Comparing The COVID-19 Response in New Hampshire and the United States

Louis Keith

11/15/2020

Introduction

The purpose of this project is to look at the Coronavirus pandemic in the state of New Hampshire and compare it to the country as a whole. There are numerous questions that could be answered by looking at the data, and we will explore each of them one by one.

Importing Libraries

The very first thing that needs to be done is importing libraries that will be useful to this analysis. Readxl will allow importing from excel sheets and dplyr will provide some functions for tidy data analysis. Janitor will also provide similar functionality. ggplot2 will be very helpful for creating plots of all types.

```r
library(readxl)
library(dplyr)
library(janitor)
library(ggplot2)
```

Importing the Data

Next, we will import the data. Both data sets come from the COVID tracking project, one has detailed data for just New Hampshire, the other for the entire country.

```r
new_hampshire_history <- read_excel("C:/Users/louis/Documents/Clarkson/Fall S
emester 2020/STAT 383 Probability and Statistics/Final Project/new-hampshire-
history.xlsx")
national_history <- read_excel("C:/Users/louis/Documents/Clarkson/Fall Semest
er 2020/STAT 383 Probability and Statistics/Final Project/national-history.xl
sx")
```

One thing that was forgotten in the presenting data part of the project was the total populations of both NH and the country as a whole. Ideally there would be data for each day, but since we don’t have that, current populations will be stored as constants. This will be for the purpose of normalizing for population.

```r
nh_pop = 1359711
us_pop = 331002651
```

Both populations were pulled from google.

Preliminary Look

Open both data frames and take a look at the data.

```r
View(new_hampshire_history)
View(national_history)
```

Both data sets are formatted similarly. They each contain one row for each day since the beginning of the data collection began. The data for New Hampshire begins on March 4th while the national data begins on January 22nd. A logical first step would be to truncate the data in the national data set up to March 4th so there is an apples to apples comparison.

Cleaning

There are also a myriad of columns populated by NA values that we will get rid of as they are of no use to our analysis. The data quality grade also isn’t useful for us, and for the New Hampshire data it is superfluous to state it came from NH. Hospitalized and hospitalized_cumulative have identical values, so one of them can go too.

```
new_hampshire_history = new_hampshire_history %>% clean_names()
nhf = new_hampshire_history %>% select(-c(data_quality_grade, state, death_co
nfirmed, death_probable, in_icu_currently, negative_tests_antibody, negative_
tests_people_antibody, negative_tests_viral, on_ventilator_cumulative, on_ven
tilator_currently, positive_tests_antibody, positive_tests_antigen, positive_
tests_people_antibody, positive_tests_people_antigen, positive_tests_viral, t
otal_test_encounters_viral, total_tests_people_antigen, total_tests_antigen,
total_test_encounters_viral_increase, positive_score, hospitalized_cumulative
))
nhf$date <- as.Date(nhf$date)
```

Now that the NH data only contains numeric columns and a properly stored date column, we can do the same for the national data. This data comes much cleaner, so the only thing that needs to be done is to convert the date and clean the names. We do want to only keep the data after March 3rd though.

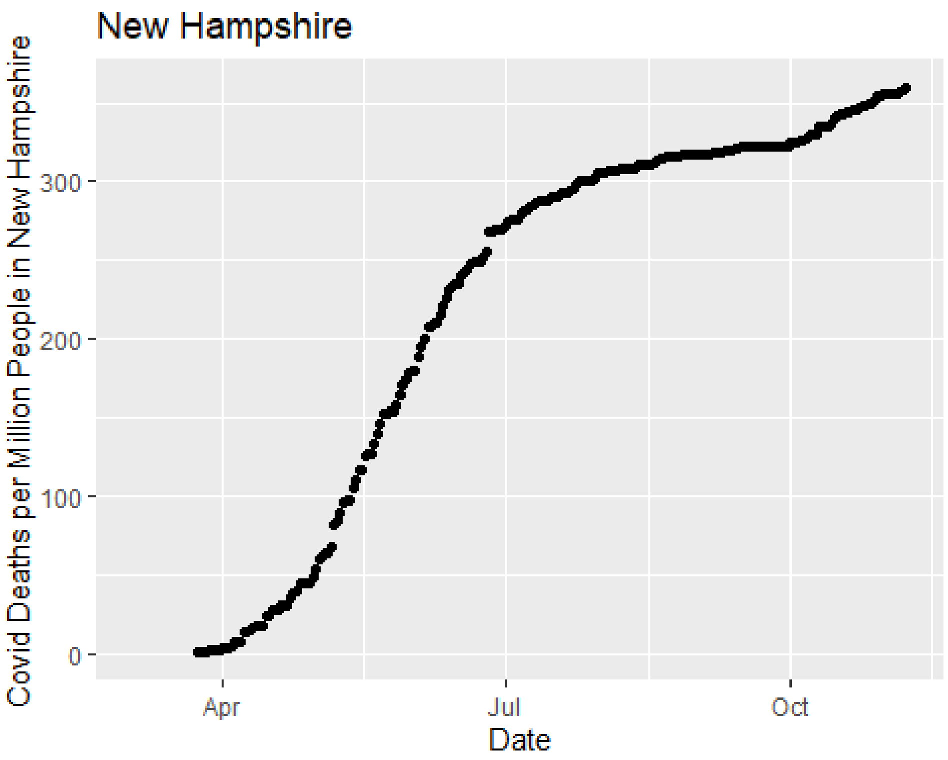
```
nf = national_history %>% clean_names()
nf$date <- as.Date(nf$date)
nf = nf %>% filter(date >= as.Date('2020-03-04'))
```

At this stage, both data sets should only contain usable dates as well as numeric columns.

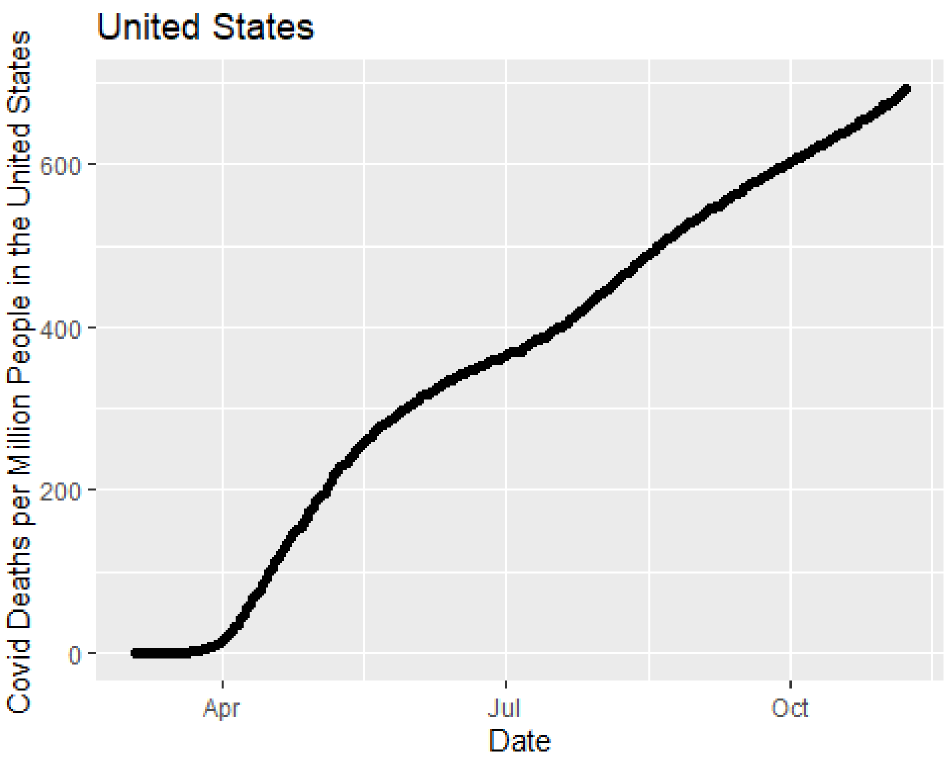
Preliminary Visualizations

Now that the data cleaning step is done, some exploratory visualizations can be done to look at the relationships between data. Some of the most interesting graphs will be deaths over time normalized per million people.

```
ggplot(data = nhf, aes(x = date, y = (death / nh_pop) * 1000000)) + geom_poin
t() + xlab("Date") + ylab("Covid Deaths per Million People in New Hampshire")
+ ggtitle("New Hampshire")
```



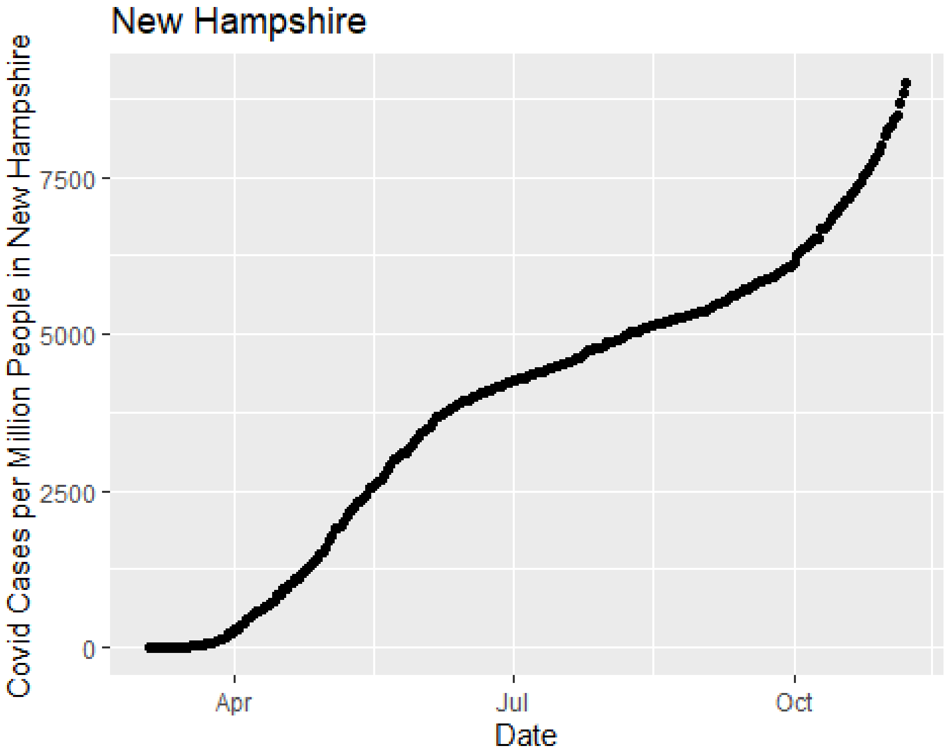
```
ggplot(data = nf, aes(x = date, y = (death / us_pop) * 1000000)) + geom_point
() + xlab("Date") + ylab("Covid Deaths per Million People in the United State
s") + ggtitle("United States")
```



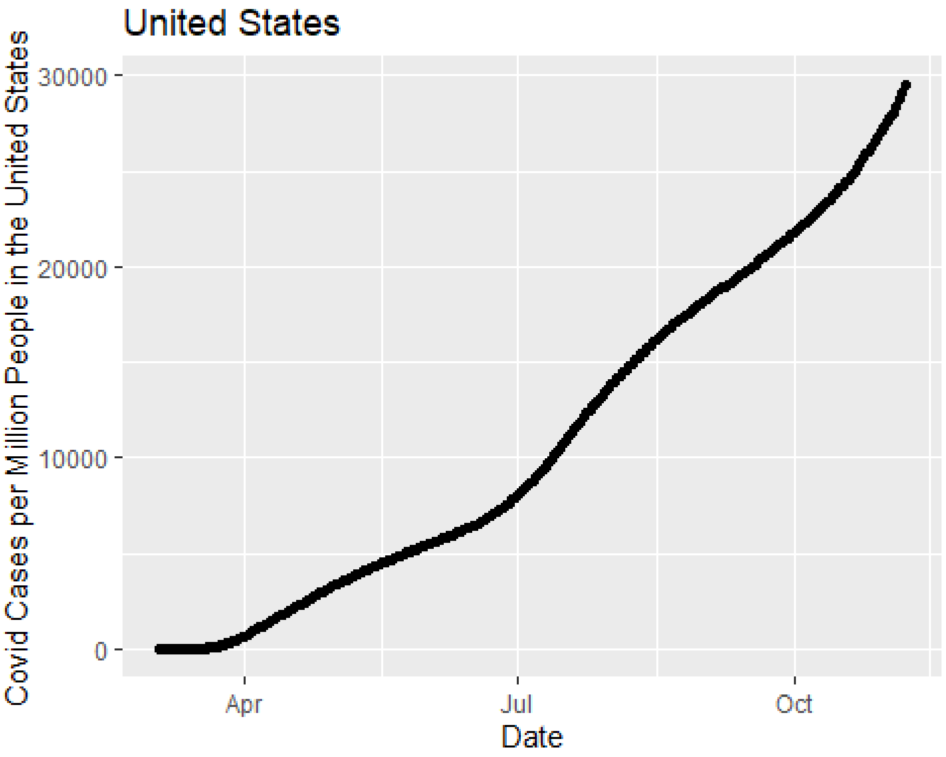
Having done absolutely no statistical analysis at this point, the graph shows that up to November, about 2.5x more people adjusted for population have died in the United States than in New Hampshire. It also appears that the graph for the United States appears roughly linear while the graph for New Hampshire appears to show evidence of “flattening the curve”. Again, there are no hard conclusions to be drawn from this, but it suggests that New Hampshire might be handling the pandemic better than the country as a whole.

We can make the same graphs for the total positive cases so far.

```r
ggplot(data = nhf, aes(x = date, y = (positive / nh_pop) * 1000000)) + geom_p
oint() + xlab("Date") + ylab("Covid Cases per Million People in New Hampshire
") + ggtitle("New Hampshire")
```



```r
ggplot(data = nf, aes(x = date, y = (positive / us_pop) * 1000000)) + geom_po
int() + xlab("Date") + ylab("Covid Cases per Million People in the United Sta
tes") + ggtitle("United States")
```



Similar to the analyses for the previous set of graphs, no hard conclusions can be drawn from either graph. However, recent reports of spiking cases can be pretty clearly noted in both graphs. It appears to be particularly bad in New Hampshire. This will be something that is explored in a later question.

Question 1: Is New Hampshire Handling the Pandemic Better or Worse Than the United States?

Given the preliminary result from above, a one sided test for whether or not the death rate in New Hampshire adjusted for population is lower than the same for the United States seems to make sense.

The hypotheses would be:

$$H_0: p_{deaths-US} - p_{deaths-NH} = 0$$

$$H_a: p_{deaths-US} - p_{deaths-NH} > 0$$

We can determine whether or not to reject the null hypothesis by creating a one sided confidence interval, with the upper bound being 1, the equation for the lower bound is:

$$p_2^{hat} - p_1^{hat} - Z_\alpha \sqrt{\frac{p_2^{hat}(1-p_2^{hat})}{n_2} + \frac{p_1^{hat}(1-p_1^{hat})}{n_1}}$$

Plugging the values we have on November 7th into this equation, we will use a 99% confidence interval:

```
# the value for death at the top of the table (latest value) divided by the t
otal US population
p2 = nf$death[1]/us_pop
# doing the same for the NH values.
p1 = nhf$death[1]/nh_pop
# calculating Z statistic
Z = qnorm(0.99)
lower_bound_proportion_of_deaths = p2 - p1 - Z*sqrt((p2*(1-p2)/us_pop) + (p1*
(1-p1)/nh_pop))
```

The resulting confidence interval is (0.000294944,1) and does not contain 0, this provides sufficient evidence to reject the null hypothesis.

Since we reject the null hypothesis, we conclude that New Hampshire is likely to have a smaller proportion of deaths than the rest of the country. This is evidence that the New Hampshire COVID response has indeed been better than the United States as a whole.

Deaths are not the entire story however. To conclude that New Hampshire is definitely doing better, it is a good idea to perform the same analysis for cases in general.

$$H_0: p_{cases-US} - p_{cases-NH} = 0$$

$$H_a: p_{cases-US} - p_{cases-NH} > 0$$

The same exact confidence interval equation can be used, and the upper bound will still be 1.

```
# the value for death at the top of the table (latest value) divided by the t
otal US population
p2 = nf$positive[1]/us_pop
# doing the same for the NH values.
p1 = nhf$positive[1]/nh_pop
# calculating Z statistic
Z = qnorm(0.99)
lower_bound_proportion_of_cases = p2 - p1 - Z*sqrt((p2*(1-p2)/us_pop) + (p1*(
1-p1)/nh_pop))
```

The resulting confidence interval is (0.020298,1) and does not contain 0, this provides sufficient evidence to reject the null hypothesis.

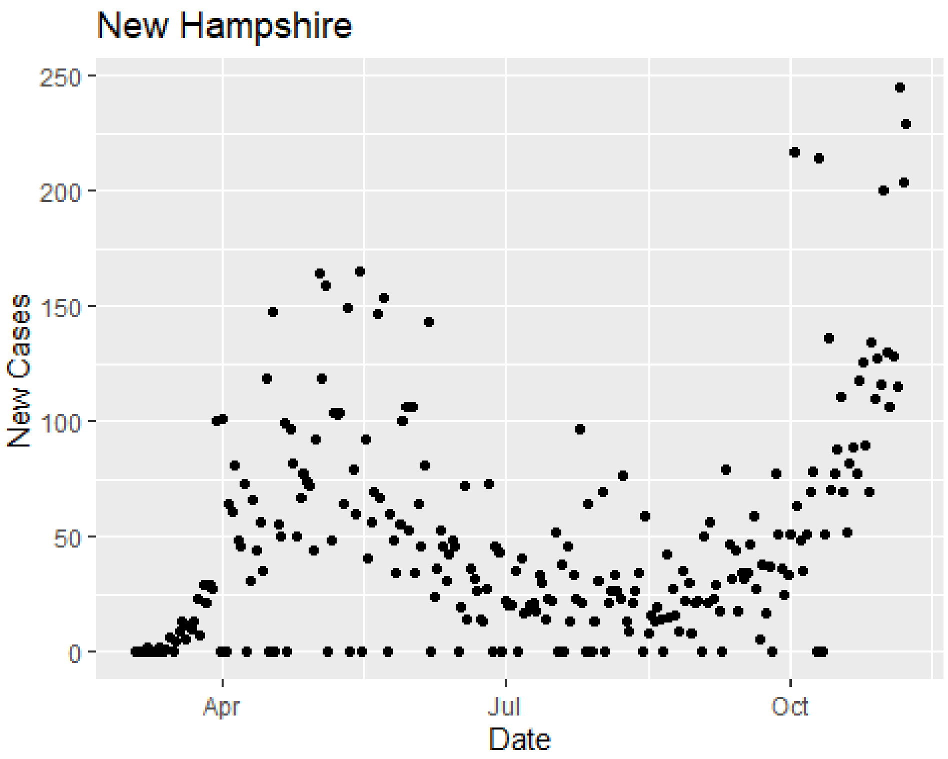
Since we reject the null hypothesis, we conclude that New Hampshire is likely to have a smaller proportion of positive cases than the rest of the country. This too is evidence that the New Hampshire COVID response has been better than the United States as whole.

Together, positive cases and deaths show that New Hampshire has responded better than the average within the United States.

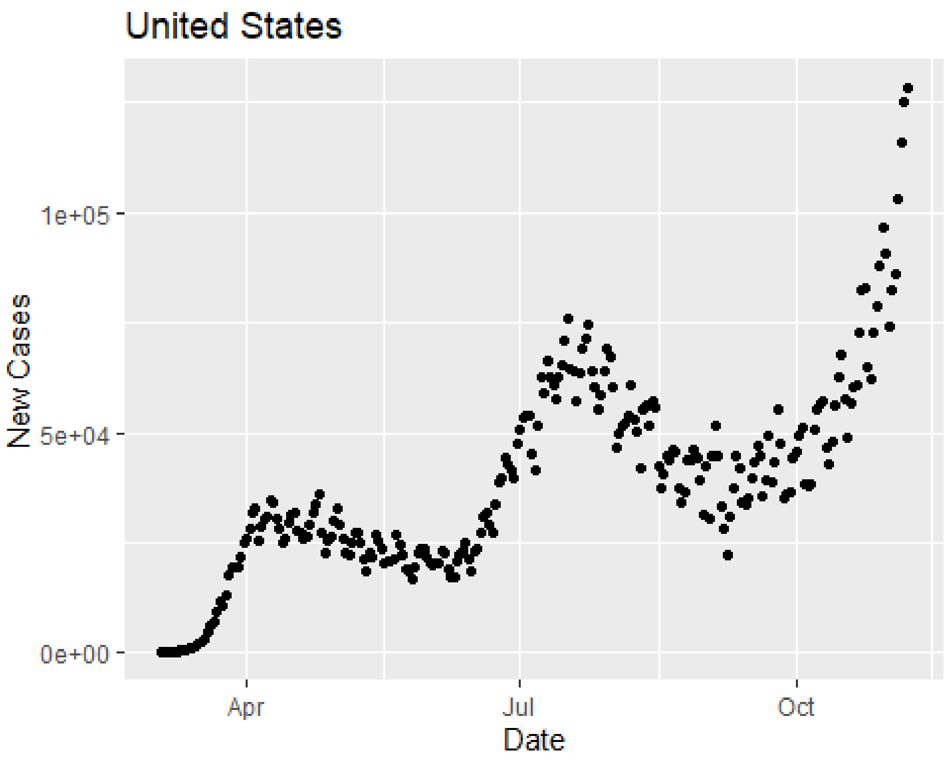
Question 2: Have There Been Several Waves? Which Wave are we in? Are They Getting Smaller/Larger?

In order to answer this question, it is a good idea to look at the positive cases each day.

```
ggplot(data = nhf, aes(x = date, y = positive_increase)) + geom_point() + xla
b("Date") + ylab("New Cases") + ggtitle("New Hampshire")
```



```
ggplot(data = nf, aes(x = date, y = positive_increase)) + geom_point() + xlab
("Date") + ylab("New Cases") + ggtitle("United States")
```



Looking at the national data, it seems very obvious that there are three peaks, and we are in the middle of a third wave. This is consistent with reports of spiking cases in many states. The New Hampshire data is also very interesting. It appears that we are in the middle of a second wave, with the first coinciding mostly with the space between the first and second waves in the country as a whole. There are also quite a few days where there are zero cases recorded. This is probably an error in the data collection. It is important to note that the waves appear to be getting larger in both New Hampshire and the country as a whole.

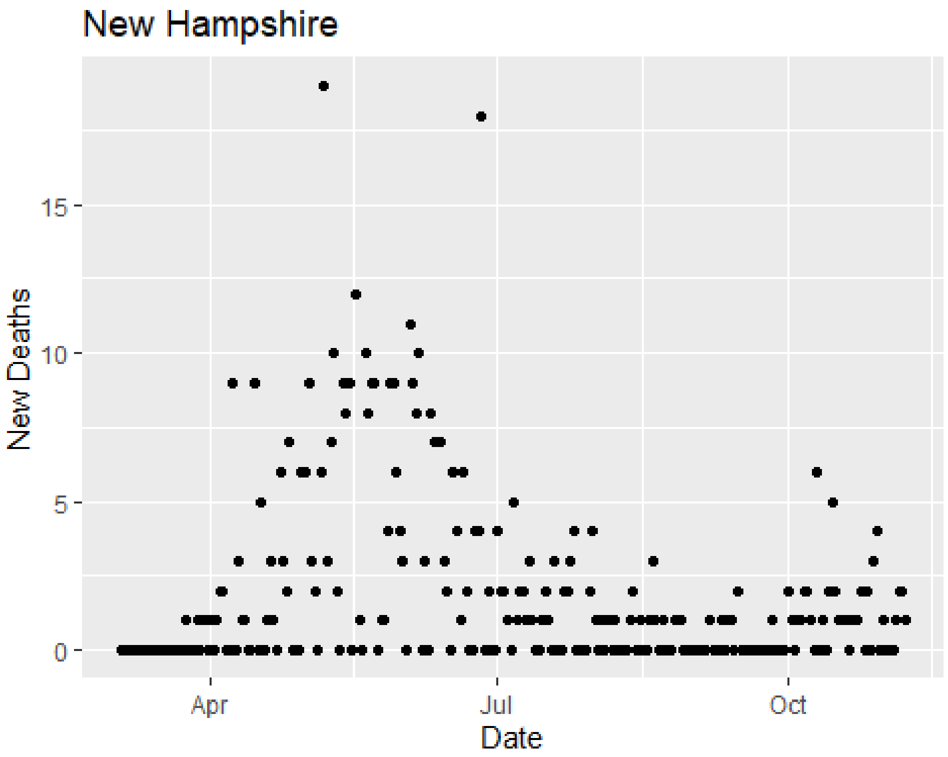
Looking at the New Hampshire graph, it would make sense that the cases that were meant to be recorded on the days with zero new cases were lumped into the next day. This explains why there is a subset of points that look like they are about twice the height of the rest of the curve, and others with zero cases. This is quite unfortunate, but if that assumption is correct, then the residuals will be worse but the overall fit should remain mostly the same.

Overall, a tentative conclusion we can draw from this is that the waves in New Hampshire and the United States do not line up very well. They have not been synced up, and there have been only two in New Hampshire while there were three across the nation as a whole.

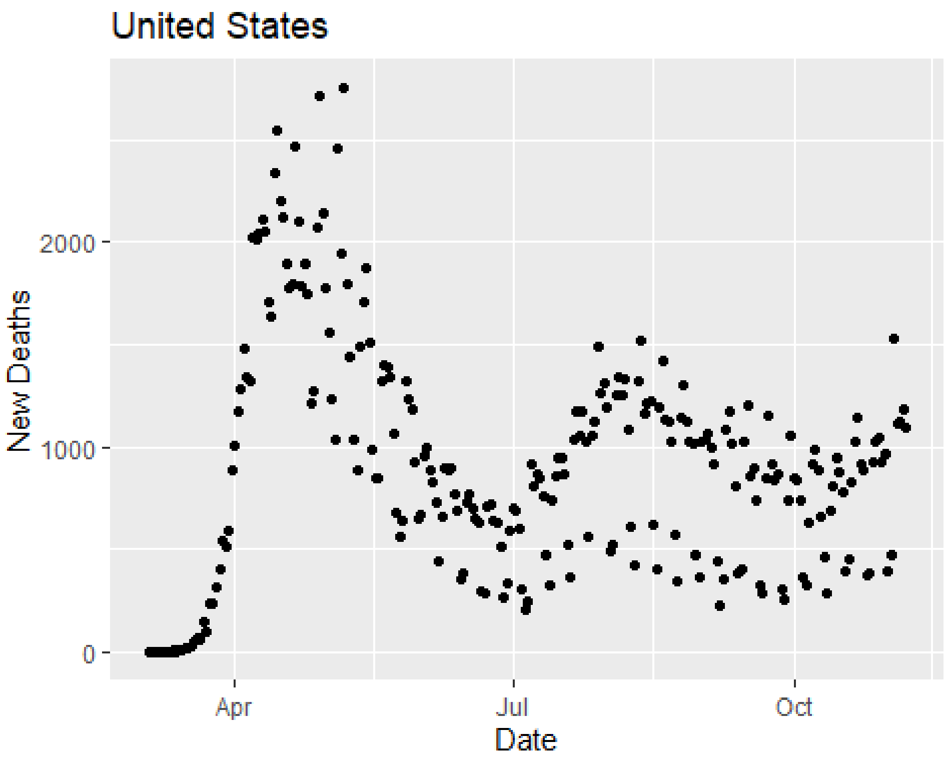
Hard statistical analysis on this problem is not trivial, so for the purposes of this project it will not be performed. This is not one of the three statistical analyses that will be done, but it is very interesting information nonetheless.

Nonetheless, we can repeat this analysis for the deaths.

```
ggplot(data = nhf, aes(x = date, y = death_increase)) + geom_point() + xlab("
Date") + ylab("New Deaths") + ggtitle("New Hampshire")
```



```
ggplot(data = nf, aes(x = date, y = death_increase)) + geom_point() + xlab("D
ate") + ylab("New Deaths") + ggtitle("United States")
```



This appears to roughly support the story that the cases have told. The United States had a spike just before New Hampshire, and then another two waves, but the separation appears less than for the cases. The number of deaths for the second and third waves for the United States appear to show up less well on the deaths. This could be because treatments have improved since the early months of the pandemic, reducing the number of deaths relative to cases. This will be a subject of further analysis.

There also appears to be some odd separation between two distinct bands in the United States data. This could be explained by an error in the data, but the mechanism is unknown.

The New Hampshire data on its own is more difficult to examine since there are so few deaths, but it mirrors the same trends that were observed for the national data. The second wave seems less pronounced in the death data, potentially because of better treatments.

As with the analysis for cases, this on its own does not provide the kind of data to make any conclusions. but it may offer grounds for explanation for things that will be seen later in the analysis.

Question 3: Do Cases in New Hampshire Correlate with Cases in The United States? How about deaths?

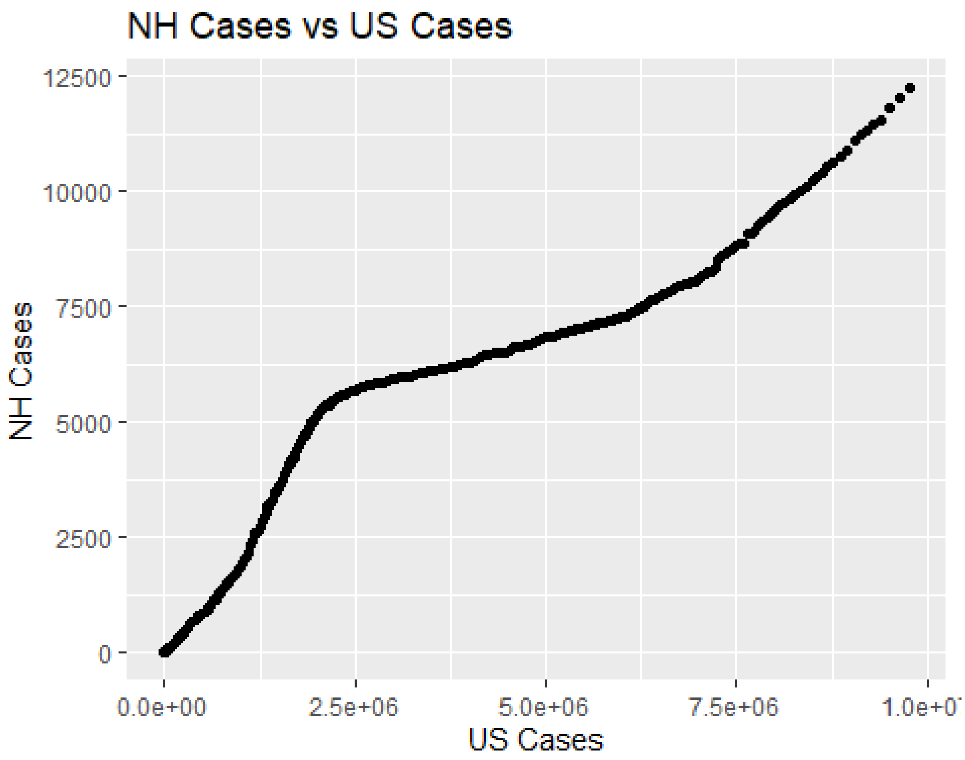
This was not in the initial submission for questions, but this is something that conclusive statistical analysis can be done on, so it will be explored.

First, we need to create a data frame that contains both columns of data that we wish to analyze. This can be done by selecting columns from both of the full data frames available and then joining them by the date column.

```r
nh_cases_and_deaths = nhf %>%
  # grab the columns we want from the New Hampshire data and put them in another data frame
  select(date, positive, death) %>%
  # rename the columns so that they are distinct when joined
  mutate(nh_positive = positive, nh_death = death) %>%
  # get rid of the original names for those columns.
  select(-c(positive, death))

# do the same for the national data
us_cases_and_deaths = nf %>%
  select(date, positive, death) %>%
  mutate(us_positive = positive, us_death = death) %>%
  select(-c(positive, death))

# create combined data frame
cdf = full_join(nh_cases_and_deaths, us_cases_and_deaths, by = 'date')
# there are NAs at the beginning of the NH data set for deaths where there should probably be zeros. In order to plot those points we can replace them
cdf = cdf %>% replace(., is.na(.), 0)
```

Now we can easily create a graph to explore the relationship between cases in the united states and cases in New Hampshire.

```r
ggplot(data = cdf, aes(x = us_positive, y = nh_positive)) + geom_point() + xlab("US Cases") + ylab("NH Cases") + ggtitle("NH Cases vs US Cases")
```



This is a very interesting plot that tries to predict the number of cases in NH based on the number of cases in the US. It appears that there was a period during which New Hampshire had more cases than would be suggested by just predicting off of national data. We can fit a linear model to explore this relationship further.

$$H_0: \beta_1 = 0$$

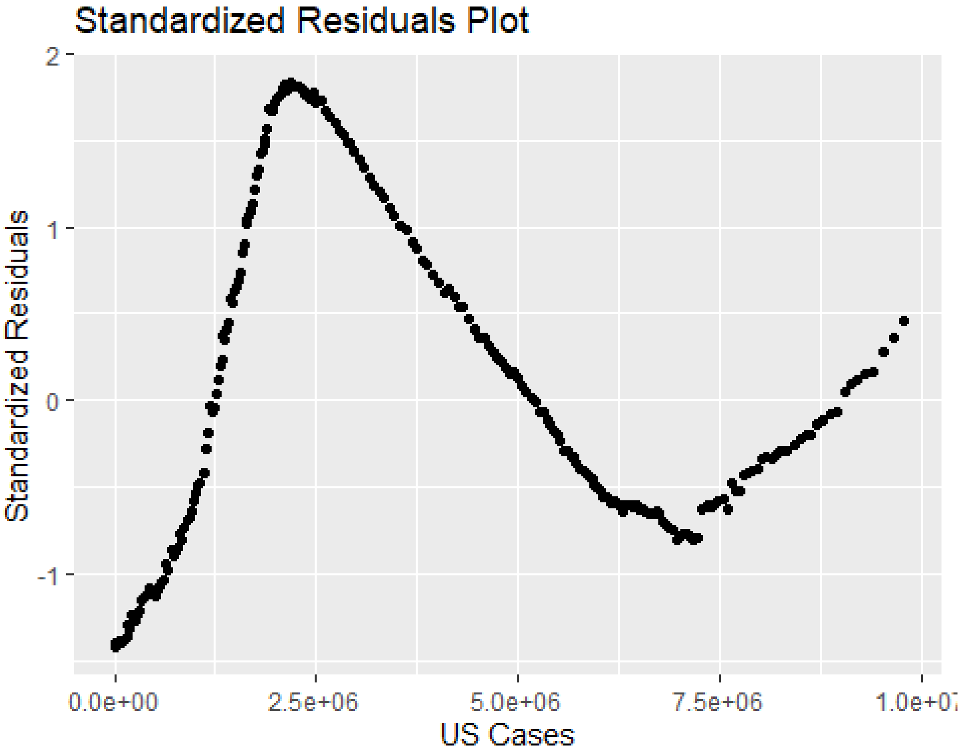
$$H_0: \beta_2 \neq 0$$

```
nh_cases_vs_us_cases = lm(data = cdf, nh_positive ~ us_positive)
summary(nh_cases_vs_us_cases)

##
## Call:
## lm(formula = nh_positive ~ us_positive, data = cdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1355.8  -656.5  -223.0   702.1  1749.7
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.357e+03  9.822e+01   13.82   <2e-16 ***
## us_positive 1.070e-03  2.103e-05   50.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 952 on 247 degrees of freedom
## Multiple R-squared:  0.913,  Adjusted R-squared:  0.9126
## F-statistic:  2591 on 1 and 247 DF,  p-value: < 2.2e-16
```

From this we can certainly conclude that there is a strong statistical relationship between these two variables, as is to be expected. Both p-values for the intercept are about low as R can output, with an R-squared of 0.913. We can safely reject the null hypothesis. The most interesting aspect of this graph is the big hump in the data, we can see what this looks like on the standardized residual plot.

```
ggplot(data = cdf, aes(x = us_positive, y = nh_cases_vs_us_cases$residuals/sq
rt(anova(nh_cases_vs_us_cases)$"Mean Sq"[2]))) + geom_point() + xlab("US Case
s") + ylab("Standardized Residuals") + ggtitle("Standardized Residuals Plot")
```



There is an incredibly obvious pattern here, which shows that there is problem with the linear model. However, not a single data point breaks the 2 mark on the standardized residuals, which shows that the fit itself isn't too bad. There is a problem with the shape of the graph though. Instead of coming up with an alternate relationship, another explanation can be offered.

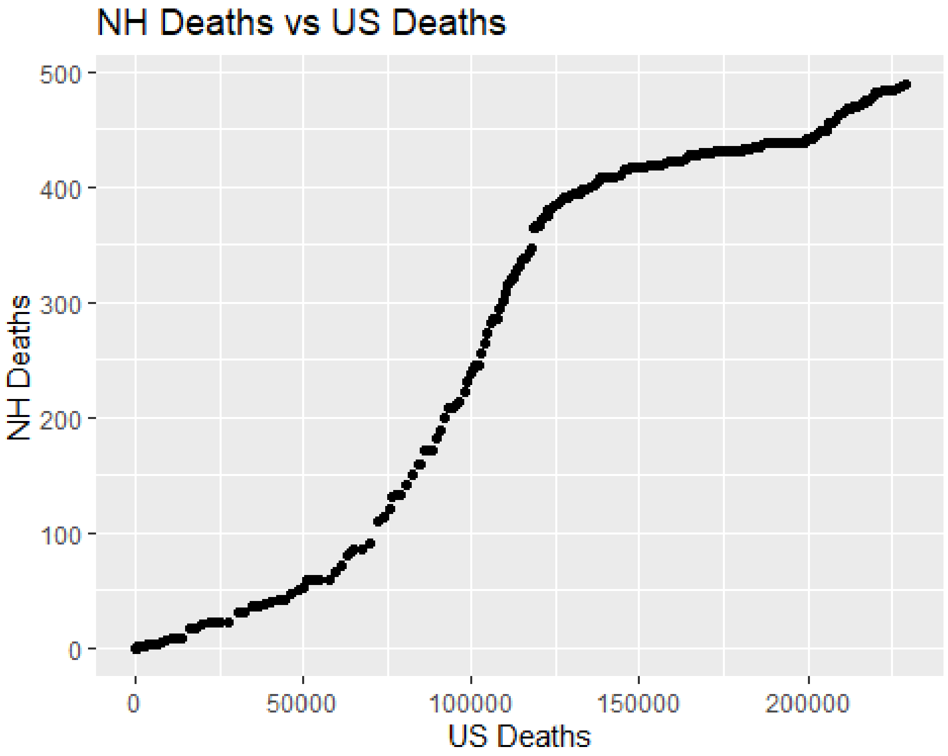
From the question 2 analysis we can see that New Hampshire was in the middle of its first wave while the greater United States was recovering from its first wave. During this period, New daily cases were decreasing in the United States while spiking in New Hampshire. This offers explanation for this hump, because it appears to happen around the same place in the data (May and June). In the residuals, there is a time immediately after the hump that the cases appear to go below the y = 0 line. This can also be explained by the fact that the

United States was entering its second wave while New Hampshire was recovering from its first wave, making the cases in New Hampshire lower than what would be predicted by the model. Finally, the line moves back above the y = 0 line at the end of the data set because New Hampshire appears to be spiking right now more severely than the average of the rest of the nation.

While this is not exactly a desirable result in the residuals, it combines extremely well with the Question 2 analysis to offer a very good look at what is happening in New Hampshire and how it related to the United States. It offers statistical backing for the earlier claim that the waves in New Hampshire are not synchronized with the waves across the rest of the nation.

We can repeat this analysis for the deaths.

```r
ggplot(data = cdf, aes(x = us_death, y = nh_death)) + geom_point() + xlab("US Deaths") + ylab("NH Deaths") + ggtitle("NH Deaths vs US Deaths")
```



This is very interesting, as at first glance it does not seem to match up with the cases data. It does however make a great deal of sense once one considers that if someone dies of COVID, it is often after a lengthy weeks-long battle. This explains why the exact same hump appears in the data but delayed by several weeks. Many of those cases that were observed in May and June are resulting in deaths that occur later on in July and August.

We can predict that the residuals will likely tell a similar story, but shifted to the right. First we must make the model.

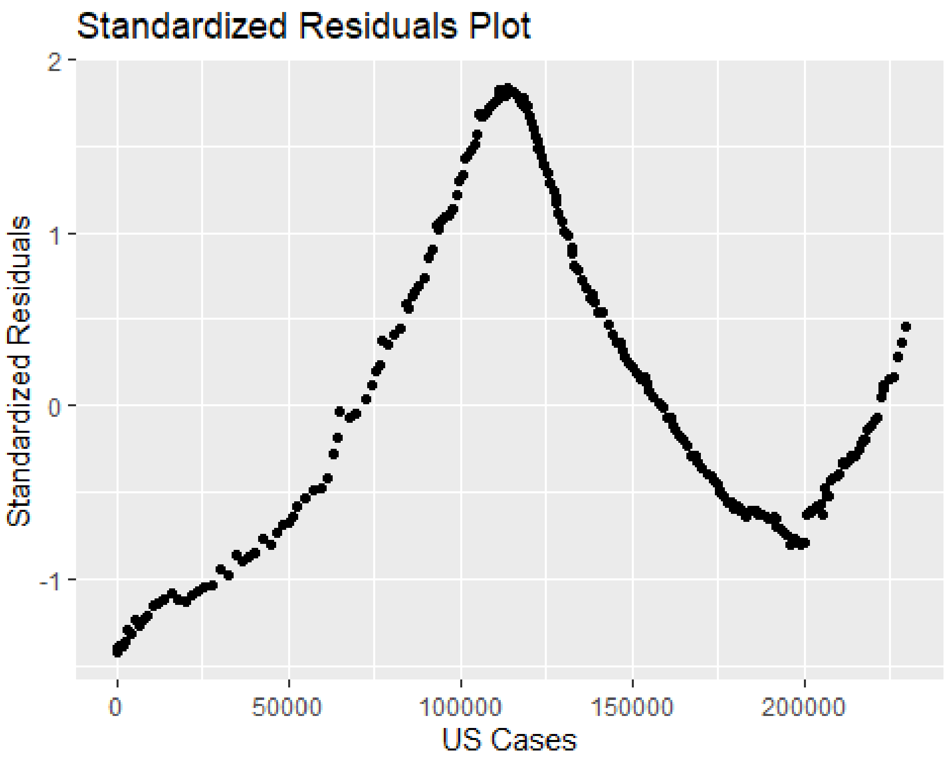
$$H_0: \beta_1 = 0$$

$$H_0: \beta_2 \neq 0$$

```
nh_cases_vs_us_cases = lm(data = cdf, nh_positive ~ us_positive)
summary(nh_cases_vs_us_cases)

##
## Call:
## lm(formula = nh_positive ~ us_positive, data = cdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1355.8  -656.5  -223.0   702.1  1749.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.357e+03  9.822e+01   13.82   <2e-16 ***
## us_positive 1.070e-03  2.103e-05   50.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 952 on 247 degrees of freedom
## Multiple R-squared:  0.913,  Adjusted R-squared:  0.9126
## F-statistic:  2591 on 1 and 247 DF,  p-value: < 2.2e-16
```

Again, the null hypothesis can be rejected and it can be concluded that there is a strong relationship. The R-squared is the exact same at 0.913. The p-values are very low. All of this is expected, what will be more interesting is the residuals.

```
ggplot(data = cdf, aes(x = us_death, y = nh_cases_vs_us_cases$residuals/sqrt(
anova(nh_cases_vs_us_cases)$"Mean Sq"[2]))) + geom_point() + xlab("US Cases")
+ ylab("Standardized Residuals") + ggtitle("Standardized Residuals Plot")
```



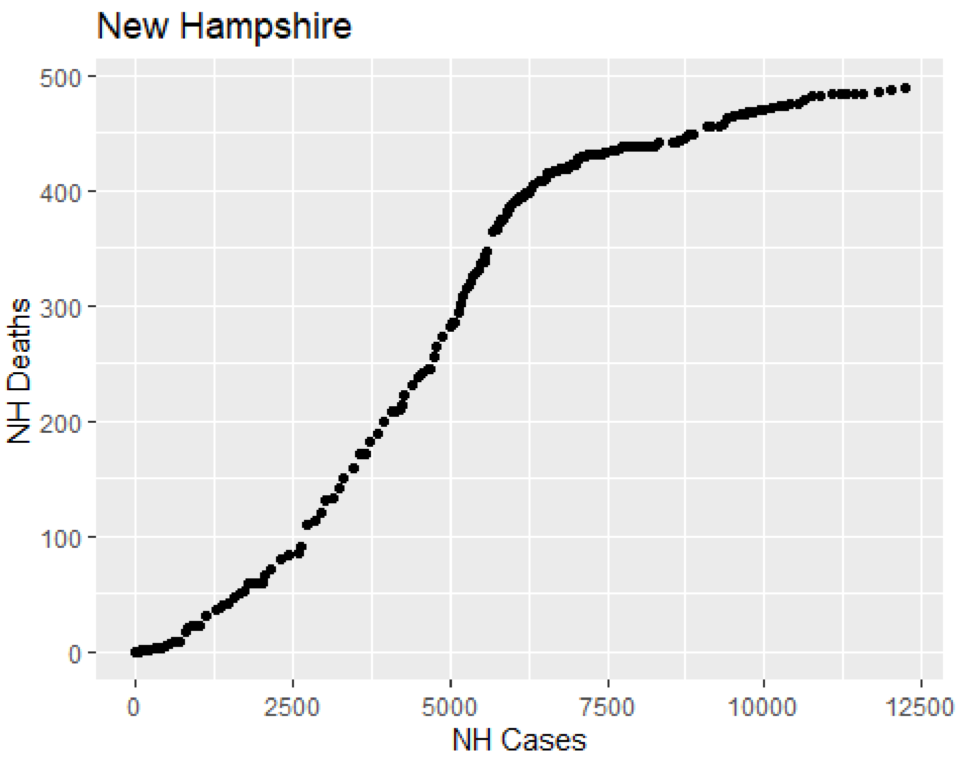
The result is exactly as predicted after seeing the initial graph. This tells the same story as the cases graph but shifted a little bit to the right because the death from any given case is likely to happen several weeks to a month later.

Question 4: Are Treatments Improving Over Time?

This is the further analysis for the observation that was made at the end of the second question. Is the case fatality rate dropping over time? This was also not a question that was initially asked when planning out this project, but one that arose out of looking at the data.

We can begin this analysis like we have the others, coming up with some preliminary graphs. We can re-use the cdf data frame from earlier.

```
ggplot(data = cdf, aes(x = nh_positive, y = nh_death)) + geom_point() + xlab(
"NH Cases") + ylab("NH Deaths") + ggtitle("New Hampshire")
```



This is interesting, it tells a similar story as the graph relating NH deaths and US deaths. A possible explanation for this is that it coincides with the majority of deaths from the first wave. Perhaps New Hampshire hospitals were overwhelmed and unable to give everyone sick the best treatments that were available. The curve appears to flatten afterwards relative to before the hump so perhaps treatments improved as a result. The linear model is the next step.

$$H_0: \beta_1 = 0$$

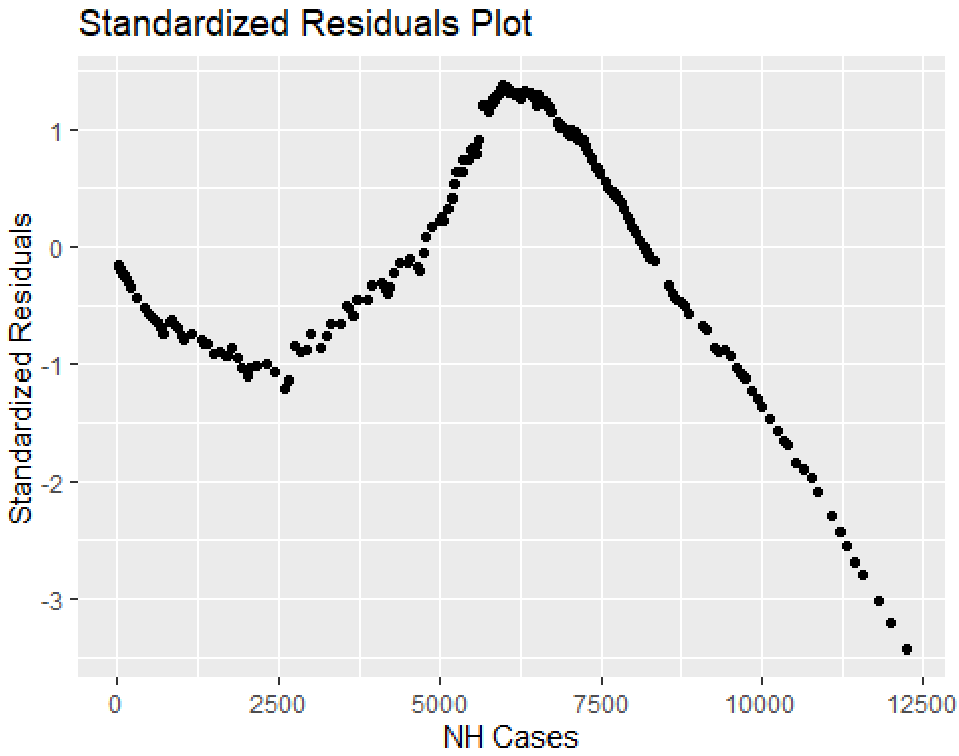
$$H_0: \beta_2 \neq 0$$

```
nh_model = lm(data = cdf, nh_death ~ nh_positive)
summary(nh_model)

## 
## Call:
## lm(formula = nh_death ~ nh_positive, data = cdf)
## 
## Residuals:
##       Min       1Q   Median       3Q      Max 
## -164.732  -32.345   -7.046   44.267   66.506 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 6.9400327  5.8722199   1.182    0.238    
## nh_positive 0.0528382  0.0009472  55.786   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.03 on 247 degrees of freedom
## Multiple R-squared:  0.9265, Adjusted R-squared:  0.9262
## F-statistic:  3112 on 1 and 247 DF,  p-value: < 2.2e-16
```

We can reject the null hypothesis. Interestingly, the intercept is not significant, but the slope is extremely significant and the R-squared is very high. The most interesting part of this model is that the slope that is given is 0.0528, meaning that on average across the whole pandemic, the case fatality rate has been about 5.28%.

We can look at the residuals plot.

```
ggplot(data = cdf, aes(x = nh_positive, y = nh_model$residuals/sqrt(anova(nh_
model)$"Mean Sq"[2]))) + geom_point() + xlab("NH Cases") + ylab("Standardized
Residuals") + ggtitle("Standardized Residuals Plot")
```

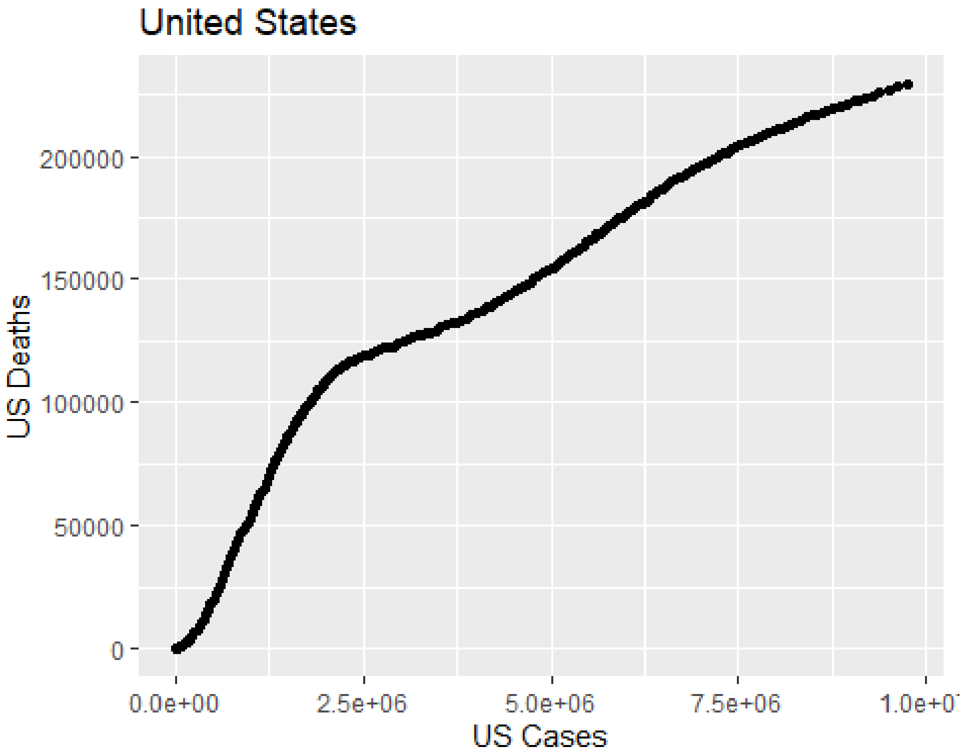


While this is a concerning residuals plot, it bodes well for the conclusions that are being developed thus far. The line (and thus the y = 0 on this plot) is pulled up by all the deaths that occurred in the hump, the hump itself is explained above, and the fall to below -3 can be explained by the aforementioned flattening of the curve because of treatments developed during the first wave.

The slope of the line puts the case fatality rate in New Hampshire at about 5.28% averaged across the pandemic, while the final numbers on November 7 (489/12241) put the cumulative case fatality rate at about 4%. This provides further support that perhaps the case fatality rate is falling over time.

We can and should repeat this analysis for the entire United States.

```
ggplot(data = cdf, aes(x = us_positive, y = us_death)) + geom_point() + xlab(
"US Cases") + ylab("US Deaths") + ggtitle("United States")
```



This graph makes a lot of sense, knowing what we know from previous analysis, we can explain that the initial hump is due to the first wave in the May and June months, and a strain on the health care system as beds fill up. Additionally, the graph is appearing to flatten out as time goes on, supporting the notion that the case fatality rate is falling.

$$H_0 : \beta_1 = 0$$

$$H_0 : \beta_2 \neq 0$$

Again, we make the linear model.

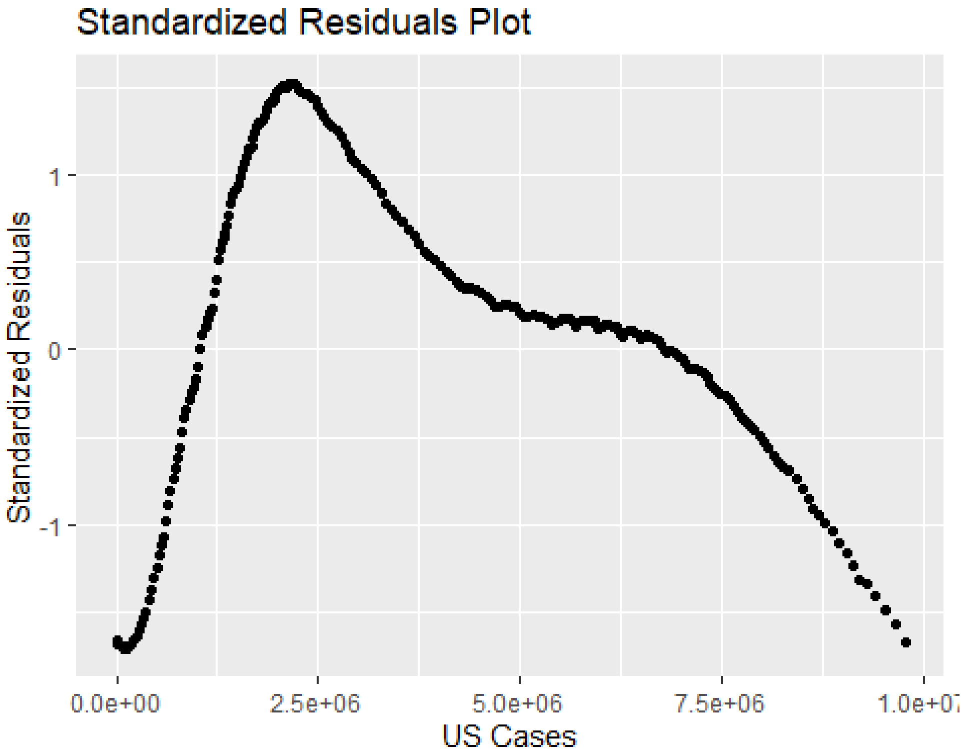
```
us_model = lm(data = cdf, us_death ~ us_positive)
summary(us_model)

##
## Call:
## lm(formula = us_death ~ us_positive, data = cdf)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33578 -13312   2642  15224  29911
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.272e+04  2.024e+03   16.17   <2e-16 ***
## us_positive 2.350e-02  4.334e-04   54.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19620 on 247 degrees of freedom
## Multiple R-squared:  0.9225, Adjusted R-squared:  0.9222
## F-statistic:  2940 on 1 and 247 DF,  p-value: < 2.2e-16
```

As expected, the null hypothesis can be rejected, the p-values for both the intercept and slope are incredibly small. The R-squared is again very high. The most interesting take-away from this is that the case fatality rate (slope) is 2.35%, half that of New Hampshire.

We can look at the residuals plot.

```
ggplot(data = cdf, aes(x = us_positive, y = us_model$residuals/sqrt(anova(us_
model)$"Mean Sq"[2]))) + geom_point() + xlab("US Cases") + ylab("Standardized
Residuals") + ggtitle("Standardized Residuals Plot")
```



This again repeats the same story that was told in New Hampshire but to a less extreme degree. None of the residuals ever get below -2. Interestingly, the line begins well below y = 0 and flies upwards, this is likely as hospitals begin to run out of resources in the early phases of the pandemic. The slow fall later on is likely because treatments improve over time.

The slope of the line puts the case fatality rate in the United States at about 2.35% averaged across the pandemic, while the final numbers on November 7 (229238/9761481) put the cumulative case fatality rate at about 2.34%. This does not provide any evidence that the case fatality rate has fallen over time. This is very interesting because the cumulative case fatality rate can be visualized as a line from the origin through the very last point, practically the entire data set lies above this line. It is expected that the slope of the model is a fair bit higher, but this does not happen here.

These numbers seem close enough that there doesn't seem to be any point in testing if the difference is statistically significant.

Question 5: Is the Case Fatality Rate Higher in New Hampshire than Average?

This analysis will be very similar to Question 1, but it makes sense to compare New Hampshire to the rest of the country since the case fatality rates came out so different. We will use the cumulative rates for both.

$$H_0: p_{fatality-rate-NH} - p_{fatality-rate-US} = 0$$

$$H_a: p_{fatality-rate-NH} - p_{fatality-rate-US} > 0$$

The equation for calculating the confidence is the same as outlined far above. Yet again this is a one sided confidence interval with the upper bound being 1.

$$p_2^{hat} - p_1^{hat} - Z_\alpha \sqrt{\frac{p_2^{hat}(1 - p_2^{hat})}{n_2} + \frac{p_1^{hat}(1 - p_1^{hat})}{n_1}}$$

Plugging the values we have on November 7th into this equation, we will again use a 99% confidence interval:

```r
# the value for death at the top of the table for NH divided by the number of
positive results at the top of the table for NH.
n2 = cdf$nh_positive[1]
p2 = cdf$nh_death[1]/n2
# doing the same for the NH values.
n1 = cdf$us_positive[1]
p1 = cdf$us_death[1]/n1
# calculating Z statistic
Z = qnorm(0.99)
lower_bound_nh_vs_us = p2 - p1 - Z*sqrt((p2*(1-p2)/n2) + (p1*(1-p1)/n1))
```

The lower bound is 0.01234, meaning the interval is (0.01234,1), which does not contain 0. This means that New Hampshire is in fact likely to have a higher case fatality rate than the rest of the country. This is a very interesting result when combined with the conclusion in Question 1, where it was determined that New Hampshire had a better response than the rest of the country. The state had a lower proportion of cases and deaths, but a greater proportion of those cases result in death. This means that it isn't so cut and dry that New

Hampshire responded better than the rest of the country. This might be explained by things like demographic differences (perhaps an above-average age of residents), but more data is needed to determine that conclusively.

Conclusion

Overall, this has been an extremely insightful analysis. This has not exactly fit the formatting outlined in the rubric, but each question is somewhat structured like it's own miniature paper with its own methods, analyses, and conclusions (even if not labelled as so). Questions 1 and 5 showed that New Hampshire has overall responded better than the rest of the country, but has had a higher case fatality rate. This could be for demographic reasons, or it could be for other factors that are outside the scope of this paper to explore. Question 2 did not include any hard statistical analysis but did offer great insight in other parts by exploring the wave structure of pandemics. This was instrumental for explaining what was happening and why residuals plots looked the way they did. The idea for Question 3 came directly from looking at the results of the question prior, and confirmed what was expected: That cases within New Hampshire very closely correlated with cases in the rest of the country. The idea for Question 4 also came directly from the results of Question 2, and explored the idea that the case fatality rate is dropping. While no significant evidence was found of this happening, the trend seemed clear from the graphs. Perhaps another method of analysis would have revealed this to be true, or perhaps the effect is so slight as to not be significant.

This has given an excellent understanding of the way that the pandemic has developed over time in both the United States as a whole and in New Hampshire. It was a very interesting subject to look into.