**DTC**

Ryan Bolen, Katie Yamabe, Megan Mixon, Hunter Elliott

# Load relevant libraries  
  
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

library(scales)

Attaching package: 'scales'

The following object is masked from 'package:fixest':  
  
 pvalue

# Load various data sources   
  
main\_data <- read\_dta("finaldata.dta.gz")  
industry\_names <- read.csv('indnames.csv')  
  
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 to get industry name into each observation  
  
main\_data <- main\_data %>% right\_join(industry\_names, join\_by(ind))

**Data Cleaning and Assumptions:**

For this analysis, we got rid of any observations that did not contain an industry specified in the raw data, cutting down observations to avoid going past the year 2022 (after COVID).  Then we joined the main data frame with the data set to categorize 'general retail industries'. We realized that we needed to change continuous weeks of unemployment to 0 if it equals 999 to avoid washing out all variance and driving the average of unemployed time per category way up. We filtered out individuals that are not in the labor force, creating a true or false variable for whether individuals were unemployed or not, as well as creating another true/false binary variable for females and males. We created a binary variable indicating whether the observation date was recorded before or after the first lock downs associated with COVID-19 which we utilized in various regressions. We then plotted this to visualize the average duration of time unemployed for a given industry each year and month.

The data available for this study was only provided monthly, so we assumed each observation was reported on the first day of each month. To best answer the research questions, we are focusing on framing the data in two categories: pre-covid (months prior to March 2020) and post-covid (months post-March 2020). To best characterize these periods to achieve a true "before and after effect", we omit all observations in March 2020. We did this because this time-period is considered "during COVID" and contains a very small portion of the data sample, thus not distinctly fitting into pre or post-COVID time frames.

Before running the regressions to begin answering the questions, we decided to graph the categories 'arts', 'dining' 'entertainment', and 'recreation' because we wanted to visualize the average unemployment rate for a given industry from the years 2019-2023. We assume that industry type was omitted when recorded as NA so that we only measure the industries where we can categorize the industry name.

# 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 %>% 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,   
  
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() + xlab("Date") + ylab("Unemployment Rate") + ggtitle("Unemployment Rate by Industry")

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



* **How has COVID-19 affected the health of the retail industry, as measured by employment?**

To answer "How has COVID-19 affected the health of the retail industry, as measured by employment", we made one regression model.  This model(retail\_health) has 'average duration of unemployment' as the dependent variable and the independent variable is 'post\_covid'. We chose to evaluate this variable because it allows us to observe how the duration of unemployment is associated with dates post covid relative to individuals working specifically within the 'Retail Trade' category. When using etable(), the results show that the occurrence of COVID-19 is associated with an approx. 0.4522 average increase in the duration of unemployment within the retail industry.However, this coefficient is not significant, likely because once we grouped all the observations into months retrieving the unemployment rate, the sample size rapidly shrunk. Overall, COVID-19 affected the health of the retail industry, as measured by employment through a 45.22% unit average increase in the duration of unemployment for individuals working in the retail industry.

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

* **How has retail fared relative to other industries?**

To answer, "how the retail industry fared relative to other industries", we made two regression models with 'unemployment rate' as the dependent variable. The first model(retail\_comparison) includes 'post\_covid' and 'retail', among the data set containing the percentage of unemployed individuals within theconstruction, dining and entertainment, manufacturing, retail trade, and wholesale trade.We chose to use these variables to show how the unemployment rate changed within the retail industry during post covid when compared to other industries.

We then used etable() to see the results of the regression, showing that the occurrence of COVID-19 had a 0.0026 negative impact on the overall employment rate post covid; however, the retail industry specifically did not have as much of a negative effect as other industries. That said, a one-unit change in the unemployment rate is associated with a 0.0556 unit decrease in the retail industry post-covid.

However, we believe that adding an interaction term between 'post\_covid' and 'retail' would aid in evaluating a possible difference in unemployment rates after COVID-19 for the retail industry compared to other industries. In the second regression model(retail\_comparison\_interaction), the dependent variable is 'unemployment rate', with independent variables being 'post\_covid' and 'retail', along with an interaction term of 'post\_covid' and 'retail'. With this, we are also assuming that 'retail\_trade' is the reference category and therefore is omitted from the model to avoid multicollinearity. When reviewing results in etable(), we can see that adding an interaction term allows the model to estimate the differential effect of COVID-19 on the unemployment rate in the retail industry relative to other industries.

Overall, the occurrence of COVID-19 had a 0.0032% unit decrease on the overall unemployment rate; however, the retail industry specifically did not have as much of a negative effect as other industries, as it had a 0.0071% unit change increase in unemployment rate in the retail industry relative to other industries during post covid.

When evaluating these results, we look back on our first plot in the analysis, showing the average duration of time individuals were unemployed for a given industry each year, to aid in explaining why we are seeing these results. Perhaps we see these results because the 'arts' and 'dining and entertainment' industries had the worst effects during the time of COVID-19, due to government health policies. Due to people not being able to fill restaurants to full capacity, the dining industry did not require as much staff during COVID-19, resulting in more layoffs and shutdowns. In addition, the entertainment industry also had the most unemployed people who were working in that sector due to gathering restrictions, which caused events such as concerts and movie production to come to a halt.

However, we are likely to see more positive unemployment rate results in retail relative to other industries because the retail industry was likely already showing unemployment rates due to online shopping and inflation.  When compared to the dining and entertainment industry, not many individuals worked in retail before COVID.  Though retail was not essential during COVID-19, it did not produce as bad an impact relative to other industries because other industries did not have a lot of traffic anyway.

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

**3. Retail needs to worry about who has money to spend - what has changed about who is working and earning money?**

To answer this question, we made a regression model(earnings\_model) with household income as the dependent variable. We chose to recode education('educ') to a factor variable with distinct levels to more accurately measure an individual's education status to see how it would impact their household income post covid. The model's independent variables include 'post\_covid', 'female', 'age', and 'educ\_factor'. We chose these variables because they allow us to examine individuals measured by household income among different categories from all industries, showing us what has changed about who has money to spend during that time. We wanted to include two demographic variables, 'age' and 'sex', to see who was making money and not making money among those groups. We assume that the data consisting of individuals who did not report the duration of time they were unemployed was 0 and so we manipulated the data to show that the duration of unemployment time was 0 for n/a observations. We also filtered out observations where people reported that they were not in the work force, under the assumption that they were not actively looking for work and not part of unemployment measurements. In addition, we chose to log the regression because household income is skewed, making for a very broad data set to analyze.  We then adjusted for heteroskedasticity that was present within the regression due to age and time variables, indicating that the variance of household income is not constant across time, age group, and industry. We present the results using etable(), showing that a .01 unit change in X is associated with a 3.18% change in household income. This tells us that during post covid, household income increased by 3.18%.  When we look at the coefficients, we see that the household income for females decreased by 7.66%.  Individuals with a bachelor's degree or higher had a 38.87% increase in household income, whereas individuals with education lower than a bachelor's degree experienced a decrease in household income post covid. Perhaps this is because most jobs that allowed people to work from home required a bachelor's degree or higher, allowing them to keep their jobs, relative to in-person jobs, requiring less than a bachelor's degree, being in higher demand post covid.

Individuals with some high school education experienced a 51.74% decrease in household income while individuals with no schooling experienced a 42.17% decrease in household income. That said, those unable to get a job due to lack of education post-covid experienced a decrease in household income. Perhaps this is because the jobs that didn't require higher education were in high demand, filling up fast due to most people in the industries not having enough education for the higher paying jobs. To visualize this regression, we created a histogram using ggplot(). We can see that the household income on the number of individuals in all industries portrays a very skewed pattern, resulting in a large right tail. This outcome further justifies why we needed to do a log on household income, along with heteroskedasticity-robust standard errors.

# 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,vcov = 'hetero')

earnings\_model  
Dependent Var.: log(hhincome)  
   
Constant 11.44\*\*\* (0.0023)  
post\_covidTRUE 0.0318\*\*\* (0.0012)  
femaleTRUE -0.0766\*\*\* (0.0010)  
age 0.0017\*\*\* (3.58e-5)  
educ\_factorBachelor'sdegreeorhigher 0.3887\*\*\* (0.0016)  
educ\_factorHighschoolgraduate -0.2130\*\*\* (0.0016)  
educ\_factorNoschooling/Primary -0.4217\*\*\* (0.0142)  
educ\_factorSomecollege -0.0504\*\*\* (0.0018)  
educ\_factorSomehighschool -0.5174\*\*\* (0.0080)  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
S.E. type Heteroskedast.-rob.  
Observations 2,671,644  
R2 0.10433  
Adj. R2 0.10433  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Plot showing how unevenly distributed income is in our data set (right skewed):  
  
brk <- seq(0,500000,by = 20000)  
  
ggplot(data = main\_data %>% filter(hhincome > 0),mapping = aes(x = hhincome)) + geom\_histogram(binwidth = 10000) + xlab("Household Income") + ylab("Count") + ggtitle("Income Distribution") + scale\_x\_continuous(breaks = brk,labels = label\_dollar(suffix = "", prefix = "$"),limits = c(0,500000)) + theme(axis.text.x = element\_text(angle = 90))

Warning: Removed 49419 rows containing non-finite values (`stat\_bin()`).

Warning: Removed 2 rows containing missing values (`geom\_bar()`).



**Concluding Statements**

Overall, the majority of individuals who are working are the ones who are employed with a bachelor's degree or higher during post covid, having a 38.87% increase in household income. To conclude, the regression shows that mostly males who have a bachelor's degree or higher would receive more household income post covid; therefore, the retail industry needs to focus on this demographic relative to individuals who have an education lower than a bachelor's degree to increase profits in the retail industry.