Question 1 - Econ Final Project

Ben Carlson

5/24/2022

```
knitr::opts_chunk$set(echo = TRUE)
library("dplyr")
library("readr")
library("readxl")
library("purrr")
library('tidyverse')
library('lubridate')
library(ggplot2)
library(vtable)
library(fixest)
library(ipumsr)
library(stringr)
library(rdrobust)
library(jtools)
```

Read in Data

```
ddi <- read_ipums_ddi("cps_00002.xml")
data <- read_ipums_micro(ddi)
```

```
## Use of data from IPUMS CPS is subject to conditions including that users should
## cite the data appropriately. Use command `ipums_conditions()` for more details.
```

Filter only the retail industry

```
data2 <- data %>% filter(IND1990 %in% c("580","581","582","591","601","610","611",
                                        "612","620","621","622","623","630","631","632", "633","640","641","642","650","651"
,"652","660","661","662","663","670","671","672","681","682","691"))
```

Count the occurances of each industry for every year and month

This will tell us how many people are employed in retail industry jobs for each month

```
data3 <- data2 %>% group_by(YEAR,MONTH,IND1990) %>% count(IND1990)
```

Rename column

```
colnames(data3)[colnames(data3) == 'n'] <- 'MONTHEMPLY'
```

Combine all the counts for each year and month

This will tell us the total retail employees for each month

```
data4 <- data3 %>% group_by(YEAR,MONTH) %>% summarize(TOTALMONTHLYRETAILEMPLY = sum(MONTHEMPLY))
```

```
## `summarise()` has grouped output by 'YEAR'. You can override using the
## `.groups` argument.
```

Formate Month and Year into a Date

```
data4$Date <- str_c(data4$YEAR,'-' ,data4$MONTH)
vtable(data4)
```

data4

Name	Class	Label	Values
YEAR	numeric	Survey year	Num: 2014 to 2022
MONTH	haven_labelled	Month	Num: 1 to 12
TOTALMONTHLYRETAILEMPLY	integer	NULL	Num: 7206 to 11676
Date	character	NULL	

```
data4 <- data4 %>% mutate(Date = ym(Date))
```

Make true or false for if the date was after Covid

Make a new column that makes date a numeric value, this will make it possibe to run a regression discontinuity

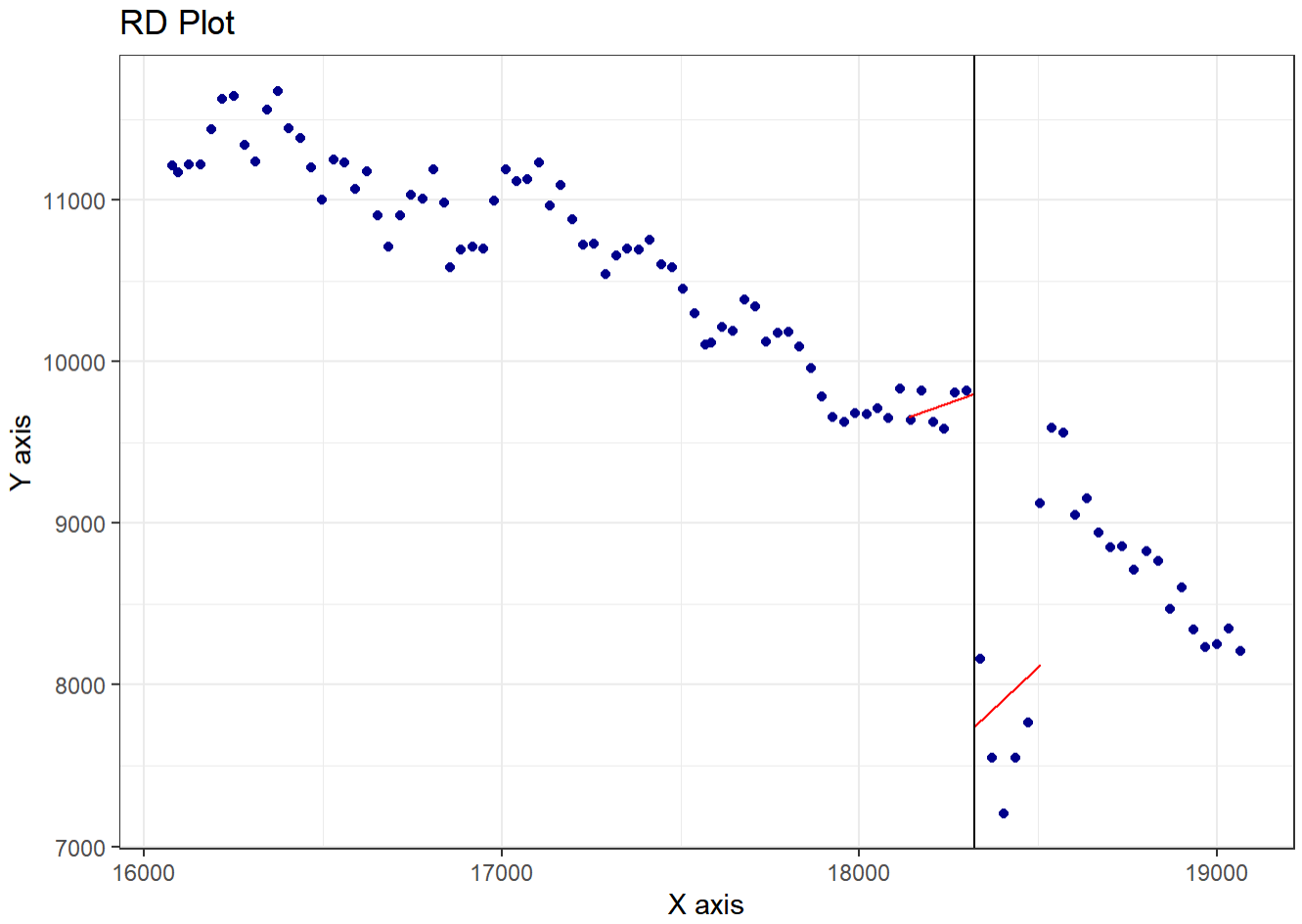
```
data4 <- data4 %>% mutate(AfterCovid = Date >= ym('2020-3'))
data4 <-data4 %>% mutate(date_number = as.numeric(Date))
```

Let's Run a Interrupted Time series design for the effect of Covid, and graph it, we will have March 2020 be the cutoff,

which is represented by the number 18322, and we will see the effect with a six month window on both sides of the cutoff.

(Interrupted time series is the same format in rdrobust, just using time as a cutoff)

```
rdo1 <- rdrobust(data4$TOTALMONTHLYRETAILEMPLY, data4$date_number, c = 18322, p = 1, h = 184, kernel = "uniform")
rdplot(data4$TOTALMONTHLYRETAILEMPLY, data4$date_number, c = 18322, p = 1, h = 184, kernel = "uniform")
```




```
summary(rdo1)
```

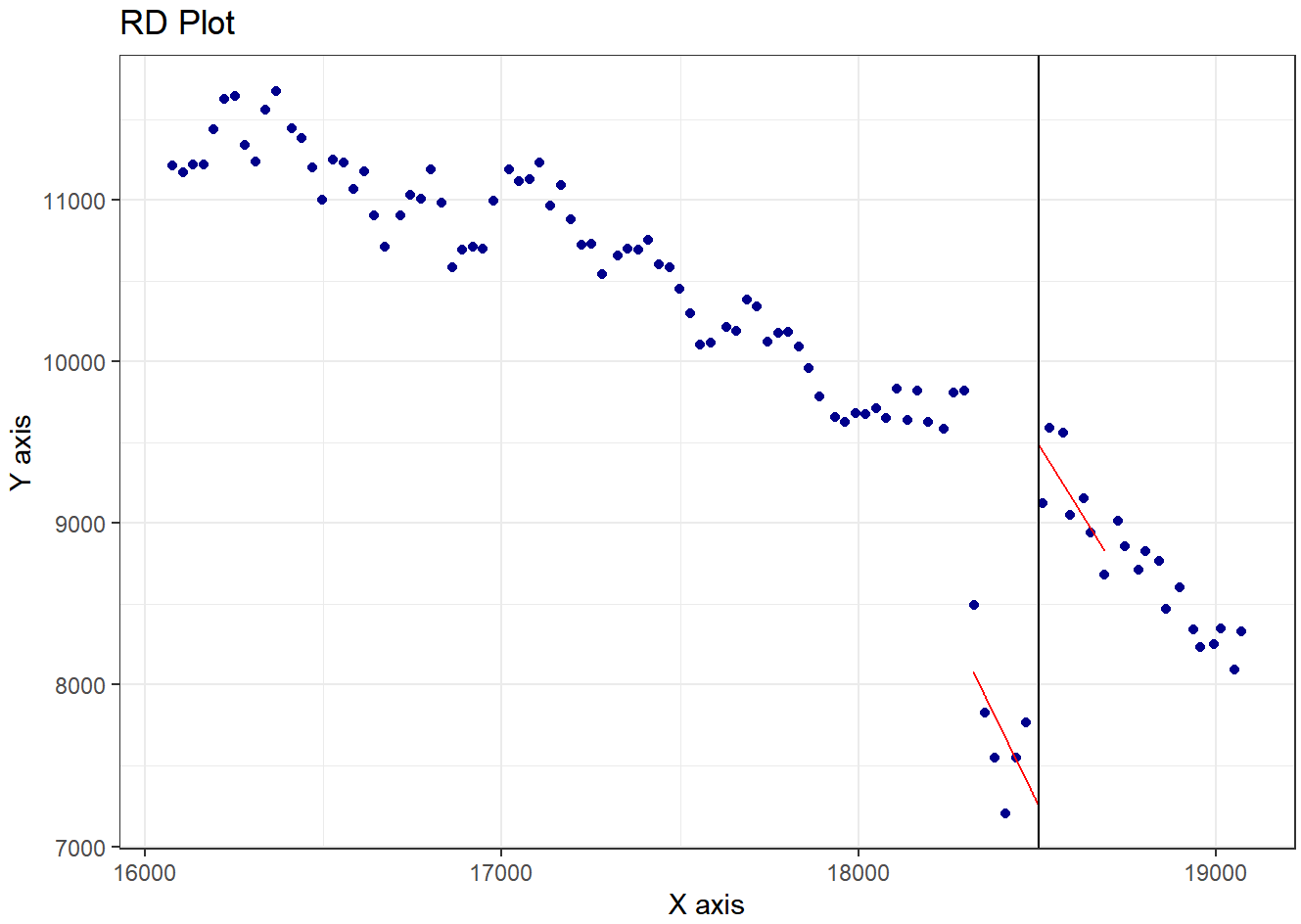
```
## Call: rdrobust
##
## Number of Obs.                  100
## BW type                      Manual
## Kernel                      Uniform
## VCE method                       NN
##
## Number of Obs.                   74           26
## Eff. Number of Obs.               6            7
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                 184.000      184.000
## BW bias (b)                 184.000      184.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                      74           26
##
## ===============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ===============================================================================
##   Conventional -2059.284   482.059    -4.272     0.000 [-3004.102 , -1114.465]
##         Robust         -         -    -1.962     0.050 [-2720.201 , -1.100]
## ===============================================================================
```

Analysis:

We see some really interesting results from the graph and the following summary table. In the table we see that on average that after the cutoff employment fell by 2,059 employees for the six months after compared to the six months before. Our graph gives us a good visualization of the change before and after the cutoff. Our Y Axis is the number of employees for each month, and the X-axis is the numeric representation of those months. We see that from February to March, Employment drastically fell by more than 1,000, and fell even more in April and May, it's lowest losing more than 2,500 employees from February.

However what I find the most fascinating is the quick recovery that seems to happen in September. This huge jump calls for doing another interrupted time series design, but this time for the cutoff of employment increasing, by doing so we will see the impacts of reopening or the reduction of restrictions in regards to Covid. # # # #### Let's Run the same design but this time for the jump in September after Covid.

```
rdo2 <- rdrobust(data4$TOTALMONTHLYRETAILEMPLY, data4$date_number, c = 18506, p = 1, h = 184 , kernel = "uniform")
rdplot(data4$TOTALMONTHLYRETAILEMPLY, data4$date_number, c = 18506, p = 1, h = 184, kernel = "uniform")
```




```
summary(rdo2)
```

```
## Call: rdrobust
##
## Number of Obs.                  100
## BW type                      Manual
## Kernel                      Uniform
## VCE method                       NN
##
## Number of Obs.                   80           20
## Eff. Number of Obs.               6            7
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                 184.000      184.000
## BW bias (b)                 184.000      184.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                      80           20
##
## ===============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ===============================================================================
##   Conventional  2233.426   405.421     5.509     0.000  [1438.815 , 3028.038]
##         Robust         -         -     1.289     0.197  [-455.259 , 2203.709]
## ===============================================================================
```

Analysis:

Now we see that the Coefficient in this model has changed to be positive, as there is now an increase employees. However in this case the increase is 2233.43 employees for the six months, this means that jump up back to pre-covid employment was stronger than the initial decrease caused by COVID. However observing the graph I notice that the slope of the monthly employees seems to be trending in more steep negative direction than it was before covid, lets investigate.

Let’s see if Covid has made impacted the regression line.

lin4 represents a year and a half before covid (March)

lin5 represents a year and a half after covid employment recovered (September)

```
lin4 <- data4[c(56:74),] %>% mutate(date_number1 = date_number - 17743)
lin5 <- data4[c(81:99),]
```

Lets run two simple feols regressions

```
reg1 <- feols(TOTALMONTHLYRETAILEMPLY ~ date_number, data = lin4)
reg2 <- feols(TOTALMONTHLYRETAILEMPLY ~ date_number, data = lin5)
```

Compare both models

```
export_summs(reg1,reg2)
```

	Model 1	Model 2
(Intercept)	23669.20 ***	51965.93 ***
	(3844.42)	(4294.25)
date_number	-0.77 **	-2.30 ***
	(0.21)	(0.23)
nobs	19	19
r.squared	0.43	0.86
adj.r.squared	0.40	0.85
within.r.squared		
pseudo.r.squared		
sigma	155.04	166.03
nobs.1	19.00	19.00
AIC	247.47	250.07
BIC	249.36	251.96
logLik	-121.73	-123.03

*** p < 0.001; ** p < 0.01; * p < 0.05.

Analysis:

For my final Analysis we see that theses two models are very different. The date number coefficient is what is most important, as that tells us on average how much employment is changing per day. Before Covid, we see the downtrend that we were observing before in the graph, with on average .77 people leaving the retail industry each day. However after Covid, this number increased almost three fold, with on average 2.3 people leaving per day after employment recovered in September.

Overall looking over all of the analysis I conducted, a clear picture has been made in regards to Covid's impact on retail employment. Initially the impacts of Covid caused a huge decrease in employment, however within the short term of six months Employment levels bounced back to pre-covid levels within one month. However in the long term of a year and a half, we see that the rate of people leaving the retail industry has increased almost by 300%.

Assumptions:

There are plenty of assumptions that we in the group are making with out analysis, the first one is that the effects of Covid started on March 2020, and that Covid caused the huge decrease of employment in March. However we know that it was around mid march when lockdown started to happen, which in turn lead to people not being able to work, many in the retail industry. We are also assuming that September was when covid restrictions where lifted to allow people back to work in retail. This is hard to prove as so may different states had their own rules. However the huge outlier of the significant jump in our data from August to July provides evidence that this would be the month to choose to do a R.D of the impact of the 'end of covid'.

Question 2 Regression

Joy Kung

2022-06-02

```
library(vtable)
```

```
## Loading required package: kableExtra
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter()     masks stats::filter()
## ✕ dplyr::group_rows() masks kableExtra::group_rows()
## ✕ dplyr::lag()        masks stats::lag()
```

```
library(jtools)
library(fixest)
library(readxl)
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(purrr)
```

Load the clean data

```
industrydata4 <-read_csv("~/Desktop/R-ECONO/industrydata4.csv")
```

```
## New names:
## Rows: 200 Columns: 5
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: "," chr
## (1): industry2 dbl (2): ...1, avg_monthly_emp lgl (1): after_covid date (1):
## date
## 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`
```

```
regression_data <- rename(industrydata4, Industry = industry2)
show(regression_data)
```

```
## # A tibble: 200 × 5
##     ...1 date       after_covid Industry     avg_monthly_emp
##    <dbl> <date>     <lgl>       <chr>                  <dbl>
##  1     1 2014-01-01 FALSE       Others                 5415.
##  2     2 2014-01-01 FALSE       Retail Trade          11213
##  3     3 2014-02-01 FALSE       Others                 5402.
##  4     4 2014-02-01 FALSE       Retail Trade          11169
##  5     5 2014-03-01 FALSE       Others                 5334.
##  6     6 2014-03-01 FALSE       Retail Trade          11217
##  7     7 2014-04-01 FALSE       Others                 5336.
##  8     8 2014-04-01 FALSE       Retail Trade          11221
##  9     9 2014-05-01 FALSE       Others                 5392.
## 10    10 2014-05-01 FALSE       Retail Trade          11438
## # … with 190 more rows
```

Question2: How does retail do compare with other industries?

Categorical Regression:

Retails Industries vs. Other Industries Before and After COVID

In our model, other industries include Entertainment and Recreational Services, Finance, Insurance, Real Estate, Manufacturing, Personal Services, Professional and Related Services,Public Administration, Wholesale Trade. As retails trade industry and other industries are categorical variable, we decided to use categorical regression for the question.

```
regression1<-feols(avg_monthly_emp~Industry*after_covid, data = regression_data)

etable(regression1)
```

```
##                                          regression1
## Dependent Var.:                        avg_monthly_emp
##
## (Intercept)                          5,175.9*** (52.08)
## IndustryRetailTrade                  5,468.0*** (73.65)
## after_covidTRUE                       -958.4*** (102.1)
## IndustryRetailTrade x after_covidTRUE -1,172.5*** (144.4)
## ____________________________________ ___________________
## S.E. type                                           IID
## Observations                                        200
## R2                                              0.97337
## Adj. R2                                         0.97296
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We use March as our COVID cutoff because the COVID outbreak in the U.S started around that time, and our data does not specify the actual employment. Thus, in this model, we assume the COVID started on March 1st. In this model, we only include 8 other industries that we were able to find the complete employment data from CPS. This leads to the assumption that we see 8 other industries as the rest of the industries.

The intercept coefficient shows other industries had a average monthly employment of 5175.9 before COVID outbreak.The IndustryRetailTrade coefficient shows that average monthly employment of retail trade industry is 5458 higher than the other industries before COIVD. This indicates that retail trade industry is doing two times better than the rest industries overall in term of employment.

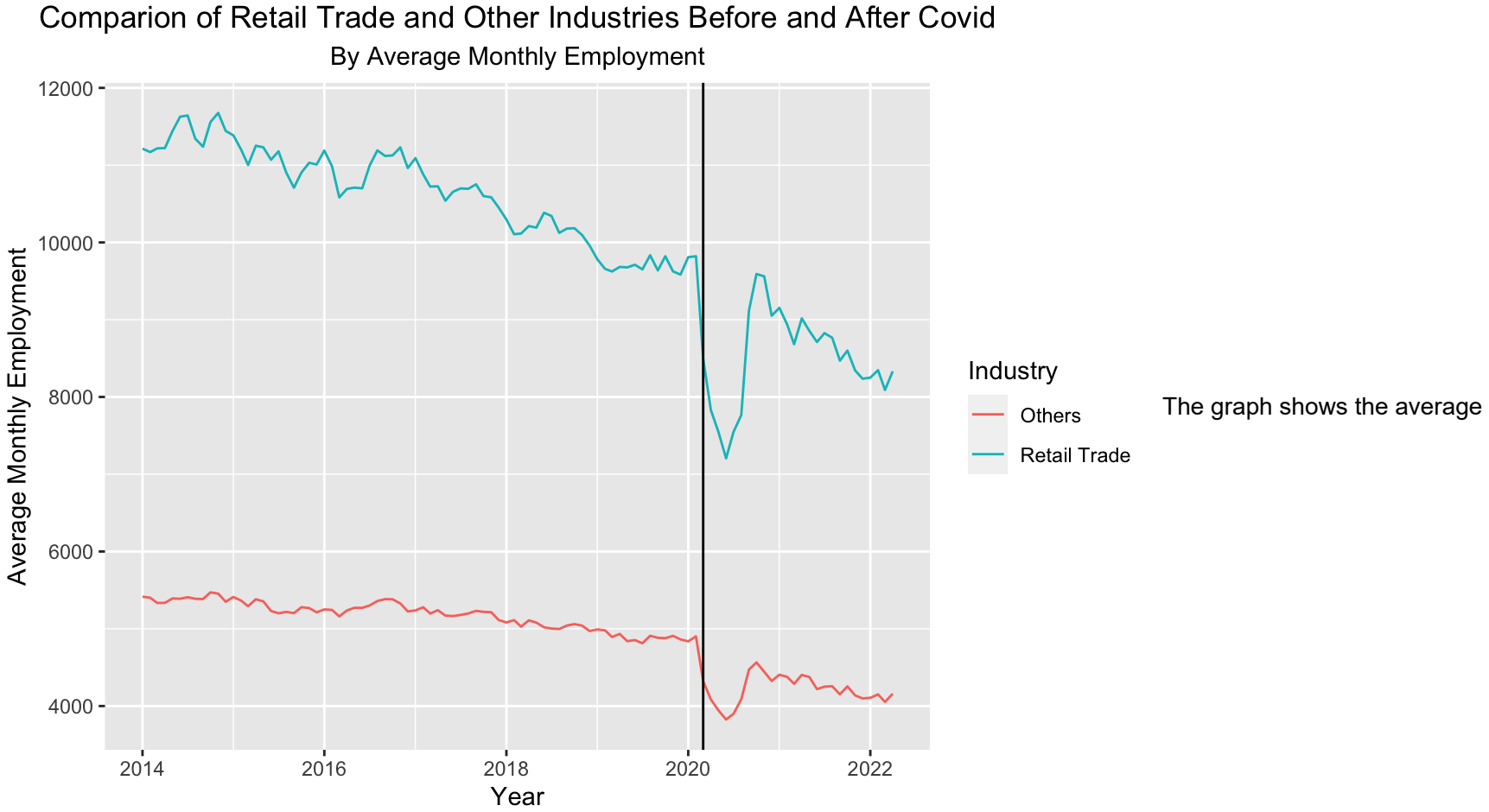
After the COVID outbreak, the coefficient of after_covidTRUE dropped by 958.4, which shows COVID caused many people who work in other industries to lose jobs. The coefficient IndustryRetailTrade x after_covidTRUE of -1172.5 indicates COVID had a worse impact on Retail Trade than other industries in terms of employment. The regression result shows that Retail Industry had higher employment than other industries before COVID occurred. However, the retail trade industry has the most people losing jobs after COVID. The reason of this could be the implementation of COVID quarantine and guidelines force retails stores to close, and other cause can be more people tend to shop online.

##Graph

```
regression1_graph <- regression_data%>%
  group_by(Industry, month=floor_date(date,'month'))%>%
  summarize(avg_monthly_emp)
```

```
## `summarise()` has grouped output by 'Industry'. You can override using the
## `.groups` argument.
```

```
ggplot(regression1_graph,aes(x=month,y=avg_monthly_emp, color=Industry))+geom_line()+
  geom_vline(xintercept = as.Date('2020-03-01'))+
  labs(title = "Comparion of Retail Trade and Other Industries Before and After Covid",
       subtitle = "By Average Monthly Employment")+
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5)) +
  xlab("Year") +
  ylab("Average Monthly Employment")
```



monthly employment of retail trade and other industries. We can see retail has higher employment than other industries the whole time., and a gradual downward trend in the retail industry over time before COVID. The gradual change can be inferred as changes in industry trend in general, or other factors (consumer, environment, etc) cause change in industry development trends.

The vertical line is our COVID outbreak time in our model, which is March First 2020. The spike drop at the beginning of 2020 March shows COVID reduced employment of all the industries, and it started recover before 2021. We can see retail industries have larger unemployment after COVID.

Question 3 Memo

Kennedi Finnes

6/10/2022

Research Question

To answer question 3 “what has changed about who is working and earning money?” This analysis focused on observing how the average age across varying industries changed before and after COVID. This was motivated by the phenomena of increased retirement after large unemployment events that may decrease the average age of employment in certain industires.

Data

The section below details the process for obtaining and cleaning the data necessary to answer the research question in R.

```
#Read in data
model3_data <- read_dta('cps_00003.dta.gz')
vtable(model3_data)
```

model3_data

Name	Class	Label	Values
year	numeric	survey year	Num: 2009 to 2022
serial	numeric	household serial number	Num: 1 to 99461
month	haven_labelled	month	Num: 1 to 12
hwtfinl	numeric	household weight, basic monthly	Num: 0 to 26194.133
cpsid	numeric	cpsid, household record	Num: 0 to 20220406880900
asecflag	haven_labelled	flag for asec	Num: 1 to 2
hflag	haven_labelled	flag for the 3/8 file 2014	Num: 0 to 1
asecwth	numeric	annual social and economic supplement household weight	Num: 52.51 to 28654.31

Name	Class	Label	Values
pernum	numeric	person number in sample unit	Num: 1 to 16
wtfinl	numeric	final basic weight	Num: 0 to 43347.56
cpsidp	numeric	cpsid, person record	Num: 0 to 20220406880904
asecwt	numeric	annual social and economic supplement weight	Num: 50.17 to 44423.83
age	haven_labelled	age	Num: 0 to 85
empstat	haven_labelled	employment status	Num: 0 to 36
ind	numeric	industry	Num: 0 to 9890
classwkr	haven_labelled	class of worker	Num: 0 to 29
incwage	numeric	wage and salary income	Num: 0 to 99999999

```
#Clean Data
model3_data <- model3_data %>%
  filter(year %in% c("2020","2021","2022"))
retail <- c(4670:5790)
health <- c(7860:8470)
trans <- c(6070:6390)
model3_data2 <- model3_data %>%  mutate(Industry = case_when(ind %in% retail  ~ "Retail",
                                                             ind %in% health ~ "Healthcare",
                                                             ind %in% trans ~ "Transportation"))
model3_data3 <- model3_data2 %>% filter(Industry %in%  c("Retail",
                                                         "Healthcare",
                                                         "Transportation"))
model3_data3$Date <- str_c(model3_data3$year,'-' ,model3_data3$month)
model3_data3 <- model3_data3 %>% mutate(Date = ym(Date))
model3_data3 <- model3_data3 %>% mutate(date_number = as.numeric(Date))

model3_data4 <- model3_data3 %>% group_by(date_number, Industry) %>% summarize(Avgagemonth = mean(age))
```

```
## `summarise()` has grouped output by 'date_number'. You can override using the
## `.groups` argument.
```

```
model3_data4 <- model3_data4 %>% mutate(AfterCovid = (date_number >= 18322))
```

Reaserch Design and Regression Results

To asnwer the research question, this analysis chose to use a regression discontinuity approach to observe the changes in average industry age before and after COVID while simultaneaously controlling for other variables that could impact the average age of the industry.

```
model_3_A <- rdrobust(model3_data4$Avgagemonth, model3_data4$date_number,  c = 18322, p = 1, h = 365, kernel = "uniform")
```

```
## [1] "Mass points detected in the running variable."
```

```
summary(model_3_A)
```

```
## Call: rdrobust
##
## Number of Obs.                   84
## BW type                      Manual
## Kernel                      Uniform
## VCE method                       NN
##
## Number of Obs.                    6           78
## Eff. Number of Obs.               6           39
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                 365.000      365.000
## BW bias (b)                 365.000      365.000
## rho (h/b)                     1.000        1.000
## Unique Obs.                       2           26
##
## ===============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
## ===============================================================================
##   Conventional     0.503     2.367     0.212     0.832    [-4.137 , 5.142]
##         Robust         -         -    -0.420     0.675    [-5.147 , 3.331]
## ===============================================================================
```

Interpreting Regression Results and Addressing Assumptions

From the summary above, being after COVID is associated with a 0.5 increase in the average age. However, there are several assumptions that this design makes that could influence this result. The main assumption is that nothing else that influences average age within an industry is also jumping at the cutoff. In this case there may be an unobserved variable we need to control for.

Conculsion

Examining average age within an industry before and after COVID, this regression noted an increase in the average age. However, this result is likely not very generalizeable. This question may have been better answered using a regression design with a categorical *Industry* variable.