001: CEFs, inference, simulation, *etc.*

EC 607

Due *before* midnight on Sunday, 09 May 2021

DUE Upload your answer on Canvas *before* midnight (PDT) on Sunday, 09 May 2021.

IMPORTANT Your submission should be a PDF that includes

1. your typed responses/answers to the problems (along with any figures/tables)
2. R code you used to generate your answers

Your answers must be **in your own words** (they should not be identical to anyone else's words).

It's fine if work with other people, but if it becomes clear that you are copying others' work, you will fail the course.

OBJECTIVE This problem set has three purposes: (1) reinforce the metrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality within econometrics.

Part 1/3: CEFs and regression

Let's start with generating data. We want a nonlinear CEF, define our data-generating process (DGP) as

$$y_i = 3 + \mathbb{I}(x_i < 5)\left(x_i^2 + 1\right) + \mathbb{I}(x_i \geq 5)\left(-0.25 * x^2 + 25\right) + u_i$$

where

- $\mathbb{I}(x)$ denotes an indicator function that takes a value of 1 whenever x is true.
- x_i is distributed as a continuous uniform random variable taking on values from $[0, 10]$. I'm going to round x_i to 1 decimal.
- u_i is a heteroskedastic disturbance that follows a normal distribution with mean zero and standard deviation $0.5 + |5 - x|$.

Notice that this DGP is really just two separate DGPs determined by whether x_i is above or below 5 (plus the disturbance u_i).

01. Time to generate data. Given this is the first problem of your first problem set, I'll give you some code (for free).

```
# Load packages
library(pacman)
p_load(tidyverse, estimatr, huxtable, magrittr, here)
# Set a seed
set.seed(12345)
# Set sample size to 1,000
n = 1e3
# Generate data
dgp_df = tibble(
  x = runif(n = n, min = 0, max = 10) %>% round(1),
  u = rnorm(n = n, mean = 0, sd = 0.5 + abs(5 - x)),
  y = (x < 5) * (x^2 + 1) + (x >= 5) * (-0.25 * x^2 + 25) + u
)
# Summarize the dataset
dgp_df %>% summary()
```

```
#>        x                u                  y
#>  Min.   : 0.00   Min.   :-15.340   Min.   :-13.53
#>  1st Qu.: 2.70   1st Qu.: -1.637   1st Qu.:  4.75
#>  Median : 5.20   Median : -0.024   Median : 10.25
#>  Mean   : 5.14   Mean   : -0.084   Mean   :  9.93
#>  3rd Qu.: 7.60   3rd Qu.:  1.554   3rd Qu.: 15.62
#>  Max.   :10.00   Max.   : 15.159   Max.   : 25.42
```

Run this code. Your output should match my output (and you should understand what's going on).

02. Create a scatter plot of your dataset (*e.g.*, using `geom_point` from `ggplot2`).

03. Derive the CEF and add it to your scatter plot.

Hint: Keep in mind the definition of the CEF (the expected value of y given x).

Hint: You can plot a function in `ggplot2` using `stat_function`.

04. Regress y on x. Calculate standard errors assuming **homoskedasticity**. Report your results.

05. Do heteroskedasticity-robust standard errors "matter" here? Why? Explain your reasoning.

06. Add your regression line to your scatter plot. You can do this in `ggplot2` using `geom_abline()` and `geom_smooth()` (among other options).

07. For each of our values of x ([0, 10] rounded to one decimal), calculate the sample mean of y conditional on x and the number of observations for each x.

Now run a regression using this sample-based CEF: Regress the conditional mean of $y \mid x$ on x, weighting by the number of observations. Do your results from this CEF regression match your results in **04**? Should they for this sample? Comment on the point estimates and the standard errors—and explain why each should or should not match.

Hint: You can use the `weights` argument in `lm()` and `lm_robust()` to run a weighted regression.

08. Does OLS provide a decent linear approximation to the CEF in this setting? Under what conditions would this linear approximation of the CEF be helpful? Under what conditions would it be less helpful?

Part 2/3: Inference and simulation

Now it's time for a good, old-fashioned simulation.

Now imagine you're working on a project, and it occurs to you that

1. You have a pretty small sample size (but could spend a lot of money to get bigger n).
2. It's unlikely that your disturbance is actually normally distributed.
3. You might have an endogenous treatment D_i but have a sense of how treatment comes about.

Given that the small-sample properties of OLS generally use *well-behaved disturbance* and the large-sample properties are, by definition, for **big** n, you are wondering how well OLS is going to perform. Plus, you are really concerned about the endogenous treatment but optimistic that you know how the treatment is endogenous. Can we recover the *true* treatment effect?

This is the perfect scenario for a simulation.

I'll walk you through some of the steps of the simulation. But you have to write your own code.

Let's start by defining the DGP (using notation from class)

$$\begin{aligned}
\text{Y}_{0i} &= X_i + u_i \\
\text{Y}_{1i} &= \text{Y}_{0i} + W_i + v_i \\
\text{D}_i &= \mathbb{I}(X_i + \varepsilon_i > 10) \\
\text{Y}_i &= \text{Y}_{0i} + \text{D}_i \tau_i
\end{aligned}$$

where

- $X_i \sim$ Normal with mean 10 and standard devation 3
- $W_i \sim$ Normal with mean 3 and standard devation 2
- $u_i \sim$ Uniform $\in [-10, 10]$
- $v_i \sim$ Uniform $\in [-5, 5]$
- $\varepsilon_i \sim$ Uniform $\in [-1, 1]$

10. Derive an expression for τ_i (individual i's treatment effect).

11. What assumptions does the expression for the treatment effect in **10** depend upon?

12. Based upon **10**, what is the average treatment effect in this population? (Your answer should be a number.)

13. If we regress Y_i on D_i should we expect to recover the average causal effect of treatment (D_i)? Explain.

14. Would conditioning on X and/or W help the regression in **13**? Explain.

15. Now back to R: Write some R code that generates a 1,000-observation sample from the DGP.

16. For your sample, what is the correlation between Y_{0i} and D_i? What about Y_{1i} and D_i? What do these correlations tell you?

17. Using your sample, calculate the average treatment effect (ATE), the average treatment effect on the treated (TOT or ATT), and the average treatment effect for the untreated. Why do these quantities differ?

18. Run four regressions:

1. Regress Y_i on D_i
2. Regress Y_i on D_i and X_i
3. Regress Y_i on D_i and W_i
4. Regress Y_i on D_i, X_i, and W_i

Do the results of these regressions match your expectation for recovering the ATE or ATT? Explain.

19. Now wrap your code from **15** and **18** into a function. This function will be a single iteration of the simulation. The function should output the estimated treatment effect in each of the four regressions in **18**.

Hint 1: Help your future self by writing this function so that you can easily change the sample size.

Hint 2: Use `tidy()` from the `broom` package to easily convert regression results into a data frame.

Hint 3: Label the output of the four regressions so that you can distinguish between each specification.

20. Run a simulation with at least 500 iterations. Each iteration should

- take a new **15-observation** sample from our DGP
- output **four treatment-effect estimates** (one for each regression in **18**)
- output **four standard errors** (one for each estimate)

Summarize your results with a figure (*e.g.*, `geom_density()`) and/or a table.

Hints: The `apply()` family (*e.g.*, `lapply()`) works well for tasks like this, as does the `map` family from the `purrr` package (see the `future_map` family from the `furrr` package for parallelization). Also: The notes from class.

21. Are any of the estimation strategies (the four regressions) providing *reasonable* estimates of the average treatment effect?

22. With 15 observations, do you think think you have enough power to *detect* a treatment effect? Explain.

23. Increase the sample size to 1,000 observations per sample and repeat the simulation (including graphical/table summary). Does anything important change for causal estimates (*e.g.*, centers of the distributions) or inference (*e.g.*, rejection rates)?

24. Would getting even bigger data help the regressions that appear to be biased? *Related:* Is it worth paying for a bigger sample in this setting? Explain.

25. Should we control for W_i? Explain.

26. Draw the DAG for this DGP. What are the pathways from D to Y? How do you close the open pathways to get to the causal effect of D on Y?

Hint: Check out the ggdag package for drawing DAGs in R.

Part 3/3: Function time

27. Write your own function(s) that (1) produce the OLS-based coefficients for a regression and (2) produce the homoskedasticity-based standard errors for the coefficients. Confirm that your function is "working" by using your function to re-estimate the regression you ran in question **04** above.

You should be able to do most of this by converting your dataset to matrices (`as.matrix()` or `matrix()`) and then applying a little matrix math. In R, `%*%` is matrix multiplication, `solve()` produces the inverse of an invertible matrix, `crossprod()` calculates cross products, and `diag()` allows you to define a diagonal matrix or access the diagonal of an existing matrix.

Part 4/3: Bonus!

B01. Does anything important change if $\text{D}_i = \mathbb{I}(X_i + W_i + \varepsilon_i > 13)$?

B02. Repeat the simulation steps—but use a Normal distribution for u, v, and ε (try to match the mean and variance). What changes (now that we're using a very well-behaved distribution)?

B03. Repeat the simulation steps—but use a very poorly behaved distribution for u, v, and ε (try to match the mean and variance, if they are defined). What changes?

B04. When we regress Y_i on D_i (and potentially controls), are we estimating the ATE or the ATT?