Homework 5

STUDENT NAME

Due 8/3 11:59pm

NOTE: If you would like to answer the written portions using RMarkdown, simply type your answers below the questions. If you would like to print out the pdf to do the written portions, you should add `\vspace{50mm}` to add space in the knitted pdf for written work (adjust the 50mm as necessary). Alternatively, you can do the written portions on a separate paper.

Part 1: Theory

1

Suppose you collect data from people in a single city over three years, recording how many times they got sick that year and average time spent exercising each day. You collect data from 100 people, with 3 periods of data. For 25% of the people, data on the 2nd year is missing; for everyone else, there is data on all 3 periods.

1.1

Is this a balanced panel or an unbalanced panel? Explain the difference.

1.2

If you were to estimate a first-differences regression with this data, how many observations would you have? Show your work.

1.3

If you were to estimate a two-way fixed effects regression (with fixed effects for individuals and periods), how many observations would you have? Show your work.

1.4

If you were to implement the two-way fixed effects regression by using dummy variables (and the model included a constant term + the average exercise variable), how many degrees of freedom would that regression have?

1.5

Could the weather (which could prevent people from exercising and make them more likely to get sick) cause an omitted variable bias problem in the two-way fixed effects regression? Why or why not?

1.6

Could the individual's diet cause an omitted variable bias problem in the two-way fixed effects regression? Why or why not?

1.7

Could a pre-existing health condition cause an omitted variable bias problem in the two-way fixed effects regression? Why or why not?

2

You want to know if the construction of new housing in a downtown neighborhood raises or lowers the price of rent in the surrounding area. A naive person might try to answer this by estimating the regression $AverageNeighborhoodRent = \beta_0 + \beta_1 NumberofNewUnitsPastYear + u$, which would be biased and capture reverse causality (neighborhoods with rising rents are more attractive for developers). You decide to resolve this using an instrument, the number of building fires in the neighborhood that year. Parcels with burned buildings are more likely to be rebuilt into new developments than parcels with functional existing structures, so this instrument is relevant.

2.1

Name a potential way that this instrument might violate the exclusion restriction.

2.2

Assume for now that this instrument *does* satisfy the exclusion restriction, but is only weakly relevant, say $\text{Corr}(z, x) = 0.02$. *Under the assumption that the exclusion restriction holds perfectly*, what is the consequence of this in terms of a) bias and b) variance of the parameter estimates?

2.3

Now assume that the exclusion restriction is *close* to holding, i.e. $\text{Corr}(z, u)$ is *close* to zero but not exactly zero. Would you expect this to cause a large or a small bias? Why?

2.4

Write the first-stage and second-stage for the 2SLS regression. Label which is which.

2.5

Write the regression and hypothesis test you would use in order to test whether the original “naive” regression had an endogeneity problem.

Part 2: Application

For these problems, you will be using pre-processed data from my paper on the recreational value of rare species. The data describe the number of visits made by birdwatchers to various United States Forest Services ranger districts in the Intermountain West over 2008-2021. In 2017, a new bird species unique to the Minidoka

Ranger District was officially recognized. My paper (and this portion of the assignment) attempts to detect the causal effect of this recognition on visits to the Minidoka Ranger District.

```
library(tidyverse)
```

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

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

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

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

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

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

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

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(fixest)
```

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

```
library(readxl)
data <- read_csv("rd4_visits.csv")
```

```
## Rows: 896 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): district_name, forest_name
## dbl (3): year, visits, district_num
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1

Generate a new variable for defining the treated unit (1 for Minidoka Ranger District, 0 for all other ranger districts), a new variable for the post-treatment period (1 for all years $\geq$ 2017), and their interaction.

2

Plot a line plot for *only the pre-treatment period* showing visits over time for each ranger district. Make sure the Minidoka Ranger District stands out somehow (e.g. by making it a different color or line size) to make it easy to compare its trajectory to the rest.

3

Based on your plot, though there are a few districts with noisy visitation, there should be two ranger districts whose pre-period trends are clearly and distinctly not parallel with the rest. Name these districts and remove them from the data.

4

Estimate the basic form of the difference in differences regression (using indicators for $Post$ and $Treat$ rather than fixed effects).

5

Explain (in words, not math) what the parallel trends assumption would *mean* in this context. I.e. what has to be true about the visitation to these ranger districts for the parameter estimate on $Treat \times Post$ in the previous question to be interpreted causally?

6

Interpret (including sign, significance, and size) the parameter estimate that corresponds to the causal treatment effect.

7

Estimate the model again, this time using two-way fixed effects (the parameter estimate of interest will be the same in this case).

8

Conduct the following placebo test: Filter the actually-treated unit out of the dataset, which should leave you with 59 ranger districts. Write code that will save the coefficients from 59 different regression estimations, each of which uses a placebo treatment for a different control group that starts in 2017, just like the real treated unit. Plot a histogram of these coefficients. How does the parameter estimate from the actually treated group compare?