**Unemployment and COVID-19 in the Greater Boston Area**

MA 214 Term Project

The Outliers

Christopher Balboni, Alec Lanza, Connor Long, Nicole Lyu

Professor Vaughan

13 December 2020

**MA214 Final Project**

***(1) Title and Abstract/Executive Summary***

Throughout the course of our research into unemployment in the Greater Boston Area and the impact the COVID-19 Pandemic had on unemployment levels in the state of Massachusetts, we have come to two conclusions. For this project, we were required to perform a simple statistical analysis, and a more complex analysis. We performed our analysis because as Bentley business students, we are very interested in unemployment levels as it is an important indicator of local economic health. We wanted to know if we could use COVID-19 case levels as an accurate predictor of unemployment. From our simple analysis, we have concluded that unemployment levels in the cities of Boston and Waltham are different, with Boston having higher unemployment. We came to this conclusion by utilizing unemployment data collected from the State of Massachusetts Department of Labor website. Using this data, we created a corresponding confidence interval to determine if the numbers are different. For our complex analysis, we wanted to explore if the number of reported COVID-19 cases is a good predictor of the level of unemployment. We concluded that our model is fundamentally flawed due to the nature of the data, which made any tool used from a linear regression model invalid. We attempted to perform a linear regression analysis on the data, however when we were working through the standard assumption checks, we noticed that the COVID-19 case data comes from a time-series dataset. When performing the Durbin-Watson Test of autocorrelation, we determined that there was in fact correlation between the data points, and that we were unable to perform our analysis.

***(2) Introduction***

The purpose of this report is to provide insight into the unemployment market in the state of Massachusetts. This report was written by a team of four students enrolled in Bentley University’s MA-214-HB1 course. As part of our course curriculum, we were tasked with creating a statistical report based on a prevalent and practical issue. We chose to focus our report on unemployment in the Greater Boston area, and the impact that the COVID-19 Pandemic had on unemployment in Massachusetts. As students who are currently studying in Waltham, and presumably going to find work in the Greater Boston area in the future, having an understanding of the unemployment market in our region is important to us.

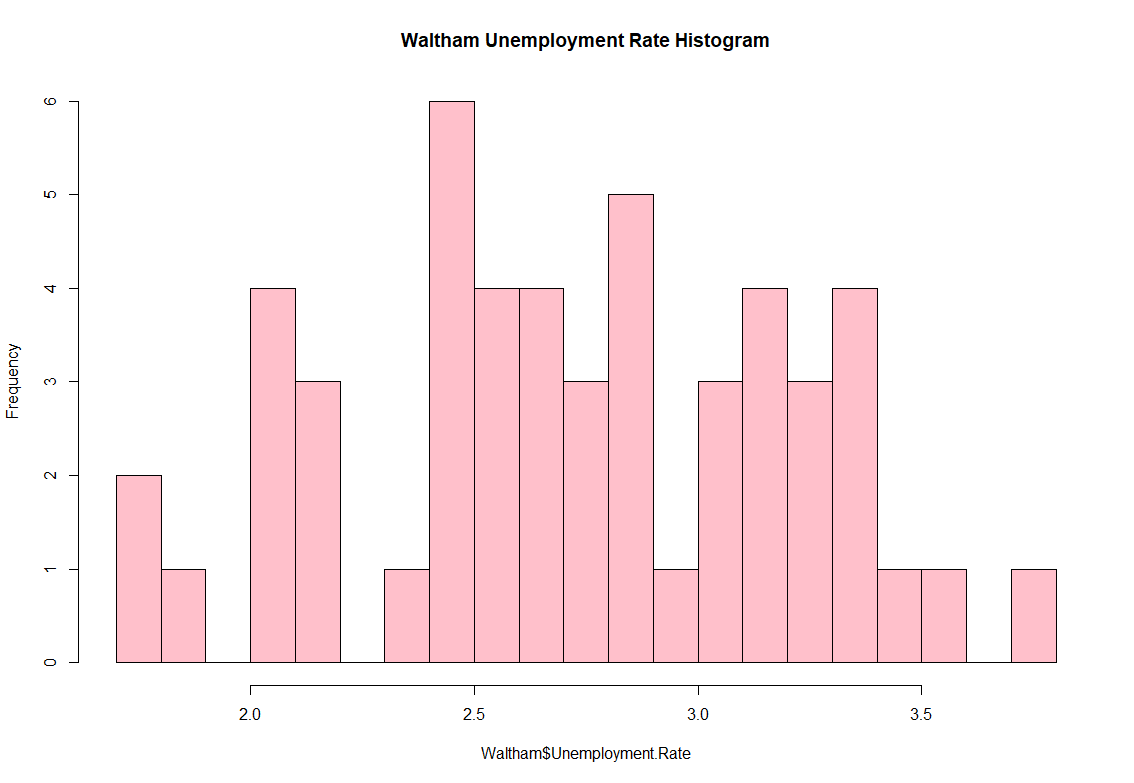
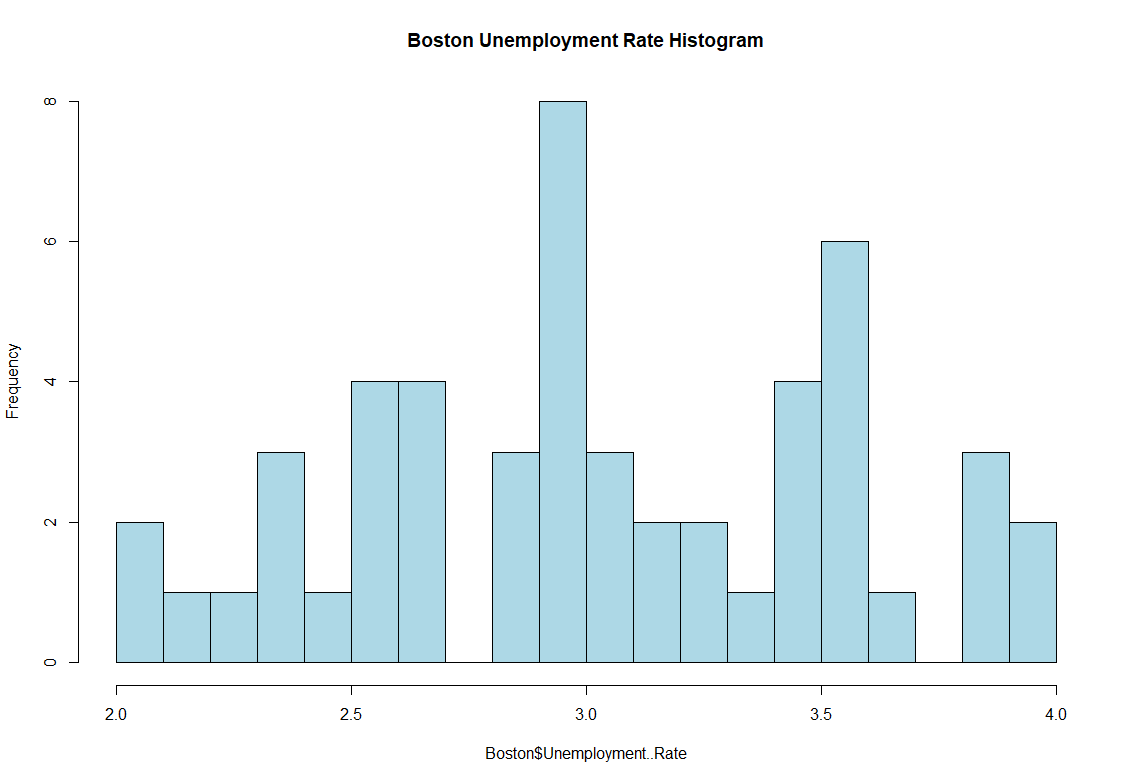
This report has three main sections. We will begin by providing data characteristics where we detail the data we collected, as well as where we collected it from. We will then provide base level information about the shape, center, and spread of this data, which will be useful to us later in the report. To continue, we will then perform an in-depth analysis of our data, where we perform two separate statistical tests on our data. We first conduct a two-populations means testing, where we will determine whether there is a difference in unemployment between the cities of Waltham and Boston during non-lockdown and non-quarantine times. We will follow that with an advanced linear regression model, attempting to model how the number of COVID-19 cases in the state of Massachusetts impacts the level of unemployment in the state. To finish, we will provide summaries of our findings and concluding remarks.

***(3) Data Characteristics***

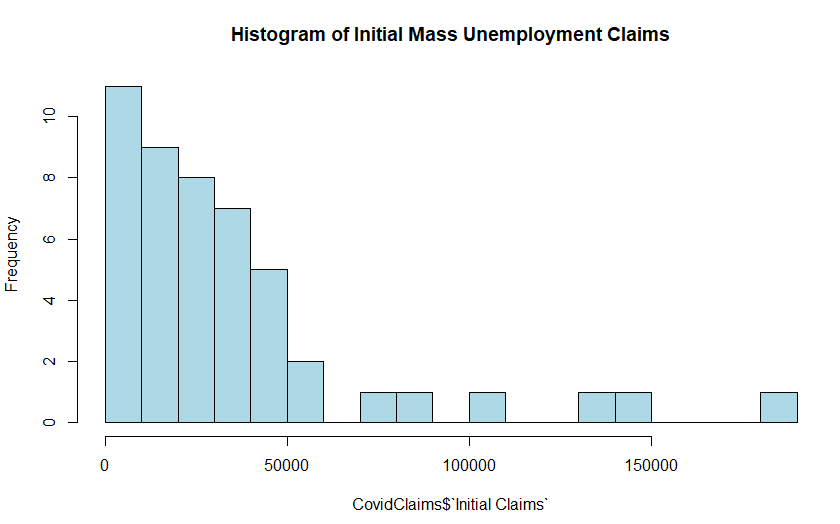
Descriptive Analysis

We constructed histograms of both samples for Boston and Waltham’s unemployment rates. As the employment data is taken every month, the data is time series. The data is also quantitative, continuous, ratio scale, and observational. Boston’s Unemployment rate has a mean of 3.076, a median of 3, and a standard deviation of 0.495. It has a reasonably normal distribution with most data points in the middle. Waltham’s unemployment rate has a mean of 2.773, a median of 2.8, and a standard deviation of 0.495, making it very slightly left skewed, but still fairly normal.

Unemployment Rates Histograms:

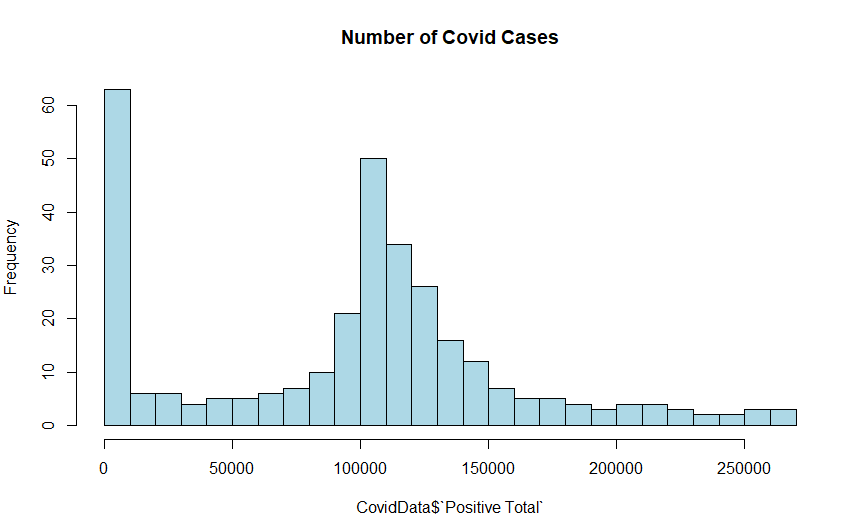


Next we took a look at the initial unemployment claims for the entire state of Massachusetts. This is once again quantitative, continuous, ratio scale, observational, and time series. This data has a mean of 36495.6, a median of 26255.5, and a standard deviation of 37731.07. This data appears skewed to the right perhaps due to the exponential growth of COVID-19 cases.



As for the COVID-19 cases, the data is quantitative, discrete, ratio scale, observational, and also time series as it is taken daily. The number of COVID-19 cases per day has a mean of 94,821.74, a median of 104,443.5, and a standard deviation of 64,487.72. This data does seem fairly left skewed due to a large number of instances of low cases in the histogram, those points are worth noting when we do our tests later.

Covid Cases Histogram:



***(4) In Depth Analysis***

**2- Population Test-**

Purpose- The purpose of our analysis is to see if there is a true difference in average monthly unemployment rate of two cities in Massachusetts during “normal” times (pre-pandemic). We chose Boston and Waltham because as a Bentley student, these are the two places where we spend most of our time.

Method Used-

Creating a confidence interval on the data of 2-population means will tell us whether there is a true difference between the unemployment rate of Boston and Waltham. We collected our data from the Massachusetts state government website, specifically the labor market information page. Although the data is available up until October 2020, we chose to only collect data before April 2020. This is because COVID-19 dramatically increased the unemployment rates and thus created significant outliers for both samples. Since our goal is to figure out the average unemployment rate of two cities under “normal” conditions, we gathered 51 data points from January 2016 to March 2020. We get the results of a 95% confidence interval by using the t.test() function in R.

Identify the problem

According to the inferential framework, first, we identified this is a estimation problem because we are not making any decisions with the data, simply testing whether there is a difference. Then, we found the sample is composed of quantitative data, and this is a means problem.

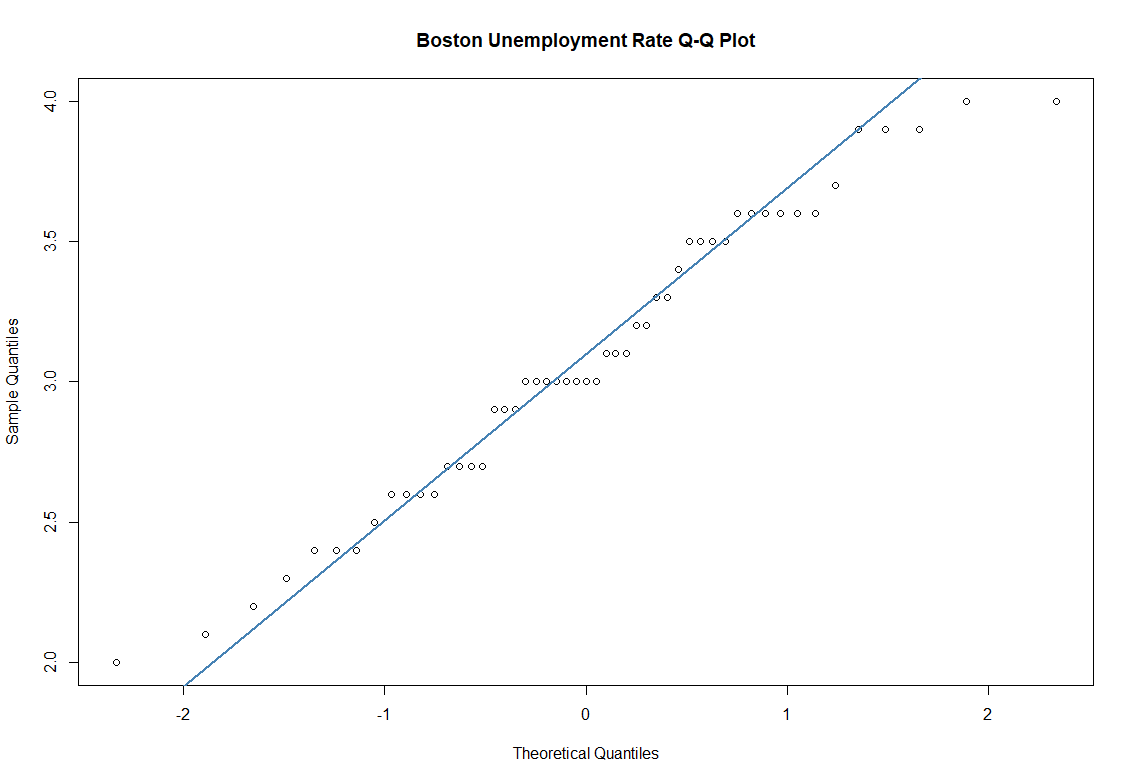
This is a 2-population means problem, thus:

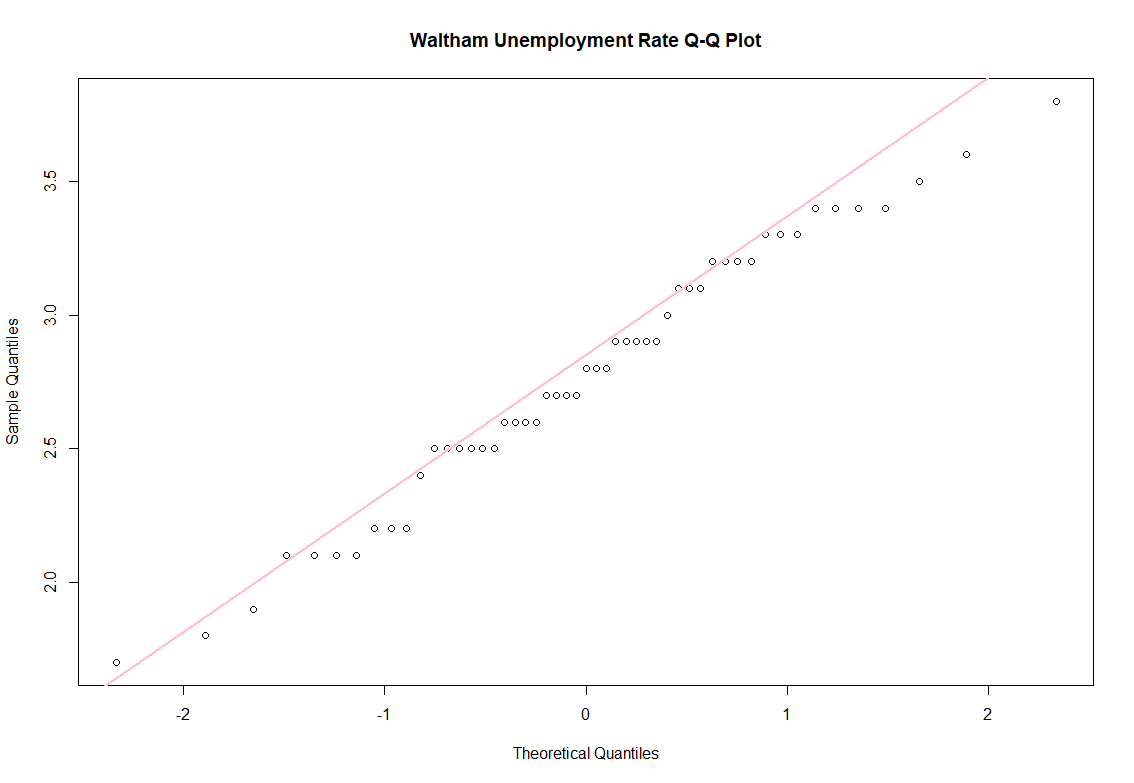
Our target is difference in average unemployment rate, µD = µBoston − µWaltham.

Estimator is sample difference in average unemployment rate, X Boston − Y Waltham = 3.076471 - 2.772549 = 0.3039216.

Checking normality-

For a sample size bigger than 30, the central limit theorem indicates that the sampling distribution of a population can be approximated by a normal probability distribution. In this case, both samples have 51 data points, so the central limit theorem applies. We construct Q-Q Plots to double check if the populations are normally distributed.





Most dots lie on the line, so we can conclude that both populations are normal.

Interpretation of Test Result-

We can find the confidence interval bounds by doing t.test( ). The result above tells us that our 95% 2-sided confidence interval for the difference in the mean unemployment rate in Boston versus Waltham is (0.104, 0.503). 0 is not included in the interval and therefore we reject the Null Hypothesis, so there is a true difference in Boston unemployment rate and Waltham unemployment rate. Because both bounds of the confidence internal are positive, and the interval does not include 0, it means that the average Boston unemployment rate minus that of Waltham is bigger than 0. We finally can conclude that Boston has a true higher unemployment rate than Waltham during “normal” times.

**Linear Regression Test-**

Purpose- The purpose of our analysis is to see if the number of COVID-19 cases reported in a week can be a predictor for the number of unemployment claims filed that week. Performing a linear regression test will give us more insight into whether there is a correlation between weekly COVID-19 cases and unemployment claims and the strength of the correlation between the two variables. Based on these observations, we can determine just how effective weekly unemployment claims can be predicted by positive weekly COVID-19 cases.

Method Used- Since we are intending to use simple linear regression, we are assuming:

Yi = β0 + β1Xi + Ei

Where:

* Yi= The predicted weekly number of unemployment claims
* Xi= The weekly number of positive COVID-19 cases
* Β0 and β1= The intercept and slope of the function
* Ei= The error term

To use a simple linear regression, we have collected data from 2 observational studies. The data regarding weekly positive COVID-19 cases comes directly from the Massachusetts Department of Health, and the weekly unemployment claims come from United States Department of Labor. With our data, we estimated parameters and coefficients of the model by using the “Method of Observed Least Squares (OLS), which is used to “fit” a linear equation to the observed data points. This was done in R Studio using the lm() function. All R code used in this section can be found in the appendix under *Linear Regression Test*, note 3.

Linear Model Utility Assessment-

This section will walk through how useful the model is our model can be at predicting and describing our data. The two tools for evaluating model utility are as follows:

1. *Coefficient of Determination:*

When doing a linear regression test, the goal of the test is to reduce the total variation in the response. Measuring variation involves 3 methods, which are defined in (1) in the *Linear Regression test* section of the appendix.

When dividing 2 of the method, the sum of squares due to regression over the sum of squares total (SSR/SST), we get the coefficient of determination (R^2), which is used to summarize how much of the variation can be explained by weekly COVID-19 cases. By using the lm() function and displaying the results of this function with the summary() function in R Studio, the resulting R^2 value was:

R^2= 0.2669

This means about 26.69% of the variability we see in the weekly unemployment claims is explained by the variability we see in the weekly positive COVID-19 cases reported. We recognize that the R^2 value is low, which tells us the linear relationship between the response Yi and the predictor Xi is weak, but this does not mean using weekly COVID-19 cases as a predictor is not important.

1. *Mean Squared Error (MSE)/Root MSE (RMSE):*

Another metric we can use to evaluate our model is its predictive performance. To do this, we look at the square root of the MSE (refer to (2) in the *Linear Regression test* section for more information on MSE), which is as follows:

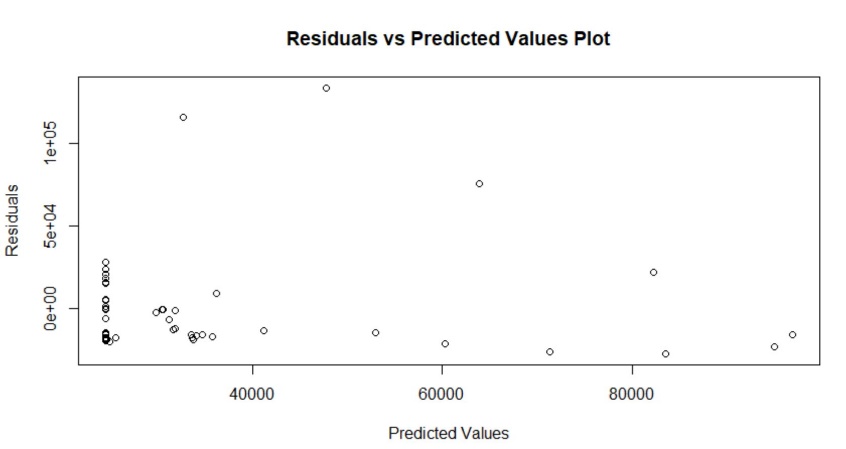
RMSE= MSE^½

This can be used as an estimator for the standard deviation, which is a tool to to evaluate how useful our model is. Using R Studio, we can obtain the RMSE by using the summary(fit) function, which we determined to be 32660. This means after taking our predictor, the amount of positive weekly COVID-19 cases, into account, the errors have a standard deviation of 32660 unemployment claims. To put the standard deviation into context, unemployment claims range from 4712 to 181423 claims. The size of the standard deviation tells us if our predictions are generally correct. If the standard deviation is too large, predictions are meaningless. Size is based on context, so in this case we can conclude that predictions can be reasonably correct since the standard deviation compared to the size of the claims is not small, but it is also not exceptionally large.

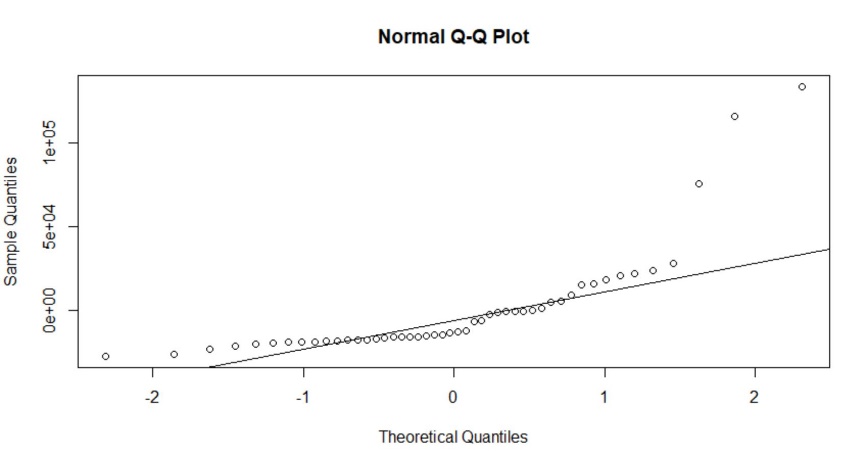
Assumptions-

After checking for the utility of our model, we must then check the assumptions made for a linear regression test:

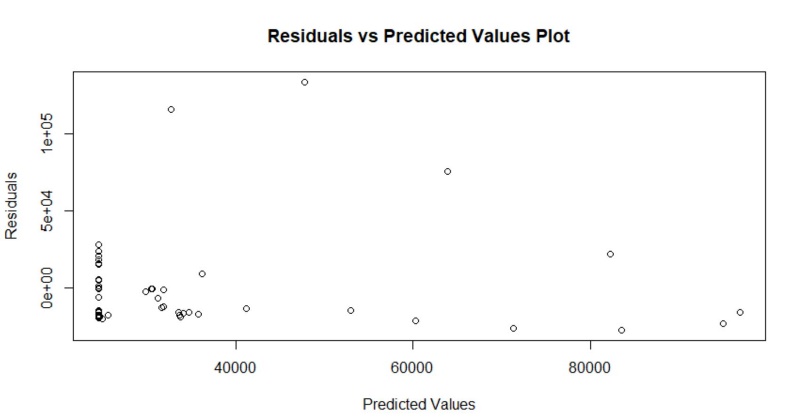
1. *The Model Fits “Well”-*

This assumption is checked by examining a residual vs. predicted values plot in R Studio as shown on the left. From observing the plot, we can see there is no clear pattern, so we can conclude the hypothesized relationship is reasonable.

1. *Distribution of Errors (Ei ~ Normal)-*

This assumption is to check the normality of the errors, which can be done by examining a QQ plot of our residuals. From the QQ plot made on R Studio, we see that the residuals follow a relatively straight line, but we also not there is a big deviation in the top right corner, and a slight deviation in the bottom left corner. 

1. *Homogeneity (Constant Standard Deviation)-*

To assess if the standard deviation is constant, we again look at our residual vs predicted values plot. From our observation we note that the residuals are not necessarily evenly spread out, but there is also not a clear pattern the residuals follow so we determine the assumption is reasonable. 

1. *Independence of Errors-*

The data we used is time series data, which means the observations collected come in a clear order. In the case of our data, the observations are made weekly. Since we used time series data, we must test to see if there is autocorrelation, which is done by using the Durbin-Watson test in R Studio. The code for this test can be referenced (3) under the *Linear Regression Test* section of the appendix. This test structures the hypotheses as:

H0= There is no autocorrelation vs Ha: There is autocorrelation

The resulting p-value from performing the Durbin-Watson test at the .01% significance level on R Studio was 0, which means there is evidence of autocorrelation in our data. To solve this issue, we changed the predictor and response data to percent changes because this change has the potential to remove some of the autocorrelation. After doing this and performing the Durbin-Watson test at the .01% significance level again, we still received a p-value of 0. Due to this reason, we cannot use our model. Refer to the concluding remarks under the *Linear Regression Conclusion section* for limitations and improvements of the test performed.

***(5) Summary and Concluding Remarks***

*2-Population Conclusion-*

Unemployment rate is an important indicator of local economic health. We first found out the average unemployment rate in Boston is higher than Waltham before the pandemic by doing a 2-population means test. The data we used was attained through Massachusetts government labor market info, with no target on one socioeconomic group or another. As such, the data is unbiased and provides us with a fair model for the general public. Our test result might indicate Waltham has a more stable economy than Boston, and further investigation is needed as to specific reasons for this difference and data might need to be adjusted seasonally.

*Liner Regression Conclusion-*

By failing the Durbin Watson test for autocorrelation, we cannot proceed any further with our model. This means we are unable to interpret relationships and make inferential statements between the number of positive COVID-19 cases and weekly unemployment claims. We also cannot estimate the average number of unemployment claims based on a given number of positive COVID-19 cases or predict a range of new unemployment claims based on the number of positive COVID-19 cases. This is because all of the tools mentioned are based on the assumption checks made, and when one assumption fails, it makes the results from any tool worthless. We believe there were limitations relating to the amount of data and difference in how the data was taken. To improve this study, we think it should be performed on a smaller scale, which would reduce the limitations mentioned. The data we have now is extremely limited due to the time span of COVID-19. For a detailed explanation on the limitations and improvements of this study, refer to (4) and (5) of the *Linear Regression Test* in the Appendix.

***(6) Appendix and Reference***

**Appendix-**

*Descriptive Analysis Code Used-*

BostonMean<-mean(Boston$Unemployment..Rate)BostonMean

BostonMedian<-median(Boston$Unemployment..Rate)BostonMedian

sd(Boston$Unemployment..Rate)

hist(Boston$Unemployment..Rate,col="lightblue", main="Boston Unemployment Rate Histogram",breaks=20)

WalthamMean<-mean(Waltham$Unemployment.Rate)

WalthamMedian<-median(Waltham$Unemployment.Rate)

sd(Waltham$Unemployment.Rate)

hist(Waltham$Unemployment.Rate,col="pink", main="Waltham Unemployment Rate Histogram", breaks=20)

CovidClaimsMean<-mean(CovidClaims$InitialClaims)

CovidClaimsMedian<-median(CovidClaims$InitialClaims)

sd(CovidClaims$InitialClaims)

hist(CovidClaims$`Initial Claims`, col = "lightblue", breaks = 20, main = "Histogram of Initial Mass Unemployment Claims")

CovidDataMean<-mean(CovidData$PositiveTotal)

CovidClaimsMedian<-median(CovidData$PositiveTotal)

sd(CovidData$PositiveTotal)

hist(CovidData$PositiveTotal, col = "lightblue", breaks = 20, main = "Number of Covid Cases")

*2-Population Means Test-*

(1) The R Code Used:

qqnorm(Boston$Unemployment..Rate, pch = 1, main="Boston Unemployment Rate Q-Q Plot")

qqline(Boston$Unemployment..Rate, col = "steelblue", lwd = 2)

qqnorm(Waltham$Unemployment.Rate, pch = 1, main="Waltham Unemployment Rate Q-Q Plot")

qqline(Waltham$Unemployment.Rate, col = "pink", lwd = 2)

Estimator <- BostonMean-WalthamMeanEstimator

t.test(Boston$Unemployment..Rate,Waltham$Unemployment.Rate)

*Linear Regression Test-*

(1) Measuring variation involves 3 methods:

* + Sum of Squares due to Regression (SSR)- SSR is the amount of variation that the model can explain by the variation in the predictor, weekly COVID-19 cases.
  + Sum of Squares due to Error (SSE)- SSE is the amount of variation that the model cannot explain by weekly COVID-19 cases.
  + Sum of Squares Total (SST)- SST is the total amount of variation in the response Yi, the weekly number of unemployment claims.

(2) The square root of the mean squared error is developed as follows:

Ei = Yi–

Where:

Ei= Residual

Yi= Observed weekly number of unemployment claims

= The predicted weekly number of unemployment claims

When squaring these values, we get the sum of squares due to error:

SSE=

Then we take the average of the SSE to get the MSE, which is then used to get the root mean squared error (RMSE):

MSE= SSE/(n-2)

RMSE= MSE^½

(3) R code used:

fit <- lm(InitialClaims~WeeklyCases, data = RUpload)

Summary(fit)

plot(predict(fit), residuals(fit), main= “Residuals vs Predicted Values Plot”, xlab= “Predicted Values”, ylab= “Residuals)

qqnorm(residuals(fit), main= “Normal Q-Q Plot of Residuals”)

library(car)

durbinWatsonTest(fit)

(4) Limitations Explanation:

There are many limitations of this study that should be noted. The first limitation to be considered is the amount of data we have. For the data used in the linear regression test, the range of data was from the beginning of January to the end of November. Since the data for unemployment claims was taken weekly, we had to bunch the COVID-19 daily cases into weeks, which gave us 48 observations. If COVID-19 continues to affect the United States as it has in the past year, this would give us more observations to base the assumptions of our test on. If the assumptions were to pass, more observations will make the tools used in a linear regression model more effective. Another limitation of this study is the rapid increase in testing throughout the period the data was taken from. The COVID-19 pandemic occurred abruptly. Due to this reason, no one had any tools even detect the virus. The testing of the virus is pivotal to this study because that is how the data is collected. January, February, March and even April most likely do not have accurate representations of the amount of people who have tested positive because there was simply no mass production of COVID-19 tests. Testing kits were sparse, and only reserved for people who showed clear symptoms of the virus. Additionally, since the outbreak, numerous different tests have become available to the public for use. This means all COVID-19 cases were not a result of the same testing kit used. The lack of consistency with testing can undoubtedly skew results because some test kits are more accurate than others.

(5) Improvements Explanation:

To improve this study right now, we believe we need to collect data on a smaller scale rather than observing data for an entire state. To do this, we should select a populated town and observe the amount of positive COVID-19 cases and unemployment claims for the week. By doing this, we think this will reduce the effect of variability in testing limitation we described above. In a town, there is most likely 1 or 2 testing sites. These testing sites are likely to be run by the same lab, which means they are using the same type of testing kit. The results of the cases reported would be jumbled with results of other testing kits, which would make the results more consistent, and maybe even more accurate. As a result, the more consistent data may be a better indicator of weekly unemployment claims. Our only other recommendation is to wait to conduct this test on a larger scale. As time progresses, testing will become more standardized and accurate, which will again give more consistent data to compare to the unemployment claims. Additionally, this will give rise to more data being produced.

**References-**

|  |  |
| --- | --- |
| **Title** | **Links** |
| Department of Labor | <https://oui.doleta.gov/unemploy/claims.asp> |
| Massachusetts COVID-19 Data | <https://www.mass.gov/info-details/covid-19-response-reporting> |
| Massachusetts Unemployment Claims | [https://lmi.dua.eol.mass.gov/LMI/LaborForceAndUnemployment#](https://lmi.dua.eol.mass.gov/LMI/LaborForceAndUnemployment) |