Multiple Linear Regression Analysis for Human Life Expectancy

Deepak Kumar Swain, x19216769@student.ncirl.ie, MSc Data Analytics, National College of Ireland

***Abstract*— Human lifespan is dependent on several factors which can either increase the lifespan or decrease it. In this paper, an attempt is made to use multilinear regression model to analyze life expectancy of human based of few factors.**

Objective

In this project, multiple regression model will be applied on a dataset collected from World Health Organization website using SPSS and R. The objective of this project is to analyze the relationships between the features and response using multiple linear regression. This is not only limited to getting a satisfactory R squared value for the model. After applying the regression model, an evaluation will be done to validate whether the assumptions for multilinear regression are being violated or not. If the assumptions are not satisfied, further analysis and transformation will be performed if necessary. At the end, required model selection methods are applied to find the best subset of predictors for the model.

About the Dataset:

The dataset is created from five different datasets available on WHO portal which can be easily accessible from following URLs. The final dataset is containing the values of different features and response variable for all the countries available in the WHO portal for the year of 2015.

Source:

https://apps.who.int/gho/data/node.main.11?lang=en

https://apps.who.int/gho/data/node.main.688?lang=en

https://apps.who.int/gho/data/node.main.525?lang=en

https://apps.who.int/gho/data/node.main.NCDCHILDBMIMINUS2C?lang=en

https://apps.who.int/gho/data/node.main.MHSUICIDE?lang=en

Variable	Description	Type
Country	All the country names	Strings
Year	Year	Numerical, constant
LifeExpectancy	Life expectancy at birth (in years)	Continuous Numerical
AdultMortalityRate	Adult mortality rate (probability of dying between 15 and 60 years per 1000 population)	Continuous Numerical
InfantMortalityRate	Infant mortality rate (probability of dying between birth and age 1 per 1000 live births)	Continuous Numerical
CrudeSuicideRate	Crude suicide rates (per 100 000 population)	Continuous Numerical
NCDMortalityRate	Age-standardized Noncommunicable Disease mortality rate (per 100 000 population)	Continuous Numerical

In this dataset, country and year do not have much meaning for the required model as the year is constant and for each row, there is a different country and does not show any correlation with other variables. So, the columns which are going to be considered for the model are 'Life Expectancy', 'Adult Mortality Rate', 'Infant Mortality Rate', 'Crude Suicide Rate' and 'NCD Mortality Rate'.

Multiple Linear Regression:

To apply multiple regression on the dataset, we are going to take the variable 'LifeExpectancy' as the independent variable and the remaining four variables as independent variables. The reason for taking 'LifeExpectancy' as the dependent variable is due to its high correlation with other variables. Theoretically, the independent variable should not have any correlation among them but in practical scenarios, we will see correlation between the features. Another logic behind choosing the dependent variable is due to the general tendency of lifespan with respect to various factors.

After applying an initial regression model by taking the dependent and independent variables in SPSS, following results transpired.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.992[a]	.984	.984	.9793

a. Predictors: (Constant), NCDmortalityrate, Crudesuiciderate, Infantmortalityrate, Adultmortalityrate

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	86.607	.318		272.542	.000
	Adultmortalityrate	-.046	.002	-.528	-25.153	.000
	Infantmortalityrate	-.126	.007	-.351	-17.556	.000
	Crudesuiciderate	.075	.013	.061	5.865	.000
	NCDmortalityrate	-.009	.001	-.185	-14.227	.000

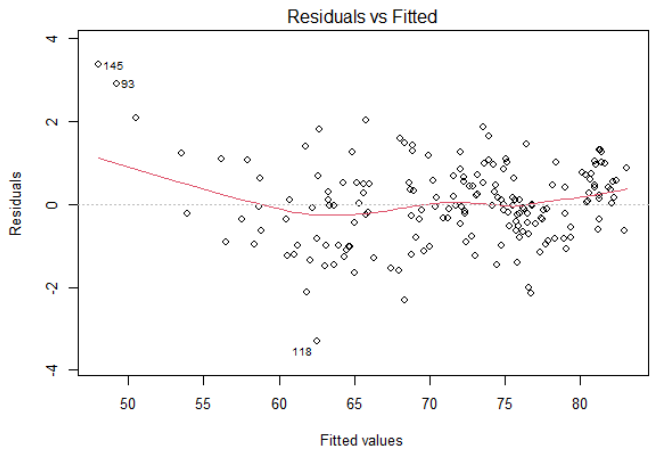
a. Dependent Variable: Lifeexpectancy

The R squared value is showing 0.984 which is the proportion [1] of the variance of 'Life Expectancy' that is explained by the four features in the model. The coefficient table is also showing that the features are having significance values that are less than 0.05. This means all the features are significantly contributing to the model and cannot be removed. So, the R squared value and the significance levels are showing a sign of a good model. Also, the adjusted R squared value is showing the same as R squared value which means the R squared value did not increase due more than one predictor (In this case, four predictors). However, this is not enough to decide that the multiple regression model is good enough. It needs to be a generalized model without violating certain assumptions. So, further analysis is done below to test it.

I. Gauss-Markov Assumptions:

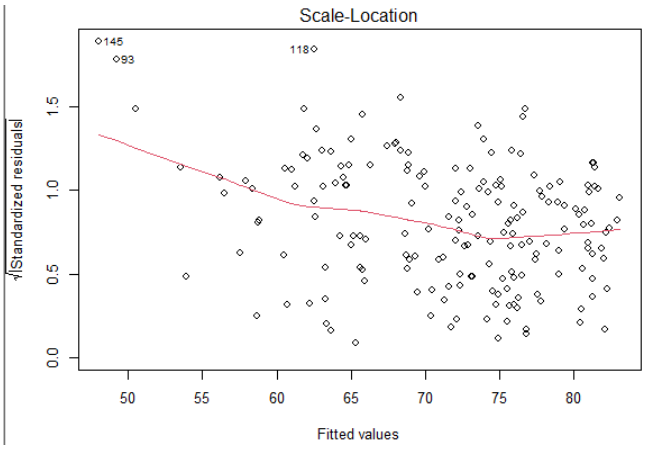
The crucial part while applying linear regression model is that the ordinary least square estimates [3] are best linear unbiased estimators (BLUE). The model must satisfy the Gauss-Markov assumptions. These assumptions are being validated using diagnostic plots in R.

Assumption 1 - Correct Functional Form (In this case, Linearity): One of the most critical assumptions in a multiple linear regression model is that the dependent variable [3] has linear relationships with the features, in fact, this linear relations should be straight lines without showing any biases. This assumption can be validated in the 'Residuals Vs Fitted Value' plot.



From the above plot, it is evident that there is no systematic pattern or relation between residuals and the predicted values. Also, there is no evidence of any nonlinear pattern like parabolic, exponential, quadratic, gaussian, double gaussian. This indicates that, the model is satisfying the assumption of correct functional form. So, there was no need to perform any kind of transformations on features.

Assumption 2 – Homoscedasticity: This assumption states that the noises in a linear regression model [3] should have constant variance. This can be validated from the following diagnostic plots.



From the above diagnostic plot, it is clearly visible that the plot is a random band around a horizontal line, i.e., there are equal variances across the line. The plot is not showing any signs of funnel like shape (no fan in or fan out). Few outliers are also available in the plot. However, these outliers do not have major impact here.

An ncvTest was also performed on the dataset which was supposed to give a non-significant p-value to satisfy the assumption. But, for the model ncvTest score was showing a significant value. However, it is known that ncvTest can sometimes be oversensitive with respect to the feature values. So, we may not take this as a violation to the assumption.

Since the diagnostic plot is not showing any violation to the assumption, no transformation is required.

Assumption 3 – No Autocorrelation between residuals: This is also an important assumption which states that the errors should be independent of each other. Generally, in time series data, we may see violation to this assumption as the residuals can be correlated to each other. But we may also face this violation in multivariate datasets if the residuals are correlated. So, it is always good to check whether a model is violating this assumption or not. Durbin-Watson test has been performed to test the serial correlation between the residuals.

```
> durbinWatsonTest(lmfit1)
 lag Autocorrelation D-W Statistic p-value
   1      0.01371134       1.960917   0.786
 Alternative hypothesis: rho != 0
```

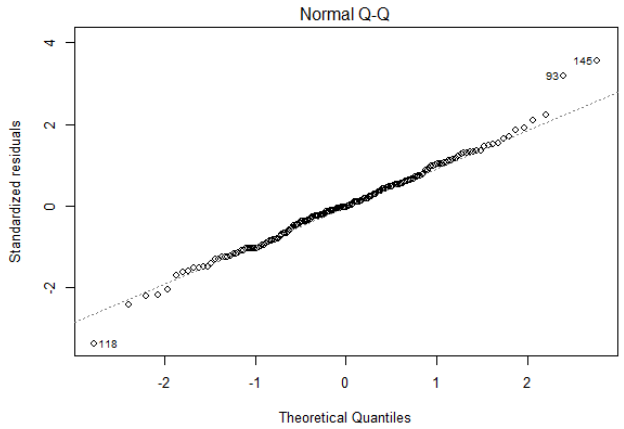
The formula used to calculate D-W Statistic is:

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Where, e_t is the residual figure and T is the number of observation.

When the D-W statistic value is close to to 2, it indicates that there is no correlation between the error terms. For the corrent dataset, D-W statistics value is 1.960917 which is close to 2. This indicates that the model is not violating the assumption.

Assumption 4 - Normal Distribution of Errors: This assumption states that the error values should be normally distributed with a mean of 0 if the respose variable is normally distributed.



This assumption has been validated using one of the diagnostics plot known as 'Normal Q-Q plot' which signifies that the standardized residual points should fall on a straight 45 degree line [3]. From the above plot, it is evident that the model is not violating the assumption. With this, all the Gauss-Markov assumptions are satified.

II. Assumption of Multicollinearity:

A multiple linear regression model cannot be considered as a good model if there is high multicollinearity [3] between the features.

Correlations

		Adultmortalityrate	Infantmortality rate	Crudesuicide rate	NCDmortalityrate
Adultmortalityrate	Pearson Correlation	1	.857**	-.052	.677**
	Sig. (2-tailed)		.000	.483	.000
	N	182	182	182	182
Infantmortalityrate	Pearson Correlation	.857**	1	-.244**	.608**
	Sig. (2-tailed)	.000		.001	.000
	N	182	182	182	182
Crudesuiciderate	Pearson Correlation	-.052	-.244**	1	-.104
	Sig. (2-tailed)	.483	.001		.163
	N	182	182	182	182
NCDmortalityrate	Pearson Correlation	.677**	.608**	-.104	1
	Sig. (2-tailed)	.000	.000	.163	
	N	182	182	182	182

**. Correlation is significant at the 0.01 level (2-tailed).

The above correlation table is computed using SPSS by taking all the features. For the model, it is seen that there are correlations between the features. For example, the correlation between 'Adult Mortality Rate' and 'Infant Mortality Rate' is 0.857 which indicates that these two features are highly correlated to each other. However, we can not decide from this correlation plot that these are causing casualities in the model.

So, to test the molticollinearity, a measure called Variation Inflation Factor (VIF) is used. This test basically measures if any feature varies due to other features. The VIF test takes one feature as an dependent variable and the other features as independent variables and then checks if there are any variance in the dependent variables due to the independent ones. This process contines for each independent variables. The VIF factor is calculated using following formula.

$$VIF_j = \frac{1}{1 - R_j^2}$$

Where R_j^2 is the coefficient [1] of determination when the regression is done between feature j and the remaining features.

Model		Collinearity Statistics	
		Tolerance	VIF
1	(Constant)		
	Adultmortalityrate	.205	4.874
	Infantmortalityrate	.226	4.420
	Crudesuiciderate	.844	1.184
	NCDmortalityrate	.536	1.866

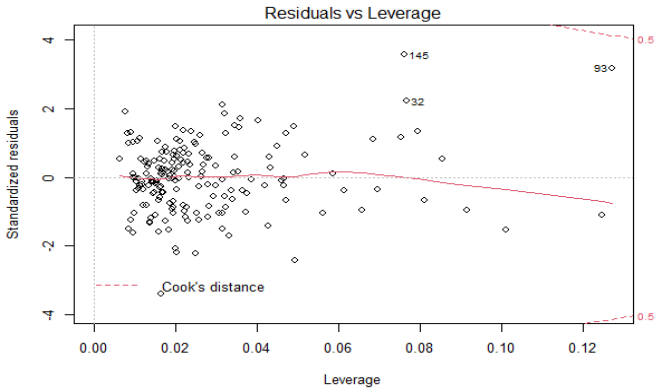
a. Dependent Variable: Lifeexpectancy

The above figure depicts the VIF and multicollinearity tolerance values for the model which was calculated using SPSS. For an ideal model, the VIF values for all the features should be below 5 (In fact, it is someties argued to be below 10). From the table, it is clearly seen that the VIF values for Adultmortalityrate, Infantmortalityrate, Crudesuiciderate and NCDmortalityrate are 4.874, 4.420, 1.184 and 1.866 respectively. This is satisfying the assumption of no multicollinearity.

There is another factor called 'Collinearity Tolerance' which is nothing but the inverse of VIF. Sometimes, it can be argued that, the collinearity tolerance values should be close to 1 or atleast mor than 0.5. But, if the VIF value is high, the collinearity tolerance will be low. Since, VIF values are satisfying the assumption, collinearity tolerance values can be ignored.

III. Assumption on Outliers:

Outliers are the data points that are found at an abnormal distance from other observations in a dataset. These outliers usually have high residuals as these are not predicted by a model properly. These outliers can influece our models drastically. So, to check if there are any influential data points that can impact the performance of a model, an estimate factor called "Cook's distance" is calculated. As per the assumption, no data points should have cook's distance value as 1 or more. Cook's distance 1 or more leads to violation of the assumption.

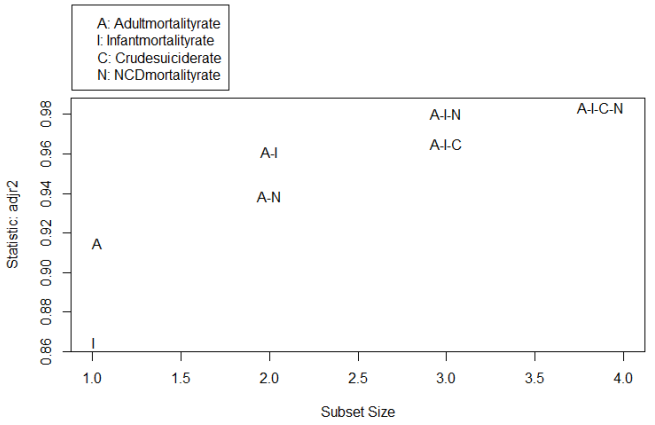


The above leverage plot is explaining about the cook's distance for data points of th model. It is clearly seen that no data points are having cook's value 1 or more. In fact, cook's distances are not exceeding 0.5 which is indicating that there are no influencial points in the model. Hence, the assumption is not getting violated.

The model has satisfied all the abobe mentioned assumptions for multiple linear regression.

Final Model Selection:

At the initial stage all the four variables were taken as predictors which gave the adjusted R squared value as 0.984. This is a pretty satisfactory value for a model. However, there are inbuilt packages which can be used to to find the best subset of feature variables for the model. In the dataset, there are four features available which means there can be 16 linear models built from it. To find the best subset, a subset plot was drawn in R which was showing the corresponding adjusted R squared values.



As per the above plot, the model gives the best adjusted R squared value when all the four variables are used as predictors.

There were other classic approaches [3] like 'Forward selection', 'Backward selection', 'Mixed selection' available to decide the best subset of predictors. In forward selection, we start with a null model which only has the intercept. Then we keep adding the predictors one by one. On the other hand, the backward selection is the opposite where it starts taking all the predictors in the model and then removes the ones which are not significant. Mixed selection is a mixture of forwad and backward selection where it starts with a null model and then keeps adding predictors one by one on the basis of best fitting value. While adding the predictors, if any of the considered predictors shows insignificant value, that predictor gets removed from the model.

These three classic methods are very useful when there are large number of predictors. But the current model has only four features. So, the subset plot was used to decide the best subset of the predictors for the model.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.992[a]	.984	.984	.9793

a. Predictors: (Constant), Crudesuiciderate, Adultmortalityrate, NCDmortalityrate, Infantmortalityrate

Conclusion:

The final model showed the R squared and adjusted R squared values as 0.992 and 0.984 respectively. This indicates that 99.2% of the variance for 'LifeExpectancy' were being explained by the features 'Adult Mortality Rate', 'Infant Mortality Rate', 'Crude Suicide Rate' and 'NCD Mortality Rate' which is a descent value for a model. Hence, it will be wise to consider multiple linear regression for the dataset.

References:

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, "An Introduction to Statistical Learning With Applications in R", 8th ed, pp. 71-81, 2017

[2] Andy Field, "Discovering Statistics Using SPSS", 3rd ed, SAGE Publications Inc, pp. 197-263, 2009

[3] Douglas A. Lind, William C. Marchal, SamuelA. Wathen, "Basic Statistics for Business & Economics", 5th ed, 2006, pp. 375-447