Terminal Assignment (TABA): Logistic Regression And Time Series

Aafaq Iqbal khan
National College of Ireland
x20108851

ABSTRACT: In the terminal assignment of the Statistics of Data Analytics we worked on three datasets: overseas trip, house registration, and child birth data. On each dataset we apply different number of statistical models to reach the final model for each dataset. We analyze the two time series data and forecast their future values of three period ahead. One time series has both trend and seasonality while the other has only trend in data. Then estimate a logistic regression model to understand the relationships between the low weighted born babies characterizes. We use R for the statistical analysis of time series dataset as R has advanced visualization and data wrangling tools. For logistic Regression and Principal Component Analysis, we use SPSS software.

I. INTRODUCTION

This section gives the introduction and data description of each datasets that are used in the study. Each dataset explained in separate part.

Part A: House Registration

The house registrations data contains 43 records from 1978 to 2019. The csv file is read by appropriate library in R studio for further analysis. Then it is converted into time series object having starting year from 1978 and ended at 2019. The frequency of the time series is one because it is yearly data. The fig. 1 shows the time series plot of the house registration dataset. As we can see that there is no any seasonality in time series but there is a trend. From start of the series till 2005 there is an upward trend but after 2005 an abrupt decrease in registrations. We need to consider this while modeling the time series. It will be discussed later in this section.

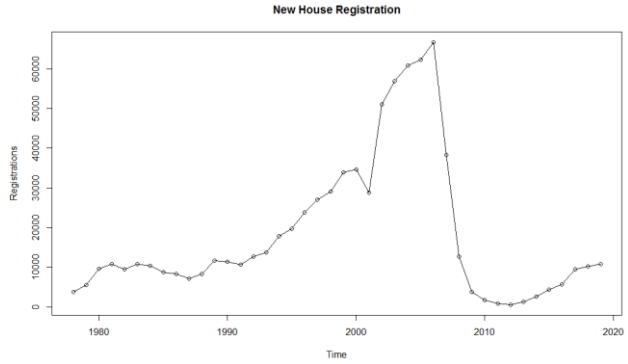


Fig. 1 Time series plot

Stationary Time Series Assumption:

To process the time series it should be stationary. It means time series should not have any trend or seasonality within the data. As from fig. 1 there is a clear trend so we need to remove this trend for further analysis. For that purpose, Difference function (diff) is used. The fig. 2 shows the results, the trend has been removed and now data points are around the horizontal line. So data is stationary now with no trend. This implies that data is independent of time. We tried the second order difference too but the result is almost identical so for simplicity we select the first order difference to remove the trend.

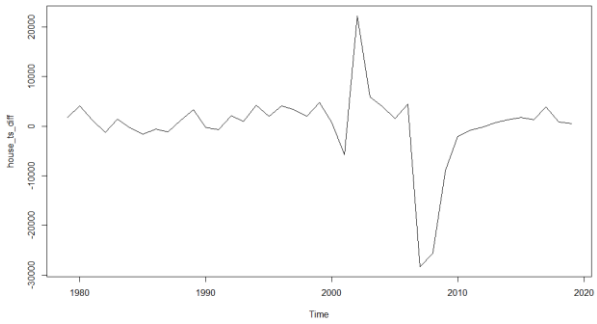


Fig. 2 Stationary time series

Part B: Overseas Trip

The overseas trip data contains 33 records from quarter 1 of 2012 to quarter 4 of 2019. The data has the information of the overseas trip to Ireland by non-residents during the mentioned time period. The data is given in csv file and it read in R studio for further analysis. Then convert the data into time series starting year from 2012 and ended at 2019. The frequency of the time series is four because it is quarterly data. The fig. 3 shows the time series plot of the overseas dataset. As we can see that there is a seasonality in time series as well as a trend. At start the seasonality difference is small but as it goes with upward trend the difference in seasonality is slightly increasing.

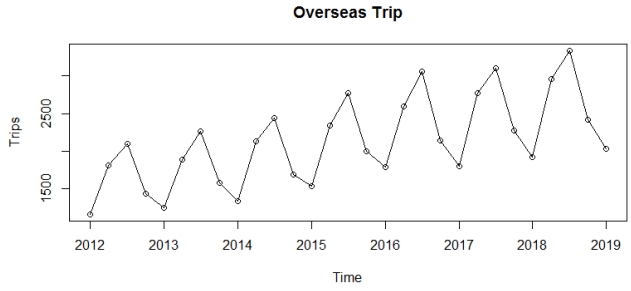


Fig. 3 Time series plot

Stationary Time Series Assumption

The data need to be checked whether it is stationary or not. As from fig. 3 there is a clear trend and seasonality so we need to remove this trend for further analysis. We perform augmented dickey-fuller test to check that data is stationary or not. As in fig. 4we have non-significant p-value 0.906 so our data is not stationary.

H_0	Time series is not stationary
H_1	Time series is stationary.

```
        Augmented Dickey-Fuller Test

data:  trip_ts
Dickey-Fuller = -1.1057, Lag order = 3, p-value = 0.906
alternative hypothesis: stationary
```

Fig. 4 Augmented Dickey-Fuller test

We use decompose function to decompose the time series into three parts: trend component, seasonal component, and irregular component (fig.5). As the seasonal fluctuations depend on the level of the time series so we use multiplicative decomposition. The classical decomposition have some disadvantages like few first and last missing values so we use other more robust method: Seasonal and Trend decomposition using Loess (STL). But the issue is the STL only works with additive models so we need to change our multiplicative seasonality to additive, we do with the log transformation.

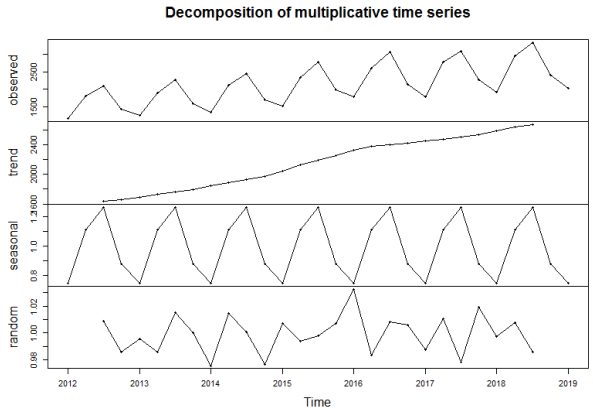


Fig. 5 Decomposition

By using seasonal plot graph, we can observe the variation or trend according to seasonality as we have four quarters as seasons from year 2012 to 2019. To remove the trend form time series we use diff() function with difference parameter value 1. The fig. 6 and 7 show the seasonal plots of time series with trend and without trend.

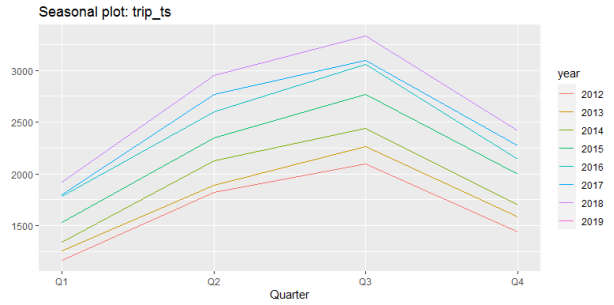


Fig. 6 Seasonal Plot with trend

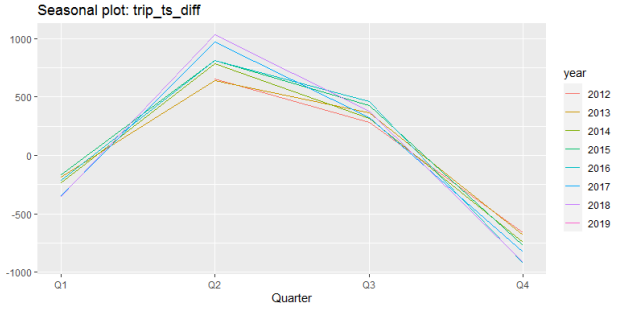


Fig 7. Seasonal Plot without trend

The fig.8 shows that quarter 3 and 2 have higher average number of overseas trips to Ireland by non-residents.

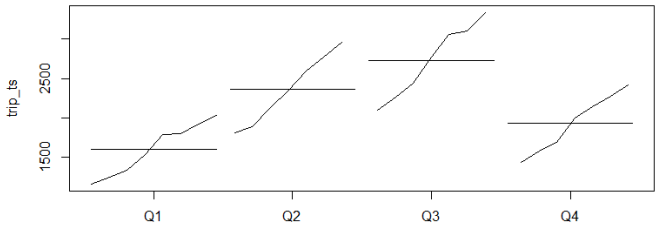


Fig. 8 Month plot

II. MODEL SELECTION AND DIAGNOSTICS ASSESSMENT:

Part A: House Registration

As our time series is non seasonal and has only trend so we use only non-seasonal time series models. In this part, different models will be applied on data and evaluate their performance. For better evaluation of the models we split the data into train and test. Data point from 1978 to 2012 are used for learning the models and test data ranging from 2013 to 2019 will be used to test the models on unseen data.

Naïve Model:

It forecast the next values on the basis of last value in time series. There are two types of models: *naïve* and *snaive*, we use naïve because our data have only trend not seasonality. The fig.9 shows that Root mean square error (RMSE) value is 8275 and 6728 on training and testing data respectively. As we can see in fig. 10 all predicted values are just equal to their previous values.

```
                    ME      RMSE      MAE       MPE     MAPE
Training set -149.6364 8275.963 4567.152 -17.39563 34.7574
Test set     5707.0000 6728.637 5707.000  83.43910 83.4391
```

Fig.9 Naïve Model summary

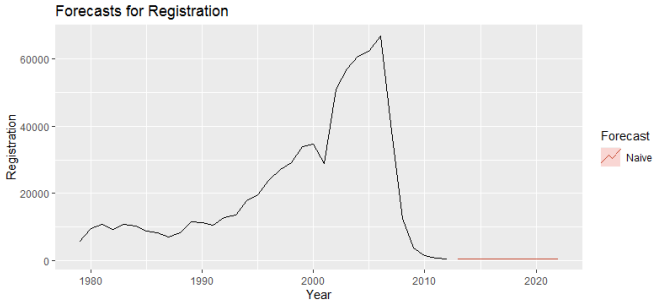


Fig. 10 Naïve Prediction Visualization

Exponential Smoothing Models

Initially we apply Holt model on our training data and evaluate them. The RMSE of the holt model on training and test data are 8283 and 26159 in turn. The high RMSE on the test data depicts that our model is poor to predict on unseen data. We give try to simple exponential smoothing model as well by eliminating the trend within the data by taking differences. For this purpose, we use order 1 difference and split them into test and train data as well. The RMSE scores do not improve much having the values 8235 on training and 10623 on test data. The fig. 11 summarize the prediction of the both models.

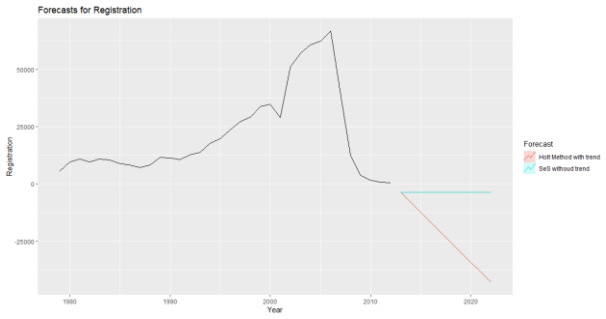


Fig. 11 Holt and SES Prediction Visualization

After that we apply the ets() function on the data to find the best exponential smoothing model by giving "ZZZ" parameter. It returns the parameter (M,N,N) it means that error type is multiplicative and without the seasonality and trend. The fig. 12 shows the summary, The RMSE on test data has been reduced which is a good thing so we check the other evaluations parameter and assumption for ETS (Error, Trend, and Seasonal) model.

	ME	RMSE	MAE	MPE	MAPE
Training set	-243.6381	8173.839	4531.518	-18.65672	35.50857
Test set	5706.9793	6728.620	5706.979	83.43855	83.43855

Fig.12 ETS model summary

The fig. 13 illustrates the residual plots of the ETS model, in Autocorrelation function (ACF) there is negative spike at lag 5 which is also significant as well. This shows that our model is not a final one. The residual distribution shows there are extreme values in residuals.

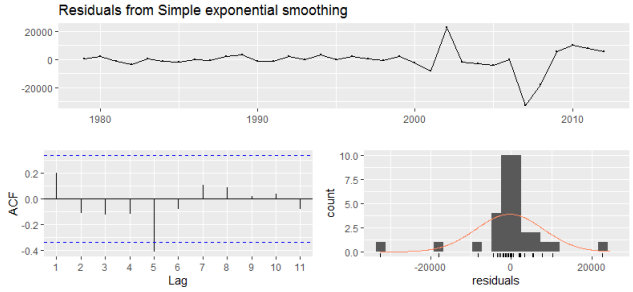


Fig. 13 ETS model Residual plot

The Q-Q plot of the residuals in fig. 14 tells us that residual is not normally distributed as it is depicting in distribution of the residuals.

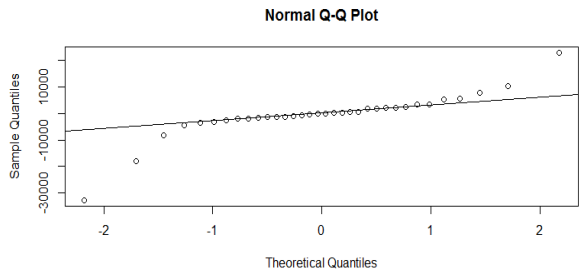


Fig. 14 Q-Q plot of residuals

To check whether the residuals of the model is stationary or not we performed Box-Ljung test on the residuals. As per Box-Ljung test residuals are stationary if we have non-significant p-value. As we have p-value 0.223 so residuals of the ETS model are stationary.

H_0	Residual are stationary
H_1	Residuals are not stationary.

```
        Box-Ljung test

data:  m_ses$residuals
X-squared = 1.4842, df = 1, p-value = 0.2231
```

Fig. 15 Box Ljung Test

Challenge

In time series, we have an upward trend but there is a sudden decrease in trend at the end of the time series. This sudden decrease may be the cause of high RMSE because our model is trained on upward trend data. We tried to reduce the effect of this sudden decrease in trend by taking the log of the time series and plot the log transformed time series but the pattern of the time series does not changed much (fig. 16). So we continue with the original data and drop the log transformed time series for further analysis.

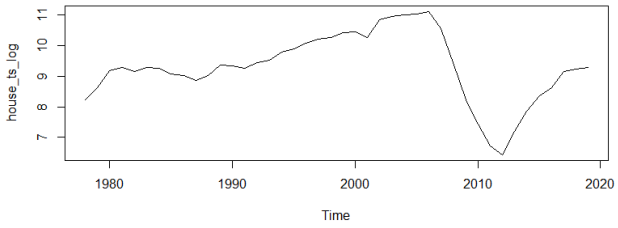


Fig. 16 Log transformed time series.

ARIMA MODEL

Initially, we apply the ARIMA model manually by giving the appropriate value of the p, q, and d to find the best prediction ARIMA model.

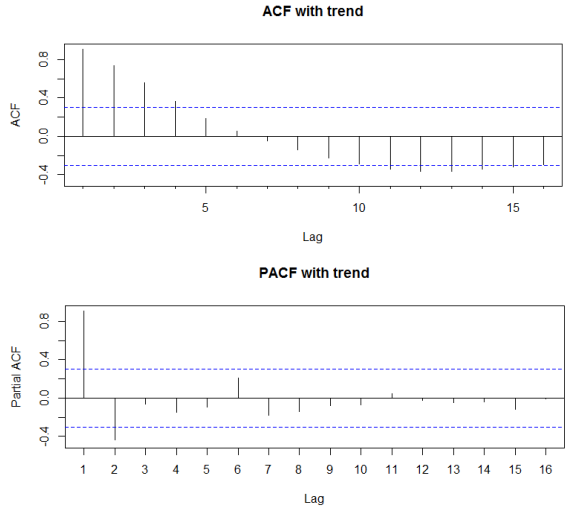


Fig. 17 ACF and PACF plots

Firstly, autocorrelation (ACF) and partial autocorrelation (PACF) graphs are plotted to find the value of 'p' and 'q'. In fig. 17 we can see that PACF has one significant positive spike and one negative spike at lag 1 and 2 respectively. In ACF, at starting there are 4 significant spikes but they are geometrically decaying after lag 1. So as ACF has a geometric decline and PCAF has two significant spikes so it is Auto Regressor AR (1) model.

So we apply ARIMA model with order (2,0,0) it means p is 2 as per PACF plot, difference we use is 0, and q will be

zero. The summary of the model is given in fig 18. It performed well on test data and have improved RMSE of 5317.

	ME	RMSE	MAE	MPE	MAPE
Training set	264.378	7020.728	4068.252	-23.55391	39.49965
Test set	-5062.630	5317.717	5062.630	-100.24324	100.24324

Fig. 18 Model Summary ARIMA (2,0,0)

But when we plot the residuals of the model (fig. 19), we can see a significant negative lag at 5. We need to remove this we can remove this by giving q =5.

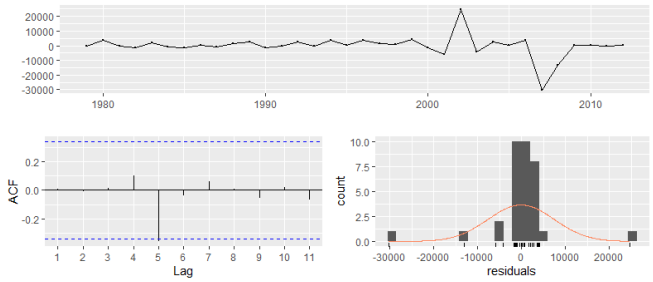


Fig. 19 Residual plot of ARIMA (2,1,0)

The fig. 20 shows that spike at lag 5 has been removed and the RMSE of the model are 5847 and 7341 for training and testing data in turn.

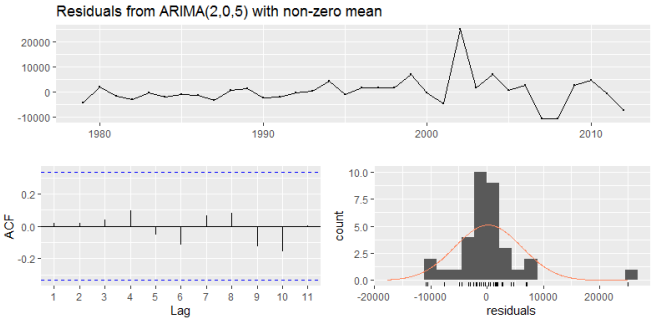


Fig. 20 Residual plot of ARIMA (2,0,5)

The Q-Q plot in fig. 21 shows that residual is not much normally distributed due to the some extreme values.

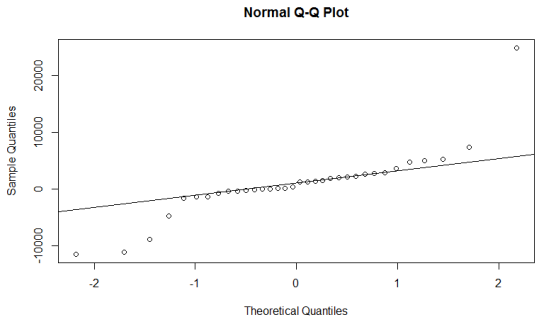


Fig. 21 Q-Q plot of ARIMA (2,1,7)

As the p-value in Box-Ljung test (Fig. 22) is non-significant that is 0.9589 so the assumption of the stationary residuals is satisfied.

```
        Box-Ljung test

data:  fitF$residuals
X-squared = 0.0026513, df = 1, p-value = 0.9589
```

Fig. 22 Box Ljung Test

Part B: Overseas Trip

The overseas trip time series is a seasonal data with having trend as well, so we use seasonal time series models. We try non-seasonal models as well by removing the seasonal component from the time series. In this part, different models will be applied on data and evaluates their performance. For better evaluation of the models we split the data into train and test. Data point from 2012 Q1 to 2017 Q4 are used for learning the models and test data ranging from 2018 Q1 to last will be used to test the models on unseen data.

Snaïve Model

For overseas time series we use snaive model because of the seasonality in data. In snaive model, the forecast value will be equal to last observed value from the same season of the year. The fig. 23 shows that Root mean square error (RMSE) value is 191.4 and 188.4 on training and testing data respectively. The difference of RMSE and MAPE of training and testing are almost same so can safely say our model is not overfitting.

	ME	RMSE	MAE	MPE	MAPE
Training set	170.745	191.4607	170.745	8.077620	8.077620
Test set	182.980	188.4195	182.980	7.398614	7.398614

Fig.23 Snaïve Model summary

Exponential Smoothing Models

Initially, we apply Holt-Winter model on our training data and evaluate it. The Holt-Winter model is useful when data has both trend and seasonal components The RMSE of the Holt-Winter model on training and test data are 47.1 and 42.1 in turn. There is big improvement in the RMSE and MAPE than simple Snaive model. So Holt-Winter model performs well on time series.

	ME	RMSE	MAE	MPE	MAPE
Training set	-1.766807	47.16768	40.54411	-0.2066754	2.048138
Test set	-28.609180	42.19357	31.64246	-1.1465633	1.272186

Fig. 24 Snaïve Model summary

Now we try to forecast the time series by removing the seasonal component from the data and apply Holt model on it. Then we compare the results between Holt-Winter model and Holt model without seasonality. We remove the seasonal component by seasadj() function in R. The fig. 25 show the time series without seasonality.

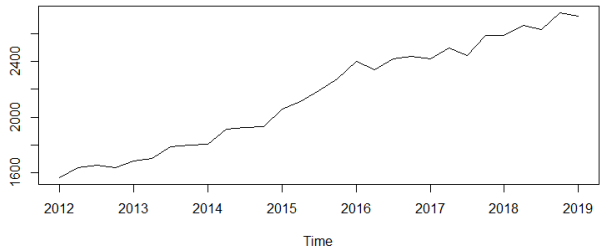


Fig. 25 Without seasonal component

We again split the resultant time series into train and split and apply Holt model on it. The RMSE on train is 49.7 and on test is 38.6. The fig. 26 shows the forecast result of the both Holt-Winter and Holt model. This infer that Holt-winter performs well because it consider the seasonality part as well.

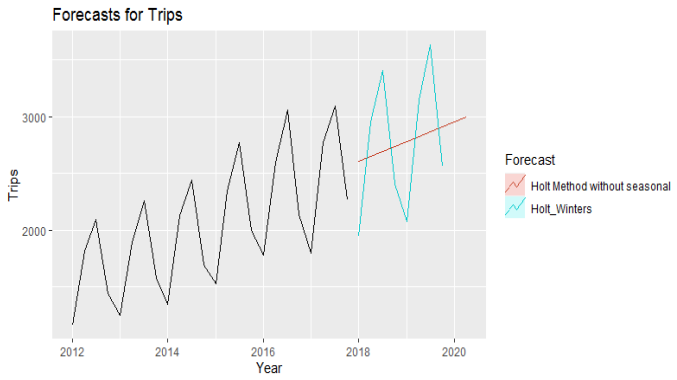


Fig. 26 Holt-Winter and Holt prediction

The residual plots in fig. 27 shows no significant lag in autocorrelation function graph and the residual distribution is also good with having no extreme values. The stationary assumption of the residuals is also satisfied as the p value of the Box-Ljung test is 0.887 which is non-significant so we do not reject null hypothesis.

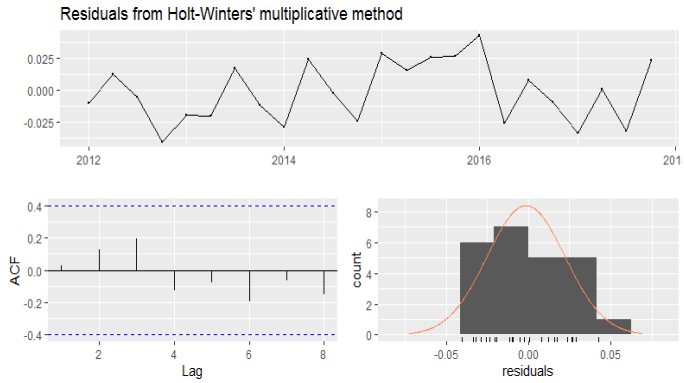
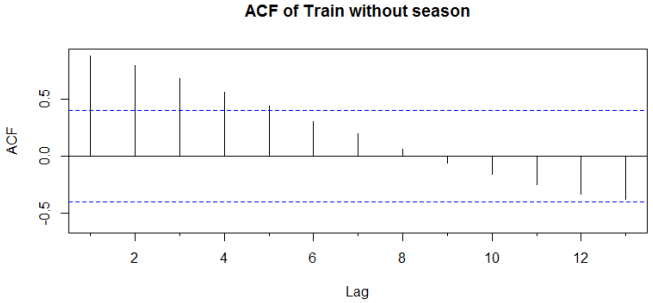


Fig. 27 Residuals plot of Holt-Winter Model

ARIMA MODEL

We apply three ARIMA model on time series, firstly we manually select the P and Q values from ACF and PACF and the D value from ndiffs() function in R. Secondly, we auto ARIMA without seasonality by giving FASLE to seasonal parameter. Lastly, auto Seasonal ARIMA model is use with seasonal component.

For first ARIMA model, the appropriate value of the p and q from ACF and PACF are p= 1 and q=0. As there is a geometric decay in ACF and significant spike at lag 1 in PACF so we use AR (1) model. The order of ARIMA model is (1,1,0) here p=1 , d=1 (find by ndiffs function), and q=0. The resultant RMSE on train and test data are 68 and 88 respectively.



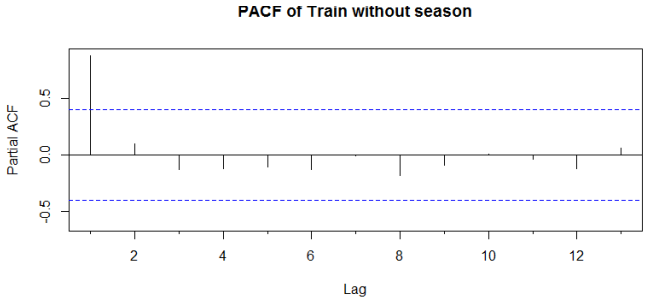


Fig. 28 ACF and PACF plots

Now we try our second ARIMA model, by using auto ARIMA model without seasonality. The auto ARIMA suggests the same order that we used earlier that is (1 ,1, 0) but Auto ARIMA suggest to use the order with enabling the drift. This model performs well on testing and training data. The fig. 29 summarize the evaluation. It has good RMSE score on both test and train data.

	ME	RMSE	MAE	MPE	MAPE
Training set	0.4338002	48.75616	40.04317	-0.04156571	1.9191638
Test set	-13.2841297	31.63162	25.43687	-0.50136579	0.9489146

Fig. 29 Model Summary Auto ARIMA (1,1,0)

The fig. 30 plots the forecast of both ARIMA models on time series without the seasonal component. The Auto ARIMA model with drift performs well. The residuals of the model is also stationary as the p-value of the Box- Ljung test is 0.88 which is non-significant.

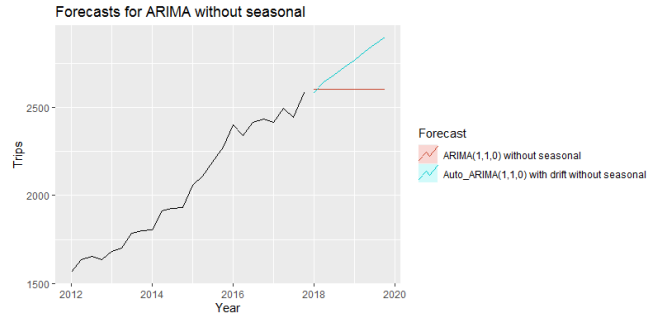


Fig. 30 ARIMA models forecast

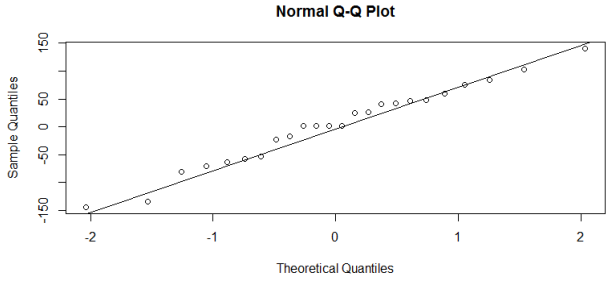
At the last, we use auto ARIMA model with consideration of the seasonal component as well. The order we get from the model is [ARIMA (1,0,0) (0,1,0)[4] with drift]. The RMSE both on test and train are little bit higher than we got from ARIMA models without seasonality. The fig. 31 summarize the model evaluation parameter.

```
fitc <- auto.arima(train, seasonal = TRUE)
arima_f_C =forecast(fitc)
accuracy(arima_f_C, test)
```

	ME	RMSE	MAE	MPE	MAPE
Training set	2.154051	69.77208	55.84005	-0.2152261	2.691669
Test set	-7.163904	53.26764	45.85145	-0.6872255	1.885583

Fig. 31 Model Summary Auto ARIMA with seasonal

The model has no issue with residuals. The distribution of the residual are fine and there is a significant lag in ACF graph of residuals. The fig. 32 shows Q-Q plot of the residuals which shows a striagth diagonal line that depicts that residual are normally distributed. The p-value of Box-Ljung test is also non-significant so residual are stationary as well.




```
        Box-Ljung test

data:  arima_f_C$residuals
X-squared = 0.05151, df = 1, p-value = 0.8205
```

Fig.32 Q-Q plot and Box-Ljung Test

Part C: Child Birth Dataset

Binary Logistic Regression is a Classification Algorithm in supervised learning technique. Binary Logistic Regression is use to predict dichotomous dependent variable having only 2 possible outcomes with the help of one or more independent predictors. As our dependent variable is also dichotomous i.e. '1' for low birth weight and '0' for normal birth weight so binary logistic regression is appropriate. The objective of the study is to find the best model that describe the relationship between the dependent variable and minimum number of predictors. Moreover, dimensionality reduction technique Principal Component Analysis (PCA) is also used to reduce independent variables into factors based on the Eigenvalue.

Dataset:

The dataset used is related to the 'Child birth survey in Ireland'. The original dataset consists of 42 records and 16 columns that have the different information of the child birth in Ireland e.g. gestational Age, mother and father age, length of the new born baby etc.

Checking the Data Assumptions

Dichotomous Target Variable: our dependent variable "lowbwt" has categorical values '0' and '1', which satisfies this assumption.

***Collinearity and Multi-Collinearity*:** Before performing the analysis, the possibility of existence of any multicollinearity with in the data should be checked. There should be not any multicollinearity with in the predictors. The fig. 33 shows the collinearity tolerance and statistics VIF of each independent variables. As we can see from the figure, all variables have collinearity tolerance above 0.10 and the statistics VIF less than 10, it means that there is no any multicollinearity with in data so absence of multicollinearity assumption is satisfied.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	2.244	1.691		1.327	.196		
	Length	-.059	.026	-.492	-2.317	.028	.280	3.571
	Birthweight	-.021	.148	-.035	-.139	.890	.197	5.078
	Headcirc	-.009	.026	-.063	-.356	.724	.403	2.481
	Gestation	-.024	.028	-.177	-.855	.400	.296	3.383
	smoker	.112	.133	.160	.841	.408	.350	2.857
	mage	-.015	.021	-.242	-.719	.478	.111	8.997
	mnocig	-.006	.005	-.209	-1.126	.270	.366	2.735
	mheight	.010	.011	.187	.942	.354	.319	3.134
	mppwt	-.010	.009	-.199	-1.127	.270	.405	2.469
	fage	.001	.014	.014	.051	.960	.163	6.140
	fedyrs	.000	.024	.001	.008	.993	.589	1.697
	fnocig	.005	.003	.242	1.574	.127	.535	1.871
	fheight	.008	.008	.150	.985	.333	.544	1.840
	mage35	.500	.246	.420	2.029	.052	.295	3.390

a. Dependent Variable: lowbwt

Fig. 33 Multicollinearity Check

The fig. 34 shows the correlation coefficient of predictors. All predictors are not strongly correlated to each other this is also suitable for our analysis.

Correlation Matrix

	Constant	Length	Birthweight	Headcirc	Gestation	smoker	mage	mnocig	mheight	mppwt	fage	fedyrs	fnocig	fheight	mage35
Constant	1.000	.245	-.388	-.346	.514	-.546	-.382	.032	-.835	.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	.763	.034
Headcirc	-.346	.007	-.101	1.000	-.353	.406	.196	.086	.097	.162	-.279	.076	-.263	-.067	-.260
Gestation	.514	-.153	-.559	-.353	1.000	-.668	.106	.234	-.285	.229	-.456	.197	.453	-.543	.284
smoker	-.546	-.235	.614	.406	-.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	-.128	-.017	-.368	-.235	-.347	-.094	-.388
mheight	-.835	-.455	.407	.097	-.285	.494	.310	-.128	1.000	-.757	.141	-.469	-.563	.451	-.101
mppwt	.610	.093	-.312	.162	.229	-.224	-.455	-.017	-.757	1.000	.013	.451	.418	-.355	.008
fage	-.170	-.201	.420	-.279	-.456	.208	-.670	-.368	.141	.013	1.000	-.165	-.120	.581	-.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	.763	-.067	-.543	.398	-.010	-.094	.451	-.355	.581	-.771	-.706	1.000	-.190
mage35	.370	-.121	.034	-.260	.284	-.143	-.375	-.388	-.101	.008	-.046	.271	.567	-.190	1.000

Fig. 34 Correlations

Sample Size: The sample size use for logistic regression should be appropriate in size. Ideally, there should be 20 records per predictor in model. But in our data, we have 42 cases in total with 15 predictors. A small number of rows with large number of predictors may be a problematic issue for the performance of the model. So, sample size assumption is not fulfilled if we use all available predictors in model to predict the target variable.

Check for Outliers: The presence of outliers in data can effects the prediction power of the logistic regression model. So data was checked for outlier's detection and no outlier detection was found.

Like linear regression, Logistic regression has no assumptions about the normal distribution of data. Now we apply three approaches to reach the our best model.

A. Technique 1: Raw Model

Initially, we use all the variables as predictors except the "*id*" and as well as "lowbwt" which is our target variable. Next, we use different evaluation test and techniques to measures the model accuracy and assumptions verifications. At first, we verify the size and nature of the data, in the fig. 35 we can see that there are no missing values and the total cases that were used to train the model is 42. The fig 36 confirm the dichotomous encoding of our target variable.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	42	100.0
	Missing Cases	0	.0
	Total	42	100.0
Unselected Cases		0	.0
Total		42	100.0

Fig. 35 Case Processing Summary

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Fig. 36 Dependent Variable Encoding

The fig. 37 shows the result of the Omnibus Tests of Model Coefficients that explains how good our logistic regression model performs with none of the independent variable entered the model referred to as the 'goodness of fit' test. The p-value

is 0.002 which is significant which indicate that our model is better than the model with not having any predictor.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	34.450	14	.002
	Block	34.450	14	.002
	Model	34.450	14	.002

Fig. 37 Omnibus Test

This Hosmer and Lemeshow test is use to check the assumption of Goodness-of-fit of the logistic regression model. It tells how well our data fits the model. A logistic regression with null hypothesis of Hosmer and Lemeshow test is considered a good model. In fig. 38 the p-value is non-significant so our model has a good fit.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	8	1.000

Fig. 38 Hosmer and Lemeshow test

The Cox & Snell R Square and the Nagelkerke R Square is used to distraction in response variable or the variation in target variable explained by the model with set of predictors. In fig. 39, Cox & Snell R Square has value 0.56 and Nagelkerke R Square with value of 1.00 tells us that 56% to 100% variance in the dependent variable is explained by the set of predictors.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.560	1.000

Fig. 39 Model Summary

In our data, there are 36 cases with normal birth weight babies and 6 bases have low birth weight issue. Our model classify the all cases correctly as model has 100% accuracy shown in fig. 40.

Classification Table[a]

			Predicted		
			lowbwt		Percentage Correct
	Observed		0	1	
Step 1	lowbwt	0	36	0	100.0
		1	0	6	100.0
	Overall Percentage				100.0

a. The cut value is .500

Fig. 40 Classification Table

Our raw model has 100% accuracy along with 56% to 100% variance in response variable is explained by the all predictors but as we can see in fig. 41 none of the predictor has significant p-value it means that with having all variables in the model no predictor is significant to predict the outcome. So we should try to use less number of predictors to build the model that has all significant independent variables and fulfill all the assumptions as well.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Length	-7.134	4923.455	.000	1	.999	.001
	Birthweight	-22.242	28359.429	.000	1	.999	.000
	Headcirc	2.828	4135.322	.000	1	.999	16.904
	Gestation	1.279	5968.790	.000	1	1.000	3.594
	smoker	4.147	22865.360	.000	1	1.000	63.275
	mage	-3.350	3722.529	.000	1	.999	.035
	mnocig	-.651	996.798	.000	1	.999	.521
	mheight	.537	1664.341	.000	1	1.000	1.711
	mppwt	-.616	1569.976	.000	1	1.000	.540
	fage	.624	2840.621	.000	1	1.000	1.866
	fedyrs	1.524	5765.585	.000	1	1.000	4.590
	fnocig	.508	948.748	.000	1	1.000	1.662
	fheight	.506	1460.868	.000	1	1.000	1.658
	mage35	74.712	50240.019	.000	1	.999	2.800E+32
	Constant	153.468	231028.139	.000	1	.999	4.469E+66

Fig. 41 Variables in Equation

Type B: Manually selected Predictors

When we explore the data some of the columns are providing same information so we need to exclude the predictors that have repetitive information for the model. For example "*BirthWeight* " and our response variable "lowbwt" carry same information. Those babies that have lower weight than 2.7 kg is classified as '1' in 'lowbwt' variable. So we need to exclude the "BirthWeight" columns from model. Similarly, "*mage*" and "*mage35*" have same details, those mother who have age more than 35 is categorize "1" in "mage35" column. So, we use only one from them we select "mage35" that have only 2 values 1 and 0.

In "*smoker*" variable, 1 represents mother smokes and 0 represents not a smoker. The variable "*mnocig*" tells the number of cigarettes consumed by the mother in a day. Both "smoke" and "*mnocig*" have equivalent information so we select only "*smoker*" variable for our model as the important thing to get to know whether a mother smokes or not. The length of the baby is also play a vital role in determine the weight of the baby so we include "*length*" as well in our model. On intuitive basis, we select "*gestation*", "*mppwt*" (Mothers pre-pregnancy weight) and "fnocig" (number of cigarettes consumed by the father in a day)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Length	-.691	1.098	.396	1	.529	.501
	Gestation	-1.343	1.411	.906	1	.341	.261
	mppwt	-.351	.339	1.072	1	.301	.704
	smoker(1)	-8.016	12.681	.400	1	.527	.000
	fnocig	.004	.072	.003	1	.954	1.004
	mage35	1.988	18.480	.012	1	.914	7.299
	Constant	103.242	77.151	1.791	1	.181	6.880E+44
Step 2[a]	Length	-.709	1.068	.441	1	.506	.492
	Gestation	-1.344	1.405	.915	1	.339	.261
	mppwt	-.352	.341	1.068	1	.301	.703
	smoker(1)	-8.174	12.783	.409	1	.523	.000
	mage35	1.935	18.602	.011	1	.917	6.927
	Constant	104.352	75.429	1.914	1	.167	2.086E+45
Step 3[a]	Length	-.728	1.058	.474	1	.491	.483
	Gestation	-1.370	1.414	.940	1	.332	.254
	mppwt	-.356	.347	1.048	1	.306	.701
	smoker(1)	-8.380	13.256	.400	1	.527	.000
	Constant	106.496	75.145	2.008	1	.156	1.781E+46
Step 4[a]	Gestation	-1.868	1.028	3.300	1	.069	.154
	mppwt	-.401	.282	2.019	1	.155	.670
	smoker(1)	-7.561	8.078	.876	1	.349	.001
	Constant	91.721	50.626	3.282	1	.070	6.821E+39

Fig. 42 Variables in Equation

So we include all the above discussed variables in model as predictors and use backward conditional Logistic Regression approach. This will exclude the irrelevant predictor

iteratively from the model and give us a good model. The fig. 42 shows that there are 4 iterations in the backward conditional model and it eliminates the irrelevant predictors from model at each step and at the end we have only 3 predictors "*gestation", "mppwt*", and "*smoker*".

Cox & Snell R square and Nagelkerke R square are used to determine the range of the percentage of variance in dependent variable explained by the predictors. It is explained from 48% to 86%, which is lower than our raw model which has the range from 56% to 100%.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	6.120[a]	.491	.877
2	6.124[a]	.491	.877
3	6.137[a]	.490	.876
4	6.673[b]	.484	.865

Fig. 43 Model Summary

The accuracy of the model with only three predictors is 96%. The true positive rate (sensitivity) of the last model is 83% and true negative rate (specificity) is 100%.

Classification Table[a]

			Predicted		
			lowbwt		Percentage Correct
	Observed		0	1	
Step 1	lowbwt	0	36	0	100.0
		1	1	5	83.3
	Overall Percentage				97.6
Step 2	lowbwt	0	36	0	100.0
		1	1	5	83.3
	Overall Percentage				97.6
Step 3	lowbwt	0	36	0	100.0
		1	1	5	83.3
	Overall Percentage				97.6
Step 4	lowbwt	0	36	0	100.0
		1	1	5	83.3
	Overall Percentage				97.6

Fig. 44 Classification table.

PRINCIPLE COMPONENT ANALYSIS:

In backward conditional logistic regression approach, we get the final model that has only three predictors and also has a good accuracy 97%. We will try to find better model by doing principle component analysis on all available variables excluding "*birthWeight*" because it is equivalent to our response variable. The PCA is a dimensionality reduction technique that reduce the dimensionality of the data by finding the important factors within the data.

First of all check whether our data is suitable for PCA or not for that purpose we use Kaiser-Meyer-Olkin Measure of sampling adequacy (KMO) and Bartlett's Test of Sphericity. In fig. 45 we can see that KMO has value of 0.552 but for the PCA it should be near or greater than 0.6. As 0.552 is near to 0.6 so can assume that condition is satisfied. Moreover, Bartlett's Test has significant p-value that mean data is suitable for PCA.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.552
Bartlett's Test of Sphericity	Approx. Chi-Square	249.411
	df	78
	Sig.	.000

Fig. 45 KMO and Barlett's Test

In the fig. 46 the information about the communalities are given as we can see all variables have extraction value larger than 0.5 so it is good for our PCA.

Communalities

	Initial	Extraction
Length	1.000	.804
Gestation	1.000	.806
smoker	1.000	.795
mppwt	1.000	.782
fnocig	1.000	.791
mage35	1.000	.575
Headcirc	1.000	.661
mnocig	1.000	.777
mheight	1.000	.898
fage	1.000	.860
fedyrs	1.000	.692
fheight	1.000	.596
mage	1.000	.886

Fig.46 Commonalities Table

The fig. 47 tells us that only 5 factors are enough to explain the 76.33% of variance in the data. For factor analysis the explained variance should be at least 60% our data fulfilling the requirement.

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.259	25.068	25.068	3.259	25.068	25.068	2.487	19.128	19.128
2	2.596	19.968	45.036	2.596	19.968	45.036	2.206	16.969	36.098
3	1.950	15.001	60.037	1.950	15.001	60.037	2.080	16.003	52.101
4	1.100	8.463	68.499	1.100	8.463	68.499	1.826	14.044	66.145
5	1.019	7.837	76.337	1.019	7.837	76.337	1.325	10.192	76.337
6	.834	6.415	82.752						
7	.714	5.492	88.244						
8	.529	4.069	92.313						
9	.335	2.580	94.893						
10	.226	1.735	96.628						
11	.216	1.659	98.287						
12	.160	1.234	99.521						
13	.062	.479	100.000						

Fig. 47 Total Variance Explained

As per the scree plot (fig. 48), the first 3 factors explain most the variance in data than rest of the factor. So for simplicity we can take only 3 factors rather than all factor for our model building step. The total variance explained by the first three factors are 61% which is greater than 60 percent threshold.

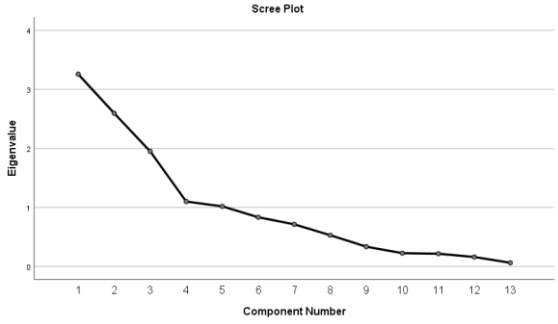


Fig. 48. Scree plot

Logistic Regression on PCA factors:

We apply binary logistic regression on first three factors that we got from PCA. In fig. 49 the accuracy of the model is 95 % with sensitivity 83% and true negative rate (specificity) 97%. The accuracy is lower than our manually selected predictor model which have 97% accuracy.

Classification Table[a]

			Predicted		
			lowbwt		Percentage Correct
	Observed		0	1	
Step 1	lowbwt	0	35	1	97.2
		1	1	5	83.3
	Overall Percentage				95.2

Fig. 49 Classification table

The value of the Cox $ Snell R-squared is 0.373 and Nagelkerke R-squared is 0.667 (fig. 50). The variance in target variable is explained by the model in the range of 37% to 66.7%.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	14.818[a]	.373	.667

Fig. 50 Model Summary

The variables in logistic regression equation is mentioned in fig. 51 only one factor has significant P-value.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	REGR factor score 1 for analysis 1	-.528	.594	.791	1	.374	.590
	REGR factor score 2 for analysis 1	-2.527	.975	6.721	1	.010	.080
	REGR factor score 3 for analysis 1	1.249	.817	2.335	1	.126	3.486
	Constant	-3.956	1.505	6.909	1	.009	.019

Fig. 51. Summary

III. SUMMARY AND FORECAST

Part A: House Registration

By evaluating the models on the basis of Root mean square error (RMSE), the model which has low RMSE value is best performing models. We consider the RMSE value on both train and test data while selecting the final model. The Table.1 summarizes the all models and the ARIMA (2,0,0) has lower RMSE on testing data and ARIMA (2,0,5) has lease RMSE on Training data. For simplicity, we select ARIMA (2,0,0) having lowest RMSE on test data.

Model	Naive	Holt	SES without trend	ETS	ARIMA (2,0,0)	ARIMA (2,0,7)
RMSE Train	8275	8283	8235	8173	7020	5847
RMSE Test	6728	26159	10623	6728	5317	7341

Table 1. House Models summary

It is use to forecast the house registrations in three period ahead. The fig. 52 shows the forecast of upcoming years 2020, 2021, and 2022. We can see there is a smooth upward trend in house registrations. The dark gray and light gray represents the 80% and 95% confidence interval. As a conclusion, from 2020 to 2022 there is a gradual increase in the number of registrations of house.

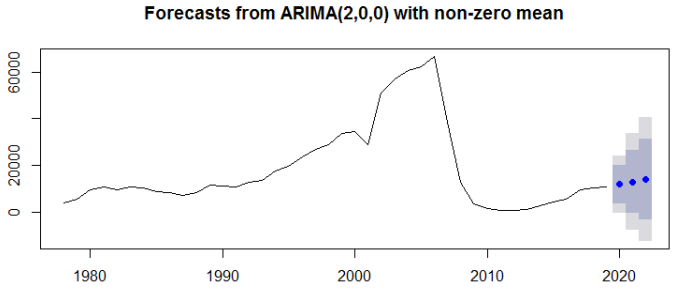


Fig. 52 three period ahead forecast

Part B: Overseas Trip

Again, we select the best model on the basis of the lowest RMSE of the models and we consider the RMSE value both on test and train data while deciding the final model. The Table. 2 summarizes the all models that we used on overseas trip time series. The Holt-Winter has lower RMSE on training data and Auto ARIMA without seasonal having order (1,1,0) has least RMSE on Test data. As we can see, the RMSE difference on training data between Holt-winter and Auto ARIMA without seasonal is not much bigger so we can safely use Auto ARIMA without seasonal as our final model because it has good RMSE on both test and train data and as explained in pervious section it also has stationary residuals and normal distribution in residuals as well.

Model	Snaive	Holt-Winter	Holt without Seasonal	ARIMA (1,1,0)	Auto ARIMA without season	Seasonal Auto ARIMA
RMSE Train	191.4	47.1	49.7	68.4	48.3	69.1
RMSE Test	188.4	42.1	38.6	88.2	31.6	53.3

Table 2. Overseas Trip Models summary

Our final model Auto ARIMA without seasonal will be use to forecast the overseas trips to Ireland by non-residents three period ahead. The fig. 53 shows the forecast of upcoming quarters 2020 Q1, Q2, and Q3. We can see there is a smooth increase in seasonality with level of overseas trip time series. The dark gray and light gray represents the 80% and 95% confidence interval.

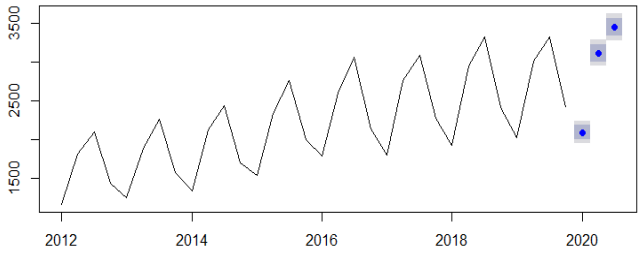


Fig. 53 Three period ahead forecast

Part C Child Birth Data

After applying three approaches for logistic regression, as per in table 3, we conclude that the logistic regression model with selected predictors performs better the accuracy of the raw model is 100 percent but does not have any significant predictor. So we go with the selected predictor model that have slightly less accuracy than raw model but overall it performs well.

Model	Accuracy	Cox $ Snell R-squared	Nagelkerke R-squared	Hosmer and Lemeshow Test	# of significant predictors
Raw Model	100	0.56	1	1	0
Selected Predictor	97	0.48	0.86	0.74	1
After PCA	95	0.37	0.66	0.73	1

Table 3 Logistic Regression models summary

The equation of the final model is

$$Log(p/1-p) = 91.721 - 1.868 * Gestation - 0.401 * mppwt - 7.561 * smoker$$

The interpretation of the equation is that, with the one unit increase in gestation the log odds will decrease by the 1.868. As we can see in equation, smoker has the higher coefficient so it is the most dominated factor. As smoker variable has only two levels: 1 means mother smokes and 0 means mother is not a smoker. If a mother smokes than the smoker variable will be present in the equation and cause the 7.56 unit change. If mother does not smoke than smoker variable will be the part of the equation.

BIBLIOGRAPHY

[1] Cock, D. D. (2011). Ames, iowa: Alternative to the boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3)

[2] Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL: http://CRAN.R-project.org/doc/Rnews/

[3] Thomas Lumley using Fortran code by Alan Miller (2009). leaps: regression subset selection. R package version 2.9. http://CRAN.R-project.org/package=leaps

[4] H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009