# Term Project: Countries

Riley Adams

12/17/2020

## Introduction

How long can you expect to live? It is a well known fact that, on average, humans live longer than they used to. However, the average length of life will vary from country to country. What aspects of a country contribute to how long its citizens will live, on average?

This report is concerned with the average life expectancy in a given country, and what variables can be used to estimate it. We have built a regression model from an 80% subsample of the dataset called "Countries." Our sample, called "country80" contains data on 149 countries. Of these 149 countries, we begin with the following data:

LifeExpectancy – Average life expectancy in years

Country – list of countries in the data set

Code – three letter code for each country

LandArea - the land area of each country (in square kilometers)

Population - the population of each country (in millions)

Rural - the percentage of population living in rural areas

Health - percentage of government expenditures directed towards health care

Internet - percentage of population with internet access

BirthRate - amount of births per thousand people in a country

ElderlyPop - percentage of population age 65 or older

CO2 - carbon dioxide emissions in metric tons per capita

Cell - Cell phone subscriptions per 100 people

In this report, we will use statistical methods to decide which of these metrics in the data sample can best be used to predict the average life expectancy of a country. We will build a multiple linear regression model using only the most important predictors. The importance of predictors will be narrowed down using step-wise model selection procedures along with partial F-tests. In these tests we will observe the criteria of adjusted $R^2$ and Mallows $C_p$ to choose the strongest subsets of predictors. We will also rely on t-tests to assess significance of a predictor after all other predictors have been adjusted for. Furthermore, we will scrutinize the predictors to determine if multicollinearity is occurring among them and remove predictors as deemed necessary to avoid this.

At each new adjustment of the model, we will conduct a residual analysis. This will help us determine normality in the error for the predictors regressed on life expectancy, as well as linearity and constant variance in our residual values. Based on inference from the residual analysis, we will transform the model as needed until a sound model is achieved.

# Results

The most effective, parsimonious model which was achieved has 4 predictors: Rural, Health, Internet and Birth Rate. The model is as follows:

$$\hat{Y} = \sqrt{\beta_0 + \beta_{\text{rural}} X_1 + \beta_{\text{health}} X_2 + \beta_{\text{internet}} X_3 + \beta_{\text{births}} X_4}$$

where $\hat{Y}$ is the estimated mean life expectancy in years and the values of the regression coefficients are as follows:

$$\hat{Y} = \sqrt{6377.839 + (-8.883) X_1 + (24.001) X_2 + (9.325) X_3 + (-76.120) X_4}$$

The multiple regression model suggests that Life Expectancy varies linearly with the square root of the sum of Rural, Health, Internet and Birthrate; each multiplied by their respective coefficients (above).

Each coefficient above shows the amount by which the square of the mean life expectancy will change when its respective predictor variable $X_i$ is changed by 1 unit, while all other $X_j$, $i \neq j$ are held constant. In this case, the intercept $\beta_0 = 6377.839$ has no meaning.

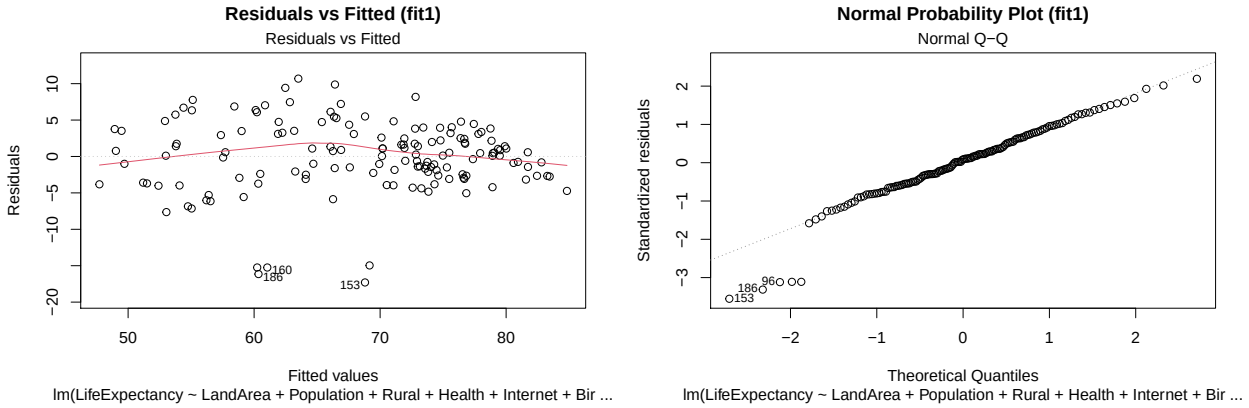
We can infer from this model that having a higher percentage of a country's population living in rural areas will negatively impact the life expectancy, as will the rate of births. We can see that a higher birth rate lowers life expectancy. On the other hand, the strongest predictors of longer lives are health care expenditures (as a percentage of government spending), and what proportion of the population has access to the internet.

The model has an adjusted R-squared value of 0.7891 and all predictors are significant with respect to the t-test with $\alpha = .05$, except Health, which has a p-value = .05736. Since the value was borderline, health was deemed significant per the model building process outlined below.

# Model Building

Initially, we started with scatterplots of all numerical predictors against Life Expectancy. From simple observation, the strongest positive linear relationships were with Health and Cell, neither being overtly obvious in its linearity. The strongest negative linear trends were found against Rural and against BirthRate.

We then fit a full model of Life Expectancy regressed on all numerical predictors, and plot the residual vs fitted as well as the normal probability plot.



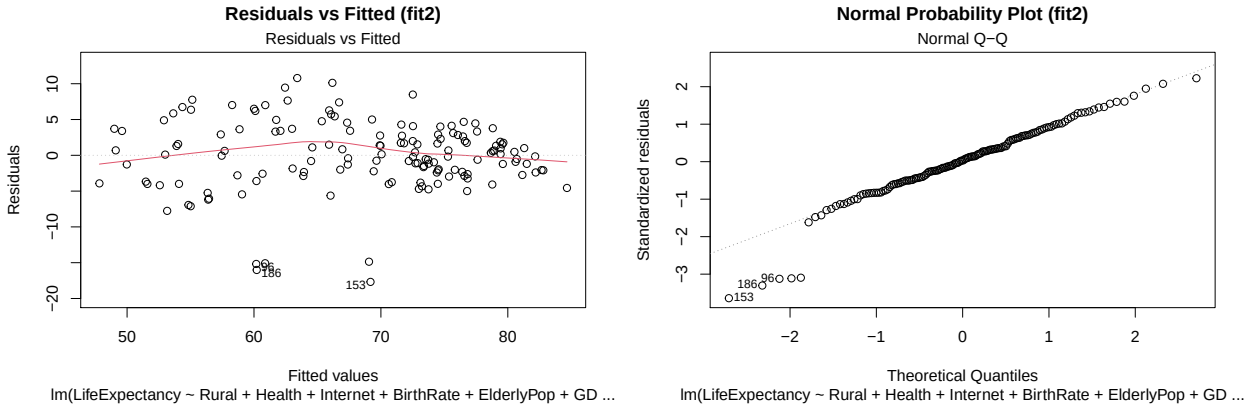

In the initial model (called fit1), our residual analysis tells us that there is non linearity due to our residuals being too high in the mid range, and too low on the high end. Furthermore, the variance in the residuals is nonconstant. The QQ plot indicates the the residuals are not quite normally distributed as the values on the low end are quite low, and they bend down at the top a little bit.

Next, we use the step-wise model selection procedure of backward elimination. This uses partial F-tests and, in this case, the $R^2_{adj}$ criterion to start with the full model and then one-by-one eliminate the least significant

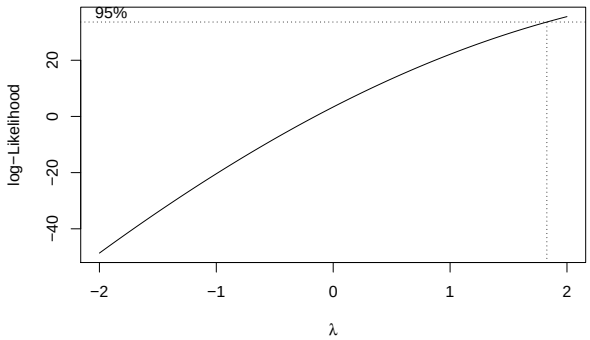
predictor and show us the remaining predictors in the would be model as well as the corresponding $R^2_{adj}$. We choose the model with the highest $R^2_{adj}$.

```
##   (Intercept) LandArea Population Rural Health Internet BirthRate ElderlyPop
## 1           1        0          0     0      0        0         1          0
## 2           1        0          0     0      0        1         1          0
## 3           1        0          0     1      0        1         1          0
## 4           1        0          0     1      0        1         1          1
## 5           1        0          0     1      1        1         1          1
## 6           1        0          0     1      1        1         1          1
## 7           1        0          0     1      1        1         1          1
## 8           1        0          0     1      1        1         1          1
##   CO2 GDP Cell adjusted r^2
## 1   0   0    0    0.7327963
## 2   0   0    0    0.7564876
## 3   0   0    0    0.7649131
## 4   0   0    0    0.7699997
## 5   0   0    0    0.7751014
## 6   0   1    0    0.7767840
## 7   0   1    1    0.7777896
## 8   1   1    1    0.7773005
```

Per backward elimination procedure, the suggested model is LifeExpectancy ~ Rural, Health, Internet, BirthRate, ElderlyPop, GDP, Cell. We fit this model and call it "fit2".



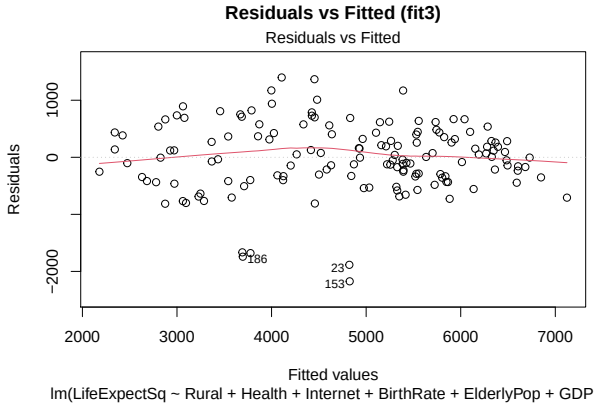

We notice that fit2 still has the same issues that fit1 did in its residuals. We have nonlinearity and nonconstant variance, making this model a good candidate for a boxcox transformation. We obtain a boxcox plot to determine the trasformation we will make to our response variable (Life Expectancy).

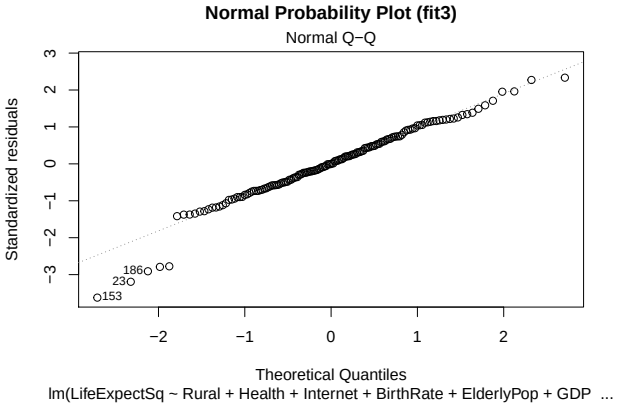



Since the boxcox plot has its maximum value near $\lambda = 2$, we raise $\hat{Y}$ to the 2nd power. Fitting a new model (called fit 3) as such:

$$\hat{Y} = \sqrt{\beta_0 + \beta_{\text{rural}}X_1 + \beta_{\text{health}}X_2 + \beta_{\text{internet}}X_3 + \beta_{\text{births}}X_4 + \beta_{\text{elder}}X_5 + \beta_{\text{gdp}}X_6 + \beta_{\text{cell}}X_7}$$

This new model, fit3, has $R^2_{adj} = 0.7979$ which is better than the previous model, fit2, $R^2_{adj} = 0.7778$. We carry out residual analysis.




We notice that we now have linearity in the residuals and the variance has gotten closer to constant, although still not quite there. The lower end of the QQ plot has tucked in a bit and the rest is quite close the line, showing we gotten closer to a normal distribution for our residuals. We run another model selection process to see if any predictors have become insignificant due to our transformation of the model. This time we use backward elimination with $R^2_{adj}$ and exhaustive search with Mallow's $C_p$ as criteria, to weed out weaker predictors.

```
##   (Intercept) Rural Health Internet BirthRate ElderlyPop GDP Cell adjusted r^2
## 1           1     0      0        0         1          0   0    0    0.7364894
## 2           1     0      0        1         1          0   0    0    0.7756859
## 3           1     1      0        1         1          0   0    0    0.7852103
## 4           1     1      1        1         1          0   0    0    0.7891219
## 5           1     1      1        1         1          1   0    0    0.7932579
## 6           1     1      1        1         1          1   1    0    0.7972095
## 7           1     1      1        1         1          1   1    1    0.7979232
```

```
##   (Intercept) Rural Health Internet BirthRate ElderlyPop GDP Cell Mallows' Cp
## 1           1     0      0        0         1          0   0    0   46.385719
## 2           1     0      0        1         1          0   0    0   18.956288
## 3           1     1      0        0         1          0   1    0   12.743486
## 4           1     1      1        0         1          0   1    0   10.826770
## 5           1     1      1        1         1          1   0    0    9.278258
## 6           1     1      1        1         1          1   1    0    7.497954
## 7           1     1      1        1         1          1   1    1    8.000000
```

The backward elimination method suggests we keep Cell because it has a higher $R^2_{adj}$, but just barely. Mallows $C_p$ is closer to $p = 7$ when we eliminate Cell from the model. Furthermore, the t-test for cell suggests it is not significant when all other predictors are deemed significant. At this point Cell is deemed insignificant, and dropped from the model.

We build a new model with 6 predictors, which we call "fit4":

$$\hat{Y} = \sqrt{\beta_0 + \beta_{\text{rural}}X_1 + \beta_{\text{health}}X_2 + \beta_{\text{internet}}X_3 + \beta_{\text{births}}X_4 + \beta_{\text{elder}}X_5 + \beta_{\text{gdp}}X_6}$$

For fit4, we have $R^2_{adj} = 0.7972$, just slightly reduced from fit3. However, we now have significance in the t-tests for all predictor variables, but GDP is borderline. We carry out residual analysis and find we have a little more "wobbling" in the QQ plot. In fit5 through fit7, we attempt various transformations on the predictors but none improve the model. We come back to fit4 and reevaluate our decision to keep GDP in the model.

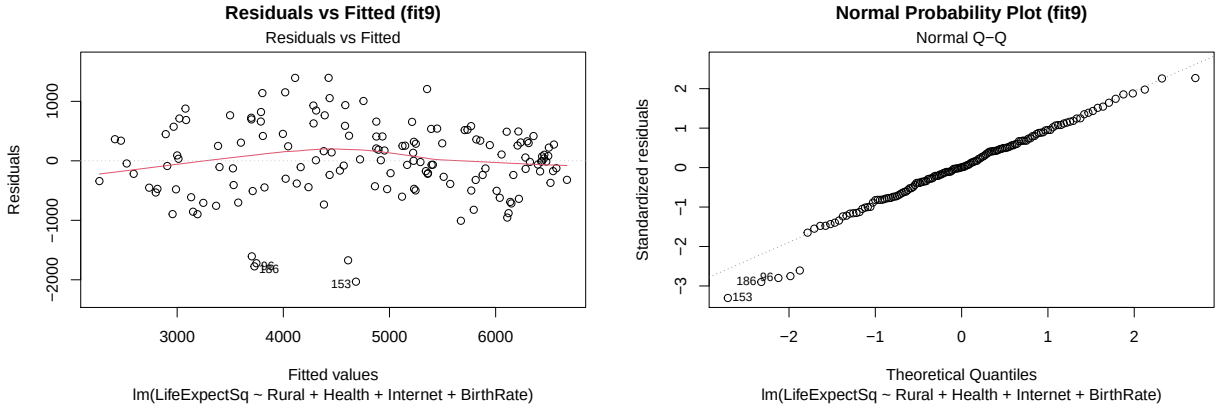
We build a new model:

$$\hat{Y} = \sqrt{\beta_0 + \beta_{\text{rural}}X_1 + \beta_{\text{health}}X_2 + \beta_{\text{internet}}X_3 + \beta_{\text{births}}X_4 + \beta_{\text{elder}}X_5}$$

The residual analysis of this model is looking much better, but we notice that elderlypop appears to be collinear with internet. It is also the only predictor with p-value $> .05$. Suspecting multicollinearity in the model, we asses the variance inflation factors of the model, and see that the VIF for elderlypop is higher than all others at 4.12, approaching a standard cutoff value of 5. Taking all these reasons into account, we drop elderlypop and test out what becomes our final model, fit9:

$$\hat{Y} = \sqrt{\beta_0 + \beta_{\text{rural}}X_1 + \beta_{\text{health}}X_2 + \beta_{\text{internet}}X_3 + \beta_{\text{births}}X_4}$$

```
##
## Call:
## lm(formula = LifeExpectSq ~ Rural + Health + Internet + BirthRate,
##     data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2031.67  -382.93     4.78   409.11  1394.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6377.839    287.014  22.221  < 2e-16 ***
## Rural         -8.883      3.109  -2.857  0.00491 **
## Health        24.001     12.527   1.916  0.05736 .
## Internet       9.325      3.236   2.882  0.00457 **
## BirthRate    -76.120      7.347 -10.361  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 621.5 on 143 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7949, Adjusted R-squared:  0.7891
## F-statistic: 138.5 on 4 and 143 DF,  p-value: < 2.2e-16
```




Notice, in fit9 the residuals now have linearity and arguably constant variance. The points of the QQ plot reside almost entirely on the normal line, save for a very few outliers which are now much closer. It is reasonable to say this model achieves normally distributed, linear residuals with constant variance. And one can reasonably deem all predictors significant, per the t-test, while using scrutiny to recognize that health is still significant despite having p-value just barely over .05. (Subsequent models were tested with Health omitted and were not strong models.)

# Summary

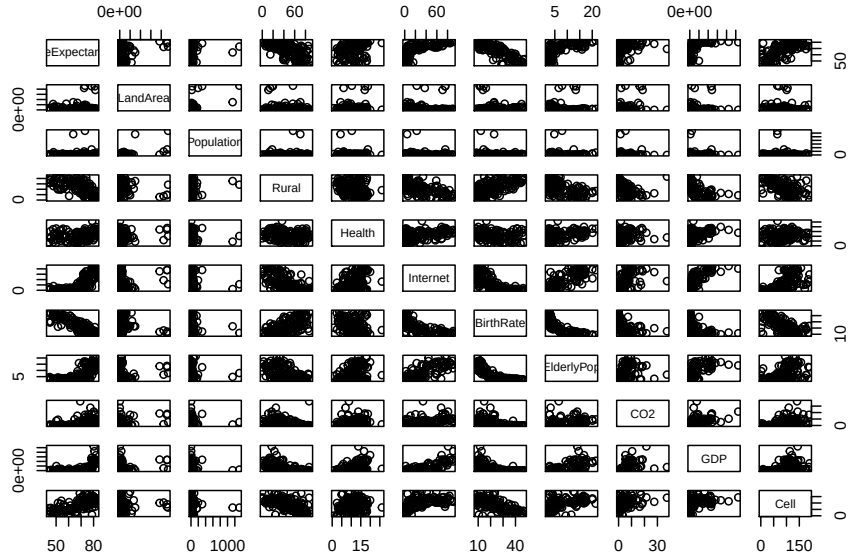
In this project, we used an 80% subsample of the "countries" data to build a multiple regression model. Our final model, called "fit9", predicts life expectancy in a given country by using the metrics of: proportion of population living in rural areas, proportion of government spending on health care, percentage of population with internet access, and rate of births.

We arrived at this model using a holistic model selection approach. We used step-wise model selection algorithms with built in partial F-tests, including "backwards elimination" and "best subsets" exhaustive search. We compared $R_a^2 dj$ and Mallow's $C_p$ in candidate models, as well as significance levels of predictors in the t-tests. We assessed normality, linearity and constant variance in residuals of models, and made transformations, where deemed appropriate per the boxcox transformation method. We combined all of the above with assessment of multicollinearity in the predictor variables by means of scrutinizing the scatter plots, and observing variance inflation factors. All in all, 12 models were created (fit1 through fit12) and fit9, the strongest model, was selected. Fit9:

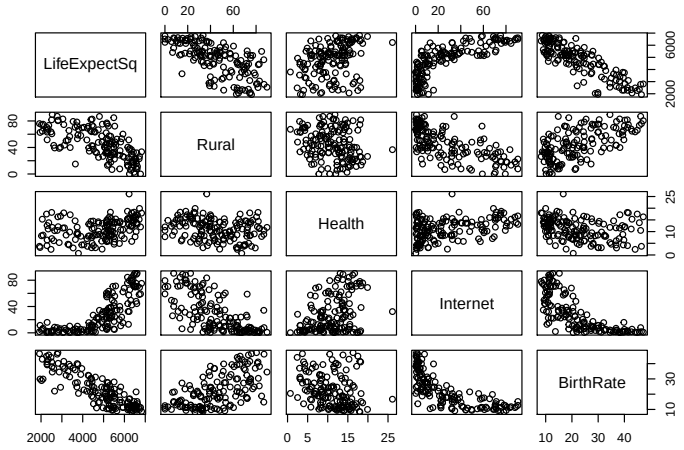
$$\hat{Y} = \sqrt{\beta_0 + \beta_{\text{rural}} X_1 + \beta_{\text{health}} X_2 + \beta_{\text{internet}} X_3 + \beta_{\text{births}} X_4}$$

# Appendix

## pairwise plots for initial data




## pairwise plots for final data




## Code for all models fitted and model building process

```
# set up for project / take 80% subsample
country <- read.csv("~/UCD/Fall 2020/STA 108/data/countries.csv")
set.seed(918343985)
country80 <- country[sample(187, 187 * .8),]
rm(country)
attach(country80)

## fit model [1] -----
fit1 <- lm(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet + BirthRate + ElderlyPop + CO2 + GDP + Cell, data = country80)
summary(fit1)
# Residual plots
plot(fit1, which = 1)
```

```r
#Normal Probability Plot
plot(fit1, which = 2)
```

```r
# Model Selection ----
library(leaps)
## Forward selection
fit1_forward <- regsubsets(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet + BirthRate + ElderlyPop + CO2 + GDP + Cell,
                           data = country80, method = "forward")
cbind(summary(fit1_forward)$which, "adjusted r^2" = summary(fit1_forward)$adjr2)
## Backward elimination
fit1_backward <- regsubsets(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet + BirthRate + ElderlyPop + CO2 + GDP + Cell,
                            data = country80, method = "backward")
cbind(summary(fit1_backward)$which, "adjusted r^2" = summary(fit1_backward)$adjr2)


## fit model [2] -----
fit2 <- lm(LifeExpectancy ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cell, data = country80)
summary(fit2)
# Residual plots
plot(fit2, which = 1)
```

```r
#Normal Probability Plot
plot(fit2, which = 2)
```

```r
library(MASS)
# Boxcox for fit2
boxcox(fit2)
```

```r
## square Y
LifeExpectSq <- (LifeExpectancy)^2

## fit model[3]
LifeExpectSq <- (LifeExpectancy)^2
fit3 <- lm(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cell, data = country80)
summary(fit3)
# Residual plots
plot(fit3, which = 1)
```

```r
#Normal Probability Plot
plot(fit3, which = 2)
```

```r
# Model Selection [3] ----
## Forward selection
fit3_forward <- regsubsets(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cell,
                           data = country80, method = "forward")
cbind(summary(fit3_forward)$which, "adjusted r^2" = summary(fit3_forward)$adjr2)
## Backward elimination
fit3_backward <- regsubsets(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cell,
                            data = country80, method = "backward")
cbind(summary(fit3_backward)$which, "adjusted r^2" = summary(fit3_backward)$adjr2)
# Mallow's Cp
fit3_subset <- regsubsets(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cell,
                          data = country80, method = "exhaustive")
cbind(summary(fit3_subset)$which, "Mallows' Cp" = summary(fit3_subset)$cp)


## fit model [4]
fit4 <- lm(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP, data = country80)
summary(fit4)
# Residual plots
plot(fit4, which = 1)
```

```r
#Normal Probability Plot
plot(fit4, which = 2)
```

```r
# Pairs scatterplot analysis [4]
pairs(cbind(LifeExpectSq, Rural, Health, Internet, BirthRate, ElderlyPop, GDP))
```

```r
cor.test(Rural,BirthRate)


## fit model [5]
InternetSqrt <- Internet^(.5)
ElderSqrt <- ElderlyPop^(.5)
GDPsqrt <- GDP^(.5)
fit5 <- lm(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate + ElderSqrt + GDPsqrt, data = country80)
summary(fit5)
# Residual plots
plot(fit5, which = 1)
```

```r
#Normal Probability Plot
plot(fit5, which = 2)
```

```r
# Model Selection [5] ----
## Forward selection
fit5_forward <- regsubsets(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate + ElderSqrt + GDPsqrt,
                           data = country80, method = "forward")
cbind(summary(fit5_forward)$which, "adjusted r^2" = summary(fit5_forward)$adjr2)
## Backward elimination
fit5_backward <- regsubsets(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate + ElderSqrt + GDPsqrt,
                            data = country80, method = "backward")
cbind(summary(fit5_backward)$which, "adjusted r^2" = summary(fit5_backward)$adjr2)
# Mallow's Cp
fit5_subset <- regsubsets(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate + ElderSqrt + GDPsqrt,
                          data = country80, method = "exhaustive")
cbind(summary(fit5_subset)$which, "Mallows' Cp" = summary(fit5_subset)$cp)
```

```r
## fit model [6] -----
fit6 <- lm(LifeExpectSq ~ Health + InternetSqrt + BirthRate + ElderSqrt + GDPsqrt, data = country80)
summary(fit6)


## fit model and summary output [7] -----
fit7 <- lm(LifeExpectSq ~ InternetSqrt + BirthRate + ElderSqrt + GDPsqrt, data = country80)
summary(fit7)
# Residual plots
plot(fit7, which = 1)
```

```r
#Normal Probability Plot
plot(fit7, which = 2)
```

```r
shapiro.test(resid(fit7))


## fit model [8] -----
fit8 <- lm(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop, data = country80)
summary(fit8)
# Residual plots
plot(fit8, which = 1)
```

```r
#Normal Probability Plot
plot(fit8, which = 2)
```

```r
shapiro.test(resid(fit8))


## fit model [9] -----
fit9 <- lm(LifeExpectSq ~ Rural + Health + Internet + BirthRate, data = country80)
summary(fit9)
# Residual plots
plot(fit9, which = 1)
```

```r
#Normal Probability Plot
plot(fit9, which = 2)
```

```r
shapiro.test(resid(fit9))
pairs(cbind(LifeExpectSq, Rural, Health, Internet, BirthRate))
```

```r
library(car)
vif(fit8)
vif(fit9)


## fit model [10] -----
fit10 <- lm(LifeExpectancy ~ Rural + Health + Internet + BirthRate + ElderlyPop, data = country80)
summary(fit10)
# Residual plots
plot(fit10, which = 1)
```

```r
#Normal Probability Plot
plot(fit10, which = 2)
```

```r
shapiro.test(resid(fit10))


## fit model [11] -----
fit11 <- lm(LifeExpectSq ~ Rural + Internet + BirthRate, data = country80)
summary(fit11)
# Residual plots
plot(fit11, which = 1)
```

```r
#Normal Probability Plot
plot(fit11, which = 2)
```

```r
shapiro.test(resid(fit11))
pairs(cbind(LifeExpectSq, Rural, Internet, BirthRate))
```

```r
## fit model and summary output [12] -----
fit12 <- lm(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate, data = country80)
summary(fit12)
# Residual plots
plot(fit12, which = 1)
```

```r
#Normal Probability Plot
plot(fit12, which = 2)
```

```r
shapiro.test(resid(fit12))
pairs(cbind(LifeExpectSq, Rural, Health, InternetSqrt, BirthRate))
```