Statistics for Data Analysis

Time Series and Binary Logistic Regression

David Casado   
*Statistics for Data Data Analytics  
National College of Ireland  
Dublin, Ireland*  
*x19204221@student.ncirl.ie*

*Abstract*—this study is composed by two independent parts which will offer two different statistical studies. Part A will focus on time series and forecast prediction on two datasets, ‘OverseasTrips’ and ‘NewHousesRegistrations’ from the Central Statistics Office website. Part B will present a study using Binary Logistic Regression that will include different aspects like the descriptive analysis of the variables contained on the dataset ‘ChildBirths’ to the final binary regression model build.

Keywords—Time series, decomposition, ARIMA, exponential smoothing, ETS, Binary Logistic Regression, correlation, AIC, prediction

# PART A – OverseasTrips

## Introduction

OverseasTrips data set represents the number of overseas trips to Ireland made by non-residents on quarterly basis starting from Q1 2012 and finishing on Q4 2019. Each observation in the dataset represents the total of trips in thousands at the end of each quarter.

The objective of this study is to design a time series forecasting model that can extract information from historical information of the series and predict future behavior of the time series based on past patterns.

## Exploratory Analysis

Data preparation. No outliers or missing data.

Library fpp2 will be installed which will install packages forecast v8.14, ggplot2 v3.3.3, fma v2.4 and expsmooth v2.3 which contains the main functions required for time series analysis [1]

Data integrity of the file s checked after is loaded into the dataset called *OverseasTrips*. No outliers or missing values are observed in the data. Data values are displayed in correct time order and there is not data missing. Therefore, data is considered to be ready to start the analysis and it contains observations representing the total number of trips to Ireland made by non-residents from Q1 2012 to Q4 2019.

Figure 1 represents the time series and it will help to obtain a better understanding on how the data changes over time. It can be observed that the total number of trips is increasing in numbers every year although it fluctuates during the year. The time series of *OverseasTrips* have a positive trend and a strong seasonality. Seasonality pattern is regular with highest peaks on Q3 on each cycle of the season and drops or valleys in Q1. These peaks and valleys data points are predictable. In addition to seasonality, the time series has an increasing or positive trend. The red line represent the predicted values of the linear model with the time as independent variable.

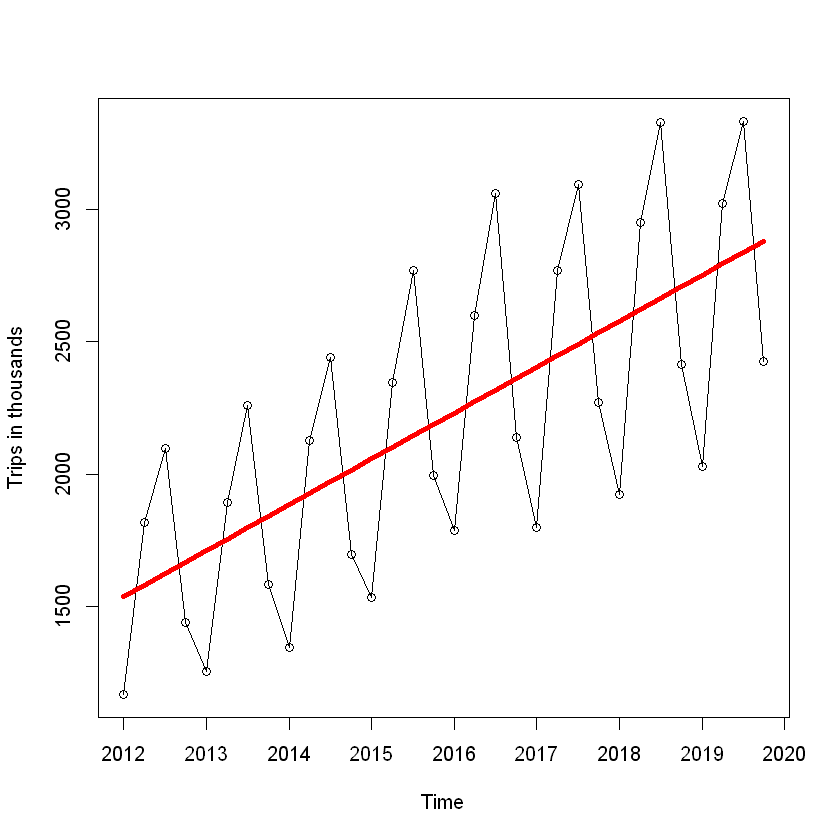


Figure I‑1- OverseasTrips data plot

In order to obtain a model valid for forecasting, the time series needs to be stationary. This means that the data over the time has a constant mean and that the variability of the time series is constant over time. Trend and seasonality will be considered when choosing the forecast method.

## Data Decomposition

Decomposition of a time series leads to identification and extraction of the individual components. Primary objective of decomposition is to study the components of the time series, not forecasting. However, a forecasting model will be built on top of the decomposed series.

In the time series, each observation or data point can be expressed as a sum or a product of 3 components: Seasonality (St), Trend (Tt) and Error (Et). Decomposition methods are useful to analyze the trend a seasonal pattern in a time series. The objective of the decomposition is to disintegrate the components of the time series into the trend component, seasonal component and the error or reminder component. Most popular methods of decomposition are multiplicative and additive. Additive decomposition assumes that the seasonal fluctuations are constant and do not depend on level of the time series. In the other hand, the multiplicative method is recommended when the seasonal fluctuations increases accordingly to the level

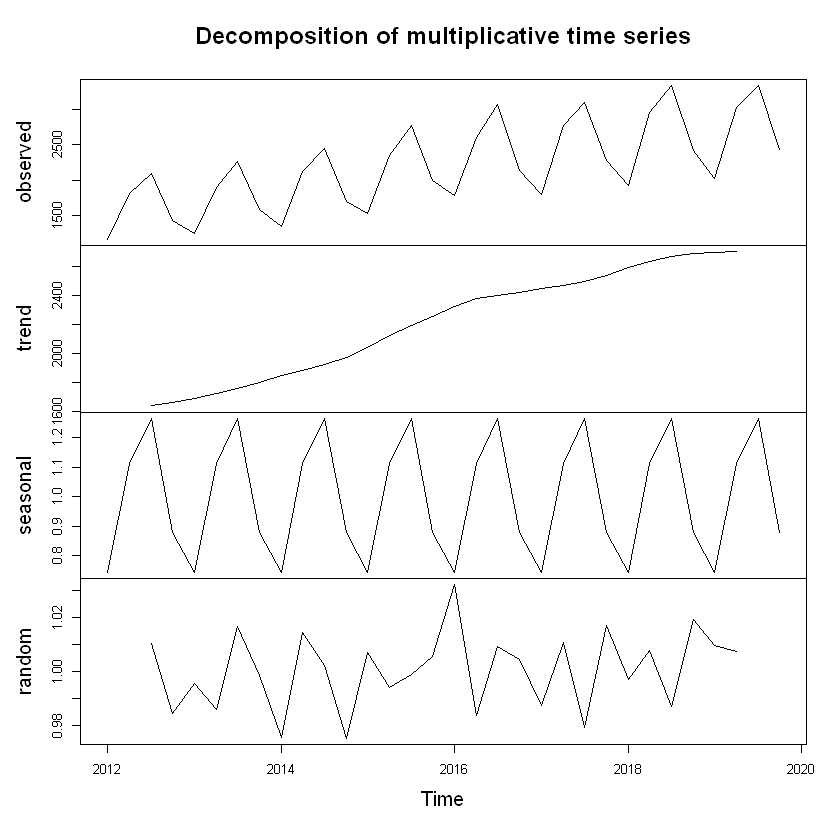


Figure I‑2 Decomposition of Multiplicative Time Series

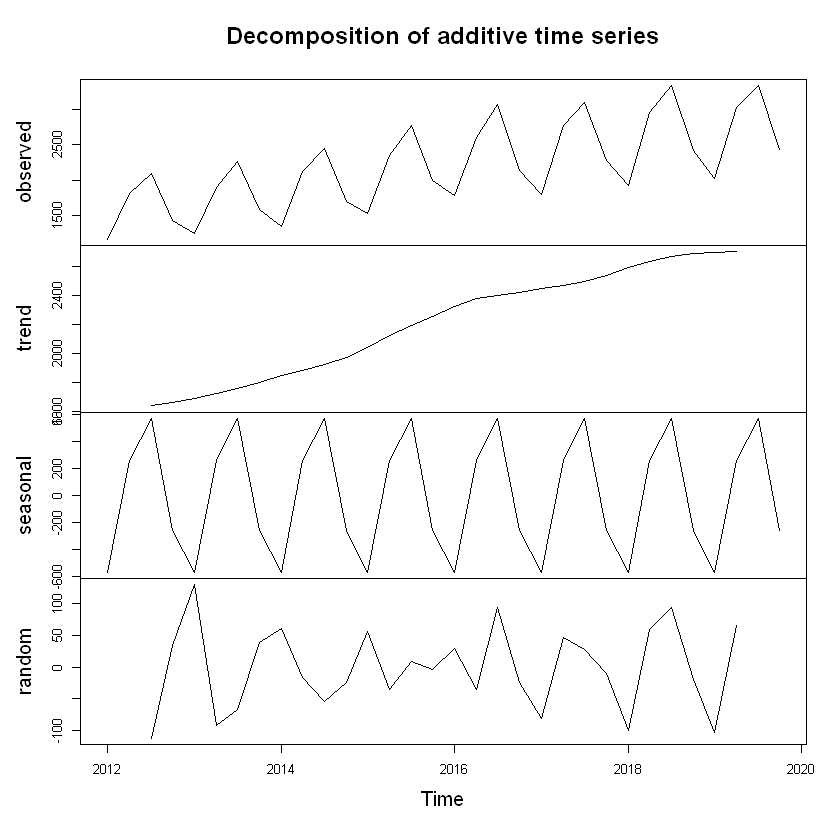


Figure I‑3 Decomposition of Additive Time Series

Before applying a statistical model to a time series it needs to be stationary. When the data has trend and seasonality, there could be a combine effect between the autocorrelations for small large for small lags and seasonal lags [1].

The fact that the seasonality variation increase over time lead to believe that a multiplicative model could be more suitable on this instance. In addition, a multiplicative model can be considered equivalent to an additive model of the logarithmic time series.

## Exponential Smoothing

One of the most popular and accurate forecasting methods is Exponential Smoothing introduced in 1950 by Brown, Holt and Winters [1]. A smoothed version of the *OverseasTrips* curve which damp down fluctuations can help to distinguish any pattern in the data. Using exponential smoothing methods is equivalent to use weighted average of past observations.

The construction of a Simple Exponential Smoothing or a Holt Model it is not recommended on this instance as they can only deal with time series with no clear trend or seasonal pattern on the first and only with trend but no seasonality on the latter. As previously mentioned, the *OverseasTrips* time series have both trend and seasonality and a method that can deal with both need to be selected for forecasting. An extension of Holt Method is the Holt-Winters Method which enables the model to be able to deal with seasonality in addition to trend and level.

Figure 4 represents two forecast implemented using two variations of the Holt-Winters Method: Additive and Multiplicative.

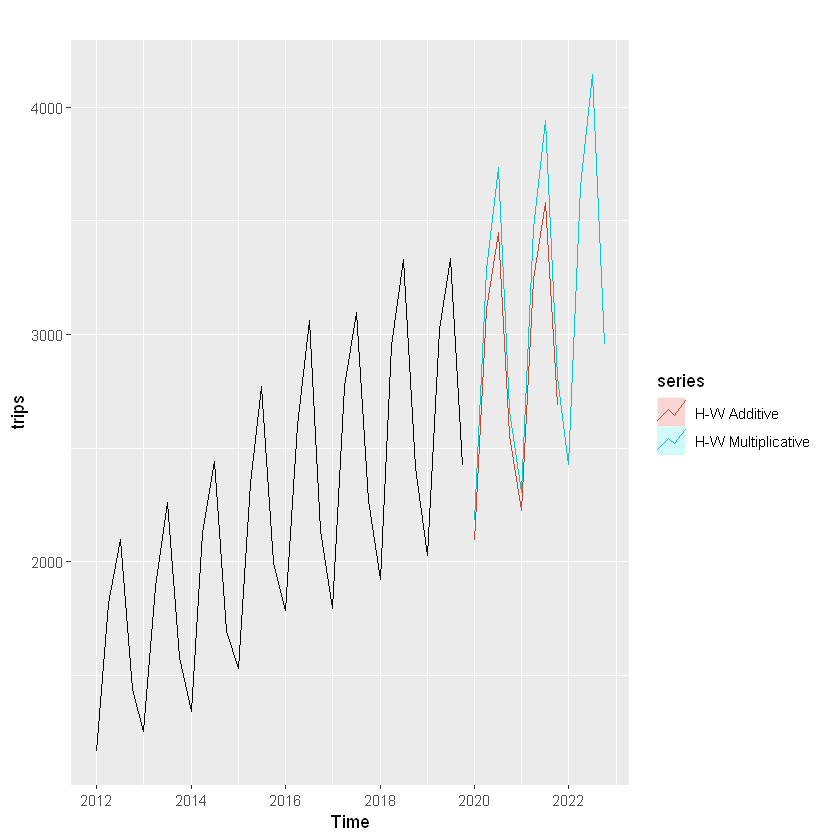


Figure I‑4 Comparison between Holt-Winters Additive and multiplicative methods.

It can be observed that both representations have small differences on upper and lower values while they are similar in between. The residuals of both models will give more insight on which of both models is better choice for forecasting. The mean value of the residuals for the additive model is 63.38 (absolute value) while the mean value of the multiplicative model is 0.027 (absolute value). The additive model has a AIC=406.57 and RMSE=76.60 and the multiplicative model has a AIC=410.07 and RMSE = 72.4. Furthermore, the multiplicative model has a smaller mean value for the residuals which in addition to small RMSE make the multiplicative option a slighter better one. This will be equivalent to run function ETS(M,A,M) in R which will return the same value for the mean of residuals of 0.02 which can be considered as marginally better. In addition, it has a lowest value for AIC 378.82 obtained yet which can lead to a conclusion that it will be a strong candidate for final forecast.

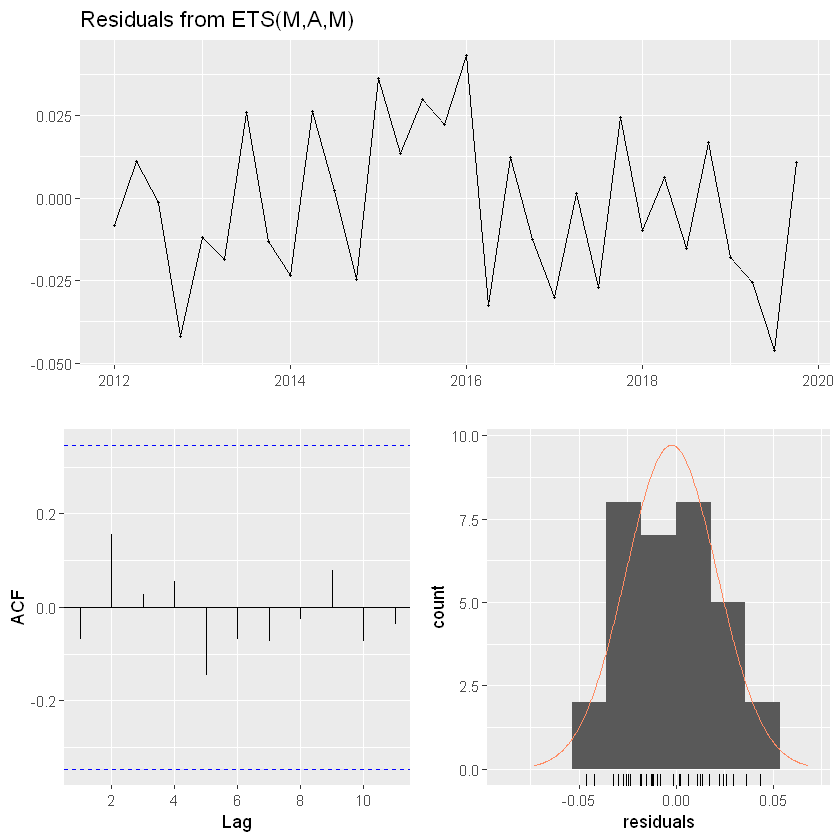


Figure I‑5- Residuals for ETS(M,A,M)

ETS(M,A,M) will deal with errors with a multiplicative approach, trend with additive and finally, seasonality with a multiplicative type. Figure 5 shows the residuals for the model. The fact the correlation do not pass the blue dotted line indicate that this are not significant different from zero. This can be confirmed with the Ljung-Box test that, which with a p-value = 0.376 (df=3, lag=11), confirms that the autocorrelation between residuals do not significantly differ from zero and that these can be considered as white noise.

ARIMA Model

The results that have been obtained previously using ETS function are reasonably good as ETS focus on trend and seasonality of the time series. On the other hand, ARIMA models (Autoregressive Integrated Moving Average) uses the autocorrelations in the data as approach to forecast time series [1]. The ARIMA model will need to be able to deal with the seasonality that there is present on the dataset. For this instance the ARIMA model will be written as follows:

Where *m* will represent the seasonal period (4 in this instance) and *(p, d, q)* will look after the Non-Seasonal part of the model and *(P, D, Q)* after the seasonal part of the model [1]. D represents the seasonal differencing which in this case is equal to one. Differencing will assist to convert a non-stationary time series into a stationary time series by reducing or even eliminating the trend and seasonality by computing the differences between consecutives observations. By differencing with D=1 the time series will be become stationary. This is confirmed by the Augmented Dickey-Fuller Test equals to -49.412 with a p-value < 0.01. MA(1) factor deals with the forecast errors and it is the value of *q*.

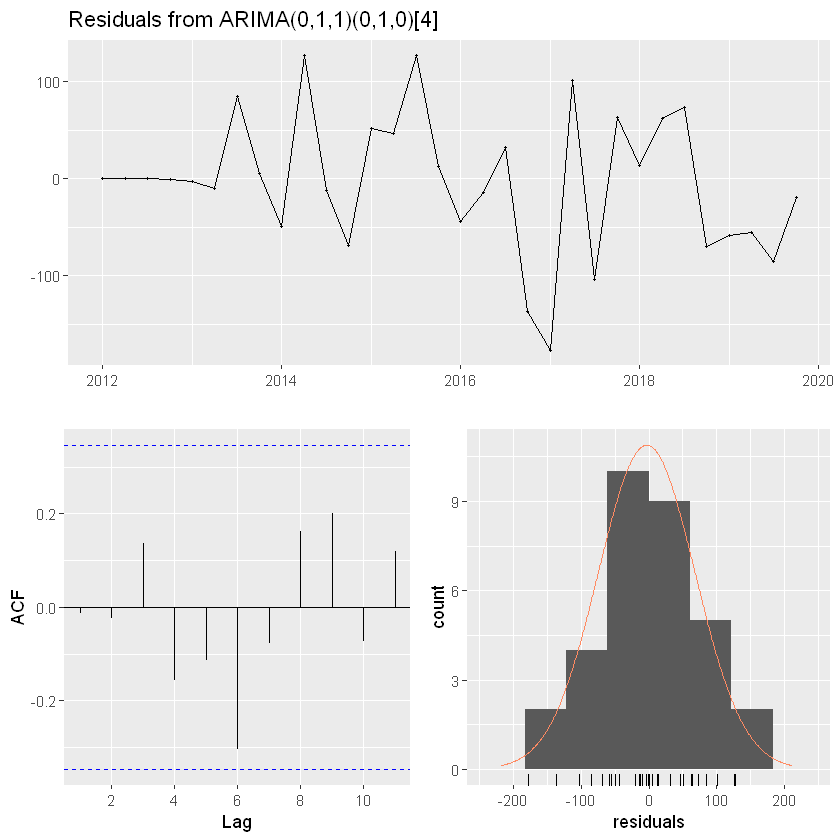


Figure I‑6 Residuals for ARIMA (0,1,1)(0,1,0)4

Figure 6 shows that the residuals for the ARIMA model are normally distributed and the ACF plot indicates that there is no correlation and the residual can be considered no significant different from 0 as the Ljung-Box test has a p-value = 0.301 (df=5, lag=6) which confirms that residuals are similar to white noise. Finally, this ARIMA model has an AIC = 314.98 which is lower that the previously calculated value using ETS function.

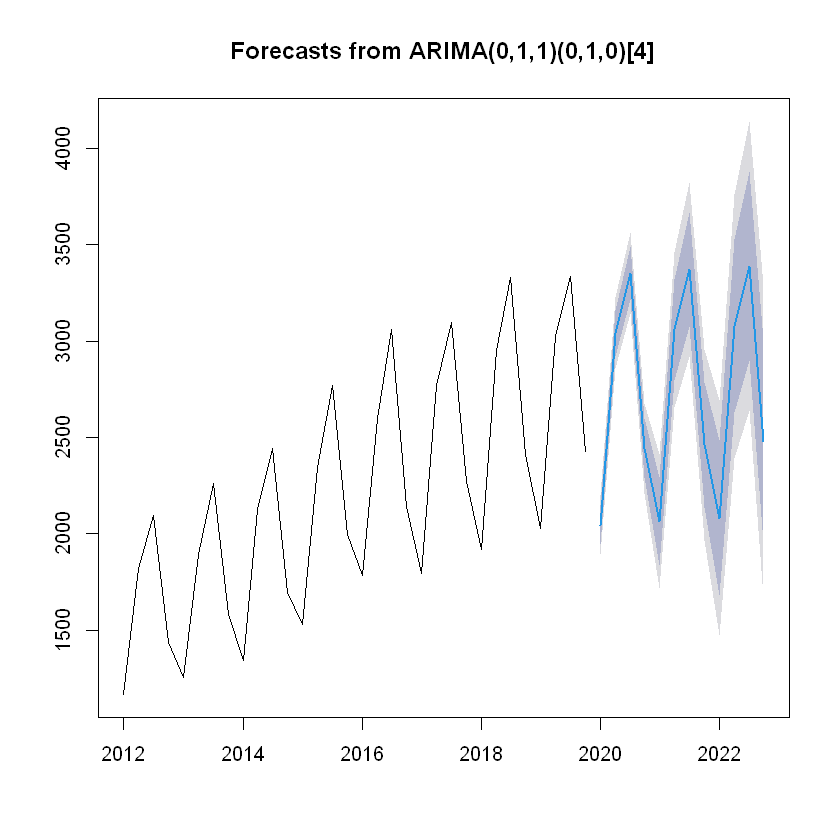


Figure I‑7 Forecast from ARIMA

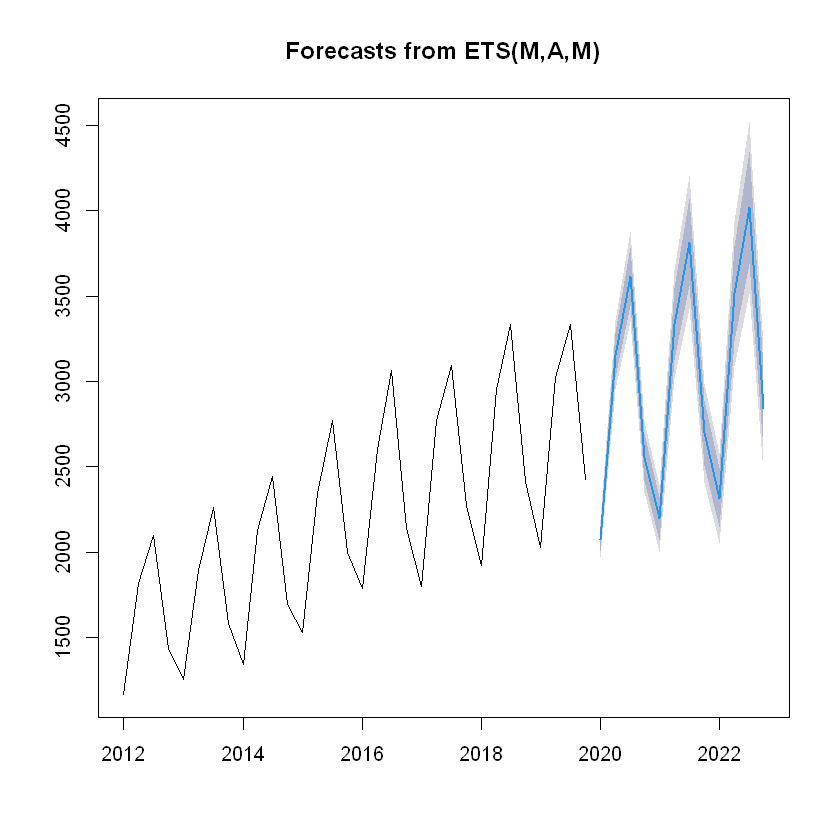


Figure I‑8 Forecast from ETS

RMSE value from ETS model is 54.69 while it is higher for the ARIMA model at 70.31. This difference in addition to the visual inspections of the corresponding forecast plots represented on Figure 7 and Figure 8 of both models lead to believe that the forecast model from ETS function will provide a slighter better forecast that the ARIMA version.

# PART A - NEW HOUSE REGISTRATIONS

## Introduction

NewHousesRegistration dataset represents an annual series of new houses registration in Ireland from 1978 to 2019. The time series has a frequency of one and each observation represents the total annual number of houses registered per year. A number of visualisations will be performed on this section in order to gain more insights hidden in the data. Then, the section will be completed by performing forecast analysis using different methods and comparing the results among them. Exploratory Analysis

Creating a time series plot will help to find pattern hidden in the data which will permit to undertand how the time series acted in the past and wether similar patter is expected in the future. Figure 9 represents the time series plot. The time series has an overall positive trend with a very strong fluctuation around year 2007 and from there a big and strong negative down point which have a “reset-like” effect on the time series as it goes down to a very low value similar to the start of the time series and then a positive trend starts again. At a glance, there a not seasonal pattern that needs to be dealt immediately although the presence of a cycle pattern will need to be investigated as there curve present up and downs that might indicate the presence of a cycle pattern.

The three highest values for Houses registered corresponds to years 2004, 2005 and 2006 with a total number of 60782, 62284 and 66649. There three observation are highlighted as outliers although will remain in the dataset for the analysis as they are considered legitimate observations.

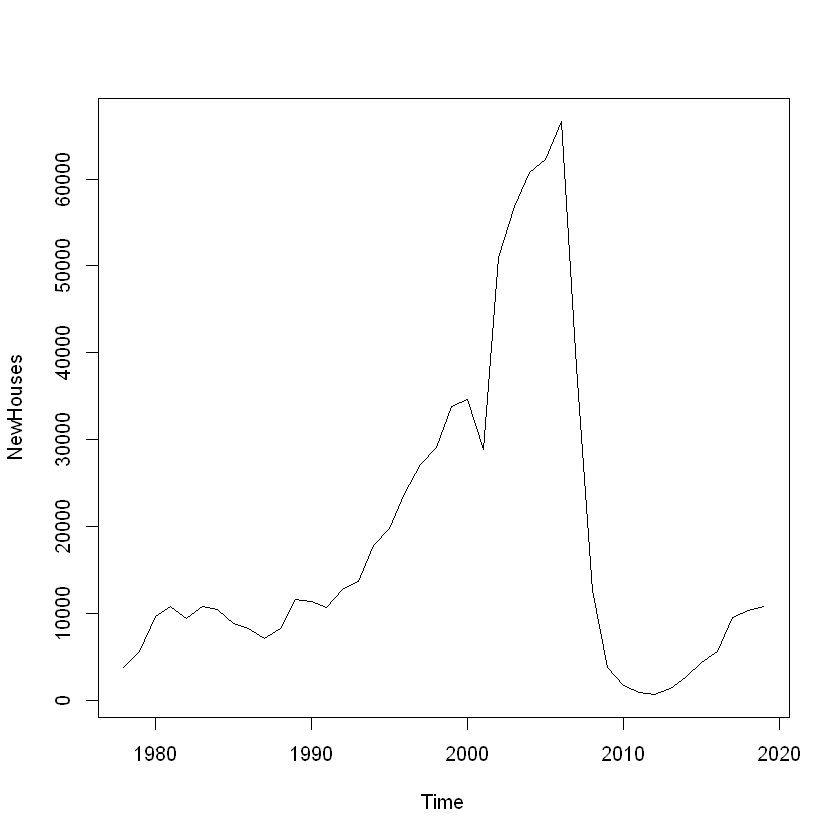


Figure II‑1 - New Houses Registration Time Series Plot

## Moving Averages

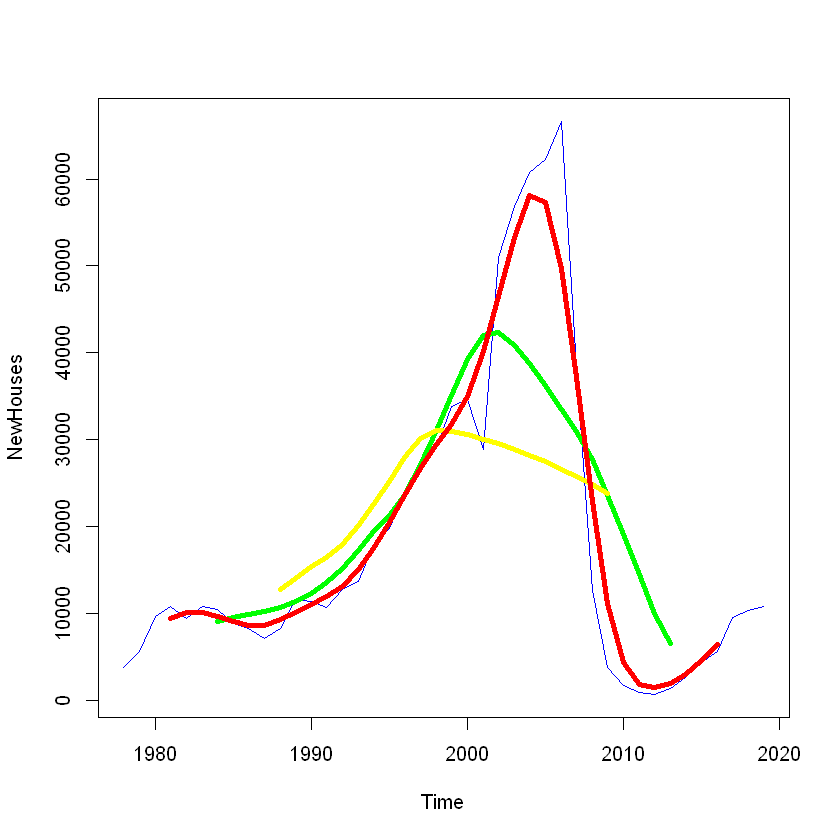
Time series like this have important irregular or error component which responds to events that are difficult to predict like strikes, economic recessions etc. In order to detect any pattern in the data a smoothed version of the curve will be plotted with the aim of damping down the fluctuations or changes. 

Figure II‑2 Moving Averages Curves.

Red: 2x4 MA Green: MA(12) Yellow: MA(20)

One of the most popular methods for smoothing a time series is Moving Average [1]. Moving Average of order q, MA(q), is a linear combination of the current observation with the q most recent observations. The result of applying Moving Average method on a time series is some missing points at the beginning and at the end to allow for the calculations.

## Forecasting

Some forecast models will be chosen on this section and the corresponding forecast predictions will be produced for a period of three years.

Naïve Method

This method is very simple as just takes all forecast predictions to be the same value of the last observation. This model works well when data follows a random walk [1].

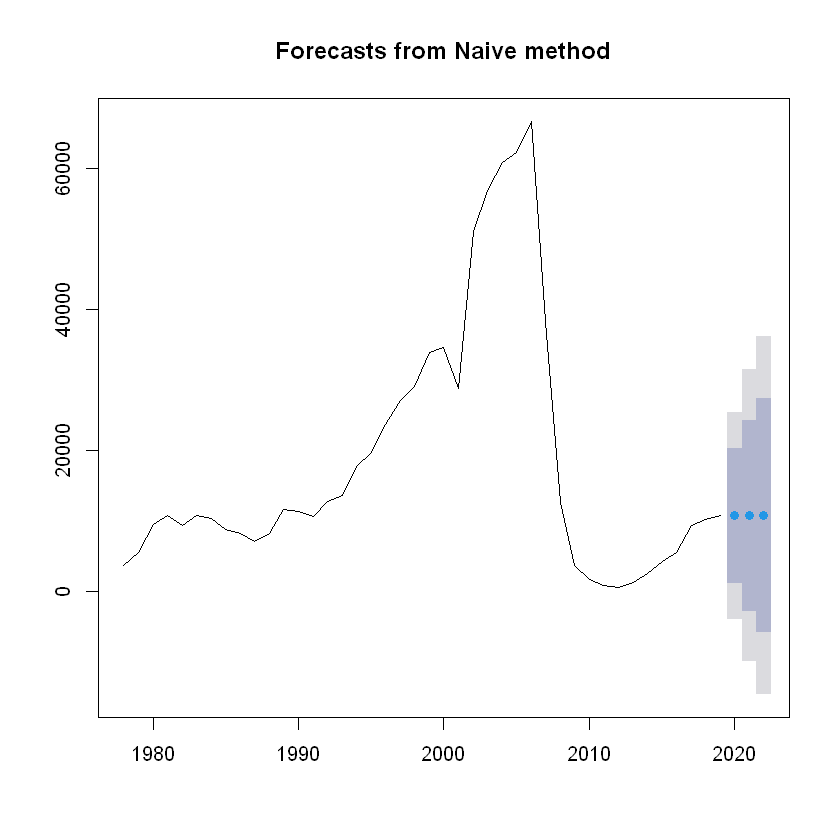


Figure II‑3 Naive Method

There are others “simple” methods like Mean Model and Drift Model which will predict future values based on the average of the historical data for the former or future values will increase or decrease over time on the later. RMSE stands for Root Mean Squared Error and provides a measurement of the errors made by the observations. On this cases RMSE for the Mean model is 17881.98, the Drift Model has 7464.78 and the Naïve has 7466.73.

ETS Function

ETS models are a more flexible method which provides the ability of provide different options which indicates how to handle different component of the time series. Figure 12 represent the forecast prediction using an ETS model that handle errors as Additive, the trend type as Additive too and finally it consider that there is not seasonality (None).

As previously mentioned, RMSE reflects the average error performed by the model in calculating each of the forecast observations. On the other hand, AIC (Akaike’s Information Criterion) is used to compare the fit of different models of the same data set. It is usually considered that the lower the AIC, the better model when working with ETS function.

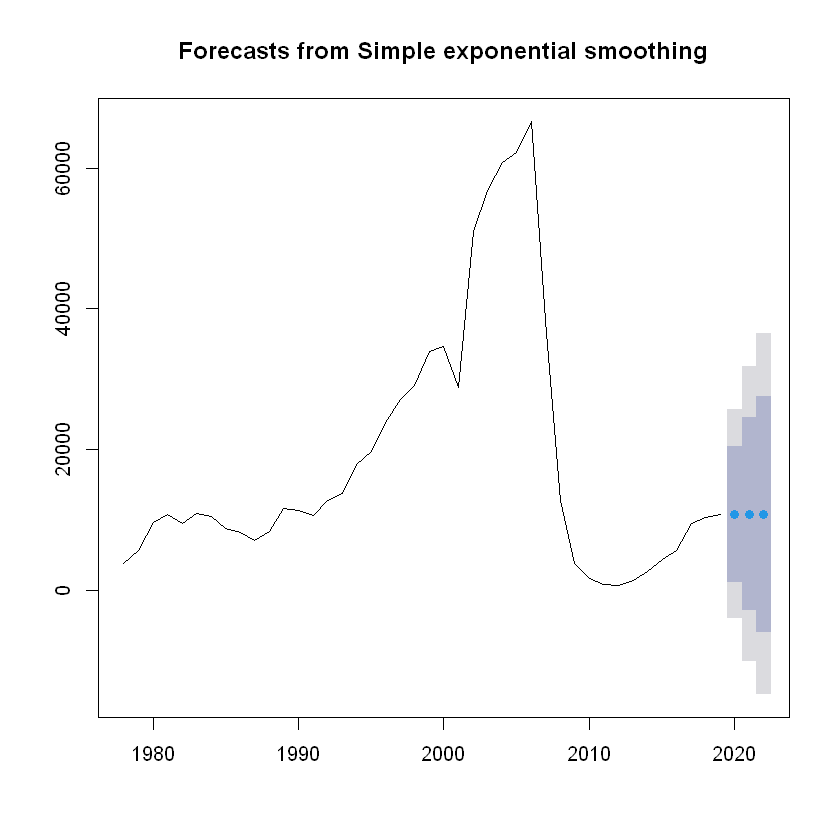


Figure II‑4 Simple Exponential Smoothing

Simple Exponential Smoothing method, SES for short, is a method that provide good forecasting result when working with data with no trend nor season. The fact that there is trend in the dataset might be the reason that this model do not provide the best results. It is worth to mention that the forecasting model ETS(A,N,N) is equivalent to an Simple Exponential Smoothing model. As it can be observed on comparison Table 1, the AIC value produced by this two models is very similar

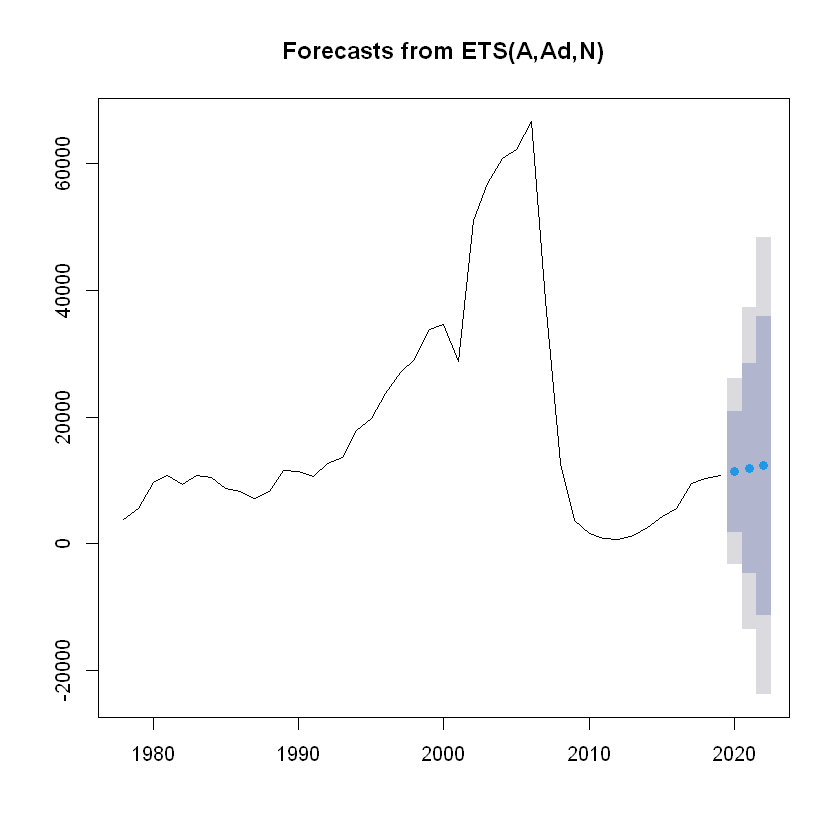


Figure II‑5 ETS - A, Ad, N

|  |  |  |
| --- | --- | --- |
|  | RMSE | AIC |
| SES | 7378.82 | 911.11 |
| ETS(A,N,N) | 7413.07 | 911.5062 |
| ETS(A,A,N) | 6993.5 | 912.61 |
| ETS(M,A,N) | 7395.98 | 863.47 |

Table II‑1 Comparison of models

On this instance and based on the AIC values of the different model, the model ETS(M,AN) is the most appropriate. Table 2 represents the values of the forecast for three periods ahead for each of the models considered in this section.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Fcst Val 1 | Fcst Val 2 | Fcst Val 3 |
| SES | 10783.75 | 10783.75 | 10783.75 |
| ETS(A,N,N) | 10783.95 | 10783.95 | 10783.95 |
| ETS(A,A,N) | 11411.76 | 11913.93 | 12315.66 |
| ETS(M,A,N) | 11057.77 | 11331.56 | 11605.36 |

Table II‑2 Forecasted values per model

## Transformations

It could be interesting to see how the forecast of the time series could have looked if the “big” fluctuation would have not happened around 2007. The consequences of the Celtic Tiger years in Ireland had their peak around 2008 when a global economic recession hit dramatically Ireland the number of unemployment increased rapidly [2]. If a new time series is created from 1978 to 2006, it could be observed how the forecast was at the time prior to the fluctuation around 2007 and 2008.

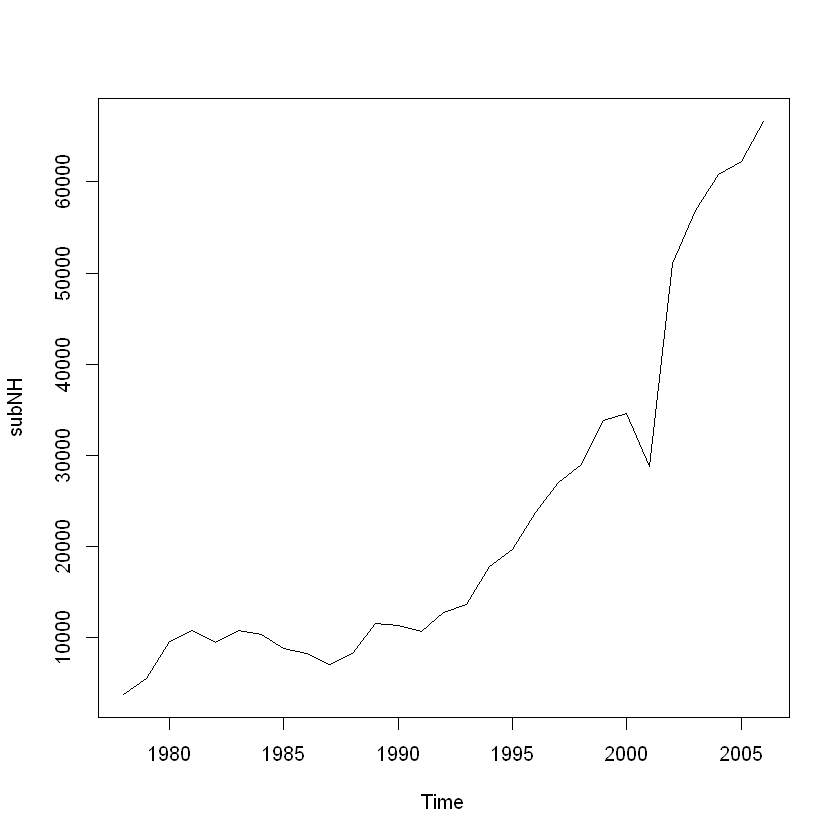


Figure II‑6 Time Series 1978 - 2006 Prior “big” fluctuation

Holt (1957) worked on an extension of the Simple Exponential Smoothing method for data with trend and level but without seasonality [1]. Holt’s model gives dynamic adaptation of the slope over time base on previous values (class comment) A dampened version of the Holt’s method will have an effect where will push the forecast line to the horizontal over a larger period of time. Forecast points seems to have a tendency to go up which lead to conclusions like that the logarithmic version of the model could potentially offer a more accurate representation.

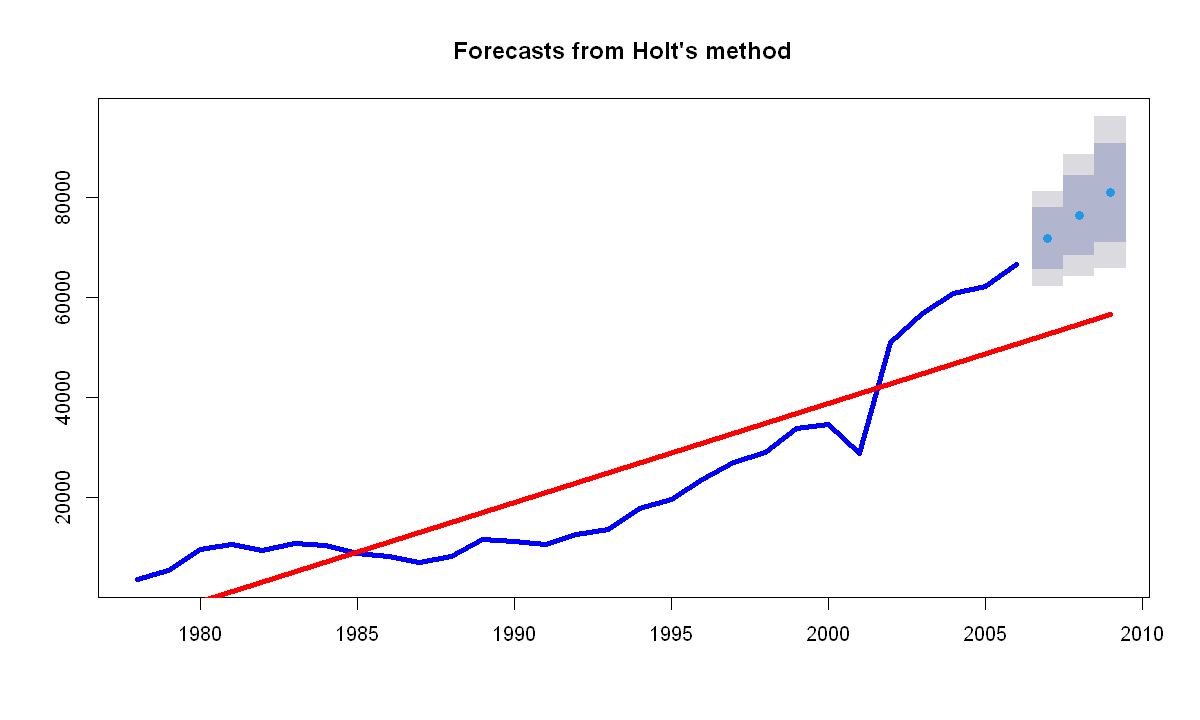


Figure II‑7 Holt's Method with Linear model

Logarithmic Version

One step further on the analysis of this new adapted dataset from 1978 until 2006 is to use the logarithmic version of the same. Based on the shape of the curve, there are indications that lead to believe that an algorithmic version of the model could have a better fit. Figure 16 represents a logarithmic model of the dependent variable (number of houses registered) and a linear combination of the independent variable (time). It can be observed at a glance that this new curve has a better fit with the original data that the linear model represented on Figure 15.

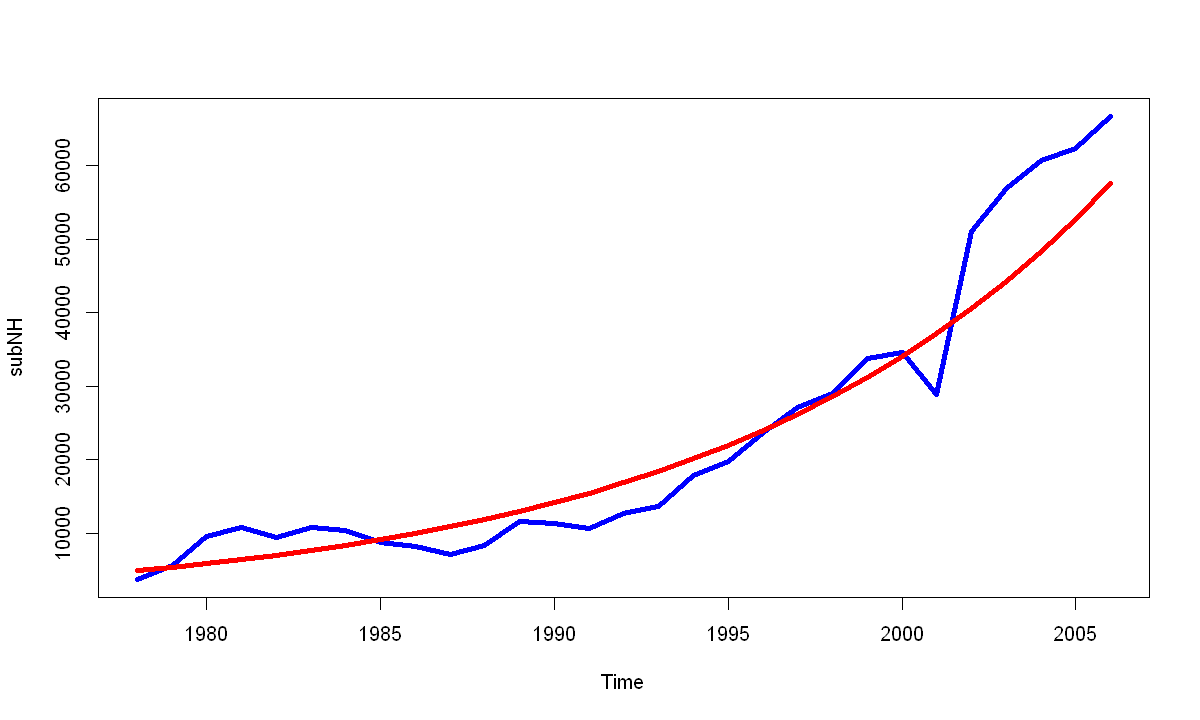


Figure II‑8 Logarithmic Linear Model

## Conclusion

The suspicion that the logarithmic model could deliver a better performance can be confirmed by looking at the residuals for each model. The mean of residuals for the lineal model are close to 2.0 while the mean of the residuals of the logarithmic lineal model is 1.0. The residuals have been reduced to the half which is a clear statement that the logarithmic linear model provides a better fit with the original data. Visualization of this model is represented on Figure II-9 where the forecast points are relative close to the red line.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Fcst Val 1 | Fcst Val 2 | Fcst Val 3 |
| Holt Model | 71750 | 76334 | 80919 |
| Log Holt Model | 73829 | 81783 | 90593 |

Table II‑3 Holt model and Log Holt Model forecated values

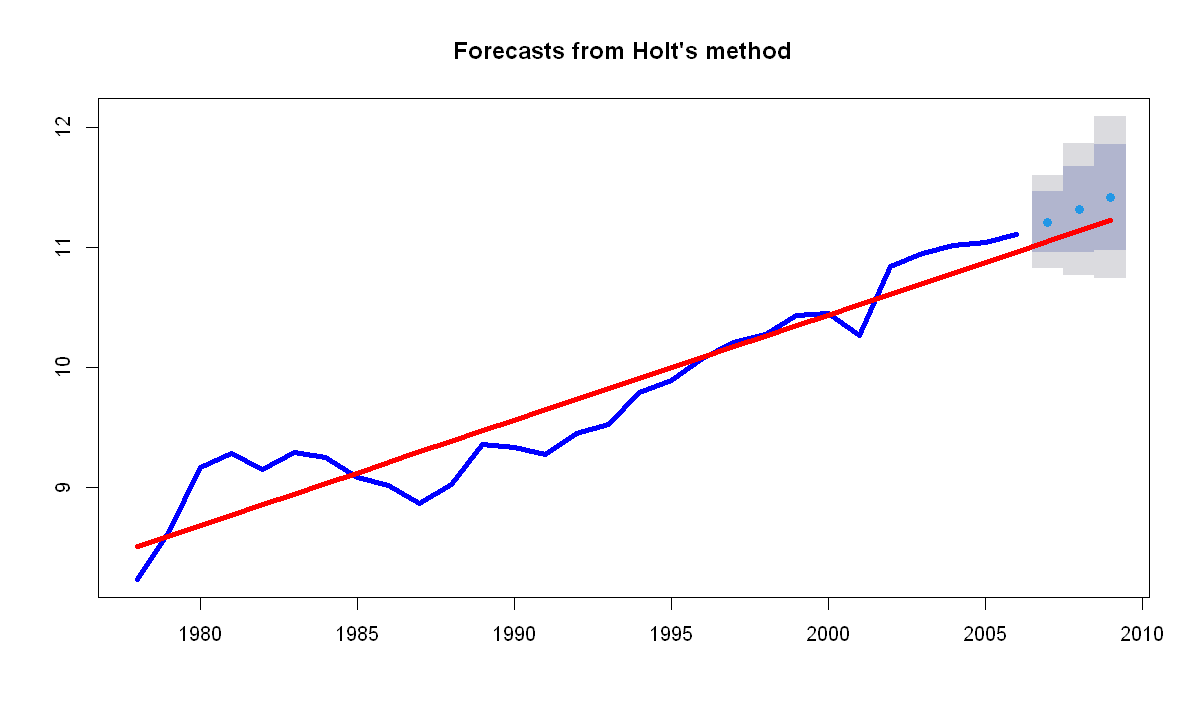


Figure II‑9 Log Holt Model and Log- Lineal Mode

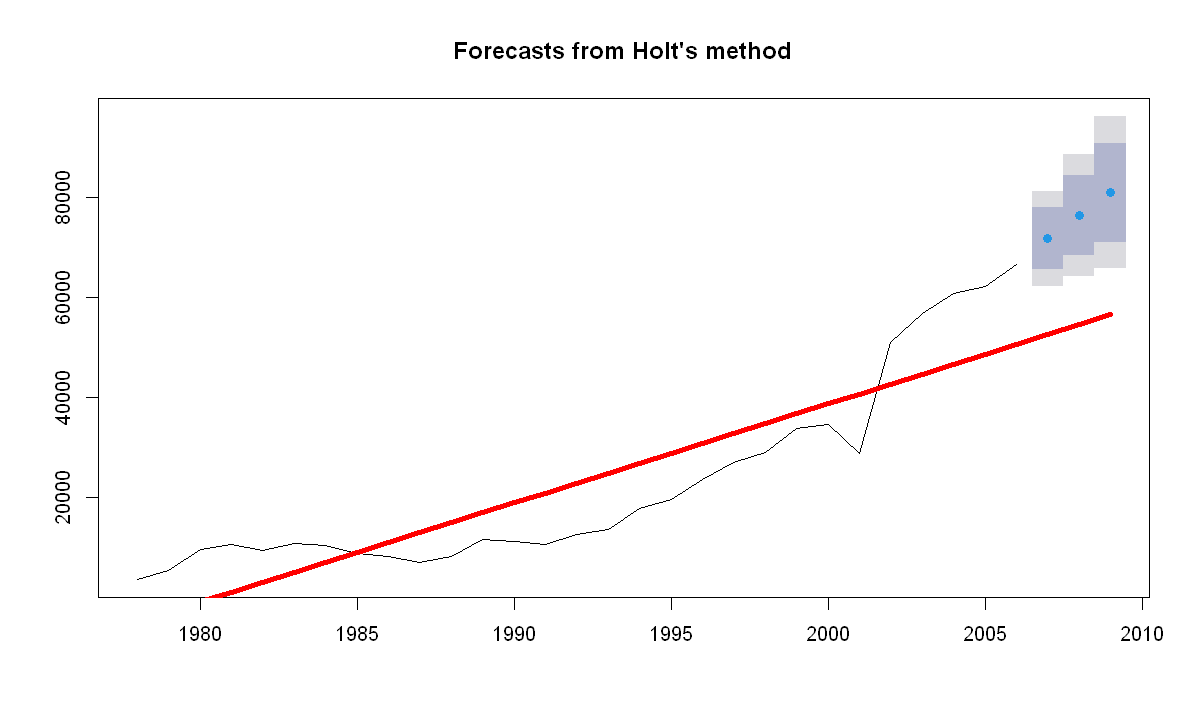


Figure II‑10 Forecast from Holt's method compared with original data

## Additional modeling with ARIMA

Lastly, to finalize this section, the time series forecast with be analyzed with an Non-Seasonal ARIMA models which will use ARIMA(2,0,0) for the model with the original data and ARIMA (0,1,0) for the model with data until the fluctuation. ARIMA(0,1,0) it is considered similar to a Random Walk with drift where the autoregressive part is 0, the integration =1 and the moving average is 0. Visualization for the forecast of these models are represented on Table II-4 and Table II-5 with the corresponding ACF plot.

AIC has been used to determine which model is better, but it need to be highlighted that this method is only valid when comparing model of the same class. ARIMA candidate models cannot be, or should not be, compared with ETS candidate model using their corresponding AIC. The reason behind this is that both models uses different methods to compute the likelihood. However, when working with non-seasonal data this model can be compared [1].

|  |
| --- |
|  |
|  |

Table II‑4 ARIMA (2,0,0) and ACF Plot

|  |
| --- |
|  |
|  |

Table II‑5 ARIMA(0,1,0) and ACF Plot

# PART B – BORN IN THE US

## Introduction

The file Childbirths data set contains data on childbirths in a city in the US. The dataset contains a total of 16 variables and 42 observations in total. At a glance, the data do not contain any strange or missing values and it is considered to be in a good status to start a more detailed analyst. There is a column which represents the observation id that will be removed from the dataset as it is not meaningful.

## Descritpive Analysis

As previously mentioned, there are 16 variables in the dataset which represents different parameters taken at childbirth, in a given hospital, it is assumed, in the US. The measurements of the baby taken are: **length** in centimetres, weight of the baby in Kilograms (**BirthWeight**), **circumference** of the head in centimetres (**HeadCirc**) and the total length of the **gestation** in weeks.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Length | BirthWeight | HeadCirc | Gestation |
| Mean | 51.33 | 3.31 | 34.6 | 39.19 |
| Median | 52.00 | 3.29 | 34.0 | 39.50 |
| Range | 43-58 | 1.92 – 4.57 | 30.39 | 33-45 |
| SD | 2.93 | 0.60 | 2.39 | 2.64 |

Table III‑1 Variables measure of central tendency and dispersion

At the same time, the dataset contain data related to the child parents. Table III-2 represent some parameters relative to the mum with the following units: **mage** (age of the mother), **mnocig** (number of cigarettes per day), **mheight** (height of the mother in cm) and **mppwt** (weight of the mother before pregnancy in Kg).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | mage | mnocig | mheight | mppwt |
| Mean | 25.55 | 9.42 | 164.5 | 57.50 |
| Median | 24.00 | 4.50 | 164.5 | 57.00 |
| Range | 18 - 41 | 0 - 50 | 149 -181 | 45 - 78 |
| SD | 5.66 | 12.51 | 6.50 | 7.19 |

Table III‑2 Variables measure of central tendency and dispersion

Table III-3 represents the measurements taken relative to the father of the baby: **fage** (age of the father), **fedyrs** (number of years in education of the father), **fnocig** (number of cigarettes smoked per day) and **fheight** (height of the father)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | fage | fedyrs | fnocig | Fheight |
| Mean | 28.9 | 13.67 | 17.19 | 180.5 |
| Median | 29.5 | 14.00 | 18.50 | 180.5 |
| Range | 19 - 46 | 10 – 16 | 0 – 50 | 0 – 50 |
| SD | 6.86 | 2.16 | 17.30 | 6.97 |

Table III‑3 Variables measure of central tendency and dispersion

In addition to this variables, there is an additional variable that represents whether the mum is or is not as smoker. This is represented by a binary variable called **smoker** and Figure 1 shows a boxplot representing a visual interpretation of how the weight of the mum vary depending whether the mum is a smoker or not.

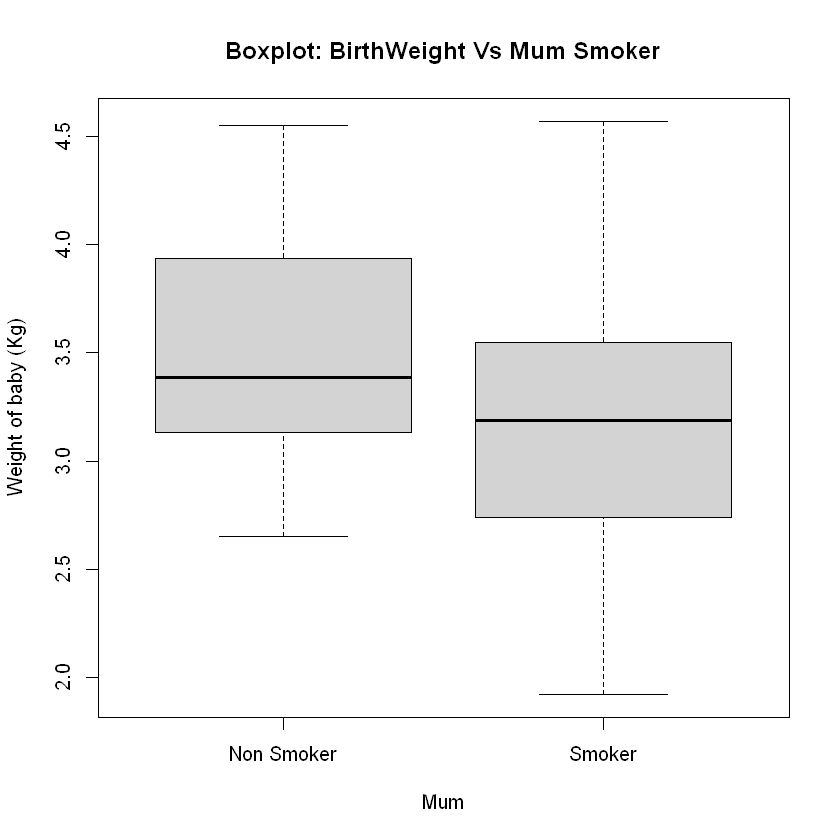


Figure III‑1 BoxPlot Baby Weight ~ Mum Smoker

The dataset contain another binary variable called **lowbwt** which represents with a 0 if the baby is considered to have a low birth weight and 1 if otherwise. This variable behaviour is represented with a boxplot on Figure III-2.

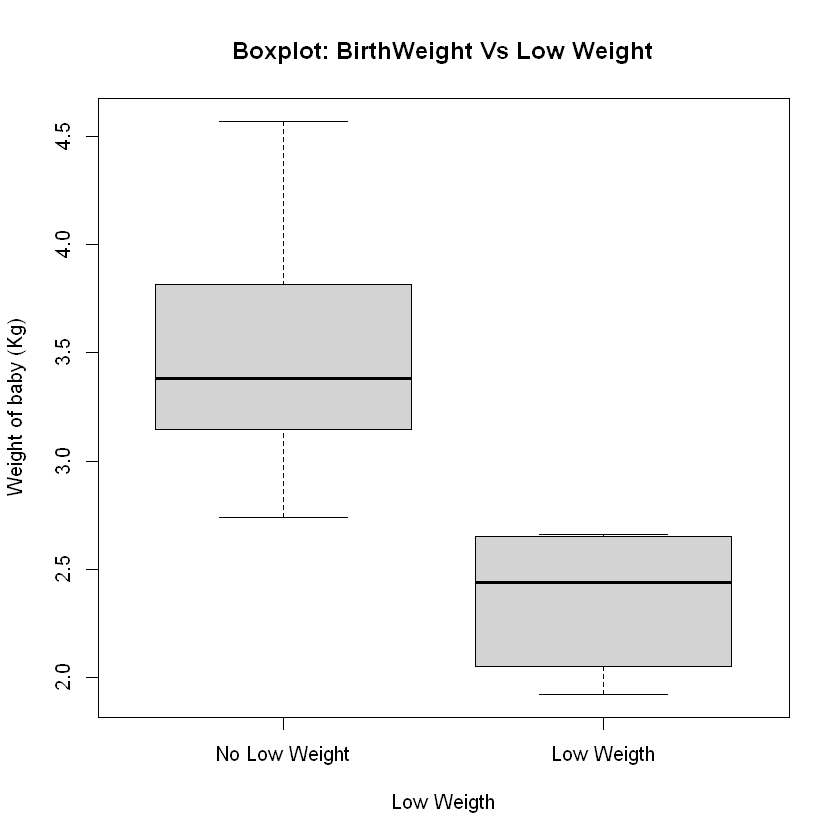


Figure III‑2 Boxplot Baby Weight ~ Low Weight

Lastly, another binary variable in the dataset is **mage35** and represent with a 0 if the mum is not over 35 years of age and with 1 if it is. Figure III-3 represents how the weight of the baby vary depending on whether the mother is over 35 years of age. In addition, it is worth to mention that an observation is considered as outlier which represents the lower value of the range for the variable BirthWeigth (see Table III-1) with value 1.92 kg. This observation will remain in the dataset as it is considered a legitimated observation.

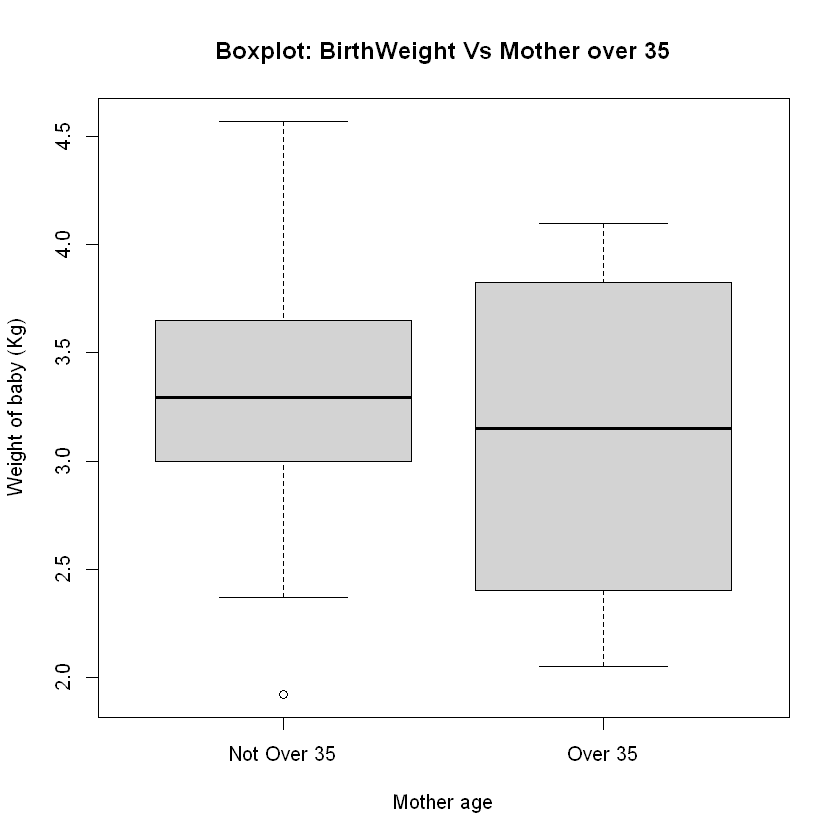


Figure III‑3 Baby Weight ~ Mother over 35 years

## Model Building.

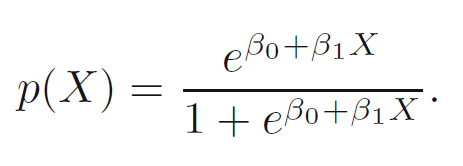
A logistic regression model will be chosen on this section in order to predict the result of an observation. On this instance, the result variable will be qualitative dichotomous so the final model will be classified as Binomial Logistic Regression.

*Why will Lineal regression not be used?*

Using Linear Regression would imply to set some ordering system on the outcomes. For instance, three possible categories could be encoded to for the response variable: 1 if stroke, 2 if drug overdose and 3 if epileptic seizure. The problems is that this system assume that the difference will be one between these three categories while there is no particular reason this would be the case. In general, a qualitative variable cannot be translated to a quantitative variable which can be used for linear regression [3].

## Logistic Regression.

A question that might arise by looking at the dataset is which weight is considered to be low weight at the facility where the child birth happened. In order to find out the cut off weight of the dataset to be considered as low weight, a logistic regression model will need to be built with a response variable with two possible variables as in Equation1 figure.



Equation 1Logistic Regression Function

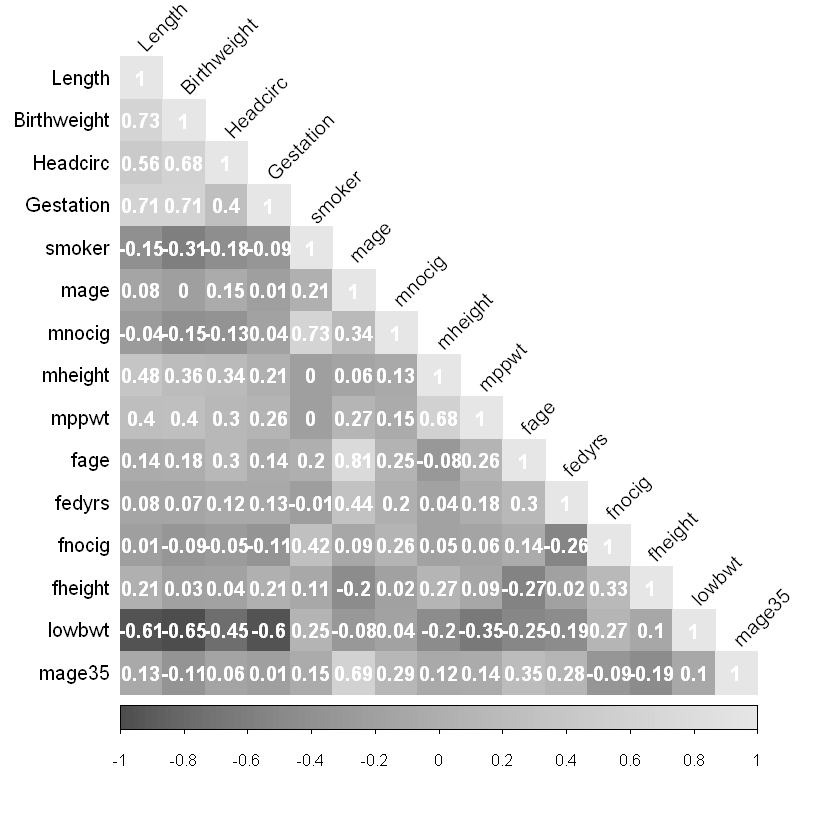


Figure III‑4 Correlation between childbirths dataset variables

Variables that will be in the model can be selected by StepWise algorithm.

Akaike information criterion (AIC) small values indicates a model with a low test error [3]. Basically, the AIC will be calculated based on the number of independent variables that has been selected for the model and the maximum likelihood estimate of the model. A number of model will be created with different number if variables and the model with the lowest AIC will be a string candidate to be the one which delivers better performance. A model with variables BirthWeigth and Gestation variables has previously been discarded as both variables have a strong correlation (0.6) and a AIC Value of 6. Model selected with only BirthWeigth and a AIC value of 4 variable seems to be the most adequate to find out what the criteria is to classify a baby as having low weight.

A prediction matching the model response variable using the predicted values calculated by the system will be designed. Table III-1 represents the values predicted compared with the actual values in the dataset. On this case, for those observations which has the childbirth as low weight, the model had predicted accordingly. Same apply for the six observations which have been marked as having low weigth that have been predicted accurately by the model.

|  |  |  |
| --- | --- | --- |
|  | Predicted | |
| Actual | No | Yes |
| No | 36 | 0 |
| Yes | 0 | 6 |

Table III‑1 Response prediction

With this prediction system the cut off point for a baby to be qualified as low weight is 2.69 Kg or lower while a baby with a weight of 2.70 Kg will not be classified as having a low weight. The coefficients for the model are β0 = 1283.537 and β1 = -475.35.

It make sense that BirthWeigth is the only variable used to determine whether the baby has a low weight. But, what would happened if the weight of the baby would not be available? Another model would need to be built and alternative parameters would need to be considered.

An empty model will be the starting point an then variables will be added which cause a reduction on the AIC value of the model.

|  |  |  |
| --- | --- | --- |
| Model StepWise | Variable | AIC |
| Glm0 | lowbwt ~ 0 | 39.46 |
| Glm1 | + fheight | 34.87 |
| Glm2 | + Gestation | 18.25 |
| Glm3 | + Length | 13.745 |
| Glm4 | + mage35 | 11.57 |
| Glm5 | + mppwt | 8 |

Table III‑2 Logistic Regression models: AIC Variation

Table III-2 show some interesting results on how the model AIC value vary as relevant variables get added to it. Main jump happen with a model with variable fheight and Gestation as the AIC decreases from 34.87 to 15.85. Variable Length gets added into consideration with a marginal decreased of the AIC values. The fact that the AIC variation, although positive, is small and it bring the cost of adding a variable which is strongly correlated with Gestation (0.71 as shown in Figure III-4) lead to the conclusion that Length variable will not be added to the model as it could impact to the interpretability of the model. Lastly, variable mage35 gets added to the model as as it reduces de AIC value. Fedrys variable will not be added to the model as it only add a marginal decrease on the AIC value. Final model chosen for the study is:

*FinalModel = glm(lowbwt ~ Gestation + fheigth + mppwt)*

## Conclusion.

The results of this alternative Multiple Logistic Regression replicates are that for the 36 cases which have been predicted as not having lower weight, the actual original data was zero (no low weight). On the other hand, for the 6 cases which had been predicted as having low weight, the original data set had them as one (low weight).

With an AIC value of 8, this new model provides an alternative procedure to classify whether a baby has been born with low weight in the scenario that the baby weight variable is not available for some reason.

##### References

1. Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 05/05/2021.
2. https://www.imf.org/en/Home Status Available. Accessed on 06/05/2021
3. G. James, D. Witten, T. Hastie, R. Tibshirani. “An Introduction to Statistical Learning with Applications in R” Springer, 2013.