(A)

w203_sp_21_sec_04 Lab 2, Final Report

[redacted]

75 / 75

QUESTION 1

1 Lab 2 Final Report 75 / 75

✓ - 0 pts Correct

- How does the CARES act and other pandemic-related job aid affect this story?
- Are there some sectors that you think will be more strongly, or more directly affected?
- You've done a good job of working with dimension reduction techniques. However, there are time-based and aggregation-based considerations that you're bringing on.

↳ Is it possible - aggregation of unemployment claims going to be able - EVER - to tell a timely-detailed story?

Page 1

Red Light, Green Light: The effects of State-Level Public Health Policies to Reduce COVID-19 Spread on Unemployment

Kanika Mahajan, Malachy Moran, Natali Ojeda, and Inigo Verduzco

4/19/2021

Contents

this feels like a speculation

1 Introduction: Holy Whac-A-Moley

During the COVID-19 pandemic, State governments have faced a critical policy question: Should they focus on reducing the spread of COVID-19 by imposing strict public health policies that reduce contagion (business closures, stay-at-home orders, travel restrictions, etc.)? Or, should they focus on reducing the potentially unprecedented economic impacts of the pandemic by not closing their economies, at the cost of higher COVID-19 cases?

The answers to this policy dilemma is not straightforward. Focusing exclusively or zealously on reducing COVID-19 cases and spread can have important economic consequences due to reduced economic activity, businesses going bankrupt and people becoming unemployed. On the other hand, focusing exclusively on minimizing economic costs can lead to higher COVID-19 cases, saturated medical facilities, and increasing mortality due to COVID-19.

In most cases, due to a more active public health stance to reduce cases or fears of over saturation of medical facilities (e.g. ICU beds), State governments have had to opt for policies to limit activities and business closures. But, at what economic cost?

While we agree that addressing the COVID-19 pandemic should remain a priority for State governments, understanding the full extent these policies have had is important. This information can allow policymakers to make public policy decisions with full knowledge of the potential costs and benefits of these measures.

In this paper we aim to identify the effects of public health policies to contain the COVID-19 pandemic on economic outcomes. Given the unprecedented levels of unemployment that have been observed since the beginning of the pandemic, and the economic implications they entail to individuals and local economies, we are particularly interested in exploring the effect that public health policies to curb COVID-19 infection rates have had on unemployment levels across the 50 U.S. states.

While we consider that our theoretical model of transmission mechanisms from public health policies correctly represents existing causal pathways, our estimation strategy and model yielded estimates that are not statistically significant. Given the challenges that we found in our data (e.g. cross section of only 50 states), we have reasons to believe that, even if we had found statistically significant estimates, these would likely be biased. As a result, our study cannot provide substantive evidence to support a causal relationship. However, this does not mean that the relationship does not exist; only that given the limitations of our data and available statistical models, we cannot derive conclusive results regarding the causal relationship and its magnitude between unemployment and COVID policies. Given this, the models and theory outlined in this paper remain important, as they can provide guidance to other researchers about avenues of inquiry to attempt or to avoid.

This rest of this paper is structured as follows. Section 2 states our main research question, describes the theoretical causal model we assume to identify the causal relationship of interest, and our overall study design. In Section 3, we briefly describe our data and variables, including the different transformations we performed to the data to get the final set of variables we used in our estimation model. Section 4 describes our main statistical models. The following section, Section 5, presents results from our regressions and tests on the assumptions of our statistical models. Finally, in Section 6 we provide our main conclusions and briefly discuss the implications of our results.

2 Research Question, Theoretical Model and Study Design

2.1 Research Question

Our main goal for this research is to examine whether there is any statistically significant relationship between state-level policies to curb the COVID-19 pandemic and economic outcomes in the 50 U.S. states. In particular, we are interested in exploring whether state-level business closure policies to curb the COVID-19 pandemic led to higher state-level unemployment levels in the U.S.

good.

how strong a recommendation are you willing to make from a null finding?

The implications of these findings could be useful to policymakers trying to decide on the right mix of policies to contain COVID-19 spread while mitigating the economic costs to their constituents and local economies. If the evidence shows that there is no meaningful relationship between pandemic-related public health policies and unemployment levels, policymakers may want to consider very aggressive pandemic containment policies, as they are unlikely to affect economic outcomes like unemployment. If, on the other hand, the evidence suggests a strong positive association between pandemic-related public health policies and unemployment levels (i.e. more restrictions leading to more unemployed individuals), policymakers may have to ponder the optimal level and stringency of these policies to try to minimize their effects on economic outcomes and unemployment levels.

2.2 Underlying Causal Model

In this section we establish the causal relationships between public health policies to limit the spread of COVID-19 (COVID policies), and unemployment levels. We also discuss any possible interference from omitted variables. We will be constructing the causal models using the data sources section, and the DAGitty web interface.

2.2.1 Simplified Model

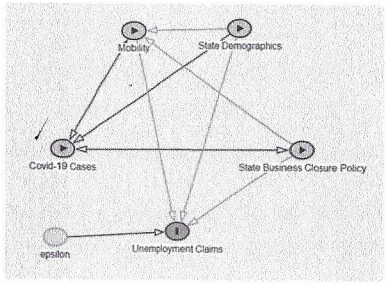
We argue that the causal path between COVID policies and unemployment levels goes as follows: restrictions on mobility and business operations (hours, number of clients, full closures, etc.) as well as people's natural risk calculations of potential infection (e.g. how likely you are of being infected if you go eat at a restaurant) reduce the number of jobs in the market.

This happens because, COVID policies (e.g. mandatory closures and stay-at-home orders) constrain business operations, reducing overall revenues. And, with less people consuming goods and services due to fears of getting infected with COVID while eating out or buying in shops, there is also a demand-side effect, with the net effect being a reduction in revenues. This leads businesses to reduce hiring or even furlough or fire workers to reduce costs as their revenues go down.

Finally, less demand for work and a relatively stable supply of labor in the short-term, leads to higher unemployment levels (unemployment claims) - in the longer-term this relationship may not bee as straightforward as some people may leave the work force altogether.

Our proposed simplified causal model is shown below. We will traverse around the model from the bottom in a clockwise direction and explain each piece.

This is a nice treatment for this class



Unemployment Claims

Do you know what/whether this well-fitness is present in the data that you have? What is the right time scale?

The primary variable we are attempting to explain is the number of unemployment claims filed during the current month, standardized by a state's population. We see this being affected by the other variables in a few ways.

State business closure policies, such as closing restaurants or gyms should result in higher unemployment claims, as business owners are forced to lay-off or furlough workers to reduce costs.

Unemployment claims will also increase with decreased mobility (which can be seen in an increase in "residence mobility" scores, indicating people are staying home). This would indicate that people are not out purchasing things and driving revenue into businesses, even if they might **not** be restricted by government policy. This in turn would cause businesses to lay off workers.

State demographics have an impact on the number of unemployment claims, though it is unclear in which direction it will influence them. States with a high proportion of children too young for the workforce, or retirees who have left the workforce, would expect to see less of an impact from COVID-19. On the other hand, states with a less educated population who are more likely to have service jobs that cannot be done from home, might see a greater impact.

Good argument here.

Importantly, we assume that COVID-19 case counts do not directly impact unemployment claims. This seems likely for a number of reasons. For example, it is illegal for businesses to fire workers for becoming sick with COVID, or for taking sick time because of COVID. All other possible impacts of COVID-19 are thought to be mediated through our other variables. One could argue that higher infection rates may result in more unemployment if high infection rates translate into more sick workers and, thus, less revenue for businesses (who cannot provide the goods and services because of this). While this may be true in cases where demand is not affected, since demand for goods and services has gone down in parallel, we think that this channel is highly unlikely to come at play if at all.

why not? what interveeds?

COVID-19 Cases

COVID-19 cases, while not expected to have a direct impact on unemployment claims, are still an import variable to account for in our model, as they do impact our other variables of interest.

COVID case counts are expected to both affect and be affected by population mobility. Increased mobility outside the home is expected to increase the number of cases by increasing the opportunities for transmission. Likewise, increased COVID cases are thought to drive down mobility by scaring people into staying home more.

COVID cases are expected to be impacted by state demographics, but it is unclear in which direction. Certain populations, such as the elderly, are at increased risk of catching a serious enough case of COVID that they seek-out testing. However, other age groups, such as the young, are more likely to catch asymptomatic COVID, which could result in them spreading the disease further.

State business closure policies are thought to both affect and be affected by COVID-19 case counts. State and local governments are known to use infection rates as a major factor when deciding whether or not businesses should close. The intended effect of closing down businesses is to reduce cases, which eventually leads to the re-opening of businesses, so the effect is cyclical in nature.

Mobility

Population mobility is thought to have a direct impact on unemployment claims as already mentioned. It is important to note that we believe mobility will have an impact on unemployment that is distinct from business closures.

State's COVID policies will have an effect on mobility, as people are more likely to stay home if they are not allowed to go to their jobs or if there is nowhere open for them to go to. Thus, increases in business closures will have a negative impact on overall mobility. However, just because a business is open doesn't mean that people will go there. Mobility therefore measures something distinct from state COVID policies.

Demographics will also have an impact on mobility, but the direction of that impact is unclear. At risk populations, such as the elderly, will be more likely to stay home even if businesses are open unless they

I think that "demographics" as a concept is too broad to be very useful.

absolutely need to leave. States with a less educated population however are more likely to see increased mobility, as those with service jobs may be forced to be more mobile when businesses re-open.

State Demographics

As was already ~~been stated in the previous sections,~~ state demographics are expected to have an impact on unemployment claims, mobility and COVID cases.

State Business Closure Policy

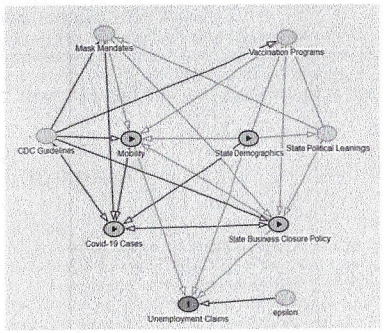
The effects of state closure policies on unemployment is the main effect that we wish to study. ~~As already stated~~ it is expected to affect unemployment claims, COVID case counts and mobility. And, we expect it to be affected by COVID case counts.

Epsilon

Our error term (epsilon) is everything else that could affect unemployment claims and that is independent of any of the other variables in our model. This could include factors such as mechanization of jobs, outsourcing, employee relocation, families having children, or voluntary worker non-participation in the workforce. We assume that they are no factors in the epsilon term that directly effect both the outcome variable and the explanatory variables.

2.2.2 Omitted Variables

No chain of causality, of course, is ever this simple, and we have identified several omitted variables that could have an effect on our final model. They are outlined in the causal diagram below. We will go through each in turn and discuss how they might affect our final explanatory variables.



CDC Guidelines, Mask Mandates and Vaccination Programs

Though not necessarily mandatory, the guidelines on policy put out by the Centers for Disease control do have an impact on several variables in our data set and are not otherwise being accounted for in our models. Properly implemented guidance on workplace safety from the CDC would allow the population to go back to work safely without increasing COVID cases. This would tend to drive up mobility, spur fewer shutdowns and drive down COVID cases (which also increases mobility and decreases business closures). As such, we expect effective CDC risk mitigation guidelines to bias the coefficient of state business policy and mobility away from zero, and bias the effect of COVID-19 towards zero. Vaccination programs have almost identical effects.

The same is generally true of mask mandates, although their effect is not as clear. Mandating masks allows more people to go to work and businesses safely, without a rise in COVID cases. The effect on mobility may be mitigated by some people being less likely to go out if they must wear a mask to do so. Thus we expect mask mandates to bias the coefficient for COVID policies towards zero, and COVID-19 cases away from zero, but to have little effect on mobility.

State Political Leanings

While we would hope that the political leanings of a state would have little to no effect on COVID outcomes, we know that this is likely not the case. More conservative states are less likely to close down businesses regardless of COVID-19 statistics and other measurable factors. This would decouple policy to some extent from demographics and COVID-19. This has the overall effect of biasing the effect of COVID policies towards zero. COVID cases would rise, businesses would not close, but we would still see the unemployment counts rise as case counts exerted their effects through mobility. However, the effect may not be as large as with COVID policies as businesses may still operate (even if at reduced capacity), leading to less overall unemployment because of it.

Final Model Thoughts

There are several unaccounted for variables in the design that we were unable to measure with our available data. While this may be a measurement problem, we believe that it is unlikely (but not impossible) that there are other variables that have both a direct impact on unemployment and an explanatory variable for which we have not already accounted.

2.3 Study Design

Our main interest is in determining the effect of COVID policies on unemployment levels at the state level, across all states in the US. In particular, we are interested the short-term effects these policies may have on unemployment levels. Since official numbers of unemployment rates tend to be published with some lag, and capture the effects after some time has elapsed, we focus on initial unemployment claims as the our main dependent variable of interest. While this variables does not fully capture the actual unemployment levels, it is typically uses as a good immediate proxy for total unemployment levels.

One interesting aspect of the evolution of unemployment levels during the COVID-19 pandemic in the US is the unprecedented levels of unemployment observed at the start of the pandemic (March-April 2020). This coincided with the first wave of COVID-19 infections in the US. However, it is unclear whether this was a one-off correction (possibly an overreaction by the market) or a more permanent causal relationship between COVID-19 infections, COVID policies and unemployment levels.

Given the causal model presented above, we would expect that higher unemployment levels are observed during peak infection levels. Since we do not have enough information on COVID policies and infections for the first wave (the cases were limited to very few states) of infections in March of 2020, we test our model against data from the most severe infection wave to-date, January 2021. This coincides with the 3rd infection wave, by far the one with the highest infection numbers since the start of the COVID-19 pandemic in the US. As an additional test to our model, we also run all tests on data from the second infection wave, for the month of July 2020.

Since data for our different variables of interest is published at different time intervals, we aggregated our data to the monthly level. However, our analysis abstracts away from time dynamics and simply tests the relationship for a specific month for all 50 states (cross section). As mentioned above, our main results are for January 2021 but, we also ran our estimation model for July 2020 to test our model against a different month.

Can/could you expand this to an arbitrary month?

Does your model actually have this expectation? What you've just written in the PP before is not captured in the model you have drawn.

3 Data and Variables of Interest

In this section we provide some information on the main datasets and transformations we performed to obtain the variables we used in our statistical specifications. We discuss data sources as well as the methodology we followed in our transformations.

The state COVID policy data we will analyze comes from the COVID-19 US State Policy Database (CUSP). The state policies we will be considering will be: stay at home/shelter in place issued and business closures (non-essential businesses, gyms, movie theaters, bars, K-12 schools). We will get data on COVID infection rates within each state from the New York Times Database on COVID case rates. The mobility data comes from [Google's COVID - 19 Community Mobility Reports] (https://www.google.com/covid19/mobility/). Finally, we will use demographic data from the American Community Survey (ACS) form the U.S. Census Bureau, including total population, gender distribution and median age to control for other variables that may be related to unemployment levels at the state level. Washington DC, while originally included in all the data sets, was removed. The reasons for this are two-fold: as will be explained in the relevant sections, it was an outlier for many of our variables, it is also an outlier in terms of how its government operates, being run more or less directly by congress.

3.1 Government COVID Policies

3.1.1 Data

For our data on government policy, we began with the CUSP, compiled by the Boston University School of Public Health. This dataset consists of 222 variables across the 50 states plus Washington D.C.. The data set gives the start and end dates of various types of government policies that a state could have put in place to limit the spread of COVID-19. The dataset also contains some indicator variables for nuances like religious exemptions and unique identifiers for each state.

As the rest of our data was rolled up into a monthly total, whatever variables were to be used for policy would also need to be at a monthly level. There are essentially two possible options for ways to accomplish this task. The first would be to count up the number of days in a month where the policy was in effect, and code it as a discrete variable. The second would be to decide a cutoff point for the number of days during a single month that a policy needed to be in effect, and code an indicator variable that takes on the value "1" if the policy was in place for at least that many days and "0" otherwise.

The difficulty with using a count of "total days a policy is in effect" is that it implies a linear relationship that cannot be strongly supported. For example, there is no reason to think that closing restaurants for dine in service for two days has twice the effect on unemployment as closing them down for only one. Likewise, there are confounding factors that could affect this relationship. It is easy to imagine a situation where a restaurant owner, facing a month long shut down, decides to pivot to online ordering and delivery, keeping cooks and other employees on staff who would have been furloughed during a shutdown that only lasted three weeks.

We then decided to use the cut-off date and indicator method. The natural number to use for this cut-off was half the length of the month, this allows for easy interpretability of our indicator variables. If the policy was in place for at least half of the month it will be coded as a "1", otherwise it will be coded as a "0". The indicator therefore codes whether the month was more characterized by the policy being in place or not in place.

Given that we will only be using a single month's worth of data, and will therefore have only 50 data points, it was not possible to use all, or even most, of the possible policy variables. A three step process was undertaken to reduce the number of variables under consideration

First, many of the 222 variables in the data set are simply coded for multiple implementations of the same policy. For example, a state may have ordered the closure of bars, reopened them, and then re-closed them again during a spike in cases. Several states had gone through as many as three rounds of the same policy

similarity - does a closure in the second half of the month affect the first half of the month?

At this point, when you're building scales like this: show the reader that this scale is "convergently valid": that it is producing numbers that make sense.

the coordinates of each state along these two dimensions, and it is these coordinates that will serve as the variables for our analysis.

3.2 State Characteristics

3.2.1 Data

We included state characteristics and demographic variables in our model to control for them and statistically remove their effect from the rest of the variables. We utilized the U.S. Census Bureau Data to get eight variables we considered to be relevant in explaining unemployment across the 50 states plus the District of Columbia: Total population, total area (in square miles), number of people with education less than high school as an education indicator, people less than 18 years old to remove the effect of this population in our indicators, number of females as a gender indicator, median age in the state, number of whites as a race indicator and number of hispanic or latino as an ethnicity indicator.

To remove the effect of the absolute values in each state, we created six variables in percentages: % of females % of non-white population % of hispanic population % of U.S. citizens 18+ years old Population density: Total population / Total area (in square miles) % of education "less than high school" in population 18+ years old

3.2.2 Variables

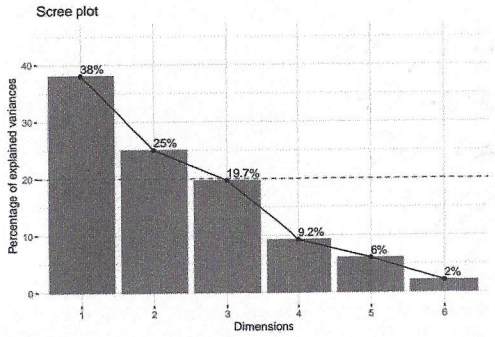
Keeping in mind we have only 50 states plus D.C., we wanted to keep the least number of variables in an effort to keep our degrees of freedom as high as possible. As stated in the data section, we kept 6 continuous variables from the U.S. Census Bureau and we wanted to reduce the number of them by keeping as much variance in the data as possible. For that purpose, we used PC (Principal Components), a dimensionality reduction technique that projects each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. Basically, a change of basis of the data. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data. The ith principal component can be taken as a direction orthogonal to the first i-1 principal components that maximizes the variance of the projected data.

We used the library devtools and ggplot_pca from the AMR package in R to get the principal components:

```
## Importance of components:
##                           PC1    PC2    PC3     PC4    PC5     PC6
## Standard deviation     1.5109 1.2250 1.0864 0.74370 0.6005 0.34998
## Proportion of Variance 0.3805 0.2501 0.1967 0.09218 0.0601 0.02041
## Cumulative Proportion  0.3805 0.6306 0.8273 0.91949 0.9796 1.00000

## Standard deviations (1, .., p=6):
## [1] 1.5109359 1.2250400 1.0863708 0.7436954 0.6004814 0.3499821
##
## Rotation (n x k) = (6 x 6):
##                                PC1        PC2        PC3          PC4         PC5
## female_p             -0.2220729  0.6001382 -0.3502413  0.462995624 -0.46967213
## non_white_p          -0.4756633  0.2436446 -0.2129312 -0.775108488  0.03108741
## hispanic_p           -0.4402909 -0.4839353 -0.2219672  0.369464109  0.32425885
## non_ctzen18plus_p    -0.5689604 -0.3186402 -0.1430987  0.044624549 -0.25673154
## pop_density          -0.3599253  0.4836834  0.3929593  0.215279891  0.62132243
## high_schl18plus_p     0.2780465  0.1039078 -0.7796331 -0.001172105  0.47047340
##                                PC6
## female_p             -0.1813614
## non_white_p          -0.2593887
## hispanic_p           -0.5301260
## non_ctzen18plus_p     0.6974043
## pop_density           0.2229352
## high_schl18plus_p     0.2876091
```


As observed above, with the three first components, we can explain about 83% of the variance in the data.

Scree plot below represent the proportion of variance explained by each principal component. We know that the amount of variance we would expect to have in a given dimension by pure chance is 1/(6-1)=1/5 or about 20% since we have six dimensions. Scree plot presents that threshold in red. Because of it, we only considered the first three components. The third one was selected as well due to its close proximity to the threshold.



To understand the elements contained in each principal component, we will visualize in the graphs below how the samples relate to one another in our PCA (which samples are similar and which are different) and will simultaneously reveal how each variable contributes to each principal component. Distance along X-axis indicates correlation with the first dimension, and distance along Y-axis indicates correlation with the second dimension. Principal components are linear combinations of all the individual variables and are typically named after the concepts having the highest correlations.

The left graph shows that percentage of U.S. citizens having 18+ years old (non_ctzen18plus_8) followed by percentage of non-white population (non_white_p) as well as percentage of hispanic population (hispanic_p) have the highest correlation in absolute value with the first principal component. All of these correlations are negative, hence, we named the principal component "white non-hispanic citizens" (white_nonhisp_cit). The graph to the left shows a similar view but it includes which point corresponds to which state. For example, some states known for having a high percentage of hispanic/latino population are California (CA) and Texas (TX) and the arrow of percentage of hispanic population is precisely pointing towards those states.

during the over one-year long period that the data set covers. These columns could be collapsed together into a single variable.

Second, some policies contained in the dataset were not expected to have a significant impact on unemployment claims. These included policies such as mask-mandates and the prioritization of vaccines. While mask-mandates may allow some businesses to remain open, most of this information could be more thoroughly captured through other variables. The same applies to vaccine prioritization policies. Policies not thought to directly contribute to explaining the unemployment rate were dropped from the data set.

Finally, since we will be examining the month of January 2021, any policy which 2 or fewer states had in place during at least half of January were dropped for lack of variation.

This left us with 6 policies left which comprised our dataset:

1. Was there a curfew?
2. Were restaurants closed?
3. Were gyms closed?
4. Were movie theaters closed?
5. Were bars closed?
6. Were casinos closed?

3.1.2 Variables

State COVID policy data is provided as dates on which different states had policies in effect and dates they were dropped. Using this information every state will be categorized as either a 1 (policy in effect) or 0 (policy not in effect) for every month as long as the policy was in effect for at least 15 days that month.

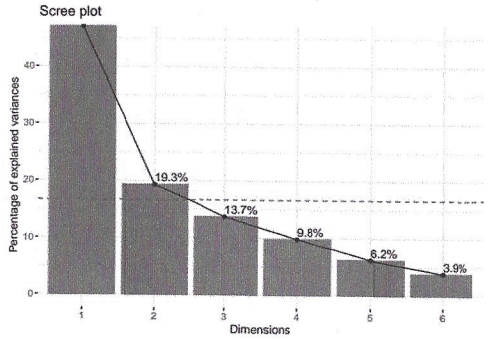
In an effort to further reduce the number of variables for our model (given our limited total number of observations for a single month, i.e. 50), we will be employing Multiple Correspondence Analysis (MCA). MCA is the generalization of the PCA (Principal Component Analysis) method for dimensional reduction that is applicable to categorical or indicator variables (Kassambara 2017). As with PCA, we will be attempting to create orthogonal components, called dimensions, that capture as much of the variation contained within our indicators as possible. This will be accomplished using the `FactoMineR` package in R.

In order to keep the degrees of freedom high for our model high, we will be attempting to reduce our six active indicator variables to no more than three dimensions. After running our MCA, we find that we have several good candidates.

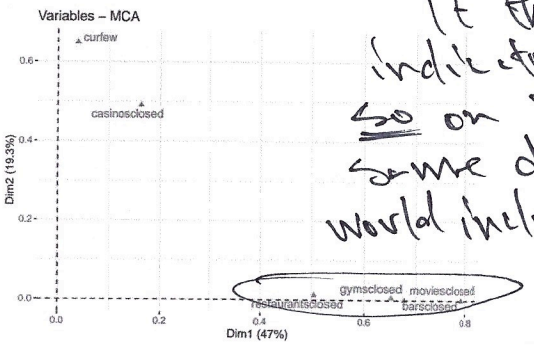
We are able to account for 80% of the variance in our indicators within the first three components. Given that we have six total dimensions, the amount of variance we would expect to have in a given dimension by pure chance is $\frac{1}{6}$ or about 16%. If we plot the amount of variance we can capture on each dimension in a scree plot with a cutoff line at 16% we can see that only our first two dimensions reach this threshold.

this is reasonable — but why are you trying to preserve degrees of freedom? what do you want to achieve?

what does it mean to be "more thoroughly captured?



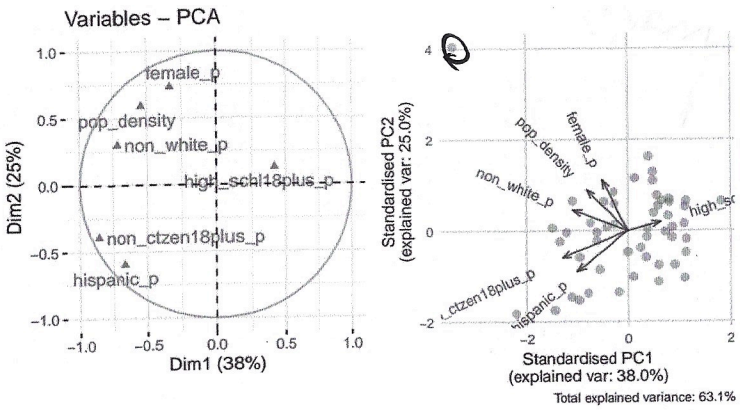
As such, only our first two dimensions will be considered going forward. This will still allow us to account for 66% of the variance in our data. In order to understand and better interpret these two dimensions, it is important to understand which of our indicator variables has the most influence on the coordinates in that dimension. We can see this most clearly by plotting the correlation between each variable and our dimensions on a scatter plot. Distance along X-axis indicates correlation with the first dimension, and distance along Y-axis indicates correlation with the second dimension.



if these indicators are so on the same dimension, would including a single proxy just do the job?

As we can see from our plot, the two dimensions that we have from our MCA divide our indicators very cleanly into two distinct groups. The closure of restaurants, gyms, movie theaters and bars are very strongly correlated with dimension one, but not at all correlated with dimension two. On the other hand, the implementation of a curfew and the closure of casinos are strongly correlated with dimension two, and only slightly correlated with dimension one.

This gives us our interpretation of the two dimensions. Dimension one is a measure of general business closures, so we will call that dimension `bus_closed`. Dimension two is a measure of late night or entertainment venues are opened, so we will call that dimension `nightlife_closed`. Our MCA gives us access to



Following a similar methodology, we obtained and calculated the second and third principal components, and named them "Non-hispanic females in high density areas" (nonhisp_fem_highdens) and "Less educated population in high density areas" (lesseduc_highdens).

As observed below when obtaining the statistics of these newly created variables, we noticed that all of them have outliers. The first variable, in the extreme left values, and the second and third variables in the extreme right values.

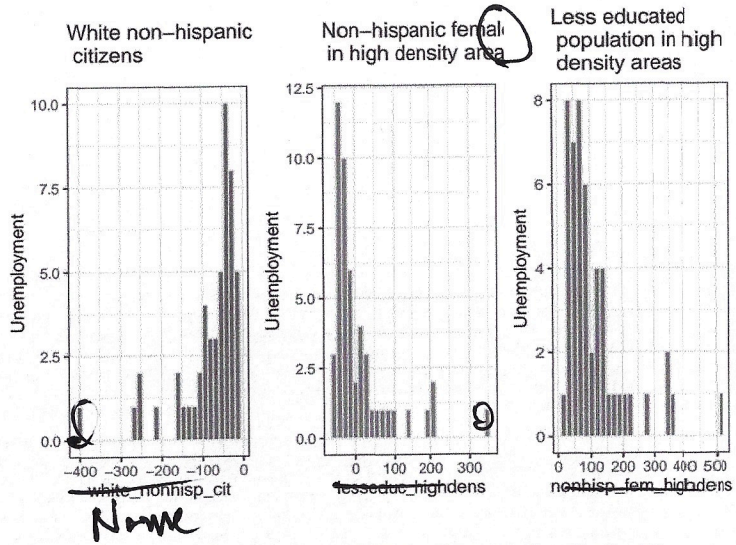
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -3776.68   -101.44    -52.01   -154.22    -33.02    -11.74
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -60.98  -36.80  -20.37   89.13   25.89 4024.87
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    23.57   51.65   82.09  212.21  134.02 5060.44
```

We noticed that the District of Columbia was causing all of these outliers and decided to remove this observation. As a result, we got the histograms of our 3 new variables below:

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Distribution of the first variable is skewed to the left and the other two are skewed to the right.

3.3 Mobility

We are interested in using mobility data as one of the control variables in our model because as policies dictate rules around mobility, mobility data gives us a reflection of the actual movement of our population during COVID times. We would expect that restricted mobility or high residential mobility would in general increase unemployment as people are staying at home more.

3.3.1 Data

We used [Google's COVID - 19 Community Mobility Reports] (https://www.google.com/covid19/mobility/), to extract data on movement of people in response to policies during COVID to contain the virus. The data consists of mobility trends over time for 50 states and the District of Columbia across 6 different categories where each category consists of places with similar characteristics impacted by similar policies. The 6 broad categories were:

- Retail and recreation
- Groceries and pharmacies
- Parks
- Transit stations
- Workplaces
- Residential

The 2021 US mobility data had data which was available for 92 days from January 2021 - April 2nd 2021 at both county and state levels. For each date, mobility data was reported as a positive or negative percentage change from the baseline. Google's database defines baseline day as the median value from the 5-week period between Jan 3 – Feb 6, 2020.

We used January 2021 data for our analysis. As our unit of analysis for this project has been at the level of state and monthly, we aggregated daily data for the month of January by taking average of daily percentage changes for every state. The mobility dataset contained data at all three levels: national, state and county level. We created an indicator variable that identified state level data and extracted data for the month of January filtering on the indicator variable. As part of the data cleaning process we had to re-format the date variable. Missing data was minimal and only 5 observations were missing for the park's mobility variable and were not included in calculating monthly average for states.

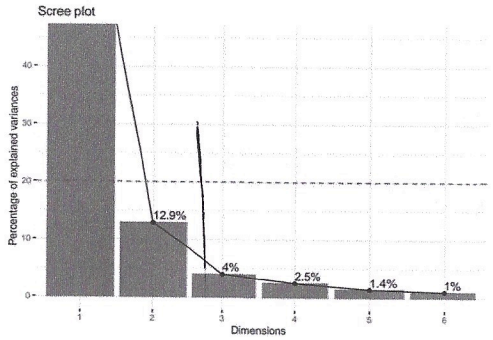
3.3.2 Variables

Considering we are planning to include data on government policy, state census and covid cases, we again decided to do Principal Component Analysis (PCA) on mobility data as well to reduce the number of variables we will be using, in an effort to keep our degrees of freedom high. As mentioned before PCA helps in identifying the coordinates which represent the most significant variance in our dataset. Each principal component (PC) sums up some percentage of the total variation in the dataset and are represented as eigenvectors with eigenvalues where the eigenvector with the highest eigenvalue represents the first principal component and the one capturing the maximum variance. Variables highly correlated contribute strongly to a principal component.

All variables of interest in our mobility dataset are continuous and we used library devtools, the promp() function and ggplot_pca in R to build the principal components and conduct the PCA analysis. On running the analysis we had 6 Principal Components as we had 6 variables to begin with. The results in the output below show that PC1 alone explains 78.3% of the variance.

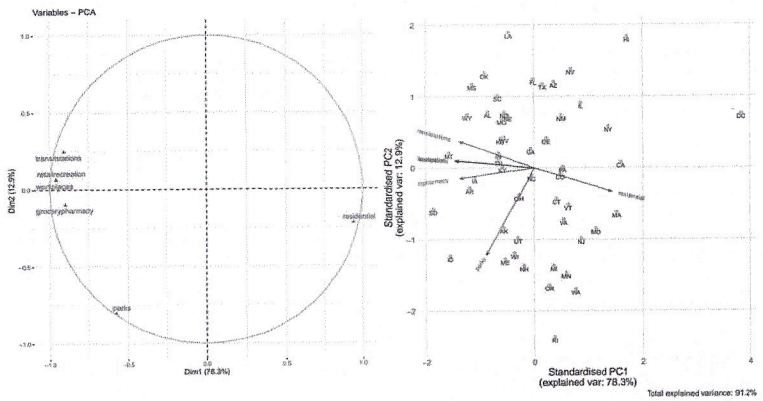
```
##   |
##   |
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.167 0.8789 0.48715 0.38588 0.29057 0.24273
## Proportion of Variance 0.783 0.1287 0.03955 0.02482 0.01407 0.00982
## Cumulative Proportion  0.783 0.9117 0.95129 0.97611 0.99018 1.00000

## Standard deviations (1, .., p=6):
## [1] 2.1674893 0.8788770 0.4871499 0.3858770 0.2905670 0.2427342
##
## Rotation (n x k) = (6 x 6):
##                             PC1         PC2         PC3         PC4         PC5
## retailrecreation -0.4426478  0.07671655 -0.1358158  0.30600288  0.79677826
## grocerypharmacy  -0.4161234 -0.11160678 -0.8399229 -0.09669224 -0.29368376
## parks            -0.2671361 -0.91444816  0.2607091 -0.11465835  0.05411761
## transitstations  -0.4193163  0.27926718  0.2269460 -0.81995956  0.05417050
## workplaces       -0.4415992  0.07033084  0.2676420  0.38535634 -0.51408198
## residential       0.4344037 -0.25001295 -0.2915093 -0.25106400  0.09354577
##                             PC6
## retailrecreation -0.22635856
## grocerypharmacy   0.11539163
## parks             0.09153368
## transitstations  -0.13932368
## workplaces       -0.56181111
## residential      -0.76943183
```


Given that we have 6 dimensions in our dataset the amount of variance we would expect to have in a given dimension by pure chance is 1/(6-1) or about 20%. Using this as our threshold we observed that PC2 explained only 12.9% additional variance (below our threshold) in the dataset therefore we decided to just include PC1 in our final model.

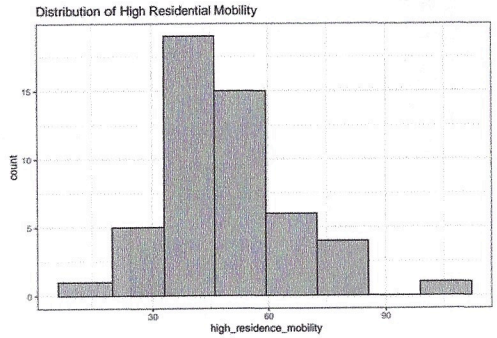


Next we studied how each of the variables contributes to our first principal component which we plan to include in our model. Looking at the results in the output and the ellipse figure we noticed that residential, workplace and retail and recreation mobility had the highest absolute correlation with PC1. Additionally residential mobility was positively correlated with PC1 and the remaining categories were negatively correlated. GIven the magnitude and direction of each variable with the first component one could infer that PC1 largely represents people staying at home or high residential mobility and less outside home mobility. Based on this concept we named our PC1 as high residential mobility.

The ggplot_pca below shows the residential vector points in a completely opposite direction compared to other vectors. The distance along x-axis represents the correlation with PC1 and y-axis represents correlation with PC2. States exhibiting higher residential mobility like California (CA), New York (NY), Massachusetts (MA) are in the direction of residential vector (towards the right) and states with less residential and high mobility in outside categories like Montana (MT), Iowa (IA), South Dakota (SA) are in the direction of vectors pointing to the outside categories on the left. This distribution is somewhat expected as one would expect political leanings of a state would affect state policies and how people think and follow mobility restrictions.



We further looked at the distribution of our newly created high residential variable and observed that it was not perfectly normal or bell shaped. However there was no major skewness and or long tails.



3.4 COVID-19 Cases

We included information on new COVID-19 cases at the state level as a proxy for new COVID-19 infections. The data comes from a live dataset compiled from state and local governments and health departments by the New York Times, the Coronavirus (Covid-19) Data in the United States dataset at the county level.

this structure of headings doesn't make great sense.

This dataset records all new official cases and deaths in a day by US county for each day since January 21st, 2020, to today (new data is added at the end of each day).

3.4.1 Data

As mentioned above, we use data on new official COVID-19 cases for each county in the US form the New York Times. The dataset records the number of new reported cases per day for each county reporting a case (with missing values for a county if no new cases are recorded). It was compiled by a group of several journalists from a wide set of government and other sources and includes both probable and confirmed cases and deaths.

3.4.2 Variables

The variable we use in our estimation is total number of new COVID-19 cases per 100,000 persons in a given state, for a particular month (January 2021 or July 2020). Since there is no accurate way to measure new COVID-19 infections, which may be asymptomatic or go unreported even if an infected individual presents symptoms, we use total number of reported cases as our proxy measure for total new COVID-19 infections. This variable will likely underestimate true infections levels but, given the difficulties in measuring this variable, it is the best alternative we have.

The raw data on COVID-19 cases includes only the number of reported cases in a given county, if there were any reported cases. To get a monthly total of new cases, we aggregated across days and counties in each particular state. We then dividied this variable by the number of 100,000 persons in each state (using population totals from the U.S. Census American Community Survey) Our resulting variable measures the total number of new COVID-19 cases in a state in a given month.

We call this variable "new infections" but, do note that what it actually measures is new reported COVID-19 cases.

3.5 Unemployment

Our main goal is to estimate the effects of COVID policies on unemployment levels. To measure unemployment levels we used unemployment insurance claims as a proxy for unemployment levels.

3.5.1 Data

We use data on initial unemployment claims from the US Department of Labor's Employment and Training Administration. In particular we use weekly initial unemployment claims which record the total number of new unemployment insurance claims filed in a given week. This data is published in ETA's weekly releases of unemployment insurance claims and is available at the national and state levels.

3.5.2 Variables

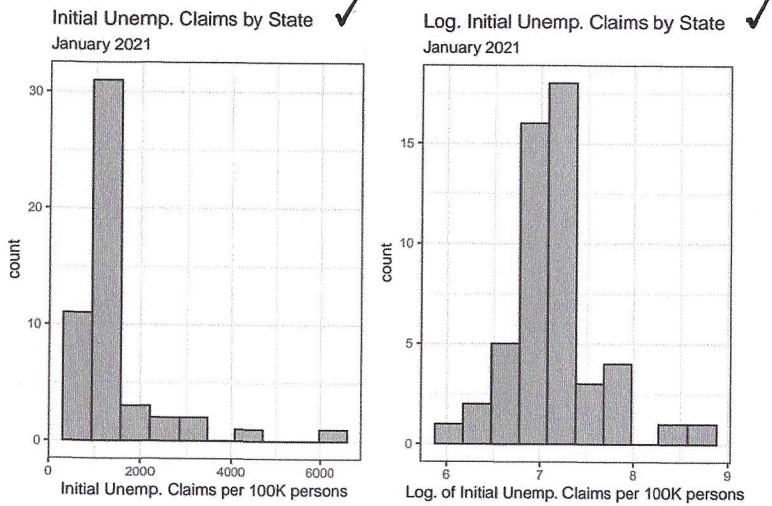
As mentioned before, we use weekly initial unemployment insurance claims as our measure of unemployment levels. We chose this variable as it is a commonly used measure of actual unemployment levels at the weekly level. It provides information on the number of new unemployment insurance claims filed each week at the state and national levels. As such, it is as close to real-time as possible compared to other measures of unemployment levels.

Since the original data is at the weekly level and our unit of analysis is at the month level, for each state in the US we aggregated each months new unemployment insurance claims into a monthly total. Since totals will depend on the size of the working population in a give state, and to keep this measure consistent with

this is a big aggreg. flow

our measures for other variables (like new COVID-19 infections, for example), we standardize the number of new unemployment insurance claims to total per month per 100,000 persons, by dividing by the number of 100,000 persons living in each state.

Note that initial unemployment insurance claims do not equal actual unemployment numbers for several reasons. For example, not all unemployed individuals are eligible to claim unemployment benefits. Thus, the actual number of unemployed individuals in a given state may be higher than that reflected in the initial unemployment claims data. In other cases, people may wait before filing a claim even after being unemployed for different reasons. Despite these shortcomings, this variable is a good and immediate proxy of the number of newly unemployed individuals in a state.



As the figure above shows, the initial unemployment claims by state (left panel) has a very skewed distribution (right skew). Taking natural logarithms, as shown in the right panel helps in making the distribution of this variable closer to normal. For this reason, we will use this transformation in our models.

4 Statistical Model

To estimate the effect of COVID policies on unemployment insurance claims we model the natural logarithm of unemployment insurance claims per 100,000 persons in a state as a linear function of our measure of COVID policies (as described in section 3.1.2), state characteristics (as described in section 3.2.2), our measure of mobility (as described in section 3.3.2), and total COVID-19 cases per 100,000 persons in the state.

We estimate five different models, starting from a very simple model including only an intercept term with no additional variables, adding new sets of variables (measuring the different relevant variables identified in our theoretical causal model as relevant).

The five models we ran are:

1) Model 1: Natural logarithm of initial unemployment claims on a constant

$$log(unemployment)_s = \beta_0 + \epsilon$$

2) Model 2: Model 1, adding COVID policy variables

$$log(unemployment)_s = \beta_0 + \beta_1 BusinessClosures_s + \beta_2 NightlifeClosures_s + \epsilon$$

3) Model 3: Model 2, adding measure of mobility

$$log(unemployment)_s = \beta_0 + \beta_1 BusinessClosures_s + \beta_2 NightlifeClosures_s + \beta_3 High\_Residential\_Mobility_s + \epsilon$$

4) Model 4: Model 3, adding state-level characteristics

$$log(unemployment)_s = \beta_0 + \beta_1 BusinessClosures_s + \beta_2 NightlifeClosures_s + \beta_3 High\_Residential\_Mobility_s + \beta_4 White\_Nonhispanic\_Citizens_s + \beta_5 Less\_Educated\_in\_HighDensity_s + \beta_6 Nonhispanic\_Females\_in\_HighDensity_s + \epsilon$$

5) Model 5: Model 4, adding new COVID-19 cases variable

$$log(unemployment)_s = \beta_0 + \beta_1 BusinessClosures_s + \beta_2 NightlifeClosures_s + \beta_3 High\_Residential\_Mobility_s + \beta_4 White\_Nonhispanic\_Citizens_s + \beta_5 Less\_Educated\_in\_HighDensity_s + \beta_6 Nonhispanic\_Females\_in\_HighDensity_s + \beta_7 New\_Infections_s + \epsilon$$

where the subscript s denotes a state of the US.

Our main parameters of interest are the coefficients on COVID policies, β_1 and β_2. These coefficients measure the marginal change of an additional unit of each of these variables on the log of new unemployment insurance claims.

We acknowledge that, given our variable transformation, these estimates are not easily interpretable as they are combinations of different underlying policy variables in non-linear ways. However, the can potentially be informative of the overall sign and statistical significance of the relationship we look to estimate.

5 Results

In this section, we present the main regression results for each of the specifications described in section 4 (models 1 to 5).

5.1 Regression Results

Below, we present our main regression results for each of the five models specified in the previous section. The dependent variable is the natural logarithm of new unemployment insurance claims. These regressions exclude the District of Columbia as we identified DC to be an outlier that was heavily skewing our estimates.

We estimate our coefficients using OLS regression and R's `lm` package, specifying robust standard errors (as a conservative way in case of potential heteroskedasticity issues).

Either: ① match these to the variable names in the models on page 19. (preferred) or write the whole concept in English.

Table 1: Main models (log-linear). January 2021.

	Dependent variable:				
	ln_uic				
	(1)	(2)	(3)	(4)	(5)
bus_closed		0.052	−0.066	−0.005	0.007
		(0.096)	(0.087)	(0.109)	(0.117)
nightlife_closed		−0.087	−0.025	0.064	0.040
		(0.183)	(0.171)	(0.179)	(0.176)
high_residence_mobility			0.012***	0.010	0.006
			(0.005)	(0.007)	(0.010)
white_nonhisp_cit				0.003	−0.001
				(0.006)	(0.008)
lesseduc_highdens				−0.003	−0.006
				(0.009)	(0.011)
nonhisp_fem_highdens				0.005	0.006
				(0.010)	(0.010)
new_infections					−0.0001
					(0.0002)
Constant	7.133***	7.133***	6.528***	6.294***	6.388***
	(0.068)	(0.071)	(0.237)	(0.858)	(0.947)
Observations	50	50	50	50	50
R^2	0.000	0.012	0.120	0.150	0.167
Adjusted R^2	0.000	-0.030	0.062	0.031	0.028
Residual Std. Error	0.479 (df = 49)	0.486 (df = 47)	0.463 (df = 46)	0.471 (df = 43)	0.472 (df = 42)
Note:					*p<0.1; **p<0.05; ***p<0.01

Looking at our regression results (Table 1) one thing stands out: none of our estimates of interest - for *bus_closed* and *nightlife_closed* are statistically significant in any of our specifications (2 to 5). And, the coefficients are somewhat unstable across specification, switching from negative to positive across specifications. In model 5 (the specification we consider to be the model that would best capture potential effects) the coefficients for our COVID policy variables are both positive, as we would expect given our theoretical model; business closures and late night and entertainment venues closures are associated with higher unemployment claims. However, even in this case, these coefficients are not statistically significant at any conventional level.

Given the very small sample size we have and the trade-offs between the number of observations we have and the limited potential to include more variables in our regressions (as more parameters to estimate means less degrees of freedom), these results are, perhaps, not too surprising.

As an additional check on our estimates, we run the same set of regressions on data for July 2020 (the time when the second major wave of COVID-19 infections hit the US). We present these results in the table below.

Table 2: Main models (log-linear). July 2020.

	Dependent variable:				
	ln_uic				
	(1)	(2)	(3)	(4)	(5)
bus_closed		0.051	−0.069	−0.085	−0.075
		(0.115)	(0.100)	(0.107)	(0.111)
nightlife_closed		0.048	−0.005	−0.040	−0.036
		(0.147)	(0.122)	(0.123)	(0.134)
high_residence_mobility			0.005***	0.0002	−0.0002
			(0.002)	(0.004)	(0.004)
white_nonhisp_cit				−0.014	−0.014
				(0.011)	(0.011)
lesseduc_highdens				−0.028**	−0.026*
				(0.013)	(0.014)
nonhisp_fem_highdens				0.012	0.011
				(0.007)	(0.007)
new_infections					0.0001
					(0.0002)
Constant	7.086***	7.086***	7.137***	4.862***	4.880***
	(0.062)	(0.064)	(0.058)	(1.015)	(1.013)
Observations	50	50	50	50	50
R^2	0.000	0.008	0.194	0.342	0.349
Adjusted R^2	0.000	-0.034	0.142	0.250	0.240
Residual Std. Error	0.433 (df = 49)	0.440 (df = 47)	0.401 (df = 46)	0.375 (df = 43)	0.377 (df = 42)
Note:					*p<0.1; **p<0.05; ***p<0.01

The results for the July 2020 cross-section (Table 2) also show no statistically significant relationship between unemployment insurance claims and COVID policies. In this case, however, for models 3 to 5, the coefficients on these variables are consistently negative. This differs from our results for January 2021, where these coefficients were less stable, switching from positive to negative.

However, overall, the main takeaway from these regressions is that our models fail to find any statistically significant relationship between COVID policies and unemployment insurance claims.

Note that this does **not** mean that a causal effect does not exist. Our results simply tell us that, even if there were a causal relationship, our model is not capable of detecting it. Instead, our models seem to be capturing noisy or weak signals and lack the power to detect anything but really strong signals (if at all).

As we discuss in the following section, and given our small sample size, using OLS under the classical linear model assumptions may not be fully warranted. Those findings support our view that given our very limited dataset, we are unable to precisely estimate our coefficients of interest.

5.2 Testing Classical Linear Model Assumptions

Assumption 1: Independent and Identically Distributed (I.I.D.)

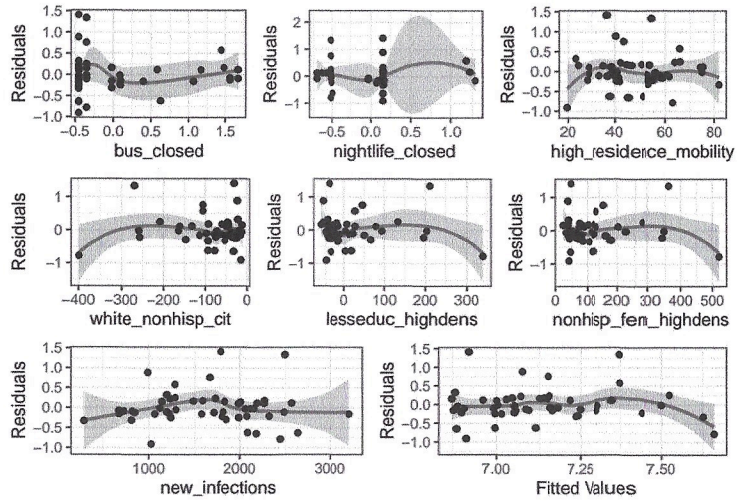
Unfortunately, given the nature of the data, our model fails the first requirement of any classical model, that the data be I.I.D.. Simply by nature of trying to design a model based off of the 50 United States, we run into several issues with independence. Firstly, the states are all related by location. States near to each other are likely to be affected by the same events, such as snowstorms that limit mobility. States in the same region are also likely to have similar demographics in terms of racial and age diversity. There is also a free flow of individuals between states, which further complicates matters.

What this translates to in our data is that, rather than having even 50 independent observations, we instead have clusters of semi-dependent data. This results in even less power than our low sample size is already giving us. Some strategies to control this in future modeling exercises could include creating a regional indicator such as "Southwest, Southeast, Midwest..." that would group states that share similar characteristics. This would help to draw out some of the dependence between the data points into a measurable form.

Another option would be to gather data at a more granular level, such as county level. This would still have the same issues as the larger data set, but the additional data points would help to offset the power loss caused by the dependence between observations. Random sampling within this county level data would be the best case scenario for minimizing dependence, but may not be worth the trade-off in reduced sample size. Unfortunately, even with all of these precautions in place, it would not solve the dependence problems that come from drawing our data out of a single, interconnected country.

Assumption 2: Linear Conditional Expectations

The second assumption required for a classical linear model is that there is a linear conditional expectation in the explanatory variables. The best way to check for this assumption is through visually inspecting plots of the model residuals on the y-axis against each explainer as the x-axis. A plot of the fitted points against the residuals is also a useful tool, especially when there are too many variables to plot, as it captures the model as a whole. Looking at only fitted values against residuals can hide issues in individual components though, so we prefer to look at each explainer individually if able. With seven explanatory variables, we are just on the edge of being able to do so concisely. Had we included more variables this would be impractical. The diagnostic plots are below.

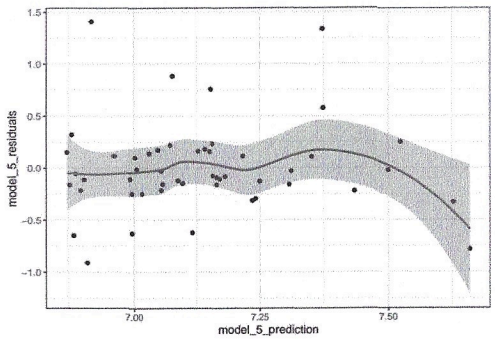


It appears that we have problems of non-linear conditional expectation for several of our variables. Our two MCA generated variables from the Policy dataset appear to be more or less linear, although `bus_closed` appears to spike upwards at low values, and nightlife_closed has large gaps that make assurances about linearity uncertain. The Mobility variable and the 3 Demographic variables all suffer from non-linearity at the extremes. In the case of the demographic variables, we have one point that is separated from the others. Upon examination of the data set, this point belongs to New Jersey. While linearity might be improved if we dropped New Jersey, it would severely hinder the applicability of our model in explaining causality for all states. Finally, our measure of new infections appears to have linear conditional expectation.

Our model as a whole fails the visual test for linearity at higher values, with our model consistently over predicting these points. Given the weight of evidence, we must conclude that this assumption is not met. What that means for our model is that while it remains the best linear predictor for the data it was given, the coefficients are not reflective of the true relationship between the variables and the outcome. Efforts could be made to linearize some of these relationships, but this would further cloud the meaning of our already complicated PCA/MCA variables. A better strategy would be to re-specify some of the variables, such as by using a different measure for policy instead of a simple indicator.

Assumption 3: Homoskedasticity

This assumption refers to the variance of the residual, or error term, in a regression model being constant. That is, the error term does not vary much as the value of the predictor variable changes. A way to validate this assumption is to perform an ocular test of the residuals vs. predictions plot. This way, we can assess if the dispersion of the observations in the plot are constant or not. The diagnostic plot can be found below:



It can be observed in the plot above that majority of time, the variance is constant for a sizable portion of the observations. However, the right hand side of the plot shows an increase of the dispersion in the residuals and hence, a non-constant variance. As a consequence of this, the estimated standard errors could be not correct and because of this, we should be careful with the conclusions observed with the confidence intervals and hypothesis tests as they may not be completely reliable.

Assumption 4: No perfect collinearity

This assumption refers to the following: If the correlation between two or more regressors is perfect, that is, one regressor can be written as a linear combination of the other(s), we have perfect multicollinearity. While strong multicollinearity in general is unpleasant as it causes the variance of the OLS estimator to be large, the presence of perfect multicollinearity makes it impossible to solve for the OLS estimator, i.e., the model cannot be estimated in the first place.

When there is a presence of perfect collinearity, R studio drops the variables having this issue when a model when run. In this case, none of the variables is dropped when running the model. Hence, there is no perfect multicollinearity.

In the case of strong collinearity, we calculated the paired-correlations of the variables contained in our model. We have variables coming from 4 different clusters of variables: Government Policies, Demographics, Mobility and Number of Infections. Since we calculated Principal Components and by definition, they are orthogonal to each other, we will only calculated the correlation coefficients between the variables from different clusters:

- Business closed and high resident mobility:

```
## [1] 0.4608983
```

- Business closed and white non-hispanic citizens:

```
## [1] -0.02318216
```

- Business closed and non-hispanic females in high density areas:

```
## [1] -0.04588159
```


- Business closed and less educated population in high density areas:

```
## [1] -0.02391711
```

- Business closed and number of new COVID-19 cases:

```
## [1] -0.04573385
```

- Nightlife closed and high resident mobility:

```
## [1] -0.1547569
```

- Nightlife closed and white non-hispanic citizens:

```
## [1] 0.3727877
```

- Nightlife closed and non-hispanic females in high density areas:

```
## [1] -0.4134839
```

- Nightlife closed and less educated population in high density areas:

```
## [1] -0.415214
```

- Nightlife closed and number of new COVID-19 cases:

```
## [1] -0.1659136
```

- High_residence_mobility and white non-hispanic citizens:

```
## [1] -0.5535642
```

- High_residence_mobility and non-hispanic females in high density areas:

```
## [1] 0.4734751
```

- High_residence_mobility and less educated population in high density areas:

```
## [1] 0.4737685
```

- High_residence_mobility and number of new COVID-19 cases:

```
## [1] 0.01544052
```

- White non-hispanic citizens and number of new COVID-19 cases:

```
## [1] -0.3857487
```

- Non-hispanic females in high density areas and number of new COVID-19 cases:

```
## [1] 0.3643304
```

- Less educated population in high density areas and number of new COVID-19 cases:

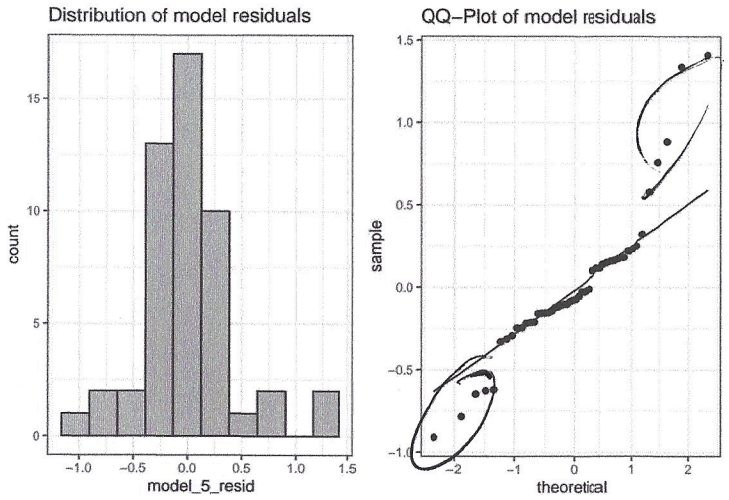
```
## [1] 0.3429254
```

All correlations are less than |0.6|. The highest correlations in absolute value occur between high residence mobility and demographic variables with the correlation of |0.55| between this variable and proportion of white non-hispanic citizens being the highest. We consider that given these values, we cannot say there is almost perfect collinearity.

Assumption 5: Normally distributed Errors

The last assumption is normally distributed errors. To assess this assumption we are building histograms and qq-plots of our residuals from our final model. We also conducted a Shapiro-Wilk test of normality. The Shapiro-Wilk test was statistically significant with a p value of 0.0002 therefore we reject the null hypothesis for this test and our residuals are not from a normal distribution. The histogram of residuals below show a narrow distribution in the center with slightly heavy right and left tails. Also the qqplot of model residuals shows deviations from normality both on the left and right tail ends. Looking at these results unfortunately, we conclude that our model does not meet the assumption of normally distributed errors. Having not met this assumption threatens the validity of the t-tests testing whether our model coefficients are significantly different from zero and calculations of confidence intervals.

```
##
##  Shapiro-Wilk normality test
##
## data:  model_5_resid
## W = 0.88876, p-value = 0.0002087
```

6 Conclusions and Discussion

6.1 Conclusions

In this study, we set out to test whether COVID policies led to higher unemployment. We did this by constructing a set of COVID 19 policy variables using MCA, as well as a set of covariates using PCA, and used initial unemployment insurance claims as a proxy measure of unemployment. We tested our model against data from two months that coincided with surges in COVID 19 cases in the US: January of 2021, and July of 2020. We fail to detect any significant causal effect between these variables.

As a result, our study is inconclusive about the causal relationship between COVID policies and unemployment insurance claims levels (as measured and implemented in our study). Note that this does not mean that the causal relationship does not exist, and it certainly does not suggest that the relationship does exist. At most, it means that our model and data cannot detect this effect, even if it truly exists.

6.2 Discussion

Given the limitations in our data – mainly the inherently small number of observations we have from a cross section of 50 US states, and modelling constraints, it is not very surprising that our estimation models failed to detect a statistically significant effect for our variables of interest (COVID policies).

While we think that our theoretical causal model does a good job of capturing the existing causal paths, sample size limitations, and the challenges that this implies for accurate estimation of a causal relationship using a simple OLS regression under classical linear model (CLM) assumptions make it difficult for us to be

able to find any effect. Furthermore, we cannot rule out important violations to the CLM that would bias our results. For example, despite our best efforts to compress as much information in as few independent variables as we could, it was difficult to strike the right balance between the available number of observations vis-à-vis the number of covariates. For example, variables that measure how well business set-up or prepare for pandemic operations may mediate our relationship of interest (resulting in omitted variable bias away from zero in our estimates).

The challenges that a small sample presented are difficult to address. With cross-sections of 50 observations, even under an ideal scenario where half of US states were randomly selected to implement COVID policies and the other half to not implement them (and with full compliance), it would be hard to measure all but *very* large effects.

Doing some basic calculations, with a setup as described in the paragraph above (and no covariates), we would only have enough power (at 80% power and 95% significance) to detect effect sizes of 0.8 standard deviations (a *very* large effect).

The discussion above drives home the fact that we would need to increase our sample size to be able to detect smaller (and perhaps more likely) causal effects. Possible options include using differenced panels with state and month (or even week) effects, and then use these fixed state and time effects to help us estimate causal effects of COVID policies, if we have enough cross-state and cross-time to get these estimates. Further refinement could also be performed on the policy variables themselves, perhaps measuring them in terms of length and adding a variable measuring how many times the policy had been implemented. We leave these explorations as potential extensions for future research projects.

7 References

1. Ankan, A., Wortel, I., & Textor, J. (2021, February 16). Testing graphical causal models using the r package "dagitty". Retrieved April 11, 2021, from https://currentprotocols.onlinelibrary.wiley.com/doi/full/10.1002/cpz1.45
2. Google. COVID-19 Community Mobility Reports. U.S. data. Retrieved March 17, 2021, from https://www.google.com/covid19/mobility/
3. Hayden, Luke (2018, August 9). Principal components analysis in R. Retrieved April 3, 2021, from https://tinyurl.com/2669t34w
4. Kassambara. (2017, September 24). MCA - multiple Correspondence analysis in R: Essentials. Retrieved April 11, 2021, from http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/
5. The New York Times. (2021). Coronavirus (Covid-19) Data in the United States. Retrieved April 10, 2021, from https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv
6. US Department of Labor (2021). Unemployment Insurance Weekly Claims Report. Employment and Training Administration. Retreived April 4, 2021, from https://oui.doleta.gov/unemploy/csv/ar539.csv

1 Lab 2 Final Report 75 / 75

✓ - 0 pts Correct

- Overall - there is so much good work in here.
- There are some small odds-and-ends that I'd like to see cleaned up from a presentation standpoint.
- The aggregation - to state-month - is probably the best option available to you, but still leaves me wonder whether any technique would have been able to find a relationship.