Statistical Analysis of Time-Series Models and Logistic Regression Models

Raunak Milind Sathe
x20118350
National College of Ireland
Dublin, Ireland

***Abstract*—In this study, Time Series and Logistic Regression models were built using the Overseas Trips, New House Registration and Child birth datasets. Simple models, Exponential Smoothing models and SARIMA models were built on the Overseas trips' dataset. ARIMA was used for alongside simple models and exponential smoothing models for New House Registration dataset. Forecasts were then built for 3 periods ahead. Three Logistic Regression models were built using the Child birth dataset to identify the best model. Principal component analysis method was also used for dimension reduction and model building. This was done programmatically using R and Python.**

Keywords—Time Series, Logistic Regression, R, Python

I. PART A – TIME SERIES

A. Overseas Trips

A. Introduction - The dataset contained information on the overseas trips to Ireland by non-residents. This was a univariate dataset with a total of 33 rows. The data started from the year 2012 till 2019 and was on a quarterly basis.

B. Objective -

The objective of this time series analysis is to find the most suitable model from three different categories of time series models and forecast the series for a period of 3 quarters ahead.

C. Exploratory Data Analysis –

Before modelling the time series, it is critical to do some explanatory data analysis to understand the data. The most important factors to understand from a time series data whether there is trend and/or seasonality to the data. Below shows the time series plot

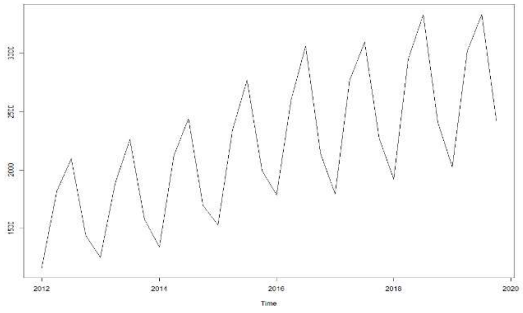

Fig 1 – Time series plot

Here are some initial observations from the plot –

- The time series starts from 2012 and ends in 2019 and the data is quarterly.
- As we can see there is a clear trend to the time series which is upwards.
- Also, there seems to be seasonality to the time series.
- The variance in seasonality is perhaps increasing with time.

Next, I have plotted the season plot to further understand the time series in respect of seasonality.

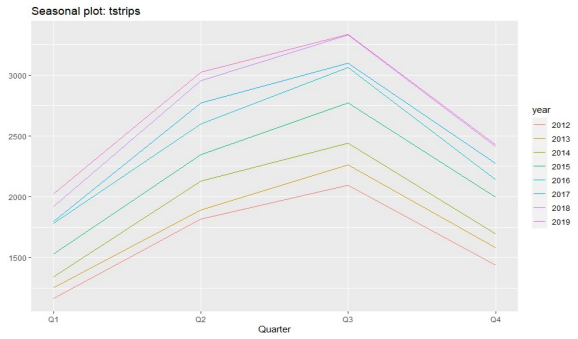

Fig 2. Seasonality of time series

As we can see from the above plot, there is a repetitive peak in the third quarter and a repetitive low in the first quarter. Next, I have decomposed the time series into its trend, seasonality and random component.

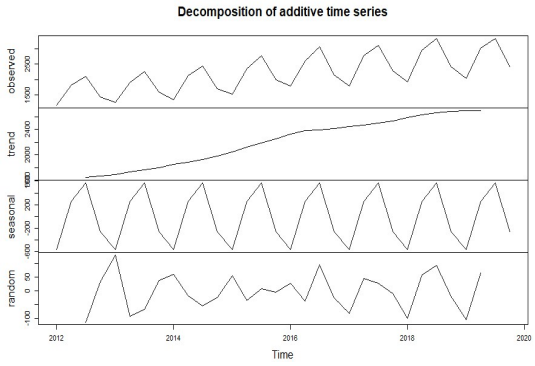

Fig 3. Decomposition of time series

As we can observe in the image above, there is a clear trend in the time series. There also appears to be a clear seasonality to the time series. basis. This concludes the exploratory data analysis.

D. Model Building Process –

In the model building process, I have built and compared 3 category of models – Simple Models (Mean model, Naïve model and Seasonal Naïve model), Exponential Smoothing Model (Holt-Winter model) and SARIMA models. Various different variations of the exponential Smoothing Model and SARIMA models were tried which have been explained in detail in the below sections.

1. Simple Models –

If we have 2 models with similar performance but differ in complexities, we will pick the model with the lowest complexity as that model will tend to hold up well over time. Some of the simplest time series models are the Mean model, naïve model and the seasonal naïve model. I have plotted the mean, naïve and the seasonal naïve models and compared these models on the basis of Root Mean Square Error (RMSE). The model with the lowest RMSE will act as a baseline model against which all future, more complicated models will be evaluated.

1. Mean Model – The mean model is a simple model where the forecasts are equal to the average or mean of the past data. The model was built with an RMSE value is 598.416
2. Naïve Model – In this model, the forecast values are simply equal to the previous values. The naïve model output for this time series is RMSE is 638.5391.
3. Seasonal Naïve model – In this method we set the forecast equal to the value observed during the same season the previous year. The Seasonal Naïve model output for this time series is RMSE is 176.6505.

The following table summarizes the simple model outputs–

Model	RMSE
Mean model	598.416
Naïve model	638.5391
Seasonal naïve model	176.6505

Table 1 – Outputs of simple models

As we can see from the table above, the Seasonal Naïve model has the lowest RMSE value of 176.6505. As I will be using the RMSE values to compare the different categories of models, we can safely say that the Seasonal Naïve model is the best performing model from the simple models' category. The following plot shows the forecast for 3 periods.

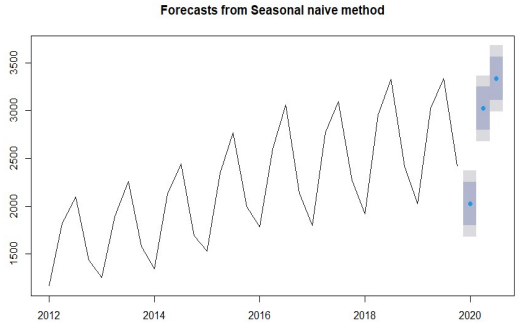

Fig 4. Forecast Seasonal Naïve

2. Exponential Smoothing Models –

When we talk about Time-Series models, two critical aspects need to be taken into account and accordingly models will be built – Trend and Seasonality. From fig 2 and fig 3 we can almost say for sure that there exists trend and seasonality. *However, these are just visual observations and not statistical conclusions*. Although there are multiple tests to test for seasonality and trend, *I have taken another way*. Since the data is very small, to confirm the trend and seasonality I have built a simple exponential smoothing model (no trend and no seasonality), a Holt model (only trend no seasonality) and a simple additive Holt-Winter model (includes both the trend and seasonality component). *Ideally, the Holt winter model should have a better performance i.e., lower RMSE as other models will fail to take into account the trend and/or seasonality.*

Another important factor to consider is whether the time series is additive or multiplicative. As observed in the fig 1, there appears to be an increasing variance with respect to time. As I cannot confirm just from visual observations whether it is additive or multiplicative, I will be using both in model building.

Model 1 – Simple Exponential Smoothing model (no trend, no seasonality)

The Simple Exponential Smoothing model is a model that doesn't take into account any trend or seasonality. The output of the Simple Exponential Smoothing model is a corrected AIC value of 518.2487 and a RMSE value of 521.98.

Model 2 – Holt model – AAN (only trend, no seasonality)

The holt model takes into account only the trend and no seasonality. The output of the model is a corrected AIC value of 513.5090 and a RMSE value of 445.15.

Model 3 – Holt-Winter model – AAA (both trend and seasonality)

The Holt-Winter model is a model that takes into account both the trend and seasonality component of the data. The output is a corrected AIC value of 414.7587 and RMSE value of 76.76. Ideally, we should see a better output for this model. *I have built models 1, 2 and 3 just for the sake of confirming this*. The following table summarizes the outputs of these models –

Model	Trend and seasonality component	Corrected AIC	RMSE
Simple Exponential Smoothing model	None	518.2487	521.98
Holt model	Trend=A, no q	513.5090	445.15
Holt-Winter model	**Trend=A, Seasonality=A**	**414.7587**	**76.76**

Table 2 – Confirmation that there exists a trend and seasonality

As we can see from the above table, the Holt-Winter model had a lower corrected AIC and RMSE than the simple exponential smoothing model and Holt model. Thus, we have confirmed that there is a definite trend and seasonality to the dataset. Hence, I will now focus on the *different variations of Holt-Winter models and find a model with the lowest corrected AIC and lowest RMSE*. Model 3 will act as the baseline model against which all possible Holt-Winter models will be evaluated.

Model 4 (MAA) – In this step, I have substituted the multiplicative value for the random component, Additive values for both the Trend and Seasonality. The output is a corrected AIC value of 413.1186 and a RMSE value of 77.4.

Model 5 (MMM)– in this step, I have substituted values for M for random, trend and the seasonality component. The corrected AIC was 389.9692 and the RMSE was 59.09.

Model 6 (MAM) – In this model, I have substituted the multiplicative value in the random component place, Additive value for the Trend component and Multiplicative value for the Seasonality component. The corrected AIC was 387.0085 and the RMSE value was 54.7. The following table summarizes the outputs of models 3, 4, 5 and 6.

Holt-Winter Models	Corrected AIC	RMSE
AAA	414.7587	76.76
MAA	413.1186	77.4
MMM	389.9692	59.09
MAM	387.0085	54.7

Table 3 – outputs of different Holt-Winter models

As we can see from the table above, the MAM model is the best performing model with a corrected AIC value of 387.0085 and a RMSE value of 54.7. Thus, in the category of exponential smoothing models, we compared Simple Exponential Smoothing models, Holt models and Holt-Winter models and we can safely say that the Holt-Winter model with the variation of (M, A, M) is the best performing model from this category. The following diagram shows a forecast for 3 periods

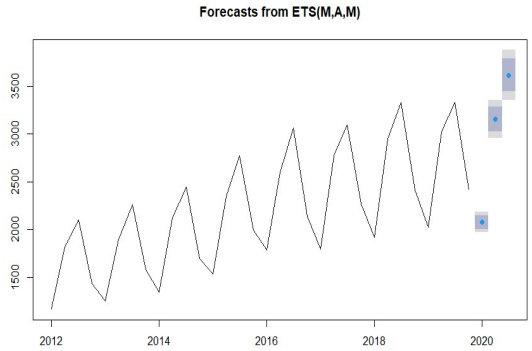

Fig 5. Forecast ETS (M, A, M)

3. Seasonal ARIMA model –

Step 1 – find the level of differencing (ordinary and seasonal) –

Ordinary - By running the ndiff function, I found the number of ordinary differences needed here is 1. Hence d=1.

Seasonal differencing – by running the nsdiffs function, I found the number of seasonal differences to use is also 1. Thus D=1.

Step 2 – Analysis and model building process –

The Seasonal ARIMA model can be used when there is a seasonality to the time series. SARIMA has to be used when there is a seasonal decline seen at fixed intervals or lags. For the model building process, it is important to plot the ACF and the PACF plots.

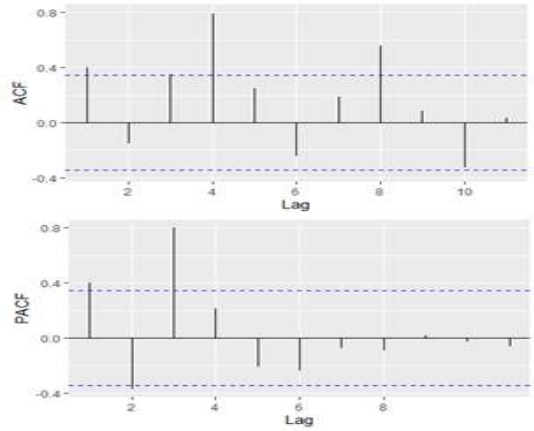

Fig 6. ACF and PACF plots

As we can see in the ACF plot, there is gradual decline taking place at 2 sets of lags – the first is happening at 1, 3, 5, 7 and the second is happening at 4, 8. As the lag at 4 is highly prominent than the other lag, I will be taking this lag into account

Model 1 – (0, 1, 0), (0, 1, 0) at lag 4 – the first variation was built using the values of 0, 1, 0, 0, 1, 0 in the place of p, d, q, P, D, Q respectively. The AIC was 315.99 and RMSE was 74.2759.

Model 2 - (1, 1 ,0), (0, 1, 0) at lag 4 – The second variation was built using the values of 1, 1, 0, 0, 1, 0 in the place of p, d, q, P, D, Q respectively. The AIC was 315.87 and RMSE was 70.84983.

Thus, we can see that the model 2 has the lowest AIC and RMSE.

Step 3 – finally, let us check the residuals. The following diagram shows the distribution of the residuals.

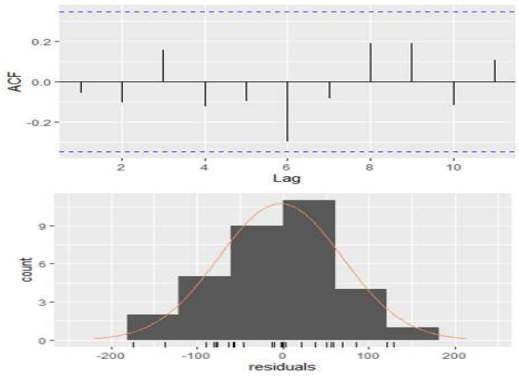

Fig 7. Residual plots

As we can see from the above diagrams, the residuals are below the threshold and are normally distributed. The Ljung-Box test determines whether the auto-correlations are zero or not.

```
        Box-Ljung test

data:  fit7$residuals
X-squared = 0.11943, df = 1, p-value = 0.7297
```

As we can, the results are not significant, suggesting the autocorrelations don't differ from zero.

Hence, we can say that the model is working correctly and that we have not missed out on any information. The following plot shows the forecasts for a period of 3.

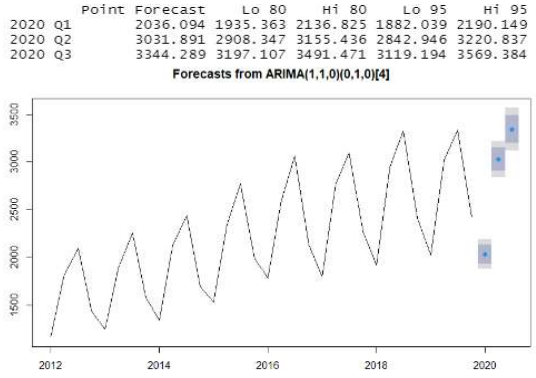

Fig 8. Forecasts for a period of 3 quarters.

E. Results –

The following table summarizes the best performing model from each category –

Model Category	Corrected AIC	**RMSE**
Simple model - Seasonal naïve model	-	**176.6505**
Exponential Smoothing model – Holt-Winter model (MAM)	387.0085	**54.7**
SARIMA model - (1, 1 ,0), (0, 1, 0) at lag 4	315.87	**70.84983**

Table 4 – outputs of best performing models from each category

In order to evaluate the models, only RMSE can be used across multiple categories and not corrected AIC. The Holt-Winter model ((MAM) has the best RMSE of 54.7 followed closely by the SARIMA model - (1, 1 ,0), (0, 1, 0) at lag 4 with an RMSE of 70.84. I will be considering the Holt-Winter model (MAM) as the most suitable model for this dataset for the following reasons –

1. The Holt-Winter model (MAM) has lower RMSE.
2. Although, the SARIMA models give more importance to the most recent values, the time series is fairly stable in terms of variance and looks constant throughout. Hence, the forecast values don't necessarily depend heavily on the most recent values.

B. New House Registration

A. Introduction - The dataset contained information on the new house registrations in Ireland. This was a univariate dataset with a total of 42 rows. The data started from the year 1978 till 2019 and was on a yearly basis.

B. Objective - The objective of this time series analysis is to find the most suitable model from three different categories of time series models and forecast the series for a period of 3 quarters ahead. quarterly basis.

C. Exploratory Data Analysis –

As usual before building the models, it is important to do the exploratory data analysis to extract information that maybe a significant contributor to the model. The initial step is to plot the time series –

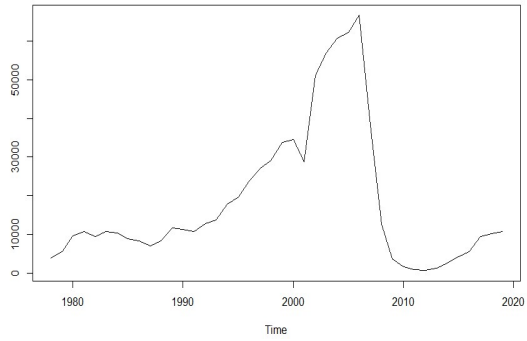

Fig 9. New House registration plot

This plot is very confusing to analyse. There seems to be an upward trend up till 2008. After that there is a sudden drop in the registrations. This makes sense as this time period relates to the economic depression of 2008. The registrations have again started going up over the last few years. There doesn't seem to be any evidence of seasonality.

D. Model Building –

1. Simple models –

These are the simplest time series models and are the easiest to plot. These models can be used baseline models against which all future more complex models will be judged. The following are the simple models –

Simple mean model – the forecast values are simply equal to the mean of the previous values. The model output for this time series is RMSE is 17881.98.

Naïve model (Random Walk model) – In this model, the forecast values are simply equal to the previous values. The naïve model output for this time series is RMSE is 7466.737.

Seasonal Naïve model – In this method we set the forecast equal to the value observed during the same season the previous year. The Seasonal Naïve model output for this time series is RMSE is 7466.737.

The following table summarizes the above simple model outputs –

Model	**RMSE**
Mean model	17881.98
Naïve model	7466.737
Seasonal naïve model	7466.737

Table 4 – outputs of simple models

As we can see from the table, the naïve model is the best performing model. I have decided not to go with the seasonal naïve model as there is no seasonal component to the time series and it might not hold up in the longer run. The following

diagram shows the forecast for a period of 3 for the naïve model –

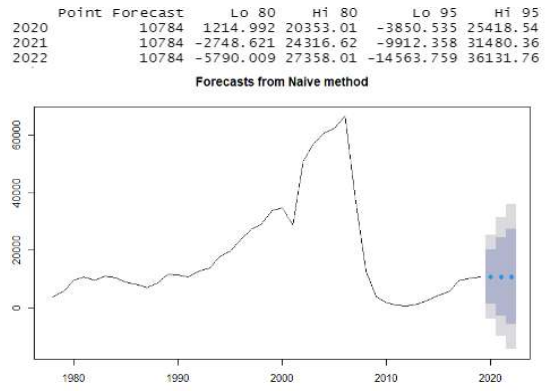

Fig 10. Naïve model forecast

2. Exponential Smoothing model –

The plot seems very confusing especially the trend component. By just simply observing the plot, we can say that there is no seasonality. When I used the seasonplot function, I got an error that the dataset is not seasonal. The sudden drop in the trend seems to be due to the global market crash of 2008. Hence, the initial observations are not providing any conclusive results regarding the presence of the trend component to the data.

Simple Exponential Smoothing model – The model was thus built without any trend and seasonality component with an AIC value of 911.7487 and an RMSE value of 7378.822.

Holt model (only Trend) –

Model 1 (AAN) - In the first variation of this model, I substituted additive values for the random component and the Trend component. The AIC value was 915.0120 and the RMSE value is 6993.505.

Model 2 (MMN) – in the second variation of the model, I substituted multiplicative values for the random and trend components. The AIC value was 872.1890 and the RMSE value was 8072.683.

Model 3 (MAN) – in the third variation, I substituted multiplicative value for the random component and additive value for the trend component. The AIC value was 865.1418 and the RMSE value was 7395.984.

The following table summarizes the outputs of the above models –

Model	Corrected AIC	RMSE
Simple Exponential Smoothing model	911.7487	7378.822
Holt model (AAN)	915.0120	6993.505
Holt model (MMN)	872.1890	8072.683
Holt model (MAN)	865.1418	7395.984

Table 5 – outputs of the Exponential smoothing models

Although the corrected AIC value for the Holt model (MAN) is the lowest, I decided to keep the Holt model (AAN) as the best performing model for the category Exponential Smoothing model as this model has the lowest RMSE value and ultimately the different categories will be compared based on RMSE and not AIC. The following diagram shows the forecast for a period of 3 years for the Holt-model (AAN) –

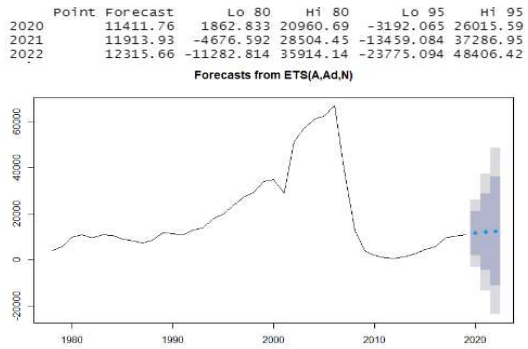

Fig 11. ETS (A, A, N) forecast

3. ARIMA models –

Step 1 – analysing the stationarity of the time series –

In this step, I have first used the augmented Dickey-Fuller test which showed that the time series is not stationary. The p-value in the Dickey-Fuller test only showed significance after taking a difference of 3. However, the ndiffs function in R was showing that 0 levels of differences are needed. Hence, I decided to plot 2 models with d=0 and d=3.

Step 2 – calculating the AR and MA values

To do this, I first plotted the ACF and PACF plots. I plotted 2 plots using values of d=0 and d=3.

Model 1 – (p=2, d=0, q=0)

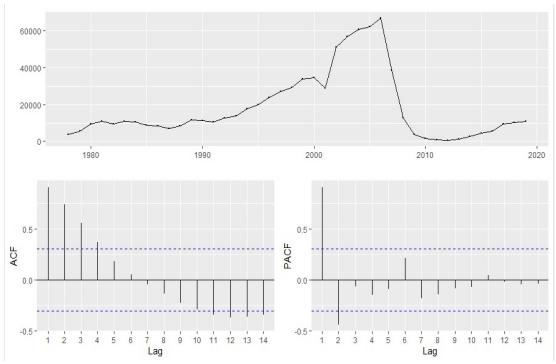

Fig 12. ACF and PACF plots

As we can see from the above plots, there is a clear gradual decline in the ACF plot and sudden spike in the PACF plot. This is a classic autoregressive representation. Hence, I have taken values of p=2, d=0, q=0. The corrected AIC value is 865.94 and the RMSE value is 6342.208.

Model 2 – (p=3, d=3, q=0)

As mentioned before I used 3 orders of difference to get a significant p-value score in the adf test. As the plots were now too complex to interpret, I relied on the auto.arima function for this model. The auto.arima function suggested a p score of

3 and a q score of 0. The corrected AIC value was 829.16 and a RMSE score of 8742.945.

The following table summarizes the outputs of the 2 models –

Model	Corrected AIC	RMSE
Model 1 - p=2, d=0, q=0	865.94	6342.208
Model 2 - p=3, d=3, q=0	829.16	8742.945

Table 6 – outputs of ARIMA models

Although, model 2 had a better corrected AIC, the difference was not too great. However, the difference in the RMSE values is very significant. Hence, I decided to select model 1 with p=2, d=0, q=0 as the most suitable model for low complexity and significant difference in RMSE.

Step 3 – checking residuals –

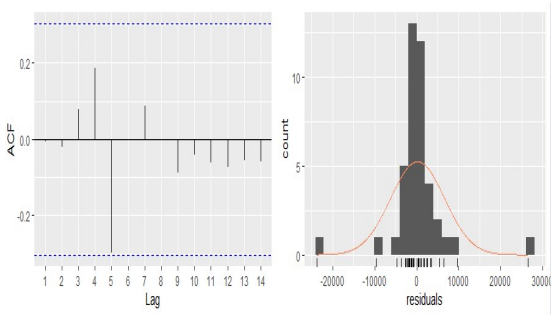

Fig 13. Residuals plots

As we can see from the plot, the residuals are not beyond the threshold and are also normally distributed. Hence, we can say that the model is working and that we have not missed out on any significant information from the time series.

Step 4 – Forecasting –

After confirming that this is the best performing ARIMA model, I plotted the forecast for 3 periods –

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2020	11818.59	3383.902	20253.27	-1081.151	24718.33
2021	12957.23	-1109.161	27023.62	-8555.457	34469.91
2022	13994.20	-3917.264	31905.66	-13399.019	41387.42

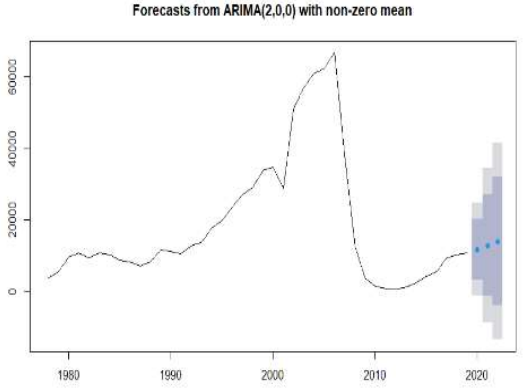

Fig 12. ARIMA (2, 0, 0) model forecast

E. Results –

The following table summarizes the best performing models from each of the three categories –

Model	Corrected AIC	**RMSE**
Simple model - Naïve model	-	**7466.737**
Exponential Smoothing model - Holt model (AAN)	915.0120	**6993.505**
ARIMA model - p=2, d=0, q=0	865.94	**6342.208**

Table 7 – outputs of best performing models from each category

As these are different categories, AIC value cannot be used. Only RMSE value can be used here. I have selected the ARIMA model as the best performing model for the following reasons –

1. ARIMA model has the lowest RMSE value.
2. ARIMA model gives more importance to the most recent data. The significant drop in house registrations is a result of the global economic crisis of 2008 and not a repetitive process. Hence, the most recent values should be given more weightage.

II. Part B – Logistic Regression

A. Introduction – The dataset used was the child births dataset from a US city. The dataset contained three dichotomous variables – lowbwt (low birth weight), smoker and mage35 (mother age above or below 35). The dataset contained a total of 42 entries and 16 variables. The dependent variable lowbwt is a dichotomous variable. The variables Length, Birthweight, Head circumference, Gestation, mother age, mnocig (no of cigarettes smoked by mother), mheight (mother's height), mppwt, fage, fnocig, fedyrs, fheight, are all continuous variables. The variables smoker, and mage35 are dichotomous variables.

I decided to use the lowbwt variable as the dependent variable. This was a dichotomous variable with values 0=no and 1=yes.

B. Objective – The objective of this analysis is to build a suitable logistic regression using suitable variables.

C. Exploratory Data Analysis –

As usual before building the model, some data exploration was carried out. On primary examination, some variables appeared to be problematic. After carrying out some primary exploratory data analysis, it yielded the following result –

- Birthweight – the target variable lowbwt was just a dichotomous representation of the variable Birthweight. The Birthweight value of smaller than 2.70 was listed as 0 and greater than 2.70 was listed as 1. If this variable was included in the model building, it would've given a 100% accuracy incorrectly. Hence, this variable was removed.
- Smoker and mnocig – On closer examination, the two variables were again essentially the same. They

portrayed the same information. Value of 1 for smoker was the equivalent of number of cigarettes smoked by the mother greater than 0. Hence, one of these variables was removed, but the question was which? After plotting heat map, it turned out mnocig was a more significant contributor towards the target variable and hence, it was kept.

- Mage and mage35 – the 2 variables were once again giving the same information. A mage value of less than 35 was directly equivalent to 0 for mage35 and vice versa. Hence, one of these variables had to be removed. It turned out that the mage35 was a less significant contributor to the target variable and hence, it was dropped.

The distribution of the variables is shown below –

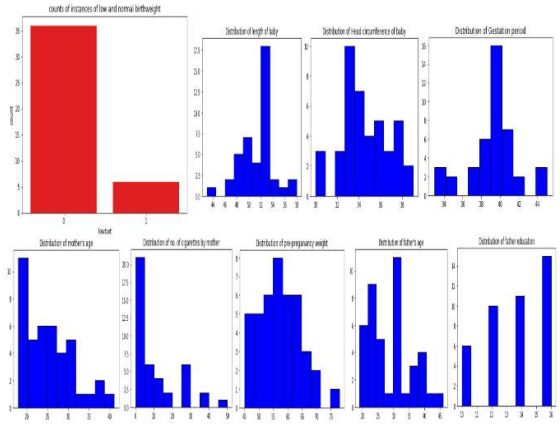

Fig 13. Variables distribution

Above shown are the distributions of the variables lowbwt, Length of baby, Head circumference, Gestation, mother's age, no of cigarettes by mother, pre-pregnancy weight, father's age and father's education in that particular order.

The first plot in red is the distribution of the dependent variable. As we can see the number of low birth weights are very low in comparison to the babies not born with low weight. The remaining independent variables show a relatively normal distribution.

D. Assumptions –

1. The target variable is a dichotomous variable

2. There exists a relationship between the dependent variable and the independent variables –

Correlation Matrix

		Constant	Length	Birthweight	Headcirc	Gestation	smoker	mage	mnocig	mheight	mppwt	fage	fedyrs	fnocig	fheight	mage35
Step 1	Constant	1.000	.245	-.388	-.346	.514	-.546	-.382	.032	-.635	.610	-.170	.470	.718	-.563	.370
	Length	.245	1.000	-.486	.007	-.153	-.235	.034	.129	-.455	.093	-.201	.548	.378	-.551	-.121
	Birthweight	-.388	-.486	1.000	-.101	-.559	.614	-.131	-.383	.407	-.312	.420	-.508	-.428	.703	.034
	Headcirc	-.346	.007	-.101	1.000	-.353	.405	.196	.086	.097	.162	-.279	.076	-.263	-.067	-.260
	Gestation	.514	-.153	-.559	-.353	1.000	-.668	.106	.234	-.265	.229	-.456	.197	.453	-.543	.284
	smoker	-.546	-.235	.614	.405	-.668	1.000	-.003	-.522	.494	-.224	.208	-.146	-.415	.398	-.143
	mage	-.382	.034	-.131	.196	.106	-.003	1.000	.463	.310	-.455	-.670	-.434	-.516	-.010	-.375
	mnocig	.032	.129	-.383	.086	.234	-.522	.463	1.000	-.126	-.017	-.368	-.235	-.347	-.094	-.388
	mheight	-.635	-.455	.407	.097	-.265	.494	.310	-.126	1.000	-.757	.141	-.469	-.563	.451	-.101
	mppwt	.610	.093	-.312	.162	.229	-.224	-.455	-.017	-.757	1.000	.013	.451	.418	-.355	.009
	fage	-.170	-.201	.420	-.279	-.456	.208	-.670	-.368	.141	.013	1.000	-.165	-.120	.591	-.046
	fedyrs	.470	.548	-.508	.076	.197	-.146	-.434	-.235	-.469	.451	-.165	1.000	.782	-.771	.271
	fnocig	.718	.378	-.428	-.263	.453	-.415	-.516	-.347	-.563	.418	-.120	.782	1.000	-.706	.567
	fheight	-.563	-.551	.703	-.067	-.543	.398	-.010	-.094	.451	-.355	.591	-.771	-.706	1.000	-.190
	mage35	.370	-.121	.034	-.260	.284	-.143	-.375	-.388	-.101	.068	-.046	.271	.567	-.190	1.000

E. Model Building process –

The null hypothesis states that the independent variables have very little or no effect on the target variable. A p-value less than 0.05 is statistically insignificant. When the p-value is lower than 0.05 we can reject the null hypothesis and keep the variable in the dataset. When the p-value is greater than 0.05, we fail to reject the null hypothesis. This means we exclude the variable from the equation.

Model 1 – In this model I have taken into account the characteristics of the mother and the father and also the combinations of them. Variable used were *mage, mnocig, mheight, fage, fedyrs, fnocig and fheight*. The reasons for shortlisting these variables were intuition and the correlation heatmap, which showed high interaction between some of these variables as well strong correlation with the target variable.

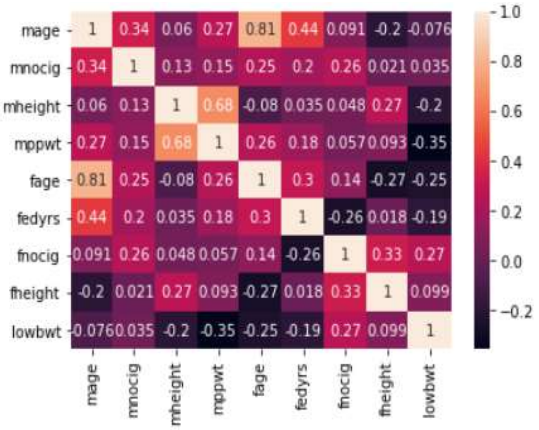

Fig 14. Correlation heatmap (lowbwt, mage, mnocig, mheight, mppwt, fage, fedyrs, fnocig and fheight)

As can be seen from the image above, variables such as fage, mppwt, fnocig are somewhat related to the lowbwt. Mage is not related to lowbwt surprisingly, but has a very high correlation with fage.

Step 1 – First, I entered the variable mage into the regression, it turned out to be insignificant. Then I entered the variable fage into the regression which again turned out to be insignificant.

Step 2 – Then, I used the interaction term of the fage and mage which turned out to be significant. Hence, I kept the interaction term and removed the others.

Step 3 – Then, I kept the fage term and added the product interaction of $(fage)^2$ and $(mage)^2$. All these variables and their interaction terms showed significance. Hence, they were kept.

Step 4 – Then I added the variable fedyrs into the equation as a product of the term acquired in step 3. This again showed very high significance.

Step 5 – Then I tried adding other variables such as fheight, fnocig, mheight and mnocig, but they all showed insignificance. This wasn't surprising as the relation with the dependent variable is very low.

Hence, the final model was built with the product interaction terms of fage, fedyrs, $(fage)^2$ and $(mage)^2$ which was I(fage^2)*fage*I(mage^2)*fedyrs. The AIC value of this model was 376.44.

Model 2 – For this model I have used the variables associated with the baby's characteristics and one important mother's characteristic such as Length, Gestation, Headcirc and mppwt. The reasons were again intuition and the correlation heatmap.

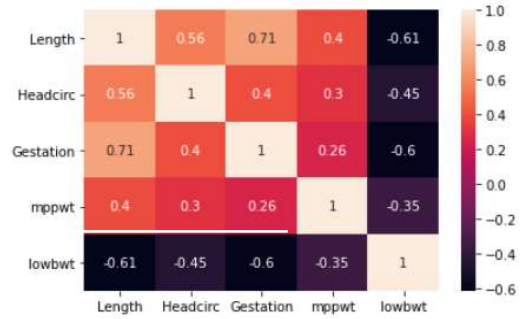

Fig 15. Correlation heatmap (lowbwt, Length, Headcirc, Gestation, mppwt)

As we can see, variables Length, Headcirc, Gestation and mppwt are very highly related to the lowbwt in a negative way i.e., an increase in Length, Gestation, Headcirc and mppwt will not lead to low birthweight.

Step 1 – Looking at the correlation heatmap, the variables Length and Gestation have a correlation coefficient of 0.71, which is very high. Since, Length has a higher relation with lowbwt, I decided to take the Length variable first. The Length variable turned out to be significant. The AIC value was 18.694.

Step 2 – Next, since there was high correlation between Length and Gestation, using those variables individually was not an option. Hence, I took an interaction of these terms. The interaction turned out to be significant. However, the AIC value increased to 19.16. Hence, this term was not included.

Step 3 – Next, I used the variable mppwt in the equation. Taking an interaction of Length:mppwt:Gestation proved very significant, as it lowered the AIC to 16.92. However, the variables Length, mppwt and Gestation no longer showed significance individually. Hence, the individual variables were dropped.

Step 4 – Next, I added the Headcirc to the model individually as well as a part of the interaction. The interaction of Length:mppwt:Gestation:Headcirc showed high significance. This further lowered the AIC to 15.449. The variable Headcirc was not showing significance individually anymore hence, it was dropped.

Hence, after going through all the steps the model 2 was built with just the single term of Length:mppwt:Gestation:Headcirc and an AIC of 15.449.

Model 3 – After taking into account all the terms in the dataset on the basis of their significance, I used PCA to lower the dimensions of the data. I reduced the variables into 3 components. The components were as follows –

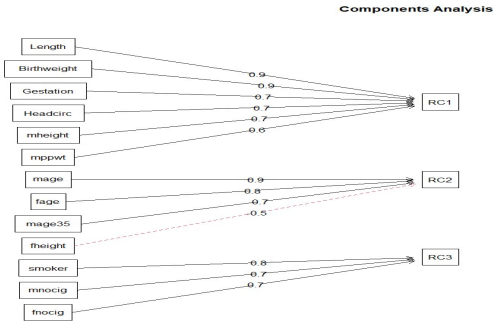

Fig 16. Component analysis

Step 1 – First, I built the model using just the first component (RC1). This component showed significance and the AIC was 17.871.

Step 2 – Next, I added the second principal component to the model (RC2). However, the components no longer showed significance. I also used the interaction term of the 2 components, but it was still insignificant. I also tried adding the square terms of the components, but still did not show any significance. Hence, the second component was not included.

Step 3 – Next, I add the third component to the model (RC3). The component was not significant at the 0.05 level. I also added the interaction of the first and the third component but it was still insignificant.

Step 4 – finally, I added all three components but still did not show any significance. However, when I added the interaction term RC1:RC2:RC3, it showed significance at the 0.05 level. The AIC was 31.099.

Hence, I decided to keep the model with just the first principal component and leave out the rest of the components as well as the interaction terms. Thus, the AIC of the third model was 17.871.

F. Result –

The following table summarizes the outputs of the 3 models –

Model	Model terms	AIC
1	I(fage^2)*fage*I(mage^2)	376.44
2	mppwt:Gestation:Length:Headcirc	15.449
3	RC1:RC2:RC3	31.099

Table 8 – outputs of Logistic Regression models

Taking a look at the above table and the correlations, I have decided to conclude model 2 as the best performing model for the following reasons –

1. The model 2 has the lowest AIC.
2. Taking a look at the correlation diagrams, all the terms in the model 2 had much stronger relationship with the dependent variable. Thus, they will be good predictors going forward as well.
3. The model 3 requires PCA before building the model. The model 1 requires a greater number of variables that are not even significantly related to the dependent variable. Hence, these models are more complex and misleading than model 2.

III. Conclusion

PART A –

1. Overseas Trips time series – Three different categories of time series models were plotted – Simple models, Exponential smoothing models (Holt-Winter model) and SARIMA models. The Holt-Winter (MAM) model is the most suitable model with a corrected AIC value of 387.0085 and a RMSE value of 54.7

2. New House Registration – Three different categories of time series models were plotted - Simple models, Exponential Smoothing models and ARIMA models. The ARIMA model

- p=2, d=0, q=0 was the best performing with a RMSE score of 6342.208 and the most suitable model for this time series and accordingly a forecast was built for 3 periods.

PART B –

A Logistic Regression model was built using the child births dataset. The model with the variables Length, mppwt, Gestation and Headcirc was built. The model was built using the interaction term of Length:mppwt:Gestation:Headcirc with an AIC of 15.44.

IV. Reference

[1] R for Data Science, Book by Grolemund and Hadley wickham

[2] Mcleod, Ian & Yu, Hao & Mahdi, Esam. (2012). Time Series Analysis with R. Elsevier B.V., Handbook in Statistics. 30. 661-712. 10.1016/B978-0-444-53858-1.00023-5.