Problem Set 5

Alex, Micah, Scott, and David

12/01/2021

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.0.5
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.0.5
```

```
library(lmtest)
```

```
## Warning: package 'zoo' was built under R version 4.0.5
```

```
library(AER)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.0.5
```

1. Vietnam Draft Lottery

Questions to Answer

A. 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 <- 'fill this in'
```

Narrative: ...

B. 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.

Narrative: ...

C. 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?)

```r
model_education <- 'fill this in'
```

Narrative: . . .

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

```r
model_income <- 'fill this in'
```

Narrative: . . .

E. 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 <- 'fill this in'
```

Narrative: . . .

F. 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?

Narrative: . . .

G. 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.)**

```r
model_differential_attrition <- 'fill this in'
```

Narrative: . . .

H. 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?

Narrative: . . .

2. 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.

3. Optional Online advertising natural experiment.

Questions to Answer

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'
```

Narrative: ...

B. 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'
```

Narrative: ...

C. 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: ...

D. 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'
```

Narrative: ...

E. 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'
```

Narrative: ...

F. 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: ...

G. 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
```

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

Narrative: . . .

H. 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.

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

Narrative: . . .

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: . . .

J. 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.

Narrative: . . .