STAT 443: Predicting Lebron James' 3-pointers

Jimin Choi, Cindy Trinh, Aasritha Kosaraju, Elijah Kim

Winter 2024

Contents

Tasks

Jimin Choi: coding for data preparation, checking nonstationarity of data, regression and residual diagnostics, fitting the ARMA models, and testing (APSE calculations).

Cindy Trinh: analysis/write up for Introduction, Dataset Descriptive Analysis, Preparing Data, Checking for Nonstationarity, Model Proposal

Aasritha Kosaraju:analysis/write up for Model Diagnostics, Fit and Prediction Quality, Forecasting and Conclusion

Elijah Kim: coding for differencing, HW smoothing, forecasting; and analysis for residual diagnostics for ARMA models

Introduction

Lebron James is the oldest active NBA player at the age of 39. He has been playing in the NBA for over 21 years, ever since he finished high school. Since most professional basketball athletes retire in their mid thirties, we are interested in seeing if his performance in professional basketball shooting has been declining with age. A 3-pointer is the hardest shot to make in a game and is therefore a good reflection of the shooting skills of a professional basketball athlete. We are then motivated to predict the total number of 3-pointers that Lebron will score in a game.

Dataset Descriptive Analysis

The dataset that we will be using contains 451 rows (excluding the header with variable names) and 29 columns. Each row represents a game, while each column represents a statistic, like number of 3-pointers (3P) or number of field goals (FG). In the interest of modelling total 3-pointers per game by Lebron as a time series, we will only be using the column with statistic "3P" from the dataset, which represents 3-pointers. We will also treat game number as our unit of time, which is the order in which he plays a game.

Preparing Data

We first clean up the data by removing games where Lebron did not play, which are all "NA"'s in the game (G) column. We also create an index to assign each game a number in chronological order. With the dataset ready to use, we divide it into a training set and a test set with a respective split of 80% and 20%. We will use the training set to perform some preliminary analysis of stationarity and subsequently fit potential models. We will use the test set for assessing the forecasting power of our final models.

Stationarity Check

With our training data ready for use, we can perform some tests to inform our choice between stationary and nonstationary models. We first plot the time series of the training data (Figure 1), noting that it looks random. Next, we plot the acf (Figure 2)and pacf (Figure 3) plots. The ACF plot shows no periodicity or trend in the spikes, which suggests that the data is stationary. The data also appears uncorrelated because less than 5% of the spikes occur outside the confidence band. The PACF is not relevant here since the ACF shows that the data is stationary. Before we conclude on the stationarity of the data, we will perform further testing and apply smoothing/differencing to increase the robustness of our claim.

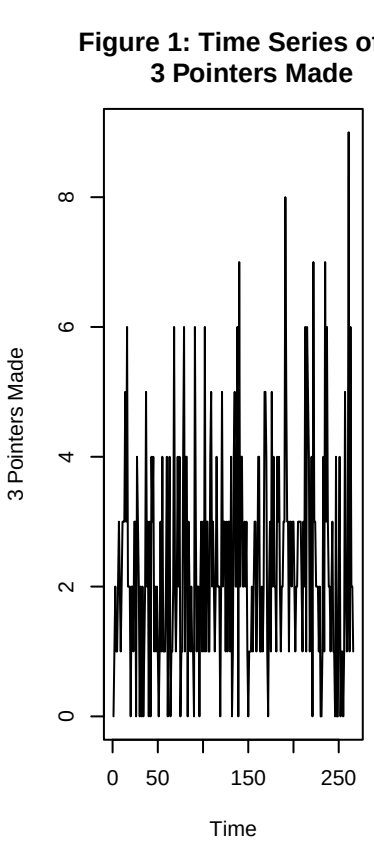


Figure 1: Time Series of 3 Pointers Made

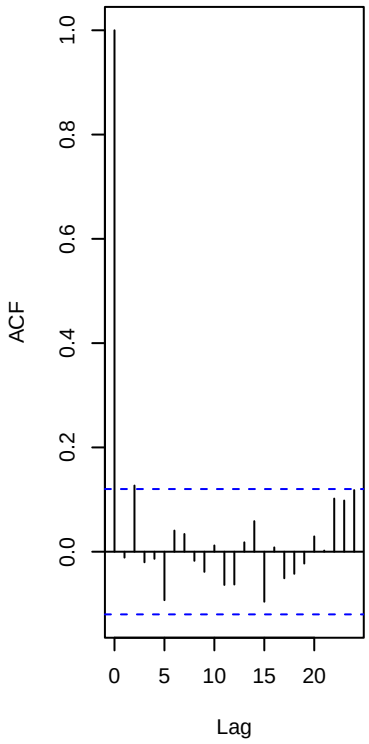


Figure 2: ACF Plot of 3 Pointers Made

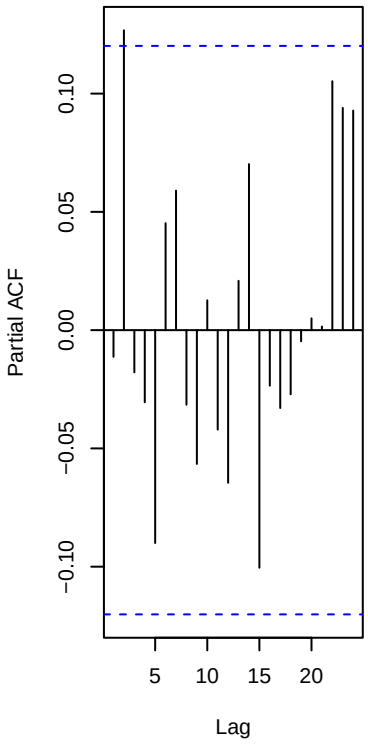


Figure 3: PACF Plot of 3 Pointers Made

Further Testing on Stationarity of Data

We will perform a Fligner-Killeen test to test for homogeneity of variance, with segments of 4, 8, and 12. All 3 tests show high p-values, implying that we have no evidence against the homogeneity of variance. The data has constant variance.

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  train$X3P and seg
## Fligner-Killeen:med chi-squared = 1.9037, df = 3, p-value = 0.5926
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  train$X3P and seg
## Fligner-Killeen:med chi-squared = 4.8291, df = 7, p-value = 0.6808
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  train$X3P and seg
## Fligner-Killeen:med chi-squared = 8.4451, df = 11, p-value = 0.673
```

We will also perform the Dicker-Fuller test to test for non-stationarity. This test shows a small p-value of 0.01, implying that we have strong evidence against the hypothesis of non-stationarity. The time series is stationary.

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  train$X3P
## Dickey-Fuller = -5.8505, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary
```

Differencing

We will start with first order differencing. We immediately see that the resulting acf plot (Figure 4) looks slightly worse than the nondifferenced acf plot (Figure 2). The differenced acf plot has a much larger spike at lag 1, although it continues to suggest that the data is stationary. Looking at the plot of differenced data against time (Figure 5), we see that the points randomly occur about 0 and have constant variance. Therefore, we can conclude the residuals are white noise. However, due to a worse-performing acf plot, differencing is unnecessary.

Figure 4: 1 Time Differenced ACF Plot of 3PM

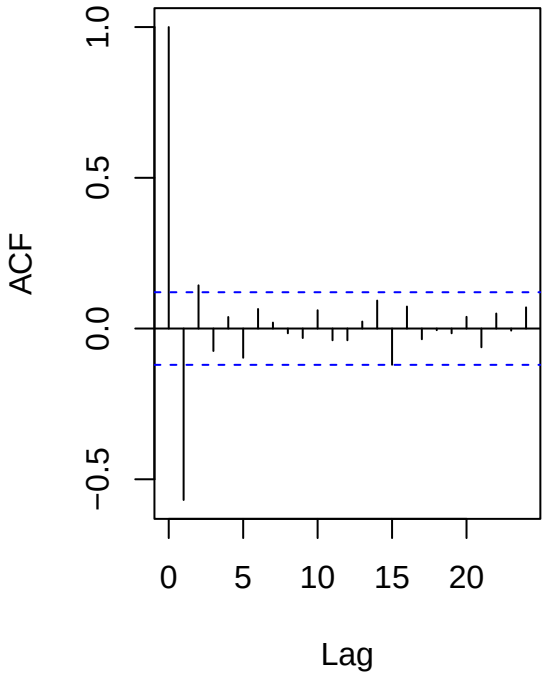
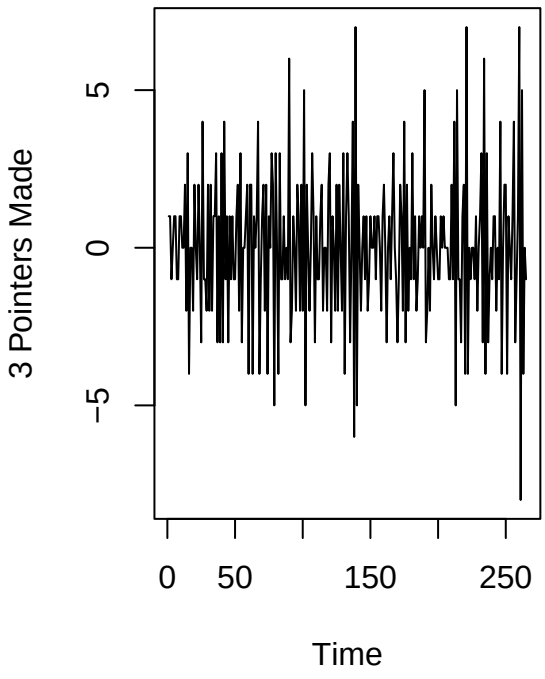


Figure 5: 1 Time Differenced Plot of 3PM



Regression

We will fit a regression model of polynomial degree up to 10. With an AIC selection criterion, we pick a regression model with the lowest AIC of 1025.905. This model has a linear trend (B0 + B1t).

```
##           [,1]
##  [1,] 1025.905
##  [2,] 1026.750
##  [3,] 1027.670
##  [4,] 1029.665
##  [5,] 1031.281
##  [6,] 1033.059
##  [7,] 1028.809
##  [8,] 1030.555
##  [9,] 1031.906
## [10,] 1031.803
```

We then perform a series of residual diagnostics to evaluate the fit of our selected model.

Figure 6: Normal QQ Plot

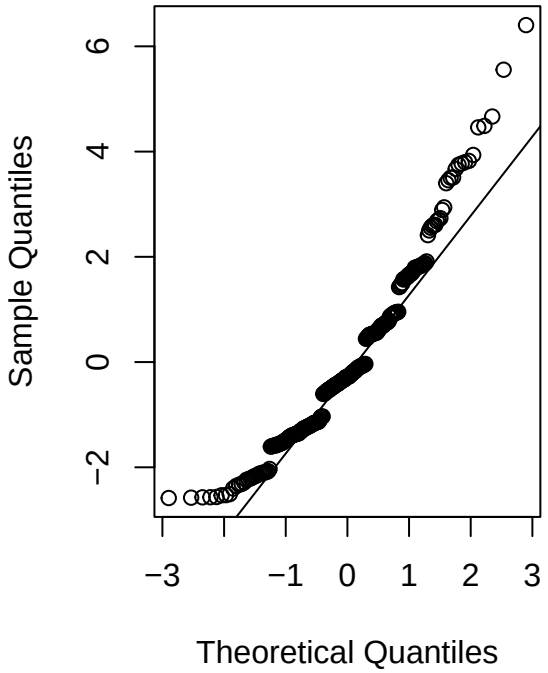
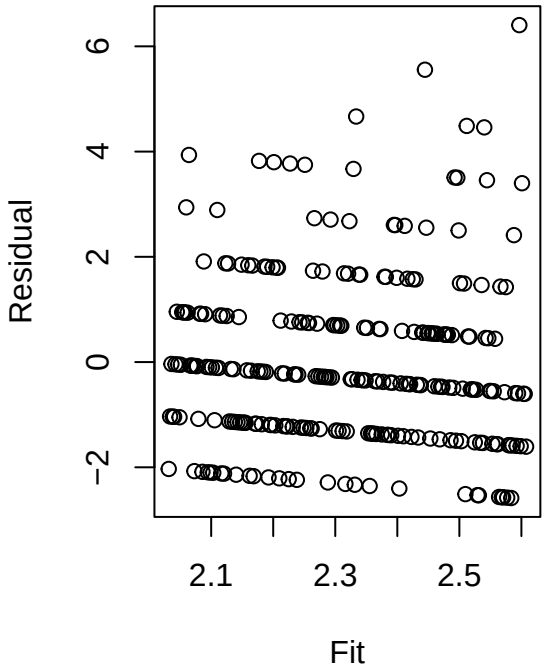


Figure 7: Plot of Fitted vs Residuals



First, we generate a qqplot (Figure 6) and note that the residuals do not fall approximately along the straight line. This violates the normality assumption of residuals. Furthermore, the fitted vs residual plot (Figure 7) do not randomly occur about 0. This implies that the residuals do not have mean 0.

Second, we perform a Shapiro-Wilk Test to test for normality of residuals.

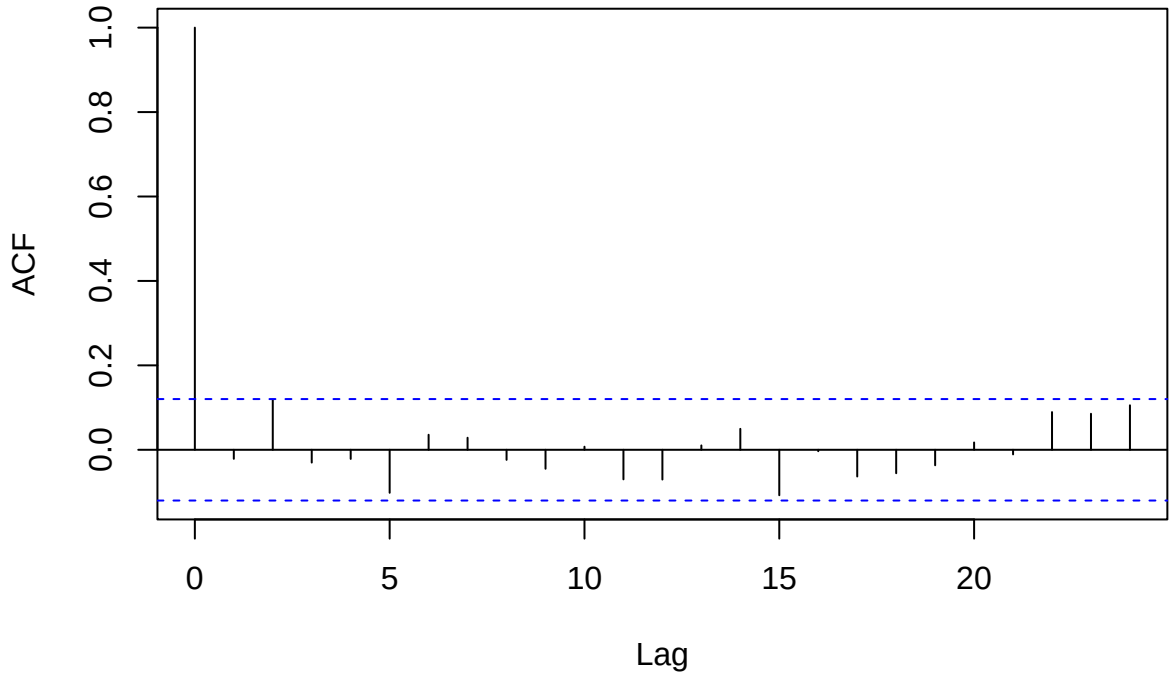
```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  reg$residuals
## W = 0.94403, p-value = 1.548e-08
```

Having observed a low p-value in the Shapiro-Wilk normality test, we have evidence against the residuals being normal.

Third, we generate an acf plot (Figure 8) to check for correlated residuals.

Figure 8: ACF plot of Regression residuals



Since about 95% of the spikes sit within the confidence band, the assumption of independence is not violated.

Fourth, we check for randomness of the residuals.

```
##
## Attaching package: 'randtests'

## The following object is masked from 'package:tseries':
##
##     runs.test

##
##  Difference Sign Test
##
## data:  reg$residuals
## statistic = -4.558, n = 266, p-value = 5.165e-06
## alternative hypothesis: nonrandomness
```

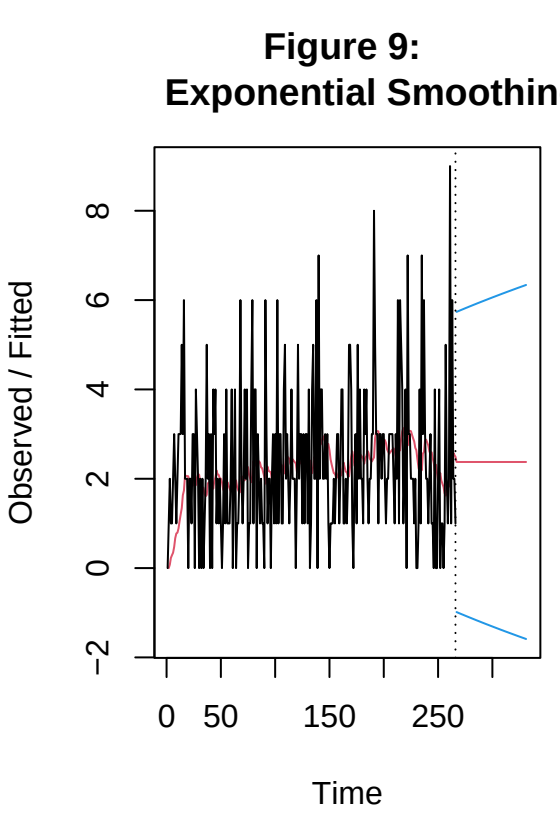
```
##
##  Runs Test
##
## data:  reg$residuals
## statistic = -0.12286, runs = 133, n1 = 133, n2 = 133, n = 266, p-value
## = 0.9022
## alternative hypothesis: nonrandomness
```

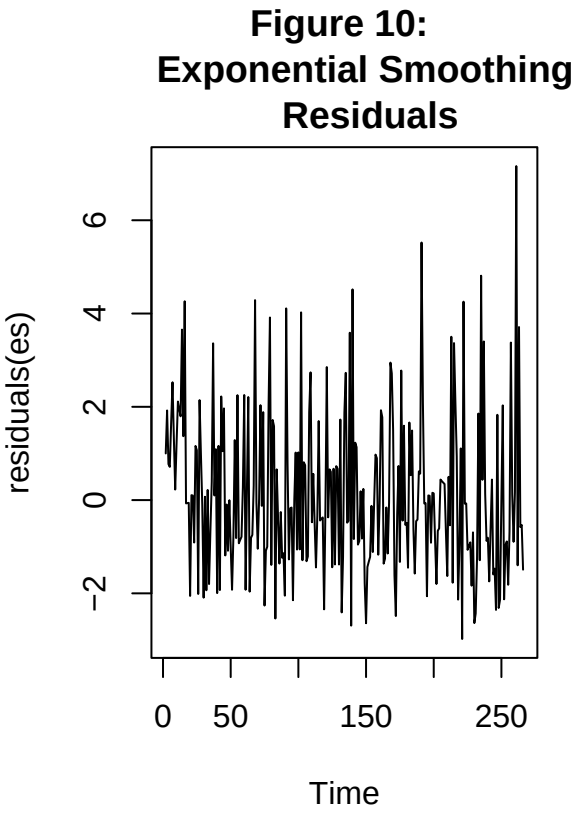
Under the difference sign and runs tests, we have evidence against the hypothesis that the residuals are random.

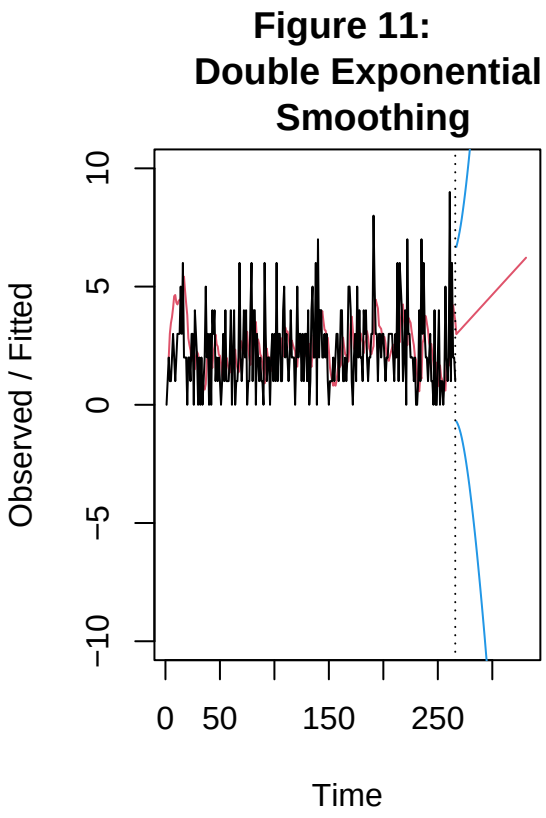
We have learned that the residuals are not random and not normal, although they are uncorrelated. However, they are not white noise because they do not have zero mean. Regression is therefore unnecessary.

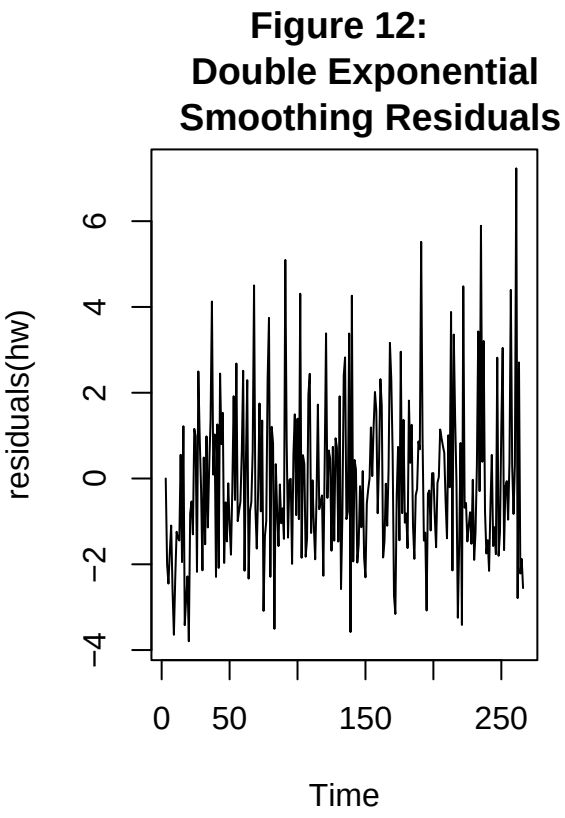
HW smoothing

We will apply simple exponential smoothing and double exponential smoothing. We do not need to consider HW smoothing with seasonality because our data is not seasonal.









Plotting the Holt Winters models (Figures 9 and 10), we can see that the fit to the training data is questionable. We can also see that the predictions using smoothing are far worse. We also observe the residuals resulting from either smoothing method is not white nose because the residual plots (Figures 10 and 12) do not show zero mean. HW smoothing is unnecessary.

We have seen through the results of further testing and applying smoothing methods and differencing that our data is robustly stationary.

Model Proposal

Having proven that our data is stationary, we will propose a class of stationary models. This class of models is ARMA(p,q), where p and q range from 0 to 2.

Model Diagnostics

In this step, we perform several checks to ensure the adequacy of our models. From the QQ-plots in Figures 13-18(refer to Appendix), 95% of the data points do not seem to be falling within the blue bands. The blue bands represent a 95% confidence interval around the 45 degree normal distribution line. This leads to a suspicion of the data not being normally distributed. To further investigate normality, the Shapiro Wilk test of Normality will be performed.

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

```
##
## data:  resid(arma10$fit)
## W = 0.91579, p-value = 4.313e-11
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(arma01$fit)
## W = 0.91502, p-value = 3.743e-11
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(arma11$fit)
## W = 0.9232, p-value = 1.757e-10
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(arma21$fit)
## W = 0.94232, p-value = 1.035e-08
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(arma12$fit)
## W = 0.94194, p-value = 9.471e-09
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(arma22$fit)
## W = 0.95378, p-value = 1.797e-07
```

Inspecting the results of the test, we notice that the p-values are < 0.001 indicating that there is very strong evidence against the null hypothesis. So, we can conclude that there is non-normality in the residuals.

Next we will be inspecting the correlation in the residuals. Examining the Auto-Correlaton Function(ACF) plots in Figure 13-18, we notice that 95% of the spikes fall within the blue bands. Additionally, while we see one or two spikes in Figure 13, Figure 14, and Figure 18 that are on and/or slightly past the blue bands, the residuals seem to be uncorrelated. We conclude the few spikes to be false positives.

In order to further test for correlation, we will also be using the Ljung-Box statistic plots in Figures 13-18. Looking at Figures 16, Figure 17, and Figure 18, we see that we have relatively large p-value. These p-value fail to reject the null hypothesis, meaning that the residuals for our model are uncorrelated. When we look at the plot in Figure 13, there is some concern over the small p-value at lag 2. This could mean that the residuals are correlated at this lag. Similar concerns are present in Figure 14 and Figure 15 in lags < 5.

Overall, we conclude that while none of the models meet the normality assumption for residuals, the ARIMA(2,0,1) model in Figure 16, ARIMA(1,0,2) model in Figure 17, and ARIMA(2,0,2) model in Figure 18 pass our test for autocorrelation and show that the residuals are not correlated.

Fit and Prediction Quality

In this step, we will be taking all the models that have passed the residuals diagnostic test and now test for fit and prediction quaility. For checking fit, we utilize information criteria such as AIC, AICc, and BIC. Lower values of these criteria indicate a better fit.

```
## [1] "ARIMA(1,0,0)"
```

```
##      AIC     AICc      BIC
## 3.866888 3.867059 3.907303
```

```
## [1] "ARIMA(0,0,1)"
```

```
##      AIC     AICc      BIC
## 3.866914 3.867085 3.907329
```

```
## [1] "ARIMA(1,0,1)"
```

```
##      AIC     AICc      BIC
## 3.873236 3.873581 3.927123
```

```
## [1] "ARIMA(2,0,1)"
```

```
##      AIC     AICc      BIC
## 3.865430 3.866006 3.932789
```

```
## [1] "ARIMA(1,0,2)"
```

```
##      AIC     AICc      BIC
## 3.864626 3.865203 3.931985
```

```
## [1] "ARIMA(2,0,2)"
```

```
##      AIC     AICc      BIC
## 3.855405 3.856273 3.936236
```

We see that ARIMA(2,0,2) has the lower AIC and the lowest AICc, ARIMA(1,0,0) has the lowest BIC.

In addition to testing for fit quality, we also are interested in testing for prediction and seeing the prediction power of our models. To test the prediction power, we will be using

$$APSE = MSE_{prediction} = \frac{1}{n} \sum_{n \in test} (y_i - \hat{y}_i)^2 = 1$$

Figure 19: ARIMA(1,0,0)

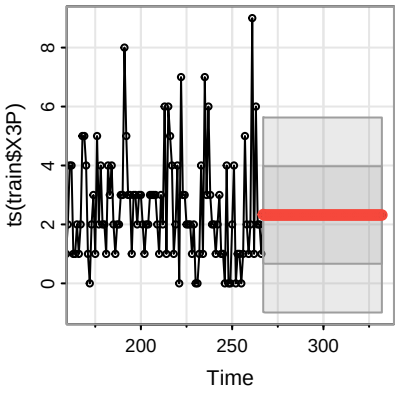


Figure 20: ARIMA(0,0,1)

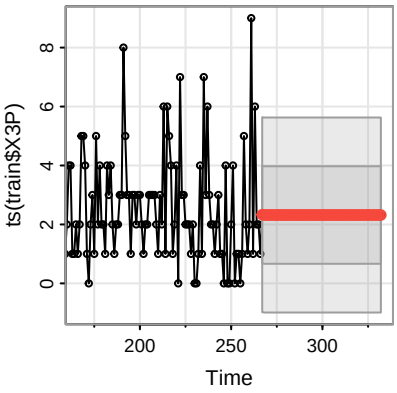


Figure 21: ARIMA(1,0,1)

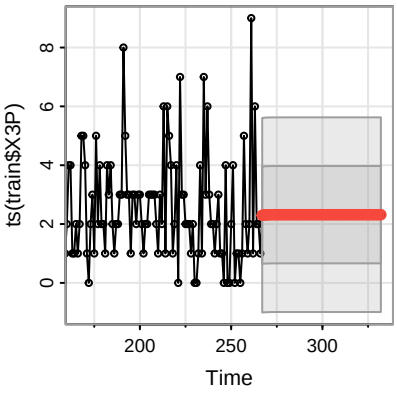


Figure 22: ARMA(2,0,1)

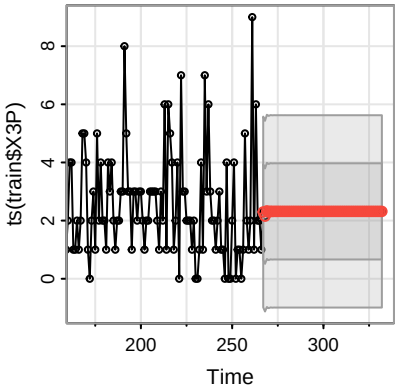


Figure 23: ARIMA(1,0,2)

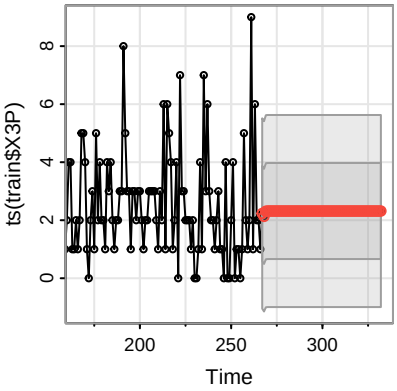
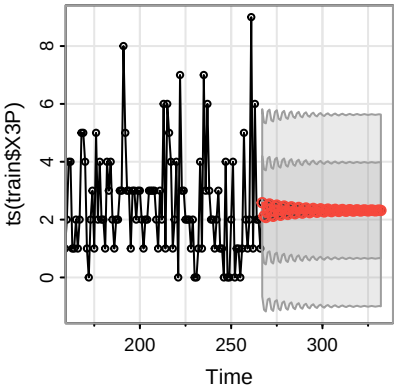


Figure 24: ARIMA(2,0,2)




```
## [1] "APSE for ARIMA(1,0,0):  2.46158163208422"

## [1] "APSE for ARIMA(0,0,1):  2.46173919026684"

## [1] "APSE for ARIMA(1,0,1):  2.4617962876551"

## [1] "APSE for ARIMA(2,0,1):  2.45137774219332"

## [1] "APSE for ARIMA(1,0,2):  2.45379319212815"

## [1] "APSE for ARIMA(2,0,2):  2.45509643391664"
```

Looking at the APSE values for all our models, we see that ARIMA(2,0,1) has the lowest APSE. This means that ARIMA(2,0,1)'s prediction are closer to the actual values of our data.

From our fit and prediction quality check, we have seen that ARIMA(2,0,2) and ARIMA(1,0,0) have the best fit but ARIMA(2,0,1) have the lowest APSE so the best prediction power. Therefore, to move forward with forecasting, we will be using ARIMA(2,0,1).

Forecasting

In this step, we will be using our selected model to predict future data. We will be combining our test and train datasets to use the full information of our data to project the number of 3-pointers Lebron James will be making as his career progresses.

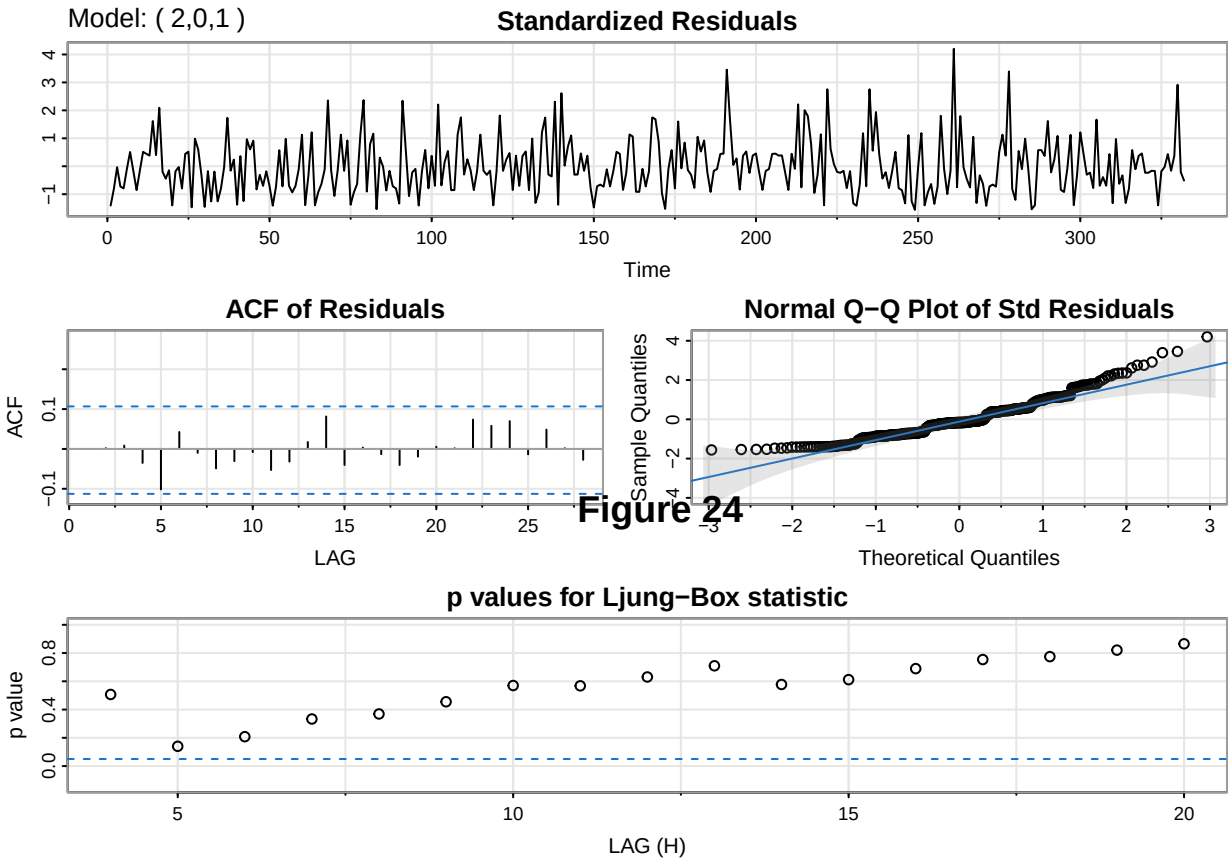

Model: (2,0,1)
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q−Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung−Box statistic
p value
LAG (H)

Figure 24

Figure 25

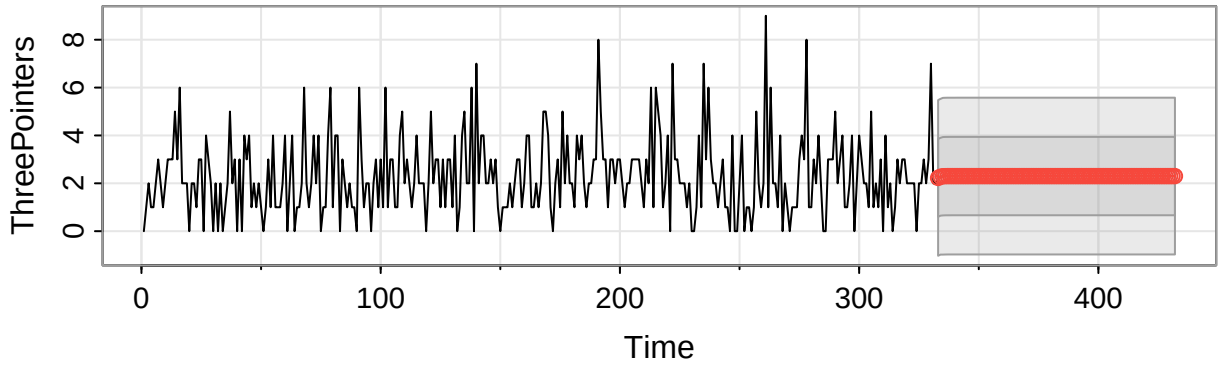
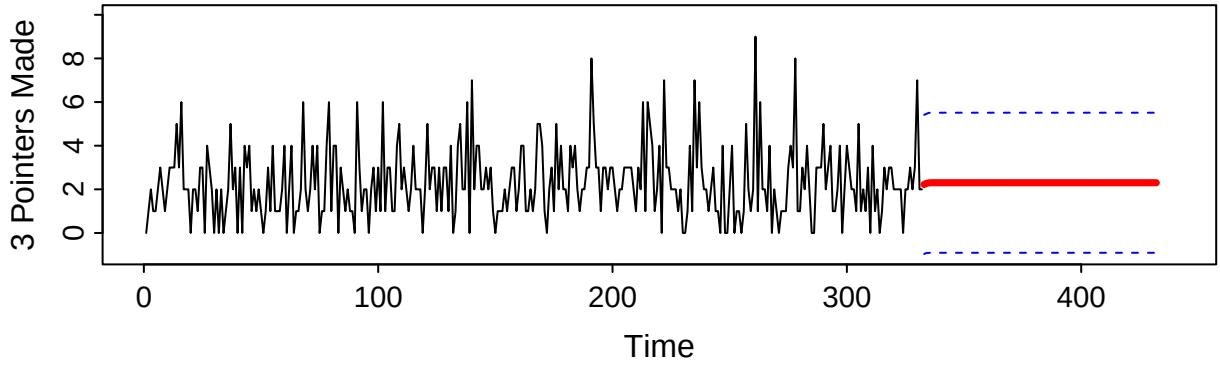


Figure 26: Times Series of 3 Pointers Made with Forecasts



Looking at our forecast plot in Figure 26, we see that the model shows Lebron James to make about two to three 3-pointers per page for the next few years. Our margin of error is given by the blue lines. Looking at the residuals diagnostics in Figure 24, we see that while our residuals violate the normality assumption, they are uncorrelated.

Conclusion

In this analysis, we aimed to predict the total number of 3-pointers that Lebron James would score in a game, considering his age and career trajectory.

Initially, we explored various models, including regression, Holt-Winters smoothing, and differencing for ARIMA models, to capture the underlying patterns and trends in Lebron James' 3-point performance. We saw that there were no trends and seasonality in our data. We also saw that our variance was constant. Since we had a constant mean and constant variance, we concluded that our data was stationary and proceed to propose a few stationary models.

We proposed a class of stationary models, ARIMA(p,d,q), where p and q ranged from 0 to 2, and evaluated their adequacy through diagnostics, fit and prediction assessments. Despite seeing that the residuals were not normal, we identified ARIMA(2,0,1) as the best model to use for our forecasting since it had a good combination of fit and prediction quality. With ARIMA(2,0,1) selected, we proceeded to forecast Lebron James' future 3-point performance. Our forecast suggests a consistent level of performance, with Lebron making approximately two to three 3-pointers per game over the next few years. .

In conclusion, our analysis provides a robust framework for predicting Lebron James' 3-point performance, leveraging time series modelling techniques. As Lebron continues his career, our model serves as a tool for fans, analysts, and decision-makers to anticipate his future contributions on the court.

Appendix

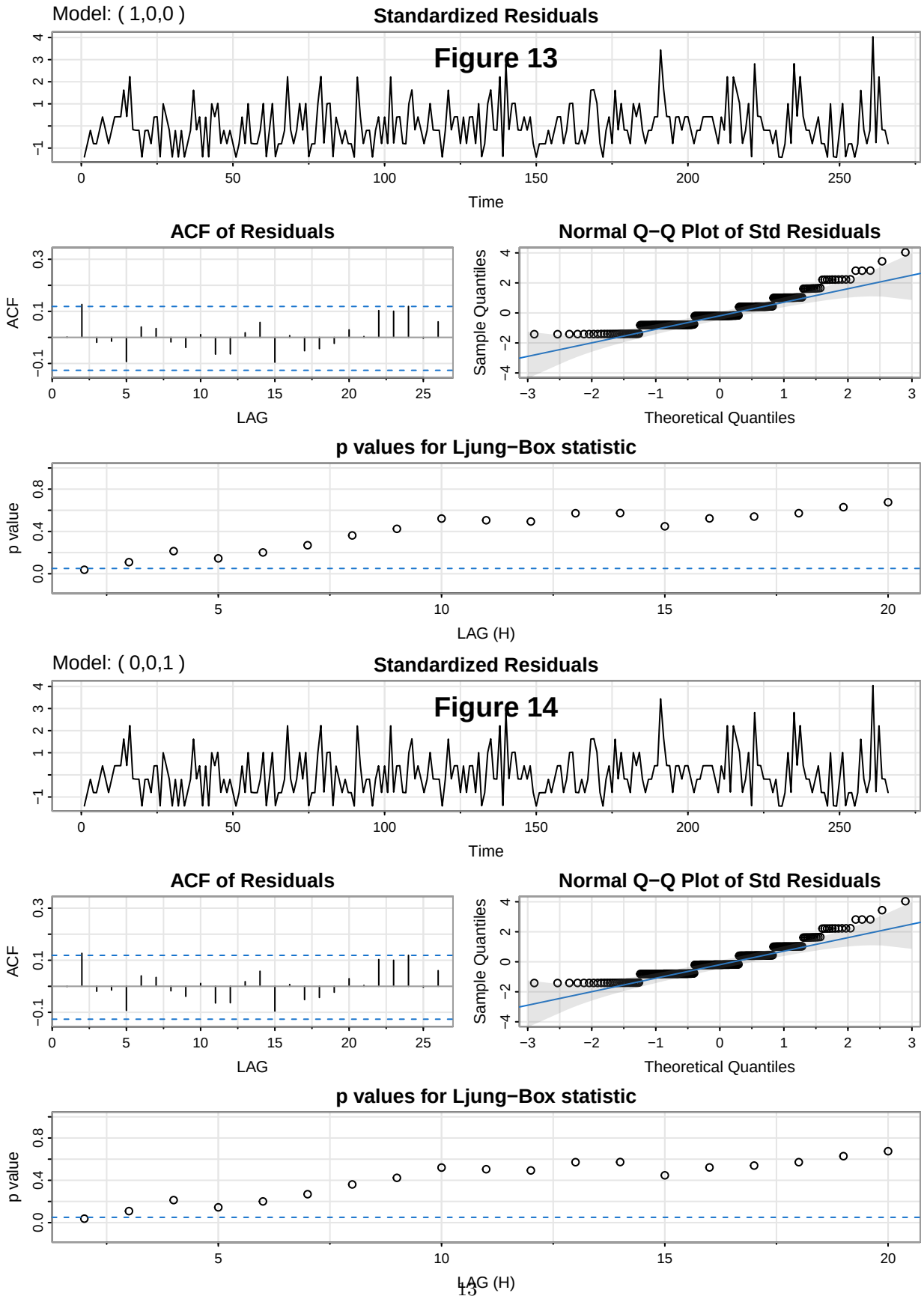


Figure 13

Figure 14

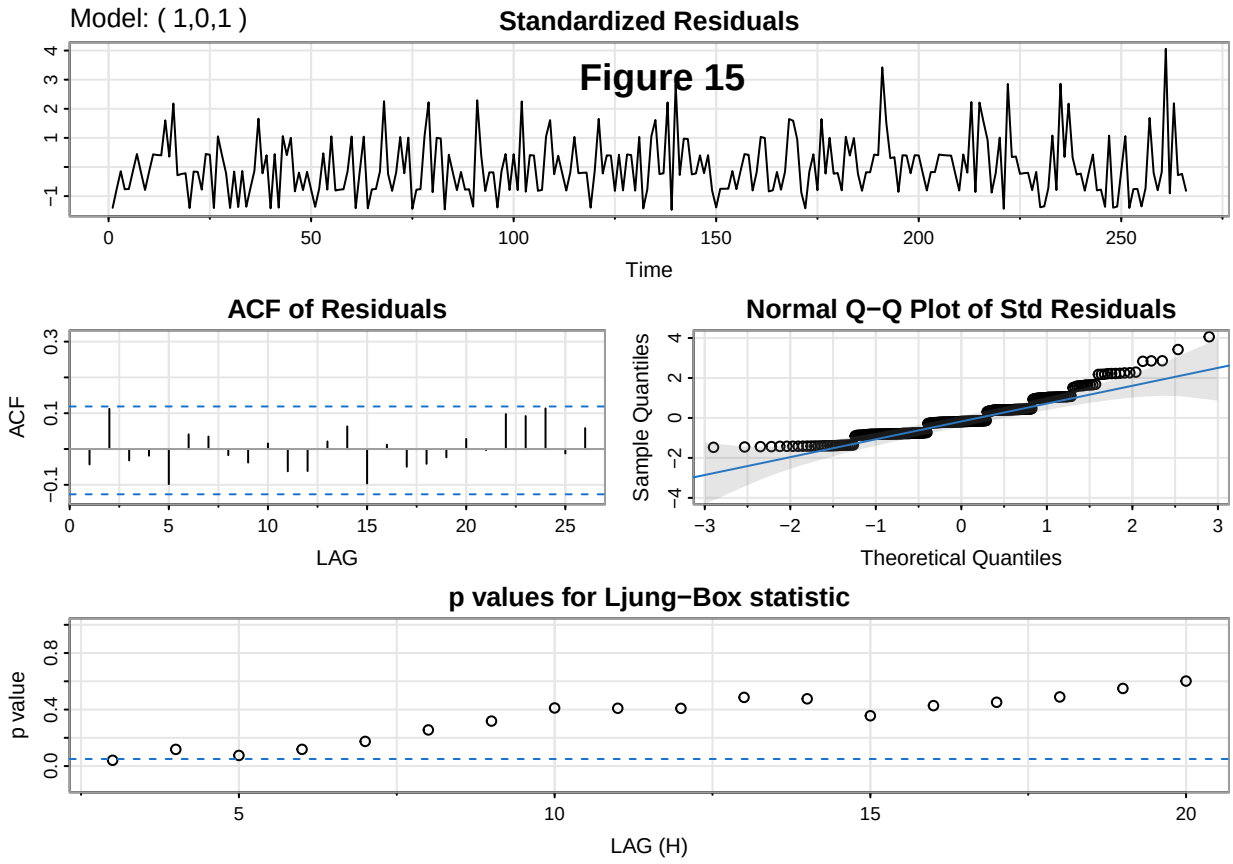

Model: (1,0,1)
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q−Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung−Box statistic
p value
LAG (H)

Figure 15

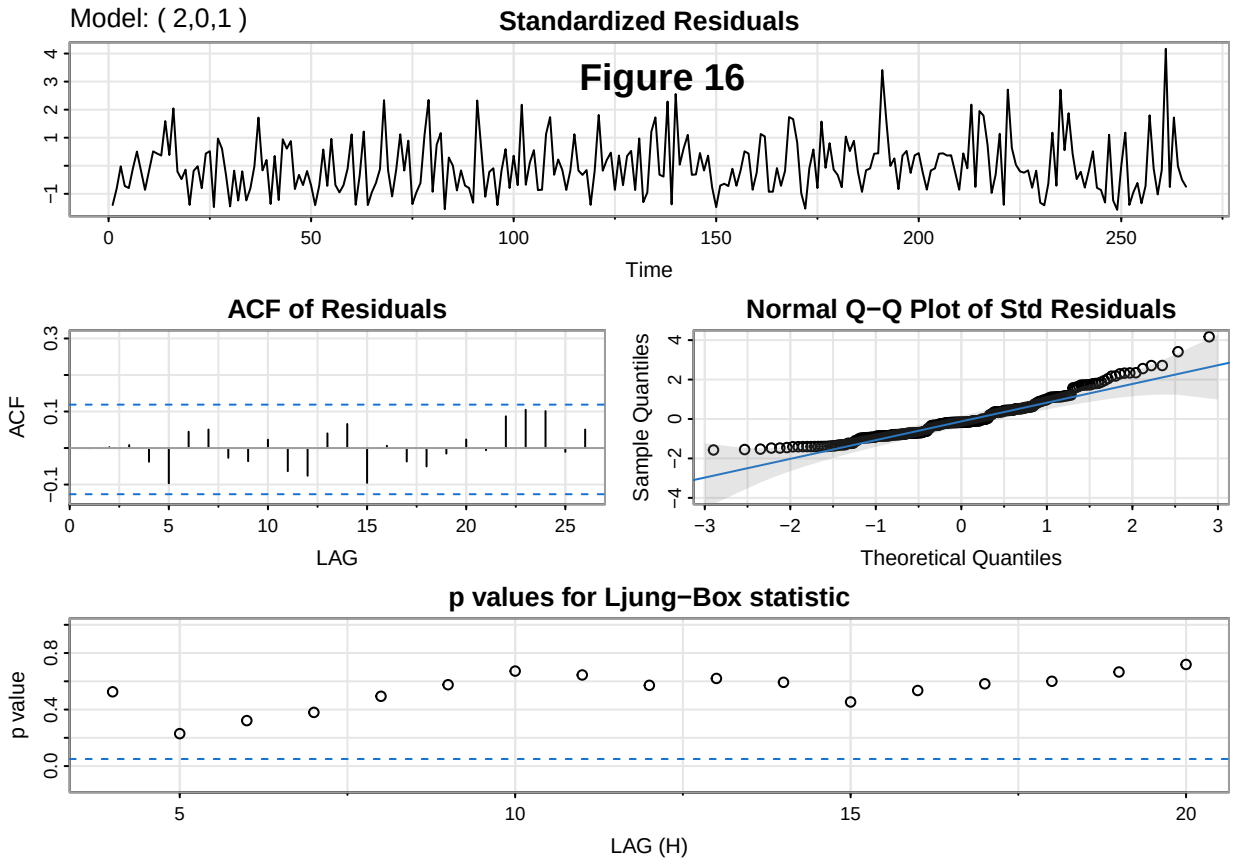

Model: (2,0,1)
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q−Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung−Box statistic
p value
LAG (H)

Figure 16

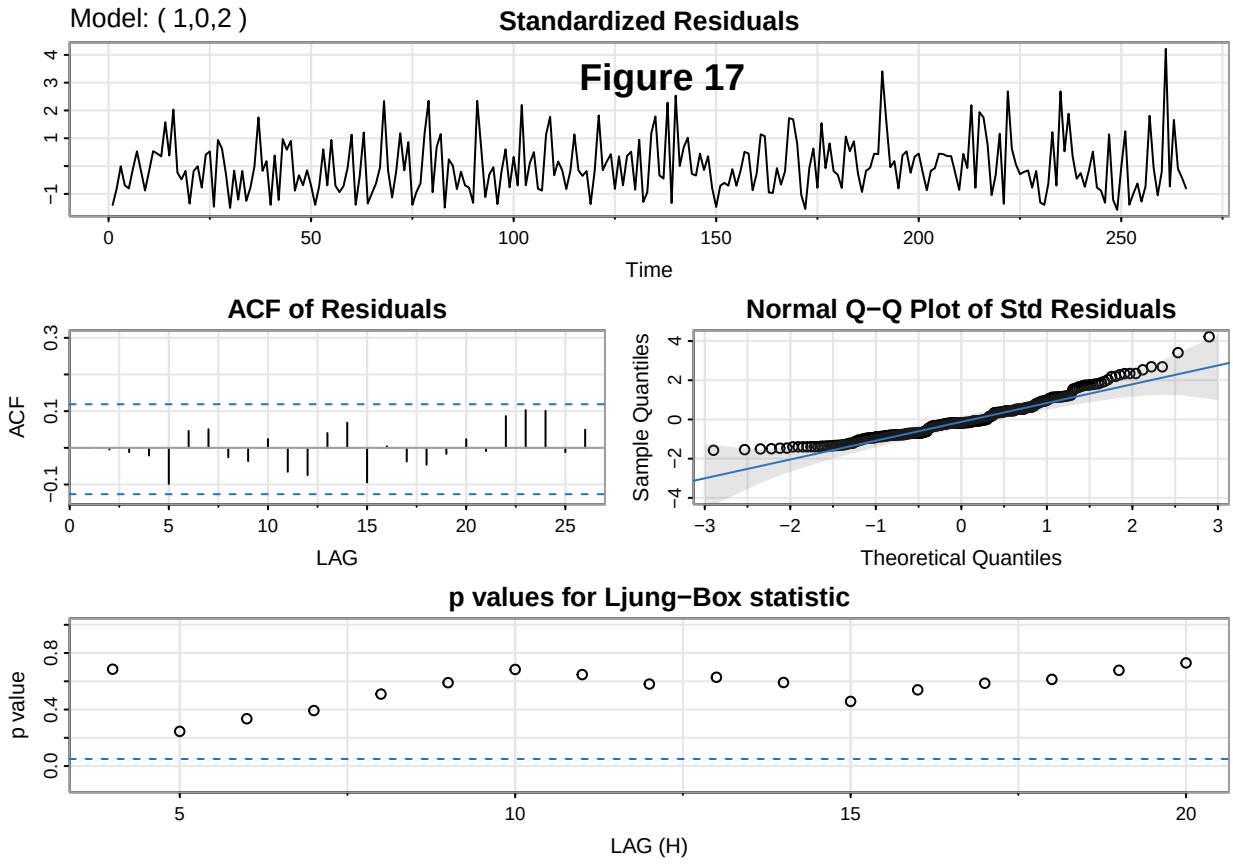

Model: (1,0,2)
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q−Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung−Box statistic
p value
LAG (H)

Figure 17

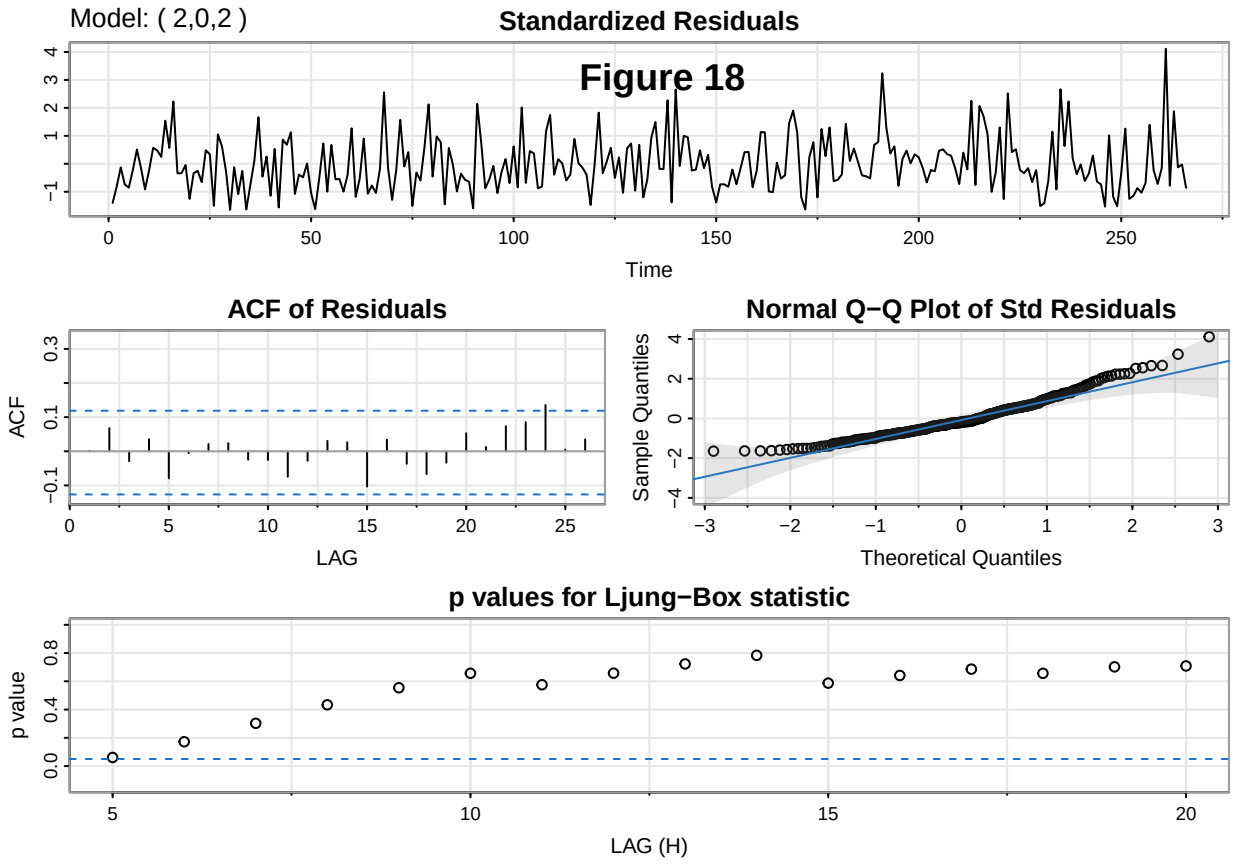

Model: (2,0,2)
Standardized Residuals
Time
ACF of Residuals
ACF
LAG
Normal Q−Q Plot of Std Residuals
Sample Quantiles
Theoretical Quantiles
p values for Ljung−Box statistic
p value
LAG (H)

Figure 18