# Statistics for Data Analytics – CA2


**Student ID- x18127339**

**Student Name- Punit Lohani**

**MSc in Data Analytics**

**Cohort- B**


**Submitted to- Tony Delaney**

# Table of Contents

# Multiple Regression

## Introduction

Multiple Regressions can be used to analyze the relation between the one continuous dependent and the independent variables. Multiple Regressions helps to answer any particular question while performing the research. With the help of this analysis it is checked that how the several independent variables that are taken into consideration are affecting the dependent variable. (Pallant, 2005)

## Objective of the Analysis

The objective of the multiple regression analysis is to analyze and predict how the population using improved drinking water sources and solid fuels affects the Healthy life expectancy at Birth?

With the help of these datasets it is checked that how a newly born can expect to live a healthy life if they are provided with the environment having access to clean drinking water and the percentage of consumption and use of solid fuels in the households. Also check will be performed to see that out of the two independent variables taken for the analysis, which factor will influence the dependent variable to a great extent.

## Datasets

The datasets have been downloaded from http://data.un.org .

## Data Navigation Approach

Health > WHO Data > World Health Statistics > Mortality and global health estimates > Healthy life expectancy (HALE) at birth (years)

Health > WHO Data > World Health Statistics > Risk factors > Population using improved drinking -water sources (%)

Health > WHO Data > World Health Statistics > Risk factors > Population using solid fuels (%)

Three different datasets were obtained on the basis of countries and their respective descriptions and then merged into a single file to make it ready for the run in SPSS and proceed with the Multiple regression analysis.

## Variable Type

Independent Variables:

- Percentage of population using improved drinking water sources.
- Percentage of population using solid fuels.

Dependent Variable:

- Healthy life expectancy at Birth.

## Assumptions

- ### Descriptive Statistics

**Descriptive Statistics**

| | Mean | Std. Deviation | N |
|---|---|---|---|
| Healthy_life_exp | 61.24 | 8.227 | 172 |
| Water | 87.87 | 14.976 | 172 |
| Fuel | 35.59 | 37.088 | 172 |

The Descriptive statistics table mentioned above is representing and giving information about the sample size, mean and standard deviation of the dependent as well as independent variables used in the analysis. The sample size is enough for the generation of the result.

- ### Multicollinearity Check

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95.0% Confidence Interval for B | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound | Tolerance | VIF |
| 1 | (Constant) | 43.449 | 3.358 | | 12.938 | .000 | 36.820 | 50.079 | | |
| | Water | .244 | .034 | .443 | 7.204 | .000 | .177 | .310 | .470 | 2.128 |
| | Fuel | -.101 | .014 | -.456 | -7.416 | .000 | -.128 | -.074 | .470 | 2.128 |

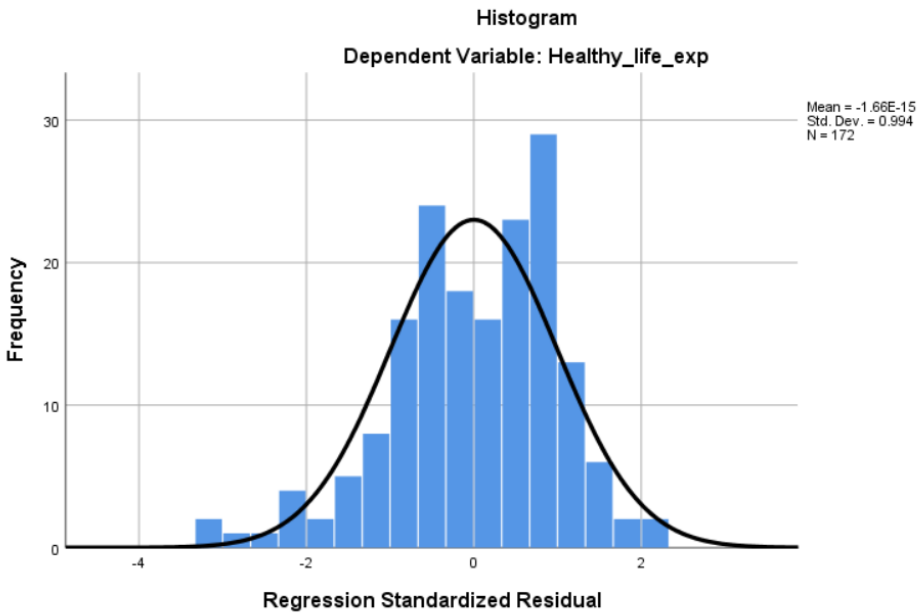
a. Dependent Variable: Healthy_life_exp

Below is the analysis figured out from the Coefficients table:

The values of Collinearity Tolerance and Statistics VIF values are 0.470 and 2.128 respectively. These values are checked to identify the possibility of multicollinearity. In ideal scenario the Collinearity

Tolerance value must be greater than 0.10 while the Statistics VIF value must be less than 10. In the above scenario both the conditions are satisfying.

- ### Check for Normality




The Histogram represents that the data is normally distributed and the bell shaped curve can be clearly seen.

- ### Check for Outliers

Scatter plot is taken into consideration for performing the check on the existence of outliers. Outliers are detected if the cases have standardized residual existing less than -3.3 and more than 3.3. In the above case there are no outliers and the same have been described later in the report.

## SPSS Output

The output related with the SPSS output is mentioned below:

- ## Model Summary

**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 | .836[a] | .699 | .696 | 4.538 | .699 | 196.564 | 2 | 169 | .000 | 2.202 |

a. Predictors: (Constant), Fuel, Water

b. Dependent Variable: Healthy_life_exp

The model summary table is representing the information as described below:

- The R Square value in the model summary table is 0.699. This value provides the information that how the two independent variables are explaining the variance in the dependent variable. On expressing the value in percentage it explains that 69.9% of the variance in the dependent variable which is Healthy life expectancy at birth.

- The Adjusted R square value is 0.696 which provides the better estimates of the true value.
- The Durbin-Watson value in the above case is 2.202, and the value in this case is greater than 2. Value greater than 2 denotes that between the adjacent residuals there is negative correlation (Field, 2009).

- ## ANOVA Table

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8094.180 | 2 | 4047.090 | 196.564 | .000[b] |
| | Residual | 3479.564 | 169 | 20.589 | | |
| | Total | 11573.744 | 171 | | | |

a. Dependent Variable: Healthy_life_exp

b. Predictors: (Constant), Fuel, Water

ANOVA table is referred when we want to know that the result is statistically significant or not. Significant value in the above case is 0.000 which represents statistical significance. Ideally the value must be less than 0.05, so here the conditions are met.

- ## Coefficients

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | 95.0% Confidence Interval for B | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Lower Bound | Upper Bound | Tolerance | VIF |
| 1 | (Constant) | 43.449 | 3.358 | | 12.938 | .000 | 36.820 | 50.079 | | |
| | Water | .244 | .034 | .443 | 7.204 | .000 | .177 | .310 | .470 | 2.128 |
| | Fuel | -.101 | .014 | -.456 | -7.416 | .000 | -.128 | -.074 | .470 | 2.128 |

a. Dependent Variable: Healthy_life_exp

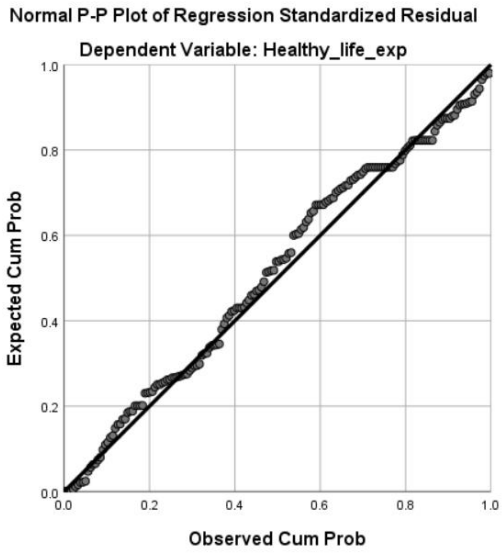
Below are the analysis figured out from the Coefficients table:

1) Unstandardized B value for the independent variables i.e. water and fuel are 0.244 and -0.101 and the constant value is 43.449 respectively. These values are used in the creation of regression equation.
2) Beta values of independent variables under Standardized coefficients are checked. Basically Beta values are used for the comparison and it represents the positive as well as negative correlation. In the above case, Beta values as 0.443 and -0.456 respectively.

3) Another check is performed on the Significance value corresponding to these Beta values. Significance values must be less than 0.05, Here both the values are 0.000, hence they make unique and statistically significant contribution to the prediction of Healthy life expectancy at birth.
4) The values of Collinearity Tolerance and Statistics VIF values are 0.470 and 2.128 respectively. These values are checked to identify the possibility of multicollinearity. In ideal scenario the Collinearity Tolerance value must be greater than 0.10 while the Statistics VIF value must be less than 10. In the above scenario both the conditions are meeting.

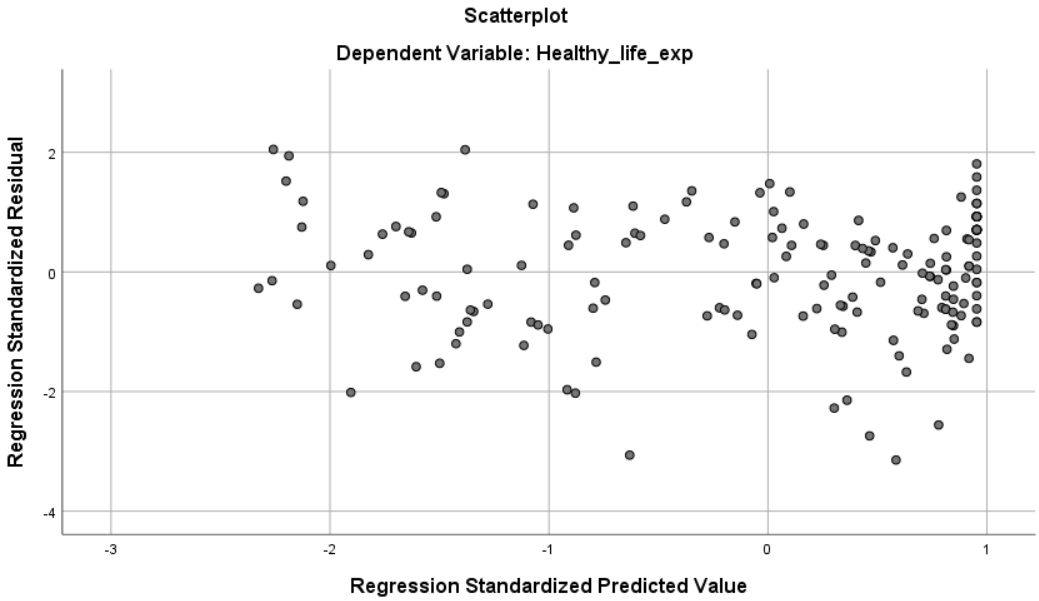
- ## Charts

### ➢ Normal P-P Plot



Ideally, in the Normal P-P Plot, our points must lie in the straight diagonal line from base left to upper right and from the normality there should not be much deviation. As it can be seen above these conditions are met, hence the Normal P-P plot is fine.

### ➢ Scatter Plot




Ideally, the standardized residuals must be roughly distributed in a rectangular manner and it should not be curvilinear and the maximum must not be on either one side. Also, scatter plot is used to detect the existence of outliers. Outliers are detected if the cases have standardized residual existing less than -3.3 and more than 3.3 . In this scenario as can be viewed from the above scatter plot most of the values are in between the range of -2 to 1. So, no outliers can be observed.

## • Conclusion

Since the two independent variables used in our analysis are reasonably correlated with each other and on further checking the significance values corresponding to them are 0.000 respectively which means they are making unique and statistically significant contribution to the prediction of Healthy life expectancy at birth. Furthermore the Beta values were checked. For the first independent variable i.e. population using improved drinking water the value is 0.443 and for the another independent variable i.e. population using solid fuels the value is -0.456 respectively. Positive Beta value indicates positive correlation while the negative Beta value indicates negative correlation with the dependent variable. If the population is having access to improved drinking water then it would more likely to contribute towards the healthy life expectancy at birth. So it can be concluded that population using improved drinking water is highly correlated with the Healthy life expectancy at birth. In addition, if other related factors are included then they can play a crucial role in analysis of the complete model in future.

# Binary Logistic Regression

## Introduction

Binary logistic regression is used to prognosticate the categorical outcome with two or more than two categories. Using binary logistic regression a model is developed to check the relation between the dependent and independent variable (Pallant, 2005).

Here the relationship between Sobriety check with the Random breathe check, fines, and license suspended are checked.

## Overview

In order to get familiarize with the terminologies that are used above, a brief description is presented below:

- Sobriety checkpoints are the location where the police officers block the road in order to perform a screening on the drivers of the vehicle which are passing through so that they can suspect the person who is driving under the influence of alcohol.
- Ensuring the road safety, the police officer conducts Random breath testing on driver that gives an idea about the level of alcohol consumed. If the level crosses the legal limit then the person can be found guilty and the necessary actions can be taken against them.
- The police personnel can also impose fine on the drivers if they observe that the alcohol limit has crossed.
- The driver needs to present the driving license if they are stopped by the police. Driving under the influence of alcohol can result in loss of license for a period of time.

## Objective of the Analysis

The main objective of our analysis is to predict the relationship between the Sobriety checkpoints with the random breathe testing, fines and license suspension. If the Sobriety checkpoints are encouraged by the countries then which factor will contribute most so as to ensure the road safety?

## Datasets

Datasets have being downloaded from http://www.who.int/gho/en/.

## Data Navigation Approach

Health Topics > Data > GHO themes > Non communicable diseases and mental health > Alcohol > Alcohol Control Policies > Drink Driving

In total there were four datasets that were obtained as a part of the analysis and then merged into a single file to make it ready for the run in SPSS.

Below are the list of Independent and dependent variables which are categorical in nature:

**Independent Variables**

1. Random breath testing
2. Fines
3. License suspension

**Dependent Variable**

1. Sobriety checkpoints

## Assumption

- ### Check for Multicollinearity

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Tolerance | VIF |
| 1 | (Constant) | .035 | .153 | | .225 | .822 | | |
| | License_suspension | .184 | .117 | .132 | 1.568 | .119 | .821 | 1.217 |
| | Random_Breathe_Test | .216 | .088 | .203 | 2.445 | .016 | .850 | 1.177 |
| | Fines | .204 | .147 | .108 | 1.383 | .169 | .954 | 1.048 |

a. Dependent Variable: Sobriety_check

Before proceeding with logistic regression,there is a need to identify the possibility of multicollinearity. Preliminary check is performed where the collinearity tolerance and statistics VIF values for all the three independent variables are checked. Collinearity tolerance must be greater than 0.10 while the statistics VIF value must be less than 10. In the above case the conditions are meeting as needed.

- ### Check for Outliers

Additionally, performed the check on outliers in the case wise list and no outliers were found which is mentioned later in the report.

It is quite clear that all the preliminary conditions are met which indicates to proceed with the logistic regression analysis.

## SPSS Output

- ### Case Processing Summary

**Case Processing Summary**

| Unweighted Cases[a] | | N | Percent |
|---|---|---|---|
| Selected Cases | Included in Analysis | 158 | 100.0 |
| | Missing Cases | 0 | .0 |
| | Total | 158 | 100.0 |
| Unselected Cases | | 0 | .0 |
| Total | | 158 | 100.0 |

a. If weight is in effect, see classification table for the total number of cases.

The Case Processing Summary table mentioned above is representing that there are 158 cases in the sample and no missing cases are encountered.

- ### Dependent Variable

**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| 0 | 0 |
| 1 | 1 |

Here , encoded Yes as 1 and No as 0 for the dependent variable and the same has been accepted by the SPSS.

- ## Block 0

  The Block 0 represents the result of the analysis done and it does not includes the independent variables used in the model.

**Block 0: Beginning Block**

**Classification Table[a,b]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Sobriety_check | | Percentage Correct |
| | | | 0 | 1 | |
| Step 0 | Sobriety_check | 0 | 0 | 75 | .0 |
| | | 1 | 0 | 83 | 100.0 |
| | Overall Percentage | | | | 52.5 |

a. Constant is included in the model.

b. The cut value is .500

From the above classification table it can be figured out that there are overall 52.5% of the countries have the encouragement of having the sobriety checkpoints.

- ## Block 1

In the Block 1, the testing of model is done. Overall model performance can be figured out with Omnibus Tests of Model Coefficients. It is also known as goodness of fit test.

**Block 1: Method = Enter**

**Omnibus Tests of Model Coefficients**

| | | Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | Step | 17.060 | 3 | .001 |
| | Block | 17.060 | 3 | .001 |
| | Model | 17.060 | 3 | .001 |

The Significant value must be less than 0.05. In this case the value is 0.001 which is satisfying the ideal scenario. The Chi-square value is 17.06 with 3 degree of freedom. This shows that the Block 1 is better model than Block 0.

- **Hosmer and Lemeshow Test**

**Hosmer and Lemeshow Test**

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 1 | 3.194 | 3 | .363 |

For Hosmer and Lemeshow test, the Chi-square value is 3.194 at significance of 0.363 and 3 degree of freedom. To support the model the Significance value must be greater than 0.05 and the same has been satisfied as shown in the above case.

- **Model Summary**

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 201.569[a] | .102 | .137 |

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

In the Model Summary table we check for the values of Cox and Snell R square and Nagelkerke R square. As seen in the above case, the Cox & Snell R Square value is 0.102 and the Nagelkerke R Square value is 0.137 respectively. The minimum and the maximum value ranges from 0 to 1. So, the set of variables are representing the variability between 10.2 percent to 13.7 percent respectively.

- ## Classification Table

**Classification Table[a]**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | Sobriety_check | | Percentage Correct |
| | | | 0 | 1 | |
| Step 1 | Sobriety_check | 0 | 39 | 36 | 52.0 |
| | | 1 | 24 | 59 | 71.1 |
| | Overall Percentage | | | | 62.0 |

a. The cut value is .500

From the Block 0 classification table it was seen that overall 52.5% of the countries have the encouragement of having the sobriety checkpoints.

By comparing this classification table with the classification table of Block 0 , it can be observed that on including the independent variables in the model there are the signs of improvement as for Block 0 the overall percentage was 52.7% and now for Block 1 it is 62.0%

- ## Variables in the Equation

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower | Upper |
| Step 1[a] | License_suspension | .885 | .543 | 2.650 | 1 | .104 | 2.422 | .835 | 7.027 |
| | Random_Breathe_Test | .923 | .382 | 5.847 | 1 | .016 | 2.516 | 1.191 | 5.316 |
| | Fines | 1.046 | .728 | 2.062 | 1 | .151 | 2.845 | .683 | 11.852 |
| | Constant | -2.263 | .830 | 7.436 | 1 | .006 | .104 | | |

a. Variable(s) entered on step 1: License_suspension, Random_Breathe_Test, Fines.

The table represents all the variables and their contribution. To figure out which variable contributes significantly to the predictive ability of the model we check for the Significance value and this value should be less than 0.05. Since the value of Random Breathe Test is less than 0.05, hence it can be said that it is contributing more significantly to the predictive ability of the model as compared to other variables i.e. License suspension and fines.

The B value in the above case is positive for License suspension, Random Breathe Test and Fines which means they are having direct relationship with the Sobriety checkpoints.

- **Casewise List**

**Casewise List**[a]

a. The casewise plot is not produced because no outliers were found.

The casewise plot is not produced as there were no outliers.

## Conclusion

Logistic regression was performed to ensure contribution of the factors that were taken into consideration for the encouragement of sobriety checkpoints by country so as to ensure the road safety. The model had three independent variables i.e. license suspension, Random breathe test and fines. On going through the outcomes of the analysis it was figured out of the three factors that were taken, only Random breathe test contributed significantly to the predictive ability of the model therefore it is the strongest predictor of the encouragement of the sobriety checkpoints. Furthermore the better model for the detailed analysis can also be developed in future if more independent factors are included.

BIBLIOGRAPHY

Pallant, J. (2007) *SPSS Survival Manual,* 3rd edition. Maidenhead: Open University Press, McGRaw-Hill

Field, A. (2005) *Discovering Statistics Using SPSS,* 2nd edition. London: SAGE Publication