Time Series Analysis

Unemployment Rate in Australia

Project Report

Time Series Project

Arvindh Bharadwaj Venkatesan

2024-06-16

Table of Contents

Introduction

Australia has witnessed a number of economic fluctuations in recent decades, all of which have had a substantial influence on the labor market. Notable among these is the global financial crisis of 2008-2010, which had a significant impact on unemployment rates, particularly among young people aged 15 to 24. During economic downturns, this population is more vulnerable.

This time series study seeks to estimate the unemployment rates of Australians aged 15 to 24 from November 2002 to November 2011. The goal is to evaluate historical trends and anticipate future unemployment rates for this demographic while taking into account key economic events that occurred during this time period. By studying this data, we want to generate insights that might help predict future unemployment patterns and influence policy choices. The analysis will also involve projecting unemployment rates for the next ten months after November 2011.

Methodology

The data selected for this assignment was sourced from the Australian Bureau of Statistics' Unemployment Rate report [1]. It represents the unemployment rate of people aged 15 to 24 in Australia from November 2002 to November 2011. According to the source, employment in Australia typically increases over the Christmas/New Year period, which may present challenges for perfect modeling. In this report, we will first explore the data in depth and then proceed with appropriate modeling steps.

The data did not have any missing values, and the only pre-processing required was parsing numbers from strings to numeric format, which was done in Excel. The Exploratory Data Analysis (EDA) involved generating time series plots to visualize the unemployment rate over time and calculating descriptive statistics such as mean, median, and standard deviation to summarize the data. To account for significant economic interventions, the unemployment data was split into two periods: 2001 to 2008 and 2008 to 2011. This was done to capture the impact of the Global Financial Crisis.

For each period, stationarity was ensured using the the Augmented Dickey-Fuller (ADF) test was applied, and seasonal differencing was used to remove seasonal trends. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots were examined to identify significant lags, aiding in determining the initial values for the SARIMA model parameters. An initial SARIMA model was then specified based on insights from the ACF and PACF plots, incorporating parameters for autoregressive (AR), differencing (I), and moving average (MA) components for both the regular and seasonal parts of the series. Model parameters were estimated using the Maximum Likelihood Estimation (MLE) method, and several candidate models were fitted with variations in the number of AR and MA terms.

Residual analysis was performed to check for any remaining autocorrelation, ensuring that residuals behaved like white noise. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to compare the goodness-of-fit for different models,

with lower values being preferred. The best-fitting model for each period was then used to generate forecasts for future unemployment rates, producing forecasts for the next 10 months from December 2011. The accuracy of these forecasts was evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

The null hypothesis H_0 and alternative hypothesis H_1 for the significance tests used in this analysis are as follows:

Shapiro-Wilk Test for Normality

- **Null Hypothesis (H0)**: $\rho \geq 0.05$ (Distribution of the series is normal)
- **Alternative Hypothesis (H1)**: $\rho < 0.05$ (Distribution of the series is not normal)

Correlation Analysis

- **Null Hypothesis (H0)**: $r = 0$ (There is no association)
- **Alternative Hypothesis (H1)**: $r \neq 0$ (A non-zero correlation could exist)

Augmented Dickey-Fuller Unit-Root Test for Stationarity

- **Null Hypothesis (H0)**: $\rho \geq 0.05$ (Time series data is non-stationary)
- **Alternative Hypothesis (H1)**: $\rho < 0.05$ (Time series data is stationary)

Phillips-Perron Unit-Root Test for Stationarity

- **Null Hypothesis (H0)**: $\rho \geq 0.05$ (Time series data is non-stationary)
- **Alternative Hypothesis (H1)**: $\rho < 0.05$ (Time series data is stationary)

The Phillips-Perron (PP) test is another method used to test for the presence of a unit root in a time series, similar to the ADF test but with adjustments for serial correlation and heteroskedasticity.

KPSS Test for Stationarity

- **Null Hypothesis (H0)**: $\rho \leq 0.05$ (Time series data is stationary)
- **Alternative Hypothesis (H1)**: $\rho > 0.05$ (Time series data is non-stationary)

Box-Ljung Test for the Residual Analysis

- **Null Hypothesis (H0)**: $\rho \geq 0.05$ (Error terms are uncorrelated)
- **Alternative Hypothesis (H1)**: $\rho < 0.05$ (Error terms are correlated)

- All statistical analyses were conducted at a significance level of 5%. RStudio, along with additional packages, was utilized to perform the computations and generate visual representations presented throughout this report

1: Monthly Australia Unemployment Rate from 2002 to 2011 among people aged 15 – 24

1.1. Descriptive Analysis

```
# Read Data

# Data Retrieval
unemployment_rate <- read.csv("Unemployment rate of persons aged 15-24 - Orig
inal.csv",col.names =c('Date','unemp_rate'))
head(unemployment_rate)
```

```
##     Date unemp_rate
## 1 Nov-02       11.4
## 2 Dec-02       12.5
## 3 Jan-03       13.6
## 4 Feb-03       13.5
## 5 Mar-03       13.3
## 6 Apr-03       12.5
```

```
# Summary Statistics

summary(unemployment_rate$unemp_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.40    9.80   10.80   10.74   11.50   13.60
```

The Summary statistics of the unemployment rate data shows a slight variation in the min and max levels. There was a noticeable range in the data, with the unemployment rate ranging from 7.40% (min) to 13.60% (max). These numbers show how this age group's unemployment rate changed throughout the specified period of time.

Distribution of the data could be found using histogram.

```
hist(unemployment_rate$unemp_rate, breaks = 15, col = "lightblue", main = "Hi
stogram of Unemployment rate of persons aged 15-24", xlab = "Unemployment Rat
e", ylab = "Frequency")
```

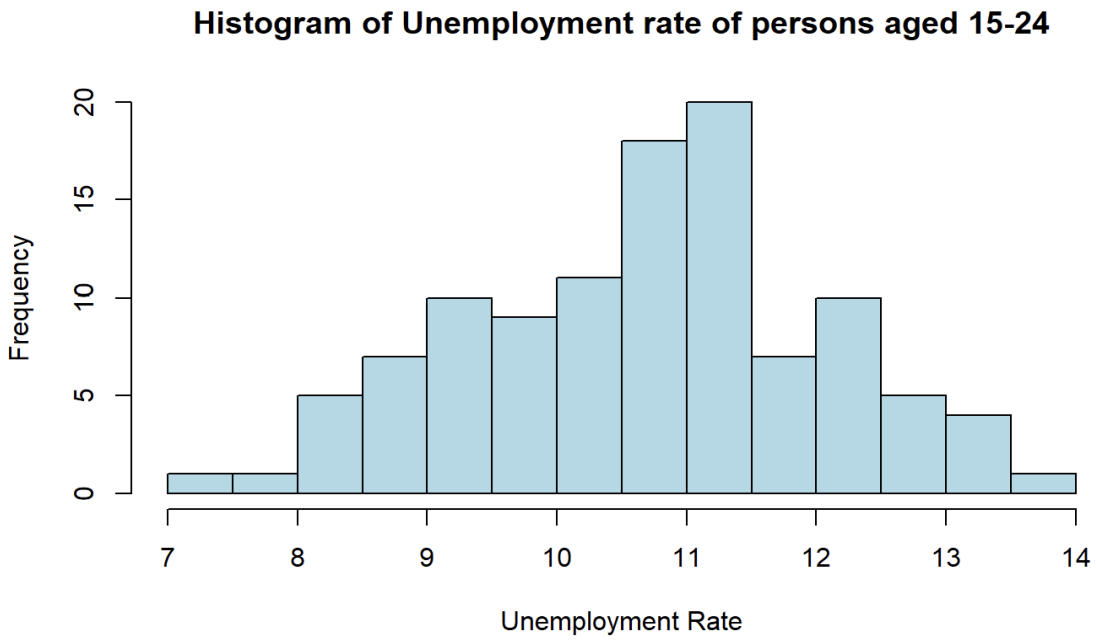

Figure 1: Histogram of Unemployment rate of persons aged 15-24

According to Figure 1, the histogram of the given data appears to be fairly normal, with a slight left skewness. This skewness might be attributed to a change point in the data, as shown in Figure 2.

```
# Convert data into a time series object
unemployment.ts <-  as.vector(unemployment_rate$unemp_rate) # Converting to a
vector
unemployment.ts <-  ts(unemployment.ts, start = c(2002,11), frequency = 12)
class(unemployment.ts)

## [1] "ts"

# Plot with time series plot with mean line
plot(unemployment.ts,ylab='Unemployment Rate',xlab='Year',type='o', main = "U
nemployment rate of people aged 15-24")
mean_unemployment <- round(mean(unemployment.ts),6)
abline(h = mean_unemployment, col = "red", lty = 2)
text(x = max(time(unemployment.ts)), y = mean_unemployment, labels = paste("M
ean:", mean_unemployment), pos = 3, col = "red")
```

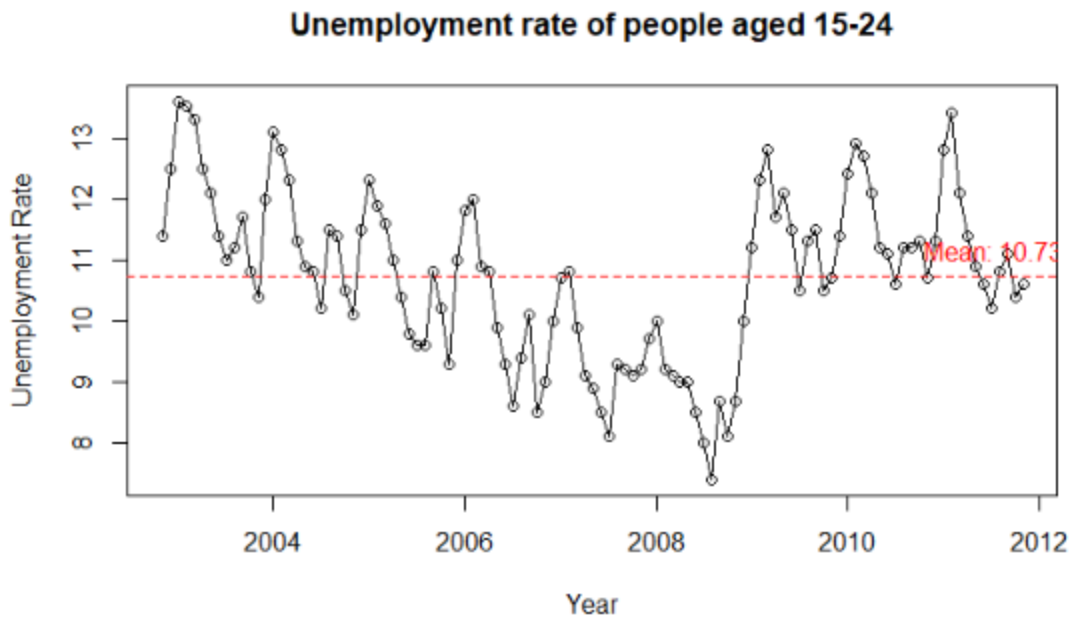

Figure 2: Time Series Characteristics Plot

Time Series Plot Characteristics (Figure 2):

1. **Trend**: The time series plot reveals a clear trend. Initially, there is a decreasing trend, followed by a recovery that stabilizes over time.

2. **Seasonality**: Based on visual interpretation, there is a clear indication of seasonality in the data.

3. **Changing Variance**: There is slight changing variance.

4. **Behavior**: The plot predominantly demonstrates moving average (MA) behavior, with a few auto regressive (AR) points in it.

5. **Intervention/ Change Point**: There is noticeable increase in the unemployment rate observed after mid-2008 could be attributed to significant economic events such as the "Labour Market Downturn" in Australia [2] and the global impact of the Great Recession on the international and Australian economy [3]. These events likely serve as key reasons for the change point in our data during that period.

6. **Auto Correlation**: The scatter plot of Unemployment rate to its first lag as shown in Figure 3 indicates strong positive auto correlation between observation of successive seasons.

```
# First Lag Correlation
y = unemployment.ts
x = zlag(unemployment.ts)
# Plot for First Lag
plot(y=unemployment.ts,x=zlag(unemployment.ts),ylab='Unemployment Rate', xlab
```

```
='Previous Observation Unemployment Rate' , main = "Scatter plot of Unemploym
ent rate of people aged 15-24")
```

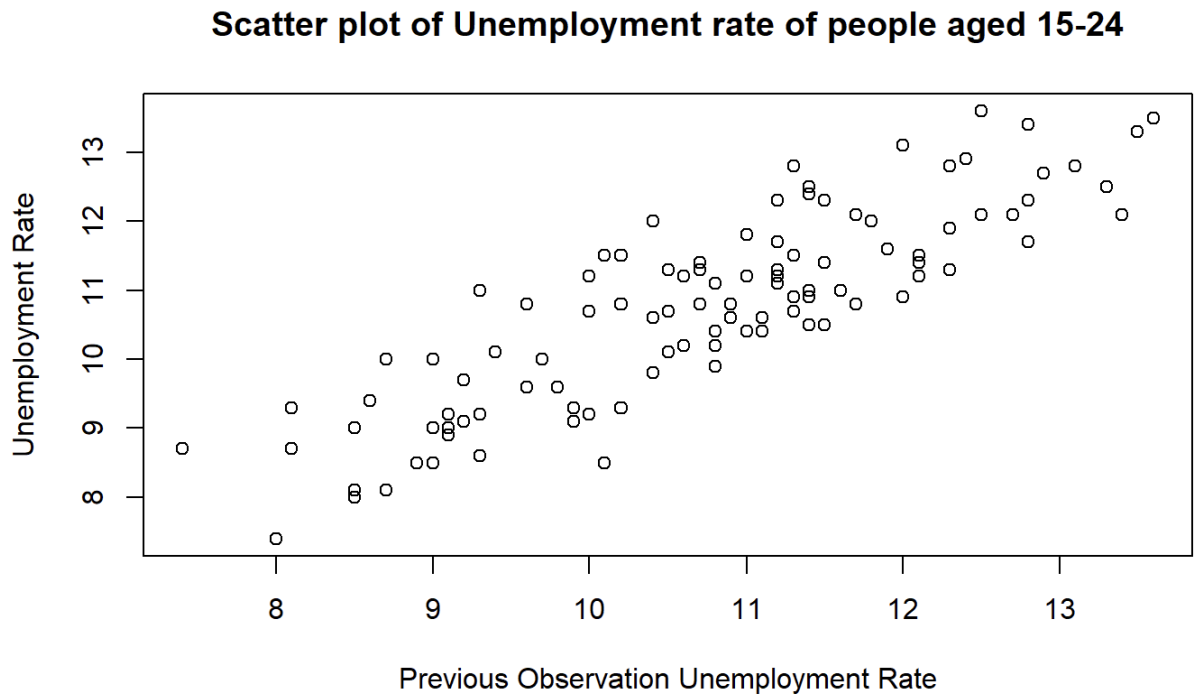

Figure 3: Scatter plot of Unemployment rate of people aged 15-24

1.1.1. Checking for Seasonality

Visually identifying seasonal patterns in the time series plot can be further validated using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.

We will utilize custom functions that we have created and provided in the course to enhance code readability. Please refer to the "Appendix - Useful User-defined Functions" for details.

```
# ACF and PACF
plot_acf_pacf(unemployment.ts) # clear seasonality
```

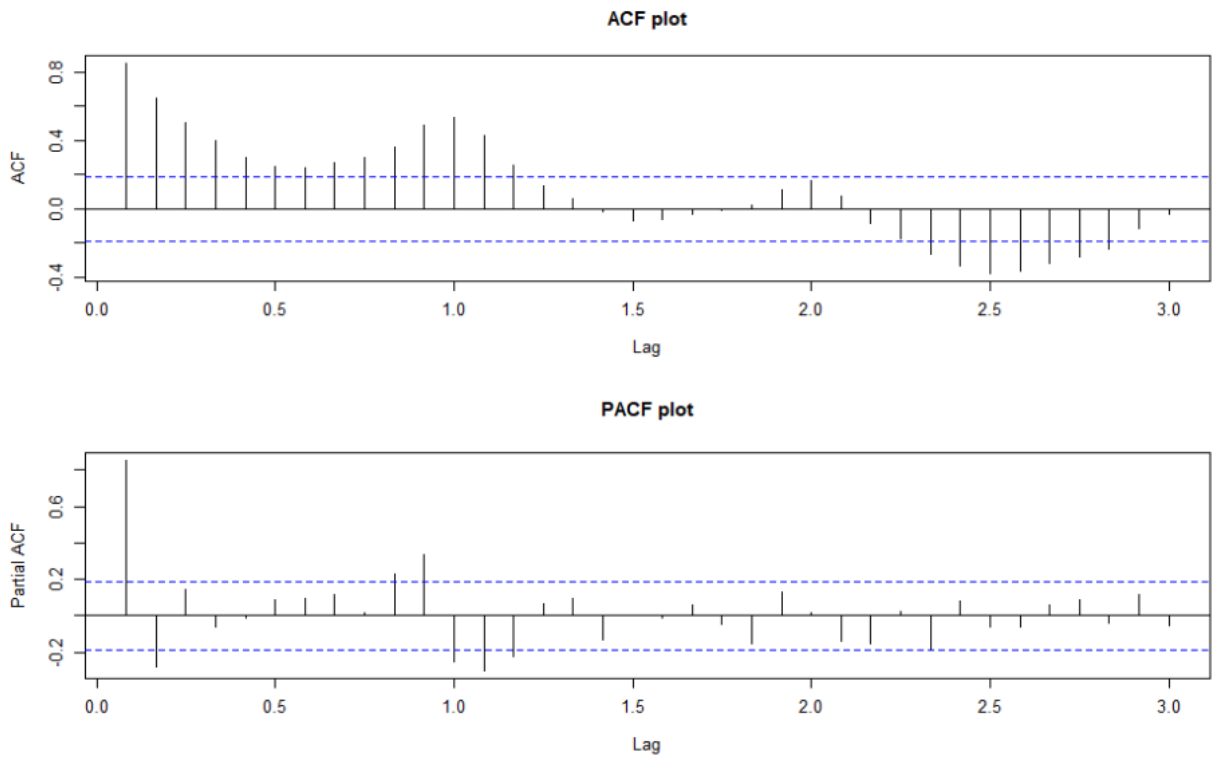

Fig 4: ACF and PACF plots of unemployment TS object

ACF and PACF plots in Figure 4 show a clear wave pattern, indicating seasonality. Additionally, the decay in the seasonal lags suggests a trend in the data, highlighting its non-stationarity.

1.1.2. Test for Normality

```
qq_shapiro_function(unemployment.ts) # custom function
```

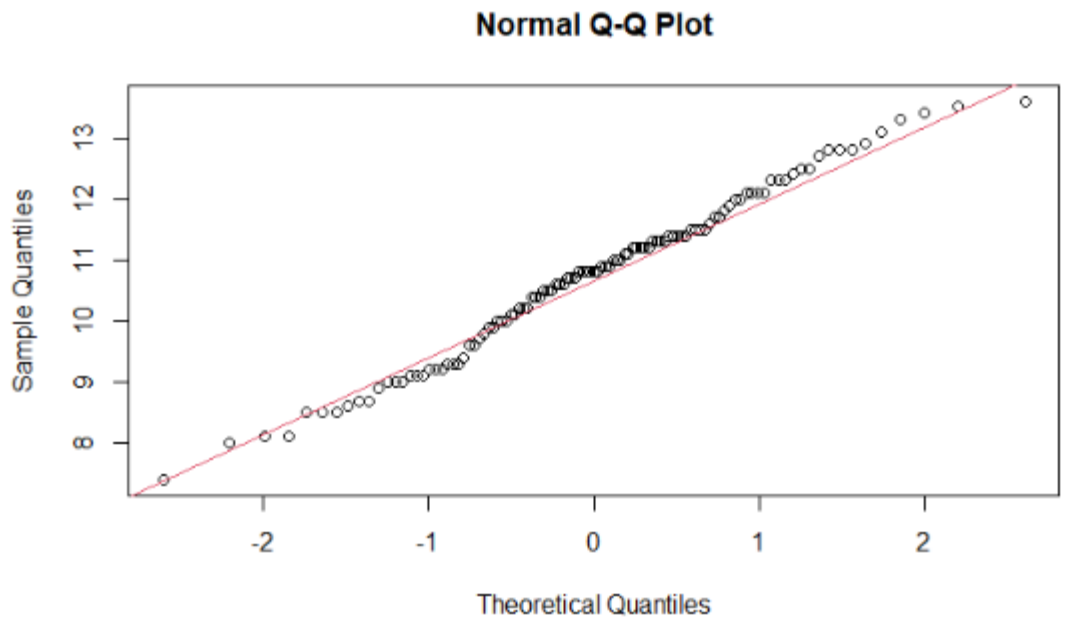

Figure 5: Q-Q Plot for Normality

```
##
##  Shapiro-Wilk normality test
##
## data:  ts_object
## W = 0.98752, p-value = 0.4098
```

In the Q-Q plot (Figure 5), the data point are mostly aligning with the center line (expect for the both ends) which suggest the data is following normal distribution.

- **Null Hypothesis (H0)**: $\rho \geq 0.05$ (Distribution of the series is normal)
- **Alternative Hypothesis (H1)**: $\rho < 0.05$ (Distribution of the series is not normal

With a p-value of 0.4098 from the Shapiro-Wilk test at the 0.05 significance level, we do not have sufficient evidence to reject the null hypothesis. Therefore, based on this test, the data is likely normally distributed.

1.1.3. Test for Stationarity

ADF, PP, and KPSS tests are can used to determine whether a time series is stationary or not.

```
ts_stationary_tests(unemployment.ts)
```

```
## $ADF_Test
##
##  Augmented Dickey-Fuller Test
##
```

```
## data:  ts_object
## Dickey-Fuller = -3.15, Lag order = 4, p-value = 0.09982
## alternative hypothesis: stationary
##
##
## $PP_Test
##
##  Phillips-Perron Unit Root Test
##
## data:  ts_object
## Dickey-Fuller Z(alpha) = -17.996, Truncation lag parameter = 4, p-value
## = 0.09268
## alternative hypothesis: stationary
##
##
## $KPSS_Test
##
##  KPSS Test for Level Stationarity
##
## data:  ts_object
## KPSS Level = 0.39158, Truncation lag parameter = 4, p-value = 0.08079
```

Even though the KPSS test suggests otherwise, the high p-values from the ADF and PP tests indicate that our data is non-stationary. Considering our data exhibits seasonality, we can rely on the ADF and PP tests along with the time series plot to confirm its non-stationarity.

Given our understanding that the data is non-stationary, seasonal, and exhibits a change point, we will develop SARIMA-based models to capture these characteristics. We will refine and improve these models to assess their fit, and select the best model for forecasting. If the results show unsatisfactory large confidence intervals due to the change point, we will consider splitting the data and re-fitting the models accordingly.

1.2. SARIMA Modelling

1.2.1. Model Specifications

Finding Seasonal Components (P, D, Q)

To begin, we will explore seasonal components (P, D, Q) and seek optimal values that can effectively capture both seasonal patterns and trends. If significant variance remains, we will apply appropriate transformation techniques as necessary.

```
# First Differencing (D=1)

m1.unemployment = Arima(unemployment.ts,order=c(0,0,0),seasonal=list(order=c(
0,1,0), period=12))
res.m1 = residuals(m1.unemployment);
plot_residuals_acf_pacf(res.m1,"(0,0,0),(0,1,0)")
```

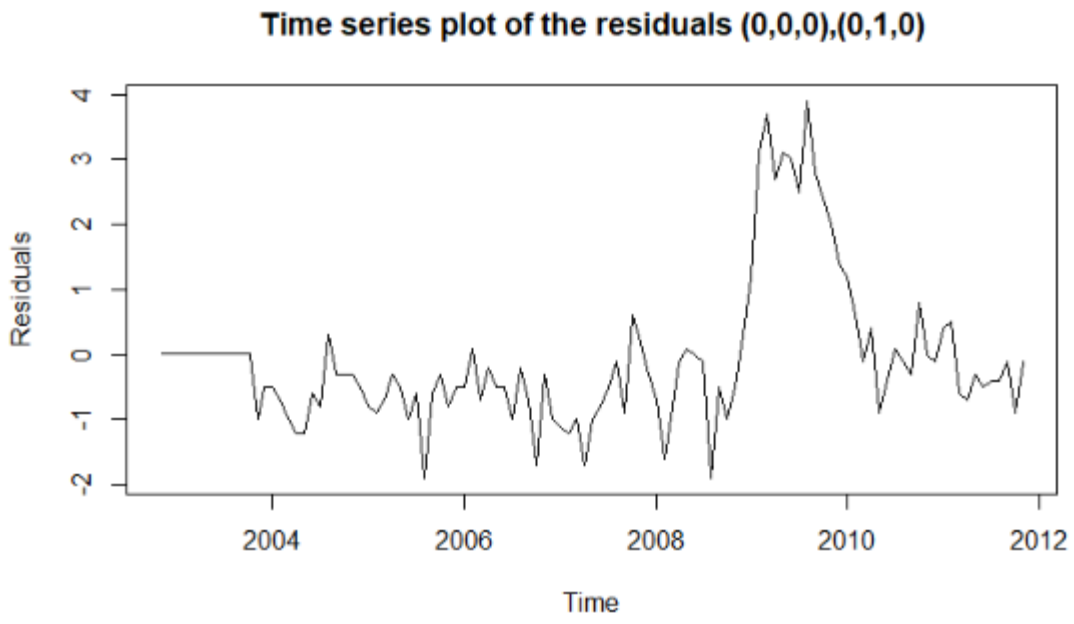

Figure 6: Residual Time Series plot after first seasonal Differencing

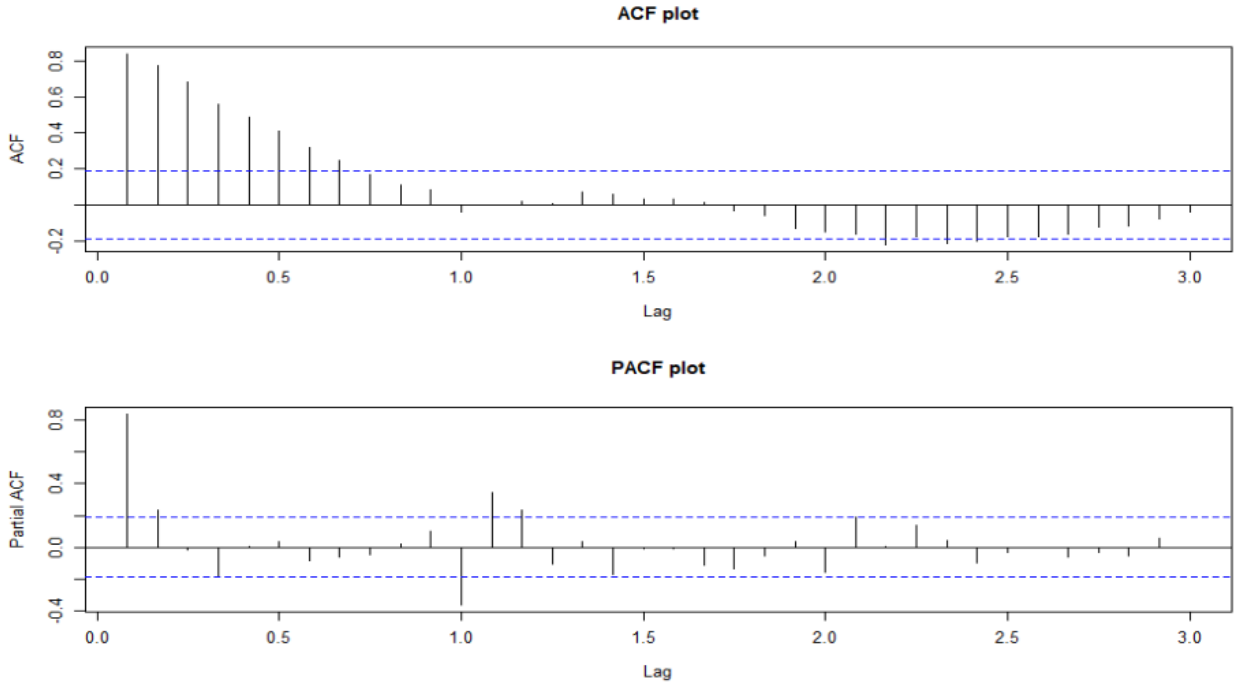

Figure 7: ACF & PACF Residuals plot of Model (0,0,0),(0,1,0):

Seasonal differencing successfully removed the trend. Based on Figure 7, where the ACF shows a decaying pattern, we can set Q = 0 . Similarly, from the PACF plot which indicates a lag at season = 1, we can set P = 1.

```
#So, we will add the SARMA(1,1,0) component and see if we get rid of seasonal
component.
```

```
m2.unemployment = Arima(unemployment.ts,order=c(0,0,0),seasonal=list(order=c(
1,1,0), period=12))
res.m2 = residuals(m2.unemployment);
plot_residuals_acf_pacf(res.m2, "(0,0,0),(1,1,0)")
```

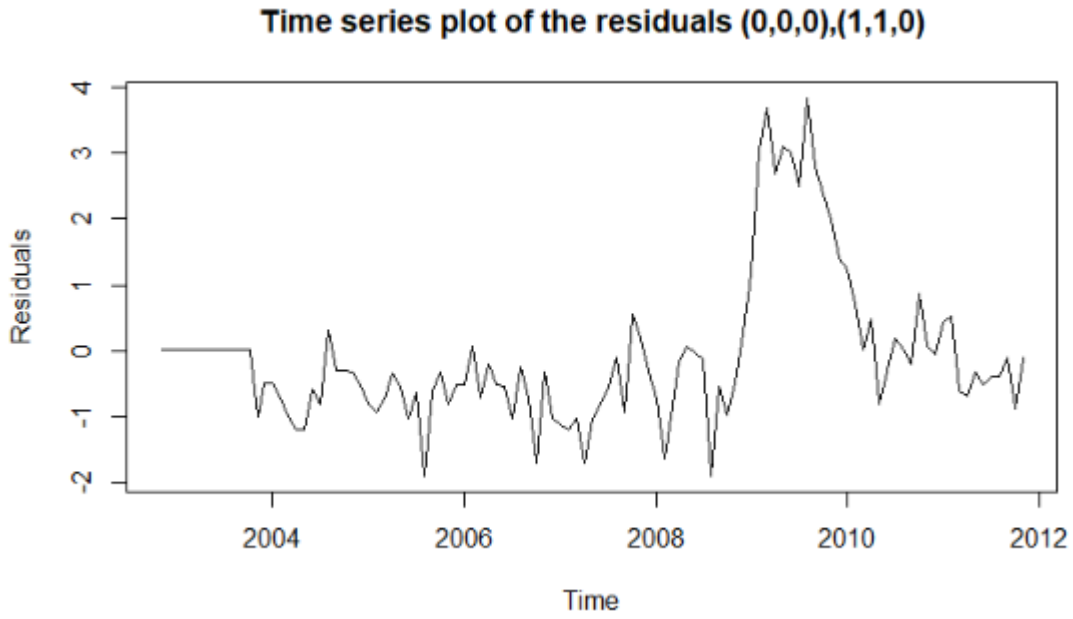

Figure 8: Residual Time Series plot of pdq(0,0,0) and PDQ(1,1,0)

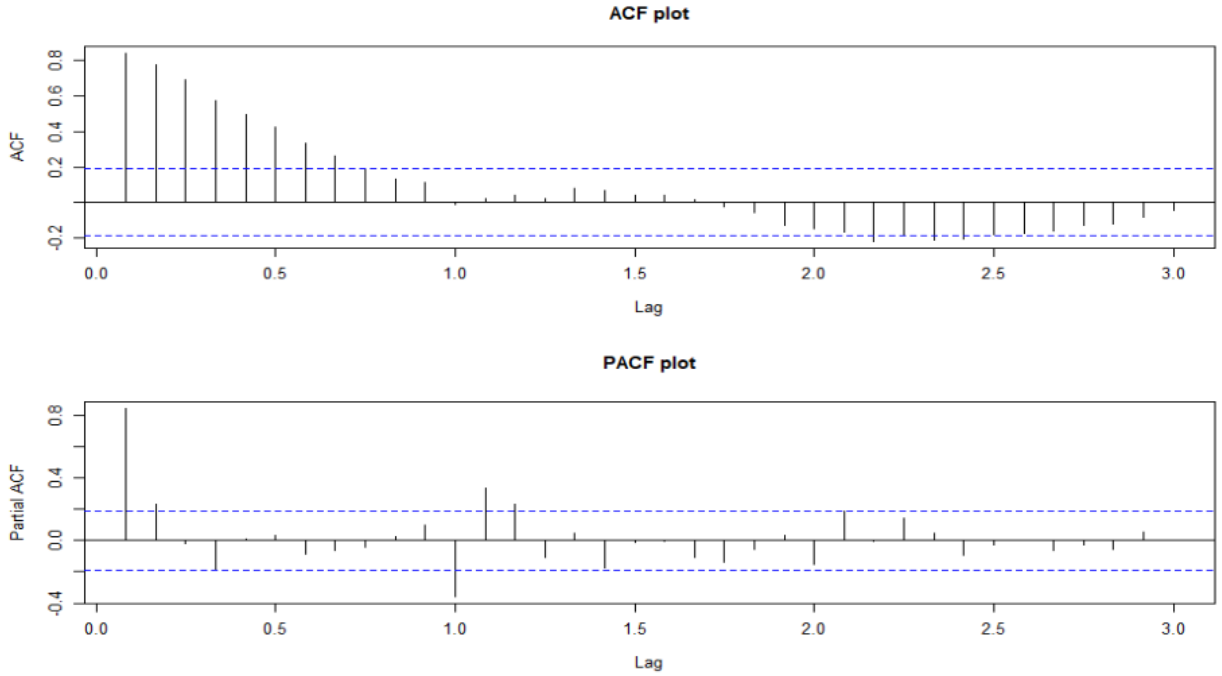

Figure 9: Residual ACF & PACF plot of pdq(0,0,0) and PDQ(1,1,0)

No significant improvement was noted. The PACF reveals a seasonal lag at season = 1, potentially attributable to a change point. To not lose further observations and

acknowledging the presence of residual white noise, we will proceed with this seasonal component and conduct further analysis on seasonality in the residual analysis section.

1.2.2. Finding ARIMA Components (p,d,q)

To address change in variance, we can employ transformation techniques before determining ARIMA components to see any potential improvements.

1.2.3. Box-Cox Transformation

```
# Box-cox transformation
BC <- BoxCox.ar(unemployment.ts)
```

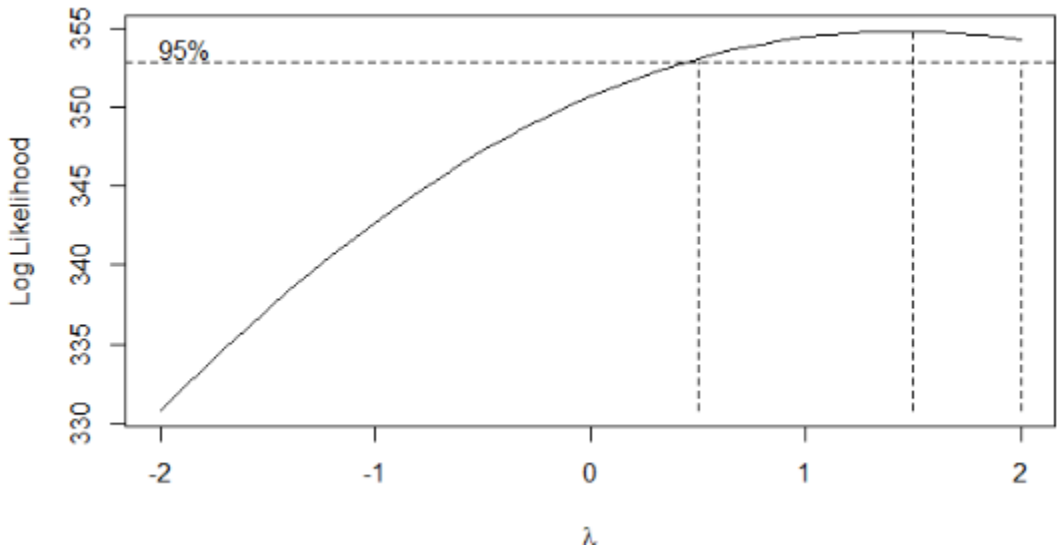

Figure 10: Box-Cox Transformation Log-Likelihood Plot and Confidence Intervals

```
BC$ci

## [1] 0.5 2.0

lambda <- BC$lambda[which(max(BC$loglike) == BC$loglike)]
lambda

## [1] 1.5
```

The obtained lambda value of 1.5 suggests that a Box-Cox transformation could be suitable

```
BC.unemployment.ts = (unemployment.ts^lambda-1)/lambda

m3.unemployment = Arima(BC.unemployment.ts,order=c(0,0,0),seasonal=list(order
=c(1,1,0), period=12))
res.m3 = residuals(m3.unemployment);
plot_residuals_acf_pacf(res.m3,"BC (0,0,0),(1,1,0)")
```

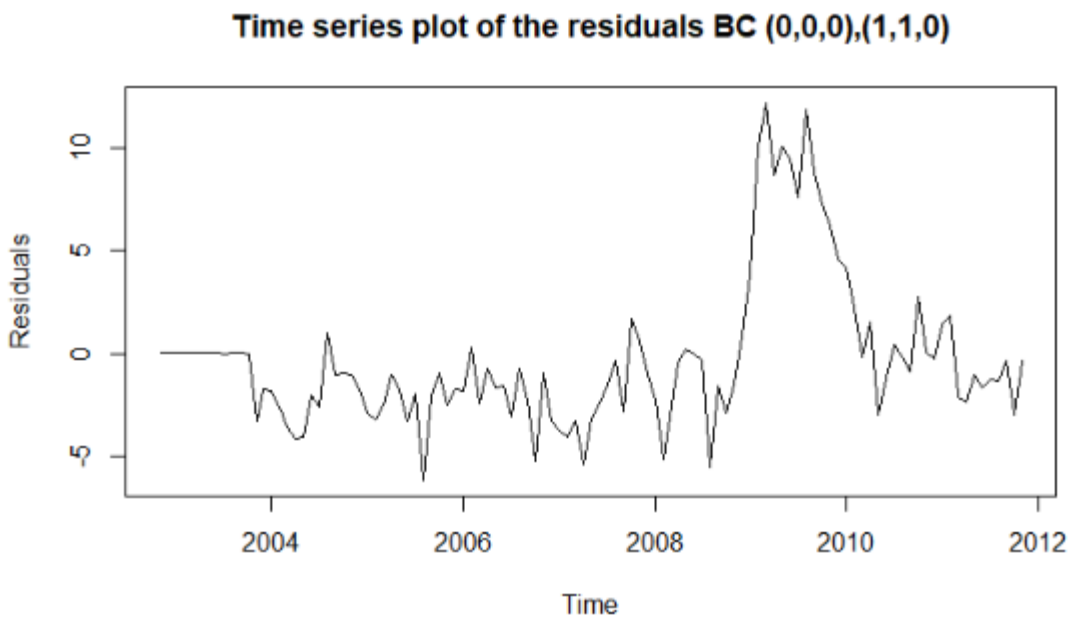

Figure 11: Residual Time Series plot for box-cox transformed data of model (0,0,0),(1,1,0)

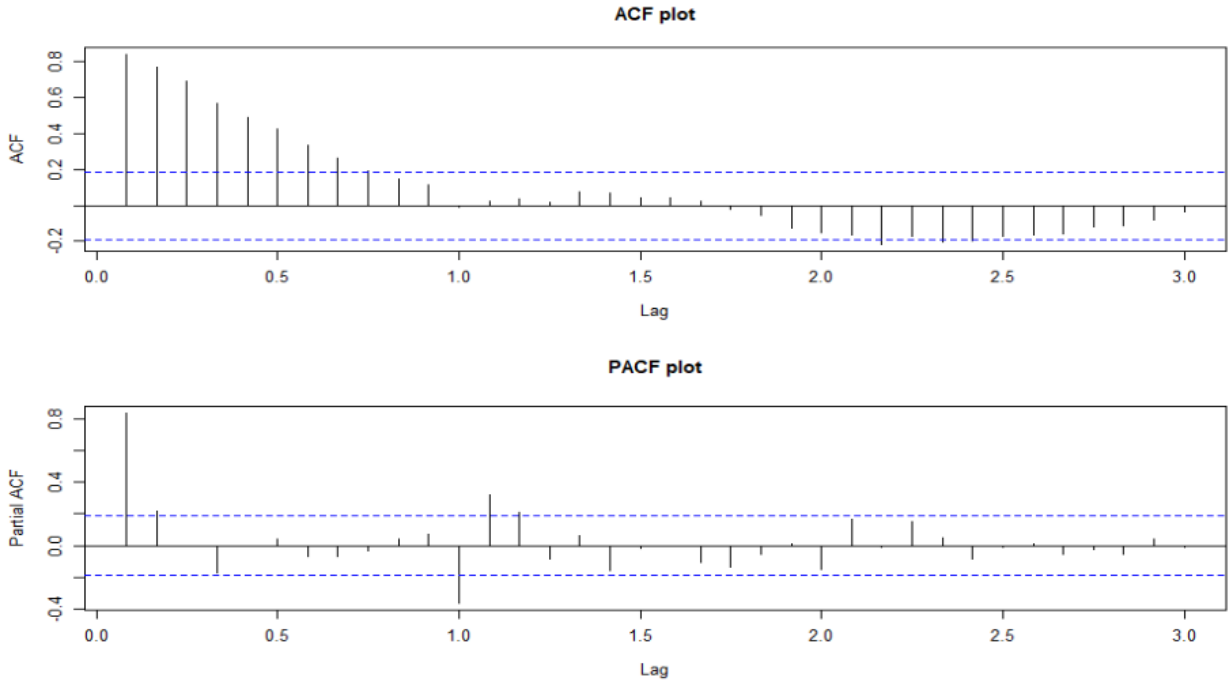

Figure 12: ACF & PACF Residuals plot for box-cox transformed data of model (0,0,0),(1,1,0)

Now, we still observe a decaying pattern in our ACF plot. To capture any remaining non-seasonal trends, we will apply ordinary differencing (d). This is also evident from the p-values of the tests below: ADF test has a p-value greater than 0.1, while KPSS test has a p-value less than 0.1

```
# Stationarity test
ts_stationary_tests(res.m3)
```

```
## $ADF_Test
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_object
## Dickey-Fuller = -2.6977, Lag order = 4, p-value = 0.2872
## alternative hypothesis: stationary
##
##
## $PP_Test
##
##  Phillips-Perron Unit Root Test
##
## data:  ts_object
## Dickey-Fuller Z(alpha) = -18.775, Truncation lag parameter = 4, p-value
## = 0.08053
## alternative hypothesis: stationary
##
##
## $KPSS_Test
##
##  KPSS Test for Level Stationarity
##
## data:  ts_object
## KPSS Level = 0.48939, Truncation lag parameter = 4, p-value = 0.04406
```

```
# Now setting up the ARIMA Component with First Differencing (d=1)

m4.unemployment = Arima(BC.unemployment.ts,order=c(0,1,0),seasonal=list(order
=c(1,1,0), period=12))
res.m4 = residuals(m4.unemployment);
plot_residuals_acf_pacf(res.m4,"(0,1,0),(1,1,0)")
```

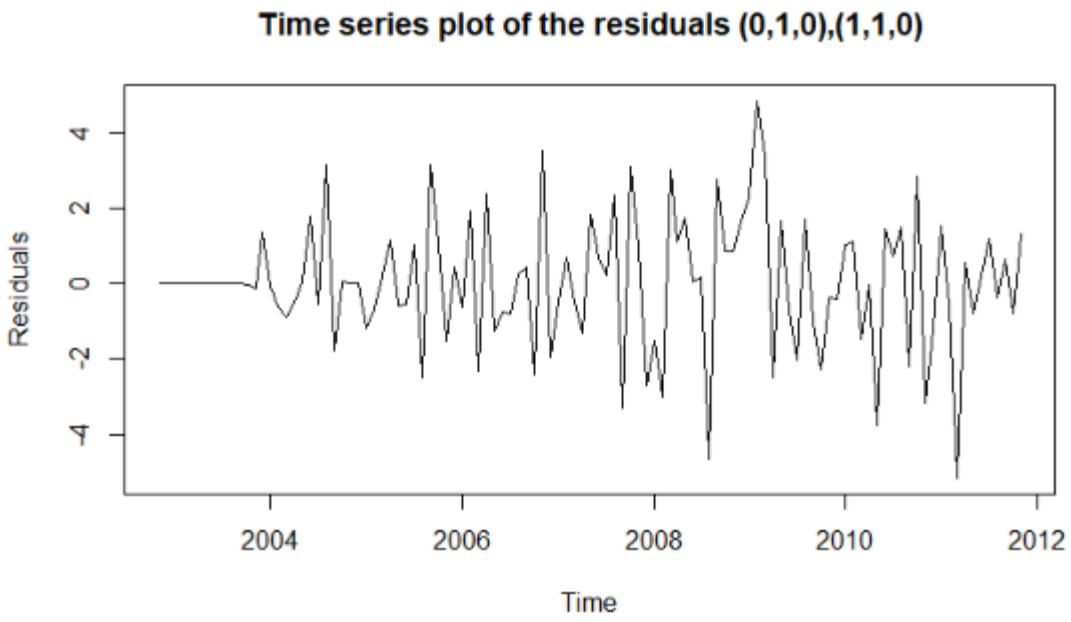

Figure 13: Residual Time Series plot for box-cox transformed data of model (0,1,0),(1,1,0)

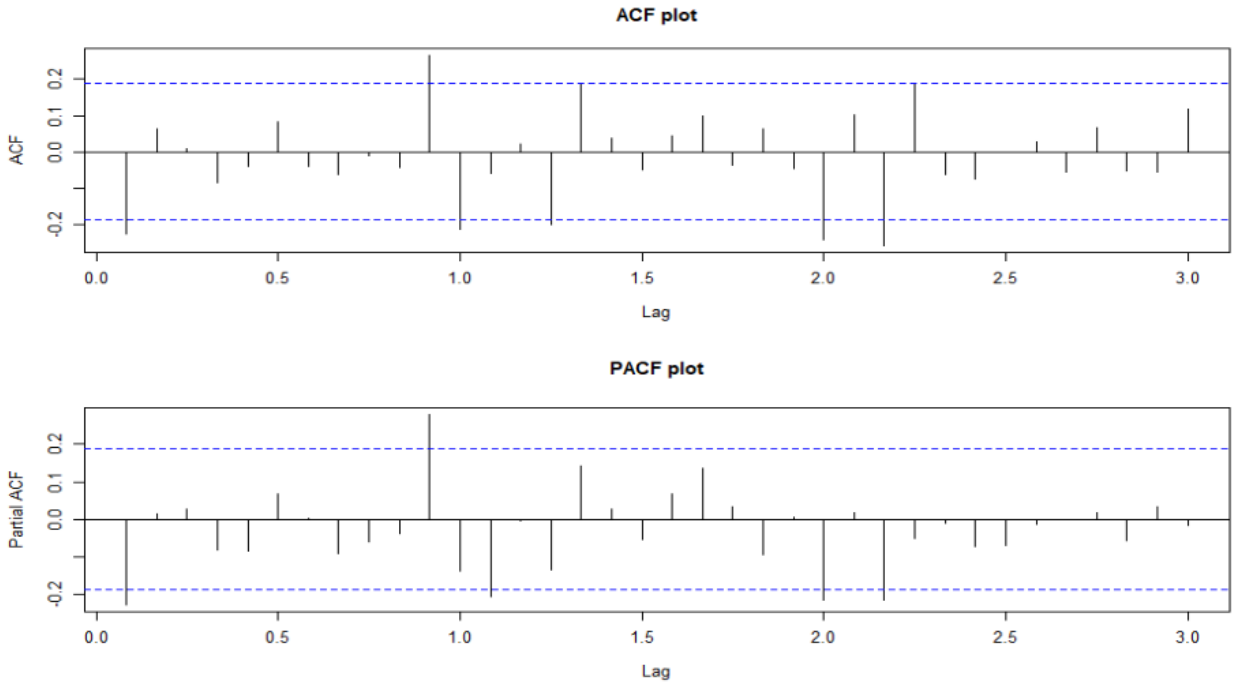

Figure 14: ACF & PACF Residuals plot for box-cox transformed data of model (0,1,0),(1,1,0)

Differencing successfully removed the decaying pattern in the ACF plot. Based on the obtained ACF and PACF plots in Figure 14, we can set p = 2 (PACF) and q = 3 (ACF).

```
m5.unemployment = Arima(BC.unemployment.ts,order=c(2,1,3),seasonal=list(order
=c(1,1,0), period=12))
```

```
res.m5 = residuals(m5.unemployment);
plot_residuals_acf_pacf(res.m5,"(2,1,3),(1,1,0)")
```

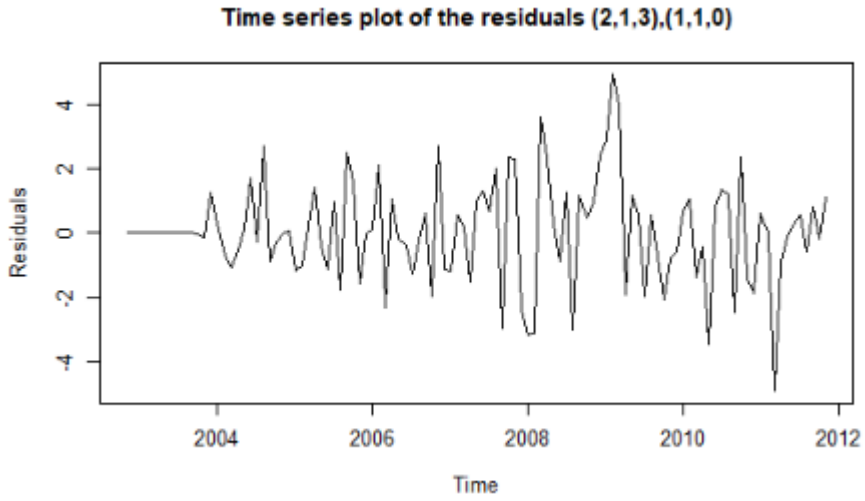

Figure 15: Residual Time Series plot for box-cox transformed data of model (2,1,3),(1,1,0)

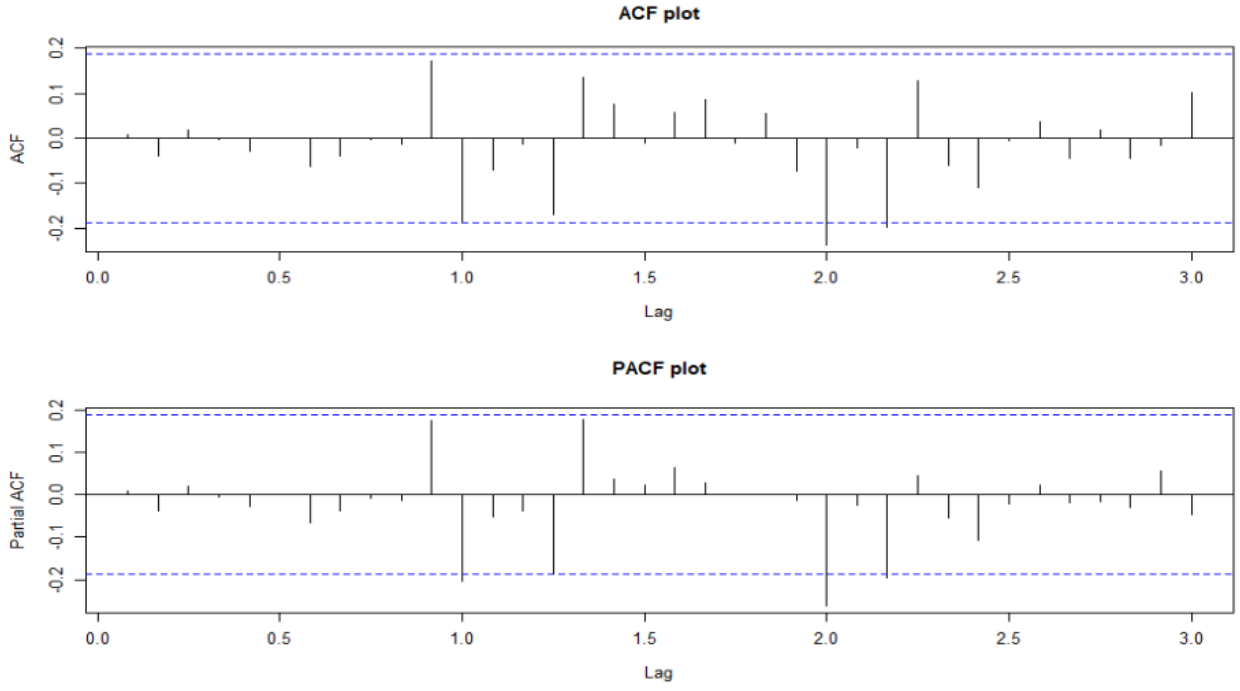

Figure 16: ACF & PACF Residuals plot for box-cox transformed data of model (2,1,3),(1,1,0)

There is still a slight seasonal lag, which could be due to a change point or simply white noise. Therefore, we will proceed with fitting this model and conduct further diagnostic checks to determine its adequacy.

1.2.4. EACF

EACF is used to find more tentative models

```
eacf(res.m4) #residuals of m4 is being used for EACF since there is leftover signal in them.
```

```
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o x  x  o  o
1 o o o o o o o o o o o  x  o  o
2 x o o o o o o o o o o  x  o  o
3 x o o o o o o o o o o  o  o  o
4 x x x x o o o o o o o  o  o  o
5 x x x x o o o o o o o  o  o  o
6 o x x o o o o o o o o  o  o  o
7 o x x o o o o o o o o  o  o  o
```

Figure 17: EACF Table

Based on the results of EACF and considering the topmost left zeros, the selected models are SARIMA(0,1,1)x(1,1,0)_12, SARIMA(0,1,2)x(1,1,0)_12, SARIMA(1,1,1)x(1,1,0)_12 and SARIMA(1,1,2)x(1,1,0)_12

1.2.5. BIC Table

BIC table can be used to further expand the set of possible models

```
# BIC table
bic_table = armasubsets(y=res.m4,nar=5,nma=5,y.name='p',ar.method='ols')

## Reordering variables and trying again:

plot(bic_table)
```

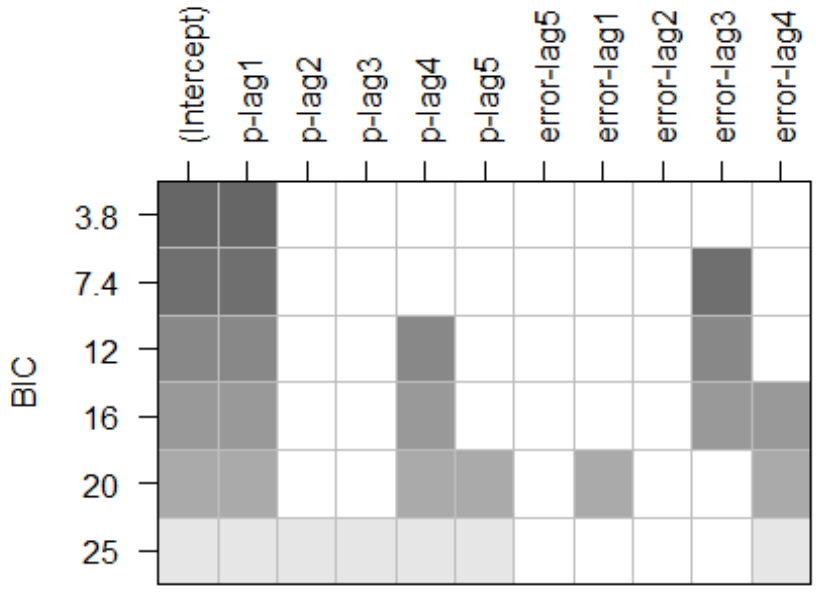

Figure 18: BIC Table

According to BIC table (Figure 18), feasible models are SARIMA(1,1,0)x(1,1,1)_12 and SARIMA(1,1,3)x(1,1,1)_12

Based on all the analyses conducted, the potential models are:

- SARIMA(0,1,1)x(1,1,0)_12
- SARIMA(0,1,2)x(1,1,0)_12
- SARIMA(1,1,1)x(1,1,0)_12
- SARIMA(1,1,2)x(1,1,0)_12
- SARIMA(2,1,3)x(1,1,0)_12
- SARIMA(1,1,0)x(1,1,1)_12
- SARIMA(1,1,3)x(1,1,1)_12

1.3. Model Fitting and Diagoniostics Checking

```
# SARIMA(0,1,1)x(1,1,0)_12
m5_011.unemployment = Arima(BC.unemployment.ts,order=c(0,1,1),seasonal=list(o
rder=c(1,1,0), period=12),method = "ML")
coeftest(m5_011.unemployment)

##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.212411   0.095744 -2.2185   0.02652 *
```

```
## sar1 -0.476371   0.086863 -5.4842 4.154e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m5_011.unemployment)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98124, p-value = 0.128
```

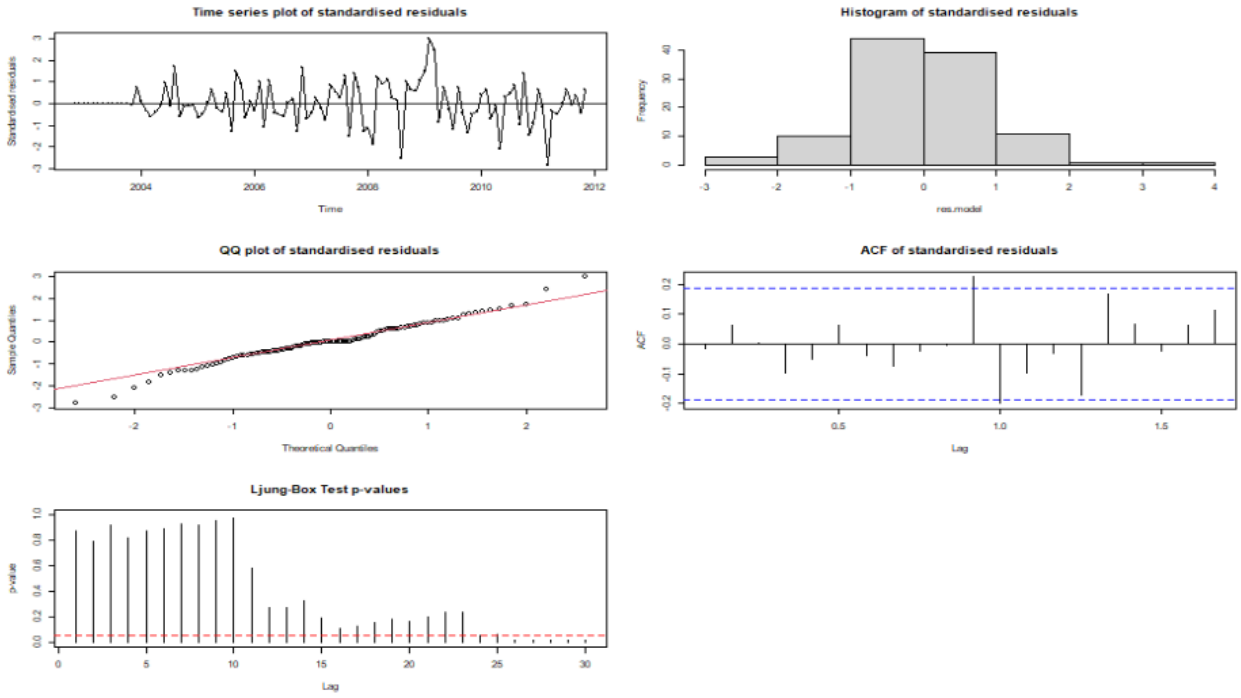

Figure 19: Residual Analysis for SARIMA(0,1,1)x(1,1,0) Model with Maximum Likelihood Estimation

```
m5_011.unemploymentCSS = Arima(BC.unemployment.ts,order=c(0,1,1),seasonal=list(order=c(1,1,0), period=12),method = "CSS")
coeftest(m5_011.unemploymentCSS)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.203021   0.096457 -2.1048   0.03531 *
## sar1 -0.510940   0.089754 -5.6927 1.251e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m5_011.unemploymentCSS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.96617, p-value = 0.007143
```

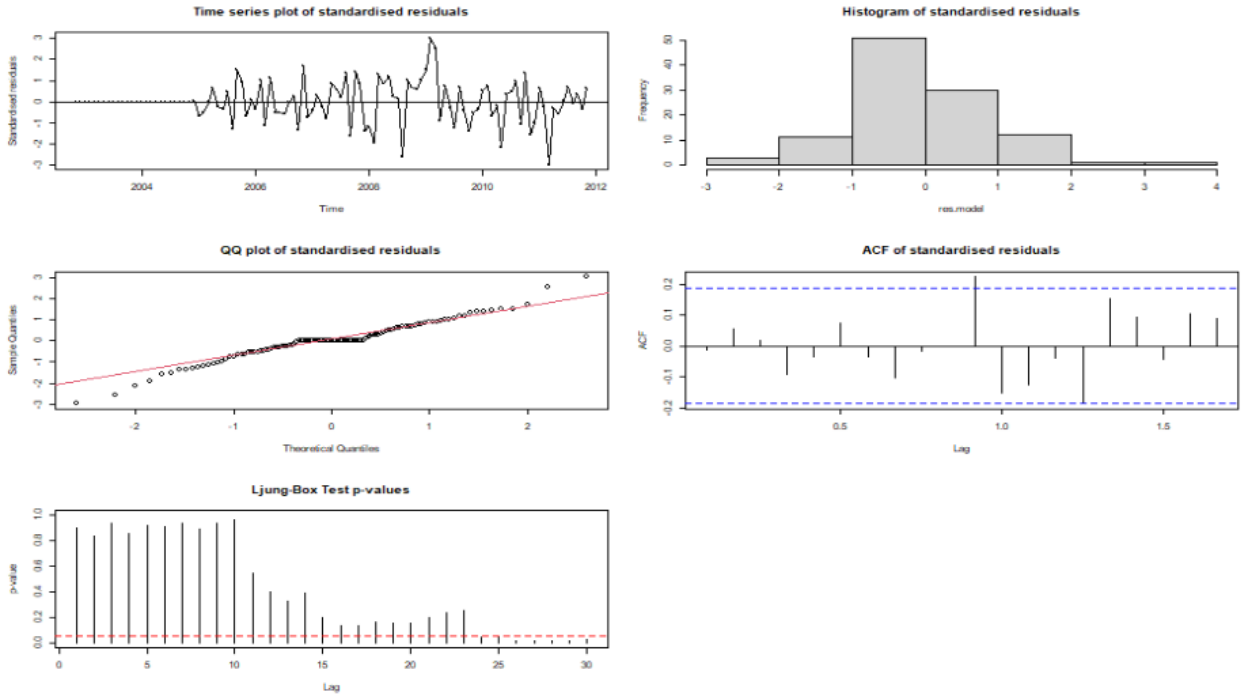

Figure 20: Residual Analysis for SARIMA(0,1,1)x(1,1,0) Model with Conditional Sum-of-Squares Estimation

Both ML and CSS models perform well in capturing seasonality for SARIMA(0,1,1)x(1,1,0)_12, as evidenced by statistically significant p-values for both seasonal and MA components. Most of the residuals appear to be white noise. The histogram is symmetric in both cases, though there are deviations greater than ±3 and both tails are skewed, likely due to the change/intervention point in our data.

The Shapiro-Wilk test yielded a p-value of 0.128 in ML and 0.007 in CSS, indicating insufficient evidence to reject the null hypothesis at an alpha (α) level of 0.05 in ML but not in CSS. This is further supported by the Q-Q plot of standardized residuals in both cases, where ML showing little deviation from the center line and CSS showing slight deviations. We observe only late auto correlations in the ACF plot, and the Ljung-Box test p-values suggest that most observed lags are not significant, at least until the first 25 lags (in both cases). Hence, SARIMA(0,1,1)x(1,1,0)_12 (especially ML) appears to be a suitable model for forecasting. We will further evaluate this based on additional evaluation metrics later.

```
# SARIMA(0,1,2)x(1,1,0)_12
m5_012.unemployment = Arima(BC.unemployment.ts,order=c(0,1,2),seasonal=list(o
rder=c(1,1,0), period=12),method = "ML")
coeftest(m5_012.unemployment)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.221404   0.101428 -2.1829   0.02905 *
## ma2   0.075819   0.114625  0.6615   0.50832
## sar1 -0.475885   0.087373 -5.4466 5.134e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m5_012.unemployment)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98083, p-value = 0.1183
```

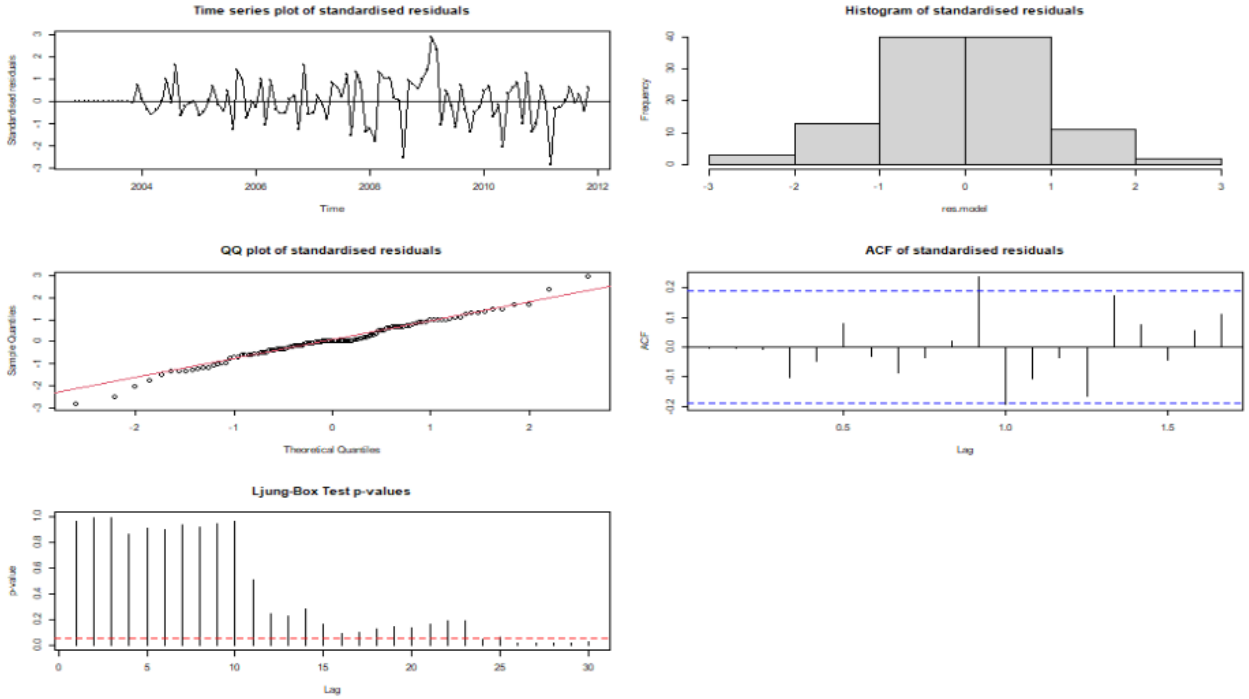

Figure 21: Residual Analysis for SARIMA(0,1,2)x(1,1,0) Model with ML

```
m5_012.unemploymentCSS = Arima(BC.unemployment.ts,order=c(0,1,2),seasonal=lis
t(order=c(1,1,0), period=12),method = "CSS")
coeftest(m5_012.unemploymentCSS)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.212076   0.102207 -2.0750   0.03799 *
```

```
## ma2   0.066742   0.113349  0.5888   0.55598
## sar1 -0.511085   0.090376 -5.6551 1.558e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m5_012.unemploymentCSS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.96474, p-value = 0.005503
```

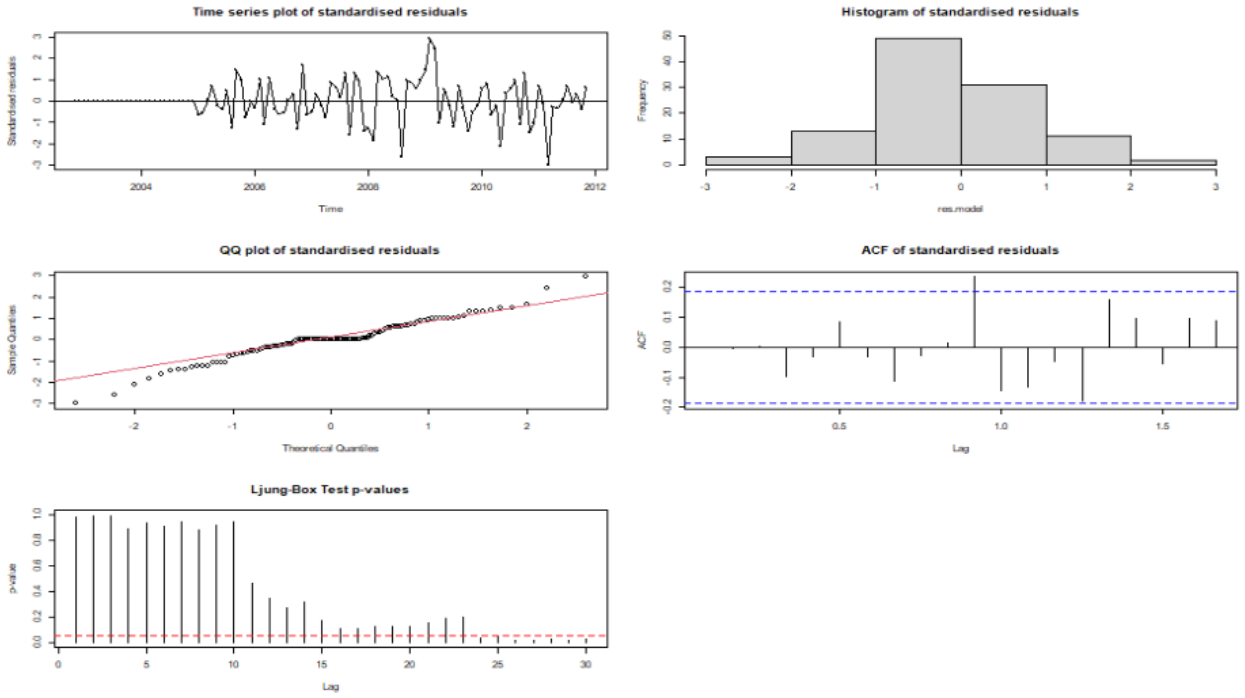

Figure 22: Residual Analysis for SARIMA(0,1,2)x(1,1,0) Model with CSS

Increasing the q value by 1 from the previous model made the histogram more symmetric, with the range within ±3. However, this adjustment might result in over fitting, as the coefficient for ma2 in both ML and CSS models is not statistically significant. Other results remain similar to those of the SARIMA(0,1,1)x(1,1,0)_12 model, with the ML model providing more normally distributed residuals compared to the CSS model. Additionally, the late auto correlations are not significantly different in either model.

```
# SARIMA(1,1,1)x(1,1,0)_12
m5_111.unemployment = Arima(BC.unemployment.ts,order=c(1,1,1),seasonal=list(o
rder=c(1,1,0), period=12),method = "ML")
coeftest(m5_111.unemployment)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.265803   0.368947 -0.7204    0.4713
## ma1   0.040292   0.379614  0.1061    0.9155
## sar1 -0.477736   0.087255 -5.4752 4.371e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res.m5 = residuals(m5_111.unemployment);
residual.analysis(model = m5_111.unemployment)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98049, p-value = 0.1106
```

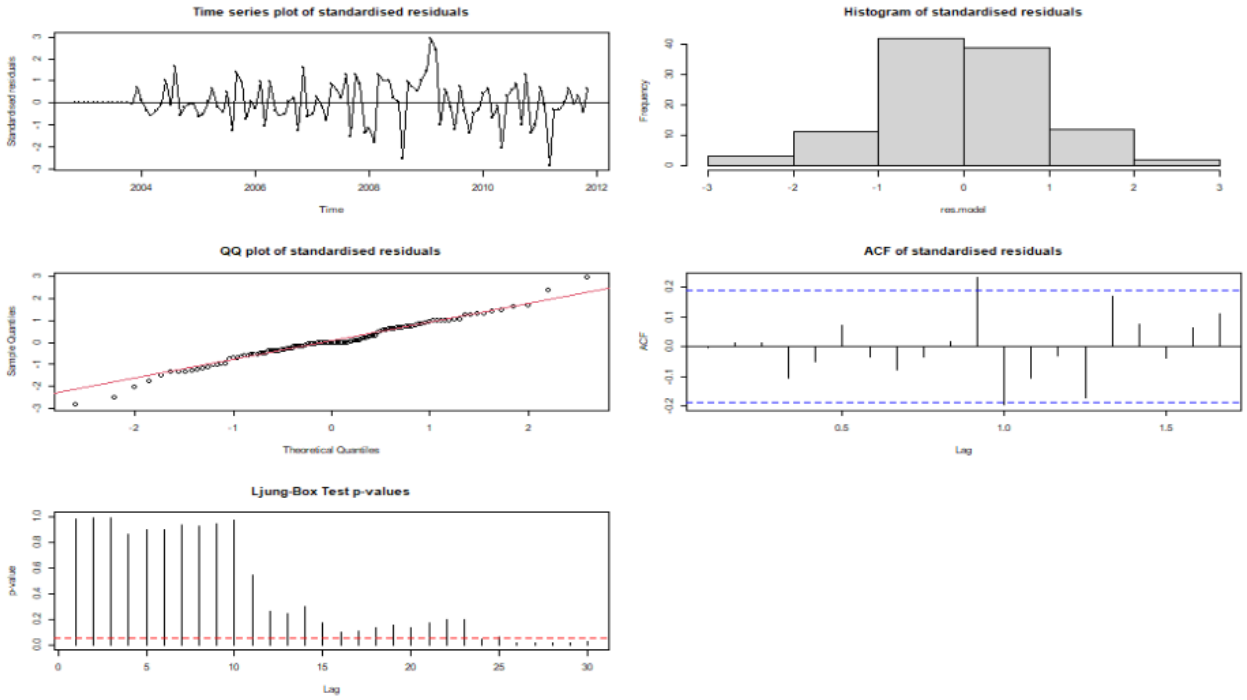

Figure 23: Residual Analysis for SARIMA(1,1,1)x(1,1,0) Model with ML

```
m5_111.unemploymentCSS = Arima(BC.unemployment.ts,order=c(1,1,1),seasonal=lis
t(order=c(1,1,0), period=12),method = "CSS")
coeftest(m5_111.unemploymentCSS)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
```

```
## ar1  -0.242887   0.392690 -0.6185    0.5362
## ma1   0.028508   0.402445  0.0708    0.9435
## sar1 -0.513103   0.090920 -5.6434 1.667e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res.m5 = residuals(m5_111.unemploymentCSS);
residual.analysis(model = m5_111.unemploymentCSS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.96462, p-value = 0.005381
```

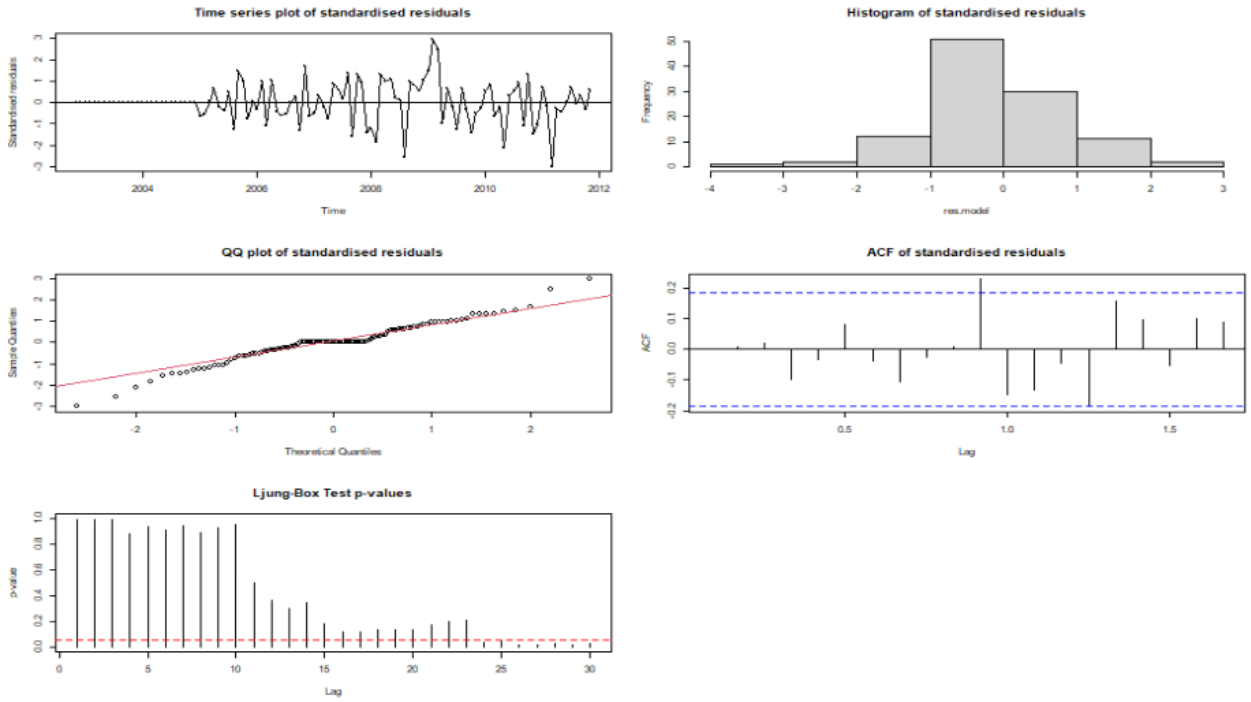

Figure 24: Residual Analysis for SARIMA(1,1,1)x(1,1,0) Model with CSS

For the SARIMA(1,1,1)x(1,1,0)_12 model, both ML and CSS methods do not show significant ARIMA components due to insignificant coefficient values. Compared to CSS, the ML model provides more normalized residuals and histograms with non-extreme, symmetric values. As with previous models, we observe only late lags in the ACF, and the Ljung-Box test indicates that the observed auto correlations are not significant.

```
# SARIMA(1,1,2)x(1,1,0)_12
m5_112.unemployment = Arima(BC.unemployment.ts,order=c(1,1,2),seasonal=list(o
rder=c(1,1,0), period=12),method = "ML")
coeftest(m5_112.unemployment)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.095820   0.658068 -0.1456    0.8842
## ma1  -0.127856   0.652235 -0.1960    0.8446
## ma2   0.057853   0.182822  0.3164    0.7517
## sar1 -0.476749   0.087557 -5.4450 5.179e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res.m5 = residuals(m5_112.unemployment);
residual.analysis(model = m5_112.unemployment)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98058, p-value = 0.1126
```

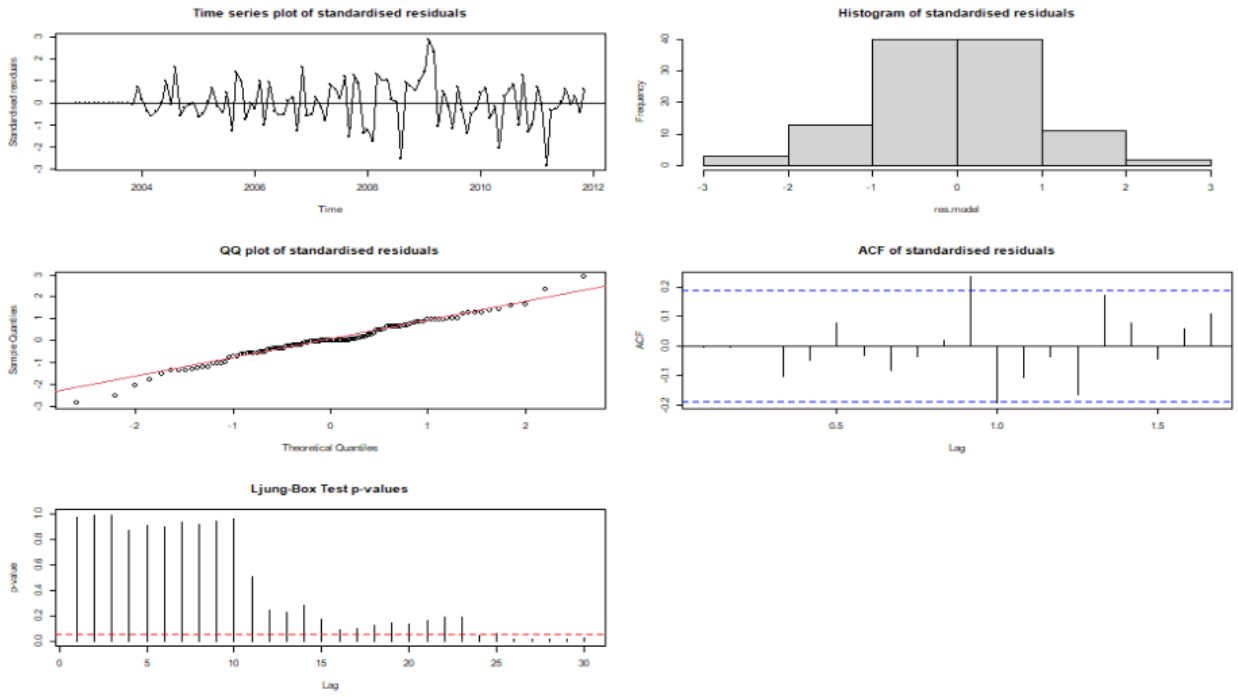

Figure 25: Residual Analysis for SARIMA(1,1,2)x(1,1,0) Model with ML

```
m5_112.unemploymentCSS = Arima(BC.unemployment.ts,order=c(1,1,2),seasonal=lis
t(order=c(1,1,0), period=12),method = "CSS")
coeftest(m5_112.unemploymentCSS)
```

```
##
## z test of coefficients:
##
```

```
##       Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.074081   0.739416 -0.1002    0.9202
## ma1  -0.139199   0.735436 -0.1893    0.8499
## ma2   0.053122   0.189103  0.2809    0.7788
## sar1 -0.511503   0.091558 -5.5866 2.315e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
res.m5 = residuals(m5_112.unemploymentCSS);
residual.analysis(model = m5_112.unemploymentCSS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.96437, p-value = 0.005149
```

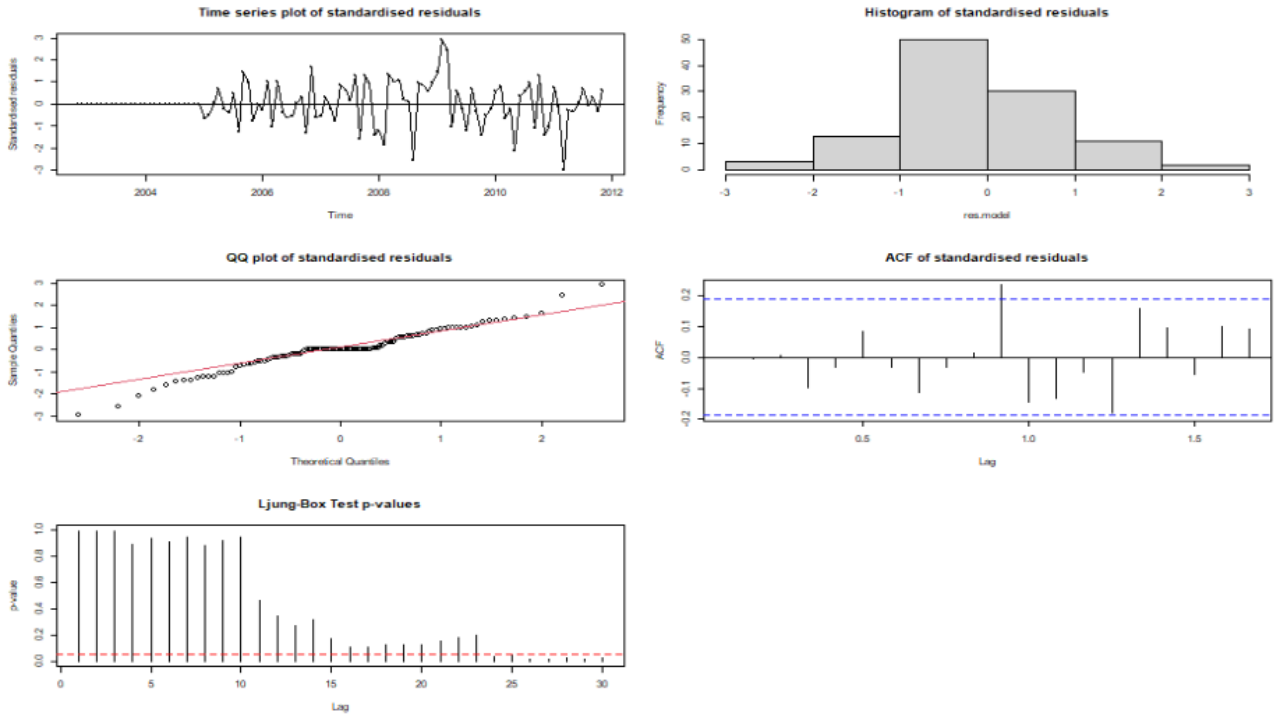

Figure 26: Residual Analysis for SARIMA(1,1,2)x(1,1,0) Model with CSS

Similarly, none of the ARIMA coefficients are significant in this model. This suggests that these model variants may not be suitable for our data.

```
# SARIMA(2,1,3)x(1,1,0)_12
m5_213.unemployment = Arima(BC.unemployment.ts,order=c(2,1,3),seasonal=list(o
rder=c(1,1,0), period=12),method = "ML")
coeftest(m5_213.unemployment)
```

```
##
## z test of coefficients:
```

```
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1   0.312938   0.133004  2.3528 0.0186304 *
## ar2  -0.749860   0.114468 -6.5508 5.723e-11 ***
## ma1  -0.561546   0.163148 -3.4420 0.0005775 ***
## ma2   1.032346   0.094801 10.8897 < 2.2e-16 ***
## ma3  -0.279397   0.131507 -2.1246 0.0336222 *
## sar1 -0.471212   0.089884 -5.2424 1.585e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m5_213.unemployment)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98699, p-value = 0.3747
```

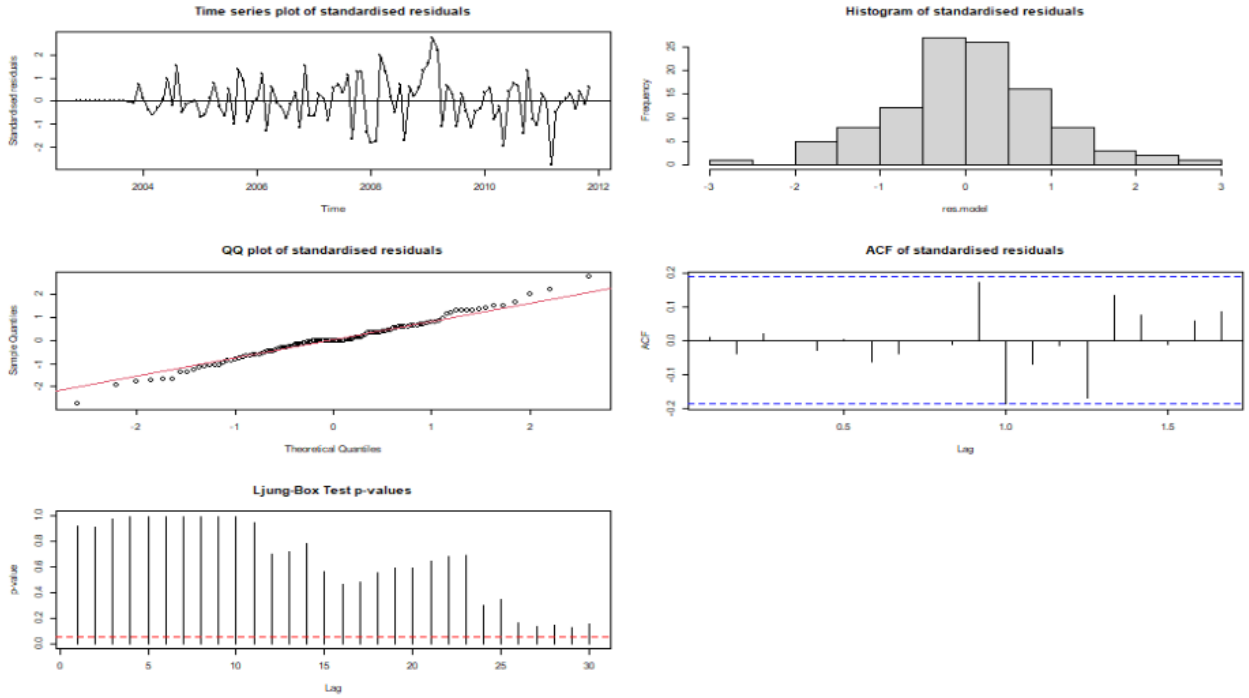

Figure 27: Residual Analysis for SARIMA(2,1,3)x(1,1,0) Model with ML

```
m5_213.unemploymentCSS = Arima(BC.unemployment.ts,order=c(2,1,3),seasonal=lis
t(order=c(1,1,0), period=12),method = "CSS")
coeftest(m5_213.unemploymentCSS)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
```

```
## ar1   0.6520711  0.0070615  92.342 < 2.2e-16 ***
## ar2  -0.7168944  0.0167126 -42.895 < 2.2e-16 ***
## ma1  -1.0046903  0.0279814 -35.906 < 2.2e-16 ***
## ma2   1.4295166  0.0127585 112.044 < 2.2e-16 ***
## ma3  -0.6255693  0.0181781 -34.413 < 2.2e-16 ***
## sar1 -0.5677354  0.0139103 -40.814 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m5_213.unemploymentCSS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.93456, p-value = 4.414e-05
```

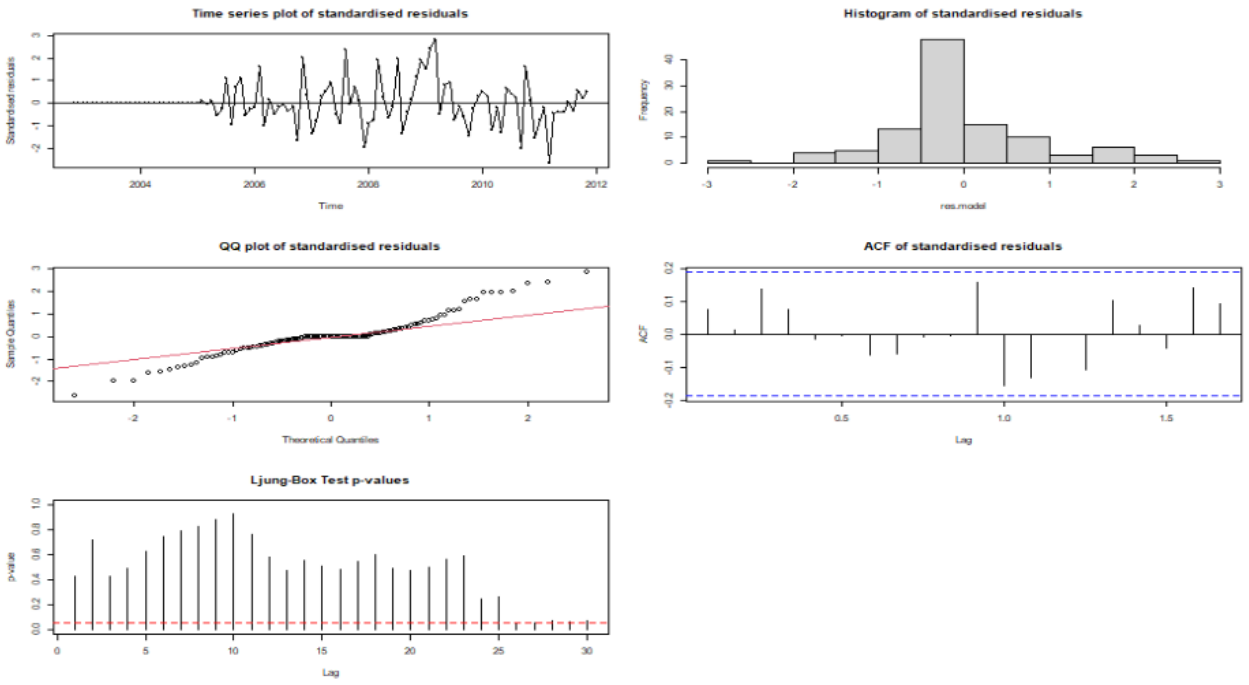

Figure 28: Residual Analysis for SARIMA(2,1,3)x(1,1,0) Model with CSS

For the SARIMA(2,1,3)x(1,1,0)_12 model, which was selected through visual inspection of ACF and PACF lags prior to using BIC and EACF, the ML model provides more statistically significant, normally distributed residuals. The Q-Q plot shows that the distribution of residuals in the ML model aligns better compared to its CSS counterpart. This ML model produces better white noise in the residuals than any other models we have developed so far. Additionally, the histogram is more symmetrical, and the range is within ±3, although there is an outlier in the left tail, likely due to a change point in the series.

Except for season 1, there are no significant auto correlations left in the ACF plot, and even the first season lag is not significant, as all p-values in the Ljung-Box test are greater than the significance level. Therefore, this model shows potential to be a good fit for forecasting.

```
# SARIMA(1,1,3)x(1,1,0)_12
m5_113.unemployment = Arima(BC.unemployment.ts,order=c(1,1,3),seasonal=list(o
rder=c(1,1,0), period=12),method = "ML")
coeftest(m5_113.unemployment)

##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1   0.938648   0.049357 19.0173 < 2.2e-16 ***
## ma1  -1.197530   0.124706 -9.6028 < 2.2e-16 ***
## ma2   0.305347   0.163020  1.8731   0.06106 .
## ma3  -0.107811   0.120289 -0.8963   0.37011
## sar1 -0.468416   0.088915 -5.2682 1.378e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m5_113.unemployment)
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.97551, p-value = 0.04183
```

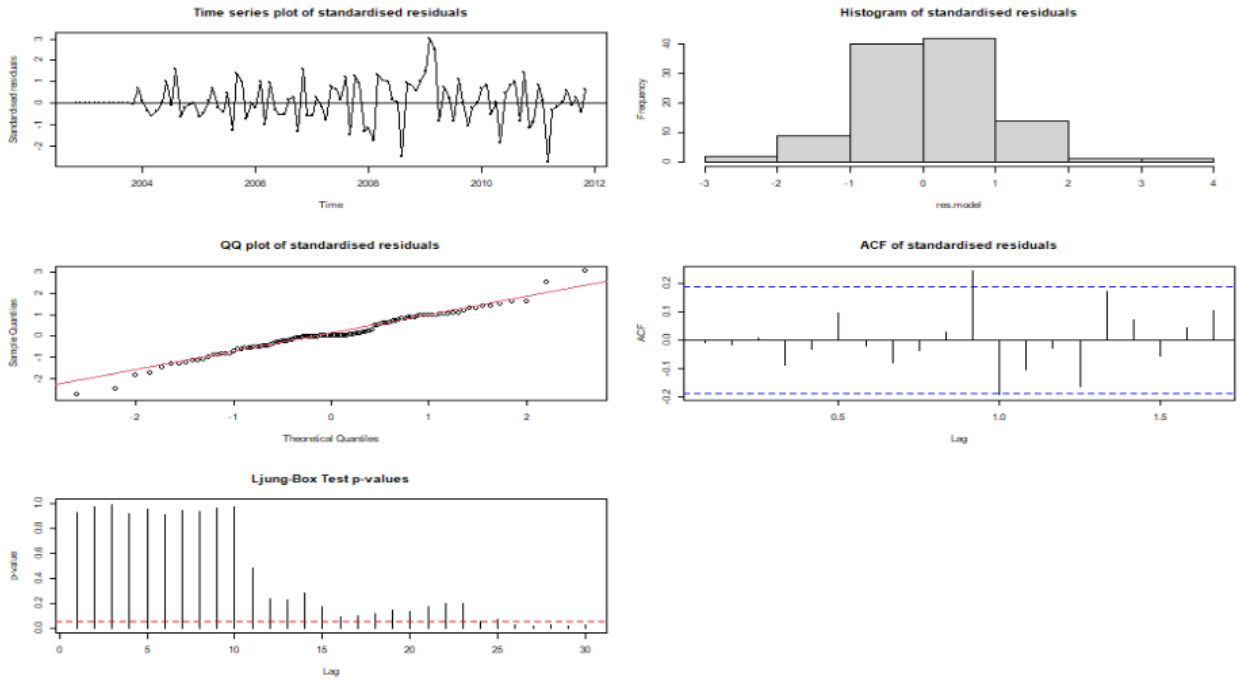

Figure 29: Residual Analysis for SARIMA(1,1,3)x(1,1,0) Model with ML

```
m5_113.unemploymentCSS = Arima(BC.unemployment.ts,order=c(1,1,3),seasonal=list(order=c(1,1,0), period=12),method = "CSS")
coeftest(m5_113.unemploymentCSS)

## Warning in sqrt(diag(se)): NaNs produced

##
## z test of coefficients:
##
##      Estimate Std. Error z value Pr(>|z|)
## ar1   0.84596        NaN     NaN      NaN
## ma1  -1.17387        NaN     NaN      NaN
## ma2   0.28770    0.17265  1.6663  0.09565 .
## ma3  -0.18335        NaN     NaN      NaN
## sar1 -0.43236        NaN     NaN      NaN
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m5_113.unemploymentCSS)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.95064, p-value = 0.0004974
```

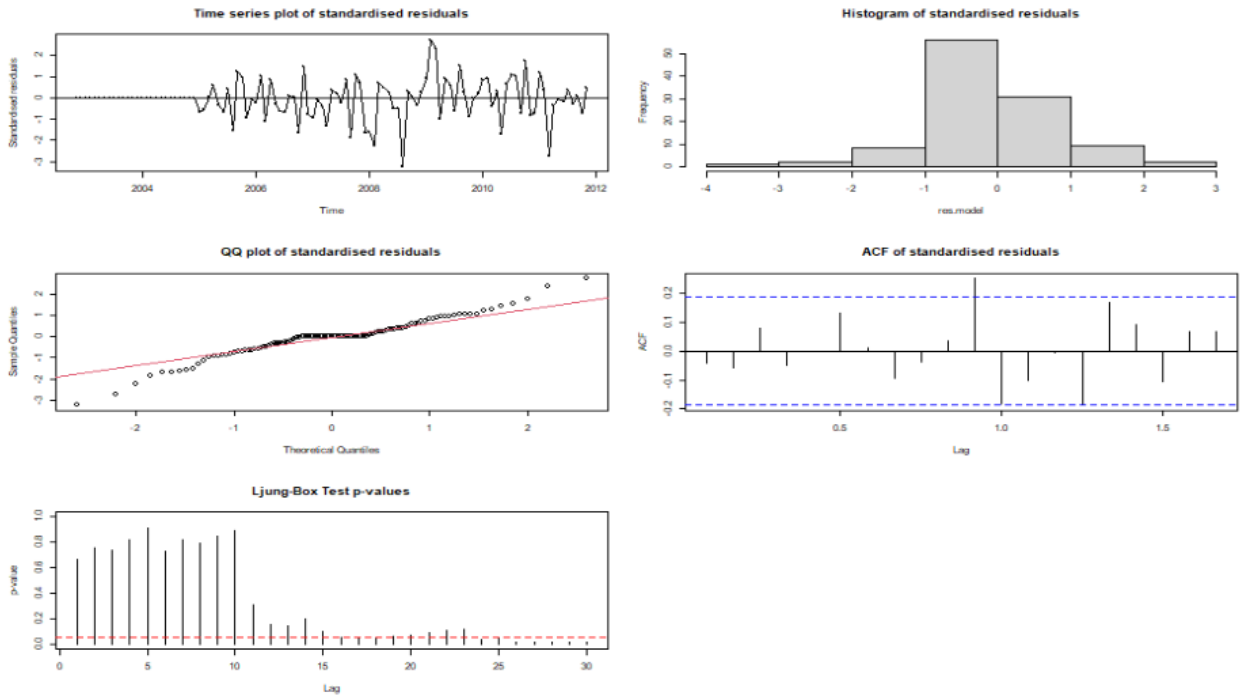

Figure 30: Residual Analysis for SARIMA(1,1,3)x(1,1,0) Model with CSS

Reducing the p-value by 1 affected the ML model, rendering MA coefficients ma2 and ma3 insignificant. NaN values appeared in the CSS method, possibly due to values like zero or less than zero. Therefore, since CSS-ML checking would only yield the ML result without CSS, this step was omitted. SARIMA(1,1,3)x(1,1,0)_12 model may not be suitable for our data.

```
# SARIMA(1,1,0)x(1,1,0)_12
m5_110.unemployment = Arima(BC.unemployment.ts,order=c(1,1,0),seasonal=list(o
rder=c(1,1,0), period=12),method = "ML")
coeftest(m5_110.unemployment)

##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.227653   0.099947 -2.2777   0.02274 *
## sar1 -0.476779   0.086815 -5.4919 3.976e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m5_110.unemployment)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98037, p-value = 0.1081
```

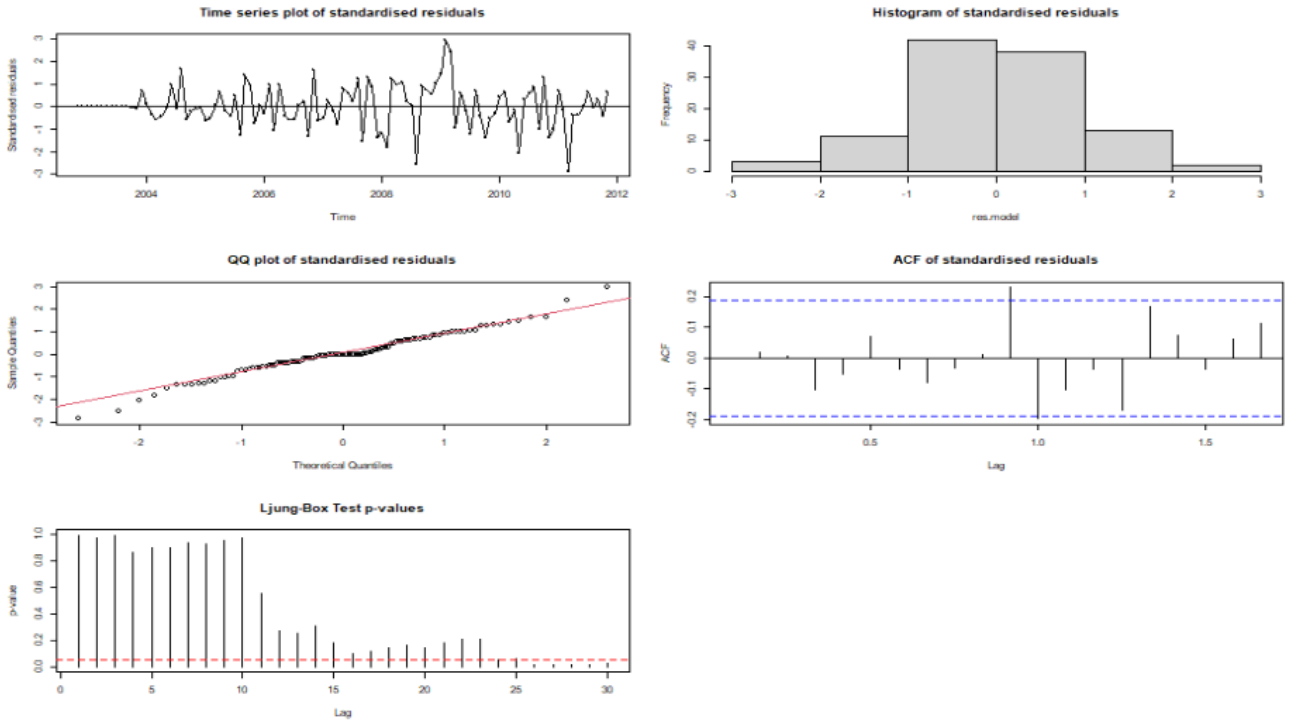

Figure 31: Residual Analysis for SARIMA(1,1,0)x(1,1,0) Model with ML

```
m5_110.unemploymentCSS = Arima(BC.unemployment.ts,order=c(1,1,0),seasonal=lis
t(order=c(1,1,0), period=12),method = "CSS")
coeftest(m5_110.unemploymentCSS)

##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1  -0.215743   0.100557 -2.1455   0.03192 *
## sar1 -0.512234   0.090036 -5.6892 1.277e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m5_110.unemploymentCSS)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.96466, p-value = 0.005428
```

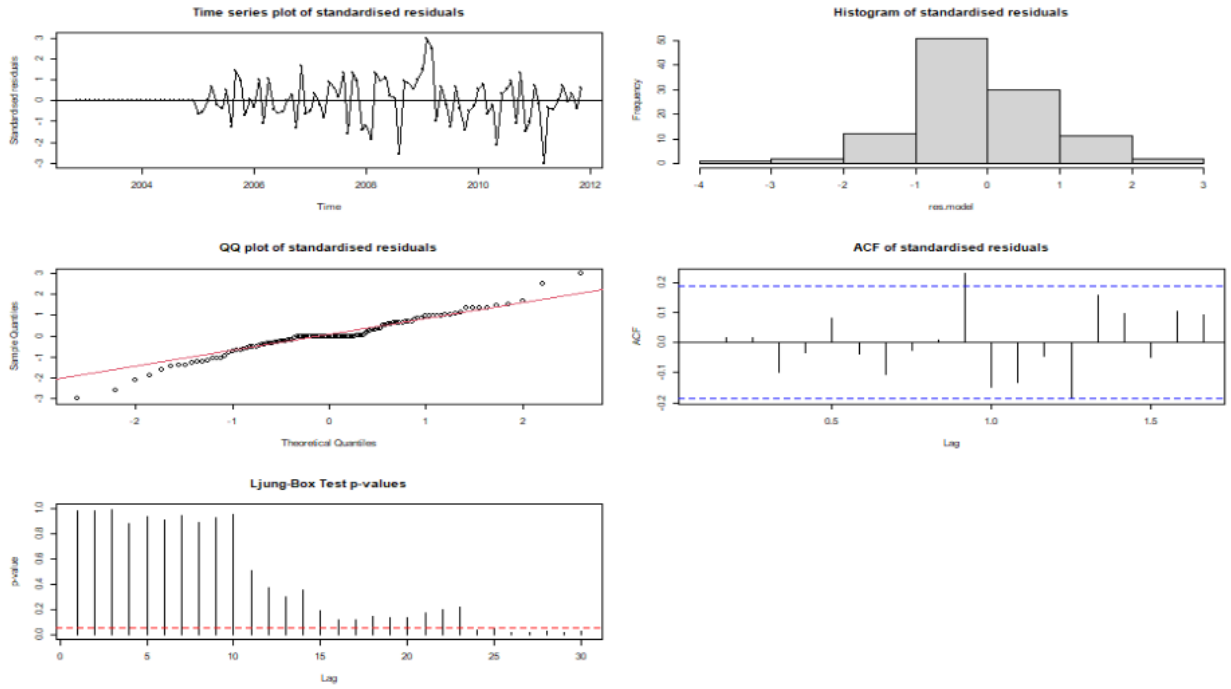

Figure 32: Residual Analysis for SARIMA(1,1,0)x(1,1,0) Model with CSS

For SARIMA(1,1,0)x(1,1,0)_12, both ML and CSS methods yield significant coefficient values (below the 0.05 significance level). The ML model generates normally distributed residuals and a more symmetric histogram within a ±3 range. In CSS, there are no seasonal lags, while in ML, there is one seasonal lag present. Most of the non-late auto correlations in the ACF plot are not significant, as indicated by the results of the Ljung-Box test.

1.4. Model Evaluation

1.4.1. AIC and BIC Score

```
# Sorting ml models using AIC

sc.AIC = AIC(m5_011.unemployment, m5_012.unemployment, m5_111.unemployment, m
5_112.unemployment,
             m5_213.unemployment, m5_113.unemployment,m5_110.unemployment)

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

	df <dbl>	AIC <dbl>
m5_110.unemployment	3	395.1072
m5_011.unemployment	3	395.4813
m5_213.unemployment	7	395.6945
m5_012.unemployment	4	397.0440
m5_111.unemployment	4	397.0968
m5_112.unemployment	5	399.0229
m5_113.unemployment	6	399.4909

7 rows

Figure 33: AIC Table

```
# Sorting ml model using BIC

sc.BIC = BIC(m5_011.unemployment, m5_012.unemployment, m5_111.unemployment, m
5_112.unemployment,
             m5_213.unemployment, m5_113.unemployment,m5_110.unemployment)

sort.score(sc.BIC, score = "bic")
```

	df <dbl>	BIC <dbl>
m5_110.unemployment	3	402.8002
m5_011.unemployment	3	403.1744
m5_012.unemployment	4	407.3014
m5_111.unemployment	4	407.3542
m5_112.unemployment	5	411.8446
m5_213.unemployment	7	413.6449
m5_113.unemployment	6	414.8770

Figure 34: BIC Table

The AIC scores are very close for the top 3 models, differing by decimal points. Based on both AIC and BIC scores, m5_110.unemployment emerged as the top model.

1.4.2. Evaluation Metrices

```
# calculating different evaluation metrics for all models

Sm5_011.unemployment <- accuracy(m5_011.unemployment)[1:7]
Sm5_012.unemployment <- accuracy(m5_012.unemployment)[1:7]
Sm5_111.unemployment <- accuracy(m5_111.unemployment)[1:7]
Sm5_112.unemployment <- accuracy(m5_112.unemployment)[1:7]
Sm5_213.unemployment <- accuracy(m5_213.unemployment)[1:7]
Sm5_113.unemployment <- accuracy(m5_113.unemployment)[1:7]
Sm5_110.unemployment <- accuracy(m5_110.unemployment)[1:7]

df.Smodels <- data.frame(
  rbind(Sm5_011.unemployment, Sm5_012.unemployment, Sm5_111.unemployment, Sm5
_112.unemployment,
        Sm5_213.unemployment, Sm5_113.unemployment, Sm5_110.unemployment)
)
colnames(df.Smodels) <- c("ME", "RMSE", "MAE", "MPE", "MAPE",
                          "MASE", "ACF1")
rownames(df.Smodels) <- c("SARIMA(0,1,1)x(1,1,0)_12", "SARIMA(0,1,2)x(1,1,0)_
12", "SARIMA(1,1,1)x(1,1,0)_12",
                          "SARIMA(1,1,2)x(1,1,0)_12", "SARIMA(2,1,3)x(2,1,0)_
12", "SARIMA(1,1,3)x(1,1,0)_12",
                          "SARIMA(1,1,0)x(1,1,0)_12")
round(df.Smodels,  digits = 3)
```

	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>	MASE <dbl>	ACF1 <dbl>
SARIMA(0,1,1)x(1,1,0)_12	0.044	1.699	1.247	0.098	5.880	0.445	-0.015
SARIMA(0,1,2)x(1,1,0)_12	0.041	1.695	1.247	0.091	5.890	0.445	-0.004
SARIMA(1,1,1)x(1,1,0)_12	0.042	1.695	1.245	0.095	5.879	0.444	-0.002
SARIMA(1,1,2)x(1,1,0)_12	0.041	1.694	1.245	0.092	5.884	0.444	-0.003
SARIMA(2,1,3)x(2,1,0)_12	0.044	1.616	1.192	0.145	5.614	0.425	0.009
SARIMA(1,1,3)x(1,1,0)_12	0.143	1.667	1.215	0.498	5.733	0.433	-0.008
SARIMA(1,1,0)x(1,1,0)_12	0.043	1.695	1.245	0.095	5.878	0.444	0.000

Figure 35: Evaluation Metrics for SARIMA Models

Considering overall metrics, SARIMA(2,1,3)x(2,1,0)_12 consistently scores the best (lowest values) across most metrics. It exhibits the lowest RMSE, MAE, MAPE, and MASE scores, while its ME is comparable with a few other models. Therefore, based on the majority of evaluation metrics, SARIMA(2,1,3)x(2,1,0)_12 is considered the best model.

So, we have narrowed down our potential models to m5_110.unemployment and Sm5_213.unemployment. Despite their similar AIC scores, Sm5_213.unemployment exhibits better overall evaluation metric scores and significantly higher coefficient values for all parameters. Therefore, we will select Sm5_213.unemployment as our preferred model in this model.

1.5. Over-parameterized model

For Sm5_213.unemployment over-parameterized models include
SARIMA(2,1,4)x(1,1,0)_12 and **SARIMA(3,1,3)x(1,1,0)_12**

```
# SARIMA(2,1,4)x(2,1,0)
m5_214.unemployment = Arima(BC.unemployment.ts,order=c(2,1,4),seasonal=list(o
rder=c(1,1,0), period=12),method = "ML")
coeftest(m5_214.unemployment)

##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1   0.277863   0.176026  1.5785   0.11444
## ar2  -0.699216   0.168059 -4.1605 3.175e-05 ***
## ma1  -0.513282   0.202763 -2.5314   0.01136 *
## ma2   0.932346   0.236340  3.9449 7.982e-05 ***
## ma3  -0.226261   0.167176 -1.3534   0.17592
## ma4  -0.070950   0.148743 -0.4770   0.63337
## sar1 -0.469415   0.090377 -5.1940 2.058e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m5_214.unemployment)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98529, p-value = 0.2765
```

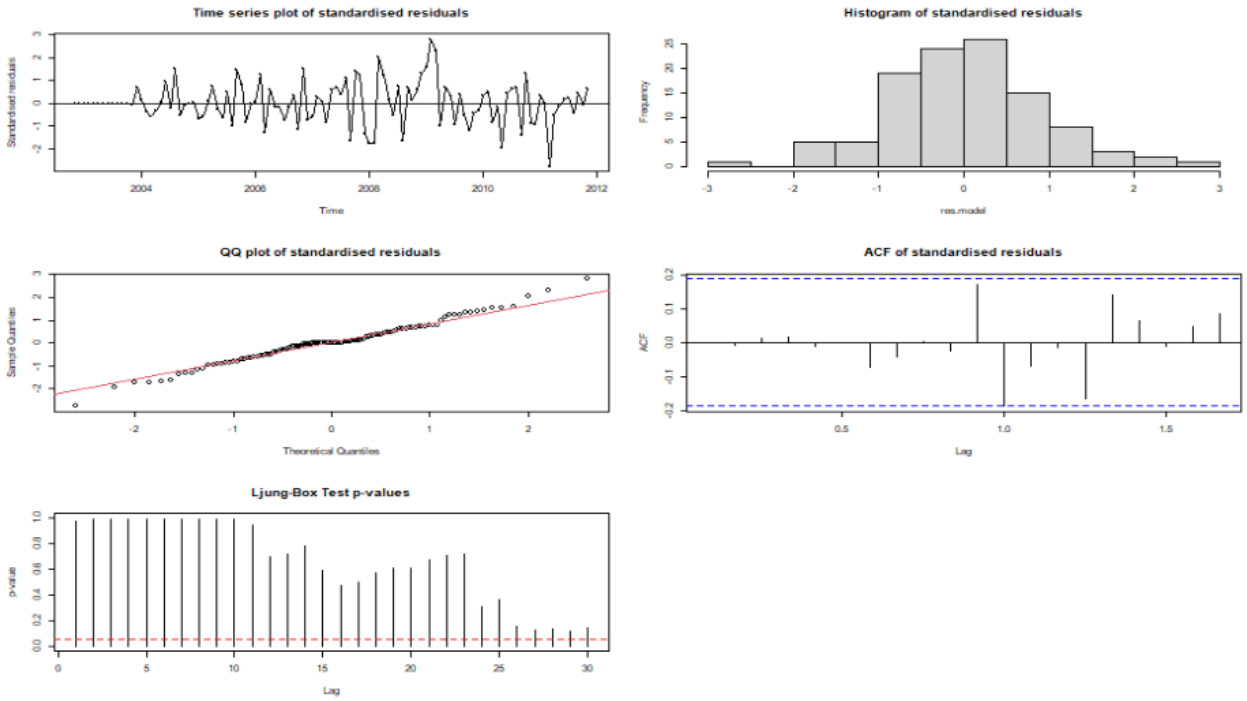

Figure 36: Residual Analysis for Overparameterized SARIMA(2,1,4)x(1,1,0) Model with ML

The residuals show no remaining seasonality, with a symmetric histogram (aside from one outlier) within a ±3 range, and they appear to follow a normal distribution according to both the Shapiro-Wilk normality test and the Q-Q plot. Additionally, there is no autocorrelation observed in the ACF plot, and the Ljung-Box test shows no significant lags. However, increasing the q value by 1 resulted in insignificant coefficients such as ar1, ma3, and ma4. This over-parametrization suggests over fitting of the model.

```
# SARIMA(3,1,3)x(2,1,0)
m5_313.unemployment = Arima(BC.unemployment.ts,order=c(3,1,3),seasonal=list(o
rder=c(1,1,0), period=12))
coeftest(m5_313.unemployment)

##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ar1   0.482697   0.440348  1.0962    0.2730
## ar2  -0.772051   0.138366 -5.5798 2.408e-08 ***
## ar3   0.145240   0.370335  0.3922    0.6949
## ma1  -0.721928   0.414271 -1.7426    0.0814 .
## ma2   1.065589   0.127982  8.3261 < 2.2e-16 ***
## ma3  -0.440406   0.408787 -1.0773    0.2813
## sar1 -0.470629   0.089899 -5.2351 1.649e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m5_313.unemployment)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.98586, p-value = 0.3064
```

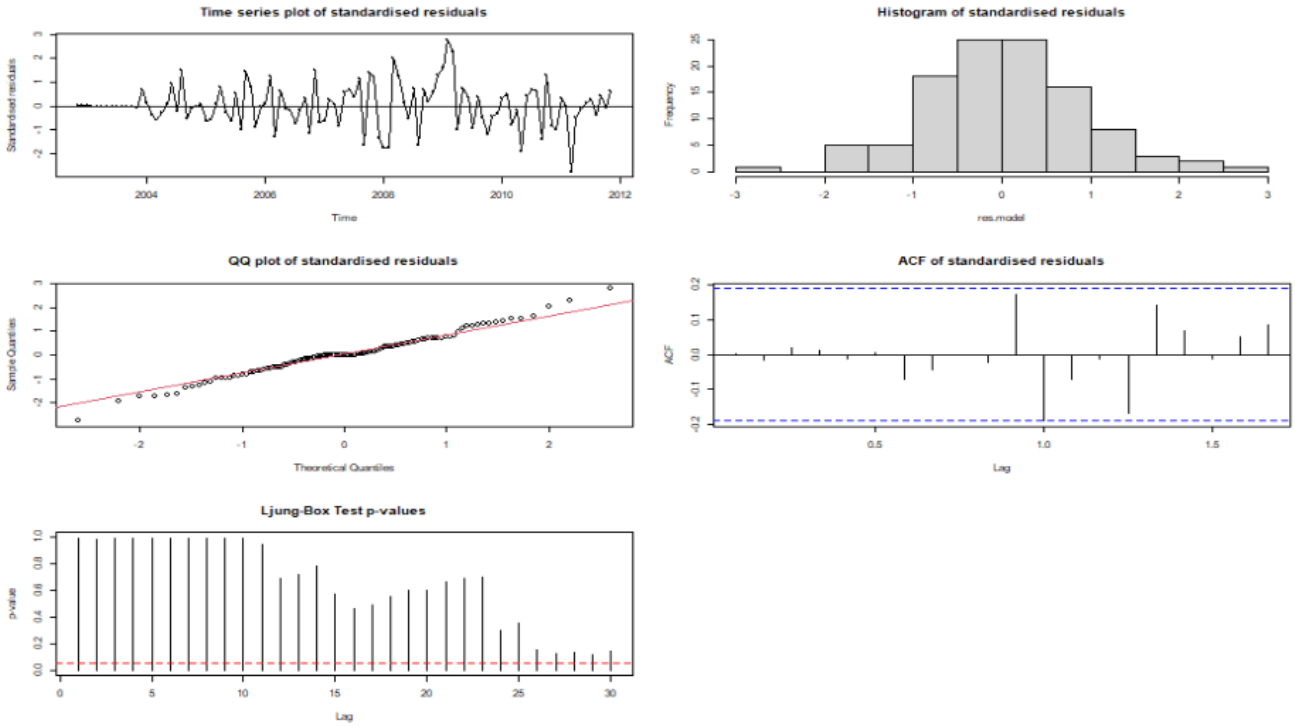

Figure 37: Residual Analysis for Overparameterized SARIMA(3,1,3)x(1,1,0) Model with ML

The residuals show no remaining seasonality, with a symmetric histogram (aside from one outlier) within a ±3 range, and they appear to follow a normal distribution according to both the Shapiro-Wilk normality test and the Q-Q plot. Additionally, there is no autocorrelation observed in the ACF plot, and the Ljung-Box test shows no significant lags. However, increasing the p value by 1 resulted in insignificant coefficients such as ar1, ar3, ma 1 and ma4. This over-parametrization suggests over fitting of the model.

1.6. Forecast

Since over-parameterized models did not improve the results, the Sm5_213.unemployment model is selected to forecast the next 10 observations of the data.

```
# Forecasting
m5_213.unemployment.forecast = Arima(unemployment.ts,order=c(2,1,3),seasonal=
list(order=c(1,1,0), period=12),
                             lambda = 1.5, method = "CSS") # raw series, CSS was
better for m5_213.unemployment.forecast (check co-efficient values)

forecast_data = forecast(m5_213.unemployment.forecast, h = 10) # next 10 obse
rvations
plot(forecast_data, xlab = "Year",ylab = "Unemployment Rate", main = "10 Mont
h Forecast for Unemployment rate of persons aged 15-24 using
     SARIMA(2,1,3)(1,1,0)_12")
```

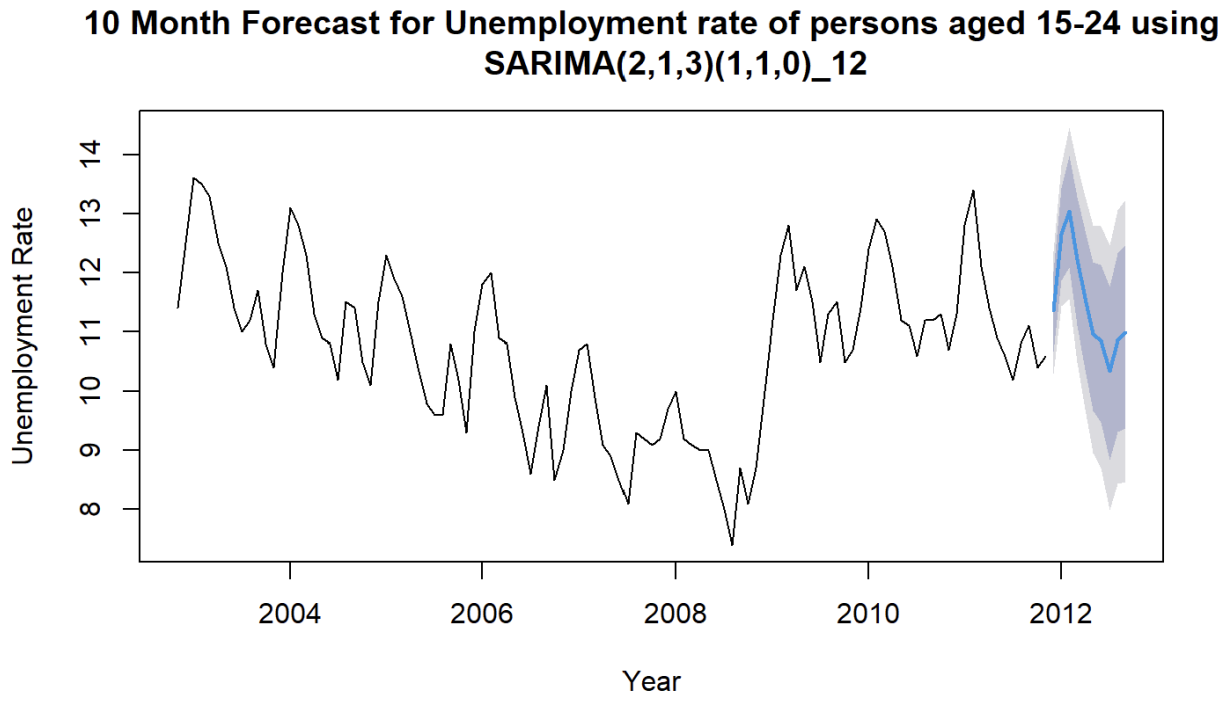

Figure 38: 10-Month Forecast for Unemployment Rate of Persons Aged 15-24 Using Best SARIMA(2,1,3)(1,1,0)

Figure (38) depicts the 10-month forecast for the original data, with the forecasted values shown in a blue line and 80% and 95% confidence intervals. The fitting appears satisfactory, with the change point slightly influencing the confidence intervals but not significantly. To further assess this, the data will be split at the change point, and a new model will be fitted to investigate potential improvements.

2: Monthly Australia Unemployment Rate from 2008 to 2011 among people aged 15 – 24

2.1. Descriptive Statistics

Given the nature of our data, which revolves around the job market and unemployment rates, it is assumed that the current trend depends more on recent years rather than historical records spanning a decade. The job market evolves constantly and undergoes changes with each generation. Therefore, even though we lose a significant portion of the data by splitting it at the change point, it is expected that this will not significantly affect the accuracy of the next 10-month forecast.

2.1.1 Splitting the data

We have split the unemployment time series data starting from the change point in November 2008, which corresponds to indices 75 to 109.

```
unemployment.ts.p2 =  ts(unemployment.ts[73:109],start=c(2008,11), frequency=
12)
unemployment.ts.p2
```

```
##       Jan  Feb  Mar  Apr  May  Jun  Jul  Aug  Sep  Oct  Nov  Dec
## 2008                                                     8.7 10.0
## 2009 11.2 12.3 12.8 11.7 12.1 11.5 10.5 11.3 11.5 10.5 10.7 11.4
## 2010 12.4 12.9 12.7 12.1 11.2 11.1 10.6 11.2 11.2 11.3 10.7 11.3
## 2011 12.8 13.4 12.1 11.4 10.9 10.6 10.2 10.8 11.1 10.4 10.6
```

```
# Plotting the time series plot after the split
plot(unemployment.ts.p2,ylab='Unemployment Rate',xlab='Year',type='o', main =
"Time series plot of Unemployment rate of persons aged 15-24 (After Split)")
```

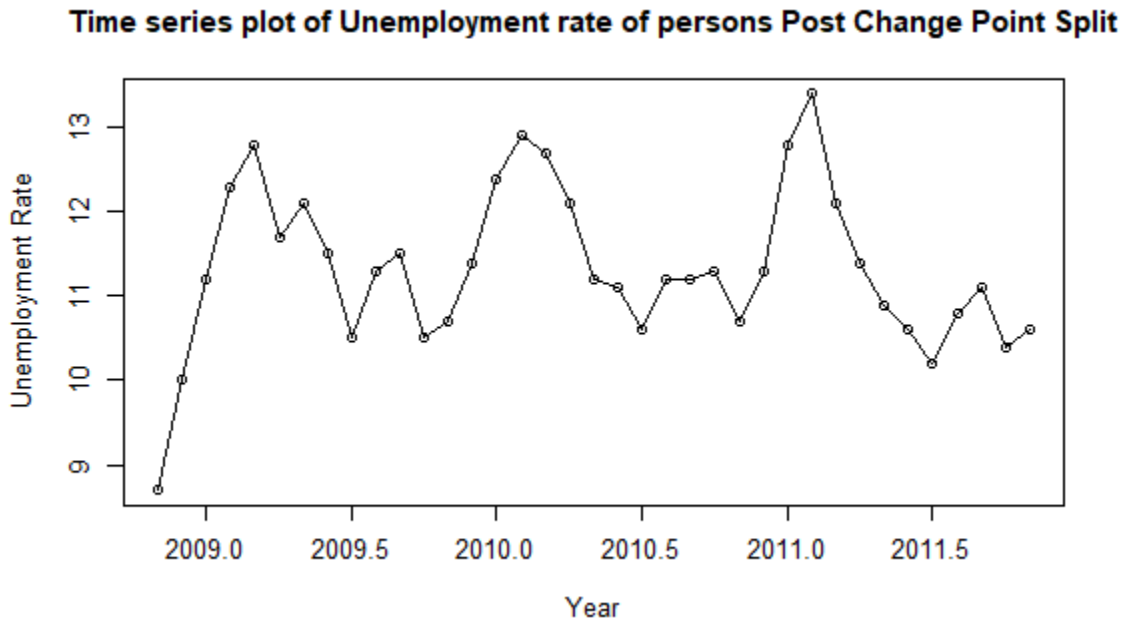

Figure 39: Time series plot of Unemployment rate of persons Post Change Point Split

Time Series Plot Characteristics after Split (Figure 39):

1. **Trend**: The time series plot seems to be fluctuating a lot throughout the years. But overall, it looks to be stable with no trend.

2. **Seasonality**: Visually, it looks to be having seasonality. But it cannot be confirmed with just this, as it follows some behaviour.

3. **Changing Variance**: There is a slight changing variance.

4. **Behavior**: The time series plot has both Auto–Regressive(AR) and Moving- Average(MA) behaviour.

5. **Intervention/ Change Point**: Since this data was split after the change point, it does not have any change points.

2.1.2. Checking for Seasonality

```
plot_acf_pacf(unemployment.ts.p2)
```

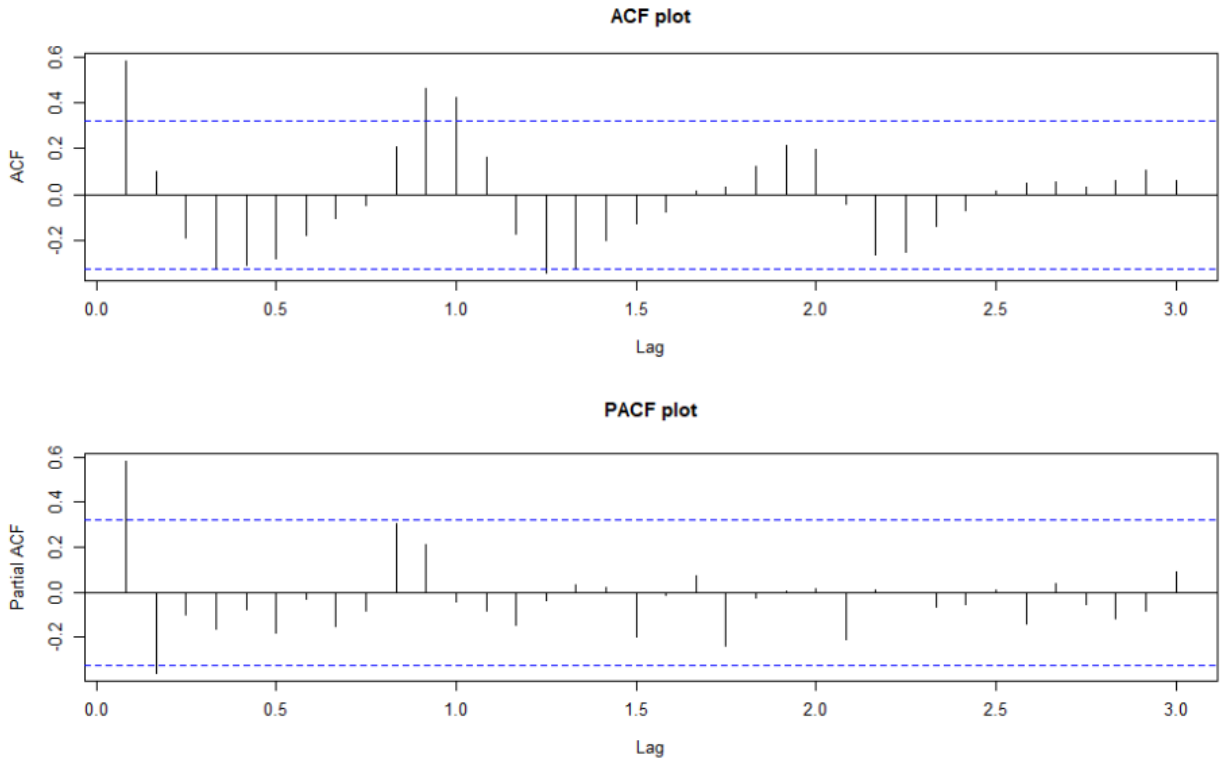

Figure 40: ACF & PACF for the splitted time series object

From this, we can confirm that the plot exhibits seasonality, as evidenced by the clear patterns in the ACF plot. Additionally, the decaying pattern suggests that the data might be non-stationary.

2.1.3. Test for Stationarity

```
ts_stationary_tests(unemployment.ts.p2)

## Warning in kpss.test(ts_object): p-value greater than printed p-value

## $ADF_Test
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_object
```

```
## Dickey-Fuller = -3.4894, Lag order = 3, p-value = 0.05972
## alternative hypothesis: stationary
##
##
## $PP_Test
##
##  Phillips-Perron Unit Root Test
##
## data:  ts_object
## Dickey-Fuller Z(alpha) = -16.586, Truncation lag parameter = 3, p-value
## = 0.09122
## alternative hypothesis: stationary
##
##
## $KPSS_Test
##
##  KPSS Test for Level Stationarity
##
## data:  ts_object
## KPSS Level = 0.090413, Truncation lag parameter = 3, p-value = 0.1
```

The decaying pattern, along with the ADF and PP test p-values being greater than 0.05, suggests that the data might be non-stationary. Although the KPSS test indicates potential stationarity, the overall evidence points towards non-stationarity.

Similar to what we did in the first half of the report, we will follow the same procedure and apply the SARIMA model to account for the seasonality in the data and also handle the non - stationarity in the the data on the way.

2.2. Model Specifications

2.2.1. Finding Seasonal Components (P, D, Q)

We will be setting up the seasonal components (P, D, Q) for our SARIMA model by fitting a model with initial seasonal difference and examining their ACF and PACF plots of the residuals.

```
# Start with the first seasonal difference
m1.unemployment.p2 = Arima(unemployment.ts.p2,order=c(0,0,0), seasonal=list(o
rder=c(0,1,0), period=12))
res.m1.p2 = residuals(m1.unemployment.p2);
plot_residuals_acf_pacf(res.m1.p2, "(0,0,0),(0,1,0)")
```

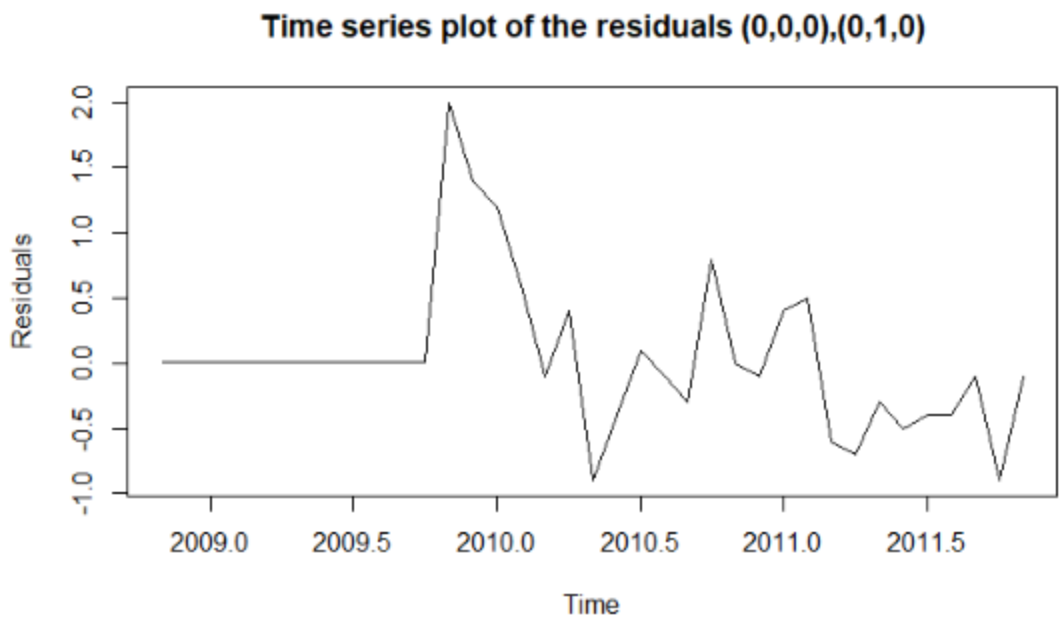

Figure 41: Residual Time Series plot for box-cox transformed data of model (0,0,0),(0,1,0)

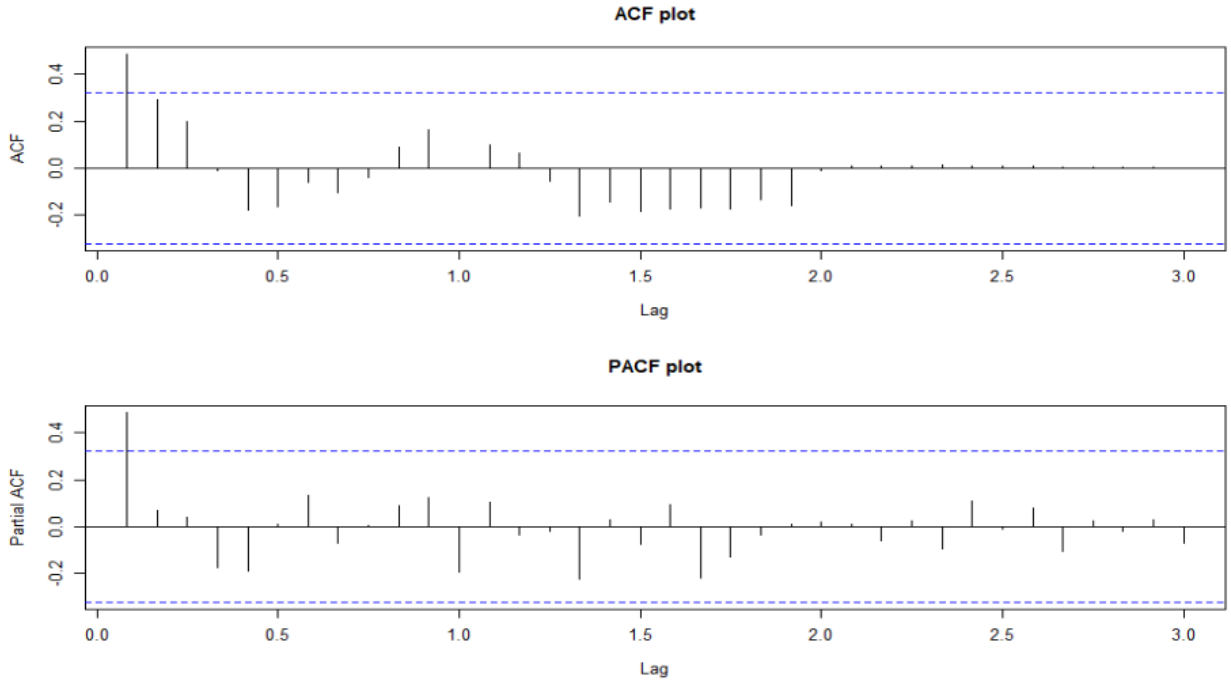

Figure 42: ACF & PACF Residuals plot for box-cox transformed data of model (0,0,0),(0,1,0)

The ACF and PACF plots of the residuals indicate that there are no significant lags, suggesting that the values of P and Q in the SARIMA model may be 0. Furthermore, after applying the

seasonal difference to the data, the seasonality in the residual plot has been removed, indicating it was successful in addressing the seasonal component of the data.

Now, we proceed with setting up the ARIMA component of the SARIMA model.

2.2.2. Finding ARIMA Components (p,d,q)

To stabilize the variance in the data before configuring the ARIMA component, we will use a Box-Cox transformation.

Box-Cox Transformation

```
BC <- BoxCox.ar(unemployment.ts.p2)
```

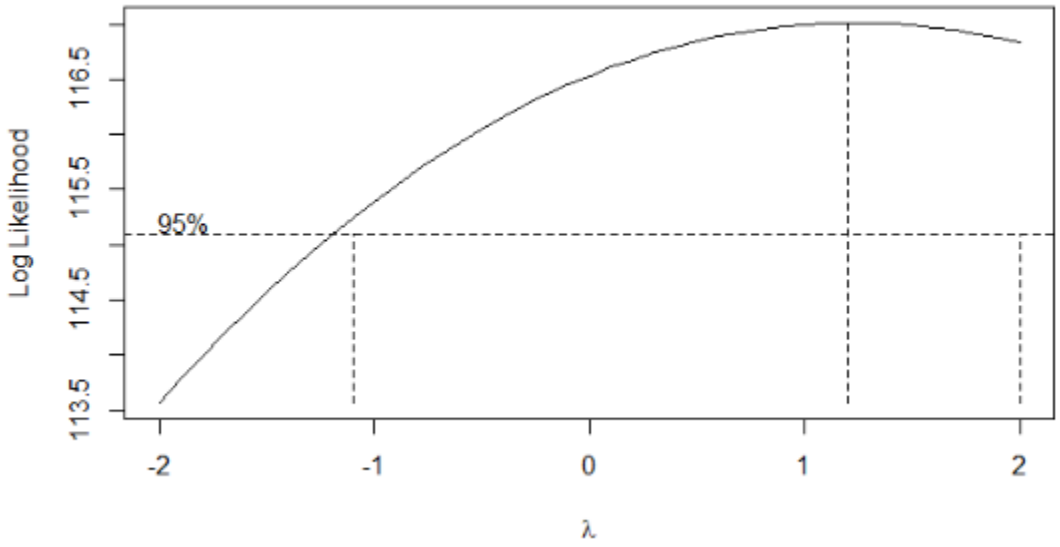

Figure 43: Box-Cox Transformation Log-Likelihood Plot and Confidence Intervals

```
BC$ci
```

```
## [1] -1.1  2.0
```

```
lambda <- BC$lambda[which(max(BC$loglike) == BC$loglike)]
lambda
```

```
## [1] 1.2
```

```
BC.unemployment.ts.p2 = (unemployment.ts^lambda-1)/lambda
```

We applied a Box-Cox transformation with a lambda value of 1.3. This transformation makes the data more consistent and better suited for ARIMA modeling.

To remove the non-stationarity from the series, we are applying an ordinary differencing with d = 1. This step helps to make the time series data stationary by removing trends and other non-stationary components.

```
m2.unemployment.p2 = Arima(BC.unemployment.ts.p2,order=c(0,1,0),
                     seasonal=list(order=c(0,1,0), period=12))
res.m2.p2 = residuals(m2.unemployment.p2);
plot_residuals_acf_pacf(res.m2.p2, "(0,1,0),(0,1,0)")
```

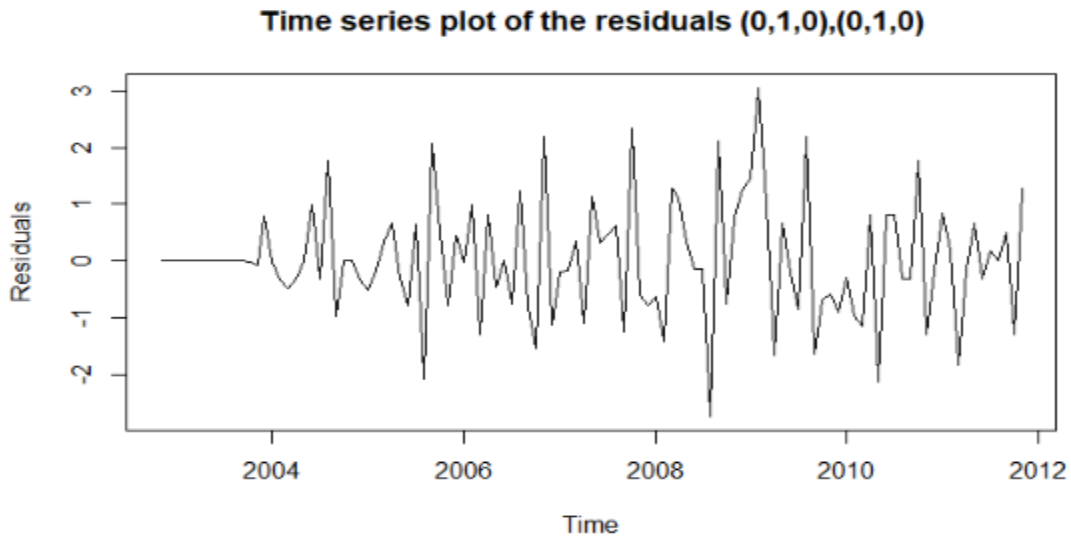

Figure 44: Residual Time Series plot for box-cox transformed data of model (0,1,0),(0,1,0)

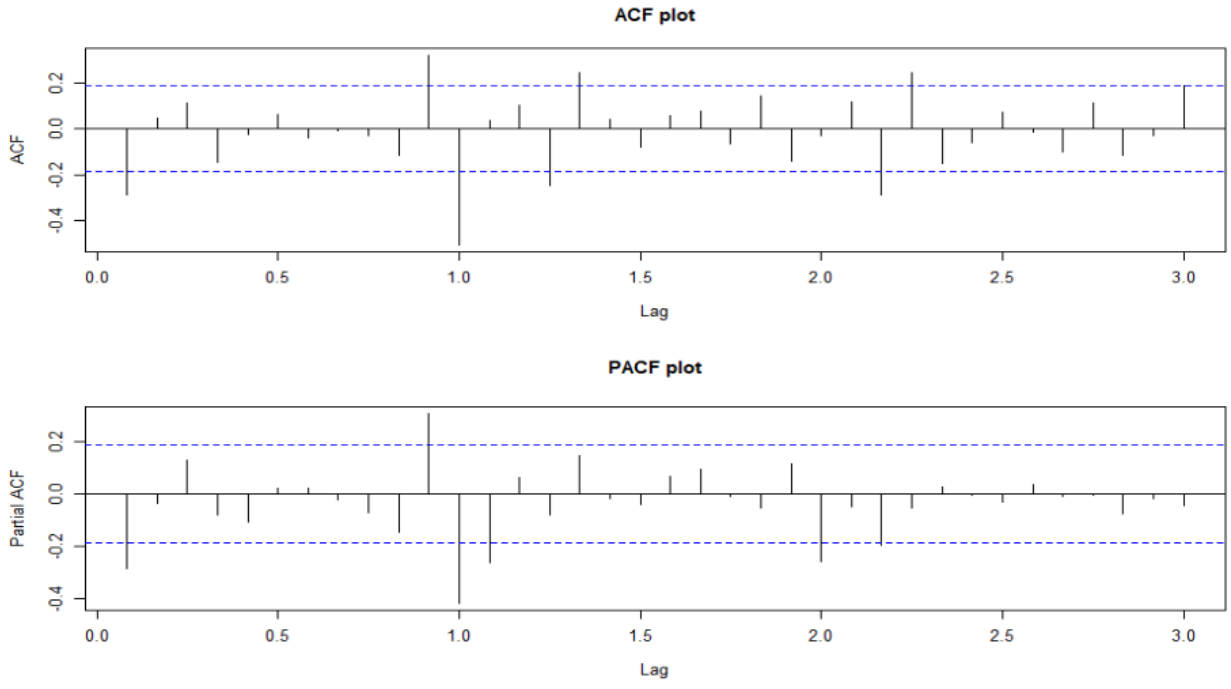

Figure 45: ACF & PACF Residuals plot for box-cox transformed data of model (0,1,0),(0,1,0)

After applying differencing, the residual plot indicates that the trend has been removed. Additionally, we observe two significant lags in both the ACF and PACF plots. We will use these values for p and q to determine the optimal ARIMA parameters.(p =2 and q =2)

```
m3.unemployment.p2 = Arima(BC.unemployment.ts.p2,order=c(2,1,2),
                    seasonal=list(order=c(0,1,0), period=12))
res.m3.p2 = residuals(m3.unemployment.p2);
plot_residuals_acf_pacf(res.m3.p2, "(2,1,2,(0,1,0)")
```

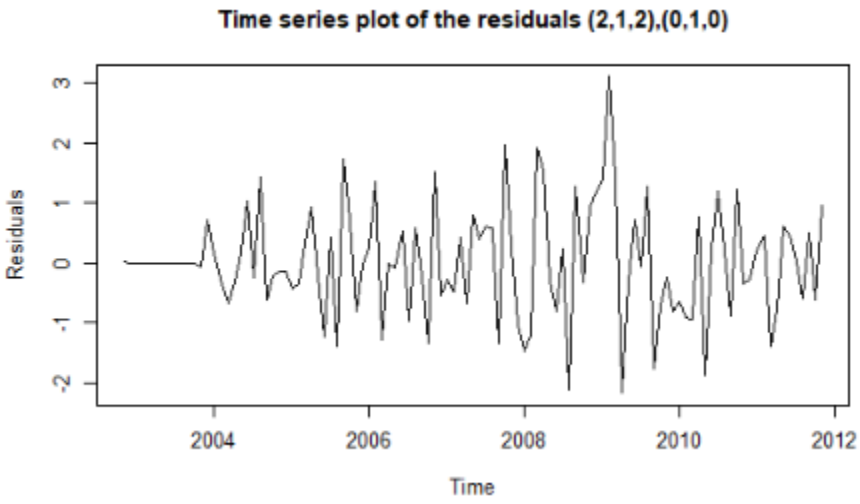

Figure 46: Residual Time Series plot for box-cox transformed data of model (2,1,2),(0,1,0)

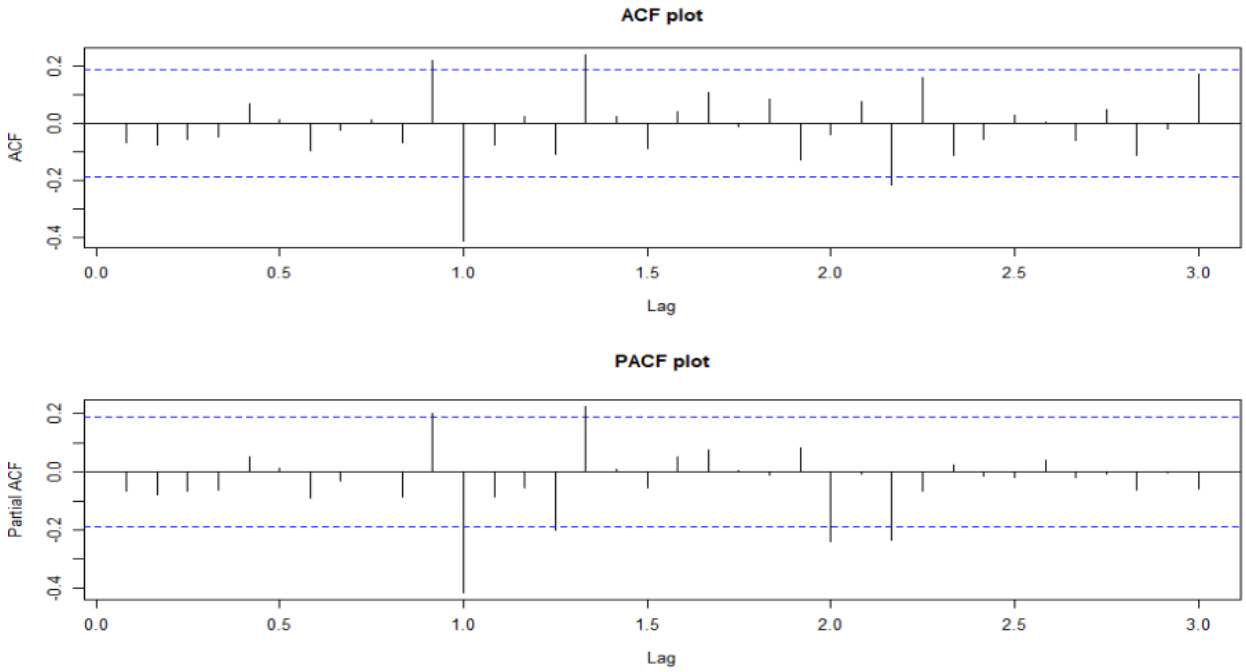

Figure 47: ACF & PACF Residuals plot for box-cox transformed data of model (2,1,2),(0,1,0)

As we can see, there is still one significant lag in the ACF and PACF plots.Therefore, we will overparameterize the p and q values to check if it helps improve the model.

We will first try improving the model by increasing the p value by 1 and observe its impact.

```
m4.unemployment.p2 = Arima(BC.unemployment.ts.p2,order=c(3,1,2),
                           seasonal=list(order=c(0,1,0), period=12))
res.m4.p2= residuals(m4.unemployment.p2);
plot_residuals_acf_pacf(res.m4.p2, "(3,1,2),(0,1,0)")
```

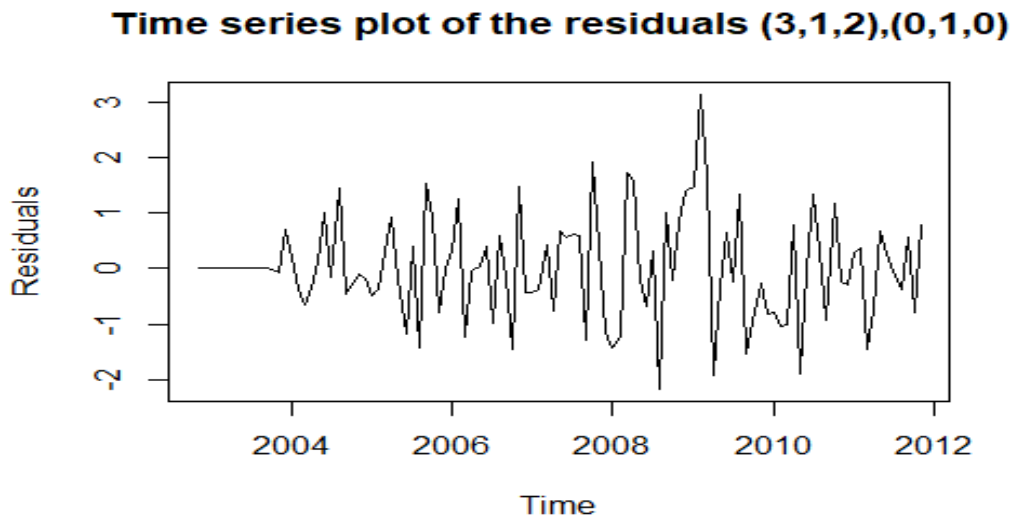

Figure 48: Residual Time Series plot for box-cox transformed data of model (3,1,2),(0,1,0)

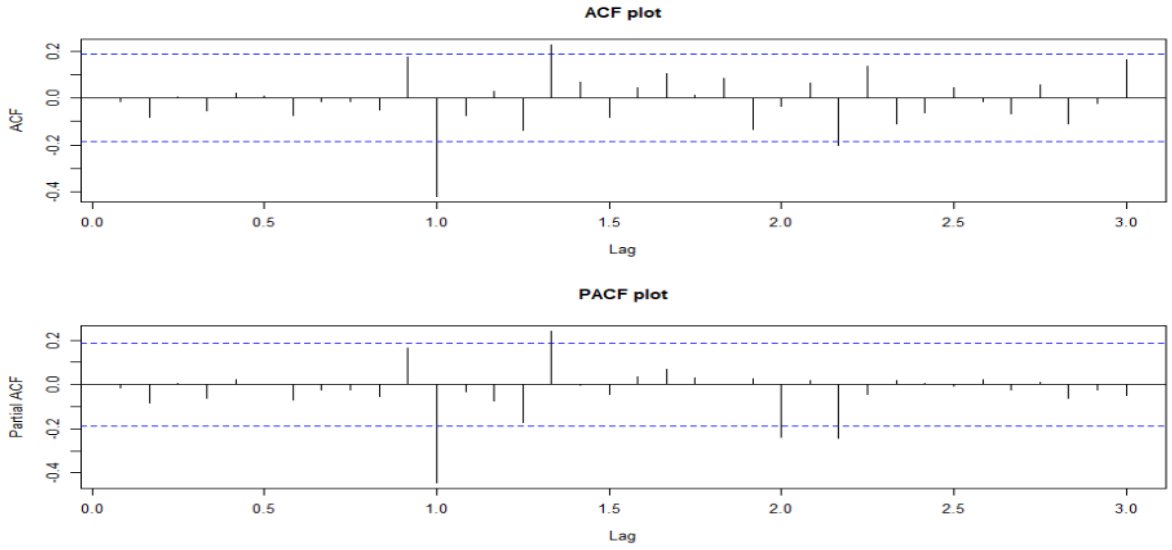

Figure 49: ACF & PACF Residuals plot for box-cox transformed data of model (3,1,2),(0,1,0)

With this adjustment, the significant lag has been removed, indicating that the overparameterized model is a better fit than the previous one. Therefore, we will use this model for fitting i.e SARIMA(3,1,2)x(0,1,0)_12.

```
ts_stationary_tests(res.m4.p2)
```

```
## Warning in adf.test(ts_object): p-value smaller than printed p-value

## Warning in pp.test(ts_object): p-value smaller than printed p-value

## Warning in kpss.test(ts_object): p-value greater than printed p-value

## $ADF_Test
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_object
## Dickey-Fuller = -4.6848, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
##
##
## $PP_Test
##
##  Phillips-Perron Unit Root Test
##
## data:  ts_object
## Dickey-Fuller Z(alpha) = -103.23, Truncation lag parameter = 4, p-value
## = 0.01
## alternative hypothesis: stationary
##
##
## $KPSS_Test
##
##  KPSS Test for Level Stationarity
##
## data:  ts_object
## KPSS Level = 0.04965, Truncation lag parameter = 4, p-value = 0.1
```

The Dickey-Fuller test yielded a statistically significant p-value of 0.01, supporting stationarity. Similarly, the Phillips-Perron test and KPSS test also confirmed the data's stationarity with p-values of 0.01 and 0.1, respectively.

2.2.3. EACF

```
eacf(res.m2.p2)
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x o o o o o o o o o x  x  o  o
## 1 o o o o o o o o o o o  x  o  o
```

```
## 2 x x o o o o o o o o o  x  x  o
## 3 x x x o o o o o o o o  x  x  o
## 4 x x x o o o o o o o o  x  o  o
## 5 x o o o o o o o o o o  x  o  o
## 6 x x o o o o o o o o o  x  o  o
## 7 x x o o o o o o o o o  x  x  o
```

Using EACF, we the optimal models are **SARIMA(1,1,1)x(0,1,0)_12, SARIMA(0,1,1)x(0,1,0)_12, SARIMA(0,1,2)x(0,1,0)_12 and SARIMA(1,1,2)x(0,1,0)_12**

2.2.4. BIC Table

```
par(mfrow=c(1,1))
bic_table = armasubsets(y=res.m2,nar=5,nma=5,y.name='p',ar.method='ols')
plot(bic_table)
```

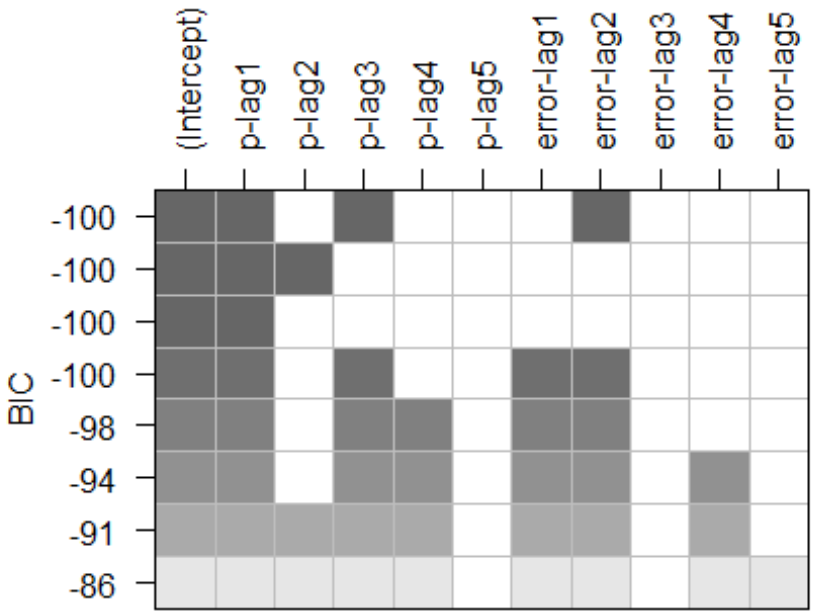

Figure 50: BIC Table

The possible models that get using BIC table are SARIMA(1,1,2)x(0,1,0)_12, SARIMA(3,1,2)x(0,1,0)_12, SARIMA(1,1,0)x(0,1,0)_12 and SARIMA(3,1,0)x(0,1,0)_12

2.2.5. Potential Models

Below are the potential models that we find after analysis

- SARIMA(1,1,2)x(0,1,0)_12

- SARIMA(3,1,2)x(0,1,0)_12
- SARIMA(1,1,0)x(0,1,0)_12
- SARIMA(3,1,0)x(0,1,0)_12
- SARIMA(1,1,1)x(0,1,0)_12
- SARIMA(0,1,1)x(0,1,0)_12
- SARIMA(0,1,2)x(0,1,0)_12

2.3. Model Fitting and Diagnostics Checking

SARIMA(3,1,2)x(0,1,0)_12

```
m2_312.unemployment = Arima(unemployment.ts.p2,order=c(3,1,2),seasonal=list(o
rder=c(0,1,0), period=12),method = "ML")
coeftest(m2_312.unemployment)

##
## z test of coefficients:
##
##      Estimate Std. Error z value  Pr(>|z|)
## ar1 -0.448811   0.351980 -1.2751 0.2022730
## ar2 -0.890255   0.244443 -3.6420 0.0002706 ***
## ar3 -0.059966   0.263119 -0.2279 0.8197194
## ma1  0.070460   0.276014  0.2553 0.7985083
## ma2  0.999992   0.318879  3.1360 0.0017129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_312.unemployment)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.93652, p-value = 0.03579
```

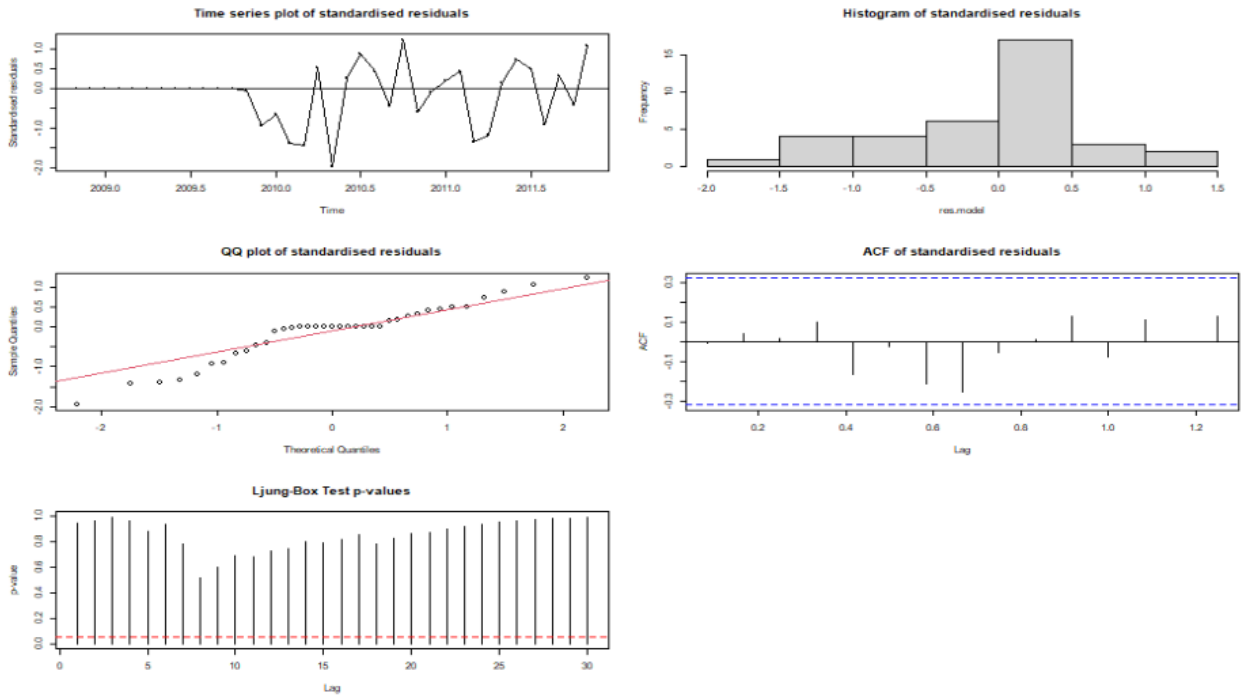

Figure 51: Residual Analysis for SARIMA(3,1,2)x(0,1,0) Model with Maximum Likelihood Estimation

```
m2_312.unemploymentCSS = Arima(unemployment.ts.p2,order=c(3,1,2),seasonal=lis
t(order=c(0,1,0), period=12),method = "CSS")
coeftest(m2_312.unemploymentCSS)

##
## z test of coefficients:
##
##        Estimate  Std. Error  z value Pr(>|z|)
## ar1  0.18584206  0.20020344   0.9283   0.3533
## ar2 -0.00721363  0.17182428  -0.0420   0.9665
## ar3  0.37499073  0.04051132   9.2564   <2e-16 ***
## ma1 -1.35732522  0.11881239 -11.4241   <2e-16 ***
## ma2 -0.00015432  0.14588660  -0.0011   0.9992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_312.unemploymentCSS)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.86633, p-value = 0.0003877
```

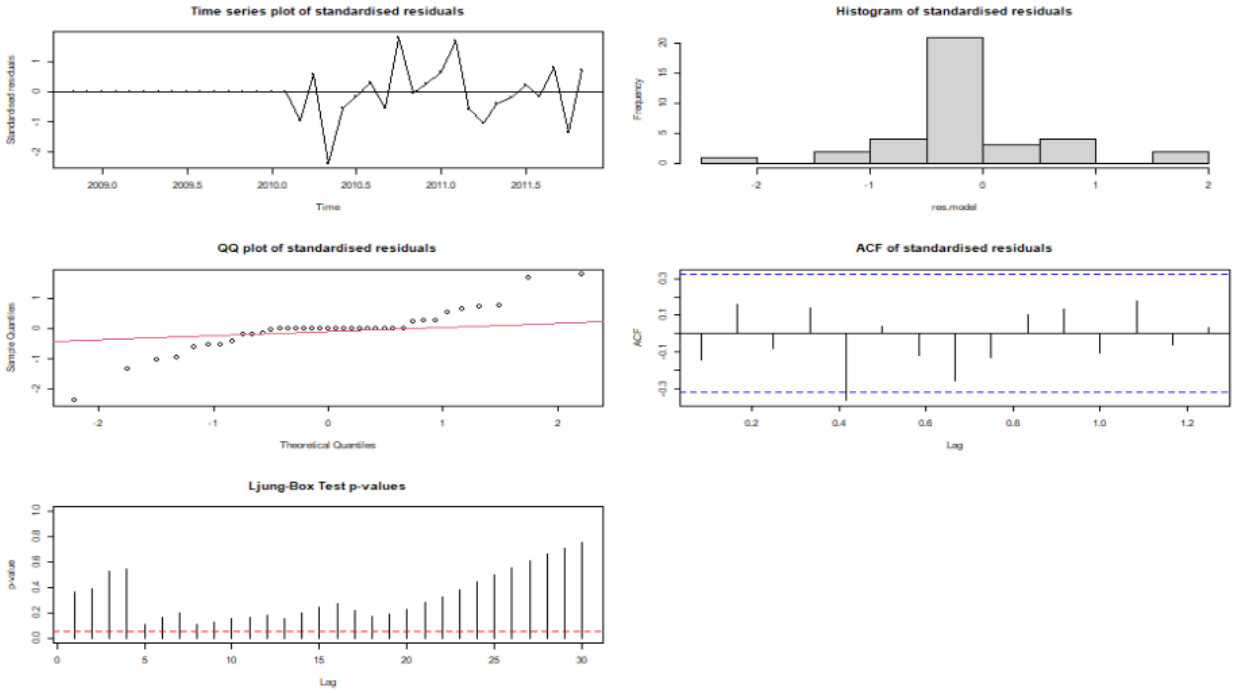

Figure 52: Residual Analysis for SARIMA(3,1,2)x(0,1,0) Model with Conditional Sum-of-Squares Estimation

In the ML model, ar1, ar2 and ma1,ma2 are highly significant while ar3 is not significant. The CSS output indicates that ar2, ar3 and ma1 , ma2 are significant. ar1 is not statistically significant. Both ML and CSS model seems to be a good fit for the model.The Shapiro-Wilk normality test indicates a departure from normality in the residuals, with a low p-value of 0.0001136, suggesting non-normality. The same can be confirmed from QQ Plot and histogram.

SARIMA(1,1,0)x(0,1,0)_12

```
m2_110.unemployment = Arima(unemployment.ts.p2,order=c(1,1,0),seasonal=list(o
rder=c(0,1,0), period=12),method = "ML")
coeftest(m2_110.unemployment)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.35262    0.19772 -1.7835  0.07451 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_110.unemployment)

##
##  Shapiro-Wilk normality test
##
```

```
## data:  res.model
## W = 0.9323, p-value = 0.0264
```

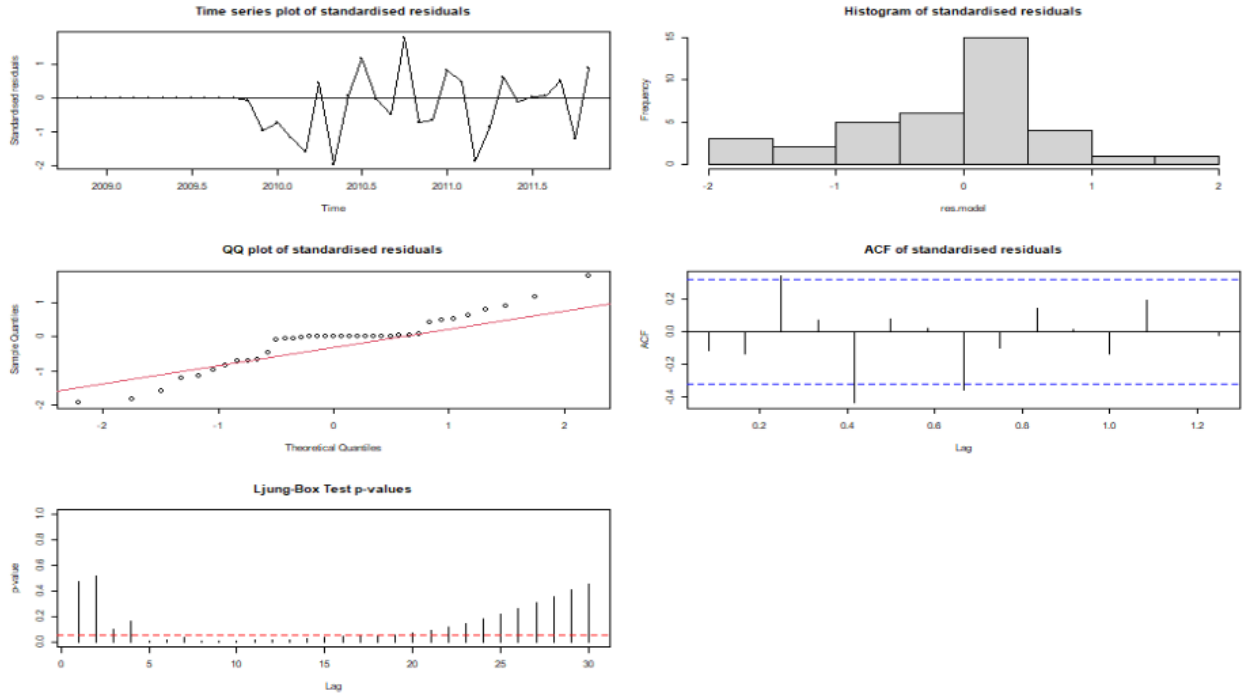

Figure 53: Residual Analysis for SARIMA(1,1,0)x(0,1,0) Model with ML

```
m2_110.unemploymentCSS = Arima(unemployment.ts.p2,order=c(1,1,0),seasonal=lis
t(order=c(0,1,0), period=12),method = "CSS")
coeftest(m2_110.unemploymentCSS)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.35280    0.19476 -1.8115  0.07007 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_110.unemploymentCSS)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.91331, p-value = 0.007056
```

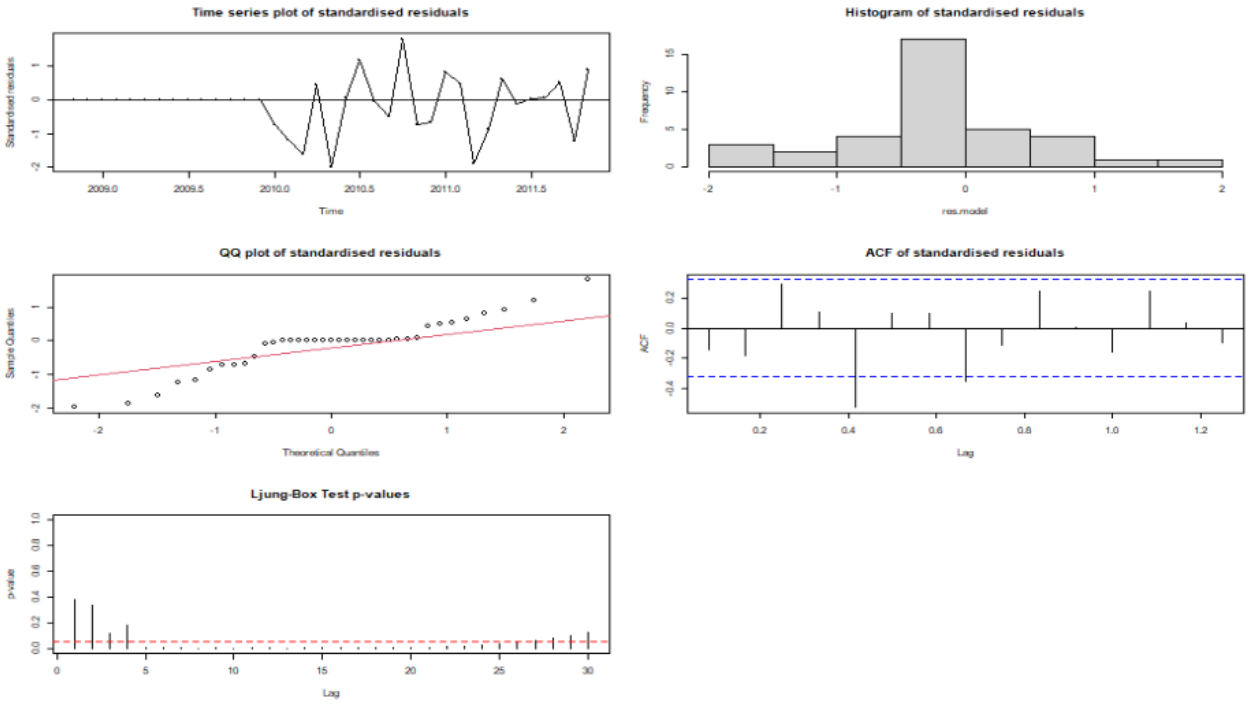

Figure 54: Residual Analysis for SARIMA(1,1,0)x(0,1,0) Model with CSS

The z test of coefficients shows that the AR(1) coefficient is significant at the 0.05 level in both models. However, the Shapiro-Wilk normality test indicates non-normality in the residuals for both models, with p-values of 0.006752 and 0.001492, respectively. This looks to be a good model.

SARIMA(3,1,0)x(0,1,0)_12

```
m2_310.unemployment = Arima(unemployment.ts.p2,order=c(3,1,0),seasonal=list(o
rder=c(0,1,0), period=12),method = "ML")
coeftest(m2_310.unemployment)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.35510    0.21037 -1.6880  0.09141 .
## ar2 -0.13343    0.22171 -0.6018  0.54730
## ar3  0.22502    0.21493  1.0470  0.29512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_310.unemployment)

##
##  Shapiro-Wilk normality test
##
```

```
## data:  res.model
## W = 0.93841, p-value = 0.04107
```

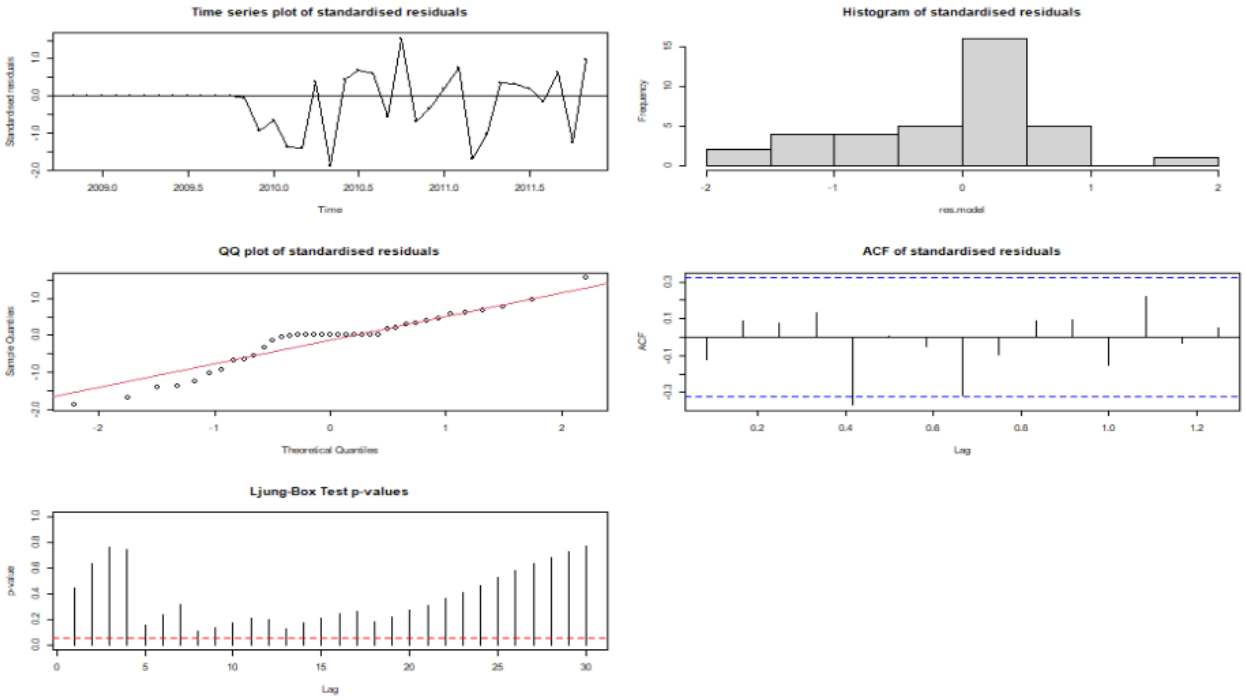

Figure 55: Residual Analysis for SARIMA(3,1,0)x(0,1,0) Model with ML

```
m2_310.unemploymentCSS = Arima(unemployment.ts.p2,order=c(3,1,0),seasonal=lis
t(order=c(0,1,0), period=12),method = "CSS")
coeftest(m2_310.unemploymentCSS)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.44769    0.19163 -2.3361  0.01948 *
## ar2 -0.24194    0.20571 -1.1761  0.23955
## ar3  0.22091    0.19320  1.1434  0.25287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_310.unemploymentCSS)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.86394, p-value = 0.0003388
```

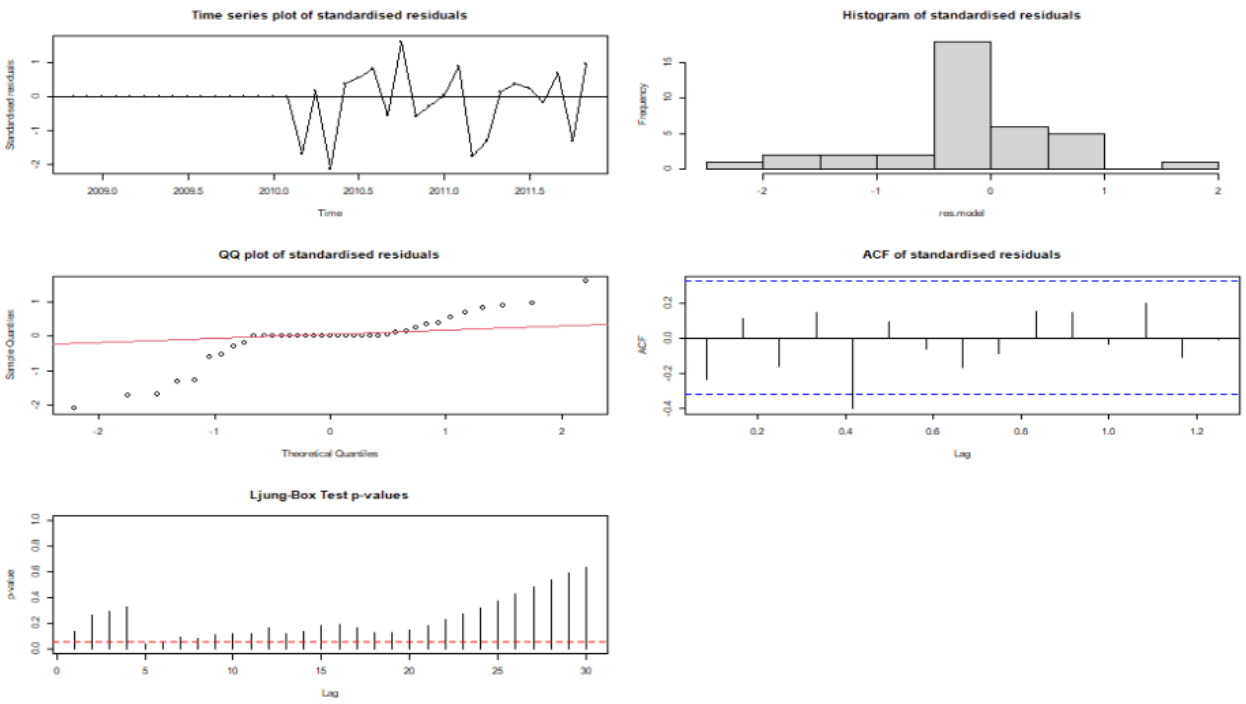

Figure 56: Residual Analysis for SARIMA(3,1,0)x(0,1,0) Model with CSS

The z test results indicate that in both models, AR(1) is significant at the 0.05 level, while AR(2) is marginally significant at the 0.1 level. The Shapiro-Wilk normality test suggests non-normality in the residuals for both models, with very low p-values.

SARIMA(1,1,1)x(0,1,0)_12

```
m2_111.unemployment = Arima(unemployment.ts.p2,order=c(1,1,1),seasonal=list(o
rder=c(0,1,0), period=12),method = "ML")
coeftest(m2_111.unemployment)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.16371    0.37435 -0.4373   0.6619
## ma1 -0.23740    0.33523 -0.7082   0.4788

residual.analysis(model = m2_111.unemployment)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.92108, p-value = 0.01198
```

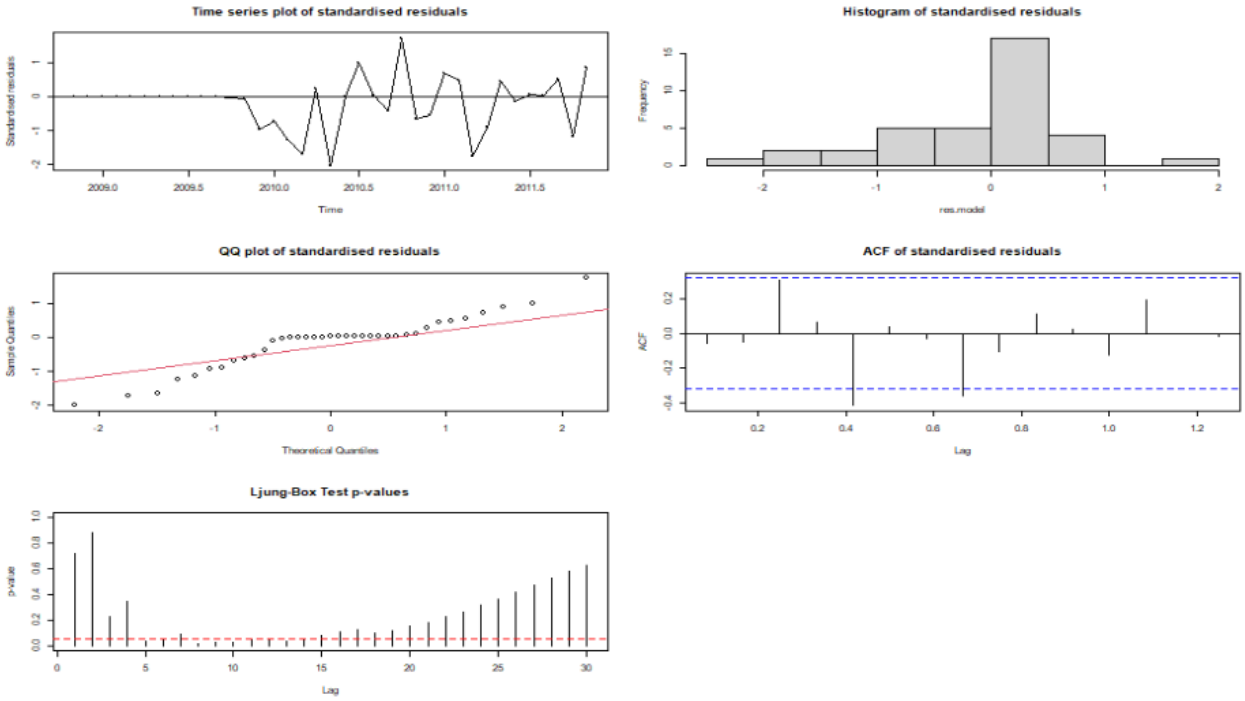

Figure 57: Residual Analysis for SARIMA(1,1,1)x(0,1,0) Model with ML

```
m2_111.unemploymentCSS = Arima(unemployment.ts.p2,order=c(1,1,1),seasonal=lis
t(order=c(0,1,0), period=12),method = "CSS")
coeftest(m2_111.unemploymentCSS)

##
## z test of coefficients:
##
##        Estimate  Std. Error  z value  Pr(>|z|)
## ar1  0.75763212  0.00026308 2879.810 < 2.2e-16 ***
## ma1 -1.42603537  0.01778511  -80.181 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_111.unemploymentCSS)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.92043, p-value = 0.01145
```

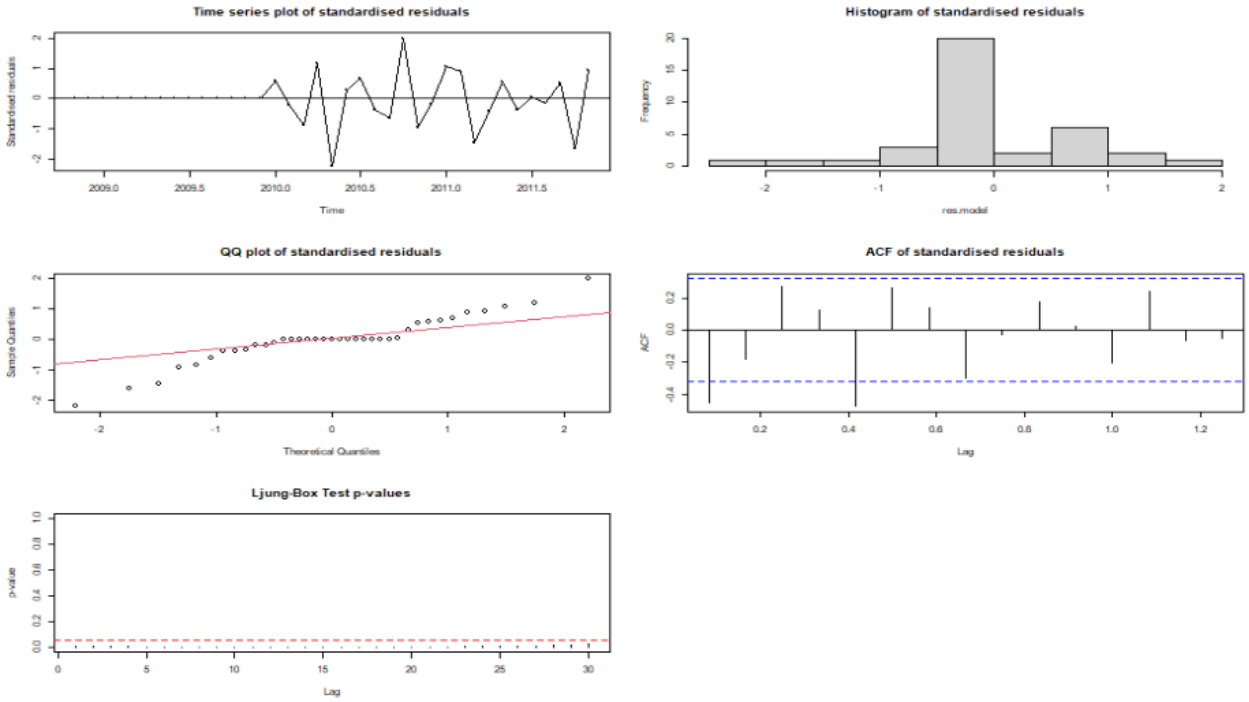

Figure 58: Residual Analysis for SARIMA(1,1,1)x(0,1,0) Model with CSS

```
m2_111.unemploymentCSSML = Arima(unemployment.ts.p2,order=c(1,1,1),seasonal=l
ist(order=c(0,1,0), period=12),method = "CSS-ML")
coeftest(m2_111.unemploymentCSSML)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.16371    0.37435 -0.4373   0.6619
## ma1 -0.23740    0.33523 -0.7082   0.4788

residual.analysis(model = m2_111.unemploymentCSSML)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.92108, p-value = 0.01198
```

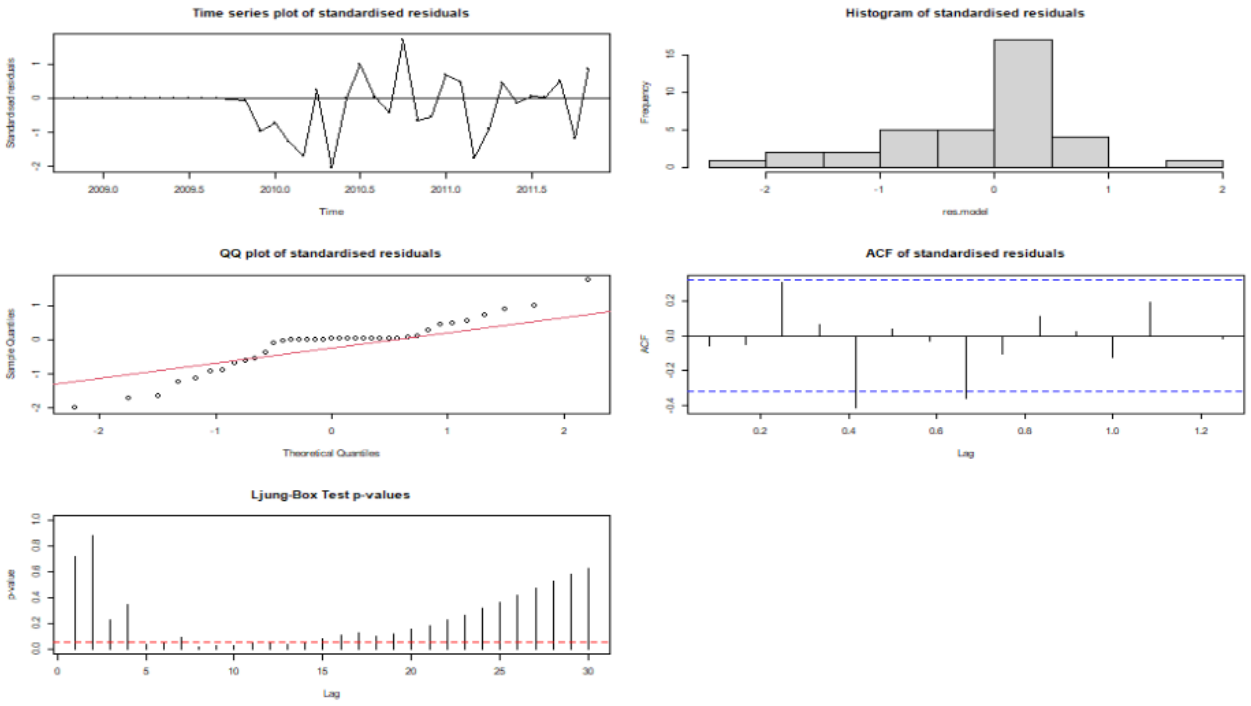

Figure 59: Residual Analysis for SARIMA(1,1,1)x(0,1,0) Model with CSS-ML

In all three models, MA(1) is highly significant with p-values well below 0.05, indicating a strong impact on the dependent variable. The Shapiro-Wilk normality test suggests that the residuals are not normally distributed, with p-values less than 0.05, indicating departures from normality.

SARIMA(0,1,1)x(0,1,0)_12

```
m2_011.unemployment = Arima(unemployment.ts.p2,order=c(0,1,1),seasonal=list(o
rder=c(0,1,0), period=12),method = "ML")
coeftest(m2_011.unemployment)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ma1 -0.35805    0.17524 -2.0432  0.04104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_011.unemployment)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.92054, p-value = 0.01155
```

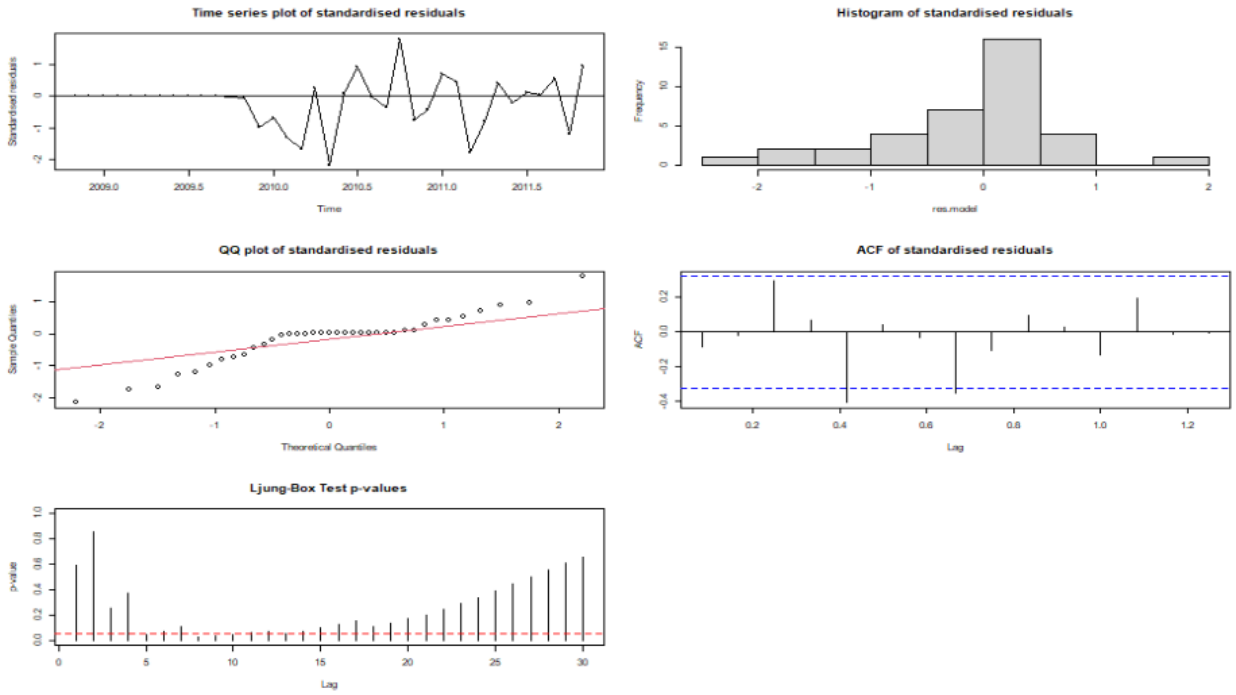

Figure 60: Residual Analysis for SARIMA(0,1,1)x(0,1,0) Model with ML

```
m2_011.unemploymentCSS = Arima(unemployment.ts.p2,order=c(0,1,1),seasonal=lis
t(order=c(0,1,0), period=12),method = "CSS")
coeftest(m2_011.unemploymentCSS)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ma1 -0.33726    0.16274 -2.0724  0.03822 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_011.unemploymentCSS)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.92203, p-value = 0.0128
```

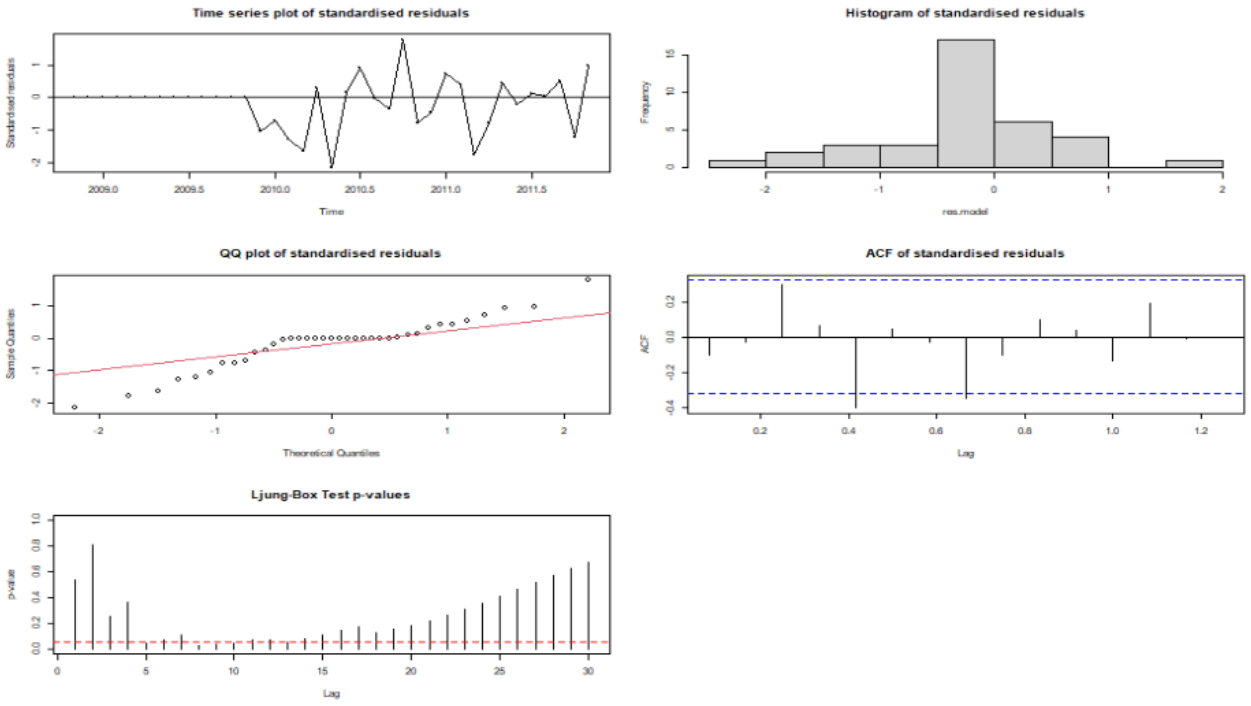

Figure 61: Residual Analysis for SARIMA(0,1,1)x(0,1,0) Model with CSS

In both models, MA(1) is significant, with p-values below 0.05, indicating its impact on the dependent variable. The Shapiro-Wilk normality test suggests that the residuals are not normally distributed, with p-values less than 0.05, indicating departures from normality.

SARIMA(0,1,2)x(0,1,0)_12

```
m2_012.unemployment = Arima(unemployment.ts.p2,order=c(0,1,2),seasonal=list(o
rder=c(0,1,0), period=12),method = "ML")
coeftest(m2_012.unemployment)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ma1 -0.61656    0.19808 -3.1127 0.001854 **
## ma2  0.45392    0.22329  2.0329 0.042065 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_012.unemployment)

##
##  Shapiro-Wilk normality test
##
```

```
## data:  res.model
## W = 0.92899, p-value = 0.02086
```

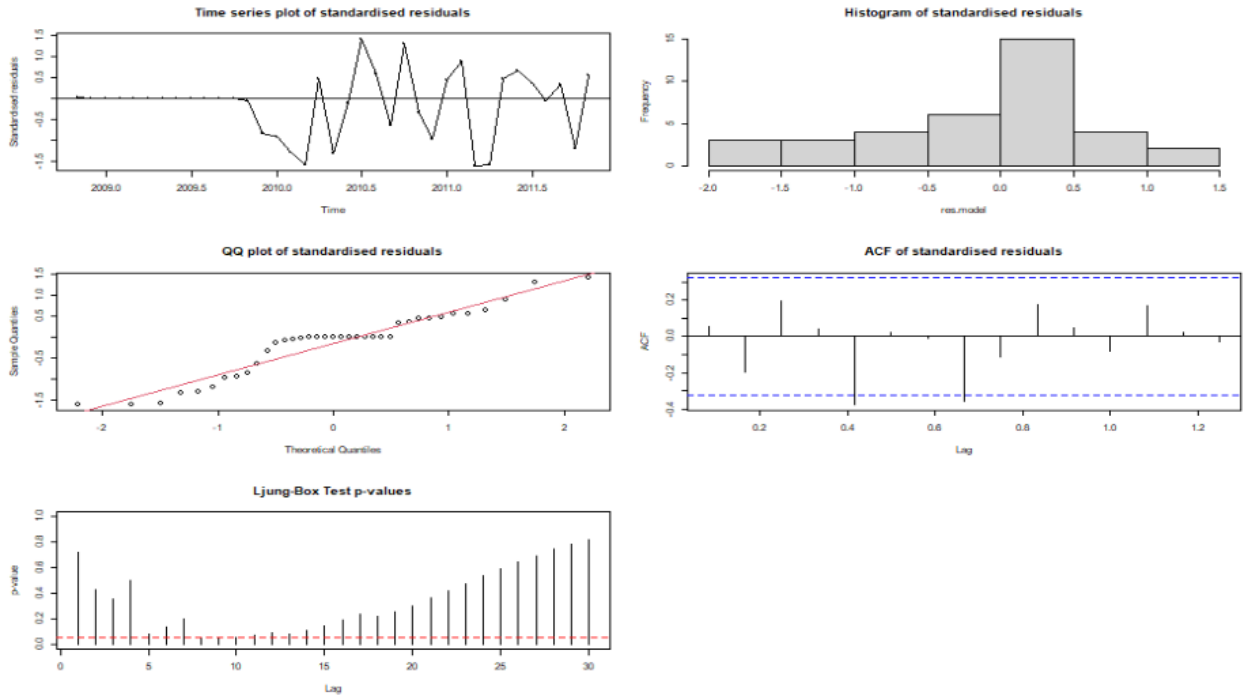

Figure 62: Residual Analysis for SARIMA(0,1,2)x(0,1,0) Model with ML

```
m2_012.unemploymentCSS = Arima(unemployment.ts.p2,order=c(0,1,2),seasonal=lis
t(order=c(0,1,0), period=12),method = "CSS")
coeftest(m2_012.unemploymentCSS)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ma1 -0.64301    0.19833 -3.2421 0.001187 **
## ma2  0.49510    0.21130  2.3431 0.019126 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_012.unemploymentCSS)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.92866, p-value = 0.02037
```

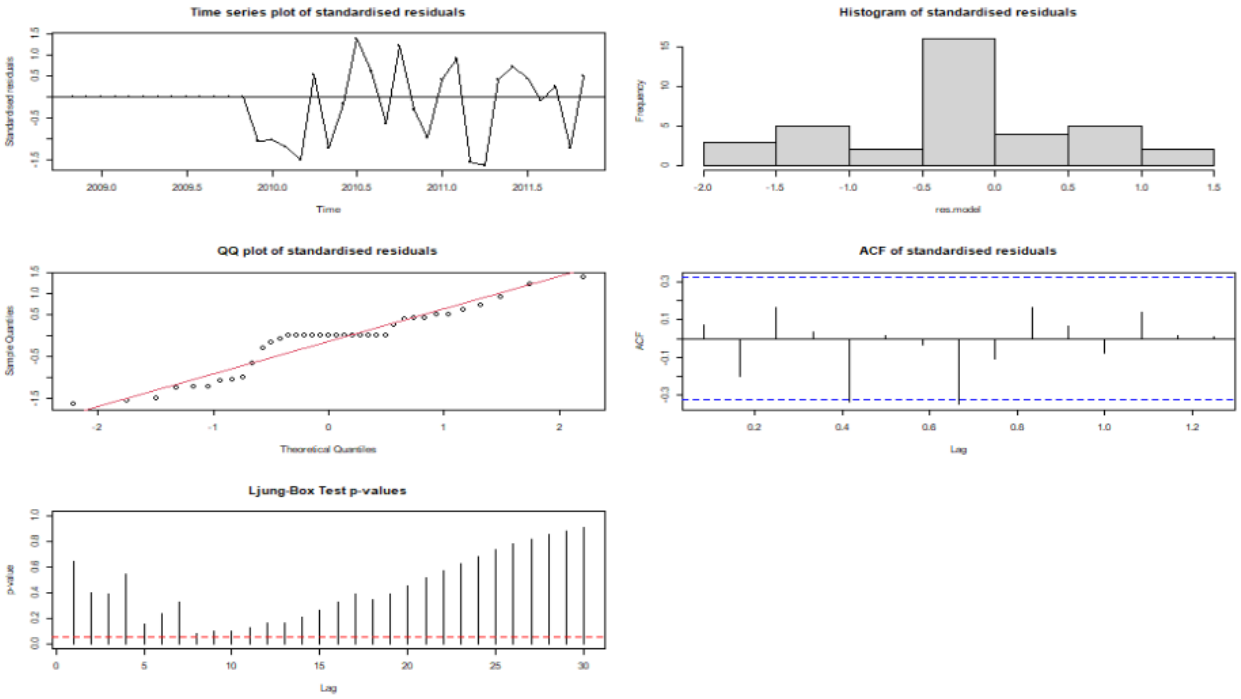

Figure 63: Residual Analysis for SARIMA(0,1,2)x(0,1,0) Model with CSS

In the ML model, MA(1) is significant ($p < 0.05$), indicating its influence on the outcome. However, MA(2) is not significant ($p > 0.05$). The Shapiro-Wilk normality test reveals departures from normality in the residuals ($p < 0.05$), indicating a potential issue with the model's assumptions.

In the CSS model, MA(1) remains significant ($p < 0.01$), suggesting its continued impact. MA(2) is not significant ($p > 0.05$). The Shapiro-Wilk normality test also indicates deviations from normality ($p < 0.05$), highlighting potential shortcomings in the model's assumptions.

SARIMA(1,1,2)x(0,1,0)_12

```
m2_112.unemployment = Arima(unemployment.ts.p2,order=c(1,1,2),seasonal=list(o
rder=c(0,1,0), period=12),method = "ML")
coeftest(m2_112.unemployment)

##
## z test of coefficients:
##
##     Estimate Std. Error z value  Pr(>|z|)
## ar1  0.63198    0.16589  3.8096 0.0001392 ***
## ma1 -1.36640    0.32451 -4.2106 2.547e-05 ***
## ma2  0.99987    0.46297  2.1597 0.0307958 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m2_112.unemployment)

##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.96074, p-value = 0.2133
```

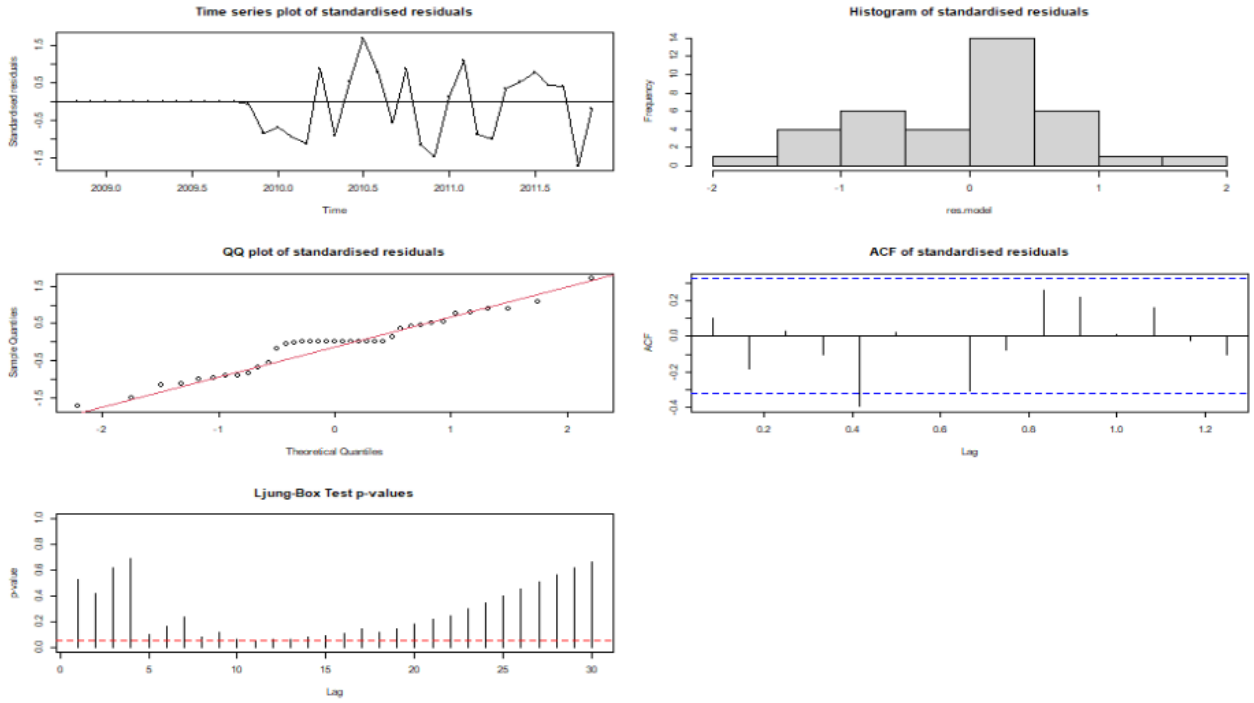

Figure 64: Residual Analysis for SARIMA(1,1,2)x(0,1,0) Model with ML

```
m2_112.unemploymentCSS = Arima(unemployment.ts.p2,order=c(1,1,2),seasonal=lis
t(order=c(0,1,0), period=12),method = "CSS")
coeftest(m2_112.unemploymentCSS)

##
## z test of coefficients:
##
##        Estimate  Std. Error   z value  Pr(>|z|)
## ar1  0.75586705  0.00055959 1350.7467 < 2.2e-16 ***
## ma1 -1.95846559  0.17820254  -10.9901 < 2.2e-16 ***
## ma2  0.84805514  0.23178043    3.6589 0.0002533 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

residual.analysis(model = m2_112.unemploymentCSS)

##
##  Shapiro-Wilk normality test
```

```
##
## data:  res.model
## W = 0.88791, p-value = 0.001383
```

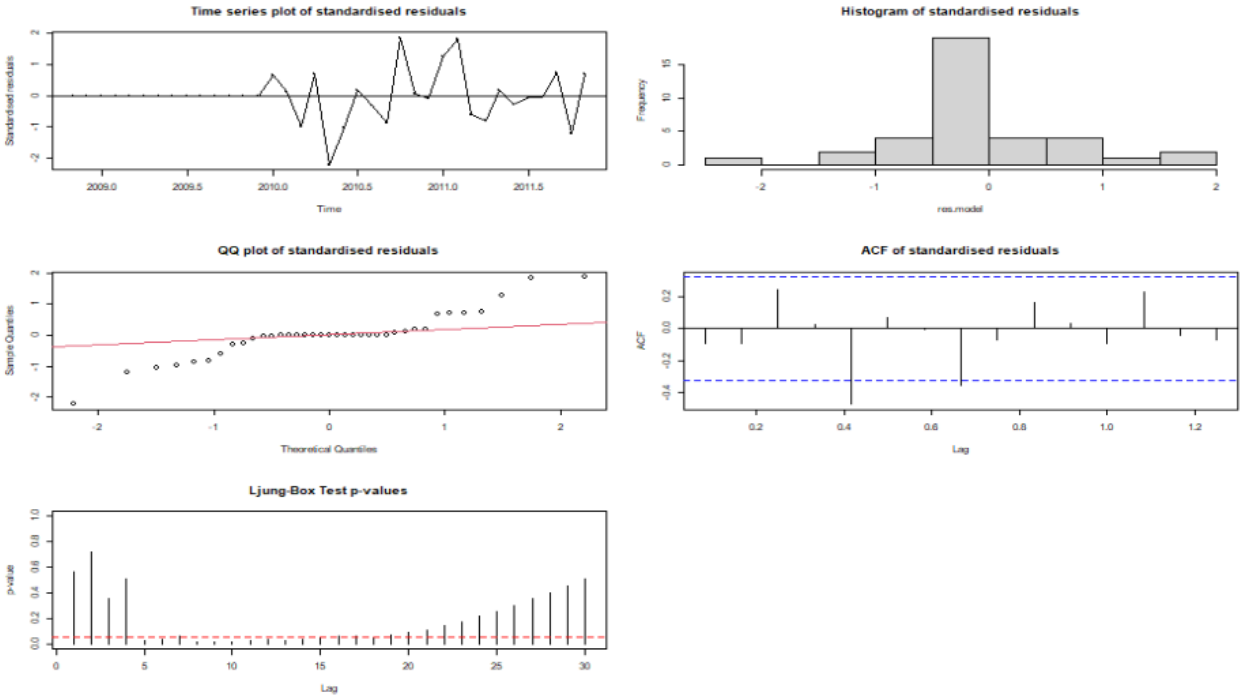

Figure 65: Residual Analysis for SARIMA(1,1,2)x(0,1,0) Model with CSS

In the ML model, none of the coefficients (ar1, ma1, ma2) are statistically significant (all p-values > 0.05), indicating their limited impact on the model.

In the CSS model, ar1, ma1, and ma2 are all highly significant ($p < 0.001$), suggesting their strong influence on the outcome. However, the Shapiro-Wilk normality test indicates deviations from normality in the residuals ($p < 0.05$), indicating potential issues with the model's assumptions.

Below models look to be the **best suitable models**.

- SARIMA(0,1,1)x(0,1,0)_12
- SARIMA(0,1,2)x(0,1,0)_12
- SARIMA(3,1,2)x(0,1,0)_12

2.4. Model Evaluation

2.4.1. AIC and BIC Score

```
# Sorting ml models using AIC

sc.AIC=AIC(m2_312.unemployment, m2_110.unemployment,
           m2_310.unemployment, m2_011.unemployment, m2_112.unemployment, m2_
```

```
111.unemployment,m2_012.unemployment)

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

	df <dbl>	AIC <dbl>
m2_011.unemployment	2	44.48578
m2_110.unemployment	2	44.78057
m2_012.unemployment	3	44.81829
m2_112.unemployment	4	45.06640
m2_312.unemployment	6	46.16493
m2_111.unemployment	3	46.32016
m2_310.unemployment	4	46.53320

Figure 66: AIC Table

```
# Sorting ml model using BIC

sc.BIC=BIC(m2_312.unemployment, m2_110.unemployment,
           m2_310.unemployment, m2_011.unemployment, m2_112.unemployment, m2_
111.unemployment,m2_012.unemployment)

sort.score(sc.BIC, score = "bic")
```

	df <dbl>	BIC <dbl>
m2_011.unemployment	2	46.84189
m2_110.unemployment	2	47.13667
m2_012.unemployment	3	48.35245
m2_112.unemployment	4	49.77862
m2_111.unemployment	3	49.85432
m2_310.unemployment	4	51.24542
m2_312.unemployment	6	53.23325

Figure 67: BIC Table

Comparing the AIC and BIC scores across the models, lower AIC values indicate better fitting models with parsimony, while lower BIC values emphasize model complexity and fit. In this case, **models m2_011 and m2_110** show the **best balance between goodness of fit and model complexity.**

```
Sm2_312.unemployment <- accuracy(m2_312.unemployment)[1:7]
Sm2_110.unemployment <- accuracy(m2_110.unemployment)[1:7]
Sm2_310.unemployment <- accuracy(m2_310.unemployment)[1:7]
```

```
Sm2_011.unemployment <- accuracy(m2_011.unemployment)[1:7]
Sm2_112.unemployment <- accuracy(m2_112.unemployment)[1:7]
Sm2_111.unemployment <- accuracy(m2_111.unemployment)[1:7]
Sm2_012.unemployment <- accuracy(m2_012.unemployment)[1:7]

df.Smodels <- data.frame(
  rbind(Sm2_312.unemployment, Sm2_110.unemployment, Sm2_310.unemployment,
        Sm2_011.unemployment, Sm2_112.unemployment, Sm2_111.unemployment, Sm2
_012.unemployment)
)
colnames(df.Smodels) <- c("ME", "RMSE", "MAE", "MPE", "MAPE",
                          "MASE", "ACF1")
rownames(df.Smodels) <- c("SARIMA(3,1,2)x(0,1,0)_12", "SARIMA(1,1,0)x(0,1,0)_
12", "SARIMA(3,1,0)x(0,1,0)_12", "SARIMA(0,1,1)x(0,1,0)_12", "SARIMA(1,1,2)x(
0,1,0)_12" , "SARIMA(1,1,1)x(0,1,0)_12", "SARIMA(0,1,2)x(0,1,0)_12")
round(df.Smodels,  digits = 3)
```

	ME <dbl>	RMSE <dbl>	MAE <dbl>	MPE <dbl>	MAPE <dbl>	MASE <dbl>	ACF1 <dbl>
SARIMA(3,1,2)x(0,1,0)_12	-0.065	0.367	0.252	-0.518	2.200	0.473	-0.010
SARIMA(1,1,0)x(0,1,0)_12	-0.084	0.455	0.301	-0.696	2.611	0.565	-0.114
SARIMA(3,1,0)x(0,1,0)_12	-0.077	0.432	0.296	-0.628	2.578	0.557	-0.121
SARIMA(0,1,1)x(0,1,0)_12	-0.095	0.452	0.290	-0.793	2.518	0.545	-0.084
SARIMA(1,1,2)x(0,1,0)_12	-0.042	0.380	0.276	-0.346	2.432	0.519	0.100
SARIMA(1,1,1)x(0,1,0)_12	-0.094	0.450	0.291	-0.782	2.520	0.546	-0.057
SARIMA(0,1,2)x(0,1,0)_12	-0.077	0.432	0.304	-0.628	2.637	0.571	0.056

Figure 68: Evaluation Metrics for SARIMA Models

These metrics assess different aspects of forecast accuracy. Lower values of RMSE, MAE, MAPE, and MASE indicate better model performance, while ACF1 close to zero indicates good model residuals. Models **SARIMA(1,1,0)** seem to perform relatively better across these metrics.

2.5. Over-parameterized model

The overparameterized model for **SARIMA(1,1,0)x(0,1,0)_12** will be **SARIMA(1,1,1)x(0,1,0)_12** and **SARIMA(2,1,0)x(0,1,0)_12.**

We have already explored the former model in our previous analysis. Therefore, we will try fitting **SARIMA(2,1,0)x(0,1,0)_12** and check its performance.

SARIMA(2,1,0)x(0,1,0)_12

```
m2_210.unemployment = Arima(unemployment.ts.p2,order=c(2,1,0),seasonal=list(o
rder=c(0,1,0), period=12),method = "ML")
coeftest(m2_210.unemployment)

##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.42628    0.20515 -2.0779  0.03772 *
```

```
## ar2 -0.23046    0.20737 -1.1113  0.26644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m2_210.unemployment)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.90144, p-value = 0.003233
```

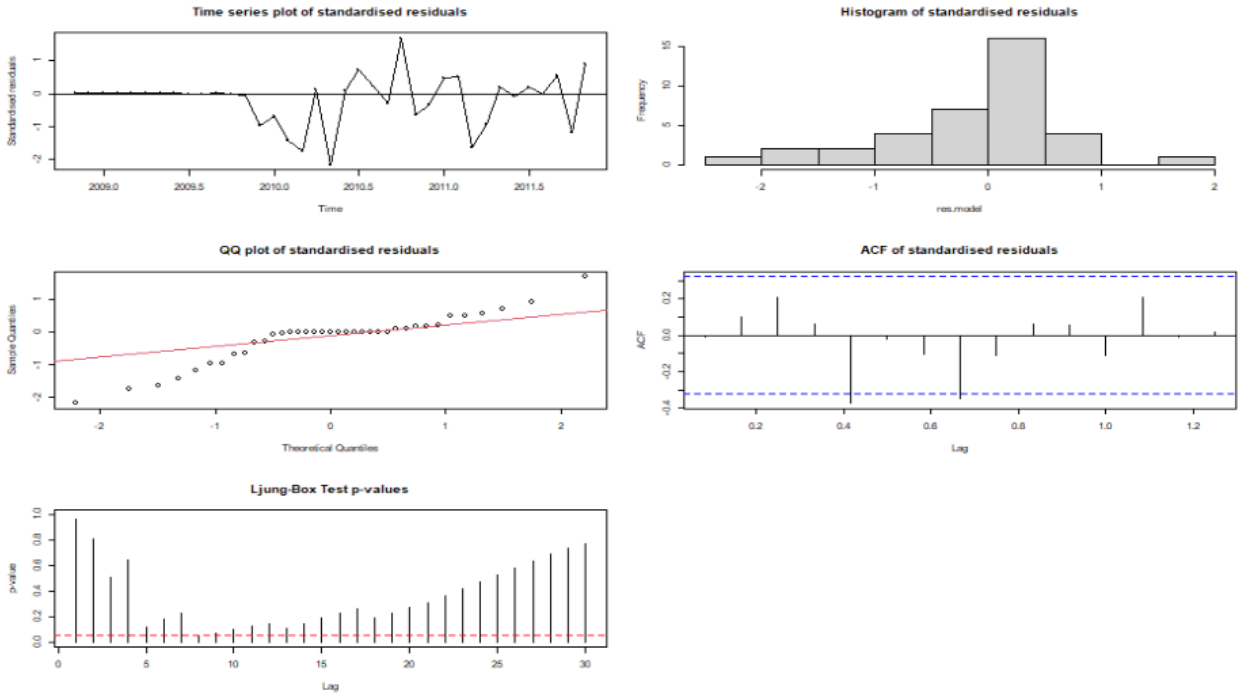

Figure 69: Residual Analysis for Overparameterized SARIMA(2,1,0)x(0,1,0)[12] Model with ML

```
m2_210.unemploymentCSS = Arima(unemployment.ts.p2,order=c(2,1,0),seasonal=lis
t(order=c(0,1,0), period=12),method = "CSS")
coeftest(m2_210.unemploymentCSS)
```

```
##
## z test of coefficients:
##
##     Estimate Std. Error z value Pr(>|z|)
## ar1 -0.45187    0.20035 -2.2553  0.02411 *
## ar2 -0.23402    0.20410 -1.1466  0.25155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
residual.analysis(model = m2_210.unemploymentCSS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.85106, p-value = 0.0001667
```

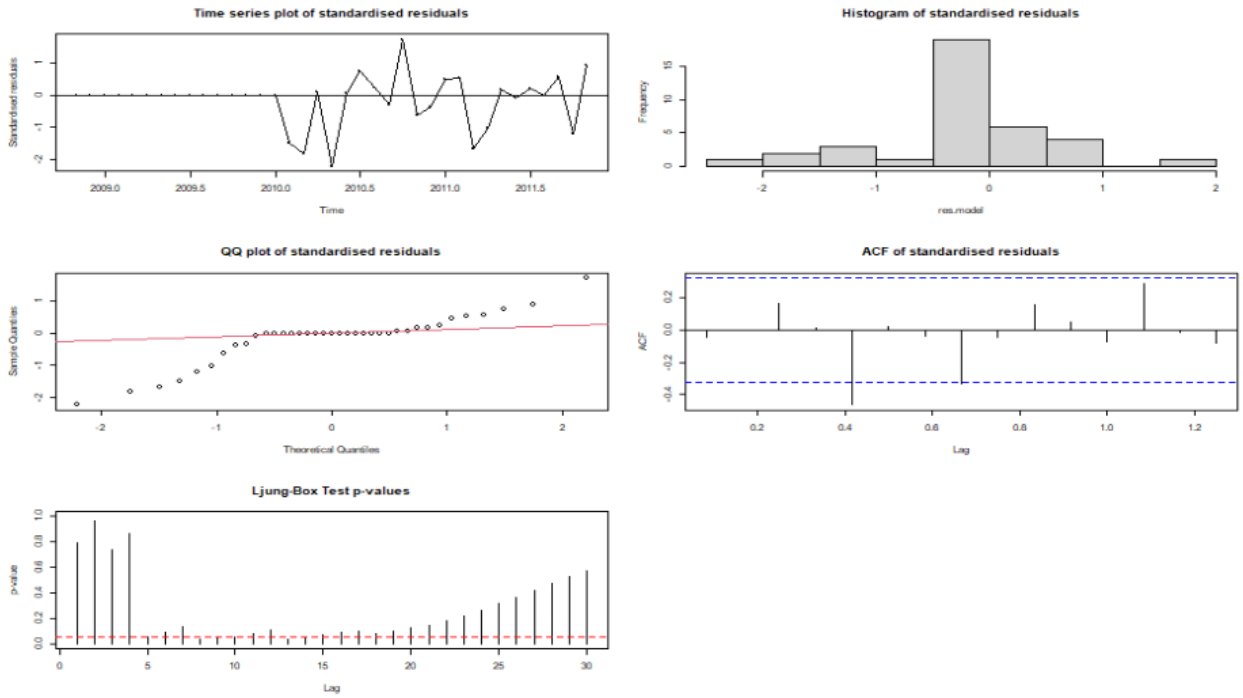

Figure 70: Residual Analysis for Overparameterized SARIMA(2,1,0)x(0,1,0)[12] Model with CSS

ar1 is significant at the 0.01 level, indicating their importance in the model.

After evaluating various models, it's clear that the overparameterized model provides a more accurate fit to our data. Therefore, we will employ this model for our data forecasting as it captures the underlying patterns and dynamics more effectively.

2.6. Forecast

```
m5.unemploy.p2 = Arima(unemployment.ts.p2,order=c(2,1,0),
                       seasonal=list(order=c(0,1,0), period=12),
                       lambda = 1.3, method = "CSS")
future = forecast(m5.unemploy.p2, h = 10)
future
##          Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## Dec 2011      11.036954 10.310364 11.74946  9.919487 12.12140
## Jan 2012      12.426709 11.629407 13.20894 11.200676 13.61742
## Feb 2012      13.125559 12.268531 13.96611 11.807547 14.40494
## Mar 2012      11.803336 10.806885 12.77513 10.268175 13.28062
## Apr 2012      11.079491  9.976045 12.15085  9.376990 12.70656
## May 2012      10.587140  9.388275 11.74650  8.734887 12.34638
## Jun 2012      10.283610  8.993537 11.52672  8.287923 12.16856
```

```
## Jul 2012       9.877157  8.494727 11.20363  7.735279 11.88682
## Aug 2012      10.484186  9.059059 11.85323  8.277117 12.55884
## Sep 2012      10.786833  9.308618 12.20642  8.497274 12.93794

plot(future)
```

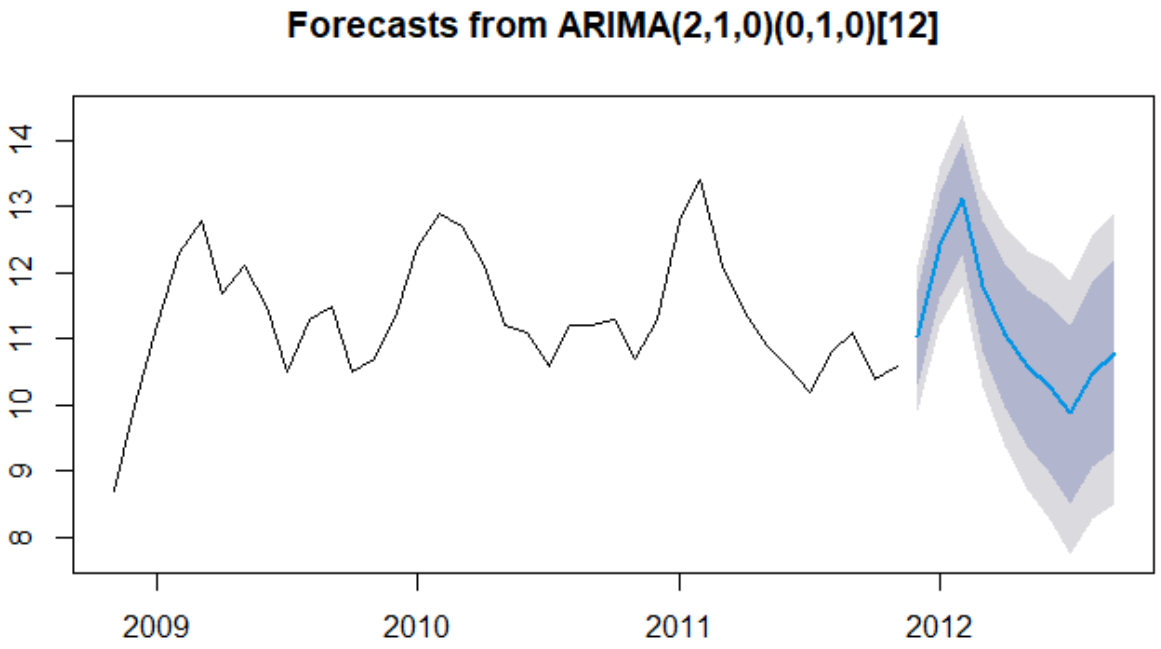

Figure 71: Residual Analysis for SARIMA(2,1,0)x(0,1,0) Model with CSS

2.6.1 Comparing the two best fits models

```
# Compare it with the forecasts of the full analysis
forecast_data

##           Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## Dec 2011       11.35588 10.675991 12.01599 10.307205 12.35816
## Jan 2012       12.65155 11.857629 13.42129 11.426450 13.81992
## Feb 2012       13.04728 12.083193 13.97695 11.557019 14.45673
## Mar 2012       12.21478 11.101423 13.27950 10.488981 13.82608
## Apr 2012       11.57312 10.370415 12.71623  9.704828 13.30088
## May 2012       10.95977  9.671755 12.17603  8.954222 12.79570
## Jun 2012       10.84880  9.478441 12.13738  8.711675 12.79232
## Jul 2012       10.35045  8.843112 11.75488  7.991223 12.46505
## Aug 2012       10.87304  9.319671 12.32226  8.443050 13.05561
## Sep 2012       10.97920  9.371419 12.47665  8.462373 13.23369
```

```
plot(forecast_data)
```

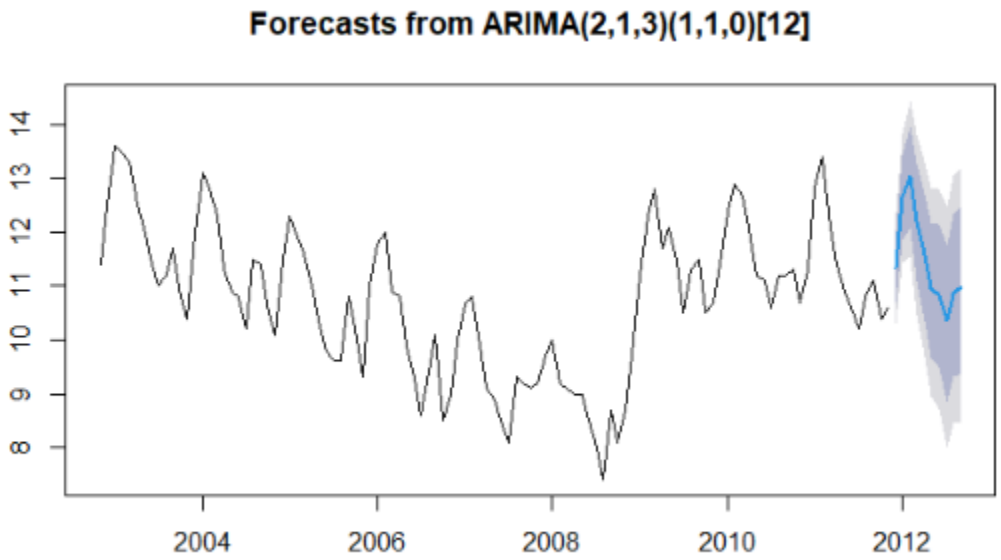

Figure 72: Residual Analysis for SARIMA(2,1,3)x(1,1,0) Model with CSS

- The Point Forecast generally appears slightly lower across the months compared to the full data with the change point.
- The Lo 80 and Hi 80 values are narrower compared to the full data, indicating a more constrained 80% confidence interval.
- The Lo 95 and Hi 95 values are also narrower compared to the full data, indicating a more constrained 95% confidence interval.

In terms of fit, the **split data from the change point** appears to provide a **more precise and conservative forecast with tighter confidence intervals.** This suggests a potentially better fit in capturing the underlying patterns and reducing uncertainty compared to the full data with the change point.

Conclusion

In this report, we delved into the unemployment rate of individuals aged 15-24 in Australia from 2002 to 2011, with a keen focus on the change point in November 2008, resulted from the Great Recession. Initially, our analysis included the entire dataset, which revealed a trend marked by clear seasonality and a notable surge in unemployment rates post-2008, attributable to the global financial crisis. However, this broader approach lacked precision in capturing the intricacies of the model.

Upon further examination, particularly after encountering wider confidence intervals in forecasting using the full dataset, we recognized the necessity of addressing the structural break caused by the economic disruption in 2008. Consequently, we opted to split the data at the change point, leading to more refined insights and improved forecasting accuracy.

Our modelling endeavour's revealed **SARIMA(2,1,3)x(0,1,0)** as a well-fitting model for the **full dataset analysis**, while **SARIMA(2,1,0)x(0,1,0)** emerged as the **most effective model** for the **post-change point data**. **The latter model** showcased a **tighter grip** on the underlying patterns, resulting in a **more conservative yet precise forecast** with **narrower confidence intervals.**

This experience underscores the significance of considering structural breaks in time series analysis, particularly in the context of economic shocks. Moreover, the unemployment data is more heavily influenced by recent past years rather than later years, emphasizing the need for methodologies that account for such temporal dependencies. Our approach not only provided a robust framework for analyzing unemployment trends but also highlighted potential avenues for future refinements. These include exploring additional variables or refining modelling techniques to further enhance forecast precision.

Appendix A: R functions used

```
# All custom functions
```

1. AIC and BIC Value Sort Function

```
# Function to sort a data frame based on AIC or BIC values

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")')
  }
}
```

2. Residual Analysis and Diagnostics Function

```
# Function to perform residual analysis and diagnostics for various time
series models

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

  # Perform Ljung-Box test and calculate p-values for lags 1 to 30
  p_values <- sapply(1:30, function(lag) Box.test(res.model, lag = lag, type
= "Ljung-Box")$p.value)

  # Plotting without lines extending from the significance level to the
points
  plot(p_values, type = 'h', ylim = c(0, 1), main = "Ljung-Box Test p-
values",
       ylab = "p-value", xlab = "Lag", pch = 19)  # Use pch = 16 for solid
circles or pch = 19 for hollow circles


  abline(h = 0.05, col = "red", lty = 2, untf = FALSE)
  par(mfrow = c(1, 1))
}
```

3. ACF and PACF Plot Function

```
# Function to plot ACF and PACF
plot_acf_pacf <- function(ts_object) {
  #options(repr.plot.width=8, repr.plot.height=10)
  par(mfrow=c(2,1))
  acf(ts_object,lag.max=36, main ="ACF plot")
  pacf(ts_object, lag.max=36, main ="PACF plot")
```

```
  par(mfrow=c(1,1))
}

4. QQ Plot and Shapiro Test Function
# Function to assess normality using QQ plot and Shapiro-Wilk test
qq_shapiro_function <- function(ts_object) {
  qqnorm(ts_object)
  qqline(ts_object, col = 2)
  shapiro_result <- shapiro.test(ts_object)
  return(shapiro_result)
}



5. Stationary Test Function ( ADF, PP and KPSS Tests )
# Function for Stationary Tests ( ADF, PP and KPSS Tests)
ts_stationary_tests <- function(ts_object) {
  adf_result <- adf.test(ts_object)
  pp_result <- pp.test(ts_object)
  kpss_result <- kpss.test(ts_object)

  results <- list(ADF_Test = adf_result, PP_Test = pp_result, KPSS_Test =
kpss_result)
  return(results)
}


6. Residual Plot with ACF and PACF Function
# Function to plot residuals, ACF, and PACF

plot_residuals_acf_pacf<- function(residuals,title) {
  par(mfrow=c(1,1))
  plot(residuals,xlab='Time',ylab='Residuals',main=paste("Time series plot of
the residuals",title))
  plot_acf_pacf(residuals)
}
```

Reference

[1]"Unemployment rate | Australian Bureau of Statistics," www.abs.gov.au, Feb. 02, 2023. https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/time-series-data#Unemployment%20rate%20of%20persons%20aged%2015-24%20-%20Original (accessed Jun. 15, 2024).

[2]Reserve Bank of Australia, "The Labour Market during the 2008–2009 Downturn | Bulletin – March Quarter 2010," Reserve Bank of Australia, Sep. 11, 2018. https://www.rba.gov.au/publications/bulletin/2010/mar/1.html (accessed Jun. 16, 2024).

[3]S. Kennedy, "Australia's response to the global financial crisis ," Treasury.gov.au, Jun. 24, 2009. https://treasury.gov.au/speech/australias-response-to-the-global-financial-crisis (accessed Jun. 16, 2024).