Modeling and Testing the Presence of Autocorrelation in Time Series Models and Residuals: The Example of the Initial Trajectory of COVID-19 Cases in the US

David Caudillo*, Matthew Bennett* & Abu Yahya**
Department of Mechanical and Civil Engineering, Department of Mathematical Sciences
Supervisor: Alessandro Maria Selvitella

Abstract

The purpose of this project is to develop an understanding of how to model and verify the models accuracy for the time evolution of the initial COVID-19 cases that have impacted the United States via statistical methods. The weeks that our group will analyze are from January 23rd, 2020 to February 12th, 2020. We will consider a time series model with exponential deterministic trend and $AR(1)$ error term in the log-scale to describe the initial evolution of COVID-19. Furthermore, we will develop a linear regression model of the data and implement a Ljung-Box test, to determine whether our model exhibits or does not exhibit a lack of fit. These tests will be performed via RStudio. The data used for this project was obtained from the Center of Disease Control and Prevention (CDC) website.

Keywords: COVID-19 , Statistical Methods, Linear Regression Model, Ljung-Box Test

Introduction

The World Health Organization (WHO) declared a pandemic on March 11, 2020 due to the newly discovered disease COVID-19. Our goal is to primarily understand the spread of Covid-19 with the progression of time. We will analyze the data, plot it and identify any patterns. We will use a Time series analysis. Time-series models can have multiple covariates;we set Y(dependent) or the number of cases and the X(explanatory) as the time/period. Our Primary focus is the 3 weeks of January. Seeing an exponential trend on the cases, we are determined to shift our focus on that spike. Using the linear relationship of the data we strive to explain and compare the data to the predicted values given by the program.

Linear Regression Model

As seen in Figure 1, the Date Vs New Cases graph is showing how the as time progresses within the first three weeks the number of new cases is still increasing at a linear rate. It can also be seen that there are two outliers on the graph where the number of new cases skyrocketed for the said days. This is almost the case with Figure 2 where the graph is showing the Date Vs. 7-Day Moving Average. The graph is more consistent in the beginning than the Date Vs. New Cases graph, but does progressively start increasing at a linear rate. Instead of having two outliers like in Figure 1, Figure 2 only has one outlier that appears during the linear trend of the graph. With both figures there is a regression line that is best fitted to the data points on the graph. This regression line provides an equation were if you were to choose a specific date it would provide you with either the number of new cases or the 7-day moving average on that day.

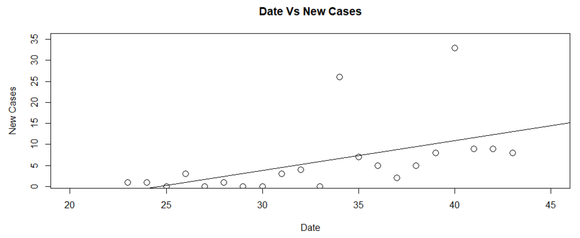

Figure 1: Date Vs New Cases Plot

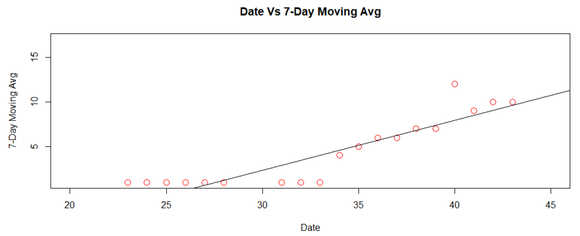

Figure 2: Date Vs 7-Day Plot

Methods

- Linear Regression Model, (1)
- Ljung-Box Test, (2)

$$E[y] = \beta_0 + \beta_1 x \tag{1}$$

$$Q(m) = n(n+2)\sum_{j=1}^{m} \frac{r_j^2}{n-j} \tag{2}$$

The Ljung-Box Test Hypothesis: If p-value ≤ 0.05

- Reject Null Hypothesis, H_O: Model does exhibit a lack of fit
- Accept Alternative Hypothesis, H_A: Model does not exhibit a significant lack of fit

Conclusion

Based on the output results of the Ljung-Box Test Analysis via RStudio, the p-value of the linear regression model is ≤ 0.05. Thus, we can reject the null hypothesis (H_O) and conclude that there is sufficient evidence to say that our linear regression model does not show a significant lack of fit. The raw script output can be seen in Figure 5.

```
>
> arma10$residuals %>%
+   Box.test(.,
+            lag = 1,
+            type = "Ljung-Box",
+            fitdf = 1)

        Box-Ljung test

data:  .
X-squared = 0.025064, df = 0, p-value < 2.2e-16
```

Figure 5: Raw script result of p-value and x^2

Ljung-Box Test Analysis

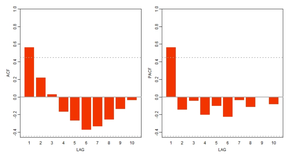

Figure 3: ACF vs PACF Residuals

As seen In Figure 3, there appears to be a spike at lag of 1 on both the ACF and PACF that crosses the confidence intervals, and much lower spike on the subsequent lags. Therefore an AR(1) model seems to be the best fit.

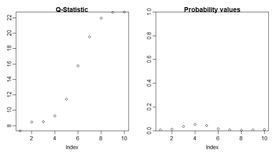

Figure 4: Q-Statistics and Probability Values

Figure 4 shows the Q-stat and P value results. The P value shows a strong auto correlation, while the Q-stat shows that there could be extreme values.

References

[1] Edward Frees. *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press, 2010.

[2] K Kotzé. Tutorial: Simulating and estimating arma models. 2021.

Acknowledgment

Our group would like to express our appreciation by thanking Dr.Selvitella for providing us with an opportunity to further explore data sciences and for his guidance throughout the time of this project.