

STATISTICS FOR DATA ANALYTICS

CONTINUOUS ASSESSMENT - REPORT

ON

MULTIPLE LINEAR REGRESSION

AND

TIME SERIES ANALYSIS

**Submitted by**

Gollamudi Sarath Chandra – x19193581

MSc Data Analytics – JAN/Group A – 2020

**Multiple Regression Model**

The Multiple regression model is a statistical technique that is used to find the relationship between the continuous dependent variable and several independent variables, as well as predictive analysis.

**Objective:**

The main objective of this is to perform the multiple linear regression on the Internet usage of individual’s dataset. There are 3 types of multiple regression which are as follows Standard, Hierarchical and Stepwise. Using the hierarchical multiple regression, the analysis has accomplished to estimate the interrelationship among the independent variables (IVs) and the percentage of influence on the dependent variable (DV). The Following are the dependent and independent variables.

**IVs**

Job Purpose 🡪 Scale

Banking 🡪 Scale

Social Network 🡪 Scale

Online Course 🡪 Scale

Weekly Once Usage 🡪 Scale

Daily Usage 🡪 Scale

**DV**

Overall usage 🡪 Scale

**1. Dataset and Research Question:**

**1.1 Data Set Used:**

The dataset is related to the Internet usage of an individual person around the European region countries. It is extracted from the Digital Economy and society tables present on the Eurostat website. The initial dataset contains 126 records out of which, as part of the data cleaning process the removal of NA values has performed in R and Excel. After the cleaning, the dataset consists of 114 records with 7 columns with which the linear regression has performed. This data has formed by merging the columns from different tables.

**Data Source:** <https://ec.europa.eu/eurostat/web/digital-economy-and-society/data/main-tables>

**1.2 Research Question:** Predicting the overall usage of the internet by an individual person around the European region using the factors which are influencing.

**2. Regression Type and Assumptions Verification:**

**2.1 Hierarchical Regression:**

The Hierarchical regression is a statistical method of finding the correlation among and testing a hypothesis about a DV and several IVs. In this method, the IVs are entered into the equation in blocks, with every IV being assessed in terms of what it adds to the prediction of the DV once the previous variables are controlled for.

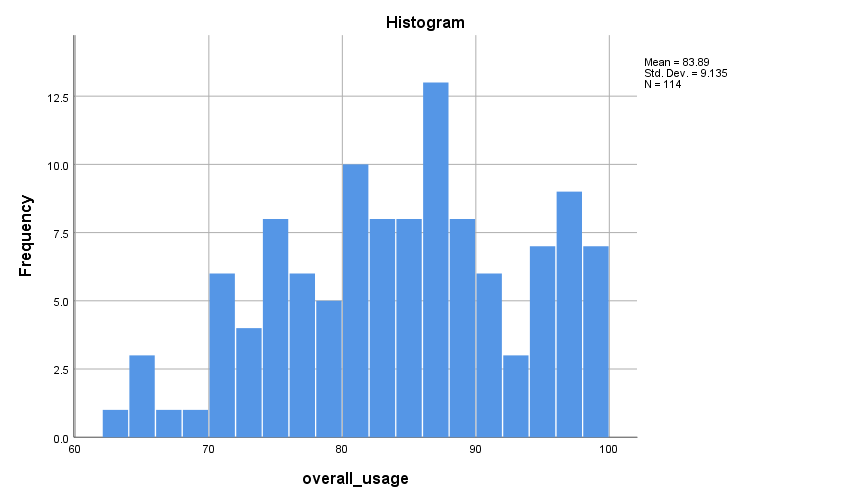
**2.2 Assumptions:**

**Sample Size:**

The sample size indicates how many samples required for the multiple regression. For calculating the sample size (Fidell, 2013, p. 123) has given a formula which is as follows N>50+8m ( where m is the total number of IVs). As per the formula, the value is N>98 and the records chosen are 114. This means that the sample size assumption is satisfied.

**Normality Test for the DV:**

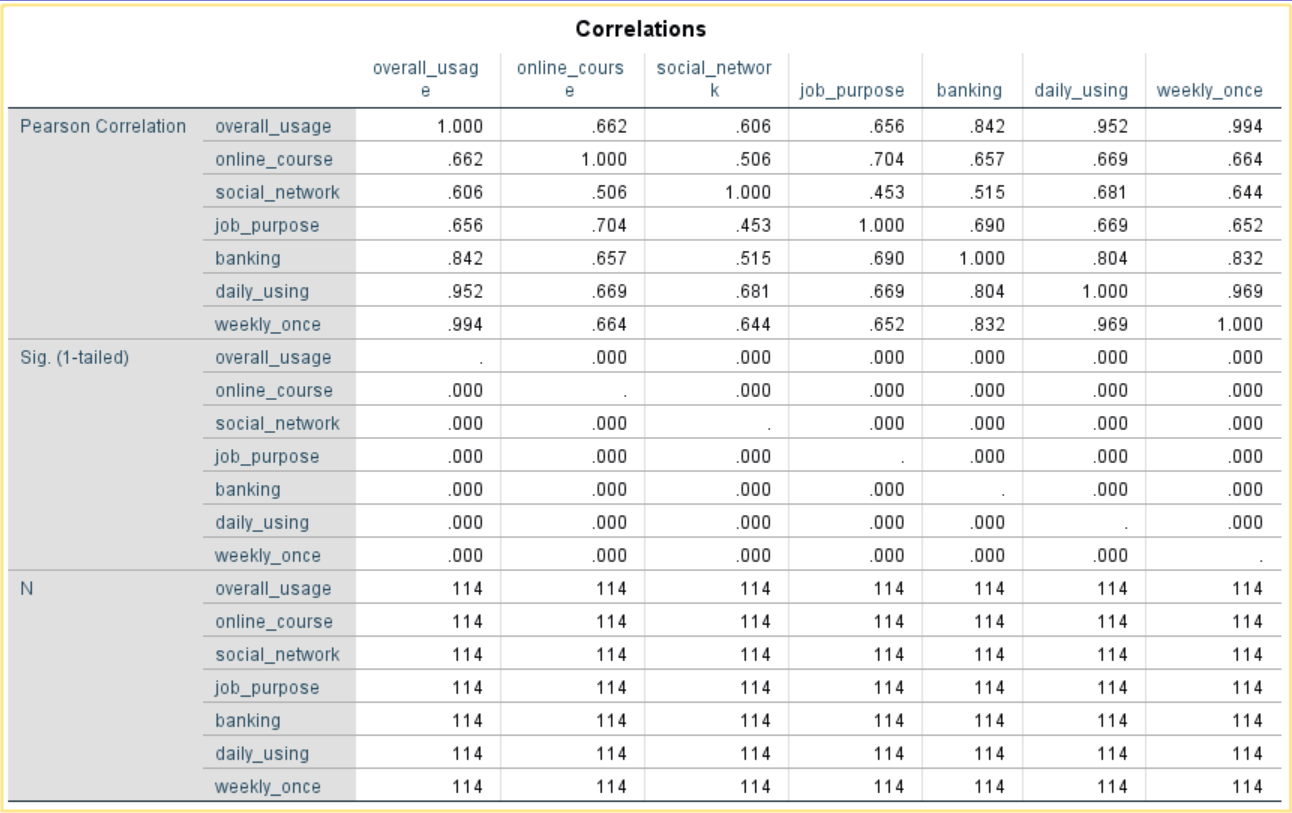
The dependent variable “overall usage of the internet” is plotted on the histogram plot and the skewness is determined by using the descriptive statistics. We could see that the data has been distributed normally and there is no skewness.



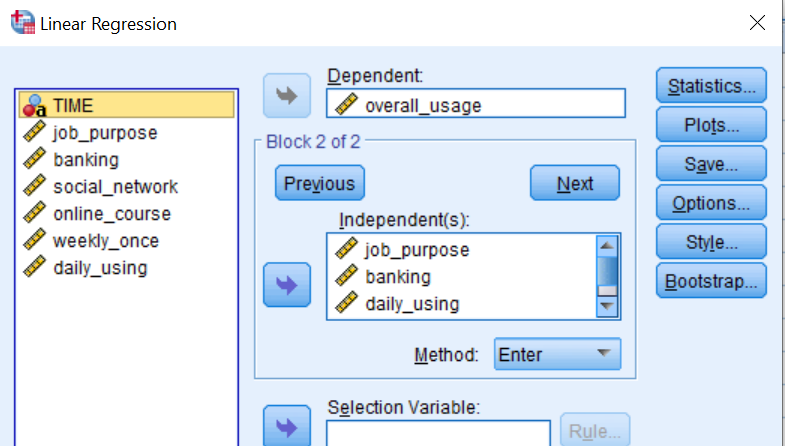
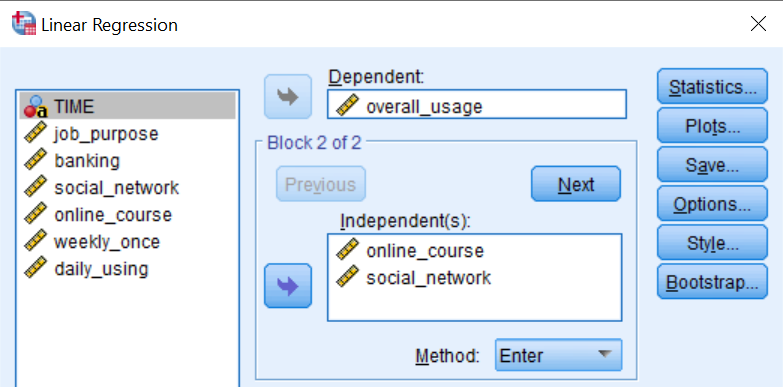
Based on the output generated from the multplie regression below steps has been explained.

**Correlations & Multicolinearity:**

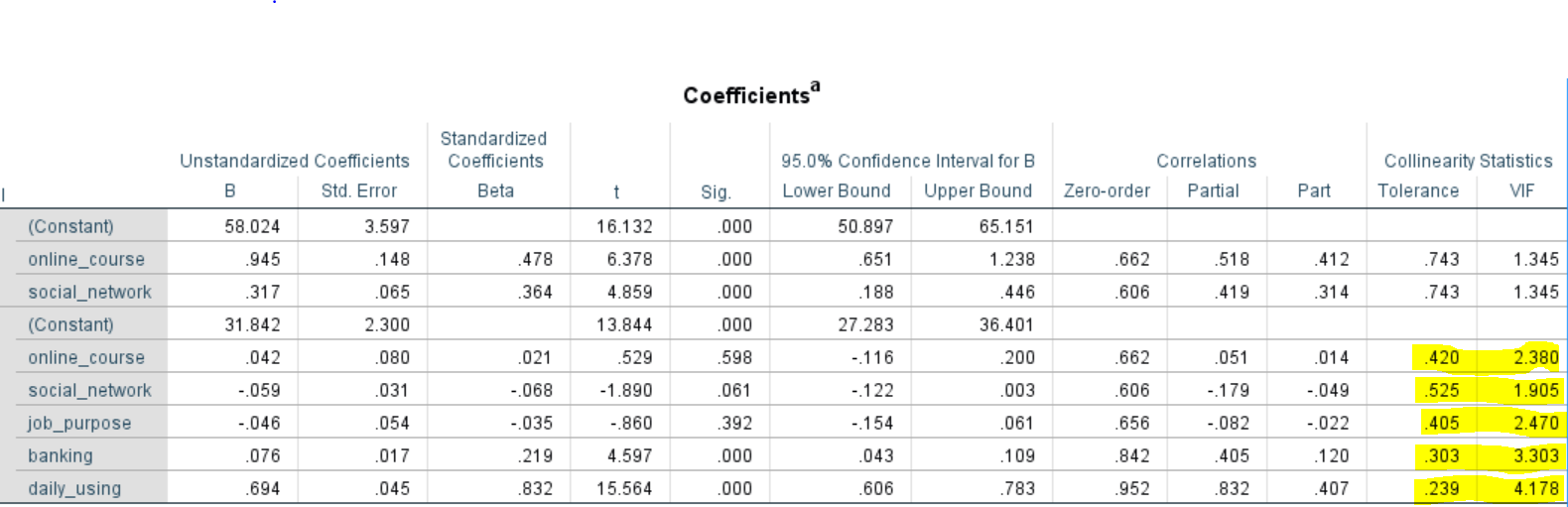
These tell the relationship between the variables. The independent variables should show the relationship of at least above .3 and close to 1 with the dependent variable. We could see in the below table that the correlation among the variables is neither less than .3 nor above 1 and statistically significant at the 0.01 level.



From the table above, we could see that the correlation between IVs: daily usage and weekly once is high with the DV: overall usage when compared to the other independent variables which are (.952 & .994). This means that we can ignore the highly correlated variable while performing the hierarchical regression. To do the hierarchical regression we need to arrange the data in the form of blocks as shown below.



We could see from the above table the 5 IVs were selected in the form of 2 blocks after ignoring the highly correlated variable (weekly once) with the DV as overall usage. In the correlation, matrix multicollinearity may not be evident. To verify this, we need to check the **Tolerance and VIF** values of the results which are present in the table called coefficients after generating the hierarchical regression output.



Tolerance is an index that tells what quantity of the variability of the required independent is not explained by other IVs in the model and the formula is as follows 1-R squared of each variable. On the other hand, VIF is just a converse of the tolerance value (1/tolerance). Moreover, the presence of multicollinearity can be identified if the tolerance value is less than .10 or VIF value is greater than 10. This means that we could see in the table above both the values are within the specified limit, so we have not breached the correlation and multicollinearity assumptions.

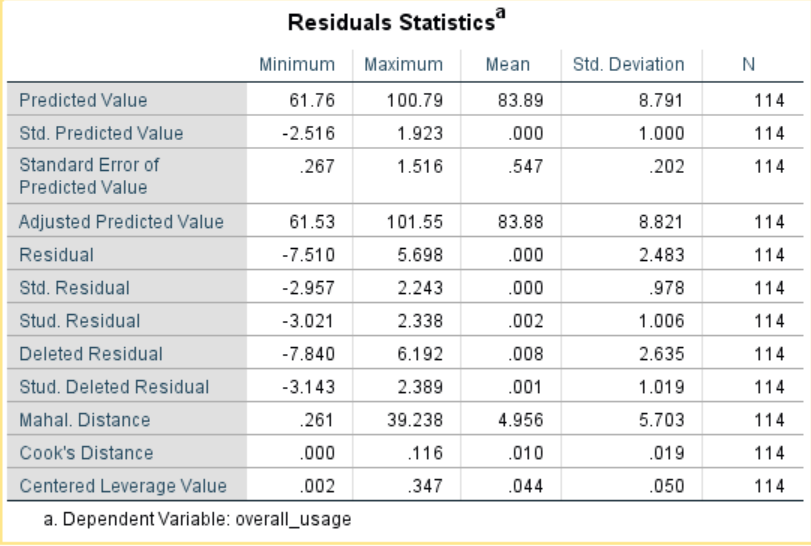
**Outliers Verification:**

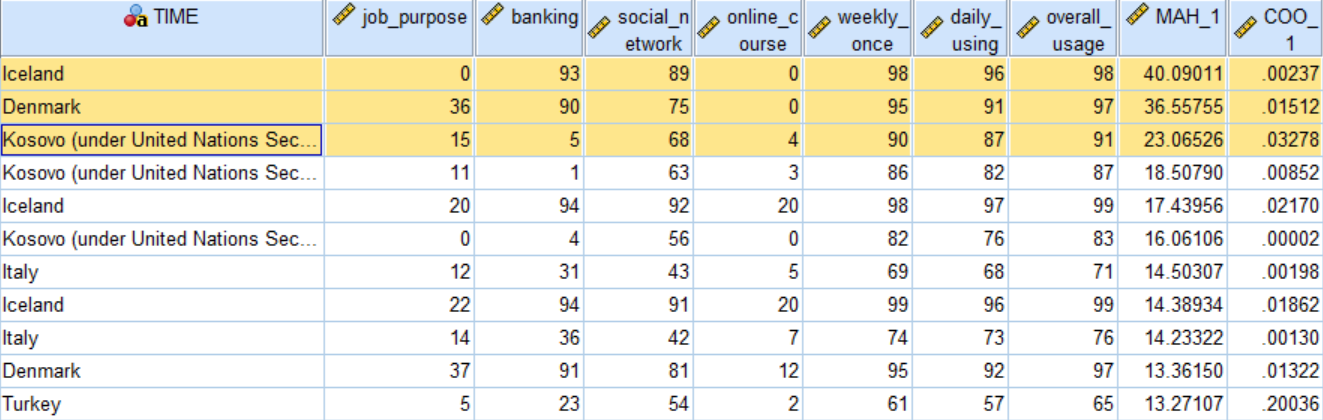
These outliers can be detected by verifying the Mahalnobis distances which are produced by the multiple regression. To identify which samples are outliers we need to determine the critical chi-square value. A list of chi-square values which are already determined in the book (Fidell, 2013, p. 123) has been taken that are as follows.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **IVs** | **CV** | **IVs** | **CV** | **IVs** | **CV** |
| 2 | 13.82 | 4 | 18.47 | 6 | 22.46 |
| 3 | 16.27 | 5 | 20.52 | 7 | 24.32 |

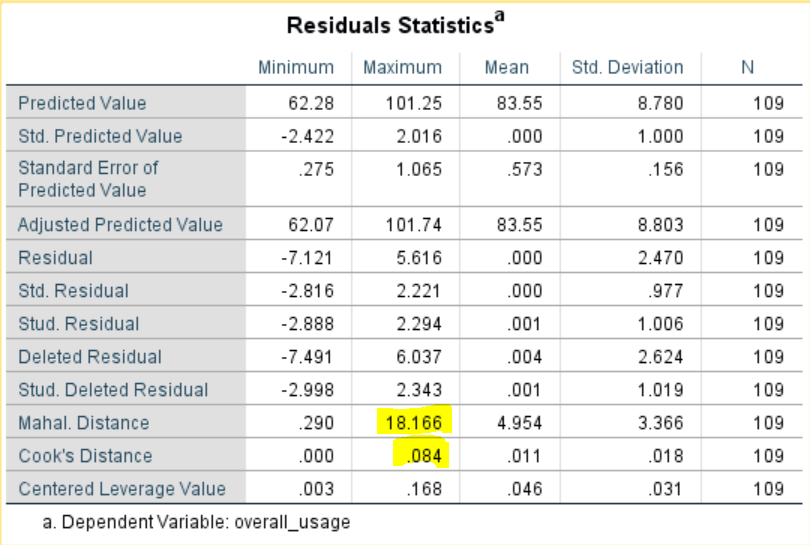
In our case, the total IVs are 6, so the Mahalnobis distance should lay under the value 22.46.

We could see in the below residual statistics table the value is 39.238 which is greater than the defined. So, to adjust this we need to remove the outlier values which are above the defined value from the sheet the data view of spss.



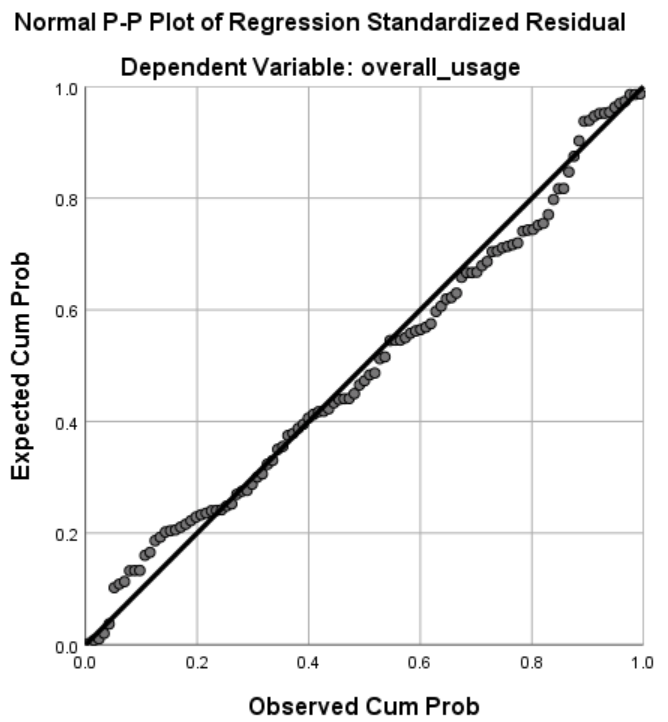
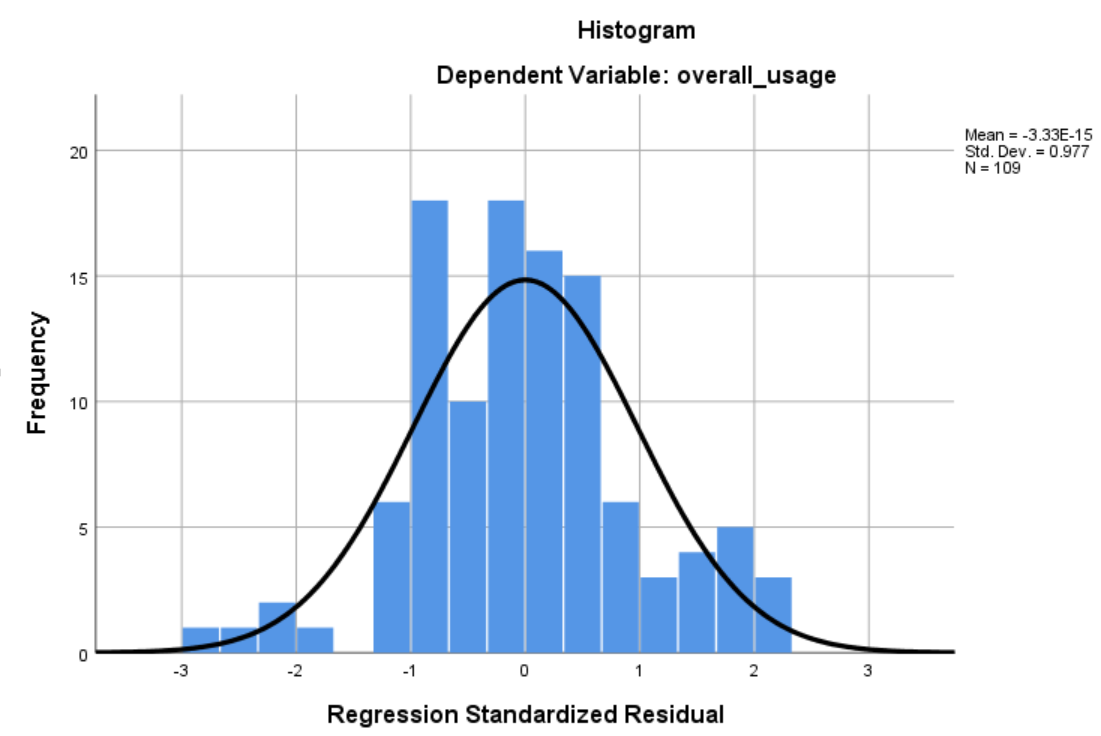


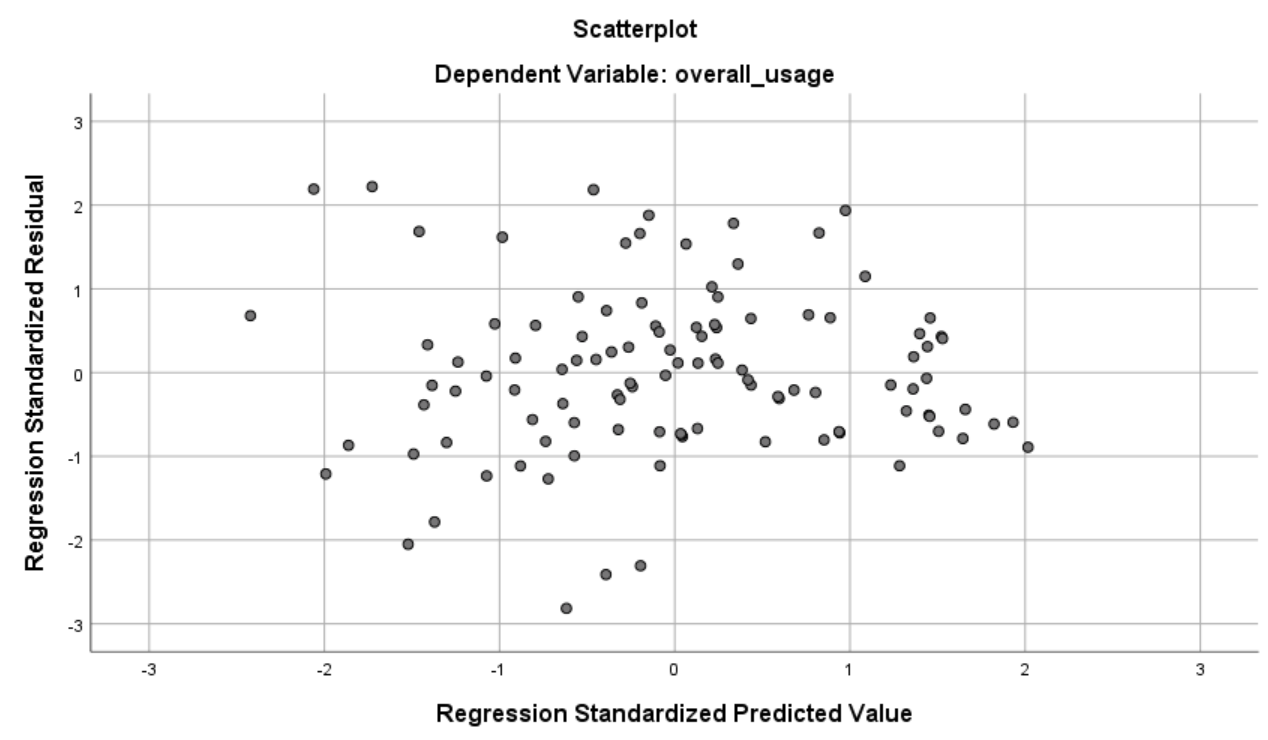
After removing the outliers, in the regression output the mahal distance value is below the critical value (18.166 < 20.52) as shown in the below table. So, we could say that there are no outliers. On the other hand, based on Cook’s distance value we can decide if there is any major problem with the data. According to (Fidell, 2013, p. 123) if the cook’s distance value is below 1 it suggests that there are no major problems in our data. We could also see that the cook’s distance is below 1 in our output. This means that no major problems.



**Homoscedasticity & Independence of Residuals:**

These assumptions can be verified by inspecting the Histogram, Normal P-P plot and the scatter plot. Below are the outputs of our hierarchical multiple regression.

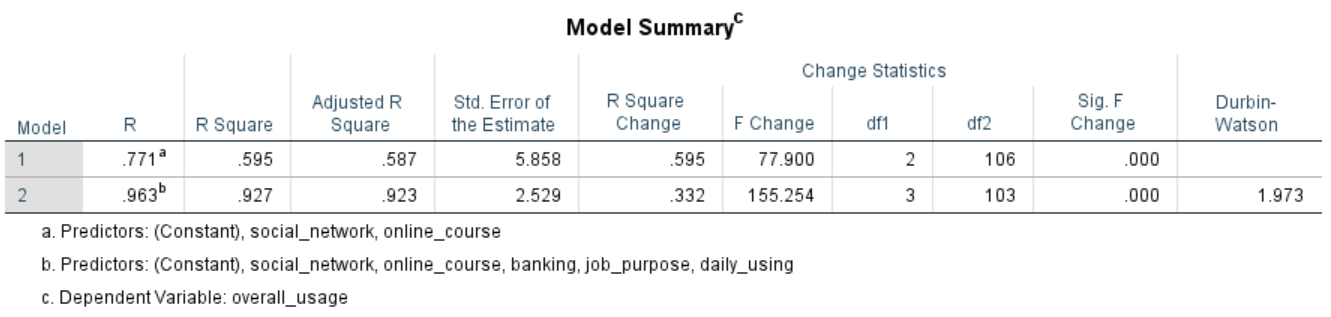




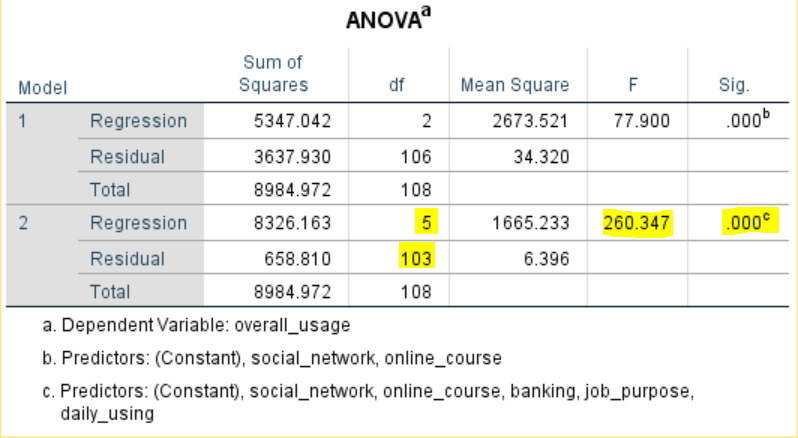
* From the **Histogram,** we could tell that our DV is normally distributed.
* In the **Normal P-P plot** it is clearly showed that there is no major deviation from the normality line and the points lie in a straight diagonal line from bottom left to top right.
* **Scatter Plot** tells that the residuals are rectangularly distributed with most of the points are close to the centre line and lies within the range -3 to 3. This means that **homoscedasticity** is not violated.

**Evaluating the model:**

The model is evaluated using the summary table of the generated output. Since we have used the Hierarchical multiple regression method for our analysis in the below table the output is in the form of 2 models.



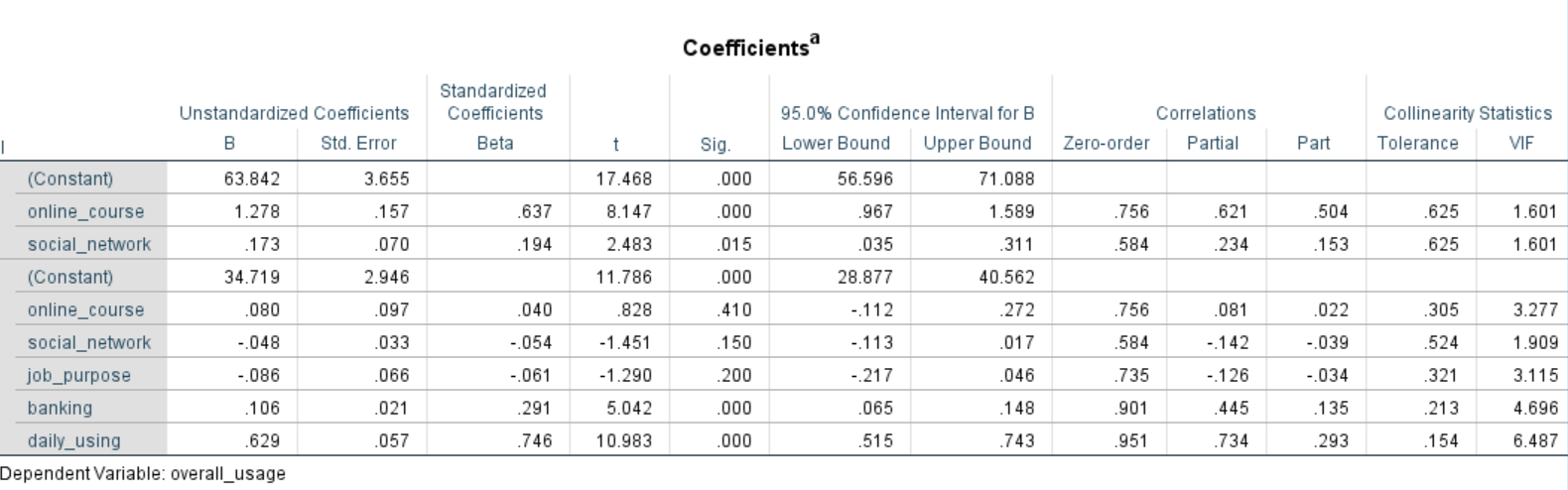
The value of R shows that the correlation of the overall usage of the internet with the predictors. In the table above the R square value which is the square root of the R of the first model (social\_network, online\_course) explains 59.5 percent of the variance (0.595\*100) between the dependent and predictor variables at a statistical significance level (.000). Whereas in the second block when we entered the other variables (banking, job\_purpose, daily\_using) the new correlation R is .963 and the model explains 92.7 percent (0.927\*100) at a statistical significance level (.000). The second block includes all the variables which are in the first block as well. The autocorrelation has measured using the Durbin-Watson test, we could see that the value is less than 2 this means that our data has no outliers.



To verify whether the F values are statistically significant or not the **ANOVA** test has performed. The table above indicates (which includes both blocks of variables) is significantly based on the F value at degrees of freedom to be **F (5, 103) = 260.347, p<0.05.**

**Evaluating the independent variables:**

To evaluate how well each of the variables makes a uniquely significant contribution to the final equation, we need to look at the below coefficients table.



From the table above we could see that in the second block which summarises results of all the variables there are only 2 variables that are contributing highly, daily\_using (beta=.746) followed by banking (beta=.291) at a statistically significant level (.000). Neither online course nor social network made a unique contribution.

**Conclusion:**

Hierarchical multiple regression was performed to investigate the overall usage of the internet by an individual person around the European region. In the first step of hierarchical regression, two predictors were entered: online\_course and social network. This model was statistically significant F (2, 106) = 77.900, p< .005 and explained 59.5 percent of the variance in DV which made a unique contribution to the model. After entry of banking, job\_purpose and daily\_using at step 2 the total variance explained by the model was 92.7% (F (5, 103) = 260.347, p<0.05). The introduction of banking, job\_purpose, and daily\_using explained an additional 33.2% of the variance in DV, after controlling for online\_course and social\_network. In the final adjusted predictor variables, daily\_using and banking were recording a higher Beta value of (.746 & .291) than the remaining IVs.

**TIME SERIES ANALYSIS**

The collection of numerical data over a period (weekly, monthly, quarterly, or yearly) is called a time series. On the other hand, the word forecasting refers to predicting future events based on historical activities.

**Objective:**

The main objective of this time series analysis is to perform a forecasting method on the “total wastewater generated” dataset of the country Republic of Moldova and predicting future events. There are several methods in R for forecasting, among them the methods that I have used are as follows Holt-winters smoothing method and ARIMA.

**1. Dataset and Research Question:**

* 1. **Dataset Used:**

The dataset is related to the total wastewater in the country “Republic of Moldova”. It is extracted from the ‘Environment Statistics Database’ on the UNdata website. The dataset consists of 20 of years data (1995-2015) with which the forecasting methods have performed.

**Data Source:** <http://data.un.org/Data.aspx?d=ENV&f=variableID%3a84>

* 1. **Research Question:**

Predicting the wastage of water based on the historical time series data by using different forecasting methods and finding the best prediction method for this data.

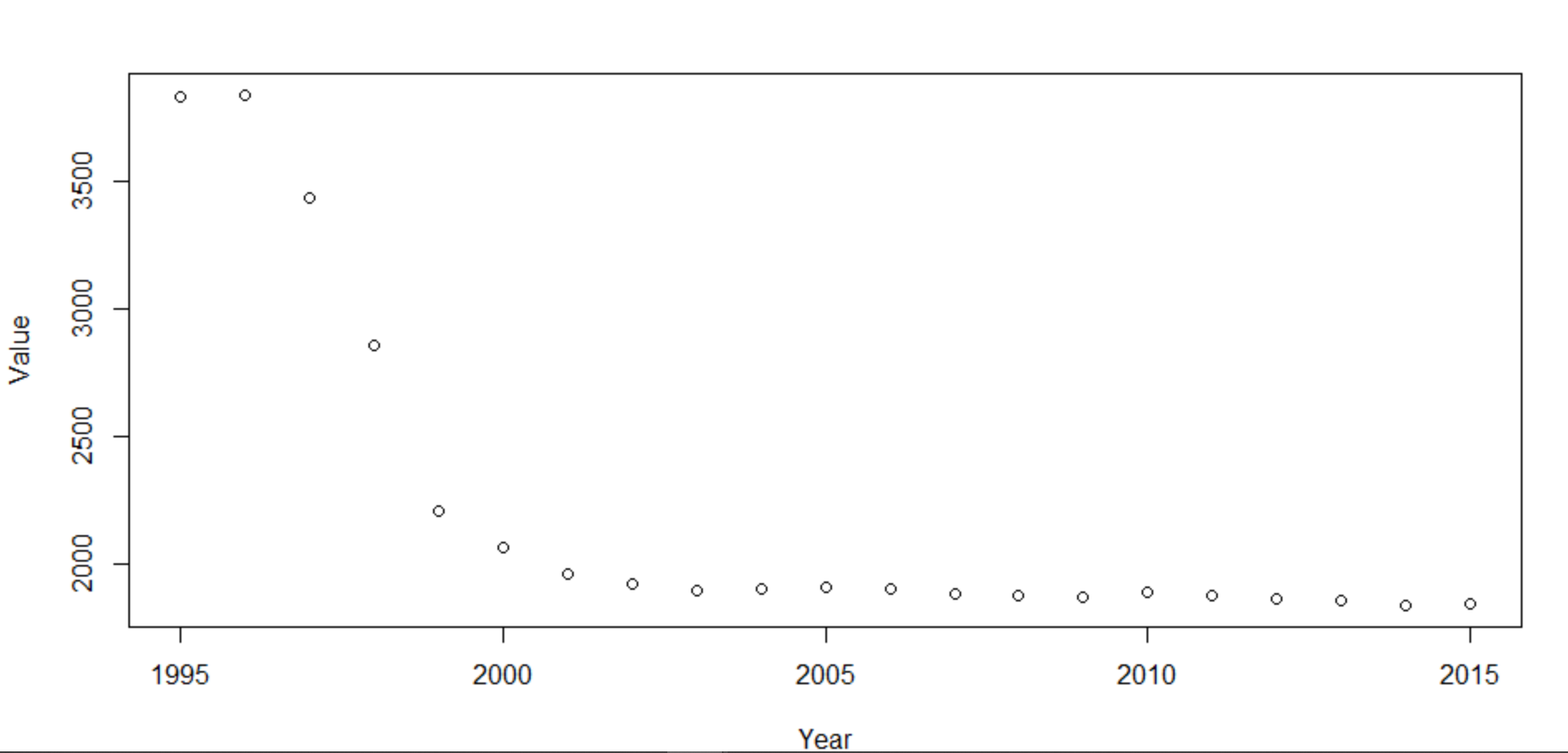
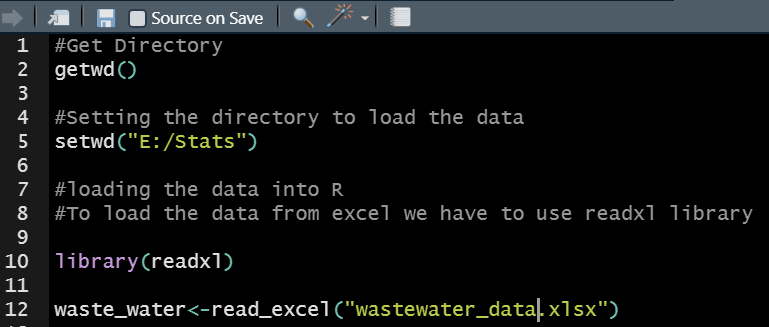
**2. Process of Forecasting:**

The forecasting process is analysed based on the below steps.

* Analysing the Data
* Verifying the data is stationary or Non-stationary
* Forecasting and Evaluating the best model

**Analysing the data:**

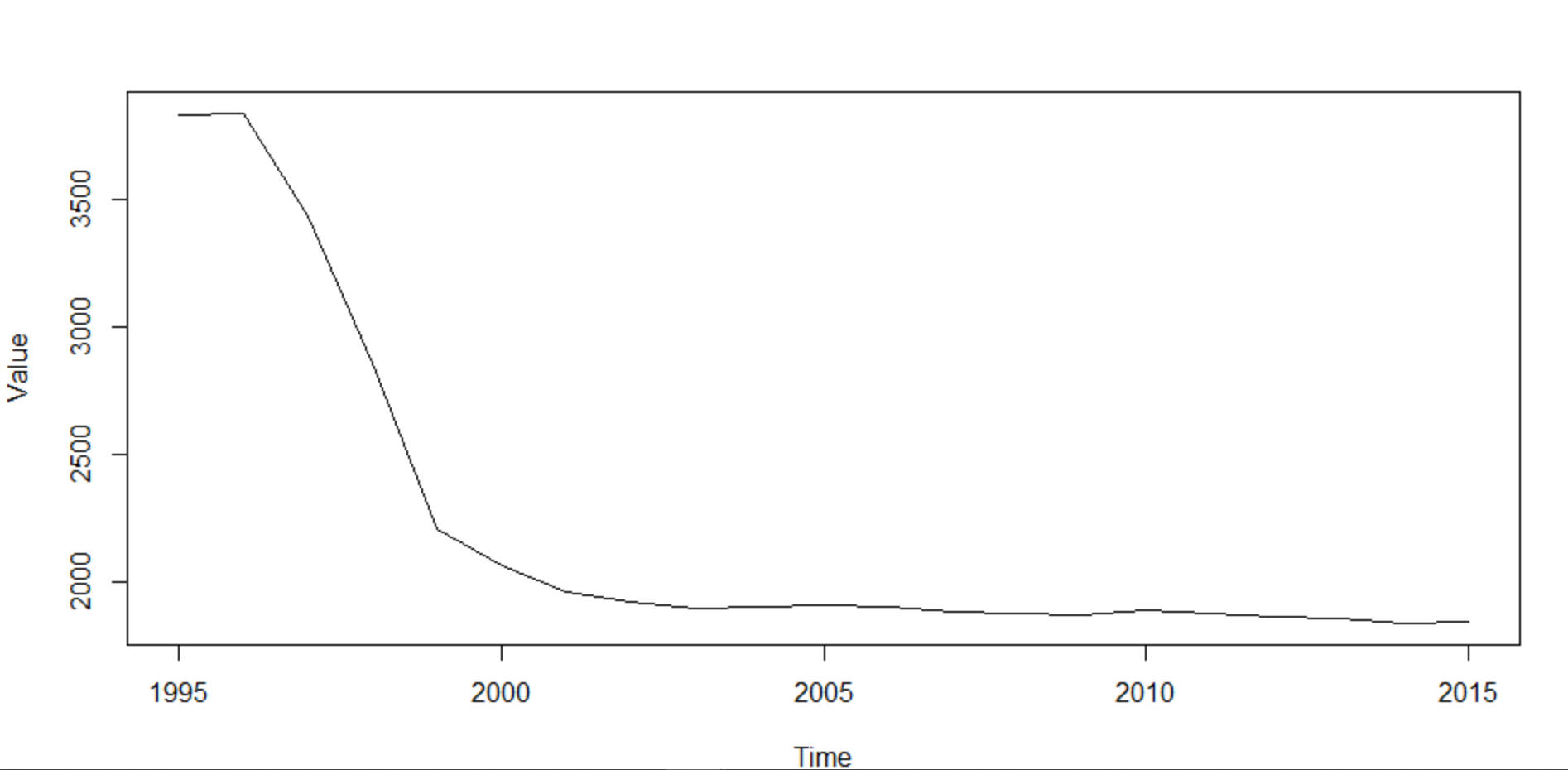
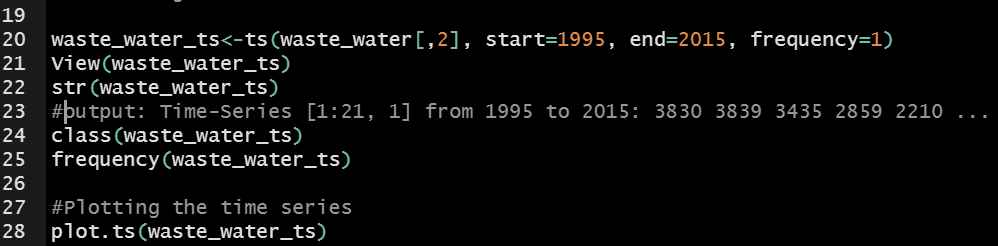
The time series mainly consists of four components which are as follows: the trend, the seasonal variation, the cyclical variation, and the irregular variation. Based on all these the pattern of this dataset has been verified after loading the data into R as shown below.



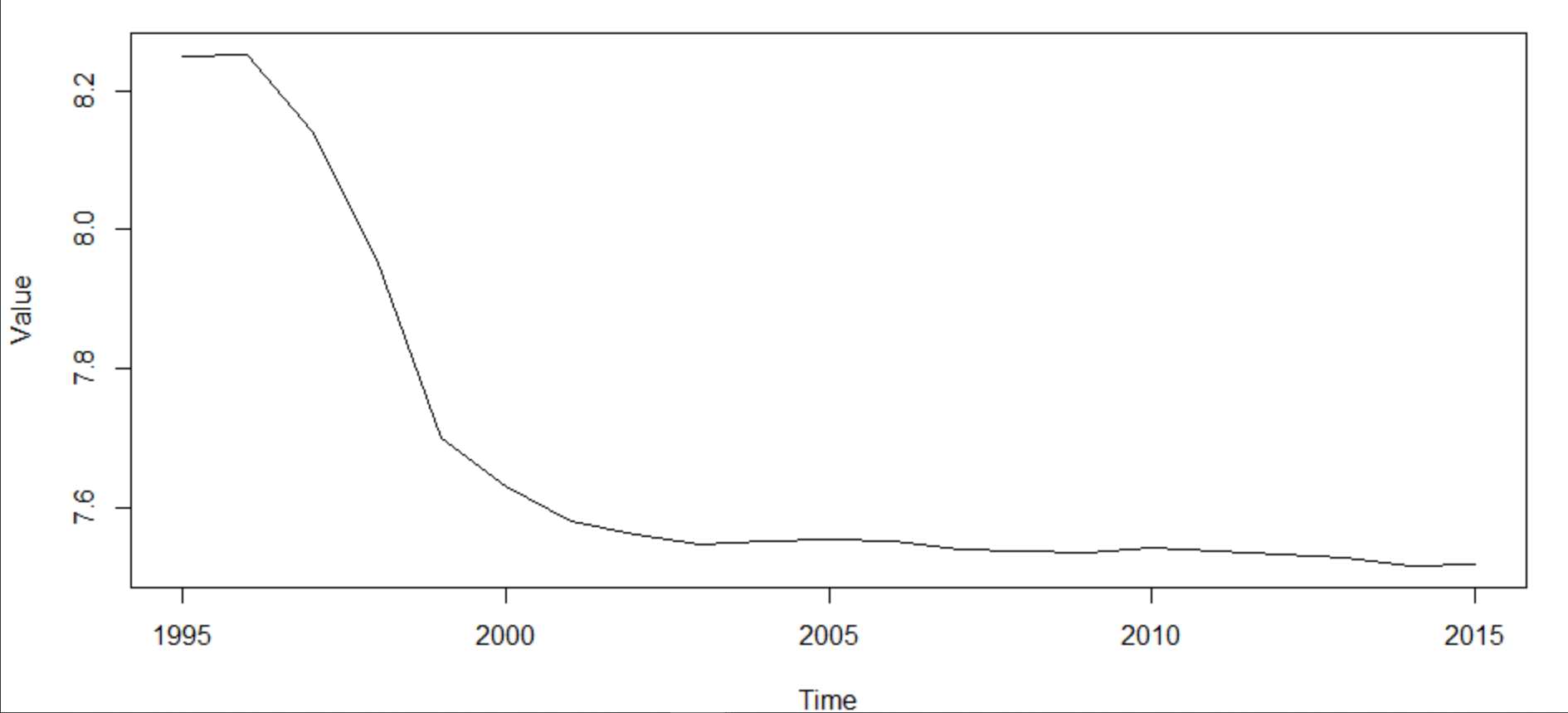
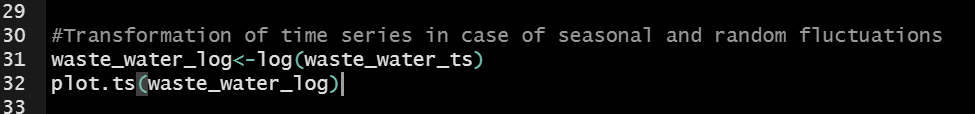
We could see that the data is loaded successfully, and the graph is generated with the year and value.

**Verifying the Data is stationary or Non-stationary:**

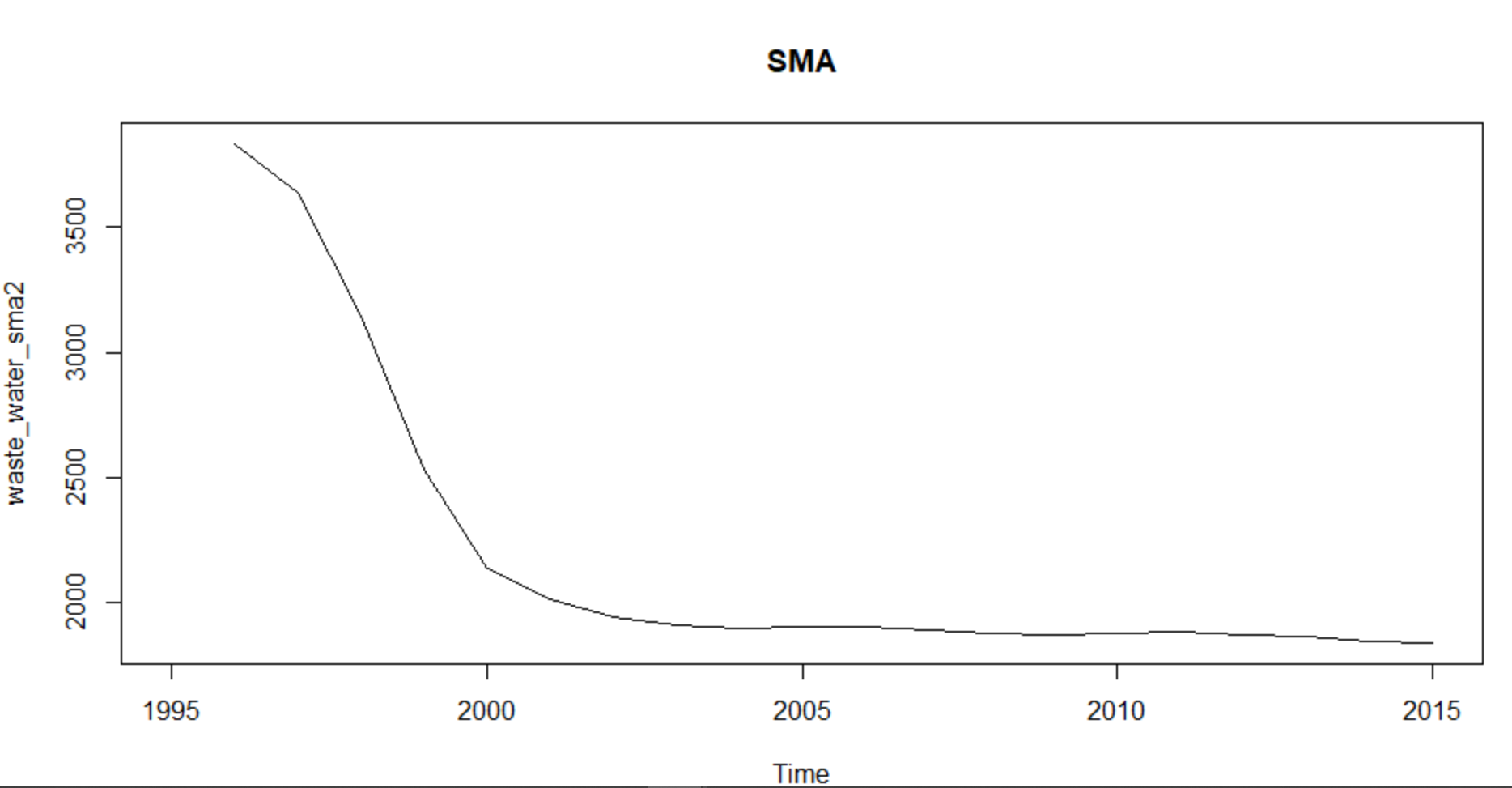
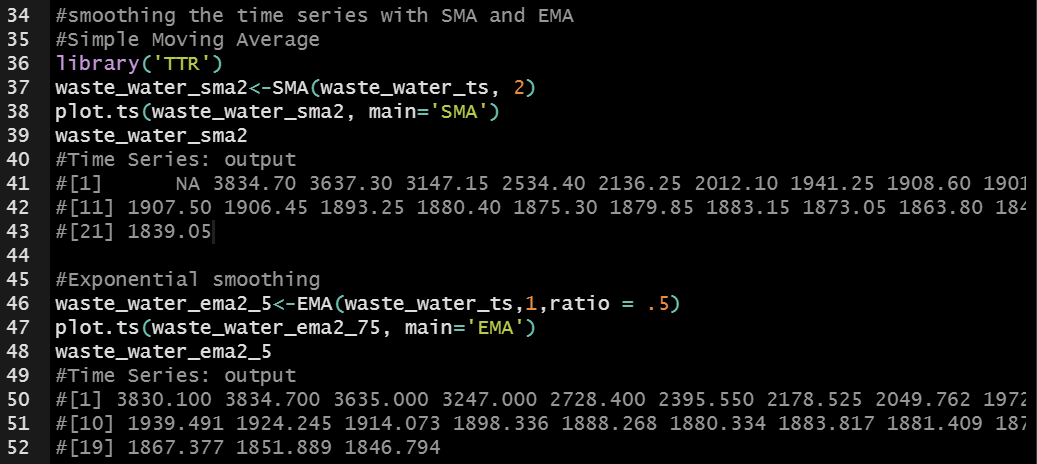
To verify the stationarity, the loaded data is converted into yearly time series data by using the function ts() in R as shown below.

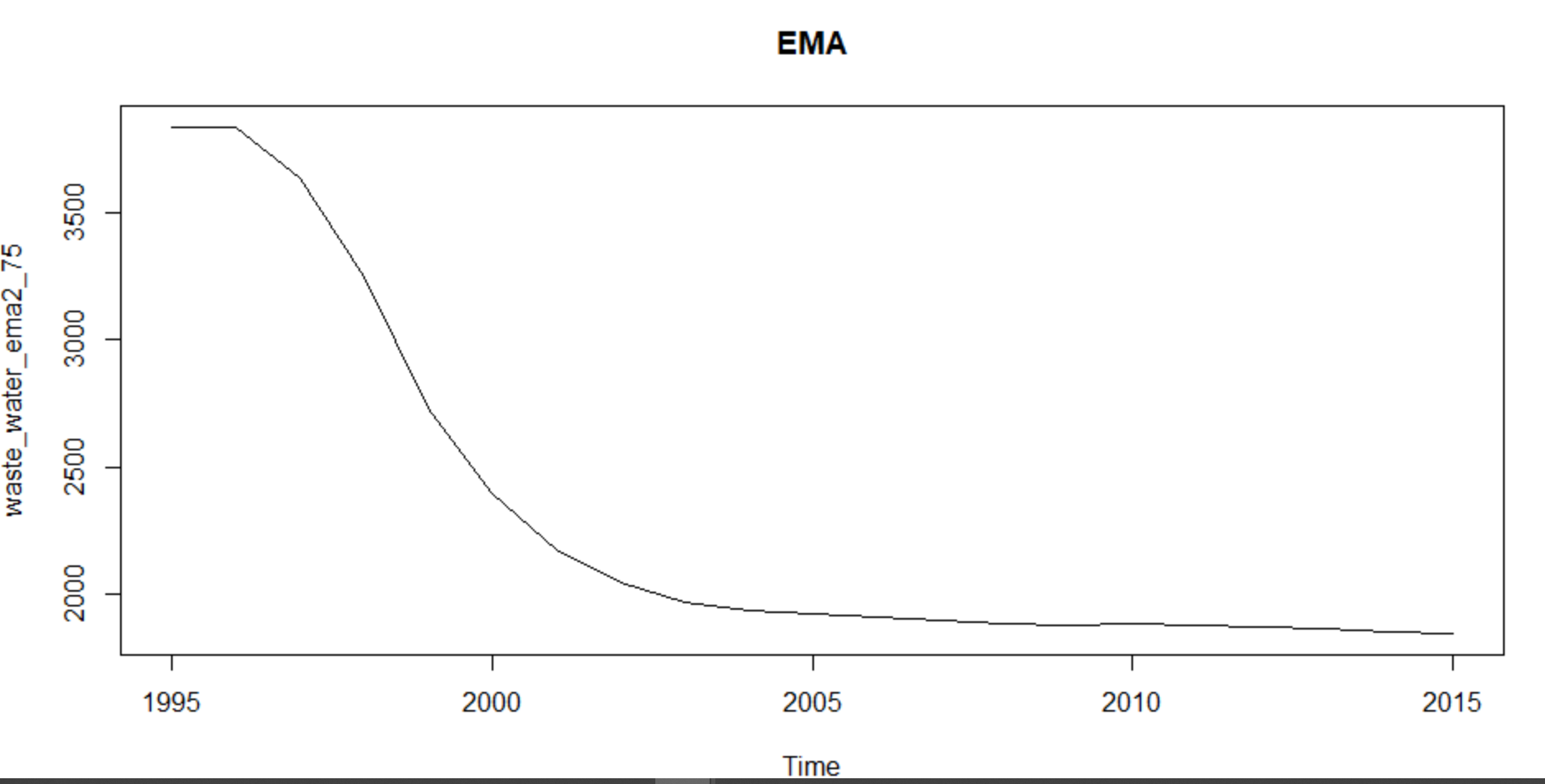


From the above graph, we could see that data is plotted in the form of time series from 1995 to 2015. We can see that over a period the data appear curvilinear. To forecast, the data should be stationary this can be done by removing the trend, seasonality, and irregularity if exists. In order to remove these, we need to use the log transformation on the time series. At first, the lm() function is applied to the time series data as shown below in R for the transformation.

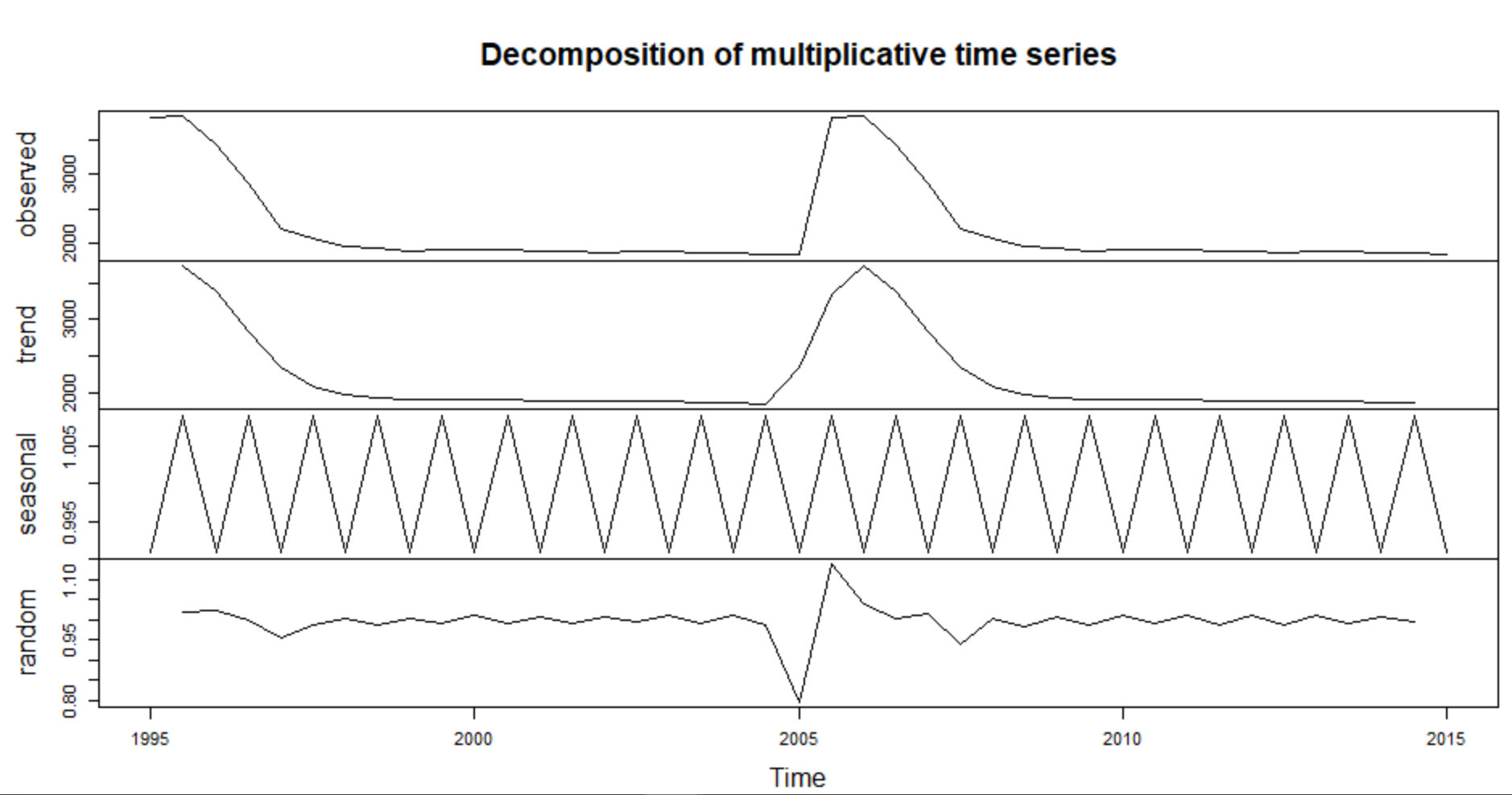
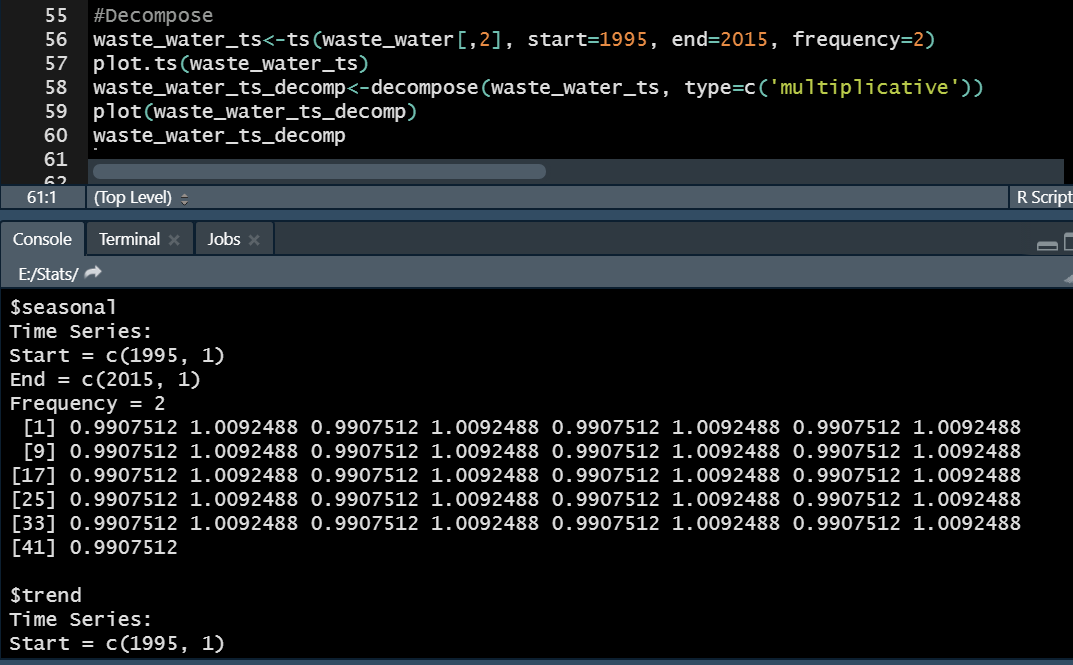


We can see in the graph above still there is some nonlinear trend in the data. To check whether the trend component present in the time series or not we need to smooth the series. In the smoothing process, one is Simple Moving Average (SMA) by taking the number of last periods and applying on the time series. The second is Exponential Moving Average (Weighted Moving Average) this can be applied by taking the last periods along with the ratio. In other words, it provides the weight to the most recent observation in the smoothing process. This ratio can be more than 0 and above 1. In R there are functions called SMA() and EMA() which are from the library TTR to perform, as shown in the below screenshots.





From the above graphs and the output on the R console screen, we could observe that the starting points are removing and there is not much change in the pattern, this means that there is no trend or seasonality present in the data. To confirm this behaviour, we need to check by decomposing the data. Since it is a yearly data the decomposition function cannot work on this, so we need to increase the time series frequency with one value. Below are the outputs.



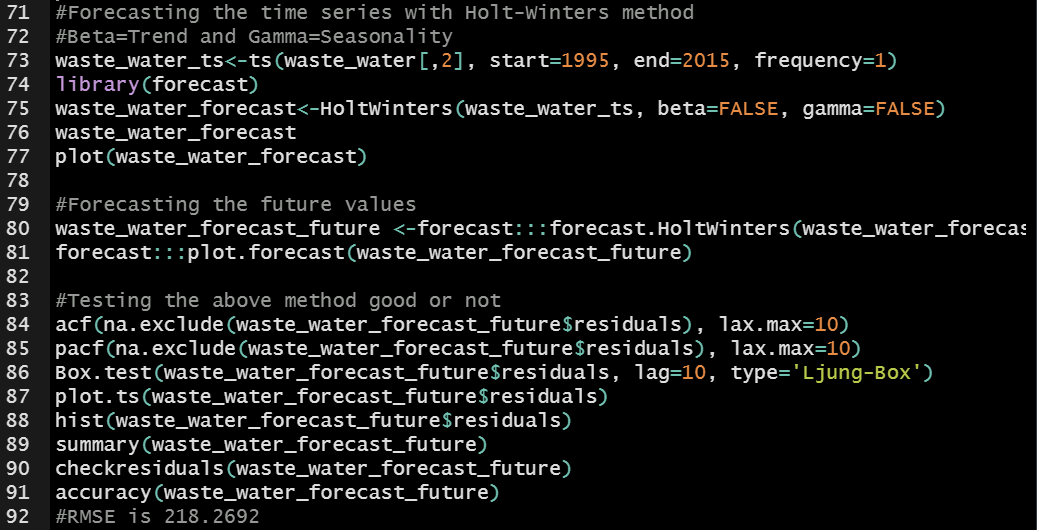
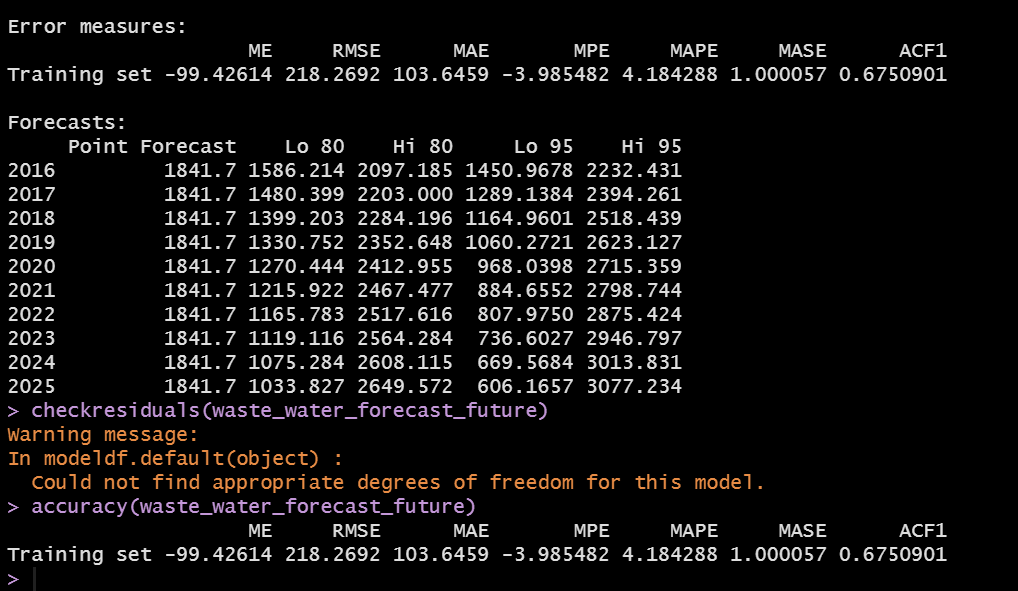
In the above output and graph, we can clearly see that there is no trend, seasonality, and irregularity in our time series. With this, we can confirm that the data is stationary and can apply the forecasting methods.

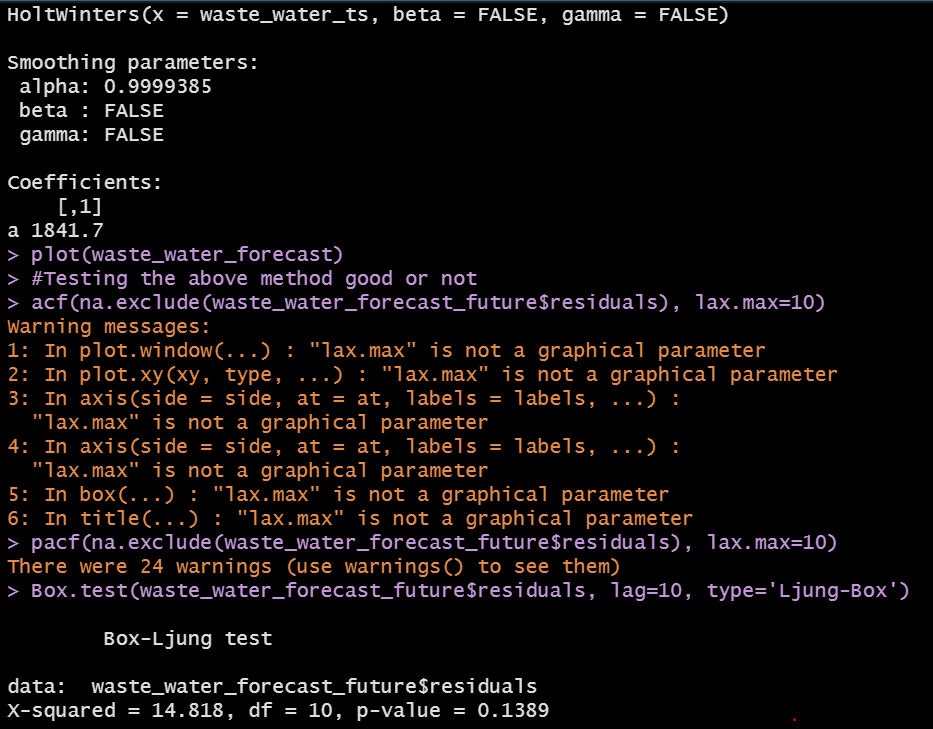
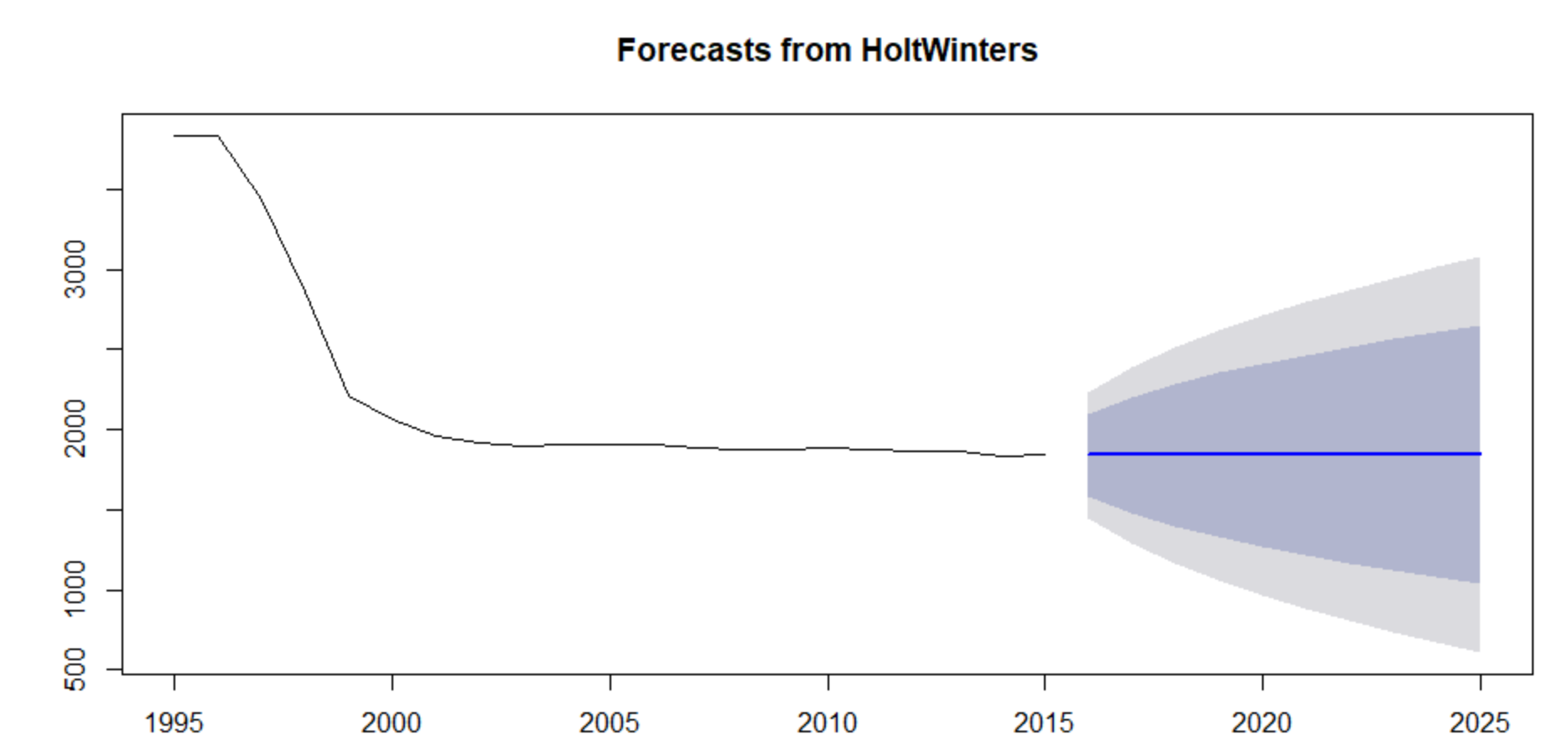
**Forecasting and Evaluating the best model:**

The methods which are used as part of forecasting process on this dataset are Holt-Winters smoothing method and ARIMA.

**Holt-Winters method:**

It is one of the most used amongst the smoothing models. In order to do this, we need to make sure that beta and gamma are given as false because we already observed that the data is stationary and there are no trends and seasonality. Below are the R code, forecasting output values, and the graph.

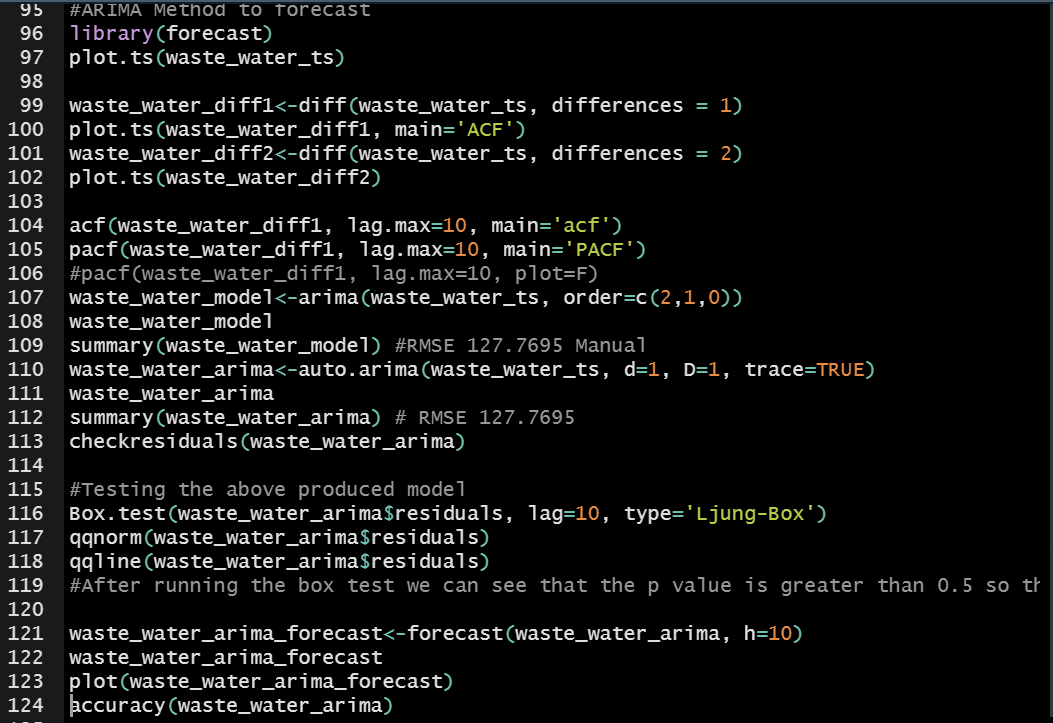
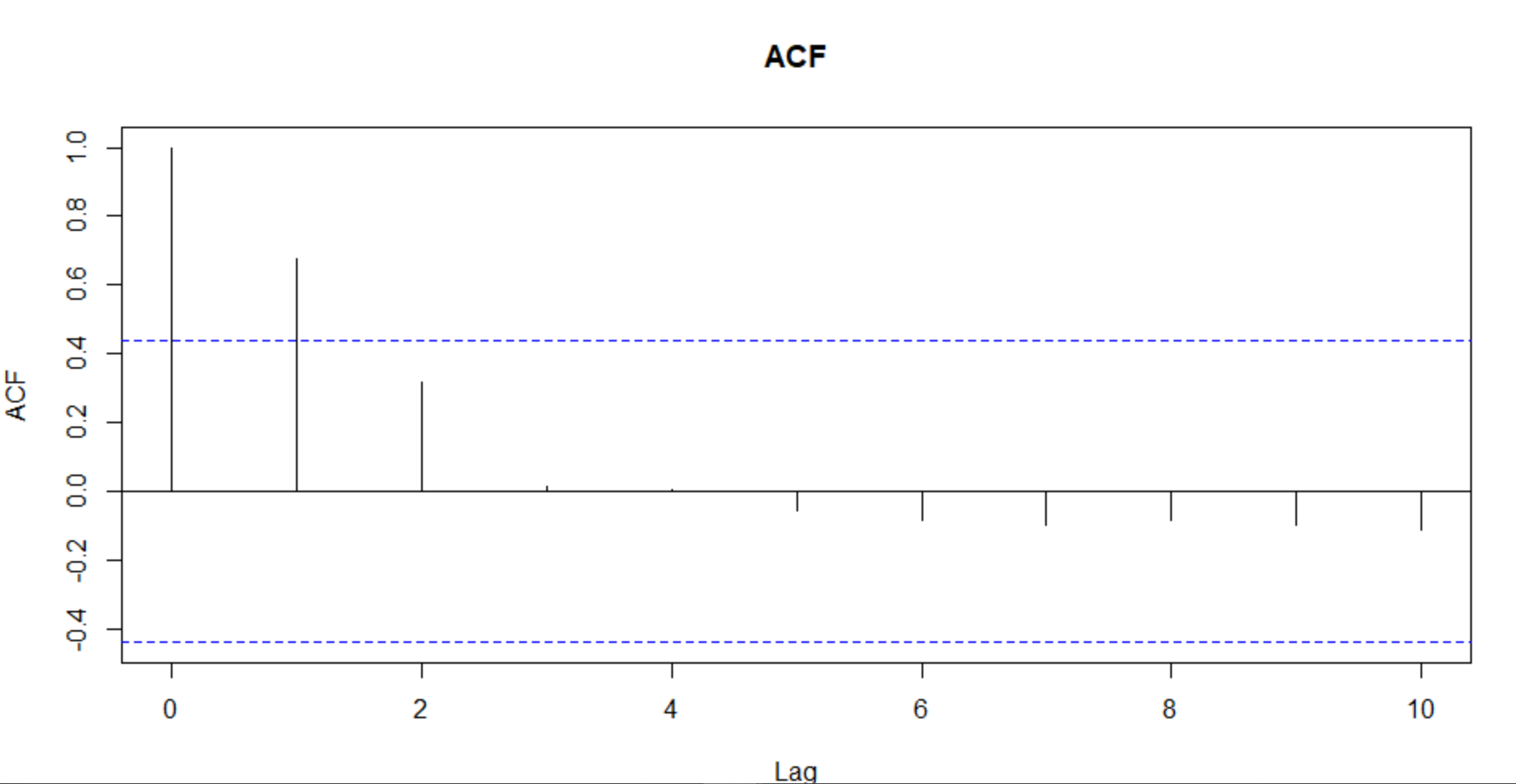
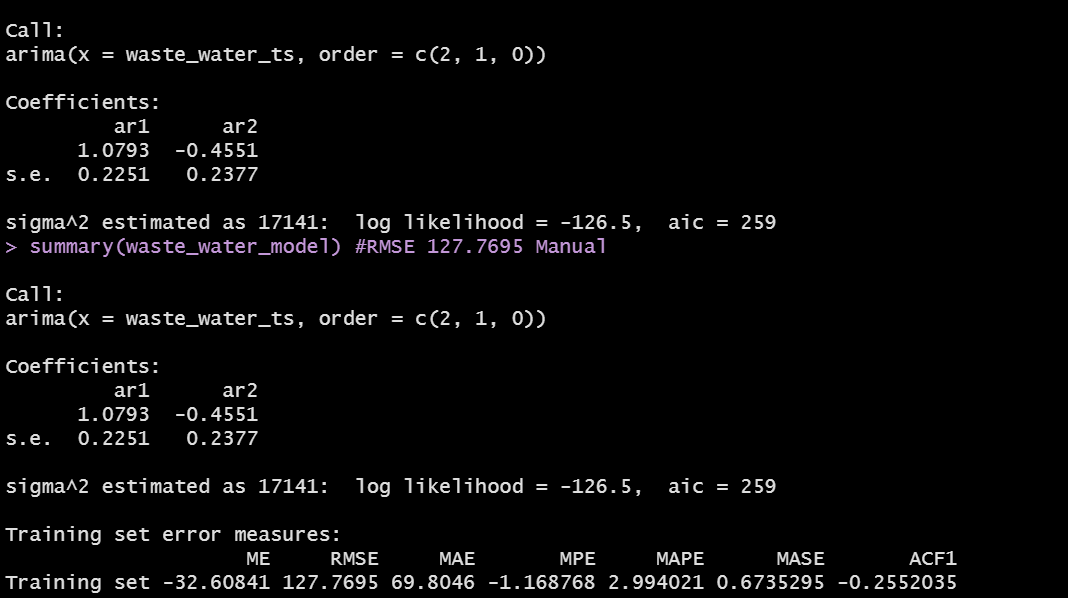
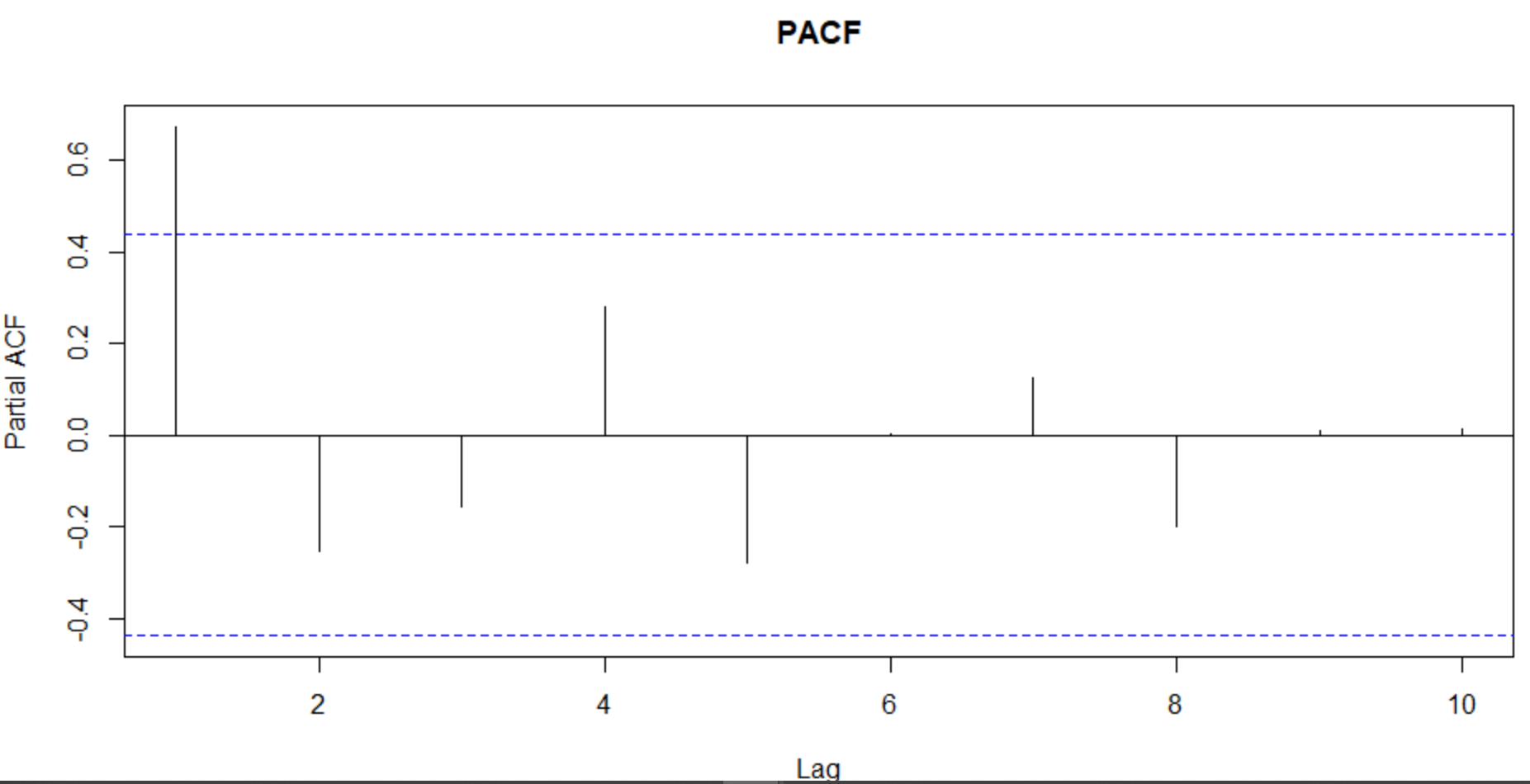
 

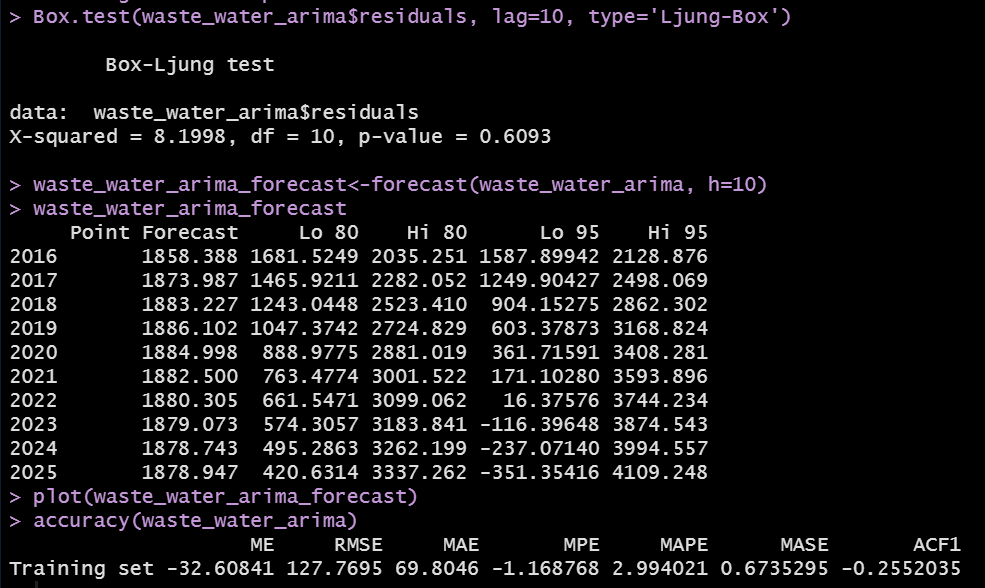
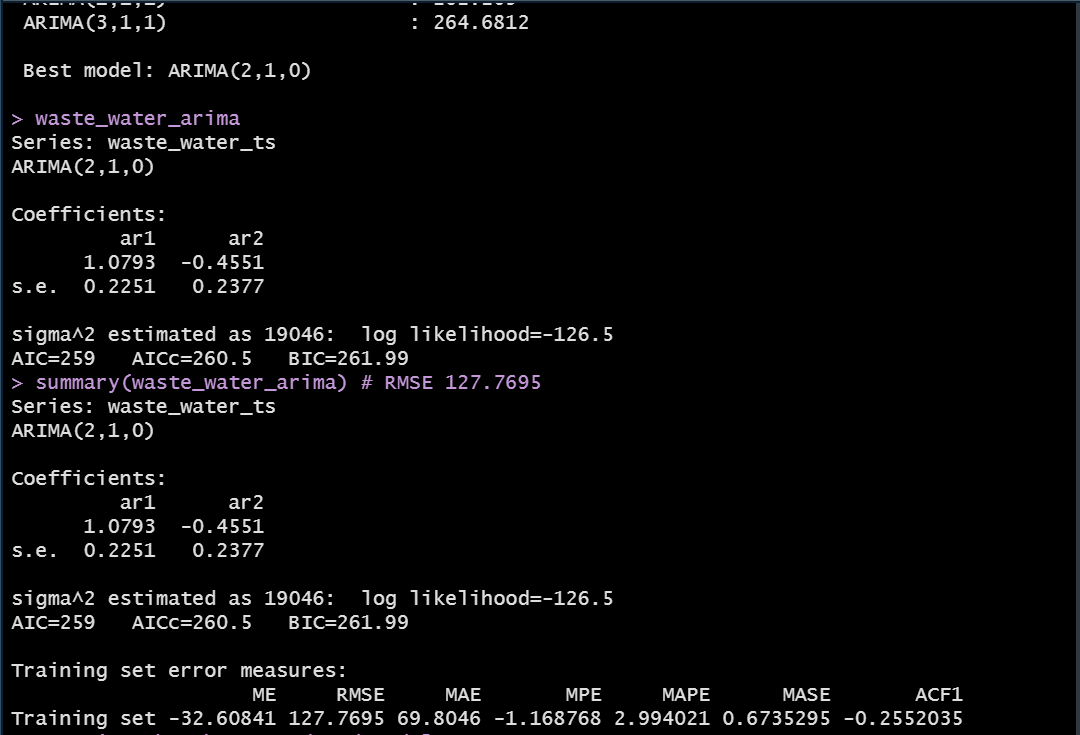
 

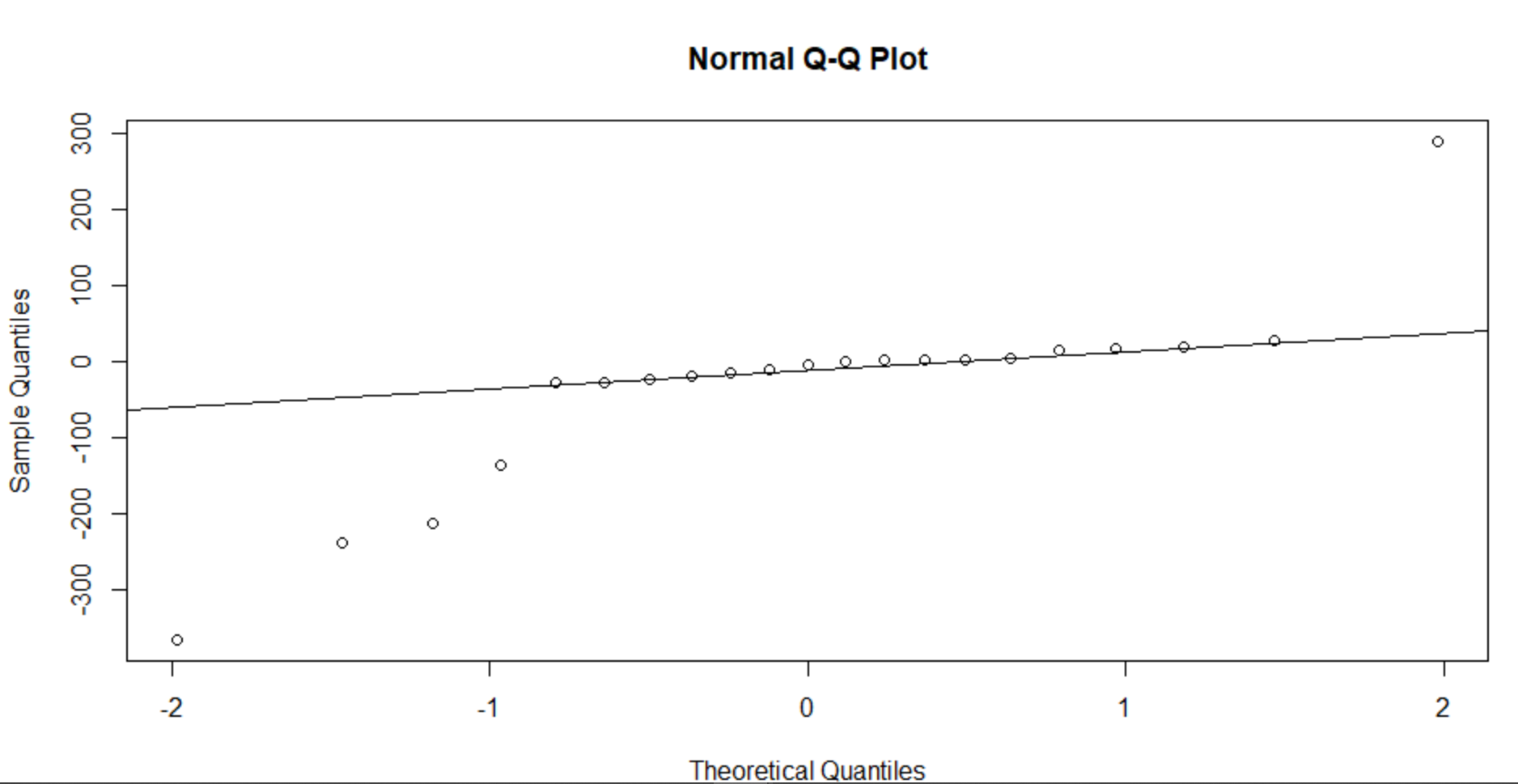
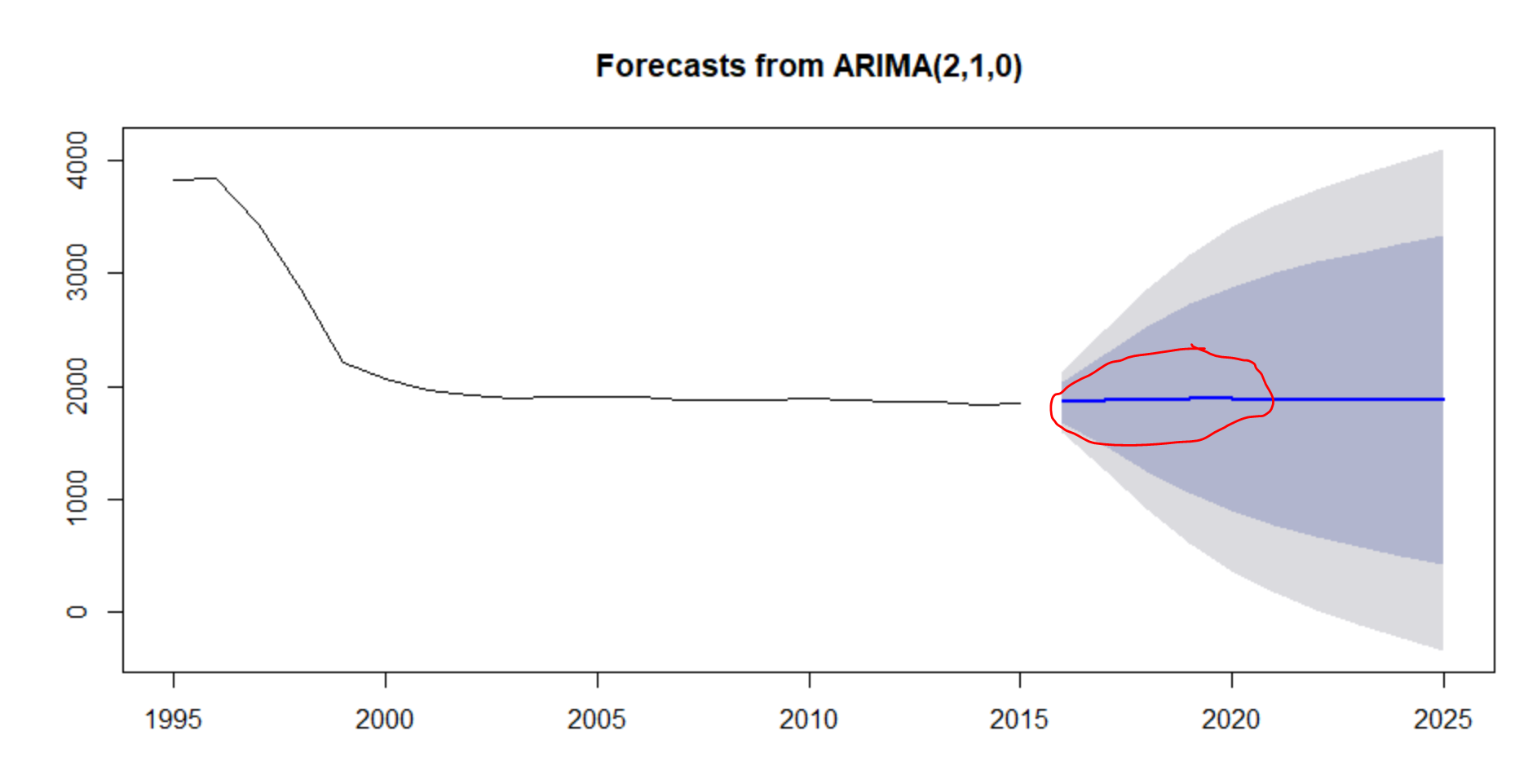
At first, the holt-winters method is applied to the time series data for smoothing it and then the forecasting function is applied to it. From the output console screen and graph, we could see that the forecasted values of the holt-winters method from 2016 to 2025 are the same (1841.7) in between Lo 80 to Hi 95 intervals. Second, to check whether the residuals of this model looks like white noise or not, **acf, pacf, and Ljung-Box** tests are applied to the forecasting output. From the output console screen, we can say that P-value is less than 0.5 and there is some correlation between the residuals After observing the output of all these we could clearly say that this method is not forecasting the values accurately. The RMSE value of this method is **218.2692** which is the accuracy of the model.

**ARIMA Method:**

The ARIMA stands for Auto Regressive Integrated Moving Average. It is one of the finest methodologies for doing the time series forecasting. Moreover, it works based on the assumption that over a period of times the current values are related or correlated with their immediate previous or n previous values. The model is classified as an **ARIMA (p, d, q)** for non-seasonal ARIMA where ‘p’ stands for the number of autoregressive terms, ‘d’ stands for the lagged difference between current and previous values needed for stationarity and ‘q’ stands for the lagged forecast errors. This analysis is performed by both manually identifying the (p, d, q) values and by using the auto.arima() function. Below are code, forecasting output values, and the graphs.



Firstly, time-series data is applied with the difference of 1 using the function diff(). Then the acf and pacf functions are applied to the difference 1 data as shown in the above code. Secondly, from the acf graph, we can clearly see that there are two spikes and it is gradually decreasing whereas in pacf there is only one spike, this means that it is an **AR(1)** signature. By observing the acf and pacf graphs at first the arima model is applied manually using the order **(2, 1, 0) as (p, d, q)** values on the data and the RMSE value obtained from this model is **127.7695**. To verify the model which we applied is correct or not, the auto.arima() function is used on the data and it will generate the best model output by comparing all the possible models as shown in the above output screenshot. The auto.arima() function gives the best model as **(2, 1, 0)** and the RMSE of this is **127.7695**. We could observe that the output of both the manual order model and the auto arima model is the same. From the output graph of **Normal Q-Q plot** we could see that the residuals are normally and independently distributed with mean zero and there is no relationship between them**.** Thirdly, to check the residuals of the model looks like white noise or not, the **Ljung-Box** test is applied. It says that the arima model appears to fit the data well. On the other hand, the P-value of the Ljung-Box test is greater than the 0.5 this means that it clearly says that there is no correlation and we can safely use this model for forecasting. At last, the function **forecast()** from the library forecast in R is applied to the residuals of the arima model to predict the next 10 years ahead. We could see in the output console screen the forecasted values are not same and it varies from year to year in between the confidence intervals **Lo 80 to Hi 95**, and the forecasting graph also plot the same where it is following the previous patterns which are shown by highlighting with the red mark.

**Conclusion:**

To conclude, after examining the different time series models two of them were selected which are Holt-Winters smoothing model and ARIMA regression model. This case study was based on the total wastewater generated data obtained from the UNdata website and included data for more than 10 years. To evaluate the efficiency of prediction the RMSE parameter was used. The obtained results show that the ARIMA model has the least value than the Holt-Winters, this means that the ARIMA model was significantly better in sense of the prediction accuracy.