|  |  |
| --- | --- |
| Unemployment rate Australia 2010-2021 | Abstract  A Research on the Time series data for unemployment rate in Australia from 2010 until 2021. We have deployed an ARIMA & ARMA-GARCH model to perform time-series Visual and Numerical analysis.  Durvesh ChattopadhyayDurvesh C |

Table of Contents

[List of Figures 2](#_Toc74517116)

[List of Tables 2](#_Toc74517117)

[1. Introduction 3](#_Toc74517118)

[2. Methodology 3](#_Toc74517119)

[3. Exploration of Data 4](#_Toc74517120)

[3.1. Time-Series Plot of Unemployment Rate 4](#_Toc74517121)

[3.2. Scatterplot of the Data and Lagged Version of Data 5](#_Toc74517122)

[Data Interpretation 6](#_Toc74517123)

[Test of Stationary 6](#_Toc74517124)

[ACF and PACF Plots 6](#_Toc74517125)

[ADF Test 6](#_Toc74517126)

[Transformation 6](#_Toc74517127)

[First-Order Differencing 7](#_Toc74517128)

[McLeod Li Test and Normal Q-Q Plot 8](#_Toc74517129)

[Model-Building 8](#_Toc74517130)

[Model Specification 8](#_Toc74517131)

[ACF and PACF Plots 9](#_Toc74517132)

[EACF (Extended Autocorrelation Function) 9](#_Toc74517133)

[BIC (Bayesian Information Criterion) 10](#_Toc74517134)

[Model fitting and Parameter estimation 11](#_Toc74517135)

[For ARIMA (0,1,1) 11](#_Toc74517136)

[For ARIMA (1,1,1) 12](#_Toc74517137)

[For ARIMA (3,1,0) 14](#_Toc74517138)

[For ARIMA (3,1,1) 15](#_Toc74517139)

[For ARIMA (3,1,2) 17](#_Toc74517140)

[For ARIMA (3,1,3) 19](#_Toc74517141)

[AIC and BIC values for the selected models: 21](#_Toc74517142)

[Overfitting Analysis: 22](#_Toc74517143)

[Results 24](#_Toc74517144)

[Limitations and Proposed Solutions: 25](#_Toc74517145)

[Conclusion: 25](#_Toc74517146)

[Reference: 25](#_Toc74517147)

[Appendix: 26](#_Toc74517148)

# List of Figures

[Figure 1: Model Selection Framework for our project. 4](#_Toc74517149)

[Figure 2: Time-Series Plot of Unemployment Rate (2010-2021) 5](#_Toc74517150)

[Figure 3: Scatter Plot of Data and Lagged Data 5](#_Toc74517151)

[Figure 4: ACF and PACF Plots of Time-Series 6](#_Toc74517152)

[Figure 5: ADF Test of Stationary on Series 6](#_Toc74517153)

[Figure 6: First-Order Differencing of Time-Series 7](#_Toc74517154)

[Figure 7: McLeod Li Test and Normality Test 8](#_Toc74517155)

[Figure 8: ACF and PACF Plots of Differenced Series 9](#_Toc74517156)

[Figure 9: EACF Plot 10](#_Toc74517157)

[Figure 10: BIC 10](#_Toc74517158)

[Figure 11: Residual Analysis of ARIMA (0,1,1) 12](#_Toc74517159)

[Figure 12: Residual Analysis for ARIMA (1,1,1) 13](#_Toc74517160)

[Figure 13: Residual Analysis of ARIMA (3,1,0) 15](#_Toc74517161)

[Figure 14: Residual analysis of ARIMA (3,1,1) 17](#_Toc74517162)

[Figure 15: Residual Analysis for ARIMA (3,1,2) 19](#_Toc74517163)

[Figure 16:: Residual Analysis for ARIMA (3,1,3) 21](#_Toc74517164)

[Figure 17: Residual Analysis for Overfitting Model 23](#_Toc74517165)

[Figure 18: Forecast for Unemployment Rate for next 10 months. 24](#_Toc74517166)

# List of Tables

[Table 1: PP Test after First-Order Differencing 7](#_Toc74517225)

[Table 2: ADF Test after First-order differencing 7](#_Toc74517226)

[Table 3: Possible ARIMA Models from ACF and PACF 9](#_Toc74517227)

[Table 4: Possible ARIMA Models from EACF 10](#_Toc74517228)

[Table 5: Possible ARIMA (p,d,q) Models 11](#_Toc74517229)

[Table 6: Z-test for coefficients for CSS and ML for ARIMA (0,1,1) 11](#_Toc74517230)

[Table 7: Shapiro-Wilk Normality Test for ARIMA (0,1,1) 11](#_Toc74517231)

[Table 8: Z-test for coefficients for CSS and ML for ARIMA (1,1,1) 13](#_Toc74517232)

[Table 9: Shapiro-Wilk Normality Test for ARIMA (1,1,1) 13](#_Toc74517233)

[Table 10: Z-test for coefficients for CSS and ML for ARIMA (3,1,0) 14](#_Toc74517234)

[Table 11: Shapiro-Wilk Normality Test for ARIMA (3,1,0) 14](#_Toc74517235)

[Table 12: : Z-test for coefficients for CSS and ML for ARIMA (3,1,1) 16](#_Toc74517236)

[Table 13: Shapiro-Wilk Normality Test for ARIMA (3,1,1) 16](#_Toc74517237)

[Table 14: Z-test for coefficients for CSS and ML for ARIMA (3,1,2) 18](#_Toc74517238)

[Table 15: Shapiro-Wilk Normality Test for ARIMA (3,1,2) 18](#_Toc74517239)

[Table 16: Z-test for coefficients for CSS and ML for ARIMA (3,1,3) 20](#_Toc74517240)

[Table 17: Shapiro-Wilk Normality Test for ARIMA (3,1,3) 20](#_Toc74517241)

[Table 18: AIC Table for selected ARIMA models 21](#_Toc74517242)

[Table 19: BIC Table for selected ARIMA models. 22](#_Toc74517243)

[Table 20: Parameter Estimation for ARIMA (4,1,2) overfitting model 22](#_Toc74517244)

[Table 21: Shapiro-Wilk Test for overfitting model 23](#_Toc74517245)

[Table 22: Forecast Table. 24](#_Toc74517246)

# 1. Introduction

The report presents time-series analysis on annual unemployment rate, an important driver of monetary policy decisions, persisting in Australia for over a decade, 2010-2021. Due to economic downturn faced over the last year, unemployment rate has been at all-time high to 7.4% which aggravates the significance of this analysis. It is an important summary statistic to monitor economic progress which necessitates making monetary decisions, such as wage rates, monetary policy settings. (Education, 2021) The long-term unemployment often increases the well-being costs of Australia, leading to public policy interventions.

The aim of this report is to identify the best fitted time-series model on the univariate time-series dataset to understand unemployment component and forecast the rate in the future for next 4 months. The report is primarily divided into three sections. First, the exploratory data analysis suggested the presence of trend, AR and MA components in the series. The stationary component was removed by performing first order differencing, evidenced by stationary tests. The report proceeds into model specification by using model specification tools like ACF and PACF plots along with EACF and BIC. The possible set of candidate stochastic trend models are fitted to the series, to identify the best fitted model, based on residual analysis. Finally, the best fitted model was used to forecast the variable for the next 10 months.

# 2. Methodology

The unemployment rate is the percentage of people who are unemployed in the labour force. It is an important summary statistic to monitor economic progress. This report acknowledges the significance of unemployment rate and performs the time-series analysis on a univariate time-series dataset on annual unemployment rate from Jan 2010 – April 2021 to identify the best fitted model for forecasting for next 10 months. The analysis is performed on R environment on the dataset extracted from Australian Macro Database.

To identify the best fitted model, we have used autoregressive moving average model (ARIMA). The report is primarily divided into two sections.

The first section of the report performs exploratory data analysis using time-series plot in combination with scatter plot to evaluate the series on 5 characteristics which shows that there is trend as well as degree of changing variance after 2020, along with AR and MA behaviour.

Thus, the nature of series calls for stabilizing variance through suitable transformations (log transformation), followed by differencing to handle the stationary component of the series. Tests like McLeod Li test and Normal Q-Q plot were used to check the presence of volatility clustering in the series to apply GARCH model. Further, plots like ACF and PACF defines possible set of parameters which leads to parameter estimation along with model diagnostics, using residual analysis including overfitting method to arrive at the suitable forecasting model.

The forecasting model was used to forecast unemployment rate for the next 10 months. The model performance was evaluated using RMSE and MSE.

The Below figure 1 is our Model Selection Framework followed:

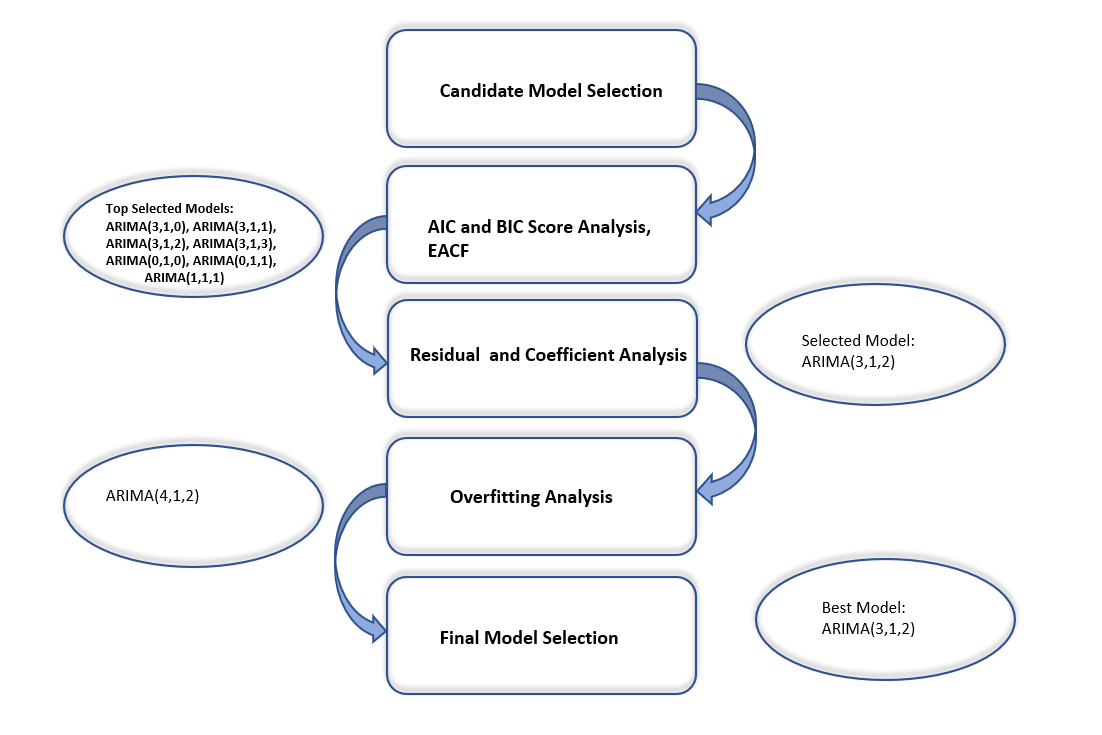


Figure 1: Model Selection Framework for our project.

# 3. Exploration of Data

The series consists of unemployment rate recordings from January 2010 – April 2021 from Australian Marco Database. It is an annual time series data with percentage of unemployed labour force in Australia over the decade. The dataset consists of the unemployment rate reaching a skyrocket 7.4% in July 2020. This could be the effects of the ongoing global pandemic. It also records the lowest unemployment rate at 4.9% in June 2011. The whole data set averages out at 5.6%. The data was converted into time-series and explored against time.

## 3.1. Time-Series Plot of Unemployment Rate

The Figure 1 captures the time-series data of unemployment rate from 2010-2021. The plot exhibits that the unemployment rate has decreased over the first two years, contrary to the next years which projects an increasing trend from 2012 – 2015. According to [Australian Bureau of Statistics](https://www.theguardian.com/business/2014/aug/07/australias-jobless-rate-hits-highest-level-in-more-than-a-decade), the increase was induced by reduction of full-time job as well as increase in participation workforce by 43,400.

* Trend – The plot shows an increasing as well as decreasing trend throughout the decade except for 2020.
* Seasonality – The series does not show any seasonality or cyclical pattern.
* Changing variance – There is no significant changing variance with the exception of 2020 which could be induced by the economic downturn due to COVID-19.
* Change Point – The series does have a change point during latter of 2020, possibly induced by COVID-19.
* Behaviour – The neighbours in time seem related to one another and nearby pattern, suggesting AR and MA behaviour.

Chart, scatter chart

Description automatically generated

Figure 2: Time-Series Plot of Unemployment Rate (2010-2021)

The non-stationary nature of the series is primarily because of two components: trend component and irregular variation. With the small degree of variance in the series and other characteristics, stochastic trend model can be used to model the series.

## 3.2. Scatterplot of the Data and Lagged Version of Data

The scatter plot also adds to exploration as it can project any latent trend or presence of correlation in the data.

Chart, scatter chart

Description automatically generated

Figure 3: Scatter Plot of Data and Lagged Data

The Figure 2 is the scatter plot of neighbours in time which shows that there exists positive correlation, an upward trend, between unemployment rate of the current year and previous year – suggesting that higher unemployment rate is followed by higher unemployment rate in next year. The correlation function shows an extremely high correlation value of 0.9318962 and confirms AR component.

# Data Interpretation

## Test of Stationary

It is necessary to ensure constant statistical properties of time series, i.e., stationary time series before proceeding to model-building strategy. We have used ACF and PACF plots as well as ADF Test to check for stationary.

### ACF and PACF Plots

Chart, histogram

Description automatically generated

Figure 4: ACF and PACF Plots of Time-Series

The ACF plot confirms the presence of autocorrelation, with significant lags, and trend in the series because of its slowly decaying nature of the plot. Also, PACF plot suggests more than one significant value, projecting trend. Thus, the series is not stationary.

### ADF Test

The Augmented Dickey-Fuller Test is the test of stationary on the dataset. The test fails to reject the null hypothesis at 5% significance level, suggesting that the data is non-stationary.

Text

Description automatically generated

Figure 5: ADF Test of Stationary on Series

## Transformation

With the presence of variance in time series data, it is necessary to stabilize variance as the non-constant variance will induce bigger errors in forecast and make it difficult to extrapolate. Thus, we have performed variance-stabilizing transformations on the series – Box-Cox Transformation and Log Transformation.

After performing Box-Cox Transformation on the series, the lambda value is close to -1 which suggests that the results are closer to the results derived from Log Transformation. Thus, we have utilized log transformation to the time-series to reduce the variance in the series.

However, the transformation did not significantly reduce the variance but has proved to be relatively effective. Thus, we will continue the analysis with the log transformed series.

## First-Order Differencing

The presence of trend in the time-series suggests non-stationary nature of the series. A stationary time series has constant statistical properties which is necessary to increase predictive power of the model. Thus, we perform different levels of degree of differencing to make the data stationary.

The Figure 4 shows the series after performing first order differencing which results in a stationary series, evidenced by ADF test and PP test. Both tests of stationary reject the null hypothesis that the series is non-stationary at 5% significance level.

Chart

Description automatically generated with low confidence

Figure 6: First-Order Differencing of Time-Series

Text

Description automatically generated

Table 1: PP Test after First-Order Differencing

Text

Description automatically generated

Table 2: ADF Test after First-order differencing

Thus, the value of d is 1 and we can proceed with model-building strategy to arrive at set of possible ARIMA(p,d,q) models.

## McLeod Li Test and Normal Q-Q Plot

McLeod Li test is performed to check for the presence of ARCH (autoregressive conditional heteroskedasticity) effects among lags. ARCH identifies nonconstant volatility in relation to the previous period’s volatility which produces heavy tails. GARCH is an extension of ARCH that estimates volatility in the series.

Chart, scatter chart

Description automatically generated

Figure 7: McLeod Li Test and Normality Test

In the McLeod Li Test, we can see that for initial few lags p-value is significant, but majority of the lags do not indicate the presence of volatility clustering as the observed p-value is above the horizontal red line. To confirm the presence of volatility clustering, we also explored the Normal Q-Q plot which indicated otherwise. Q-Q plot does not indicate heavy tails as not many points deviate from the reference line at the ends. Thus, we can say that the volatility clustering for first 2 lags in the McLeod Li test might be a result of the change point as well as the outliers in the Q-Q plot. Thus, we can say that it might not be productive to implement GARCH model because of minimal volatility in the return series, and we will proceed with ARIMA model.

# Model-Building

To identify an appropriate predictive model for the series, we have divided it into three sections:

* Model Specification– ACF, PACF, BIC and EACF
* Model Fitting and Parameter Estimation - ARIMA
* Model Diagnostics – Residual Analysis

## Model Specification

This part of the report deals with model specification by using tools like ACF and PACF plots, EACF and BIC tables to identify set of possible ARIMA (p,q,d) models to fit the stochastic trend model.

As supported by first order differencing, the value for d = 1 (order of differencing transforming series into stationary). To deal with the stochastic components of the time-series data, we will determine the possible values of p (number of Autoregressive terms) and q (number of lagged forecast errors in prediction equation) to achieve the best fitted ARIMA class model.

### ACF and PACF Plots

Chart, box and whisker chart

Description automatically generated

Figure 8: ACF and PACF Plots of Differenced Series

***ACF:*** The ACF plot does not show any significant lags beyond the confidence band. Although, lag 3 is “cut-off” lag, thus, we will consider q=0,3. Thus, IMA (1,0) and IMA (1,3) are possible set of models. Since, we see that lag 3 is just insignificant, we will consider q=1,2 so that we explore all possible scenarios to arrive at best model.

***PACF:*** The PACF plot shows there is significant lag at 3, therefore, p = 3. Thus, AR(3) is appropriate.

Thus, possible set of ARIMA (p,d,q) models from ACF and PACF plots are:

|  |
| --- |
| Possible Models |
| ARIMA (3,1,0) |
| ARIMA (3,1,1) |
| ARIMA (3,1,2) |
| ARIMA (3,1,3) |

Table 3: Possible ARIMA Models from ACF and PACF

### EACF (Extended Autocorrelation Function)

This method captures the white noise behaviour of the series. The symbols x and o mean the sample correlation of AR residuals are significantly different and not significantly different respectively.

Background pattern

Description automatically generated with medium confidence

Figure 9: EACF Plot

The plot gives p =0 and q=0 as the vertex value gives the ARMA (0,0). Considering neighbouring values and d=1, possible set of ARIMA class models are:

|  |
| --- |
| Possible Models |
| ARIMA (0,1,0) |
| ARIMA (0,1,1) |
| ARIMA (1,1,1) |

Table 4: Possible ARIMA Models from EACF

### BIC (Bayesian Information Criterion)

This model is based on MLE to identify the p and q values by determining the orders that minimises the BIC.

*BIC = -2log\*(maximum likelihood) + klog(n)*

A picture containing text, crossword puzzle

Description automatically generated

Figure 10: BIC

The BIC table gives the value if p = 3 and q=0 which leads to ARMA (3, 0). Thus, the only possible ARIMA class model is ARIMA (3,1,0)

Thus, the model specification results in following set of possible models:

|  |  |  |
| --- | --- | --- |
| Possible Set of ARIMA (p,d,q) Models | | |
| ACF and PACF Plots | **EACF** | **BIC** |
| ARIMA (3,1,0) | ARIMA (0,1,1) | ARIMA (3,1,0) |
| ARIMA (3,1,1) | ARIMA (1,1,1) |  |
| ARIMA (3,1,2) | ARIMA (0,1,0) |  |
| ARIMA (3,1,3) |  |  |

Table 5: Possible ARIMA (p,d,q) Models

## Model fitting and Parameter estimation

### For ARIMA (0,1,1)

**Parameter Estimation:** From table 6, the 'CSS' and ‘ML’ model, we can see that the p-value for all the parameters is 0.3 at 5% which is >0.05 (taking alpha level as 0.05). Hence, we fail to reject Ho, which means we can say that all the parameters of the model are not significant.

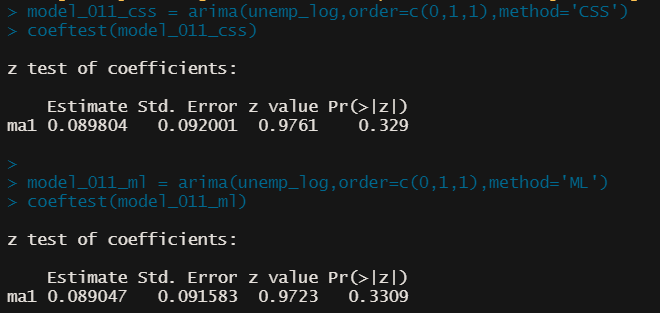


Table 6: Z-test for coefficients for CSS and ML for ARIMA (0,1,1)

#### Residual Analysis:

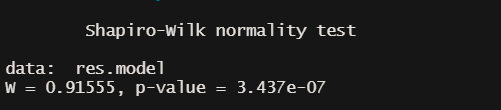


Table 7: Shapiro-Wilk Normality Test for ARIMA (0,1,1)

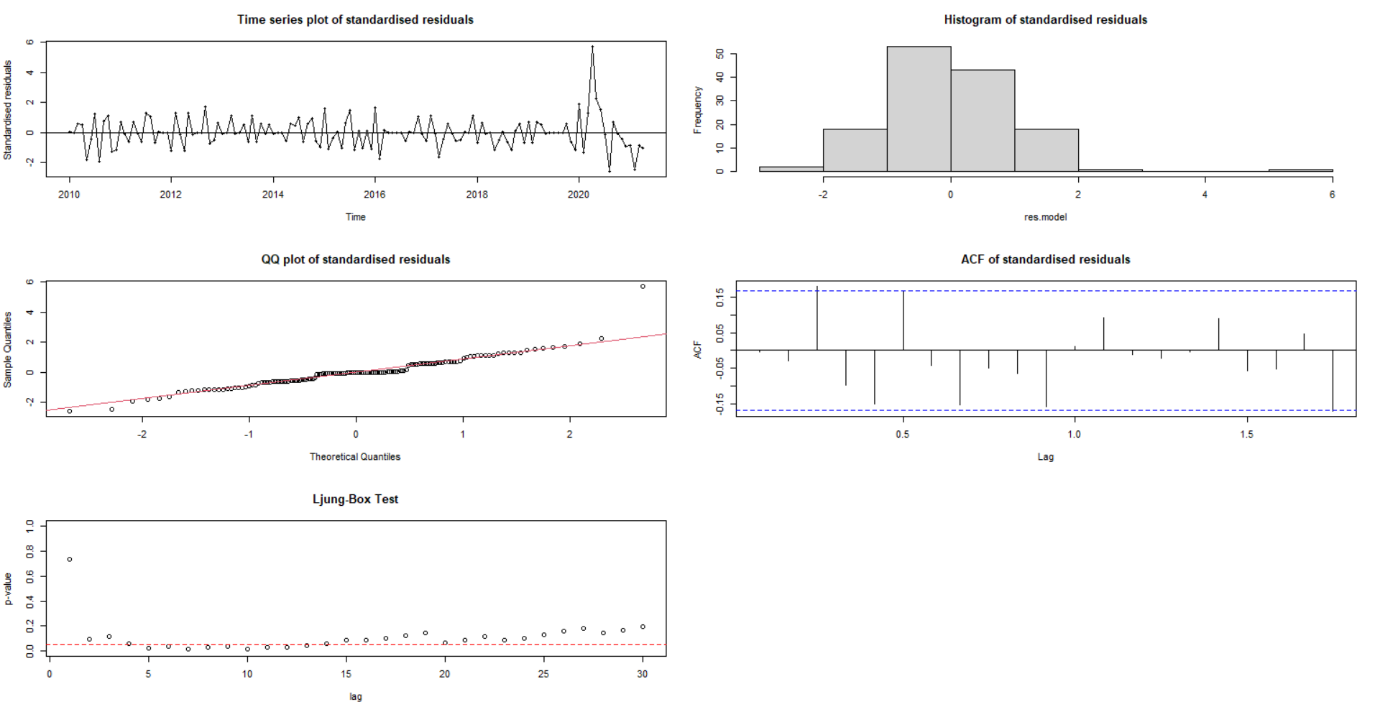


Figure 11: Residual Analysis of ARIMA (0,1,1)

Based on figure 11 the times series plot of the standardised residuals of ARIMA (0,1,1) imply that the values are distributed in a rectangle pattern, allowing the model to be adequate. In the above figure from the QQ plot we can infer that the residuals are suggested to be normally distributed but the same cannot be confirmed from the histogram and Shapiro-Wilk test, they state that the residuals are not normally distributed.

From the above figure we can see that the ACF plot of the residuals contain significant lags at 2 and 5, which states that there is some autocorrelation left in the residuals and the model cannot capture all the variations. From Ljung-Box test we can see that there is some statistical evidence of residuals being correlated.

Because of some correlation in residual analysis, and the ma parameter of the model is not significant for ARIMA (0,1,1), hence we will ignore the coefficient test of this model and this model cannot be adequate for modelling our series.

### For ARIMA (1,1,1)

**Parameter Estimation:** From table 8, the 'CSS' and ‘ML’ model, we can see that the p-value for MA1 is significant but AR1 is 0.07 which is greater than 0.05 (taking alpha level as 0.05). As one of the parameters is insignificant, we fail to reject Ho, which means we can say that the parameters of the model are not significant.

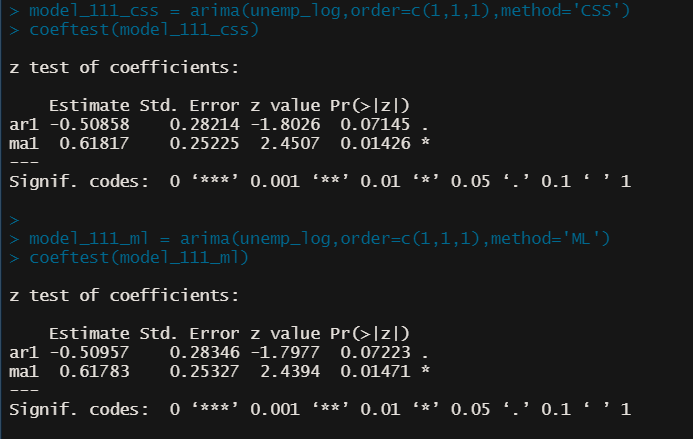


Table 8: Z-test for coefficients for CSS and ML for ARIMA (1,1,1)

**Residual Analysis:**

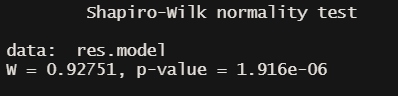


Table 9: Shapiro-Wilk Normality Test for ARIMA (1,1,1)

Diagram

Description automatically generated with medium confidence

Figure 12: Residual Analysis for ARIMA (1,1,1)

Based on figure 12 the times series plot of the standardised residuals of ARIMA (1,1,1) imply that the values are distributed in a rectangle pattern, allowing the model to be adequate. In the above figure from the QQ plot we can infer that the residuals are suggested to be normally distributed but the same cannot be confirmed from the histogram and Shapiro-Wilk test, they state that the residuals are not normally distributed.

From the above figure we can see that the ACF plot of the residuals does not contain any significant lag in it, which states that there is no autocorrelation left in the residuals and the model can capture all the variations. From Ljung-Box test we can see that there is some statistical evidence of residuals being correlated.

Because of some correlation in residual analysis, and all the parameters of the model are not significant for ARIMA (1,1,1), hence we will ignore the coefficient test of this model and this model cannot be adequate for modelling our series.

### For ARIMA (3,1,0)

**Parameter Estimation:** From table 10, the 'CSS' and ‘ML’ model, we can see that the p-value for AR3 is significant but AR1 and AR2 values is greater than 0.05 (taking alpha level as 0.05). As two of the parameters is insignificant, we fail to reject Ho, which means we can say that all the parameters of the model are not significant.

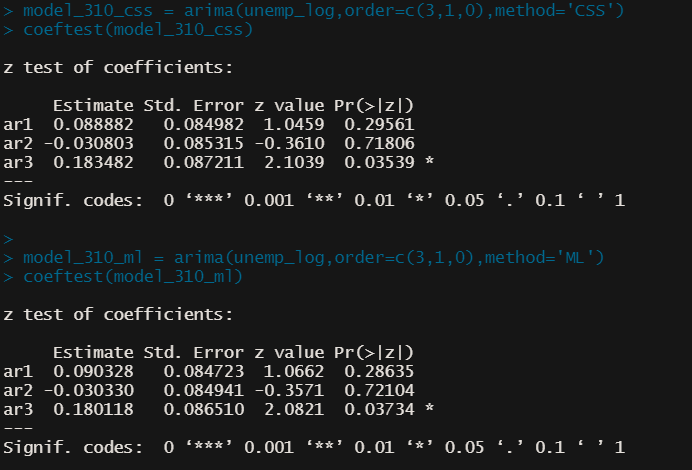


Table 10: Z-test for coefficients for CSS and ML for ARIMA (3,1,0)

**Residual Analysis:**

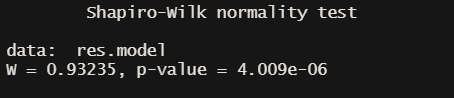


Table 11: Shapiro-Wilk Normality Test for ARIMA (3,1,0)

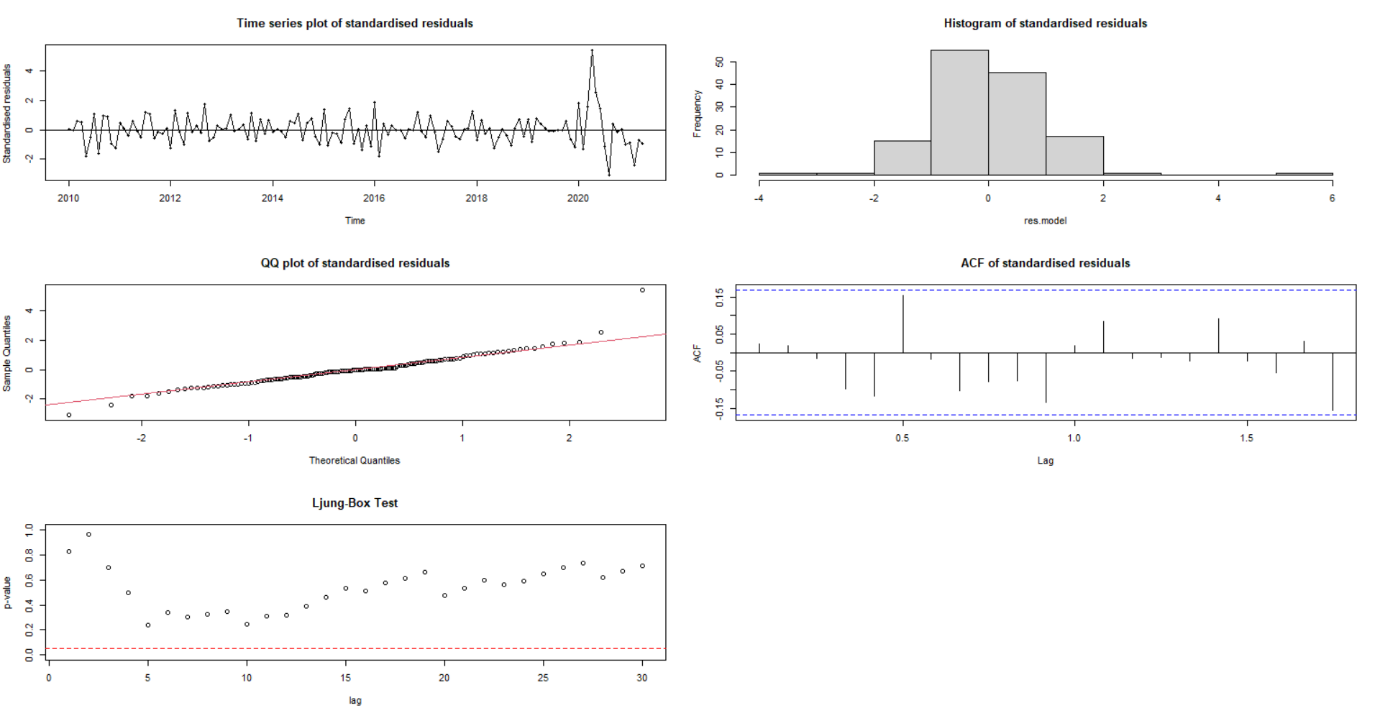


Figure 13: Residual Analysis of ARIMA (3,1,0)

Based on figure 13 the times series plot of the standardised residuals of ARIMA (3,1,0) imply that the values are distributed in a rectangle pattern, allowing the model to be adequate. In the above figure from the QQ plot we can infer that the residuals are suggested to be normally distributed but the same cannot be confirmed from the histogram and Shapiro-Wilk test, they state that the residuals are not normally distributed.

From the above figure we can see that the ACF plot of the residuals does not contain any significant lag in it, which states that there is no autocorrelation left in the residuals and the model can capture all the variations. From Ljung-Box test we can see that there is no statistical evidence of residuals being correlated.

Despite good residual analysis, since all the parameters of the model are not significant for ARIMA (3,1,0), hence we will ignore the coefficient test of this model and this model cannot be adequate for modelling our series.

### For ARIMA (3,1,1)

**Parameter Estimation:** From table 12, the 'CSS' and ‘ML’ model, we can see that the p-value for AR3 is significant but AR1,AR2 and MA1 values is greater than 0.05 (taking alpha level as 0.05). As three of the parameters is insignificant, we fail to reject Ho, which means we can say that all the parameters of the model are not significant.

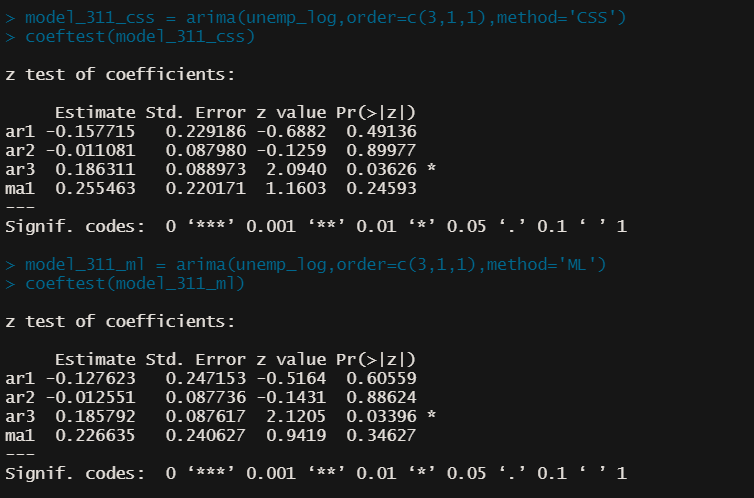


Table 12: : Z-test for coefficients for CSS and ML for ARIMA (3,1,1)

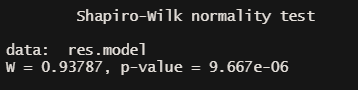


Table 13: Shapiro-Wilk Normality Test for ARIMA (3,1,1)

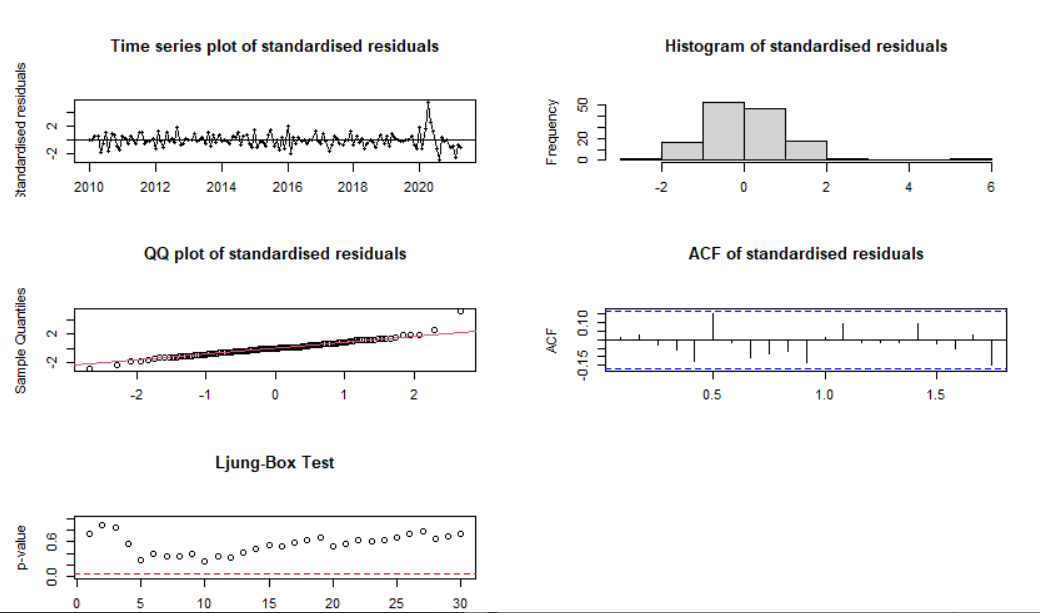


Figure 14: Residual analysis of ARIMA (3,1,1)

Based on figure 14 the times series plot of the standardised residuals of ARIMA (3,1,1) imply that the values are distributed in a rectangle pattern, allowing the model to be adequate. In the above figure from the QQ plot we can infer that the residuals are suggested to be normally distributed but the same cannot be confirmed from the histogram and Shapiro-Wilk test, they state that the residuals are not normally distributed.

From the above figure we can see that the ACF plot of the residuals does not contain any significant lag in it, which states that there is no autocorrelation left in the residuals and the model can capture all the variations. From Ljung-Box test we can see that there is no statistical evidence of residuals being correlated.

Despite good residual analysis, since all the parameters of the model are not significant for ARIMA (3,1,1), hence we will ignore the coefficient test of this model and this model cannot be adequate for modelling our series.

### For ARIMA (3,1,2)

**Parameter Estimation:** In the table 12 for the 'CSS' model we can see that the p-value for all the parameters is < 0.01 at 5% which is <0.05 (taking alpha level as 0.05). Hence, we reject H0, which means we can say that all the parameters of the model are significant. Similarly, we see that the all the parameters in 'ML' model are significant and since all the parameters are significant in both the models, we can say that the model can be a good fit for our data based on parameter estimation.

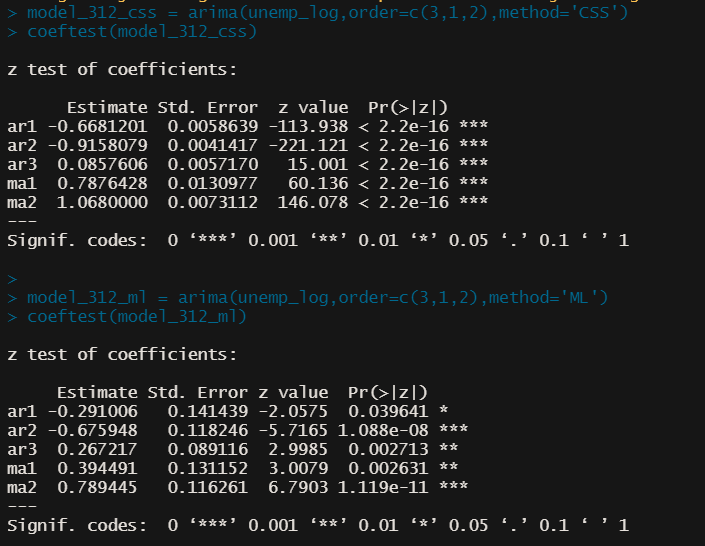


Table 14: Z-test for coefficients for CSS and ML for ARIMA (3,1,2)

**Residual Analysis:**

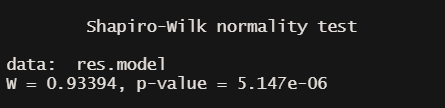


Table 15: Shapiro-Wilk Normality Test for ARIMA (3,1,2)

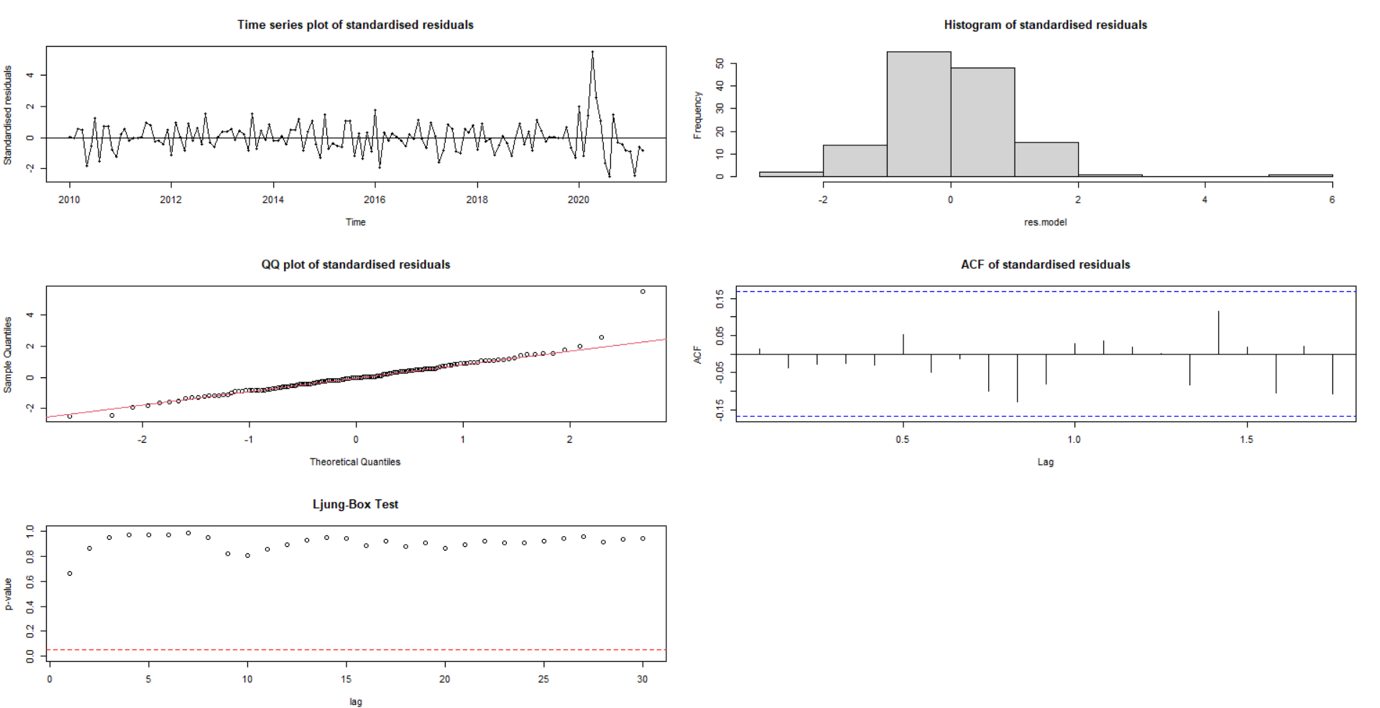


Figure 15: Residual Analysis for ARIMA (3,1,2)

Based on figure 14 the times series plot of the standardised residuals of ARIMA (3,1,2) imply that the values are distributed in a rectangle pattern, allowing the model to be adequate. In the above figure from the QQ plot we can infer that the residuals are suggested to be normally distributed but the same cannot be confirmed from the histogram and Shapiro-Wilk test, they state that the residuals are not normally distributed.

From the above figure we can see that the ACF plot of the residuals does not contain any significant lag in it, which states that there is no autocorrelation left in the residuals and the model can capture all the variations. From Ljung-Box test we can see that there is no statistical evidence of residuals being correlated.

Since all the parameters are significant for ARIMA (3,1,2) and the residual analysis look good, this model can be considered adequate for modelling our analysis.

### For ARIMA (3,1,3)

**Parameter Estimation:** We can see that NANs have been produced in the result of coefficient test of 'CSS' model; this is because negative values are found in the inverse of the Hessian matrix evaluated at the returning solution. As a result, they frequently imply that the model is inappropriate for the data; the data lacks sufficient information to generate trustworthy estimates for the chosen parameters.

For the coefficient test of ML model, we can see that ar1, ar3 and ma3 are not significant. Hence, we can say that not all parameters for this model are significant.

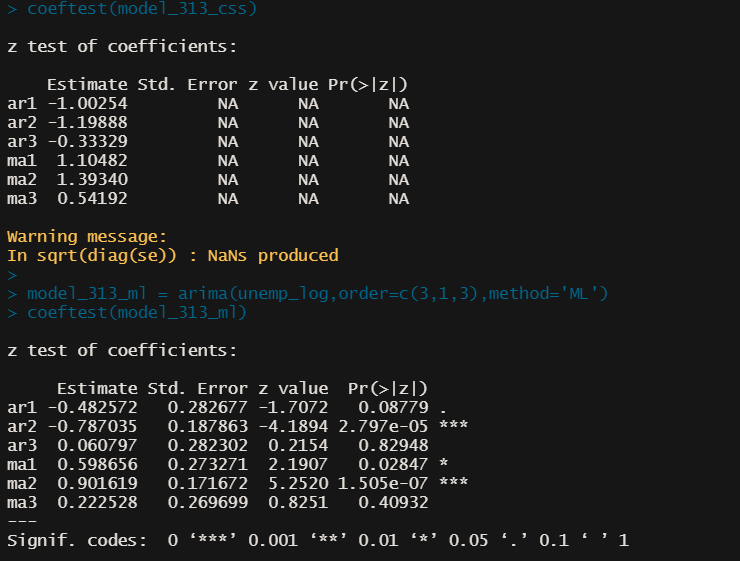


Table 16: Z-test for coefficients for CSS and ML for ARIMA (3,1,3)

#### **Residual Analysis**:

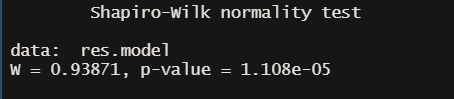


Table 17: Shapiro-Wilk Normality Test for ARIMA (3,1,3)

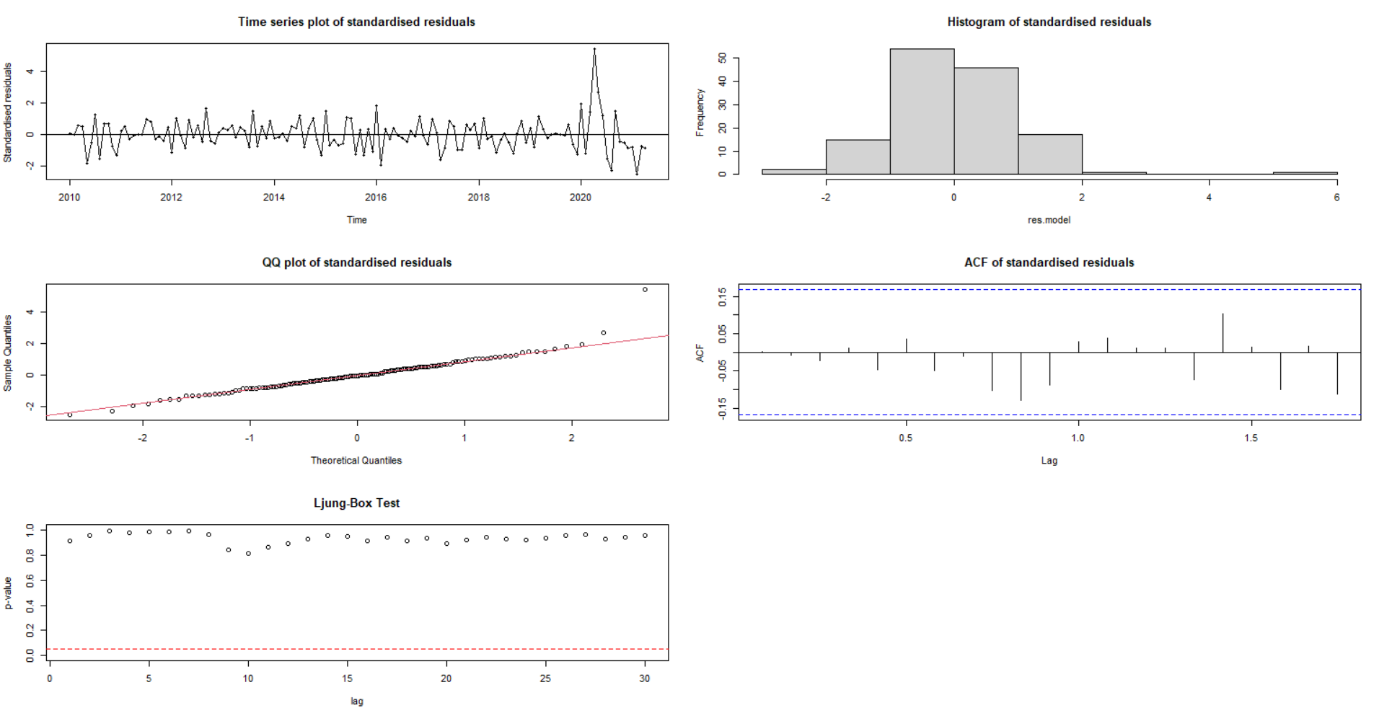


Figure 16:: Residual Analysis for ARIMA (3,1,3)

Based on figure 15 the times series plot of the standardised residuals of ARIMA (3,1,3) imply that the values are distributed in a rectangle pattern, allowing the model to be adequate. In the above figure from the QQ plot we can infer that the residuals are suggested to be normally distributed but the same cannot be confirmed from the histogram and Shapiro-Wilk test, they state that the residuals are not normally distributed.

From the above figure we can see that the ACF plot of the residuals does not contain any significant lag in it, which states that there is no autocorrelation left in the residuals and the model can capture all the variations. From Ljung-Box test we can see that there is no statistical evidence of residuals being correlated.

Since all the parameters are not significant for ARIMA (3,1,3) and our data lacks sufficient information to generate trustworthy estimates for the chosen parameters. Hence, we will ignore the coefficient test of this model and this model cannot be adequate for modelling our series.

### AIC and BIC values for the selected models:

|  |  |  |
| --- | --- | --- |
| Model Name | df | AIC |
| model\_312\_ml | 6 | -543.9904 |
| model\_313\_ml | 7 | -542.5763 |
| model\_010\_ml | 1 | -539.6620 |
| model\_310\_ml | 4 | -538.8805 |
| model\_011\_ml | 2 | -538.6312 |
| model\_111\_ml | 3 | -537.7745 |
| model\_311\_ml | 5 | -537.6337 |

Table 18: AIC Table for selected ARIMA models

|  |  |  |
| --- | --- | --- |
| Model Name | df | BIC |
| model\_010\_ml | 1 | -1536.7493 |
| model\_011\_ml | 2 | -2532.8059 |
| model\_111\_ml | 3 | -529.0366 |
| model\_310\_ml | 4 | -527.2299 |
| model\_312\_ml | 6 | -526.5145 |
| model\_311\_ml | 5 | -523.0704 |
| model\_313\_ml | 7 | -522.1877 |

Table 19: BIC Table for selected ARIMA models.

The Akaike information criterion (AIC) is a measure of error rate and, as a result, the comparative efficiency of predictive models for a given data set. AIC measures the quality of each model in relation to the other models given a set of data models. As a result, AIC provides a method for model selection. Hence, we would select the model with the lowest AIC score. Also, Bayesian information criterion (BIC) is a criterion for selecting a model from a finite set of models, the model with the lowest BIC is preferred. From the above AIC and BIC table we can see that ARIMA (3,1,2) has the lowest AIC at -543.9904 whereas ARIMA (0,1,0) has the lowest BIC at -1536.7493. But since ARIMA (0,1,0) does not specify any model we will skip it and select the next best model with the lowest BIC, the model with best parameter estimation and residual analysis. ARIMA (3,1,2) is the best model having the next lowest BIC at –526.5145.

From all the above analysis we can say the ARIMA (3,1,2) is the best model suited for our dataset.

## Overfitting Analysis:

We will conduct overfitting and do the analysis on the overfitted models just to check the neighbouring models of our selected model. The overfitted models are ARIMA (4,1,2) and ARIMA (3,1,3). Since we have already preformed residual analysis and parameter estimation on ARIMA (3,1,3), we will perform the same tests on ARIMA (4,1,2).

**Parameter Estimation:**

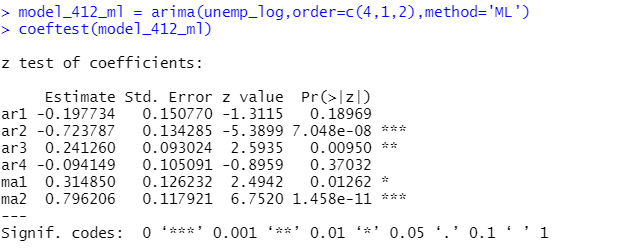


Table 20: Parameter Estimation for ARIMA (4,1,2) overfitting model

From the table we can see that ar1 and ar4 parameter has p value>0.05, hence we can say that this parameter is not significant (taking alpha level as 0.05) and hence we can say that not all parameters are significant.

**Residual Analysis:**

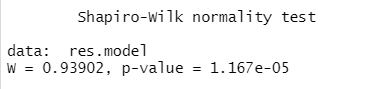


Table 21: Shapiro-Wilk Test for overfitting model

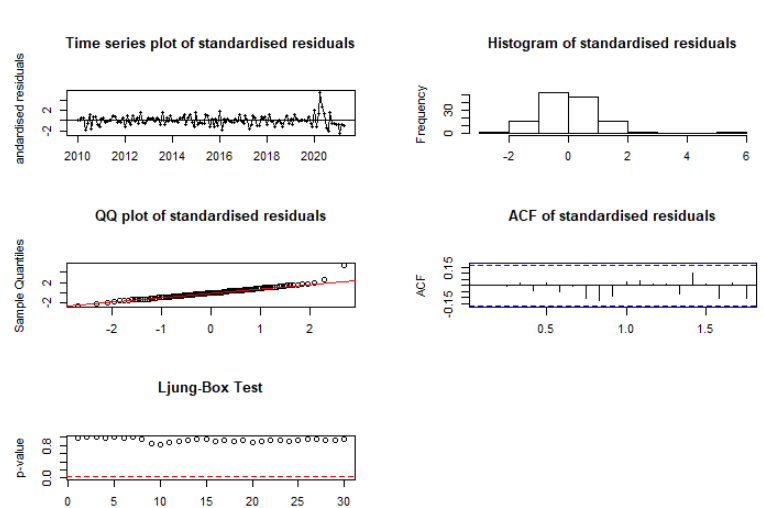


Figure 17: Residual Analysis for Overfitting Model

Based on figure 29 the times series plot of the standardised residuals of ARIMA (4,1,2) imply that the values are distributed in a rectangle pattern, allowing the model to be adequate. In the above figure from the QQ plot we can infer that the residuals are suggested to be normally distributed but the same cannot be confirmed from the histogram and Shapiro-Wilk test, they state that the residuals are not normally distributed. (D.Cryer, 2008)

From the above figure we can see that the ACF plot of the residuals does not contain any significant lag in it, which states that there is no autocorrelation left in the residuals and the model can capture all the variations. From Ljung-Box test we can see that there is no statistical evidence of residuals being correlated.

Despite good residual analysis, since all the parameters of the model are not significant for ARIMA (4,1,2), hence we will ignore the coefficient test of this model and this model cannot be adequate for modelling our series.

Finally, we will move ahead with our best selected model for forecasting i.e., **ARIMA (3,1,2).**

# Results

The Figure 30 shows the forecast plot of the rate of unemployment in Australia from, May 2021 until February 2022. The highlighted area shows the confidence interval width of the predicted values.

According to the above forecast results, the values tend to have a good forecasting result. With the confidence interval values not fluctuating more than 2 percent the forecasting plot predicts that the average rate of unemployment is at a constant rate 5.39 % and stays steady till the end of February 2022.

|  |  |
| --- | --- |
| Forecast Month, Year | Forecast Mean |
| May 2021 | 5.429564 |
| Jun 2021 | 5.422130 |
| Jul 2021 | 5.419791 |
| Aug 2021 | 5.406841 |
| Sep 2021 | 5.410205 |
| Oct 2021 | 5.417357 |
| Nov 2021 | 5.409538 |
| Dec 2021 | 5.407881 |
| Jan 2022 | 5.415558 |
| Feb 2022 | 5.412355 |

Table 22: Forecast Table.



The RMSE of 0.03 and MAE of 0.02 suggests that model is a good fit.

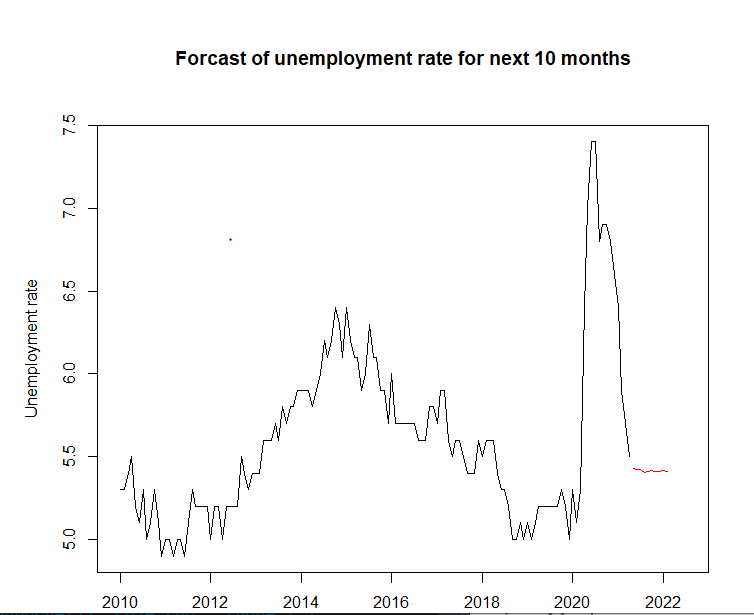


Figure 18: Forecast for Unemployment Rate for next 10 months.

# Limitations and Proposed Solutions:

In this section we will go through the potential issues identified during our research working the Unemployment data.

1. ARIMA performs better in short-term forecasting, whereas LSTM (Long short-term memory) performs better in long-term modelling. Time series forecasting approaches such as ARIMA rely on univariate data with linear connections and fixed and manually diagnosed temporal dependency. When comparing LSTM to ARIMA for deep learning issues with a large dataset, it was discovered that the average drop-in error rates attained by LSTM is somewhere between 84–87 %, showing that LSTM is better to ARIMA. (Central, 2017)
2. From the QQ plot for the coefficient and residual analysis, we can infer that the residuals are suggested to be normally distributed but the same cannot be confirmed from the histogram and Shapiro-Wilk test, they state that the residuals are not normally distributed.
3. The transforms of Box-Cox and Log were unable to stabilise the variance in the series. These assumptions, however, can be disregarded considering the high number of observations. To cope with normality and stabilising the series, other transformation methods might be investigated.
4. We could have test higher order of P=5. However, the lower orders performed well in our case.
5. From the figure 17, we can see there is a sudden rise in the unemployment rate between 2020 to 2021. These values could be treated as outliers. Outliers are difficult to predict with ARIMA since they are outside of the general trend represented by the model. (Towards Data Science , 2018)

# Conclusion:

To conclude, from the first section of the report, based on the model specifications tools. i.e., PACF and ACF plots, EACF and BIC table, we effectively found the probable ARIMA(p,d,q) models. The set of the candidate models were ARIMA (0,1,0), ARIMA (0,1,1), ARIMA (1,1,1), ARIMA (3,1,0), ARIMA (3,1,1), ARIMA (3,1,2) and ARIMA (3,1,3). Based on the residual approach, ARIMA (3,1,2) model was used to model the time-series data of unemployment rate of Australia for over a decade (2010-2021). Several different orders of p and q were selected to enhance range of models for the ordinary component. Using combination of model diagnostics techniques like AIC and BIC score along with residual analysis, we came at the best fitted ARIMA (3,1,2) model. Supported by overfitting analysis and RMSE, ARIMA (3,1,2) is the best fit model for the series. The final section of the report forecasts the unemployment rate of Australia from May 2021 – Feb 2022. The report acknowledges the short coming and proposes solutions for future research.

# Reference:

Central, D. S. (2017). Arima- Sarima- LSTM for forecasting. Retrieved from https://www.datasciencecentral.com/profiles/blogs/arima-sarima-vs-lstm-with-ensemble-learning-insights-for-time-ser)

D.Cryer, J. &.-S. (2008). *Time Series Analysis with applications in R.* Springer.

Education, E. (2021). Unemployment: Its Measurement and Types.

forecast, A. l. (2016). *Another look at measures of forecast .*

Towards Data Science . (2018). *Limitations of Arima dealing with Outliers* . Retrieved from https://towardsdatascience.com/limitations-of-arima-dealing-with-outliers-30cc0c6ddf33)

# Appendix:

This section contains the R Code.

library(TSA)

library(fUnitRoots)

library(forecast)

library(CombMSC)

library(lmtest)

library(fGarch)

library(rugarch)

library(tseries)

library(FitAR)

sort.score <- function(x, score = c("bic", "aic")){

if (score == "aic"){

x[with(x, order(AIC)),]

} else if (score == "bic") {

x[with(x, order(BIC)),]

} else {

warning('score = "x" only accepts valid arguments ("aic","bic")')

}

}

residual.analysis <- function(model, std = TRUE,start = 2, class = c("ARIMA","GARCH","ARMA-GARCH", "garch", "fGARCH")[1]){

library(TSA)

library(FitAR)

if (class == "ARIMA"){

if (std == TRUE){

res.model = rstandard(model)

}else{

res.model = residuals(model)

}

}else if (class == "GARCH"){

res.model = model$residuals[start:model$n.used]

}else if (class == "garch"){

res.model = model$residuals[start:model$n.used]

}else if (class == "ARMA-GARCH"){

res.model = model@fit$residuals

}else if (class == "fGARCH"){

res.model = model@residuals

}else {

stop("The argument 'class' must be either 'ARIMA' or 'GARCH' ")

}

par(mfrow=c(3,2))

plot(res.model,type='o',ylab='Standardised residuals', main="Time series plot of standardised residuals")

abline(h=0)

hist(res.model,main="Histogram of standardised residuals")

qqnorm(res.model,main="QQ plot of standardised residuals")

qqline(res.model, col = 2)

acf(res.model,main="ACF of standardised residuals")

print(shapiro.test(res.model))

k=0

LBQPlot(res.model, lag.max = 30, StartLag = k + 1, k = 0, SquaredQ = FALSE)

par(mfrow=c(1,1))

}

unemp <- read.csv("dataset.csv", header = TRUE)

unemp\_ts <- ts(unemp$value, start = c(2010,1), end = c(2021,4), frequency = 12)

unemp\_ts

plot(unemp\_ts, ylab='Unemployment Rate', xlab='Year', type='o', main='Time Series plot of unemployment rate in Australia')

# Checking the correlation

y = unemp\_ts

x = zlag(y)

index = 2:length(x)

plot(y=y[index], x=x[index], ylab='Unemployment Rate', xlab = "Previous year's Unemployment Rate", main = 'Scatter plot of Unemployment Rate in consecutive years')

cor(y[index], x[index])

# plotting acf plot to check for stationarity

acf(unemp\_ts, main = 'ACF for Unemployment Rate series')

# ADF test for stationarity

library(tseries)

adf.test(unemp\_ts)

# Test for Volatility

McLeod.Li.test(y=unemp\_ts, main = 'McLeod-Li test for monthly return series')

#Checking Q-Q plot to confirm Volatility

qqnorm(unemp\_diff)

qqline(unemp\_diff, col = 2, lwd = 1, lty = 2)

# Log Tranformation

unemp\_log <- log(unemp\_ts)

plot(unemp\_log, ylab='Unemployment Rate', xlab='Year', type='o', main='Log Transformed Series', ylim=c(1.2,2))

# Box-Cox Tranformation

BC <- BoxCox.ar(y = unemp\_ts, lambda = seq(-2,2,0.1))

BC$ci

mean(BC$ci)

lambda <- mean(BC$ci)

unemp\_tsBC <- ((unemp\_ts^lambda)-1)/lambda

plot(unemp\_tsBC, ylab='Unemployment Rate', xlab='Year', type='o', main='Box-Cox Transformed Series')

# First order differencing

unemp\_diff <- diff(unemp\_log, differences = 1)

plot(unemp\_diff, ylab='Unemployment Rate', xlab='Year', type='o', main='First Difference Series')

# Checking stationarity of first differencing using adf test

adf.test(unemp\_diff)

pp.test(unemp\_diff)

# Model Specification

# ACF & PACF plot

par(mfrow=c(1,2))

acf(unemp\_diff, ci.type="ma", main = "ACF of unemployment series.")

pacf(unemp\_diff, main = "PACF of unemployment series.")

# EACF plot

eacf(unemp\_diff)

# BIC plot

bic\_ms = armasubsets(y=unemp\_diff,nar=5,nma=5,y.name='test',ar.method='ols')

plot(bic\_ms)

# Model Fitting & Residual Analysis

model\_010\_css = arima(unemp\_log,order=c(0,1,0),method='CSS')

coeftest(model\_010\_css)

model\_010\_ml = arima(unemp\_log,order=c(0,1,0),method='ML')

coeftest(model\_010\_ml)

residual.analysis(model = model\_010\_ml)

model\_011\_css = arima(unemp\_log,order=c(0,1,1),method='CSS')

coeftest(model\_011\_css)

model\_011\_ml = arima(unemp\_log,order=c(0,1,1),method='ML')

coeftest(model\_011\_ml)

residual.analysis(model = model\_011\_ml)

model\_111\_css = arima(unemp\_log,order=c(1,1,1),method='CSS')

coeftest(model\_111\_css)

model\_111\_ml = arima(unemp\_log,order=c(1,1,1),method='ML')

coeftest(model\_111\_ml)

residual.analysis(model = model\_111\_ml)

model\_310\_css = arima(unemp\_log,order=c(3,1,0),method='CSS')

coeftest(model\_310\_css)

model\_310\_ml = arima(unemp\_log,order=c(3,1,0),method='ML')

coeftest(model\_310\_ml)

residual.analysis(model = model\_310\_ml)

model\_311\_css = arima(unemp\_log,order=c(3,1,1),method='CSS')

coeftest(model\_311\_css)

model\_311\_ml = arima(unemp\_log,order=c(3,1,1),method='ML')

coeftest(model\_311\_ml)

residual.analysis(model = model\_311\_ml)

model\_312\_css = arima(unemp\_log,order=c(3,1,2),method='CSS')

coeftest(model\_312\_css)

model\_312\_ml = arima(unemp\_log,order=c(3,1,2),method='ML')

coeftest(model\_312\_ml)

residual.analysis(model = model\_312\_ml)

model\_313\_css = arima(unemp\_log,order=c(3,1,3),method='CSS')

coeftest(model\_313\_css)

model\_313\_ml = arima(unemp\_log,order=c(3,1,3),method='ML')

coeftest(model\_313\_ml)

residual.analysis(model = model\_313\_ml)

# AIC & BIC Table

sc.AIC=AIC(model\_010\_ml, model\_011\_ml, model\_111\_ml,model\_310\_ml,model\_313\_ml, model\_311\_ml, model\_312\_ml)

sc.BIC=AIC(model\_010\_ml, model\_011\_ml, model\_111\_ml,model\_310\_ml,model\_313\_ml, model\_311\_ml, model\_312\_ml, k=log(length(unemp\_ts)))

sort.score(sc.AIC, score = "aic")

sort.score(sc.BIC, score = "aic")

# Overfitted models ARIMA(4,1,2) & ARIMA(3,1,3)

model\_412\_css = arima(unemp\_log,order=c(4,1,2),method='CSS')

coeftest(model\_412\_css)

model\_412\_ml = arima(unemp\_log,order=c(4,1,2),method='ML')

coeftest(model\_412\_ml)

residual.analysis(model = model\_412\_ml)

# Forecasting unemeployment of next 10 months

library(forecast)

fit = Arima(unemp\_log,c(3,1,2))

summary(fit)

forecastlog=forecast(fit,h=10)

# Converting forecasted log transformed predictions back to normal

forecastraw=exp(forecastlog$mean)

forecastraw

# Plotting the forecast

par(mfrow=c(1,1))

plot(unemp\_ts, xlim=c(2010,2022.5), ylab='Unemployment rate', xlab='Year', main='Forcast of unemployment rate for next 10 months')

lines(ts(as.vector(forecastraw), start=c(2021,5), frequency=12), col="red", type="l")