Firearm Time Series Analysis

Part 1: Introduction

In today's society, gun violence continues to be an increasingly pressing issue. In 2017 alone, the Centers for Disease Control and Prevention (CDCP) reported that there are 36,000 gun deaths per year, an average of *100 deaths per day*. CBS news reported that there have already been *346 mass shootings* in the U.S. in 2019 with about a month to go. Furthermore, the CDCP reported that *over half of suicide deaths resulted from a firearm*. This unprecedented and disturbing trend in gun violence has many citizens across the U.S. wondering what could be causing this.

In this project, we will analyze *monthly handgun sales and firearms related deaths* in California from 1980 to 1998. There are 2 columns in the data frame, with the first being gun sales and the second being firearm deaths (per 100,000).

Part 2: Material and Methods

Initial Data Inspection

As stated previously, the data set is split into two columns, gun sales and firearm deaths. There are a total of 227 total rows in the data frame, with each row representing a different month from 1980 to 1998. There are no missing values in the data set. Further inspecting the data, we note the summary statistics of each column below.

	Sales	**Deaths**
Min	59.24	0.700
Q1	106.28	1.140
Median	121.53	1.260
Mean	120.16	1.238
Q3	136.92	1.365
Max	210.43	1.650

We can think of the rows of the data frame as a collection of random variables indexed according to the order they were observed in time, so it is obvious that *we are working with time series data*. Furthermore, we can see that this time series is the only realization of the stochastic process and that there is a dependence among observations, typical characteristics of time series.

Differencing to obtain stationarity

After plotting the time series of both features (***Figure 1***), we can observe a trend in the data and possibly a seasonal effect. This trend indicates that the values of each observation are dependent on time, proving that both series are *not stationary*. This trend and lack of stationarity also indicates that both processes *do not* have mean zero and constant variance. Since the series are not stationary, we are not yet able to perform our typical time series analysis techniques we have developed throughout the course.

Although the processes are not stationary, there are several methods to turn non-stationary time series into stationary ones. One of these methods, differencing, involves subtracting the previous observation from the previous one. Differencing can be used to remove the general trend of the data, as well as any seasonal effects.

The plot in ***Figure 2*** shows the time series after performing *1 time differencing* on both processes. It appears that the first order time differencing eliminated the obvious trend that the models previously possessed. Although this initial trend may have been eliminated, the plots appear to have a seasonal trend lurking. In particular, there appears to be a seasonal trend at approximately every 12 time lags (i.e. 1 year). To now eliminate this seasonal component of the time series, we can now perform a *seasonal difference* on both models.

After applying seasonal differencing at multiples of lag 12, one year, the time series plots are displayed in ***Figure 3***. It appears now that both processes have achieved stationarity. For both models, there appears to be approximately zero mean and approximately constant variance- the two conditions for stationarity.

Determining the models through plots of ACF & PACF

After transforming the data into a stationary form, we can now determine a SARIMA model for the processes by observing the ACF and PACF.

Figure 4 in the appendix shows the ACF and PACF of the gun sales data. Determining a SARIMA model for the data can be broken down into two steps. First, we must determine the non-seasonal component of the model by observing patterns within a 12 lag period. Specifically noting the first period, it appears (approximately) that the **ACF cuts off after lag 2** and the **PACF tails off**. This indicates that the non-seasonal component of the model can be represented with an **MA(2)** process.

Similarly, the seasonal component of the model can be determined by observing the ACF and PACF of the data at every multiple of lag 12. From this, notice that the **ACF tails off** and the **PACF cuts off at lag r = 3**. Thus, we can represent the seasonal component of the model as an **AR(3)** process.

Thus, I obtained this model for sales with the following results:

Coefficient	**Estimate**	**SE**	**t.value**	**p.value**
ma1	-0.2252	0.0675	-3.3374	0.001
ma2	-0.2942	0.0682	-4.3164	0.000
sar1	-0.6154	0.0673	-9.1431	0.000
sar2	-0.5712	0.0668	-8.5468	0.000
sar3	-0.2944	0.0658	-4.4717	0.000

Information Criteria	**Value**
AIC	7.621455
AICc	7.622673
BIC	7.711214

Determining the how to *exactly* model a SARIMA model based on the ACF and PACF alone is a difficult task. As a result, I was interested in testing different models “close” to the model I chose, by alternating values of *p, q, P,* and *Q.* Trying different models with different values of *p, q, P* and *Q*, it can be observed that some of these models actually have *smaller information criteria values*. **However,** each of these models with smaller values for information criteria had at least one coefficient that was not statistically significant. In fact, the original model that we chose was one of the only models that had coefficients that were *all* statistically significant. Therefore, we can be more confident that our model for sales is an appropriate one.

Following a similar procedure to sales, we can determine a SARIMA model for deaths using the plots of the ACF and PACF shown in ***Figure 5***. The non-seasonal part shows the **ACF cuts off at lag 1** and the **PACF tails off,** which is indicative of a **MA(1)** process. Furthermore, the seasonal component shows the **ACF cuts off at lag r = 1** and the **PACF tails off**, which is *also* implying a **MA(1)** should be used.
Thus, I obtained this model for deaths with the following results:

Coefficient	**Estimate**	**SE**	**t.value**	**p.value**
ma1	-0.6286	0.0643	-9.7704	0

sma1	-1.0000	0.1064	-9.3950	0

Information Criteria	**Value**
AIC	-1.856231
AICc	-1.855991
BIC	-1.811351

Similar to the model for sales, I was interested in observing information criteria for models “close” to the one I observed. There were again models “close” to the one given that had information criteria values lower than this model. However, these models all had at least one coefficient that was statistically insignificant. In fact, the model we selected was one of the only models that had coefficients that were all statistically significant. Therefore, we can be even more confident that the model we obtained by observing the ACF and PACF was an appropriate one.

Diagnostics

After fitting a model for each of the series, it is important to check model diagnostics to ensure that the model is actually an appropriate one.

Displayed in ***Figure 6*** are diagnostic plots for sales outputted by R’s sarima( ) function. It appears that the model is a good fit. In particular, there does not appear to be a trend in the standardized residuals. Also, the ACF of the residuals does not appear to have a significant dependence structure. Although the Q-Q plot does show the existence of outliers, the sample quantiles typically tend to the theoretical quantiles, showing that the normality assumption is met. Furthermore, most of the p-values in the Ljung-Box plot appear to be statistically significant.

Similarly, displayed in ***Figure 7*** are diagnostic plots for deaths. It appears that the model selected is also a good fit. In particular, there does not appear to be a trend in the standardized residuals. Also, the ACF of the residuals does not appear to have a significant dependence structure. In contrast to the model fitted for sales, the Q-Q plot does *not* show the existence of outliers, the sample quantiles typically tend to the theoretical quantiles, showing that the normality assumption is met. Furthermore, most of the p-values in the Ljung-Box plot appear to be statistically significant.

After checking the diagnostics for each of the models, we can now be more confident that the individual models are appropriate.

Dependence among sales & deaths

After determining the SARIMA models for sales and deaths individually, we are now interested in finding a potential dependence among them. One way to determine if there is a dependency between sales and deaths, one can examine the **cross correlation function (CCF).** In general, the CCF of any two series can show whether one series "leads" the other, i.e. one process can be predicted by the other.

From the CCF plotted in ***Figure 8***, it appears that there is no truly dominant peak in the CCF at any specific time lag. It is not very clear from the CCF that one time series can predict the other since there is no obvious, significant peak. Therefore, we must obtain another method to try to discover a dependency.

Since we were unable to determine a "leading" relationship between sales and deaths, we now turn to **linear regression.** After running linear regression with sales (*Y*) and deaths (*x*), at first glance it seems like the new model is a good one, with very statistically significant p-values. Although it seems that this model is appropriate, we can observe from the time series plot of the residuals that the *errors are autocorrelated* (**Figure 9**). Therefore, we must now obtain a model that accounts for these autocorrelated errors.

After performing first order differencing and seasonal differencing at every *r = 12* lags, the residuals appear to be stationary (***Figure 10)***, meaning the transformed residuals are no longer autocorrelated. In order to model the behavior of the residuals, we again observe the plots of the ACF and PACF (***Figure 11***). Following a similar procedure to when we determined SARIMA models for sales and deaths, we observe that the transformed residuals can be modelled as an **SARIMA(0,1,1)x(2,1,0)**. Without displaying the output (code in Code Appendix), each coefficient of this model is indeed statistically significant and the diagnostic plots (***Figure 12***).

To obtain a new linear model with uncorrelated errors, we determine a SARIMA model for deaths (*Y*) given the parameters for the SARIMA model for the residuals.

Coefficient	Estimate	SE	t.value	p.value
ma1	-0.7684	0.0574	-13.3889	0.0000
sar1	-0.6591	0.0686	-9.6075	0.0000

sar2	-0.2553	0.0710	-3.5960	0.0004
xreg	**0.0010**	0.0004	2.4484	0.0152

As a result, we can observe that the results are statistically significant. In particular, we see that the regression coefficient the resulting coefficient is 0.0010, and there is no regression coefficient since we are not interested in centering the data.

$$Y_t = 0.001x_t + \varepsilon_t \, , \, \varepsilon_t \sim WN(0, \sigma^2)$$

Part 3: Summary & Conclusions

We began the analysis by first examining each time series individually. After utilizing various differencing techniques, we observed the ACF and PACF of each series to determine a Seasonal ARIMA model for each variable. We were then interested in comparing our candidate model to other models with SARIMA parameters “close” to each candidate. It appeared that there were models that had smaller information criteria values, and could potentially be a more appropriate model than the candidate. However, these models commonly had at least one statistically insignificant SARIMA coefficient, so in fact our model was more appropriate.

After examining each series individually, we now wanted to develop a linear regression model to use the time series of sales to predict the time series of death. The errors of the fitted linear model appeared to be autocorrelated, so we first had to transform the errors to remove the autocorrelation. Following a similar procedure to finding models for sales and deaths, we then determined a SARIMA model for the linear model’s residuals. Finally, we were able to fit another SARIMA model for deaths using the same parameters as the SARIMA model for the residuals to obtain a regression coefficient sales in the linear model.

One of the main goals of this project was to determine if a relationship exists between firearm deaths and gun sales. After careful examination we were able to develop a linear model, $Y_t = 0.001x_t + \varepsilon_t \, , \, \varepsilon_t \sim WN(0, \sigma^2)$, shows there is indeed a positive relationship between firearm deaths and gun sales.

It is worth mentioning that although there is a positive relationship between firearm deaths and gun sales, is not the only relationship that exists. There are many factors that go into the number of firearm deaths, not just gun sales. Thus, politicians should take gun sales into consideration when determining the various gun control laws.

APPENDIX

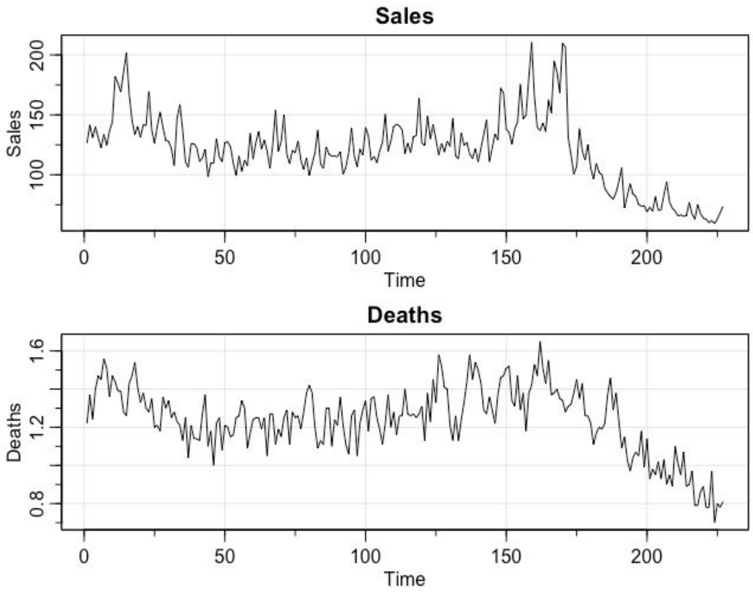

Figure 1: Time series of raw data

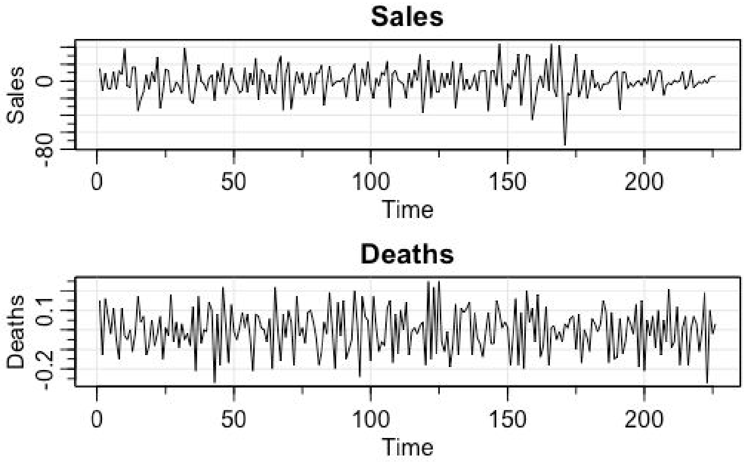

Figure 2: Time series after first-order differencing

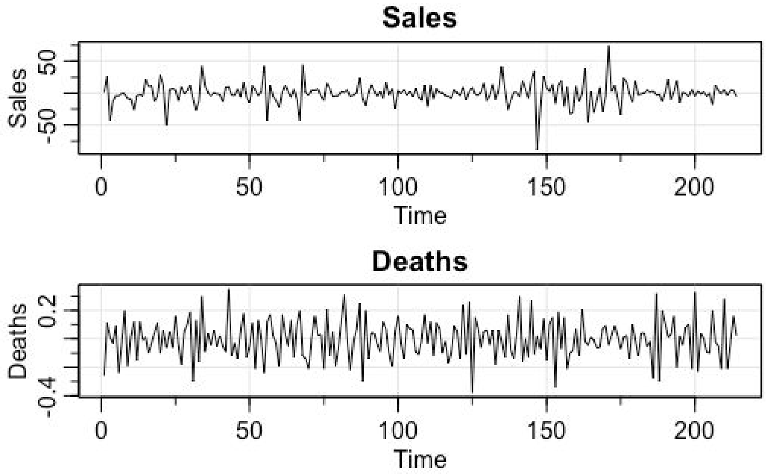

Figure 3: Time series after first order and seasonal differencing

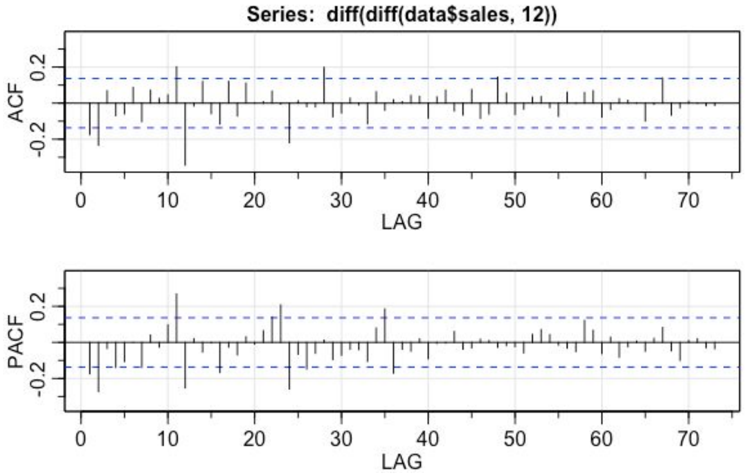

Figure 4: ACF & PACF of Sales

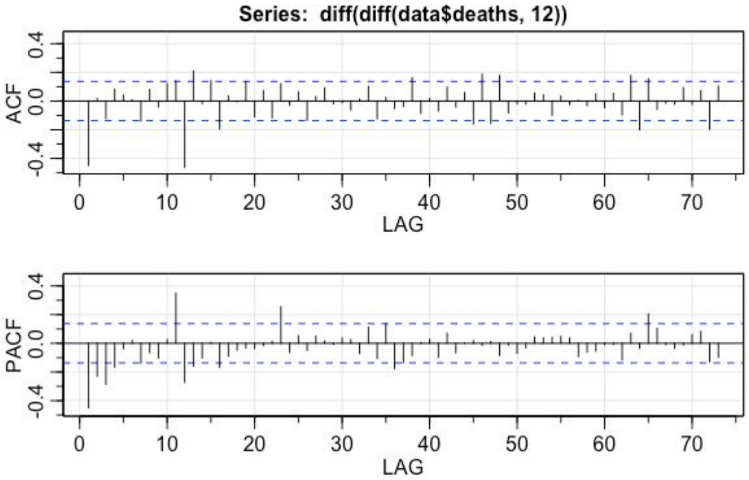

Figure 5: ACF & PACF of Deaths

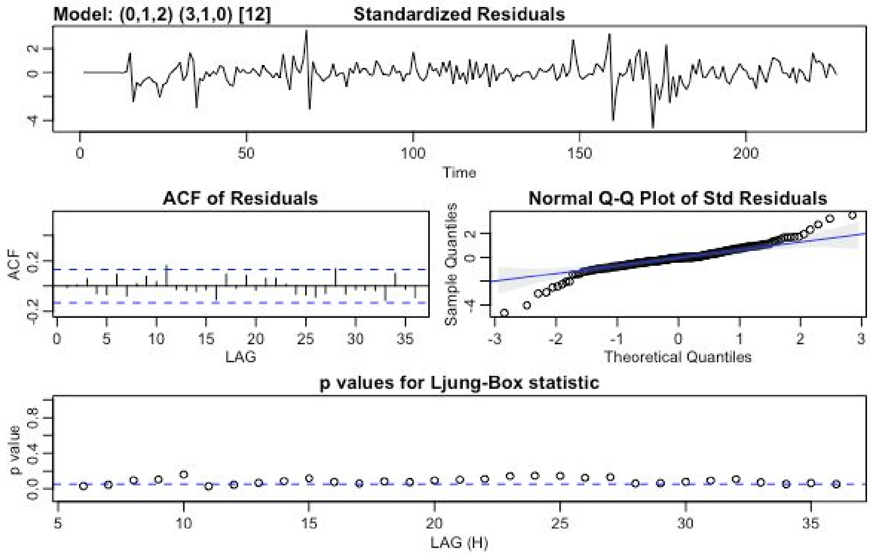

Figure 6: Diagnostic plot for Sales SARIMA model

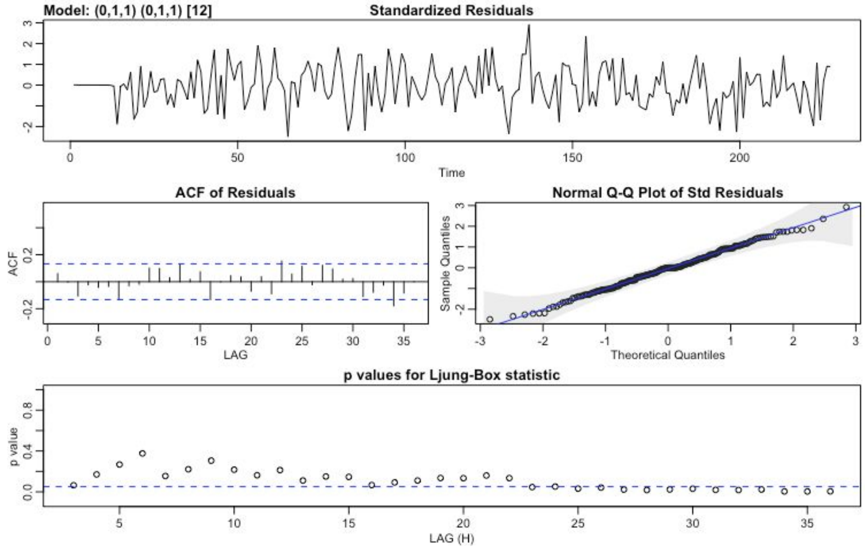

Figure 7: Diagnostic plot for Deaths SARIMA model

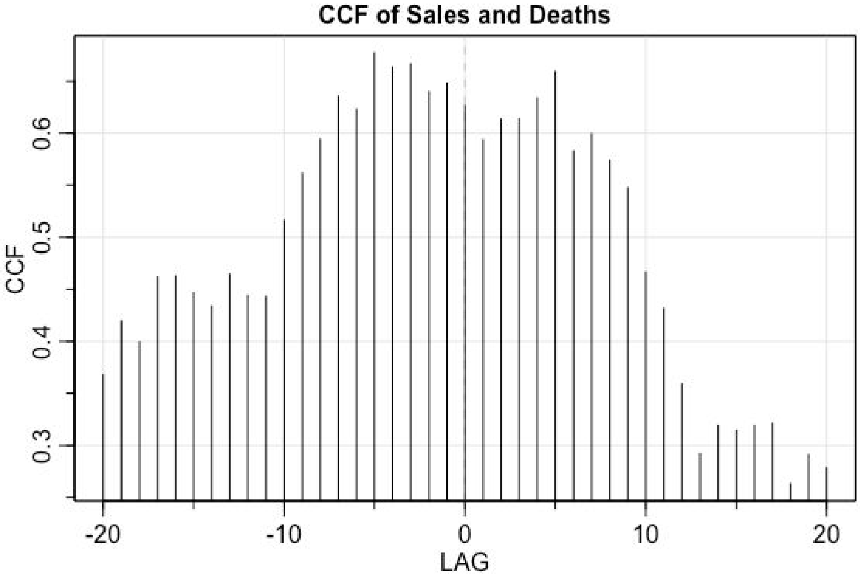

Figure 8: CCF of Sales & Deaths

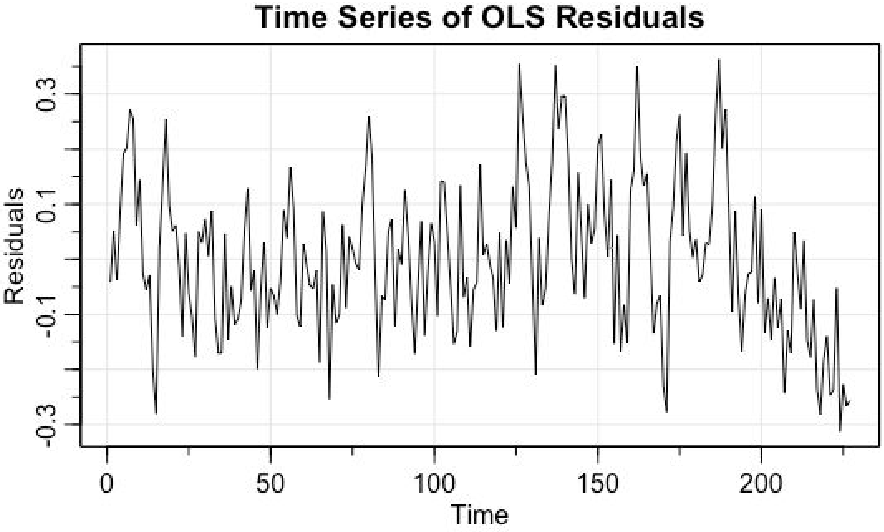

Figure 9: Time series of OLS residuals

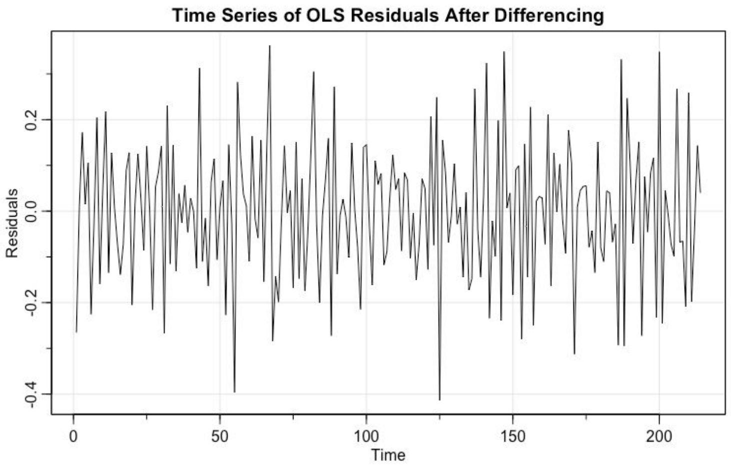

Figure 10: Time series of OLS residuals after differencing

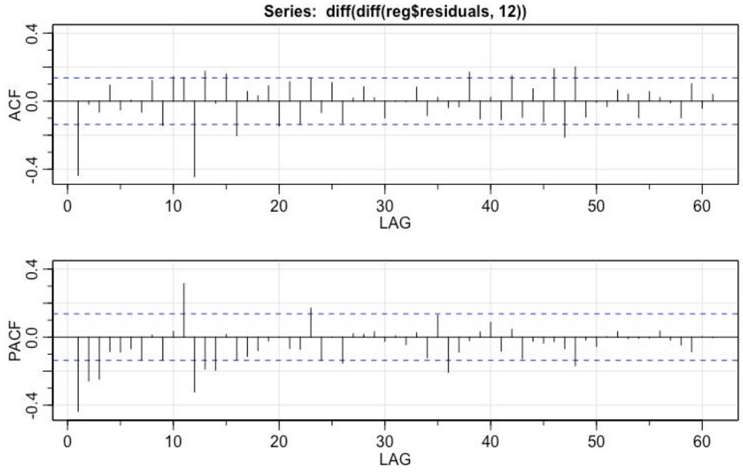

Figure 11: ACF & PACF of differenced residuals

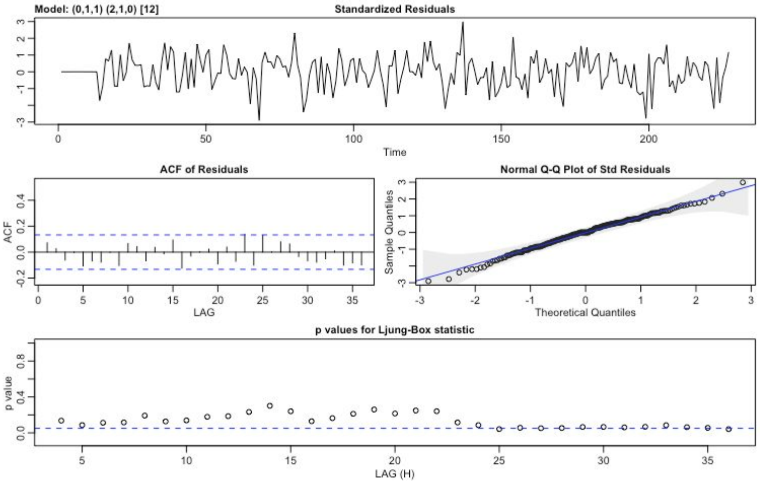

Figure 12: Diagnostics of differenced residuals