Analyzing the Impact of Vaccinations, Government Policy and Other Factors on the COVID-19 Pandemic

---

A Thesis

Presented to

The Division of Mathematical and Natural Sciences

Reed College

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Arts

Evan Pugh

December 2021

Approved for the Division
(Mathematics - Statistics)

Jonathan Wells

Acknowledgements

I would like to thank all the great teachers and advisors I've had over my many years of school. In particular I'd like to thank Kelly McConville, Jonathan Kadish and Jonathan Wells for helping me while I was working on my thesis. I would also like to thank Dr. Michael Morse, for being a tremendous help in teaching me how to deal with issues in school I've struggled with my since I was 11. I would like to thank my parents, Bill Pugh and Lisa Orange, for always wanting what was best for me and always encouraging me. And lastly, I would like to thank my brother Connor for nothing in particular, just everything in general.

Table of Contents

Abstract

The ongoing COVID-19 pandemic has been one of the most world shaking events in recent history. During this time period, there was a wide variety of responses to and consequences of the pandemic across the United States. As life slowly returns to a new normal, it is important that we don't simply move on from this, but that we learn from it. There is a very real chance that many of us will experience another global pandemic similar to this one, and making the same mistakes then that we did this time will cost lives. In this paper, I break down a wide variety of data in order to analyze the collective impact of things such as vaccinations, government policy, and the Delta variant on how this pandemic spread, as well as how those variables interact with each other. Overall, my data showed that the best approach for combating the pandemic was using government restrictions as a short term fix to prevent bad situations from getting really out of hand and to buy time for long term solutions like vaccinations to be developed.

Dedication

This paper is dedicated to everyone who ever believed I deserved another chance whenever I stumbled. Thank you for knowing I could do better.

Chapter 1

Context

The ongoing COVID-19 pandemic has been one of the biggest world altering events since the end of World War II, if not the biggest. Starting with a few cases in Wuhan, China at the end of 2019, the COVID-19 outbreak was declared a pandemic on March 11th, 2020 as worldwide spread began and by the end of the month, the number of global cases was nearly 800,000 and the United States alone had more than 140,000 confirmed cases. As of December 3rd, WHO reports that globally there have been are currently roughly 262.87 million confirmed cases and 5.2 million confirmed deaths worldwide In the US we have had over 48 million cases of COVID and are sitting at around 776,500 deaths. (World Health Organization, 2021)

The number of cases and deaths haven't been the only effects of this pandemic. In the United States alone, the number of home owners behind on their mortgage by three months or more increased 250% to over 2 million households, a level not seen since the peak of the Great Recession (Bureau, 2021), unemployment reached 14.8% (the highest level ever recorded since this was someone that began being tracked in 1948), and back in October 2020 the Congressional Budget Office projected a total of $7.6 trillion in lost output as a result of the pandemic over the next decade. (Cutler & Summers, 2020)

However, things could have been much worse. If it weren't for the unprecedented speed at which a vaccine was developed and the different measures that have been taken to curb the spread of the virus, things could have been much, much worse. And while the prospect of things returning to pre-pandemic normalcy is still a far off prospect, it does feel like the pandemic is winding down, at least in the United States, and the tension surrounding the situation is lessening.

Unfortunately, it is fairly likely that many of use experience another pandemic of a scale similar to this or the 1918 Influenza pandemic. Due to the nature of zoonotic

diseases like Influenza and COVID-19, odds of pandemics of these scales are only going to increase as humanity continues cutting down forests and building infrastructure. So we are not allowed to simply wash our hands off this mess and be done with it once things settle in to a new normal. It is our responsibility to learn what we did right and what we did wrong during this pandemic.

Since COVID-19 was officially declared a pandemic, there have been many studies analyzing the effectiveness of different measures taken to combat the virus, including vaccinations. However, most of those studies focus primarily on just one measure, or maybe a handful of measures. My goal with this thesis is to focus on analyzing the impact of our combined response to the pandemic rather than just looking at the impact of one specific policy. I think this is important because when we come up with policies for dealing with the pandemic, we aren't just implementing one policy or a small handful of measures. We are trying a ton of different things and I think it is important to examine the impact of those things as a whole rather than just as individual measures.

Chapter 2

Data

2.1 Overview

To investigate what factors have impacted the spread of the COVID-19 pandemic, I have created one large data frame using data I pulled from a wide variety of sources. The main combined dataset I'm using is a consists of daily county level from 2819 of the counties 3006 counties in the United States. The data for a county begins once it is after March 8th, 2020 and there have been COVID-19 cases in the county for at least a week. Daily observations for the county continue until the Delta variant reaches 0.40 percent of the total new cases of COVID-19 in the region at which point the data for that county stops, for reasons that I will go into later. In total, the dataset I used to build my model consists of 1027654 observations of 81 different variables. Here is a breakdown of where I got that data from and what it means.

2.2 General

Before I get into discussing specific data sources, I would like to do some general discussion about where I'm getting my data from. A lot of my data comes from the similar sources and/or is focused on the same topics. So to avoid unnecessarily repeating myself, I'd like to start off by discussing some topics that are applicable to multiple of the variables that I've collected data on and then later referencing.

So to start off with, most of my data comes directly from the government itself, generally either the CDC or the Census Bureau. The data I get from these sources is mostly variables like population, demographics, reporting on COVID, etc. The CDC and Census Bureau are well regarded as accurate data sources so for the most part I

have no reason to believe has any significant errors. However there are two different kinds of issues with the data on COVID-19 in particular that I've noticed. The first is simply the that data is more likely to be accidentally entered wrong. Compared to the types of data I previously discussed, the data I'm using that centers on COVID-19 statistics is the kind of data that often gets entered every day and even county level data is complied from multiple different sources, so it is a lot easier for some of my data to be off because someone entered a number wrong. Additionally, if someone incorrectly enters data for something that's fairly static (at least for the time frame I'm using) and can be measured again, then there's a good chance that the typo will get fixed. That's not the case with daily reporting on COVID-19 cases. If 127 cases of COVID-19 are reported that day, that's how many cases of COVID-19 are reported that day regardless of if they later find out that it should have been 137 cases of COVID. The second potential issue that data just isn't reported. So you have people taking at home COVID-19 tests that aren't counted in my data on testing, you have people who just catch COVID-19 and never feel sick enough to go to the hospital or get tested, etc. Again, this is just really something that there isn't a way around and overall the picture I get from this data is still mostly accurate. So while it is a source of concern that is worth acknowledging, it's mostly unavoidable and shouldn't cause too much trouble.

Another data source that I'm going to be frequently referring to is Johns Hopkins. Johns Hopkins is widely regarded as a quality institution and is one of the all time most cited intuitions in research papers. Additionally, a lot of the data I'm using from them is data that has been scrapped and aggregated from various government sources, like the CDC. I don't think there is much chance that either of these parties made significant errors, so I'm content with trusting this data.

Now let's get into specifics. With each data source I have I plan to discuss two things. First, I'll focus on how reliable is this data and whoever gathered it. Second, I'll discuss in what ways it is relevant to my topic and in what ways it may not be a perfect measure of what I'm looking for.

2.3 Data Sources

2.3.1 COVID-19 cases and deaths data

The primary dependent variables I'm going to be making models to calculate for are the number of COVID-19 cases by county and the number of COVID-19 deaths by

county. My source for this data is Johns Hopkins university, which has aggregated a ton of county level data on COVID-19 that is reported by the government into one file. Johns Hopkins and the governments collection of COVID-19 data are both sources I trust a lot and don't think I could find better alternatives for.

There are instances where data is incorrect, either because something was/wasn't attributed to COVID-19 by mistake or by human error in entering data. I have done my best to correct this in obvious instances where it resulted in a negative number of new cases or deaths (I'll discuss that more in my Methods Section), however that doesn't perfectly replicate what the true data is and it doesn't catch instances where there were errors but just not any large enough to cause a negative amount of new case or deaths. Regardless, I feel that it is close enough to the true values here that it is not a major concern for my model.

2.3.2 Vaccination data

Data on vaccinations is probably one of the more important pieces of data to have in this model. The variables I took from this data set are the county and state level data for percentage and number of all people who are vaccinated, percentage and number of people over 18 who are vaccinated, and percentage and number of people over 65 who are vaccinated.

This data comes from the CDC so overall I'm very confident in the accuracy of the data that is being reported, especially since there aren't really unreportable cases like there are with mild COVID-19 cases that are never detected or self administered COVID-19 tests that are used but the results of which are never reported. However, there is one pretty big issues. County level vaccination rate data is not reported for Hawaii, Texas nor California counties with less than 20,000 people. I could simply use the statewide data for each county, however vaccination data varies far to much on a county by county basis within a state for that to really work. Instead I have gone with the simple option of eliminating those counties from my data. This is not ideal, especially in the case of Texas. Texas makes up a lot of my data and was particularly interesting considering the massive spike they had late this summer and particular policy decisions that have been made there. However, it was ultimately necessary and the only alternative I could think of was managing to find a way of generating estimates of the vaccination data for my missing counties, which would be far too time consuming and might not even wind up giving me anything usable.

2.3.3 Government policy data

Now let's get into the government policy variables. This data set is from the University of Oxford and consists of categorical data where each observation is a day in a given nation or state and the variables represent a broad array of different types of policy steps a state or local government can take to combat the pandemic. Each variable is ranked on a scale from 0 to 4, with 0 being that there is no policy in place and larger numbers representing more significant policies. Some policies might not go all the way to 4 and instead cap at either 3 or 2. As an example, let's look at the school closure variable. A 0 in this column means that there has been no change implemented. A 1 means that schools have modified the way that they operate to deal with the pandemic (i.e. having a mask mandate, allowing for students to attend online if they wish, etc) but classes are still happening in the school buildings. A 2 means that school is canceled for some but not for others (i.e. just badly effected schools are closed, just high schools are closed). A 3 means that closing is required on all levels for schools in the state, although online classes may still be happening. More details on which variables mean what and how to interpret the values for them are available on the GitHub for this project. Tracker (2021)

Now this data set isn't perfect. The intricacies of broad policy decisions cannot be sorted into just a few categories and there is some missing data, especially since for my model I'm going to be converting it into numeric data. However, it is still absolutely incredible that this amount of data has been gathered considering that this values have to be changed manually by someone looking at headlines and seeing what kinds of policy decisions have been made. I'm lucky to have a dataset this good, I could never have created something anywhere near this massive on my own.

Even so, there are a fair amount of variables from this data set that I chose not to use. This was mainly due to some combination of not having enough variation in them over the course of the pandemic or the categories being too broad to capture the kind of detail I'd like. These were mostly just unfortunate issues I ran into because of the specifics about what I'm doing rather than the data set in general, such as a lack of other countries in the mix, my particular time frame meaning some data is just kinda pointless, or just stuff that has a murky relationship with what I'm trying to measure.

2.3.4 Test positive rate

The source I am using to track statewide data on COVID-19 tests is a compilation of data from state governments. This data was initially complied by the COVID-19 Tracking Project, however it was taken over by the Johns Hopkins Coronavirus Resource Center as of March 7th, 2021. This two groups have very solid track records and I am confident that they can compile data accurately, so the only potential source of error would be from the data sources themselves. From this dataset I will mainly just be uses the variables representing the total number of tests and the total number of positives. I'm including this data in addition to just the raw number of COVID-19 cases this for two reasons. First, the number of COVID-19 tests performed provides good information on how a state is ramping up capability for COVID-19 and how seriously people in that state are treating the COVID-19 pandemic. Second, the percentage of tests that are positive tells us a lot about the nature of the pandemic and our ability to deal with it. If that percentage is high, that means there are likely a lot more COVID-19 cases that are going unrecorded and more people should be getting tested.

It is worth noting that this data does not include all tests performed, since there are a large amount of at home COVID-19 tests performed and negative tests aren't reported and positive tests are merely encouraged to be reported. However, I feel that this is an unavoidable problem and despite this the testing data I have is still something I believe will prove extremely useful. Another concern I have is that my data on COVID-19 tests is statewide data rather than county wide data like the rest of my COVID-19 data. But I think this is one of the variables where not having county level data is too big of a deal and it's not a problem I have a good solution for.

2.3.5 COVID-19 Variants

My data on COVID-19 variants comes from the CDC itself and is important mainly because of the Delta variant. While other variants have been somewhat influential at one point, the Delta variant has been utterly dominant since it first appeared and brought the pandemic roaring back almost single handedly. As previously discussed, the CDC is extremely reputable and I doubt that there is any significant error in this data. However, that doesn't mean there aren't issues with it. This CDC tracker tracks the percent of COVID-19 cases that belong to different COVID-19 variants across 10 regions of the US. This is an issue because the larger a region the data covers the less likely it is for my data to be equally representative of all the area covered in it.

The difference in prevalence of the delta variant between two neighboring regions can get to be as big as 50% (although not in my data frame since I cut off my data once Delta hits 40%), meaning that two neighboring counties could be reported as having widely different values when they are actually probably fairly similar. This data is also only reported a few times a month, meaning that the data between reports has to be interpolated. However this is really the only source I have and doing anything truly complex here would have been a ton of effort for not much reward.

2.3.6 Hospitalization data

Another source of data that I've collected is data from hospitals and how they are weathering the pandemic. This is data that is reported to the government by hospitals, so I think it is pretty high quality. There are a ton of variables in here and I'm not going use all of them, but the mains ones that I have included in my data set are the ones related to how many hospitals are reporting that they have a staffing shortage and the ones related to what percentage of hospital beds are being taken up by people with COVID-19.

One problem with this data set is that it relies on hospital reporting this data themselves. So at the start of the pandemic when things weren't as bad you had fewer hospital reporting, there might be some hospitals that only occasionally reported or just occasionally reported, etc. However, it does tell me how many hospitals are choosing to report this data on this particular date and in general the hospitals that aren't reporting this data to the government are the ones not dealing with much COVID-19, so it still works as a good gauge. This is supported by the fact that the number of hospitals reporting in a state tends to only increase over time, meaning we generally aren't losing hospitals that were once reporting data. Another potential issue is that this data is by state and not by county. However I feel like there are factors that make it so it's not as important to have county level data for this variable as it is for other variables, such as the fact that it's probably fairly common for people to go to a hospital that isn't in their county.

2.3.7 Google Mobility data

The final major source of data I have included is Google's COVID-19 mobility data. The variables included in this data are based off of anonymized user location data from Google. It creates different categories for what type of place a location is, then tracks how much time people spend in those locations. Then it measures the

percent change in how much time people in a county/state are spending in locations corresponding to each category, with the baseline being based off of data from January 3rd through February 6th, 2020. I've used county level data when available, but if it wasn't available for a county I used the state level data instead. This might pose some problems since I think there is probably a fair amount of correlation between the counties that Google doesn't bother or have enough data to track, but it's still the closest representation I have.

2.3.8 Miscellaneous Data

This section is just for the sort of miscellaneous data from the Census Bureau that I have in my model that isn't really particularly related to COVID and more just for background on the counties to help the other variables work better. As previously said, Census Bureau data is very reliable so I don't have any concerns about significant errors in this data. In this category I have data on ages, population on a state and county level, what the average unemployment rate of a county was in 2020, median household income, and average household size. I initially had a lot more variables in this category, however I had to whittle down the scale of my model for practical reasons. These variables are the ones that survived.

2.4 Variable names and meanings

Now that we've covered all the data sources, let's go into the data variables that I'm actually using in my model. Below I've included a brief description of all the variables I've included in my model except for the government response variables, since those are all interlocked in complicated ways that I'll have to explain. Some of these may seem odd, and I'll explain them later in my methods section.

- Population: The population of the county (measured in thousands)
- State Pop: The population of the state the county resides in (measured in thousands)
- TotalCasesPerCapita: The total number of confirmed cases in that county per 100,000 people
- TotalDeathsPerCapita: The total number of confirmed deaths in that county per 100,000 people

- RollCasesCapita: The average number of cases per 100,000 people in the county over the past week
- RollDeathsCapita: The average number of deaths per 100,000 people in the county over the past week
- RollStateDeathsCapita: The average number of deaths per 100,000 people in the state over the past week
- CaseGrowthRate: The exponential growth rate of cases in the county between 7 days from now and 14 days from now. Multiplied by 100
- CaseGrowthRate2: The square of the CaseGrowthRate variable, divided by 100
- CaseGrowthRate3: The cube of the CaseGrowthRate variable, divided by 100,000
- VaccinatedAllPerCapita: Number of people in a county who are vaccinated per 100,000 people in the county
- Vaccinated18To64PerCapita: Number of people in a county between the age 18 and 64 who are vaccinated per 100,000 people in the county
- VaccinatedOver65PerCapita: Number of people in a county 65 year old or older who are vaccinated per 100,000 people in the county
- UnvaccinatedAllPerCapita: Number of people in a county who are unvaccinated per 100,000 people in the county
- Unvaccinated18To64PerCapita: Number of people in a county between the age 18 and 64 who are unvaccinated per 100,000 people in the county
- UnvaccinatedOver65PerCapita: Number of people in a county 65 year old or older who are unvaccinated per 100,000 people in the county
- WeeklyTestsPerCapita: Number of COVID-19 tests per 100,000 people performed over the past week in a state
- TestPositiveRate: The number of confirmed COVID-19 cases in the state during the past 7 days over the number of COVID-19 tests performed in the state that week
- LagWeeklyTestsPerCapita: The WeeklyTestsPerCapita variable from two weeks ago

- LagPositiveRate: The TestPositiveRate variable from two weeks ago
- Delta: The percentage of cases attributed to the Delta variant of COVID in that region of the US
- VaccinationRateRestOfState: The vaccination rate for the state outside of this county
- VaccinatedDiff: The Vaccination rate of a county minus the VaccinationRateRestOfState variable
- HospitalStaffingShortage: Percentage of hospitals that are reporting a staffing shortage
- HospitalBedsCovidPercent: Percentage of hospital beds that are taken up by COVID-19 patients
- retail_and_recreation_percent_change_from_baseline: Percent change from baseline of activity/movement in retail and recreation areas (restaurants, shopping centers, museums). County level when available, otherwise state level
- grocery_and_pharmacy_percent_change_from_baseline: Percent change from baseline of activity/movement in places like grocery stores, pharmacies, and farmers markets. County level when available, otherwise state level
- parks_percent_change_from_baseline: Percent change from baseline of activity/movement in places like local and national parks, public beaches and plazas. County level when available, otherwise state level
- transit_stations_percent_change_from_baseline: Percent change from baseline of activity/movement in transport hubs like subway, bus and train stations. County level when available, otherwise state level
- workplaces_percent_change_from_baseline: Percent change from baseline of activity/movement in workplaces. County level when available, otherwise state level
- residential_percent_change_from_baseline: Percent change from baseline of activity/movement in residential areas. County level when available, otherwise state level
- Over65Percent: Percentage of a county that is over 65 years old

- Age18To64Percent: Percentage of a county that is 18 to 64 years old
- Metro_2013: A dummy variable representing whether a county is part of a metropolitan area as of 2013
- Unemployment_rate_2020: The average unemployment rate of a county in 2020
- CaseGrowthPastTwoWeeks: Exponential growth rate of cases over the past two weeks, multiplied by 100 (set to 1000 if Infinite because it started at 0)
- CaseGrowthPastThreeWeeks: Exponential growth rate of cases over the past three weeks, multiplied by 100 (set to 1000 if Infinite because it started at 0)
- CaseGrowthPastFourWeeks: Exponential growth rate of cases over the past two weeks, multiplied by 100 (set to 1000 if Infinite because it started at 0)

2.5 Government Response Variables

Since a lot of my government response variables are built off of each other, I thought it'd be easier to explain them while they are isolated and on their own. So starting off, our base variables are:

- C1_School.closing: Records school and university closings
- C2_Workplace.closing: Records closings of workplaces
- C3_Cancel.public.events: Records cancellations of public events
- C4_Restrictions.on.gatherings: Records limits on public gatherings
- C5_Close.public.transport: Records closing of public transport
- C6_Stay.at.home.requirements: Records orders to "shelter-in-place" and otherwise confine to the home
- C7_Restrictions.on.internal.movement: Records restrictions on internal movement between cities/regions
- E1_Income.support: Records if the government is providing direct cash payments to people who lose their jobs or cannot work
- E2_Debt.contract.relief: Records if the government is freezing financial obligations for households

- H2_Testing.policy: Records government policy on who has access to testing
- H6_Facial.Coverings: Records policies on the use of facial coverings outside the home
- H8_Protection.of.elderly.people: Records policies for protecting elderly people (as defined locally) in Long Term Care Facilities and/or the community and home setting

  In addition to these variables, I also have lags of all of them as well. The lags that I have are simply the the same variable but with '.lag' attached and the end and the value that variable has is what the value of that variable was for the county three days ago.

  I also have the Sum variables. There is my C_Sum variable, which is the sum of all of Closure related variables (the ones that start with a C) and my H_Sum variable, which is the sum of all of my Healthcare related variables (the ones that start with an H). For both of these variables, I have lagged versions of them by 3 days, 2 weeks and 4 weeks (the same variable names but with '_Three.Days' , '_Two.Weeks' and '_Four.Weeks' attached to the end of them), as well as interaction terms with the percentage of people in that county who are vaccinated (appended with '_Vac').

  So now that we've covered all of the data sources and variables in my data set, let's discuss how I went about creating some of these variables based off of the data from my data sources.

Chapter 3

Methods

In data analysis, it is not enough to simply collect data. You have to mold it into the form you want, polish any rough spots, and create models and visualizations that turn an incomprehensible pile of data into something much simpler that can be easily interpreted. In this section, I'll be going over the methods by which I've used the data I gathered to craft new variables, as well as the steps I took to fix errors in my data and how I determined which observations I was going to use. Let's start off by talking about all the different variables I had to create from the data I had gathered.

3.1 Variable modifications

3.1.1 Cases and Deaths

I'll start off with my data on cases and deaths, since that is the data I used to create the most new variables. I did the exact same things for both cases and deaths so I'll just take about this in terms of cases but everything I'm saying also applies to deaths.

I started with just variables representing the total confirmed cases in a county. I then created a variable representing the average of COVID-19 cases over the past week, generally referred to as the "Rolling seven day average." I did this by simply subtracting what the case count in that county was 7 days ago from the current case count in the county and then dividing that by 7. It's important to prevent what I'm going to call "bleed over," which is when you have are looking at data from the beginning of your time frame and by pulling data from seven days ago you are actually pulling data from the end of your time frame in the previous county. I'll talk more about how I dealt with that issue in my section on lead and lag variables since it's also applicable there. Next, I created a "per capita" version of my rolling average.

I did this by taking my rolling average for a particular observation, dividing it by the population of that county and then multiplying it by 100,000. I also did this for state level data. This variable tells me on average how many people per 100,000 caught COVID-19 in a day over the past week. I then used this variable to create my exponential growth rate variable, which is the variable my model is actually going to be tracking.

3.2 Growth Rate

While the goal of my paper is to estimate what influences the spread of COVID-19 cases, it wasn't initially clear what I should be using as my response variable. I could just use the number of new cases, but that's not really what I'm trying to measure. If I had a rolling average of 40 new cases of COVID-19 per capita the week after I implement a new COVID-19 policy that looks like a failure, but if I had 60 cases per capita the week before I implemented that policy then that's actually really good. So I tried looking at the percentage change in weekly COVID-19 cases, but that didn't feel quite right. What I ended up settling on is a variable measuring exponential population growth. So if I had X0 cases last week and X1 cases this week, the population growth would be the value of r such that X1 = X0 * e^r. So to calculate the value for CaseGrowthRate, I used the equation (ln(X1) – ln(X0))*100, where X1 is the rolling average of COVID-19 cases this week and X0 is the rolling average of COVID-19 cases from a week ago. While multiplying it by 100 isn't really part of the equation for calculating exponential growth rate, I found it beneficial due to some weird quirks of my model. While changing the scale of this variable shouldn't have any impact on the performance of my model, doing so did significantly improve the accuracy of my model. My guess here is that it's something to do with the significant figures on numbers, which have an unusually significant impact when dealing with the data in my model that has relatively small coefficient due to how important a lot of these numbers since their impact is being converted through this exponential growth equation. Also I feel this is a situation where describing it as a percentage is valid, so multiplying it by 100% makes sense. I then did the same but for the exponential growth between one week from now and two weeks from now in order to create my predicted growth rate variable, which will be my response variable in this data set.

While I could have used just straight percentage growth here instead, I feel that exponential growth is a more accurate way of tracking the spread of COVID than a raw percentage change due to the nature of how a virus spreads. It also has the

advantage of being the same both forwards and backwards. So for instance, let's say I had 10 cases of COVID-19 one week, 20 cases the next week, and then 10 cases the week after that. If I just used percentage change, a variable representing that as a percent change would mark that as 100% and -50%. However, if I used exponential growth instead, then that would they would be marked as 69.3% and -69.3%.

3.2.1 Hospitals

One of the larger data sets that I used was a data set tracking what was going on with hospitalizations during the pandemic. I ended up using none of the actual variables from this directly and instead created two new variables for my model. The first was a variable representing the percentage of hospitals reporting that they were short staffed. This is made by taking the number of hospitals in the state that reported being short staffed and dividing it the number of hospital in the state that were reporting any data to the federal government on COVID-19. It's worth noting that the number of hospitals in a state reporting COVID-19 data to the government is not equivalent to the total number of hospital in that state since it took a while for some hospitals to report any data and I assume there are some that have still not reported data. However I feel like it is good enough considering that by the time that hospitals started getting short staffed thanks to COVID-19, the vast majority of hospitals had started reporting by the end and I assume the ones that never reported are very few and likely very small.

The other variable I made out of this data set is a variable representing the percentage of hospital ICU beds in use by people who had COVID-19. This is derived by simply dividing the total number of people in an ICU bed because of COVID-19 by the number of ICU beds total in hospitals reporting data. Similar to the last variable, this isn't exactly representative of all hospitals since not all hospitals were reporting, especially at the beginning, but I feel that because this is a percent in each hospital rather than a strict binary like "Are you short staffed or not" this variable is still pretty meaningful regardless of the total number of hospitals reporting.

3.2.2 Vaccination data

The data I received on vaccinations contained 6 variables that I ended up using: the percentage of people 65 or older who are vaccinated, the percentage of people who are 18 or older who are vaccinated, the percentage of people vaccinated for all ages the number of people 65 or older who are vaccinated, the number of people 18 or older who

are vaccinated, and the number of people vaccinated in total. This data is updated daily and I have it for every county except for Texas, Hawaii, and California counties with less than 20,000 people, which I'll discuss later. For now, let's talk about the new variables I made with this data. First I calculated the number of people between 18 and 64 who are vaccinated by subtracting the number of people 18 or over who are vaccinated from the number of people over 65 who are vaccinated. I then used my age data and population data to create some new variables: The number of people in the county who were unvaccinated, the number of people in the county who were 65 or older and unvaccinated and the number of people in the county who are 18 to 64 and unvaccinated. I am using data in this way rather than just using percentages because using percentages is complicated when dealing with subcategories. I wouldn't just be able to tell you what the effect of increasing the vaccination rate of people who are 65 years by 10% is because that is going to increase the vaccination rate of everyone by some percentage and I won't know that percentage without knowing the underlying age data.

Next I moved on to state level data. I subtracted the county level total vaccinated population from the state level total vaccinated population and then divided it by the state population minus the county population to effectively create the vaccination rate of the state minus that one county. I then subtracted this rate from the county's vaccination rate to get their difference. The goal of these two variables is to see how much it matters if a county outperforms or under-performs the rest of the state in vaccinations.

3.2.3 Delta Variant

The Delta variant of COVID-19 is the primary cause for the reigniting of the pandemic and as a result something I absolutely needed to include. The data I gathered is from the CDC and it tracks the percentage of different variants of COVID-19 across 10 different regions of the country While the CDC does report many different variants and sub-variants of COVID-19, some of which were at one point relevant, I decided that for my data I just needed to group everything into two categories: Delta Variant and non-Delta Variant. So I created a variable measuring the percentage of cases in a region caused by Delta or a Delta sub-variant and added it to my main data frame by using a crosswalk to map each region to a county. Since the CDC only reported variant data a few times a month, I then had to fill in the blanks using a function that interpolated NA values in a vector.

3.2.4 Testing Data

Another variable I collected is data on how much total COVID-19 testing each state does and how many of those tests ended up being positive. Using this data, I created two variables that I ended up including in my model. First I created a variable that showed how much testing the state had done in the past week, sort of similar to my rolling average variable although I didn't divide it by 7. I then divided this number by the state population and multiplied it by 100,000 to get the weekly tests per capita. This is important for my model because testing is an important part of managing to COVID-19 response and not doing enough can lead to things getting really bad before you even really know what's going on. The second variable I made is a positive test ratio. This variable represents the percentage of tests done in the past week that were positive and it is calculated by diving the number of confirmed cases that week. This is important to my model for the same reasons as my weekly testing rate. If the percentage of positive test results is high, that can mean that not enough tests are being done which is a problem.

3.2.5 Lags and Leads

Since policy decisions can some time takes a while to kick in and the growth rate of cases might take a while to respond to things, it is necessary to use lags and leads in my model. A lag of a variable is simply that variable but from a specific time frame ago, while a lead is that variable but from a specific time frame in the future. It's important to remember the "bleed over" effect I talked about before in my section on case growth, where you accidentally end up pulling data from a different county. To avoid this issue when using leads, what you need to do is actually create a data frame that goes beyond where the cutoff point of your model is and base the lead data off of that before trimming it from your data set. So for my expected case growth variable, the variable I'm trying to estimate with my model, I base it off of the difference in cases between 7 days from now and 14 days from now, meaning that I actually to include case data all the way up to September 15th before cutting off it off at September 1st. With lags however, I need to either include data from before where my data starts or infer that it is zero if it's something that just wasn't recorded before that point because it didn't exist, such as the number of weekly tests. My model includes a two week lag on weekly testing per capita variable and my test positive rate, as well as a three day lag on all of my government policy data, and a two week as well as a four week lag on just the C_Sum and H_Sum variables.

3.2.6 Powers

For measuring the impact of current case growth on future case growth, I didn't feel it was appropriate to use a linear data for reasons that will become apparent in the results section. So instead I made it's impact a cubic function that involves both the square of the case growth rate and the cube of case growth rate. Because my case growth rate can already get pretty big, I decided to divide the squared variable by 100 and the cubic variable by 100,000. Since this is just changing the scale of a variable it doesn't have a real influence on the how much it impacts the model, it just makes the numbers easier to deal with. Otherwise I'd have giant variables and tiny coefficients, which can cause some loss of accuracy because it can cause some significant figures get lost just by how small the coefficient would be.

3.2.7 Interaction terms

While this data set does already have a lot of things that technically count as interaction terms, such as per capita data, this section will be discussing interaction terms that aren't the kind of thing you would ever see outside of the context of a statistical model exploring how different variables interact. Specifically I'm using this to gauge the combined impact of different variables and how they interact with vaccinations. I created variables C_Sum and H_Sum, which are the sums of all the government policy variables surrounding closures and policy variables surrounding healthcare. Closure variables have a C in front of them and healthcare variables have a big H in front of them. Additionally, I also multiplied this variable by the county vaccination rate in order to create a variable with the purpose of seeing how the effect of government policies changes as the vaccination rate changes.

3.2.8 Incorrectly entered data

A noticeable issue I've encountered in my data is that for my total case, death and testing data, there would be some instances where data was entered incorrectly and then later changed such that it would create this large spike that would result in the new case/death/testing variable being negative. Not only is this obviously inaccurate in a big way that could seriously through of my model, but if my new case data was negative by a large enough amount that it meant my rolling average was negative, that would make my case growth variable NA and force me to get rid of it from the model. So what I did is that for any variable where this was an issues, I created a list of all the

points where it went negative. For cases and deaths I created some code that would basically select a range around a negative point that contained whatever mistake was made. For my testing data the list of negative points was small enough that I did it by hand. Then I would simply modify that data so that it simply followed a linear trend from the start of the range to the end of the range. There are still probably a lot of errors in this data set, but this caught the big ones that I found frustrating and I don't believe there is either a good way of dealing with any remaining issues or that it is a significant enough issue to throw off my data.

3.2.9 Missing/Removed Data

As previously mentioned, Hawaii, Texas and California counties with less than 20,000 do not have vaccination data so they have been removed. I might find a way to add these back in, but as of right now that seems fairly unlikely. I'll go more in detail here when I have finalized what I'm doing here.

Additionally, I also removed Alaska from my model entirely. This is because the way Alaska is organized is confusing. Instead of counties they have boroughs with different sorts of classes, census areas, many of my data sources break up Alaska in completely different ways. So I thought I should just remove it.

Lastly, I removed DC from my data because it counts as a "state" in my data but it only has one county and that was causing weird interactions.

One of the other big reasons why I had to remove a lot of my data from this model is the case growth variable. This variable not only requires data from 1 week ago, but it also requires that there be active new cases in that county during the past week for it to not be considered NA. This removes a lot of the data from early in the pandemic before it had reached everywhere, as well as a lot of data in really small counties. However this isn't really much of an issue since that data wouldn't really have much value in my model. Little if any government policies were in place during that time, vaccinations weren't even being talked about and obviously delta wasn't a concern.

As previously mentioned, I also removed data where the Delta variant made up over 40% of the cases in the region. Before doing this, both the Delta variant variable and my vaccination variables were having impacts that I felt was far too small. As in half of a county getting vaccinated overnight would barely move the needle on cases. After looking into my data, I determined that this was because a large spike happened because of the Delta variant right in the middle of the vaccination roll out. This was based on minor oddities with my model, such as how my model believed that a senior

getting vaccinated had 6 times the impact of a non-senior adult getting vaccinated (although both were still way lower than they should have), which I presumed was because vaccination rates among seniors was significantly less correlated with the Delta variant than vaccinations among adults. The correlation between vaccinations and the delta variant resulted in my model being unable to tell exactly how strong these variables are since they sort of cancelled each other out. I tried various ways of untangling them with interaction terms, lags and powers, but none that I tried worked. So, I decided that my model would stop tracking data for a state once the delta variant in that state reached above 40%, preventing the Delta variant from interfering too much in my model while still giving me a rough idea of what kind of impact Delta has before it truly takes off. This does mean that the picture I get of these both vaccinations and the Delta variant is less full, but I think it is worth it for having less biased results.

3.3 Potential Areas of Improvement

As with many research papers, there were a fair amount of stones left unturned that I wish I had gotten to explore some more. However, due to constraints on time, resources and data available, I wasn't able to explore them in this thesis.

The biggest thing that I wish I had is more cross-tab data, specifically for cases. All my case data told me was how many cases there were. If I had case data that broke down infections by age range and vaccination status, I feel like that would have been very significant to my overall model. And while that data does exist, in doesn't exist on a daily basis for each county so I'd either have to completely overhaul everything I've done for a new scope or interpolate based off of less granular data what those cross tabs for my individual counties on individual days would have looked like. However it's not even really clear if I'd get good results by doing that and since it would have been a massive time investment, I didn't pursue them.

Another major area where I wish I had been able to do more was with vaccinations. Both because I had to cut out a very important state from my data as well as several counties and because I wasn't able to untangle the vaccination data from the delta variant data. This is disappointing because by cutting off my data early, I lacked a lot of data on how the pandemic spreads with higher vaccination rates and can't really extrapolate as well for what things would look like at something like a 70% vaccination rate.

I also just generally wish I had more precise data in some of the situations where

I had only state or regional level data. Mainly my government response variables (which didn't have data on policies implemented by city/county governments) and my regional Delta variant data. However there was really no solution here that would not have required a lot more work than I was prepared to do during the time frame I was given, so it was a somewhat necessary loss.

Something else I would have liked to play around with more was the scaling/type of my government response variables. As I previous stated, this data is treated as numeric when it is stored as categorical, and that could potentially create discrepancies between what the impact a change in policy actually is and what my model says the impact is. However, there are ways around this. First, I could treat these variables as categorical without actually using categorical variables by creating dummy variables. Problem with this is that it means creating 34 variables just for my base policies without looking at their lagged counterparts and that wouldn't work for my summed variables. The other option is that I create some sort of scale for each of the variables so that instead of being listed as a 0, 1, 2, 3 or 4, it's listed as a value that is an appropriate scale for the impact of each level of the variable. However I was never able to figure out a good way to calculate what those values should be in a way that wasn't biased.

I also think there is room to explore models similar to this one in purpose but for a different scenario or scope. You could look at other or multiple countries, you could look at just state level data rather than county, you could look at weekly data instead of daily, you could look at deaths or hospitalizations instead of cases, you could look at data a month from now rather than 7-14 days from now, etc. However doing that would have required fairly significant overhauls of my model that I did not have time for, so it is best suited for a separate project.

Lastly, I would not be surprised if there were some minor errors lurking in my data. This data set is massive and has been through many revisions and iterations, so it is impossible for me to guarantee there aren't any mistakes in here that would skew the data. If any do come to my attention after the completion of this thesis and I have the time to adjust it, an updated version will be added to my github with an explanation for what was fixed. However, I don't think any of these mistakes would be something significant enough to throw off my model in a really meaningful way or detract from the big picture.

While there are likely many more options out there to be discussed, however I think we should now bring it back to what I actually did and turn to look at the type of model I picked for my data.

3.4 Modeling

3.4.1 OLS model

When dealing with statistical models, the default option is a linear regression, or Ordinary Least Squares (OLS). The way OLS works is that it calculates the coefficients for your model that would result in the the smallest sum of squared residuals for your model (a residual is the difference between the value your model predictions and what the true value is). While a linear regression is perfectly fine in many situations, there are instances where it is simply not up to the task, as was the case with my data. The big problem I ran into when trying creating a linear model off of this data is that it is all highly correlated. In general, methods of dealing with the COVID-19 pandemic are implemented in response to rising COVID-19 case numbers and multiple different methods will be implemented at once, meaning that they all tend to rise and fall together. This means that an OLS regression on my data would be overfitted, or extremely sensitive to small changes in data because it tries too hard to make the residuals small and it ends up creating an extremely jittery model where many of the coefficients are more influenced by random noise than they are underlying trends. For example, early on I ran an OLS regression where the coefficient for the variable measuring school closings was positive and statistically significant. When I changed the lag that variable was using, it changed to being negative and statistically significant. But when I changed the lag another variable was using, the coefficient for school closings went back to being positive and statistically significant. This is a clear case overfitting, so I decided the best course of action was to switch from an OLS regression to a ridge regression.

3.4.2 Ridge regression

A ridge regression is a particular type of regression model generally used when dealing with variables that are highly correlated. How it works is similar to a regular OLS model, but it imposes an additional penalty based on the sums of the squares of my coefficients. This value is multiplied by some value lambda to determine how big the penalty is. What this does is that it discourages your model from assigning to too large of a coefficient to a variable unless it has a really good reason to. This is good because it effectively spreads the impact of highly correlated variables out over each other rather than trying to account for all these small variations in my data by assigning large coefficients that counteract each other to some degree in order to

create a more jittery prediction that is overfitted to the random noise in my data. Theoretically I could have used a LASSO regression here instead, which is basically the same except it uses the absolute value of coefficients instead of their squares, but I decided against that. If the impact of a variable is small the a LASSO regression will just outright set the coefficient of that variable to zero, while a ridge regression won't thanks to how squaring small numbers with an absolute value less than 1 works. Because I have the impact of my stuff like government interventions and vaccinations spread out over many variables, I don't want a LASSO model picking off small parts of it just because the impact of that part is small despite proportionally being very impactful for how big of a piece of the overall pie it is.

The downside of ridge regression is that it does not give me things such as p-values or a confidence interval that I could use to discuss the likelihood of certain variables having no correlation with expected case growth. However, I don't think that's too big of a deal because I'm not really setting out to prove a relationship here. At least not for the main variables I'm going to be discussing. For many of these variables, such as the rise in the delta variant, vaccinations, or the government policies I've included, it is either already proven that these variables do have a statistically significant impact on the spread of COVID or it's just so obvious that no one has bothered to try and prove that. Their impact is a given and not something I would want to go out of my way to prove, I'm more interested in exactly how big the impact is. And while it would be nice to have 95% confidence intervals for my model's coefficients, it's not worth the costs of using a linear regression. I could remove a lot of the variables in my model to prevent overfitting concerns, but that means either tossing out data that I know for a fact is relevant and will remain lurking in my model just with other closely correlated variables as proxies. Additionally, it is less the specific values I am interested in and more the general trends of my data and it's implications. Ultimately I think ridge regression is the best option I have. So let's look at what I used this regression to create.

Chapter 4

Results

4.1 Model

Here are the results from my rough model of Exponential Case Growth between the rolling case per capita average 7 days from now and the rolling case per capita average 14 days from now. Exponential Case Growth is recorded as a percentage rather than a decimal, which effectively shifts the decimal place over by two. The "As percent" column represents how much of that variable increasing by 1 will have an impact on the overall number of cases between 7 and 14 days from now. So if the "As Percent" column contains a 0.992 for a particular variable, then increment a variable by 1 will decrease the number of COVID-19 cases that happen between 7 days from now and 14 days from now by 0.8%. For creating this model, I took 70% of my data for my training data set that I build my model off of and kept 30% of my data for later as a testing data set to calculate how effective my model is on data that isn't factored in to it.

Variable	Coefficent	As % change
Intercept	15.810148	1.171285
HospitalStaffingShortage	-2.550000	0.974798
HopsitalBedsCovidPercent	6.130000	1.063260
Population	0.000908	1.000009
State Pop	0.000108	1.000001
TotalCasesPerCapita	-0.000594	0.999994
TotalDeathsPerCapita	-0.012800	0.999872
Confirmed State Cases	-0.001570	0.999984
Confirmed State Deaths	-0.048400	0.999517
RollStateDeathsCapita	-4.330000	0.957590

	Variable	Coefficent	As % change
11	CaseGrowthRate	-0.081100	0.999189
12	CaseGrowthRateState	0.116000	1.001159
13	RollCasesPerCapita	-0.150000	0.998506
14	RollDeathsPerCapita	-1.130000	0.988813
15	State_Vaccinated_All_Percent	-6.220000	0.939702
16	State_Vaccinated18To64_Percent	-6.840000	0.933884
17	State_Vaccinated_65Plus_Percent	-1.970000	0.980444
18	WeeklyTestsPerCapita	0.000205	1.000002
19	TestPositiveRate	0.157000	1.001570
20	LagWeeklyTestsPerCapita	-0.000210	0.999998
21	LagPositiveRate	-6.640000	0.935715
22	Delta	11.000000	1.116666
23	VaccinationRateRestOfState	-6.170000	0.940180
24	VaccinatedDiff	14.900000	1.160787
25	C1_School.closing	-0.056300	0.999437
26	C2_Workplace.closing	-0.066400	0.999336
27	C3_Cancel.public.events	-0.355000	0.996456
28	C4_Restrictions.on.gatherings	-0.230000	0.997700
29	C5_Close.public.transport	-0.324000	0.996769
30	C6_Stay.at.home.requirements	-0.460000	0.995411
31	C7_Restrictions.on.internal.movement	0.329000	1.003295
32	E1_Income.support	-0.525000	0.994760
33	E2_Debt.contract.relief	-0.019200	0.999808
34	H2_Testing.policy	-0.165000	0.998349
35	H6_Facial.Coverings	-0.268000	0.997326
36	H8_Protection.of.elderly.people	0.038000	1.000380
37	C1_School.closing.lag	-0.321000	0.996792
38	C2_Workplace.closing.lag	-0.419000	0.995819
39	C3_Cancel.public.events.lag	-0.758000	0.992450
40	C4_Restrictions.on.gatherings.lag	-0.319000	0.996819
41	C5_Close.public.transport.lag	-0.401000	0.995999
42	C6_Stay.at.home.requirements.lag	-0.776000	0.992268
43	C7_Restrictions.on.internal.movement.lag	-0.301000	0.996993
44	E1_Income.support.lag	-1.220000	0.987856
45	E2_Debt.contract.relief.lag	-0.316000	0.996849
46	H2_Testing.policy.lag	-0.107000	0.998932
47	H6_Facial.Coverings.lag	-0.370000	0.996306
48	H8_Protection.of.elderly.people.lag	-0.197000	0.998032
49	C_Sum	-0.082400	0.999177

	Variable	Coefficent	As % change
50	H_Sum	-1.42e-01	0.998585
51	C_Sum_Vac	-5.53e-01	0.994482
52	H_Sum_Vac	-5.20e-01	0.994816
53	retail_and_recreation_percent_change	2.47e-02	1.000247
54	grocery_and_pharmacy_percent_change	9.52e-04	1.000010
55	parks_percent_change	1.19e-03	1.000012
56	transit_stations_percent_change	1.32e-02	1.000132
57	workplaces_percent_change	1.43e-02	1.000143
58	residential_percent_change	-5.01e-02	0.999499
59	Over65Percent	-3.74e+00	0.963285
60	Age18To64Percent	-1.27e+00	0.987404
61	Metro_2013	8.95e-03	1.000090
62	Unemployment_rate_2020	1.42e-01	1.001422
63	Median_Household_Income_2019	-1.12e-05	1.000000
64	AveHouseholdSize	6.57e-01	1.006589
65	VaccinatedAllPerCapita	-5.32e-05	0.999999
66	Vaccinated18To64PerCapita	-1.18e-04	0.999999
67	VaccinatedOver65PerCapita	-6.34e-05	0.999999
68	UnvaccinatedAllPerCapita	5.31e-05	1.000001
69	Unvaccinated18To64PerCapita	9.07e-05	1.000001
70	UnvaccinatedOver65PerCapita	2.01e-05	1.000000
71	C_Sum_Three.Days	2.67e-02	1.000267
72	C_Sum_Two.Weeks	1.24e-02	1.000124
73	C_Sum_Four.Weeks	-4.31e-03	0.999957
74	H_Sum_Three.Days	1.21e-01	1.001210
75	H_Sum_Two.Weeks	4.76e-02	1.000476
76	H_Sum_Four.Weeks	5.15e-02	1.000516
77	CaseGrowthPastTwoWeeks	-1.61e-03	0.999984
78	CaseGrowthPastThreeWeeks	-1.02e-03	0.999990
79	CaseGrowthPastFourWeeks	-1.75e-03	0.999982

Model Stats:
Lambda = 81.13908
RMSE = 3896.1
NRMSE = 3.0622
R^2 = 0.10632
IQR predictions: 19.14
IQR actual: 64.79

I picked this value of lambda mostly through trial and error and just intuition about how much I wanted to penalize large coefficients. The RMSE, NRMSE, R^2, IQR prediction and IQR actual are all calculated off of my testing data.

4.2 Intrepreting Impact

4.2.1 Moderate and Extreme Change

Now it can be hard to interpret exactly what these coefficients mean because of the different scaling on all of them. We have variables that go from 0 to 1, while other variables go from 0 to 100,000. To provide some context for what these numbers mean, I've created the following table based off of percentile measurements of my data. The "Moderate Change" column represents the difference between the 25th percentile and the 75th percentile, also known as the IQR. The "Extreme Change" column represents the difference between the 2.5th percentile and the 97.5th percentile. The "Moderate Change Impact" and "Extreme Change Impact" columns represents the change the difference these respective changes in a variable will have on my model's prediction. To fit this data on to the page without making it either super small or going off the page, I've split it up a fair amount.

Variable.name	Moderate Change	Extreme Change
HospitalStaffingShortage	0.15898703	0.35
HopsitalBedsCovidPercent	0.06587837	0.2044994
Population	71.46	842.756
State Pop	7333.01	20715.678
TotalCasesPerCapita	7805.4599	13535.20799
TotalDeathsPerCapita	137.90876	346.1417
Confirmed State Cases	433.86493	1653.139
Confirmed State Deaths	7.534999	33.64157454
RollStateDeathsCapita	0.3899388	1.4318313
CaseGrowthRate	65.83722	291.4999
CaseGrowthRateState	32.64214	106.60562
RollCasesPerCapita	28.295378	110.1676588
RollDeathsPerCapita	0.5405725	3.23477
State_Vaccinated_All_Percent	0.08595768	0.4412886
State_Vaccinated18To64_Percent	0.07144547	0.485566
State_Vaccinated_65Plus_Percent	0.2486483	0.7970161
WeeklyTestsPerCapita	1719.127	5549.1849
TestPositiveRate	0.07403994	0.33451556
LagWeeklyTestsPerCapita	1728.805	5621.2056
LagPositiveRate	0.07462901	0.34825065
Delta	0	0.1947304
VaccinationRateRestOfState	0.08626786	0.4418342
VaccinatedDiff	0.004162305	0.20086837
C1_School.closing	1	2
C2_Workplace.closing	1	3
C3_Cancel.public.events	1	1
C4_Restrictions.on.gatherings	2	4
C5_Close.public.transport	1	2
C6_Stay.at.home.requirements	0	2
C7_Restrictions.on.internal.movement	0	1
E1_Income.support	1	2
E2_Debt.contract.relief	1	1
H2_Testing.policy	1	2
H6_Facial.Coverings	1	3
H8_Protection.of.elderly.people	1	2
C1_School.closing.lag	1	2
C2_Workplace.closing.lag	1	3
C3_Cancel.public.events.lag	1	1
C4_Restrictions.on.gatherings.lag	2	4

Variable.name	Moderate Change	Extreme Change
C5_Close.public.transport.lag	1	2
C6_Stay.at.home.requirements.lag	0	2
C7_Restrictions.on.internal.movement.lag	0	1
E1_Income.support.lag	1	2
E2_Debt.contract.relief.lag	1	1
H2_Testing.policy.lag	1	2
H6_Facial.Coverings.lag	1	3
H8_Protection.of.elderly.people.lag	1	2
C_Sum	3	13
H_Sum	2	5
C_Sum_Vac	0.476	2.934
H_Sum_Vac	0.55	3.573
retail_and_recreation_percent_change_from_baseline	17	67
grocery_and_pharmacy_percent_change_from_baseline	14	52
parks_percent_change_from_baseline	85	275.4
transit_stations_percent_change_from_baseline	31	89
workplaces_percent_change_from_baseline	17	54
residential_percent_change_from_baseline	7	21
Over65Percent	0.0505076	0.1779002
Age18To64Percent	0.042003	0.1524
Metro_2013	1	1
Unemployment_rate_2020	2.7	8.3
Median_Household_Income_2019	16368	60791
AveHouseholdSize	0.25	0.94
VaccinatedAllPerCapita	5800	39700
Vaccinated18To64PerCapita	2807.787	24117.63
VaccinatedOver65PerCapita	2677.235	15967.17
UnvaccinatedAllPerCapita	5800.02	39700.01
Unvaccinated18To64PerCapita	6532.21	32001.42
UnvaccinatedOver65PerCapita	7366.45	24662.183
C_Sum_Three.Days	6	14
C_Sum_Two.Weeks	5	13
C_Sum_Four.Weeks	5	14
H_Sum_Three.Days	2	5
H_Sum_Two.Weeks	2	5
H_Sum_Four.Weeks	2	7
CaseGrowthPastTwoWeeks	97.24611	1160.9438
CaseGrowthPastThreeWeeks	123.72851	1183.2581
CaseGrowthPastFourWeeks	152.34113	1203.1622

Variable.name	Moderate Impact	Extreme Impact
HospitalStaffingShortage	0.995952027	0.99111
HopsitalBedsCovidPercent	1.00405884	1.01264
Population	1.000649002	1.00768
State Pop	1.007991398	1.02268
TotalCasesPerCapita	0.95494234	0.92281
TotalDeathsPerCapita	0.982579996	0.95678
Confirmed State Cases	0.993220774	0.97439
Confirmed State Deaths	0.996365245	0.98386
RollStateDeathsCapita	0.983350457	0.93996
CaseGrowthRate	0.946599556	0.76428
CaseGrowthRateState	1.037817047	1.12459
RollCasesPerCapita	0.958892157	0.84825
RollDeathsPerCapita	0.993936933	0.96426
State_Vaccinated_All_Percent	0.994668337	0.97293
State_Vaccinated18To64_Percent	0.995124849	0.96733
State_Vaccinated_65Plus_Percent	0.995101269	0.98438
WeeklyTestsPerCapita	1.003531199	1.01142
TestPositiveRate	1.000116135	1.00052
LagWeeklyTestsPerCapita	0.996384958	0.98827
LagPositiveRate	0.995067791	0.97715
Delta	1	1.02172
VaccinationRateRestOfState	0.99469276	0.97311
VaccinatedDiff	1.000620401	1.02969
C1_School.closing	0.999437323	0.99887
C2_Workplace.closing	0.999336656	0.99801
C3_Cancel.public.events	0.996468487	0.99647
C4_Restrictions.on.gatherings	0.995427076	0.99083
C5_Close.public.transport	0.996769205	0.99355
C6_Stay.at.home.requirements	1	0.99084
C7_Restrictions.on.internal.movement	1	1.00331
E1_Income.support	0.994787656	0.98955
E2_Debt.contract.relief	0.999808316	0.99981
H2_Testing.policy	0.998354546	0.99671
H6_Facial.Coverings	0.997340557	0.99202
H8_Protection.of.elderly.people	1.000379909	1.00076
C1_School.closing.lag	0.996812556	0.99361
C2_Workplace.closing.lag	0.995836874	0.98751
C3_Cancel.public.events.lag	0.99250663	0.99251
C4_Restrictions.on.gatherings.lag	0.993687504	0.98733

Variable.name	Moderate Impact	Extreme Impact
C5__Close.public.transport.lag	0.995998967	0.99201
C6__Stay.at.home.requirements.lag	1	0.98459
C7__Restrictions.on.internal.movement.lag	1	0.997
E1__Income.support.lag	0.988003786	0.97586
E2__Debt.contract.relief.lag	0.996858806	0.99686
H2__Testing.policy.lag	0.998934219	0.99787
H6__Facial.Coverings.lag	0.996333149	0.989
H8__Protection.of.elderly.people.lag	0.998040108	0.99608
C__Sum	0.997550191	0.98938
H__Sum	0.997204731	0.99301
C__Sum__Vac	0.997369685	0.9839
H__Sum__Vac	0.997145373	0.9816
retail__and__recreation__percent__change__from__baseline	1.004190765	1.0165
grocery__and__pharmacy__percent__change__from__baseline	1.000133345	1.0005
parks__percent__change__from__baseline	1.001008641	1.00327
transit__stations__percent__change__from__baseline	1.004080833	1.01172
workplaces__percent__change__from__baseline	1.002420816	1.00768
residential__percent__change__from__baseline	0.996502203	0.98952
Over65Percent	0.998124259	0.9934
Age18To64Percent	0.999471501	0.99808
Metro__2013	1.000089529	1.00009
Unemployment__rate__2020	1.003874131	1.01192
Median__Household__Income__2019	0.998172177	0.99322
AveHouseholdSize	1.001668861	1.00628
VaccinatedAllPerCapita	0.996920029	0.97911
Vaccinated18To64PerCapita	0.996682978	0.97186
VaccinatedOver65PerCapita	0.998303718	0.98993
UnvaccinatedAllPerCapita	1.003239996	1.02198
Unvaccinated18To64PerCapita	1.006237685	1.03038
UnvaccinatedOver65PerCapita	1.001482909	1.00496
C__Sum__Three.Days	1.00160456	1.00374
C__Sum__Two.Weeks	1.000618693	1.00161
C__Sum__Four.Weeks	0.99978446	0.9994
H__Sum__Three.Days	1.002439924	1.0061
H__Sum__Two.Weeks	1.000954647	1.00239
H__Sum__Four.Weeks	1.001034724	1.00362
CaseGrowthPastTwoWeeks	0.998436701	0.98146
CaseGrowthPastThreeWeeks	0.998737784	0.98798
CaseGrowthPastFourWeeks	0.997328668	0.97904

Now the values here for data involving vaccinations and the delta variant are slightly misleading, because it includes a lot of data from before the first vaccination and before the first recorded Delta case in the United States. So to give those particular results even more context, I've created a new table that does the same thing but looks at just the vaccination data from after the FDA approved the first COVID-19 vaccine (December 11th) and delta data from after the first case of the delta variant of COVID-19 was detected in the United States (May 20th, 2021).

```
knitr::kable(SpecialImpact[1:14,1:3], digits = 4)
```

Variable.Name	Moderate Change	Extreme Change
State_Vaccinated_All_Percent	0.2905	0.4859157
State_Vaccinated18To64_Percent	0.2896	0.5458737
State_Vaccinated_65Plus_Percent	0.6727	0.8331451
VaccinationRateRestOfState	0.2911	0.4870552
VaccinatedDiff (county - state)	0.0464	0.31217177
VaccinatedAllPerCapita	23100	45400
Vaccinated18To64PerCapita	11823	29288.63
VaccinatedOver65PerCapita	10693	18070.41
UnvaccinatedAllPerCapita	23100	45400.00323
Unvaccinated18To64PerCapita	11924	34066.37176
UnvaccinatedOver65PerCapita	10883	24475.34502
C_Sum_Vac	1.366	3.57
H_Sum_Vac	2.08	4.104
Delta numbers	0.1867	0.37083681

```
knitr::kable(SpecialImpact[1:14,c(1,4,5)], digits = 4)
```

Variable.Name	Moderate Impact	Extreme Impact
State_Vaccinated_All_Percent	0.9821	0.9702
State_Vaccinated18To64_Percent	0.9804	0.9633
State_Vaccinated_65Plus_Percent	0.9868	0.9837
VaccinationRateRestOfState	0.9822	0.9704
VaccinatedDiff (county - state)	1.0069	1.0476
VaccinatedAllPerCapita	0.9878	0.9761
Vaccinated18To64PerCapita	0.9861	0.9659
VaccinatedOver65PerCapita	0.9932	0.9886
UnvaccinatedAllPerCapita	1.0123	1.0244
Unvaccinated18To64PerCapita	1.0109	1.0314
UnvaccinatedOver65PerCapita	1.0022	1.0049
C_Sum_Vac	0.9925	0.9804
H_Sum_Vac	0.9892	0.9789
Delta numbers	1.0208	1.042

4.2.2 Aggregated impacts

Another issue with interpreting the above data is that specific variables are designed to go together, but since their impact is split up it can be hard to tell exactly what

their impact is. I've done some calculations to show what the aggregate impact of COVID-19 policies and vaccinations are. Let's start with the vaccination data. This data is broken into 5 categories: minor, mild, moderate, aggressive, severe. These represent the 2.5th, 25th, 50th, 75th, and 97.5th percentiles of various COVID-19 policies respectively. The vaccination percentage change just represents the changing of the C_Sum_Vac and H_Sum_Vac, not vaccination rates as a whole.

```
knitr::kable(Minor)
```

Minor	0% Vaccinatinated	11% Vaccinatinated	25% Vaccinatinated.25	39% Vaccinatinated
Day 0	0.9841959	0.9790361	0.9725083	0.9660240
Day 3	0.9663252	0.9612592	0.9548499	0.9484833
Week 2	0.9685271	0.9634495	0.9570256	0.9506445
Week 4	0.9694416	0.9643592	0.9579292	0.9515421

```
knitr::kable(Mild, row.names = FALSE)
```

Mild	0% Vaccinatinated	11% Vaccinatinated	25% Vaccinatinated.25	39% Vaccinatinated
Day 0	0.9575872	0.9480099	0.9359592	0.9240616
Day 3	0.9141064	0.9049640	0.8934604	0.8821030
Week 2	0.9175168	0.9083404	0.8967939	0.8853941
Week 4	0.9200787	0.9108766	0.8992979	0.8878663

```
knitr::kable(Moderate, row.names = FALSE)
```

Moderate	0% Vaccinatinated	11% Vaccinatinated	25% Vaccinatinated.25	39% Vaccinatinated
Day 0	0.9483484	0.9371854	0.9231678	0.9093599
Day 3	0.8922193	0.8817170	0.8685291	0.8555384
Week 2	0.8942523	0.8837261	0.8705081	0.8574878
Week 4	0.8950967	0.8845606	0.8713301	0.8582975

```
knitr::kable(Aggresive, row.names = FALSE)
```

Aggresive	0% Vaccinatinated	11% Vaccinatinated	25% Vaccinatinated.25	39% Vaccinatinated
Day 0	0.9317837	0.9197294	0.9046129	0.8897448
Day 3	0.8441363	0.8332159	0.8195213	0.8060517
Week 2	0.8486187	0.8376403	0.8238730	0.8103319
Week 4	0.8516816	0.8406636	0.8268466	0.8132566

```
knitr::kable(Severe, row.names = FALSE)
```

Severe	0% Vaccinatinated	11% Vaccinatinated	25% Vaccinatinated.25	39% Vaccinatinated
Day 0	0.9201194	0.9054898	0.8872061	0.8692916
Day 3	0.8167408	0.8037548	0.7875254	0.7716236
Week 2	0.8218762	0.8088085	0.7924771	0.7764754
Week 4	0.8251252	0.8120059	0.7956099	0.7795449

And lastly, my Vaccination data. For the sake of simplicity, the following calculations will be on a county where the percentage of people between the ages of 18 and 64 is 60% and the percentage of people 65 or above is 20% (pretty close to

what their medians are, 0.586 and 0.195 respectively). Additionally, the population of the state is 5 million people and the county's population is 100,000. I've also create three levels of vaccination. "Low," where 25% of the seniors are vaccinated and 10% of non-senior adults are vaccinated (11% of total population), "Medium," where 50% of the seniors are vaccinated and 25% of non-senior adults are vaccinated (25% of total population), and "High," where 75% of the seniors are vaccinated and 40% of non-senior adults are vaccinated (39% of total population). Then I also have columns for when there are no COVID-19 policies implemented, mild COVID-19 policies implemented (25th percentile that was previously described) and aggressive COVID-19 policies implemented (75th percentile described before).

```
knitr::kable(LowStateVaccinations, row.names = FALSE)
```

Low State Vaccinations	no.covid.policy	mild.covid.policy	Aggressive.COVID.policy
low	0.9476146	0.9426467	0.9353556
medium	0.9318169	0.9207513	0.9046451
high	0.9162814	0.8993634	0.8749419

```
knitr::kable(MediumStateVaccinations, row.names = FALSE)
```

Medium State Vaccinations	no.covid.policy	mild.covid.policy	Aggressive.COVID.policy
low	0.9077169	0.9029581	0.8959740
medium	0.8788445	0.8684080	0.8592989
high	0.8479583	0.8323018	0.8097013

```
knitr::kable(HighStateVaccinations, row.names = FALSE)
```

High State Vaccinations	no.covid.policy	mild.covid.policy	Aggressive.COVID.policy
low	0.8856961	0.8810528	0.8742381
medium	0.8707329	0.8603927	0.8453423
high	0.8411782	0.8256470	0.8032272

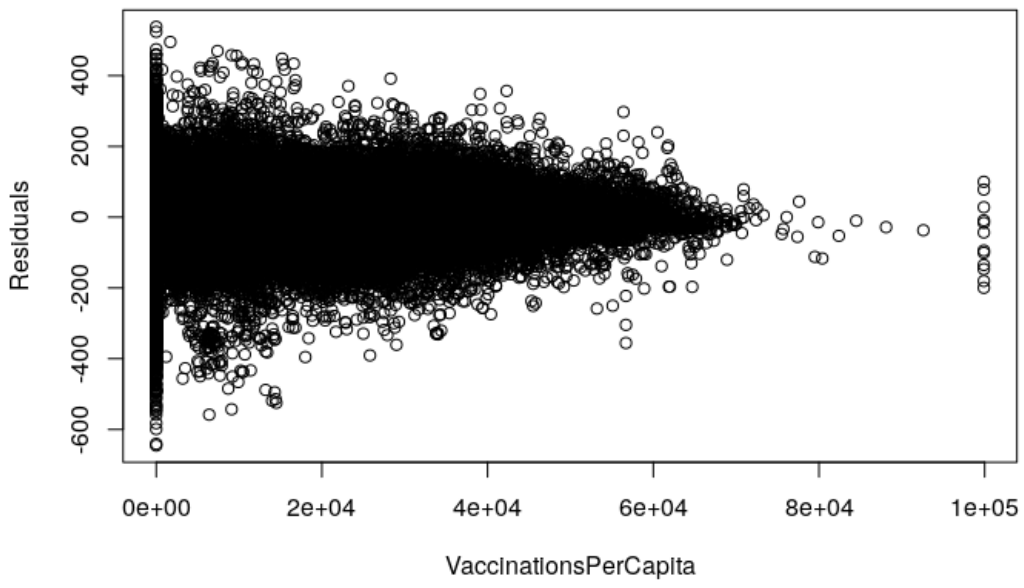
```
knitr::kable(StateVaccineImpact, row.names = FALSE)
```

County vs State	Low State Vac	Medium State Vac	High State Vac
Low County vaccinations	0.9749158	0.9338686	0.9112134
Medium County vaccinations	0.9956555	0.9390540	0.9303867
High County vaccinations	1.0168351	0.9410142	0.9334901

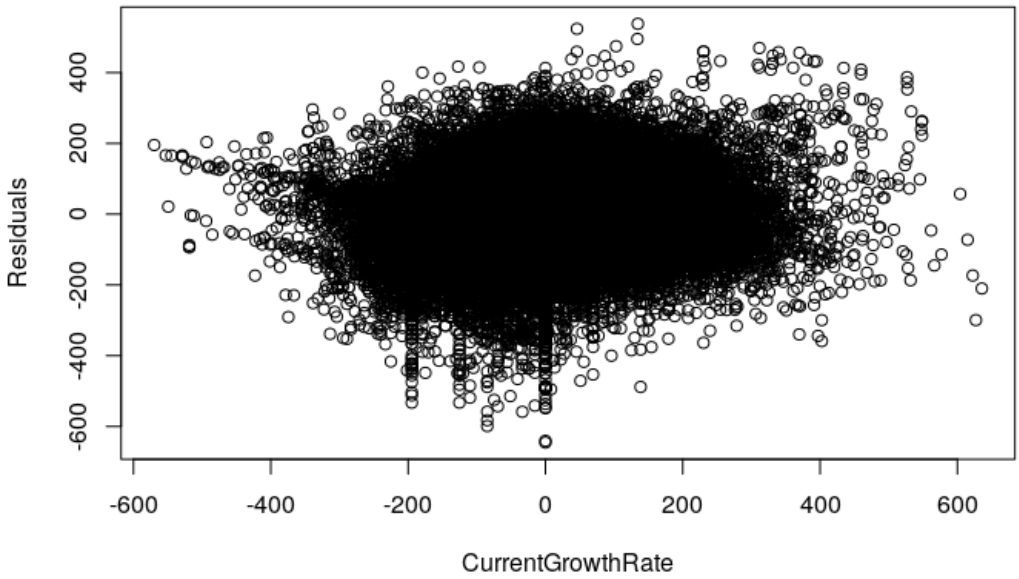
4.3 Validations

For this model I did several different kinds of validations to verify that my model wasn't significantly biased in some way. First I tried plotting some residuals. This is a technique where you look at the value of a variable and see how it lines up with the residuals from the model. If there is a noticeable pattern, that indicates there is something off about that variable's impact on the model that needs to be accounted for. When I did this with my data, the results were a bit lackluster. For the sake of brevity I've included only two of the residual plots that I looked at. One is of county vaccinations per capita among all age groups and the other is of the current growth rate.

```
knitr::include_graphics("VaccinationsResiduals.png", dpi = 125)
```




```
knitr::include_graphics("CurrentGrowthRate.png", dpi = 125)
```



While there does seem to be a significant pattern with the vaccination data, with how it narrows down as vaccinations increase and then flares out at full vaccinations, I think that's mostly just a product of the values being normally distributed and the more common vaccination levels being more likely to have more significant outliers. And for the the growth rate data, there is a bit of a parabolic bend to the residuals if you squint, but considering I've already added squared and cubed versions of the current growth rate and at my limit for how much data I could include in my model, trying to adjust get rid of this small discrepancy with even more powers seems like a task better left for future exploration.

I also ran a k-fold cross validation on my model. The way this works is that I break my model into 5 different chunks, then create five different models using each chunk as the testing data to see what kind of results I get, then averaging those coefficients together and seeing how they compare to my. While there were some significant differences in the coefficients in certain places, those differences became basically unnoticeable when converted to percentages and basically compared to the range of the variables in play. Even in our extreme cases, the impact of a variable in our regular model going from it's 2.5th percentile to 97.5th percentile was within 1% of that variable's effect in the k-fold model for that same situation, with a majority being even below 0.1%. The only two variables that changed the extreme impact by more than 0.5% was CaseGrowthRate at a 0.850% decrease and CaseGrowthRateState at a 0.954% increase. So between this and that residual plot it does seem as though there is something slightly off about the way case growth in my model works, it doesn't

seem like a significant enough issue to throw my model off by much and change the big picture view from the model. Speaking of which, let's get into analyzing this model and examining what exactly this big picture view is.

Chapter 5

Thesis Discussion

5.1 Overview

First lets talk about the stats of the overall model itself. This model has an R^2 of 10.63%, meaning that this model explains 10.63% of the variation in predicted case growth. This is somewhat low compared to what you would find from other statistical models. I've even gotten better R^2 values with this very data set back when I was working with OLS and I had fewer variables. I think this low R^2 is justified for two main reasons. First is simply the nature of this pandemic. Aside from just the fact that it is tied heavily into human behavior and that dramatically lowers what you can expect for an R^2 value, COVID-19 is special because of the potential for super spreaders. Multiple studies have found that about 80% of secondary transmissions were caused by around 10 to 20% of infected individuals.(Lewis, 2021) And it is very difficult to predict when these super spreader events are going to happen, especially for models that can only take in limited data and interpret it in limited ways. The second factor is that it's like this as a deliberate trade off I made to improve the precision of my coefficients. Decisions I made such as limiting my model to before the Delta variant dominated the US did help separate out the effects of the delta variant and my vaccination data, but at the cost of the overall predictive ability of my model. I do think this R^2 value could be improved given more time and data, but this is still something I'm satisfied with and believe is significant enough to be worth discussing.

One of the issues my model does have is it's limited spread. My predicted values for my training data have an IQR of 19.14, while the corresponding true values from my data set have an IQR of 64.79, nearly 3.5 times the IQR of my training data. Additionally, my predicted values max out at 195.8, while my true values max out at 668.4. Interestingly, this is a way bigger gap the difference between the minimum of my

predicted values and the minimum of my true values (-361.2 and -603.6 respectively). These factors combined suggests my model has a hard time predicting the more extreme changes in cases, especially spikes rather than dips. This lines up with what I discussed previously about how my model just can't predict occurrences like super spreader events. Together these lead me to believe the unexplained variation in my model has more to do with data that my model simply isn't accounting for and rather than because the coefficients my model has calculated are significantly off.

Also, something that is important to note with this data is that the impact a variable has on a model compounds with itself. The value my model is predicting, case growth, is inherently based on what the current number of cases are. So if I implement some kind of policy that decreases the growth rate that means it decreases the number predicted for next week, and if that policy remains in place, then not only does it still continue to effect my model, but the decreases that it causes over the week will result in the starting point for next week's predicted case growth to be smaller, meaning that it impacts the expected number of cases two weeks from now twice. So while something like a 5% change in case growth may not seem like much, it is a lot if that factor remains in place for a while.

5.2 Analysis of Secondary Significant Effects

So overall we found several variables that had a substantial impact on my model. In this section I'll be going over all of the ones that had more than a +/-1% influence in the extreme change impact category, except for CaseGrowthRate, aggregated COVID-19 policy, and aggregated vaccinations. Those will be saved for later since those are the most important sections.

To start off with something simple, HopsitalBedsCovidPercent. This variable represents the percentage of hospital beds that are taken up by people infected with COVID-19, and on the extreme change category it had an impact of +1.26%, although theoretically it could go as high as +6.32%. While people in hospital beds aren't out there causing more cases of COVID-19, it is possible they spread a fair amount to people before they were hospitalized and it can be a sign that the situation with the pandemic is getting really bad in a state.

Another fairly simple variable that manages to have a decent impact is state population, with an extreme impact of +2.26%. Interestingly, this variable seems to have a significantly larger impact on my model than the population of a county itself. My guess is that this is related to how some of the states that got hit the hardest by

COVID-19 are the ones with large populations. The three most populated states in my model are California, Florida and New York, and all three of those states had some really severe COVID-19 waves (if Texas was in the model it would have been second, and Texas also had a big COVID-19 wave). My model gives all the counties in those states a +4.37%, +2.35% and +2.13% percent bonus just for their populations.

Next up is total cases and deaths, both on a county and state level. Total cases per capita has an extreme impact of -7.72%, Total deaths per capita has an extreme impact of -4.32%, confirmed state cases has an extreme impact of -2.56%, and confirmed state cases has an extreme impact of -1.62%. My guess here is that these variable are representing a shift in the people who can catch COVID-19. People who have already caught COVID-19 are less likely to catch it again, and people who died from COVID-19 can't catch it again for obvious reasons. There might also just be a time element here where these variables are acting as a proxy for how far into the pandemic we are and just the longer we've been dealing with the pandemic the better we are at dealing with it.

The results of my COVID-19 testing data are a bit weird when you first look at them. My weekly testing per capita variable has an extreme impact of 1.14%, the test positive rate has an extreme impact of 0.052%, the lag of weekly testing per capita has an extreme impact of -1.17% and the lag of the is the test positive rate is -2.29%. The impact of testing per capita makes sense, as testing capacity increases we are better equipped to deal with the pandemic. But the positive test rate having a small (but positive) impact while the lag of the test positive rate having a negative impact on case growth is pretty odd considering that a high test positive rate is considered a major sign that things are going wrong. My guess is that the test positive rate being high is an indication that we are about to turn a corner on this situation soon, and that's what the lag is capturing. These variables also have overlap with the H2_Testing variable and therefore all the H_Sum variants and the 3 day lag of H2_Testing, so that potentially explains some of why this variable is weird.

The delta variant is a bit weird in this data set because I've deliberately structured my data so that it doesn't include data from when the delta variant makes up 40% or more of the cases of Delta in the region. This was necessary to untangle it from my vaccination data, and it paid off here because it resulted in a Delta variable coefficient that was roughly 6 times bigger than what it was in my old model that didn't make this change. This gives us a Delta variant extreme impact of +4.02%. If you extrapolate with the Delta variant to values above 0.4 and look at the difference between when the Delta variant wasn't in the US at all and when it became basically

the only variant in the country, we get an impact of +12.52%. Now we shouldn't take this variable as seriously as we take variables that aren't extrapolated, but I do think it's worth mentioning that an impact of +11.6% seems to be pretty close to the opposite of the impact of our combined vaccination data, where a county that falls into the -11.5% to -16.8% range when you're looking at the more "middle of the pack" data from my previous combined vaccination data in the results chapters. So it feels almost like the impact of people getting vaccinated has been canceled out by the impact of the Delta variant (and vice versa), which fits with how the data was extremely difficult to untangle.

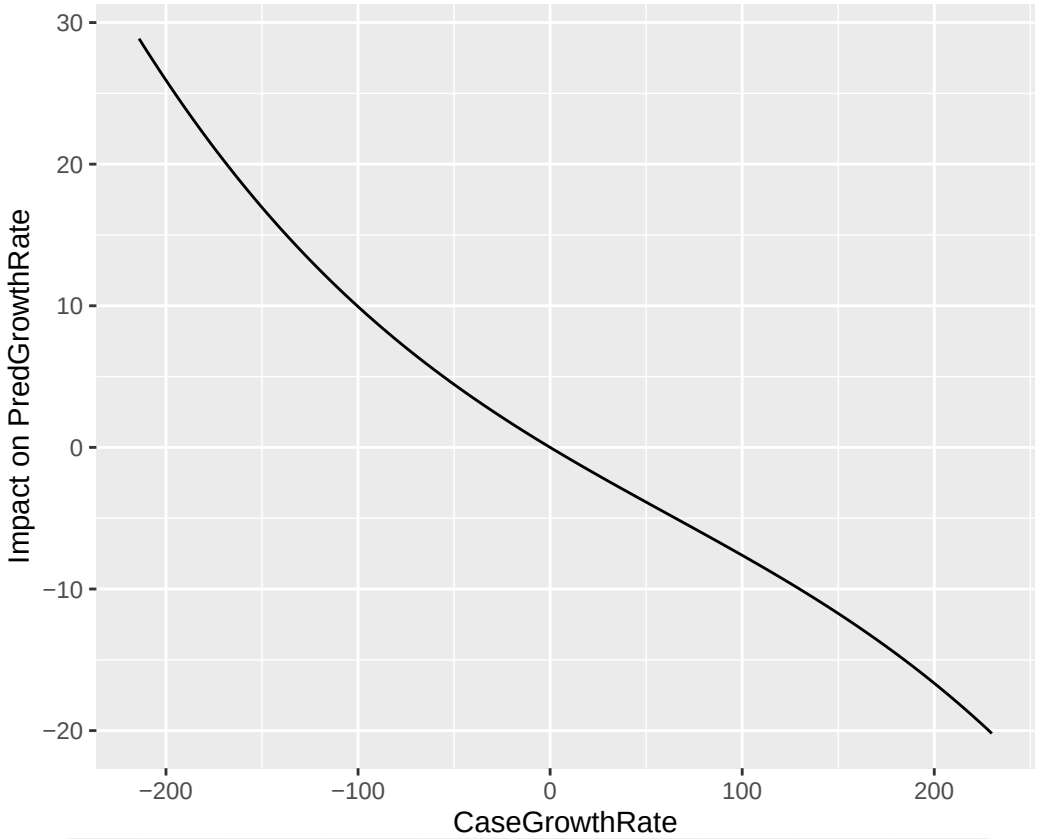
The mobility data I've included is another example of data where each variable doesn't really carry much weight on it's own but combined together it does actually have a fairly significant impact. In order their isolated extreme impacts are retail and recreation at +1.65%, transit stations at +1.25%, residential areas at -0.93%, workplaces at +0.77%, parks at +0.33%, and groceries and pharmacies at 0.05%. This results in a combined extreme impact of people staying more at their homes rather than going to other locations of -4.83%. There is probably a fair amount of overlap between this and variables such as stay at home orders, cancellation of public events and workplace closings, so there is some fuzziness here, but it's still a good tool and can account for things those other variables don't, such as how rigidly are people actually following stay at home instructions.

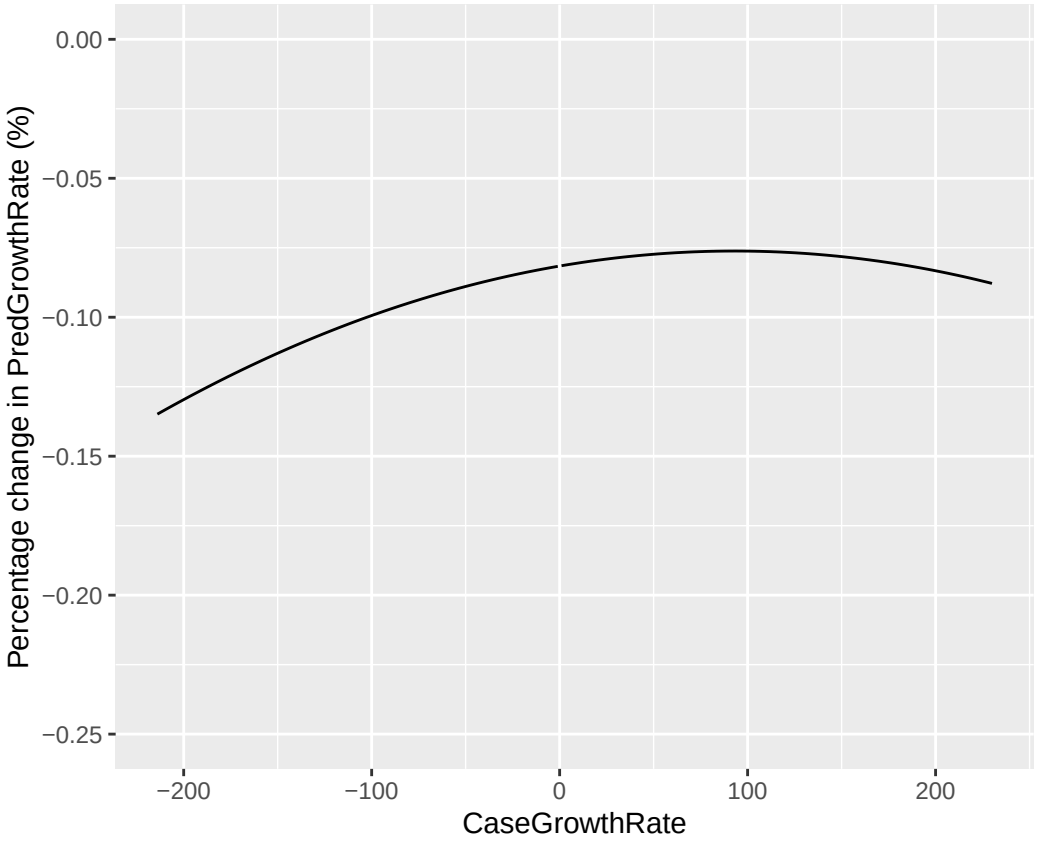
5.3 Analysis of primary effects

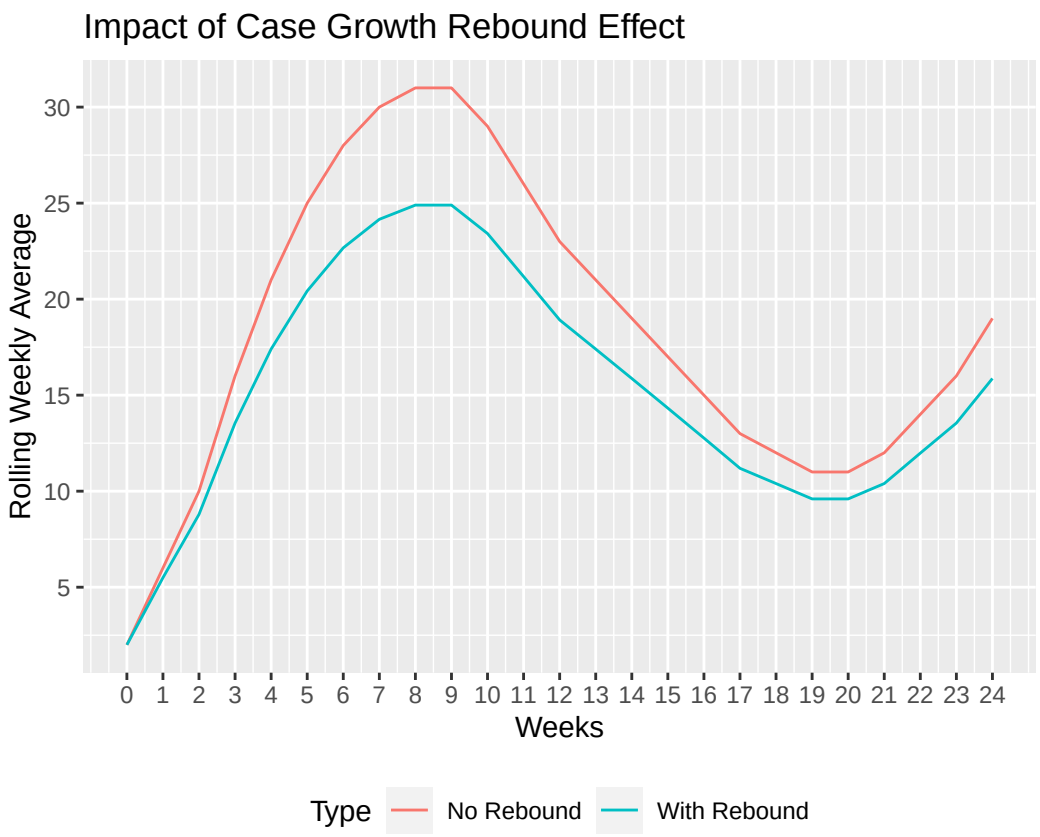
Overall the three variables are the ones that I feel are the most important and merit the most talking about. Let's start with CaseGrowthRate, since this is relevant to discussing the other two later.

5.3.1 Case Growth Rate Impact

Since I'm including not just a variable representing the case growth rate, but also the case growth rate squared and cubed and they are multiplied by these different powers of ten, it can be a little tricky to see how they interact. To help visualization things, here are two graphs showing the impact of CaseGrowthRate on PredGrowthRate by showing it as just a raw change and by showing it as the percent change in the predicted growth rate.


30
20
10
0
−10
−20
Impact on PredGrowthRate
−200
−100
0
100
200
CaseGrowthRate


0.00
−0.05
−0.10
−0.15
−0.20
−0.25
Percentage change in PredGrowthRate (%)
−200
−100
0
100
200
CaseGrowthRate



So basically, if cases are currently on the rise, then based off of the CaseGrowthRate variable we'd expect it to start decreasing, and if cases are slowing down, then based off of the CaseGrowthRate variable we'd expect it to start increasing again. Additionally, this effect is increases the further away from 93.15 the current case growth rate gets (to clarify, it still always pulls variables to 0, it's just that the pull is weakest at 93.15). At 93.15 it only decreases the expected case growth by 7.6%, the effect remains under 10% for about 95% of the data, and for a few outliers it can get up to around to over a 20% decrease (that's only for around 0.05% of the data however).

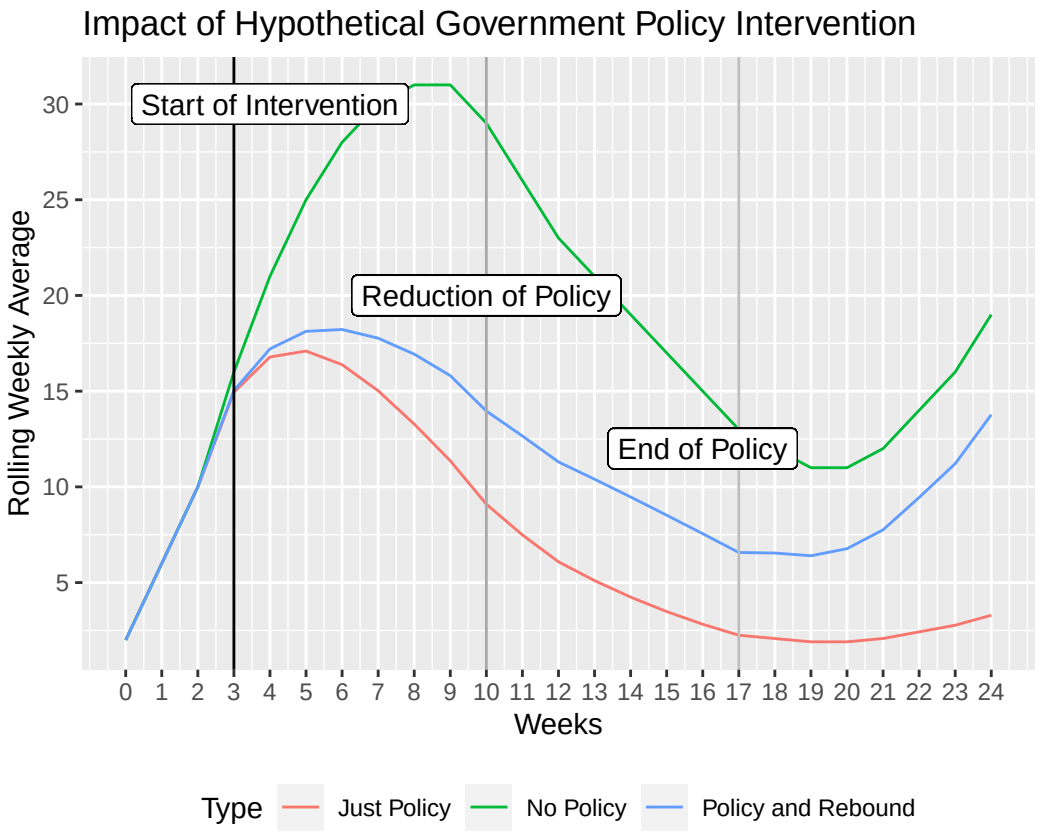
Basically this variable acts like a rubber band, pulling the expected value closer to 0 and pulling harder for more extreme values. What this also means is that if there is something that changes the case growth rate, the impact of that change is also slowly reduced to 0. If you have something at the start of 2020 that decreases the case growth for a week by 40% but does nothing for the rest of the pandemic, that doesn't mean you're going to have 40% less cases for the entire rest of the pandemic. Which is something we need to keep in mind when discussing our next variables of interest, government policy.

The flat rolling data for cases (as well as deaths) also has a pretty important impact. The rolling cases per capita average has an extreme impact of -14.5%, rolling

death per capita average has an extreme impact of -3.53%, and rolling deaths per capita average for the state has an extreme impact of -5.67%. My interpretation is that this represents how there is a sort of soft cap to how high rolling averages can get and this controls from that outside of how quickly COVID-19 is growing. The higher a rolling average gets, the more likely it is to be reaching a peak and starting to turn around or for it to already be in active decline.

5.3.2 Government Policy Impact

While most of the individual impacts of the Government policy variables are small and understated, their impact together is truly immense. Their impact over time is interesting as well. Below I've laid out a hypothetical situation for a COVID-19 response from the a local government and tracked it over 20 weeks. The state government decides to implement aggressive restrictions starting two weeks into this time frame, then reduces it after 7 weeks, and lifts it all together after 16 weeks. There are no vaccinations. I imputed what the scenario would be like if there was no response from the government, then calculated how the situation would change if the government did respond before adjusting for the passage of time, then I added in the how in the rubber band effect. I sort of ball-parked these variables, couldn't really come up with a good way to actually use my model to calculate them.




```
MediumRestriction[25]/CaseLoad[25]
```

```
[1] 0.724724
```

```
GovtPolicy[25]/CaseLoad[25]
```

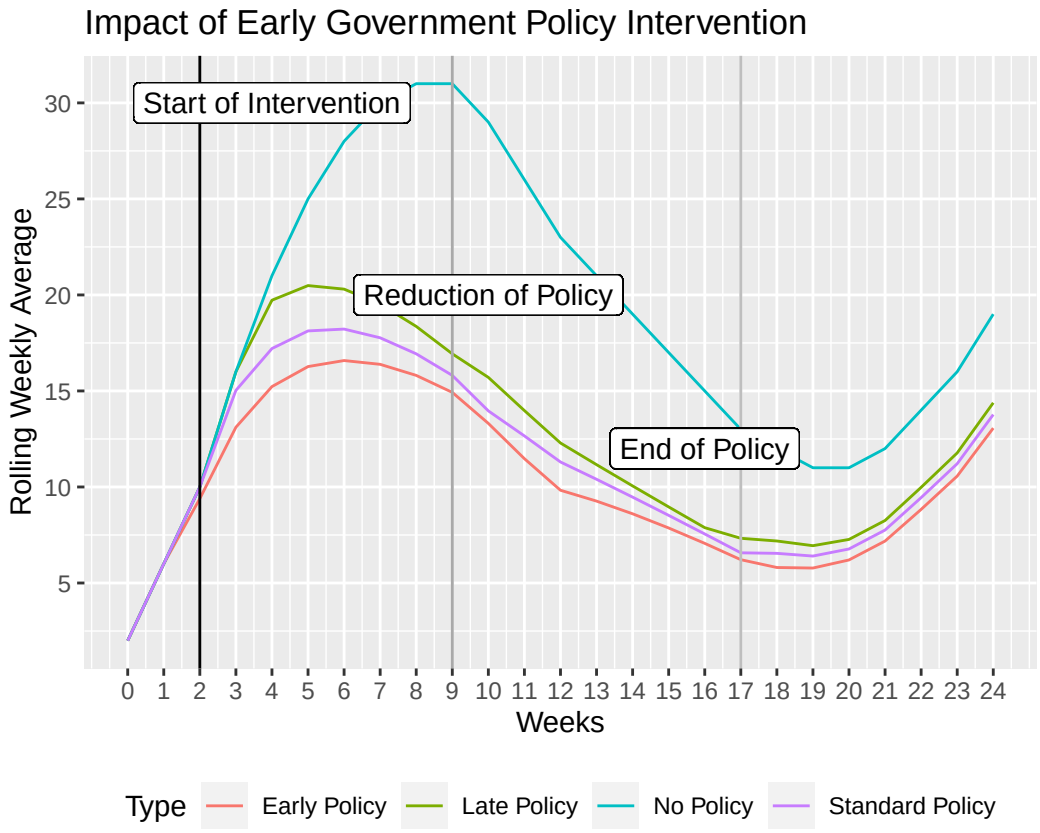
```
[1] 0.1732811
```

```
sum(MediumRestriction)/sum(CaseLoad)
```

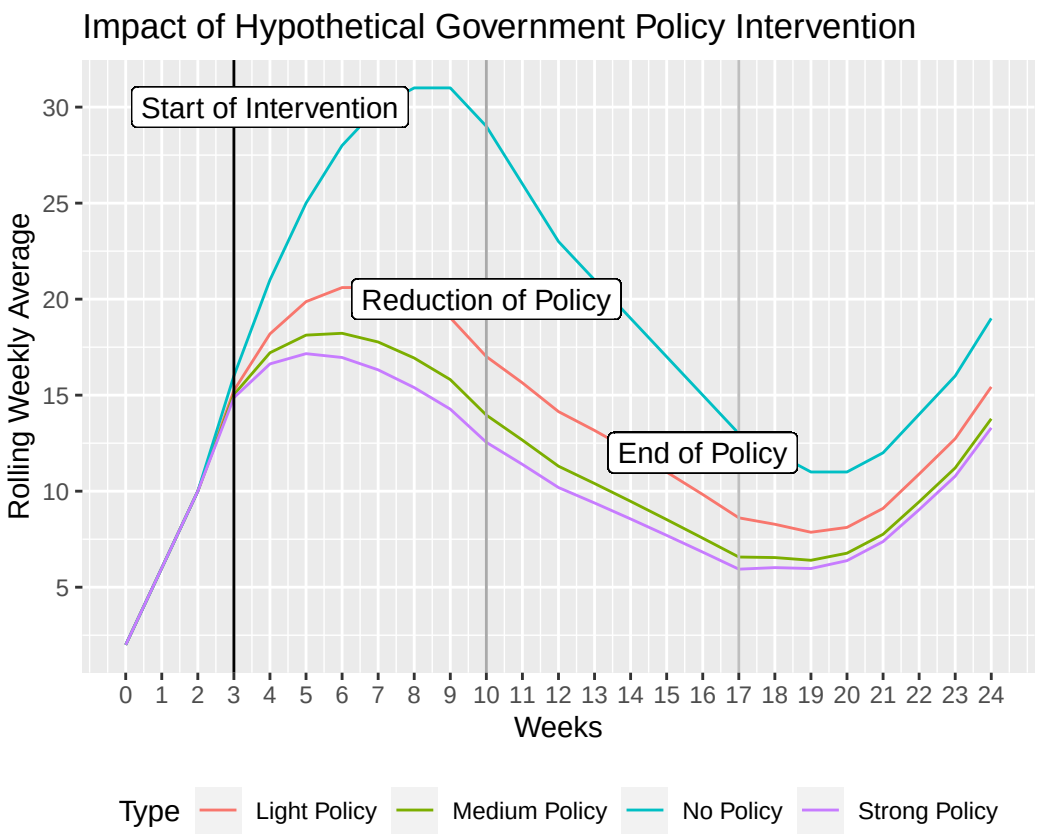
```
[1] 0.6101408
```

So not only does government policy have a pretty big impact, but so does the time impact of this rubber banding effect. Without the decay of the policy's impact over time, cases after twenty weeks would be only 17.3% of what they would have been without any intervention. However, once you include the effect of time, it goes up to around 72.5% of the cases. Which is still really impactful. Overall this policy intervention, accounting for the time lag, decreased the total number of cases over this time span to 61.0% of what they otherwise would have been.

It's also important when these policies are implemented and for how long. Cutting off a bad situation in the larval stage and keeping in place aggressive policies until the peak is behind can prevent things from getting out of hand, while implementing a policy too late and stopping or lessening it too early. Below I've graphed two slight variations of this same situation, one where the aggressive restrictions are implemented a week earlier and left in place before reduced to more mild restrictions one weeks later then removed a week later, and another where the aggressive restriction is implemented a week later and are reduced to mild restrictions a week earlier, and removed a week earlier.



Let's also look at what happens when we both decrease the policy levels in our equations to Moderate and Minor as well as increase them to Severe and Moderate.




```
1- sum(EarlyRestriction)/sum(CaseLoad)
```

```
[1] 0.4393654
```

```
1- sum(LateRestriction)/sum(CaseLoad)
```

```
[1] 0.3395527
```

```
1- sum(MediumRestriction)/sum(CaseLoad)
```

```
[1] 0.3898592
```

```
1- sum(EarlyRestriction)/sum(LateRestriction)
```

```
[1] 0.1511289
```

```
1- EarlyRestriction/LateRestriction
```

```
 [1] 0.00000000 0.00000000 0.06065225 0.18055187 0.22811975 0.20584918
 [7] 0.18313679 0.16062339 0.13892772 0.11858085 0.15248212 0.17917087
[13] 0.20011600 0.16999691 0.14429634 0.12239834 0.10376368 0.15212237
[19] 0.19206370 0.16747176 0.14694024 0.12967137 0.11493432 0.10243908
[25] 0.09161317
```

```
1- EarlyRestriction[25]/LateRestriction[25]
```

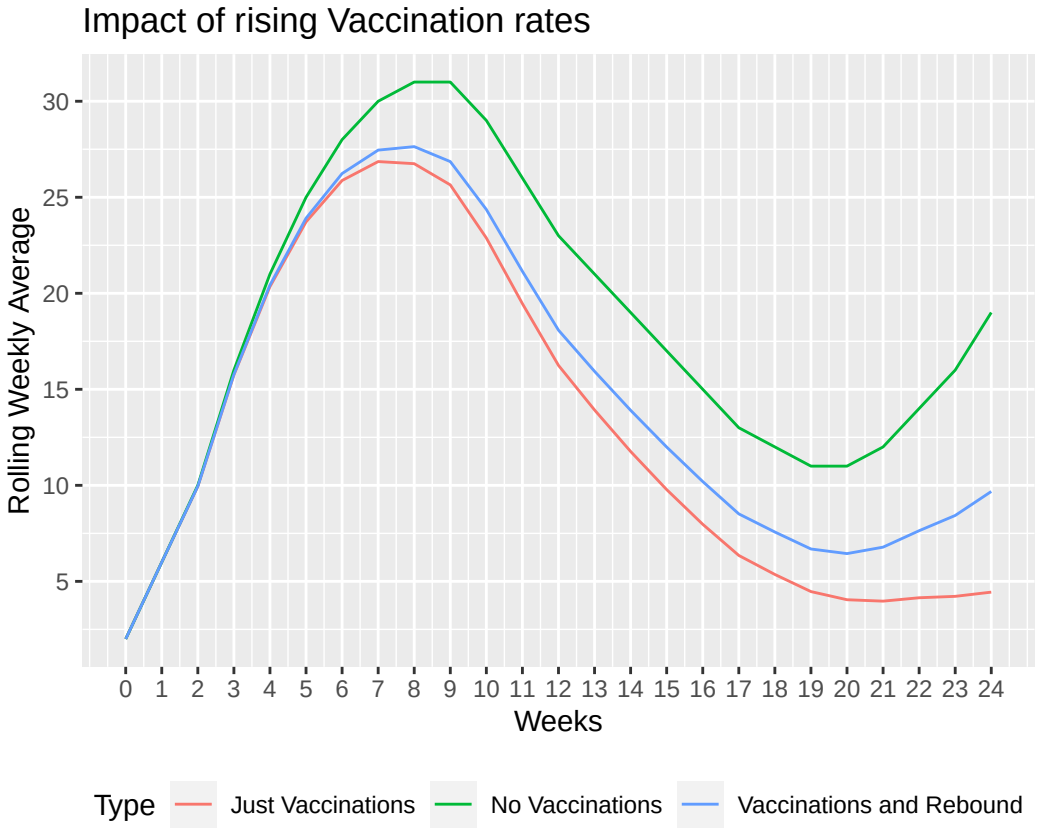
```
[1] 0.09161317
```

In comparison to the 38.99% decrease in total cases that we had in the original version, the early version decreased total cases by 43.93% while the late version reduced total cases by 33.95%. This also means that the early version had 14.8% fewer cases than the late version. So while late is far better than never, being early is still really important. What is interesting though is how this gap between the early version and the late version changes over time. At week 4 there is a 22.8% difference between the two, but by the end that's gone down to just 9.16%. And that gap is only going to continue to narrow as time goes on thanks to the rubber band effect. It seems like this policy restrictions are good at slowing down the pandemic while they are in place, but they don't really have any kind of long lasting impact once they get lifted. Which leads in nicely to examining what the impact of vaccinations are.

5.3.3 Vaccination data

While definitely vaccinations are definitely similar to government interventions in that both play a vital role in dealing with the pandemic, they play two very different roles. To illustrate what I mean, let's do a similar hypothetical situation to what we did with government policy but for vaccinations instead of policy. We're keeping the same time frame and case load, but now instead of looking at a changing policy response, we're going to be looking at rising vaccinations. For this we're going to assume that a county starts off with zero people vaccinated, and over the course of the 20 weeks steadily progresses to the medium vaccination level square with mild COVID-19 policy in my table for a state with a medium level of vaccinations, where vaccinations have a -12.1% impact on expected case growth. Like we did with government policy, we're gonna look at how this data changes when we just include the impact of the

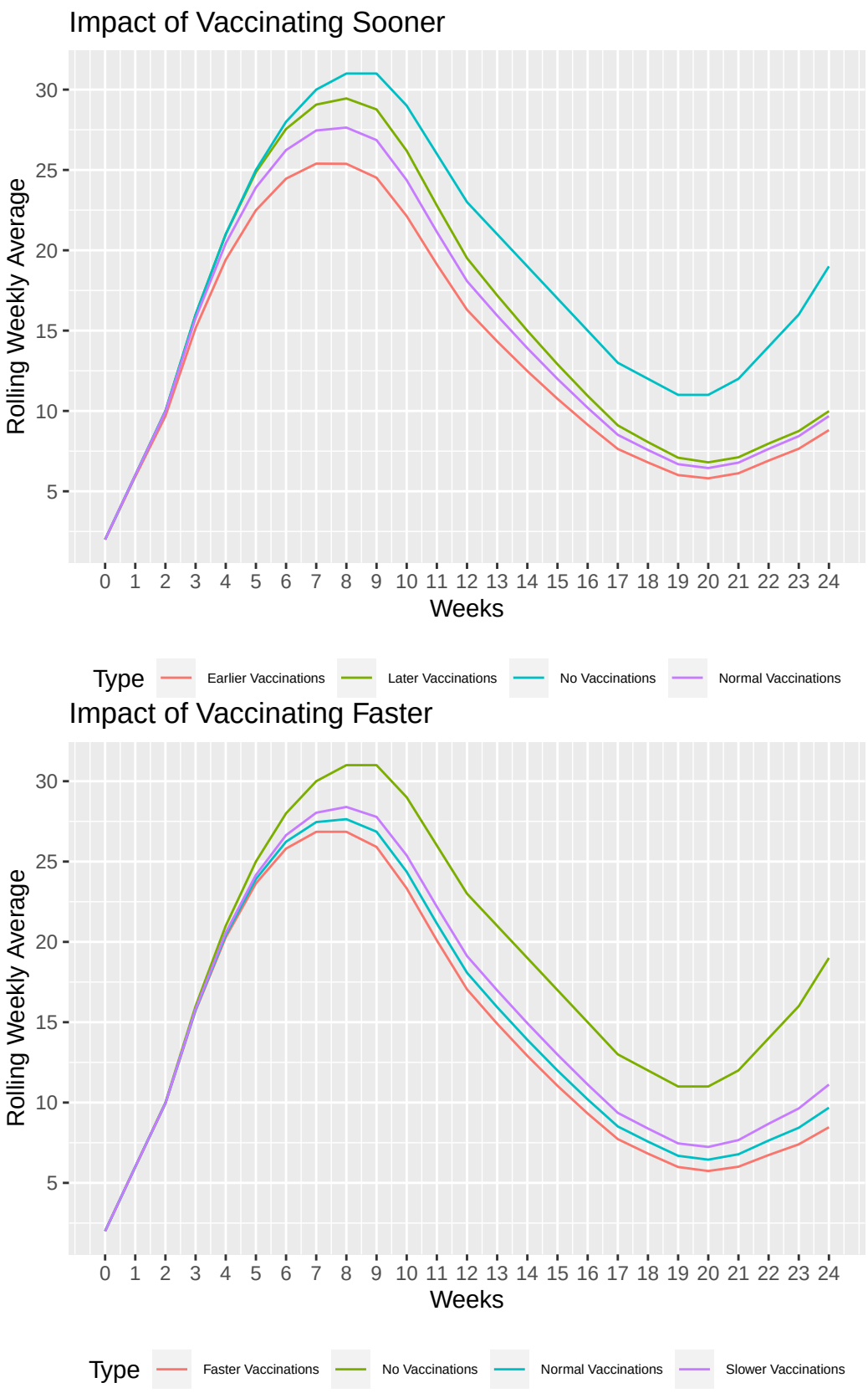
change in vaccination rates, as well as the impact of the change in vaccination rates plus that rubber band effect.




```
sum(VacAndRecoil)/sum(CaseLoad)
```

```
[1] 0.7939372
```

So overall vaccinations do a great job of eventually dealing with an ongoing pandemic and lessening the potential danger of future outbreaks by causing permanent change, however they do barely anything in the short term to actually stop the spread or minimize the peak of the cases. Now let's try a bunch of other scenarios. We are going to test out how what happens if vaccinations get started earlier versus what happens if they get started later, as well as what happens if a county manages to only get to the low vaccination level by the end of the time period compared to what will happen if the county actually manages to get to a relatively high amount of vaccinations (using the same definition of low and high from earlier).



So vaccinating earlier vs later does have an impact in the short term but not much

impact in the long term if the vaccination levels that the values end up at aren't that much different, while vaccination quicker vs slower doesn't have much impact in the short run but in the long run the faster vaccinations do pick up more of an advantage.

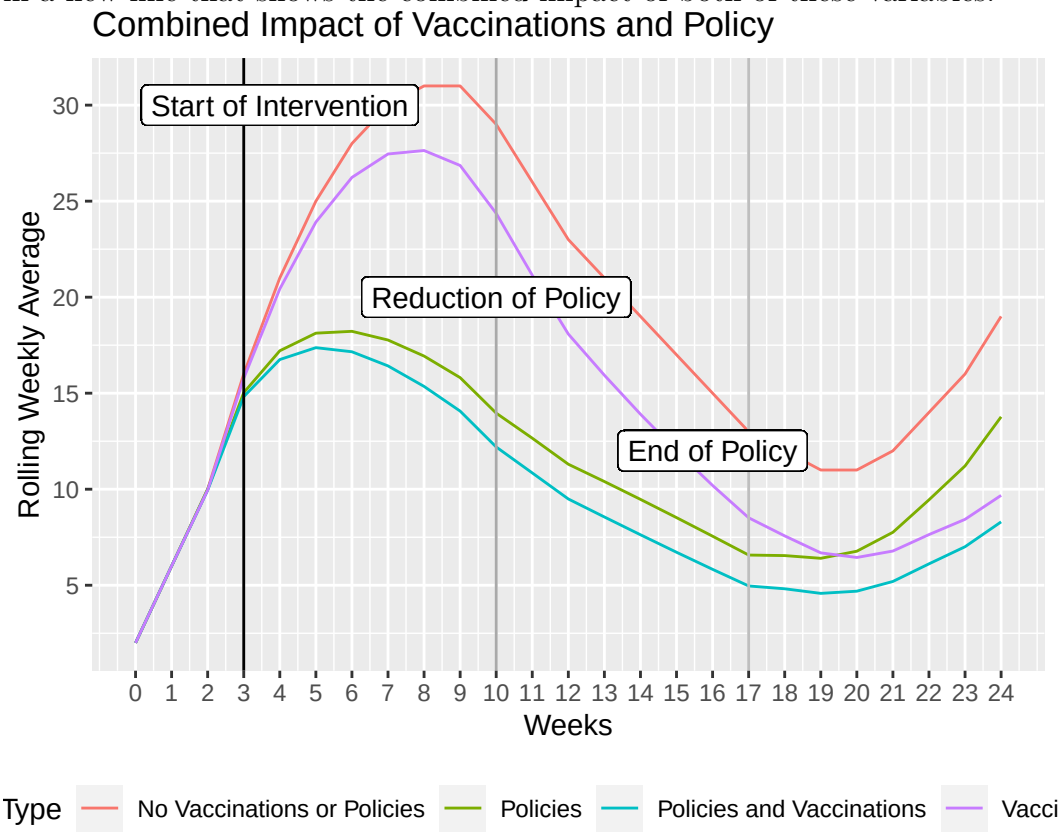
Despite the fact that government policies and vaccinations both accomplish the same general goal of reducing COVID-19 cases by a significant amount, they go about them in two different ways. Government policies are almost immediately and do a great job at reducing the worst of the pandemic to prevent things like hospitals being overrun, but eventually they get lifted and their impact while have little bearing on how things play out down the line. Conversely, it takes a while for the impact of rolling out vaccinations to have a real impact and a vaccination roll-out does little to help with an outbreak happening right now, but it doesn't fade off and instead only gets stronger over time. And this just makes sense if you think about it. With vaccinations, it's just simply the nature of vaccinations that it can't do much right away. Not only does it just take time to manufacture enough of the vaccine to get it to people and there are people who are hesitant about getting vaccinated or just aren't particularly motivated to do so, but the vaccine has to get actually developed, go through several rounds of testing, then slowly approved for different age groups. We're very lucky we got a vaccine ready as quickly as we did, and it still took us until December 2020. In comparison, a governor can pretty easily implement restrictions that will have significant results in less than a week. But on the flip side, vaccination's a lot better long term game plan for dealing with COVID-19. Whenever COVID-19 restrictions get lifted, and they eventually have to get lifted, they lose their affect on the model and so their impact fades away over the course of time thanks to the rubber band effect of the Case Growth variable. However, the number of people who are vaccinated can only increase, making it so that the impact of people being vaccinated only keeps rising and isn't ever going to be undone. While there might be something like a new variant that dramatically worsens our situation, that probably wouldn't really change the impact of vaccinations much since it's likely that any variant that causes problem would have been way more contagious and deadly in a situation where there was no vaccine available.

There's this trade-off to them that makes them fundamentally different from each so that you can't just say "Vaccinations are better at preventing the spread of COVID-19 than policy decisions" or "Government action is a more effective way of dealing with the pandemic than vaccinations." While getting everyone vaccinated is the best way to defeat the pandemic, that takes time and we need ways to deal with outbreaks until enough people are vaccinated that we can return to somewhat normal

life.

5.3.4 Combined

Now let's look at all three of these variables together. Same scenario as we've always had and we're actually going to include the trend lines from the previous sections (only the medium one for the government policy section), but now we're going to add in a new line that shows the combined impact of both of these variables.



As you can see from this graph, government policies and vaccinations are great complements to each other. The combined version prevents that early peak from ever getting too high and it builds up enough momentum with vaccinations so that when the policy is lifted, cases remain low. Let's also look at some numbers from this data.

```
1- sum(MediumRestriction)/sum(CaseLoad)
```

```
[1] 0.3898592
```

```
1- sum(VacAndRecoil)/sum(CaseLoad)
```

```
[1] 0.2060628
```

```
1- sum(VacAndPolicy)/sum(CaseLoad)
```

```
[1] 0.482909
```

```
1- MediumRestriction[25]/CaseLoad[25]
```

```
[1] 0.275276
```

```
1- VacAndRecoil[25]/CaseLoad[25]
```

```
[1] 0.4905048
```

```
1- VacAndPolicy[25]/CaseLoad[25]
```

```
[1] 0.5631444
```

These values do a great job at highlighting what I previously discussed. First, government restrictions do the bulk of the heavy lifting when it comes to lowering the case load in the short term and flattening the curve. Second, vaccinations are the primary key to lowering cases in the long term and preventing them from creeping back up. And third, implementing both together gives us the best of both worlds and is always better than just one or the other.

5.4 Conclusion

The main take away from this model is fairly simple. Government policy and vaccinations are both very important to fighting against COVID-19 but in two very different ways. The fact that government policy decisions aren't permanent, have a cost to keeping in place and have an impact that fades over time means that it's more of a stop gap rather than a permanent solution. It's main function is to prevent worst case scenarios and buy time for vaccines, which are the real solution to dealing with the pandemic. While the closure related policies we put in place do eventually get lifted, people aren't going to become unvaccinated. While they may need a booster shot eventually, even without it they'll still have a significant resistance to COVID-19. This relationship not only makes logical sense, but it's also something many experts have told us some variation of during the pandemic. Now the focus is on using government

policies to mitigate the pandemic until we reach a certain vaccination threshold, with several states setting vaccination targets that and pledge to lift/relax the restrictions in place once that target was hit, but earlier in the pandemic a lot of the discussion was about using government response to buy time for us to build up resources that weren't quite as effective as vaccinations but were still effective once they managed to get actually put in place, such increasing our testing capacity. And I believe this is what we should focus on in the unfortunate scenario that we do go through another pandemic like this. Policies like school closures and stay at home orders are not cures for the pandemic. They are band-aid solutions meant as a way to quickly respond to a situation that has gotten out of hand and minimize damage until we can get more long term solutions online. The government shouldn't just spend a lockdown twiddling it's thumbs and waiting for things to get better, it needs to use that valuable time to ramp up other longer term solutions. This is especially the case when dealing with government policies that have real long term costs. I've generally treated government policies as one homogeneous unit while exploring this data, but that's really not the case. Some policies like requiring masks in particular public places like airports can basically be left in place indefinitely with no significant cost, but policies like limits on public gatherings and a light curfew can build serious resentment towards those implementing these restrictions if left in place for too long, and policies like school closures and workplace closures build resentment even quicker and can have major negative consequences for many people even if only in place for just a little while. There's also my testing data variable, which honestly might be closer to functioning like vaccinations than other government policies I looked at due to the fact that once a county or state has reached a certain testing capacity, it's rare that they dip below that threshold again. Like how vaccination levels never go down because people don't become unvaccinated. And I barely scratched the service on how these government policies change as more people get vaccinated, I really only explored the effect of vaccinations on government policy as a whole. It feels like there is a bottomless rabbit hole of compelling follow-up questions to this model, but I think it's time to return to main point of the results I have rather than speculating endlessly about tangents.

In summary, vaccinations and government interventions are the excellent weapons we have against fighting the COVID-19 pandemic and have opposing strengths and weaknesses that make them truly shine when used in conjunction with each other. A competent pandemic response plan needs to pay attention to both of these tools and use them appropriately in order to complement each other. This is important not only so we can prepare for future pandemics, but also to judge those that were

in power during this pandemic. Many mistakes were made over this pandemic, and it's important to recognize who was doing and saying things that is backed up by the data I have here, and who made errors in judgement that cost people their lives.

But let's end this on a such a dour note. For whatever it's worth, I am genuinely hopeful that lessons have been learned from this pandemic will stick for a long time. Going into this pandemic it feels like we had been lulled into a false sense of security regarding the possibility of a pandemic on this scale. Diseases like the 2009 H1N1 virus or the Ebola outbreak came and went, and while they were certainly tragic, they didn't bring the entire globe to a halt the way COVID-19 did. We stopped worrying about the possibility of something like this happening and let many things slip that we really shouldn't have. But after these past two years, no one is going to want to go through a repeat of this again and I expect that should something like this happen in again, we will be far more ready.

References

Age Data. (2021). U.S. Census Bureau. Retrieved from https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html

Bureau, C. F. P. (2021). Housing insecurity and the COVID-19 pandemic. Retrieved from https://files.consumerfinance.gov/f/documents/cfpb_Housing_insecurity_and_the_COVID-19_pandemic.pdf

Civic Impact, J. H. C. for. (2021). U.s. Testing data. Retrieved from https://github.com/govex/COVID-19/tree/master/data_tables/testing_data

COVID-19 Data Repository. (2020). U.S. Department of Agriculture Economic Research Service. Retrieved from https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx

COVID-19 Data Repository. (2021). Center for Systems Science; Engineering (CSSE) at Johns Hopkins University. Retrieved from https://github.com/CSSEGISandData/COVID-19

COVID-19 Reported Patient Impact and Hospital Capacity by State Timeseries. (2021). HealthData.gov. Retrieved from https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh

Cutler & Summers. (2020). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7604733/

Disease Control, C. for. (2021). Monitoring variant proportions. Retrieved from https://covid.cdc.gov/covid-data-tracker/#variant-proportions

Google COVID-19 Community Mobility Reports. (2021). Google LLC. Retrieved from https://www.google.com/covid19/mobility/

Labor Statistics, U. S. B. of. (2021). Economic news release, state employment and unemployment (monthly). Retrieved from https://www.bls.gov/web/laus.supp.toc.html

Lewis, D. (2021). Superspreading drives the COVID pandemic — and could help to tame it. Retrieved from https://www.nature.com/articles/d41586-021-00460-x

Pugh, E. (2021). My Github Page. Retrieved from https://github.com/EvanPugh?tab=projects

Reporting COVID-19 Vaccination Demographic Data. (2021). Center for Disease Control. Retrieved from https://www.cdc.gov/coronavirus/2019-ncov/vaccines/distributing/demographics-vaccination-data.html

Tracker, O. C. G. R. (2021). Blavatnik school of government, university of oxford. Retrieved from https://github.com/OxCGRT/covid-policy-tracker

US Census Bureau COVID-19 Site. (2021). Average household size and population density - county. Retrieved from https://covid19.census.gov/datasets/average-household-size-and-population-density-county/explore?showTable=true

World Health Organization. (2021). WHO coronavirus (COVID-19) dashboard. Retrieved from https://covid19.who.int/