Inspira-Journal of Commerce, Economics & Computer Science (JCECS)
ISSN : 2395-7069 General Impact Factor : 2.4668, Volume 04, No. 01, January-March., 2018, pp. 100-112

PREDICTING THE OCCURRENCE DEATHS DUE TO CARDIO-VASCULAR DISEASE (CVD)

Er. Ravija Sharma*
Dr. R.K. Sharma**

ABSTRACT

Cardiovascular disease is the leading cause of deaths worldwide, though since the 1970s, cardiovascular mortality rates have declined in many high-income countries. At the same time, cardiovascular deaths and disease have increased at a fast rate in low- and middle-income countries. Although cardiovascular disease usually affects older adults, the antecedents of cardiovascular disease, notably atherosclerosis, begin in early life, making primary prevention efforts necessary from childhood.

***KEYWORDS**: Cardiovascular Disease (CVD), Cardiovascular Deaths, Cardiovascular Mortality Rates.*

Introduction

Cardiovascular disease (also called heart disease) is a class of diseases that involve the heart, the blood vessels (arteries, capillaries, and veins) or both. Cardiovascular disease refers to any disease that affects the cardiovascular system, principally cardiac disease, vascular diseases of the brain and kidney, and peripheral arterial disease. The causes of cardiovascular disease are diverse but atherosclerosis and/or hypertension are the most common. Additionally, with aging come a number of physiological and morphological changes that alter cardiovascular function and lead to subsequently increased risk of cardiovascular disease, even in healthy asymptomatic individuals. As per World Health Organization's report, around 17.7 million people died due to CVDs in 2015, which is about 31% of all the worldwide demises. Out of these, 7.4 million deaths were due to coronary heart disease and 6.7 million were due to stroke. And more than 3/4th of CVD deaths happened in low and middle income countries. People in low-income countries & middle-income countries often do not have the benefit of integrated primary health care programs for early detection and treatment and hence, are at a higher risk compared to people in high-income countries. Leading causes for CVDs include behavioral risk factors such as consumption of tobacco, unhealthy diet and obesity, physical inactivity and harmful use of alcohol.

Research Objectives

- To study the factors responsible for deaths due to cardiovascular diseases by building an effective predictive model.
- Estimation of number of deaths due to CVDs based on the variation in the independent variables.
- To find relationship between the health index of a nation with respect to its economic development.

Data Collection & Sampling Technique

For addressing the research questions, data is required at a country level; hence, the best way to collect the data is secondary research. There are reports published by credible organizations like WHO, UN, World Bank, Central Intelligence Agency etc. containing data on some of the important parameters likeGDP, per-capita alcohol consumption, health care expenditure, CVD related deaths, literacy rate, percentage of urban population etc.

* M.Tech.(Quality Management), Birla Institute of Technology & Science, Pilani, Rajasthan, India.

** Associate Professor, Department of EAFM, Government Shakambhar P.G. College Sambhar Lake, Jaipur, Rajasthan, India.

Variable Selection for Answering the Research Questions

Based on the secondary research on CVDs as well as some primary research done with some of the expert cardiologists in Jaipur, it has become conclusive that effective health care measures can drastically reduce the number of deaths due to CVDs. Availability of proper health care facilities in a country is dependent on multiple factors like its economic development, government's health care programs, urbanization etc. Timely diagnosis and treatment can help us drastically reduce death rates due to CVD, which is also dependent on the awareness level in the country. Literacy rate can be considered as a good proxy for the awareness level in the country. CVDs are a lifestyle disease and will be higher if the number of obese population in a country is high. Similarly, smoking and consumption of alcoholic beverages can increase the changes of CVDs.

Research Hypothesis

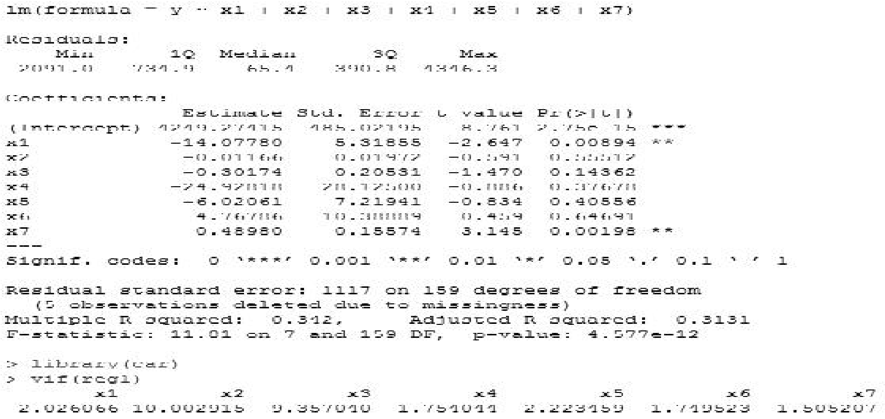
- **Deaths due to Cardiovascular Disease are Dependent on GDP per capita (PPP):** A higher GDP per capita implies higher per capita spending power on health facilities which might lead to a lower death rate due to cardiovascular diseases.
- **Deaths due to Cardiovascular Disease are Dependent on % of Urban Population:** A higher degree of urbanization can mean that a higher percentage of people live in towns and cities with greater availability of qualified doctors and good medical facilities which can prevent deaths.
- **Deaths due to Cardiovascular Disease is Dependent on Per Capita Govt. Spending on Health as % Of GDP:** The higher a government spends on health, the better are the medical facilities available to the population and the lower is the expected fatality rates due to cardiovascular disease.
- **Deaths Due to Cardiovascular Disease are Dependent on Literacy Rate:** A higher literacy rate can mean a higher degree of awareness about health risks and healthy diet which should reduce the deaths due to cardiovascular diseases.
- **Deaths Due to Cardiovascular Disease are Dependent on % of Obese Population:** Obesity is directly related to cardiovascular diseases and hence, to deaths resulting from these.
- **Deaths Due to Cardiovascular Disease are Dependent on Per Capita Alcohol Consumption Per Year (in Liters):** Alcohol consumption is also known to increase the risks of cardiovascular ailments which can ultimately result in death.

Forming the Regression Model

The methodologies used are predictive techniques like **Linear Regression, Robust Regression, and Generalized Additive Models**. The first step was to make hypothesis based on the problem statement to predict the number of deaths due to CVDs and figure out the parameters that influence it. Below is a step-by-step analysis of the predictive models built for the research purpose.

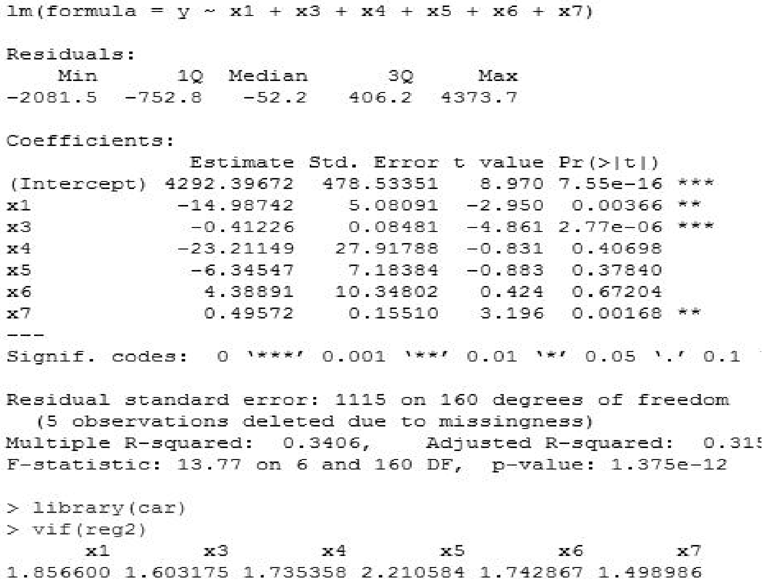
Step 1: Removal of Multicollinearity from the data

Multi-collinearity means, one or more of the predictor variables being linearly dependent on few other predictor variables. There is a very important consequence to this in linear regression as it leads to ambiguous values of regression co-efficient. We will first need to remove variables from our consideration which have high VIF values and then proceed to build our regression model. Using VIF criterion to eliminate Multicollinearity (if present) by dropping variables. Drop one variable at a time until Multicollinearity is eliminated. Here are the results from the Multicollinearity analysis:



```
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7)

Residuals:
    Min      1Q  Median      3Q     Max
-2091.0  -734.9    65.4   390.8  4346.3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4249.27415  485.02195   8.761 2.75e-15 ***
x1           -14.07780    5.31855  -2.647  0.00894 **
x2            -0.01166    0.01972  -0.591  0.55512
x3            -0.30174    0.20531  -1.470  0.14362
x4           -24.92818   28.12500  -0.886  0.37678
x5            -6.02061    7.21941  -0.834  0.40556
x6             4.76786   10.38889   0.459  0.64691
x7             0.48980    0.15574   3.145  0.00198 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1117 on 159 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R squared:  0.312,     Adjusted R squared:  0.3131
F-statistic: 11.01 on 7 and 159 DF,  p-value: 4.577e-12

> library(car)
> vif(reg1)
       x1        x2        x3        x4        x5        x6        x7
 2.026066 10.002915  9.357040  1.754044  2.223459  1.749523  1.505207
```

Dropping GDP (per capita) from the Multicollinearity Analysis, we get the Final Multicollinearity Analysis as Follows:



```
lm(formula = y ~ x1 + x3 + x4 + x5 + x6 + x7)

Residuals:
    Min      1Q  Median      3Q     Max
-2081.5  -752.8   -52.2   406.2  4373.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4292.39672  478.53351   8.970 7.55e-16 ***
x1           -14.98742    5.08091  -2.950  0.00366 **
x3            -0.41226    0.08481  -4.861 2.77e-06 ***
x4           -23.21149   27.91788  -0.831  0.40698
x5            -6.34547    7.18384  -0.883  0.37840
x6             4.38891   10.34802   0.424  0.67204
x7             0.49572    0.15510   3.196  0.00168 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 1115 on 160 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.3406,    Adjusted R-squared:  0.31
F-statistic: 13.77 on 6 and 160 DF,  p-value: 1.375e-12

> library(car)
> vif(reg2)
      x1       x3       x4       x5       x6       x7
1.856600 1.603175 1.735358 2.210584 1.742867 1.498986
```

Step 2: Treatment of Outliers from the Data

Remove the outliers using the criteria: *Standard Residual >2.0***.**

Outliers are removed one after another.

We have removed *18 observations* which constituted *approximately 10%* of the total observation points. The following countries turned out to be outliers:

Regression Analysis: Urban Population (versus Literacy, Health Expenditure, ...

The regression equation is:

Urban Popu (%) = 11.5 + 0.352 Literacy + 0.00653 Health Expenditure (per capita)
- 0.107 Alcohol Consumption (Per Capita + 0.473 Obesity (%)

Predictor	Coef	SE Coef	T	P	VIF
Constant	11.500	7.356	1.56	0.120	
Literacy	0.3522	0.1079	3.26	0.001	1.993
Health Expenditure (per capita)	0.006527	0.001242	5.26	0.000	1.356
Alcohol Consumption (per capita)	-0.1071	0.4222	-0.25	0.800	1.584
Obesity (%)	0.4728	0.1523	3.11	0.002	1.608

S = 17.7650   R-Sq = 43.2%   R-Sq (adj) = 41.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	39586.0	9896.5	31.36	0.000
Residual Error	165	52073.0	315.6		
Total	169	91659.0			

Source	DF	Seq SS
Literacy	1	26676.5
Health Expenditure (per capita)	1	9854.3
Alcohol Consumption (per capita)	1	11.3
Obesity (%)	1	3043.9

Run regression and check for influential observations, outliers and high leverage observations.

Step 3a: If observations marked RX are seen then one of these observations is deleted and regression is rerun. Go back to Step 2.

We removed 18 observations which constituted approximately 10% of the total observation points. The following countries turned out to be outliers**:**

Angola	Iraq	Micronesia, Fed. Sts.	Samoa
Belarus	Kazakhstan	Namibia	Solomon Islands
Bulgaria	Kyrgyz Republic	Qatar	Turkmenistan
Guatemala	Malta	Russian Federation	Ukraine
United Kingdom	Uzbekistan		

A closer analysis of the names of the countries in the above list yields an interesting observation. Most of these countries are from the Russian and the eastern European region. Some of the outliers are very small countries with small populations. The other outliers are from the Middle East and Africa. If we talk of countries with a relatively larger size and a significant size of population, none of the outliers are from the rest of Asia and Europe, South America, North America or from Asia-Pacific and Pacific regions. This probably implies that genetics and regional lifestyles and even weather may play a very important role in understanding fatality rates due to cardiovascular diseases. These factors need to be closely examined an independent variable accounting for the region/ race should be included to make the regression analysis more robust. However, for our analysis we proceed without such a variable.

Regression Analysis: CVD deaths versus Literacy, Health Expenditure, ...

```
The regression equation is
CVD deaths = 3945 - 2.67 Literacy - 0.454 Health Expenditure (per capita)
+ 7.6 Alcohol Consumption (Per Capita + 1.9 Obesity (%)
             - 9.25 Urban Popu (%)
        Predictor                        Coef        SE Coef     T        P        VIF
        Constant                         3944.9      484.0       8.15     0.000
        Literacy                         -2.672      7.273       -0.37    0.714    2.121
        Health Expenditure (per capita)  -0.45351    0.08763     -5.18    0.000    1.583
        Alcohol Consumption (per capita) 7.59        27.58       0.28     0.784    1.585
        Obesity (%)                      1.89        10.23       0.18     0.854    1.702
        Urban Popu (%)                   -9.248      5.085       -1.82    0.071    1.760

        S = 1160.36   R-Sq = 28.2%   R-Sq (adj) = 26.0%
```

Analysis of Variance

```
        Source                           DF       SS          MS          F        P
        Regression                       5        86545113    17309023  12.86    0.000
        Residual Error                   164      220816864  1346444
        Total                            169      307361977

        Source                           DF       Seq SS
        Literacy                         1        21441937
        Health Expenditure (per capita)  1        60439746
        Alcohol Consumption (Per Capita  1        126101
        Obesity (%)                      1        83787
        Urban Popu (%)                   1        4453541
```

Unusual Observations

```
Obs     Literacy         CVD deaths    Fit       SE Fit      Residual    St Resid
76      78               5984.0        3139.2    195.3       2844.8      2.49R
82      100              6748.0        3209.2    140.1       3538.8      3.07R
94      100              1386.0        272.4     400.9       1113.6      1.02 X
99      92               5740.0        2360.9    201.0       3379.1      2.96R
103     89               843.0         3508.1    307.8       -2665.1     -2.38R
107     89               7491.0        3355.5    164.9       4135.5      3.60R
```

114	100	1063.0	586.0	384.2	477.0	0.44 X
125	96	4450.0	2133.2	264.3	2316.8	2.05R
127	100	6296.0	3039.9	209.9	3256.1	2.85R
129	99	1385.0	3566.0	421.7	-2181.0	-2.02RX
139	95	6006.0	3574.5	281.0	2431.5	2.16R
153	99	3251.0	3548.3	449.1	-297.3	-0.28 X
157	100	7355.0	3237.0	202.2	4118.0	3.60R
162	99	1525.0	183.8	450.3	1341.2	1.25 X

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

Step 4: Keep doing this until there is no influential observations/ outliers/ high leverage observations or 10% of data has been removed. Multiple iterations yield the following regression analysis:

[Total Data points removed: 18 (10% of Observations)]

Regression Analysis: CVD deaths versus Literacy, Health Expenditure, ...

The regression equation is

CVD deaths = 4744 - 14.8 Literacy - 0.333 Health Expenditure (per capita)
+ 7.1 Alcohol Consumption (Per Capita + 18.8 Obesity (%)
- 16.2 Urban Popu (%)

Predictor	Coef	SECoef	T	P	VIF
Constant	4743.6	318.1	14.91	0.000	
Literacy	-14.822	4.889	-3.03	0.003	2.213
Health Expenditure (per capita)	-0.33334	0.05826	-5.72	0.000	1.639
Alcohol Consumption (per capita)	7.13	18.89	0.38	0.706	1.647
Obesity (%)	18.771	7.300	2.57	0.011	1.839
Urban Popu (%)	-16.171	3.615	-4.47	0.000	1.938

S = 742.608 R-Sq = 53.4% R-Sq(adj) = 51.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	92250326	18450065	33.46	0.000
Residual Error	146	80514232	551467		
Total	151	172764559			

Source	DF	Seq SS
Literacy	1	41252393
Health Expenditure (per capita)	1	38767152
Alcohol Consumption (per capita)	1	321883
Obesity (%)	1	871921
Urban Popu (%)	1	11036978

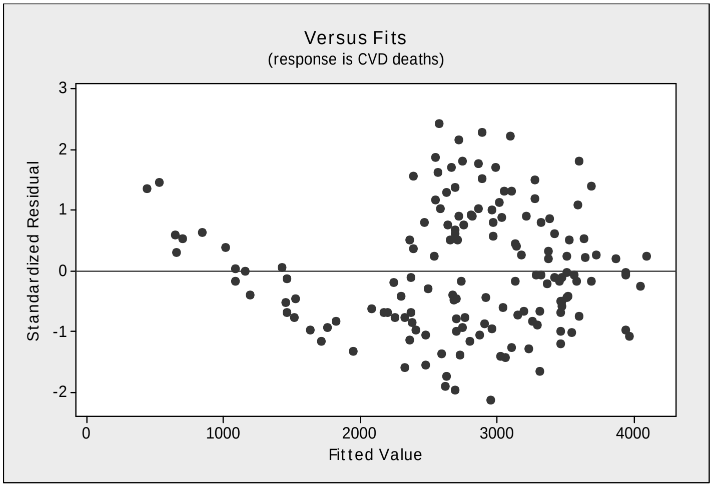
Unusual Observations

Obs	Literacy	CVD deaths	Fit	SE Fit	Residual	St Resid
6	100	4306.0	2722.6	130.6	1583.4	2.17R
9	100	4571.0	2889.2	105.0	1681.8	2.29R
47	84	1382.0	2950.2	94.8	-1568.2	-2.13R
52	94	4719.0	3099.5	135.8	1619.5	2.22R
87	100	1386.0	442.6	261.1	943.4	1.36 X
95	97	4344.0	2575.2	134.2	1768.8	2.42R
139	99	3251.0	3968.6	334.1	-717.6	-1.08 X
145	99	1525.0	539.1	298.0	985.9	1.45 X

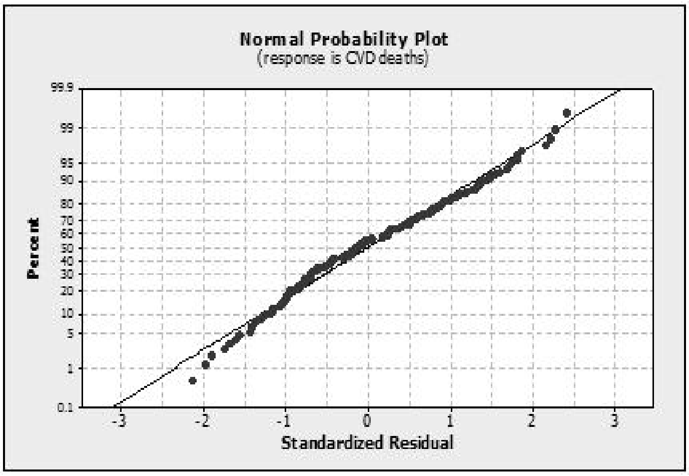
R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

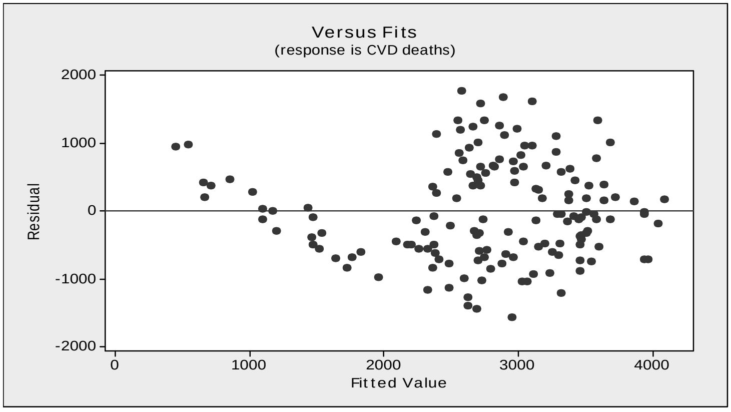
Step 5: Check for Regression Assumptions of Linearity, Normality and Homoscedasticity Using Residual Plot and Normal Probability Plots



Linearity Assumption: Violated, for values of Yi <2000, but beyond that the residual values do not seem to depend on the predicted number of deaths per 100000 due to cardiovascular diseases.



Normality Assumption: The assumption holds somewhat as the std. residual plot is a straight line



Homoscedasticity: Violated, for values of Yi <2000, but beyond that the regression works fine

Step 6: If homoscedasticity assumption is suspected to be violated then check for the same using the B-P test. (The p-value of the B-P test should be greater than 0.1 for homoscedasticity assumption to hold.) B-P Test Was not performed.

Step 7: If some of the assumptions of linear regression model is found to be violated then youmay try to use transformations to rectify the same. If NO transformation can be found which can correct for the violations then STOP. (Linearity assumptions and homocedasticity assumptions violated only for Yi < 2000. But the majority of our data points had Yi> 2000).

Step 8: If all the regression assumptions are met then look at the p-value of the F-test. If it is not significant (i.e. p-value >0.1) then STOP

The P-value was found to be 0.

Step 9: If the p-value of the F-test is significant then look at the p-values of the individual coefficients. If all the coefficient p-values are significant (i.e. ≤ 0.1) then go to Step 12.The p-value of the F-test is found to be equal to and is thus significant. When we look at p-values of the individual co-efficients and find out that the p-value for the factor- alcohol consumption is not significant

The p-value of the F-test is found to be equal to 0 and is thus significant.

When we look at p-values of the individual co-efficients and find out that the p-value for the factor- alcohol consumption is not significant

Step 10: If some of the p-values are not significant (i.e. > 0.1) then choose one of the variables with non-significant p-value and drop it from the model and run regression again.

We decide to drop the factor of alcohol consumption and redo the regression process. Here are the results after dropping alcohol consumption as an input variable:

We decide to drop the factor of alcohol consumption and redo the regression process. Here are the results after dropping alcohol consumption as an input variable:

Regression Analysis: CVD deaths versus Literacy, Health Expenditure, ...

```
The regression equation is:
        CVD deaths = 4724 - 14.2 Literacy - 0.325 Health Expenditure (per capita)
        + 19.0 Obesity (%) - 16.3 Urban Popu (%)
Predictor                      Coef      SE Coef   T       P       VIF
Constant                       4724.3    313.0     15.09   0.000
Literacy                       -14.223   4.611     -3.08   0.002   1.980
Health Expenditure (per capita) -0.32506 0.05381   -6.04   0.000   1.406
Obesity (%)                    18.974    7.259     2.61    0.010   1.829
Urban Popu (%)                 -16.267   3.595     -4.52   0.000   1.928

S = 740.439   R-Sq = 53.4%   R-Sq(adj) = 52.1%
```

Analysis of Variance

```
Source                         DF        SS          MS         F       P
Regression                     4         92171789    23042947   42.03   0.000
Residual Error                 147       80592770    548250
Total                          151       172764559

Source                         DF        Seq         SS
Literacy                       1         41252393
Health Expenditure (per capita) 1        38767152
Obesity (%)                    1         928231
Urban Popu (%)                 1         11224013
```

Unusual Observations

Obs	Literacy	CVD deaths	Fit	SE Fit	Residual	St Resid
6	100	4306.0	2690.8	99.4	1615.2	2.20R
9	100	4571.0	2878.2	100.7	1692.8	2.31R
47	84	1382.0	2963.6	87.7	-1581.6	-2.15R
52	94	4719.0	3128.5	111.5	1590.5	2.17R
87	100	1386.0	445.4	260.2	940.6	1.36 X
95	97	4344.0	2602.6	112.5	1741.4	2.38R
133	99	866.0	665.0	238.1	201.0	0.29 X
139	99	3251.0	3995.6	325.4	-744.6	-1.12 X
145	99	1525.0	570.5	285.3	954.5	1.40 X

R denotes an observation with a large standardized residual.

X denotes an observation whose X value gives it large leverage.

Step 11: Repeat Step 10 until you get the p values of all the coefficients significant (i.e. p value ≤ 0.1) We now have the p-values for all the co-efficient less than 0.1

Step 12. Check if the R^2 and R^2 (adj.) are close. If yes, your model has been built. You can now use the model appropriately based on the R2/R2 (adj.) value.

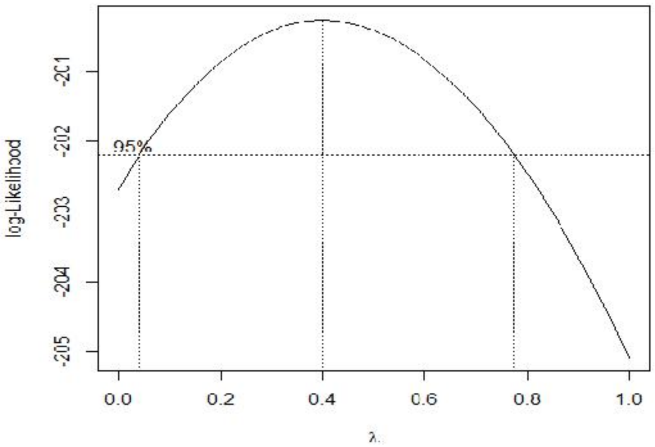
The values of R^2 and R^2 (adj.) are and respectively and thus, are quite close. We can conclude that we have built a model which explains 53.4% of the change in the number of deaths due to cardiovascular diseases using the factors: Literacy (%), Health Expenditure per capita, % of obese population and % of urban population.

Use of Advanced Regression Techniques

- **Use Box Tidwell power Transformation**

In statistics, a power transform is a family of functions that are applied to create a monotonic transformation of data using power functions. This is a useful data transformation technique used to stabilize variance, make the data more normal distribution-like, improve the validity of measures of association such as the Pearson correlation between variables and for other data stabilization procedures. In our case We obtain an error(NA/NAN/Inf) in the transformation hence we go for a Box-Cox transform.

Use Box -Cox power Transformation



Hence we take value of Lambda as 0.4

The Model Obtained from this Method is:

$y=b_0+b_1x_1+b_2X_2$

Where,

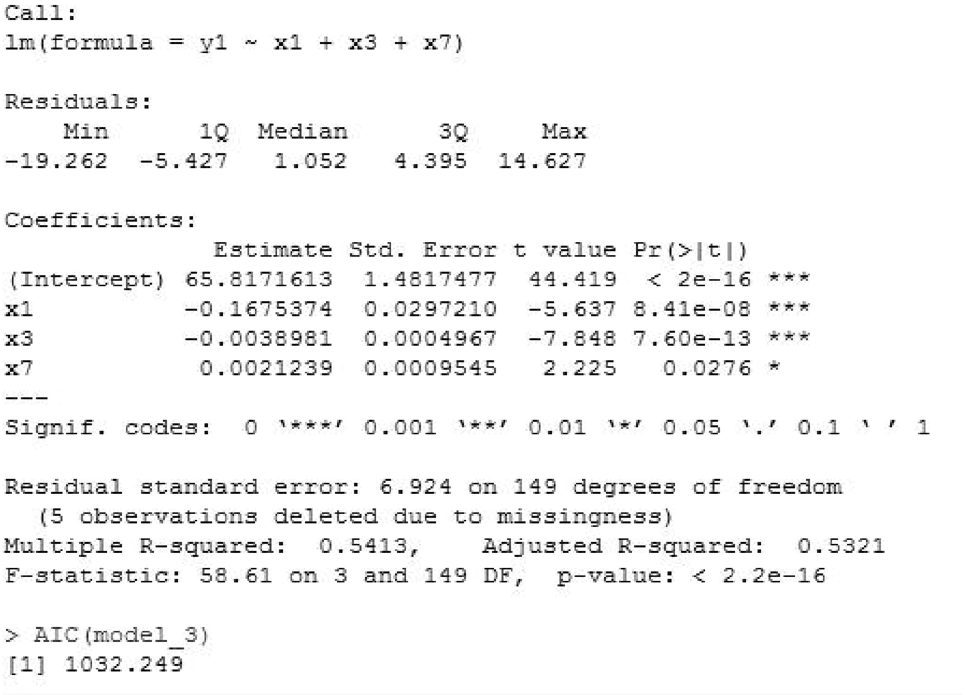
Y = ((No. of Deaths)$^{0.4}$ -1)/(0.4)

X_1 = (%Urban Population)

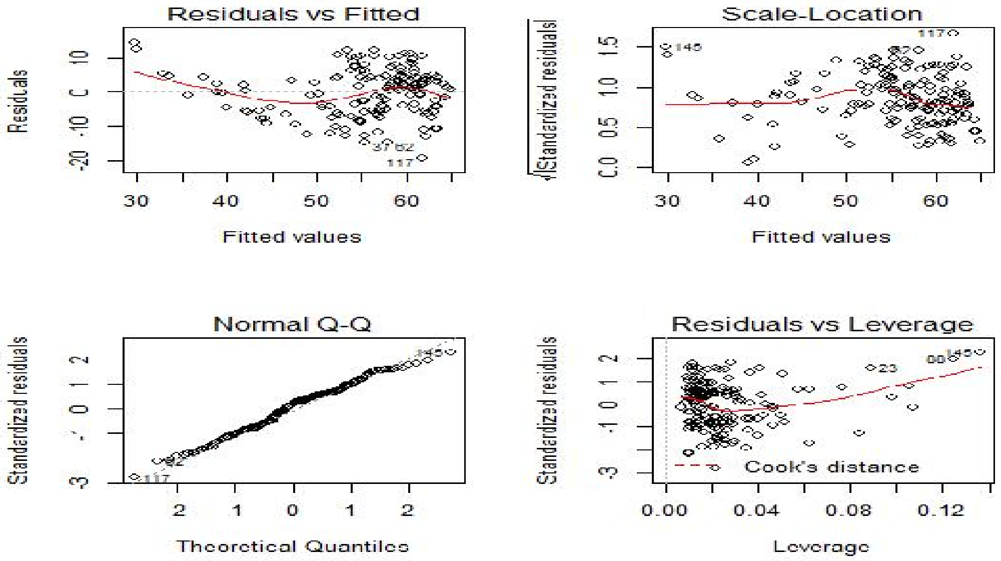
X_2 = (Per Capita Govt. Health expenditure)

X_3 = (No. of cigarettes smoked per capita annually)

The model performance obtained is as follows:



```
Call:
lm(formula = y1 ~ x1 + x3 + x7)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.262  -5.427   1.052   4.395  14.627 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 65.8171613  1.4817477  44.419  < 2e-16 ***
x1          -0.1675374  0.0297210  -5.637 8.41e-08 ***
x3          -0.0038981  0.0004967  -7.848 7.60e-13 ***
x7           0.0021239  0.0009545   2.225   0.0276 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.924 on 149 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.5413,    Adjusted R-squared:  0.5321 
F-statistic: 58.61 on 3 and 149 DF,  p-value: < 2.2e-16

> AIC(model_3)
[1] 1032.249
```

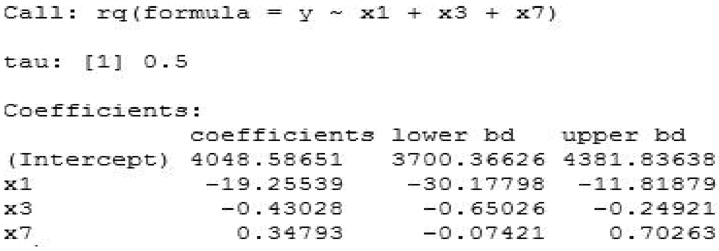


- **Robust Regression**

Robust regression is a form of regression analysis designed to circumvent some limitations of traditional parametric and non-parametric methods. Regression analysis seeks to find the relationship between one or more independent variables and a dependent variable. Certain widely used methods of regression, such as ordinary least squares, have favourable properties if their underlying assumptions are true, but can give misleading results if those assumptions are not true; thus ordinary least squares is said to be not robust to violations of its assumptions. Robust regression methods are designed to be not overly affected by violations of assumptions by the underlying data-generating process.

- **LAD:**

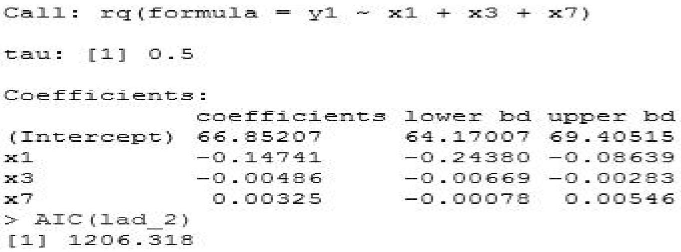
Least absolute deviations (LAD), also known as least absolute errors (LAE), least absolute value (LAV), least absolute residual (LAR), sum of absolute deviations, or the L1 norm condition, is a statistical optimality criterion and the statistical optimization technique that relies on it. Similar to the popular least squares technique, it attempts to find a function which closely approximates a set of data. In the simple case of a set of (x,y) data, the approximation function is a simple "trend line" in two-dimensional Cartesian coordinates. The method minimizes the sum of absolute errors (SAE) (the sum of the absolute values of the vertical "residuals" between points generated by the function and corresponding points in the data). The least absolute deviations estimate also arises as the maximum likelihood estimate if the errors have a Laplace distribution.



```
Call: rq(formula = y ~ x1 + x3 + x7)

tau: [1] 0.5

Coefficients:
            coefficients lower bd   upper bd
(Intercept) 4048.58651   3700.36626 4381.83638
x1           -19.25539    -30.17798  -11.81879
x3            -0.43028     -0.65026   -0.24921
x7             0.34793     -0.07421    0.70263
```

AIC in this case comes out to be **2809.**

- **LAD(with BoxCox Transformation)**

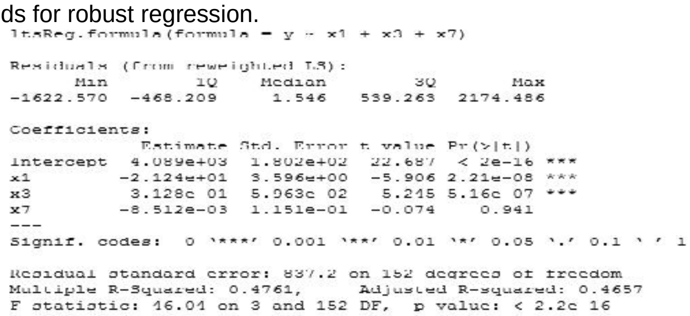


```
Call: rq(formula = y1 ~ x1 + x3 + x7)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 66.85207     64.17007 69.40515
x1          -0.14741     -0.24380 -0.08639
x3          -0.00486     -0.00669 -0.00283
x7           0.00325     -0.00078  0.00546
> AIC(lad_2)
[1] 1206.318
```

AIC in this case is **1206.318**

- **LTS**

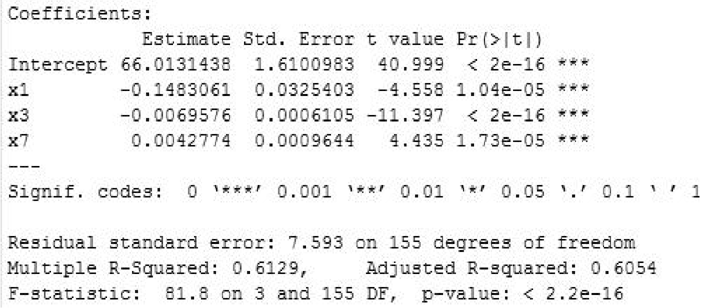
Least trimmed squares (LTS), or least trimmed sum of squares, is a robust statistical method that fits a function to a set of data whilst not being unduly affected by the presence of outliers. It is one of a number of methods for robust regression.



```
ltsReg.formula(formula = y ~ x1 + x3 + x7)

Residuals (from reweighted LS):
      Min        1Q    Median        3Q       Max
-1622.570  -468.209     1.546   539.263  2174.486

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept  4.089e+03  1.802e+02  22.687  < 2e-16 ***
x1        -2.124e+01  3.596e+00  -5.906 2.21e-08 ***
x3         3.128e-01  5.963e-02   5.245 5.16e-07 ***
x7        -8.512e-03  1.151e-01  -0.074    0.941
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 837.2 on 152 degrees of freedom
Multiple R-Squared: 0.4761,     Adjusted R-squared: 0.4657
F-statistic: 16.04 on 3 and 152 DF,  p-value: < 2.2e-16
```

Adjusted R^2 is 46.57%

- **LTS with BoxCox**

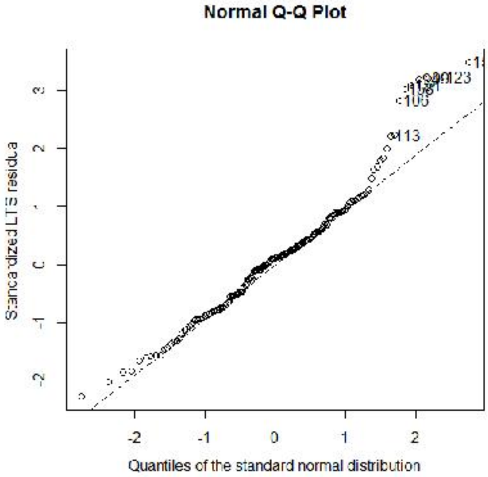


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept 66.0131438  1.6100983  40.999  < 2e-16 ***
x1        -0.1483061  0.0325403  -4.558 1.04e-05 ***
x3        -0.0069576  0.0006105 -11.397  < 2e-16 ***
x7         0.0042774  0.0009644   4.435 1.73e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.593 on 155 degrees of freedom
Multiple R-Squared: 0.6129,     Adjusted R-squared: 0.6054
F-statistic:  81.8 on 3 and 155 DF,  p-value: < 2.2e-16
```

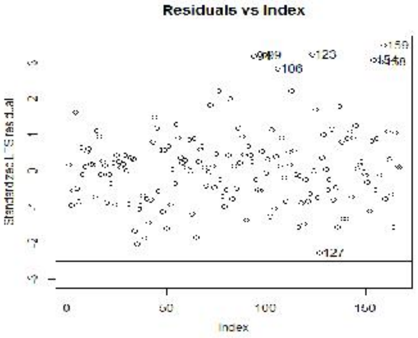
Since the adjusted R^2 is the maximum (60.54%) we will use this model.

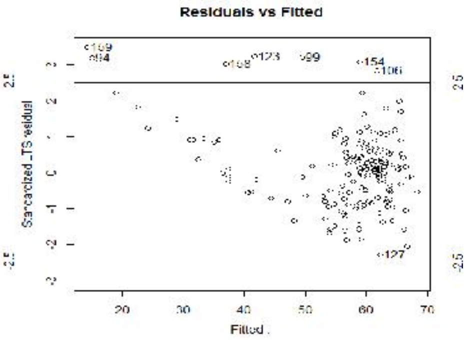
No outliers are removed as this is a robust regression.

The various plots are as follows:



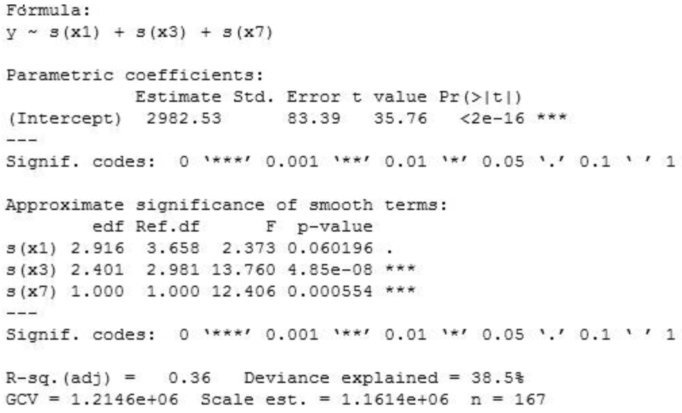
We see that majority of the points follow normality assumption. Those which don't have very little impct due to **Robust regression. Hence normality condition is satisfied.**



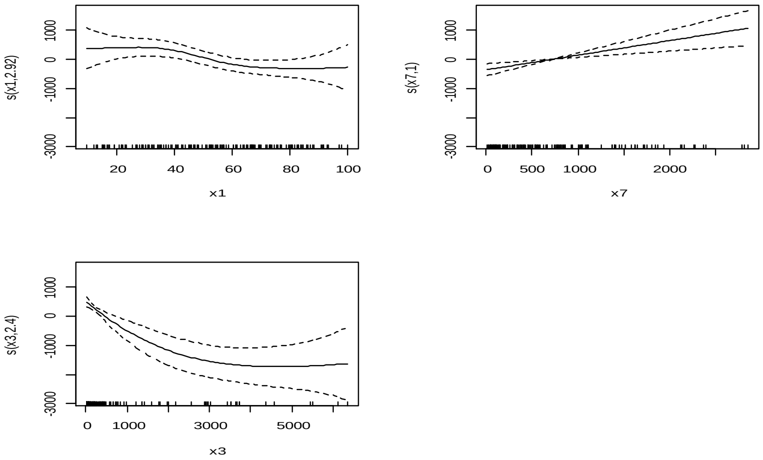


Generalized Additive Model (GAM) Implementation

In statistics, a generalized additive model (GAM) is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference about these smooth functions.

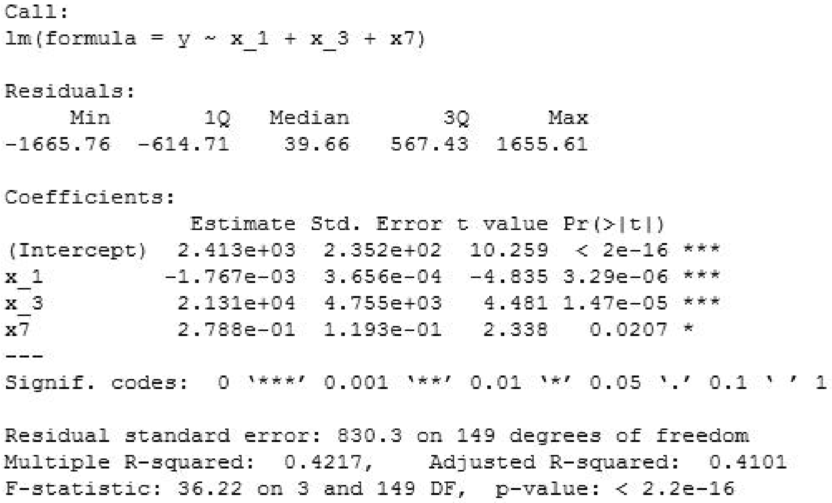


```
Fórmula:
y ~ s(x1) + s(x3) + s(x7)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2982.53      83.39   35.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
        edf Ref.df      F  p-value
s(x1) 2.916  3.658  2.373 0.060196 .
s(x3) 2.401  2.981 13.760 4.85e-08 ***
s(x7) 1.000  1.000 12.406 0.000554 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.36   Deviance explained = 38.5%
GCV = 1.2146e+06  Scale est. = 1.1614e+06  n = 167
```



The Model that we Obtained is:

Death = B0+B1(% Urban Population)3+ B2*((Per Capita Government_Health_expenditure)-.5)+B3*(No. of cigarettes smoked per capita annually). The results are:



```
Call:
lm(formula = y ~ x_1 + x_3 + x7)

Residuals:
     Min       1Q   Median       3Q      Max
-1665.76  -614.71    39.66   567.43  1655.61

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.413e+03  2.352e+02  10.259  < 2e-16 ***
x_1         -1.767e-03  3.656e-04  -4.835 3.29e-06 ***
x_3          2.131e+04  4.755e+03   4.481 1.47e-05 ***
x7           2.788e-01  1.193e-01   2.338   0.0207 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 830.3 on 149 degrees of freedom
Multiple R-squared:  0.4217,    Adjusted R-squared:  0.4101
F-statistic: 36.22 on 3 and 149 DF,  p-value: < 2.2e-16
```

Summary of all the Models used

	AIC	**R-Sq**
Normal	2380.726	50.05
Box Cox	1032.249	53.21
LAD (w/o BoxCox)	2809.21	
LAD (BoxCox)	1206.318	
LTS		47.6
LTS(Box Cox)		60.54
GAM	2497.07	41.01

Hence we will go with an LTS robust regression model with BoxCox transform

Conclusion

We have managed to build different models linear regression model which explain the variation in deaths due to cardiovascular diseases across countries. We have found that per capitagovernment health expenditure, % of urban population and number of cigarettes smoked per capita per year are the relevant factors in our model. The regression equation is as follows:

- Using Adjusted R-sq as the benchmark among linear models, LTS robust regression model with BoxCox transform gives the best fit
- Value of Adjusted R-sq even for LTS Box Cox is 60.54%, indicating model is cannot be used for giving significantly accurate prediction of cardiovascular deaths (per 100000 population). The final regression equation is:

((No. of CV Deaths/100,000 Population)0.4 -1)/(0.4)= 66.01-.1*(%Urban Population)-.007*(Per Capita Health expenditure by the Government)+.004*(No. of Cigarettes Smoked per capita per year)

It is apparent that the number of deaths increases with smoking and decrease with greater per capita government health expenditure and urbanization.Interestingly, GDP, alcohol consumption and obesity do not come out as influential in the linear models, indicating that obesity percentage does not seem to be a major determinant of cardiovascular deaths as per the model (on an average).

Value of Adjusted R-sq could have been higher if parameters such as age structure, genetics, lifestyle, pollution etc. could have been included; factors like lifestyle and genetics are not available at country level.

References

~ Health Expenditure by countries (US$ GDP): http://data.worldbank.org/indicator/SH.XPD.PCAP?page=1

~ Urban Population (as % of Population) by countries : http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?page=1

~ Literacy rate (%):

~ http://data.un.org/Data.aspx?d=SOWC&f=inID%3A74

~ Obesity (% of Obese Population :

~ https://www.cia.gov/library/publications/the-world-factbook/rankorder/2228rank.html

~ Alcohol Consumption:

~ http://apps.who.int/gho/data/node.main.A1030?lang=en

~ GDP per Capita: http://data.worldbank.org/indicator/NY.GDP.PCAP.CD?page=1

❑❑❑