Statistics CA

Multiple Linear Regression

&

Logistic Regression

By

Student ID - X18184928

Name -Terrance Thomas

Word Count: 1316

Page Count: 11

Dataset Background

- The mortality rate dataset is derived using multiple datasets from http://www.who.int/gho/en/ for Multiple Linear Regression and Logistic Regression Analysis.
- The following are the links for the variables used in the project
  http://apps.who.int/gho/data/node.main.11?lang=en
  http://apps.who.int/gho/data/node.main.A997?lang=en
  http://apps.who.int/gho/data/view.main.GSWCAH01v
  http://apps.who.int/gho/data/view.main.MHSUICIDEASDRv
  http://apps.who.int/gho/data/node.main.A860?lang=en
- The dependent and the independent variable have been merged using R. The data has been cleaned using R. The merging has done based on the common countries present in all the data used.
- Junk characters and NA values have been removed from the data set so that the results which are obtained are accurate. Post cleaning the resultant data set is of 171 rows and 6 columns.
- All the different data was combined based on the country for the year of 2016. All the common countries among the data was used and the countries which were not common was not used for the dataset. There is 1 dependent variable and 4 independent variables as shown below:

Dependent Variable	
Variable Name	**Measure**
Mortality Rate	Scale

Independent Variable	
Variable Name	**Measure**
Road Traffic Death	Scale
Suicide Rate	Scale
Maternal Mortality	Scale
NCD Mortality	Scale

	Country	Mortality Rate	RoadTrafficDeath	Suicderates	Maternalmortality	NCDmortality
1	Afghanistan	245	15.1	6.4	673	851.6
2	Albania	96	13.6	5.6	16	552.2
3	Angola	238	23.6	8.9	246	539.9
4	Antigua and Barbuda	120	7.9	.5	43	548.4
5	Argentina	111	14.0	9.1	40	424.8
6	Armenia	116	17.1	5.7	26	611.6
7	Australia	61	5.6	11.7	6	292.6
8	Austria	62	5.2	11.4	5	335.2
9	Azerbaijan	118	8.7	2.6	26	655.5
10	Bangladesh	130	15.3	6.1	186	512.6

Multiple Linear Regression

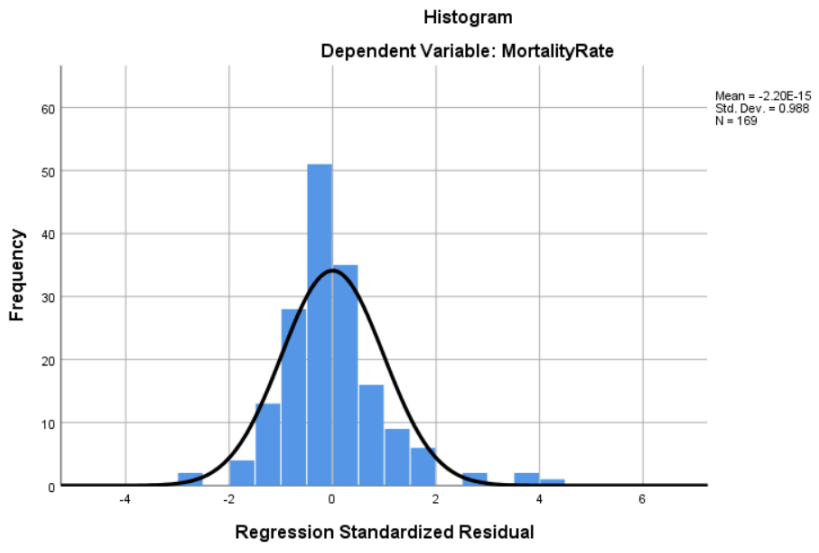
- Multiple Linear Regression is used to predict the value of a dependent variable based on 2 or more independent variables.
- In the current dataset Mortality Rate is the dependent variable and there are 4 independent variables (Road Traffic Death, Suicide Rate, Maternity Mortality and NCD mortality).
- To perform a standard multiple linear regression 3 main steps has been followed viz Checking the assumptions, Regression model Analysis and Conclusion.
- **Step 1: Checking the Assumptions:**

1. Sample Size:

  - ✓ The formula for the sample size N > 50 + (8 * m); where m is the number of independent variables (Tabachnick and Fidell, 2013,p.123).
  - ✓ Based on the above formula the sample size N > 82 as the m = 4 for the dataset; The sample size considered is 171 which satisfies the condition and acceptable.

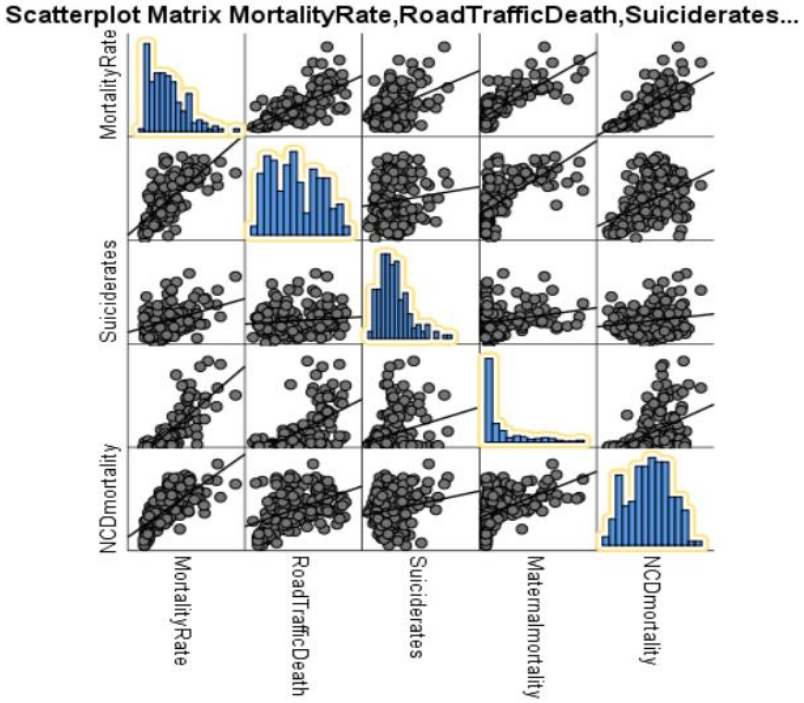
2. **Normality of Dependent Variable:**

  - ✓ The below histogram shows the Normal Distribution of Regression Residuals. The data used for the regression analysis are normally distributed.



3. Linear Relationship between Dependent and Independent Variables

- ✓ The Scatter plot Matrix is as follows:



- ✓ From the above image it is visible that there is a Linear Relationship between 'Adult Mortality Rate' and the dependent variables. However, 'Maternal Mortality Rate' are not normally standard so the log was used as shown below.

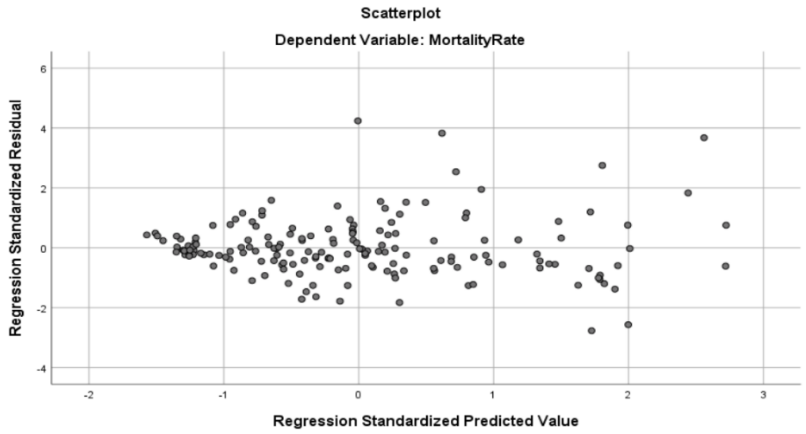
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-43.413	10.966		-3.959	.000
	RoadTrafficDeath	2.545	.377	.276	6.744	.000
	Suiciderates	3.189	.477	.210	6.686	.000
	Maternalmortality	.167	.017	.408	9.732	.000
	NCDmortality	.187	.019	.343	9.609	.000

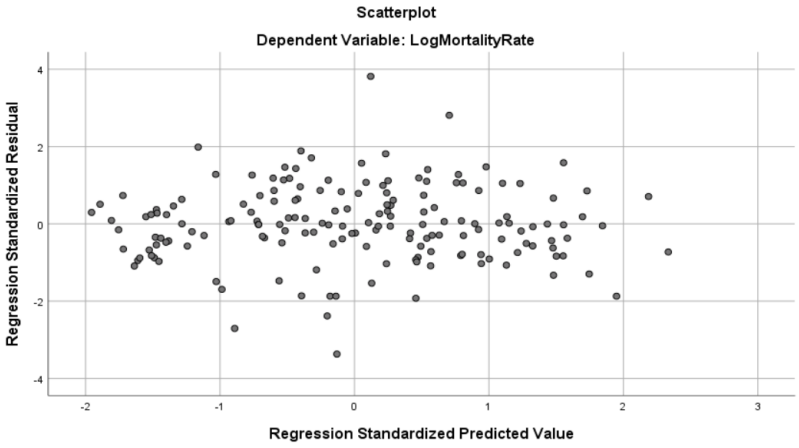
a. Dependent Variable: MortalityRate

4. Test for Heteroscedasticity:

- ✓ Residual Scatterplot does not meet the criteria of Heteroscedasticity as shown below:

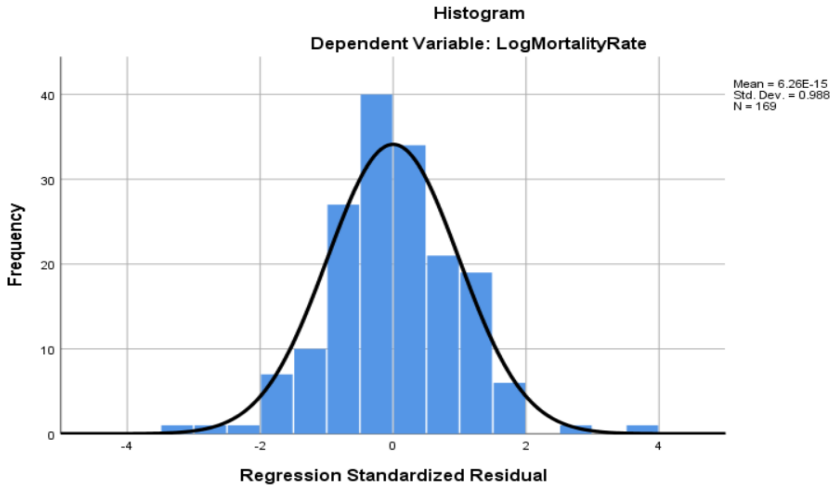


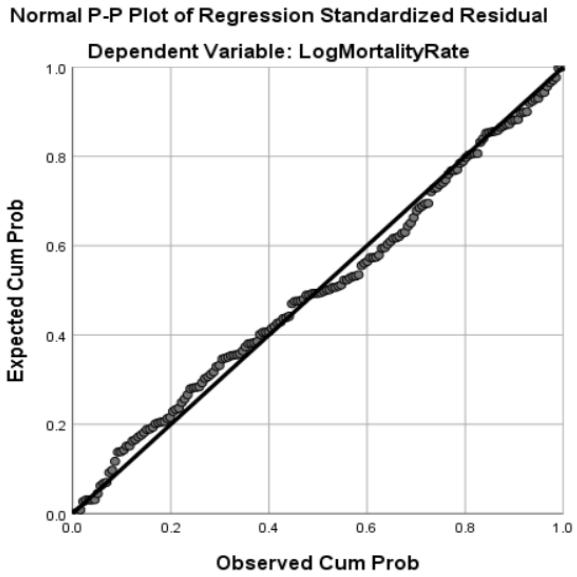
- ✓ Hence, 'Adult mortality Rate' was transformed by taking the log. Now the scatterplot meets the criteria of Heteroscedasticity as shown below:



5. Error Normally Distributed:

- ✓ Histogram and P-P plots of standardized Residuals is normally distributed as shown below:





6. Multicollinearity Test:

- ✓ The Variance Inflation Factor (VIF) test has a value less than 10 which means the predictor is not collinear with other predictors as shown below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3.262	.056		58.098	.000					
	RoadTrafficDeath	.013	.002	.223	5.312	.000	.749	.383	.141	.402	2.490
	Suiciderates	.017	.003	.178	6.595	.000	.319	.458	.175	.968	1.033
	NCDmortality	.001	.000	.380	11.145	.000	.777	.656	.296	.609	1.643
	LogMaternalmortality	.139	.015	.427	9.043	.000	.854	.577	.241	.317	3.153

a. Dependent Variable: LogMortalityRate

7. Influential Data Points:

- ✓ Cook's distance is less than 1 which means there is no influential data points in the model as shown below:

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.9336	6.1228	4.9316	.51042	169
Std. Predicted Value	-1.955	2.334	.000	1.000	169
Standard Error of Predicted Value	.016	.058	.031	.008	169
Adjusted Predicted Value	3.9318	6.1329	4.9316	.51064	169
Residual	-.63074	.71400	.00000	.18492	169
Std. Residual	-3.370	3.815	.000	.988	169
Stud. Residual	-3.397	3.917	.000	1.006	169
Deleted Residual	-.64102	.75256	-.00005	.19156	169
Stud. Deleted Residual	-3.513	4.101	.000	1.017	169
Mahal. Distance	.177	14.967	3.976	2.839	169
Cook's Distance	.000	.166	.007	.018	169
Centered Leverage Value	.001	.089	.024	.017	169

a. Dependent Variable: LogMortalityRate

- **Step 2: Multi Linear Regression Model Analysis:**

1. **Correlations:**
   - ✓ As shown in the below figure there is a strong correlation between Dependent variable and Road Traffic Death, NCD mortality Rate, Log of Maternity Morality.

Correlations

		LogMortalityRate	RoadTrafficDeath	Suiciderates	NCDmortality	LogMaternalmortality
Pearson Correlation	LogMortalityRate	1.000	.749	.319	.777	.854
	RoadTrafficDeath	.749	1.000	.115	.462	.773
	Suiciderates	.319	.115	1.000	.173	.116
	NCDmortality	.777	.462	.173	1.000	.616
	LogMaternalmortality	.854	.773	.116	.616	1.000
Sig. (1-tailed)	LogMortalityRate	.	.000	.000	.000	.000
	RoadTrafficDeath	.000	.	.068	.000	.000
	Suiciderates	.000	.068	.	.012	.067
	NCDmortality	.000	.000	.012	.	.000
	LogMaternalmortality	.000	.000	.067	.000	.
N	LogMortalityRate	169	169	169	169	169
	RoadTrafficDeath	169	169	169	169	169
	Suiciderates	169	169	169	169	169
	NCDmortality	169	169	169	169	169
	LogMaternalmortality	169	169	169	169	169

2. **Model Summary:**
   - ✓ The R square value is .884 while the adjusted R square is .881 which are almost same. Considering the R square value, it means 88.4% of variations in dependent variables accounted for independent variable.

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	.940[a]	.884	.881	.18716	.884	312.382	4	164	.000	2.084

a. Predictors: (Constant), LogMaternalmortality, Suiciderates, NCDmortality, RoadTrafficDeath

b. Dependent Variable: LogMortalityRate

3. **Coefficients:**
   - ✓ Y = a + b1X1 + b2X2 + …… + bnXn where,
     Y = Dependent Variable
     X1, X2, …., Xn = Independent Variable
     a = Intercept on Y axis
     b = Coefficient of Independent Variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3.262	.056		58.098	.000					
	RoadTrafficDeath	.013	.002	.223	5.312	.000	.749	.383	.141	.402	2.490
	Suiciderates	.017	.003	.178	6.595	.000	.319	.458	.175	.968	1.033
	NCDmortality	.001	.000	.380	11.145	.000	.777	.656	.296	.609	1.643
	LogMaternalmortality	.139	.015	.427	9.043	.000	.854	.577	.241	.317	3.153

a. Dependent Variable: LogMortalityRate

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43.769	4	10.942	312.382	.000[b]
	Residual	5.745	164	.035		
	Total	49.514	168			

a. Dependent Variable: LogMortalityRate

b. Predictors: (Constant), LogMaternalmortality, Suiciderates, NCDmortality, RoadTrafficDeath

- ✓ From the above table we get the Regression Equations as
  Log Adult Mortality = 3.262 + (0.013 x Road Traffic Deaths) + (0.017 x Suicide Rate) + (0.01 x NCD mortality) + (0.139 x Log Maternity Death)
- ✓ The equation signifies that for every 1 unit increase in Log of mortality rate (dependent variable) the independent variables Road Traffic Death increases by 0.013, Suicide Rate increases by 0.017, NCD mortality increase by 0.01 and Maternal mortality (Log) increases by 0.139.

- **Step 3: Conclusion:**
  - ✓ Multiple Linear Regression was used to predict the Adult Mortality Rate against Road Traffic Death, Suicide Rate, NCD mortality and Log of Maternal Mortality.
  - ✓ Regression output had R square as 0.884 ie 88.4% of variations was explained. Log Adult Mortality = 3.262 + (0.013 x Road Traffic Deaths) + (0.017 x Suicide Rate) + (0.01 x NCD mortality) + (0.139 x Log Maternity Death).
  - ✓ The dependent and 1 independent variable were transformed, and log of those values were used to overcome Heteroscedasticity.
  - ✓ Overall the model met the conditions and assumptions of Multiple Linear Regression.

Binary Logistic Regression

- Binary Logistic Regression will be used to predict if Adult Mortality rate is less than or greater than a certain percentage considering Road Traffic Death, Suicide Rate, Maternal Mortality and NCD Mortality.
- We have used the dataset used for Multi Linear Regression. Almost 50% of the values are near to the median value ie 146 of the dependent variable (Mortality Rate).
- So, we will convert the continues variable of the dependent variable to binary numbers viz 0 and 1.
  0 = Values from 0 to 145
  1 = Values 146 to highest
- **Step 1: Checking the Assumptions:**
  1. Dependent variables outcome is mutually exclusive as we converted continues variables into dichotomous variables.
  2. Sample size is 171 rows which is enough to perform Binary Logistic Regression.
  3. Multicollinearity Test:
     - ✓ There is a good correlation between the dependent and independent variables.

- ✓ The correlation between the dependent variables is less which is good for the model as shown below:

Correlation Matrix

		Constant	RoadTrafficDeath	Suiciderates	Maternalmortality	NCDmortality
Step 1	Constant	1.000	-.780	-.769	-.422	-.951
	RoadTrafficDeath	-.780	1.000	.575	.155	.628
	Suiciderates	-.769	.575	1.000	.373	.630
	Maternalmortality	-.422	.155	.373	1.000	.322
	NCDmortality	-.951	.628	.630	.322	1.000

- ➢ **Step 2: Binary Logistic Regression Analysis:**
  1. **Block 0 (Beginning Block):**
     - ✓ This block is the null hypothesis where there is no model applied without entering any independent variables.
     - ✓ The accuracy of the null hypothesis is 50.3% as shown below:

Classification Table[a,b]

			Predicted		
			BinaryMortalityRate		Percentage Correct
	Observed		.00	1.00	
Step 0	BinaryMortalityRate	.00	0	84	.0
		1.00	0	85	100.0
	Overall Percentage				50.3

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.012	.154	.006	1	.939	1.012

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	RoadTrafficDeath	63.472	1	.000
		Suiciderates	18.782	1	.000
		Maternalmortality	59.219	1	.000
		NCDmortality	77.407	1	.000
	Overall Statistics		105.969	4	.000

  2. **Block 1 (All Independent Variables Used):**
     - ✓ This is the alternative hypothesis where all the independent variables are used in the model.
     - ✓ **Omnibus Test:** Model is significant as $p < 0.05$

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	192.343	4	.000
	Block	192.343	4	.000
	Model	192.343	4	.000

- ✓ **Model Summary:** The model summary output tells us that model is a good fit. The Cox and Snell R Square value is 0.680 while the

Nagelkerker R Square Value is 0.906. 90.6% of variance is explained by independent variables account for dependent variables. -2 Log likelihood value is low which states that the model is a good fit.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	41.935[a]	.680	.906

a. Estimation terminated at iteration number 9 because parameter estimates changed by less than .001.

- ✓ **Hosmer and Lemeshow Test:** This test also suggest that the model is a good fit. The significance value of p = 0.951 which is greater than 0.05 stating it is a good fit.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.722	8	.951

- ✓ **Classification Table:** The figure below shows that the model has an accuracy of 97% when all the independent variables are used on the model.

Classification Table[a]

			Predicted		
			BinaryMortalityRate		Percentage Correct
	Observed		.00	1.00	
Step 1	BinaryMortalityRate	.00	81	3	96.4
		1.00	2	83	97.6
	Overall Percentage				97.0

a. The cut value is .500

- ✓ **Variables in the equation:** All the independent variables show high significance to the model.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	RoadTrafficDeath	.222	.077	8.281	1	.004	1.248
	Suiciderates	.412	.114	12.982	1	.000	1.510
	Maternalmortality	.037	.011	11.923	1	.001	1.037
	NCDmortality	.023	.006	16.635	1	.000	1.023
	Constant	-23.014	5.258	19.155	1	.000	.000

a. Variable(s) entered on step 1: RoadTrafficDeath, Suiciderates, Maternalmortality, NCDmortality.

3. Conclusion:

- ✓ The model used for Binary Logistic Regression is a good fit.
- ✓ All the independent variables are significant to the model.
- ✓ The model has an accuracy of 97%.
- ✓ Hosmer and Lemeshow states that model is 95.1% fit.

References

Maternal Death, *World Health Organization.* [Online]
Available at: http://apps.who.int/gho/data/view.main.GSWCAH01v
[Accessed 22 November 2019].

Road Traffic Deaths, *World Health Organization.* [Online]
Available at: http://apps.who.int/gho/data/node.main.A997?lang=en
[Accessed 22 November 2019].

NCD Mortality, *World Health Organization.* [Online]
Available at: http://apps.who.int/gho/data/node.main.A860?lang=en
[Accessed 22 November 2019].

Adult Mortality Rate, *World Health Organization.* [Online]
Available at: http://apps.who.int/gho/data/node.main.11?lang=en
[Accessed 22 November 2019].

Suicide Rate, *World Health Organization.* [Online]
Available at: http://apps.who.int/gho/data/view.main.MHSUICIDEASDRv
[Accessed 22 November 2019].