SOUTH DAKOTA
STATE UNIVERSITY

Regression Analyisis: Final project

Group members: Cole Ryan, Ebise Abdi, Theophilus Anim Bediako

May 5, 2022

Contents

List of Figures

Introduction

The number of years a person can expect to live is known as life expectancy. Life expectancy is an important indicator for gauging population health. It is based on an estimate of the average age individuals of a particular population will be when they die (Ortiz-Ospina, 2013). In practice, estimating life expectancy entails predicting the likelihood of surviving successive years of life. It shows the average death age of a population.

This project uses multiple linear regression to model the relationship between life expectancy and the 27 potential independent variables. Life expectancy at birth is the response or dependent variable. The explanatory or independent variables include life expectancy at 60, adult mortality, infant mortality, age 1-4 mortality, alcohol, BMI, age 5-19 thinness, age 5 – 19 obesity, hepatitis, measles, polio, diphtheria, basic water, doctors, hospitals, GNI capita, GGHE D, CHE GDP, population, HIV/AIDS, poverty, education expenditure, adult literacy, mean years of schooling. Other variables include country, country code, region, and year.

The original data is owned by The World Health Organization (WHO). The World Health Organization (WHO) through the Global Health Observatory (GHO) data repository keeps track of all countries' health status as well as many other related parameters. The datasets are made available to be used by researchers for health data analysis. The main data format is by year and country. All other variables are dependent on these two variables.

Records were obtained from 183 countries from 2000 to 2016 (17 years). The dataset consists of 3111 rows and 32 columns. The first three columns are categorical variables (country, country code, and region). The 5th column is the dependent variable, life expectancy. The rest of the columns are the independent variables. The researchers found 15246 missing values.

The researchers generated a random sample of size 51 out of 183 countries. The samples were drawn based on the number of countries in each region.

Variable Descriptions

Life expectancy at birth: It represents the overall mortality level of a population.

Adult Mortality: It represents the probability that a person between age 15 and age 60 dies per 1000 population.

Infant mortality: It represents the number of deaths of young children under 1 year per 1000 population.

Age 1_4 mort: It represents the death rate between ages 1 and 4.

Alcohol: It represents the amount of alcohol consumed per adult of 15years and above.

BMI: It represents the average body mass index of a given population.

Thinness 5-19 years: It represents the prevalence of thinness among children from 5 to 19 years.

Age 5-19 obesity: It represents the prevalence of obesity among children from 5 to 19 years.

Hepatitis: It represents vaccine coverage for a one-year-old. It is expressed as a percentage.

Measles: It represents the reported number of cases per 1000 population.

Polio: it represents the polio immunization coverage among 1-year-olds.

Diphtheria: It represents infant immunization coverage.

Basic Water: Percentage of population using at least basic drinking-water services.

GGHE-d: It represents the domestic general government health expenditure as a percentage of gross domestic product (GDP).

CHE GDP: It represents current health expenditure as a percentage of GDP.

UNE infant: Infant mortality rate (per 1000 live births).

UNE GNI: GNI (gross national income) per capita, measured in dollars.

Research Problem

What are the factors affecting life expectancy? Our project focuses on whether the 16 variables we have selected affect life expectancy. If not, what are the significant factors contributing to life expectancy. Several can be the outcomes of the research; for example, if alcohol consumption is negatively related to life expectancy, then a country that wants to improve life expectancy would have to implement policies to reduce alcohol consumption.

Statistical Theories

Ordinary Least Squares Regression(OLS): attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Generally, multiple linear regression is represented as

$$y_i = \beta 0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta p x_{ip} + \epsilon_i$$

$ = 1,2,. . . ,n $ where y_i is the dependent or predicted variable β_0 : intercept $\beta_k, k = 1, \ldots, p-1$: the change in the mean response with a unit increase in a variable X_k, when all other predictors are held constant. The OLS regression is used under the following assumptions: linear relationship,normality,, no autocorrelation, homoscedastic (constant variance in residuals), more observations (n) than features(p) and no or little multicollinearity.

Typically, when the assumptions underpinning OLS regression are satisfied, the model coefficients are unbiased and have the smallest variance among all possible linear estimators. Due to the large volume of datasets in today's world, our OLS assumptions are often violated. It is common for OLS models to overfit the training sample in such situations. Overfitting implies high variance. When these concerns arise, regularized regression (also known as penalized models or shrinkage methods) can be used to manage parameter estimates as an alternative to OLS regression. Regularized regression puts constraints on the magnitude of the coefficients and will progressively shrink them towards zero, reduce the variance and decrease our sample error (Boehmke, 2018). The common regularization methods are ridge regression, the least absolute shrinkage and selection operator(lasso) regression and the elastic net. (Zou & Hastie, 2005). The objective function of regularized regression methods is very similar to OLS regression; however, we add a penalty parameter (P).

Regularized method	Objective function
Ridge regression	$\text{minimize}\{SSE + \lambda \sum_{j=1}^{p} \beta_j^2\}$
Lasso regression	$\text{minimize}\{SSE + \lambda \sum_{j=1}^{p} \lvert\beta_j\rvert\}$
Elastic nets	$\text{minimize}\{SSE + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} \lvert\beta_j\rvert\}$

The ridge regression approach pushes variables to approximately but not equal to zero, the lasso method will actually push coefficients to zero. Elastic nets combine ridge and lasso procedures.

Data Exploration

The data set contains 3111 observations of 31 independent variables. We've 15.8% of missing values of the data set. At the next step, we dropped the predictor variables with more than 5 missing values. For the predictor variables with less missing values, we replaced the missing value with the mean value based on the continent. The sample data is chosen randomly, about 1/3 of the countries from each region. The sample data contains 51 of observations of 17 variables.

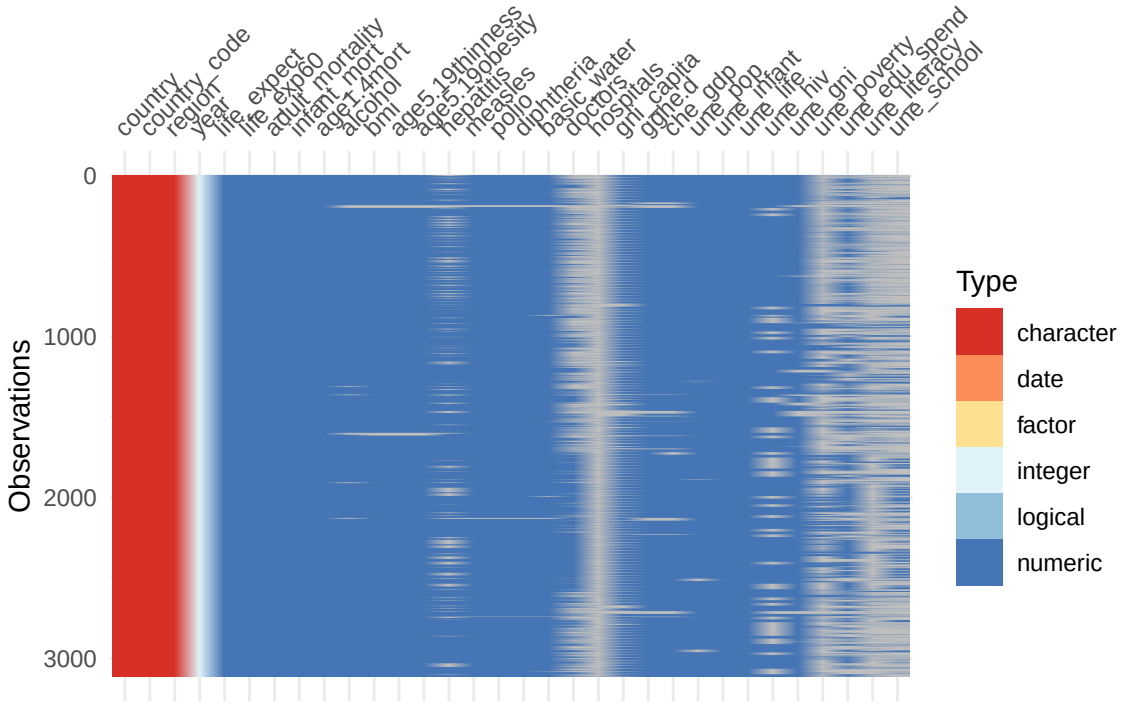


Figure 1: Data visualization

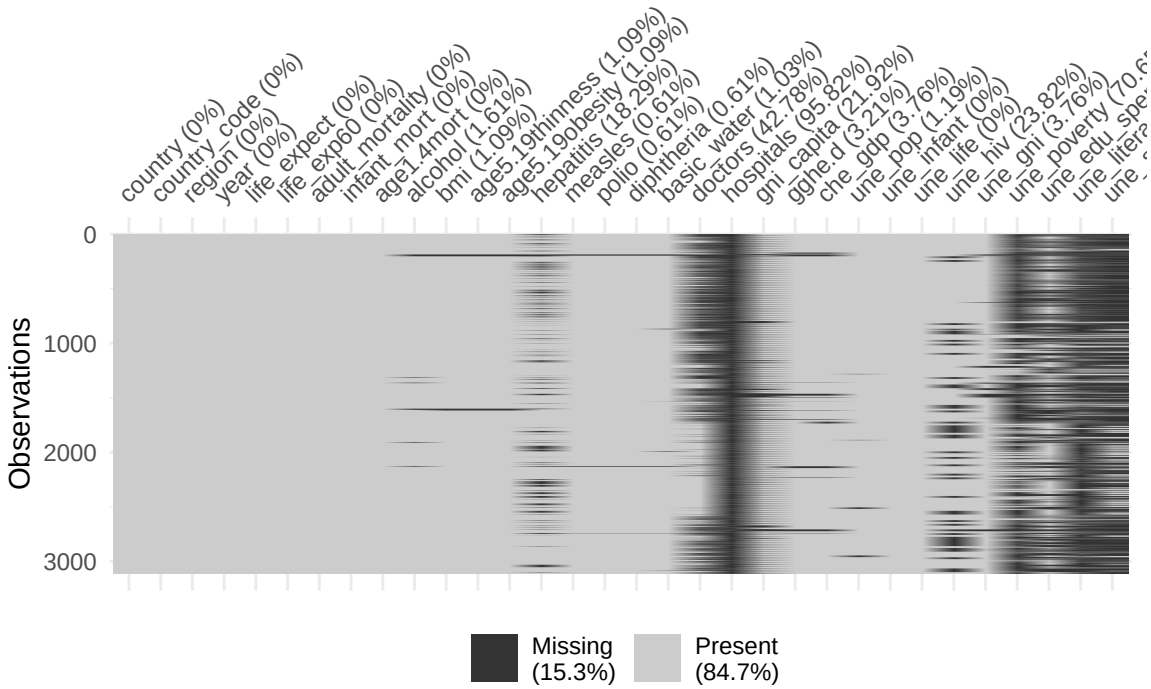
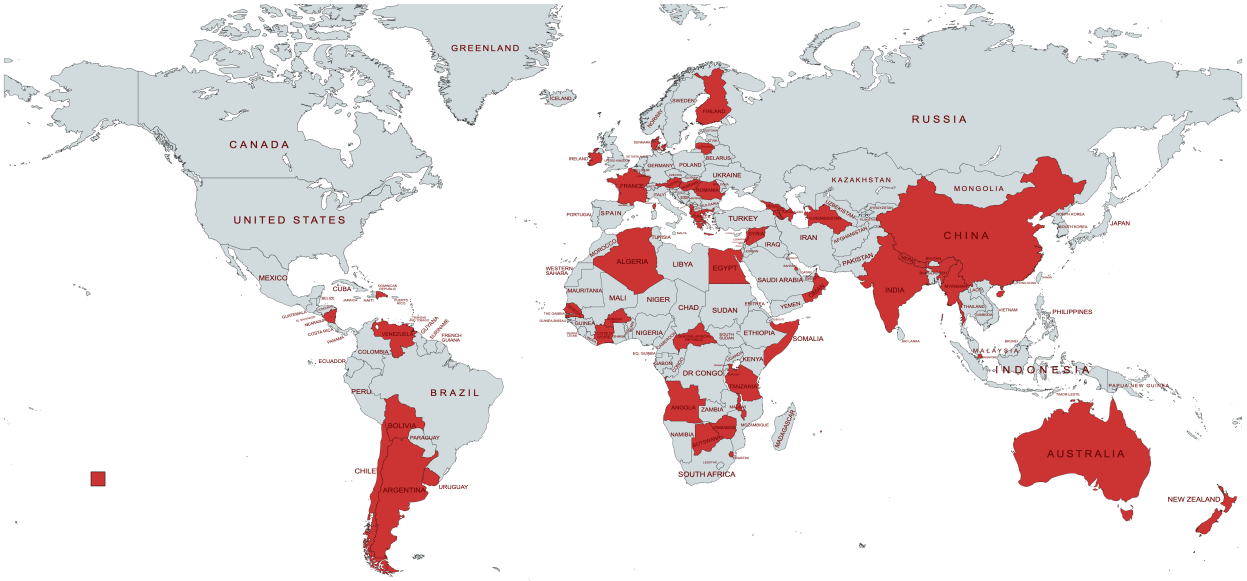


Figure 2: Missing values and present values


GREENLAND
CANADA
UNITED STATES
MEXICO
CUBA
VENEZUELA
COLOMBIA
ECUADOR
PERU
BRAZIL
BOLIVIA
PARAGUAY
CHILE
ARGENTINA
URUGUAY
FRENCH GUIANA
RUSSIA
KAZAKHSTAN
MONGOLIA
CHINA
JAPAN
NORTH KOREA
SOUTH KOREA
IRAN
IRAQ
TURKEY
SYRIA
AFGHANISTAN
PAKISTAN
INDIA
SAUDI ARABIA
YEMEN
OMAN
UNITED ARAB EMIRATES
QATAR
KUWAIT
BAHRAIN
JORDAN
ISRAEL
LEBANON
EGYPT
LIBYA
ALGERIA
TUNISIA
MOROCCO
WESTERN SAHARA
MAURITANIA
MALI
NIGER
CHAD
SUDAN
NIGERIA
GUINEA
ETHIOPIA
SOMALIA
ERITREA
KENYA
DR CONGO
UGANDA
RWANDA
BURUNDI
TANZANIA
ANGOLA
ZAMBIA
MALAWI
MOZAMBIQUE
ZIMBABWE
NAMIBIA
BOTSWANA
SOUTH AFRICA
LESOTHO
MADAGASCAR
PHILIPPINES
VIETNAM
THAILAND
MALAYSIA
INDONESIA
PAPUA NEW GUINEA
AUSTRALIA
NEW ZEALAND
SWEDEN
FINLAND
NORWAY
ICELAND
IRELAND
POLAND
BELARUS
UKRAINE
GERMANY
FRANCE
SPAIN
PORTUGAL

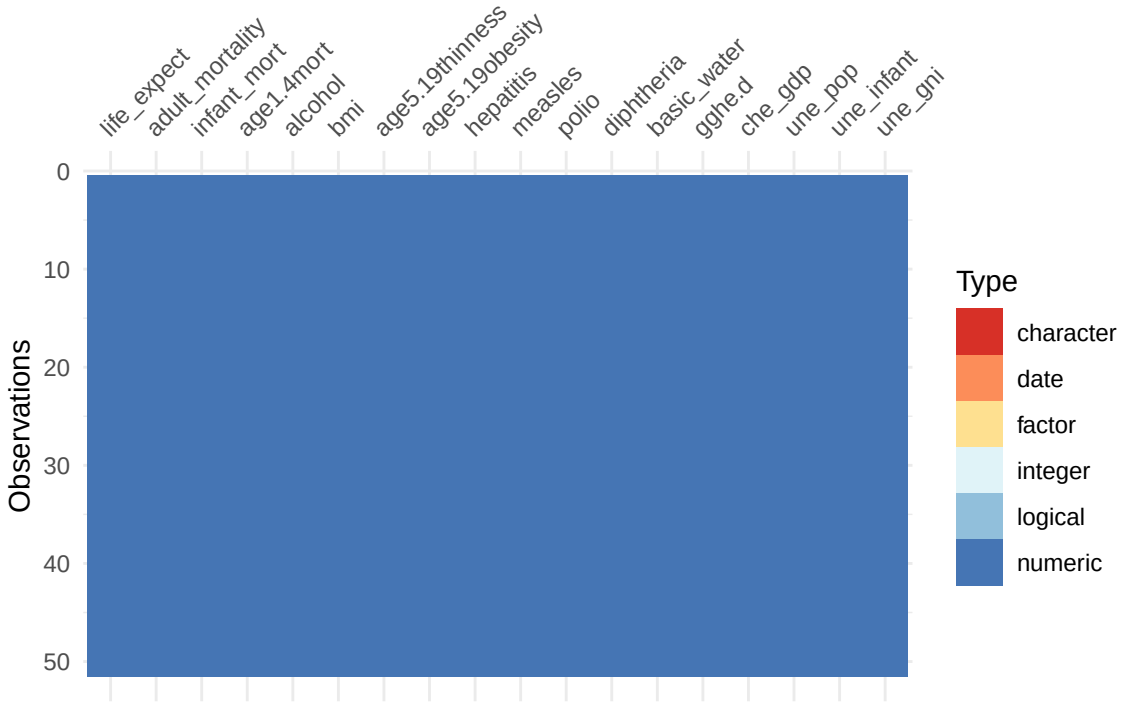


Figure 3: Data visualization afte NA values replacement

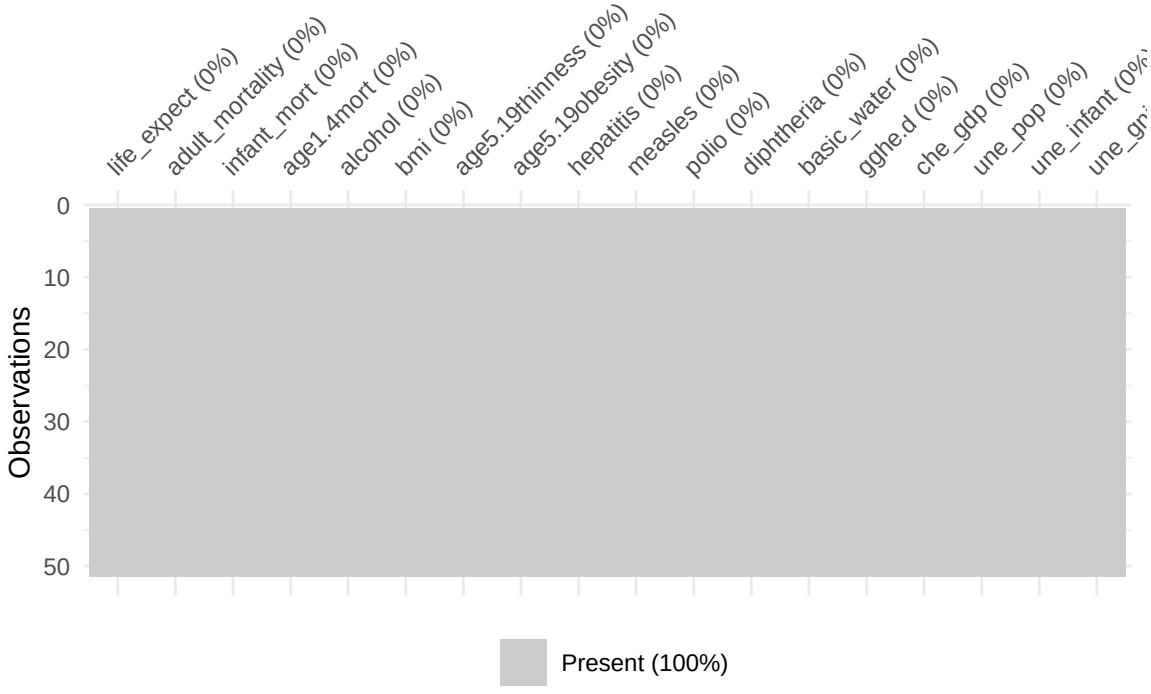


Figure 4: Data visualization after NA values replacement

The boxplot shows that some variables have outliers, we will have to pay attention later to see if the outliars are influential.

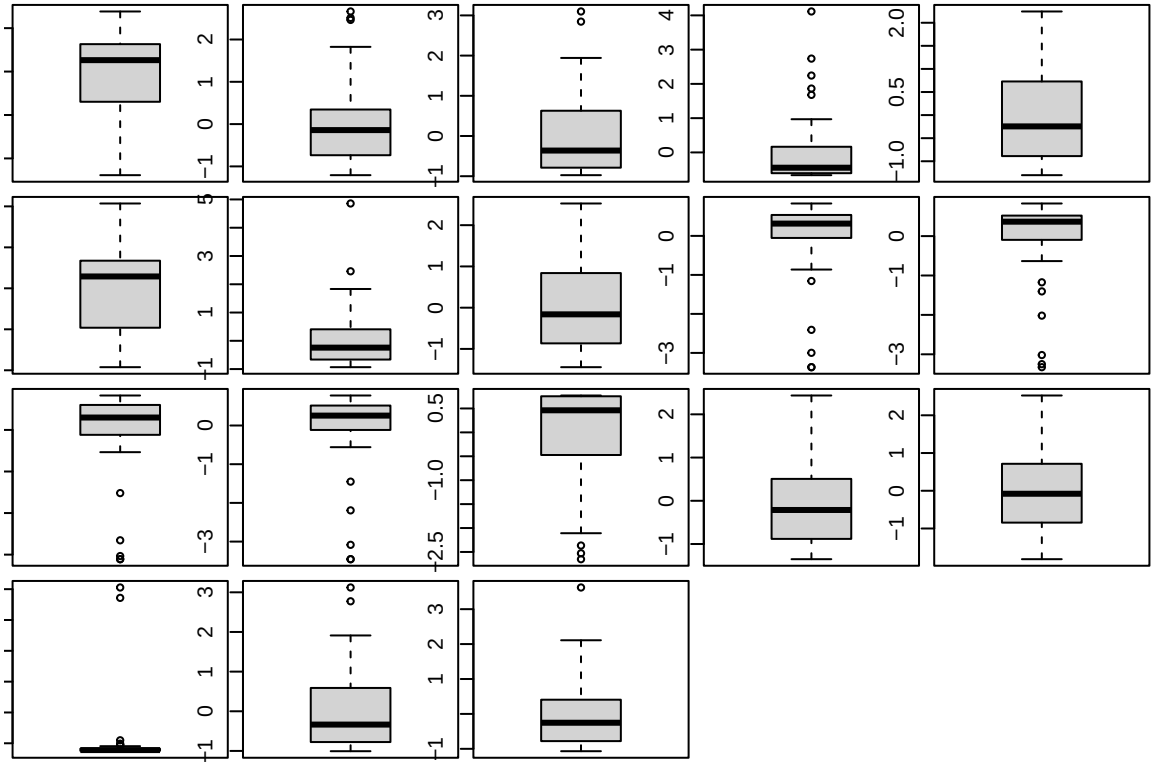


Figure 5: The boxplot of the variables

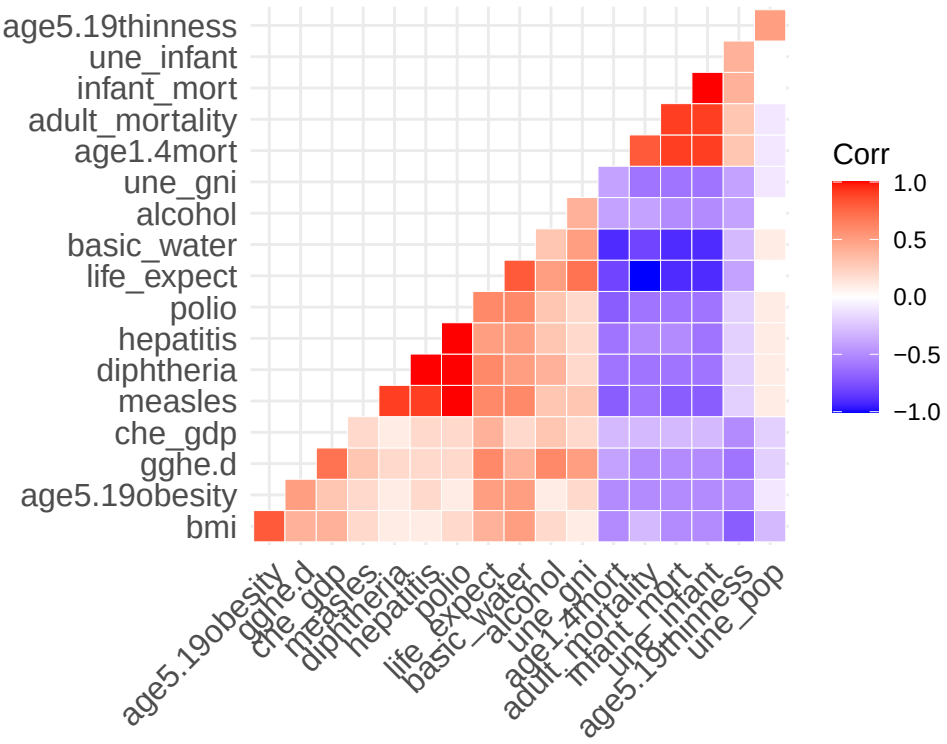


Figure 6: The correlation plot

The correlation plot indicates that some of our independent variables are highly correlated with each other. Moving forward we will have to address this in the models that we create.

Building Models

Our goal is to build a multiple linear regression model, in the general form:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \quad i = 1, ..., 51 \quad p - 1 = 16$$

where Y_i represent the response variable life expectancy,

$x_{i,p-1}$ are the predictor variables, β_0 is the intercept and the $\beta_i s$ are the coefficients of the predictor variables.

```
##     (Intercept) adult_mortality     infant_mort      age1.4mort         alcohol
##    8.553337e+01   -5.100266e-02    2.770517e+01   -5.435995e+02    1.234664e-01
##             bmi age5.19thinness  age5.19obesity       hepatitis         measles
##   -2.875720e-01   -1.096567e-01    4.093226e-02   -8.221617e-02    3.841618e-02
##           polio      diphtheria     basic_water          gghe.d         che_gdp
##   -1.317340e-01    1.675124e-01    1.969535e-02    1.921876e-01    7.232512e-02
##      une_infant         une_gni
##   -5.161504e-02    5.119731e-05
```

These would be our β_i estimates for our full model.

We have seen from the correlation plot that some of the variables may be correlated. We can use the VIF to remove variables until our model has a low amount of co-linearity. Below are our final VIF values and a model with reduced co-linearity.

```
## adult_mortality         alcohol             bmi age5.19thinness  age5.19obesity
##        5.092370        2.327883        6.071015        3.045392        4.564709
##       hepatitis     basic_water          gghe.d         che_gdp      une_infant
##        1.808439        5.424924        4.544564        2.976844        8.046981
##         une_gni
##        2.707457
```

We can build a reduced model from these variables. The coefficients in this model are more meaningful because we have greatly reduced multicolinearity.

```
##     (Intercept) adult_mortality         alcohol             bmi age5.19thinness
##    8.223102e+01   -4.961617e-02    1.149366e-01   -3.210618e-01   -9.527984e-02
##  age5.19obesity       hepatitis     basic_water          gghe.d         che_gdp
##    6.681421e-02   -1.394584e-03    6.389430e-02    1.733538e-01    6.121673e-02
##      une_infant         une_gni
##   -7.885933e-02    3.356904e-05
```

Using this new model with a reduced number of variables, we can use the step function with the AIC criteria to find a new 'best' model:

```
##     (Intercept) adult_mortality      une_infant         une_gni          gghe.d
##    7.568976e+01   -5.159683e-02   -7.786532e-02    3.967758e-05    2.603400e-01
##     basic_water         alcohol
##    4.823635e-02    1.113588e-01
```

Some of the data appears to not be linearly related to life expectancy. We can try a transformation to see if we can get a better model.

From the box-cox method, we can try the transformation $Y^{*1} = Y^{1/2}$ to get model

```
##     (Intercept) adult_mortality      une_infant         une_gni          gghe.d
##    8.725455e+00   -3.133905e-03   -5.119934e-03    1.729819e-06    1.381294e-02
##     basic_water         alcohol
##    2.968970e-03    6.144604e-03
```

We can also try the transformation $Y^{*2} = Y^{-1/2}$ to get model

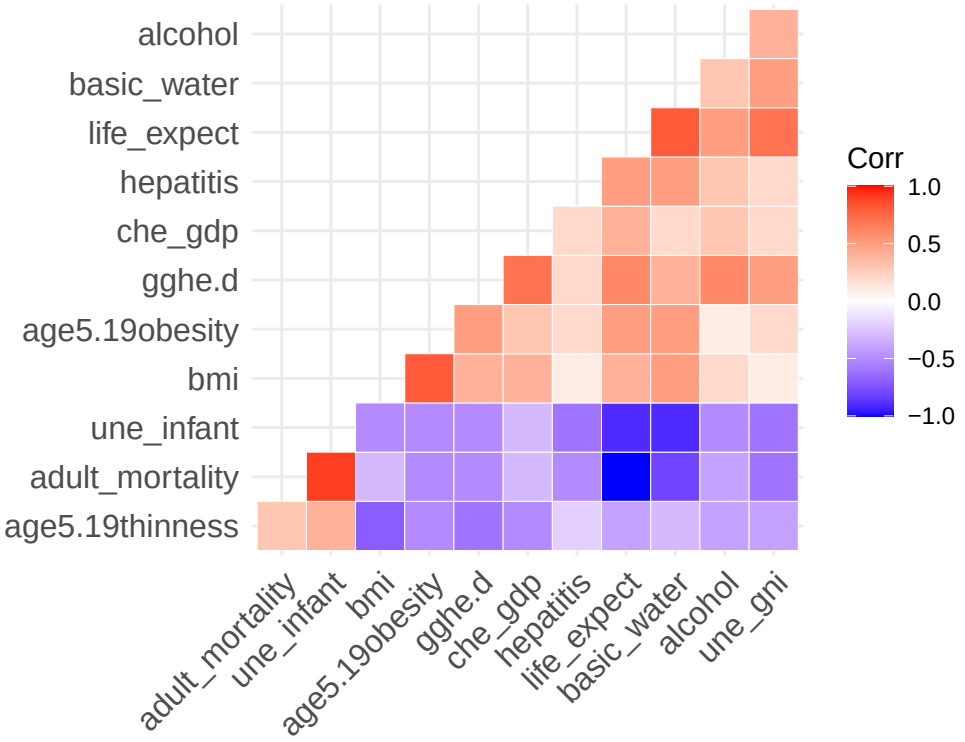


Figure 7: Correlation plot after using AIC criterion

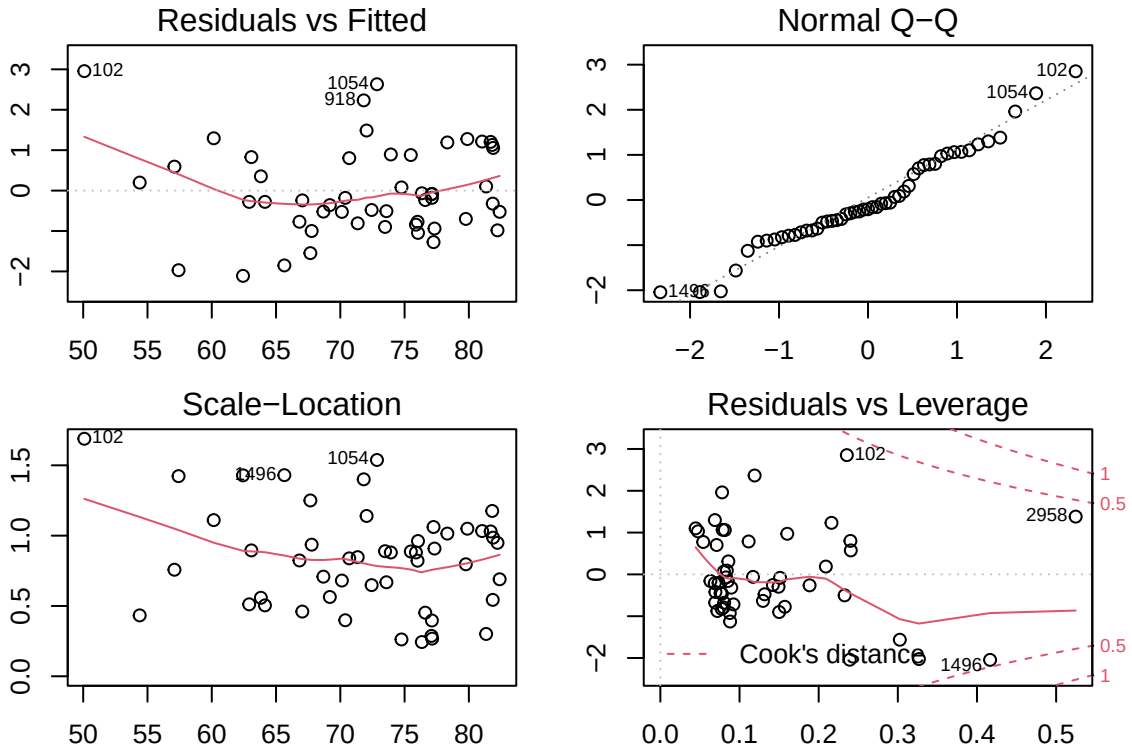


Figure 8: Residual Analysis plot

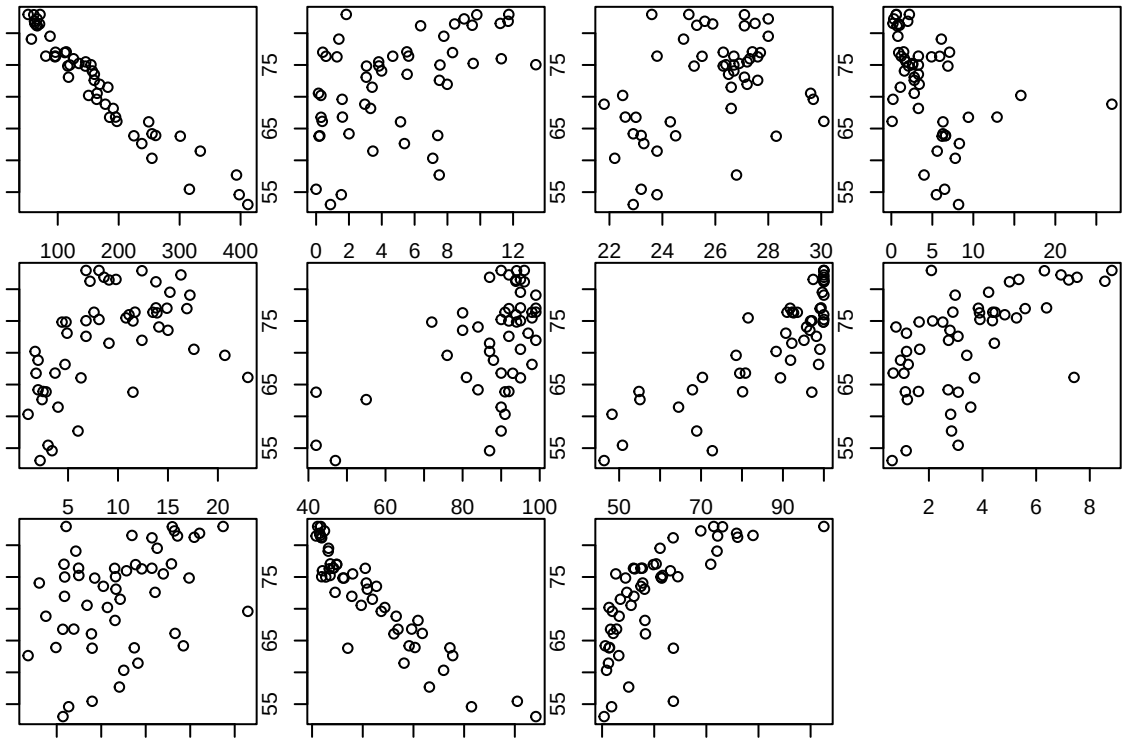


Figure 9: Regression function of the full model

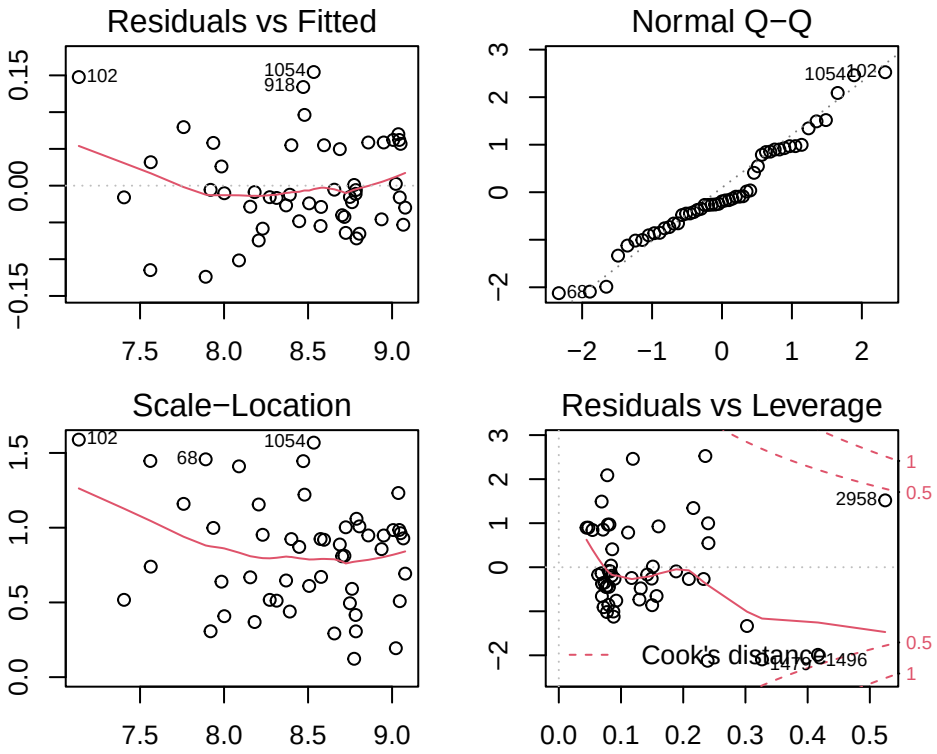


Figure 10: Residual Analysis, Yˆ(1/2)

```
##     (Intercept) adult_mortality      une_infant          gghe.d     basic_water
##    1.137187e-01    4.697523e-05    8.781982e-05   -1.602073e-04   -4.577766e-05
##         alcohol
##   -7.640322e-05
```

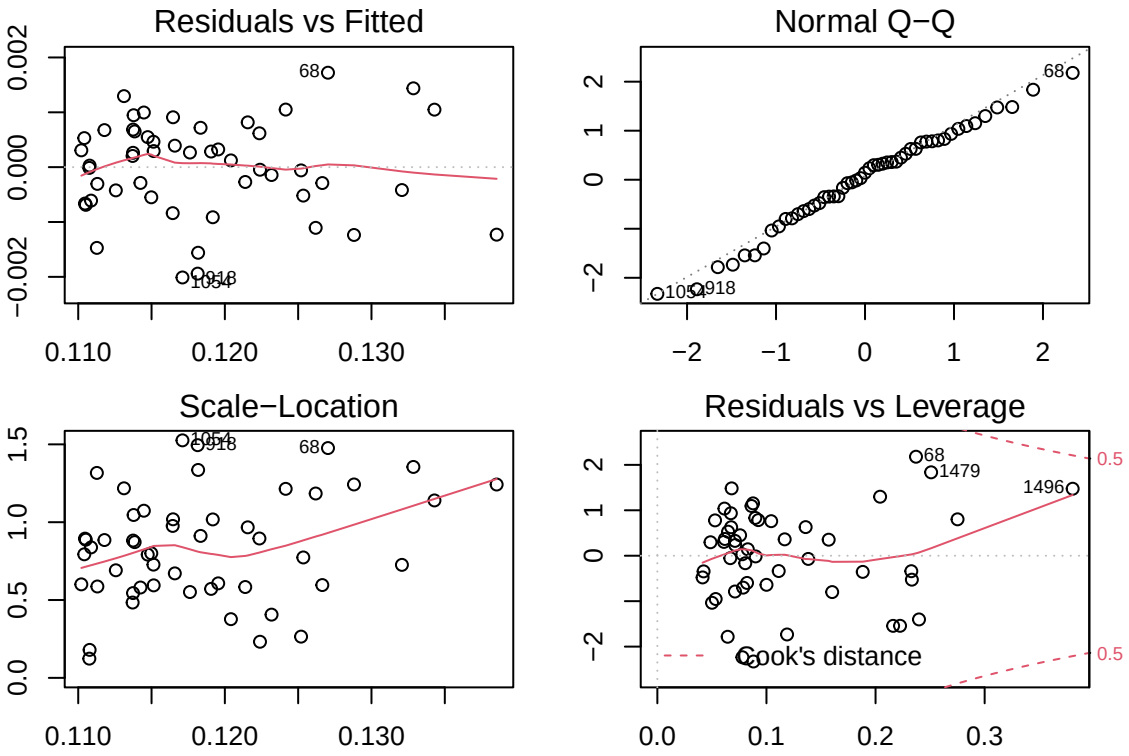


Figure 11: Residual Analysis, Yˆ(-1/2)

The third model, with transformation $Y^{*2} = Y^{-1/2}$, appears to best satisfy the assumptions for linear regression.

We can also check the r-squared values for the models.

```
##          Y Yˆ(1/2) Y(-1/2)
## r2  0.980   0.983   0.985
## r2a 0.978   0.980   0.983
```

For r-squared and adjusted r-squared, the $Y^{*2} = Y^{-1/2}$ model performs the best. We can also check adjusted r-squared. This model also meets the assumptions best as well, so we will consider it against other models we can make.

Other approaches can be used for regression as well. We can use ridge, lasso, and elastic net regression to build other models.

```
## [1] 0.0001623146
```

This gives a lambda value for our lasso regression of ~.000162.

We can use this to build the following model:

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)      1.084210e-01
## adult_mortality  4.151231e-05
```

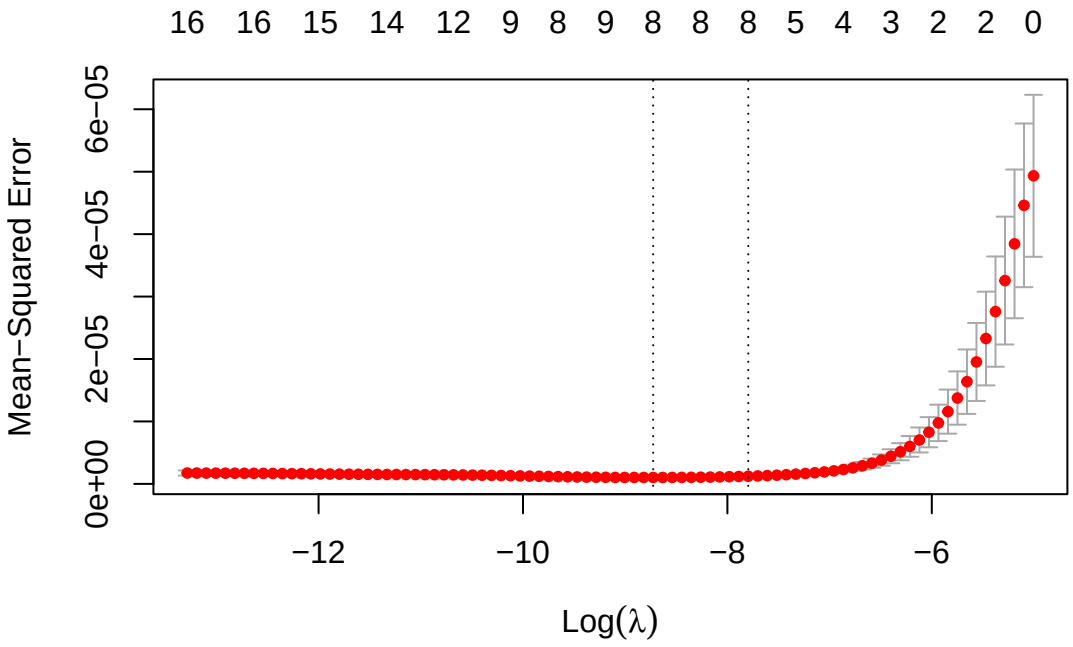


Figure 12: Lambda Values vs. MSE

```
## infant_mort      3.121427e-02
## age1.4mort       4.463024e-01
## alcohol         -7.923413e-05
## bmi              1.820175e-04
## age5.19thinness  4.947725e-05
## age5.19obesity  -3.462936e-05
## hepatitis        3.728275e-05
## measles         -9.258475e-06
## polio            6.787652e-06
## diphtheria      -3.666511e-05
## basic_water     -2.050711e-05
## gghe.d          -1.517454e-04
## che_gdp         -3.058817e-05
## une_infant       2.034600e-05
## une_gni         -1.892274e-08
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)      1.105951e-01
## adult_mortality  4.496616e-05
## infant_mort      5.234111e-02
## age1.4mort       3.968518e-01
## alcohol         -5.835341e-05
## bmi              .
## age5.19thinness  .
## age5.19obesity   .
## hepatitis        .
```

```
## measles          .
## polio            .
## diphtheria       .
## basic_water     -5.246513e-06
## gghe.d          -1.507866e-04
## che_gdp          .
## une_infant       .
## une_gni         -1.570917e-08
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## (Intercept)      1.098764e-01
## adult_mortality  4.207620e-05
## infant_mort      3.417330e-02
## age1.4mort       4.355047e-01
## alcohol         -6.680046e-05
## bmi              1.028813e-04
## age5.19thinness  3.898768e-05
## age5.19obesity  -9.644686e-06
## hepatitis        1.494542e-05
## measles          .
## polio            .
## diphtheria      -1.676160e-05
## basic_water     -1.913230e-05
## gghe.d          -1.821826e-04
## che_gdp         -4.186429e-07
## une_infant       1.844085e-05
## une_gni         -1.858511e-08
```

Now we can compare our full model to our model we build with reduced variables and the step function, and the model built with lasso.

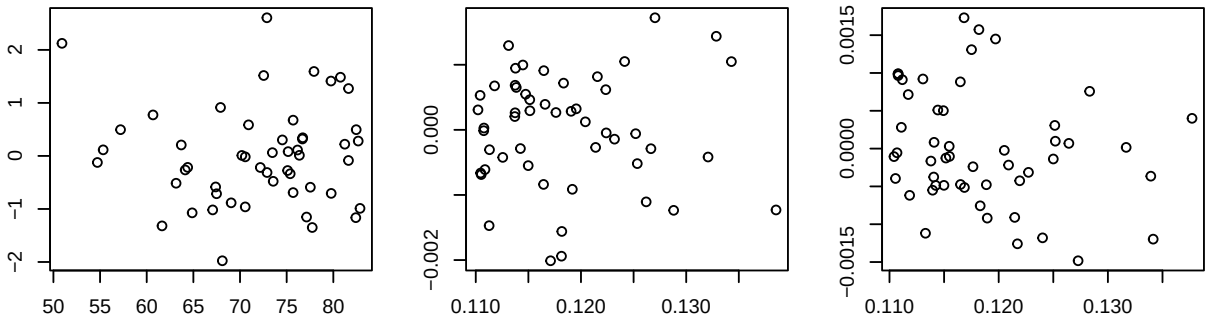


Figure 13: Residual Plots for Ridge, Lasso, & Elastic-Net

All of these models appear to be heteroscedastic, so we will compare all three of them.

R-squared for full model, the model built with AIC criteria on the transformation $Y = Y^{-1/2}$, and the ridge, lasso, and elastic net models built on the same transformation.

```
##     full transform  ridge lasso elastic
## 1 0.9862    0.9849 0.9898 0.988  0.9894
## 2 0.9797    0.9833 0.9850 0.986  0.9853
```

This shows that the ridge regression and elastic nets models are providing the best results based solely on

R-squared. When looking at adjusted R-squared what we can see is that the full model is unlikely to be our most suitable option.

We can also look at adjusted R-squared for the models.

Now that we have built some models, we can compare how they perform when making predictions on new data. We can use the 132 countries that were not part of our original sample.

```
##    full transform ridge lasso elastic
## 1 2.162     1.462 1.266 1.334   1.257
```

Conclusion

These are the MSE for the full model, the model build with AIC criteria, the ridge regression model, the lasso model, and the elastic-net (the last four all built using the $Y = Y^{-1/2}$ transformation) when tested on the non-sample data. We can see from this that the elastic-net has the lowest MSE, and is probably the best model that we have built for making new predictions.

Appendix

```
summary(raw2016)
```

```
##    country           country_code          region                year
##  Length:183         Length:183         Length:183         Min.   :2016
##  Class :character   Class :character   Class :character   1st Qu.:2016
##  Mode  :character   Mode  :character   Mode  :character   Median :2016
##                                                           Mean   :2016
##                                                           3rd Qu.:2016
##                                                           Max.   :2016
##
##   life_expect      life_exp60    adult_mortality   infant_mort
##  Min.   :52.94   Min.   :13.35   Min.   : 49.20   Min.   :0.001470
##  1st Qu.:66.07   1st Qu.:17.35   1st Qu.: 96.04   1st Qu.:0.006873
##  Median :73.40   Median :19.58   Median :147.17   Median :0.014905
##  Mean   :71.79   Mean   :19.76   Mean   :163.34   Mean   :0.024079
##  3rd Qu.:76.87   3rd Qu.:21.96   3rd Qu.:219.39   3rd Qu.:0.039000
##  Max.   :84.17   Max.   :26.39   Max.   :483.49   Max.   :0.095255
##
##    age1.4mort          alcohol            bmi        age5.19thinness
##  Min.   :0.000070   Min.   : 0.000   Min.   :20.60   Min.   : 0.100
##  1st Qu.:0.000275   1st Qu.: 1.305   1st Qu.:23.80   1st Qu.: 1.600
##  Median :0.000615   Median : 3.938   Median :26.20   Median : 3.500
##  Mean   :0.002090   Mean   : 4.787   Mean   :25.65   Mean   : 4.759
##  3rd Qu.:0.003115   3rd Qu.: 7.509   3rd Qu.:27.10   3rd Qu.: 6.600
##  Max.   :0.014615   Max.   :20.182   Max.   :32.20   Max.   :26.900
##                     NA's   :1        NA's   :2       NA's   :2
##  age5.19obesity     hepatitis        measles          polio
##  Min.   : 1.000   Min.   :26.00   Min.   :37.00   Min.   :44.00
##  1st Qu.: 3.500   1st Qu.:82.00   1st Qu.:82.50   1st Qu.:82.50
##  Median : 8.100   Median :93.00   Median :93.00   Median :93.00
##  Mean   : 8.255   Mean   :86.97   Mean   :86.82   Mean   :87.81
##  3rd Qu.:11.500   3rd Qu.:96.75   3rd Qu.:97.00   3rd Qu.:97.00
##  Max.   :26.700   Max.   :99.00   Max.   :99.00   Max.   :99.00
##  NA's   :2        NA's   :9
##    diphtheria     basic_water        doctors         hospitals     gni_capita
##  Min.   :19.00   Min.   : 38.85   Min.   : 0.128   Min.   : NA   Min.   : NA
##  1st Qu.:84.50   1st Qu.: 78.20   1st Qu.: 4.461   1st Qu.: NA   1st Qu.: NA
##  Median :93.00   Median : 95.15   Median :18.948   Median : NA   Median : NA
##  Mean   :88.01   Mean   : 86.64   Mean   :19.576   Mean   :NaN   Mean   :NaN
##  3rd Qu.:97.00   3rd Qu.: 99.25   3rd Qu.:31.697   3rd Qu.: NA   3rd Qu.: NA
##  Max.   :99.00   Max.   :100.00   Max.   :79.541   Max.   : NA   Max.   : NA
##                                   NA's   :73       NA's   :183   NA's   :183
##      gghe.d           che_gdp          une_pop            une_infant
##  Min.   : 0.3795   Min.   : 2.312   Min.   :     94.5   Min.   : 1.60
##  1st Qu.: 1.6730   1st Qu.: 4.450   1st Qu.:   2654.4   1st Qu.: 6.50
##  Median : 2.9941   Median : 6.379   Median :   9445.6   Median :14.80
##  Mean   : 3.4408   Mean   : 6.547   Mean   :  40726.3   Mean   :23.00
##  3rd Qu.: 4.4884   3rd Qu.: 8.147   3rd Qu.:  28481.9   3rd Qu.:36.55
##  Max.   :10.9497   Max.   :17.197   Max.   :1414049.4   Max.   :88.30
##  NA's   :7         NA's   :8        NA's   :2
##     une_life         une_hiv          une_gni         une_poverty
##  Min.   :51.59   Min.   : 0.100   Min.   :   740   Min.   : 0.100
##  1st Qu.:66.15   1st Qu.: 0.100   1st Qu.:  4098   1st Qu.: 0.900
```

```
##  Median :73.10   Median : 0.400   Median : 12150   Median : 2.200
##  Mean   :71.70   Mean   : 1.905   Mean   : 19047   Mean   : 9.159
##  3rd Qu.:77.07   3rd Qu.: 1.400   3rd Qu.: 26310   3rd Qu.: 6.050
##  Max.   :83.98   Max.   :28.000   Max.   :122670   Max.   :70.300
##                  NA's   :43       NA's   :7        NA's   :149
##  une_edu_spend    une_literacy     une_school
##  Min.   :1.497   Min.   :22.31   Min.   : 1.530
##  1st Qu.:3.743   1st Qu.:88.99   1st Qu.: 8.668
##  Median :4.693   Median :94.35   Median :10.961
##  Mean   :4.581   Mean   :89.61   Mean   :10.477
##  3rd Qu.:5.394   3rd Qu.:97.07   3rd Qu.:12.552
##  Max.   :7.976   Max.   :99.99   Max.   :14.207
##  NA's   :82      NA's   :150     NA's   :123
```

References

- Boehmke, B.(2018). UC Business Analytics R Programming Guide. University of Cincinnati.
- MMattson. (2020, October 6). Who national life expectancy. Kaggle. Retrieved April 7, 2022, from https://www.kaggle.com/datasets/mmattson/who-national-life-expectancy/metadata
- Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie (2013) - "Life Expectancy". Published online at OurWorldInData.org. Retrieved from: https://ourworldindata.org/life-expectancy
- Michael H. Kutner, Chris Nachtsheim, John Neter (2004). Applied Linear Regression Models, 4th edition. McGraw-Hill/Irwin. ISBN: 0072386916, 9780072386912.