5-6-2021

Time Series Forecasting of CoViD-19 Deaths in Massachusetts

Andrew Disher

Time Series Forecasting of CoViD-19 Deaths in Massachusetts

Andrew Disher

Submitted in Partial Completion of the
Requirements for Departmental Honors in Mathematics

Bridgewater State University

May 6, 2021

Dr. Wanchunzi Yu, Thesis Advisor
Dr. Kevin Rion, Committee Member
Dr. Laura K. Gross, Committee Member

Time Series Forecasting of CoViD-19 Deaths in Massachusetts

Andrew Disher* Wanchunzi Yu †

Abstract

The aim of this study was to use data provided by the Department of Public Health in the state of Massachusetts on its online dashboard to produce a time series model to accurately forecast the number of new confirmed deaths that have resulted from the spread of CoViD-19. Multiple different time series models were created, which can be classified as either an Auto-Regressive Integrated Moving Average (ARIMA) model or a Regression Model with ARIMA Errors. Two ARIMA models were created to provide a baseline forecasting performance for comparison with the Regression Model with ARIMA Errors, which used the number of CoViD-19 patients in hospitals as an exogenous variable to help make forecasts. These models were successfully constructed, passed all diagnostic tests and, after comparing the models' one week forecasts with a variety of forecast error measures, the Regression Model with ARIMA Errors was found to be a superior method to forecast new confirmed deaths of CoViD-19 in Massachusetts.

Key Words: Time Series Forecasting, CoViD-19, ARIMA, Regression Model with ARIMA Errors

1. Introduction

The CoViD-19 pandemic is one of the greatest challenges the world has faced in the twenty first century. As the pandemic has evolved, the efforts of public and private institutions to monitor the spread of the virus evolved as well. The ability to record and store data has improved to a large degree and the ability to use such data to model various trends of the pandemic has increased in turn. Over the course of the pandemic, federal and state governments have relied on data and the interpretations posed by statisticians and public health experts to make informed decisions. Thus, the importance of data driven decision making cannot be understated.

The pandemic has brought many challenges. One such challenge is the ability to understand exactly how the virus has affected the population of the United States. There are many ways to measure this, such as exploring how government imposed lock downs have affected people's economic and psychological well being. Such effects are valuable areas of study, but this study aims to understand how the most severe consequence of the virus has changed over time. This consequence is death, and it is paramount that we be able to use past information on how many people the virus has killed to gain an understanding of how many people we expect to die in the future.

This study will focus on using the data provided by the Massachusetts CoViD-19 dashboard to forecast this terrible consequence of the virus. The state of Massachusetts was selected as the subject since the dashboard data it provides is relatively clean and the number of variables it chose to record data on is plentiful. Additionally, the state dashboard provides multiple types of data in numerous formats, such as daily, weekly, and monthly data. Ultimately, daily data was used to create the models described in the study, but the ability to model data in a daily, weekly, and monthly format is particularly useful since data aggregation can alter the structural trends within time series data sets.

*Bridgewater State University, 121 Summer Street, Bridgewater, MA 02325
†Bridgewater State University, 121 Summer Street, Bridgewater, MA 02325

2. Approaches to Time Series Analysis

Time series analysis and forecasting is a field of statistics that has been closely analyzed for decades due to its many applications. Therefore, there are a variety of approaches available to take when attempting to model a time series. In this paper, we discuss the use of two specific time series models and go on to discuss further modeling options and their advantages and disadvantages. We begin our discussion by briefly reviewing the Auto-regressive Integrated Moving Average (ARIMA) time series model, its extension that allows it to incorporate seasonality (SARIMA), and another extension that takes into account exogenous predictor variables, known as the Linear Regression Model with ARIMA Errors.

2.1 Auto-regressive Integrated Moving Average (ARIMA) Models

2.1.1 ARIMA Model Overview

The ARIMA model is separated into three overarching components: the autoregressive component (AR), the moving average component (MA) and the integrated component (I). The combination of the three components culminates in a time series model that can take information from the series of interest regarding its past values and errors, and can compensate for a lack of stationarity in the data. Withholding the integration component, the model can be represented as follows:

$$Y_t = \delta + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

where Y_t is the time series we wish to model, δ is a constant, ε_t is the error term for the series, $\phi_1, \ldots, \phi_p$ are the auto-regressive weights assigned to the series' past values, and $\theta_1, \ldots, \theta_q$ are the moving average weights assigned to the past errors of the series.

This is known as an ARMA model. The errors are taken to be Gaussian white noise, i.e. they should be normally distributed, have constant variance, and retain no autocorrelation from the modeled series. The assumptions regarding normality and constant variance can be rectified through various transformations, such as the Box-Cox tranformation. This is a variance stabilization transformation that helps to alleviate issues of heteroscedasticity in the data and has often been found to coerce the distribution of the data into one that is better approximated as normal.

It is also possible to rewrite this ARMA model representation in a more compact way using the back shift operator B as follows:

$$\Phi(B) Y_t = \delta + \Theta(B) \varepsilon_t$$

where

$$\Phi(B) = 1 - \phi_1 B - \ldots - \phi_p B$$

and

$$\Theta(B) = 1 - \theta_1 B - \ldots - \theta_q B.$$

In some situations this model may be adequate, however there is often the issue of nonstationarity to consider when modeling time series data. Stationarity is defined as a property of a time series that requires the series to exhibit a constant mean and constant variance. This assumption is vital to the ARMA model yet is often not satisfied when working with data. To compensate for this lack of stationarity, it is common to introduce another transformation to the series known as differencing. To difference a series, we simply take each of the series values and subtract from them the value that previously occurred within the series. This operation is represented as the expression

$$\Delta Y_t = (1 - B)Y_t = Y_t - Y_{t-1}$$

and it can be incorporated in the representation of the ARMA model by applying it to the series Y_t such that the new model can be written as

$$\Phi(B)(1 - B)^d Y_t = \delta + \Theta(B)\varepsilon_t$$

where d is the order of differencing. It is not uncommon to difference a series multiple times, however the order of differencing should not exceed two in most situations. The constant term δ serves a much different purpose when a series is differenced. In an ARMA model, the constant is related to the mean of the process whereas in an ARIMA model the constant is used as a deterministic trend component in an otherwise stochastic process. Most of the time the constant is removed from the model, however it can be useful if there is a need to model the drift of a time series. The use of the differencing component is said to integrate the model and the new model is thus referred to as the Auto-regressive Integrated Moving Average (ARIMA) model.

This model has been studied thoroughly in books such as (Wei, 2006) and (Montgomery, Jennings, & Kulahci, 2016) and has been implemented in a variety of situations. It is useful when there is a need to model a time series using nothing but the information contained within the time series past values and errors, and as a result is a relatively flexible model.

2.1.2 Seasonal ARIMA Models

In many situations, an ARIMA model is a sufficient approach to model the information in a time series. However, it is common for a time series to exhibit a cyclical or seasonal pattern over time. Over time, the time series may behave a certain way that is observed at specific points in time relative to, for example, the time of the year. This kind of seasonality is seen often in time series of weather data where average rainfall may be seen to increase during wet seasons unique to the geography the data comes from. It can also be seen frequently in economic data that is usually measured by quarter. In cases such as these, the ARIMA model allows for an extension that is able to capture seasonal changes in the data. This involves incorporating additional auto-regressive and moving average terms that use information from values in the series that occurred during the seasonal period of change. For example, if we were discussing rainfall data in the Northeast region of the United States, we realize that the rainfall generally increases during the months of March, April, and May. When we desire to predict the rainfall during these months, it is advantageous to consider the amount of rain that fell in those months of the previous years.

Seasonality can be represented with the following modified ARIMA model:

$$\Phi(B)\Phi(B^s)(1 - B)^d(1 - B^s)^D Y_t = \Theta(B)\Theta(B^s)\varepsilon_t$$

where the expression $(1 - B^s)^D$ represents the seasonal difference of the series of period s and order D, $\Phi(B^s)$ is the seasonal auto-regressive component, and $\Theta(B^s)$ is the seasonal moving average component. The period s indicates the seasonal pattern in the data, so if we attempt to model quarterly economic data it would likely be necessary to use a period of $s = 4$. It is rare for the order D of the seasonal differencing expression to exceed 1, however it is needed in some situations. The assumptions regarding stationarity, normality, etc. for the ARIMA model and its extension, the Seasonal ARIMA (SARIMA) model, are the same, and the model extensions can also be observed in the works of (Wei, 2006) and (Montgomery et al., 2016).

2.2 Regression Model with ARIMA Errors

Another model that is useful in predicting future values of a time series is known as a Regression Model with ARIMA Errors. It is a multivariate time series model that takes advantage of exogenous predictor variables to help improve forecast accuracy. Before introducing the model, we begin by discussing one of its components, the linear regression model.

2.2.1 Linear Regression

Linear regression is a useful tool that allows statisticians to explore relationships between multiple variables. It is a deterministic model that attempts to explain how changes in its predictors affect its response variable. The model can represented as

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$$

where Y is the response variable, $X_1, \ldots, X_p$ are the p predictor variables, α is a constant, $\beta_1, \ldots, \beta_p$ are the coefficients for the predictor variables, and ε is the error term following a normal distribution with constant variance and the errors are independent and identically distributed.

Typically, the observations for each of the variables involved are taken at a singular point in time, thus eliminating any time dependence they may have. However, if the linear regression model is fit to data that are taken at different points in time, then the residuals of the model fit may exhibit a form of autocorrelation. In this scenario, the model assumptions have been violated since the residuals are no longer independent. Unlike the assumption about normality, which is relatively robust, the assumption about independent observations is a rather important one that, if not met, will result in an inadequate model that will make unreliable predictions.

Therefore, time must be taken into account when using a linear regression model to fit time series data. There are many ways to do this. One of them is to use time as a predictor variable and model a deterministic trend using

$$Y_t = \alpha + \beta_0 t + \varepsilon$$

where t is the time step, whether it be measured in days, months, years, etc. With this model, we are eliminating the time dependent structure that may have been in the residuals. This model can also be extended to capture seasonality within the data using seasonal dummy variables. In addition, exogenous predictor variables can also be included, as well as many other modifications that can handle outliers, lagged dependence, etc.

However, this model includes only a way to capture a deterministic trend and assumes that the trends within the data stay constant. This is different from the stochastic ARIMA models discussed earlier, and does not perform well when the data have a dynamic trend.

2.2.2 Regression Model with ARIMA Errors

Another modification is necessary to allow the regression model to handle changes in trend over time. Instead of including time as a predictor variable, it is sufficient to fit a regression model with an ARMA time series structure within its residuals. This model can be represented as

$$Y_t = \alpha + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_p X_{p,t} + \varepsilon_t,$$

where

$$\Phi(B)\varepsilon_t = \Theta(B)a_t,$$

and a_t is the error term for the modeled errors of the regression model. a_t is normally distributed with constant variance.

The ARMA structure imposed on the residuals is appropriate when differencing the original series Y_t is not required. When Y_t requires differencing to achieve stationarity, it is also necessary to difference the exogenous predictor variables and the error term ε_t. Doing so fits an ARIMA model to the errors instead. Lastly, the previously described extension of ARIMA models that addresses seasonal patterns can also be implemented here, such that our general model, as described in (Hyndman & Athanasopoulos, 2018), is

$$Y_t' = \alpha + \beta_1 X_{1,t}' + \beta_2 X_{2,t}' + \ldots + \beta_p X_{p,t}' + \varepsilon_t'$$

where

$$Y_t' = (1-B)^d(1-B^s)^D Y_t,$$

$$X_{i,t}' = (1-B)^d(1-B^s)^D X_{i,t},$$

$$\varepsilon_t' = (1-B)^d(1-B^s)^D \varepsilon_t.$$

Additionally,

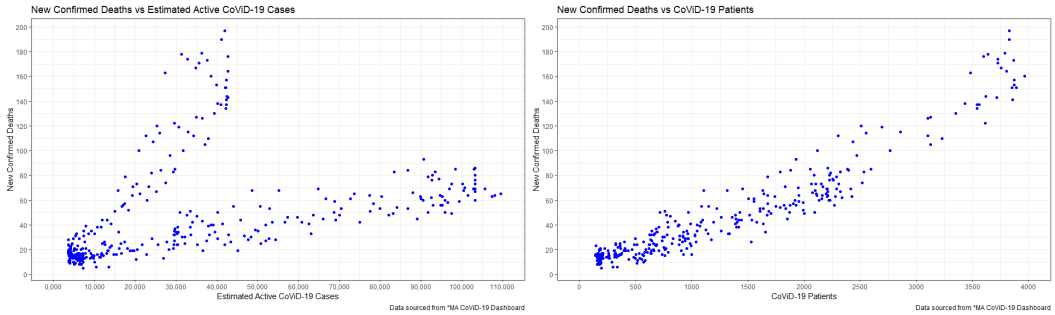
$$\Phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D\varepsilon_t = \Theta(B)\Theta(B^s)a_t.$$

3. Brief Exploratory Analysis

Due to the sense of urgency imposed by the coronavirus pandemic and the the cumulative efforts of the many private, state and federal institutions across the state, the Massachusetts CoViD-19 dashboard has a variety of data available on it. In many cases the data has been recorded on a daily basis and in other cases it has been recorded weekly. It is the former that we will be analyzing in this study and specifically we have chosen the number of deaths that have resulted from the virus to be our response variable. This is because it was valuable to analyze data that correspond to the most severe consequence of the virus.

The modeling process begins with some instigation into the nature of the variables that may be of value in these analyses. Since our response variable Y_t is chosen to be the number of new CoViD-19 deaths in Massachusetts by day, it is necessary to explore the relationships that exist between Y_t and possible explanatory variables. After exploring the many data sets available on the MA dashboard, it was decided that two possibly useful explanatory variables existed. The first possible variable is the number of active CoViD-19 cases by day, which would intuitively make sense because when the number of active cases is large, the more likely it is that a larger number of people will die as a result. The second variable that could explain the number of deaths is the number of people who are currently hospitalized due to the virus.

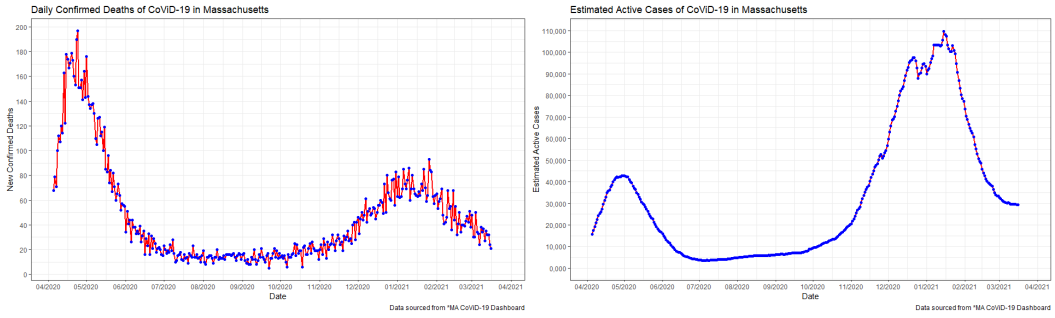
To understand the relationships between these variables, we begin by plotting their pairwise coordinates using scatter plots and calculating their correlations. In Figure 1(b), the scatter plot shows an obvious linear trend with a positive correlation. This indicates that the number of CoViD-19 patients in the MA hospitals could be useful in explaining the number of new deaths in the state, an intuitive conclusion. In Figure 1(a), however, there are two obvious linear patterns that emerge. It appears that the number of of deaths has positive correlation with the number of active cases in both patterns, yet one pattern



(a) (b)

Figure 1: Deaths vs Estimated Active Cases (a) and Deaths vs Patients (b)

seems to suggest a much stronger correlation than the other. This phenomenon occurs because the scatter plots, although useful in exploring relationships, do not incorporate the time component that is vital in time series analysis. The reality of the situation is that the trend changes over time, which can be visualized in Figure 2. In both of these time series, there is a noticeable spike around late March of 2020 and another between January through February of 2021. However, the first spike in cases seemed to correspond with a much larger increase in deaths than the second spike. This is responsible for the two different patterns that occurred in Figure 1(a). A possible explanation for this is that, as the Massachusetts pandemic response evolved to better combat the virus, hospitals were able to administer care more quickly and efficiently, perhaps due to organisational improvements or the increased availability of medical supplies. This reality makes the use of the estimated number of active CoViD-19 cases an unattractive candidate for an explanatory variable, unless a piece-wise time series model was implemented such that we selected two time periods to model separately. This would actually harm any models we wished to create since it would decrease the number of observations to work with.



(a) (b)

Figure 2: New Confirmed Deaths Time Series (a) and Est. Active Cases Time Series (b)

Returning our focus to Figure 1(b), we see that the relationship between the number of CoViD-19 patients in MA hospitals follows one defined linear trend. Comparing the patients time series in Figure 3 to the deaths time series in Figure 2(a) also suggests that the two variables tend to follow each other rather closely. In addition to the scatterplots and the time series graphs already observed, we calculated the correlations between each of the variables. It was found that the correlation between new confirmed deaths and estimated active cases was approximately 0.43, whereas the correlation between deaths and CoViD-19 patients was about .95. For these reasons, we will incorporate the number of CoViD-19

patients into our preliminary regression model as an explanatory variable X_t and omit estimated active number of CoViD-19 cases.

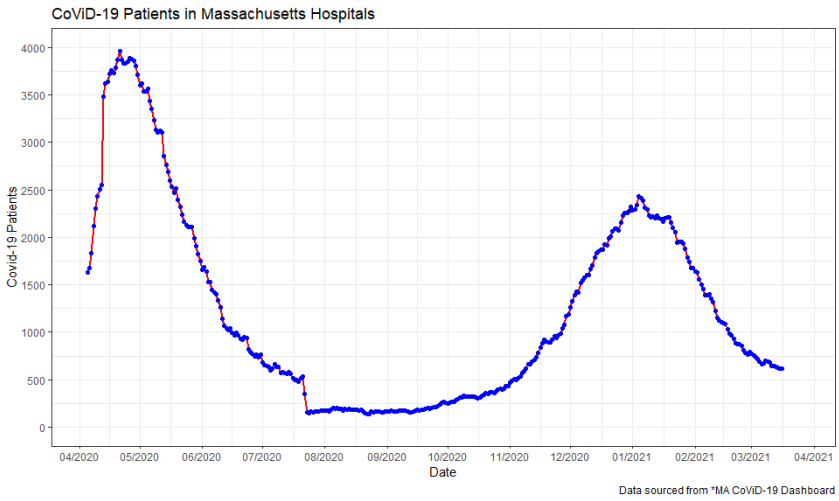


Figure 3: Patients Time Series

4. Regression Model with ARIMA Errors

4.1 Fitting a Linear Regression Model

Now armed with the knowledge to create the models discussed earlier and a general idea of which of the MA CoViD-19 dashboard variables we wish to utilize in our models, it is time to begin the model fitting process. To begin, we will first outline the step by step procedures that were taken to create the regression model with ARIMA errors.

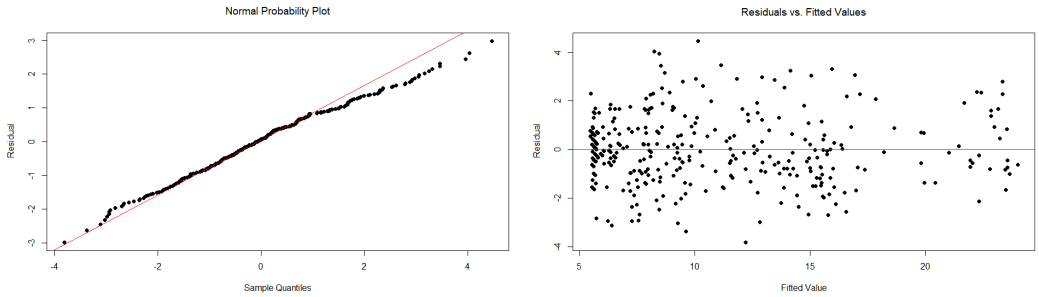
First, a linear regression model was fit using new confirmed CoViD-19 deaths as our response variable Y and current number of patients in MA hospitals due to CoViD-19 as our explanatory variable X. Our regression model will take the form of

$$Y = \alpha + \beta_1 X + \varepsilon.$$

It was found after examining diagnostic plots for this model that the plot of residuals against the fitted values of the model exhibited signs of non-constant variance. Thus, a Box Cox transformation was applied to the response variable Y such that or new model can be represented as

$$\frac{Y^{1/2} - 1}{\frac{1}{2}} = \alpha + \beta_1 X + \varepsilon.$$

Further diagnostic plots were produced and they can be seen in Figure 4. There appears to be constant variance now according to the plot of residuals against fitted values. However, the results of the normal probability plot indicate the distribution of the residuals may not be entirely normal. The heavy tails are cause for concern, however the normality issue will be later improved with further modifications to the model. The coefficient estimate for β_1 was tested for significance using a t-test, and the resulting P-value was close to zero, which is far less than the .05 threshold used. Additionally, the t-statistic was greater than 65. Both provide strong evidence that the estimate is significantly different from zero and that it is appropriate to include the current number of CoViD-19 patients in MA hospitals as an explanatory variable in the model.



(a) (b)

Figure 4: Diagnostic Plots for Linear Regression Model

Now, the residuals of a linear regression model are also assumed to be independent of one another. As discussed earlier, if a regression model is fit to data that were all collected at the same point in time, then it would be safe to conclude that this assumption holds. However, we have just fit linear regression model to two time series variables. Therein arises the need for further inspection of the residuals by plotting the residuals' sample autocorrelation function (ACF).

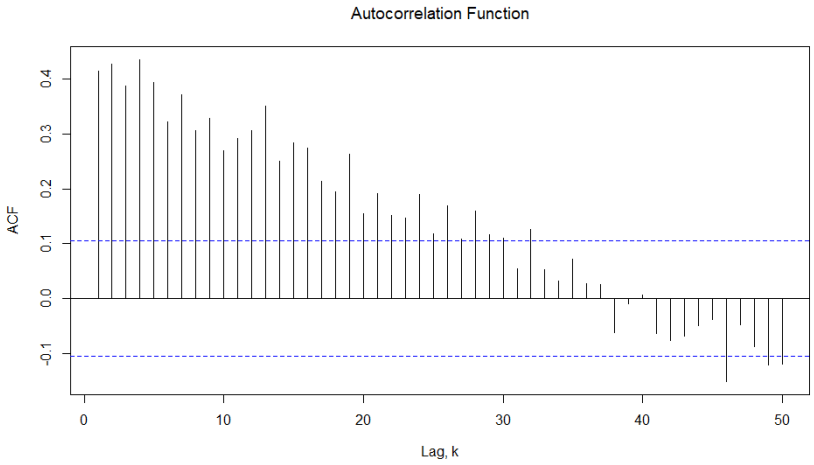


Figure 5: ACF of Residuals from Linear Regression Model

The ACF in Figure 5 shows a slowly decaying trend. This means that current values of the series are highly correlated with the past values of the series, i.e. the residuals of the linear regression model created are highly dependent upon one another. This is proof that the assumption about independence that the linear regression model demands is violated. Therefore, we must further modify our model to correct this issue.

4.2 Modeling the Residuals of the Linear Regression Model

In order to correct for the dependency of the residuals, we employ the methods discussed in section 2.2.2 regarding fitting an ARIMA model to the residuals of the regression model. To do so, we first plot the residuals of the linear regression model against time in Figure 6. The time series of the residuals (ε_t) exhibits a nonstationary trend, since the mean of the series is changing over time. The first order difference was applied and the newly differenced series can be seen in Figure 7.

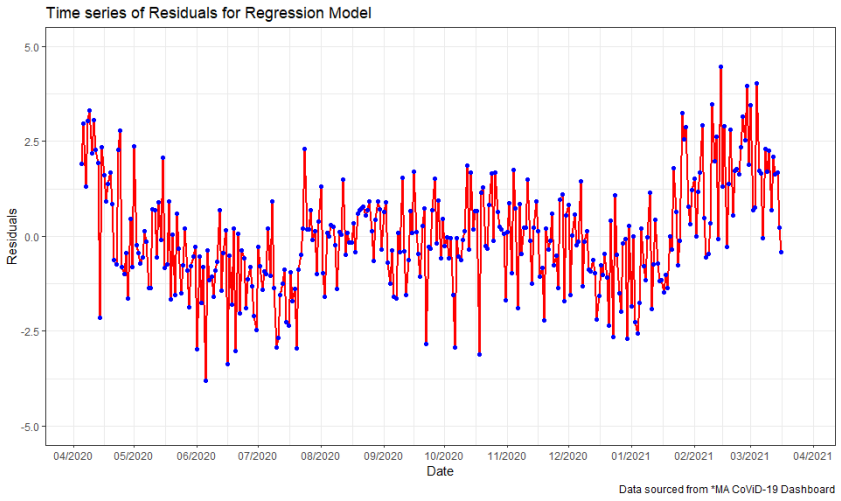


Figure 6: Time Series of ε_t

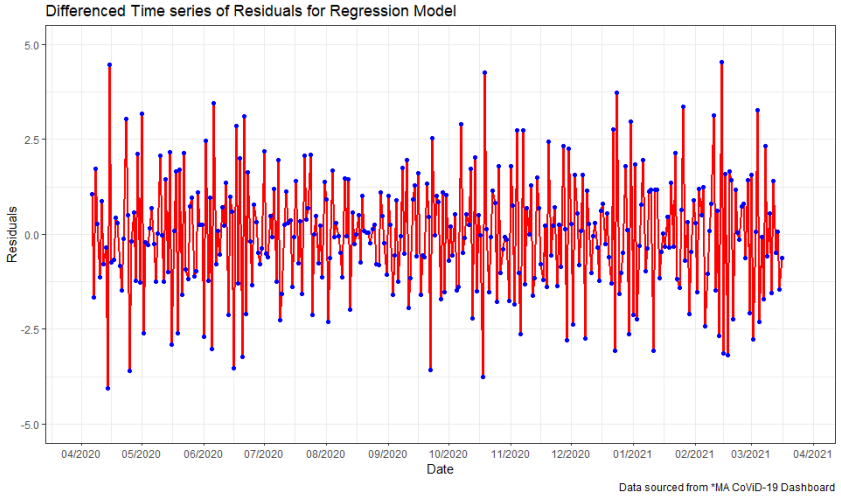


Figure 7: Time Series of ε_t with First Order Differencing

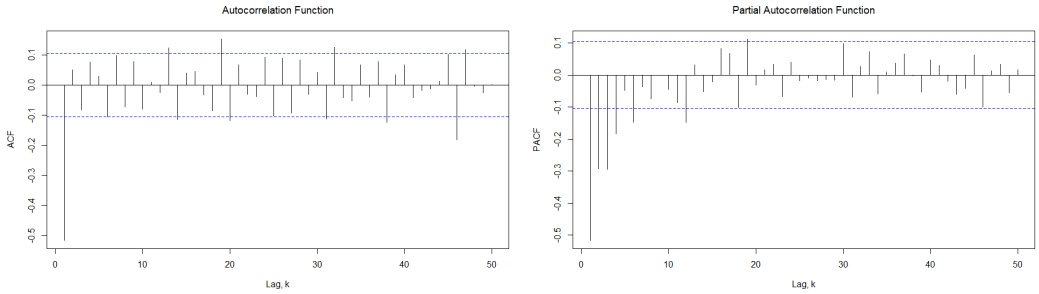
The newly differenced time series of the residuals fluctuates around a constant mean and appears to retain a relatively constant variance as well. To ensure that no further differencing or variance stabilization operations are necessary, we inspect the sample autocorrelation and partial autocorrelation functions. The ACF and PACF in Figure 8 appear to indicate stationarity since the slowly decaying trend has been removed. Furthermore, the two plots indicate that there remains some amount of autocorrelation leftover, which must be addressed by introducing either auto-regressive terms, moving average terms, or a combination of the two via an ARIMA model.

The exponential decay evident in the PACF and the singular significant autocorrelation in the ACF at lag 1 both indicate that an ARIMA(0,1,1) model is an appropriate fit the the data. The model can be represented mathematically as

$$(1 - B)\varepsilon_t = \theta_1(1 - B)\varepsilon_{t-1} + a_t$$

where a_t is the error term that follows a normal distribution with constant variance.

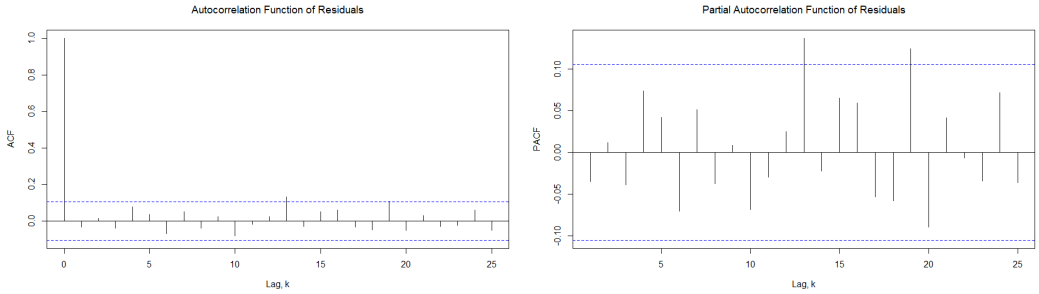
The resulting model's residual ACF and PACF can be seen in Figure 9, which indicates that all of the autocorrelation has been removed from the series, aside from a few that are



(a) (b)

Figure 8: ACF and PACF of the Differenced ε_t Series

borderline significant. These are most likely due to the effects of white noise. In Figure 10, the ARIMA(0,1,1) model diagnostics can be seen, which suggest the assumptions regarding normality, constant variance, etc. are valid.



(a) (b)

Figure 9: Residual ACF and PACF of ARIMA(0,1,1) Model

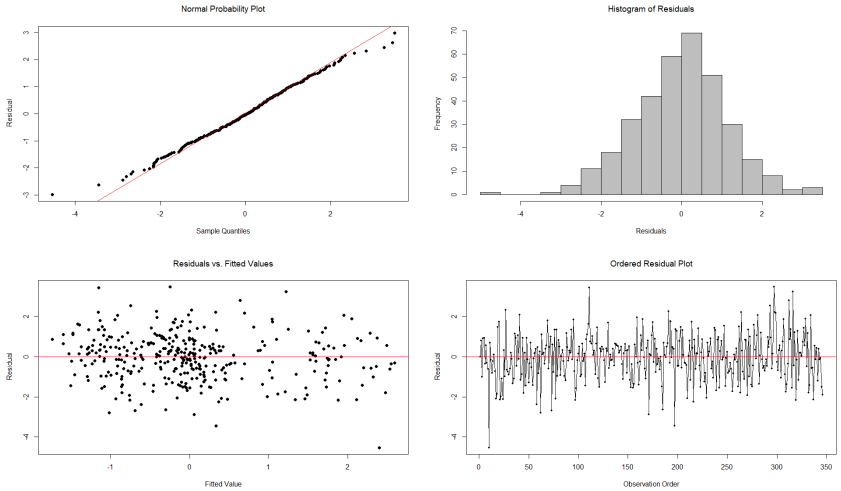


Figure 10: Diagnostic Plots for ARIMA(0,1,1) Model

The Ljung-Box statistics were calculated for each of the model's residual autocorrelations for the first 25 lags and their corresponding p-values can be seen in Figure 11. The

results of this portmanteau test further suggest that the ARIMA(0,1,1) model is an appropriate fit. The estimate of the moving average parameter θ_1 was tested for significance with a t-test and the resulting P-value was close to zero. The calculated t-statistic was less than -29, further suggesting that the use of the moving average term is appropriate to model the data.

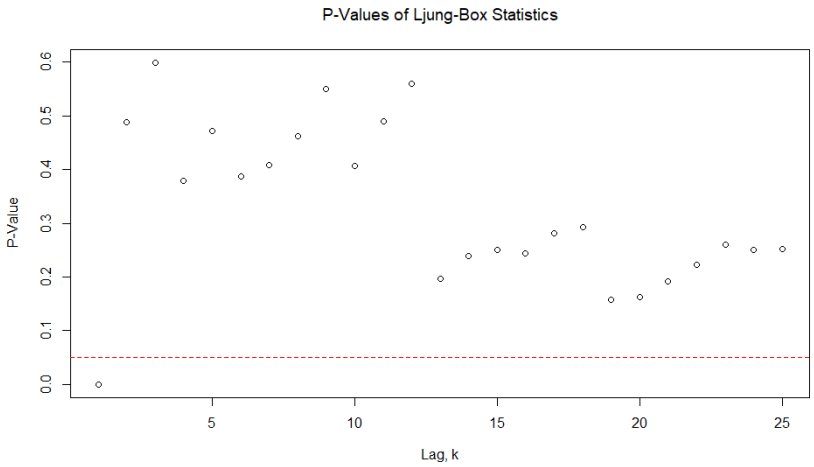


Figure 11: Ljung-Box P-Values for ARIMA(0,1,1) Model

4.3 Fitting the Regression Model with ARIMA Errors

To summarize the findings so far, we have fit a linear regression model with the new confirmed CoViD-19 deaths as the response variable and current CoViD-19 patients in the hospital as the explanatory variable such that

$$\frac{Y^{1/2}-1}{\frac{1}{2}} = \alpha + \beta_1 X + \varepsilon.$$

However, the linear regression model above failed to consider the time component that characterizes the two time series variables included. It was determined in section 4.2 that an ARIMA(0,1,1) model would be appropriate to model the residuals of the linear regression model created in section 4.1. In doing so, the previous issue of dependence between the linear regression model's residuals has been addressed. To reflect the use of the Box Cox transformation that was applied to Y we now define

$$Y' = \frac{Y^{1/2}-1}{\frac{1}{2}}.$$

Consequently, the new model has now become

$$(1-B)Y_t' = \beta_1(1-B)X_t + (1-B)\varepsilon_t$$

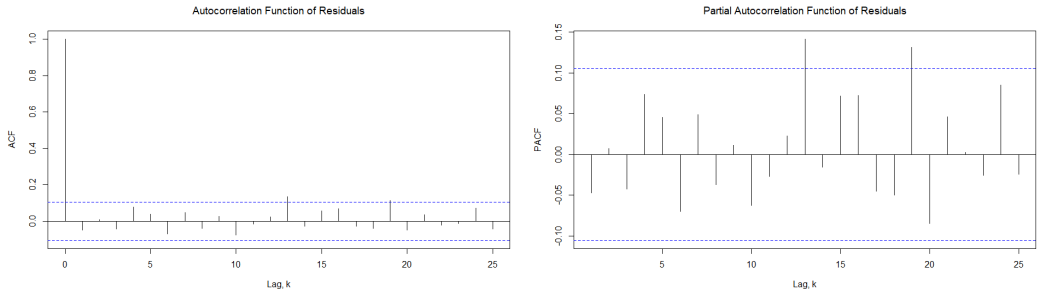
where $(1-B)\varepsilon_t = \theta_1(1-B)\varepsilon_{t-1} + a_t$ and a_t is the error term that follows a normal distribution with constant variance.

This is the form that the regression model with ARIMA errors will take. After fitting the model, it is necessary once again to check the parameter estimates for significance and to observe the behavior in the diagnostic plots. The results in Table 1 suggest that the

parameter estimates are highly significant at the .05 level and that they are appropriate for the model.

Table 1: Regression Model with ARIMA Errors Parameter Estimates

Parameter	Estimate	Std. Error	t-statistic	P-value
θ_1	-0.8437	0.0268	-31.490	$<$ 2.2e-16
β_1	0.0041	0.0003	15.531	$<$ 2.2e-16



(a) (b)

Figure 12: Residual ACF and PACF of Regression Model with ARIMA Errors

The normal probability plot and the histogram of residuals in Figure 13 suggest that it is possible that the distribution of the residuals can be approximated by a normal distribution. However, the normal probability plot appears to have some extremely heavy tails. The assumption regarding normality in regression models is accepted to be robust, however we saw that it was necessary to conduct a formal test to validate the assumption. The Shapiro-Wilk test for normality was applied to the residuals from the model, and the resulting P-value was .5042, which lies comfortably above the .05 threshold used. The residuals plotted against the fitted values of the models and the ordered residual plots suggest that our constant variance assumption also holds. Figure 14 shows that the Ljung-Box test for serial correlation in the residuals produces P-values well above the .05 threshold for each of the first 25 lags. These diagnostics, as well as the residual ACF and PACF in Figure 12, show that the model chosen properly fits the data and that no assumptions made were violated. Therefore, our final model can be stated as

$$(1-B)Y_t' = .0041(1-B)X_t + (1-B)\varepsilon_t$$

where $(1-B)\varepsilon_t = -.8437(1-B)\varepsilon_{t-1} + a_t$.

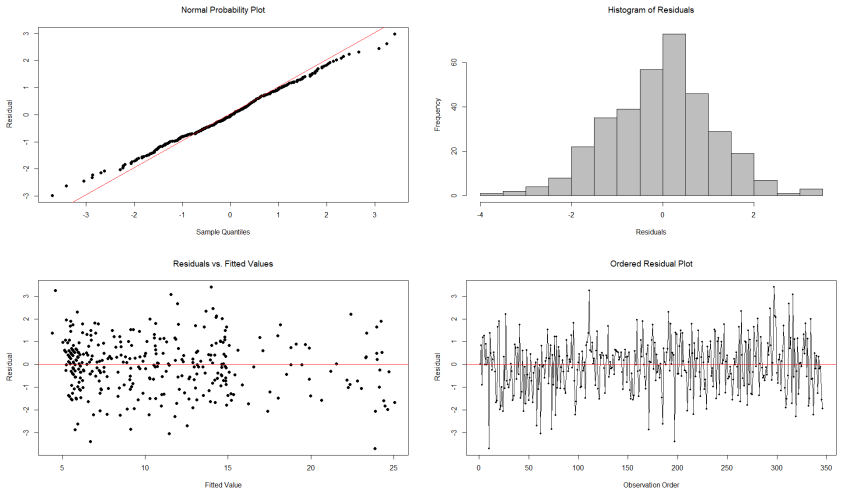


Figure 13: Diagnostic Plots for Regression Model with ARIMA Errors

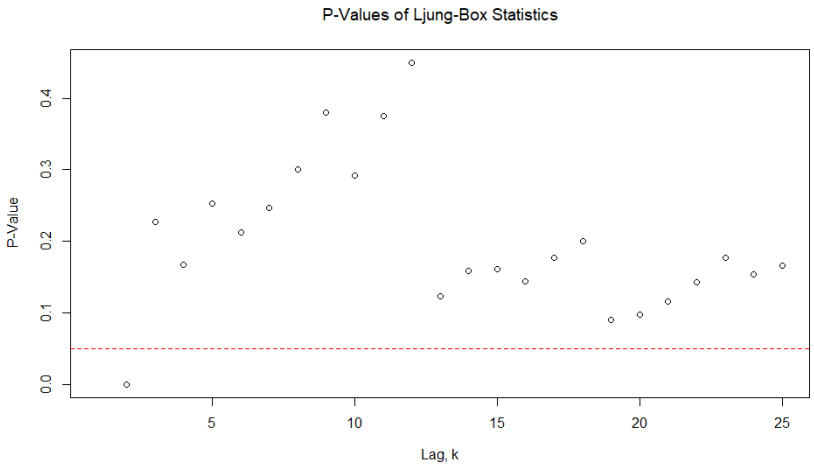


Figure 14: Ljung-Box P-Values for Regression Model with ARIMA Errors

5. Modeling Deaths with ARIMA For Comparison

The regression model with ARIMA errors appeared to be a suitable method for fitting the new confirmed deaths data. However, it required the use of an exogenous variable, the number of people currently hospitalized due to CoViD-19. Acknowledging that the data for both variables was readily available through the data publicly released on the Massachusetts CoViD-19 dashboard, it might be advantageous to model the deaths using the bare bones ARIMA model structure. If the plain ARIMA model outperforms the regression model in terms of forecasting, then there would be no need to go to such lengths to create a more complicated model that possesses fewer degrees of freedom. As such, we proceeded to create an ARIMA model for comparison.

The time series of the number of new confirmed CoViD-19 deaths in Massachusetts from Figure 2(a) was determined to exhibit signs of non-constant variance and a Box-Cox transformation was applied such that the deaths series Y_t' becomes

$$Y_t' = \frac{Y_t^{1/2} - 1}{\frac{1}{2}}.$$

The transformed series can be seen plotted in Figure 15.

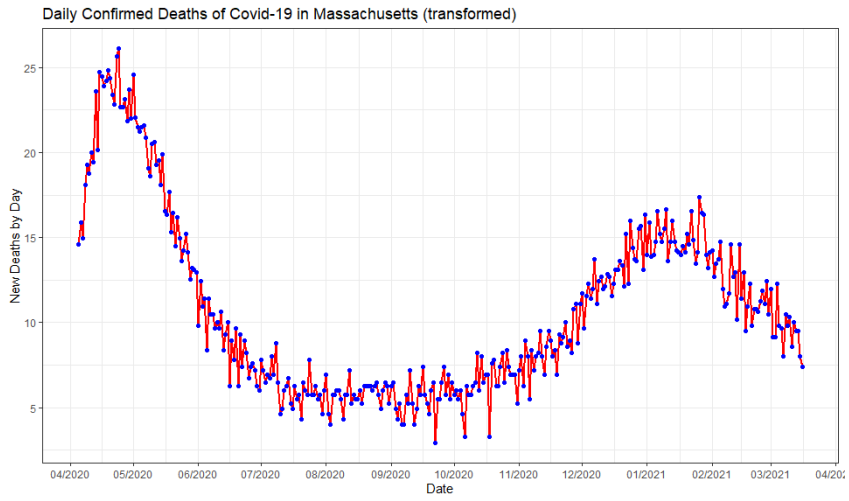


Figure 15: New Confirmed Deaths Time Series Transformed ($\lambda = .5$)

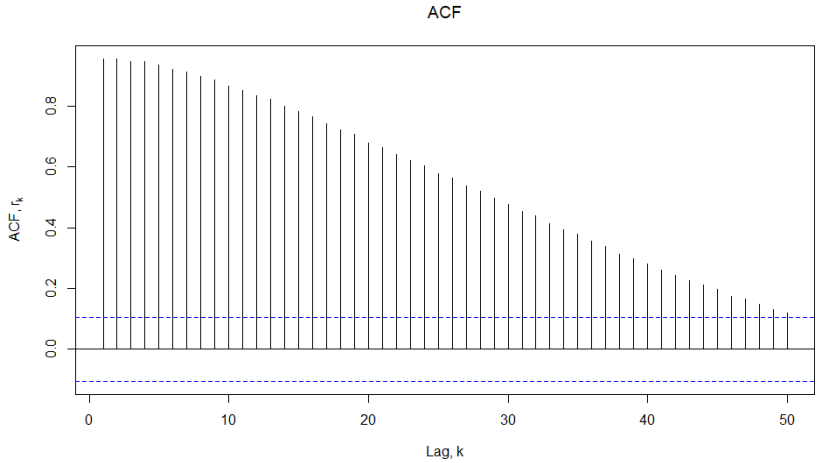


Figure 16: ACF for Transformed Time Series of New Confirmed Deaths

The sample autocorrelation function for the transformed series is plotted in Figure 16, which clearly exhibits a nonstationary trend since the autocorrelations decrease slowly in a persistent fashion. Therefore, the series requires differencing to help alleviate this issue. The differenced series can be seen plotted in Figure 17. The ACF and PACF of the differenced series can be observed in Figure 18.

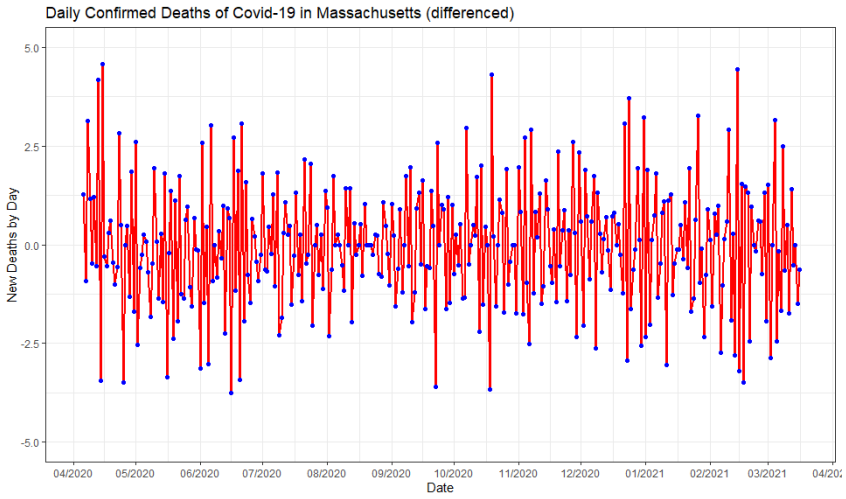
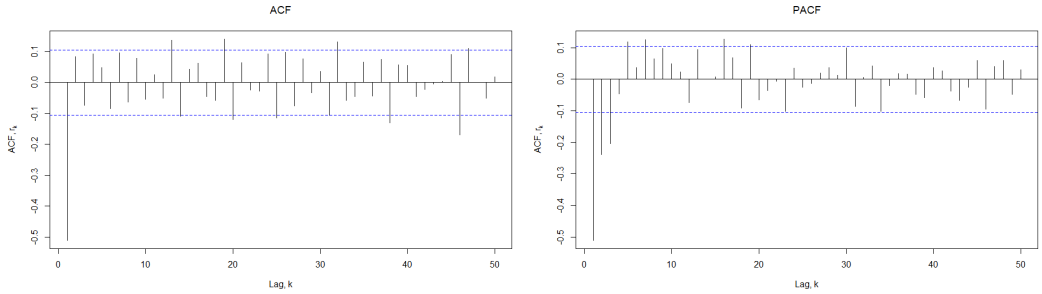


Figure 17: Differenced Time Series of New Confirmed Deaths



(a) (b)

Figure 18: ACF and PACF of Differenced Deaths Time Series

The differenced series is now stationary, judging by the time series plot and its corresponding ACF and PACF. The patterns in the ACF and PACF are similar to those discovered in the same plots for the differenced residuals of the regression model created in Section 4.2. The PACF experiences a pattern of exponential decay and the partial autocorrelations become insignificant after the first 3 lags. There is only one highly significant autocorrelation at lag 1 in the ACF, while the rest are either insignificant or borderline significant. Therefore, an ARIMA(0,1,1) model seems appropriate. The hypothesized model can then be characterized as

$$(1 - B)Y_t' = \varepsilon_t + \theta_1 \varepsilon_{t-1}.$$

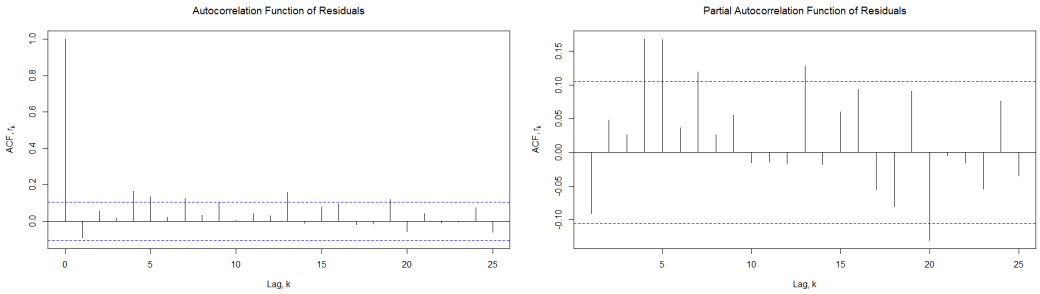
After fitting the model, the parameter estimate for θ_1 was found to be highly significant. This can be seen in Table 2, where the t-statistic was calculated to be -16.553 and the resulting P-value was close to zero.

Table 2: ARIMA(0,1,1) Deaths Model Parameter Estimates

Parameter	Estimate	Std. Error	t-statistic	P-value
θ_1	-0.5877	0.0355	-16.553	$<$ 2.2e-16

Additionally, the residual ACF and PACF plots in Figure 19 indicate that most of the

autocorrelation from the series has been successfully removed and the series reduced to white noise. There are a few autocorrelations that are borderline significant in the residual ACF, however there are two partial autocorrelations in the residual PACF that are cause for concern. However, there are no distinguishable patterns to draw insight from in the two plots to determine if additional terms are necessary. It was found possible that if we added 5 auto-regressive terms to the model, making it an ARIMA(5,1,1) model, that the residual PACF was reduced to white noise. Despite this finding, this act is highly inadvisable since it is risking the danger of overfitting the data. It is generally better to under-model the training data, since it will most likely perform better when forecasting than a model that overfits, even though it performs better on the training data.



(a) (b)

Figure 19: Residual ACF and PACF for ARIMA(0,1,1) Model

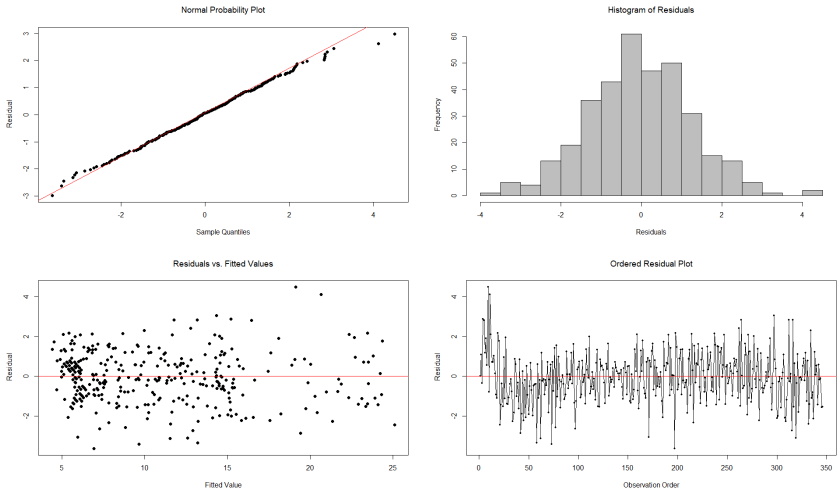


Figure 20: Diagnostic Plots for ARIMA(0,1,1) Model

The diagnostic plots for the ARIMA(0,1,1) model can be seen in Figure 20 where the assumptions regarding normality and constant variance appear to be satisfied. The Ljung-Box statistics were also calculated for the first 25 residual autocorrelations, and their respective P-values can be seen plotted in Figure 21. The Ljung-Box P-values suggest that the model has failed to remove the autocorrelations that were previously seen in the residual ACF and PACF earlier. Although it is not advised, an ARIMA(5,1,1) model was also fit to the deaths time series data in an attempt to rectify this issue. The model parameter estimates can be seen in Table 3, and the diagnostic plots can be found in the Appendix. Aside from

the nearly significant AR(1) parameter estimate, they suggest complete model adequacy, with no assumptions violated.

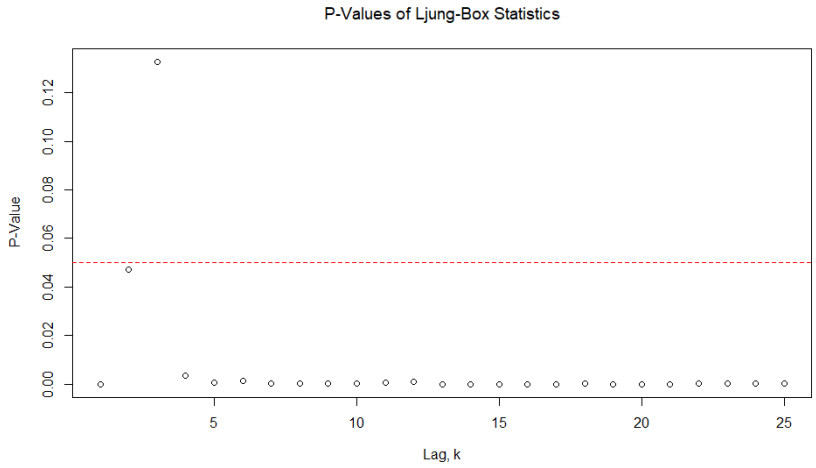


Figure 21: Ljung-Box P-Values for the ARIMA(0,1,1) Model

It should be noted that the Akaike Information Criterion and the Bayesian Information Criterion were both calculated for the two models. The AIC of the more complex ARIMA(5,1,1) model is 1135.98, while the AIC of the ARIMA(0,1,1) model is 1156.79. The ARIMA(5,1,1) has a BIC of 1162.909, while the ARIMA(0,1,1) has a BIC of 1164.487. These values both indicate that the ARIMA(5,1,1) model is a more appropriate fit, although in the next section we will compare the forecast accuracy of the three finalized models side by side.

Table 3: ARIMA(5,1,1) Deaths Model Parameter Estimates

Parameter	Estimate	Std. Error	t-statistic	P-value
θ_1	-0.8609	0.0461	-18.6817	$<$ 2.2e-16
ϕ_1	0.1291	0.0678	1.904	0.05774
ϕ_2	0.2092	0.0579	3.613	0.00035
ϕ_3	0.1301	0.0541	2.403	0.01681
ϕ_4	0.2428	0.0528	4.599	5.979e-06
ϕ_5	0.1744	0.0561	3.111	0.00202

6. Forecasting and Model Comparison

When modeling time series data, there are often two approaches that are usually taken. The first approach to time series modelling is performed with the goal of using the model to explain the data as precisely as possible. The goal of the second approach is to produce a model that yields forecasts that are as accurate as possible. Since the purpose of this study is to produce a model that will forecast the number of CoViD-19 deaths in Massachusetts, the second approach will be used. Therefore, the models will be compared using a variety of forecast error measures that will ultimately decide which model is superior.

In Figure 22, the fitted values of the three models created in the previous sections are overlaid with the training data that they were fit to. The regression model with ARIMA(0,1,1) errors can be seen in red, the ARIMA(0,1,1) model in blue, and the ARIMA(5,1,1) model

in orange. This graph shows that each of the three models perform reasonably well when trying to capture the trends within the training data, which ranges from April 5, 2020 to March 16, 2021. From this angle, the three models are essentially indistinguishable from one another, which suggests that any of the three could be used to accomplish the task of modeling the data.

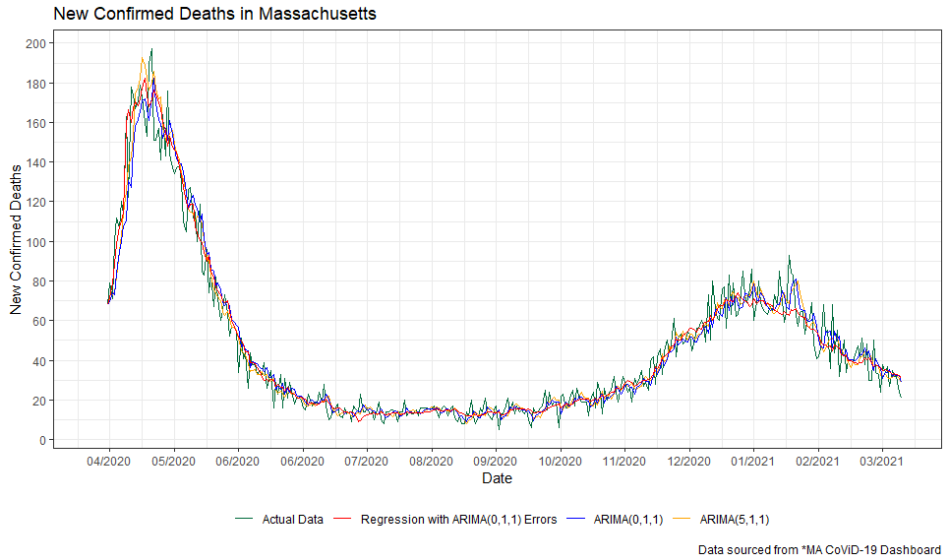


Figure 22: Overlay of Time Series Models and Training Data

As mentioned earlier, however, the goal of the study is to compare the accuracy of the forecasts and decide which model performs best. To this end, the model forecasts were obtained and plotted against one another, as well as the forecast test data, which ranges from March 17, 2021 to March 23, 2021. The forecasts can be seen in Figure 23.

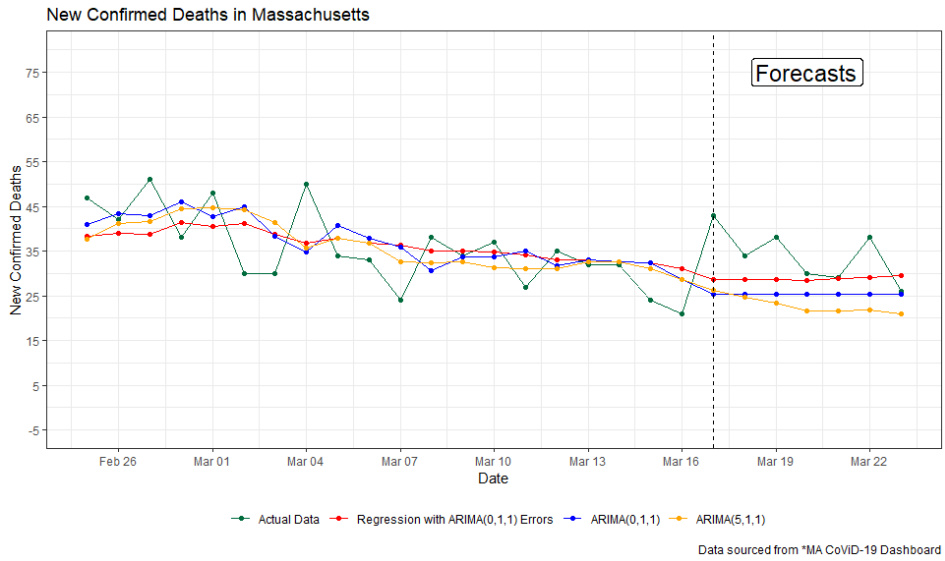


Figure 23: Comparison of Model Forecasts (March 17 - March 23, 2021)

Despite the fact that the three models seem to perform equally well on the training data, the regression model with ARIMA(0,1,1) errors certainly performs the best out of the three models when attempting to forecast the test data for the next 7 days. Additionally, the more complex ARIMA(5,1,1) model actually performs far worse than the ARIMA(0,1,1) model even though both the AIC and BIC values of the former were lower than the those of the latter. The model diagnostics of the ARIMA(5,1,1) model also appeared suggest that the model was superior to the ARIMA(0,1,1) model since it was able to eliminate all residual autocorrelation. This shows how easy it is to overfit a model, a mistake that can easily be made when fitting models in many subfields of statistics, not just time series.

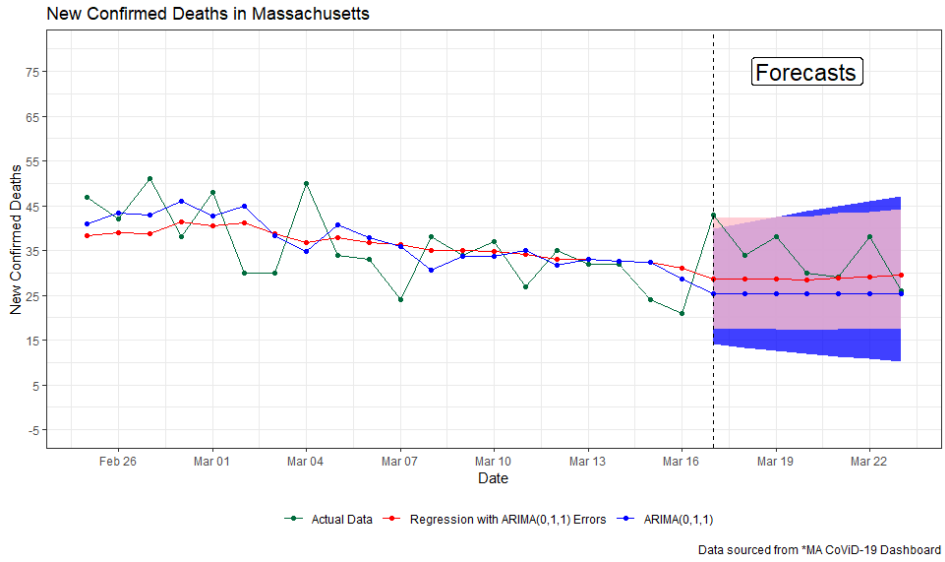


Figure 24: Comparison of Model Forecasts with Prediction Intervals (March 17 - March 23, 2021)

We finish our visual model comparison by removing the complex ARIMA(5,1,1) model and plotting the ARIMA(0,1,1) model, as well as the regression model with ARIMA(0,1,1) errors in Figure 24. The prediction intervals have also been included this time. Again, the conclusion to be made from this visual inspection is that the regression time series model outperforms the traditional ARIMA model not only because the forecast point estimates follow the test data more closely, but also because the prediction intervals are narrower. If the test data was extended to include another week of data, the prediction limits of the ARIMA model would continue to grow wider, making it a less attractive model. Both models fail to detect the data point that occurs on March 17, 2021, however the clear winner in terms of model accuracy is the regression model with ARIMA(0,1,1) errors. This conclusion is made more concrete when viewing the table of forecast error measures in Table 4.

Table 4: Comparison of Out of Sample Forecasts Error Measures

Error Measure	Reg. with ARIMA(0,1,1) Errors	ARIMA(0,1,1)	ARIMA(5,1,1)
RMSE	7.7026	10.2368	11.9078
MAE	6.1526	8.5819	11.0738
MAPE	16.578 %	23.173 %	31.369 %
MASE	0.8585	1.1975	1.5452

The 4 error measures used are the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE) and the Mean Absolute Scaled Error (MASE). Each of these measures are defined such that the model that the lowest value belongs to is deemed superior. The ARIMA(0,1,1) model outperforms the ARIMA(5,1,1) model in every measure provided, but the regression model with ARIMA(0,1,1) errors outperforms both in every measure. Therefore, the best model is the regression model with ARIMA(0,1,1) errors.

7. Post-Analysis Discussion

The results of the model fitting processes described in the previous sections of the paper indicate that, when attempting to model the number of deaths caused by CoViD-19 in Massachusetts, the regression model with ARIMA errors provides a distinct advantage in terms of forecast accuracy. This is due to its ability to incorporate exogenous explanatory variables in its design, which also provides insight on the relationships between the response variable and its predictors. In the context of this study, the number of patients hospitalized by CoViD-19 provides additional information that a traditional univariate ARIMA model does not have access to. This relationship is intuitive to some degree, since it is likely that more people will die when a greater number of people are being hospitalized.

However, sometimes the relationship between a response variable and a potential predictor is not as obvious, or maybe the strength of the relationship is not altogether understood. For example, it was mentioned earlier that the number of active CoViD-19 cases was also related to the number of deaths from the virus, but over time the trajectory and strength of the relationship changed. It was noted that this might have been due to the increasing ability for hospitals to administer care to those sick individuals. This distinct possibility could be explored further if additional data was available for analysis. Examples of data that could be relevant in this situation include records of the medical supplies available to hospitals to treat CoViD-19 patients. Early in the pandemic, there was a supply issue regarding the number of ventilators that each hospital was able to use to treat severe cases of the virus, which might have been a contributing factor that would explain why the first extreme spike in cases in Figure 2 corresponded with a large spike in deaths. The second spike in cases that occurred in 2021 was much larger than the first spike in cases, yet it was accompanied by a spike in deaths that was much smaller than the previous spike in deaths, and this may be because medical supplies, like ventilators, were more plentiful later on in the pandemic as our ability to combat the virus improved.

This is but one example of an additional variable that could help to explain the changes in the number of people dying from CoViD-19, and the identification of further variables should be of great importance in the planning of future efforts to minimize the consequences of infectious disease outbreaks. If such variables can be identified and if another infectious disease were to spread across the world as CoViD-19 has, then informative data dashboards created by state and federal organizations will be able to provide information on such variables that will undoubtedly aid the effort to fight the disease.

A final area of discussion that will conclude this study is the mention of other multivariate time series modeling approaches that could potentially perform better than the regression model with ARIMA errors. This type of model is in fact a special case of the Transfer Function Noise model. The transfer function noise model explores the relationship between exogenous variables, like the patients data used in this study, and the response variable. However, the regression with ARIMA errors model used in this study did not incorporate lagged predictors. It is possible to do so, however the identification of lagged predictors is much easier in a transfer function modeling approach and such transfer func-

tion models tend to produce more accurate forecasts. The form this model takes can be represented as

$$Y_t = v(B)X_t + N_t,$$

where Y_t is a singular response variable, $v(B)$ is the transfer function, X_t is a singular exogenous variable used to predict Y_t and N_t is the error term. The transfer function $v(B)$ can also be defined as

$$v(B) = \frac{\omega(B)B^b}{\delta(B)},$$

where $\omega(B) = \omega_0 - \omega_1 B - \ldots - \omega_s B^s$, $\delta(B) = 1 - \delta_1 B - \ldots - \delta_r B^r$, and b is a a delay parameter that represents the time lag that elapses before the impulse of the input variable produces an effect on the output variable. This is the Transfer Function Noise model as described by (Montgomery et al., 2016) and (Wei, 2006).

Additionally, time series models such as the Vector Auto-Regressive Integrated Moving Average (VARMA) model would prove to be even more advantageous. These models use information regarding multiple time series variables not to forecast a singular response variable, but to forecast each of the variables included. Including multiple variables not only improves forecast accuracy but also provides information on how each of the variables are expected to change over time. In the context of this study, such a model would not only predict the number of new CoViD-19 deaths in Massachusetts but also how many patients we expect there to be in the state's hospitals. The model can be represented as

$$\boldsymbol{\Phi}(B)\boldsymbol{D}(B)\boldsymbol{Y_t} = \boldsymbol{\Theta}(B)\boldsymbol{\varepsilon_t}$$

where $\boldsymbol{\Phi}(B)$ is the auto-regressive matrix polynomial of order p, $\boldsymbol{\Theta}(B)$ is the moving average matrix polynomial of order q, $\boldsymbol{Y_t}$ is the m-dimensional vector process being modeled, $\boldsymbol{\varepsilon_t}$ is the m-dimensional vector of error terms, and $\boldsymbol{D}(B)$ is the differencing operator matrix with m by m dimensions. This is the integrated VARMA model as described by (Tsay, 2014).

The use of these models would provide a plethora of information for decision makers to work with, and it is in this fashion that we aim to continue this study.

Appendix

A.
Model Diagnostics for ARIMA(5,1,1) model in Section 5

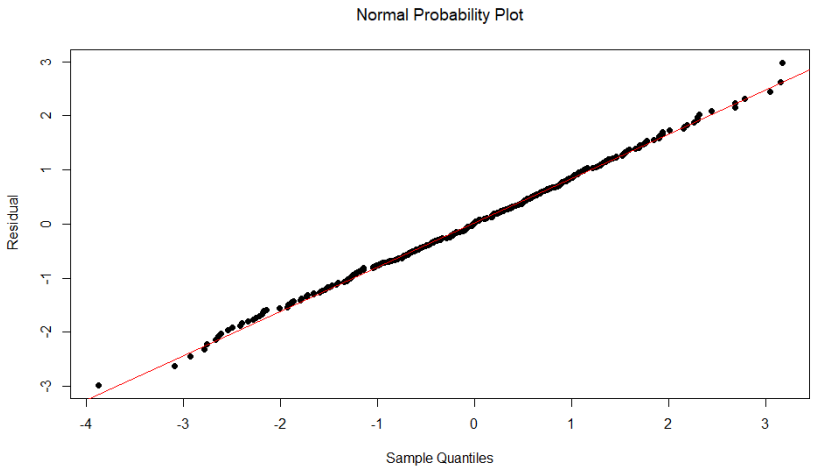


Figure 25: Normal Probability Plot for the ARIMA(5,1,1) Model

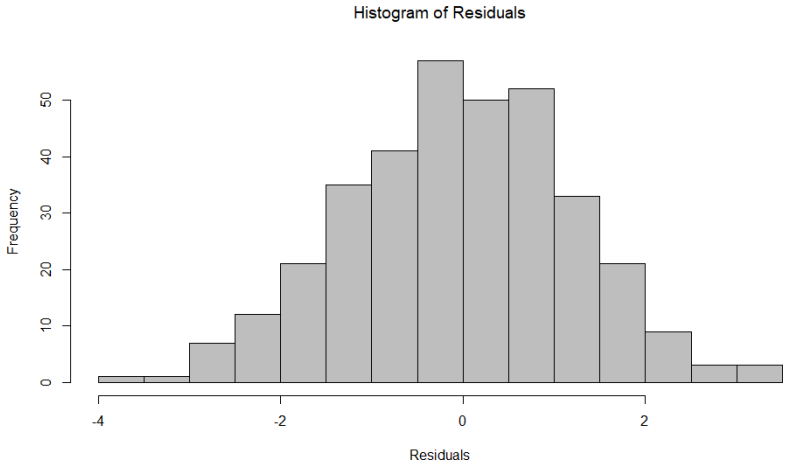


Figure 26: Histogram of Residuals for the ARIMA(5,1,1) Model

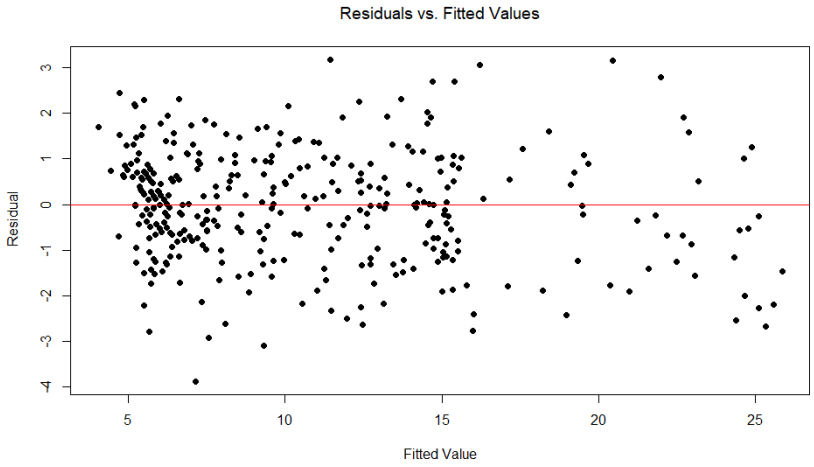


Figure 27: Residuals vs Fitted Values for the ARIMA(5,1,1) Model

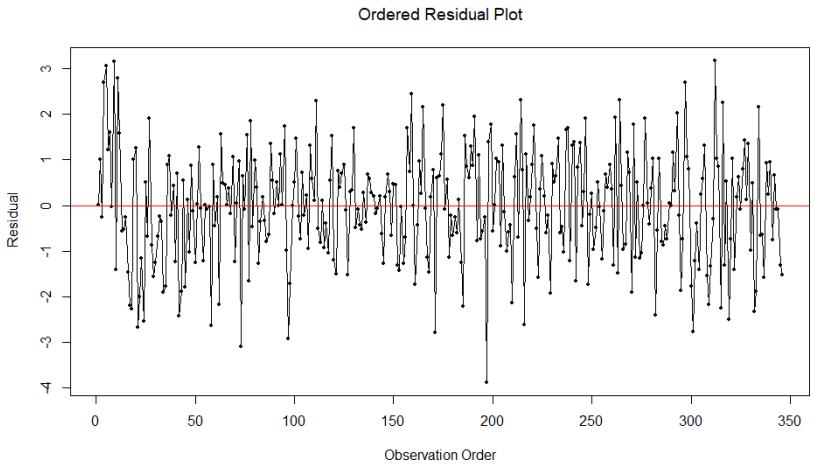
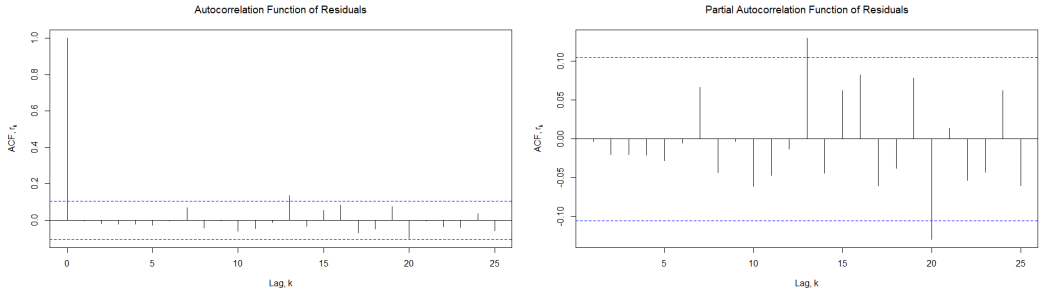


Figure 28: Ordered Residuals for the ARIMA(5,1,1) Model



(a) (b)

Figure 29: Residual ACF and PACF for ARIMA(5,1,1) Model

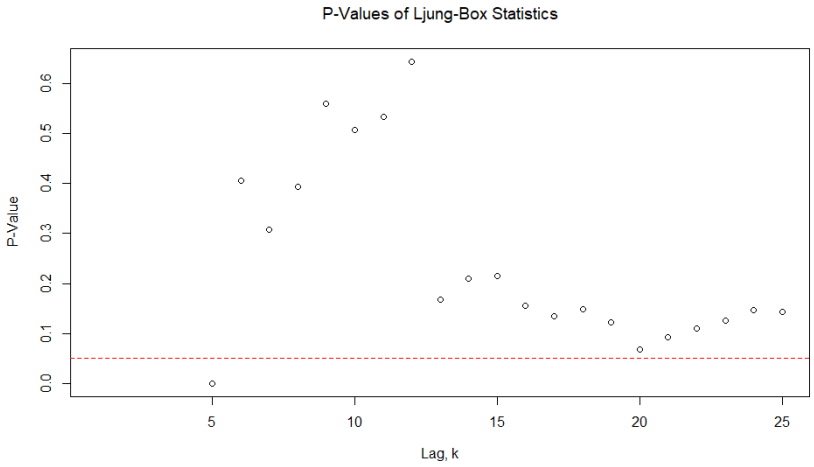


Figure 30: Ljung-Box P-Values for the ARIMA(5,1,1) Model

B.
Inverse Roots of Auto-Regressive and Moving Average Polynomials

These plots ensure that the time series models applied throughout the paper do not contain a unit root, which would indicate that the model is not fully stationary and requires further differencing procedures to make it so. If the inverse roots are within the unit circle, then the model is deemed stationary.

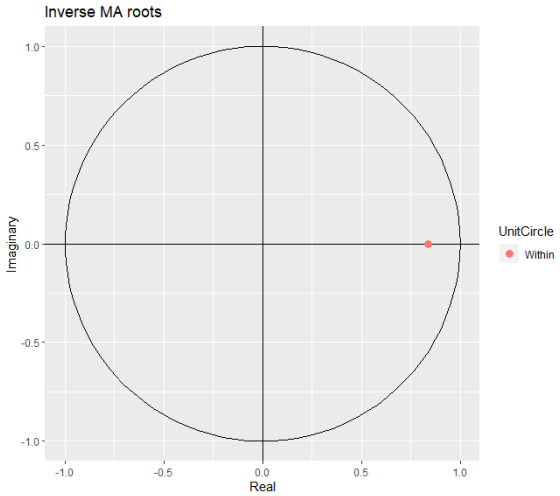


Figure 31: Inverse Roots for ARIMA(0,1,1) Model for the Regression Residuals

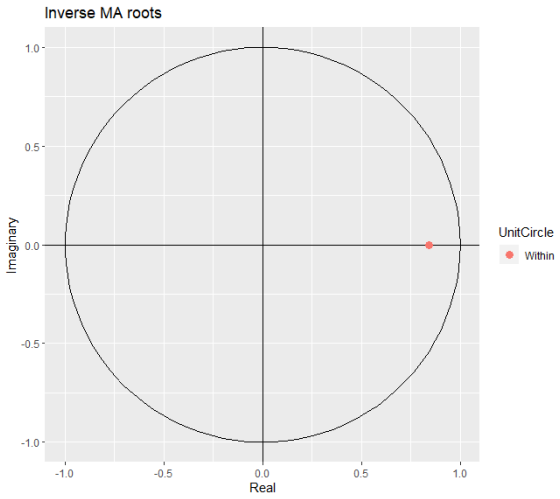


Figure 32: Inverse Roots for Regression Model with ARIMA(0,1,1) Errors

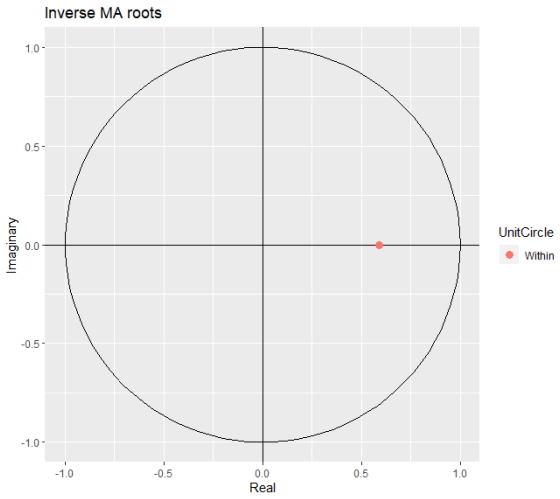


Figure 33: Inverse Roots for ARIMA(0,1,1) Model for Deaths

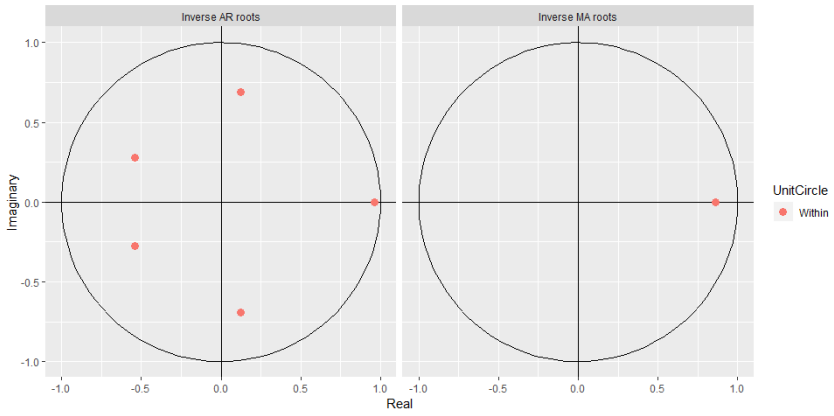


Figure 34: Inverse Roots for ARIMA(5,1,1) Model for Deaths

References

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting principles and practice, 2nd edition*. https://otexts.com/: Otexts.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2016). *Introduction to time series analysis and forecasting, 2nd edition*. Hoboken, NJ United States: John Wiley and Sons, Inc.

Tsay, R. S. (2014). *Multivariate time series analysis with R and financial applications*. Hoboken,NJ United States: John Wiley and Sons, Inc.

Wei, W. S. (2006). *Time series analysis: Univariate and multivariate methods, 2nd edition*. Upper Saddle River, NJ United States: Pearson Education, Inc.