DTC

Ryan Bolen, Katie Yamabe, Megan Mixon, Hunter Elliott

library(haven)  
library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(tidyr)  
library(ggplot2)  
library(fixest)

Warning: package 'fixest' was built under R version 4.3.2

library(lubridate)

Attaching package: 'lubridate'

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

library(vtable)

Warning: package 'vtable' was built under R version 4.3.2

Loading required package: kableExtra

Warning: package 'kableExtra' was built under R version 4.3.2

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':  
  
 group\_rows

You can add options to executable code like this

main\_data <- read\_dta("finaldata.dta.gz")  
industry\_names <- read.csv('indnames.csv')  
  
# If you are having trouble loading everything, comment the next two lines out. Only time the hh income variable from this df is used is in the 3rd regression.   
  
  
income\_var <- read\_dta('IPUMS.dta.gz')  
income\_var <- income\_var %>% select(cpsid,hhincome)  
  
main\_data <- main\_data %>% right\_join(income\_var,join\_by(cpsid))

Warning in right\_join(., income\_var, join\_by(cpsid)): Detected an unexpected many-to-many relationship between `x` and `y`.  
ℹ Row 1 of `x` matches multiple rows in `y`.  
ℹ Row 90003 of `y` matches multiple rows in `x`.  
ℹ If a many-to-many relationship is expected, set `relationship =  
 "many-to-many"` to silence this warning.

main\_data <- main\_data %>% filter(!is.na(hhincome))  
  
# Join up main data frame with the csv frame  
  
main\_data <- main\_data %>% right\_join(industry\_names, join\_by(ind))

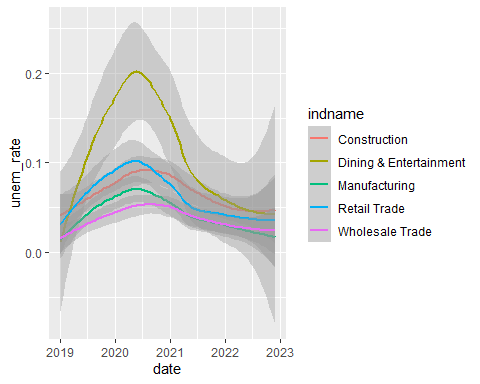
# Get rid of observations without an industry specified in raw data and cutting down to not go past 2022  
  
main\_data <- main\_data %>% filter(!is.na(ind))  
main\_data <- main\_data %>% filter(year < 2023)  
  
  
  
# Join up main data frame with the data set we created to categorize 'general retail industries' together.   
  
  
#main\_data <- main\_data %>% left\_join(big\_tib, join\_by(ind))  
  
main\_data <- main\_data %>% mutate(date = make\_date(year,month))  
  
# Need to change continuouweeks unemployed to 0 if it equals 999 - if we did not do this, it would wash out all the variance/drive the average of unemployed time per category way up.   
  
main\_data$durunemp <- replace(main\_data$durunemp,main\_data$durunemp > 998, 0)  
  
# Creating a variable to show weather or not the date is pre or post covid  
  
main\_data <- main\_data %>% mutate(post\_covid = ymd(date) > ymd('2020-03-01'))  
  
# Filtering out individuals that are not in the labor force. Creating true/false variable for whether they are unemployed or not  
  
main\_data <- main\_data %>% filter(empstat <30) %>% mutate(unemployed = empstat > 12)  
  
# Make a date variable by combining month and date  
  
main\_data <- main\_data %>% filter(ymd(date) != ymd('2020-03-01'))  
  
# Make a binary variable (T or F) for Female or Male  
  
main\_data <- main\_data %>% mutate(female = sex == 2)

# creating a variable for weather or not we are after or before covid  
# Getting an actual date variable created and grouping observations by date,   
  
#data\_by\_month <- main\_data %>% group\_by(indname,year,month) %>% summarise(avg\_time\_unempl = mean(durunemp)) %>% mutate(date = make\_date(year,month))  
  
percent\_unempl <- main\_data %>% group\_by(indname,date) %>% summarize(tot = n(),unempl = sum(unemployed)) %>% mutate(unem\_rate = unempl/tot) %>% mutate(retail = indname == 'Retail Trade') %>% mutate(post\_covid = ymd(date) > ymd('2020-04-01'))

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

#Rename the category for arts, dining entertainment, recreation because it is too long of a name and makes the graph look silly  
  
percent\_unempl$indname <- replace(percent\_unempl$indname,percent\_unempl$indname == 'Arts, Entertainment, and Recreation, and Accommodation and Food Services','Dining & Entertainment')  
  
# Plot to visualize the average duration of time un-emlpoyed for a given industry in a given year/month.  
  
ggplot(data = percent\_unempl %>% filter(indname %in% c('Retail Trade','Wholesale Trade','Manufacturing','Construction','Dining & Entertainment')),mapping = aes(x = date,y = unem\_rate, color = indname)) + geom\_smooth()

`geom\_smooth()` using method = 'loess' and formula = 'y ~ x'



# Modeling  
  
# How has COVID affected the health of the retail industry, as measured by employment?  
  
# Model 1 - regressing dependent variable is average duration of unemployment.   
  
retail\_health <- feols(durunemp ~ post\_covid, data = main\_data %>% filter(indname == 'Retail Trade'))  
  
etable(retail\_health)

retail\_health  
Dependent Var.: durunemp  
   
Constant 0.8580\*\*\* (0.0314)  
post\_covidTRUE 0.4522\*\*\* (0.0353)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 285,157  
R2 0.00057  
Adj. R2 0.00057  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# The occurence of covid 19 is assosciated with an approximate 0.37 average increase in the duration of weeks unemployed for the retail industry.

# How has retail fared relative to other industries?  
  
retail\_comparison <- feols(unem\_rate~post\_covid + retail,data = percent\_unempl)  
etable(retail\_comparison)

retail\_comparison  
Dependent Var.: unem\_rate  
   
Constant 0.1179\*\*\* (0.0206)  
post\_covidTRUE -0.0026 (0.0257)  
retailTRUE -0.0556 (0.0481)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 376  
R2 0.00360  
Adj. R2 -0.00174  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# The occurence of covid 19 had a negative on employment rate overall but the retail industry did not have as much of a negative effect as other industries.  
  
#However, it's crucial to add an interaction term between post\_covid and retail to see if the change in unemployment rates after COVID-19 was different for the retail industry compared to other industries.  
retail\_comparison\_interaction <- feols(unem\_rate ~ post\_covid + retail + post\_covid \* retail, data = percent\_unempl)  
etable(retail\_comparison\_interaction)

retail\_compariso..  
Dependent Var.: unem\_rate  
   
Constant 0.1182\*\*\* (0.0211)  
post\_covidTRUE -0.0032 (0.0267)  
retailTRUE -0.0600 (0.0791)  
post\_covidTRUE x retailTRUE 0.0071 (0.0996)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 376  
R2 0.00361  
Adj. R2 -0.00442  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#allows the model to estimate the differential effect of COVID-19 on the unemployment rate in the retail sector relative to other sectors.  
  
# Assuming 'retail\_trade' is the reference category and is omitted from the model to avoid multicollinearity

# Retail needs to worry about who has money to spend - what has changed about who is working and earning money (this regression needs to take into account all industries - need an income variable - family household income is categorized as a range of incomes and it's not in a particularly useful order)?  
  
# Recode 'educ' to a factor with more interpretable levels  
main\_data <- main\_data %>%  
 mutate(educ\_factor = case\_when(  
 educ %in% 0:5 ~ "No schooling/Primary",  
 educ %in% 6:11 ~ "Some high school",  
 educ %in% 12:73 ~ "High school graduate",  
 educ %in% 74:81 ~ "Some college",  
 educ %in% 91:92 ~ "Associate degree",  
 educ %in% 100:125 ~ "Bachelor's degree or higher",  
 TRUE ~ "Other/Unknown"  
 ))  
  
# Convert to a factor for regression analysis  
main\_data$educ\_factor <- as.factor(main\_data$educ\_factor)  
  
earnings\_model <- feols(log(hhincome)~post\_covid + female + age + educ\_factor, data = main\_data)

Warning in log(hhincome): NaNs produced

NOTE: 7,176 observations removed because of NA and infinite values (LHS: 7,176).

etable(earnings\_model)

earnings\_model  
Dependent Var.: log(hhincome)  
   
Constant 11.44\*\*\* (0.0024)  
post\_covidTRUE 0.0318\*\*\* (0.0012)  
femaleTRUE -0.0766\*\*\* (0.0010)  
age 0.0017\*\*\* (3.46e-5)  
educ\_factorBachelor'sdegreeorhigher 0.3887\*\*\* (0.0017)  
educ\_factorHighschoolgraduate -0.2130\*\*\* (0.0017)  
educ\_factorNoschooling/Primary -0.4217\*\*\* (0.0118)  
educ\_factorSomecollege -0.0504\*\*\* (0.0019)  
educ\_factorSomehighschool -0.5174\*\*\* (0.0076)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type IID  
Observations 2,671,644  
R2 0.10433  
Adj. R2 0.10433  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Assumptions:

Observations where industry type was recorded as na were omitted so that we are only measuring industries where we can categorize the industry name.

People not in the labor force are not part of the data.

People who did not report the duration of time that they were unemployed were manipulated to show that their duration of unemployment time was 0. We also filtered out people people that weren’t in the labor force and we assume that they are either unemployed or employed.

!!!!!!!

Might be good to use this assumption:

The data available for this study is provided on a monthly basis. In order to best answer the research questions here, we are focusing on framing the data in two categories: pre covid (prior to March 2020) and post covid (post March 2020). In order to best characterize these periods and get a true ‘before and after effect’ we omit observations in March 2020. (This is ‘during covid’ and is a very small portion of the sample data and does not distinctly fit into pre or post covid time frames).

# Parking Lot  
  
# Updated Model with Education  
#retail\_health\_educ <- feols(durunemp ~ post\_covid + age + educ\_factor, data = main\_data %>% filter(indname == 'Retail Trade'))  
#etable(retail\_health\_educ)  
  
#In this setup, each industry's impact on the unemployment rate relative to retail both before and after the onset of COVID-19 can be assessed through the coefficients of the interaction terms.  
  
# Using 'indname' factor in the feols regression model  
# model\_industry\_comparison <- feols(unem\_rate ~ post\_covid + indname + post\_covid \* indname, data = percent\_unempl)  
# etable(model\_industry\_comparison)