W203
Lab 2

Causal Study of Shelter-in-Place Orders on Unemployment

Robert Hosbach
Peter Kim
Heather Reed

Berkeley SCHOOL OF INFORMATION

Table of Contents

List of Figures

List of Tables

1. Introduction

Canceling large public gatherings. Asking students to stay home from school. Closing down borders. Many places around the world have implemented such drastic steps in response to the coronavirus outbreak that originated in China. The United States (U.S.) was also hard-hit by the virus beginning in early 2020, and 39 states decided to implement stay-at-home and shelter-in-place (SAH/SIP) orders in response to the rapidly increasing number of people infected with COVID-19. Shortly thereafter, the unemployment rate in the U.S. rose to a record high 14.7 percent in April 2020 [2]. Naturally, this poses the question: Did state SAH/SIP orders cause higher unemployment?

In order to assess the effects of SAH/SIP orders on unemployment, our team used the COVID-19 U.S. State Policy Database compiled by the Boston University School of Public Health in order to determine what social distancing measures each state implemented in response to the COVID-19 pandemic, as well as when these measures were implemented [4]. We used this state policy data along with the Unemployment Insurance Weekly Claims data from the U.S. Department of Labor in order to analyze how social distancing measures affected unemployment claims at the state level [3].

The COVID-19 policy data revealed that all but 11 of the 51 states (including the District of Columbia) implemented their SAH/SIP measures in the narrow period between March 19, 2020 and April 7, 2020 (inclusive). We also discovered that most states reached a peak in unemployment in the months from April to May 2020. Given that the period of peak unemployment closely followed the implementation of SAH/SIP orders for each state, we theorized that the SAH/SIP orders may have caused the high state-level unemployment.

2. Model-Building Process

The goal of our model was to explain our causal theory that the SAH/SIP orders issued by states in response to the COVID-19 pandemic caused an increase in unemployment. In order to model this, we utilized COVID-19 data regarding SAH/SIP orders and business sector closure orders issued by states in response to the pandemic as well as weekly unemployment data for each state.

2.1 Exploratory Data Analysis

Our dependent variable is the percent change in average unemployment rates between February 2020 (the baseline) and April 25, 2020 to June 13, 2020.[1] We chose the average unemployment rate over the month of February for each state as our baseline because every

[1] These dates were selected based on the weeks available in the unemployment data set. Weeks ending February 8th, 15th, 22nd, and 29th were available to characterize the month of February. Weeks ending April 25th through June 13th were also available, which is the reason such detailed dates were selected.

state that issued a SAH/SIP order did so on or after March 19, 2020 (see Figure 1). Therefore, February was the last full month before any SAH/SIP orders were established. We chose the average unemployment rate from late-April to mid-June 2020 because that time period captures the peak of nearly every state's maximum unemployment rate for 2020. Thus, we expect to observe a substantial difference in unemployment rates between the two periods in each state that could potentially be caused by the establishment of SAH/SIP orders.

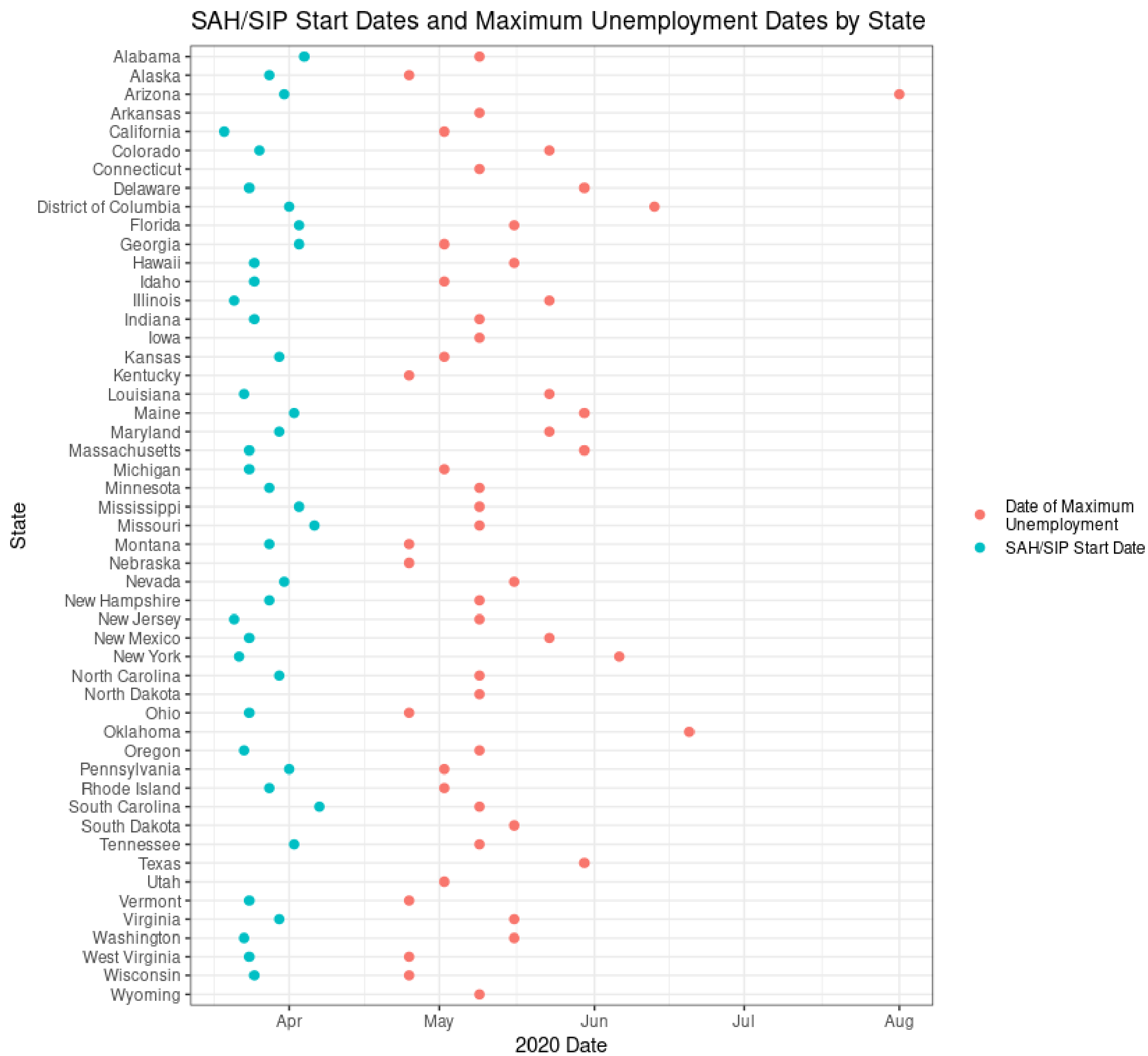

Figure 1. Start dates of SAH/SIP orders for states that established one and dates of maximum unemployment rate by state

Figure 2 shows a histogram and normal Q-Q plot of our response variable, which is non-normally distributed with a right skew. This is due in part to only having 51 data points, which correspond to the 50 states and the District of Columbia. Transforming the change in unemployment rate variable to be symmetrically distributed (ideally normally-distributed) could improve the fit of a linear model. Based on the skew in the data, we determined that a log transformation would be the most appropriate to make the distribution more symmetric and improve our model. However, we made the decision not to perform this variable transformation due to interpretability concerns. Because the response variable captures the percent change in average unemployment, conducting a log transformation would lessen the interpretability of our model.[2] We ultimately decided that the loss in interpretability was not worth the slight improvement that may be achieved by log-transforming the variable.

[2]Typically, log-transforming the dependent variable is useful in that the model can then be interpreted as the percent increase/decrease in the dependent variable for every one-unit increase in the independent variable. In our model, this interpretation would mean the percent change in the percent change of unemployment between our two periods of comparison, which is difficult to understand.

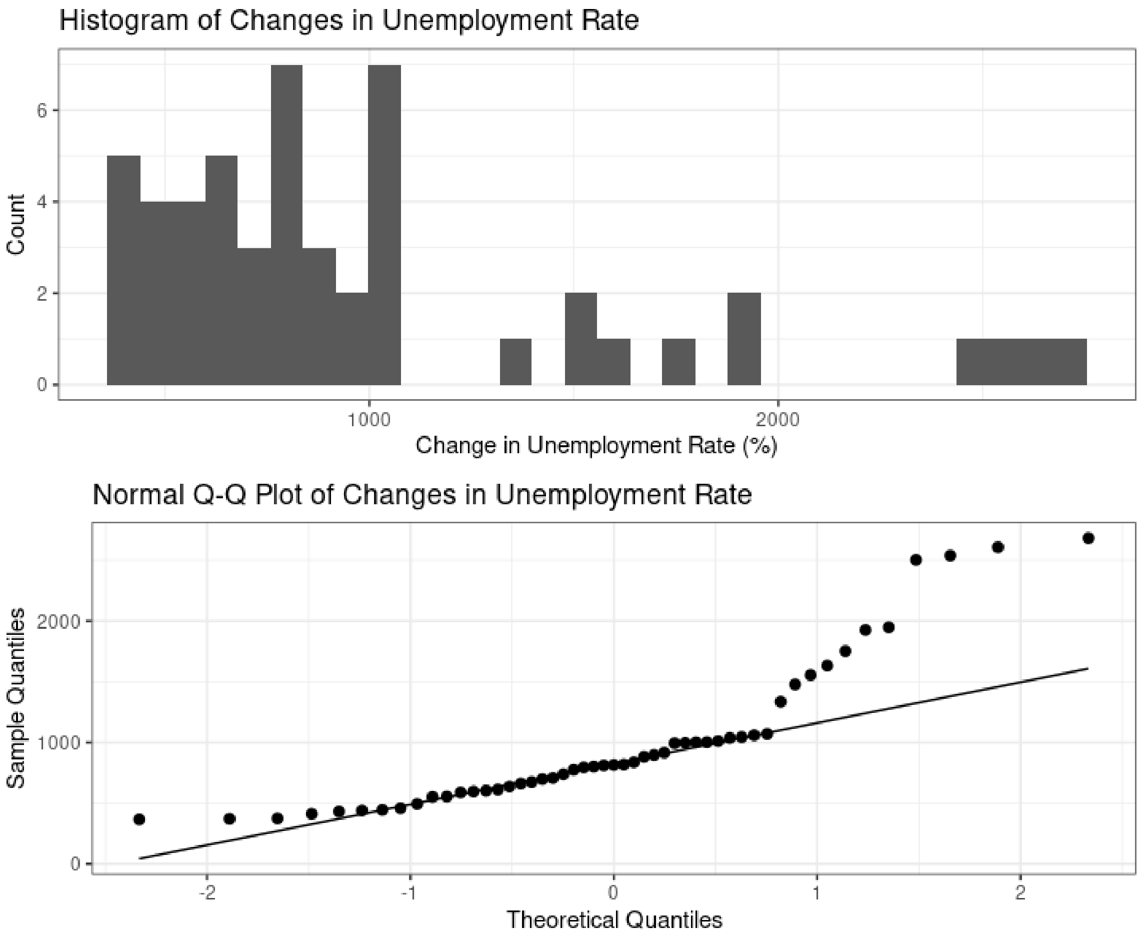

Figure 2. Histogram (top) and normal Q-Q plot (bottom) of the change in average unemployment rate between February 2020 and late-April to mid-June 2020

Our primary explanatory variable is a binary indicator for whether or not a state implemented a SAH/SIP order. Of the 51 states, 40 implemented a SAH/SIP order and 11 did not. We also developed subsequent models using binary indicator variables pertaining to specifics of the SAH/SIP orders and metric variables describing closures of different sectors of business in response to the pandemic. The first two models only include the binary indicator explanatory variables; therefore, we did not consider any variable transformations. The third model included three metric variables providing counts of the number of business sectors that closed during each of three closure phases. While the distributions for the variables are skewed, zero is a common value for all three variables because there are states that did not conduct more than one phase or two phases of closures, which makes log-transformation a poor option. Other options, such as polynomial transformations, were not considered due to interpretability challenges.

Figure 3 shows a plot of the Pearson correlation coefficients for each pair of variables we included in our three models. The strongest correlation is between the indicator variables for a state having a SAH/SIP order or not and whether movement restrictions were imposed along with that order. This is the only variable pairing from our models that has a correlation magnitude of at least 0.5. While it is somewhat surprising that we do not observe higher correlation magnitudes between some of the variables in our models, this may be a result of half of the variables being binary indicators.

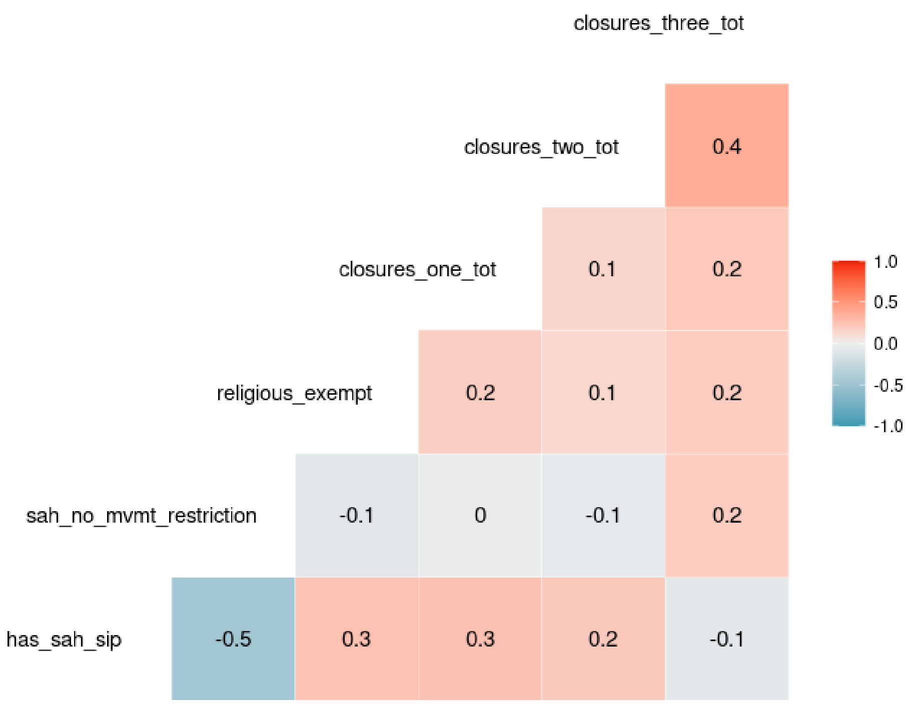

Figure 3. Pearson correlation coefficients for each pair of variables included in the three models

Next, we detail each of the three models we investigated in this causal analysis.

2.2 Model 1 (Limited Model)

$$\begin{aligned}\Delta unemployment = \beta_0 & \\ & + \beta_1 \times \texttt{State has SAH/SIP Order?(Y/N)} \\ & + error\end{aligned}$$

In our limited model, we modeled the change in unemployment ($\Delta unemployment$) using only our key explanatory variable (`State has SAH/SIP Order?(Y/N)`), which is a binary indicator representing whether or not the state ever established a SAH/SIP order. Due to our key explanatory variable being an indicator variable, we knew it would be difficult to visualize the model in terms of linearity and residual plots.

2.3 Model 2 (Main Model)

$$\begin{aligned}\Delta unemployment = \beta_0 \\ &+ \beta_1 \times \texttt{State has SAH/SIP Order?(Y/N)} \\ &+ \beta_2 \times \texttt{State has SAH/SIP Order Religous Exemptions?(Y/N)} \\ &+ \beta_3 \times \texttt{State has SAH/SIP Order Movement Restrictions?(Y/N)} \\ &+ error\end{aligned}$$

For our main model, we incorporated additional binary indicator explanatory SAH/SIP order variables regarding religious exemptions and movement restrictions for citizens. Due to these variables providing a more complete picture of how the SAH/SIP orders affected each state and the ability for citizens to work and maintain jobs, we believed that adding these covariates would help us better achieve our modeling goals and explain the relationship between the SAH/SIP orders and the change in unemployment rate. Additionally, adding two new variables created more diversity of outcomes compared to our limited model, which could only predict two values for the change in unemployment for every state because only the single binary indicator variable was used. By adding these two new indicator variables, we increased the number of possible predicted values for each state from two to eight.

2.4 Model 3

$$\begin{aligned}\Delta unemployment = \beta_0 \\ &+ \beta_1 \times \texttt{State has SAH/SIP Order?(Y/N)} \\ &+ \beta_2 \times \texttt{State has SAH/SIP Order Religous Exemptions?(Y/N)} \\ &+ \beta_3 \times \texttt{State has SAH/SIP Order Movement Restrictions?(Y/N)} \\ &+ \beta_4 \times \texttt{\# of closures in phase 1} \\ &+ \beta_5 \times \texttt{\# of closures in phase 2} \\ &+ \beta_6 \times \texttt{\# of closures in phase 3} \\ &+ error\end{aligned}$$

For our third model, we incorporated data regarding business closures in response to the COVID-19 pandemic. We wanted to see if these closures played a role in the change in unemployment rates, in addition to the SAH/SIP-specific variables we included in the first two models. We operationalized the closure data by counting the number of business sectors that were closed as a result of the pandemic. These counts were grouped by what round (or phase) of closures the state was on. The rounds or phases of closures are relative to each state and are based on when that state initiated their first, second, and third phase of closures. Of the 51 states (including the District of Columbia), every state had a first phase of business sector closures, 17 states (33%) had a second phase of closures, and three states (6%) had a third phase of closures. Our hope in including the closure data was that including metric variables that counted the number of business sectors that were closed would increase the variation in our data and provide a better linear model.

3. Regression Table

Table 1 is a regression table that facilitates easy comparison between the three linear causal models we developed. The standard errors provided with the coefficients are robust standard errors.

Table 1. Regression Table

	Dependent variable:		
	Change in Unemployment Rate		
	(1)	(2)	(3)
SAH/SIP Order? (Y/N)	425.419***	510.672***	669.695***
	(123.911)	(161.013)	(195.415)
SAH/SIP No Movement Restrictions? (Y/N)		306.735	330.723**
		(188.965)	(146.595)
SAH/SIP Religious Exemptions? (Y/N)		65.536	92.840
		(196.834)	(198.039)
Closures: # Sectors in Phase 1			−71.356
			(62.561)
Closures: # Sectors in Phase 2			−112.195***
			(42.134)
Closures: # Sectors in Phase 3			270.265
			(190.976)
Constant	664.735***	547.237***	1,041.552**
	(67.810)	(105.891)	(488.734)
Observations	51	51	51
R^2	0.085	0.106	0.190
Adjusted R^2	0.066	0.049	0.079
Note:			*p<0.1; **p<0.05; ***p<0.01

After creating our limited model, we found that our key explanatory variable was statistically significant at a 99% confidence level. However, the adjusted R-squared for the limited model was only 0.066. This is expected due to only including one explanatory variable in this model. With this in mind, we developed another model that incorporated other aspects of the COVID-19 SAH/SIP orders to see if that provided improvement.

After evaluating our main model, we found that the two additional indicator variables are not statistically significant. Our key explanatory variable, whether or not had a SAH/SIP in place, remained statistically significant at a 99% confidence level. Accordingly, our adjusted R-squared value decreased from 0.066 in our limited model to 0.049 for this main model. We observe a decrease in the adjusted R-squared value for this model because we have included two additional variables that are not statistically significant, and the adjusted R-squared metric penalizes the inclusion of non-significant variables in a model. An F-test comparing the limited model to our main model provides a p-value of 0.58, indicating that there is not enough evidence to suggest that adding the additional explanatory variables in the main model provides a statistically significant improvement over the limited model. Nevertheless, we kept the additional explanatory variables in our main model because they are not highly correlated with our key explanatory variable, yet the additional variables are directly related to established SAH/SIP orders. Moreover, a variance inflation factor (VIF) analysis shows that the factors for the explanatory variables in the main model are no larger than 1.36. This indicates no cause for concern.

Our third model results indicate that there are three statistically significant explanatory variables. Similar to the limited and main models, the SAH/SIP order variable was statistically significant at a 99% confidence level. In addition, the binary indicator variable for SAH/SIP movement restrictions and the metric closure variable for phase 2 closures were statistically significant at a 95% and 99% confidence level, respectively. The adjusted R-squared for model 3 is 0.079, which exceeds that of the other two models (albeit only marginally). A VIF analysis shows that the factors for the explanatory variables in this model are no larger than 1.52, which indicates no cause for concern.

Table 1 shows that in all three of our models our key explanatory variable that represents whether a state has a SAH/SIP order in effect was statistically significant at a 99% confidence level. Our main model and model 3 both had additional statistically significant variables at or above a 95% confidence level. However, the highest adjusted R-squared value we achieved for the three models was 0.079, indicating that the majority of the variation in our data could not be explained by our model. Finally, the estimated coefficient for our primary variable ranges from 425% to 670% across the three models.

Practically, these results indicate that there are more factors that contributed to the change in unemployment during the COVID-19 period than just whether or not a state had a SAH/SIP order and associated religious exemptions and movement restrictions in place. For example, many restaurants switched to takeout and delivery only when SAH/SIP and closure orders were implemented. A restaurant owner's decision to stay open and continue to operate vs. close their business entirely and laying off their employees likely took into account operating expenses, feasibility of altering their business model, and many other factors beyond just whether or not there was a SAH/SIP order in place. There are likely many nebulous factors that contributed to a rise of the unemployment rate during this period of time since virtually every sector of business was affected. With our limited and primarily binary indicator variables, it was difficult to capture the complexity of how the pandemic affected each state's unemployment rate.

4. Limitations of Model

In this section we describe the limitations of our main model, which incorporates three binary explanatory variables: a binary indicator of whether or not the state established a SAH/SIP order, and two more binary indicator variables for religious exemptions and movement restrictions for citizens under SAH/SIP. We ultimately selected model 2 as our main model for two reasons: 1) the two additional explanatory variables we have included in model 2 (relative to model 1) are clearly related to our causal theory, whereas the extra variables in model 3 are more tangential in nature; and 2) model 2 is the only of our three models that arguably meets the classical linear model (CLM) assumption of linear conditional expectation (LCE). We provide more detail on the latter reason below.

Our model is based on 51 observations, which means that the five assumptions of the CLM must be considered: 1) independent and identically distributed (IID) data, 2) linear conditional expectation (LCE), 3) no perfect collinearity, 4) homoskedastic errors, and 5) normally distributed errors. We discuss each of these assumptions in turn.

4.1 IID Data

This assumption relies more on the experimental design than the actual data points. Ideally, a truly random sample is drawn from the population of interest, with no dependence between data points. In our analysis, we have data for all 51 states in the U.S., which comprises the entire population of interest. Therefore, our data should be distributed according to the population. However, because we are working with state-level data, the independence assumption is challenged. Geographic clustering effects are prominent due to the fact that states interact with one another, especially their neighbors, economically (commerce and trade) as well as by portions of the population in one state frequently traveling to and from other neighboring states. Politics can lead to clustering effects as well. Due to state governors having broad policy control over how their state will respond to situations like the COVID-19 pandemic, states with governors associated with the same political party will often respond similarly. These are but a subset of the many reasons for which the data we used to construct not only our main model, but all three of our models, is not IID.

4.2 Linear Conditional Expectation

To assess the LCE assumption for our main model, we plotted our model residuals against the model's predictions and looked for a horizontal trendline at zero (see Figure 4). The model achieves a roughly horizontal line at zero for the predictions below 1,125% unemployment increase. However, one prediction exists near 1,375% unemployment increase that has a positive residual of nearly 500%. This is the prediction for Hawaii, and it causes the mostly-horizontal trendline to curve upward in that portion of the plot. It is possible that the large prediction and associated residual for Hawaii are due to the fact that Hawaii is the only analyzed state that is an island, as opposed to being land-locked. Therefore, the spread and impact of the virus in Hawaii may very well have differed from that of the other mainland states.

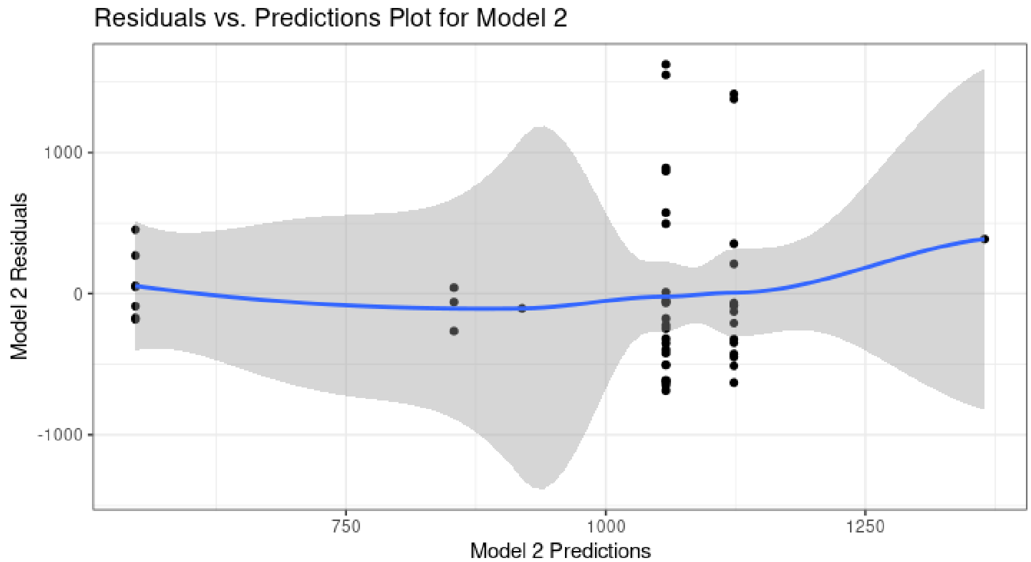

Figure 4. Plot of the residuals vs. predicted values for the main model (model 2)

While the trendline is not perfectly linear at zero, especially for the single highest prediction value, the trendline is nearly linear at zero for 50 out of the 51 model predictions. Therefore, we conclude that there is not strong evidence to suggest that our model fails to meet the LCE assumption.

4.3 No Perfect Collinearity

By virtue of the statistical software we used (R) retaining all three of the explanatory variables we specified for our main model, we meet the assumption of no perfect collinearity.

4.4 Homoskedastic Errors

We assessed error homoskedasticity for our main model in two ways. First, a visual observation of Figure 4 indicates that the error variance may be higher for higher predicted values, which suggests heteroskedasticity. We also used the Breusch-Pagan statistical test, which produced a p-value of 0.34 for our model. Given the p-value is greater than 0.05, there is not enough evidence to reject the null hypothesis that there is no evidence for heteroskedasticity. This test does not prove that the model errors are homoskedastic; but, the results do indicate that the model errors are not clearly heteroskedastic. Nevertheless, because our visual examination of Figure 4 suggested that the errors may be heteroskedastic, we used robust standard errors in our analysis. Using robust standard errors is the primary way to alleviate concerns about failing to meet the homoskedastic errors assumption.

4.5 Normally Distributed Errors

Figure 5 shows the histogram of the errors from our main model. The histogram shows that the errors are not normally distributed, but instead are right-skewed. A Shapiro-Wilks test performed on the model residuals returns a p-value of 1.46×10^{-5}, meaning that enough evidence exists in the residuals to reject the null hypothesis that the residuals are normally distributed.

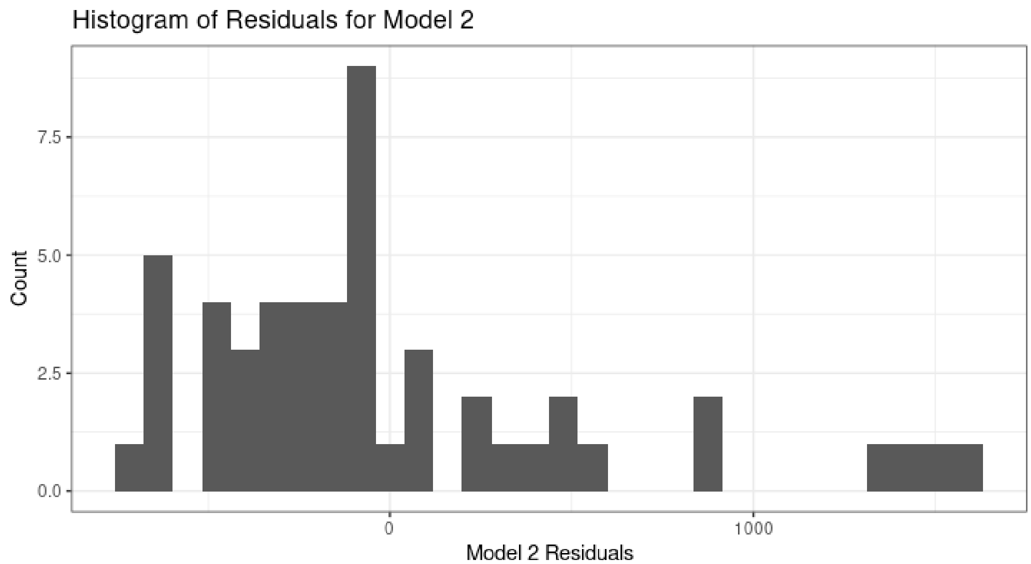

Figure 5. Histogram of residuals for our main model

We analyzed the five assumptions of the CLM and found that the IID and normally distributed errors assumptions are clearly violated. In contrast, our model appears to generally meet the LCE assumption and the model definitely meets the assumption of no perfect collinearity. While our model may not have strictly homoskedastic errors, we used robust standard errors in our analysis to mitigate any concerns with not meeting this assumption. Failure to meet some of these assumptions means that one must be cautious when interpreting the results of our model. Because our data are not independent, we cannot claim that the estimates for our model coefficients are unbiased (to claim this, we would need the IID, LCE, and no perfect collinearity assumptions to all be met). Moreover, because our model errors are not normally distributed, one should not trust any hypothesis tests associated with the model coefficients. Failure to meet all five CLM assumptions means that such hypothesis tests could produce false-positive and false-negative results at a higher rate than expected. Therefore, care must be taken when interpreting the results of our model.

5. Discussion of Omitted Variables

We developed a model to investigate our causal theory that stay-at-home and shelter-in-place orders caused increased unemployment. Our main model—model 2 in the regression table provided previously—includes three indicator explanatory variables: whether or not the state ever established a SAH/SIP order (this is our primary explanatory variable), whether or not the state restricted the movement of the general public, and whether or not the state provided an exemption from the SAH/SIP order for religious gatherings. However, there are other explanatory variables for which we do not have the data to quantify, but that can impact both unemployment and whether or not a state implements a SAH/SIP order. The omission of these variables from our model can bias our model's results. Here we discuss five omitted variables and how we expect the omission of each variable would bias our model results.

5.1 State Leadership Degree of Conservative Leaning

The effect that politics has on COVID-19 response is evident. For example, some states relaxed COVID-19-related restrictions earlier or took a somewhat laissez faire approach to handling the pandemic, whereas other states established strict and long-lasting restrictions (including SAH/SIP and business closures). In general, it appeared that states with conservative-leaning politics put less restrictions on their citizens and businesses in response to COVID-19. This notion is supported by the fact that Trump won the electoral votes in the 2020 Presidential Election for 10 of the 11 states that never established a SAH/SIP order.[3] Some have also noted that states with Republican governors tend, at least in recent history, to have lower unemployment rates than states with Democratic governors [1]. While it is difficult to adequately capture the complexity of a state's political environment as it pertains to unemployment and COVID-19 response in a variable, one could use past Presidential or governor election results as a proxy indicator variable, if required.

With this in mind, it is plausible that given less restrictions in more conservative-leaning states, there would be a smaller increase in unemployment between our comparison periods as well for states with a higher degree of conservative leaning. This indicates a positive omitted variable bias (OVB), with a magnitude that is difficult to discern. Because our estimated coefficient for the SAH/SIP indicator is also positive, the direction of the bias is away from zero.

5.2 Unemployment Benefits

The level of unemployment benefits a state provided during the pandemic is another important omitted variable. This variable is related to state-level politics, but state-level politics is a more holistic analysis of a state's political climate, whereas unemployment benefits considers this one specific government-provided benefit. We can reason that offering more

[3] Connecticut is the exception.

or better unemployment benefits during the pandemic would lead to more unemployment, as the part of the workforce that qualifies for the benefits would have more incentive to quit working and use those benefits. Similarly, with more or better unemployment benefits available, it is plausible that state governmental leadership would feel more comfortable issuing SAH/SIP orders because those in the population who lose their jobs due to the SAH/SIP order would be able to use those unemployment benefits. Therefore, the OVB for unemployment benefits is expected to be positive. Because our estimated coefficient for the SAH/SIP indicator is also positive, the direction of the bias is away from zero.

5.3 Sensationalism of Media Content Received by Populace

People receive news from a variety of print, radio, and online news organizations. While these organizations ostensibly tarry truth, they also clamor for our attention with striking headlines and often-overdramatic characterizations of what has taken place. Since early 2020, the media in the US and around the world has covered the spread and impact of the virus in great detail. While this variable is difficult to quantify because it would require knowledge of the most-watched media outlets and their content in a given state, we can still reason about how the content of COVID-19 media coverage could bias our model results. For instance, media coverage focusing on the dangers of the virus in a given state and the benefits of social distancing would likely provide more basis for a government to establish a SAH/SIP order with less resistance from the public. One can also argue that similar media coverage might lead to a larger increase in unemployment between our comparison periods due to more of the workplace not feeling safe in their workplaces or public spaces in general. Therefore, media coverage focused more on the dangers of the virus and the benefits of social distancing could lead to higher unemployment and a better likelihood of a SAH/SIP order, indicating positive OVB of unknown magnitude. Because our estimated coefficient for the SAH/SIP indicator is also positive, the direction of the bias is away from zero.

5.4 Community and Social Network Content

One's community and social network can have a major influence on how that person perceives events like the COVID-19 pandemic. Similar to professional media coverage, the content of experiences and opinions shared by a person's family, close friends, and even distant friends and professional contacts (e.g., from social networking sites) could affect a person's willingness to continue working during the pandemic. To the extent that the close communities and social networks of people in a given state tend more toward bad experiences with the virus (e.g., many infections and deaths) and the call for more government intervention to slow the spread of the virus, this could cause increased unemployment in the state. This same pressure on the government to act and recount the horrors of the virus could also create a more receptive environment for government interventions like SAH/SIP orders. Therefore, the OVB from communities and social networks is expected to be positive, meaning the direction of bias is away from zero.

5.5 Misreporting of COVID-19 Statistics

It is possible that certain counties or municipalities in a state could intentionally misreport COVID-19 cases or deaths in an attempt to lessen restrictions in the region (by under-reporting), garner more support and supplies from the governor (by over-reporting), or a variety of other reasons. If over-reporting was occurring in a state, this could have insinuated a surge in cases that could lead to higher unemployment rates (as a result of workers being fearful of continuing to work in public spaces) as well as a higher likelihood of states implementing SAH/SIP orders to control the spread of the virus. Therefore, the OVB of this omitted variable is expected to be positive, meaning the direction of bias is away from zero.

The direction of bias for all five of the omitted variables considered here is expected to be away from zero. This is problematic for two reasons. First, because the direction of the bias is away from zero in all cases, this means that the true coefficient on our primary variable (the indicator of whether or not a state has implemented a SAH/SIP order) should be something less than the positive value we are currently estimating. Second, the magnitude of the bias in each of the cases is unknown and difficult to reason about. Therefore, it is possible that the collective bias from these five omitted variables could indicate that the true coefficient on our primary variable is actually not positive, but zero or even negative. We have no way of knowing for sure.

6. Conclusion

Based on our analysis, we determined that none of our three linear models answer our research question of whether or not the SAH/SIP orders issued by states in response to COVID-19 caused higher unemployment. Even though the key explanatory variable was statistically significant in all three models, the models were unable to capture most of the variance in the data. Given that our explanatory variables were largely binary indicator variables, our ability to explain the changes in the unemployment rate was limited. We expect that there are other factors associated with the SAH/SIP orders and the state response to the pandemic that affect fluctuations in the unemployment rate.

While our linear models failed to adequately capture the variance in the data, we identified key issues with the analysis that would have brought even good results into question. Most importantly, we did not have IID data to use in our modeling, which means we cannot guarantee that our estimated model coefficients are unbiased. Moreover, we analyzed five omitted variables, all of which suggested that the true coefficient for our key explanatory variable may be zero or even negative (not positive, as all three of our models estimated). These issues, among others, would remain even if we had been able to develop a linear model that fit our data well.

References

[1] A GOP-led edge: Red states see less unemployment, more economic growth. https://thehill.com/opinion/finance/477923-a-gop-led-edge-red-states-see-less-unemployment-more-economic-growth, 2020. [Online; accessed 2-April-2021].

[2] Unemployment rate rises to record high 14.7 percent in April 2020, The Economics Daily. https://www.bls.gov/opub/ted/2020/unemployment-rate-rises-to-record-high-14-point-7-percent-in-april-2020.htm, 2020. [Online; accessed 1-April-2021].

[3] Unemployment Insurance Weekly Claims Data, United States Department of Labor. https://oui.doleta.gov/unemploy/claims.asp, 2021. [Online; accessed 2-April-2021].

[4] Jones D Bor J Lipson S Jay J Raifman J, Nocka K and Chan P. COVID-19 US state policy database. http://www.tinyurl.com/statepolicies, 2020. [Online; accessed 2-April-2021].

Appendix A: R Code

```
library(tidyverse)
library(lubridate)
library(sandwich)
library(lmtest)
library(stargazer)
library(gridExtra)
library(GGally)
library(car)

# Set working directory and read in data files
setwd("~/Lab 2")
stay_at_home <- read.csv("stay_at_home_20210402.csv")
unemployment <- read.csv("unemployment_data_20210402.csv")
closures <- read.csv("closures_and_reopening_20210402.csv")

# Rename important columns to be more useful
stay_at_home <- stay_at_home %>%
  rename(state = State,
         sah_sip_start = Stay.at.home.shelter.in.place,
         religious_exempt =
            Religious.Gatherings.Exempt.Without.Clear.Social.Distance.Mandate.,
         sah_no_mvmt_restriction =
            Stay.at.home.order.issued.but.did.not.specifically.restrict.movement.of.the.gene
         sah_sip_end = End.stay.at.home.shelter.in.place)
unemployment <- unemployment %>%
  rename(state = State,
         week_end = Filed.week.ended,
         unemp_percent = Insured.Unemployment.Rate,
         claims_new = Initial.Claims,
         claims_cont = Continued.Claims)
closures <- closures %>%
  rename(state = State)

# Filter out unnecessary columns and values
stay_at_home <- stay_at_home %>%
  select(state, sah_sip_start, religious_exempt,
         sah_no_mvmt_restriction, sah_sip_end) %>%
  filter(state != "Total")
unemployment <- unemployment %>%
  select(state, week_end, unemp_percent, claims_cont, claims_new) %>%
  filter(state != "Virgin Islands", state != "Puerto Rico")
closures <- closures %>% filter(state != "Total")
```

```r
# Wrangle columns into the correct formats
stay_at_home <- stay_at_home %>%
  mutate(sah_sip_start = mdy(sah_sip_start),
         sah_sip_end = mdy(sah_sip_end),
         has_sah_sip = !is.na(sah_sip_start),
         religious_exempt = as.logical(religious_exempt))
unemployment <- unemployment %>%
  mutate(week_end = mdy(week_end),
         claims_cont = as.numeric(gsub(",", "", claims_cont)),
         claims_new = as.numeric(gsub(",", "", claims_new)))

# Wrangle unemployment data prior to merging
baseline_unemp <- unemployment %>%
  filter(week_end >= mdy("02/01/2020"), week_end <= mdy("02/29/2020")) %>%
  group_by(state) %>%
  summarise(unemp_feb = mean(unemp_percent))
later_unemp <- unemployment %>%
  filter(week_end >= mdy("04/25/2020"), week_end <= mdy("06/13/2020")) %>%
  group_by(state) %>%
  summarise(unemp_mayjune = mean(unemp_percent))
max_unemp_date <- unemployment %>%
  group_by(state) %>%
  arrange(desc(unemp_percent)) %>%
  summarise(max_unemp_date = first(week_end))
unemployment <- left_join(baseline_unemp, later_unemp, by = "state") %>%
  left_join(., max_unemp_date, by = "state") %>%
  mutate(delta_unemp_perc = (unemp_mayjune - unemp_feb) / unemp_feb * 100)

closures <- closures %>%
  mutate(schools_1 = ifelse(closures$Closed.K.12.public.schools != "0", 1,
      0),
         daycare_1 = ifelse(closures$Closed.day.cares != "0", 1, 0),
         nursing_1 = ifelse(closures$Banned.visitors.to.nursing.homes !=
            "0", 1, 0),
         nonessential_1 =
            ifelse(closures$Closed.other.non.essential.businesses != "0",
            1, 0),
         gyms_1 = ifelse(closures$Closed.gyms != "0", 1, 0),
         restaurants_1 = ifelse(closures$Closed.restaurants != "0", 1, 0),
         movies_1 = ifelse(closures$Closed.movie.theaters != "0", 1, 0),
         bars_1 = ifelse(closures$Closed.bars != "0", 1, 0),
         casinos_1 = ifelse(closures$Closed.casinos != "0", 1, 0),
         overnight_1 = ifelse(closures$Closed.businesses.overnight != "0",
            1, 0),
```

```r
        gyms_2 = ifelse(closures$Closed.gyms.x2 != "0", 1, 0),
        restaurants_2 = ifelse(closures$Closed.restaurants.x2 != "0", 1,
           0),
        movies_2 = ifelse(closures$Closed.movie.theaters.x2 != "0", 1, 0),
        bars_2 = ifelse(closures$Closed.bars.x2 != "0", 1, 0),
        casinos_2 = ifelse(closures$Closed.casinos.x2 != "0", 1, 0),
        hair_2 = ifelse(closures$Closed.hair.salons.barber.shops.x2 !=
           "0", 1, 0),
        restaurants_3 = ifelse(closures$Closed.restaurants.x3 != "0", 1,
           0),
        bars_3 = ifelse(closures$Closed.bars.x3 != "0", 1, 0))

closures_summarized <- data.frame(
  state = closures$state,
  closures_one_tot = rowSums(closures %>% select(ends_with("_1"))),
  closures_two_tot = rowSums(closures %>% select(ends_with("_2"))),
  closures_three_tot = rowSums(closures %>% select(ends_with("_3")))
)

data <- left_join(stay_at_home, unemployment, by = "state") %>%
  left_join(., closures_summarized, by = "state")


# Start creating more variables of interest
data <- data %>%
  mutate(
    sah_no_mvmt_restriction = ifelse(
      sah_no_mvmt_restriction == "0", FALSE, TRUE))

# Create models
model_1 <- lm(delta_unemp_perc ~ has_sah_sip,
              data=data)
model_2 <- lm(delta_unemp_perc ~ has_sah_sip + sah_no_mvmt_restriction +
                religious_exempt,
              data=data)
model_3 <- lm(delta_unemp_perc ~ has_sah_sip + sah_no_mvmt_restriction +
                religious_exempt + closures_one_tot + closures_two_tot +
                closures_three_tot,
              data=data)

# Regression table
se.model1 = coeftest(model_1, vcov = vcovHC)[ , "Std. Error"]
se.model2 = coeftest(model_2, vcov = vcovHC)[ , "Std. Error"]
se.model3 = coeftest(model_3, vcov = vcovHC)[ , "Std. Error"]
```

```
stargazer(model_1, model_2, model_3,
         title = "Regression Table",
         align = TRUE,
         dep.var.labels = c("Change in Unemployment Rate"),
         covariate.labels = c(
           "SAH/SIP Order? (Y/N)",
           "SAH/SIP No Movement Restrictions? (Y/N)",
           "SAH/SIP Religious Exemptions? (Y/N)",
           "Closures: # Sectors in Phase 1",
           "Closures: # Sectors in Phase 2",
           "Closures: # Sectors in Phase 3"),
         omit.stat = c("LL", "ser", "f"),
         se=list(se.model1, se.model2, se.model3),
         no.space=TRUE,
         type="latex")

# F-test between models
anova(model_1, model_2)
anova(model_1, model_3)
anova(model_2, model_3)

# VIF analysis


# Add residuals and predictions to ‘data‘
data <- data %>%
  mutate(
    model_1_preds = predict(model_1),
    model_1_resids = resid(model_1),
    model_2_preds = predict(model_2),
    model_2_resids = resid(model_2),
    model_3_preds = predict(model_3),
    model_3_resids = resid(model_3)
  )

# Breusch-Pagan test
bptest(model_2)

# Shapiro-Wilks test
shapiro.test(data$model_2_resids)

# Plots for paper and presentation
# Dot plot of SAH/SIP and max unemployment dates
data %>%
  select(state, sah_sip_start, max_unemp_date) %>%
```

```
  gather(metric, value, -state) %>%
  ggplot(aes(x = value, y = fct_rev(state), color = metric)) +
  geom_point(size = 2) +
  theme_bw() +
  labs(x = "2020 Date", y = "State",
       title = "SAH/SIP Start Dates and Maximum Unemployment Dates by
           State") +
  scale_color_discrete(labels = c("Date of Maximum\nUnemployment",
      "SAH/SIP Start Date")) +
  theme(legend.title = element_blank())

# Histogram of change in unemployment rate
resp1 <- data %>%
  ggplot(aes(x = delta_unemp_perc)) +
  geom_histogram() +
  labs(x = "Change in Unemployment Rate (%)", y = "Count",
       title = "Histogram of Changes in Unemployment Rate") +
  theme_bw()
# Q-Q plot of change in unemployment rate
resp2 <- data %>%
  ggplot(aes(sample = delta_unemp_perc)) +
  stat_qq(size = 2) +
  stat_qq_line() +
  theme_bw() +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles",
       title = "Normal Q-Q Plot of Changes in Unemployment Rate")
grid.arrange(resp1, resp2, nrow = 2, ncol = 1)

# Residual-predictions plot for main model
data %>%
  ggplot(aes(model_2_preds, model_2_resids)) +
  geom_point() +
  stat_smooth() +
  theme_bw() +
  labs(x = "Model 2 Predictions", y = "Model 2 Residuals",
       title = "Residuals vs. Predictions Plot for Model 2")

# Residual histogram for main model
data %>%
  ggplot(aes(model_2_resids)) +
  geom_histogram() +
  theme_bw() +
  labs(x = "Model 2 Residuals", y = "Count",
       title = "Histogram of Residuals for Model 2")
```

```r
# Pearson correlation plot of all variables
data %>%
  mutate(has_sah_sip = as.numeric(has_sah_sip),
         sah_no_mvmt_restriction = as.numeric(sah_no_mvmt_restriction),
         religious_exempt = as.numeric(religious_exempt)) %>%
  select(has_sah_sip, sah_no_mvmt_restriction, religious_exempt,
         closures_one_tot, closures_two_tot, closures_three_tot) %>%
  ggcorr(method = c("pairwise", "pearson"),
         label = TRUE,
         hjust = 0.9,
         layout.exp = 1)
```