Kevin Xu Term Project Fall Quarter 2020 STA108 Professor Drake

Introduction

The “countries80” dataset I explored contains 148 observations. (80% subset of dataset “countries” with 186 observations) It contains variables: ‘Country’ (Name), ‘Code’ (Land code), ‘LandArea’ (Sq Km), ‘Population’ (Millions), ‘Rural’ (%Population living in Rurality), ‘Health’ (% Government expenditure on Healthcare), ‘Internet’ (% Population w/ Internet Access), ‘BirthRate’ (Births Per 1000 People), ‘ElderlyPop’ (% Population at least 65 years old), ‘LifeExpectancy’ (Average Life Expectancy in years), ‘CO2’ (Emissions in metric tons per capita), ‘GDP’ (per capita), ‘Cell’ (Cell Phone subscriptions per 100 people).

I am interested in obtaining an accurate and precise parsimonious multiple linear regression model that describes the relationship of response variable, Life Expectancy, on predictor variables: *Land Area, Population, Rural, Health, Internet, Birth Rate, Elderly Population, CO2, GDP, Cell*. Predictor variables *Country name* and *Country code* are omitted from regression model as a regression of Life Expectancy on qualitative identifiers is not meaningful.

My model selection process is: (1) Multicollinearity diagnostic and remedial measure of predictor variables using Pairwise Scatterplot, Correlation Matrix, t-test analysis with subsets of predictor variables, VIF of candidate predictor variables considered for model selection. (2) Model Selection: Predictor variable selection using Stepwise procedure, R^2_{adj} criterion, and t-test P-value inspection of each predictor variable. (3) Regression assumptions diagnostics (Residual analysis and QQ plot analysis). (4) Remedial measures through Transformations to the model. (5) Model appropriateness assessment through F-test ($F^* = \frac{MSR}{MSE}$), Inspection of Coefficient of Determination (R^2), and Inspection of significance of $|t^*|$ of predictor variables in model.

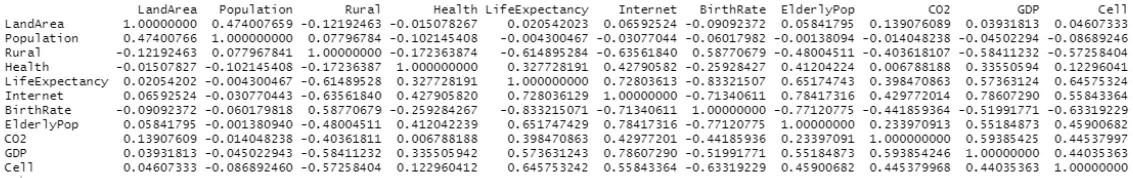
Results:

The final parsimonious model selected is: $Y_i{}^3 = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate,i} + \beta_{GDP}X_{GDP,i} + \beta_{Cell}X_{Cell,i} + \varepsilon_i$. Predictor variable $X_{BirthRate}$has a strong negative effect on response variable LifeExpectancy. As births per 1000 people increases in a given country, Life Expectancy decreases. Predictors $X_{Health}, X_{GDP}, X_{Cell}$ have positive effects on Life Expectancy; positive effects of $X_{Health}\ and\ X_{GDP}$ on Life Expectancy indicate that as % government expenditure towards healthcare increases and/or as GDP per capita increases in a given country, Life Expectancy also increases. The positive effect of X_{Cell} on Life Expectancy suggests that as number of phone subscriptions in a given country increases, Life Expectancy also increases. I would have suspected that predictor X_{Cell} would be collinear with X_{GDP} but the VIF table Figure 1(e) for predictor X_{Cell} says otherwise. The effects of those Predictors can be shown in Figure 3.3 regression coefficients as well the Scatterplot matrix and Correlation matrix.

Model Building:

<u>(1) Checking for Multicollinearity</u>:

Figure 1(b): Correlation Matrix between all predictor variables



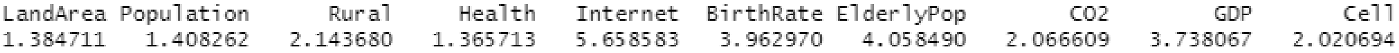
```
                  LandArea   Population       Rural       Health LifeExpectancy    Internet   BirthRate  ElderlyPop          CO2         GDP        Cell
LandArea        1.00000000  0.474007659 -0.12192463 -0.015078267    0.020542023  0.06592524 -0.09092372  0.05841795  0.139076089  0.03931813  0.04607333
Population      0.47400766  1.000000000  0.07796784 -0.102145408   -0.004300467 -0.03077044 -0.06017982 -0.00138094 -0.014048238 -0.04502294 -0.08689246
Rural          -0.12192463  0.077967841  1.00000000 -0.172363874   -0.614895284 -0.63561840  0.58770679 -0.48004511 -0.403618107 -0.58411232 -0.57258404
Health         -0.01507827 -0.102145408 -0.17236387  1.000000000    0.327728191  0.42790582 -0.25928427  0.41204224  0.006788188  0.33550594  0.12296041
LifeExpectancy  0.02054202 -0.004300467 -0.61489528  0.327728191    1.000000000  0.72803613 -0.83321507  0.65174743  0.398470863  0.57363124  0.64575324
Internet        0.06592524 -0.030770443 -0.63561840  0.427905820    0.728036129  1.00000000 -0.71340611  0.78417316  0.429772014  0.78607290  0.55843364
BirthRate      -0.09092372 -0.060179818  0.58770679 -0.259284267   -0.833215071 -0.71340611  1.00000000 -0.77120775 -0.441859364 -0.51991771 -0.63319229
ElderlyPop      0.05841795 -0.001380940 -0.48004511  0.412042239    0.651747429  0.78417316 -0.77120775  1.00000000  0.233970913  0.55184873  0.45900682
CO2             0.13907609 -0.014048238 -0.40361811  0.006788188    0.398470863  0.42977201 -0.44185936  0.23397091  1.000000000  0.59385425  0.44537997
GDP             0.03931813 -0.045022943 -0.58411232  0.335505942    0.573631243  0.78607290 -0.51991771  0.55184873  0.593854246  1.00000000  0.44035363
Cell            0.04607333 -0.086892460 -0.57258404  0.122960412    0.645753242  0.55843364 -0.63319229  0.45900682  0.445379968  0.44035363  1.00000000
```

Collinearity of a predictor variable implies that information it provides about the response variable, Life Expectancy, is redundant in the presence of other predictor variables. Predictor variables (Rural and Birthrate), (Internet and Health), and (Internet and ElderlyPop) appear to be correlated as shown in the **Pairwise Scatterplot** in *Appendix Figure 1(a)*. This was further investigated by the **Correlation Matrix** in *Figure 1(b):* $r_{Rural,Birthrate} = 0.58770679, r_{Internet,Health} = 0.42790582, r_{Internet,ElderlyPop} = 0.78417316.$

Comparisons of t statistic (t^*), regression coefficients ($b_k,, k{:}\ Health, Internet, Elderly\ Pop\ for\ respective\ tests$), and standard error ($s(b_k$) to determine multicollinearity:

Analysis from t-tests is inconclusive: Analysis of predictors $X_{ElderlyPop}$ and X_{Health} show that P-value for test statistic, t^*, for X_{Health} is significant ($4.78 * 10^{-5}$) in the model $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \varepsilon_i$, but is insignificant (0.302) in model $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{ElderlyPop}X_{ElderlyPop,i} + \varepsilon_i$. This is shown in *Figure 1(c) in the Appendix.* Analysis of predictors $X_{Internet}$ and X_{Health} show that test statistic, t^*, for health is significant ($4.78 * 10^{-5}$) in the model $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \varepsilon_i$, but is highly insignificant (0.753) in the model $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{Internet}X_{Internet,i} + \varepsilon_i$ *Figure 1(d).*

Figure 1(e): VIF



```
LandArea Population      Rural     Health   Internet  BirthRate ElderlyPop        CO2        GDP       Cell 
1.384711   1.408262   2.143680   1.365713   5.658583   3.962970   4.058490   2.066609   3.738067   2.020694 
```

Inspection of VIF (Variance Inflation Factor) of predictor variables to determine multicollinearity:

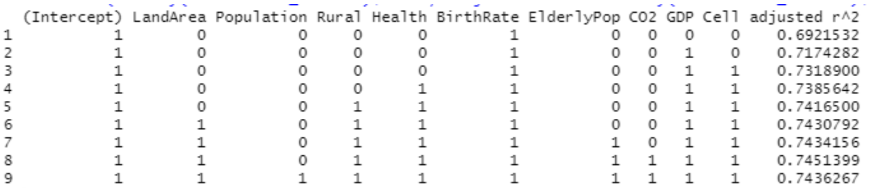
For a given predictor variable (p), VIF assesses multicollinearity by measuring how much variance of a regression coefficient is inflated due to collinearity in the model. Large values of VIF ($VIF(\widehat{\beta_k}) > 5, k{:}\ LandArea, Population, \dots, Cell$), indicate a problematic amount of multicollinearity. *Figure 1(e)* shows VIF values for each predictor variable in the preliminary model with all predictors: It shows that predictor $X_{Internet}$ has a high VIF value of 5.658583, indicating a problematic amount of collinearity. *Predictor variable $X_{Internet}$ will no longer be considered for model selection*.

(2) Model Selection: Predictor variable selection:

Note: Candidate predictor variables considered for variable selection excludes $X_{Internet}$ due to the collinearity it possesses.

R^2_{adj} criterion predictor variable selection using **Forward Selection**: The best model indicated by criterion R^2_{adj} has the highest R^2_{adj} value. The last column in Figure 1.1 shows the R^2_{adj} values for each model. The best model has $R^2_{adj} = 0.7451399$ under Forward Selection. Eight predictor variables have been selected: $X_{Land\ Area}, X_{Rural}, X_{Health}, X_{BirthRate}, X_{ElderlyPop}, X_{CO2}, X_{GDP}, and\ X_{Cell}$.

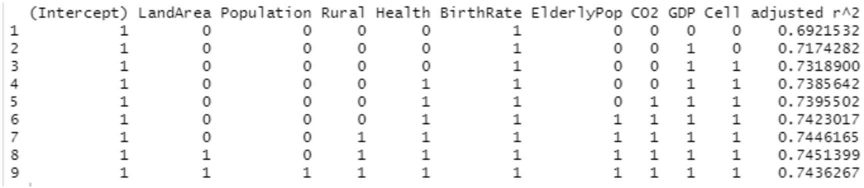
Figure 2.1: Forward Selection using R^2_{adi} criterion



```
  (Intercept) LandArea Population Rural Health BirthRate ElderlyPop CO2 GDP Cell adjusted r^2
1           1        0          0     0      0         1          0   0   0    0    0.6921532
2           1        0          0     0      0         1          0   0   1    0    0.7174282
3           1        0          0     0      0         1          0   0   1    1    0.7318900
4           1        0          0     0      1         1          0   0   1    1    0.7385642
5           1        0          0     1      1         1          0   0   1    1    0.7416500
6           1        1          0     1      1         1          0   0   1    1    0.7430792
7           1        1          0     1      1         1          1   0   1    1    0.7434156
8           1        1          0     1      1         1          1   1   1    1    0.7451399
9           1        1          1     1      1         1          1   1   1    1    0.7436267
```

R^2_{adj} criterion predictor variable selection using **Backward Elimination**: The best model under Backward Elimination is identical to that of Forward Selection. The best model has $R^2_{adj} = 0.7451399$. Predictor variable $X_{Population}$ has been eliminated.

Figure 2.2: Backward Elimination using R^2_{adi} criterion



```
  (Intercept) LandArea Population Rural Health BirthRate ElderlyPop CO2 GDP Cell adjusted r^2
1           1        0          0     0      0         1          0   0   0    0    0.6921532
2           1        0          0     0      0         1          0   0   1    0    0.7174282
3           1        0          0     0      0         1          0   0   1    1    0.7318900
4           1        0          0     0      1         1          0   0   1    1    0.7385642
5           1        0          0     0      1         1          0   1   1    1    0.7395502
6           1        0          0     0      1         1          1   1   1    1    0.7423017
7           1        0          0     1      1         1          1   1   1    1    0.7446165
8           1        1          0     1      1         1          1   1   1    1    0.7451399
9           1        1          1     1      1         1          1   1   1    1    0.7436267
```

Checking the significance of predictor variables after Stepwise procedure in t-tests: (taking $\alpha = 0.05$) This is shown in *Appendix Figure 2.3.* It shows that predictor variables $X_{Land\ Area}, X_{Rural}, X_{ElderlyPop}, X_{CO2}$ are insignificant with P-values 0.25846, 0.10528, 0.14185,0.16512 respectively. These predictors do not contribute significantly to the prediction of response variable Life Expectancy behavior and will be eliminated from the model.

From Forward selection and Backward elimination procedures, using R^2_{adj} criterion, the predictor variable $X_{Population}$ is eliminated. Additionally, from checking significance of remaining predictor variables after Stepwise procedure, $X_{Land\ Area}, X_{Rural}, X_{ElderlyPop}, X_{CO2}$ are eliminated. Predictor variables selected for my model will be: $X_{Health}, X_{BirthRate}\ X_{GDP}, and\ X_{Cell}$

(3) Regression Assumption Diagnostics:

Checking Linearity and Constant Variance Regression Assumptions:

Residual plots against each predictor variable will be utilized to visually check for noncompliance of the constant variance and linearity regression assumptions of the multiple linear regression model. They are shown in Figure *2.4.1*. Note: 4 outliers were removed in the residual plot against predictor X_{GDP}, they are shown in *Appendix Figure 2.3(a)*

Predictor	Constant Error Terms Variance Assessment	Linearity Assessment
X_{Health}	Approx. Constant Error Variance	Linear
$X_{BirthRate}$	Nonconstant Error Variance	Nonlinear
X_{GDP}	Nonconstant Error Variance	Approx. Linear
X_{Cell}	Approx. Constant Error Variance	Approx. Linear

Predictors that display nonconstant error variance are those that contain observations that shift in deviation from the residual mean $e = 0$ as the predictor level of $X_k(k: Health, BirthRate, GDP, Cell)$ increases. $X_{BirthRate}, X_{GDP}$ display nonconstant error variance: as predictor level increases, deviation of residual observations from residual mean decreases between interval $(10 \leq X_{BirthRate} \leq 20)$ and increases between interval $(20 \leq X_{BirthRate} \leq 35)$ for predictor $X_{BirthRate}$. Deviation of residual observations from residual mean decreases in the entire interval for predictor X_{GDP}.

Predictors that display nonlinearity are those that depart from the residual mean $e = 0$ in a systematic fashion as the predictor level of $X_k(k: Health, BirthRate, GDP, Cell)$ increases. This can be shown by uneven number of observations above and below the residual mean line $e = 0$ at various X_k levels. $X_{BirthRate}$ exhibits nonlinear qualities: residual observations systematically shift above the residual mean in interval $(10 < X_{BirthRate} < 20)$.

Residual Plots from Figure 2.4.1 shows that:

- Predictors $X_{BirthRate}, and\ X_{GDP}$ display noncompliance with the constant error terms variance regression assumption.
- Predictor $X_{BirthRate}$ displays noncompliance with the linearity regression assumption.

Additionally, *Figure 2.4.2 (*Residual plot of residuals (e_i) against fitted Y-hat ($\widehat{Y_i}$) values) shows that the model: $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate,i} + \beta_{GDP}X_{GDP,i} + \beta_{Cell}X_{Cell,i} + \varepsilon_i$, is neither in compliance with the constant error terms variance assumption nor linearity assumption of the multiple linear regression model: Residual observations decrease in deviation from the residual mean $e = 0$, in interval $(50 \leq \hat{Y} \leq 65)$ and increase in interval $(65 \leq \hat{Y} \leq 80)$. Residual observations systematically increase in the interval $(50 \leq \hat{Y} \leq 65)$ and systematically decrease in the interval $(65 \leq \hat{Y} \leq 80)$.

Figure 2.4.1: *Residual Plots of Residuals against individual predictor variables*

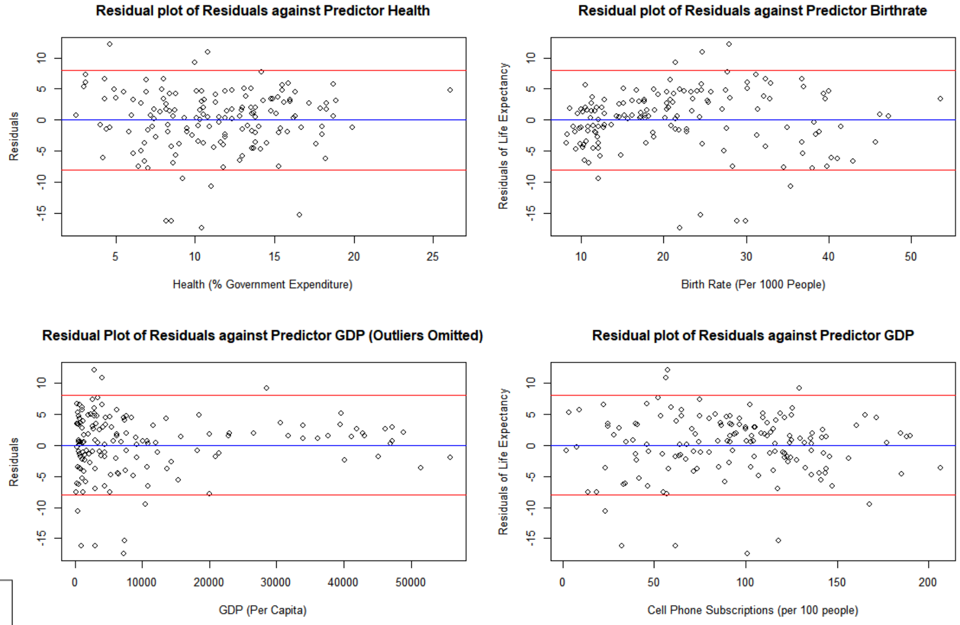
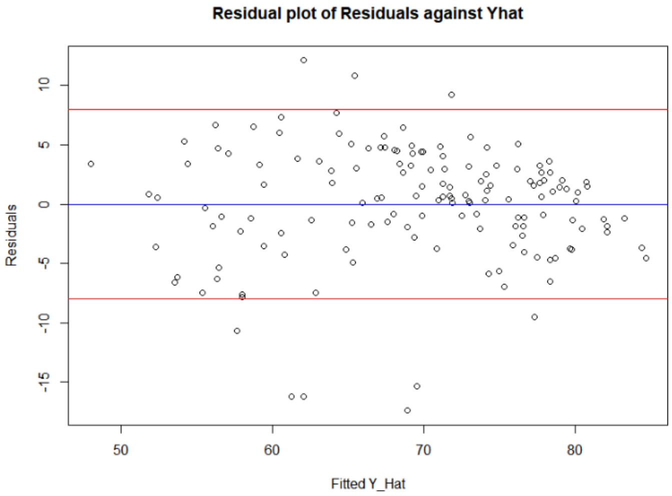


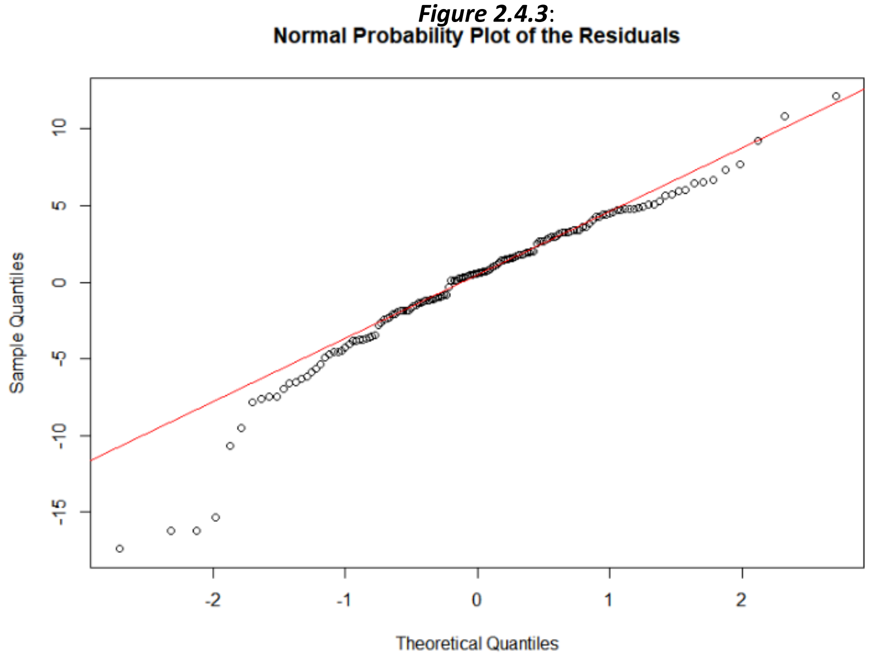
Figure 2.4.2: *Residual Plot of Residuals against Fitted $\hat{Y}$ values*



Checking Normality Regression Assumption:

The Normal Probability Plot of the Residuals, that shows residuals plotted against expected residual values under normality, will be utilized to visually check for noncompliance with the normality assumption of multiple linear regression. This is shown in *Figure 2.4.3.*

Figure 2.4.3:



The red fitted line in *Figure 2.4.3* shows the expected value of residuals when in compliance with the normality assumption. As indicated by the left tail at theoretical quantile values of less than -1, residual observations drastically depart from the expected residual values under normality. Residuals from the model are not in compliance with the normality assumption of the multiple linear regression model.

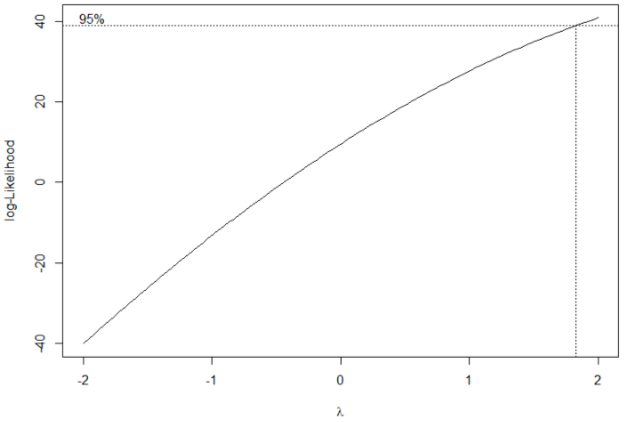
Regression Assumptions Diagnostics Conclusions:

Analysis shows that model: $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate,i} + \beta_{GDP}X_{GDP,i} + \beta_{Cell}X_{Cell,i} + \varepsilon_i$ is not in compliance with the linearity, constant error variance, or normality assumptions of the multiple linear regression model. Additionally, predictor variables $X_{BirthRate}, and\ X_{GDP}$ display noncompliance with the constant error terms variance regression assumption, and predictor variable $X_{BirthRate}$ displays noncompliance with the linearity regression assumption

(4) Possible Transformations:

Y-transformations will be considered first to the model to correct nonconstant error variance and nonnormality. After trial and error, and using Box-Cox procedure, the cubic Y-transformation ($Y' = Y^3$) was selected. The Box-Cox plot, shown in *Figure 2.5,* indicates an optimal λ value for Y^λ somewhere around the value of 2, after comparisons of QQ plots and residual plots between $Y' = \log(Y)$, $Y' = Y^2$, and $Y' = Y^3$, $Y' = Y^3$ is selected on the basis that it mitigates effects of nonlinearity, nonconstant error variance and nonnormality. The processes of selection are shown in *Appendix Figure 2.5(a)*.

Figure 2.5: Box Cox Plot of log-likelihood against λ values under the model: $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate.i} + \beta_{GDP}X_{GDP.i} + \beta_{Cell}X_{Cell.i} + \varepsilon_i$



The effects of a cubic Y-transformation on normality and error variances were investigated in Figure 2.5.1 and 2.5.2 respectively.

QQ plots: After transformation, residual observations are generally closer to the fitted line (expected residual values under normality). There are five observations that significantly deviate from expected residual value under normality at Theoretical Quantiles < -1.5 in the untransformed model, they are closer to the fitted line after transformation.

Residual plots: After transformation, residual observations are generally more evenly distributed throughout the plot and do not appear to systematically depart from e=0; as $\hat{Y}$ increases, residual deviation from e=0 is approximately constant. Error terms variance of the transformed model appears to be approximately constant and residual observations appear to be approximately linear.

Figure 2.5.1: Comparison of QQ Plots of Pre Y-Transformation and Post Y-Transformation

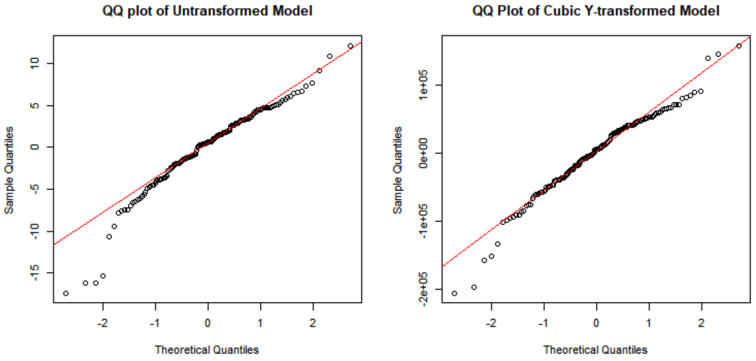
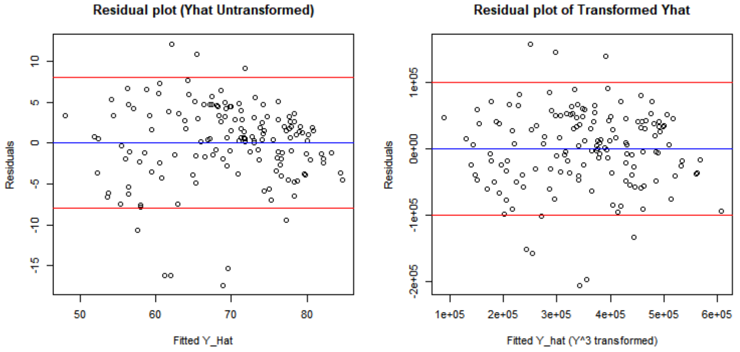


Figure 2.5.2: Comparison of Residual Plots of Residuals against Fitted $\hat{Y}$ of Pre Y-transformation and Post Y-Transformation



X-transformations are applied on predictor variables that demonstrate approximately constant error terms variance and nonlinearity. Since no predictor variables appear to exhibit these characteristics, X-transformations will not be applied the model.

The current model is:

$$Y_i^{\,3} = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate,i} + \beta_{GDP}X_{GDP,i} + \beta_{Cell}X_{Cell,i} + \varepsilon_i$$

(5) Testing Appropriateness of Model:

F-test:

As a preliminary test of appropriateness, assuming that the model complies with linearity, constant error variance and normality assumptions of the multiple linear regression model, I tested the F-statistic of the current model.

$F^* = \frac{\frac{SSR}{p-1}}{\frac{SSE}{n-p}} = 125.8015$ has a P-value: $8.287228 * 10^{-46}$. Since the P-value for the F-statistic is highly significant, the model appears to be appropriate for the data at hand. Calculations are shown in *Appendix Figure 3.1*.

Inspection of Coefficient of Determination (R^2):

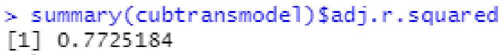
The current model appears to be appropriate due to the high value of $R^2 = 0.7725184$- this is shown in *Figure 3.2*. The R^2 of the current model also appears to be higher than the preliminary model with 4 predictors (the same predictor variables as the current model) as can be indicated from row 4 of *Figure 2.1* from Stepwise Procedure of Predictor Variable selection: where $R^2 = 0.7385642$.

($R^2{}_{Current\ model} = 0.7725184 > R^2{}_{Preliminary\ model} = 0.7385642$)

Current model: $Y_i{}^3 = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate,i} + \beta_{GDP}X_{GDP,i} + \beta_{Cell}X_{Cell,i} + \varepsilon_i$

Preliminary model: $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate,i} + \beta_{GDP}X_{GDP,i} + \beta_{Cell}X_{Cell,i} + \varepsilon_i$

Figure 3.2- R^2 of current model.

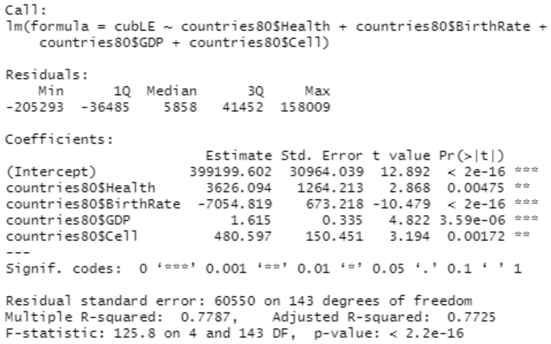


```
> summary(cubtransmodel)$adj.r.squared
[1] 0.7725184
```

Inspection of Significance of Predictor Variables in model:

The current model appears to only possess highly significant predictor variables. P-values of $|t^*|$ of predictors $X_{Health}, X_{BirthRate}, X_{GDP}, X_{Cell}$ are all less than $\alpha = 0.05$ and significantly contribute to modelling the behavior of response variable Life Expectancy. This is shown in *Figure 3.3*.

Figure 3.3- Inspection of significance of predictor variables in Current model



```
Call:
lm(formula = cubLE ~ countries80$Health + countries80$BirthRate +
    countries80$GDP + countries80$Cell)

Residuals:
    Min      1Q  Median      3Q     Max
-205293  -36485    5858   41452  158009

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           399199.602  30964.039  12.892  < 2e-16 ***
countries80$Health      3626.094   1264.213   2.868  0.00475 **
countries80$BirthRate  -7054.819    673.218 -10.479  < 2e-16 ***
countries80$GDP            1.615      0.335   4.822 3.59e-06 ***
countries80$Cell         480.597    150.451   3.194  0.00172 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60550 on 143 degrees of freedom
Multiple R-squared:  0.7787,	Adjusted R-squared:  0.7725
F-statistic: 125.8 on 4 and 143 DF,  p-value: < 2.2e-16
```

Summary

After adjustments, my model building process was:

(1) Multicollinearity diagnostic and remedial measure using Pairwise Scatterplot, Correlation Matrix, t-test analysis with subsets of predictors, VIF. $X_{Internet}$ was discovered to exhibit collinearity and was removed from model consideration.

(2) Model Selection: Predictor variable selection using Stepwise procedure (forward selection and backward elimination with R^2_{adj} criterion), and t-test P-value inspection of each predictor variable. The selected predictor variables are: $X_{Health}, X_{BirthRate}\ , X_{GDP}, and\ X_{Cell}$. Predictor variables $X_{LandArea}, X_{Population}, X_{Rural}, X_{ElderlyPop}, X_{CO2}$ are the next predictor variables eliminated from model consideration.

(3) Regression assumptions diagnostics (Residual analysis and QQ plot analysis). Nonlinearity, nonconstant error variance, and nonnormality characteristics are exhibited by the preliminary model $Y_i = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate,i} + \beta_{GDP}X_{GDP,i} + \beta_{Cell}X_{Cell,i} + \varepsilon_i$. These characteristics were identified by observing residual plots against individual predictor variables, residual plot against fitted $\hat{Y}$ values, and Normal Probability Plot of Residuals.

(4) Remedial measures through Transformations. Y-transformation: $Y' = Y^3$ was selected. Between alternatives, $Y' = Y^2\ and\ Y' = \log(Y)$: $Y' = Y^3$ corrected regression assumption violations the best. Y-Transformed model exhibited approximately constant error variance, approximately linear and less deviation from normality. X-transformation was not applicable as none of the predictor variables exhibit nonlinearity and approximately constant error variance.

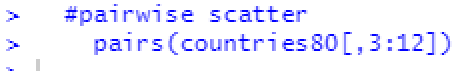
(5) Model appropriateness assessment through F-test ($F^* = \frac{MSR}{MSE}$, Inspection of Coefficient of Determination (R^2), and Inspection of significance of $|t^*|$ of predictor variables in model. F-test statistic of the final model is 125.8015 with a highly significant P-value of $8.287228 * 10^{-46}$, R^2 of the final model is 0.7725184 showing improvement after remedial measures, and all predictor variables had highly significant P-values of less than 0.005. The final model appears to be parsimonious and appropriate at predicting the behavior of Life Expectancy within the 148 countries that were included in the “countries80” subsampled dataset under the assumption that normality, linearity, and constant error variance hold.

Major findings about the data: predictor variable $X_{BirthRate}$ is most relevant with predicting LifeExpectancy. It has a strong negative effect. This can be shown in Figure 3.3 regression coefficients as well the Scatterplot matrix and Correlation matrix. This may be consistent with common sense: perhaps countries with lower life expectancy have a higher birthrate to ensure survival of some offspring, and countries with better general life expectancy have a lower birthrate as survival of offspring is likely. Predictors $X_{Health}, X_{GDP}, X_{Cell}$ have positive effects on Life Expectancy; positive effects of $X_{Health}\ and\ X_{GDP}$ on Life Expectancy indicate that as % government expenditure towards healthcare increases and/or as GDP per capita increases in a given country, Life Expectancy will also increase. The positive effect of X_{Cell} on Life Expectancy suggests that as number of phone subscriptions in a given country increases, so does Life Expectancy.

The final parsimonious model selected is: $Y_i{}^3 = \beta_0 + \beta_{Health}X_{Health,i} + \beta_{BirthRate}X_{BirthRate,i} + \beta_{GDP}X_{GDP,i} + \beta_{Cell}X_{Cell,i} + \varepsilon_i$.

Appendix

Figure 1(a): Scatterplot Matrix



```
>   #pairwise scatter
>     pairs(countries80[,3:12])
```

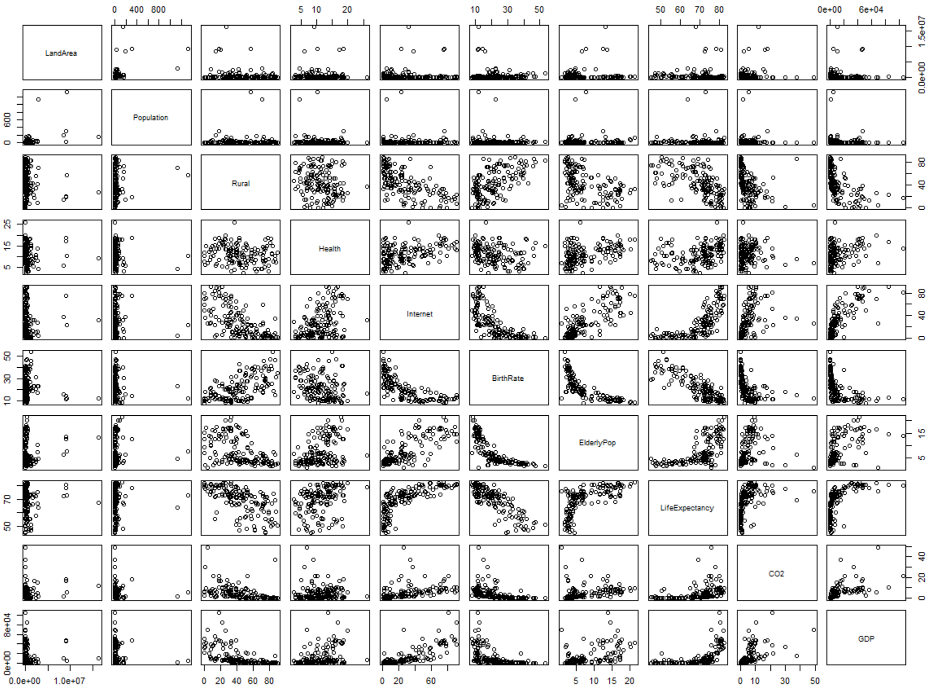
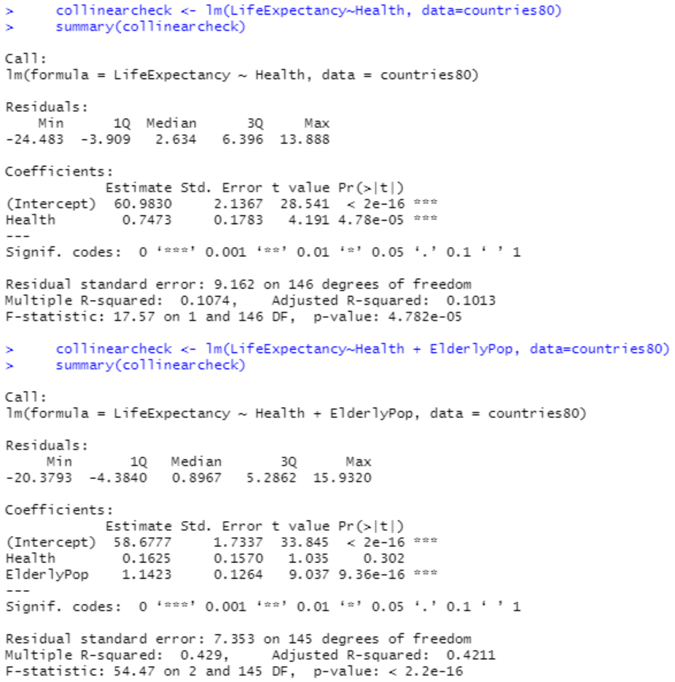
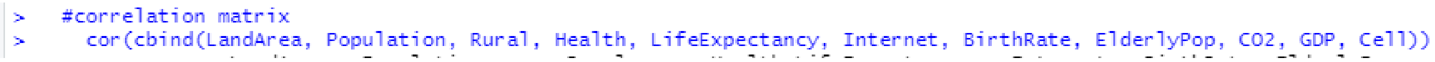


Figure 1(c): Multicollinearity inspection: t-tests of subset models of suspected collinear predictors.



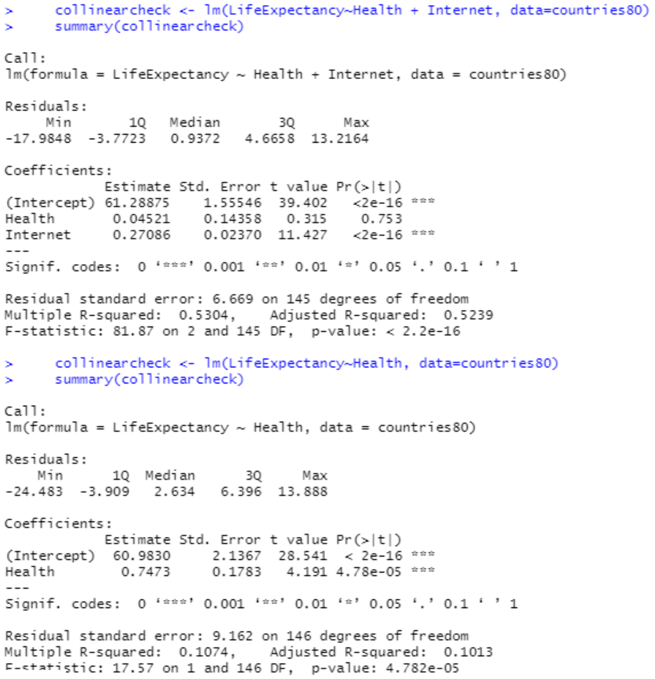
```
>     collinearcheck <- lm(LifeExpectancy~Health, data=countries80)
>     summary(collinearcheck)

Call:
lm(formula = LifeExpectancy ~ Health, data = countries80)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.483  -3.909   2.634   6.396  13.888 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  60.9830     2.1367  28.541  < 2e-16 ***
Health        0.7473     0.1783   4.191 4.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.162 on 146 degrees of freedom
Multiple R-squared:  0.1074,	Adjusted R-squared:  0.1013 
F-statistic: 17.57 on 1 and 146 DF,  p-value: 4.782e-05

>     collinearcheck <- lm(LifeExpectancy~Health + ElderlyPop, data=countries80)
>     summary(collinearcheck)

Call:
lm(formula = LifeExpectancy ~ Health + ElderlyPop, data = countries80)

Residuals:
     Min       1Q   Median       3Q      Max 
-20.3793  -4.3840   0.8967   5.2862  15.9320 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  58.6777     1.7337  33.845  < 2e-16 ***
Health        0.1625     0.1570   1.035    0.302    
ElderlyPop    1.1423     0.1264   9.037 9.36e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.353 on 145 degrees of freedom
Multiple R-squared:  0.429,	Adjusted R-squared:  0.4211 
F-statistic: 54.47 on 2 and 145 DF,  p-value: < 2.2e-16
```

Figure 1(b): Correlation Matrix



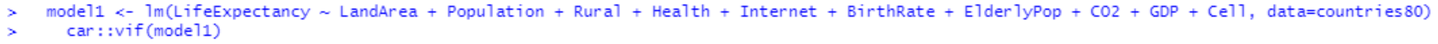
```
>   #correlation matrix
>     cor(cbind(LandArea, Population, Rural, Health, LifeExpectancy, Internet, BirthRate, ElderlyPop, CO2, GDP, Cell))
```

Figure 1(c): Multicollinearity inspection: t-tests of subset models of suspected collinear predictors.



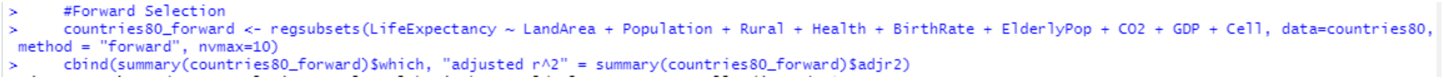
```
>     collinearcheck <- lm(LifeExpectancy~Health + Internet, data=countries80)
>     summary(collinearcheck)

Call:
lm(formula = LifeExpectancy ~ Health + Internet, data = countries80)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.9848  -3.7723   0.9372   4.6658  13.2164 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 61.28875    1.55546  39.402   <2e-16 ***
Health       0.04521    0.14358   0.315    0.753    
Internet     0.27086    0.02370  11.427   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.669 on 145 degrees of freedom
Multiple R-squared:  0.5304,	Adjusted R-squared:  0.5239 
F-statistic: 81.87 on 2 and 145 DF,  p-value: < 2.2e-16

>     collinearcheck <- lm(LifeExpectancy~Health, data=countries80)
>     summary(collinearcheck)

Call:
lm(formula = LifeExpectancy ~ Health, data = countries80)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.483  -3.909   2.634   6.396  13.888 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  60.9830     2.1367  28.541  < 2e-16 ***
Health        0.7473     0.1783   4.191 4.78e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.162 on 146 degrees of freedom
Multiple R-squared:  0.1074,	Adjusted R-squared:  0.1013 
F-statistic: 17.57 on 1 and 146 DF,  p-value: 4.782e-05
```

Figure 1(e): VIF calculation



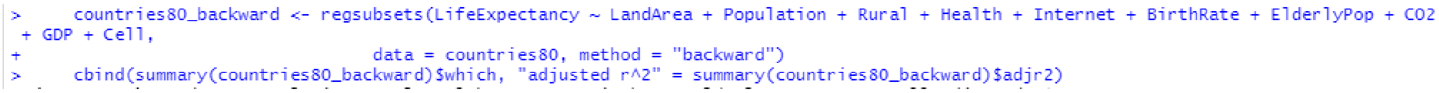
```
>    model1 <- lm(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet + BirthRate + ElderlyPop + CO2 + GDP + Cell, data=countries80)
>      car::vif(model1)
```

Figure 2.1: Forward Selection



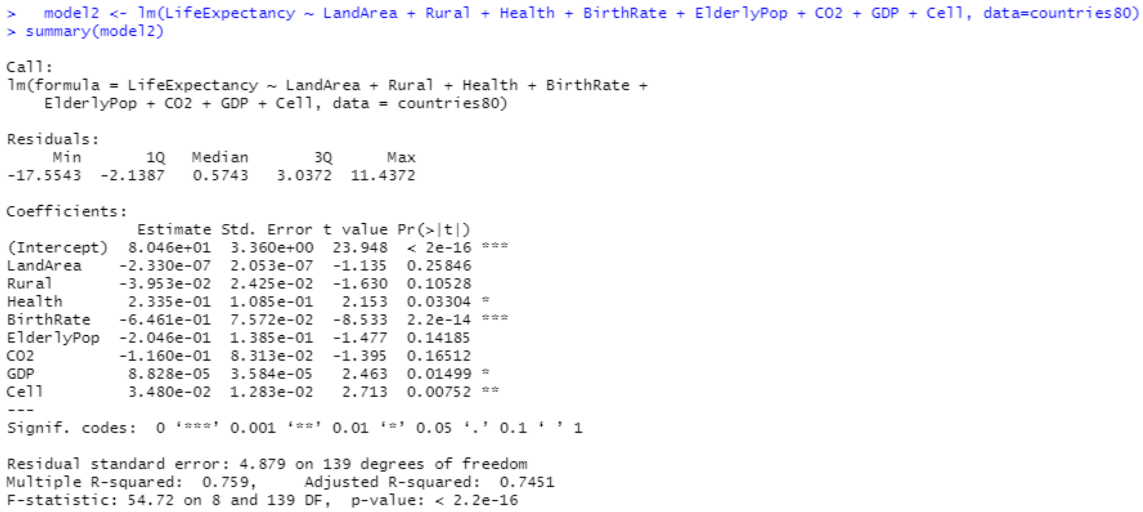
```
>     #Forward Selection
>     countries80_forward <- regsubsets(LifeExpectancy ~ LandArea + Population + Rural + Health + BirthRate + ElderlyPop + CO2 + GDP + Cell, data=countries80,
 method = "forward", nvmax=10)
>     cbind(summary(countries80_forward)$which, "adjusted r^2" = summary(countries80_forward)$adjr2)
```

Figure 2.2: Backward Elimination



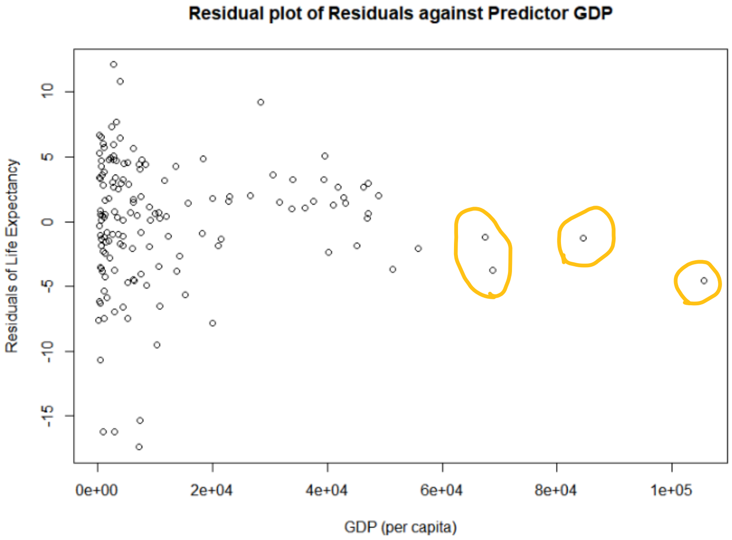
```
>     countries80_backward <- regsubsets(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet + BirthRate + ElderlyPop + CO2
 + GDP + Cell,
+                                  data = countries80, method = "backward")
>     cbind(summary(countries80_backward)$which, "adjusted r^2" = summary(countries80_backward)$adjr2)
```

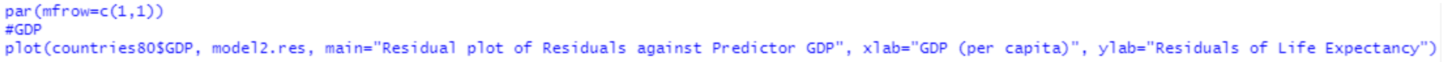
Figure 2.3: Checking significance of predictor variables after Stepwise Procedure



```
>   model2 <- lm(LifeExpectancy ~ LandArea + Rural + Health + BirthRate + ElderlyPop + CO2 + GDP + Cell, data=countries80)
> summary(model2)

Call:
lm(formula = LifeExpectancy ~ LandArea + Rural + Health + BirthRate + 
    ElderlyPop + CO2 + GDP + Cell, data = countries80)

Residuals:
     Min       1Q   Median       3Q      Max 
-17.5543  -2.1387   0.5743   3.0372  11.4372 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.046e+01  3.360e+00  23.948  < 2e-16 ***
LandArea    -2.330e-07  2.053e-07  -1.135  0.25846    
Rural       -3.953e-02  2.425e-02  -1.630  0.10528    
Health       2.335e-01  1.085e-01   2.153  0.03304 *  
BirthRate   -6.461e-01  7.572e-02  -8.533  2.2e-14 ***
ElderlyPop  -2.046e-01  1.385e-01  -1.477  0.14185    
CO2         -1.160e-01  8.313e-02  -1.395  0.16512    
GDP          8.828e-05  3.584e-05   2.463  0.01499 *  
Cell         3.480e-02  1.283e-02   2.713  0.00752 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.879 on 139 degrees of freedom
Multiple R-squared:  0.759,	Adjusted R-squared:  0.7451 
F-statistic: 54.72 on 8 and 139 DF,  p-value: < 2.2e-16
```

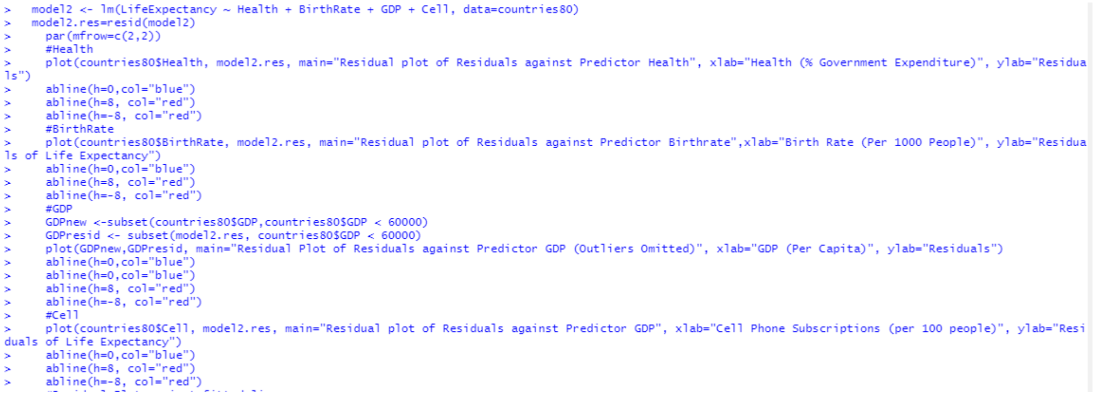
Figure 2.3(a): Omitting 4 observations from residual plot of residuals against predictor GDP.





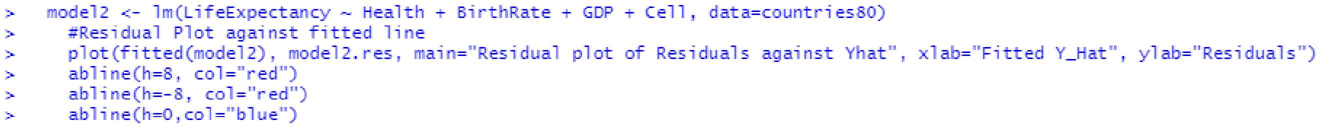
```
par(mfrow=c(1,1))
#GDP
plot(countries80$GDP, model2.res, main="Residual plot of Residuals against Predictor GDP", xlab="GDP (per capita)", ylab="Residuals of Life Expectancy")
```

Figure 2.4.1: Residual Plots against selected predictor variables.



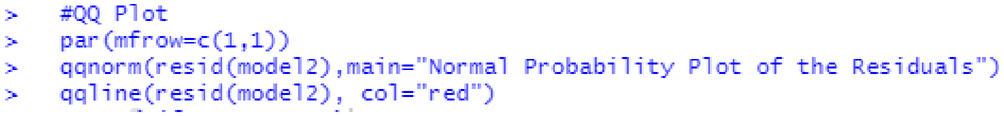
```
>   model2 <- lm(LifeExpectancy ~ Health + BirthRate + GDP + Cell, data=countries80)
>   model2.res=resid(model2)
>     par(mfrow=c(2,2))
>     #Health
>     plot(countries80$Health, model2.res, main="Residual plot of Residuals against Predictor Health", xlab="Health (% Government Expenditure)", ylab="Residua
ls")
>     abline(h=0,col="blue")
>     abline(h=8, col="red")
>     abline(h=-8, col="red")
>     #BirthRate
>     plot(countries80$BirthRate, model2.res, main="Residual plot of Residuals against Predictor Birthrate",xlab="Birth Rate (Per 1000 People)", ylab="Residua
ls of Life Expectancy")
>     abline(h=0,col="blue")
>     abline(h=8, col="red")
>     abline(h=-8, col="red")
>     #GDP
>     GDPnew <-subset(countries80$GDP,countries80$GDP < 60000)
>     GDPresid <- subset(model2.res, countries80$GDP < 60000)
>     plot(GDPnew,GDPresid, main="Residual Plot of Residuals against Predictor GDP (Outliers Omitted)", xlab="GDP (Per Capita)", ylab="Residuals")
>     abline(h=0,col="blue")
>     abline(h=0,col="blue")
>     abline(h=8, col="red")
>     abline(h=-8, col="red")
>     #Cell
>     plot(countries80$Cell, model2.res, main="Residual plot of Residuals against Predictor GDP", xlab="Cell Phone Subscriptions (per 100 people)", ylab="Resi
duals of Life Expectancy")
>     abline(h=0,col="blue")
>     abline(h=8, col="red")
>     abline(h=-8, col="red")
```

Figure 2.4.2: Residual plot against Y_hat



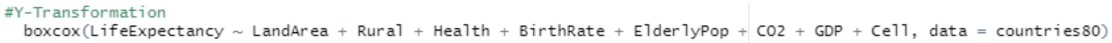
```
>   model2 <- lm(LifeExpectancy ~ Health + BirthRate + GDP + Cell, data=countries80)
>     #Residual Plot against fitted line
>     plot(fitted(model2), model2.res, main="Residual plot of Residuals against Yhat", xlab="Fitted Y_Hat", ylab="Residuals")
>     abline(h=8, col="red")
>     abline(h=-8, col="red")
>     abline(h=0,col="blue")
```

Figure 2.4.3: QQ plot of preliminary model



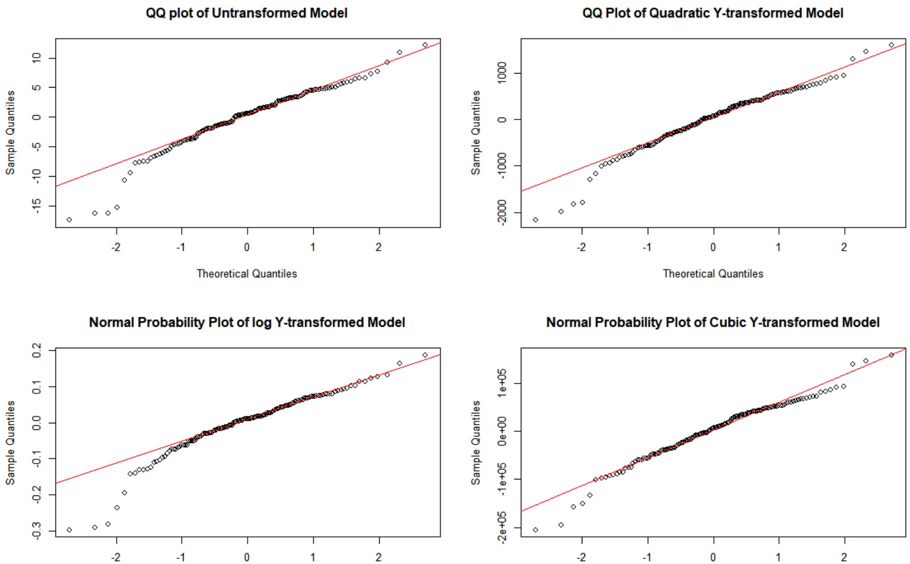
```
>   #QQ Plot
>   par(mfrow=c(1,1))
>   qqnorm(resid(model2),main="Normal Probability Plot of the Residuals")
>   qqline(resid(model2), col="red")
```

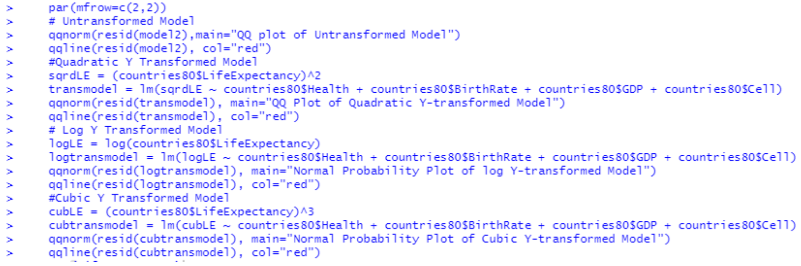
Figure 2.5: Box Cox Procedure



```
#Y-Transformation
  boxcox(LifeExpectancy ~ LandArea + Rural + Health + BirthRate + ElderlyPop + CO2 + GDP + Cell, data = countries80)
```

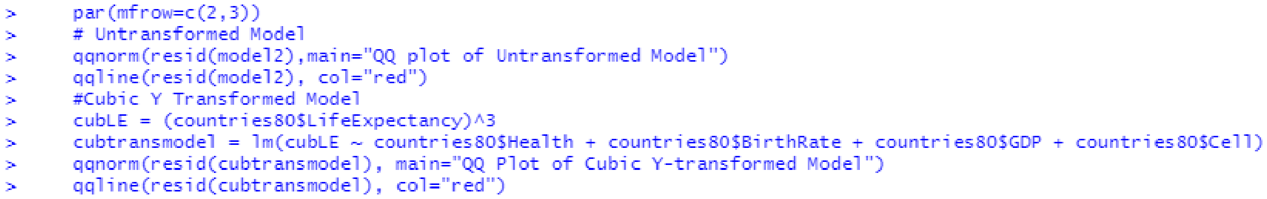
Figure 2.5(a): Comparison of QQ plots of Untransformed vs Quadratic, Log and Cubic Transformed model





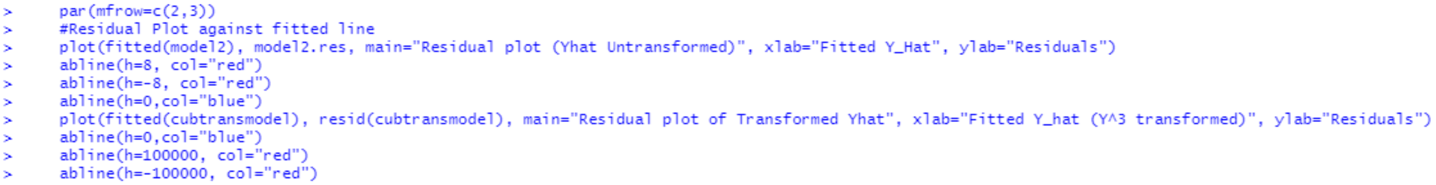
```
>     par(mfrow=c(2,2))
>     # Untransformed Model
>     qqnorm(resid(model2),main="QQ plot of Untransformed Model")
>     qqline(resid(model2), col="red")
>     #Quadratic Y Transformed Model
>     sqrdLE = (countries80$LifeExpectancy)^2
>     transmodel = lm(sqrdLE ~ countries80$Health + countries80$BirthRate + countries80$GDP + countries80$Cell)
>     qqnorm(resid(transmodel), main="QQ Plot of Quadratic Y-transformed Model")
>     qqline(resid(transmodel), col="red")
>     # Log Y Transformed Model
>     logLE = log(countries80$LifeExpectancy)
>     logtransmodel = lm(logLE ~ countries80$Health + countries80$BirthRate + countries80$GDP + countries80$Cell)
>     qqnorm(resid(logtransmodel), main="Normal Probability Plot of log Y-transformed Model")
>     qqline(resid(logtransmodel), col="red")
>     #Cubic Y Transformed Model
>     cubLE = (countries80$LifeExpectancy)^3
>     cubtransmodel = lm(cubLE ~ countries80$Health + countries80$BirthRate + countries80$GDP + countries80$Cell)
>     qqnorm(resid(cubtransmodel), main="Normal Probability Plot of Cubic Y-transformed Model")
>     qqline(resid(cubtransmodel), col="red")
```

Figure 2.5.1: Comparison of QQ plots of Untransformed model and Cubic-Transformed model.



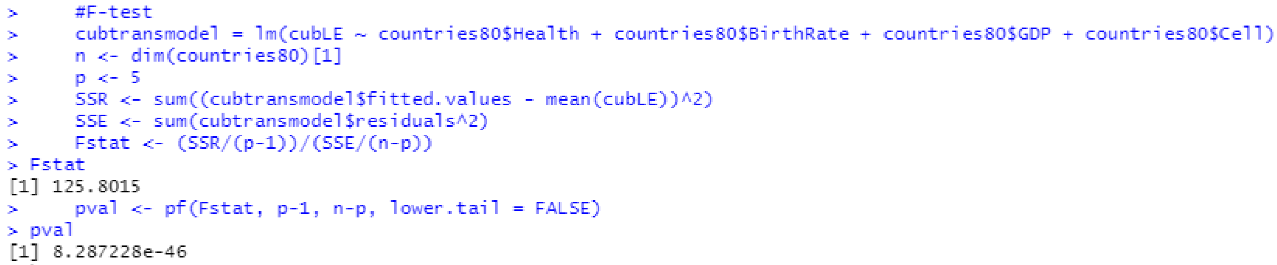
```
>     par(mfrow=c(2,3))
>     # Untransformed Model
>     qqnorm(resid(model2),main="QQ plot of Untransformed Model")
>     qqline(resid(model2), col="red")
>     #Cubic Y Transformed Model
>     cubLE = (countries80$LifeExpectancy)^3
>     cubtransmodel = lm(cubLE ~ countries80$Health + countries80$BirthRate + countries80$GDP + countries80$Cell)
>     qqnorm(resid(cubtransmodel), main="QQ Plot of Cubic Y-transformed Model")
>     qqline(resid(cubtransmodel), col="red")
```

Figure 2.5.2: Comparison of Residual plots (Residual against Y_Hat) of Untransformed model and Cubic-Transformed model.



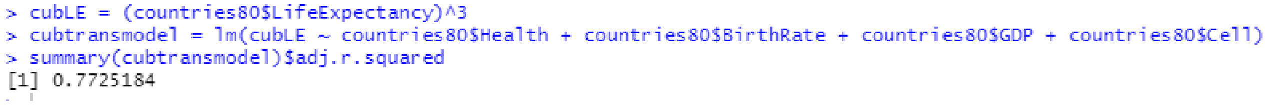
```
>     par(mfrow=c(2,3))
>     #Residual Plot against fitted line
>     plot(fitted(model2), model2.res, main="Residual plot (Yhat Untransformed)", xlab="Fitted Y_Hat", ylab="Residuals")
>     abline(h=8, col="red")
>     abline(h=-8, col="red")
>     abline(h=0,col="blue")
>     plot(fitted(cubtransmodel), resid(cubtransmodel), main="Residual plot of Transformed Yhat", xlab="Fitted Y_hat (Y^3 transformed)", ylab="Residuals")
>     abline(h=0,col="blue")
>     abline(h=100000, col="red")
>     abline(h=-100000, col="red")
```

Figure 3.1: F-test to test appropriateness of model



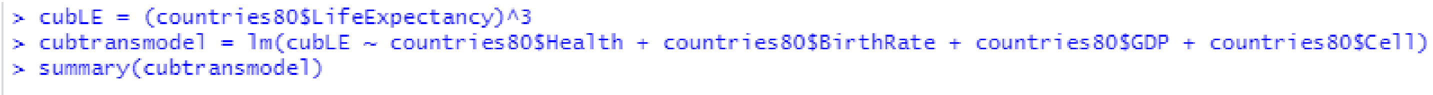
```
>     #F-test
>     cubtransmodel = lm(cubLE ~ countries80$Health + countries80$BirthRate + countries80$GDP + countries80$Cell)
>     n <- dim(countries80)[1]
>     p <- 5
>     SSR <- sum((cubtransmodel$fitted.values - mean(cubLE))^2)
>     SSE <- sum(cubtransmodel$residuals^2)
>     Fstat <- (SSR/(p-1))/(SSE/(n-p))
> Fstat
[1] 125.8015
>     pval <- pf(Fstat, p-1, n-p, lower.tail = FALSE)
> pval
[1] 8.287228e-46
```

Figure 3.2: R^2_{adj} of Final model



```
> cubLE = (countries80$LifeExpectancy)^3
> cubtransmodel = lm(cubLE ~ countries80$Health + countries80$BirthRate + countries80$GDP + countries80$Cell)
> summary(cubtransmodel)$adj.r.squared
[1] 0.7725184
```

Figure 3.3: Checking significance of predictor variables in Final model.



```
> cubLE = (countries80$LifeExpectancy)^3
> cubtransmodel = lm(cubLE ~ countries80$Health + countries80$BirthRate + countries80$GDP + countries80$Cell)
> summary(cubtransmodel)
```