Problem Set 5

Alex, Micah, Scott, and David

04/05/2025

```
library(data.table)

library(sandwich)
library(lmtest)

library(AER)

library(ggplot2)
library(patchwork)
```

Vietnam Draft Lottery

Observational esteimate

Suppose that you had not run an experiment. Estimate the "effect" of each year of education on income as an observational researcher might, by just running a regression of years of education on income (in R-ish, `income ~ years_education`). What does this naive regression suggest?

```
model_observational <- lm(income ~ years_education, data = d)
summary(model_observational)
```

```
##
## Call:
## lm(formula = income ~ years_education, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -91655 -17459   -837  16346 141587
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -23354.64    1252.74  -18.64   <2e-16 ***
## years_education   5750.48      83.34   69.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26590 on 19565 degrees of freedom
## Multiple R-squared:  0.1957, Adjusted R-squared:  0.1957
## F-statistic:  4761 on 1 and 19565 DF,  p-value: < 2.2e-16
```

Answer: The naive regression suggests that each additional year of education is associated with an increase of $5,750.48 in income. However, this is an observational estimate and may be biased due to confounding factors.

Evaluating observational estimate

Continue to suppose that we did not run the experiment, but that we saw the result that you noted in part 1. Tell a concrete story about why you don't believe that observational result tells you anything causal.

Answer: This observational result may be confounded. For example, individuals who pursue more education may also come from higher socioeconomic backgrounds, possess stronger motivation, or have access to better opportunities. These unobserved factors could inflate the relationship between education and income, making it difficult to interpret the coefficient as causal.

Natural experiment effect on education

Now, let's get to using the natural experiment. Define "having a high-ranked draft number" as having a draft number between 1-80. For the remaining 285 days of the year, consider them having a "low-ranked" draft number). Create a variable in your dataset called `high_draft` that indicates whether each person has a high-ranked draft number or not. Using a regression, estimate the effect of having a high-ranked draft number on years of education obtained. Report the estimate and a correctly computed standard error. (*Hint: How is the assignment to having a draft number conducted? Does random assignment happen at the individual level? Or, at some higher level?)

```
d$high_draft <- ifelse(d$draft_number <= 80, 1, 0)

model_education <- lm(years_education ~ high_draft, data = d)
summary(model_education)
```

```
##
## Call:
## lm(formula = years_education ~ high_draft, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5601 -1.4343 -0.4343  1.5657  5.5657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.43431    0.01691  853.40   <2e-16 ***
## high_draft   2.12576    0.03790   56.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.117 on 19565 degrees of freedom
## Multiple R-squared:  0.1385, Adjusted R-squared:  0.1384
## F-statistic:  3145 on 1 and 19565 DF,  p-value: < 2.2e-16
```

Answer: Having a high draft number increases years of education by 2.13 years (Estimate: 2.12576, SE: 0.03790, $p < 2e\text{-}16$). This is a large and statistically significant effect, supporting the idea that people used education as a way to avoid the draft.

Natural experiment effect on income

Using linear regression, estimate the effect of having a high-ranked draft number on income. Report the estimate and the correct standard error.

```
model_income <- lm(income ~ high_draft, data = d)
summary(model_income)
```

```
##
```

```
## Call:
## lm(formula = income ~ high_draft, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -67399 -21140  -3002  18005 151306
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  60761.9      235.9  257.56   <2e-16 ***
## high_draft    6637.6      528.7   12.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29530 on 19565 degrees of freedom
## Multiple R-squared:  0.007992,   Adjusted R-squared:  0.007941
## F-statistic: 157.6 on 1 and 19565 DF,  p-value: < 2.2e-16
```

Answer: Having a high draft number increases income by $6,637.60 (Estimate: 6637.6, SE: 528.7, p < 2e-16). This is statistically significant and captures the total (reduced form) effect of draft status on income.

Instrumental variables estimate of education on income

Now, estimate the Instrumental Variables regression to estimate the effect of education on income. To do so, use `AER::ivreg`. After you evaluate your code, write a narrative description about what you learn.

```r
model_iv <- ivreg(income ~ years_education | high_draft, data = d)
summary(model_iv)
```

```
##
## Call:
## ivreg(formula = income ~ years_education | high_draft, data = d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -78140 -18762  -2145  16461 147217
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      15691.6     3416.4   4.593  4.4e-06 ***
## years_education   3122.4      229.6  13.601  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27260 on 19565 degrees of freedom
## Multiple R-Squared: 0.1548,  Adjusted R-squared: 0.1548
## Wald test:   185 on 1 and 19565 DF,  p-value: < 2.2e-16
```

Answer: The IV regression estimates that each additional year of education increases income by $3,122.40 (Estimate: 3122.4, SE: 229.6, p < 2e-16). This is a local average treatment effect (LATE) — it captures the causal effect of education for individuals whose schooling choices were affected by their draft number. This estimate is lower than the observational estimate, likely because it removes upward bias from confounding.

Evaluating the exclusion restriction

Give one reason this requirement might not be satisfied in this context. In what ways might having a high draft rank affect individuals' income **other** than nudging them to attend more school?

Answer: The exclusion restriction might not be satisfied if draft number affected income in ways unrelated to education. For instance, avoiding the draft might have allowed individuals to avoid physical/mental trauma, or enter the labor market earlier, both of which could directly influence income regardless of educational attainment.

Differential attrition

Conduct a test for the presence of differential attrition by treatment condition. That is, conduct a formal test of the hypothesis that the "high-ranked draft number" treatment has no effect on whether we observe a person's income. **(Note, that an earning of $0 *actually* means they didn't earn any money – i.e. earning $0 does not mean that their data wasn't measured. Let's be really, really specific: If you write a model that looks anything like, `lm(income == 0 ~ .)` you've gone the wrong direction.)**

```
model_differential_attrition <- lm(!is.na(income) ~ high_draft, data = d)
summary(model_differential_attrition)
```

```
##
## Call:
## lm(formula = !is.na(income) ~ high_draft, data = d)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.800e-17 -1.800e-17 -1.800e-17 -1.800e-17  2.775e-13
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  1.000e+00  1.585e-17  6.309e+16   <2e-16 ***
## high_draft  -1.771e-17  3.552e-17 -4.990e-01    0.618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.984e-15 on 19565 degrees of freedom
## Multiple R-squared:    0.5,  Adjusted R-squared:    0.5
## F-statistic: 1.957e+04 on 1 and 19565 DF,  p-value: < 2.2e-16
```

Answer: The estimate is $-1.77e{-}17$ with a p-value of 0.618, indicating no statistically significant difference in missingness (or presence) of income data between high and low draft groups. This suggests that differential attrition is not a concern in this dataset.

Evaluate differential attrition

Tell a concrete story about what could be leading to the result in part 7. How might this differential attrition create bias in the estimates of a causal effect?

Answer: Since there's no evidence of differential attrition, it's unlikely to bias our causal estimates. However, if attrition were present (e.g., low draft individuals missing income data due to military service or injury), it could lead to biased estimates by systematically excluding lower earners from the analysis.

Think about Treatment Effects

Throughout this course we have focused on the average treatment effect. *Why* we are concerned about the average treatment effect. What is the relationship between an ATE, and some individuals' potential outcomes? Make the strongest case you can for why this is a *good* measure.

The Average Treatment Effect (ATE) is a central concept in causal inference because it provides a summary measure of how a treatment affects the population on average. It is defined as the difference between the average potential outcome if everyone were treated and the average potential outcome if no one were treated:

ATE = E[Y(1)−Y(0)] Where Y(1) is the potential outcome under treatment and Y(0) is the potential outcome under control.

This concept is important because we never observe both outcomes for the same individual — we only see one or the other. But the ATE allows us to summarize the expected effect across the entire population, helping policymakers or researchers understand the overall benefit or harm of a treatment or policy.

The ATE is a good measure because: 1. It helps us decide whether a treatment should be scaled up. If the average effect is positive and meaningful, it suggests the intervention is beneficial at scale. 2. While treatment effects can vary across individuals, the ATE provides a clear, interpretable number that summarizes the overall impact. 3. When treatments are randomly assigned, the ATE can be estimated without bias, making it a powerful tool in experimental design. 4. Even when we care about treatment effect heterogeneity (e.g., by gender, income), the ATE serves as a starting point or baseline comparison.

In sum, the ATE connects the individual-level potential outcomes with population-level decisions. Even though it doesn't capture every nuance of individual responses, it offers a reliable, policy-relevant summary of treatment impact.

Optional Online advertising natural experiment.

Cross table of total_ads and treatment_ads

A. Run a crosstab – which in R is `table` – of `total_ads` and `treatment_ads` to sanity check that the distribution of impressions looks as it should. After you write your code, write a few narrative sentences about whether this distribution looks reasonable. Why does it look like this? (No computation required here, just a brief verbal response.)

```
cross_tab <- 'fill this in'
```

Answer: ...

Placebo test

A colleague of yours proposes to estimate the following model: `d[ , lm(week1 ~ tretment_ads)]` You are suspicious. Run a placebo test with `week0` purchases as the outcome and report the results. Since treatment is applied in week 1, and `week0` is purchases in week 0, *should* there be an relationship? Did the placebo test "succeed" or "fail"? Why do you say so?

```
model_colleague <- 'fill this in'
```

Answer: ...

What has gone wrong?

Here's the tip off: the placebo test suggests that there is something wrong with our experiment (i.e. the randomization isn't working) or our data analysis. We suggest looking for a problem with the data analysis. Do you see something that might be spoiling the "randomness" of the treatment variable? (Hint: it should be present in the cross-tab that you wrote in the first part of this qusetion.) How can you improve your analysis to address this problem? Why does the placebo test turn out the way it does? What one thing needs to be done to analyze the data correctly? Please provide a brief explanation of why, not just what needs to be done.

Answer: ...

Conduct proposed solution and re-evaluate placebo test

Implement the procedure you propose from part 3, run the placebo test for the Week 0 data again, and report the results. (This placebo test should pass; if it does not, re-evaluate your strategy before wasting time proceeding.) How can you tell this this has fixed the problem? Is it possible, even though this test now passes, that there is still some other problem?

```
model_passes_placebo <- 'fill this in'
```

Answer: ...

Estimate treatment effect with proposed solution

Now estimate the causal effect of each ad exposure on purchases during Week 1. You should use the same technique that passed the placebo test in part 4. Describe how, if at all, the treatment estimate that your model produces changes from the estimate that your colleague produced.

```
model_causal <- 'fill this in'
```

Answer: ...

Defend you method

Upon seeing these results, the colleague who proposed the specification that did not pass the placeo test challenges your results – they make the campaign look less successful! Write a short paragraph (i.e. 4-6 sentences) that argues for why your estimation strategy is better positioned to estimate a causal effect.

Answer: ...

Intertemporal substitution?

One concern raised by David Reiley is that advertisements might just shift *when* people purchase something – rather than increasing the total amount they purchase. Given the data that you have available to you, can you propose a method of evaluating this concern? Estimate the model that you propose, and describe your findings.

```
# Use the chunk to show your work
```

```
model_overall <- 'fill this in'
```

Answer: ...

Weekly effects

If you look at purchases in each week – one regression estimated for each outcome from week 1 through week 10 (that's 10 regression in a row) – what is the relationship between treatment ads and purchases in each of those weeks. This is now ranging into exploring data with models – how many have we run in this question alone!? – so consider whether a plot might help make whatever relationship exists more clear.

```
# write whatever you want to estiamte this
```

Answer: ...

Evaluating what is happening in the data

I. What might explain this pattern in your data. Stay curious when you're writing models! But, also be clear that we're fitting a **lot** of models and making up a theory/explanation after the fact.

Answer: ...

Evaluate whether there are non-linear relationships

We started by making the assumption that there was a linear relationship between the treatment ads and purchases. What other types of relationships might exist? After you propose at least two additional non-linear relationships, write a model that estimates these, and write a test for whether these non-linear effects you've proposed produce models that fit the data better than the linear model.

Answer: ...