DATA512 Final Project

Harry McNinson

December 12, 2022

```
library(nlme)
```

OBJECTIVE

The Snohomish County of Washington State implemented masking policies during the outbreak of the Covid pandemic. The aim was to reduce the progression of confirmed COVID-19 cases within the County. The purpose of this notebook was to evaluate the impact that the masking policy had on the progression of the COVID-19 virus and the effect that this policy had on unemployment rate.

DATA PREPARATION AND LOADING

In this step, I used the RAW_us_confirmed_cases.csv file from the Kaggle repository of John Hopkins University COVID-19 data, the CDC dataset of masking mandates by county, and the New York Times mask compliance survey data. Links to these datasets will be provided in the ReadME file of these projects.

I changed the granularity of the covid confirmed cases for Snohomish County in Washington State from day to month, which is the first column of the dataset below. The second column indicates the number of infections in each of the months of interest. The mask policy (level column) started in July 2020. The data indicates a change over time after the policy was introduced. It will be interesting to find out the level shift, whether it was a trend change, and what we can expect in the future. To answer all these, I will use interrupted time series analysis on this dataset to perform this experiment.

```
# You can also do this in RStudio using the "Import Dataset" button
data <- read.csv("../data/snohomish_monthly_infections.csv",header=T)
print(data)
```

```
##         month infections time level trend trendsq
## 1   2/29/2020          1    1     0     0       0
## 2   3/31/2020       1229    2     0     0       0
## 3   4/30/2020       2433    3     0     0       0
## 4   5/31/2020       2967    4     0     0       0
## 5   6/30/2020       3546    5     0     0       0
## 6   7/31/2020       5095    6     1     1       1
## 7   8/31/2020       6309    7     1     2       4
## 8   9/30/2020       7144    8     1     3       9
## 9  10/31/2020       9191    9     1     4      16
## 10 11/30/2020      15057   10     1     5      25
## 11 12/31/2020      22393   11     1     6      36
## 12  1/31/2021      27744   12     1     7      49
## 13  2/28/2021      30208   13     1     8      64
```

```
## 14  3/31/2021      32093   14      1      9      81
## 15  4/30/2021      35943   15      1     10     100
## 16  5/31/2021      38873   16      1     11     121
## 17  6/30/2021      40232   17      1     12     144
## 18  7/31/2021      42759   18      1     13     169
## 19  8/31/2021      51116   19      1     14     196
## 20  9/30/2021      59052   20      1     15     225
## 21 10/31/2021      65719   21      1     16     256
```

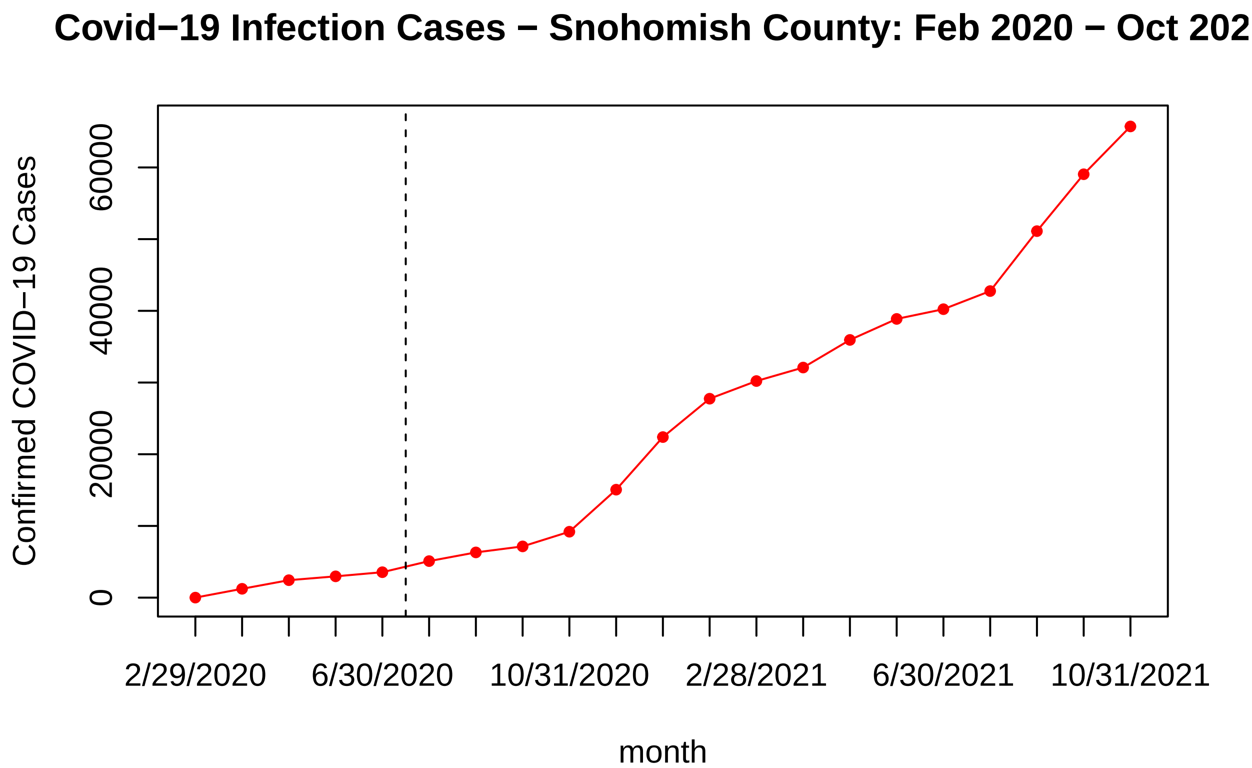
VISUALLY INSPECT DATA

This is to help look for any data quality issues that might be a problem and assess whether linear trends will be realistic. There are four things that I am interested in looking out for in the data:

1. Wild points
2. Linear trends
3. Co-interventions
4. Other data quality issues

The first section of the code below produces the initial line plot. I used the R command plot, which is the command to produce a chart. The two variables I plot against one another are time on the x-axis and infections on the y-axis. That is the outcome variable we are interested in. I named the y-axis "confirmed COVID-19 cases". The highest value in the data set is 65719, so I set limits from 0 to 66000. That means we get the full range of values. I labeled the X-axis month and created a line plot with the color red. In my code x-axis equals no (xaxt='n'); I do not want R to put numbers on the x-axis. And the reason is that I do not wish to see the numbers 1 through 21 on the x-axis. I want the actual months. So I'm saying to R, in the first instance, don't plot the x-axis. I added it using the following line of code. Next, I'm adding points, so I use the R command points. And I'm putting dots on each of the data points here to make it easier to follow. I use Abline, which produces a line on a plot at point 5.5. From our data, the first point after the mask policy mandate is 5095. This means that if I plot this line at 5094.5, that's going to leave the last point on the left and the next point on the right. The LTY equals 2 gives a dashed line, and this shows up in black on the chart.

```r
# Plot outcome variable versus time
plot(data$time, data$infections,
     main = "Covid-19 Infection Cases - Snohomish County: Feb 2020 - Oct 2021",
     ylab = "Confirmed COVID-19 Cases",
     ylim = c(0, 66000),
     xlab = "month",
     type = "l",
     col = "red",
     xaxt = "n")


# Add x-axis month labels
axis(1, at=1:21, labels = data$month)

# Add in the points for the figure
points(data$time, data$infections, col="red", pch=20)


# Label the intervention start month
abline(v=5.5, lty=2)
```



The plot shows a non-linear trend. There are two strategies to address this: quadratic model terms or differencing outcomes. In my experiment, I choose to use the quadratic model terms. Using the quadratic model term allows the linear shape to have some curve. Differencing outcomes can be used when a control group is used in the research project.

To set up the quadratic time trend, for intervention status j, at time t:

$outcome_{jt} = \beta_0 + \beta_1.time_t + \beta_2.level_j + \beta_3.trend_{jt} + \beta_4.trend^2_{jt} + \epsilon_{jt}$

β_0 = existing level, β_1 = existing trend, β_2 = level change, $\beta_3 + \beta_4$ = trend change

The addition of the quadratic term will make it harder to interpret. This is because I am not going to have a straight number that I can look at for the trend change, but I can add the two of them together, which is where predicting changes becomes quite useful.

I updated my data by adding a new column which is the trend squared.

PRELIMINARY ANALYSIS

In this step, I modeled a standard linear regression with a time series specification.

```
# A preliminary OLS regression
model_ols = lm(infections ~ time + level + trend + trendsq, data=data)

# See summary of model output
summary(model_ols)
```

```
##
```

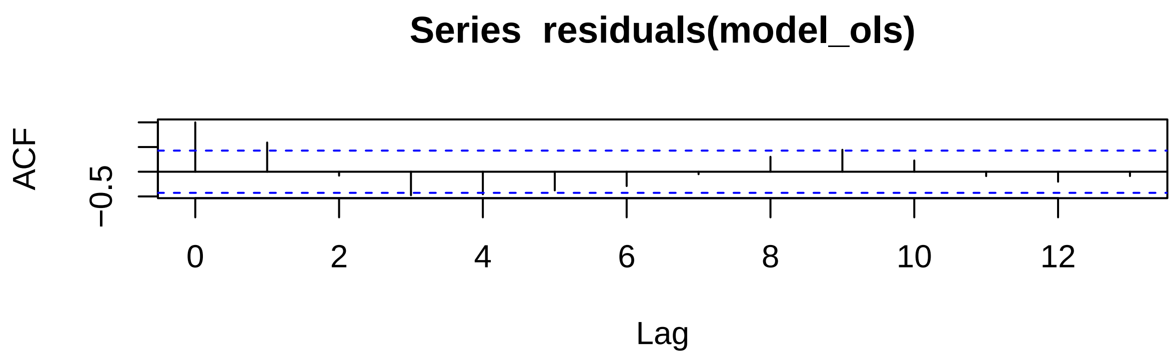
```
## Call:
## lm(formula = infections ~ time + level + trend + trendsq, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5373  -1167     49   1560   4359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -613.20    2847.31  -0.215   0.8322
## time          882.80     858.49   1.028   0.3191
## level       -3098.21    3131.28  -0.989   0.3372
## trend        1881.45    1063.76   1.769   0.0960 .
## trendsq        68.01      35.92   1.893   0.0765 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2715 on 16 degrees of freedom
## Multiple R-squared:  0.9861, Adjusted R-squared:  0.9826
## F-statistic: 283.6 on 4 and 16 DF,  p-value: 1.246e-14
```

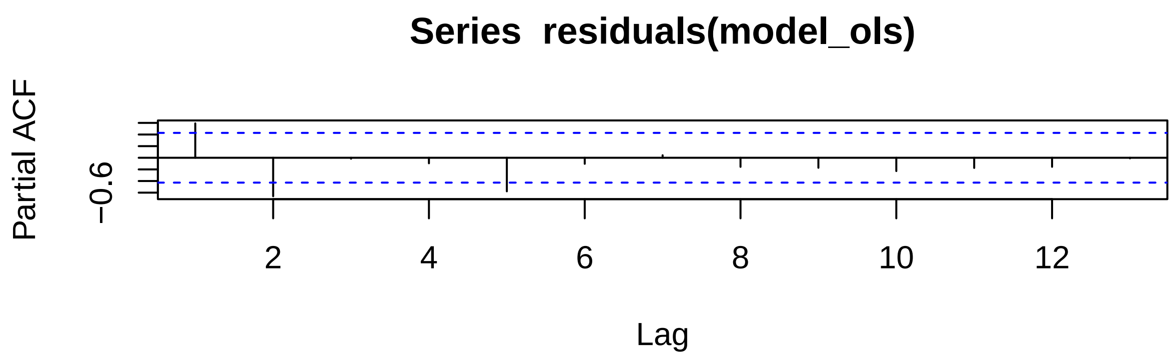
The results from the estimate column show that the existing trend increased over time by approximately 883 per month. So every month, we see that the actual infection rate was increasing before. That value, however, is not statistically significant. So it is not statistically differentiable from zero. The next variable level shows us the projected impact of the masking policy mandate on the level of confirmed COVID-19 cases over time. And we can see a dramatic and substantial drop of approximately 9736 positive points per month in confirmed COVID-19 cases. This is also not statistically significant. Also, we can see that there is also a *trend* and *trendsq* increase of approximately 1881 and 68, respectively. However, that is not statistically different from 0 because of the p-values. This result shows a change in the level, which persists over time, and a corresponding upward change in the trend.

It is important to take the model and assess whether or not there are issues with autocorrelation before running the final model.

AUTOCORRELATION

```r
## Checking for autocorolatoin

# ACF Plots
par(mfrow=c(2,1))

#Produce plots
acf(residuals(model_ols))
acf(residuals(model_ols), type='partial')
```





I used the generalized least squares (gls) approach to run the final model. This is similar to linear regression but allows the inclusion of an autoregressive or moving average process in the model itself. The gls() function is in the library NLME, or the non-linear mixed effects library. This library was loaded at the start of the code.

I created the model called model_p0 with an ar3 process, or p equals 3. I put a tilde in front of the time since that is the variable the autocorrelation operates over. I finally set the method to maximum likelihood.

FINAL MODEL

```
##Use AR(0) and MR(0) in the plot

model_p0 <- gls(infections ~ time + level + trend + trendsq,
             data=data,
             correlation=corARMA(p=3, form=~time),
             method = "ML")
summary(model_p0)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: infections ~ time + level + trend + trendsq
##   Data: data
##        AIC      BIC    logLik
##   372.8492 382.2499 -177.4246
##
```

```
## Correlation Structure: ARMA(3,0)
##  Formula: ~time
##  Parameter estimate(s):
##       Phi1        Phi2        Phi3
##  1.23399552 -0.89050379  0.06588654
##
## Coefficients:
##                 Value Std.Error    t-value p-value
## (Intercept)  990.8524 2545.8652  0.3892006  0.7023
## time         173.5561  687.8214  0.2523273  0.8040
## level       -845.7045 1263.6060 -0.6692786  0.5129
## trend       2640.7867 1062.8053  2.4847322  0.0244
## trendsq       61.3806   30.4826  2.0136231  0.0612
##
##  Correlation:
##         (Intr) time   level  trend
## time    -0.933
## level   -0.136  0.053
## trend    0.796 -0.914 -0.307
## trendsq -0.343  0.477  0.503 -0.782
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.1724489 -0.5534847  0.2702538  0.7030935  1.6763171
##
## Residual standard error: 2399.918
## Degrees of freedom: 21 total; 16 residual
```

A couple of things to point out in the results above. The first thing that R reports is the actual model. It then goes through some statistics about the model fit. It tells the correlation structure, so ARMA(3,0). This is the autoregressive parameter. This is the moving average parameter. The formula for the correlation itself follows. R then reports all the 3 phi parameters that it's come up with. The coefficients are the results: the intercept, time, level, and trend.

It can be inferred from the coefficients that, after the mask policy's introduction, there was a sustained drop in the monthly infection rate of approximately 845 people. However, the confirmed cases increased significantly, with about 2640 confirmed cases each month.

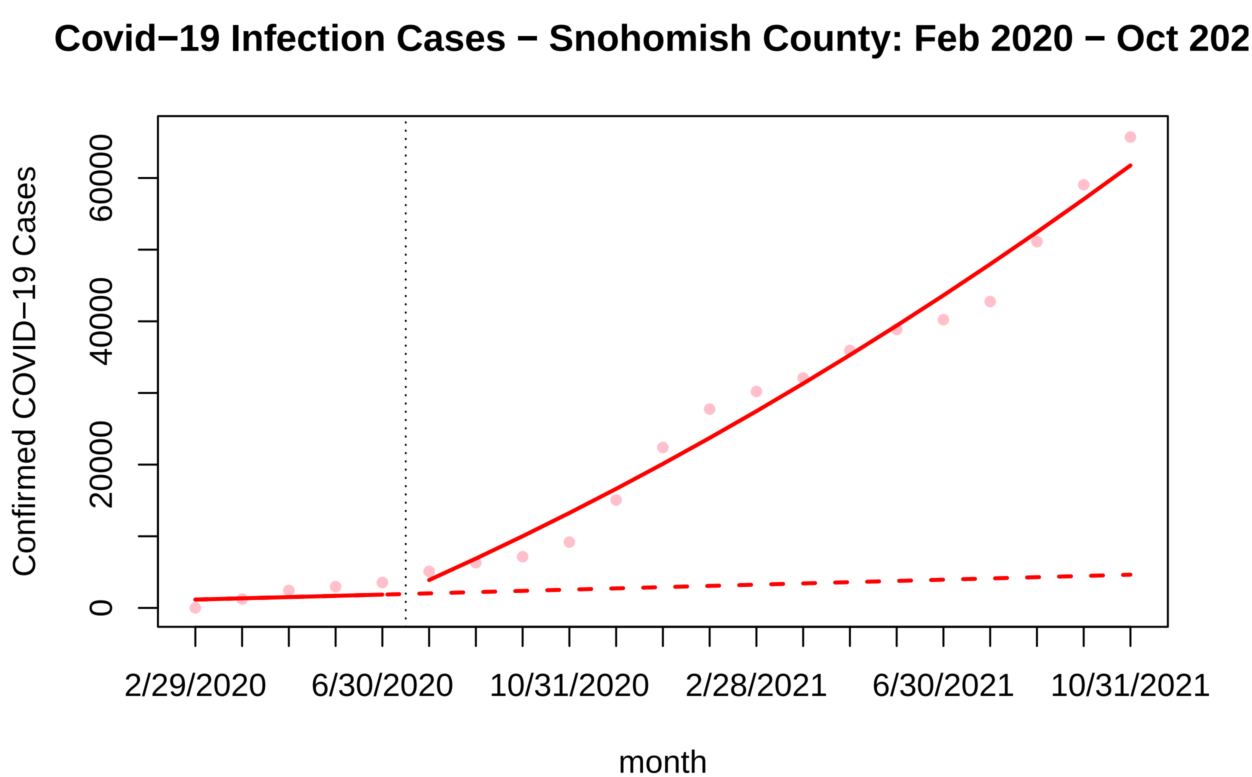
PLOTTING THE RESULTS

Now it's time to plot the results and get the raw data points. The Covid-19 confirmed cases are on the y-axis. I have set the limits on the y-axis to get all the data onto the chart. Both the y-axis and x-axis are well labeled, and I chose a circle as the plotting character. I decided to plot the pink so that the lines are a similar color to the plotted red lines but are a little bit dimmer. I completed the plot by putting in the observed lines.

I used the fitted command in R. What fitted does is it gives you the predictions out of the model. The time points for the two lines represent the pre-period and the post-period fitted lines. I used the segments command to plot the counterfactual line.

```r
#Plotting the Results

plot(data$time, data$infections,
     main = "Covid-19 Infection Cases - Snohomish County: Feb 2020 - Oct 2021",
```

```
      ylab = "Confirmed COVID-19 Cases",
      ylim =c(0, 66000),
      xlab = "month",
      pch = 20,
      col = "pink",
      xaxt="n")

# Plot dates on x-axis
axis(1, at=1:21, labels = data$month)

# Add line indicating policy start date
abline(v=5.5, lty="dotted")

# Plot the first line segment
lines(data$time[1:5], fitted(model_p0)[1:5], col="red", lwd=2)

# Plot the second line segment
lines(data$time[6:21], fitted(model_p0)[6:21], col="red", lwd=2)

#Add the counterfactual
segments(1, model_p0$coef[1]+model_p0$coef[2], 21,
         model_p0$coef[1]+model_p0$coef[2]*21,
         lty=2,
         lwd=2,
         col='red')
```



The results and the raw data points were plotted. The Covid-19 confirmed cases are on the y-axis. Both the y-axis and x-axis are well labeled, and a circle was chosen as the plotting character. I decided to plot the pink so that the lines are a similar color to the plotted red lines but are a little bit dimmer. I completed the plot by putting in the observed lines.

The fitted command in R was used. What fitted does is it gives you the predictions out of the model. The time points for the two lines represent the pre-period and the post-period fitted lines. The segments command was used to plot the counterfactual line. The plot indicates the preexisting level and trend, the modeled level and trend after, and the counterfactual, which tells what the assumption is or essentially what the outcome would have been absent from the introduction of the mask mandate policy.

Clearly, mask mandate policy was not enough by itself and therefore other interventions would be needed to be employed simultaneously to help prevent to progression of the disease.

EXTENTION PLAN

The purpose of the investigation is to steady the unemployment rate of the people living in the Snohomish County of Washington State from February 1, 2020, through to October 1, 2021, during the spread of the novel COVID-19, usually known as coronavirus. According to a report from the Census Bureau and Policy Priorities (CBPP), the COVID-19 pandemic and resulting economic fallout caused significant hardship. In the early months of the crisis, tens of millions of people lost their jobs. While employment began to rebound within a few months, unemployment remained high throughout 2020.

In April 2020, the Census Bureau began the Household Pulse Survey to collect nearly real-time data on how families were faring during this unprecedented crisis. At the end of 2021, the Census Bureau had released data from 39 Pulse surveys on household well-being. The unemployment rate jumped in April 2020 to a level not seen since the 1930s — and stood at 4.9 percent in October 2021, compared with 3.5 percent in February 2020. That official unemployment rate, moreover, understated job losses.

There were still 4.2 million fewer jobs in October 2021 than in February 2020. The majority of jobs lost in the crisis have been in industries that pay low average wages, with the lowest-paying industries accounting for 30 percent of all jobs but 59 percent of the jobs lost from February 2020 to October 2021, according to Labor Department employment data. Jobs were down nearly twice as much in low-paying industries (4.5 percent) as in medium-wage industries (2.6 percent) and roughly 15 times as much as in high-wage industries (0.3 percent) during this period. What is the story of the Snohomish County during this same period, and did the masking policy impact the rate of unemployment in anyway?

LOADING UNEMPLOYMENT DATA

The unemployment rate was already recorded in months. I prepared the data for interrupted time series analysis by adding the time, level and trend. Level represents when the mask mandate policy took effect.

```r
# You can also do this in RStudio using the "Import Dataset" button
unemployment_data <- read.csv("../data/unemployment_rate.csv",header=T)
print(unemployment_data)
```

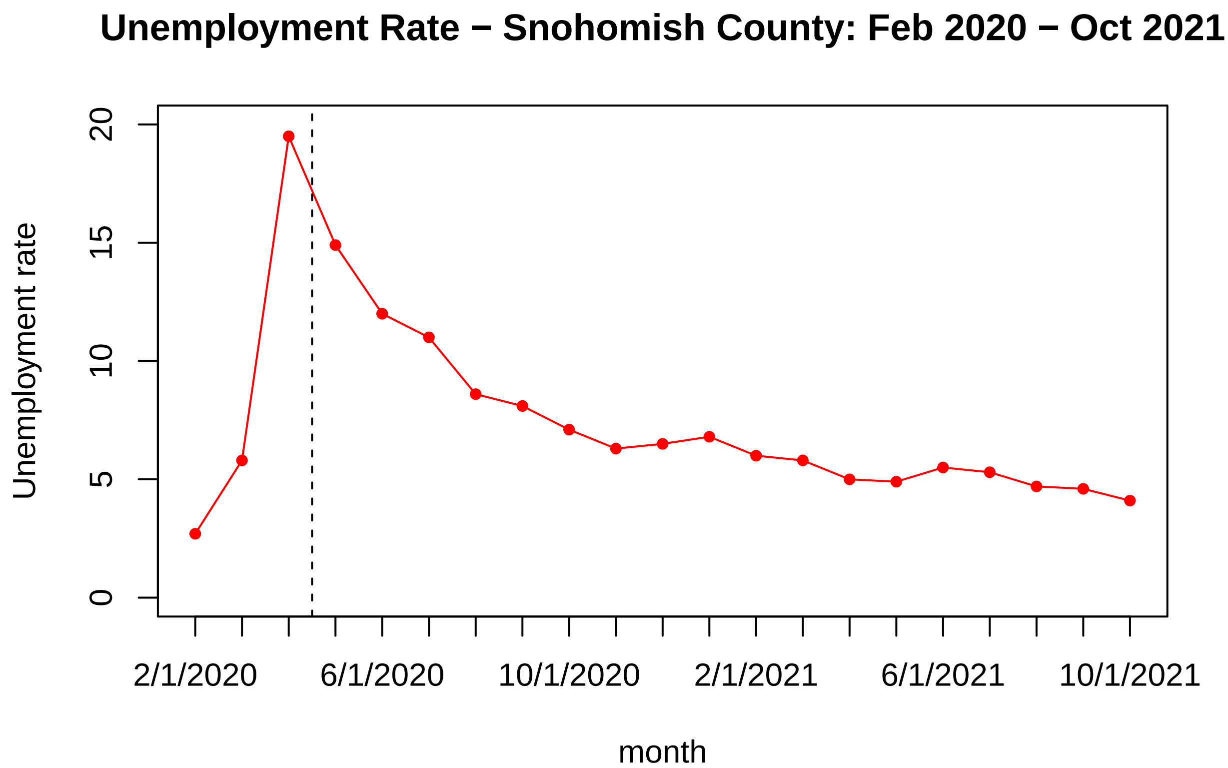
```
##        month rate time level trend
## 1   2/1/2020  2.7    1     0     0
## 2   3/1/2020  5.8    2     0     0
## 3   4/1/2020 19.5    3     0     0
## 4   5/1/2020 14.9    4     0     0
## 5   6/1/2020 12.0    5     0     0
## 6   7/1/2020 11.0    6     1     1
## 7   8/1/2020  8.6    7     1     2
```

```
## 8   9/1/2020  8.1    8     1     3
## 9  10/1/2020  7.1    9     1     4
## 10 11/1/2020  6.3   10     1     5
## 11 12/1/2020  6.5   11     1     6
## 12  1/1/2021  6.8   12     1     7
## 13  2/1/2021  6.0   13     1     8
## 14  3/1/2021  5.8   14     1     9
## 15  4/1/2021  5.0   15     1    10
## 16  5/1/2021  4.9   16     1    11
## 17  6/1/2021  5.5   17     1    12
## 18  7/1/2021  5.3   18     1    13
## 19  8/1/2021  4.7   19     1    14
## 20  9/1/2021  4.6   20     1    15
## 21 10/1/2021  4.1   21     1    16
```

VISUALLY INSPECTING UNEMPLOYMENT DATA

This is to find out if there is any problem that will affect the results of the interrupted time series analysis.

```r
# Plot outcome variable versus time
plot(unemployment_data$time, unemployment_data$rate,
     main = "Unemployment Rate - Snohomish County: Feb 2020 - Oct 2021",
     ylab = "Unemployment rate",
     ylim = c(0, 20),
     xlab = "month",
     type = "l",
     col = "red",
     xaxt = "n")


# Add x-axis month labels
axis(1, at=1:21, labels = unemployment_data$month)

# Add in the points for the figure
points(unemployment_data$time, unemployment_data$rate, col="red", pch=20)


# Label the intervention start month
abline(v=3.5, lty=2)
```



The graph above does not seem to have any data quality issues. It is good for interrupted time series analysis.

ANALYSIS

In this step, I modeled a standard linear regression with a time series specification.

$outcome_{jt} = \beta_0 + \beta_1.time_t + \beta_2.level_j + \beta_3.trend_{jt} + \epsilon_{jt}$

β_0 = existing level, β_1 = existing trend, β_2 = level change, β_3 = trend change

The model is setup as follows and the explanation is similar to the first model above.

```r
# OLS regression
model_ols2 = lm(rate ~ time + level + trend, data=unemployment_data)

# See summary of model output
summary(model_ols2)
```

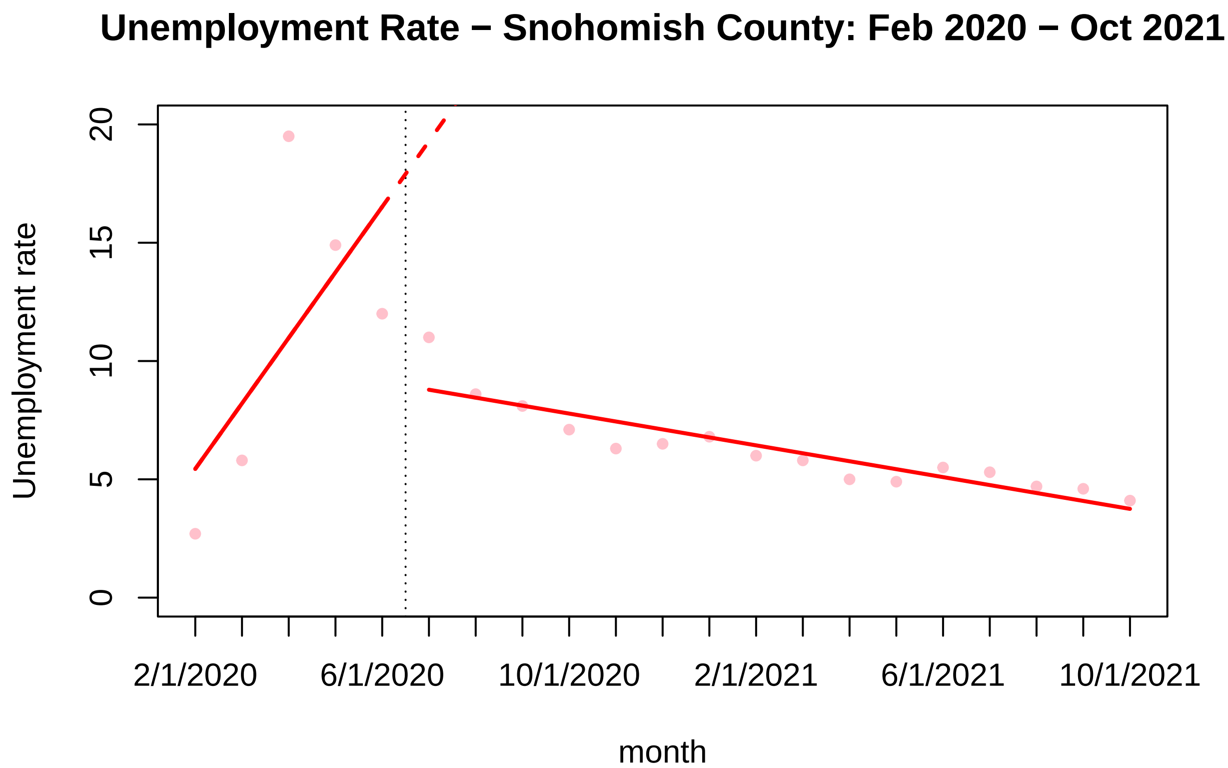
```
##
## Call:
## lm(formula = rate ~ time + level + trend, data = unemployment_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5200 -0.6796 -0.0153  0.4063  8.5200
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.6700     2.7490   0.971  0.34503
## time          2.7700     0.8289   3.342  0.00386 **
## level        -7.3975     2.4518  -3.017  0.00776 **
## trend        -3.1057     0.8410  -3.693  0.00180 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.621 on 17 degrees of freedom
## Multiple R-squared:  0.6309, Adjusted R-squared:  0.5657
## F-statistic: 9.685 on 3 and 17 DF,  p-value: 0.000588
```

From the result above, the next variable level shows us the projected impact of the masking policy mandate on the level of unemployment rate over time. And we can see a dramatic and substantial drop of approximately 7 percent in unemployment rate. This is statistically significant. Also, we can see that there is also a trend decrease of approximately 3 percent monthly. That is also statistically significant because of the p-values.

PLOTTING THE RESULTS

```r
#Plotting the Results

plot(unemployment_data$time, unemployment_data$rate,
     main = "Unemployment Rate - Snohomish County: Feb 2020 - Oct 2021",
     ylab = "Unemployment rate",
     ylim =c(0, 20),
     xlab = "month",
     pch = 20,
     col = "pink",
     xaxt="n")

# Plot dates on x-axis
axis(1, at=1:21, labels = unemployment_data$month)

# Add line indicating policy start date
abline(v=5.5, lty="dotted")

# Plot the first line segment
lines(unemployment_data$time[1:5], fitted(model_ols2)[1:5], col="red", lwd=2)

# Plot the second line segment
lines(unemployment_data$time[6:21], fitted(model_ols2)[6:21], col="red", lwd=2)

#Add the counterfactual
segments(1, model_ols2$coef[1]+model_ols2$coef[2], 21,
         model_ols2$coef[1]+model_ols2$coef[2]*21,
         lty=2,
         lwd=2,
         col='red')
```



This result shows a change in the level, which persists over time, and a corresponding downward change in the trend.

PREDICTING CHANGES

In my final analysis, I decided to look at the difference between my predicted line and counterfactual at some point in the post period, to get an idea of what the intervention or what the policy did at a particular time. Given that the policy was implemented 10months after February 2020. As you can see from the plot above, because there is trend line tilts downward and the predicted counterfactual goes upward, the difference is going to grow as time goes on.

```
# Predicted value at 10 months after the mask policy intervention
pred <- fitted(model_ols2)[15]

#Then estimate the counterfactual at the same time point
cfac <- model_ols2$coef[1] + model_ols2$coef[2]*15

# Absolute change at 10 months
pred - cfac
```

```
##        15
## -38.45485
```

```
# Relative change at 10 months
(pred - cfac) / cfac
```

```
##         15
## -0.8696258
```

For the absolute change, I tooke the predicted value and subtracted the counterfactual at that same time point. That will give an idea as to whether or not the value is increased or decreased. I took the relative change by taking the prediction minus the counterfactual and dividing it by the counterfactual, or the value I would have expected, which will give me a percentage increase or decrease relative to the counterfactual that I estimated.

After calculating in R, the prediction minus the counterfactual at point 15, gave an estimate of negative 38.45. The relative change gave a drop of 86.96%. The interpretation here is that in the 10th month after the masking policy was introduced, the average montly unemployment rate was 38.45 percent less per month than would have been expected if the masking policy was not implemented. This represented a 86.96% reduction.

DISCUSSION

Masking policy initiative introduced in the Snohomish County of Washington State, USA in July 2020 did not prevent the progression of confirmed Covid-19, however, this policy impacted the unemployment rate positively in this same region. Using an interrupted time series study design, I evaluated the impact of the policy on the progression of confirmed Covid-19 cases and unemployment rate in the Snohomish County over twenty-one months. Throughout the study period I observed an exponential increase in the number of reported Covid-19 cases. My study showed that the implementation of the masking policy had neither an immediately discernible (a level change) nor a sustained (a trend change) effect on the rate of Covid-19 confirmed cases. The results from this study were not statistically significant. There was however, a statistically significant changes in the rate of unemployment after the masking policy was introduced. The level and trend for unemployment changed significantly.

In my final analysis, I decided to look at the difference between my predicted line and counterfactual at some point in the post period, to get an idea of what the intervention or what the policy did at a particular time. Predicted value at ten months after the mask policy intervention. As you can see from Figure 2 above, because post trend line tilts downward and the predicted counterfactual goes upward, the difference is going to grow as time goes on. For the absolute change, I took the predicted value and subtracted the counterfactual at that same time point. That gave an idea as to whether the value is increased or decreased. I took the relative change by taking the prediction minus the counterfactual and dividing it by the counterfactual, or the value I would have expected, which will give me a percentage increase or decrease relative to the counterfactual that I estimated.

After calculating in R, the prediction minus the counterfactual at point 15 (in terms of the time variable), gave an estimate of negative 38.45. The relative change gave a drop of 86.96%. The interpretation here is that in the 10th month after the masking policy was introduced, the average monthly unemployment rate was 38.45 percent less per month than would have been expected if the masking policy was not implemented. This represented an 86.96% reduction.

LIMITATIONS

One of the limitations of this study is that only clinically diagnosed Covid-19 cases were counted. At the same time, there might be positive Covid-19 cases not diagnosed if patients avoided tests or did not have severe symptoms. Even so, this issue covers the entire studied period, not just the period after the mask mandate policy order. Therefore, undiagnosed cases would not significantly affect the intervention results.

CONCLUSION

This research used interrupted time series to analyze the impact of masking policies on the progression of confirmed Covid-19 in the Snohomish County, Washington State, USA. It also ascertained how the unemployment rate changed when the masking policy was implemented. Quadratic term was added to capture the non-linearity in the Covid-19 prevalence model. Several contributions distinguish this study from previous papers. For instance, this study did not consider any NPIs or vaccination policy/people which affected the dynamic of the Covid-19 epidemy. Also, this paper is human centered in the sense that, it puts real people at the forefront in hopes of solving a real-life problem. Despite the opinions of people who stood against the mask mandate policy, ITS analysis results show the importance of the mask wearing policy in controlling the Covid-19 prevalence with respect to unemployment rate. The mask wearing policy cased a significant reduction in level (-845 people/day) before other factors caused an exponential increase in the infection rate. However, when it comes to unemployment rate, the mask policy caused a significant reduction in both level (-7.4 percent) and trend (-3.0 percent per month) of unemployment rate. Even though results from the study of the mask policy against the progression of the epidemy was not statistically significant, the mask policy in the Snohomish County in early July 2020, caused a 7% reduction in unemployment rate after the intervention .