ASM - Time Series Project

Authors: Dmitriy Chukhray, Julian Fransen

Date: 2025-01-06

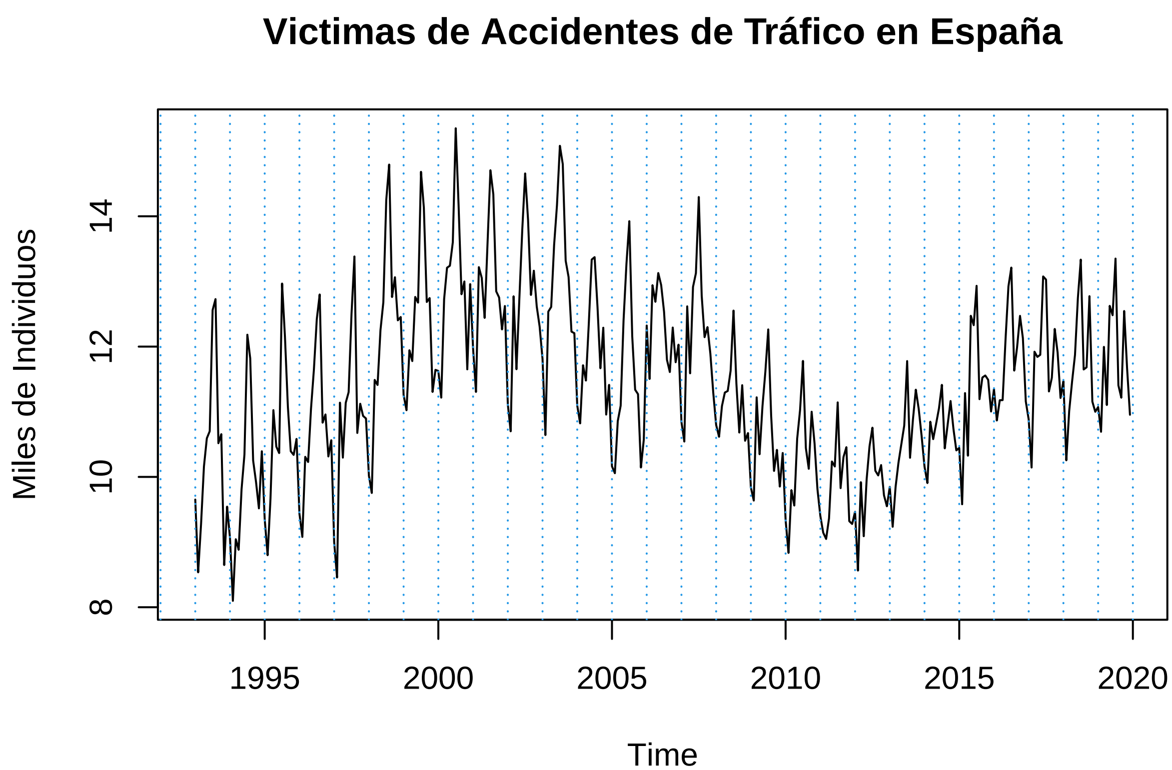
Contents

Introduction

The dataset under analysis consists of monthly data on victims of traffic accidents in Spain, including fatalities, serious injuries, and minor injuries, recorded on urban and interurban roads. In this project we will apply the Box-Jenkins ARIMA methodology to understand the time-series dynamics of these traffic incidents and to make reliable predictions for future trends. Spanning from 1993 to 2019, the dataset captures over two decades of detailed information, offering a unique opportunity to identify trends, seasonal patterns, and underlying factors that influence traffic accidents.

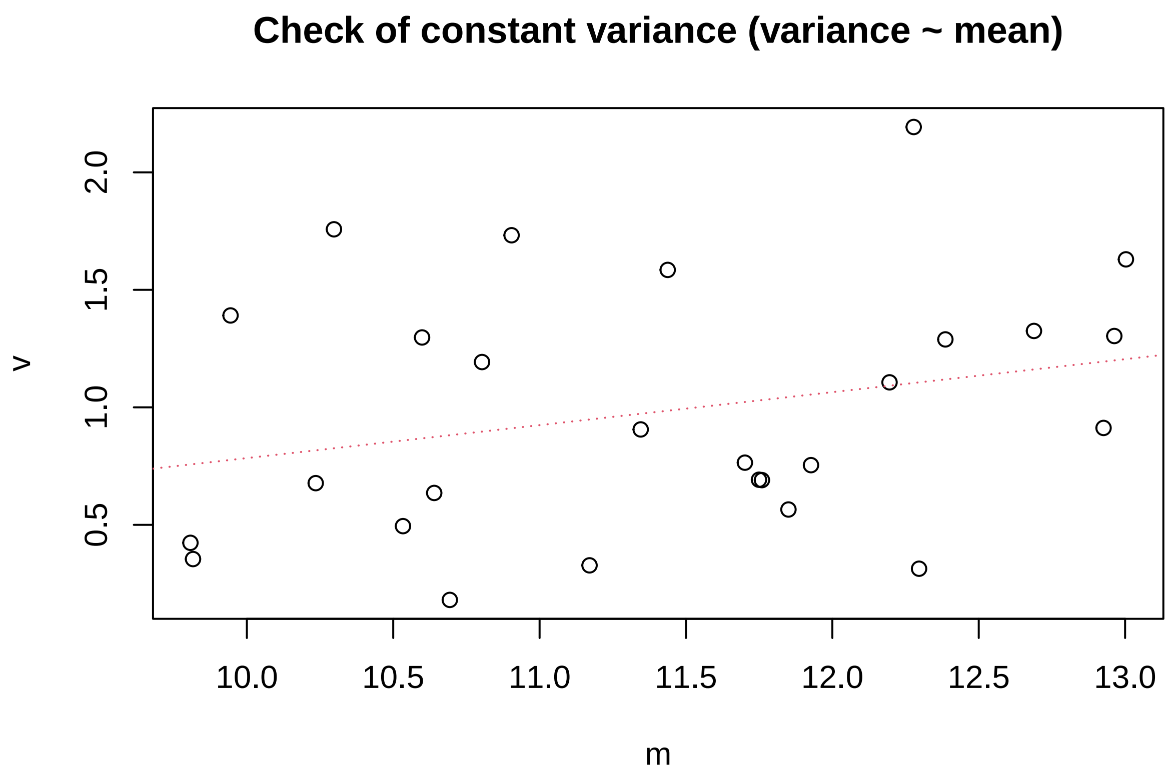
1. Identification

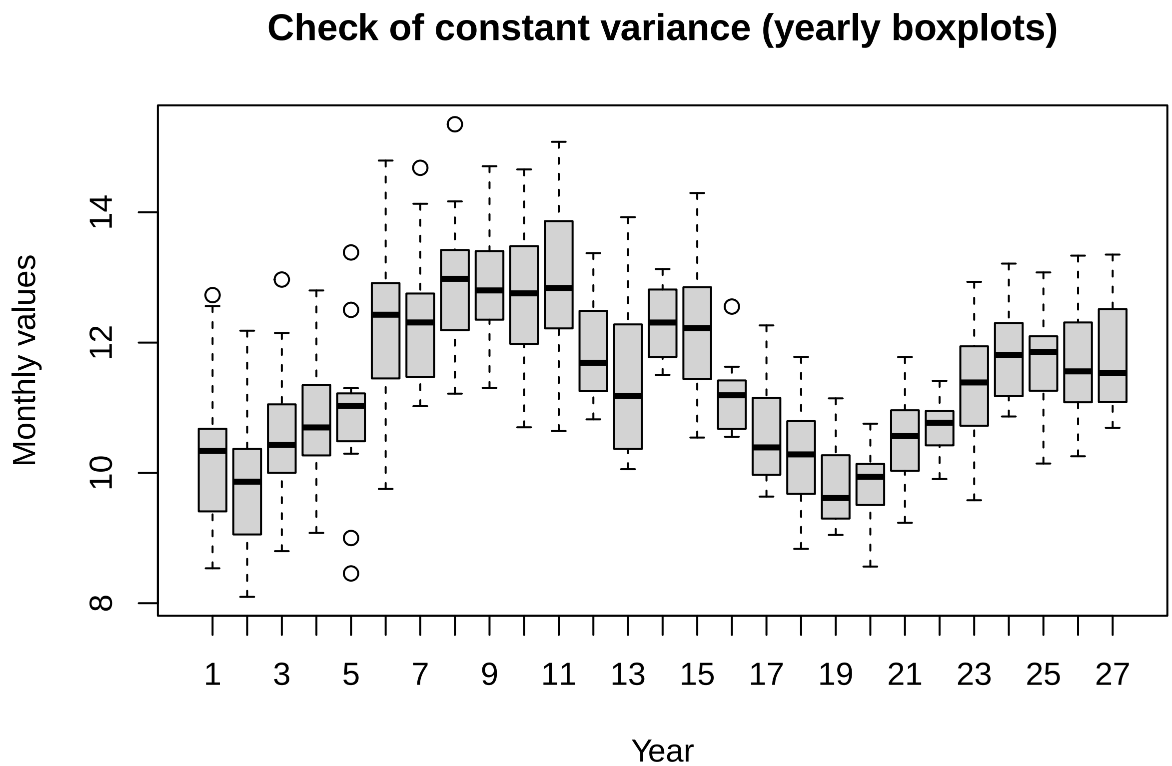
First, we load the raw data set and plot it to get a first impression of the data. Based on this, we will start with the identification of transformations.



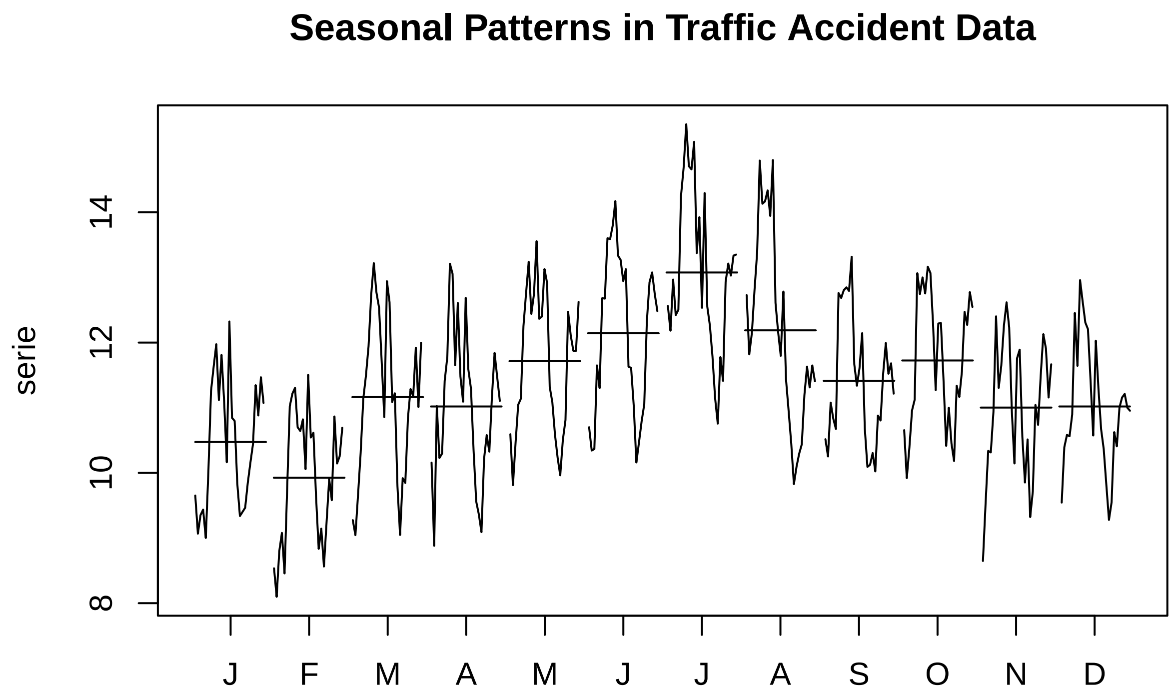
1a. Data transformations

The first step in time series analysis is to check whether transformations are necessary to stabilize variance or remove trends. We start by checking if the data shows constant variance over time. This is important because ARIMA models assume that the variance is constant. To do this, we compute the mean and variance for each year and examine their relationship. Additionally, we visualize yearly boxplots to confirm constant variance over time.

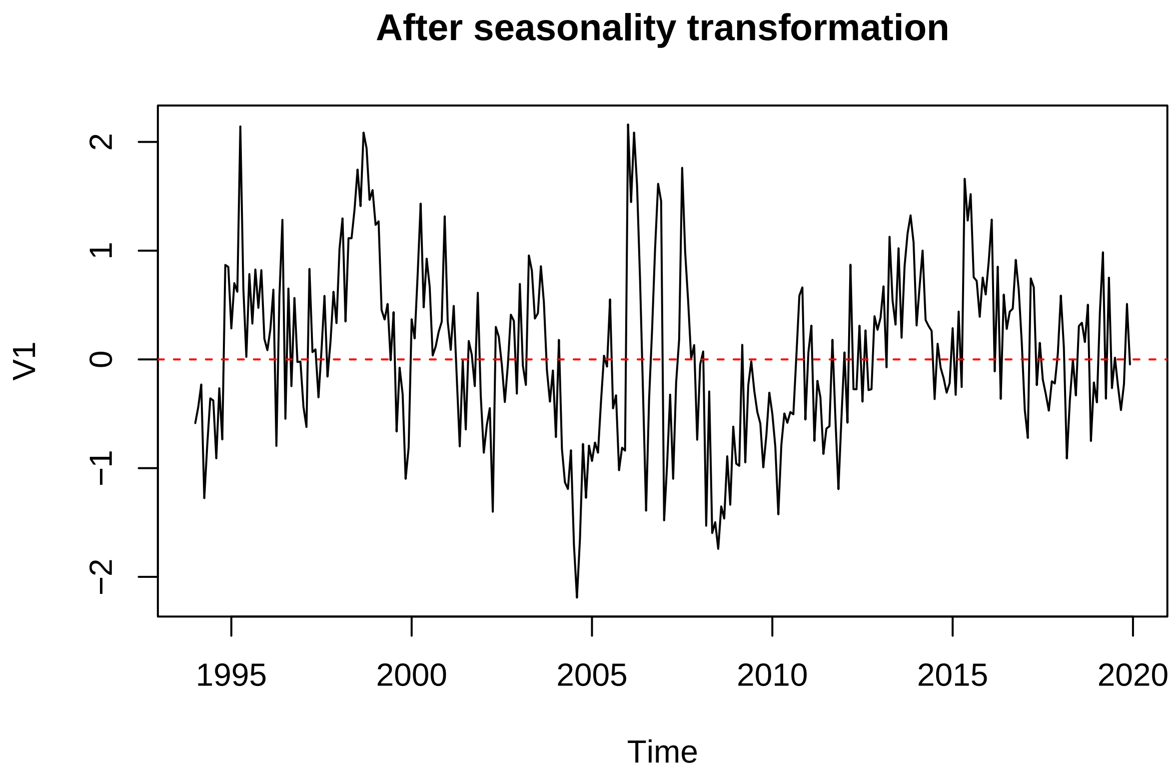




These plots show correct behaviour: `v` is basically uncorrelated with `m`, and the boxplots are similarly sized, implying constant variance. This means that a log transform is not necessary in our case. The next step is to check the existence of a seasonal pattern in the time series. To do that we are using the function `monthplot`.



In the plot above, we can clearly observe a seasonal pattern. If there were no seasonal component, the monthly means would remain at approximately the same level over time, and the shapes of the patterns for each month would not exhibit systematic repetition. These periodic fluctuations suggest that certain months consistently experience higher or lower values, driven by underlying seasonal component(s). To account for this seasonality, we apply a seasonal differencing transformation with a yearly period (12 months).



The last step of achieving the stationary of the time series is to check whether the mean is constant or not. This can be done by examining the plot of the current time series data or by straight forwardly applying regular difference and then examining the change of the time series' variance.

To verify if taking one regular difference is optimal, we calculate the variance of different transformation of the data:

```
1. original data (`serie`)
2. transformed with yearly season transformation (`d12serie`)
3. one regular difference applied to the previous transformation (`d1d12serie`)
4. another regular difference applied to the previous transformation
```

The values are here below:

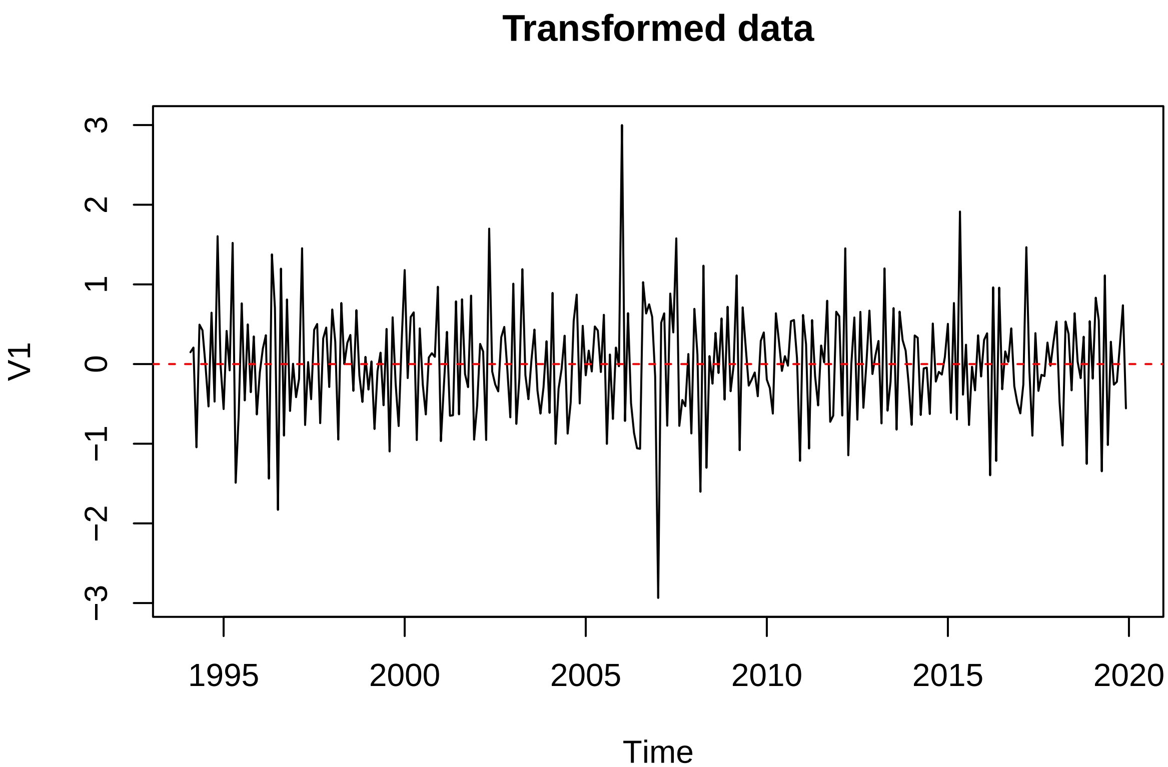
```
## Variance of original data: 1.863327
```

```
## Variance after seasonal differencing: 0.6081206
```

```
## Variance after one regular differencing: 0.4841846
```

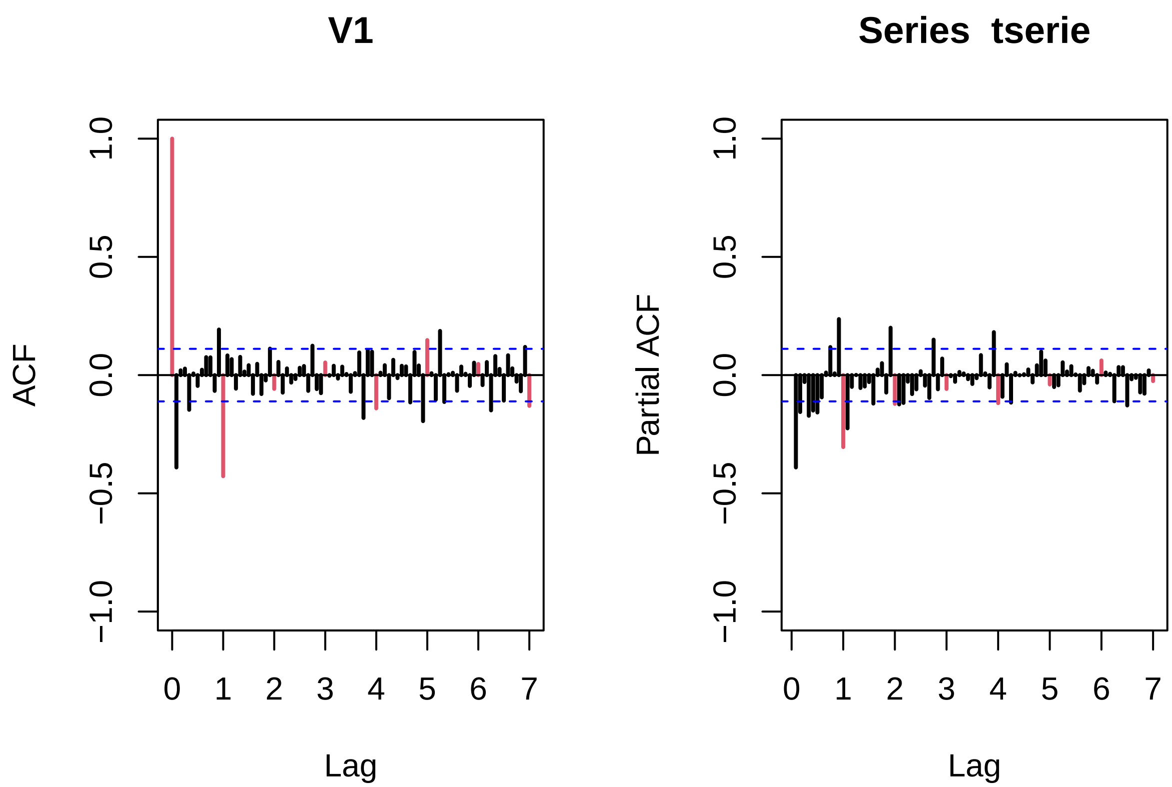
```
## Variance after two regular differencing: 1.349775
```

The total variance is minimal for one regular and one 12 month seasonal difference. This means that we should take one regular difference and no more. As we can see below, the data after transformations resembles white noise, which is the aim of data transformations in time-series data analysis. Now that we have stationary time series, we can propose ARIMA models by evaluating - among other metrics - the ACF and PACF plots.



1b.

Now we are going to take a closer look at the ACF and the PACF plots.



To propose SARIMA models we need to evaluate ACF and PACF plots separately. SARIMA model can be expressed as:

$$SARIMA(p, d, q)(P, D, Q)_s$$

Where:

- p, d, q: Non-seasonal AR, differencing, and MA terms, respectively.
- P, D, Q: Seasonal AR, differencing, and MA terms, respectively.
- s: Periodicity of the seasonal component.

Looking at the ACF plot we can clearly see decaying trend of non-seasonal lags after the lag 1, which is the biggest lag for the whole non-seasonal part and the subsequent lag experienced a sharp decline in ACF value hinting at possible MA(1) model (parameter q = 1). The same things can be said about seasonal lags in the ACF plot. There is a decaying trend of seasonal lags after the seasonal lag 1, which is the biggest seasonal lag for the whole seasonal part and the subsequent seasonal lag experienced a sharp decline in ACF value hinting at possible MA(1) model (parameter Q = 1). Now we move on to the PACF plot, which looks more complex than usual. The clearer trend is seen in seasonal lags of the PACF. The first seasonal lag is the biggest seasonal lag for the whole seasonal part and the subsequent seasonal lag experienced a sharp decline in PACF value hinting at possible AR(1) model (parameter P = 1). Here this trend is clearer than in the case of seasonal lags in the ACF because after seasonal lag 4 values become small so that exceeding confidence intervals becomes nearly impossible. In the ACF plot we have seen some seasonal lags being slightly above confidence intervals even though their antecedent lags were way below confidence intervals. The proposal of non-seasonal AR models by solely looking at the PACF plot in this particular case seems non-trivial, hence it would be a good idea to propose several hypothetical non-seasonal AR models and later check them using several model validation techniques. The decaying trend of non-seasonal AR part is there, though it is slower and seems to start later comparing it to non-seasonal MA part. Therefore, one could argue that one possible non-seasonal AR model which is AR(2) (parameter $p = 2$) because lag 2 is followed a lag that has experienced a sharp decline in PACF value. However, lags 4,5, and 6 are definitely above confidence intervals. Here is where we can propose another non-seasonal AR mode which is AR(6) (parameter $p = 6$), because lag 6 is followed by lags that are below confidence intervals consequentially and the decline in their values is sharper. Regarding the values of parameters d and D, we don't need to propose them because we know the actual values which are $d = 1$ and $D = 1$. In the end, combining everything we have seen we can propose 9 SARIMA models in the order of increasing complexity, which are:

SARIMA Models

	Model	Description
1	SARIMA(0, 1, 1)(1, 1, 0)_{12}	Non-seasonal MA(1), Seasonal AR(1)
2	SARIMA(0, 1, 1)(0, 1, 1)_{12}	Non-seasonal MA(1), Seasonal MA(1)
3	SARIMA(0, 1, 1)(1, 1, 1)_{12}	Non-seasonal MA(1), Seasonal AR(1), Seasonal MA(1)
4	SARIMA(2, 1, 1)(1, 1, 0)_{12}	Non-seasonal MA(1), Non-seasonal AR(2), Seasonal AR(1)
5	SARIMA(2, 1, 1)(0, 1, 1)_{12}	Non-seasonal MA(1), Non-seasonal AR(2), Seasonal MA(1)
6	SARIMA(2, 1, 1)(1, 1, 1)_{12}	Non-seasonal MA(1), Non-seasonal AR(2), Seasonal AR(1), Seasonal MA(1)
7	SARIMA(6, 1, 1)(1, 1, 0)_{12}	Non-seasonal MA(1), Non-seasonal AR(6), Seasonal AR(1)
8	SARIMA(6, 1, 1)(0, 1, 1)_{12}	Non-seasonal MA(1), Non-seasonal AR(6), Seasonal MA(1)
9	SARIMA(6, 1, 1)(1, 1, 1)_{12}	Non-seasonal MA(1), Non-seasonal AR(6), Seasonal AR(1), Seasonal MA(1)

Below is the same data given in a more tabular manner.

```
##    p i q P D Q      AIC      BIC
## 10 6 1 1 0 1 1 470.7858 504.4439
## 6  2 1 1 0 1 1 471.1388 489.8378
## 12 6 1 1 1 1 1 472.6071 510.0050
## 8  2 1 1 1 1 1 473.0751 495.5138
## 2  0 1 1 0 1 1 473.8614 485.0808
## 4  0 1 1 1 1 1 475.6325 490.5917
## 7  2 1 1 1 1 0 517.4252 536.1242
## 11 6 1 1 1 1 0 517.7831 551.4413
## 3  0 1 1 1 1 0 521.1615 532.3809
## 5  2 1 1 0 1 0 580.9848 595.9440
```

```
## 9   6 1 1 0 1 0 587.1439 617.0623
## 1   0 1 1 0 1 0 594.9498 602.4294
```

From the output above we can state that our theory about possible non-seasonal AR(6) model might be correct. Some of the SARIMA models that have non-seasonal AR(6) part do very well in terms of AIC values, in fact the model SARIMA(6,1,1)(0,1,1)12 has the lowest AIC value, however BIC penalizes complexity of the models differently than AIC does and in terms of BIC values these models are average at best. One might think that we should easily choose models that would have AIC values of 2-3 points higher but with BIC values of 10 points lower at worst that the earlier mentioned SARIMA models but AIC/BIC values is not the only indicator of selecting the most optimal time-series model. After experimenting with the proposed `p`, `P`, `Q` parameters for SARIMA models, we have decided to further validate 3 better than average models. These 3 models, `mod1`, `mod2`, and `mod3`, can be seen in the outputs above. Their value of `p` is different (6,2,0) but all other parameters are identical. We were thinking about inclusion of models that have seasonal AR part with the value of 1 (P = 1) but all such models always do worse in terms of AIC/BIC compared to exactly same models but with seasonal AR part with the value of 0 (P = 0). The first model `mod1` is the most complex in terms of parameters (the highest p parameter), it has the lowest AIC, and the highest BIC. The second model `mod2` is the 2nd ranked in terms of complexity of parameters (the second highest p parameter) and it has the second highest AIC and BIC. The third model `mod3` is the least complex in terms of parameters (the lowest p parameter), it has the highest AIC, and the lowest BIC.

2. Estimation

Based on the previous analyisi, we decide to investigate three model and below are the coeficients and other information about the models.

2a.

```
##
## Call:
## arima(x = serie, order = c(6, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5      ar6      ma1     sma1
##       -0.0507  -0.0756  -0.0336  -0.1479  -0.0993  -0.1476  -0.4573  -0.7066
## s.e.   0.1745   0.0976   0.0706   0.0610   0.0646   0.0648   0.1705   0.0403
##
## sigma^2 estimated as 0.2441:  log likelihood = -226.39,  aic = 470.79
```

```
##
## Call:
## arima(x = serie, order = c(2, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##          ar1     ar2      ma1     sma1
##       0.3098  0.1070  -0.8115  -0.7035
## s.e.  0.0912  0.0738   0.0696   0.0400
##
## sigma^2 estimated as 0.2508:  log likelihood = -230.57,  aic = 471.14
```

```
##
## Call:
```

```
## arima(x = serie, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ma1     sma1
##       -0.5317  -0.7023
## s.e.   0.0599   0.0388
##
## sigma^2 estimated as 0.2565:  log likelihood = -233.93,  aic = 473.86
```

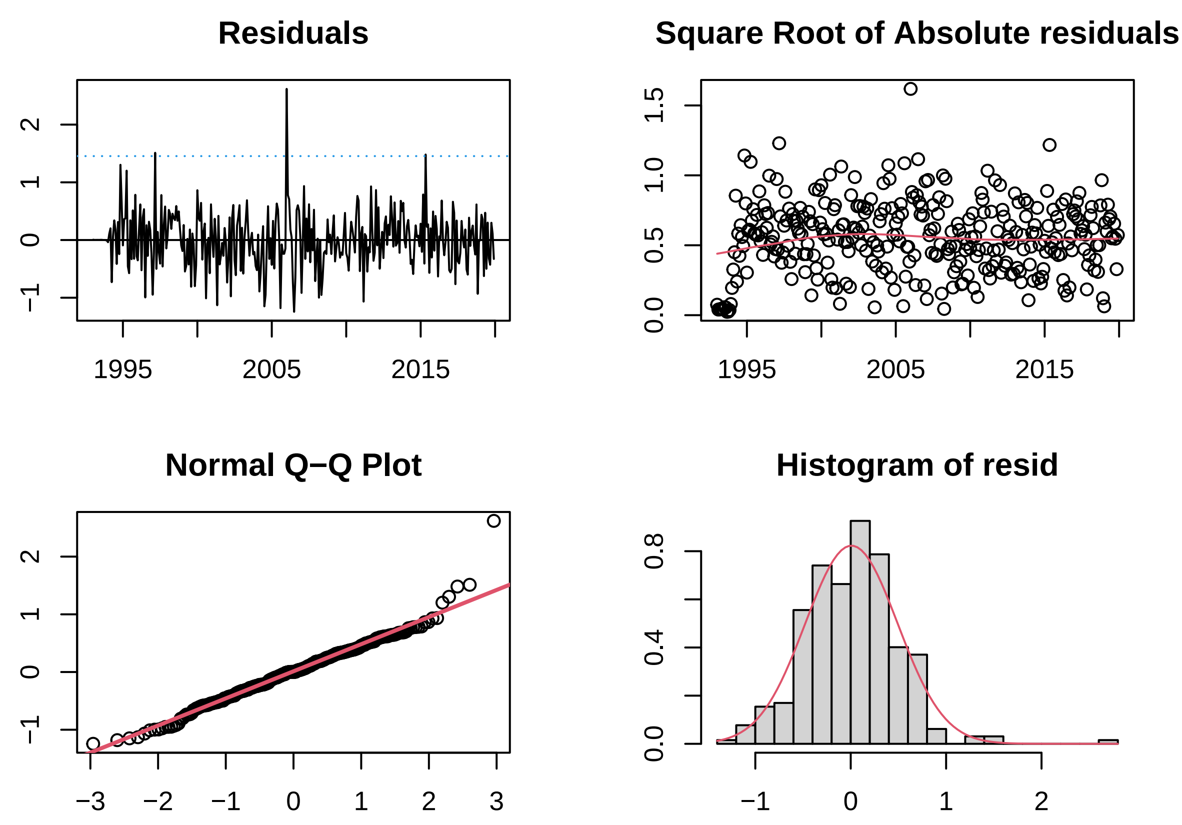
3. Validation

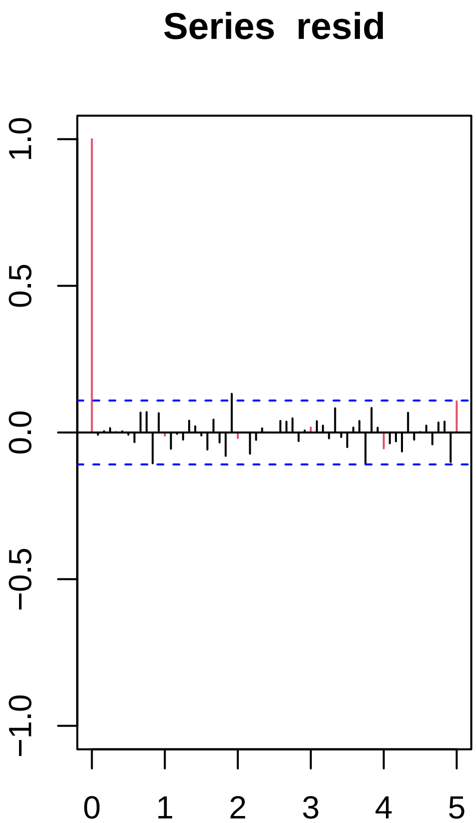
Now we will validate the proposed models, based on a thorough statistical anaylysis.

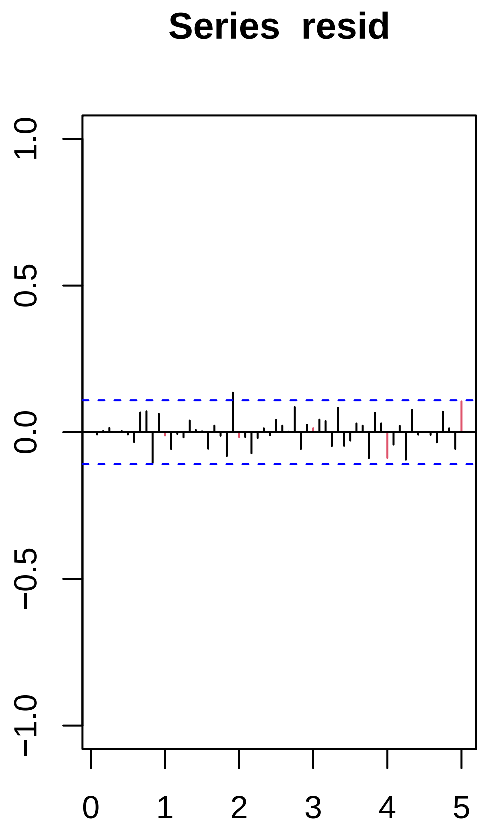
3a and 3b.

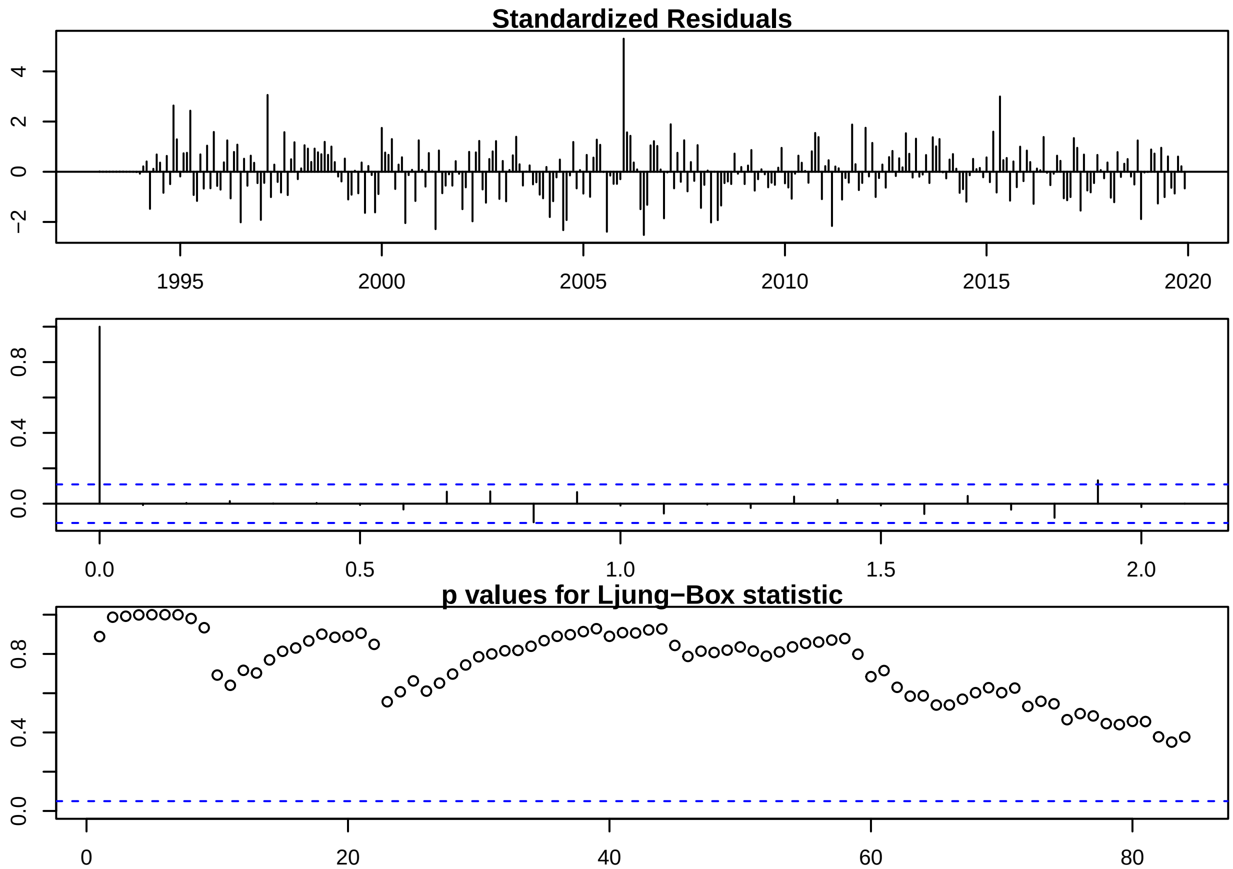
We will be taking a look at the outputs of the validation function for models `mod1`, `mod2`, and `mod3`. This function provides all the necessary things for the evaluation of a model which are complete analysis of residuals and roots of polynomials/pi and psi weights to check the causality and/or invertibility.

```
validation(mod1)
```

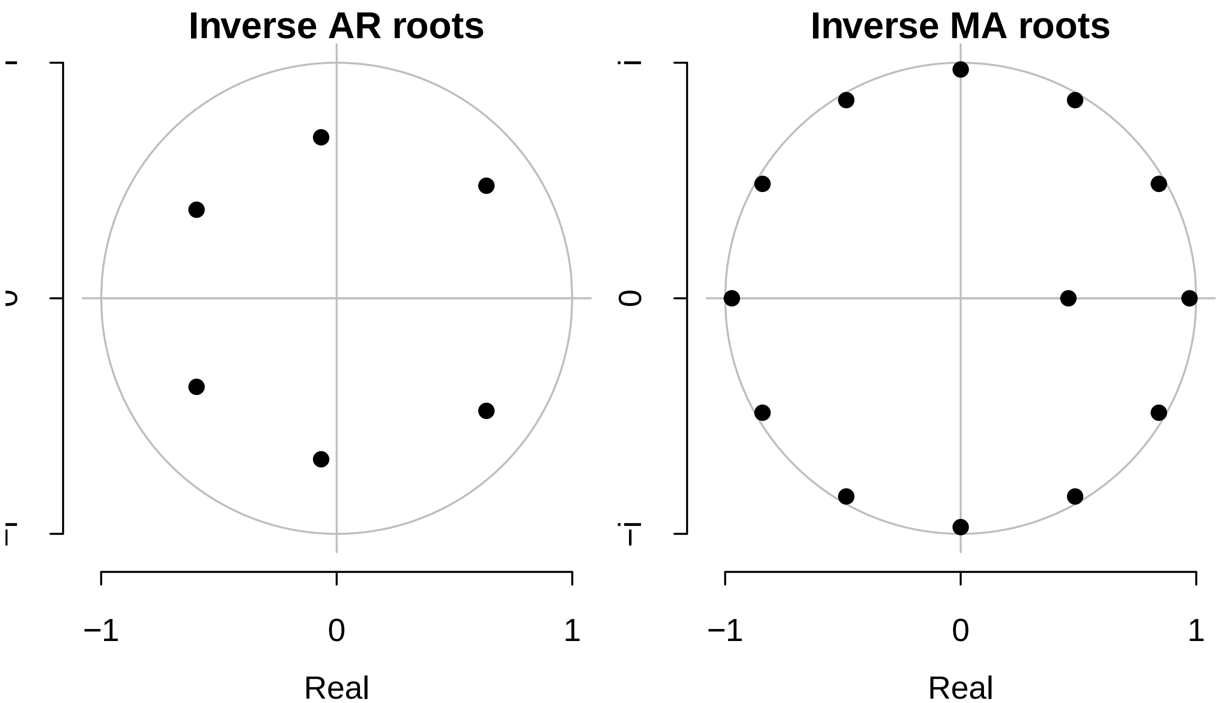



Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5




```
##
## ----------------------------------------------------------------
##
## Call:
## arima(x = serie, order = c(6, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5      ar6      ma1     sma1
##       -0.0507  -0.0756  -0.0336  -0.1479  -0.0993  -0.1476  -0.4573  -0.7066
## s.e.   0.1745   0.0976   0.0706   0.0610   0.0646   0.0648   0.1705   0.0403
##
## sigma^2 estimated as 0.2441:  log likelihood = -226.39,  aic = 470.79
##
## Modul of AR Characteristic polynomial Roots:  1.257526 1.421008 1.421008 1.257526 1.456458 1.456458
##
## Modul of MA Characteristic polynomial Roots:  1.029362 1.029362 1.029362 1.029362 1.029362 1.029362
```




```
##
## Psi-weights (MA(inf))
##
## --------------------
##        psi 1        psi 2        psi 3        psi 4        psi 5        psi 6
## -0.507935161 -0.049894722  0.007391468 -0.127446960 -0.016588240 -0.079591936
##        psi 7        psi 8        psi 9       psi 10       psi 11       psi 12
##  0.088411903  0.027576438  0.008601176  0.026744576 -0.005655552 -0.709748524
##       psi 13       psi 14       psi 15       psi 16       psi 17       psi 18
##  0.341543005  0.027683151 -0.006508706  0.088353367  0.015874260  0.059509591
##       psi 19       psi 20       psi 21       psi 22       psi 23       psi 24
## -0.059389179 -0.018531301 -0.006724334 -0.019685384  0.002659380  0.001431085
##
## Pi-weights (AR(inf))
##
## --------------------
##         pi 1         pi 2         pi 3         pi 4         pi 5         pi 6
## -0.507935161 -0.307892850 -0.174341420 -0.227608939 -0.203356551 -0.240617482
##         pi 7         pi 8         pi 9        pi 10        pi 11        pi 12
## -0.110024458 -0.050309651 -0.023004530 -0.010519023 -0.004809916 -0.708817457
##        pi 13        pi 14        pi 15        pi 16        pi 17        pi 18
## -0.359921853 -0.218022513 -0.123403073 -0.160928741 -0.143739381 -0.170044767
##        pi 19        pi 20        pi 21        pi 22        pi 23        pi 24
## -0.077754464 -0.035553912 -0.016257339 -0.007433811 -0.003399175 -0.500863415
##
## Descriptive Statistics for the Residuals
```

```
##
## ----------------------------------------
##                   resid
## nobs        324.000000
## NAs           0.000000
## Minimum      -1.244061
## Maximum       2.619121
## 1. Quartile  -0.304051
## 3. Quartile   0.329753
## Mean          0.007704
## Median        0.002521
## Sum           2.496165
## SE Mean       0.026928
## LCL Mean     -0.045272
## UCL Mean      0.060680
## Variance      0.234933
## Stdev         0.484699
## Skewness      0.422569
## Kurtosis      2.313684
##
## Normality Tests
##
## --------------------
##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.97527, p-value = 2.291e-05
##
##
##  Anderson-Darling normality test
##
## data:  resid
## A = 0.63104, p-value = 0.09927
##
##
##  Jarque Bera Test
##
## data:  resid
## X-squared = 84.073, df = 2, p-value < 2.2e-16
##
##
## Homoscedasticity Test
##
## --------------------
##
##  studentized Breusch-Pagan test
##
## data:  resid ~ I(obs - resid)
## BP = 0.00032599, df = 1, p-value = 0.9856
##
##
## Independence Tests
##
```

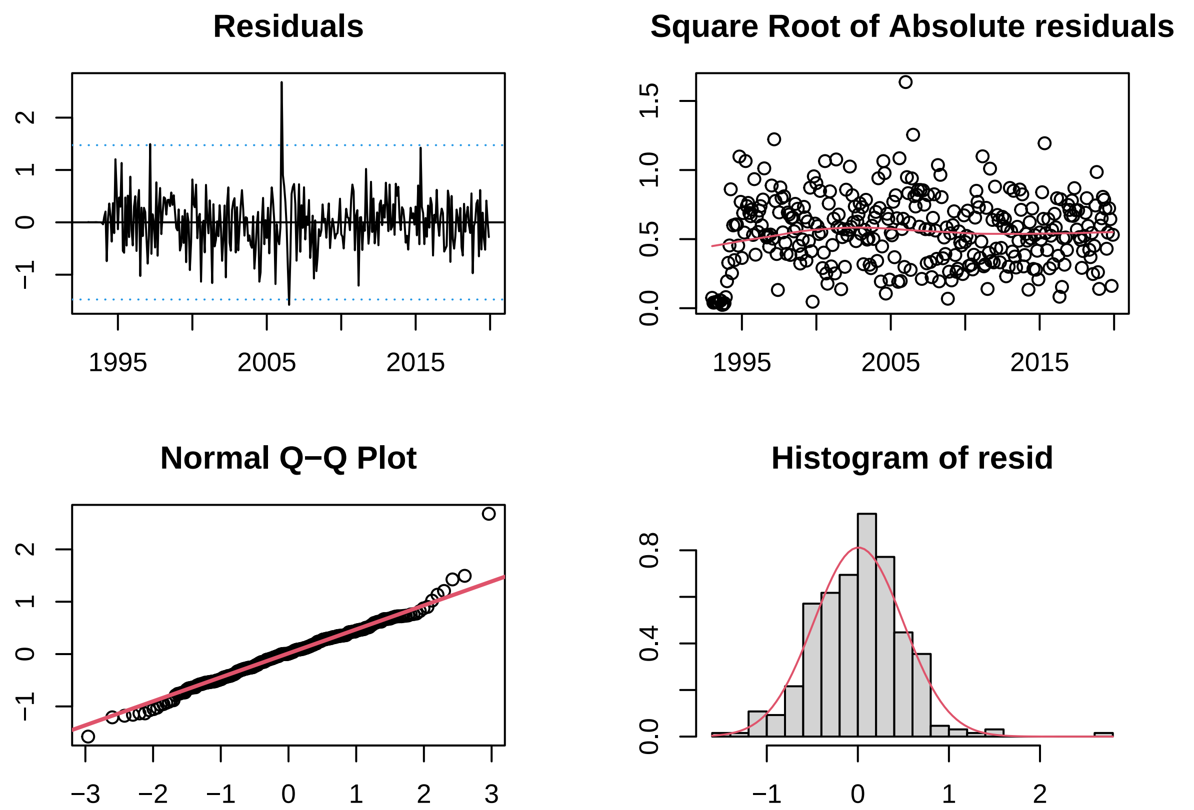
```
## --------------------
##
##  Durbin-Watson test
##
## data:  resid ~ I(1:length(resid))
## DW = 2.0149, p-value = 0.5313
## alternative hypothesis: true autocorrelation is greater than 0
##
##
## Ljung-Box test
##      lag.df   statistic   p.value
## [1,]      1  0.01964655 0.8885287
## [2,]      2  0.02575258 0.9872063
## [3,]      3  0.09764682 0.9921183
## [4,]      4  0.09821217 0.9988331
## [5,]     12  8.83949074 0.7165755
## [6,]     24 21.53189366 0.6071842
## [7,]     36 26.02163767 0.8898693
## [8,]     48 39.39546377 0.8072403
```

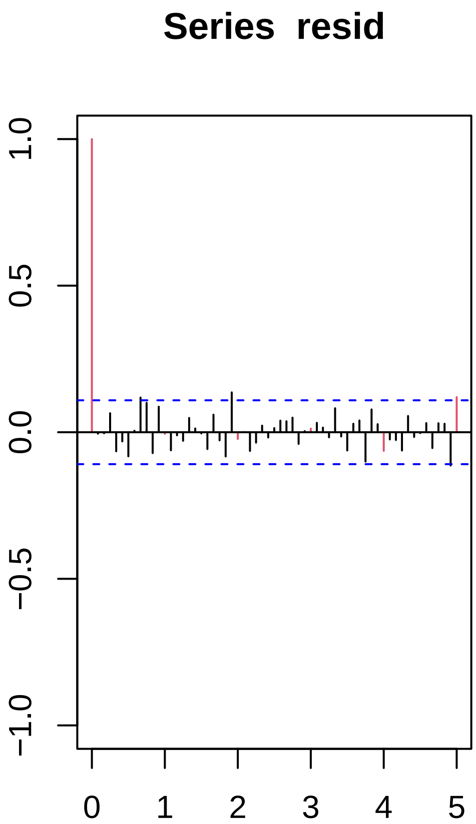
For the model SARIMA(6,1,1)(0,1,1)12

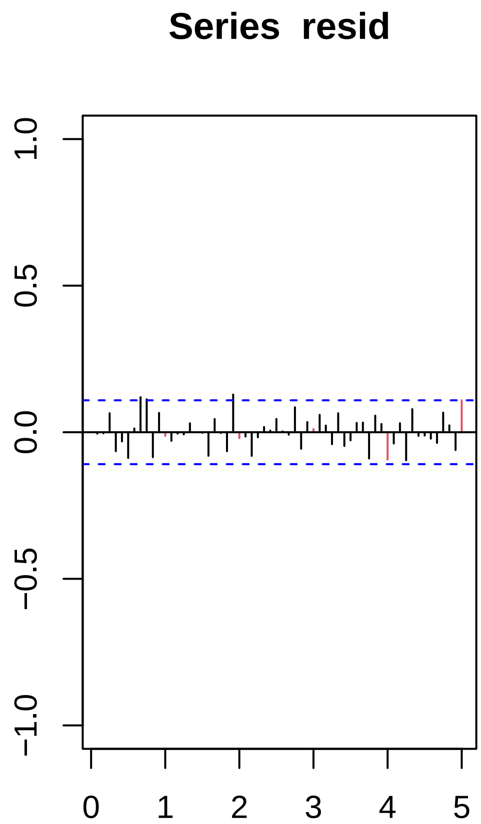
- Residuals plot
  - The residuals are centered around zero with no evident trend
  - Occasional spikes above the confidence intervals that might be explain by potential outliers
- Square Root of Absolute Residuals
  - A reasonable consistent spread, which implies constant variance over time and homoscedasticity
- Normal Q-Q Plot
  - Small deviations from the diagonal line are seen at the ends of the line, but since the trend holds and deviations clearly can't be called heavy tails they might be caused by potential outliers
- Histogram of residuals
  - Some asymmetry is visible
  - A couple of observations above the bell-shaped curve are seen at the ends of it hinting at potential outliers
- ACF and PACF of Residuals
  - There are barely any lags that are above confidence intervals and if they are they are higher by negligible value indicating there are no significant autocorrelation/partial autocorrelations in residuals
- Standardized Residuals
  - The standardized residuals fluctuate around zero without evident patterns or trends, though some standardized residuals definitely look like spikes hinting at potential outliers
- p-values for Ljung-Box Test
  - None of the p-values for the Ljung-Box test are below 0.05, indicating no significant autocorrelation in the residuals.

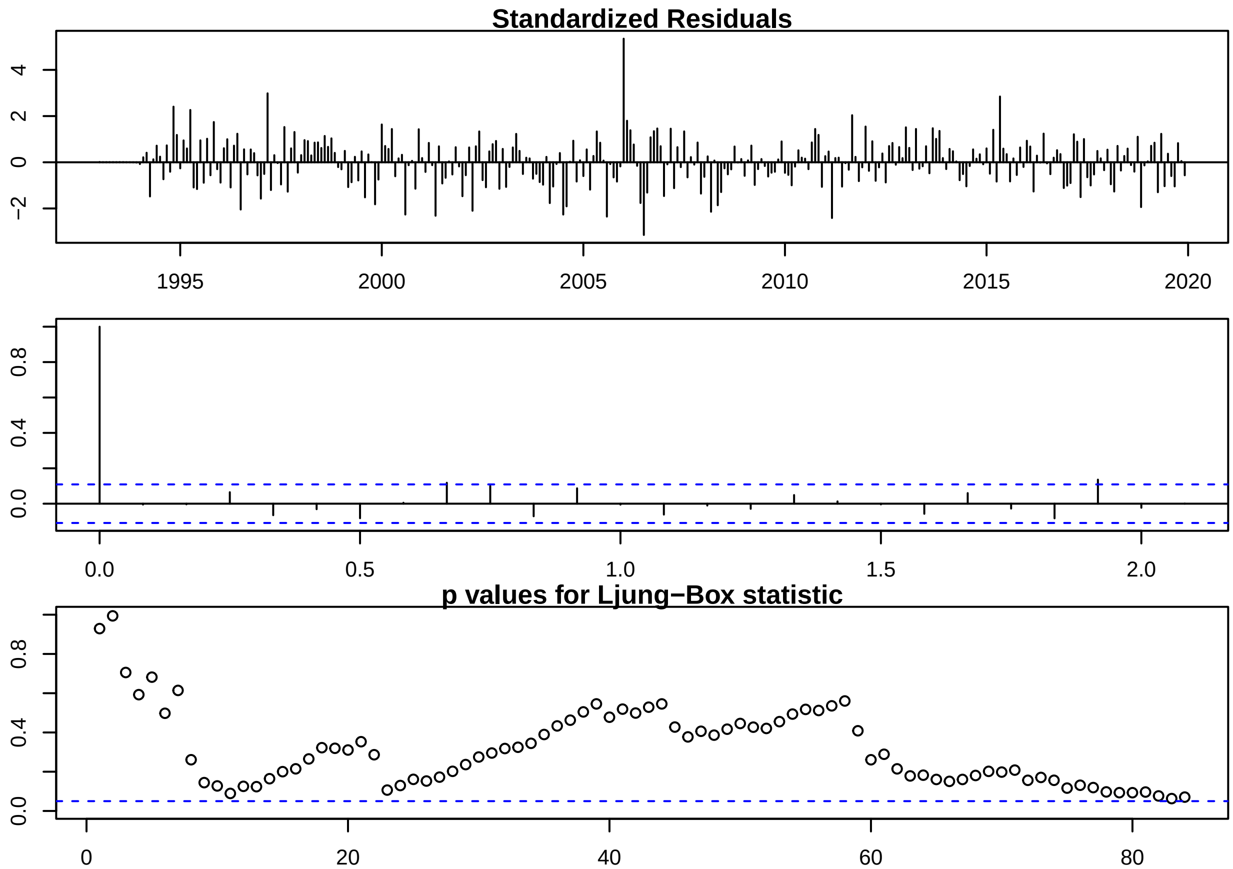
Based on the SARIMA(6,1,1)(0,1,1)12 model modules of AR/MA characteristic polynomial roots we can say that the model is both causal and invertible as all modules of polynomial roots are higher than 1.

```r
validation(mod2)
```

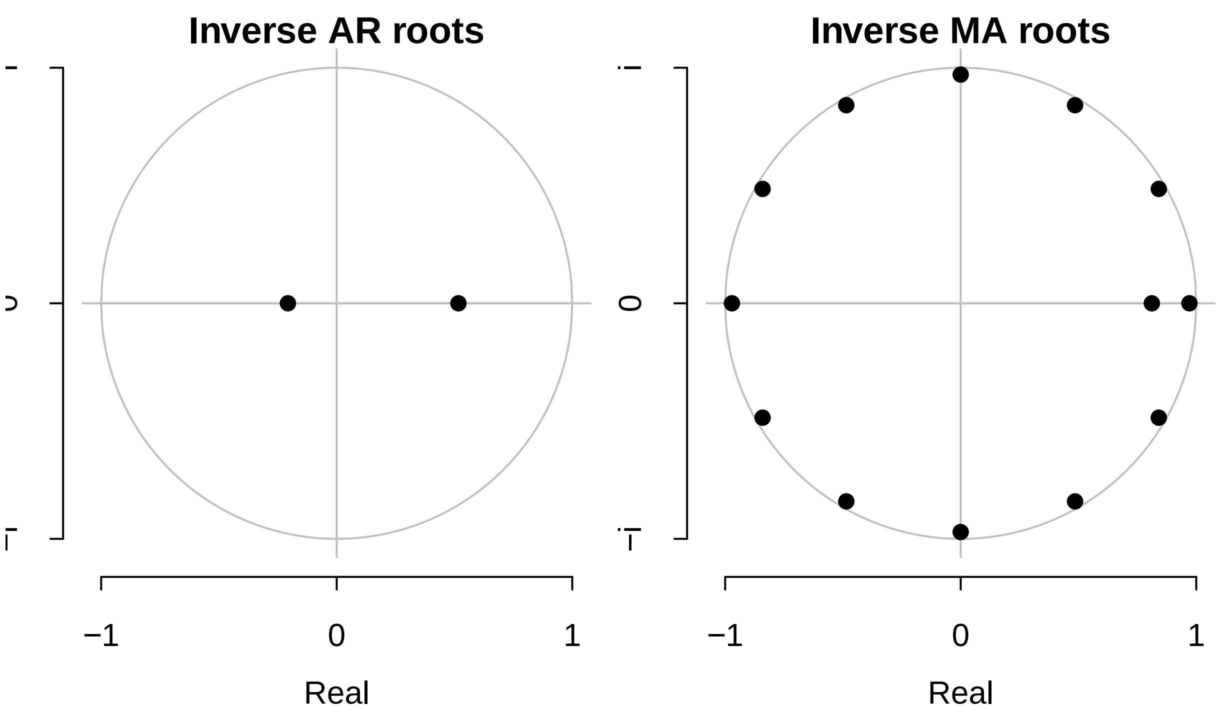



Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5




```
##
## -------------------------------------------------------------------
##
## Call:
## arima(x = serie, order = c(2, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##          ar1     ar2      ma1     sma1
##       0.3098  0.1070  -0.8115  -0.7035
## s.e.  0.0912  0.0738   0.0696   0.0400
##
## sigma^2 estimated as 0.2508:  log likelihood = -230.57,  aic = 471.14
##
## Modul of AR Characteristic polynomial Roots:  1.935002 4.830788
##
## Modul of MA Characteristic polynomial Roots:  1.029741 1.029741 1.029741 1.029741 1.029741 1.029741
```




```
##
## Psi-weights (MA(inf))
##
## --------------------
##         psi 1         psi 2         psi 3         psi 4         psi 5
## -0.5017107239 -0.0484453148 -0.0686806228 -0.0264592054 -0.0155442082
##         psi 6         psi 7         psi 8         psi 9        psi 10
## -0.0076460282 -0.0040315725 -0.0020669079 -0.0010716024 -0.0005530882
##        psi 11        psi 12        psi 13        psi 14        psi 15
## -0.0002859805 -0.7036442912  0.3528753830  0.0340416447  0.0482961835
##        psi 16        psi 17        psi 18        psi 19        psi 20
##  0.0186034185  0.0109298491  0.0053761392  0.0028347424  0.0014533107
##        psi 21        psi 22        psi 23        psi 24
##  0.0007534800  0.0003888948  0.0002010825  0.0001038971
##
## Pi-weights (AR(inf))
##
## --------------------
##        pi 1        pi 2        pi 3        pi 4        pi 5        pi 6
## -0.50171072 -0.30015897 -0.24357913 -0.19766457 -0.16040488 -0.13016863
##        pi 7        pi 8        pi 9       pi 10       pi 11       pi 12
## -0.10563190 -0.08572033 -0.06956208 -0.05644966 -0.04580892 -0.74067049
##       pi 13       pi 14       pi 15       pi 16       pi 17       pi 18
## -0.38311844 -0.23564107 -0.19122283 -0.15517741 -0.12592653 -0.10218943
##       pi 19       pi 20       pi 21       pi 22       pi 23       pi 24
## -0.08292677 -0.06729511 -0.05461001 -0.04431605 -0.03596249 -0.52409094
```

```
##
## Descriptive Statistics for the Residuals
##
## ----------------------------------------
##                   resid
## nobs        324.000000
## NAs           0.000000
## Minimum      -1.576139
## Maximum       2.679299
## 1. Quartile  -0.294924
## 3. Quartile   0.322825
## Mean          0.007232
## Median        0.008386
## Sum           2.343325
## SE Mean       0.027295
## LCL Mean     -0.046466
## UCL Mean      0.060931
## Variance      0.241388
## Stdev         0.491313
## Skewness      0.315620
## Kurtosis      2.492238
##
## Normality Tests
##
## --------------------
##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.97518, p-value = 2.203e-05
##
##
##  Anderson-Darling normality test
##
## data:  resid
## A = 0.77976, p-value = 0.04255
##
##
##  Jarque Bera Test
##
## data:  resid
## X-squared = 91.589, df = 2, p-value < 2.2e-16
##
##
## Homoscedasticity Test
##
## --------------------
##
##  studentized Breusch-Pagan test
##
## data:  resid ~ I(obs - resid)
## BP = 0.14472, df = 1, p-value = 0.7036
##
##
```

```
## Independence Tests
##
## --------------------
##
##  Durbin-Watson test
##
## data:  resid ~ I(1:length(resid))
## DW = 2.0095, p-value = 0.5119
## alternative hypothesis: true autocorrelation is greater than 0
##
##
## Ljung-Box test
##      lag.df   statistic   p.value
## [1,]      1  0.00791668 0.9291012
## [2,]      2  0.01210528 0.9939656
## [3,]      3  1.40025273 0.7054755
## [4,]      4  2.79566196 0.5925818
## [5,]     12 17.69023554 0.1254248
## [6,]     24 31.89592027 0.1295947
## [7,]     36 36.76470465 0.4332824
## [8,]     48 50.20031864 0.3862602
```

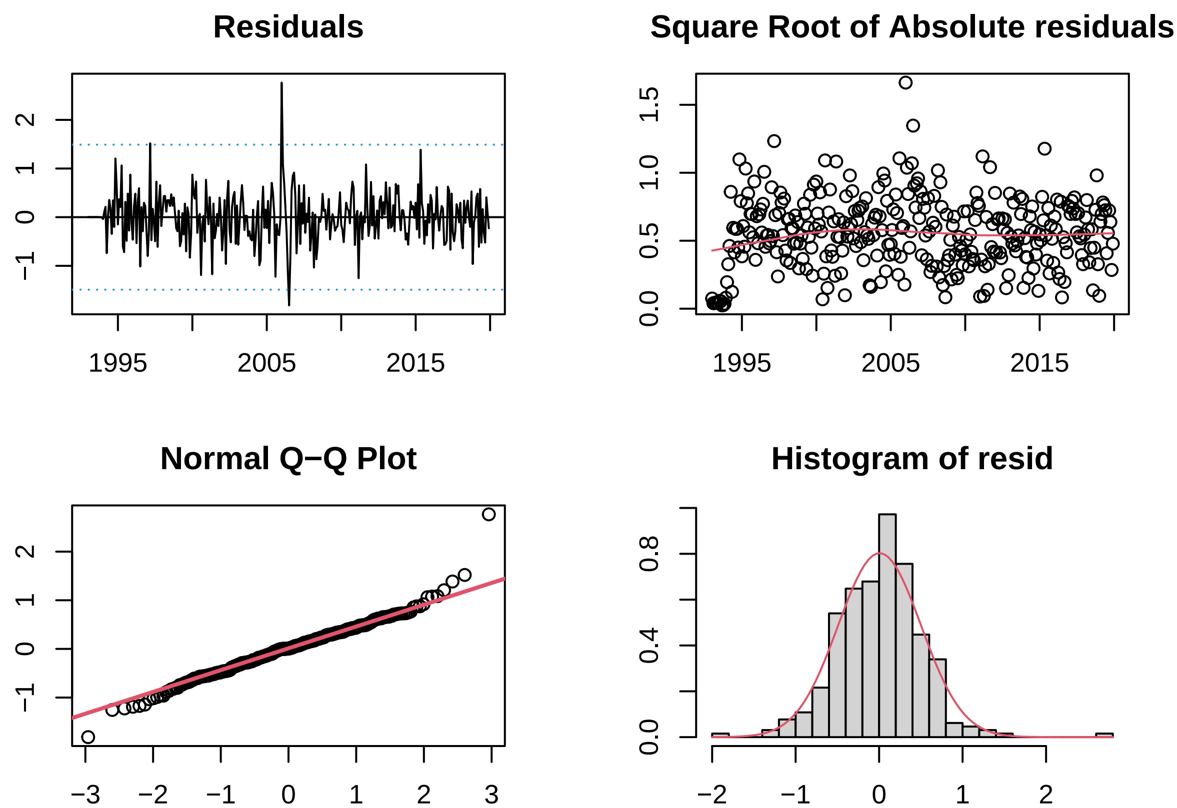
For the model SARIMA(2,1,1)(0,1,1)12

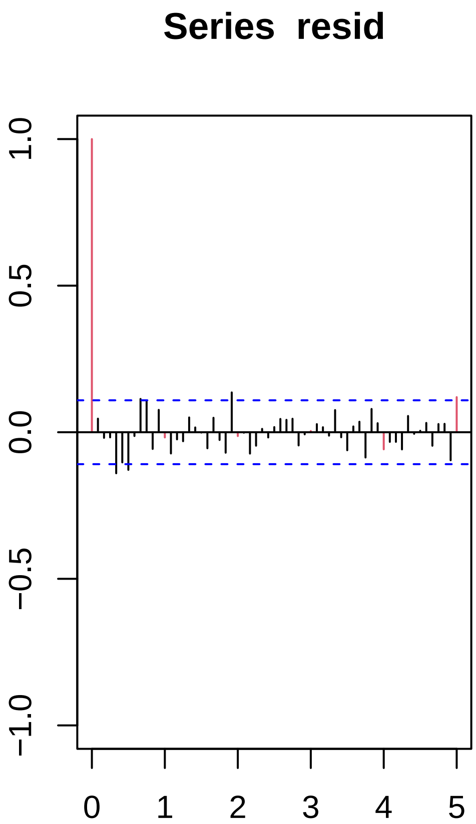
- Residuals plot
  - The residuals are centered around zero with no evident trend
  - Occasional spikes above the confidence intervals that might be explain by potential outliers
- Square Root of Absolute Residuals
  - A reasonable consistent spread, which implies constant variance over time and homoscedasticity
- Normal Q-Q Plot
  - Small deviations from the diagonal line are seen at the ends of the line, but since the trend holds and deviations clearly can't be called heavy tails they might be caused by potential outliers
  - Points at both ends of the lined distributed worse compared to SARIMA(6,1,1)(0,1,1)12 model
  - Even though they are most likely called by potential outliers it means that skewness and kurtosis values might be different for this model compared to SARIMA(6,1,1)(0,1,1)12 model
- Histogram of residuals
  - Some asymmetry is visible
  - A couple of observations above the bell-shaped curve are seen at the ends of it hinting at potential outliers
- ACF and PACF of Residuals
  - There are barely any lags that are above confidence intervals and if they are they are higher by negligible value indicating there are no significant autocorrelation/partial autocorrelations in residuals
- Standardized Residuals
  - The standardized residuals fluctuate around zero without evident patterns or trends, though some standardized residuals definitely look like spikes hinting at potential outliers
- p-values for Ljung-Box Test

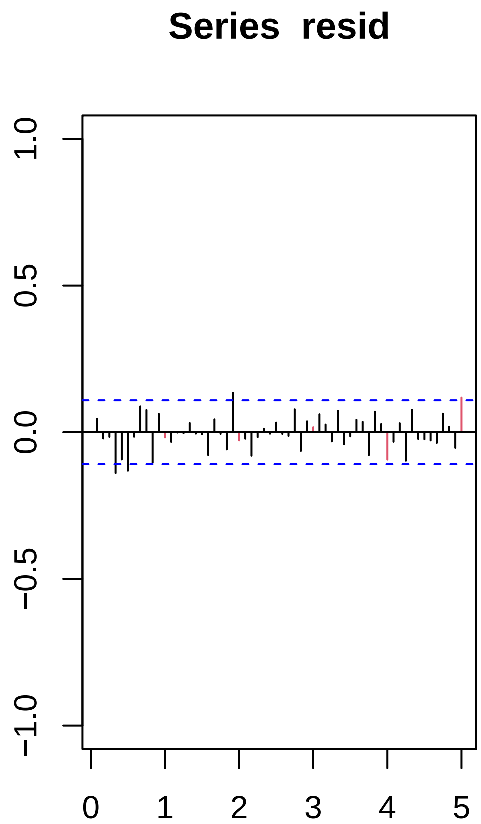
- None of the p-values for the Ljung-Box test are below 0.05 and the lowest p-value is around 0.0525, indicating no significant autocorrelation in the residuals.
- Compared to SARIMA(6,1,1)(0,1,1)12 model all p-values are significantly lower across the board, though they are still above the confidence interval

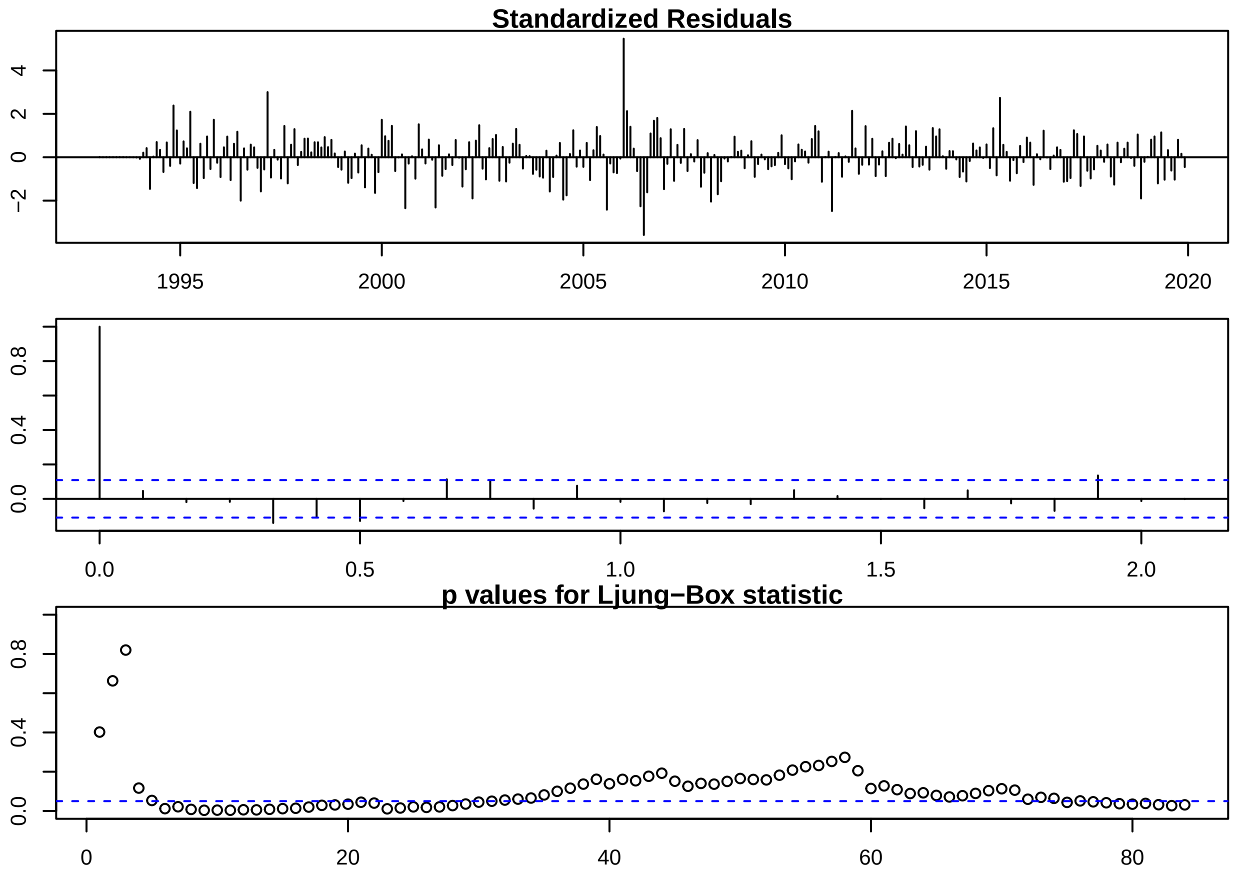
Based on the SARIMA(2,1,1)(0,1,1)12 model modules of AR/MA characteristic polynomial roots we can say that the model is both causal and invertible as all modules of polynomial roots are higher than 1.

```
validation(mod3)
```

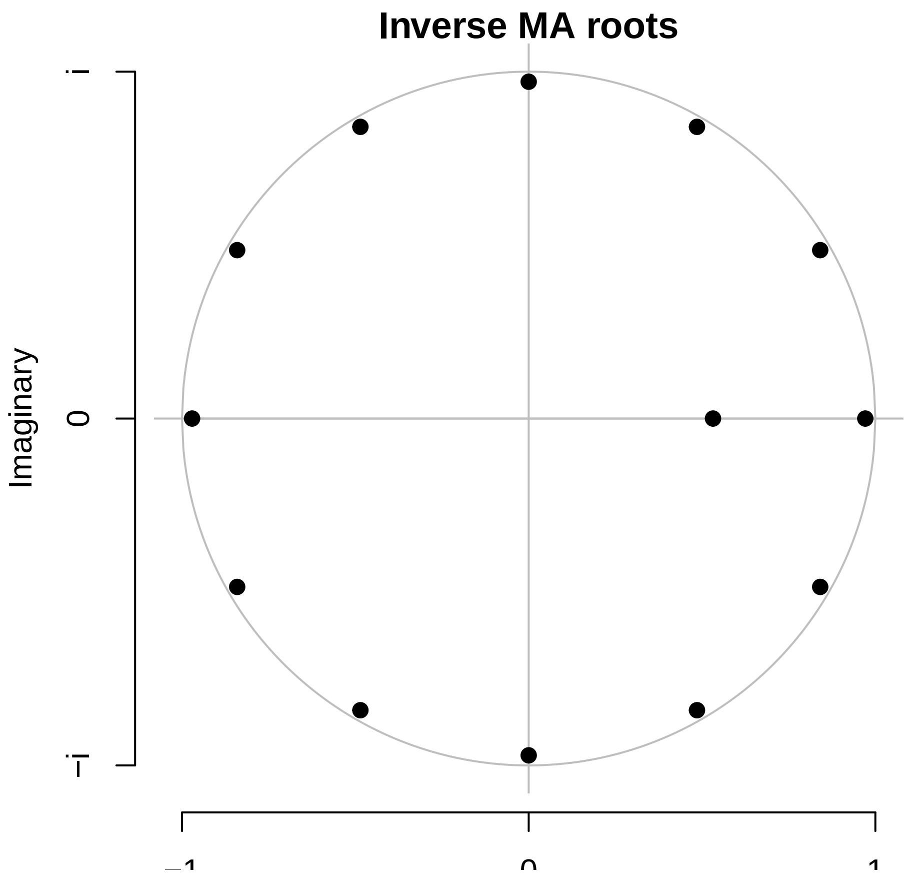



Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5


Series resid
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5




```
##
## --------------------------------------------------------------------
##
## Call:
## arima(x = serie, order = c(0, 1, 1), seasonal = list(order = c(0, 1, 1), period = 12))
##
## Coefficients:
##           ma1     sma1
##       -0.5317  -0.7023
## s.e.   0.0599   0.0388
##
## sigma^2 estimated as 0.2565:  log likelihood = -233.93,  aic = 473.86
##
## Modul of AR Characteristic polynomial Roots:
##
## Modul of MA Characteristic polynomial Roots:  1.029882 1.029882 1.029882 1.029882 1.029882 1.029882
```




```
##
## Psi-weights (MA(inf))
##
## --------------------
##      psi 1      psi 2      psi 3      psi 4      psi 5      psi 6      psi 7
## -0.5316889  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
##      psi 8      psi 9     psi 10     psi 11     psi 12     psi 13     psi 14
##  0.0000000  0.0000000  0.0000000  0.0000000 -0.7023469  0.3734301  0.0000000
##     psi 15     psi 16     psi 17     psi 18     psi 19     psi 20     psi 21
##  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
##     psi 22     psi 23     psi 24
##  0.0000000  0.0000000  0.0000000
##
## Pi-weights (AR(inf))
##
## --------------------
##          pi 1          pi 2          pi 3          pi 4          pi 5
## -0.5316889043 -0.2826930910 -0.1503047798 -0.0799153837 -0.0424901228
##          pi 6          pi 7          pi 8          pi 9         pi 10
## -0.0225915268 -0.0120116642 -0.0063864686 -0.0033956145 -0.0018054105
##         pi 11         pi 12         pi 13         pi 14         pi 15
## -0.0009599167 -0.7028572828 -0.3737014186 -0.1986928978 -0.1056428091
##         pi 16         pi 17         pi 18         pi 19         pi 20
## -0.0561691094 -0.0298644923 -0.0158786192 -0.0084424856 -0.0044887759
##         pi 21         pi 22         pi 23         pi 24
## -0.0023866324 -0.0012689459 -0.0006746845 -0.4936498983
```

```
##
## Descriptive Statistics for the Residuals
##
## ----------------------------------------
##                  resid
## nobs        324.000000
## NAs           0.000000
## Minimum      -1.813913
## Maximum       2.766604
## 1. Quartile  -0.290170
## 3. Quartile   0.314306
## Mean          0.005771
## Median        0.008353
## Sum           1.869924
## SE Mean       0.027605
## LCL Mean     -0.048537
## UCL Mean      0.060080
## Variance      0.246899
## Stdev         0.496889
## Skewness      0.335869
## Kurtosis      2.914760
##
## Normality Tests
##
## --------------------
##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.97334, p-value = 1.054e-05
##
##
##  Anderson-Darling normality test
##
## data:  resid
## A = 0.80512, p-value = 0.03683
##
##
##  Jarque Bera Test
##
## data:  resid
## X-squared = 123.75, df = 2, p-value < 2.2e-16
##
##
## Homoscedasticity Test
##
## --------------------
##
##  studentized Breusch-Pagan test
##
## data:  resid ~ I(obs - resid)
## BP = 0.087161, df = 1, p-value = 0.7678
##
##
```

```
## Independence Tests
##
## --------------------
##
##  Durbin-Watson test
##
## data:  resid ~ I(1:length(resid))
## DW = 1.9072, p-value = 0.1856
## alternative hypothesis: true autocorrelation is greater than 0
##
##
## Ljung-Box test
##      lag.df  statistic    p.value
## [1,]      1  0.7024856 0.401949748
## [2,]      2  0.8220396 0.662973808
## [3,]      3  0.9231266 0.819843553
## [4,]      4  7.3925133 0.116543473
## [5,]     12 27.7554998 0.006006002
## [6,]     24 41.3896224 0.015089534
## [7,]     36 47.2073674 0.100083041
## [8,]     48 58.7986444 0.136583780
```

For the model SARIMA(0,1,1)(0,1,1)12

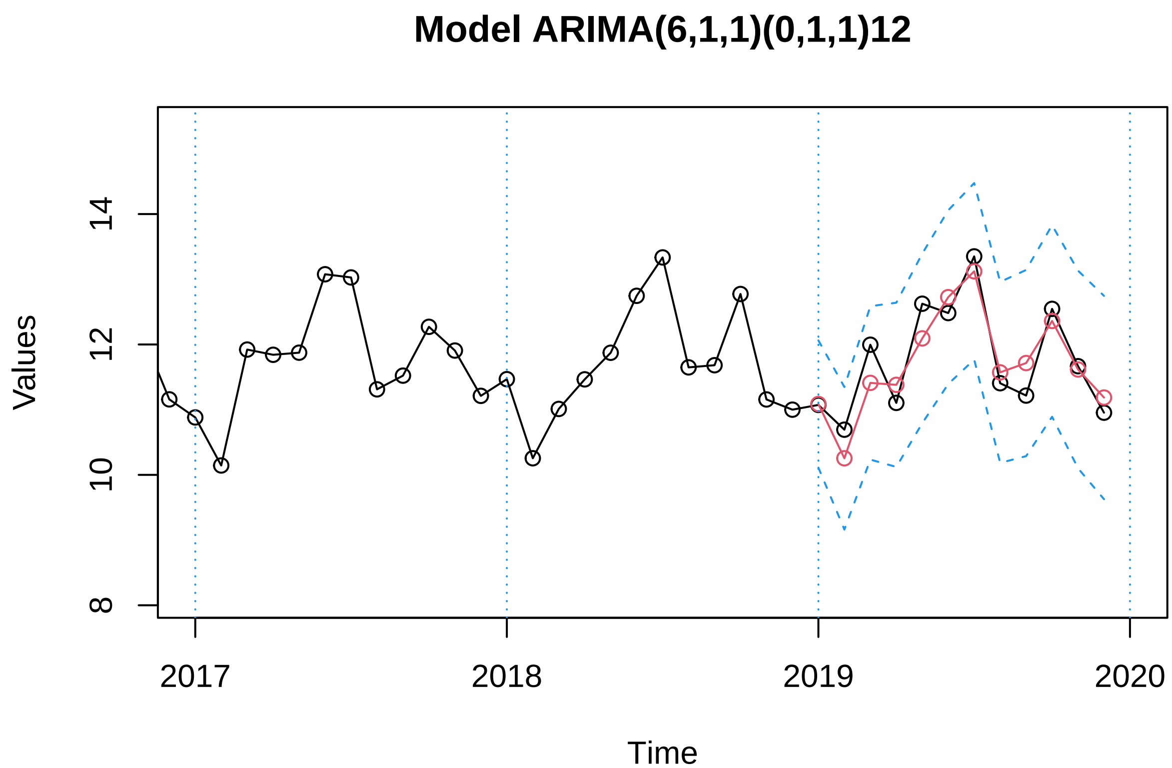
- Residuals plot
  - The residuals are centered around zero with no evident trend
  - Occasional spikes above the confidence intervals that might be explain by potential outliers
- Square Root of Absolute Residuals
  - A reasonable consistent spread, which implies constant variance over time and homoscedasticity
- Normal Q-Q Plot
  - Small deviations from the diagonal line are seen at the ends of the line, but since the trend holds and deviations clearly can't be called heavy tails they might be caused by potential outliers
  - SARIMA(0,1,1)(0,1,1)12 model has the best Normal Q-Q plot
- Histogram of residuals
  - Some asymmetry is visible
  - A couple of observations above the bell-shaped curve are seen at the ends of it hinting at potential outliers
- ACF and PACF of Residuals
  - There are barely any lags that are above confidence intervals and if they are they are higher by negligible value indicating there are no significant autocorrelation/partial autocorrelations in residuals
- Standardized Residuals
  - The standardized residuals fluctuate around zero without evident patterns or trends, though some standardized residuals definitely look like spikes hinting at potential outliers
- p-values for Ljung-Box Test
  - Around 50% of the p-values for the Ljung-Box test are below 0.05 and the lowest p-value is around close to 0 (somewhere around 0.01-0.02)

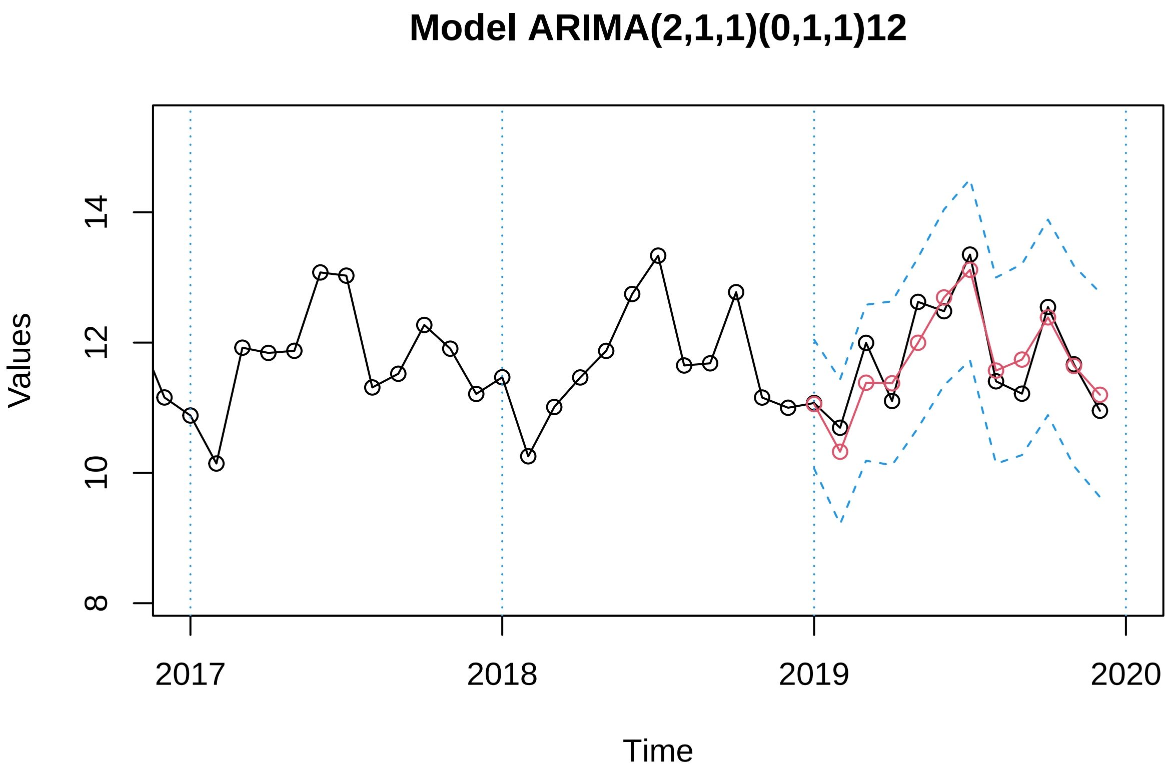
– For the most part p-values struggle to be above 0.05 for higher lags indicating significant serial autocorrelation

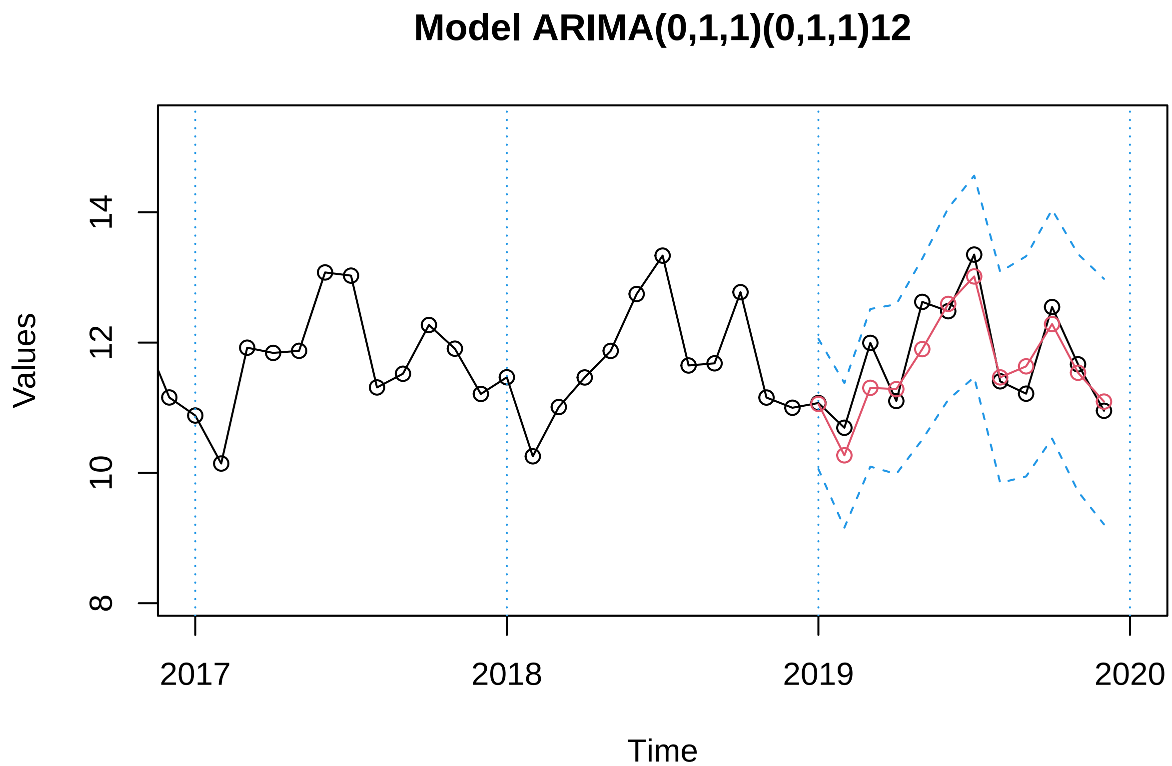
Based on the SARIMA(0,1,1)(0,1,1)12 model modules of AR/MA characteristic polynomial roots we can say that the model is both causal and invertible as all modules of polynomial roots are higher than 1.

3c.

Now we are going to check the stability of the proposed models and evaluate their capability of prediction, reserving the last 12 observations.




Model ARIMA(2,1,1)(0,1,1)12
Values
8
10
12
14
2017
2018
2019
2020
Time




```
##                  Model      RMSE       MAE      RMSPE       MAPE
##  ARIMA(6,1,1)(0,1,1) 0.3382936 0.2877056 0.02888241 0.02451385
##  ARIMA(2,1,1)(0,1,1) 0.3492316 0.2881464 0.02958092 0.02448518
##  ARIMA(0,1,1)(0,1,1) 0.3678022 0.2910628 0.03068842 0.02447672
```

As you can see all of the models manage to predict the last year somewhat good, the differences in error metrics are extremely small but they exist. By comparing all the models, model 1, which is SARIMA(6,1,1)(0,1,1)12 turns out to be the best. Additionally, the stability of all models can be affirmed, as the confidence intervals of the predictions remain well-behaved and do not widen excessively, indicating consistent performance and reliability over time.

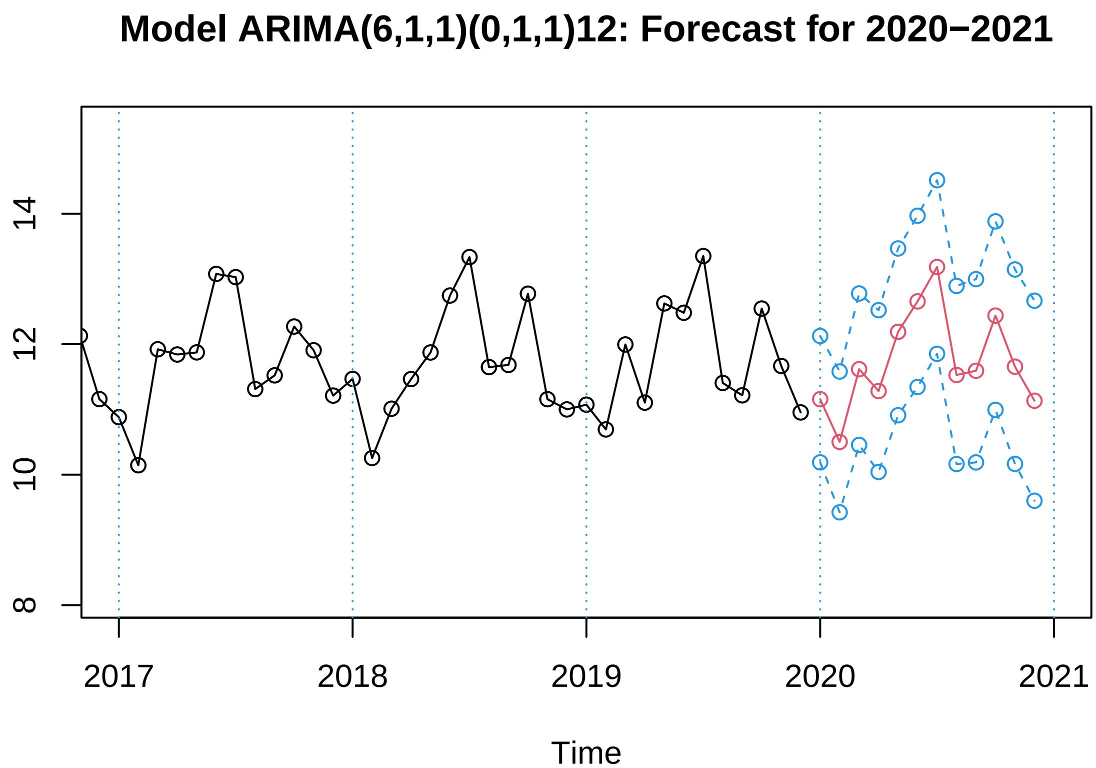
3d.

In order to select the best model for forecasting we have to evaluate validation results, stability, and capability of prediction of all three candidate models. As we have seen earlier, all three candidate models are stable, model 1 shows the best error metrics, which means we only need to look at validation results. All three candidate models are similar in terms of the following things, which are: residuals plot, square Root of absolute residuals, histogram of residuals, ACF and PACF of residuals, and standardized residuals. The only differences between three candidate models can be seen in normal Q-Q plot and p-values for Ljung-Box test. In terms of the “best” normal Q-Q plot out of all candidate models, the model 3 has it. However, in terms of the “best” p-values for Ljung-Box test out of all candidate models, the model 1 has it. Since differences in normal Q-Q plots are negligible, we have to choose the model 3 as the best because its p-values for Ljung-Box test are the best. It is very important to note, that by saying the “best” we refer to the model and data we have, the “best” model is not necessarily the best in the absolute terms.

4. Predictions

4a.

The last thing we are going to do is obtain long term forecasts for the twelve months following the last observation available using the best model for forecasting




```
##  Month Predicted  Lower_CI Upper_CI
##    Jan  11.15695 10.188671 12.12522
##    Feb  10.50197  9.422822 11.58113
##    Mar  11.61658 10.455599 12.77756
##    Apr  11.28143 10.041524 12.52133
##    May  12.18918 10.910645 13.46771
##    Jun  12.65584 11.343528 13.96815
##    Jul  13.18308 11.852664 14.51350
##    Aug  11.52804 10.163249 12.89284
##    Sep  11.59237 10.187998 12.99675
##    Oct  12.43804 10.993232 13.88285
##    Nov  11.65523 10.164962 13.14551
##    Dec  11.13346  9.600362 12.66656
```

By looking at the prediction in the future made by our best model we can affirm the stability of the model, as the confidence intervals of the predictions remain well-behaved and do not widen excessively, indicating consistent performance and reliability over time.