Name: Raj R. Kupekar
Student Id: 18186432

Statistical Analysis

Linear Regression

Objective: The aim is to apply Multiple Linear Regression on National and Global Health Risks dataset. The dependent variable in this dataset is the average of 13 international health scores for SPAR version 2018, whereas the independent variables comprises of average of 1st version from 2010 to 2017 for about 194 countries members.

The analyzed dataset is one of the 1000 health risks indicators dataset which are gathered by 11 Global Health and Development Organization. The information collected are related to the health statistics for its 194 members with an intension to analyze and step up in achieving the global health objectives. The collected data are analyzed under the Sustainable Development Goals and simultaneously tracked to achieve the specific health related targets of the SDG's

Data Transformation: In this dataset there are 9 independent variables and 1 dependent variable. As the data type of this variables are in the integer form thus; to perform Linear Regression, it is necessary to convert them in the numeric data type. This is done using as.numeric() function in r. The sapply function is used to display the missing values for each variable in horizontal form; combined with is.na() function. Likewise, all the variables are linked with missing values in them. Accordingly, the missing values are filled with the median values of each variables , thus eliminating the effect of missing values on the model. Furthermore, detection and treatment of outliers is been carried out. Boxplot is used to detect the outliers in the dataset. The outliers are eliminated using a winsorizing technique in R. The upper and lower limit outliers are treated by the given formula;

$$\text{Upper Limit} = \text{3rd Quartile} + 1.5 * \text{IQR}$$

$$\text{Lower Limit} = \text{1st Quartile} - 1.5 * \text{IQR}$$

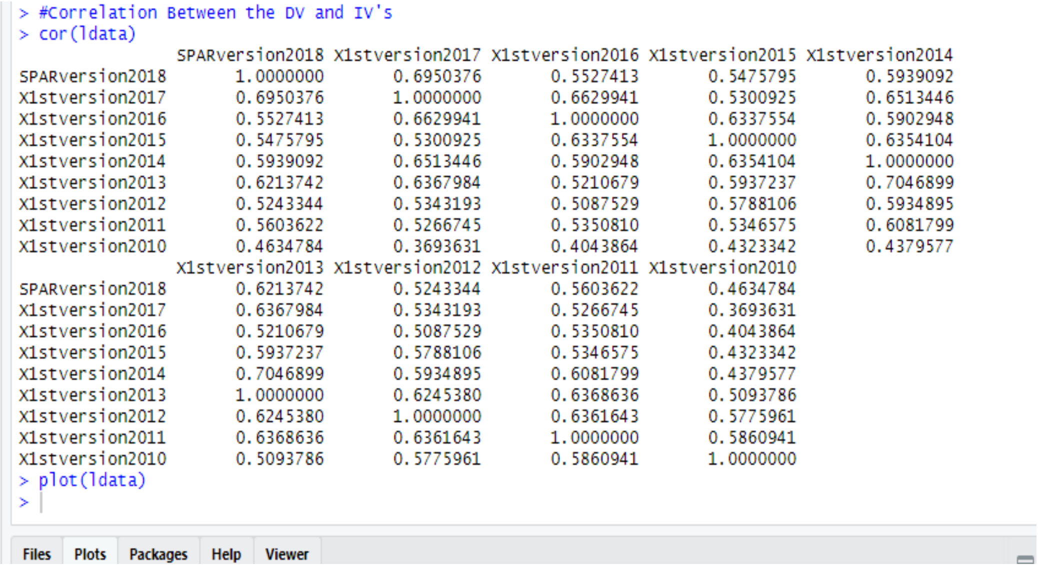
where, IQR= Inter-Quartile Range

Assumptions: Linear Regression is a linear approach to predict the linear relationship between the response variable and one or more supplementary variables. It makes assumptions on the datasets and builds an efficient predictive and analytical model. The regression has the following key assumptions.

- Absence of Multicollinearity
- Linear Relationship
- Normally Distributed
- Homoscedasticity
- No Autocorrelation

Assumption for Multicollinearity can be tested using a Correlation Matrix and Variance Inflation Factor (vif). In Correlation Matrix, if the correlation coefficients between the IV's are greater than |0.8|, than we can conclude that there is a presence of Multicollinearity between the respected IV's. In R, cor() function is used to obtain the Correlation Matrix between the IV's and DV. For National and Global Health Risks dataset, none of the IV's are having correlation coefficient above |0.8|, thus satisfying the assumptions for absence of Multicollinearity.

Another way of testing the Multicollinearity is using the Variation Inflation Factor (vif). Under the library 'car', vif can be carried out. If the vif values for variables are less than 10, then there is a clear evidence of absence of Multicollinearity. In our case, all the predictors are having vif values less than 10 thus; satisfying the assumption.



```
> #Correlation Between the DV and IV's
> cor(ldata)
                SPARversion2018 X1stversion2017 X1stversion2016 X1stversion2015 X1stversion2014
SPARversion2018       1.0000000       0.6950376       0.5527413       0.5475795       0.5939092
X1stversion2017       0.6950376       1.0000000       0.6629941       0.5300925       0.6513446
X1stversion2016       0.5527413       0.6629941       1.0000000       0.6337554       0.5902948
X1stversion2015       0.5475795       0.5300925       0.6337554       1.0000000       0.6354104
X1stversion2014       0.5939092       0.6513446       0.5902948       0.6354104       1.0000000
X1stversion2013       0.6213742       0.6367984       0.5210679       0.5937237       0.7046899
X1stversion2012       0.5243344       0.5343193       0.5087529       0.5788106       0.5934895
X1stversion2011       0.5603622       0.5266745       0.5350810       0.5346575       0.6081799
X1stversion2010       0.4634784       0.3693631       0.4043864       0.4323342       0.4379577
                X1stversion2013 X1stversion2012 X1stversion2011 X1stversion2010
SPARversion2018       0.6213742       0.5243344       0.5603622       0.4634784
X1stversion2017       0.6367984       0.5343193       0.5266745       0.3693631
X1stversion2016       0.5210679       0.5087529       0.5350810       0.4043864
X1stversion2015       0.5937237       0.5788106       0.5346575       0.4323342
X1stversion2014       0.7046899       0.5934895       0.6081799       0.4379577
X1stversion2013       1.0000000       0.6245380       0.6368636       0.5093786
X1stversion2012       0.6245380       1.0000000       0.6361643       0.5775961
X1stversion2011       0.6368636       0.6361643       1.0000000       0.5860941
X1stversion2010       0.5093786       0.5775961       0.5860941       1.0000000
> plot(ldata)
> 
```

Fig.1 Correlation Matrix

Scatter Plot of dataset depicts a visible evidence, showing an approximate positive linear relationship between the IV’s and DV. Hence, assumptions for Linear relationship is meet.

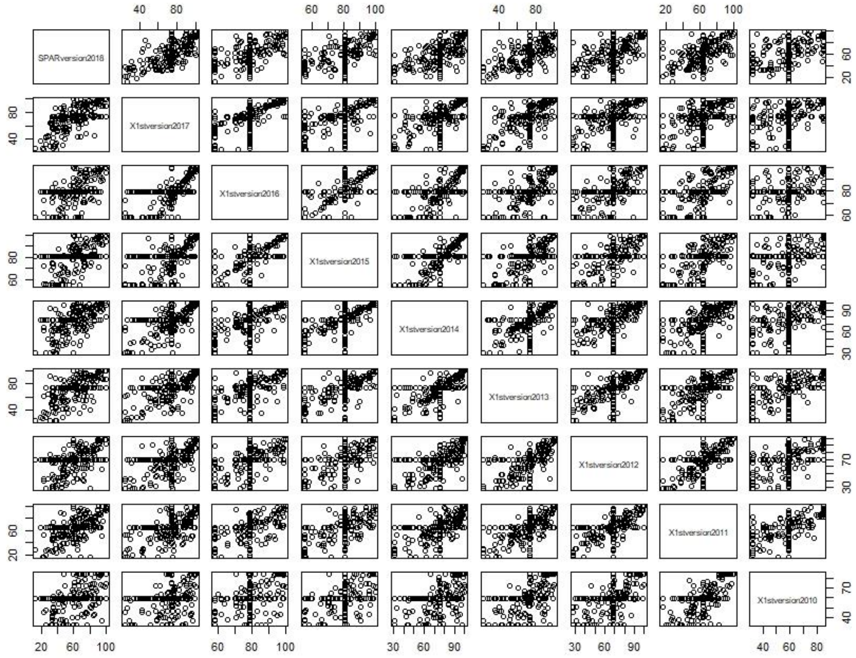


Fig.2 Scatter Plot

The Residual vs Fitted values can be used to interpret the test of regression assumptions. Ideally, the graph line for Residual vs Fitted values should be on the dotted grey line in the center with no pattern in between the data points. For the analyzed dataset, there exists a linear relationship between the DV and IV's where the red line of Residual vs Fitted values approximately approaching the grey dotted line. Assumption for Normally Distributed errors can be examined using the Normal Q-Q plot or Quantile-Quantile plot, which is a probability plot of Standardized Residuals against the Theoretical Quantiles. For this assumption to be reasonable, the residual error points should form a straight diagonal line in the graph. In our case, for Q-Q plot the points are relatively on the dotted straight line thus, qualifying the assumption. The following assumption for regression analysis is Homoscedasticity, which means that the error terms have constant variance. The best way to scrutinize this assumption is by using scatter plot of the Linear Regression model which plots a variance vs fitted values in Scale-Location graph. Evaluation of assumption is considerably analogous to Residual vs Fitted graph. If the Scale-Location graph delivers a straight horizontal line in the center with no pattern amongst the points then, the regression assumption is accomplished.

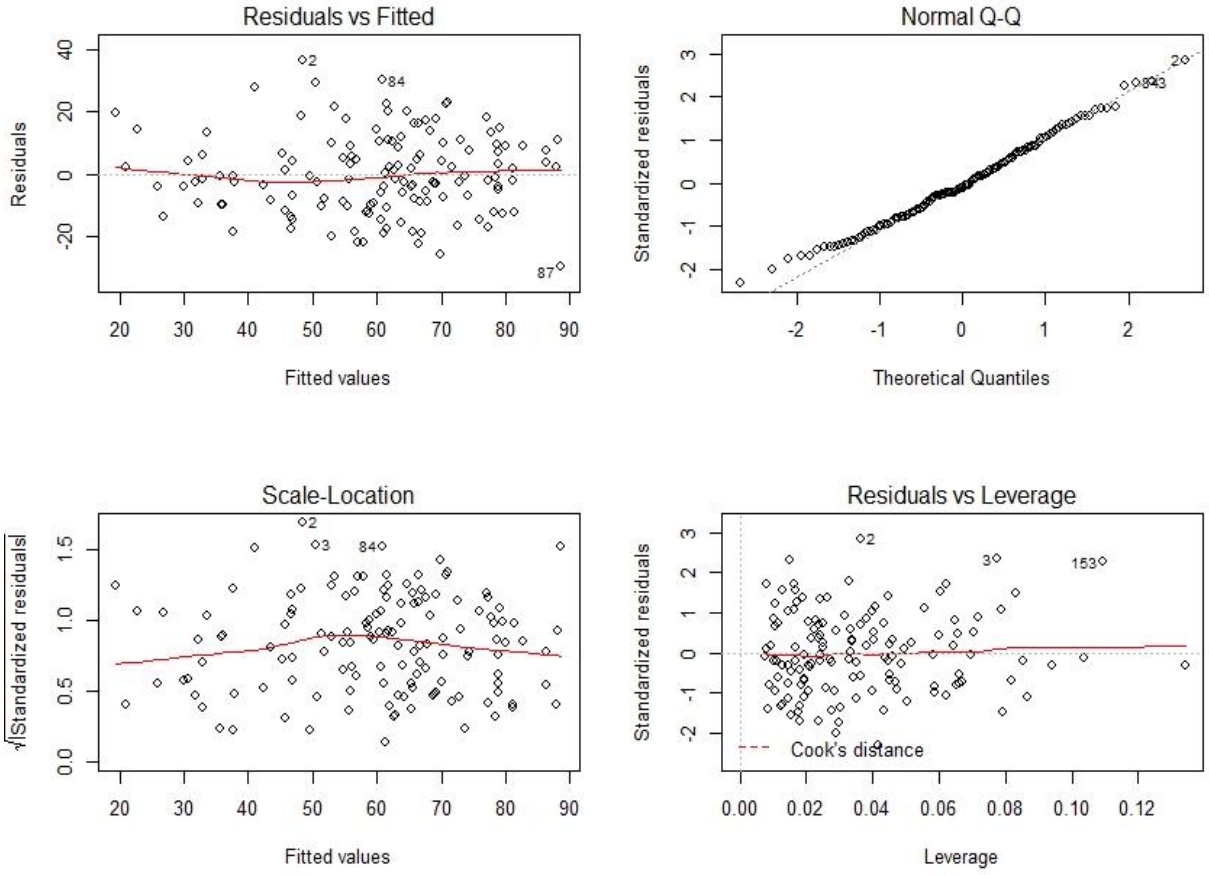


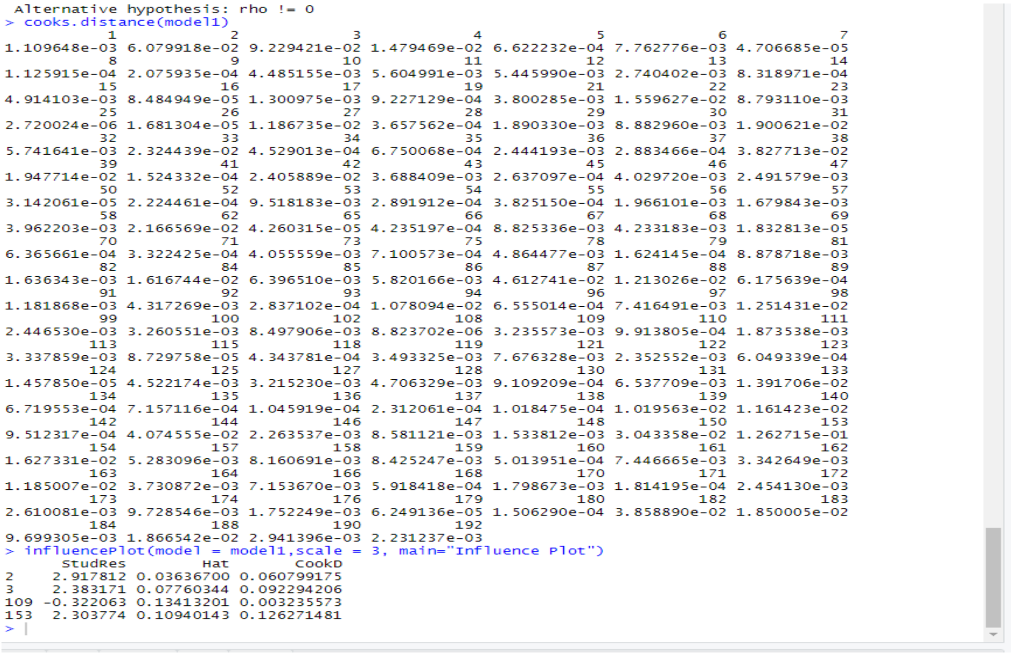
Fig.3 Scatter Plot of Regression model

Thus, the Linear Regression assumptions for Linearity, Normality Distribution and Homoscedasticity are qualified for the National and Global Health Risks dataset.

Another important assumption for Linear Regression is no Autocorrelation between the error terms. This is achieved by running a Durbin Watson Test in R using a 'lmtest' library. The assumption is achieved if the value of Durbin Watson statistics is between 1and 3 and more precisely close to 2. The value for DW statistics is 1.75, which is a clear evidence that concludes the assumption is achieved.

Cook’s distance can be used to evaluate the assumptions for no influential data points. It is carried using cooks.distance() in R under the library ‘car’. The data points with cook’s distance less than 1, demonstrates the assumption is meet.

Thus, all the important Linear Regression analysis assumptions are meet.



```
 Alternative hypothesis: rho != 0
> cooks.distance(model1)
           1            2            3            4            5            6            7 
1.109648e-03 6.079918e-02 9.229421e-02 1.479469e-02 6.622232e-04 7.762776e-03 4.706685e-05 
           8            9           10           11           12           13           14 
1.125915e-04 2.075935e-04 4.485155e-03 5.604991e-03 5.445990e-03 2.740402e-03 8.318971e-04 
          15           16           17           19           21           22           23 
4.914103e-03 8.484949e-05 1.300975e-03 9.227129e-04 3.800285e-03 1.559627e-02 8.793110e-03 
          25           26           27           28           29           30           31 
2.720024e-06 1.681304e-05 1.186735e-02 3.657562e-04 1.890330e-03 8.882960e-03 1.900621e-02 
          32           33           34           35           36           37           38 
5.741641e-03 2.324439e-02 4.529013e-04 6.750068e-04 2.444193e-03 2.883466e-04 3.827713e-02 
          39           41           42           43           45           46           47 
1.947714e-02 1.524332e-04 2.405889e-02 3.688409e-03 2.637097e-04 4.029720e-03 2.491579e-03 
          50           52           53           54           55           56           57 
3.142061e-05 2.224461e-04 9.518183e-03 2.891912e-04 3.825150e-04 1.966101e-03 1.679843e-03 
          58           62           65           66           67           68           69 
3.962203e-03 2.166569e-02 4.260315e-05 4.235197e-04 8.825336e-03 4.233183e-03 1.832813e-05 
          70           71           73           75           78           79           81 
6.365661e-04 3.322425e-04 4.055559e-03 7.100573e-04 4.864477e-03 1.624145e-04 8.878718e-03 
          82           84           85           86           87           88           89 
1.636343e-03 1.616744e-02 6.396510e-03 5.820166e-03 4.612741e-02 1.213026e-02 6.175639e-04 
          91           92           93           94           96           97           98 
1.181868e-03 4.317269e-03 2.837102e-04 1.078094e-02 6.555014e-04 7.416491e-03 1.251431e-02 
          99          100          102          108          109          110          111 
2.446530e-03 3.260551e-03 8.497906e-03 8.823702e-06 3.235573e-03 9.913805e-04 1.873538e-03 
         113          115          118          119          121          122          123 
3.337859e-03 8.729758e-05 4.343781e-04 3.493325e-03 7.676328e-03 2.352552e-03 6.049339e-04 
         124          125          127          128          130          131          133 
1.457850e-05 4.522174e-03 3.215230e-03 4.706329e-03 9.109209e-04 6.537709e-03 1.391706e-02 
         134          135          136          137          138          139          140 
6.719553e-04 7.157116e-04 1.045919e-04 2.312061e-04 1.018475e-04 1.019563e-02 1.161423e-02 
         142          144          146          147          148          150          153 
9.512317e-04 4.074555e-02 2.263537e-03 8.581121e-03 1.533812e-03 3.043358e-02 1.262715e-01 
         154          157          158          159          160          161          162 
1.627331e-02 5.283096e-03 8.160691e-03 8.425247e-03 5.013951e-04 7.446665e-03 3.342649e-03 
         163          164          166          168          170          171          172 
1.185007e-02 3.730872e-03 7.153670e-03 5.918418e-04 1.798673e-03 1.814195e-04 2.454130e-03 
         173          174          176          179          180          182          183 
2.610081e-03 9.728546e-03 1.752249e-03 6.249136e-05 1.506290e-04 3.858890e-02 1.850005e-02 
         184          188          190          192 
9.699305e-03 1.866542e-02 2.941396e-03 2.231237e-03 
> influencePlot(model = model1,scale = 3, main="Influence Plot")
      StudRes        Hat       CookD
2    2.917812 0.03636700 0.060799175
3    2.383171 0.07760344 0.092294206
109 -0.322063 0.13413201 0.003235573
153  2.303774 0.10940143 0.126271481
> |
```

Fig.4 Output for Cooks Distance

Analysis: Before performing the regression test, the dataset is divided into two subsets i.e. training and testing data. The training data comprises of 70% of the records which is used to train the model whereas the remaining records are used as the testing data to test model accuracy and to evaluate the best efficient model. In R, library ‘caret’ is used to divide the dataset using createDataPartition() function. A Linear Regression model is implemented using lm() function. The summary of the model provides the details for the significant variables denoted using the p-values. The other statistics like the RSE, R-square, Adjusted R-square, t-value and standard errors can be displayed using the summary of the model.

To remove the variables with p-values greater than .05%, stepwise regression is implemented with direction set to ‘both’. It is an iterative method which initially adds the most significant variable like forward selection and then removes any redundant variable using backward selection to build an efficient regression model. A Linear Regression model is implemented which formulates a relationship between the predictors and the response variable using the following formula.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

where, y= Response variable

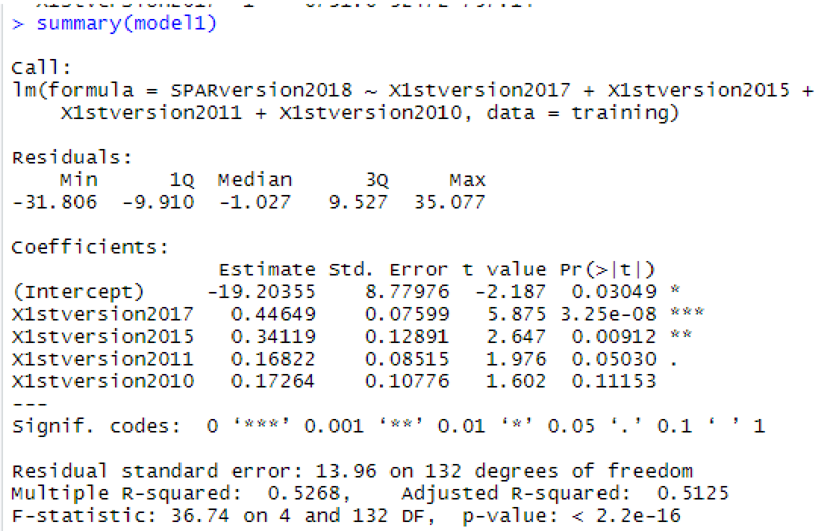
β_0= Intercept

$\beta_1 x_1$ = Coefficient of $x_1 * x_1$

$\beta_n x_n$= Coefficient of $x_n * x_n$

x_1 & x_n= Independent variables in the datasets

The summary statistics of the model is displayed in the figure below.



```
> summary(model1)

Call:
lm(formula = SPARversion2018 ~ X1stversion2017 + X1stversion2015 + 
    X1stversion2011 + X1stversion2010, data = training)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.806  -9.910  -1.027   9.527  35.077 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -19.20355    8.77976  -2.187  0.03049 *  
X1stversion2017   0.44649    0.07599   5.875 3.25e-08 ***
X1stversion2015   0.34119    0.12891   2.647  0.00912 ** 
X1stversion2011   0.16822    0.08515   1.976  0.05030 .  
X1stversion2010   0.17264    0.10776   1.602  0.11153    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.96 on 132 degrees of freedom
Multiple R-squared:  0.5268,    Adjusted R-squared:  0.5125 
F-statistic: 36.74 on 4 and 132 DF,  p-value: < 2.2e-16
```

Fig.5 Summary of Regression model

In summary statistics, we get the p-values for each independent variable which are used to determine the significant variables for the model. These p-values are calculated using the standard error and the t-statistics. Each p-value and t-statistics are associated with some hypothesis. i.e. Null Hypothesis and Alternate Hypothesis. For Null Hypothesis the beta coefficient associated with variables are equal to zero, whereas the Alternate Hypothesis is given as the beta coefficients are not equal to zero. Also, with higher value of t-statistics, it is less likely for Null Hypothesis to be true. Similarly, lower t-values corresponds to high p-values, thus interpreting the corresponding variable to be less significant. For independent variable to be significant, the p-value should be less than the statistically significant level of 0.05. In our case, the p-values are relatively below the 0.05 threshold, so we can terminate the Null Hypothesis and establish that the model is statistically significant.

Other important parameters are R^2 and Adjusted R^2 values. R^2 value explains the proportion of variance in the response variable that are explained by the predictors irrespective of the variable significant in the model. The total information in a variable is given as the amount of variance it contains, thus the value of R^2 will increase with the increase in the number of predictors. On the other hand, Adjusted R^2 penalizes if the added variable is not significant. It means that it calculates the R^2 for the added variables which improves the model. Thus, while evaluating the model Adjusted R^2 should be taken into consideration. In our case, the Adjusted R^2 is comparably about 52%. To increase this value more significant variables should be added to model.

Name: Raj R. Kupekar
Student Id: 18186432

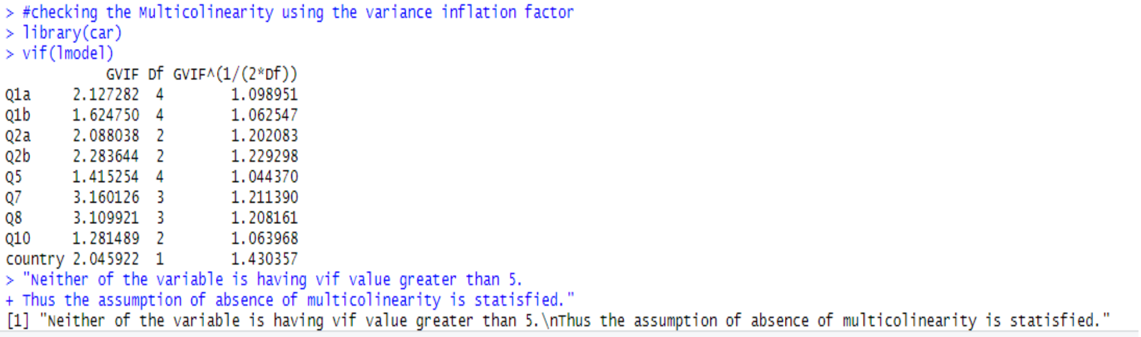
Logistics Regression

Objectives: The goal is to carry out a statistical analysis using binomial Logistics Regression. The analysis is performed on the Global Attitudes and Trends dataset which is a US-Germany relationship survey. A relationship between the DV and IV's is tested using Logistics Regression. The opinions of public regarding the alliance of US-Germany is analyzed which is dependent on other factors

The data is collected from Pew Research Center which is an unbiased body that undertakes public surveys regarding the issues, trends as well as attitudes with respect to the phenomenon happening around the world. A set of questionnaires were asked to individuals to discover their attitudes towards the alliance relationship between US and Germany. The survey was conducted either on landline or on cell phones of everyone.

Data Transformation: In Linear Regression the data type of all the variables should be in a numeric form; but this is not the case with Logistics Regression. For Logistics, only the dependent variable should be in a categorical form and rest of the variables can take any data type. Initially, the DV was in integer type so, as per the requirement it is converted to a categorical format using as.factor() function in R. Similarly, the variables holding the categorical values with data type as integer are converted to their respective factor format. To perform binomial Logistics Regression, it is clear to have only two levels in our DV. In our case, the DV is having more than two levels. Thus, the multinomial DV is remodeled into two levels using recode() function in R. The '1' and '2' levels are merged to one single level as '1', whereas the '3', '4' and '9' level is merged to one single level as '0'. No action is taken against the missing values and outliers because of their absence in the data frame.

Assumptions: Logistics Regression truly does not have much assumptions on dataset. Firstly, binary Logistics Regression requires the response variable to be binary or ordinal. This is achieved as explained in data transformation above. Also, the outcomes of response variable must be mutually exclusive which means that both the events cannot occur at the same time. To implement Logistics model the dataset should be large enough with a greater number of records. Finally, the last assumption is absence of Multicollinearity. This assumption is tested using the vif factor as tested in Linear Regression. In our case the vif values for each variable are less than 5, thus satisfying the assumption.



```
> #checking the Multicolinearity using the variance inflation factor
> library(car)
> vif(lmodel)
            GVIF Df GVIF^(1/(2*Df))
Q1a     2.127282  4        1.098951
Q1b     1.624750  4        1.062547
Q2a     2.088038  2        1.202083
Q2b     2.283644  2        1.229298
Q5      1.415254  4        1.044370
Q7      3.160126  3        1.211390
Q8      3.109921  3        1.208161
Q10     1.281489  2        1.063968
country 2.045922  1        1.430357
> "Neither of the variable is having vif value greater than 5.
+ Thus the assumption of absence of multicolinearity is statisfied."
[1] "Neither of the variable is having vif value greater than 5.\nThus the assumption of absence of multicolinearity is statisfied."
```

Fig.6 Assumption testing using VIF value

Analysis: Logistics Regression is a predictive modelling which can be used for regression as well as classification. To carry out analysis the data is divided into training and testing datasets are discussed in Linear Regression. Binomial Logistics Regression is implemented by calling the glm() function in R with family set to 'binomial'. To remove the effect of multicollinearity and insignificant variables, stepwise regression is executed as explained in Linear Regression. Thus, a binary Logistics Regression is assembled for which the dependent variable is dichotomous. It takes in a Sigmoid function which is S-curved in shape to predict the response variables. It works on the principle of Maximum Likelihood Estimation (MLE) which estimates the values for coefficient that minimizes the error terms. The MLE indeed is calculated using the odds ratio and log of odds ratio values. The odd ratio can be defined as the probability of ratio of an event occurring to the probability of event not occurring. Whereas, the log of odds or logit is given as the log of the odds ratio which are the estimates of variables. The mathematical calculation for odds ratio and logit is given as follows;

$$\text{odds ratio} = \frac{p(X)}{1-p(X)}$$

$$\text{logit} = \log\left(\frac{p(X)}{1-p(X)}\right)$$

The values of logit are the Maximum Likelihood estimates for the Logistics model. Thus, the final equation for Logistics Regression calculated using the MLE principle is given as;

$$E(y) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n}}$$

where, y= Response variable

e= Exponent

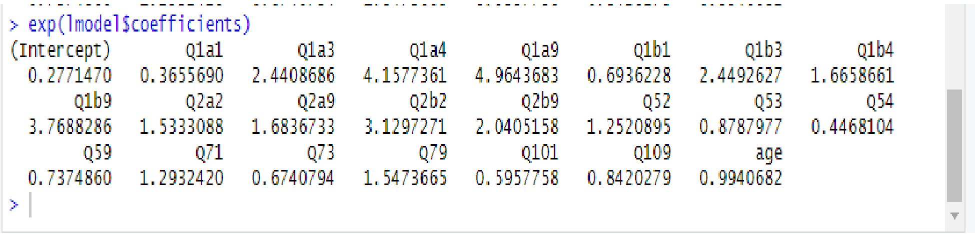
β_0= Intercept

$\beta_1 x_1$ = Coefficient of x_1* x_1

$\beta_n x_n$= Coefficient of x_n * x_n

x_1 & x_n= Independent variables in the datasets

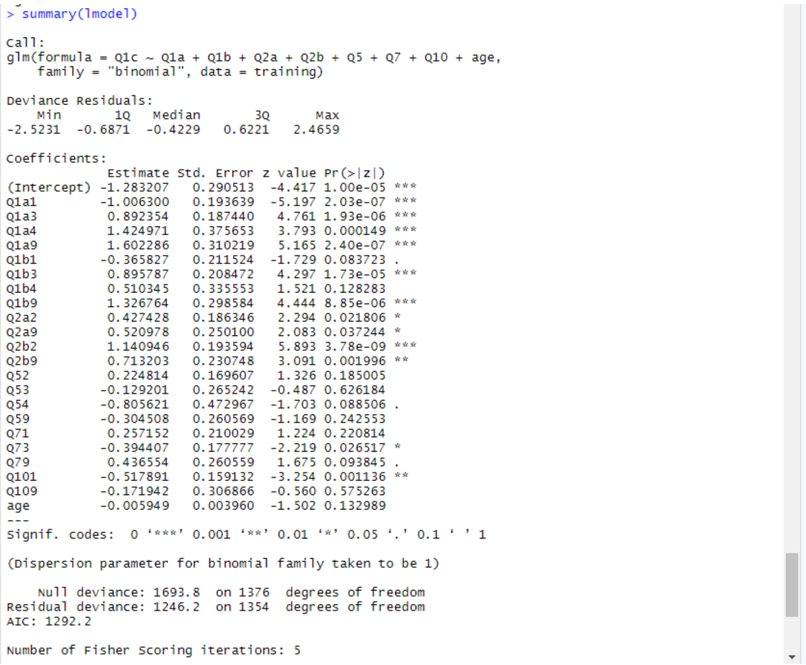
In R, Odds Ratio is calculated by taking exponent of coefficients of the Logistics model. The outputs for odd ratio are displayed below.



```
> exp(lmodel$coefficients)
(Intercept)        Q1a1        Q1a3        Q1a4        Q1a9        Q1b1        Q1b3        Q1b4 
  0.2771470   0.3655690   2.4408686   4.1577361   4.9643683   0.6936228   2.4492627   1.6658661 
       Q1b9        Q2a2        Q2a9        Q2b2        Q2b9         Q52         Q53         Q54 
  3.7688286   1.5333088   1.6836733   3.1297271   2.0405158   1.2520895   0.8787977   0.4468104 
        Q59         Q71         Q73         Q79        Q101        Q109         age 
  0.7374860   1.2932420   0.6740794   1.5473665   0.5957758   0.8420279   0.9940682 
> 
```

Fig.7 Odds Ratio of model

Before implementing the model, it is important to give the references to the categorical variables to build an efficient model. By default, the model takes the 1st value as the reference. But, if a level in variable is having a maximum count other than the default 1st level, then that level is taken as the reference value for that respective variable. The summary of the model gives us the other relevant statistics. In Logistics Regression as variance is calculated from mean and not from data, there is a possibility that variance is underestimated. Thus, to avoid this effect, the dispersion parameter can be adjusted in the summary command. By default, the dispersion parameter is taken as 1. From summary command, it is evident that the variables with significant p-values are retained in the model. But there are variables whose p-values are greater than 0.05 level. This because, while working with categorical variable even if one of the levels of that variable is significant then that variable for which the other levels are insignificant are also kept in the model. For example, in our case the p-value for Q1b4 is 0.12 which is greater than the significant level of 0.05. Practically, this variable is needed to be removed from the model. But the model keeps this variable as because the same variable with different level is more significant for the model. Thus, the model retains both the levels and builds an efficient Logistics Regression Classification.

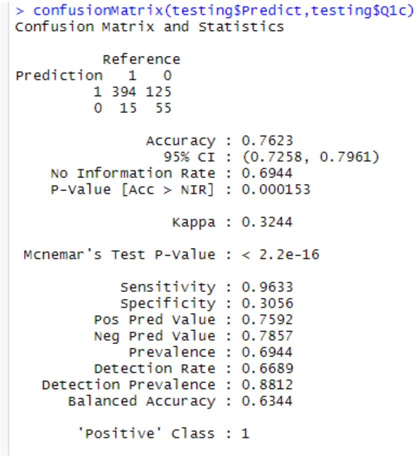


```
> summary(lmodel)

Call:
glm(formula = Q1c ~ Q1a + Q1b + Q2a + Q2b + Q5 + Q7 + Q10 + age,
    family = "binomial", data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5231  -0.6871  -0.4229   0.6221   2.4659

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.283207   0.290513  -4.417 1.00e-05 ***
Q1a1        -1.006300   0.193639  -5.197 2.03e-07 ***
Q1a3         0.892354   0.187440   4.761 1.93e-06 ***
Q1a4         1.424971   0.375653   3.793 0.000149 ***
Q1a9         1.602286   0.310219   5.165 2.40e-07 ***
Q1b1        -0.365827   0.211524  -1.729 0.083723 .
Q1b3         0.895787   0.208472   4.297 1.73e-05 ***
Q1b4         0.510345   0.335553   1.521 0.128283
Q1b9         1.326764   0.298584   4.444 8.85e-06 ***
Q2a2         0.427428   0.186346   2.294 0.021806 *
Q2a9         0.520978   0.250100   2.083 0.037244 *
Q2b2         1.140946   0.193594   5.893 3.78e-09 ***
Q2b9         0.713203   0.230748   3.091 0.001996 **
Q52          0.224814   0.169607   1.326 0.185005
Q53         -0.129201   0.265242  -0.487 0.626184
Q54         -0.805621   0.472967  -1.703 0.088506 .
Q59         -0.304508   0.260569  -1.169 0.242553
Q71          0.257152   0.210029   1.224 0.220814
Q73         -0.394407   0.177777  -2.219 0.026517 *
Q79          0.436554   0.260559   1.675 0.093845 .
Q101        -0.517891   0.159132  -3.254 0.001136 **
Q109        -0.171942   0.306866  -0.560 0.575263
age         -0.005949   0.003960  -1.502 0.132989
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1693.8  on 1376  degrees of freedom
Residual deviance: 1246.2  on 1354  degrees of freedom
AIC: 1292.2

Number of Fisher Scoring iterations: 5
```

Fig.8 Summary of glm model

The number of iterations denotes that the model underwent 5 iteration to establish the best possible estimates for MLE. The Akaike Information Criterion (AIC) is used compare the models. Lower the value of AIC the better is the model.

In R, predict() function is used for prediction with type set to ‘response’. This returns the probability values predicted by the model for the dependent variable. Hence, by setting the correct threshold values the class of the response variable can be accurately precited. In our case, the threshold is set to 0.70. Which means that if the response value is greater than 0.70, it will classify that as ‘0’ otherwise ‘1’.



```
> confusionMatrix(testing$Predict,testing$Q1c)
Confusion Matrix and Statistics

          Reference
Prediction   1   0
         1 394 125
         0  15  55

               Accuracy : 0.7623
                 95% CI : (0.7258, 0.7961)
    No Information Rate : 0.6944
    P-Value [Acc > NIR] : 0.000153

                  Kappa : 0.3244

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9633
            Specificity : 0.3056
         Pos Pred Value : 0.7592
         Neg Pred Value : 0.7857
             Prevalence : 0.6944
         Detection Rate : 0.6689
   Detection Prevalence : 0.8812
      Balanced Accuracy : 0.6344

       'Positive' Class : 1
```

Fig.9 Confusion Matrix

The actual accuracy of the model is evaluated using the confusion matrix under the ‘e1071’ library. In our case, the model accuracy is about 76 %. This accuracy can be tuned by modifying the threshold value during prediction.

The accuracy with 95% of Confidence Interval is about 72%. Similarly, the other statistics like the sensitivity, specificity, kappa value etc. can be displayed using Confusion Matrix.

Summary: Thus, model implementation and analysis of multiple Linear Regression and binomial Logistics Regression is studied. Both are Supervised Machine Learning algorithm and are used for statistical predictive modelling. The limitation in Linear Regression of linear relationship are overcomed by Logistics Regression. Also, each technique has their own advantages depending on different problem cases. In case of Linear Regression, an efficient model can be built by including significant variables in the model. Whereas, for Logistics Regression the model accuracy can be increased by altering the threshold value for probability.