Assigment - 2

DATA - Background

Here we want to study the length of time (in weeks) that an injured worker receives workers' compensation. On July, 1980, a state (Ky) raised the cap on weekly earnings that were covered by workers' compensation. An increase in the cap has no effect on the benefit for low-income workers, but it makes it less costly for a high-income worker to stay on workers' comp. Therefore, the control group is low-income workers, and the treatment group is high-income workers; high-income workers are defined as those for whom the pre-policy-change cap on benefits is binding. The data comprises of random samples both before and after the policy change, and thus we want to test whether more generous workers' compensation causes people to stay out of work longer (everything else fixed). In other words, we want to know if this new policy caused workers to spend more time unemployed. If benefits are not generous enough, then workers could sue companies for on-the-job injuries, while overly generous benefits could cause moral hazard issues and induce workers to be more reckless on the job, or to claim that off-the-job injuries were incurred while at work. The policy was designed so that the cap increase did not affect low-earnings workers, but did affect high-earnings workers, so we use low-earnings workers as our control group and high-earnings workers as our treatment group.

The main outcome variable we care about is duration (in weeks) of worker's compensation benefits. You may want take a log (and create log_duration), it because the variable is fairly skewed—most people are unemployed for a few weeks, with some unemployed for a long time.

Variables in the dataset:

A data.frame with 7150 observations on 30 variables:

durat: duration of benefits
afchnge: =1 if after change in benefits (1980)
highearn: =1 if high earner
male: =1 if male
married: =1 if married
hosp: =1 if inj. required hosp. stay
indust: industry
injtype: type of injury
age: age at time of injury
prewage: previous weekly wage, 1982 $
totmed: total med. costs, 1982 $
injdes: 4 digit injury description
benefit: real dollar value of benefit

ky: =1 for kentucky (= 0 for other state)
mi: =1 for michigan
ldurat: log(durat)
afhigh: afchnge*highearn
lprewage: log(wage)
lage: log(age)
ltotmed: log(totmed); = 0 if totmed < 1
head: =1 if head injury
neck: =1 if neck injury
upextr: =1 if upper extremities injury
trunk: =1 if trunk injury
lowback: =1 if lower back injury
lowextr: =1 if lower extremities injury
occdis: =1 if occupational disease
manuf: =1 if manufacturing industry
construc: =1 if construction industry
highlpre: highearn*lprewage

Questions: (you may consult Mixtape for codes)

1. Calculate the policy effect on duration, by hand without running any regression.
2. Calculate the policy effect on log_duration, by hand without running any regression.

   Note: For the above two questions, you need to create several variables -- before_treatment, after_treatment, before_control, after_control etc. Then calculate diff-n-diff estimate.

3. 3. Now use regression analysis to calculate the above estimates again, without any control.
4. Next, make use some controls. Think about it which ones you want to use. And then use regression adujstment procedure to evaluate the impact of policy change on the duration, and log_duration.

   Note: Clearly write your regression equations. Don't forget interprete your results properly. You may use model summary to tabulate your results from Q.3 and Q.4

5. Now perform a robustness check of your results for Q.4 above, using information from other states (i.e. Ky=0).
6. Bonus Q: It would be great if you can generate a graph like NJ-PA example discussed in the class. If you cannot no problem.

Some coding help below

Important: indust and injtype are in the dataset as numbers (1-3 and 1-8), but they’re actually categories. We have to tell R to treat them as categories (or factors), otherwise it’ll assume that you can have an injury type of 3.46 or something impossible. You can use "mutate" command as follows -

```
# Convert industry and injury type to categories/factors

mydata_fixed <- mydata %>%
  mutate(indust = as.factor(indust),
         injtype = as.factor(injtype))


library(tidyverse)     # ggplot(), %>%, mutate(), and friends
library(scales)        # Format numbers with functions like comma(), percent(), and dollar()
library(broom)         # Convert models to data frames
library(modelsummary) # Side-by-side regression tables
library(foreign)    # for importing stata data


# Load data

injury <- read_csv("all_data/injury.csv")
```

Clean the data a little so that it only includes rows from Kentucky (ky == 1), and rename some of the variables to make them easier to work with:

```
mydata <- injury %>%
  filter(ky == 1) %>%
  # The syntax for rename = new_name = original_name
  rename(duration = durat, log_duration = ldurat,
         after_1980 = afchnge)
```

Exploratory data analysis

First we can look at the distribution of unemployment benefits across high and low earners (our control and treatment groups):

```
ggplot(data = mydata, aes(x = duration)) +
  # binwidth = 8 makes each column represent 2 months (8 weeks)
  # boundary = 0 make it so the 0-8 bar starts at 0 and isn't -4 to 4
  geom_histogram(binwidth = 8, color = "white", boundary = 0) +
  facet_wrap(~ highearn)
```

The distribution is really skewed, with most people in both groups getting between 0-8 weeks of benefits (and a handful with more than 180 weeks! that’s 3.5 years!)

If we use the log of duration, we can get a less skewed distribution that works better with regression models:

```
ggplot(data = mydata, mapping = aes(x = log_duration)) +
  geom_histogram(binwidth = 0.5, color = "white", boundary = 0) +
  # Uncomment this line if you want to exponentiate the logged values on the
  # x-axis. Instead of showing 1, 2, 3, etc., it'll show e^1, e^2, e^3, etc. and
  # make the labels more human readable
  # scale_x_continuous(labels = trans_format("exp", format = round)) +
  facet_wrap(~ highearn)
```