Statistical Analysis of Time Series and Logistic Regression

Aditya Kumar Singh
Data Analytics
National College of Ireland
x20140410@student.ncirl.ie

***Abstract*—The document is a CA submission which covers the analysis of Time Series and Logistic Regression models. The first dataset contains overseas trip quarterly from 2012 to 2019. Second dataset is a yearly data of new house registration from 1978 to 2019. Third dataset contains data from children born in a city from USA. For this submission IBM SPSS 27 software and R language was used was used.**

***Index Terms*—time series, logistic regression, ARIMA model, holt winter**

I. Part A: Time Series

A time series data is set of observations sampled periodically at equal interval to a form a chronological order[1]. In this submission we have selected two time series data for analysis.

A. *OverseasTrips*

1) Dataset: The dataset contains the number of overseas trip to Ireland by non-residents. The Observation are sampled quarterly from 2012 Q1 to 2019 Q4. We plot an initial sequence chart to get the rough idea of the raw data.

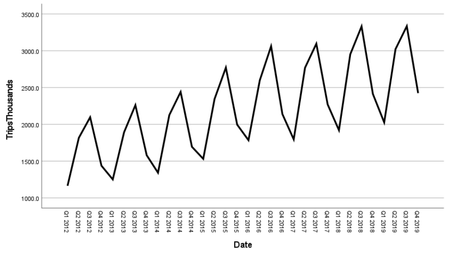


Fig. 1. Overseas Trip Sequence Chart

The figure above shows the number of trips quarterly. On initial inspection it is clear that the there is some seasonality present in the chart as well as some element of trend.

2) Initial Analysis: To check the components of raw data, initial decomposition was conducted. This is done to check Trend, Cycle and Seasonality. R provides a simple code to perform decomposition. The following graphs are obtained as shown below.

The graph above is an observed sequence chart of the raw data. We check if the data is multiplicative or additive. As seen there is a slight growth each season as compared to previous season. We can safely assume that the data shows multiplicative tendencies.

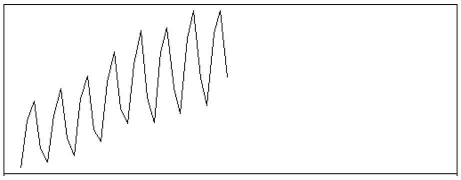


Fig. 2. Seasonal Decomposition graph 1

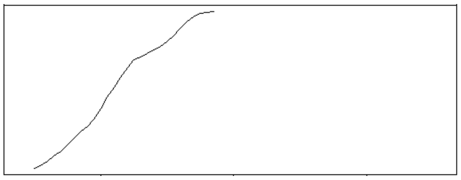


Fig. 3. Seasonal Decomposition graph 2

The graph above is plotted to find the trend in the data. As predicted from the the original sequence chart, there is an upward trend in the data. We can safely assume that the data shows trend. Another parameter that can be also checked is if there is any cycle present in the time series. There is not any cyclic tendency present.

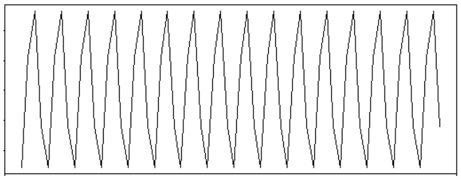


Fig. 4. Seasonal Decomposition graph 3

The graph above shows the seasonality in the data. As can be clearly seen, there is seasonality present in the data.

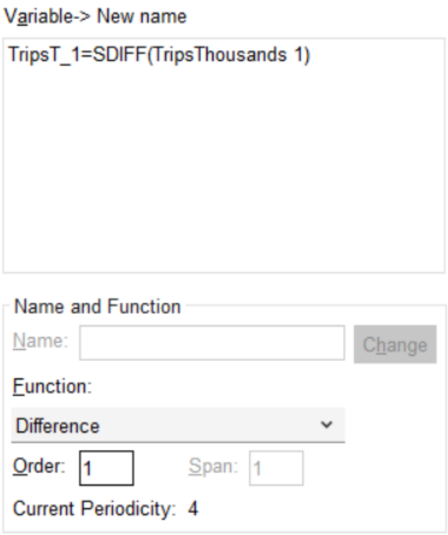
From the above graphs we have elucidated the following components of the data.

- The time series is multiplicative.
- The time series is seasonal.
- The time series has trend hence non-stationary.
- The time series is NOT cyclic.

3) Implementation: Based on the components of data, several forecasting models were selected. For this time series, the following models were finalized based on the components.

- Simple Seasonal
- Holt Winter's Multiplicative
- Seasonal ARIMA

1) Simple Seasonal:
Simple seasonal model is an appropriate model for time series that exhibit seasonal components but no trends. Special mention to the fact that simple seasoning is a good model when there is no trend.



Variable-> New name

TripsT_1=SDIFF(TripsThousands 1)

Name and Function

Name:

Change

Function:

Difference

Order: 1

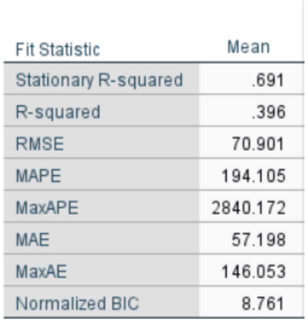
Span: 1

Current Periodicity: 4

Fig. 5. First order differencing for Simple Seasonal Model

Considering this, it was necessary to make the time series stationary before implementing the model. To make a time series stationary differencing is performed on the data. Differencing is a method of transforming the time series dependence on time. For seasonal model, first order differencing was performed before fitting the model as can be seen in the figure above which is a snippet from SPSS while performing the function. As can be seen a order one of difference was selected in the option before fitting the model.

After performing the necessary, the model was fitted to check the fit parameters. As can be seen in Figure 6, the RMSE value of 70.901 was able to be obtained. Also to notice the Mean Absolute Error(MAE) of 57.198 obtained.



Fit Statistic	Mean
Stationary R-squared	.691
R-squared	.396
RMSE	70.901
MAPE	194.105
MaxAPE	2840.172
MAE	57.198
MaxAE	146.053
Normalized BIC	8.761

Fig. 6. Model Fit Statistics for Simple Seasonal

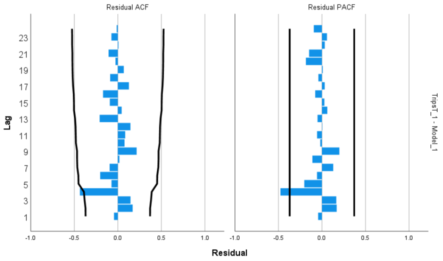


Fig. 7. Residual ACF and PACF

Plotting the Residual Auto-correlation(ACF) and Partial Auto-correlation(PACF) to check the lag cutting. For ACF, the lag cuts at 3 which suggest a moving average of order one can be performed.

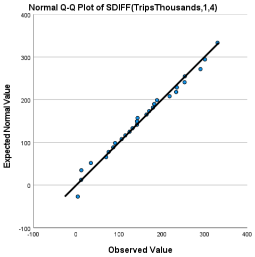
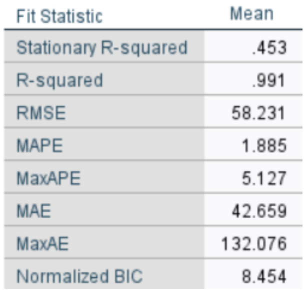


Fig. 8. Normal Q-Q for Simple Seasonal

The normal Q-Q plot was constructed to check the common distribution of the original time series and predicted data. The plot provides a good idea on the validity of the model.

2) Holt Winter's Multiplicative:
Holt Winter's is the most appropriate model for fitting a time series with seasonality and trend. The model uses the exponential smoothing to forecast seasonality ad trend[3]. For Holt Winter's model it is not necessary to perform difference or moving average as the model takes into account the following component automatically. For this submission, considering the multiplicative trend of the time series of the data, Holt's Winter Multiplicative Model is used.



Fit Statistic	Mean
Stationary R-squared	.453
R-squared	.991
RMSE	58.231
MAPE	1.885
MaxAPE	5.127
MAE	42.659
MaxAE	132.076
Normalized BIC	8.454

Fig. 9. Model Fit Statistics for Holt's Winter Multiplicative

The model was fitted to check the fit parameters. As can be seen in Figure 9, the RMSE value of 58.231 was able to be obtained. Also to notice the Mean Absolute Error(MAE) of 42.659 obtained.

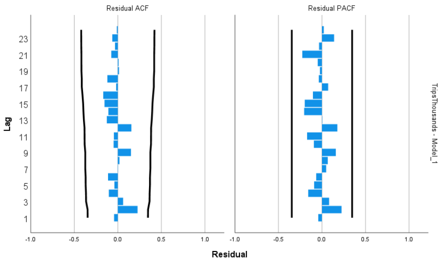


Fig. 10. Residual for Holt's Winter Multiplicative

Plotting the Residual Auto-correlation(ACF) and Partial Auto-correlation(PACF) to check the lag cutting. There is no lag cutting. Both AF and PACF show random distributed patterns which shows that this model is a good fit.

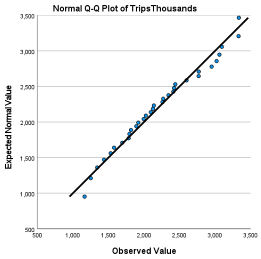


Fig. 11. Normal Q-Q for Holt's Winter Multiplicative

The normal Q-Q plot was constructed to check the common distribution of the original time series and predicted data. The plot provides a good idea on the validity of the model. The model seems to be a good fit.

3) Seasonal ARIMA

The Auto Regressive Integrated Moving Average(ARIMA) is a group of models that explain the time series based on their past values[2]. For this submission we consider the Seasonal ARIMA due to seasonality of the time serious. Important notice should be focused on the knowledge that Seasonal ARIMA removes additive seasonal effects and the time series being multiplicative it is important to obtain the parameters.

The AR(auto regressive) is a process that uses values from previous time to input as regressive equation. I(Integrated) is a differencing process to detrend the the data or as to make it stationary. ARIMA models assume the series to stationary so it is necessary to detrend the series. MA(moving average) is a process to smoothen the data. These parameters can be selected before fitting the model based on the plots. To find the most appropriate parameters for this this time series, Auto Correlation plots were constructed.

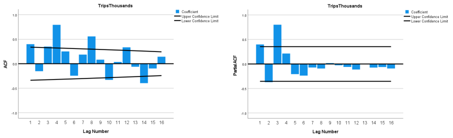


Fig. 12. ACF and PCF for ARIMA

From Figure 12, there is severe lag cutting which points towards a need for differencing.

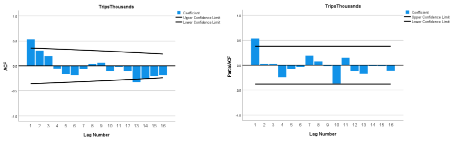
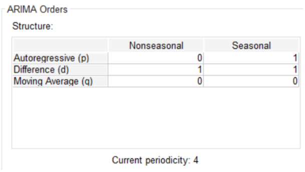


Fig. 13. ACF and PCF for ARIMA after Differencing

After making the data stationary and running the plot again, there is decrease in lag cutting but there is still a cut at lag one which means seasonal differencing parameter needs to be implemented of order one. There is also slope towards the next lag which point towards first order auto-regressive.



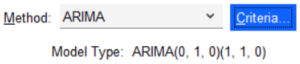
ARIMA Orders

Structure:

	Nonseasonal	Seasonal
Autoregressive (p)	0	1
Difference (d)	1	1
Moving Average (q)	0	0

Current periodicity: 4

Fig. 14. Selecting parameters for modelling



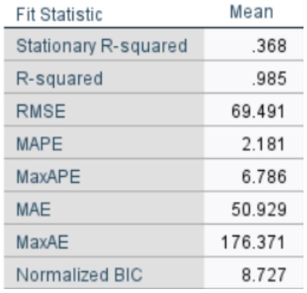
Method: ARIMA

Model Type: ARIMA(0, 1, 0)(1, 1, 0)

Fig. 15. Parameters for modelling

After doing the necessary, the model was fitted to check the fit parameters. As can be seen in Figure 16, the

RMSE value of 69.491 was able to be obtained. Also to notice the Mean Absolute Error(MAE) of 50.929 obtained.



Fit Statistic	Mean
Stationary R-squared	.368
R-squared	.985
RMSE	69.491
MAPE	2.181
MaxAPE	6.786
MAE	50.929
MaxAE	176.371
Normalized BIC	8.727

Fig. 16. Model Fit Statistics for ARIMA

Plotting the Residual Auto-correlation(ACF) and Partial Auto-correlation(PACF) to check the lag cutting. There is no lag cutting, Both AF and PACF show random distributed patterns which shows that this model is a good fit.

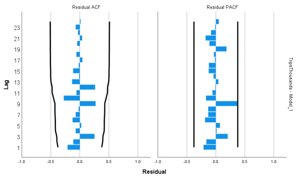


Fig. 17. Model Fit Statistics for ARIMA

The normal Q-Q plot was constructed to check the common distribution of the original time series and predicted data. The plot provides a good idea on the validity of the model. Most values lie the along the line. The model seems to be a good fit.

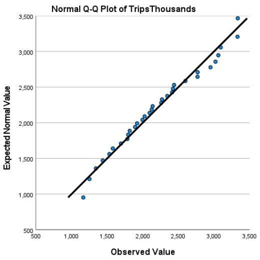


Fig. 18. Residual for ARIMA

4) Best Fit Model and Forecasting:

1) Results:
After fitting all model and comparing their fit and diagnostics, Holt's Winter Multiplicative seems to be the best fit. With the lowest RMSE and MAE of 58 and and 42 respectively it also has the best diagnostic parameters. It is chosen to forecast the time series.
Seasonal ARIMA model proved to be the second best model with RMSE score of 69 and MAE of 51. The reason for the lower performance can be attributed to the inabilty of ARIMA to remove multiplicative seasonal effects.
Simple Seasonal performed the worst among all. This perhaps due to the inability of the model to perform on trending time series. Even performing differencing before fitting the model does not improve the diagnostics.

2) Forecasting:
Holt Winter's Model is chosen to forecast.

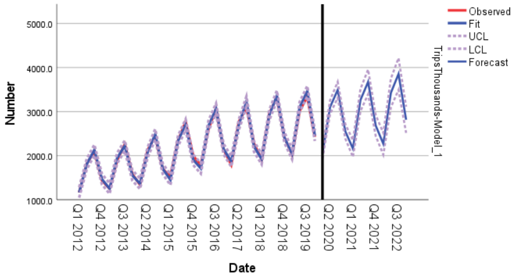
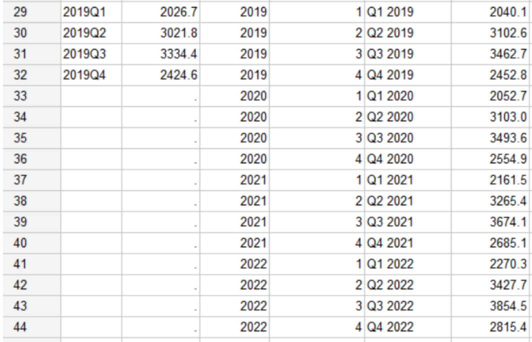


Fig. 19. Forecast Graph

The figure above shows the observed and fit line as well as the forecast. The figure below shows the predicted values for the next 3 periods.



29	2019Q1	2026.7	2019	1	Q1 2019	2040.1
30	2019Q2	3021.8	2019	2	Q2 2019	3102.6
31	2019Q3	3334.4	2019	3	Q3 2019	3462.7
32	2019Q4	2424.6	2019	4	Q4 2019	2452.8
33		.	2020	1	Q1 2020	2052.7
34		.	2020	2	Q2 2020	3103.0
35		.	2020	3	Q3 2020	3493.6
36		.	2020	4	Q4 2020	2554.9
37		.	2021	1	Q1 2021	2161.5
38		.	2021	2	Q2 2021	3265.4
39		.	2021	3	Q3 2021	3674.1
40		.	2021	4	Q4 2021	2685.1
41		.	2022	1	Q1 2022	2270.3
42		.	2022	2	Q2 2022	3427.7
43		.	2022	3	Q3 2022	3854.5
44		.	2022	4	Q4 2022	2815.4

Fig. 20. Forecast values for next 3 periods.

B. NewHouseRegistrations_Ireland

1) Dataset: The dataset contains the number of new house registration from 1978 to 2019. We plot an initial sequence chart to get the rough idea of the raw data.

The figure above shows the number of new house registration yearly. On initial inspection it is clear that the there is no seasonality nor any trend.

2) Initial Analysis: The time series shows no seasonality nor it is cyclic. The only test that id needed to be done is trend analysis to check if the series is stationary.

For this Augmented Dickey Fuller test (ADF Test) was conducted in R. The ADF is a unit root test which test

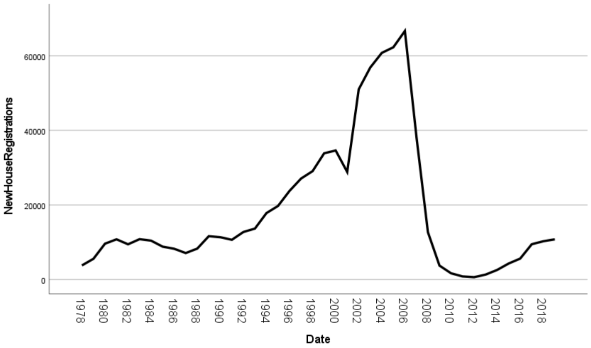
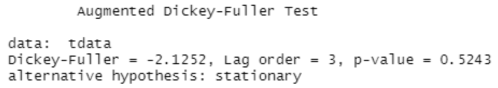


Fig. 21. New House Registration Sequence Chart

stationarity. The null hypothesis assumes the presence of unit root which equal to 1. So for the series to stationary the p-value have to less than 0.5 to reject null hypothesis.



```
        Augmented Dickey-Fuller Test

data:  tdata
Dickey-Fuller = -2.1252, Lag order = 3, p-value = 0.5243
alternative hypothesis: stationary
```

Fig. 22. Augmented Dickey Fuller test (ADF Test) result

As can be seen from figure 22, the p value greater than 0.5, hence the null hypothesis is rejected. The series is stationary.

Based on the test above the following components of the data were found out:

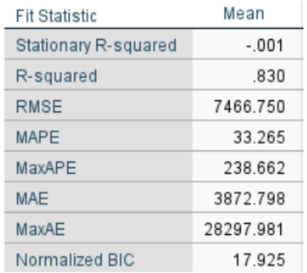
- The time series is stationary hence no trend.
- The time series is NOT seasonal.
- The time series is NOT cyclic.

3) Implementation: Based on the components of data, several forecasting models were selected. For this time series, the following models were finalized based on the components.

- Simple Non-Seasonal
- Damped Trend
- Linear ARIMA

1) Simple Non-Seasonal:
For the first model fitting we used Simple non-seasonal model.

The model was fitted to check the fit parameters. As can be seen in Figure 23, the RMSE value of 7466 was able to be obtained. Also to notice the Mean Absolute Error(MAE) of 3872 obtained.
Plotting the Residual Auto-correlation(ACF) and Partial Auto-correlation(PACF) to check the lag cutting. For ACF, there is negative lag cut at 1 which suggest a moving average of order one can be performed. The PACF sees both positive and negative lag cuts.



Fit Statistic	Mean
Stationary R-squared	-.001
R-squared	.830
RMSE	7466.750
MAPE	33.265
MaxAPE	238.662
MAE	3872.798
MaxAE	28297.981
Normalized BIC	17.925

Fig. 23. Model Fit Statistics for Simple Seasonal

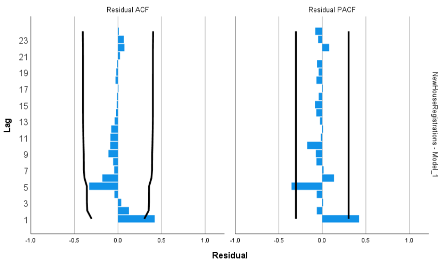


Fig. 24. Residual ACF and PACF

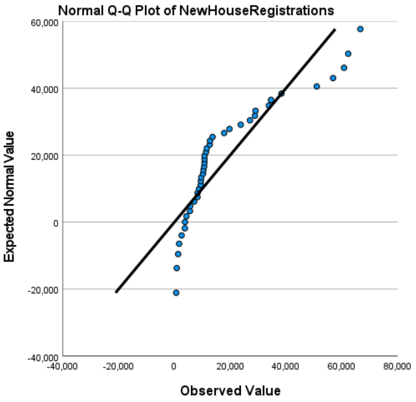
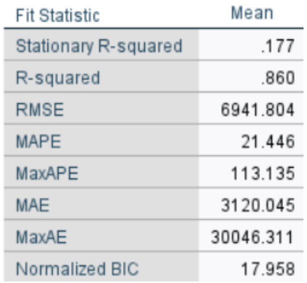


Fig. 25. Normal Q-Q for Simple Seasonal

The normal Q-Q plot was constructed to check the common distribution of the original time series and predicted data. The plot provides a good idea on the validity of the model. The model is not a good fit.

2) Damped Trend:
Damped Trend Model has a nice reputation as it is considered the best model for flexibility. It is also considered a good model for forecast accuracy.
The model was fitted to check the fit parameters. As can be seen in Figure 9, the RMSE value of 6941 was



Fit Statistic	Mean
Stationary R-squared	.177
R-squared	.860
RMSE	6941.804
MAPE	21.446
MaxAPE	113.135
MAE	3120.045
MaxAE	30046.311
Normalized BIC	17.958

Fig. 26. Model Fit Statistics for Damped Trend

able to be obtained. Also to notice the Mean Absolute Error(MAE) of 3120 obtained.

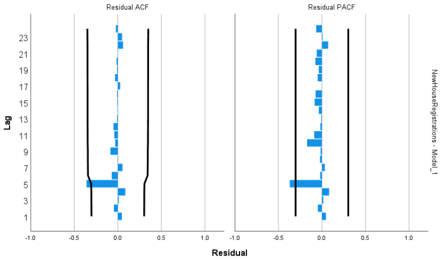


Fig. 27. Residual for Damped Trend

Plotting the Residual Auto-correlation(ACF) and Partial Auto-correlation(PACF) to check the lag cutting. There is a lag cutting at five in both AF and PACF.

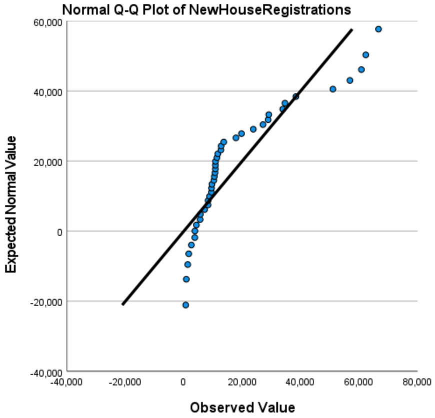


Fig. 28. Normal Q-Q for Damped Trend

The normal Q-Q plot was constructed to check the common distribution of the original time series and predicted data. The model is average fit.

3) Linear ARIMA
For this time series, Linear ARIMA model was used. To check the parameters, auto correlation plot was constructed.
It was elucidated that first order differencing was unsufficient. Second order differncing was used.

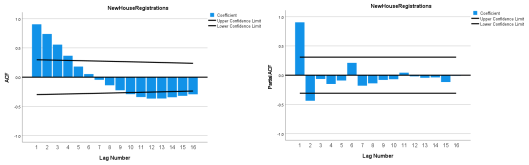


Fig. 29. ACF and PCF for ARIMA

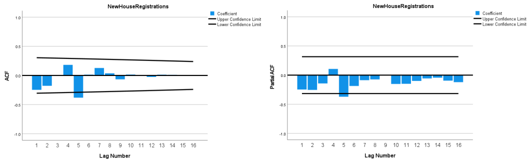
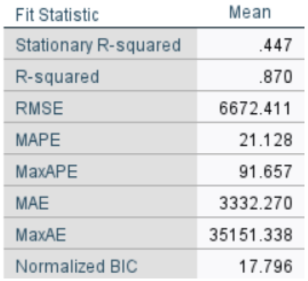


Fig. 30. ACF and PCF for ARIMA after Differencing of second order

After doing the necessary, the model was fitted to check the fit parameters. As can be seen in Figure 33, the RMSE value of 6672 was able to be obtained. Also to notice the Mean Absolute Error(MAE) of 3332 obtained.



Fit Statistic	Mean
Stationary R-squared	.447
R-squared	.870
RMSE	6672.411
MAPE	21.128
MaxAPE	91.657
MAE	3332.270
MaxAE	35151.338
Normalized BIC	17.796

Fig. 31. Model Fit Statistics for ARIMA

Plotting the Residual Auto-correlation(ACF) and Partial Auto-correlation(PACF) to check the lag cutting. There is no lag cutting, Both AF and PACF show random distributed patterns which shows that this model is a good fit.

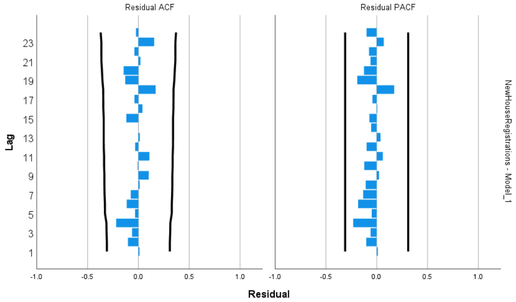


Fig. 32. Forecast Graph

The normal Q-Q plot was constructed to check the common distribution of the original time series and

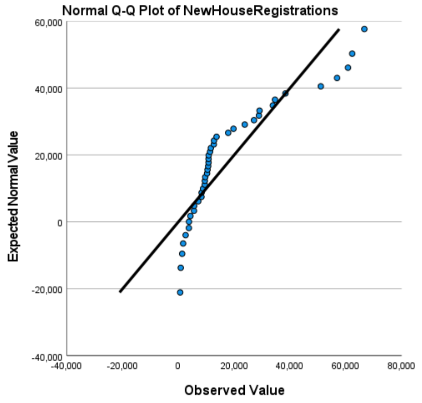


Fig. 33. Residual for ARIMA

predicted data. The plot provides a good idea on the validity of the model. The model has a reasonable fit.

4) Best Fit Model and Forecasting:

1) Results:
   After fitting all model and comparing their fit and diagnostics, Linear ARIMA seems to be the best fit. With the lowest RMSE and MAE of 6672 and a reasonable MAE of 3332 it also has the best diagnostic parameters. It is chosen to forecast the time series.
   Damped Trend model proved to be the second best model with RMSE score of 6941 and MAE of 3120. As expected of it's reputation, it performed with the lowest MAE. It has the best model fit statistics.
   Simple non-seasonal performed the worst among all. It is in line with the prediction.
2) Forecasting:
   Linear ARIMA Model is chosen to forecast.

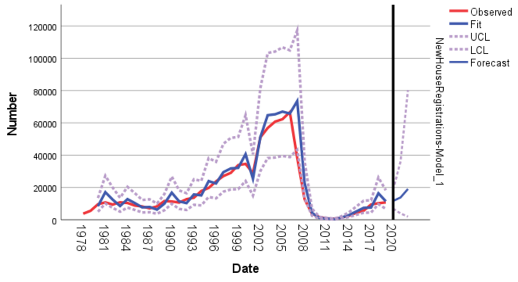
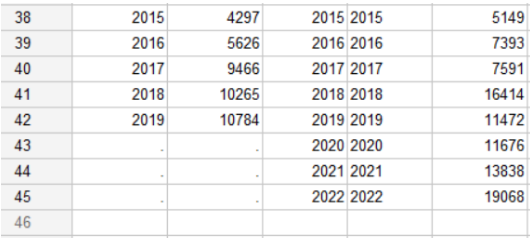


Fig. 34. Forecast Graph

The figure above shows the observed and fit line as well as the forecast. The figure below shows the predicted values for the next 3 periods.

II. Part B: Logistic Regression

When talking about binary dependent variable, logistic regression is one of the most common statistical used. This should not deter from the fact that more complex form logistic regression do exists. While comparing to linear regression model, it is to keep in mind that logistic regression the ratio in the presence of more than explanatory variable. Although the way the logistic regression works maybe similiar to multivariate linear regression, the result is observation of impact of each variable on the odds ratio of the event of interest.



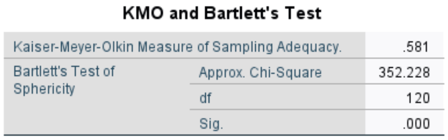
38	2015	4297	2015	2015	5149
39	2016	5626	2016	2016	7393
40	2017	9466	2017	2017	7591
41	2018	10265	2018	2018	16414
42	2019	10784	2019	2019	11472
43	.	.	2020	2020	11676
44	.	.	2021	2021	13838
45	.	.	2022	2022	19068
46					

Fig. 35. Forecast values for next 3 periods.

A. Childbirths

The data selected for this submission is a childbirth data of a particular city in USA. The dataset contains information regarding the body dimension of new born babies with data on parents like age and smoking habits. For our submission we have considered to apply binary logistic regression to find if the mother is a smoker or not.

1) Model 1: Using KMO on the original dataset very low KMO score is obtained. This is evidently pointing towards the low suitability for any detection.

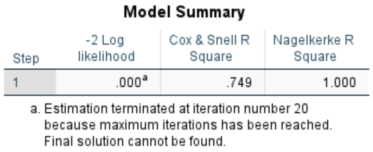


KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.581
Bartlett's Test of Sphericity	Approx. Chi-Square	352.228
	df	120
	Sig.	.000

Fig. 36. KMO test model 1

Any increase in the KMO score is a good indicator of a good model. For preliminary analysis we start with selecting all predictors. This is to get a rough idea of how to start. A good rating is above 0.8. While the score of 0.581 is considered very bad, for this submission starting with a low KMO score must be taken into account.



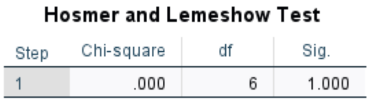
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.749	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Fig. 37. Summary Model 1

As expected, the model can be easily classified as overfitting model. All the parameters used show positive results. The Hosmer Lemeshow test shows the significance of 1. As well looking at the QQ plot below it can be seen that model is a perfect fit.



Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	6	1.000

Fig. 38. Summary Model 1

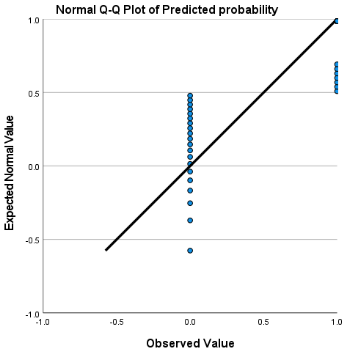
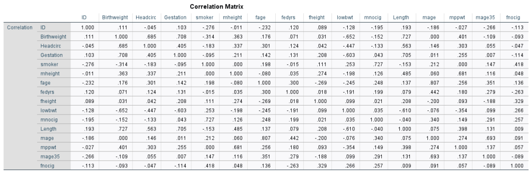


Fig. 39. QQ plot with Predicted Probablities for Model 1

2) Model 2: Taking cues from the earlier model, correlation matrix was constructed to find out the highly correlated variables from the dataset. Upon inspecting the matrix, various predictors were found to be having high correlation with the target variable. To remedy the over-fitting of the model, these variables were removed. Along with the highly correlated variable, variable with no correlation were removed as well.The following variable were removed:

- mnocig
- mheight
- mppwt
- fnocig

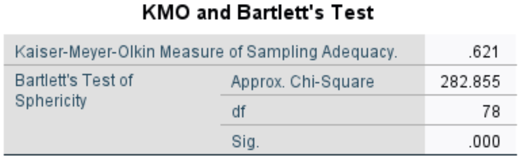


Correlation Matrix

		ID	Birthweight	Headcirc	Gestation	smoker	mheight	fage	fedyrs	fheight	lowbwt	mnocig	Length	mage	mppwt	mage35	fnocig
Correlation	ID	1.000	.111	-.045	.103	-.276	-.011	-.232	.120	.089	-.128	-.195	.193	-.186	-.027	-.266	-.113
	Birthweight	.111	1.000	.685	.708	-.314	.363	.176	.071	.031	-.652	-.152	.727	.000	.401	-.109	-.093
	Headcirc	-.045	.685	1.000	.405	-.183	.337	.301	.124	.042	-.447	-.133	.563	.146	.303	.055	-.047
	Gestation	.103	.708	.405	1.000	-.095	.211	.142	.131	.208	-.603	.043	.705	.011	.255	.007	-.114
	smoker	-.276	-.314	-.183	-.095	1.000	.000	.198	-.015	.111	.253	.727	-.153	.212	.000	.147	.418
	mheight	-.011	.363	.337	.211	.000	1.000	-.080	.035	.274	-.198	.126	.485	.060	.681	.116	.048
	fage	-.232	.176	.301	.142	.198	-.080	1.000	.300	-.269	-.245	.248	.137	.807	.256	.351	.136
	fedyrs	.120	.071	.124	.131	-.015	.035	.300	1.000	.018	-.191	.199	.079	.442	.180	.279	-.263
	fheight	.089	.031	.042	.208	.111	.274	-.269	.018	1.000	.099	.021	.208	-.200	.093	-.188	.329
	lowbwt	-.128	-.652	-.447	-.603	.253	-.198	-.245	-.191	.099	1.000	.035	-.610	-.076	-.354	.099	.266
	mnocig	-.195	-.152	-.133	.043	.727	.126	.248	.199	.021	.035	1.000	-.040	.340	.149	.291	.257
	Length	.193	.727	.563	.705	-.153	.485	.137	.079	.208	-.610	-.040	1.000	.075	.398	.131	.009
	mage	-.186	.000	.146	.011	.212	.060	.807	.442	-.200	-.076	.340	.075	1.000	.274	.693	.091
	mppwt	-.027	.401	.303	.255	.000	.681	.256	.180	.093	-.354	.149	.398	.274	1.000	.137	.057
	mage35	-.266	-.109	.055	.007	.147	.116	.351	.279	-.188	.099	.291	.131	.693	.137	1.000	-.089
	fnocig	-.113	-.093	-.047	-.114	.418	.048	.136	-.263	.329	.266	.257	.009	.091	.057	-.089	1.000

Fig. 40. Correlation Matrix

Using the new set of predictors, KMO test was performed again. The new Adequacy have increased from 0.581 to 0.621. Although not an significant increase. It is still good improvement over the abysmal later.

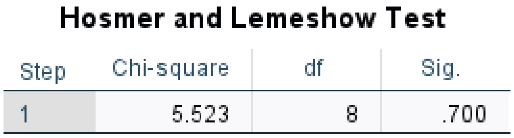


KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.621
Bartlett's Test of Sphericity	Approx. Chi-Square	282.855
	df	78
	Sig.	.000

Fig. 41. KMO test

After removing the specified predictors, another regression was performed. As expected the new model gave better results. The Hosmer And Lemeshow Test show a significance of 0.700. It is considered a good fit.

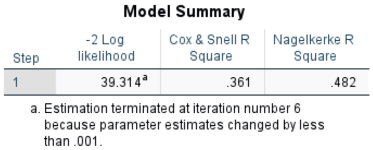


Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.523	8	.700

Fig. 42. Hosmer Lemeshow Test

The model summary also shows good results. We can safely assume that the new model is a good fit.



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	39.314[a]	.361	.482

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Fig. 43. Model Summary

The normal QQ plot for the model showing the predicted probabilities shows many point lying at the line.

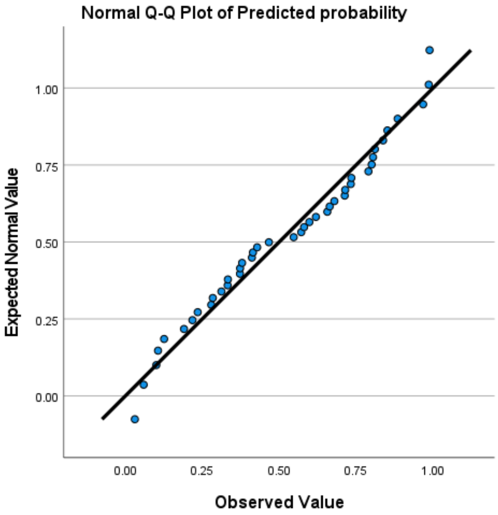


Fig. 44. QQ plot with Predicted Probabilities for Model 1

3) Summary: The effects of predictors on target variables is a significant event in Logistic progression. Hence variable selection forms the biggest process intensive task in model fitting. Variable selection either done using PCA or going through correlation matrix is a bigger challenge.

Variables that have high correlation be it negative or positive result in over-fitting models. While variables with very low correlation tending towards zero also decrease the efficiency of the model. Hence for this reason we discarded some variables which were couter effective for our approach.

The best model is the one which uses the lowest number possible of variable to give high probability percentage. The low initial KMO score of 0.581 deemed unsuitable for more predictors rejection in our case.

A comprehensive comparison between PCA and traditional method is something that can be done get more knowledge on variable selection.

References

[1] M. Ivanović and V. Kurbalija, "Time series analysis and possible applications," 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016, pp. 473-479, doi: 10.1109/MIPRO.2016.7522190.

[2] S. Noureen, S. Atique, V. Roy and S. Bayne, "Analysis and application of seasonal ARIMA model in Energy Demand Forecasting: A case study of small scale agricultural load," 2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), 2019, pp. 521-524, doi: 10.1109/MWSCAS.2019.8885349.

[3] G. Venkateswarlu and A. D. Sarma, "Performance of holt-winter and exponential smoothing methods for forecasting ionospheric TEC using IRNSS data," 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2017, pp. 1-5, doi: 10.1109/ICECCT.2017.8117892.