 17, the parameter $\hat{\theta}$ provides the causal quantity of interest.
- Let's consider each difference in turn.
  - $\hat{\gamma} + \hat{\theta}$: results of subtracting the pre-treatment from post-treatment values of the treatment group
  - $\hat{\gamma}$: result of subtracting the pre-treatment from post-treatment values of the control group.
  - $\hat{\beta} + \hat{\theta}$: the results of subtracting the post-treatment outcomes of the treatment group from the control group.
  - $\hat{\beta}$: the results of subtracting the pre-treatment outcomes of the control group from the treated group.

9.2.c. Given SUTVA, write down the key assumption required to identify the causal effect of recentralization. Interpret the assumption. Give an example, in the context of Malesky et al's study, of a confounder that would violate this assumption.

- It's the **parallel trends assumption**, which can be formally represented as:

$$\mathbb{E}\left[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{i1}(0) - Y_{i0}(0) \mid D_i = 0\right]$$

- This means that the trend/value for the treatment group would remain unchanged in the counterfactual world where they didn't receive treatment. One confounding factor that would violate the parallel trends assumption is if the councilors in the cities that centralized were already more likely to improve these services/indices than councilors in uncentralized cities. This might be the case if cities that are rapidly growing are more likely to centralize AND also dramatically improve these services.