National College of Ireland

STATISTICS FOR DATA ANALYTICS

ON

MULTIPLE LINEAR REGRESSION

&

TIME SERIES ANALYSIS

Submitted by
Abdul Azadh Abdul Saleem - x18203621
MSc in Data Analytics – JAN/Batch B - 2020/21

MULTIPLE LINEAR REGRESSION

Multiple Regression is a technique in which more than one predictor variable is used to predict the relevant response variable. The predictors are also be called as Independent variables which influences the Dependent variable which is the output. Multiple Regression is a Supervised learning technique.

OBJECTIVE:

The objective of this analysis is to perform the Multiple liner regression on the Causes of death in various NUTS2 regions for the year 2010.

Research Question: Prediction of cause of death in NUTS2 region with the help of various factors influencing it.

DATASET:

This analysis was done using the dataset which was extracted by combining tables of many datasets from the Eurostat website. This dataset has the information about the causes of death in various NUTS2 regions. Eurostat uses a set of geospatial classifications to divide the European region and to analyze, Nomenclature of Territorial Units for Statistics (NUTS) is an important method among them. This classification method divides the Europe in three different forms, NUTS1 has the major socio-economic regions, NUTS2 regions which helps to apply regional policies and NUTS3 regions for specific diagnosis. Each dataset has the record of 264 NUTS2 regions for the year 2010.

As a part of data cleaning and pre-processing, the table of data from the various dataset has been merged with the help of R language. The datasets were imported to R, with the help of DPLYR package the tables were joined using Inner join function, keeping the NUTS2 regions and the year 2010 as Key columns. In the resulting dataset, the null values were removed using TDYLR package in R. At last the pre-processed dataset resulted with 263 instances.

The link of the source from where the dataset was taken has been given below.

Data source of Independent variables:

- Death due to Transport by NUTS2 region - https://ec.europa.eu/eurostat/databrowser/view/tgs00061/default/table?lang=en
- Death due to Ischemic heart disease by NUTS2 region– https://ec.europa.eu/eurostat/databrowser/view/tgs00059/default/table?lang=en
- Death due to Cancer by NUTS2 region – https://ec.europa.eu/eurostat/databrowser/view/tgs00058/default/table?lang=en
- Death due to Accidents by NUTS2 region – https://ec.europa.eu/eurostat/databrowser/view/tgs00060/default/table?lang=en

Data source of Dependent variable:

- All cause of death by NUTS2 region- https://ec.europa.eu/eurostat/databrowser/view/tgs00057/default/table?lang=en

MEASUREMENT LEVELS OF VARIABLES:

This dataset has four independent variables(X) which has the influence on the output Death(Y), they are Accident, Cancer, Heart disease and Transport.

$$Y = b_0 + X_1b_1 + X_2b_2 + X_3b_3 + X_4b_4$$

Scale measure: ACCIDENT(X1), CANCER(X2), HEARTDISEASE(X3) & TRANSPORT(X4)

Scale measure: DEATH(Y)

Nominal measure: NUTS2region

VERIFICATIONS OF ASSUMPTIONS:

NORMALITY OF DEPENDENT VARIABLE:

As a part of verifying the assumptions, firstly the dependent variable 'Death y' is plotted in histogram which showed in Fig.1, it is slightly skewed. So, to rectify this skewness, logarithmic transformation technique was done on the dependent with log value of base 10 and the these transformed values (Death_Log10) was plotted in the histogram again as shown in Fig.2.

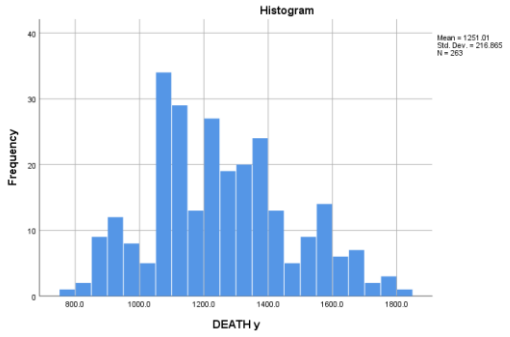


Fig.1

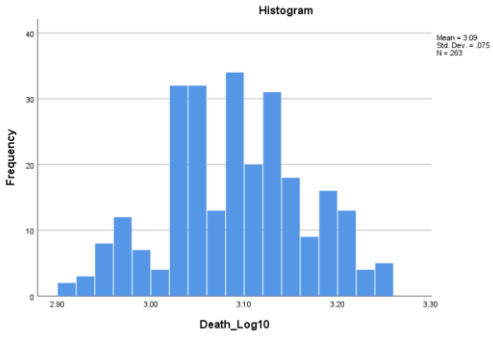


Fig.2

CORRELATION & MULTICOLLINEARITY:

The correlation table mentioned below in Fig.3 helps us to find the correlation between the dependent and the independent variables, which helps to verify this assumption.

From the table it is observed that the individual variable HEARTDISEASE shows high correlation with the dependent variable Death_Log10, with the value of 0.711 and the independent variable ACCIDENT shows the least correlation value of 0.34 with the dependent variable. The other independent variables CANCER and TRANSPORT has the values of 0.499 and 0.604 respectively. Since all the predictor variable taken shows considerable contribution towards predicting the dependent variable, none of them neglected in the analysis. For the check of Multi-collinearity between predictor variables it was observed that the variables were less correlated with each other, this shows that both these conditions were satisfied in this assumption.

Correlations

		Death_Log10	ACCIDENT	CANCER	HEART DISEASE	TRANSPORT
Pearson Correlation	Death_Log10	1.000	.340	.499	.711	.604
	ACCIDENT	.340	1.000	.228	.311	.497
	CANCER	.499	.228	1.000	.434	.114
	HEART DISEASE	.711	.311	.434	1.000	.329
	TRANSPORT	.604	.497	.114	.329	1.000
Sig. (1-tailed)	Death_Log10	.	.000	.000	.000	.000
	ACCIDENT	.000	.	.000	.000	.000
	CANCER	.000	.000	.	.000	.033
	HEART DISEASE	.000	.000	.000	.	.000
	TRANSPORT	.000	.000	.033	.000	.
N	Death_Log10	263	263	263	263	263
	ACCIDENT	263	263	263	263	263
	CANCER	263	263	263	263	263
	HEART DISEASE	263	263	263	263	263
	TRANSPORT	263	263	263	263	263

Fig.3

Based on the above examination the following Fig.4 shows the details of the independent variables utilized for this analysis.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	TRANSPORT, CANCER, HEART DISEASE, ACCIDENT[b]	.	Enter

a. Dependent Variable: Death_Log10

b. All requested variables entered.

Fig.4

Verification of VIF and Tolerance factor are also the part of this analysis using the Coefficient table as shown in Fig.5. Value of VIF should not be more than 10, in the below mentioned table of Coefficients we could see that all the values of independent variables are 1.398, 1.258, 1.382 and 1.401 which shows none of those variables exceed the value 10. Other factor is Tolerance factor, which must be more than 0.10 to satisfy this verification. From the same coefficients table, the Tolerance value of all the independent variable are more than 0.10. This shows that both VIF and Tolerance factors were satisfied in this assumption.

Coefficients[a]

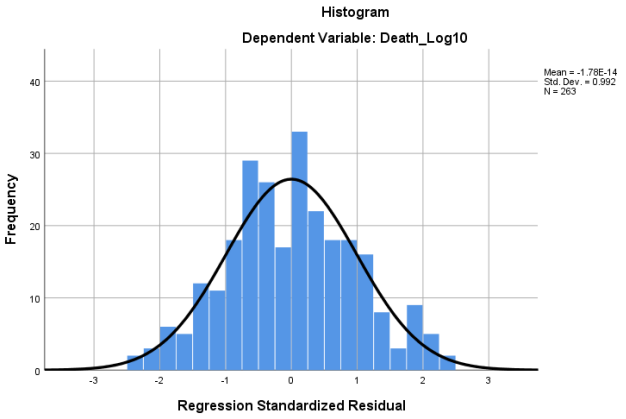
Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	2.743	.023		119.617	.000		
	ACCIDENT	-.001	.000	-.099	-2.528	.012	.715	1.398
	CANCER	.001	.000	.262	7.058	.000	.795	1.258
	HEART DISEASE	.000	.000	.474	12.159	.000	.724	1.382
	TRANSPORT	.010	.001	.467	11.911	.000	.714	1.401

a. Dependent Variable: Death_Log10

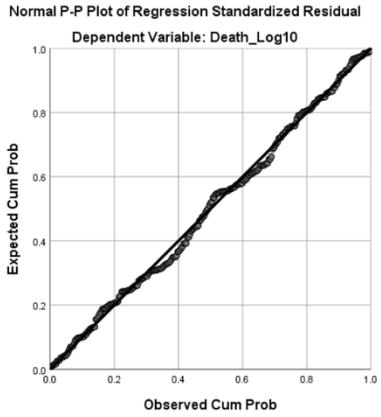
Fig.5

RESIDUAL INDEPENDENCE & HOMOSCEDASTICITY:

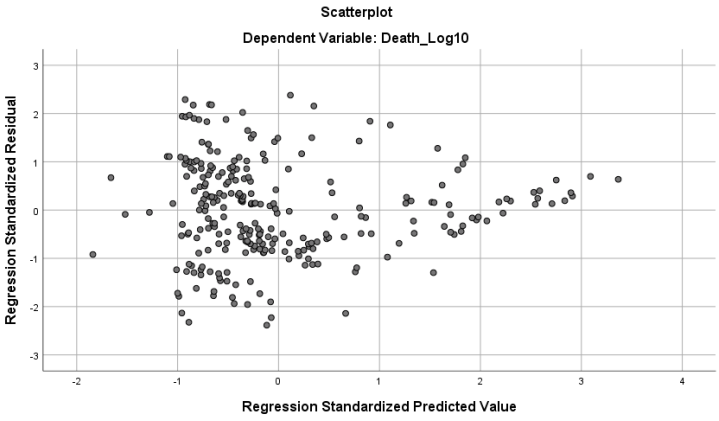
The three plots provided below helped to inspect the Homoscedasticity and Independence of Residuals of the Dependent variable.



Histogram of Frequency of Dependent variable with Regression Standardized Residuals (**Fig.6**)



Normal P-P Plot of Regression Standardized Residual (**Fig.7**)



Scatterplot of Regression standardized Residuals (**Fig.8**)

These plots help us to verify the assumptions provided below.

- Normal distribution of Regression Standardized Residual as shown in Fig.6
- The Scatter plot has the points distributed around 0, which satisfies the condition of Homoscedasticity, as shown in Fig.8.

CHECK FOR OUTLIERS:

Influence of outliers over the output has been checked with Cook's distance with the condition that it should not be more than 1 as shown in Fig.9. The below table shows that outlier values were less than 1, the table is sorted in descending order with highest value at the top and decreasing gradually. The value 0.15 is the highest Cook's distance obtained from this data. This helps us to assume that the model doesn't has any influence by the outliers.

COO_1
.15389
.12652
.09020
.08083
.04373
.04019
.03797
.03781

Fig.9

RESULT & INTERPRETATION:

The outputs obtained from the Multiple Linear Regression, helps to interpret the following information.

SUMMARY OF THE MODEL:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.846[a]	.716	.712	.04353	.716	162.933	4	258	.000	1.852

a. Predictors: (Constant), TRANSPORT, CANCER, HEART DISEASE, ACCIDENT

b. Dependent Variable: Death_Log10

Fig.10

The correlation in predicting the Cause of Death with the help of predictors can be shown as **R value in Fig.10,** which is **0.846**. The next column (**R square**) helps to know the significance of model for the analysis. This R square has the value **0.716** which shows the independent variables shows 71.6% of variability of the dependent variables. This shows that the model is significant for further analysis. **Adjusted R Square** value lies close to the value of R Square, which shows the contribution of predictors to predict the output variable is equal. **Durbin Watson** test gives the value of 1.852 which is close to 2, this provides the insight that the model has less outliers.

ANOVA:

ANOVA helps us to know the statistical significance of the model, using the F-value. **F-value** of the df(4,258) is **162.933** shown in Fig.11. Significance of F-value falls within the range 0.05 significance level, which concludes that the NULL hypothesis is not rejected, and our assumption is true.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.235	4	.309	162.933	.000[b]
	Residual	.489	258	.002		
	Total	1.724	262			

a. Dependent Variable: Death_Log10

b. Predictors: (Constant), TRANSPORT, CANCER, HEART DISEASE, ACCIDENT

Fig.11

COEFFICIENT TABLE:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	2.743	.023		119.617	.000		
	ACCIDENT	-.001	.000	-.099	-2.528	.012	.715	1.398
	CANCER	.001	.000	.262	7.058	.000	.795	1.258
	HEART DISEASE	.000	.000	.474	12.159	.000	.724	1.382
	TRANSPORT	.010	.001	.467	11.911	.000	.714	1.401

a. Dependent Variable: Death_Log10

Fig.12

The Coefficient analysis as mentioned in Fig.12, helps to get insight about how the contribution of independent variable influences dependent variable, significance level of all the independent variable in the model is less than 0.05 which proves the same. Unstandardized coefficient helps us to get knowledge about the impact of change in one independent variable, on the dependent variable. In this table if coefficient of TRANSPORT increases by one, then dependent value will have an increase of 0.10 in its outcome, likewise for the other independent variables as well. Standardized coefficient shows the outcome of standard deviation. So, if the standard deviation of TRANSPORT variable deviates by one, the standard deviation of dependent variable increases by 0.467.

CONCLUSION:

By analyzing the model by Multiple Linear Regression, it is concluded that this model is Statistically significant to predict the outcome of Cause of Death in NUTS2 regions, and this model could able to produce 70.6% accuracy in predicting the dependent variable. The assumptions we made with this model were not rejected. All the four predictors ACCIDENT, CANCER, HEART DISEASE and TRANSPORT has significant contribution in predicting the output, $F(4,263)= 162.933$, $R^2= 0.716$, $p < 0.05$.

TIME SERIES ANALYSIS

Time series data is a set of different values that a variable fluctuate over a period. Time series analysis is a technique that statistically that helps to deal with this time series data and analyzing with the insights. This technique involves the components like seasonality, trend and patterns of the time series data, which helps in predicting the future occurrences.

OBJECTIVE:

This analysis is focused on the prediction of Grocery food price index of New Zealand, for a period of five months successive to the last recorded value in the data.

Research Question: Prediction of Grocery food price index of New Zealand with the recorded data utilized, using Time series analysis.

DATASET:

The dataset utilized in this analysis is obtained from the official government repository of New Zealand. The data of Grocery Food price index is extracted from the Food price index dataset, the main reason for selecting Grocery foods price indexes for this analysis is because it has more impact on the overall food price index of New Zealand. This dataset covers a period of thirty-eight months which extends from January of 2017 till February of 2020, with values of every month.

Data Source: https://catalogue.data.govt.nz/dataset/food-price-index

DATA PRE-PROCESSING AND TRANSFORMATION:

The dataset of Food price index extracted from the repository is pre-processed with Microsoft Excel software and Grocery food index is separated and saved in .csv format which makes it suitable to import in R. This data is converted into time series data in R using ts() function and the values of start of the period and frequency where specified.

VERIFICATION OF ASSUMPTIONS:

The time series data is plotted, and it is observed that the values have a **Linear trend** and **no seasonality**. This statement has been justified by plotting with abline() function which plots a mean line for the corresponding data as shown in Fig.13, this helps to analyze the Linear trend of the data. The boxplot() plotted for this data Fig.14 helps to prove that this data doesn't have any seasonality factor.

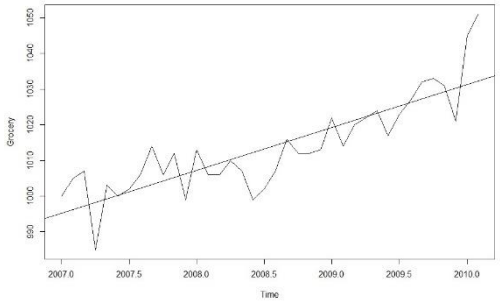


Fig.13

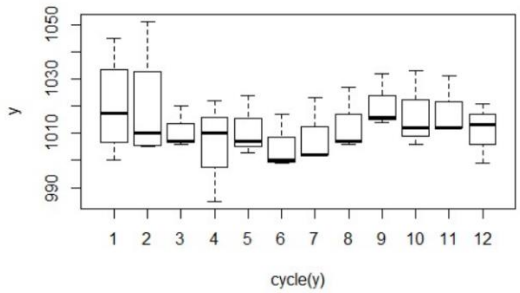


Fig.14

Seasonal plot of the time series data Fig.15 justifies the above condition. We could observe the values of the Food price index data is higher than the preceding year and the data doesn't show any seasonal pattern.

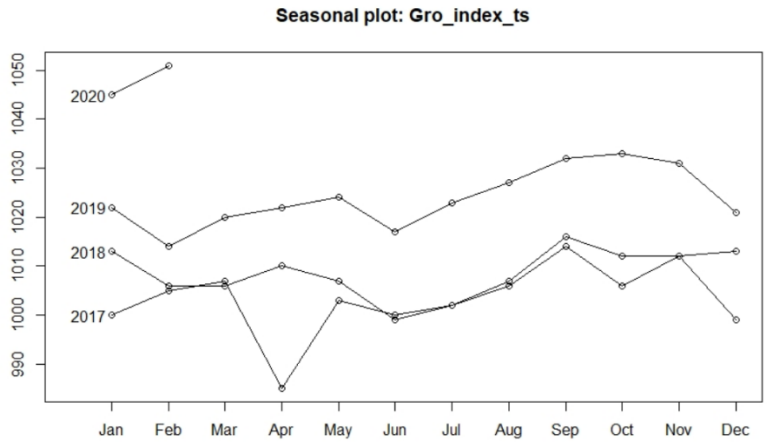


Fig.15

MODEL ANALYSIS & INTERPRETATION:

This data exhibits the property of Non-seasonality and Linearity in trend, Holt's Linear smoothing model has been chosen to forecast the future occurrences of this values. Holt's Linear smoothing model is also called as double exponential smoothing model, because it utilizes the level and trend for predicting the future occurrences. The future predictions of is obtained by two important factors α and β, where α is data smoothing factor and β is trend smoothing factor. These factors range from o to 1. Further ETS framework which comes along with forecast package, has a special function that if the model is not specified, it searches for a wide range of model for the data analyzed and finds the appropriate model for the data. Fig.16 is the output of the ets() function which is executed without specifying the model (i.e. model="ZZZ") and it automatically chose ETS(M,A,N) which means Holt's Linear smoothing model with multiplicative error model to analyze the data.

```
> Gro_hw2 = ets(Gro_index_ts, model = "ZZZ")
> Gro_hw2
ETS(M,A,N)

Call:
 ets(y = Gro_index_ts, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.0471
    beta  = 0.0471

  Initial states:
    l = 1002.1991
    b = 0.7856

  sigma:  0.0071

     AIC     AICc      BIC
293.6697 295.5447 301.8576
```

Fig.16

AIC value obtained from this model is 293.6697 which is comparatively lesser than 297.1877 and 300.4926 which are obtained from the other models Simple Exponential Smoothing model and HoltWinter's smoothi ng model. Considering Holt's Linear Smoothing model as effective model to the forecast of this data, graph

given in Fig.17 is plotted with the prediction for a period of five months following the last month of data pr ovided in the data and the values that are plotted in the forecast are provided in the Fig.18

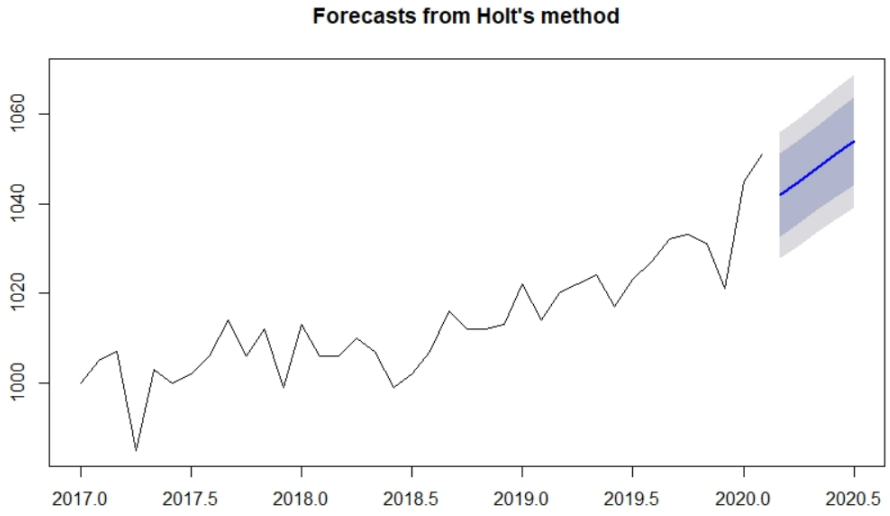


Fig.17

```
> forecast(Gro_hw2, 5)
         Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
Mar 2020       1042.089 1032.637 1051.541 1027.634 1056.544
Apr 2020       1045.111 1035.590 1054.632 1030.550 1059.672
May 2020       1048.133 1038.491 1057.774 1033.388 1062.878
Jun 2020       1051.155 1041.323 1060.986 1036.119 1066.191
Jul 2020       1054.177 1044.069 1064.284 1038.718 1069.635
> |
```

Fig.18

Accuracy of this forecast is provided in Fig.19 which shows Root Mean Square Error values is 6.79305 is shows less distraction from the idle forecast value. This proves that this model effective for forecasting the values in this data.

```
> round(accuracy(Gro_hw2),5)
                    ME    RMSE     MAE    MPE    MAPE    MASE    ACF1
Training set 1.25016 6.79305 5.24671 0.1179 0.51718 0.43863 0.12063
```

Fig.19

CONCLUSION:

This concludes that the Holt' Linear model chosen to forecast the data provided, has been efficient in prediction the future values of Food price index of New Zealand. The assumptions of Linear trend and non-seasonality observed from this data satisfied that Holt's Linear model can be used in this data to forecast the future Grocery price index effectively. Automated model selection method using ets() function proves that this model will help to obtained the best result.

REFERENCE:

[1] statistics.leard.com, (2014). Laerd Statistics Website. [online] Available at : https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php

[2] statistics.leard.com, (2014). Laerd Statistics Website. [online] Available at : https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php

[3] Jofipasi, C.A. (2017). *Selection for the best ETS (error, trend, seasonal) model to forecast weather in the Aceh Besar District*. IOP Conference Series: Materials Science and Engineering, [online] Volume 352, p. 012055. Available at: https://iopscience.iop.org/article/10.1088/1757-899X/352/1/012055 [Accessed 18 Oct. 2017].