Life Expectancy Multiple Regression

Kasia Krueger

9/14/2021

Week 2

Project Overview

This analysis is concerned with the life expectancy in years based on data collected yearly by the WHO and aggregated in Kaggle.com. The data set is from 2015 and contains 173 observations (countries).

Response Variable: Life expectancy (in years)

Quantitative Predictor Variables:
1. Adult Mortality Index - number of adult deaths for both sexes between the ages 15-60, per 1000 people
2. Body Mass Index - BMI - Average Body Mass Index of entire population
3. GDP - Gross Domestic Product per capita (in USD)
4. Schooling - Number of years of school
5. Infant death index - Number of Infant Deaths per 1000 people

Categorical Predictor Variables:

1. Country status:

- Developed
- Developing

2. Polio vaccine rates among 1-year-olds (%):

- Low (4.91 - 36.3]
- Medium (36.3 - 67.7]
- High (67.7 - 99.1]

3. Hepatitis B vaccine rates among 1-year-olds (%):

- Low (0 - 33]

- Medium (33 - 66]
- High (66 - 99.1]

4. HIV/AIDS deaths per 1,000 live births HIV/AIDS (0-4 years):

- Low (0.0908 - 3.17]
- Medium (3.17 - 6.23]
- High (6.23 - 9.31]

Summary

```r
summary(lifedata)
```

```
##    Country              Year          Status          Life.expectancy
##  Length:173         Min.   :2015   Length:173         Min.   :51.00  
##  Class :character   1st Qu.:2015   Class :character   1st Qu.:66.00  
##  Mode  :character   Median :2015   Mode  :character   Median :73.90  
##                     Mean   :2015                      Mean   :71.71  
##                     3rd Qu.:2015                      3rd Qu.:76.70  
##                     Max.   :2015                      Max.   :88.00  
##  Adult.Mortality infant.deaths     Hepatitis.B       Measles      
##  Min.   :  1.0   Min.   :  0.00   Min.   : 0.00   Min.   :    0  
##  1st Qu.: 74.0   1st Qu.:  0.00   1st Qu.:75.00   1st Qu.:    0  
##  Median :138.0   Median :  2.00   Median :92.00   Median :   16  
##  Mean   :151.4   Mean   : 22.51   Mean   :78.42   Mean   : 1515  
##  3rd Qu.:211.0   3rd Qu.: 15.00   3rd Qu.:97.00   3rd Qu.:  206  
##  Max.   :484.0   Max.   :910.00   Max.   :99.00   Max.   :90387  
##       BMI        under.five.deaths     Polio         Diphtheria      HIV.AIDS  
##  Min.   : 0.00   Min.   :   0.00   Min.   : 5.00   Min.   : 6.00   low   :160  
##  1st Qu.:23.80   1st Qu.:   0.00   1st Qu.:83.00   1st Qu.:84.00   medium: 11  
##  Median :48.60   Median :   3.00   Median :93.00   Median :93.00   high  :  2  
##  Mean   :42.32   Mean   :  29.82   Mean   :82.97   Mean   :84.48               
##  3rd Qu.:61.30   3rd Qu.:  20.00   3rd Qu.:97.00   3rd Qu.:97.00               
##  Max.   :77.60   Max.   :1100.00   Max.   :99.00   Max.   :99.00               
##       GDP            Population        Income.composition.of.resources
##  Min.   :   33.68   Length:173         Min.   :0.3470                 
##  1st Qu.:  814.55   Class :character   1st Qu.:0.5650                 
##  Median : 2954.12   Mode  :character   Median :0.7230                 
##  Mean   : 7037.71                      Mean   :0.6917                 
##  3rd Qu.: 6993.48                      3rd Qu.:0.7980                 
##  Max.   :66346.52                      Max.   :0.9480                 
##    Schooling      polio.vacc    hepB.vacc  
##  Min.   : 4.90   low   : 13   low   : 21  
##  1st Qu.:10.80   medium: 15   medium: 14  
##  Median :13.10   high  :145   high  :138  
##  Mean   :12.93                            
##  3rd Qu.:15.00                            
##  Max.   :20.40
```

1. Create a Design matrix

Give the column of 1's the name 'Intercept'

```
X <- with(lifedata, cbind(1, Adult.Mortality, BMI, GDP, Schooling, infant.deaths))
Y <- as.matrix(lifedata$Life.expectancy)
colnames(X)<-c('Intercept','Adult.Mortality','BMI','GDP','Years in School','Infant.Mortality')
```

2. Calculate Beta-hat

Using matrix algebra

```
beta.hat <- solve(crossprod(X), crossprod(X,Y))
beta.hat
```

```
##                           [,1]
## Intercept         5.765167e+01
## Adult.Mortality  -3.600419e-02
## BMI               7.190785e-03
## GDP               4.123093e-05
## Years in School   1.467075e+00
## Infant.Mortality -2.037792e-03
```

Using lm

```
life_lm <- lm(Life.expectancy ~ Adult.Mortality+BMI+GDP+Schooling+infant.deaths, lifedata)
coef(life_lm)
```

```
##     (Intercept) Adult.Mortality             BMI             GDP       Schooling
##    5.765167e+01   -3.600419e-02    7.190785e-03    4.123093e-05    1.467075e+00
##   infant.deaths
##   -2.037792e-03
```

```
summary(life_lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + BMI + GDP +
##     Schooling + infant.deaths, data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.169  -2.103   0.221   2.136   8.746
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.765e+01  1.787e+00  32.258   <2e-16 ***
## Adult.Mortality -3.600e-02  3.346e-03 -10.760   <2e-16 ***
## BMI              7.191e-03  1.574e-02   0.457    0.648
## GDP              4.123e-05  2.666e-05   1.547    0.124
## Schooling        1.467e+00  1.299e-01  11.290   <2e-16 ***
## infant.deaths   -2.038e-03  3.249e-03  -0.627    0.531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.485 on 167 degrees of freedom
## Multiple R-squared:  0.8126, Adjusted R-squared:  0.807
## F-statistic: 144.9 on 5 and 167 DF,  p-value: < 2.2e-16
```

3. Calculate the vectors of fitted values and residuals

Using matrix algebra

```
Yhat <- X %*% beta.hat # vector of fitted values
res <- Y - Yhat # vector of residuals
```

4. Calculate the residual degrees of freedom, SSE, and sigma^2

Using matrix algebra

```
SSE <- crossprod(res)
df <- nrow(X) - ncol(X) # this is n-(p+1), residual degrees of freedom
sigma.sq <- SSE/df # sigma^2
```

Using lm

```
SSE.lm <- deviance(life_lm)
df.lm <- df.residual(life_lm)
sigma.sq.lm <- deviance(life_lm) / df.residual(life_lm)
```

Values	Matrix Algebra	lm function
SSE	2028.3439019	2028.3439019
df	167	167
sigma.sq	12.1457719	12.1457719

5. Calculate R = Cor(Y,Y-hat), R2, and R2adj

```
n <- length(lifedata$Life.expectancy)
SST <- (n-1)*var(lifedata$Life.expectancy)
R <- cor(lifedata$Life.expectancy, fitted(life_lm))
Rsquared <- R^2
Rsq <- 1 - deviance(life_lm) / SST
r.squared.adj <- 1-((n-1)/df)*(1-R^2)
```

Value	Matrix Algebra	summary(life_lm)
R (Cor(Y,Y-hat))	0.9014639	0.9014639
R^2	0.8126371	0.8126371
$\mathrm{R}^2{}_{\mathrm{adj}}$	0.8070275	0.8070275

Reduced model	Adjusted R^2
life_lm, ~ Adult.Mortality	0.5748119
life_lm, ~ Adult.Mortality+Schooling	0.8066945
life_lm, ~ Adult.Mortality+GDP+Schooling	0.8085475
life_lm, ~ Adult.Mortality+GDP+Schooling+infant.deaths	0.8079365

After removing BMI from the model, when we add infant.deaths to this model (already containing the predictors Adult.Mortality+GDP+Schooling), $\mathrm{R}^2{}_{\mathrm{adj}}$ ended up decreasing by a bit. We could take this as a sign that we may be starting to overfit the model by including infant.deaths as a predictor.

6. Calculate the F-statistic in the test of overall regression

$(H0 : Beta1 = Beta2 = Beta3 = Betap = 0againstHa : notH0)$ and its associated $p - value$

```
life0 <- update(life_lm, ~ 1) # reduced model with no predictors
SSE.full <- deviance(life_lm)
SSE.reduced <- deviance(life0)
df.full <- df.residual(life_lm)
df.reduced <- df.residual(life0)
F.stat <- ((SSE.reduced - SSE.full) / (df.reduced-df.full)) / (SSE.full/df.full)
p.value <- pf(F.stat, df.reduced-df.full, df.full, lower.tail=FALSE)
```

F.stat	p.value
144.8637142	$8.0216452 \times 10^{-59}$

With such a small p-value, we will soundly reject H0 (the reduced model is adequate) in favor of Ha (the full model is adequate). ## Confirming with anova and lm:

```
anova(life0, life_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ 1
## Model 2: Life.expectancy ~ Adult.Mortality + BMI + GDP + Schooling + infant.deaths
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    172 10825.8
## 2    167  2028.3  5    8797.4 144.86 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(life_lm)$fstatistic
```

```
##    value    numdf    dendf 
## 144.8637   5.0000 167.0000
```

7. Calculate the covariance matrix Cov(Beta-hat) and the standard errors of each of the regression coefficients Beta-hat$_j$

```r
beta.Cov.unscaled <- summary(life_lm)$cov.unscaled
beta.Cov.unscaled
```

```
##                   (Intercept) Adult.Mortality           BMI           GDP
## (Intercept)      2.629877e-01   -3.326843e-04  9.558415e-05  4.954280e-07
## Adult.Mortality -3.326843e-04    9.218057e-07  1.830780e-07  8.957510e-10
## BMI              9.558415e-05    1.830780e-07  2.040806e-05 -3.719828e-09
## GDP              4.954280e-07    8.957510e-10 -3.719828e-09  5.850916e-11
## Schooling       -1.647551e-02    1.393716e-05 -7.507916e-05 -6.861886e-08
## infant.deaths   -6.102621e-05   -5.054191e-08  4.381332e-07  7.092138e-11
##                     Schooling infant.deaths
## (Intercept)     -1.647551e-02 -6.102621e-05
## Adult.Mortality  1.393716e-05 -5.054191e-08
## BMI             -7.507916e-05  4.381332e-07
## GDP             -6.861886e-08  7.092138e-11
## Schooling        1.390275e-03  2.326612e-06
## infant.deaths    2.326612e-06  8.692332e-07
```

```r
sigma.hat <- summary(life_lm)$sigma
sigma.hat
```

```
## [1] 3.485078
```

```r
var.beta <- sigma.hat^2 * diag(beta.Cov.unscaled)
var.beta
```

```
##     (Intercept) Adult.Mortality             BMI             GDP       Schooling 
##    3.194189e+00    1.119604e-05    2.478717e-04    7.106389e-10    1.688597e-02 
##   infant.deaths 
##    1.055751e-05
```

```r
se.beta <- sqrt(var.beta)
se.beta
```

```
##     (Intercept) Adult.Mortality             BMI             GDP       Schooling 
##    1.787229e+00    3.346049e-03    1.574394e-02    2.665781e-05    1.299460e-01 
##   infant.deaths 
##    3.249232e-03
```

8. Calculate the value of the t-statistic and the p-value for each of the p+1 tests H0:Beta$_j$=0 against Ha:Beta$_j$Beta0, j=0,1,.,p

```
df <- df.residual(life_lm)
t.quantile <- qt(0.95, df) # need 95th percentile for 90% confidence interval
se.beta1 <- se.beta[2] # first entry in se.beta corresponds to beta_0, not beta_1
beta1.hat <- coef(life_lm)[2] # point estimate of beta_1
beta1.hat + c(-1,1) * t.quantile * se.beta1
```

```
## [1] -0.04153865 -0.03046973
```

Confirming t-statistic interval of B$_1$ and each of the p+1 tests with lm:

```
confint(life_lm, level=0.90)
```

```
##                             5 %          95 %
## (Intercept)      5.469555e+01  6.060780e+01
## Adult.Mortality -4.153865e-02 -3.046973e-02
## BMI             -1.885015e-02  3.323172e-02
## GDP             -2.861863e-06  8.532373e-05
## Schooling        1.252140e+00  1.682009e+00
## infant.deaths   -7.412116e-03  3.336533e-03
```

BMI, GDP, and Infant deaths coefficients are not significant.

9. Create a matrix of five "extra observations" for prediction purposes, and calculate the predicted value of the response for these five observations, along with 95% prediction intervals

Life expectancy for a country with 5/1000 adult mortality deaths, average BMI of 20, $100,000 USD GDP per capita, 18 years of schooling and 0 infant deaths:

```
##       fit      lwr      upr
## 1 88.1459 79.71722 96.57459
```

Life expectancy for a country with 5/1000 adult mortality deaths, average BMI of 30, $100,000 USD GDP per capita, 18 years of schooling and 0 infant deaths:

```
##        1
## 88.21781
```

Life expectancy for a country with 5/1000 adult mortality deaths, average BMI of 30, $50,000 USD GDP per capita, 18 years of schooling and 0 infant deaths:

```
##        1
## 86.15627
```

Life expectancy for a country with 5/1000 adult mortality deaths, average BMI of 30, $50,000 USD GDP per capita, 12 years of schooling and 0 infant deaths:

```
##        1
## 77.35382
```

Life expectancy for a country with 5/1000 adult mortality deaths, average BMI of 30, $50,000 USD GDP per capita, 12 years of schooling and 5/1000 infant deaths:

```
##        1
## 77.34363
```

10. Perform a hypothesis test to test if a subset of two + of the coefficientsare equal to 0

H_0:BMI=GDP=0, H_A: not H_0

```
life.reduced <- update(life_lm, ~ BMI + GDP)
anova(life.reduced, life_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ BMI + GDP
## Model 2: Life.expectancy ~ Adult.Mortality + BMI + GDP + Schooling + infant.deaths
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    170 6965.0
## 2    167 2028.3  3    4936.7 135.48 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0:Adult.Mortality=Schooling=infant.deaths=0, H_A: not H_0

```
life.reduced1 <- update(life_lm, ~ Adult.Mortality + Schooling +infant.deaths)
anova(life.reduced1, life_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ Adult.Mortality + Schooling + infant.deaths
## Model 2: Life.expectancy ~ Adult.Mortality + BMI + GDP + Schooling + infant.deaths
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    169 2062.2
## 2    167 2028.3  2    33.828 1.3926 0.2513
```

```
summary(life.reduced1)$r.squared
```

```
## [1] 0.8095124
```

```
summary(life_lm)$r.squared
```

```
## [1] 0.8126371
```

```r
summary(update(life_lm, ~ GDP+BMI))$r.squared
```

```
## [1] 0.356627
```

11 Perform a hypothesis test where the reduced model has one or more of the regression coefficients equal to some specific value ### H_0:Schooling=4 against H_a:Schooling not equal 4

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.3
```

```r
library(carData)  # for linearhypothesis test
life.full <- lm(Y ~ Adult.Mortality+Schooling+BMI, lifedata)
linearHypothesis(life.full, c(0, 0, 1, 0), 4)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Schooling = 4
##
## Model 1: restricted model
## Model 2: Y ~ Adult.Mortality + Schooling + BMI
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    170 6777.5
## 2    169 2062.4  1    4715.1 386.37 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12 Perform a hypothesis test where some constraint is imposed on two or more of the predictors

H_0:Adult.Mortality+0.5*Schooling, H_A: not H_0

```r
life.full <- lm(Y ~ Adult.Mortality+Schooling+ BMI +GDP +infant.deaths, lifedata)

life.reduced1 <- lm(Y ~ I(Adult.Mortality+0.5*Schooling) - BMI - GDP - infant.deaths, lifedata)

anova(life.reduced1, life.full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ I(Adult.Mortality + 0.5 * Schooling) - BMI - GDP - infant.deaths
## Model 2: Y ~ Adult.Mortality + Schooling + BMI + GDP + infant.deaths
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    171 4678.4
## 2    167 2028.3  4    2650.1 54.547 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Report

1. Response and Predictor Variables

The response variable for this data set is the adult mortality in years. This data set consists of five numeric predictor variables, Adult mortality, infant deaths, BMI, GDP per capita (in U.S. dollars), and number of years of schooling.
There are 4 qualitative predictor variables: status (developed, developing), polio inoculation rate (low, medium, high), hepatitis B inoculation rate (low, medium high), and HIV/AIDS illness rate (low, medium, high) that are not included in this week's regression analysis.

I expect BMI to have a significant effect on life expectancy.

2. Estimated Regression Equation

```
##                            [,1]
## Intercept          5.765167e+01
## Adult.Mortality   -3.600419e-02
## BMI                7.190785e-03
## GDP                4.123093e-05
## Years in School    1.467075e+00
## Infant.Mortality  -2.037792e-03
```

The estimated regression equation is:
Life expectancy = 57.65 years - 0.036 adult mortality index 0.0072 BMI + 0.00004 GDP per capita + 1.47 years of schooling - 0.00204 infant mortality index.

The intercept of 57.65 years represents the life expectancy if all the predictor variables were zero.

3. Table of the T-statistics, Standard Errors, and P-values

```
summary(life_lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + BMI + GDP +
##     Schooling + infant.deaths, data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.169  -2.103   0.221   2.136   8.746
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.765e+01  1.787e+00  32.258   <2e-16 ***
## Adult.Mortality -3.600e-02  3.346e-03 -10.760   <2e-16 ***
## BMI              7.191e-03  1.574e-02   0.457    0.648
## GDP              4.123e-05  2.666e-05   1.547    0.124
## Schooling        1.467e+00  1.299e-01  11.290   <2e-16 ***
## infant.deaths   -2.038e-03  3.249e-03  -0.627    0.531
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.485 on 167 degrees of freedom
## Multiple R-squared:  0.8126, Adjusted R-squared:  0.807
## F-statistic: 144.9 on 5 and 167 DF,  p-value: < 2.2e-16
```

```
confint(life_lm, level=0.90)
```

```
##                           5 %          95 %
## (Intercept)      5.469555e+01  6.060780e+01
## Adult.Mortality -4.153865e-02 -3.046973e-02
## BMI             -1.885015e-02  3.323172e-02
## GDP             -2.861863e-06  8.532373e-05
## Schooling        1.252140e+00  1.682009e+00
## infant.deaths   -7.412116e-03  3.336533e-03
```

The f-test returned a value of 144.8637142, and a p-value of *8.0216452 x 10*$^{-59}$, indicating at least one of the predictor variables influences the response variable.

4. Interpretation of Each of the Significant Regression Coefficients

```
summary(life_lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + BMI + GDP +
##     Schooling + infant.deaths, data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.169  -2.103   0.221   2.136   8.746
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.765e+01  1.787e+00  32.258   <2e-16 ***
## Adult.Mortality -3.600e-02  3.346e-03 -10.760   <2e-16 ***
## BMI              7.191e-03  1.574e-02   0.457    0.648
## GDP              4.123e-05  2.666e-05   1.547    0.124
## Schooling        1.467e+00  1.299e-01  11.290   <2e-16 ***
## infant.deaths   -2.038e-03  3.249e-03  -0.627    0.531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.485 on 167 degrees of freedom
## Multiple R-squared:  0.8126, Adjusted R-squared:  0.807
## F-statistic: 144.9 on 5 and 167 DF,  p-value: < 2.2e-16
```

BMI, GDP, and infant deaths are not significant.

Intercept: 57.65. If all predictor variables are zero, the predicted life expectancy is 57.65 years.
Adult mortality index: - 0.036. For every 1/1000 adult deaths (ages 15-60), the life expectancy decreases by 0.036 years.
Years of schooling: 1.47. For every one year increase in schooling, life expectancy increases by 1.47 years.

5. F-statistic and Associated P-value in the Test of Overall Regression, Sigma.sq

F.stat	p.value
144.8637142	$8.0216452 \times 10^{-59}$

With such a small p-value, we will soundly reject H0 (the reduced model is adequate) in favor of Ha (the full model is adequate).

6. Descriptive Interpretation of Sigma.sq and R.sq

Sigma squared quantifies how much the responses (y) vary around the regression line. Sigma squared is measured at 12.15, indicating the predicted life expectancy differs from the actual life expectancy by 12.15 years.

R-squared measures how much of the variability in the model is explained by the dependent variables. This linear model has a R^2 value of 81.26%, indicating that 81% of the variability in the model can be explained by the dependent variables.

7. Other Hypothesis Tests

Hypothesis test of a subset of two or more predictors equal to 0

H_0:BMI=GDP=0, H_A: not H_0

```
life.reduced <- update(life_lm, ~ BMI + GDP)
anova(life.reduced, life_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ BMI + GDP
## Model 2: Life.expectancy ~ Adult.Mortality + BMI + GDP + Schooling + infant.deaths
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    170 6965.0
## 2    167 2028.3  3    4936.7 135.48 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for this hypothesis test is quite small, suggesting we should reject H0 and use the full model or at least, not this reduced model.

H_0:Adult.Mortality=Schooling=infant.deaths=0, H_A: not H~0

```
life.reduced1 <- update(life_lm, ~ Adult.Mortality + Schooling +infant.deaths)
anova(life.reduced1, life_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ Adult.Mortality + Schooling + infant.deaths
## Model 2: Life.expectancy ~ Adult.Mortality + BMI + GDP + Schooling + infant.deaths
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    169 2062.2
## 2    167 2028.3  2    33.828 1.3926 0.2513
```

The p-value for this hypothesis test is quite large (0.2513), suggesting we should retain H0 and stay with the reduced model

```
summary(life.reduced1)$r.squared
```

```
## [1] 0.8095124
```

81% of the variability in the response is explained by the predictors Adult.Mortality + Schooling+infant.deaths

```
summary(life_lm)$r.squared
```

```
## [1] 0.8126371
```

81.3% of the variability in the response is explained by all 5 predictors, then we can view the difference of 0.4% as the percentage of the variability in the response that is explained by the predictors GDP+infant.deaths when adjusting for Adult.Mortality + Schooling+infant.deaths.This illustrates that the predictors Adult.Mortality + Schooling+infant.deaths are doing the bulk of the work in reducing the unexplained variability in the response

```
summary(update(life_lm, ~ GDP+BMI))$r.squared
```

```
## [1] 0.356627
```

About 35.66% of the observed variability in the response is jointly explained by the predictors GDP+BMI (this number comes from a model where Adult.Mortality + Schooling+infant.deaths do not enter as predictors) only 0.4% of the observed variability in the response is jointly explained by the predictors GDP+BMI after adjusting for the effects of Adult.Mortality + Schooling+infant.deaths. That is, the set (GDP+BMI) is better than nothing when it comes to predicting Y, when we are already predicting Y with Adult.Mortality + Schooling+infant.deaths then these extra 2 predictors provide very little

Hypothesis test of a predictor (or several) are equal to a specific non-zero value

H_0:Schooling=4 against H_a:Schooling not equal 4

```r
life.full <- lm(Y ~ Adult.Mortality+Schooling+BMI, lifedata)
linearHypothesis(life.full, c(0, 0, 1, 0), 4)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Schooling = 4
##
## Model 1: restricted model
## Model 2: Y ~ Adult.Mortality + Schooling + BMI
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    170 6777.5
## 2    169 2062.4  1    4715.1 386.37 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The small p-value shows there is significant evidence to reject H_0 and conclude Schooling does not equal 4.

Hypothesis test of testing a linear constraint on a subset of predictors

H_0:Adult.Mortality+0.5*Schooling, H_A: not H_0

```r
life.full <- lm(Y ~ Adult.Mortality+Schooling+ BMI +GDP +infant.deaths, lifedata)

life.reduced1 <- lm(Y ~ I(Adult.Mortality+0.5*Schooling) - BMI - GDP - infant.deaths, lifedata)

anova(life.reduced1, life.full)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ I(Adult.Mortality + 0.5 * Schooling) - BMI - GDP - infant.deaths
## Model 2: Y ~ Adult.Mortality + Schooling + BMI + GDP + infant.deaths
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    171 4678.4
## 2    167 2028.3  4    2650.1 54.547 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The conclusion is to reject H0, since there is significant evidence of an observed difference, in terms of predicting the response, between the reduced and full models.

Life Expectancy - Regression Diagnostics

Kasia Krueger

9/22/2021

Week 3

Programming Assignment

```
life_lm <- lm(Life.expectancy ~ Adult.Mortality+BMI+GDP+Schooling+infant.deaths, lifedata)
```

1. Hat matrix (projection matrix) using matrix algebra

```
X <- model.matrix(life_lm) # obtain the design matrix
P <- X %*% solve(crossprod(X)) %*% t(X) #
leverage <- diag(P)
```

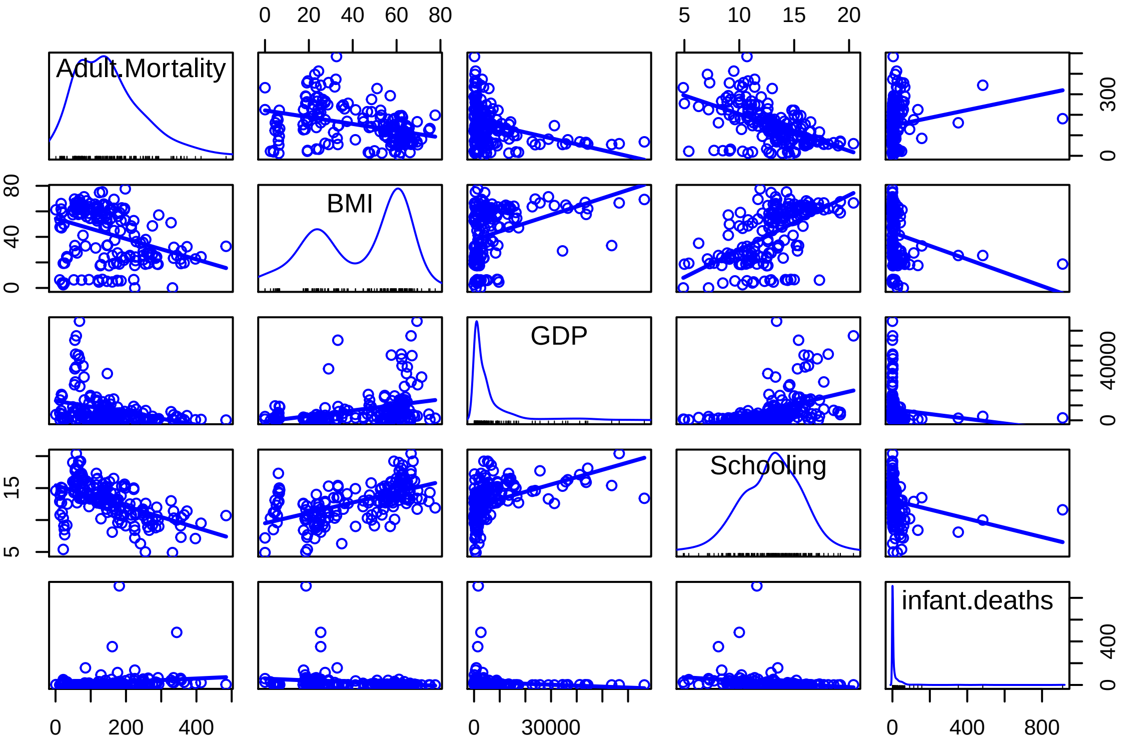
Countries with lowest and highest leverages:

```
sort(leverage)[c(1,2,172,173)]
```

```
## Saint Lucia  Uzbekistan       Qatar       India
## 0.006874469 0.008140590 0.204196350 0.671317316
```

2. Scatterplot matrix of response and five predictor variables

```
scatterplotMatrix(~Adult.Mortality+BMI+GDP+Schooling+infant.deaths, lifedata, smooth=FALSE)
```

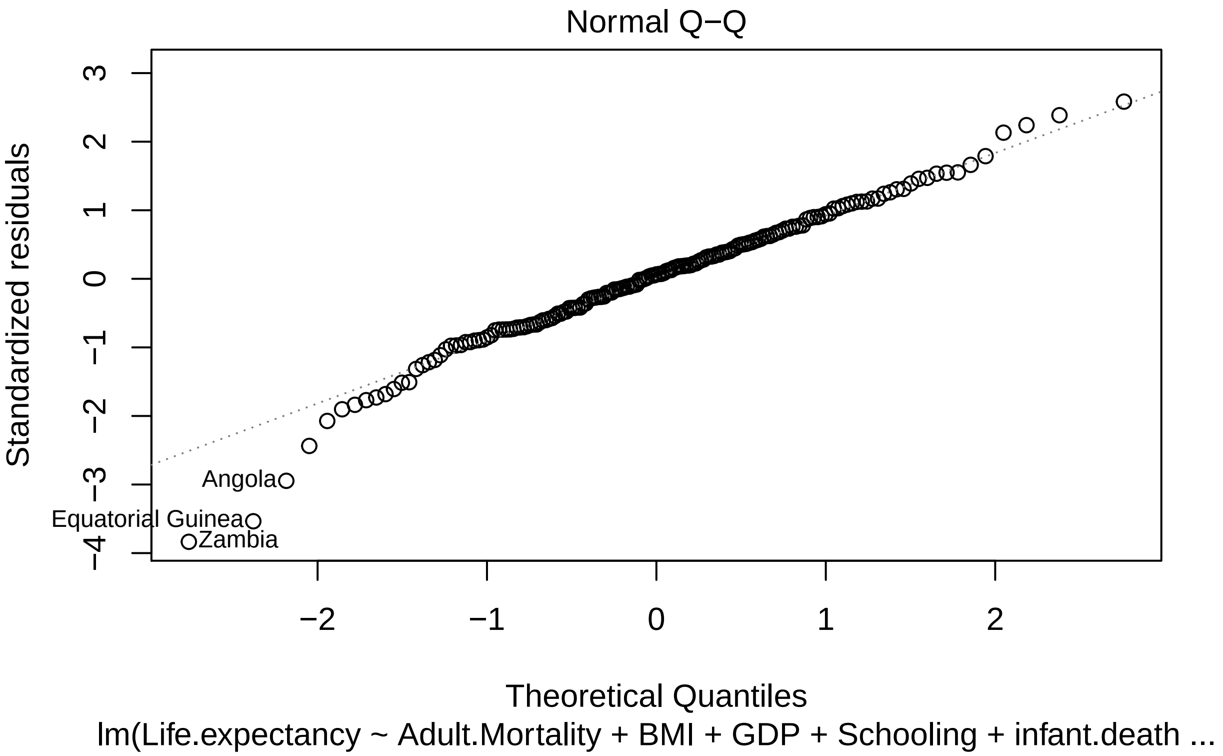


```r
summary(life_lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + BMI + GDP +
##     Schooling + infant.deaths, data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.169  -2.103   0.221   2.136   8.746
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.765e+01  1.787e+00  32.258   <2e-16 ***
## Adult.Mortality -3.600e-02  3.346e-03 -10.760   <2e-16 ***
## BMI              7.191e-03  1.574e-02   0.457    0.648
## GDP              4.123e-05  2.666e-05   1.547    0.124
## Schooling        1.467e+00  1.299e-01  11.290   <2e-16 ***
## infant.deaths   -2.038e-03  3.249e-03  -0.627    0.531
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.485 on 167 degrees of freedom
## Multiple R-squared:  0.8126, Adjusted R-squared:  0.807
## F-statistic: 144.9 on 5 and 167 DF,  p-value: < 2.2e-16
```

3. Normal q-q plot and Shapiro-Wilk test of the standardized residuals

```
res <- rstandard(life_lm) # standardized residuals
plot(life_lm, which=2)
```



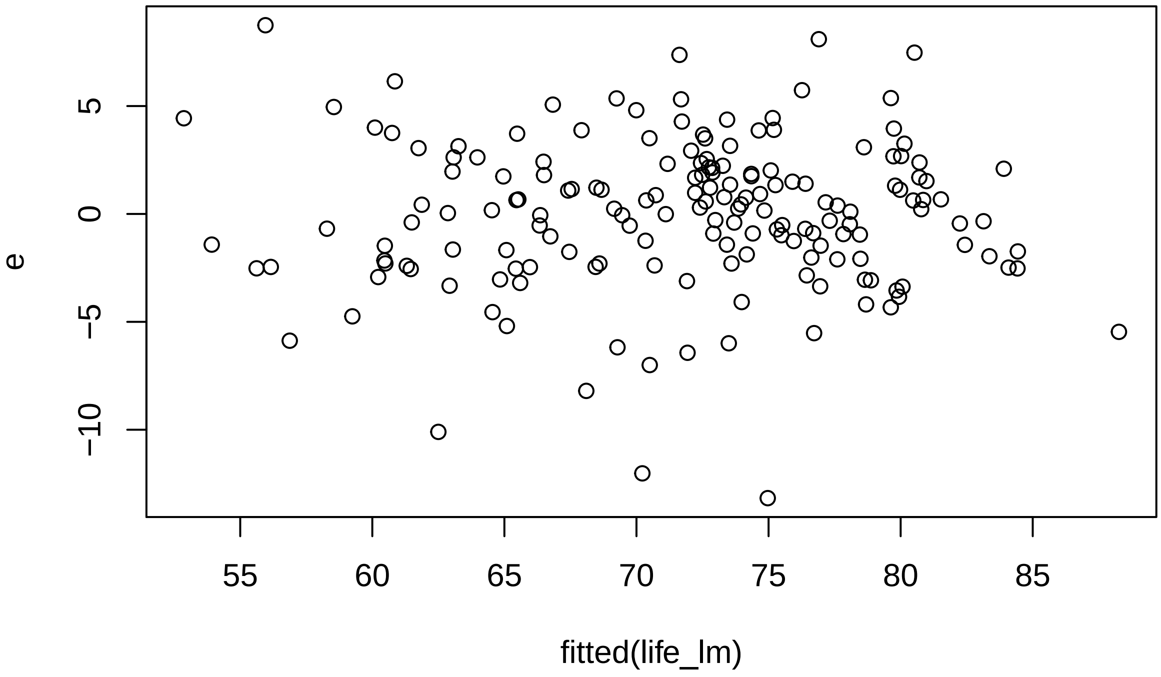
```
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97524, p-value = 0.003494
```

The plots are suggestive of the presence of some rather severe outliers, so that the normality assumption is not warranted. Also, the p-value in the Shapiro-Wilk test is only 0.003494; this small value indicates that the observed sample (of standardized residuals) is rather unlikely to occur by sheer chance if indeed the null hypothesis of a normal population is true.

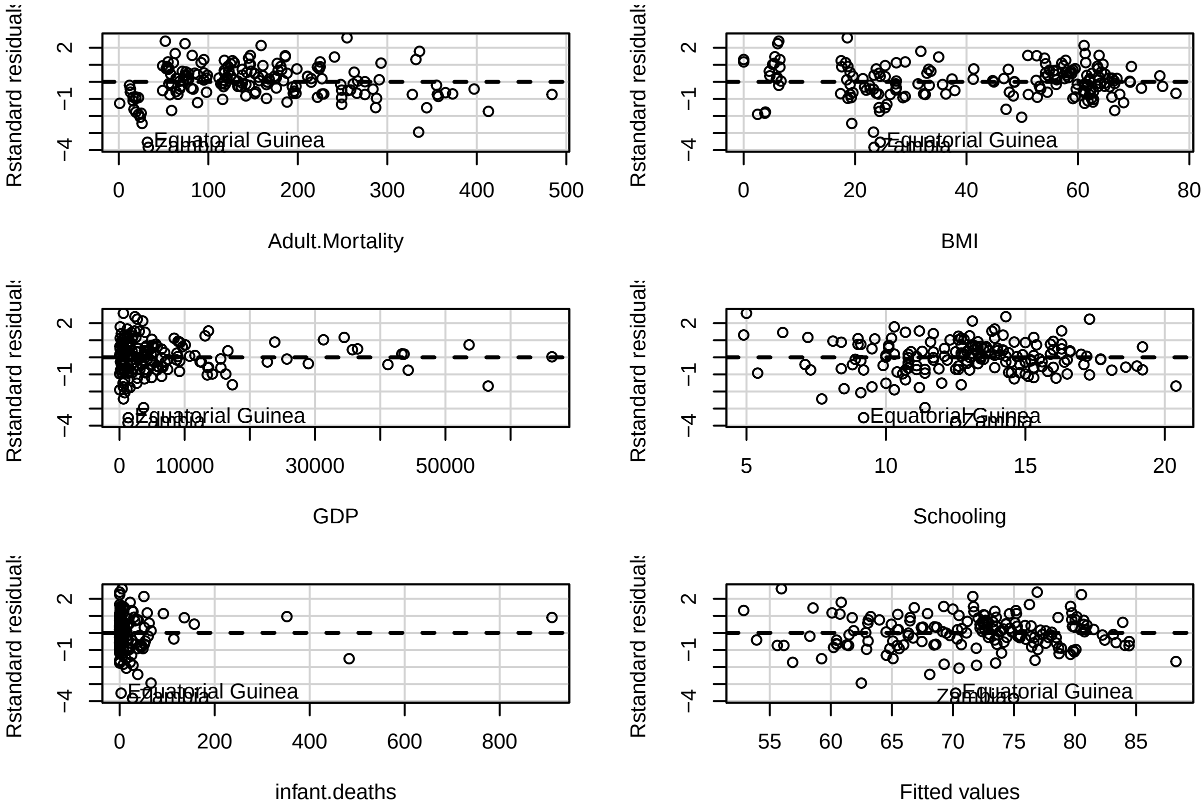
4. Plot of the standardized residuals against the fitted values

```
e <- resid(life_lm)
plot(e~fitted(life_lm))
```



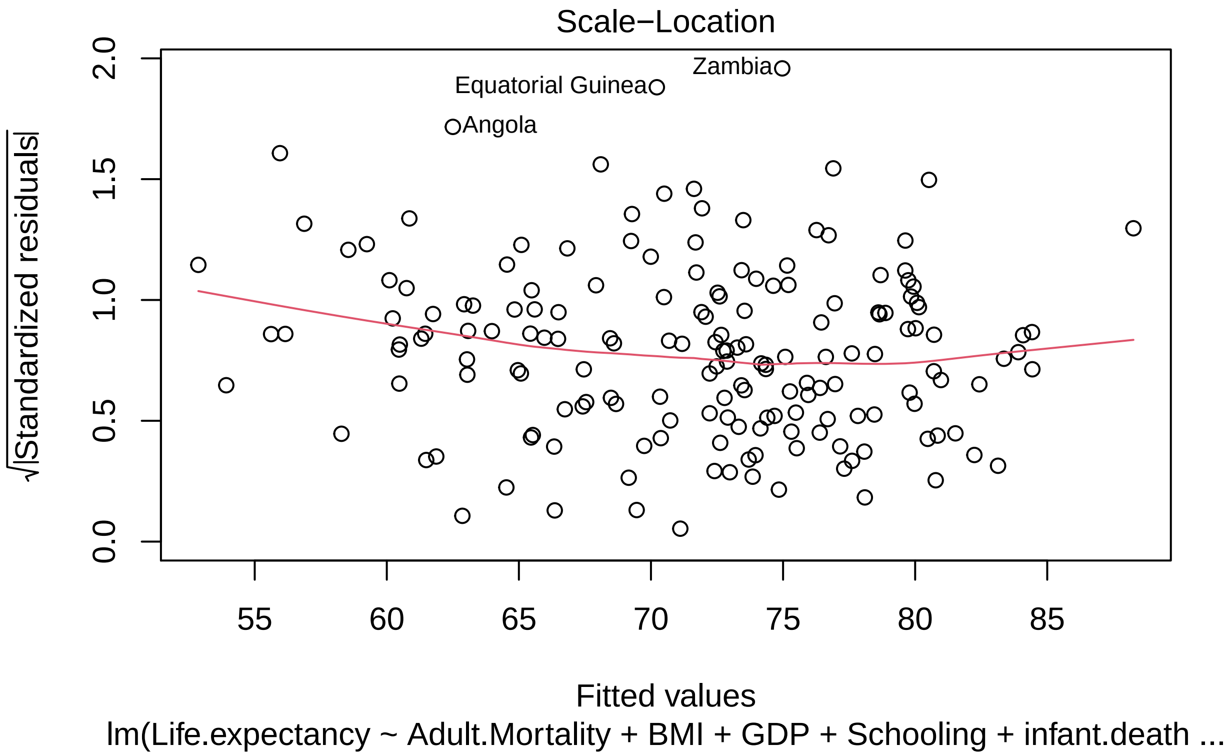
5. Plot of the standardized residuals against each of the predictor variables

```
residualPlots(life_lm, id=TRUE, quadratic=FALSE, type='rstandard', tests=FALSE)
```



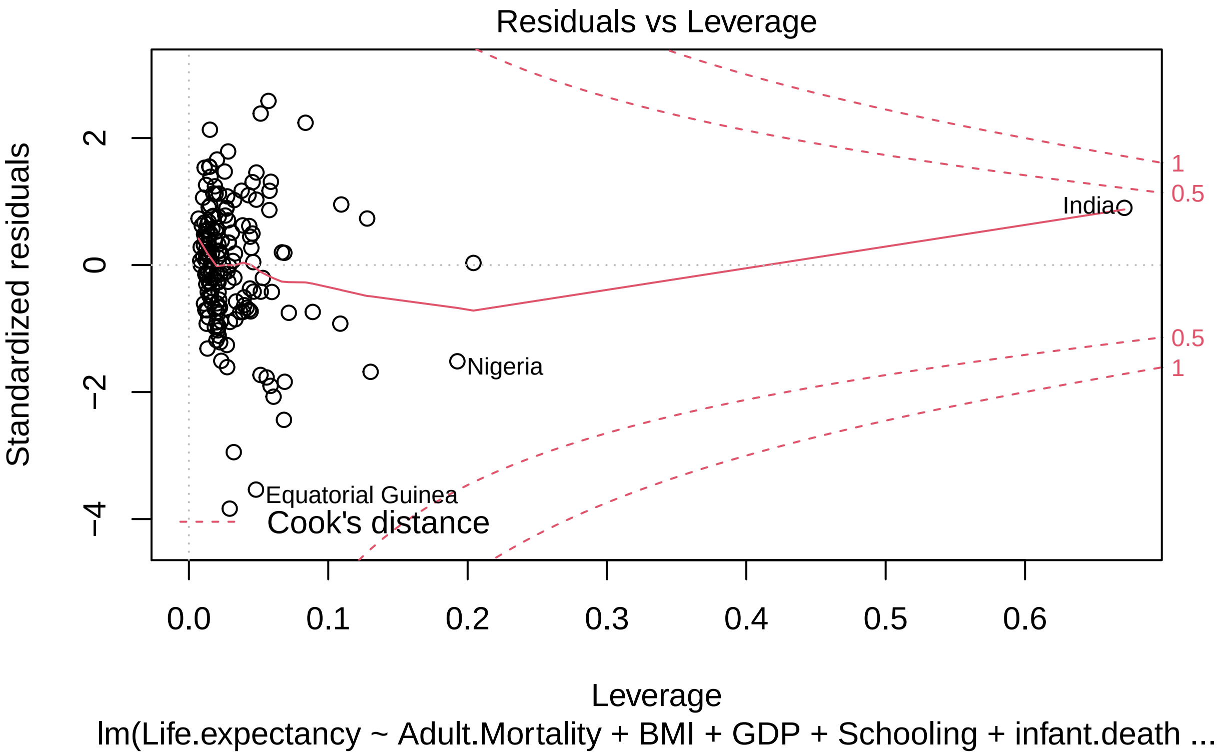
6. Location-spread plot

```
plot(life_lm, which=3)
```



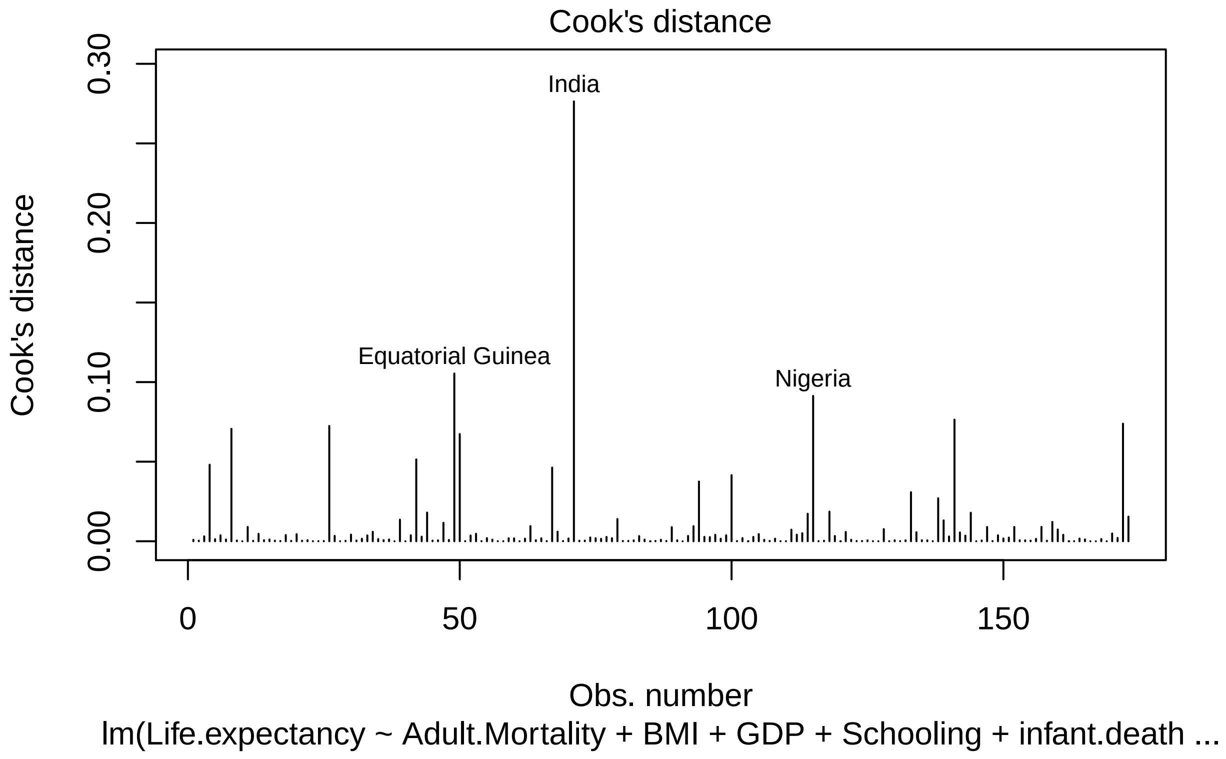
7.Residual-leverage plot

```
plot(life_lm, which=5)
```



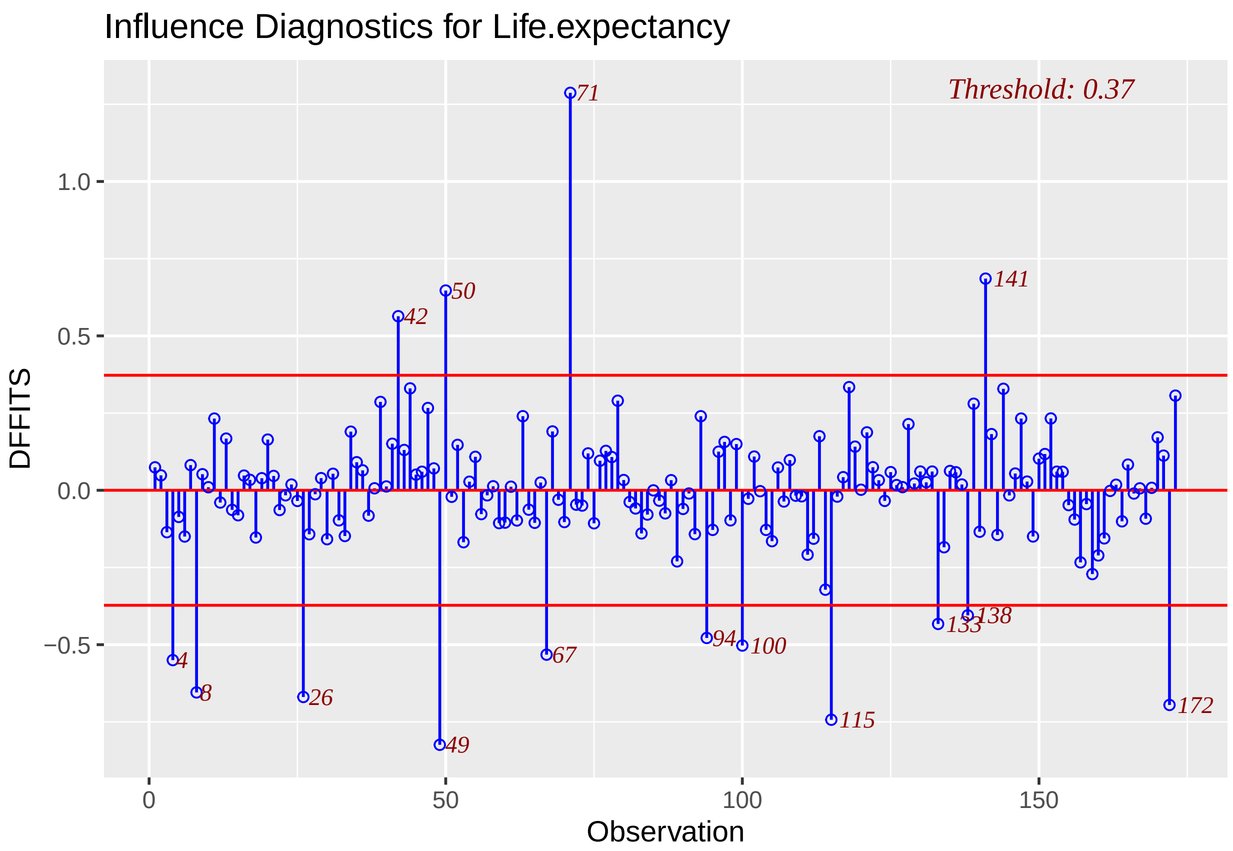
8. Index plot of cook's distance

```
plot(life_lm, which=4)
```



9. Index plot of DFFITS

```
ols_plot_dffits(life_lm)
```



```
lifedata[c(49, 71, 115, 141,172), ]
```

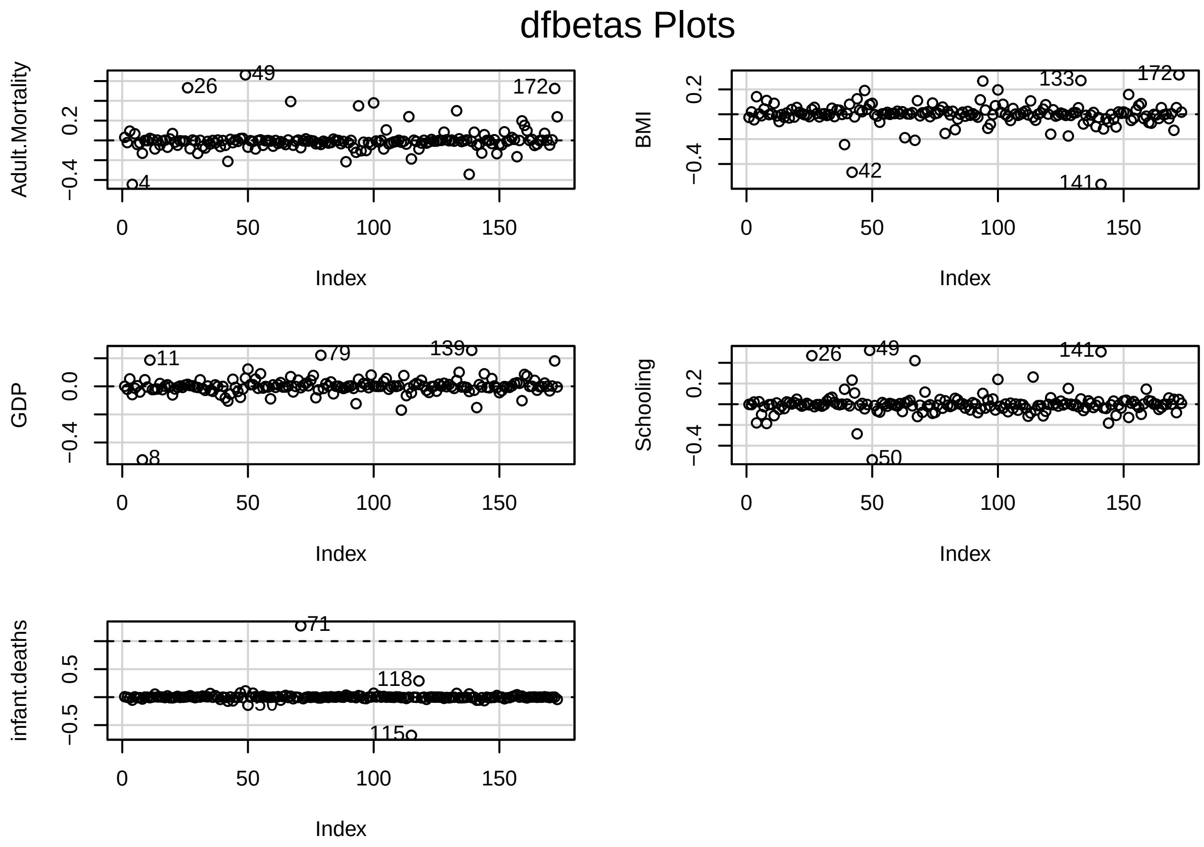
```
##                   Year     Status Life.expectancy Adult.Mortality infant.deaths
## Equatorial Guinea 2015 Developing            58.2              32             3
## India             2015 Developing            68.3             181           910
## Nigeria           2015 Developing            54.5             344           483
## Slovenia          2015  Developed            88.0              74             0
## Zambia            2015 Developing            61.8              33            27
##                   Hepatitis.B Measles  BMI under.five.deaths Polio Diphtheria
## Equatorial Guinea          16    1250 24.5                 4    17         16
## India                      87   90387 18.7              1100    86         87
## Nigeria                    49   12423 25.4               747    49         49
## Slovenia                    0      18  6.1                 0    95         95
## Zambia                      9       9 23.4                40     9          9
##                   HIV.AIDS     GDP Population Income.composition.of.resources
## Equatorial Guinea   medium 1347.31    1175389                           0.582
## India                  low 1613.19    1395398                           0.615
## Nigeria             medium 2655.16  181181744                           0.525
## Slovenia               low 2729.86     263531                           0.888
## Zambia              medium 1313.89     161587                           0.576
##                   Schooling polio.vacc hepB.vacc
## Equatorial Guinea       9.2        low       low
## India                  11.6       high      high
## Nigeria                10.0     medium    medium
## Slovenia               17.3       high       low
## Zambia                 12.5        low       low
```

10. Panel of index plots of DFBETAS

```r
dfbetas(life_lm)[71,]
```

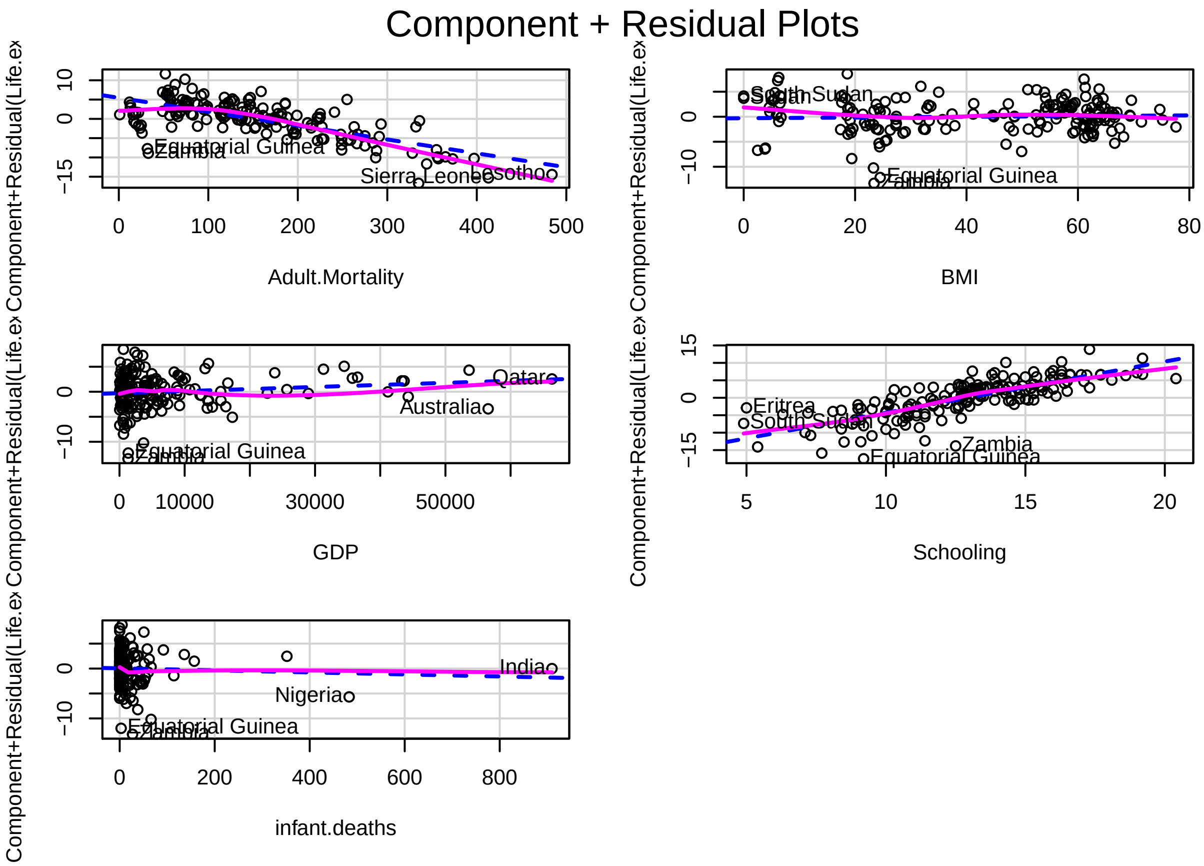
```
##      (Intercept) Adult.Mortality             BMI             GDP       Schooling
##      -0.12648273     -0.07410251      0.01117398     -0.01007439      0.11698851
##    infant.deaths
##       1.27398927
```

```r
dfbetasPlots(life_lm, id.n=4)
```



11. Panel of residual plus component plots

```r
crPlots(life_lm, id=TRUE)
```



Report Assignment

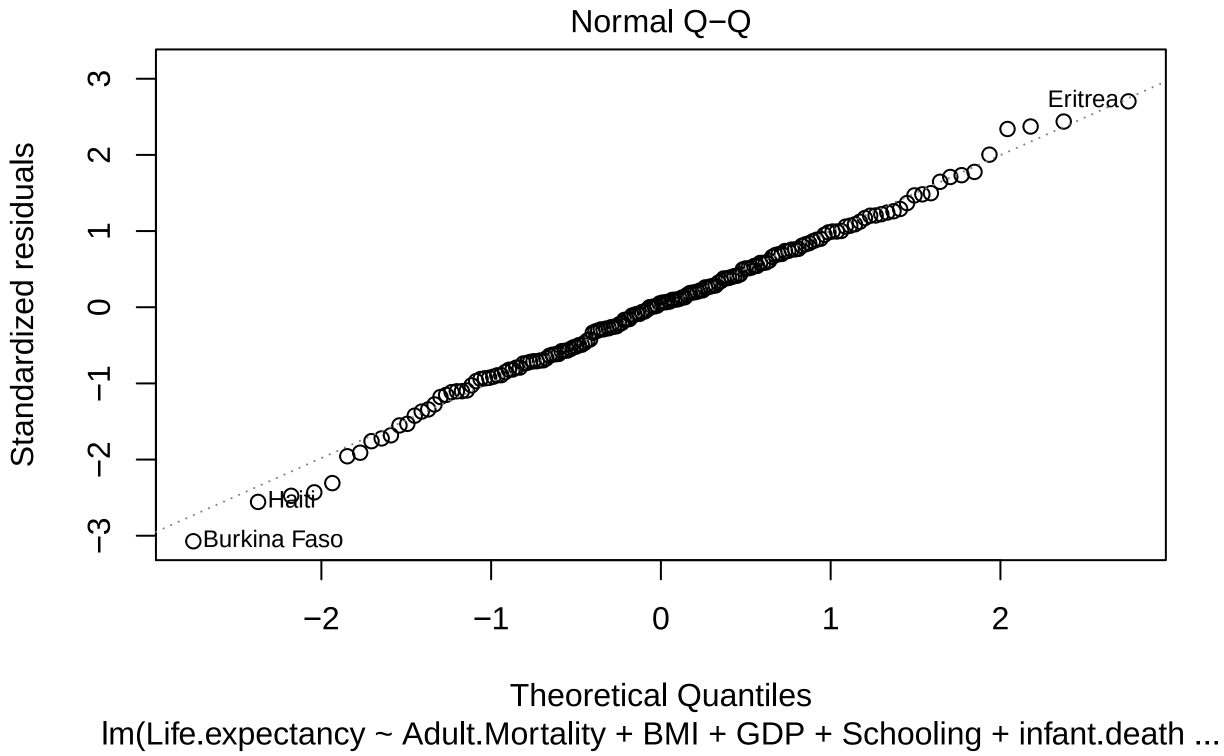
1. Interpretation of the scatterplot matrix:

The individual scatterplots of the response *Life Expectancy* with predictors Adult Mortality, BMI, GDP, Years of schooling, and infant mortality show that linear relationships and correlations exist for many of the predictors; however, the summary function shows that only adult mortality and years of schooling are significant.

2. Assessment of the normality assumption:

The population is not normal according to the Shapiro-Wilk test, p-value: 0.003494. Removing the significant outliers (Equatorial Guinea, Angola, and Zambia) increases the p-value significantly [0.712] and improves the normality assumption.

```
plot(life_lm2, which=2)
```

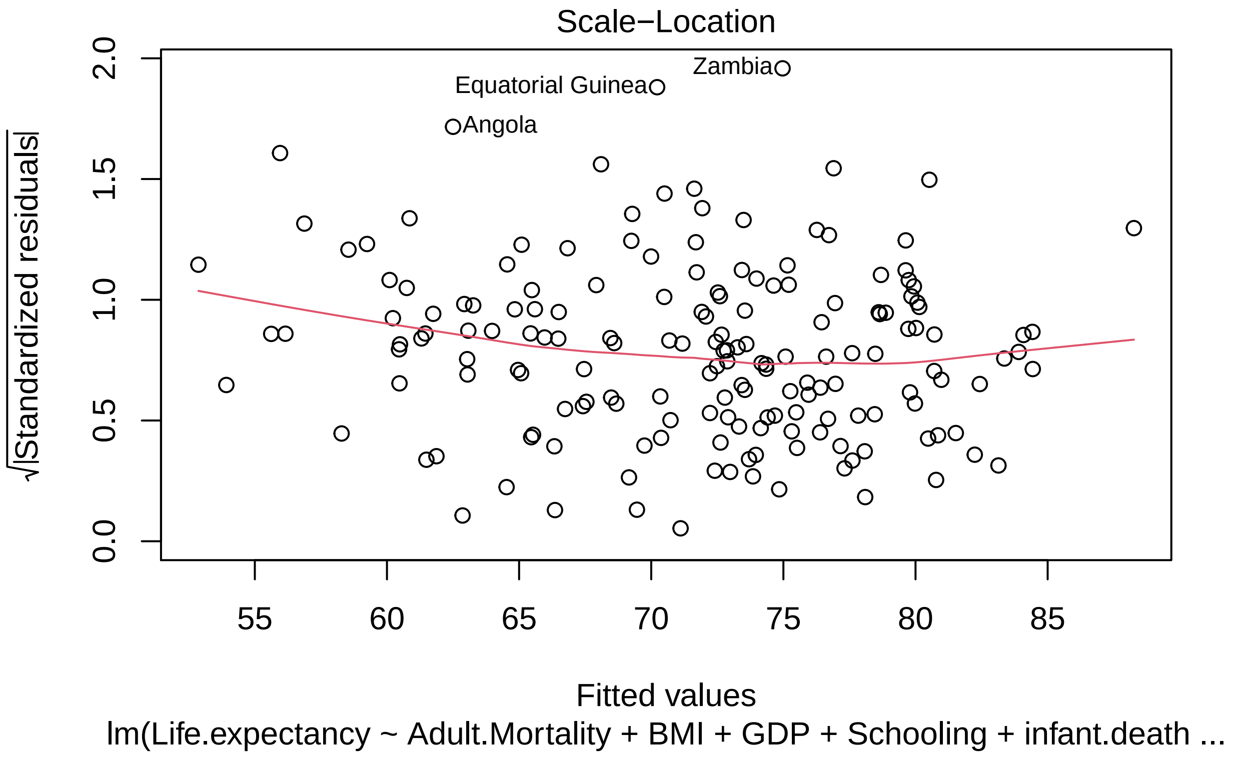


```
shapiro.test(res2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res2
## W = 0.99396, p-value = 0.712
```

3. Assessment of the linearity assumption:

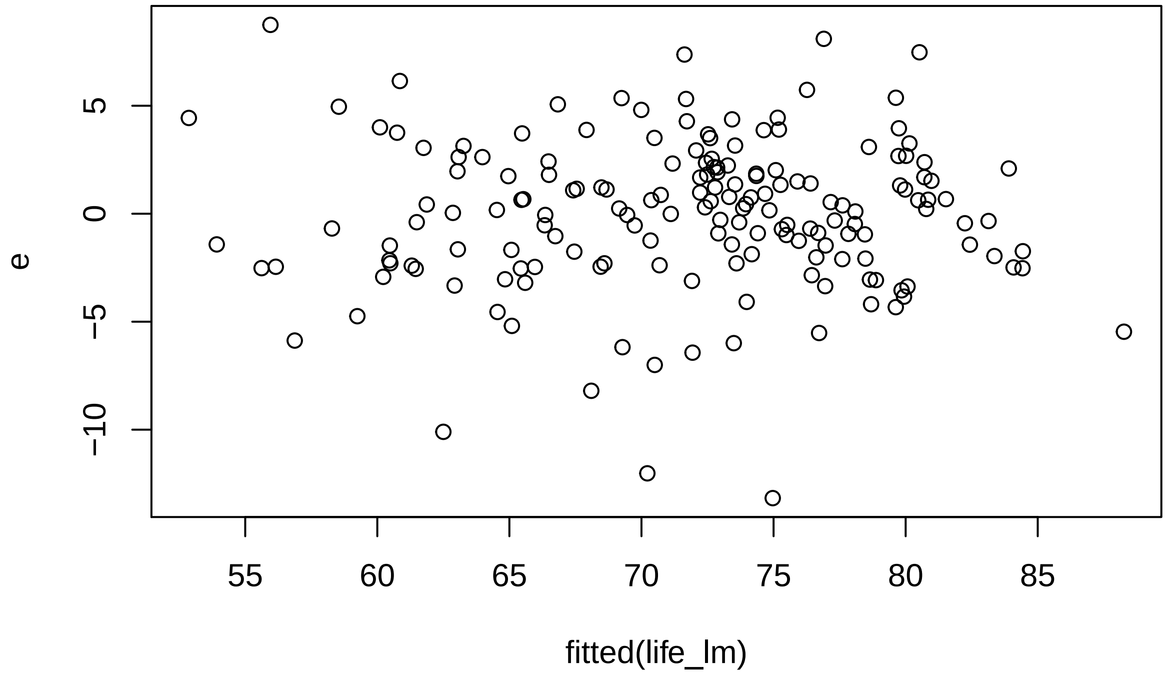
```
plot(life_lm, which=3)
```



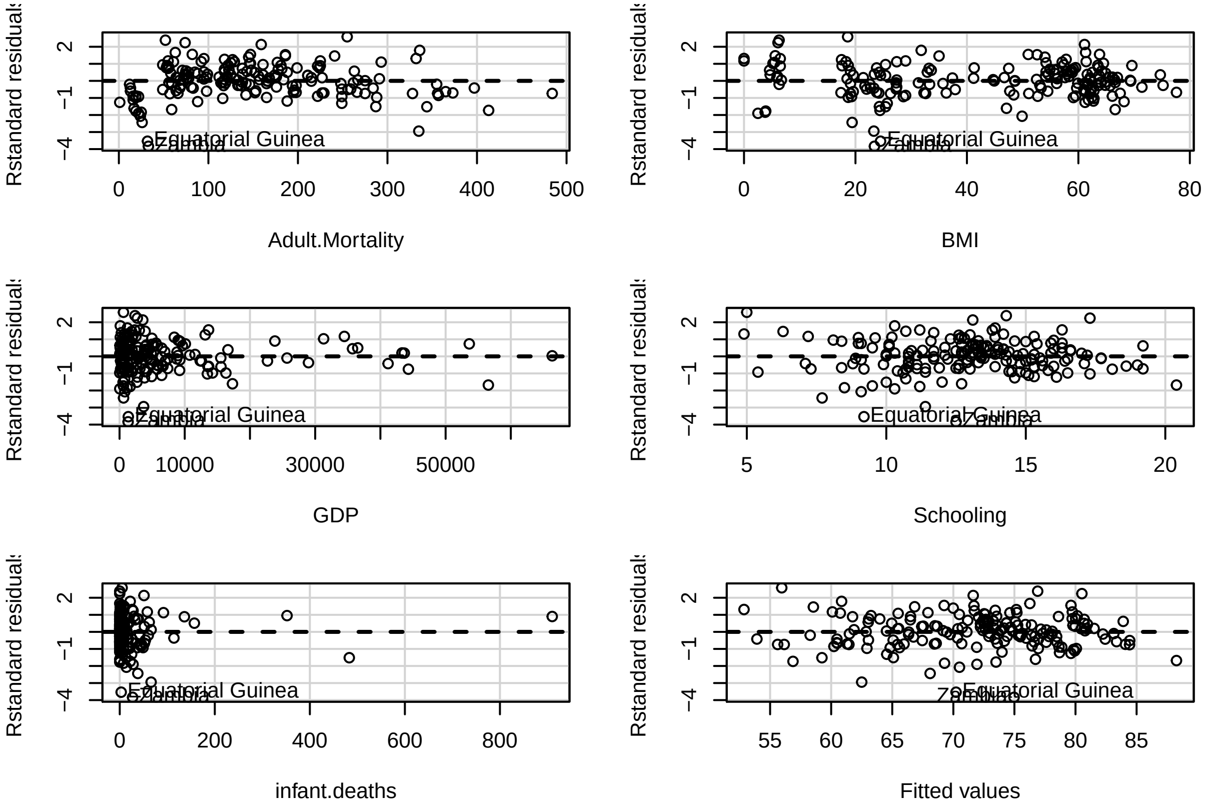
The scale-location plot looks fairly linear – there are no patterns where the standardized residuals increase with the fitted values or funnels, and the reference line is fairy horizontal.

4. Assessment of the homoscedasticity assumption:

```
plot(e~fitted(life_lm))
```



```r
residualPlots(life_lm, id=TRUE, quadratic=FALSE, type="rstandard", tests=FALSE)
```



The standardized residuals against the fitted values plot appear to be even distributed over zero – there are no patterns where the standardized residuals increase with the fitted values or funnels. The exception is for the infant deaths and GDP, which appears to have funnels.

5. List of points which appear to be outliers or high-leverage points

Outliers

The normal q-q plot shows Angola, Equatorial Guinea, and Zambia to have residuals outliers The scale-location plot shows Angola, Equatorial Guinea, and Zambia to have residual outliers.

Leverage

Cook's distance - measures the influence of an observation on the entire set of fitted values: Equatorial Guinea, India, and Nigeria
DFFITS - measures the influence of an observation on an individual fitted value: India, Nigeria, Slovenia, Zambia

6. List of points which appear to be influential points

Influence

DFBETAS - measures the influence of an observation on an individual regression coefficient. The table below summarizes the countries with highest leverage points for each of the regression coefficients from the DFBETAS plots:

Regression coefficient	Countries	High Influence Points
Adult Mortality	Equatorial Guinea, Angola, Burkina Faso, Zambia	32, 335, 26, 33
BMI	Sao Tome and Principe, Zambia, Cyprus, Slovenia	19, 33, 52, 74
GDP	Bahamas, Japan, Sao Tome and Principe, Australia	147, 55, 19, 59
Schooling	Burkina Faso, Equatorial Guinea, Cuba, Eritrea	26, 32, 92, 255
Infant mortality	India, Pakistan, Nigeria, Eritrea	181, 161, 344, 255

There appears to be an error when looking at the data for *Years of Schooling* high influence points. It is unlikely that there exists 26, 32, and certainly not 92 and 255 average years of schooling in these countries.

7. Interpretation of the residual plus component plots

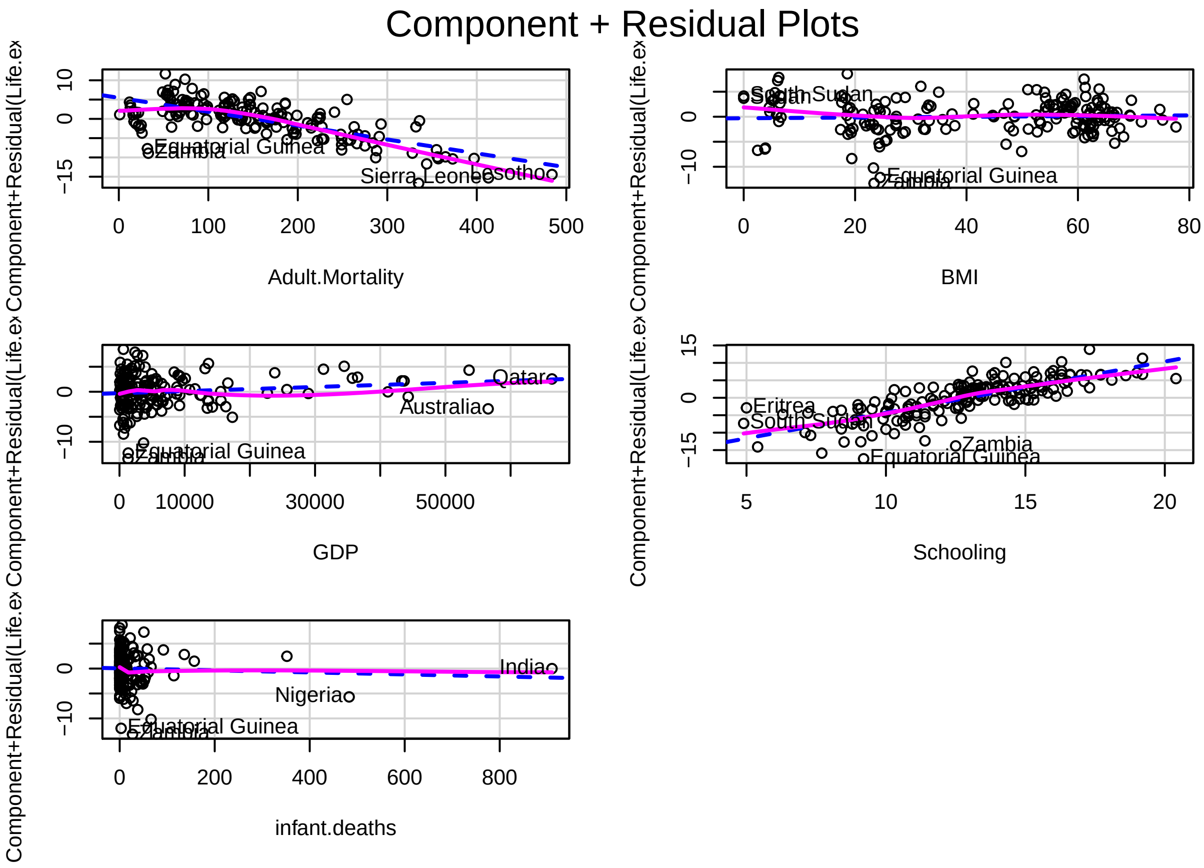
The residual plus component plots and the added variable plots mirror the statistics of the summary() function. Visually, we can see a decreasing trend for adult morality as adult deaths increase, suggesting that this variable plays a significant role in predicting life expectancy. We can also see an upward trend in Schooling, suggesting that as the number of years of schooling increases, the life expectancy increases, and this variable should be included in the model.

The regression coefficients that are not significant (BMI, GDP, infant deaths) are almost completely horizontal, with no increasing or decreasing trend on the response variable, suggesting they do not belong as a predictor in the model (they are not significant in the summary statistics as well).

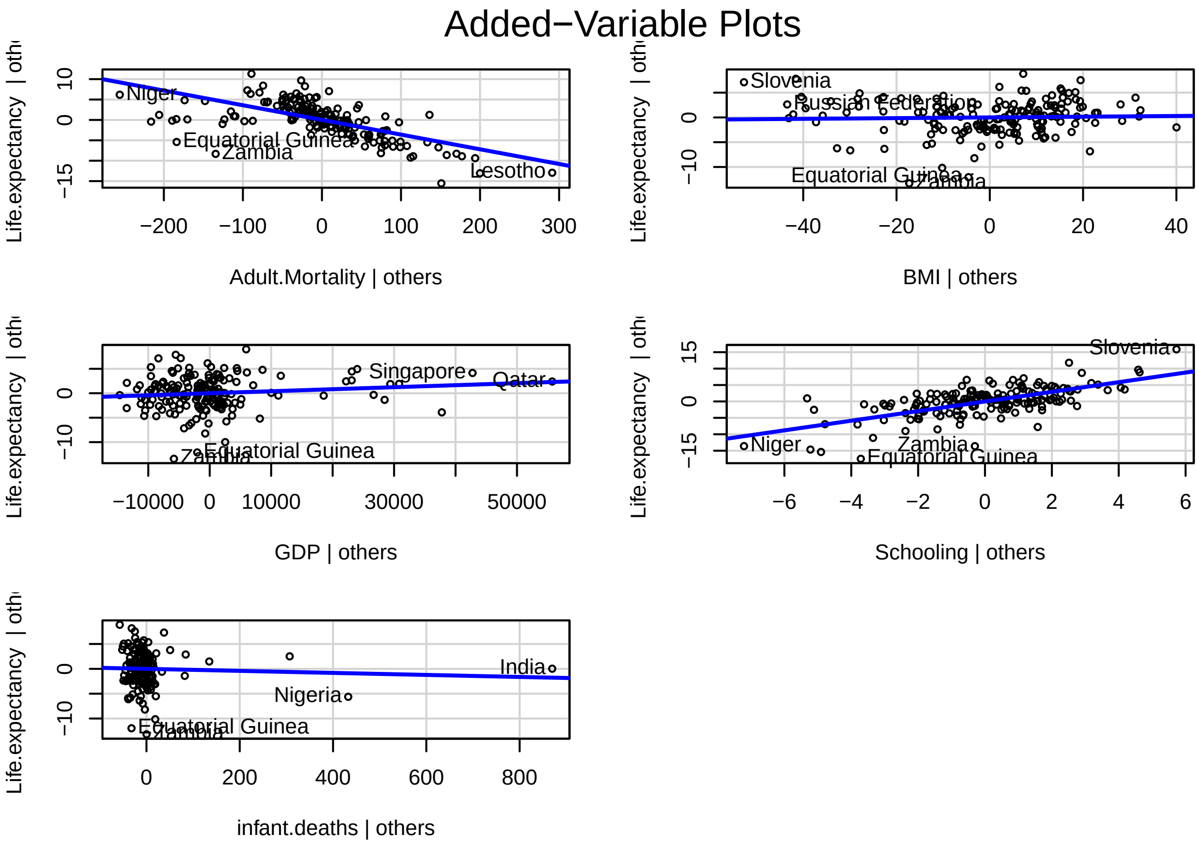
```
summary(life_lm)$coefficients
```

```
##                      Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)      5.765167e+01 1.787229e+00  32.2575688 1.201065e-73
## Adult.Mortality -3.600419e-02 3.346049e-03 -10.7602114 7.599575e-21
## BMI              7.190785e-03 1.574394e-02   0.4567335 6.484561e-01
## GDP              4.123093e-05 2.665781e-05   1.5466736 1.238349e-01
## Schooling        1.467075e+00 1.299460e-01  11.2898768 2.520101e-22
## infant.deaths   -2.037792e-03 3.249232e-03  -0.6271610 5.314102e-01
```

```
crPlots(life_lm, id=TRUE)
```



```r
avPlots(life_lm)
```


Added−Variable Plots
Life.expectancy | others
Adult.Mortality | others
Niger
Equatorial Guinea
Zambia
Lesotho
10
0
−15
−200
−100
0
100
200
300
BMI | others
Slovenia
Russian Federation
Equatorial Guinea
Zambia
0
−10
−40
−20
0
20
40
GDP | others
Singapore
Qatar
Equatorial Guinea
Zambia
0
−10
−10000
0
10000
30000
50000
Schooling | others
Slovenia
Niger
Zambia
Equatorial Guinea
15
0
−15
−6
−4
−2
0
2
4
6
infant.deaths | others
India
Nigeria
Equatorial Guinea
Zambia
0
−10
0
200
400
600
800

Life Expectancy - Categorical Variables as Predictors

Kasia Krueger

9/22/2021

Removing high influential points (errors in years of schooling and infant deaths as found in Week 3 Regression diagnostics.)

```
lifedata <- lifedata[-which(rownames(lifedata) == "Burkina Faso"),] #Probable error in years of schooli
lifedata <- lifedata[-which(rownames(lifedata) == "Equatorial Guinea"),] #Probable error in years of sc
lifedata <- lifedata[-which(rownames(lifedata) == "Cuba"),] #Error in years of schooling (92)
lifedata <- lifedata[-which(rownames(lifedata) == "Eritrea"),] # Error in years of schooling (255)
lifedata <- lifedata[-which(rownames(lifedata) == "India"),] #Probable error in infant deaths (800/1000
```

Week 4

1. Fit the multiple regression model

Using two categorical predictors and 2-3 numerical predictors, and include all pairwise interaction terms

```
lifedata$polio.vacc <- as.factor(lifedata$polio.vacc)
lifedata$HIV.AIDS <- as.factor(lifedata$HIV.AIDS)
lifedata$Status <- as.factor(lifedata$Status)


levels(lifedata$polio.vacc) <- c(’low’, ’medium’, ’high’)
levels(lifedata$HIV.AIDS) <- c(’low’, ’medium’, ’high’)

model.trtplusint <- lm(Life.expectancy ~ Adult.Mortality + Schooling
                    + Status   + polio.vacc + HIV.AIDS + Adult.Mortality:polio.vacc
                    +Adult.Mortality:Status + Schooling: Status
                       + Adult.Mortality:HIV.AIDS + polio.vacc:HIV.AIDS
                       + Schooling:polio.vacc + Schooling:HIV.AIDS
                    ,lifedata)

model.trtnoint <- update(model.trtplusint,~ .
                         -Adult.Mortality:Status
                         -Schooling:Status
                         -  Adult.Mortality:polio.vacc
                         - Adult.Mortality:HIV.AIDS
                         - polio.vacc:HIV.AIDS
                         - Schooling:polio.vacc
```

```
                          - Schooling:HIV.AIDS)
summary(model.trtplusint)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     Status + polio.vacc + HIV.AIDS + Adult.Mortality:polio.vacc +
##     Adult.Mortality:Status + Schooling:Status + Adult.Mortality:HIV.AIDS +
##     polio.vacc:HIV.AIDS + Schooling:polio.vacc + Schooling:HIV.AIDS,
##     data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3808 -1.2295  0.1008  1.5321  7.0394
##
## Coefficients: (3 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      127.09595   15.15302   8.388 3.49e-14 ***
## Adult.Mortality                   -0.13346    0.02505  -5.328 3.60e-07 ***
## Schooling                         -2.61822    0.96610  -2.710 0.007516 **
## StatusDeveloping                 -17.62631    6.22974  -2.829 0.005307 **
## polio.vaccmedium                 -57.32582   14.29304  -4.011 9.54e-05 ***
## polio.vacchigh                   -48.27696   14.01036  -3.446 0.000740 ***
## HIV.AIDSmedium                    -1.10705    8.53246  -0.130 0.896943
## HIV.AIDShigh                     -51.60133   22.42319  -2.301 0.022764 *
## Adult.Mortality:polio.vaccmedium   0.07472    0.02067   3.615 0.000411 ***
## Adult.Mortality:polio.vacchigh     0.06175    0.01997   3.093 0.002369 **
## Adult.Mortality:StatusDeveloping   0.03073    0.01575   1.952 0.052860 .
## Schooling:StatusDeveloping         0.86603    0.36029   2.404 0.017459 *
## Adult.Mortality:HIV.AIDSmedium     0.04779    0.01732   2.759 0.006518 **
## Adult.Mortality:HIV.AIDShigh       0.13285    0.05596   2.374 0.018858 *
## polio.vaccmedium:HIV.AIDSmedium   -3.18898    4.27794  -0.745 0.457175
## polio.vacchigh:HIV.AIDSmedium      3.33691    3.38301   0.986 0.325549
## polio.vaccmedium:HIV.AIDShigh           NA         NA      NA       NA
## polio.vacchigh:HIV.AIDShigh             NA         NA      NA       NA
## Schooling:polio.vaccmedium         3.49760    0.95629   3.657 0.000353 ***
## Schooling:polio.vacchigh           3.07158    0.90866   3.380 0.000924 ***
## Schooling:HIV.AIDSmedium          -1.76683    0.62570  -2.824 0.005396 **
## Schooling:HIV.AIDShigh                  NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.815 on 149 degrees of freedom
## Multiple R-squared:  0.8863, Adjusted R-squared:  0.8726
## F-statistic: 64.52 on 18 and 149 DF,  p-value: < 2.2e-16
```

```
anova(model.trtnoint, model.trtplusint)$‘Pr(>F)‘
```

```
## [1]           NA 1.378884e-05
```

Small p-value (1.378884e-05) suggests interaction effects need to stay in the model.

2. Investigating the p-values associated with each interaction term

Now, we choose which interaction effects need to stay in the model.

```
mod1.lm <- update(model.trtplusint, ~ . - polio.vacc:HIV.AIDS)
anova(mod1.lm, model.trtplusint)$‘Pr(>F)‘
```

```
## [1]      NA 0.12852
```

Large p-value (0.12852) suggests we can remove polio.vacc:HIV.AIDS from model.

```
mod2.lm <- update(mod1.lm, ~ . - Adult.Mortality:HIV.AIDS)
anova(mod2.lm, mod1.lm)$‘Pr(>F)‘
```

```
## [1]           NA 0.0003342475
```

Small p-value (0.000334) suggests Adult.Mortality:HIV.AIDS needs to stay in the model.

```
mod3.lm <- update(mod1.lm, ~ . - Schooling:HIV.AIDS)
anova(mod3.lm, mod1.lm)$‘Pr(>F)‘
```

```
## [1]         NA 0.04002118
```

Small p-value (0.04002) suggests Schooling:HIV.AIDS needs to stay in model.

```
mod4.lm <- update(mod1.lm, ~ . - Adult.Mortality:HIV.AIDS)
anova(mod4.lm, mod1.lm)$‘Pr(>F)‘
```

```
## [1]           NA 0.0003342475
```

Small p-value (0.000334) suggests we need to keep Adult.Mortality:HIV.AIDS in the model.

```
mod5.lm <- update(mod1.lm, ~ . - Adult.Mortality:Status)
anova(mod5.lm, mod1.lm)$‘Pr(>F)‘
```

```
## [1]         NA 0.04279025
```

Small p-value (0.04279) suggests we need to keep Adult.Mortality:Status in the model.

```r
mod6.lm <- update(mod1.lm, ~ . - Schooling:Status)
anova(mod6.lm, mod1.lm)$‘Pr(>F)‘
```

```
## [1]         NA 0.01918073
```

Small p-value (0.01918) suggests we need to keep Schooling:Status in the model.

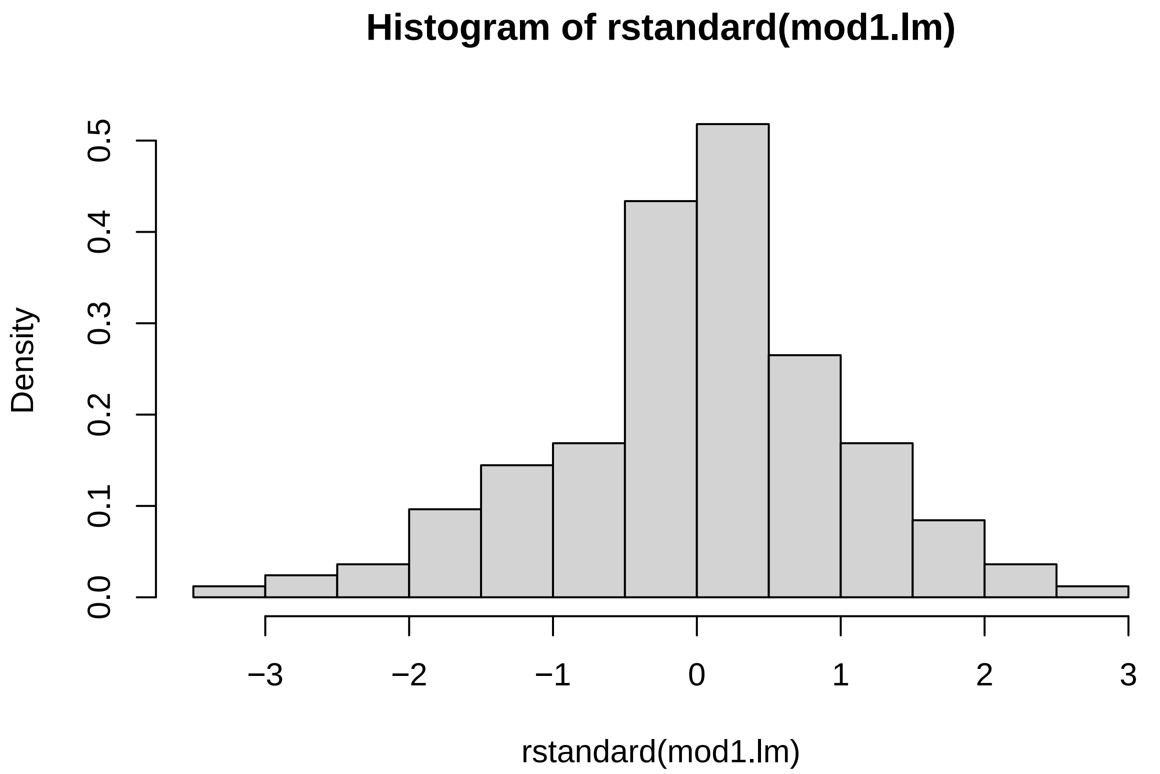
Our conclusion is to use mod1.lm.

```r
summary(mod1.lm)
```

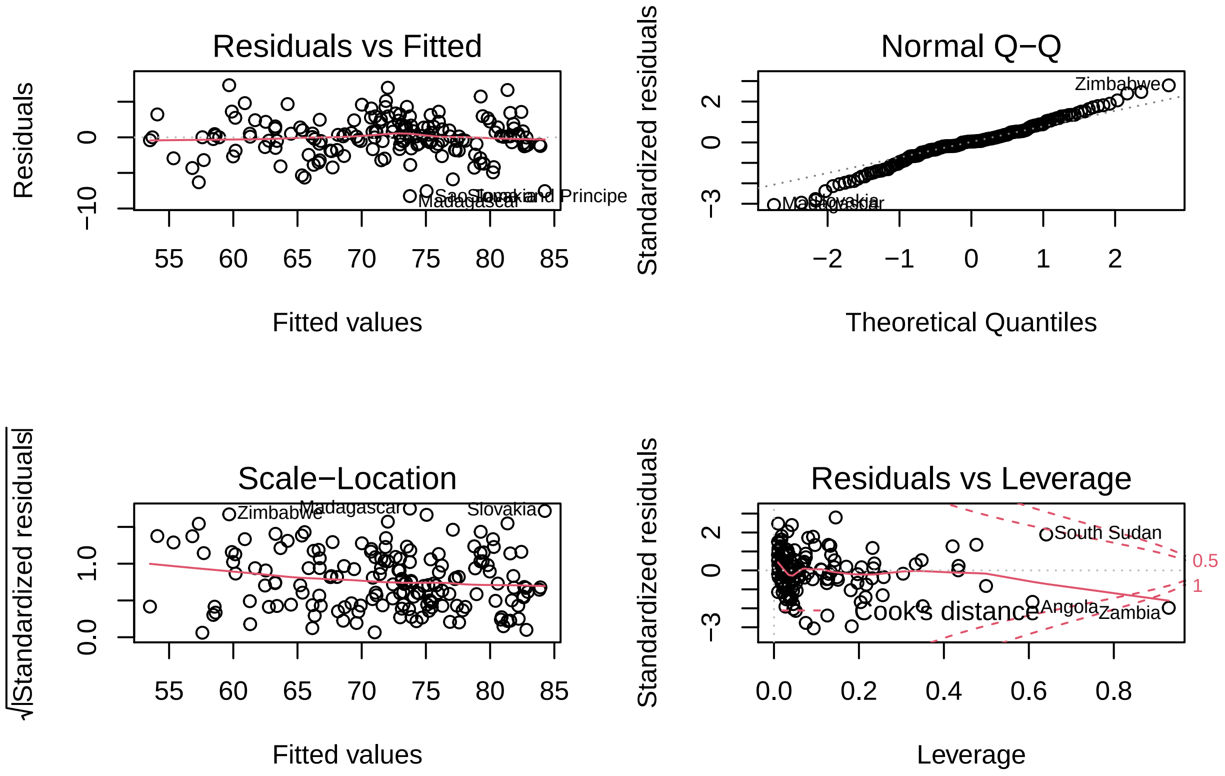
```
## 
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling + 
##     Status + polio.vacc + HIV.AIDS + Adult.Mortality:polio.vacc + 
##     Adult.Mortality:Status + Schooling:Status + Adult.Mortality:HIV.AIDS + 
##     Schooling:polio.vacc + Schooling:HIV.AIDS, data = lifedata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.2536 -1.2699  0.1178  1.4825  7.3188 
## 
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      124.03152   15.16909   8.177 1.10e-13 ***
## Adult.Mortality                   -0.13104    0.02486  -5.271 4.62e-07 ***
## Schooling                         -2.43020    0.96144  -2.528 0.012511 *  
## StatusDeveloping                 -17.71673    6.27388  -2.824 0.005385 ** 
## polio.vaccmedium                 -53.79694   14.26436  -3.771 0.000232 ***
## polio.vacchigh                   -45.21252   14.01074  -3.227 0.001534 ** 
## HIV.AIDSmedium                    -8.16404    7.84279  -1.041 0.299557    
## HIV.AIDShigh                     -51.00402   22.05354  -2.313 0.022088 *  
## Adult.Mortality:polio.vaccmedium   0.06538    0.01965   3.327 0.001102 ** 
## Adult.Mortality:polio.vacchigh     0.05932    0.01964   3.020 0.002973 ** 
## Adult.Mortality:StatusDeveloping   0.03236    0.01584   2.043 0.042790 *  
## Schooling:StatusDeveloping         0.85900    0.36284   2.367 0.019181 *  
## Adult.Mortality:HIV.AIDSmedium     0.05282    0.01439   3.671 0.000334 ***
## Adult.Mortality:HIV.AIDShigh       0.13009    0.05486   2.371 0.018997 *  
## Schooling:polio.vaccmedium         3.33096    0.94955   3.508 0.000595 ***
## Schooling:polio.vacchigh           2.88356    0.90285   3.194 0.001709 ** 
## Schooling:HIV.AIDSmedium          -1.09670    0.52944  -2.071 0.040021 *  
## Schooling:HIV.AIDShigh                  NA         NA      NA       NA    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 2.835 on 151 degrees of freedom
## Multiple R-squared:  0.8831, Adjusted R-squared:  0.8707 
## F-statistic: 71.31 on 16 and 151 DF,  p-value: < 2.2e-16
```

3. Create a set of diagnostic plots and interpret them

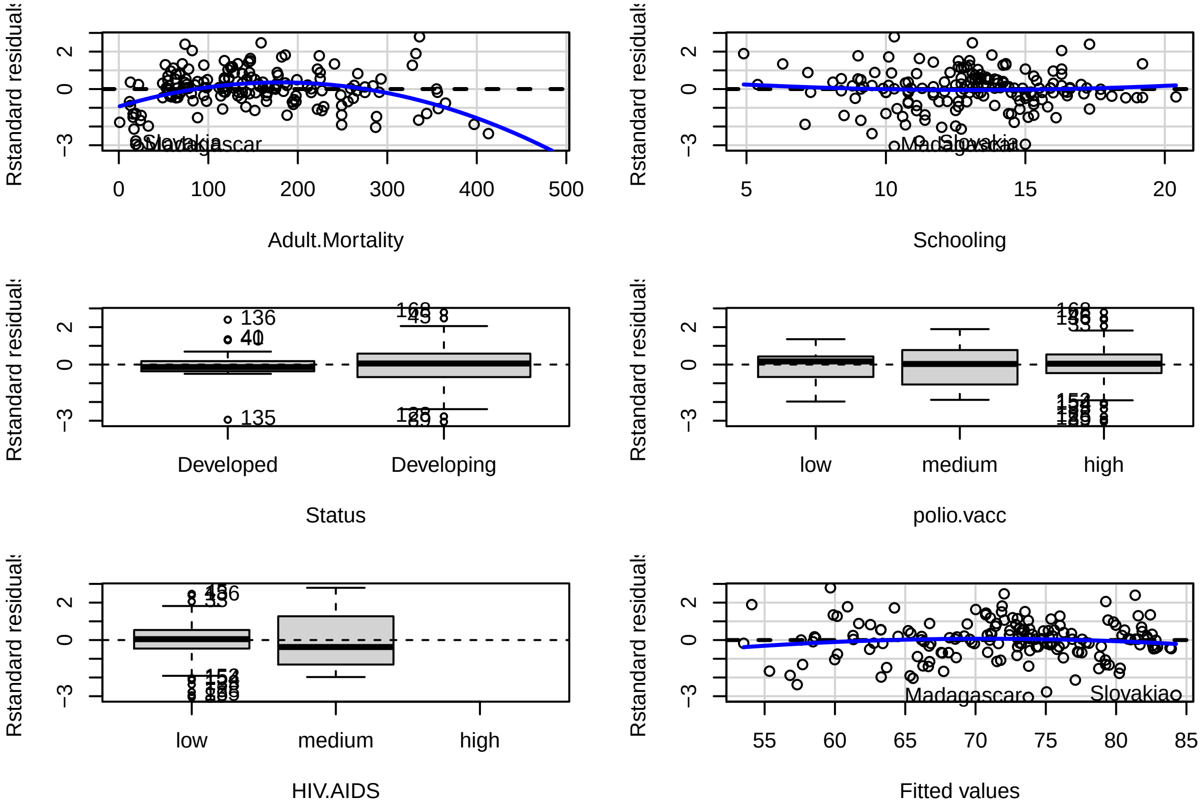
```r
hist(rstandard(mod1.lm), probability=TRUE)
```



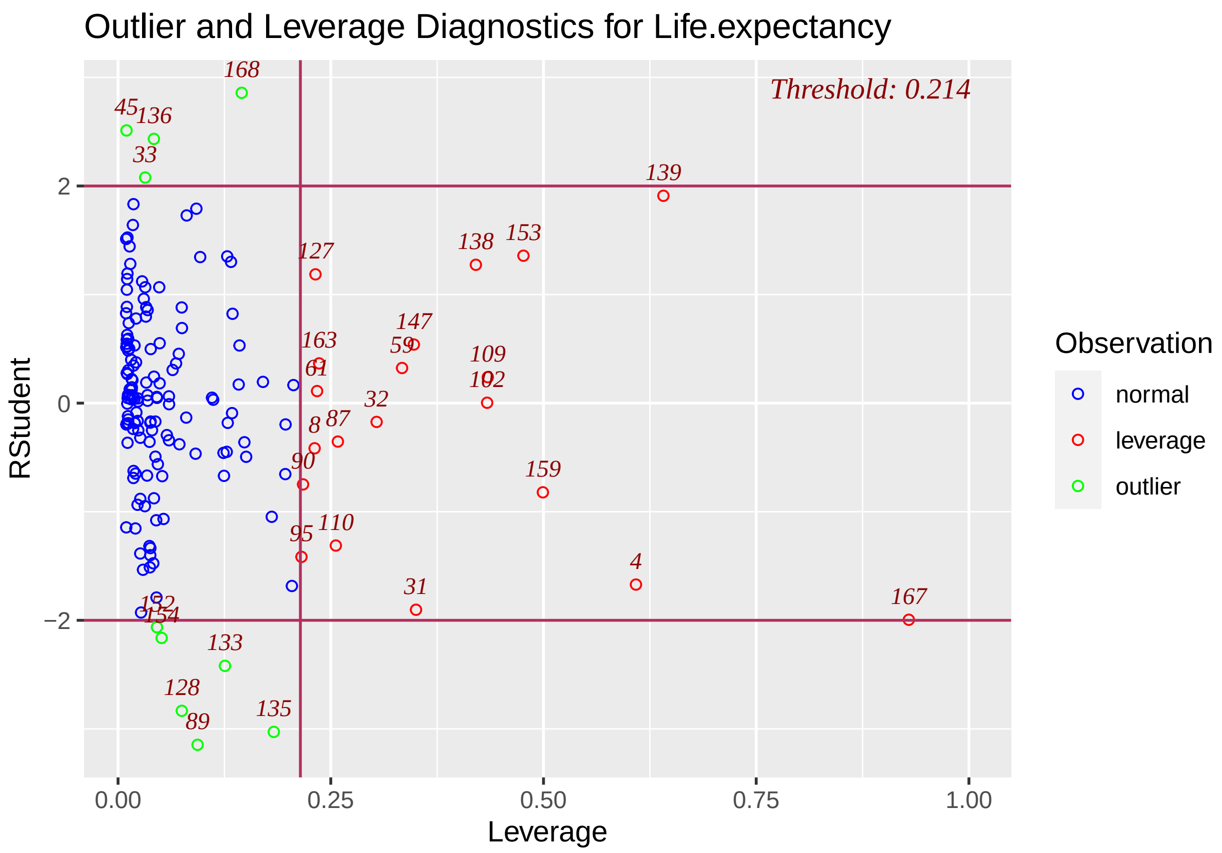
```r
par(mfrow=c(2,2))
plot(mod1.lm)
```



```
car::residualPlots(mod1.lm, id=TRUE, quadratic=TRUE, type=’rstandard’, tests=FALSE)
```



```
ols_plot_resid_lev(mod1.lm)
```



Homoscedastcity

The standardized residuals plots and scale-location plot show that the homoscedasticity assumptions appear to be met, as the residuals are scattered evenly around 0.

Normality

The QQ-plot shows the normality assumption is met, with no heavy tails.

Linearity

The linearity assumption appears to be met for the coefficients, with the exception of Adult.mortality, which appears to have a quadratic response.

Influential Observations

There are a few outliers that indicate an influential observation that can be removed as can be seen in the Residuals vs. Leverage plot and Outlier and Leverage Diagnostics plot. Zimbabwe and Madagascar appear to occur more than once as residual outliers.

4. Interpret the presence of any interaction terms in the model

```
summary(mod1.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     Status + polio.vacc + HIV.AIDS + Adult.Mortality:polio.vacc +
##     Adult.Mortality:Status + Schooling:Status + Adult.Mortality:HIV.AIDS +
##     Schooling:polio.vacc + Schooling:HIV.AIDS, data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2536 -1.2699  0.1178  1.4825  7.3188
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     124.03152   15.16909   8.177 1.10e-13 ***
## Adult.Mortality                  -0.13104    0.02486  -5.271 4.62e-07 ***
## Schooling                        -2.43020    0.96144  -2.528 0.012511 *
## StatusDeveloping                -17.71673    6.27388  -2.824 0.005385 **
## polio.vaccmedium                -53.79694   14.26436  -3.771 0.000232 ***
## polio.vacchigh                  -45.21252   14.01074  -3.227 0.001534 **
## HIV.AIDSmedium                   -8.16404    7.84279  -1.041 0.299557
## HIV.AIDShigh                    -51.00402   22.05354  -2.313 0.022088 *
## Adult.Mortality:polio.vaccmedium   0.06538    0.01965   3.327 0.001102 **
## Adult.Mortality:polio.vacchigh     0.05932    0.01964   3.020 0.002973 **
## Adult.Mortality:StatusDeveloping   0.03236    0.01584   2.043 0.042790 *
## Schooling:StatusDeveloping         0.85900    0.36284   2.367 0.019181 *
## Adult.Mortality:HIV.AIDSmedium     0.05282    0.01439   3.671 0.000334 ***
## Adult.Mortality:HIV.AIDShigh       0.13009    0.05486   2.371 0.018997 *
## Schooling:polio.vaccmedium         3.33096    0.94955   3.508 0.000595 ***
## Schooling:polio.vacchigh           2.88356    0.90285   3.194 0.001709 **
## Schooling:HIV.AIDSmedium          -1.09670    0.52944  -2.071 0.040021 *
## Schooling:HIV.AIDShigh                  NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.835 on 151 degrees of freedom
## Multiple R-squared:  0.8831, Adjusted R-squared:  0.8707
## F-statistic: 71.31 on 16 and 151 DF,  p-value: < 2.2e-16
```

- Beta0 is the intercept term and represents the life expectancy (124.03 years) if all other coefficients are zero.
- Beta1 is the slope attached to the predictor Adult Mortality, and decreases by 0.131 years for ever 1/1000 Adult deaths (ages 15-60)
- Beta2 is the slope attached to the predictor Schooling and represents a decrease of 2.43 years for every 1 year of schooling completed.
- Beta3 is the treatment effect associated with the status of a country (developed or developing). -17.72 years represents the life expectancy differential for developing countries relative to developed countries.
- Beta4 is the treatment effect associated with a country having low, medium, or high rates of vaccination for polio among 1 year olds. -53.80 years for medium rates and -45.21 years for high rates represents

the life expectancy differential for low vaccination rates relative to high or medium rates (both these coefficients have high standard errors).
- Beta5 is the treatment effect associated with a country having low, medium, or high rates of HIV/AIDS. -8.16 years for medium rates and -51.00 years for high rates represents the life expectancy differential for low HIV/AIDS rates relative to high or medium rates (with a high standard error for the HIV/AIDS "high" differential).
- Beta6 represents the interaction between Adult Mortality and vaccination for polio and the differential for low vaccination rates relative to high or medium rates.
- Beta7 represents the interaction between Adult Mortality and Status (Developing) and the differential for life expectancy in developing countries relative to developed countries -Beta8 represents the interaction between Schooling and Status (Developing) and the differential for life expectancy in developing countries relative to developed countries -Beta9 represents the interaction between Adult Mortality and HIV.AIDS and the differential for low HIV/AIDS rates relative to high or medium rates.
- Beta10 represents the interaction between Schooling and vaccination for polio and the differential for low vaccination rates relative to high or medium rates.
- Beta11 represents the interaction between Schooling and HIV/AIDS rates and the differential for low HIV/AIDS rates relative to high or medium rates.

With the addition of so many predictor variables, the intercept jumped to 124 years of life expectancy; however, this also explains why so many of the predictor variables are negative that were once positive (e.g., years of schooling having a negative effect on life expectancy).

5. Collapsing Categories

```
lifedata$polio.AIDS<- interaction(lifedata$polio.vacc
                                  , lifedata$HIV.AIDS)
polio.AIDS <- lifedata$polio.AIDS

model.diffint <- lm(Life.expectancy ~ Adult.Mortality + Schooling + polio.AIDS, lifedata) # "diffint" f
```

```
summary(model.diffint)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     polio.AIDS, data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3786 -1.6712 -0.0181  1.9342  6.9955
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              59.805939   2.167110  27.597  < 2e-16 ***
## Adult.Mortality          -0.039304   0.003694 -10.639  < 2e-16 ***
## Schooling                 1.328027   0.116060  11.443  < 2e-16 ***
## polio.AIDSmedium.low     -1.786068   1.426560  -1.252  0.21242
## polio.AIDShigh.low        1.251594   1.115385   1.122  0.26351
## polio.AIDSlow.medium     -6.823637   2.442497  -2.794  0.00586 **
## polio.AIDSmedium.medium  -0.720368   2.132718  -0.338  0.73598
## polio.AIDShigh.medium   -0.132127   1.775742  -0.074  0.94078
## polio.AIDSlow.high      -1.292640   3.410664  -0.379  0.70520
## polio.AIDShigh.high     -1.385017   3.330276  -0.416  0.67806
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.117 on 158 degrees of freedom
## Multiple R-squared:  0.8522, Adjusted R-squared:  0.8437
## F-statistic: 101.2 on 9 and 158 DF,  p-value: < 2.2e-16
```

```
newpolio.AIDS <- lifedata$polio.AIDS
levels(newpolio.AIDS)[c(4,7)] <- ’High Mortlality’ #low.medium, low.high
model.red <- update(model.diffint, ~ . -polio.AIDS + newpolio.AIDS)
summary(model.red)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     newpolio.AIDS, data = lifedata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.530 -1.695 -0.015  2.058  7.091
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  59.239487   2.135090  27.746   <2e-16 ***
## Adult.Mortality              -0.037918   0.003569 -10.624   <2e-16 ***
## Schooling                     1.350268   0.115302  11.711   <2e-16 ***
## newpolio.AIDSmedium.low      -1.661775   1.428013  -1.164   0.2463
## newpolio.AIDShigh.low         1.332055   1.117208   1.192   0.2349
## newpolio.AIDSHigh Mortlality -5.055807   2.094528  -2.414   0.0169 *
## newpolio.AIDSmedium.medium   -0.812698   2.138033  -0.380   0.7044
## newpolio.AIDShigh.medium     -0.264929   1.778465  -0.149   0.8818
## newpolio.AIDShigh.high       -1.589042   3.336964  -0.476   0.6346
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.126 on 159 degrees of freedom
## Multiple R-squared:  0.8503, Adjusted R-squared:  0.8428
## F-statistic: 112.9 on 8 and 159 DF,  p-value: < 2.2e-16
```

```
anova(model.red, model.diffint)$‘Pr(>F)‘
```

```
## [1]       NA 0.164814
```

```
linearHypothesis(model.diffint, ’polio.AIDSlow.medium = polio.AIDSlow.high ’)
```

```
## Linear hypothesis test
##
## Hypothesis:
## polio.AIDSlow.medium - polio.AIDSlow.high = 0
```

```
##
## Model 1: restricted model
## Model 2: Life.expectancy ~ Adult.Mortality + Schooling + polio.AIDS
##
##   Res.Df  RSS Df Sum of Sq      F Pr(>F)
## 1    159 1554
## 2    158 1535  1    18.921 1.9475 0.1648
```

Creating an interaction between polio vaccination rates (low, medium, high) and HIV/AIDS rates (low, medium, high) and assigning "high mortality" to "low.medium and low.high" meaning low polio vaccination rates and medium and high rates of HIV/AIDS deaths.

We see from the large p-value (0.1648) that suggests collapsing the categories does not work for the model. This is confirmed by the linearHypothesis test.

Life Expectancy - Transformations and Weighted Least Squares

Kasia Krueger

10/08/2021

Week 5

Cleaning up data

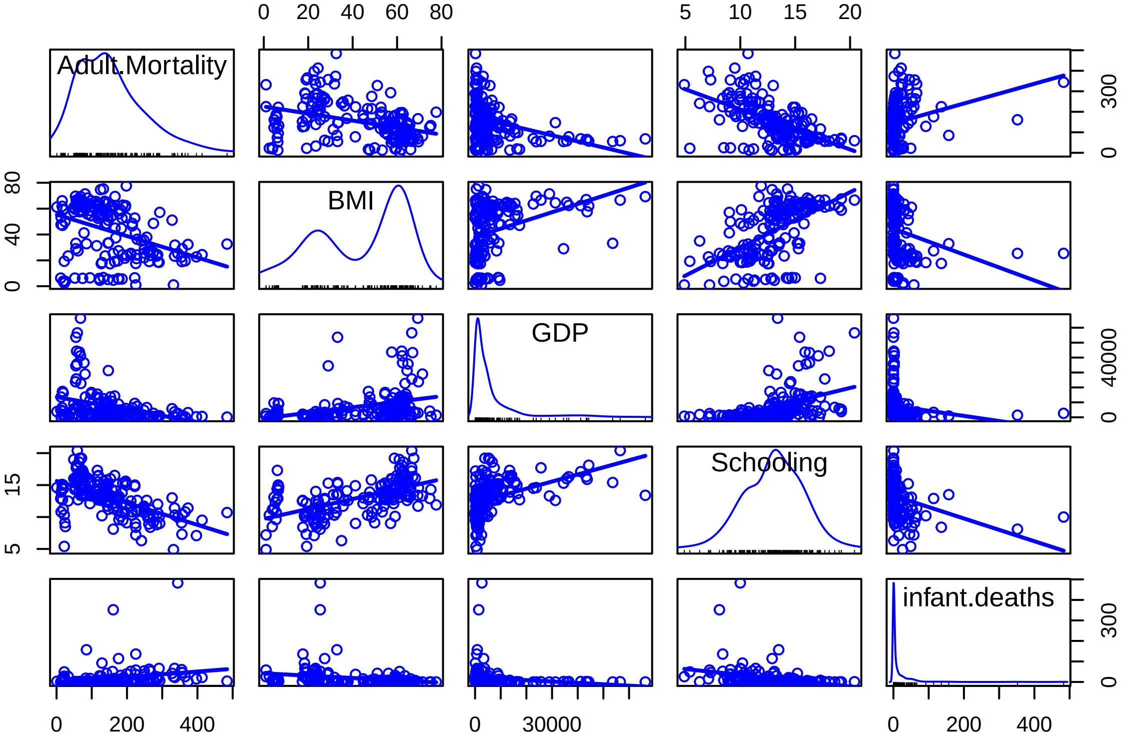
Removing high influential points (errors in years of schooling as found in Week 3 Regression diagnostics.)

```
lifedata <- lifedata[-which(rownames(lifedata) == "Burkina Faso"),]
#Possible error in years of schooling (26)
lifedata <- lifedata[-which(rownames(lifedata) == "Equatorial Guinea"),]
#Possible error in years of schooling (32)
lifedata <- lifedata[-which(rownames(lifedata) == "Cuba"),]
#Error in years of schooling (92)
lifedata <- lifedata[-which(rownames(lifedata) == "Eritrea"),]
# Error in years of schooling (255)
lifedata <- lifedata[-which(rownames(lifedata) == "India"),]
#High cook’s distance + possible error in infant deaths (800/1000)
```

```
lifedata$polio.vacc <- as.factor(lifedata$polio.vacc)
lifedata$HIV.AIDS <- as.factor(lifedata$HIV.AIDS)
lifedata$Status <- as.factor(lifedata$Status)

levels(lifedata$polio.vacc) <- c(’low’, ’medium’, ’high’)
levels(lifedata$HIV.AIDS) <- c(’low’, ’medium’, ’high’)
```

Base Model

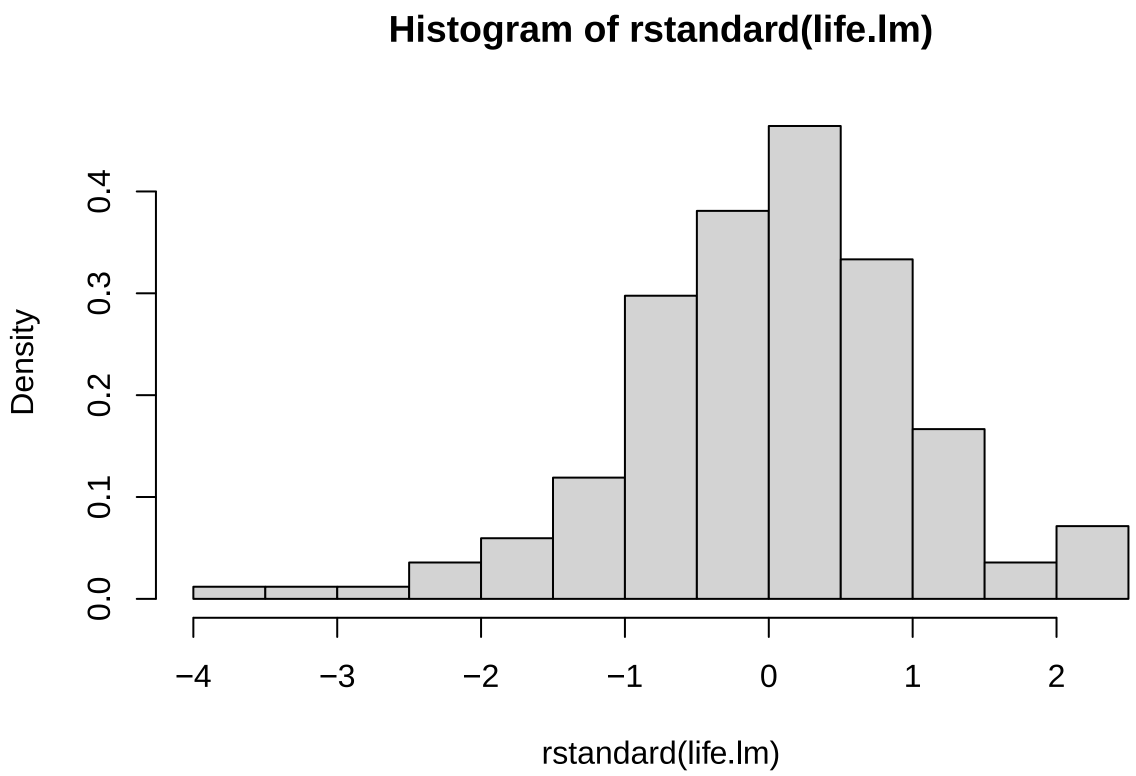
```
life.lm <- lm(Life.expectancy ~ Adult.Mortality
              + Schooling
              + BMI + GDP
              + infant.deaths
              + polio.vacc
              + HIV.AIDS
              + Status, lifedata)
scatterplotMatrix(~Adult.Mortality+BMI+GDP+Schooling+infant.deaths,lifedata, smooth=FALSE)
```



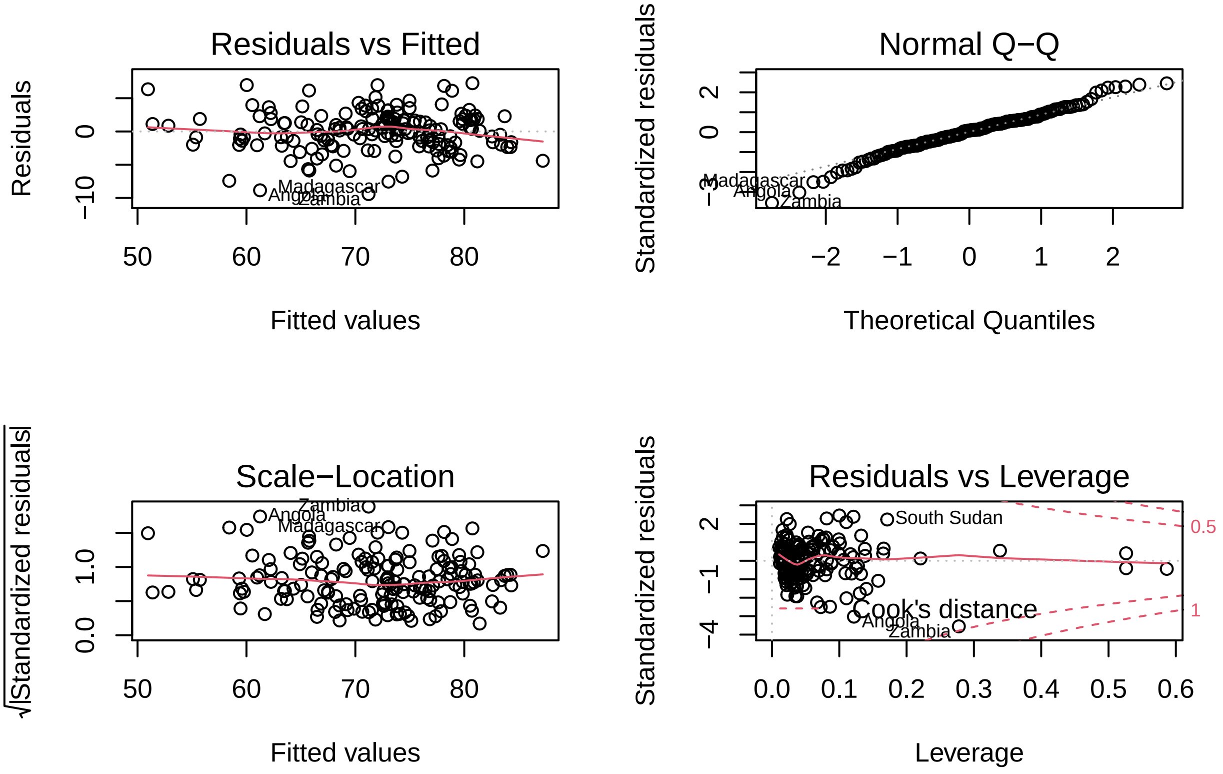
```
summary(life.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     BMI + GDP + infant.deaths + polio.vacc + HIV.AIDS + Status,
##     data = lifedata)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -9.414 -1.728  0.262  1.823  7.251
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.169e+01  2.554e+00  24.156  < 2e-16 ***
## Adult.Mortality    -3.584e-02  3.604e-03  -9.945  < 2e-16 ***
## Schooling           1.126e+00  1.510e-01   7.453 5.75e-12 ***
## BMI                 1.206e-02  1.448e-02   0.833   0.4063
## GDP                 3.168e-05  2.437e-05   1.300   0.1954
## infant.deaths      -2.716e-03  5.196e-03  -0.523   0.6018
## polio.vaccmedium   -7.009e-01  1.277e+00  -0.549   0.5838
## polio.vacchigh      2.073e+00  9.888e-01   2.097   0.0376 *
## HIV.AIDSmedium     -2.128e+00  1.172e+00  -1.815   0.0714 .
## HIV.AIDShigh       -2.457e+00  2.426e+00  -1.013   0.3127
## StatusDeveloping   -1.490e+00  7.986e-01  -1.865   0.0640 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.118 on 157 degrees of freedom
## Multiple R-squared:  0.853,  Adjusted R-squared:  0.8436
## F-statistic: 91.08 on 10 and 157 DF,  p-value: < 2.2e-16
```

```
hist(rstandard(life.lm), probability=TRUE)
```



```
par(mfrow=c(2,2))
plot(life.lm)
```



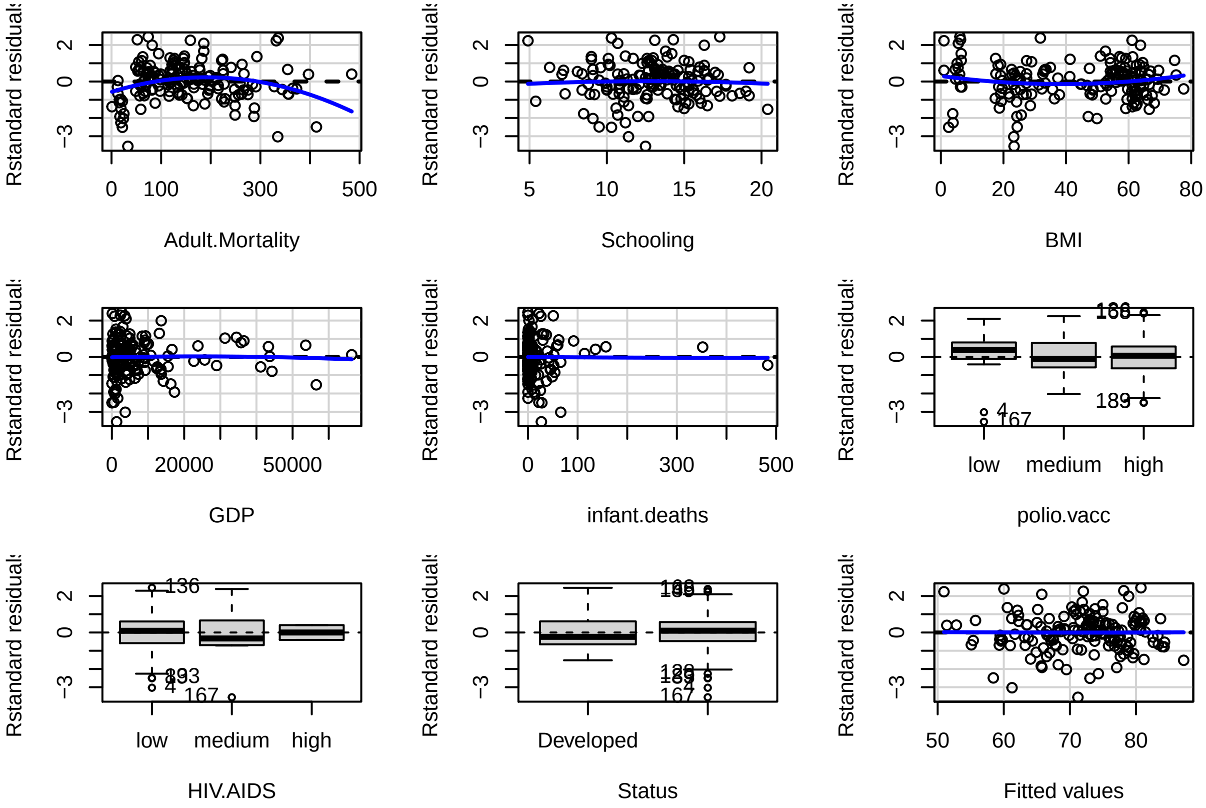
```
shapiro.test(rstandard(life.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(life.lm)
## W = 0.98233, p-value = 0.03102
```

```
ncvTest(life.lm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.057772, Df = 1, p = 0.15143
```

```
residualPlots(life.lm, type="rstandard")
```



```
##                 Test stat Pr(>|Test stat|)
## Adult.Mortality   -4.7355        4.873e-06 ***
## Schooling         -0.4623           0.6445
## BMI                1.4554           0.1476
## GDP               -0.2229           0.8239
## infant.deaths      0.0856           0.9319
## polio.vacc
## HIV.AIDS
## Status
## Tukey test        -0.0506           0.9597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary: Base Model and Diagnostics

The scatterplot matrix shows there are non-linear relationship issues with infant deaths, as well as with GDP. Adult mortality and years of schooling appear to have a strong linear relationship, as well as with BMI and years of schooling.

Homoscedastcity

The standardized residuals plots and scale-location plot show that the homoscedasticity assumptions appear to be met, as the residuals are scattered evenly around 0. The Breusch-Pagan test has a p-value of 0.15, suggesting this test matches the visual of homoscedasticity.

Normality

The histogram appears to be left-skewed with large outliers up to -4. The QQ-plot appears to have heavy tails and that the data set is not normal. The small p-value in the Shapiro-Wilk test confirms this. However, the observations in GDP appear to be right-skewed.

Linearity

The linearity assumption appears to be met for the coefficients, with the exception of Adult.mortality, which appears to have a quadratic response. The residuals vs. fitted values plot appears to be nearly horizontal, indicating linearity in the model.

Influential Observations

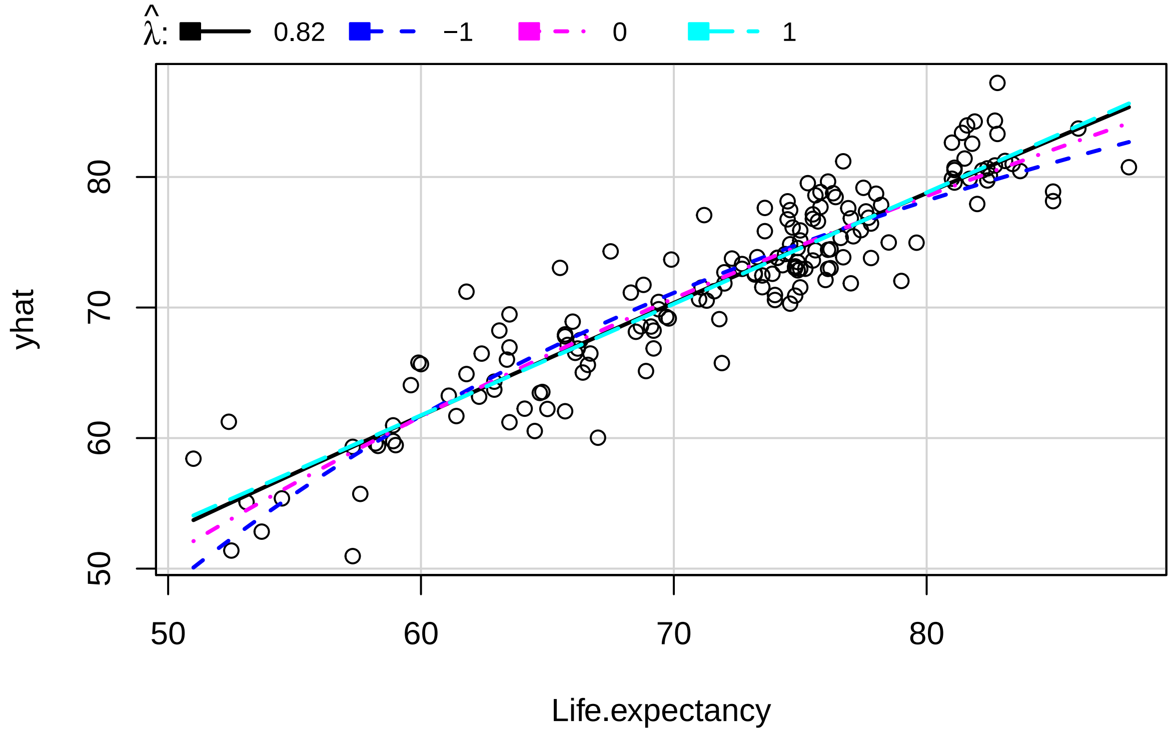
There do not appear to be influential observations. The 'Residuals vs Leverage' plot shows no observations over 0.5.

To address the linearity assumptions in the model, we test the response and the predictor variables

Addressing Assumption Violations

Test response transformation

```r
invResPlot(life.lm)
```

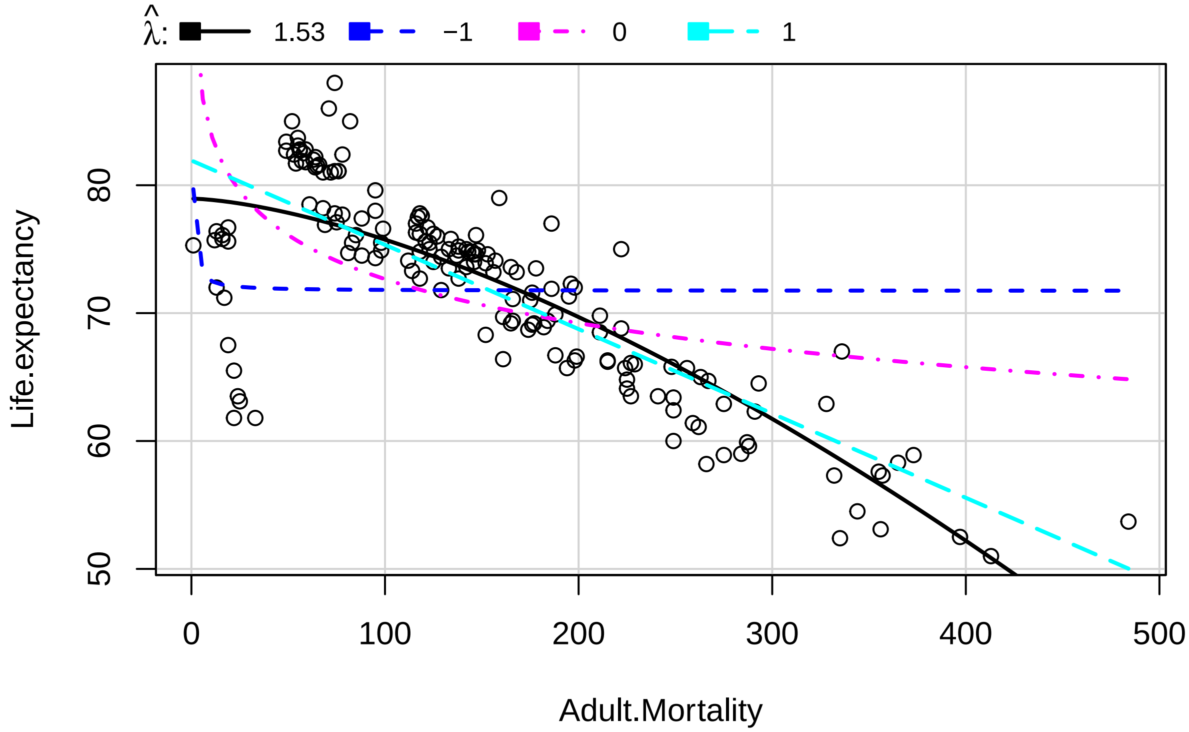


```
##       lambda      RSS
## 1  0.8169073 1300.981
## 2 -1.0000000 1430.842
## 3  0.0000000 1326.793
## 4  1.0000000 1302.245
```

Lambda near 1 suggests no transformation is needed on the response.

Test adult mortality predictor transformation

```
invTranPlot(Life.expectancy~Adult.Mortality, lifedata)$lambda
```



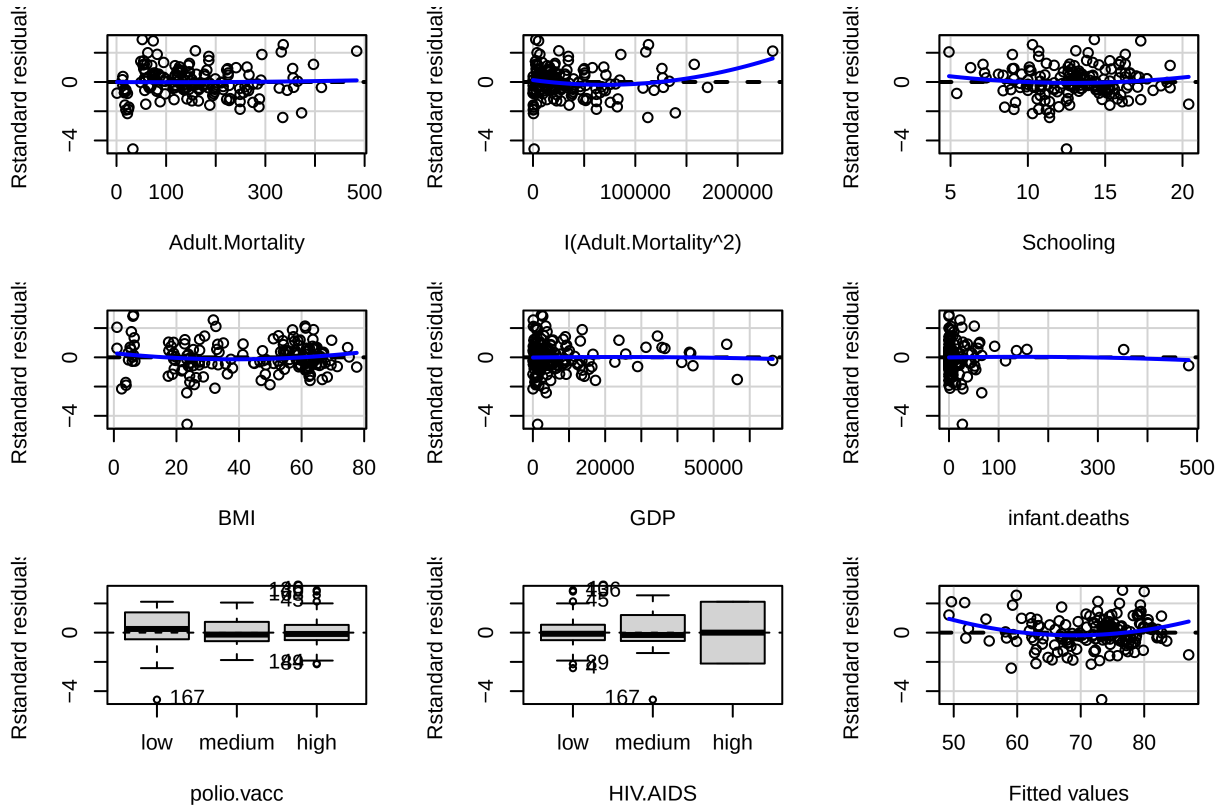
```
## [1]  1.530375 -1.000000  0.000000  1.000000
```

Lambda near 2 suggests adding a quadratic term, I(X^2), to the model

```
life.lm2 <- lm(Life.expectancy ~ Adult.Mortality+ I(Adult.Mortality^2) + Schooling + BMI + GDP + infant
summary(life.lm2)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + I(Adult.Mortality^2) +
##     Schooling + BMI + GDP + infant.deaths + polio.vacc + HIV.AIDS,
##     data = lifedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4988  -1.5141  -0.2264   1.6882   8.4153
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.743e+01  2.034e+00  28.233  < 2e-16 ***
## Adult.Mortality         4.795e-04  9.220e-03   0.052   0.9586
## I(Adult.Mortality^2) -1.122e-04  2.620e-05  -4.282 3.22e-05 ***
## Schooling               1.231e+00  1.299e-01   9.480  < 2e-16 ***
## BMI                    3.846e-03  1.375e-02   0.280   0.7800
## GDP                    5.097e-05  2.315e-05   2.202   0.0291 *
## infant.deaths         -3.730e-03  4.976e-03  -0.749   0.4547
## polio.vaccmedium       1.851e-03  1.219e+00   0.002   0.9988
## polio.vacchigh         1.678e+00  9.507e-01   1.765   0.0795 .
## HIV.AIDSmedium         5.252e-01  1.265e+00   0.415   0.6785
## HIV.AIDShigh           4.964e+00  2.881e+00   1.723   0.0868 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.983 on 157 degrees of freedom
## Multiple R-squared:  0.8654, Adjusted R-squared:  0.8569
## F-statistic:   101 on 10 and 157 DF,  p-value: < 2.2e-16
```

```
residualPlots(life.lm2, type="rstandard")
```

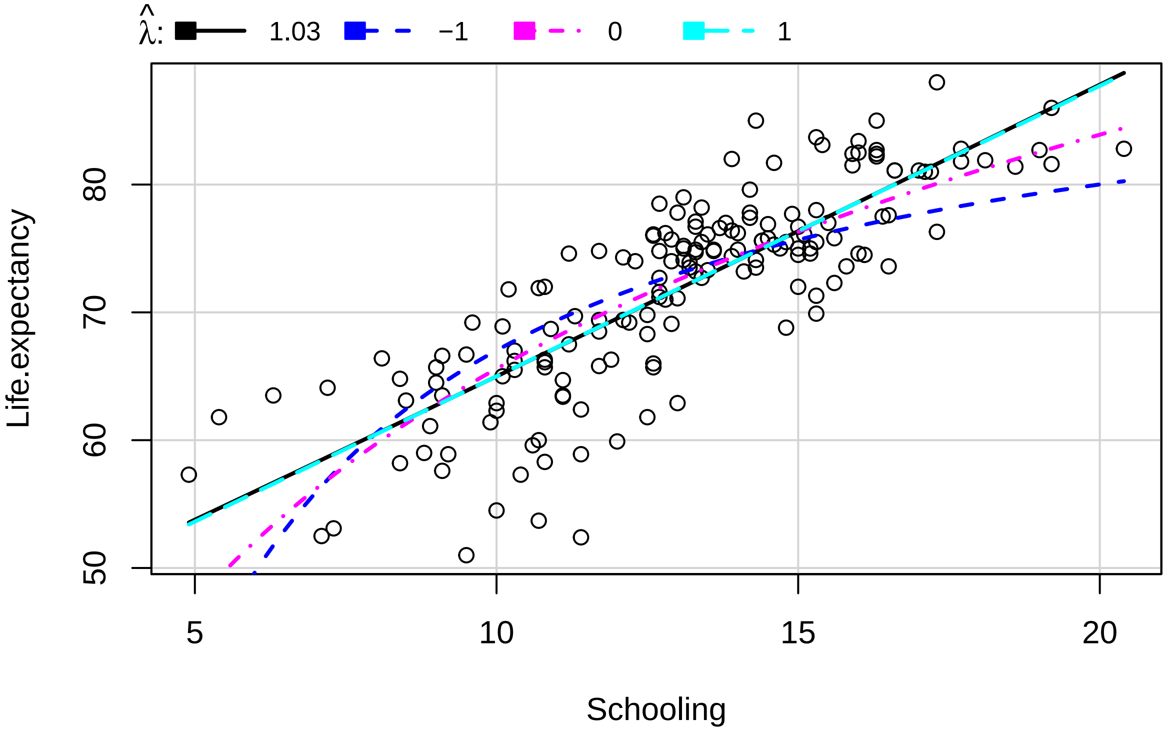


```
##                          Test stat Pr(>|Test stat|)
## Adult.Mortality            -0.5847         0.559600
## I(Adult.Mortality^2)        3.9370         0.000124 ***
## Schooling                   1.0408         0.299573
## BMI                         1.3816         0.169059
## GDP                        -0.1816         0.856100
## infant.deaths              -0.1373         0.890978
## polio.vacc
## HIV.AIDS
## Tukey test              4.4544       8.412e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding a quadratic term to the model appears to improve the linearity assumption for adult mortality.

Test schooling predictor transformation

```
invTranPlot(Life.expectancy~Schooling, lifedata)$lambda
```

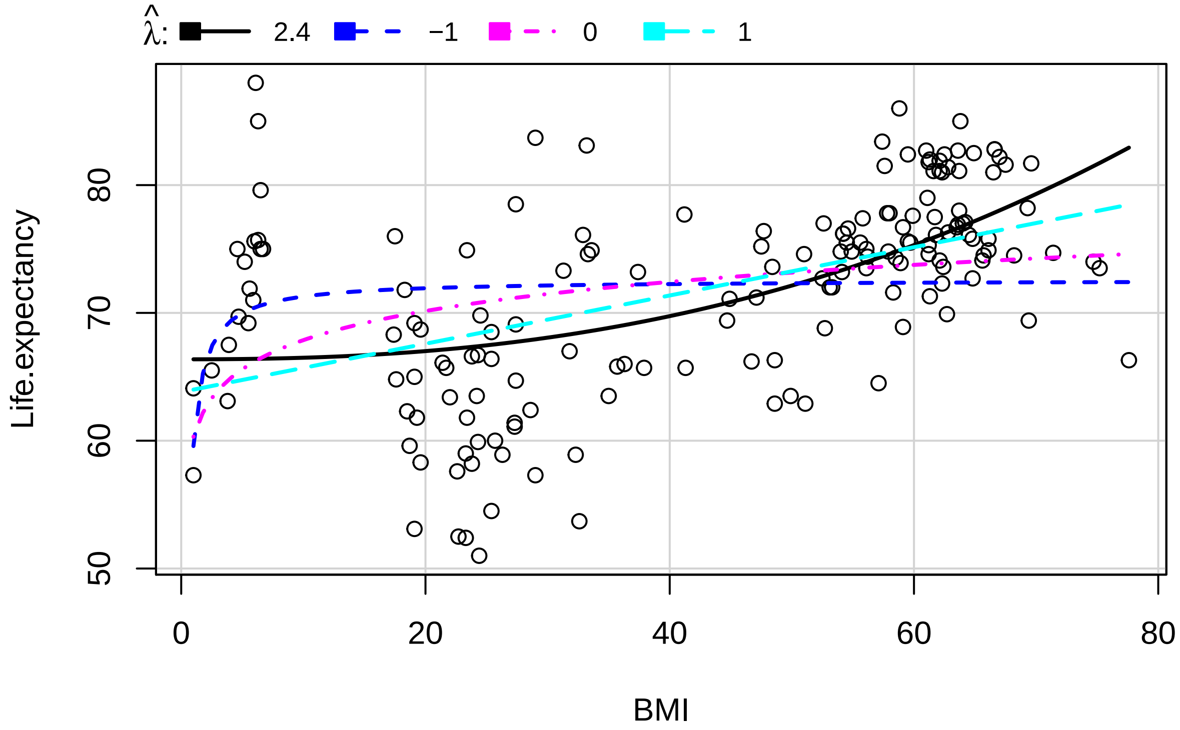


```
## [1]  1.025162 -1.000000  0.000000  1.000000
```

Lambda near 1 suggests no transformation is needed on the predictor

Test BMI predictor transformation

```
invTranPlot(Life.expectancy~BMI, lifedata)$lambda
```



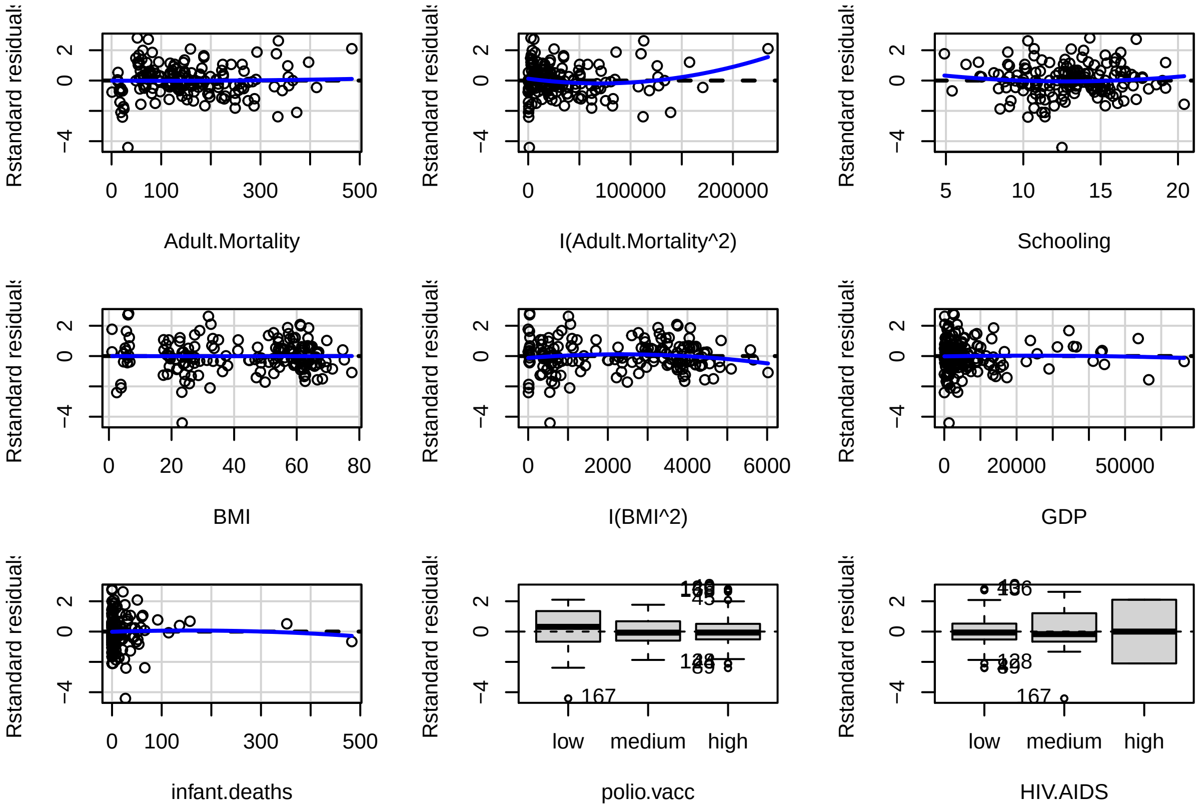
```
## [1]  2.399308 -1.000000  0.000000  1.000000
```

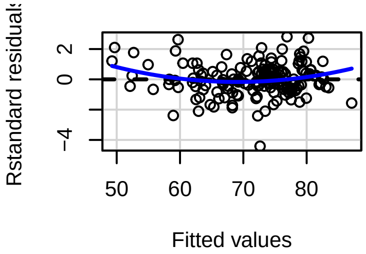
Lambda near 2 suggests adding a quadratic term, I(X^2), to the model

```
life.lm3 <- lm(Life.expectancy ~ Adult.Mortality+I(Adult.Mortality^2) + Schooling + BMI+I(BMI^2) + GDP +
summary(life.lm3)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + I(Adult.Mortality^2) +
##     Schooling + BMI + I(BMI^2) + GDP + infant.deaths + polio.vacc +
##     HIV.AIDS, data = lifedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8433  -1.5009  -0.1865   1.5479   8.1035
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.814e+01  2.091e+00  27.803  < 2e-16 ***
## Adult.Mortality       1.041e-03  9.202e-03   0.113   0.9101
## I(Adult.Mortality^2) -1.102e-04  2.617e-05  -4.211 4.28e-05 ***
## Schooling             1.216e+00  1.300e-01   9.356  < 2e-16 ***
## BMI                    -6.712e-02  5.316e-02  -1.263   0.2086
## I(BMI^2)                9.713e-04  7.030e-04   1.382   0.1691
## GDP                     4.907e-05  2.312e-05   2.122   0.0354 *
## infant.deaths          -2.874e-03  5.001e-03  -0.575   0.5663
## polio.vaccmedium        4.254e-02  1.215e+00   0.035   0.9721
## polio.vacchigh          1.885e+00  9.597e-01   1.964   0.0513 .
## HIV.AIDSmedium          4.460e-01  1.262e+00   0.353   0.7243
## HIV.AIDShigh            4.993e+00  2.873e+00   1.738   0.0841 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.975 on 156 degrees of freedom
## Multiple R-squared:  0.867,  Adjusted R-squared:  0.8577
## F-statistic: 92.49 on 11 and 156 DF,  p-value: < 2.2e-16
```

```
residualPlots(life.lm3, type="rstandard")
```



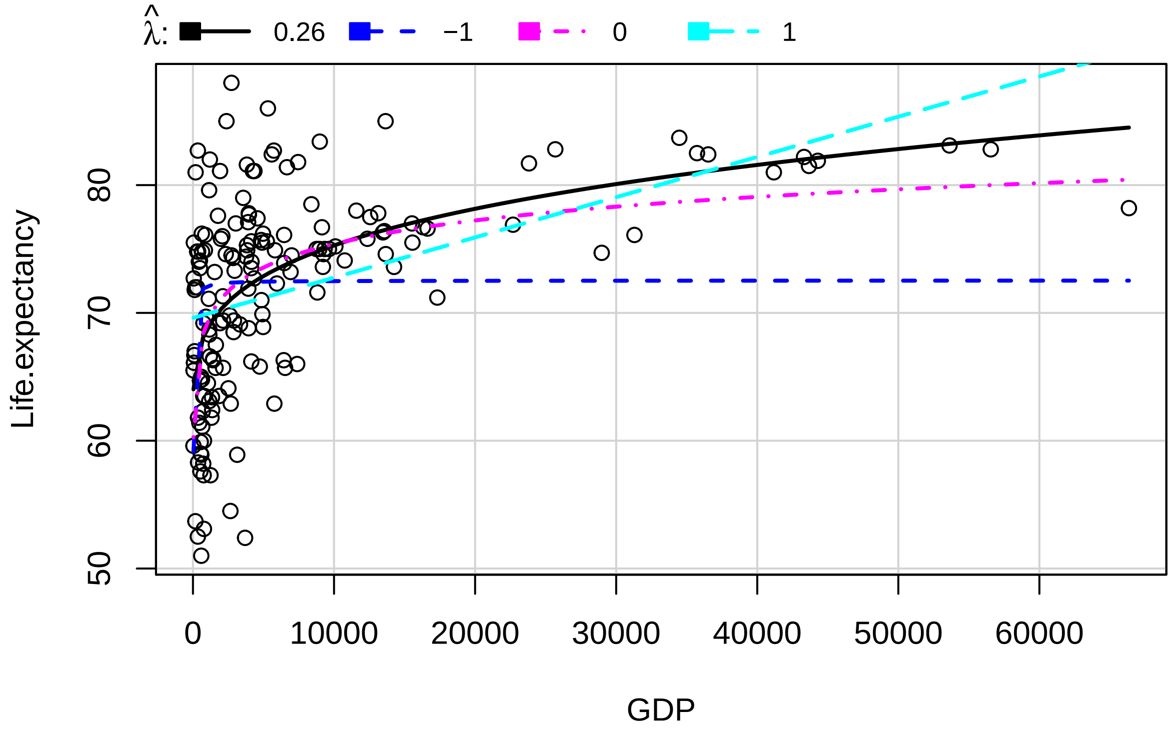


```
##                        Test stat Pr(>|Test stat|)
## Adult.Mortality          -0.5760        0.5654397
## I(Adult.Mortality^2)      3.8009        0.0002067 ***
## Schooling                 0.8352        0.4048734
## BMI                      -0.8663        0.3876537
## I(BMI^2)                 -2.2482        0.0259724 *
## GDP                      -0.2091        0.8346294
## infant.deaths            -0.2645        0.7917537
## polio.vacc
## HIV.AIDS
## Tukey test                4.1476         3.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding a quadratic term to the model appears to improve the linearity assumption for adult BMI.

Test GDP predictor transformation

```
invTranPlot(Life.expectancy~GDP, lifedata)$lambda
```



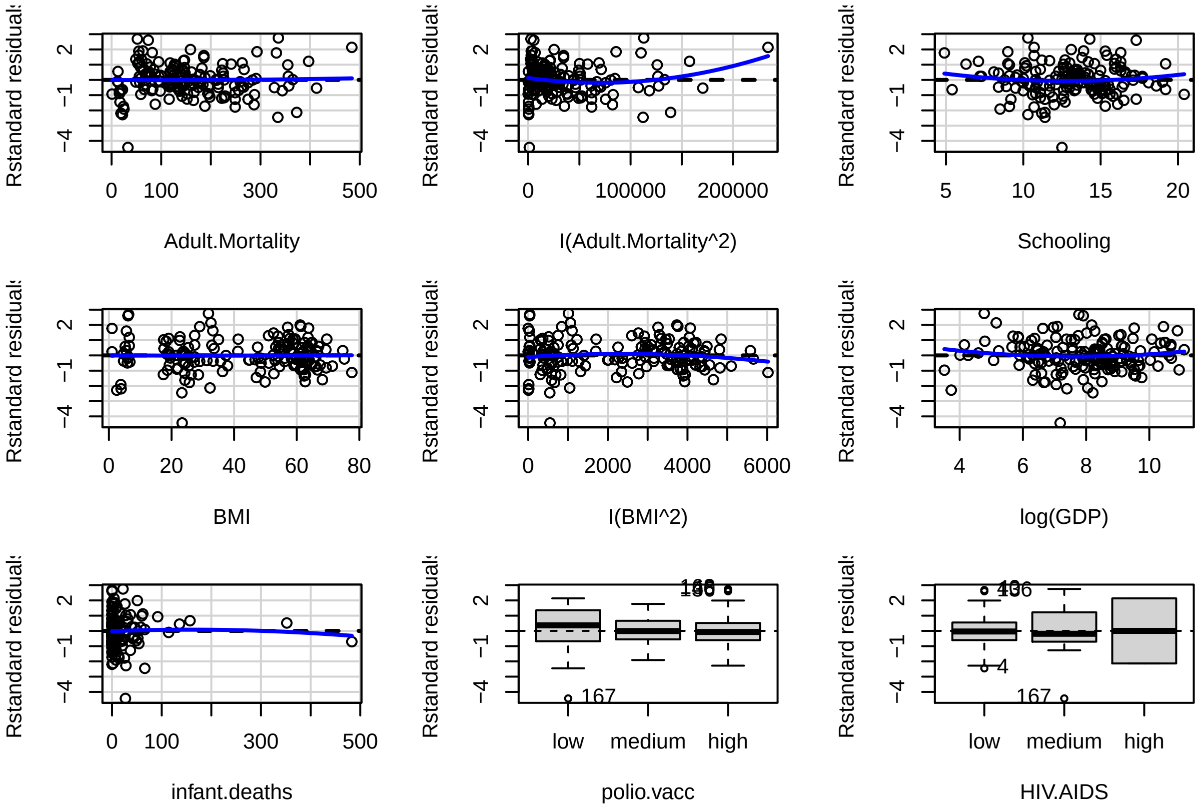
```
## [1]  0.258649 -1.000000  0.000000  1.000000
```

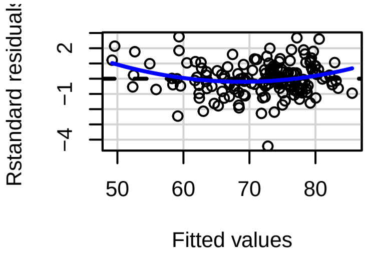
Lambda near 0 suggests trying a logarithmic transformation on the predictor

```
life.lm4 <- lm(Life.expectancy ~ Adult.Mortality+I(Adult.Mortality^2) + Schooling + BMI+I(BMI^2) + log(
summary(life.lm4)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + I(Adult.Mortality^2) +
##     Schooling + BMI + I(BMI^2) + log(GDP) + infant.deaths + polio.vacc +
##     HIV.AIDS, data = lifedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9900  -1.7695  -0.1138   1.7093   7.8163
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.681e+01  2.244e+00  25.321  < 2e-16 ***
## Adult.Mortality      -2.078e-03  9.176e-03  -0.226 0.821183
## I(Adult.Mortality^2) -1.023e-04  2.623e-05  -3.901 0.000142 ***
## Schooling             1.219e+00  1.350e-01   9.033 6.02e-16 ***
## BMI                    -6.984e-02  5.363e-02  -1.302 0.194692
## I(BMI^2)                1.027e-03  7.087e-04   1.449 0.149255
## log(GDP)                2.294e-01  1.813e-01   1.265 0.207719
## infant.deaths          -3.030e-03  5.051e-03  -0.600 0.549438
## polio.vaccmedium        2.088e-02  1.228e+00   0.017 0.986462
## polio.vacchigh          1.937e+00  9.690e-01   1.998 0.047403 *
## HIV.AIDSmedium          4.230e-01  1.274e+00   0.332 0.740270
## HIV.AIDShigh            4.716e+00  2.897e+00   1.628 0.105555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.002 on 156 degrees of freedom
## Multiple R-squared:  0.8646, Adjusted R-squared:  0.8551
## F-statistic: 90.56 on 11 and 156 DF,  p-value: < 2.2e-16
```

```r
residualPlots(life.lm4, type="rstandard")
```





```
##                         Test stat Pr(>|Test stat|)
## Adult.Mortality           -0.7264        0.4686924
## I(Adult.Mortality^2)       3.8234        0.0001903 ***
## Schooling                  1.0610        0.2903464
## BMI                       -0.9008        0.3691016
## I(BMI^2)                  -1.8725        0.0630231 .
## log(GDP)                   1.0900        0.2774088
## infant.deaths             -0.3143        0.7537139
## polio.vacc
## HIV.AIDS
## Tukey test                 4.6098        4.031e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test not done on infant deaths since there are values of zero.

Removing insignificant predictors from model and adding interactions to the model:

```
life.reduced <- update(life.lm4, ~.
                      + I(Adult.Mortality^2):polio.vacc
                      + I(Adult.Mortality^2):Status
                      + I(Adult.Mortality^2):HIV.AIDS
```

```
                       + Schooling:polio.vacc
                       + Schooling:Status
                       + Schooling:HIV.AIDS
                       + HIV.AIDS:polio.vacc
                       - Adult.Mortality
                       - BMI
                       - I(BMI^2)
                       - log(GDP)
                       - infant.deaths
                       - polio.vacc
                       - HIV.AIDS)

summary(life.reduced)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ I(Adult.Mortality^2) + Schooling +
##     I(Adult.Mortality^2):polio.vacc + I(Adult.Mortality^2):Status +
##     I(Adult.Mortality^2):HIV.AIDS + Schooling:polio.vacc + Schooling:Status +
##     Schooling:HIV.AIDS + polio.vacc:HIV.AIDS, data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4786 -1.3654 -0.1237  1.3912  6.3167
##
## Coefficients: (2 not defined because of singularities)
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            6.124e+01  1.570e+00  38.995  < 2e-16
## I(Adult.Mortality^2)                  -3.107e-04  7.323e-05  -4.243 3.80e-05
## Schooling                              1.184e+00  1.671e-01   7.082 4.83e-11
## I(Adult.Mortality^2):polio.vaccmedium  1.563e-05  2.816e-05   0.555  0.57980
## I(Adult.Mortality^2):polio.vacchigh   -1.495e-05  2.431e-05  -0.615  0.53949
## I(Adult.Mortality^2):StatusDeveloping  1.843e-04  6.949e-05   2.652  0.00885
## I(Adult.Mortality^2):HIV.AIDSmedium    9.046e-05  2.874e-05   3.147  0.00198
## I(Adult.Mortality^2):HIV.AIDShigh      6.971e-05  5.523e-05   1.262  0.20883
## Schooling:polio.vaccmedium            -1.862e-01  1.594e-01  -1.168  0.24464
## Schooling:polio.vacchigh               1.320e-01  1.279e-01   1.032  0.30364
## Schooling:StatusDeveloping            -1.688e-01  5.109e-02  -3.303  0.00119
## Schooling:HIV.AIDSmedium              -9.411e-01  2.186e-01  -4.305 2.97e-05
## Schooling:HIV.AIDShigh                -4.767e-01  8.182e-01  -0.583  0.56099
## polio.vaccmedium:HIV.AIDSmedium       -3.021e+00  3.785e+00  -0.798  0.42614
## polio.vacchigh:HIV.AIDSmedium          3.109e+00  2.975e+00   1.045  0.29770
## polio.vaccmedium:HIV.AIDShigh                 NA         NA      NA       NA
## polio.vacchigh:HIV.AIDShigh                   NA         NA      NA       NA
##
## (Intercept)                           ***
## I(Adult.Mortality^2)                  ***
## Schooling                             ***
## I(Adult.Mortality^2):polio.vaccmedium
## I(Adult.Mortality^2):polio.vacchigh
## I(Adult.Mortality^2):StatusDeveloping **
## I(Adult.Mortality^2):HIV.AIDSmedium   **
## I(Adult.Mortality^2):HIV.AIDShigh
## Schooling:polio.vaccmedium
## Schooling:polio.vacchigh
## Schooling:StatusDeveloping            **
## Schooling:HIV.AIDSmedium              ***
## Schooling:HIV.AIDShigh
## polio.vaccmedium:HIV.AIDSmedium
## polio.vacchigh:HIV.AIDSmedium
## polio.vaccmedium:HIV.AIDShigh
## polio.vacchigh:HIV.AIDShigh
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.631 on 153 degrees of freedom
## Multiple R-squared:  0.898,  Adjusted R-squared:  0.8887
## F-statistic: 96.24 on 14 and 153 DF,  p-value: < 2.2e-16
```

Small p-value (< 2.2e-16) suggests interaction effects need to stay in the model.

Now, we choose which interaction effects need to stay in the model.

```
life.reduced1 <- update(life.reduced, ~ . - I(Adult.Mortality^2):polio.vacc)
anova(life.reduced1, life.reduced)$‘Pr(>F)‘
```

```
## [1]        NA 0.3057863
```

Large p-value (0.305786) suggests we can remove I(Adult.Mortality^2):polio.vacc from model.

```
life.reduced2 <- update(life.reduced1, ~ . - I(Adult.Mortality^2):HIV.AIDS)
anova(life.reduced2, life.reduced1)$‘Pr(>F)‘
```

```
## [1]           NA 7.937699e-05
```

Small p-value (7.937699e-05) suggests we need to keep I(Adult.Mortality^2):HIV.AIDS in the model.

```
life.reduced3 <- update(life.reduced1, ~ . - Schooling:Status)
anova(life.reduced3, life.reduced1)$‘Pr(>F)‘
```

```
## [1]           NA 0.0005545418
```

Small p-value (0.0005545418) suggests we need to keep Schooling:Status in the model.

```
life.reduced4 <- update(life.reduced1, ~ . - Schooling:HIV.AIDS)
anova(life.reduced4, life.reduced1)$‘Pr(>F)‘
```

```
## [1]           NA 6.212583e-06
```

Small p-value (6.212583e-06) suggests we need to keep Schooling:HIV.AIDS from model.

```
life.reduced5 <- update(life.reduced1, ~ . - HIV.AIDS:polio.vacc)
anova(life.reduced5, life.reduced1)$`Pr(>F)`
```

```
## [1]        NA 0.3127716
```

Large p-value (0.312771) suggests we can remove HIV.AIDS:polio.vacc from model.

```
life.reduced6 <- update(life.reduced5, ~ . - Schooling:polio.vacc)
anova(life.reduced6, life.reduced5)$`Pr(>F)`
```

```
## [1]          NA 0.006823202
```

Small p-value (0.006823202) suggests we need to keep Schooling:polio.vacc in the model.

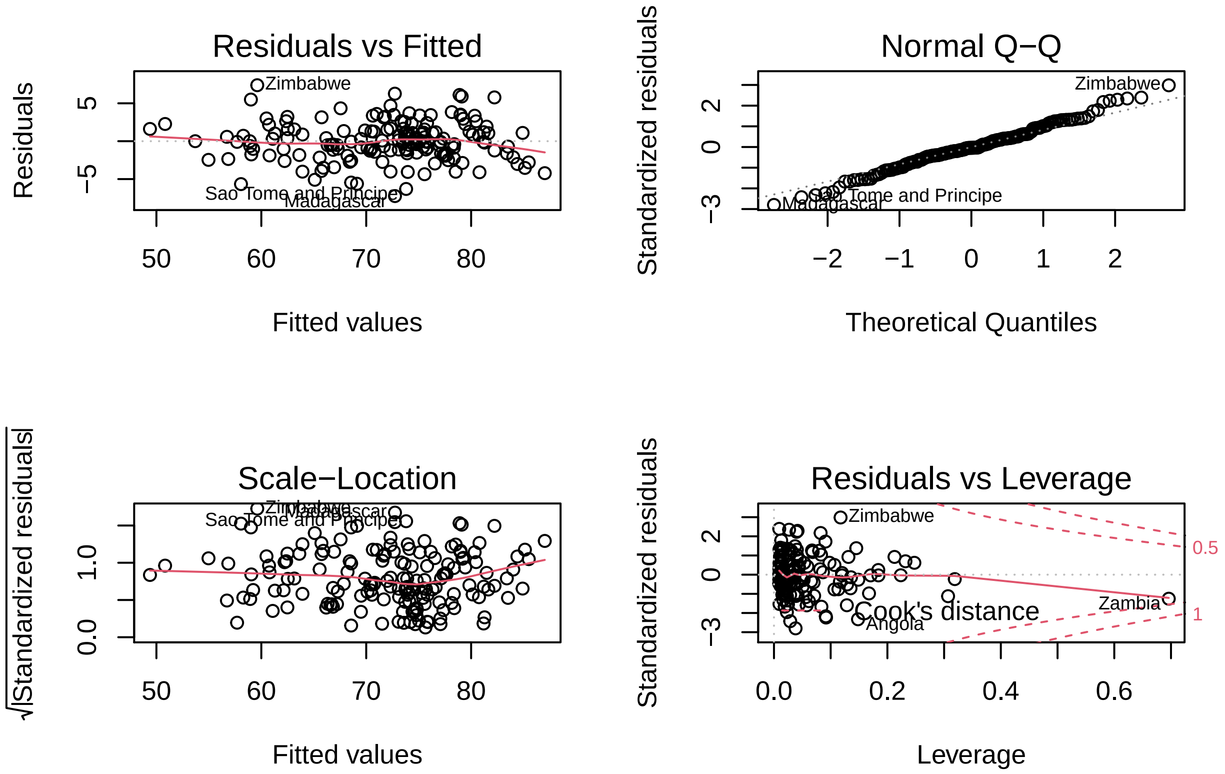
Diagnostics for Final Model:

```
summary(life.reduced5)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ I(Adult.Mortality^2) + Schooling +
##     I(Adult.Mortality^2):Status + I(Adult.Mortality^2):HIV.AIDS +
##     Schooling:polio.vacc + Schooling:Status + Schooling:HIV.AIDS,
##     data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2427 -1.4298 -0.0901  1.3582  7.3874
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            6.083e+01  1.560e+00  38.997  < 2e-16
## I(Adult.Mortality^2)                  -3.240e-04  6.935e-05  -4.672 6.37e-06
## Schooling                              1.232e+00  1.287e-01   9.572  < 2e-16
## I(Adult.Mortality^2):StatusDeveloping  1.921e-04  6.950e-05   2.765  0.00638
## I(Adult.Mortality^2):HIV.AIDSmedium    9.055e-05  1.769e-05   5.119 8.87e-07
## I(Adult.Mortality^2):HIV.AIDShigh      9.165e-05  3.653e-05   2.509  0.01314
## Schooling:polio.vaccmedium            -1.360e-01  1.050e-01  -1.295  0.19713
## Schooling:polio.vacchigh               1.075e-01  7.637e-02   1.407  0.16131
## Schooling:StatusDeveloping            -1.762e-01  5.070e-02  -3.475  0.00066
## Schooling:HIV.AIDSmedium              -8.282e-01  1.824e-01  -4.541 1.11e-05
## Schooling:HIV.AIDShigh                -8.414e-01  6.165e-01  -1.365  0.17428
##
## (Intercept)                           ***
## I(Adult.Mortality^2)                  ***
## Schooling                             ***
## I(Adult.Mortality^2):StatusDeveloping **
## I(Adult.Mortality^2):HIV.AIDSmedium   ***
## I(Adult.Mortality^2):HIV.AIDShigh     *
## Schooling:polio.vaccmedium
## Schooling:polio.vacchigh
## Schooling:StatusDeveloping            ***
## Schooling:HIV.AIDSmedium              ***
## Schooling:HIV.AIDShigh
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.637 on 157 degrees of freedom
## Multiple R-squared:  0.8949, Adjusted R-squared:  0.8882
## F-statistic: 133.6 on 10 and 157 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(life.reduced5)
```

```
## Warning: not plotting observations with leverage one:
##   84, 144
```



```r
shapiro.test(rstandard(life.reduced5))
```

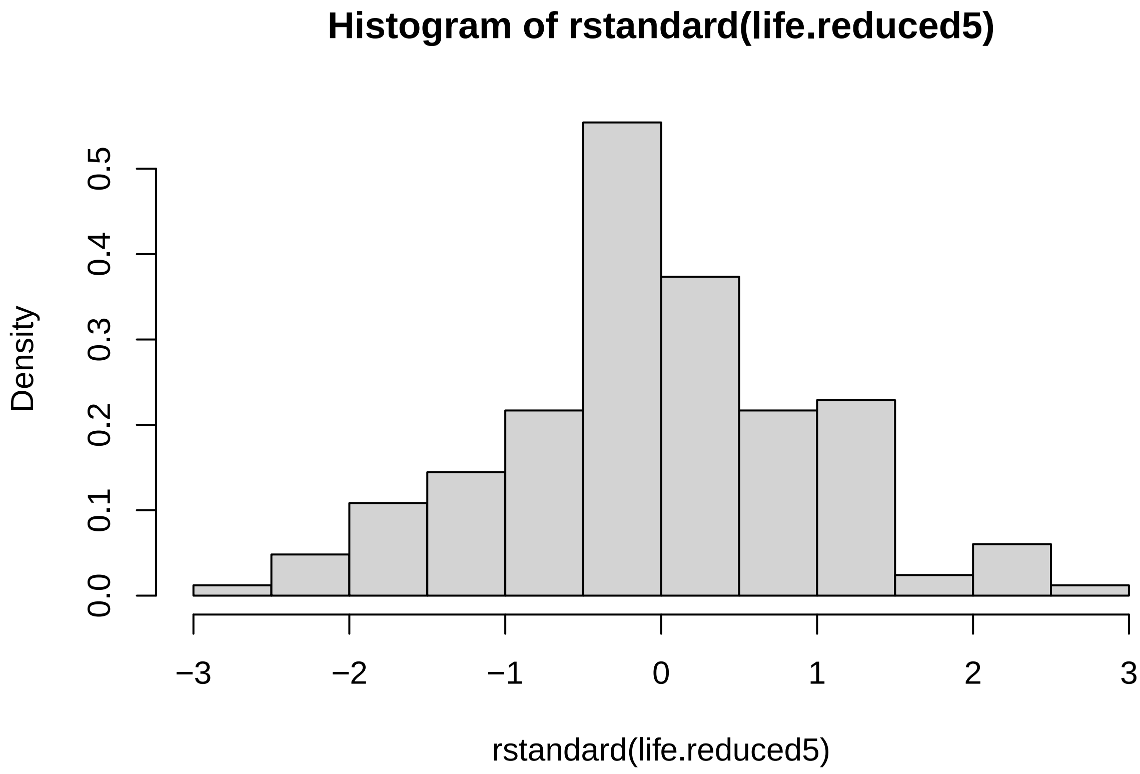
```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  rstandard(life.reduced5)
## W = 0.99195, p-value = 0.4801
```

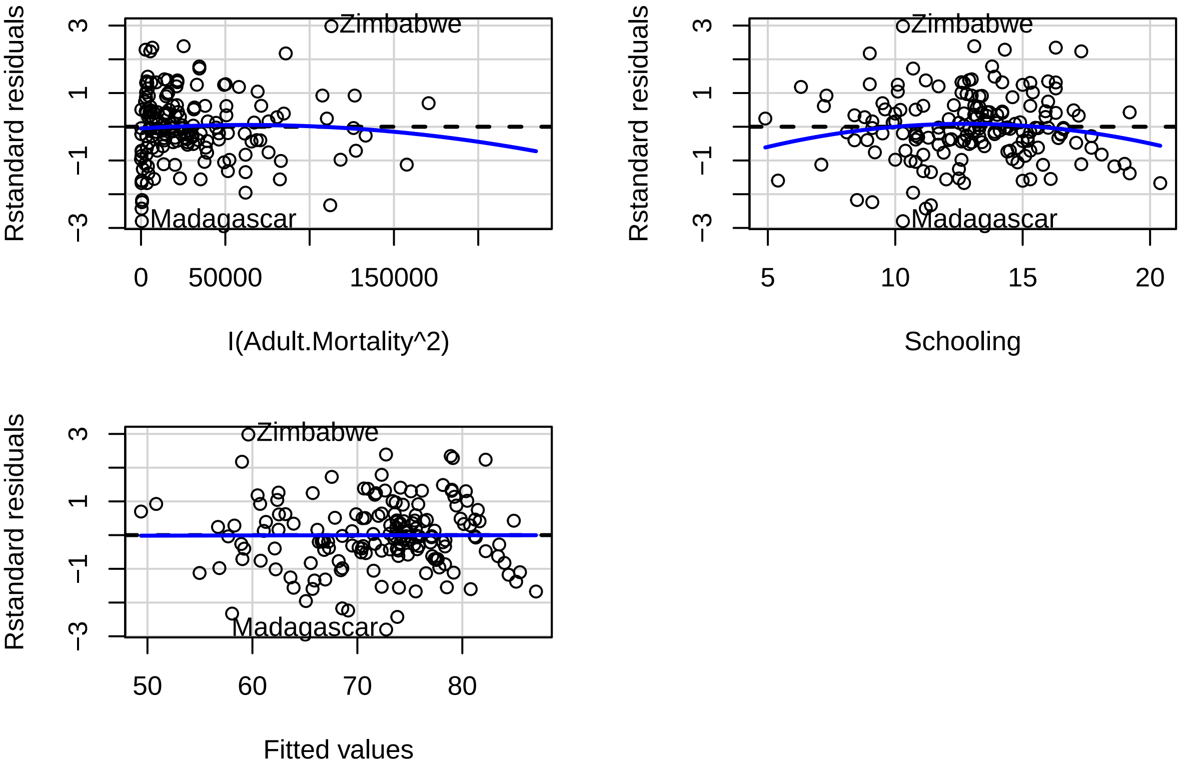
```
ncvTest(life.reduced5)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.04413981, Df = 1, p = 0.83359
```

```
par(mfrow=c(1,1))
hist(rstandard(life.reduced5), probability=TRUE)
```



```
residualPlots(life.reduced5, id=TRUE, quadratic=TRUE, type='rstandard', tests=FALSE)
```



Summary: Final Model and Diagnostics

Homoscedastcity

The standardized residuals plots and scale-location plot show that the homoscedasticity assumptions appear to be met, as the residuals are scattered evenly around 0. The Breusch-Pagan test has a p-value of 0.83359, suggesting this test matches the visual of homoscedasticity.

Normality

The histogram looks normally distributed. The QQ-plot appears to be normal with no heavy tails. The large p-value in the Shapiro-Wilk test (p-value = 0.4801) confirms this.

Linearity

The linearity assumption appears to be met for the coefficients. The residuals vs. fitted values plot appears to be nearly horizontal, indicating linearity in the model.

Influential Observations

There do not appear to be influential observations. The 'Residuals vs Leverage' plot shows no observations over 0.5.

R^2, adjusted R^2, sigma-hat

Base Model

```r
summary(life.lm)$r.squared
```

```
## [1] 0.8529651
```

```r
summary(life.lm)$adj.r.squared
```

```
## [1] 0.8435998
```

```r
summary(life.lm)$sigma
```

```
## [1] 3.118393
```

Final Model

```r
summary(life.reduced5)$r.squared
```

```
## [1] 0.8948686
```

```r
summary(life.reduced5)$adj.r.squared
```

```
## [1] 0.8881723
```

```r
summary(life.reduced5)$sigma
```

```
## [1] 2.636859
```

The model improved overall from the base model, with an improvement to the adjusted R^2 value, meaning there was not a "penalty" for adding the additional interactions to the model, as well as a reduction of the standard error, resulting in a more accurate model. R^2 improved as well, solely due to the addition of the interactions.

Life Expectancy - Collinearity and Variable Selection

Kasia Krueger

10/10/2021

Week 6

Cleaning up data

Removing high influential points (errors in years of schooling as found in Week 3 Regression diagnostics.)

```
lifedata <- lifedata[-which(rownames(lifedata) == "Burkina Faso"),]
#Possible error in years of schooling (26)
lifedata <- lifedata[-which(rownames(lifedata) == "Equatorial Guinea"),]
#Possible error in years of schooling (32)
lifedata <- lifedata[-which(rownames(lifedata) == "Cuba"),]
#Error in years of schooling (92)
lifedata <- lifedata[-which(rownames(lifedata) == "Eritrea"),]
# Error in years of schooling (255)
lifedata <- lifedata[-which(rownames(lifedata) == "India"),]
#High cook’s distance + possible error in infant deaths (800/1000)
```

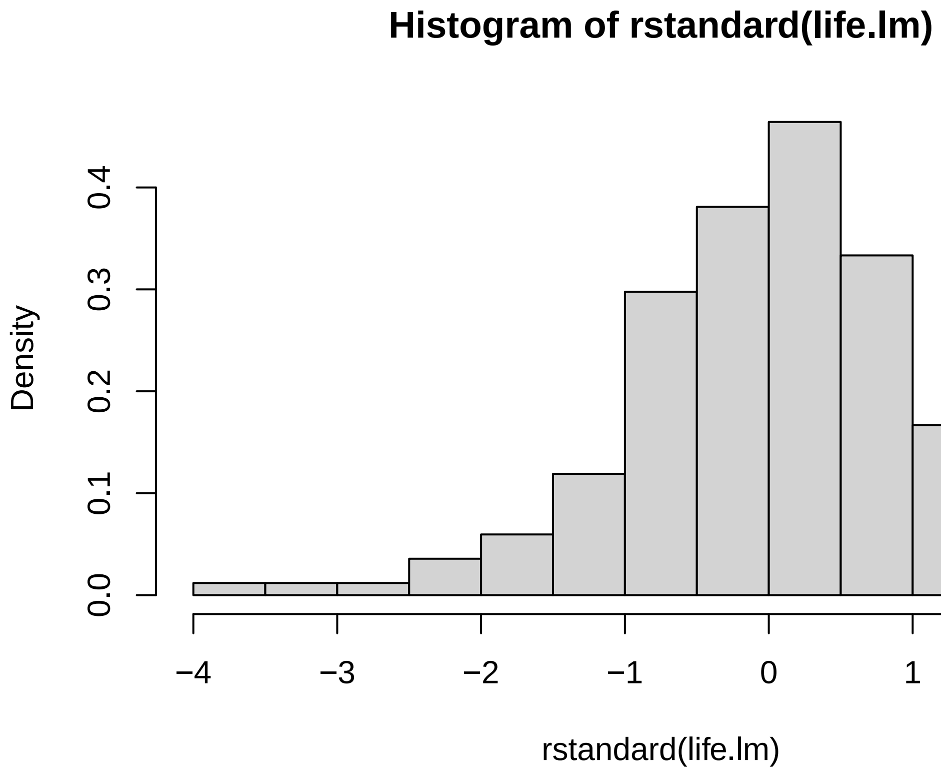
```
lifedata$polio.vacc <- as.factor(lifedata$polio.vacc)
lifedata$HIV.AIDS <- as.factor(lifedata$HIV.AIDS)
lifedata$Status <- as.factor(lifedata$Status)

levels(lifedata$polio.vacc) <- c(’low’, ’medium’, ’high’)
levels(lifedata$HIV.AIDS) <- c(’low’, ’medium’, ’high’)
```

Base Model

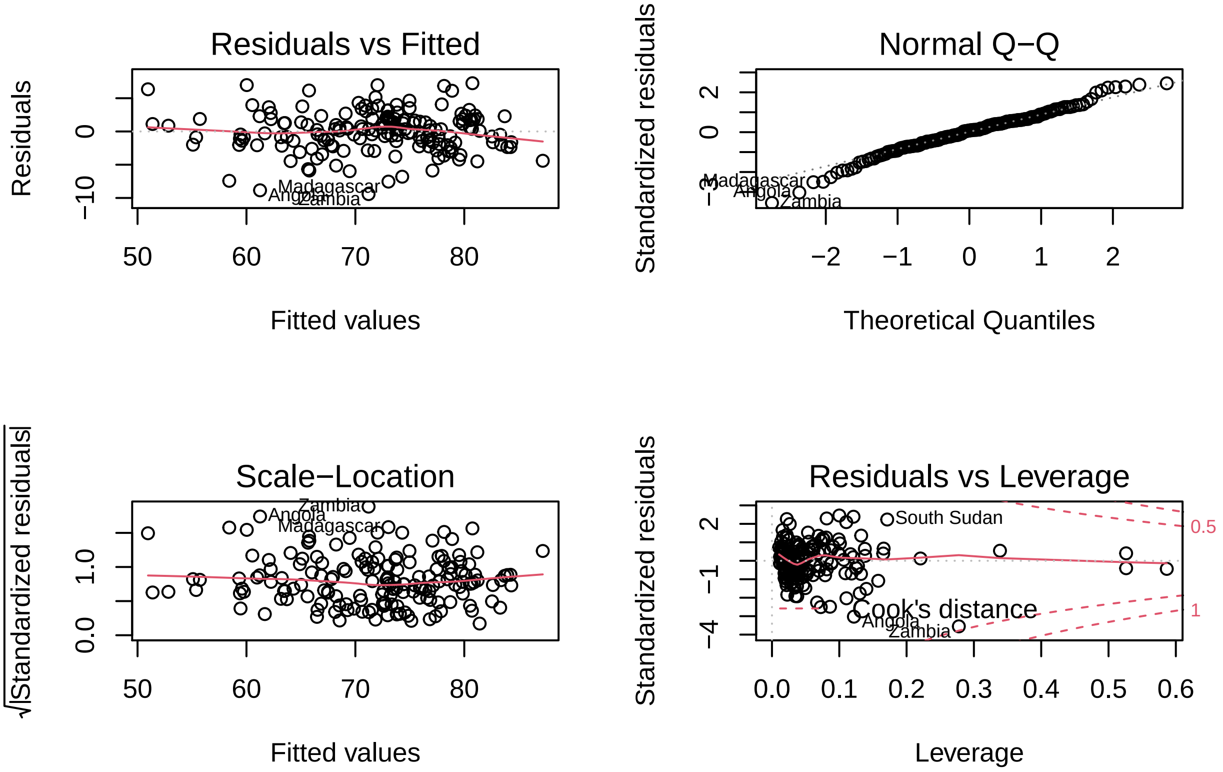
```
life.lm <- lm(Life.expectancy ~ Adult.Mortality
              + Schooling
              + BMI + GDP
              + infant.deaths
              + polio.vacc
              + HIV.AIDS
              + Status, lifedata)


summary(life.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     BMI + GDP + infant.deaths + polio.vacc + HIV.AIDS + Status,
##     data = lifedata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.414 -1.728  0.262  1.823  7.251
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.169e+01  2.554e+00  24.156  < 2e-16 ***
## Adult.Mortality   -3.584e-02  3.604e-03  -9.945  < 2e-16 ***
## Schooling          1.126e+00  1.510e-01   7.453 5.75e-12 ***
## BMI                1.206e-02  1.448e-02   0.833   0.4063
## GDP                3.168e-05  2.437e-05   1.300   0.1954
## infant.deaths     -2.716e-03  5.196e-03  -0.523   0.6018
## polio.vaccmedium -7.009e-01  1.277e+00  -0.549   0.5838
## polio.vacchigh    2.073e+00  9.888e-01   2.097   0.0376 *
## HIV.AIDSmedium   -2.128e+00  1.172e+00  -1.815   0.0714 .
## HIV.AIDShigh     -2.457e+00  2.426e+00  -1.013   0.3127
## StatusDeveloping -1.490e+00  7.986e-01  -1.865   0.0640 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.118 on 157 degrees of freedom
## Multiple R-squared:  0.853,  Adjusted R-squared:  0.8436
## F-statistic: 91.08 on 10 and 157 DF,  p-value: < 2.2e-16
```

```r
hist(rstandard(life.lm), probability=TRUE)
```



```r
par(mfrow=c(2,2))
plot(life.lm)
```



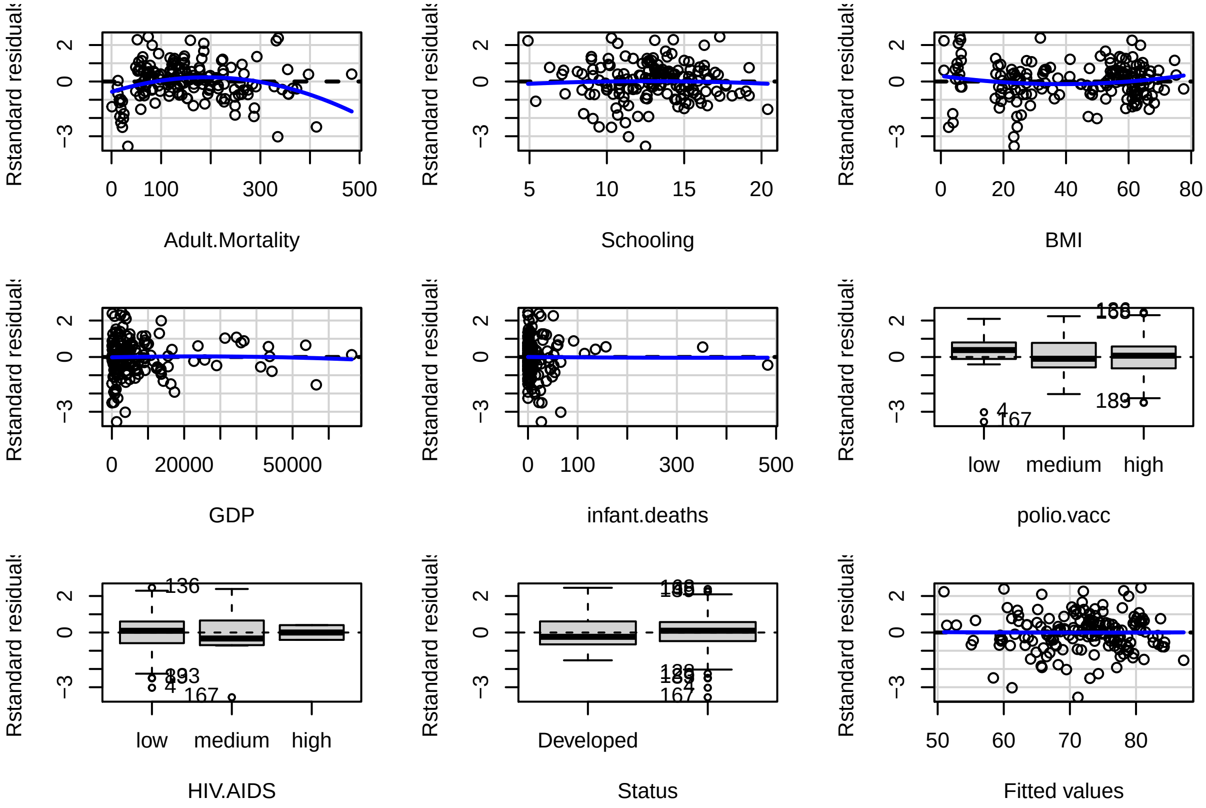
```
shapiro.test(rstandard(life.lm))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  rstandard(life.lm)
## W = 0.98233, p-value = 0.03102
```

```
ncvTest(life.lm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.057772, Df = 1, p = 0.15143
```

```
residualPlots(life.lm, type="rstandard")
```



```
##                   Test stat Pr(>|Test stat|)
## Adult.Mortality   -4.7355        4.873e-06 ***
## Schooling         -0.4623           0.6445
## BMI                1.4554           0.1476
## GDP               -0.2229           0.8239
## infant.deaths      0.0856           0.9319
## polio.vacc
## HIV.AIDS
## Status
## Tukey test        -0.0506           0.9597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Calculate VIFS
```

```
‘‘‘r
car::vif(life.lm)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## Adult.Mortality 2.050942  1        1.432111
## Schooling       3.167756  1        1.779819
## BMI             1.608322  1        1.268196
## GDP             1.328266  1        1.152504
## infant.deaths   1.188266  1        1.090076
## polio.vacc      1.470286  2        1.101160
## HIV.AIDS        1.541526  2        1.114263
## Status          1.573814  1        1.254517
```

Condition number

```
life.pc <- prcomp(~ Adult.Mortality
              + Schooling
              + BMI
              + GDP
              + infant.deaths
              - Life.expectancy, lifedata, scale=TRUE)
PC.sdev <- life.pc$sdev
PC.sdev[1] / PC.sdev
```

```
## [1] 1.000000 1.704177 1.950471 2.011996 2.713601
```

The full model did not detect any issues with collinearity, but I will explore this more during the variable selection process.

Life Expectancy - Collinearity and Variable Selection

Kasia Krueger

10/15/2021

Week 7

Cleaning up data

Removing high influential points (errors in years of schooling as found in Week 3 Regression diagnostics.)

```
lifedata <- lifedata[-which(rownames(lifedata) == "Burkina Faso"),]
#Possible error in years of schooling (26)
lifedata <- lifedata[-which(rownames(lifedata) == "Equatorial Guinea"),]
#Possible error in years of schooling (32)
lifedata <- lifedata[-which(rownames(lifedata) == "Cuba"),]
#Error in years of schooling (92)
lifedata <- lifedata[-which(rownames(lifedata) == "Eritrea"),]
# Error in years of schooling (255)
lifedata <- lifedata[-which(rownames(lifedata) == "India"),]
#High cook’s distance + possible error in infant deaths (800/1000)
```

```
lifedata$polio.vacc <- as.factor(lifedata$polio.vacc)
lifedata$HIV.AIDS <- as.factor(lifedata$HIV.AIDS)
lifedata$Status <- as.factor(lifedata$Status)

levels(lifedata$polio.vacc) <- c(’low’, ’medium’, ’high’)
levels(lifedata$HIV.AIDS) <- c(’low’, ’medium’, ’high’)
```

Full Model

```
life.lm <- lm(Life.expectancy ~ Adult.Mortality
              + Schooling
              + BMI + GDP
              + infant.deaths
              + polio.vacc
              + HIV.AIDS
              + Status, lifedata)


summary(life.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     BMI + GDP + infant.deaths + polio.vacc + HIV.AIDS + Status,
##     data = lifedata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.414 -1.728  0.262  1.823  7.251
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.169e+01  2.554e+00  24.156  < 2e-16 ***
## Adult.Mortality  -3.584e-02  3.604e-03  -9.945  < 2e-16 ***
## Schooling         1.126e+00  1.510e-01   7.453 5.75e-12 ***
## BMI               1.206e-02  1.448e-02   0.833   0.4063
## GDP               3.168e-05  2.437e-05   1.300   0.1954
## infant.deaths    -2.716e-03  5.196e-03  -0.523   0.6018
## polio.vaccmedium -7.009e-01  1.277e+00  -0.549   0.5838
## polio.vacchigh    2.073e+00  9.888e-01   2.097   0.0376 *
## HIV.AIDSmedium   -2.128e+00  1.172e+00  -1.815   0.0714 .
## HIV.AIDShigh     -2.457e+00  2.426e+00  -1.013   0.3127
## StatusDeveloping -1.490e+00  7.986e-01  -1.865   0.0640 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.118 on 157 degrees of freedom
## Multiple R-squared:  0.853,  Adjusted R-squared:  0.8436
## F-statistic: 91.08 on 10 and 157 DF,  p-value: < 2.2e-16
```

```
summary(life.lm)$sigma
```

```
## [1] 3.118393
```

Calculate VIFS

```
car::vif(life.lm)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## Adult.Mortality 2.050942  1        1.432111
## Schooling       3.167756  1        1.779819
## BMI             1.608322  1        1.268196
## GDP             1.328266  1        1.152504
## infant.deaths   1.188266  1        1.090076
## polio.vacc      1.470286  2        1.101160
## HIV.AIDS        1.541526  2        1.114263
## Status          1.573814  1        1.254517
```

Condition number

```
life.pc <- prcomp(~ Adult.Mortality
              + Schooling
              + BMI
              + GDP
              + infant.deaths
              - Life.expectancy, lifedata, scale=TRUE)
PC.sdev <- life.pc$sdev
PC.sdev[1] / PC.sdev
```

```
## [1] 1.000000 1.704177 1.950471 2.011996 2.713601
```

Summary: Collinearity in Full Model

The full model did not detect any issues with collinearity. Next, we explore this topic more during the variable selection process.

Best p-Variable Models

```
life.rs <- leaps::regsubsets(Life.expectancy ~ Adult.Mortality
              + Schooling
              + BMI + GDP
              + infant.deaths
              + polio.vacc
              + HIV.AIDS
              + Status, lifedata)

summary(life.rs)
```

```
## Subset selection object
## Call: regsubsets.formula(Life.expectancy ~ Adult.Mortality + Schooling +
##     BMI + GDP + infant.deaths + polio.vacc + HIV.AIDS + Status,
##     lifedata)
## 10 Variables  (and intercept)
##                   Forced in Forced out
## Adult.Mortality       FALSE      FALSE
## Schooling             FALSE      FALSE
## BMI                   FALSE      FALSE
## GDP                   FALSE      FALSE
## infant.deaths         FALSE      FALSE
## polio.vaccmedium      FALSE      FALSE
## polio.vacchigh        FALSE      FALSE
## HIV.AIDSmedium        FALSE      FALSE
## HIV.AIDShigh          FALSE      FALSE
## StatusDeveloping      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Adult.Mortality Schooling BMI GDP infant.deaths polio.vaccmedium
## 1  ( 1 ) " "             "*"       " " " " " "           " "
## 2  ( 1 ) "*"             "*"       " " " " " "           " "
## 3  ( 1 ) "*"             "*"       " " " " " "           " "
## 4  ( 1 ) "*"             "*"       " " " " " "           " "
## 5  ( 1 ) "*"             "*"       " " " " " "           " "
## 6  ( 1 ) "*"             "*"       " " "*" " "           " "
## 7  ( 1 ) "*"             "*"       " " "*" " "           " "
## 8  ( 1 ) "*"             "*"       "*" "*" " "           " "
##          polio.vacchigh HIV.AIDSmedium HIV.AIDShigh StatusDeveloping
## 1  ( 1 ) " "            " "            " "          " "
## 2  ( 1 ) " "            " "            " "          " "
## 3  ( 1 ) "*"            " "            " "          " "
## 4  ( 1 ) "*"            " "            " "          "*"
## 5  ( 1 ) "*"            "*"            " "          "*"
## 6  ( 1 ) "*"            "*"            " "          "*"
## 7  ( 1 ) "*"            "*"            "*"          "*"
## 8  ( 1 ) "*"            "*"            "*"          "*"
```

- Best 1-variable model is Y = β_0+$\beta_2 X_2$
- Best 2-variable model is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2$
- Best 3-variable model is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7$
- Best 4-variable model is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7 + \beta_{10} X_{10}$
- Best 5-variable model is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10}$
- Best 6-variable model is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10}$
- Best 7-variable model is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10}$
- Best 8-variable model is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10}$

Patterns:

- X_2 *Schooling* is in all the models.
- X_1 *Adult Mortality* is in all the models containing two or more variables.
- X_7 *Polio Vaccine - High* is in all the models containing three or more variables.
- X_{10} *Status - Developing* is in all the models containing four or more variables.

Forward selection, backward elimination, and stepwise selection

Forward AIC

```
mod.null <- lm(Life.expectancy ~ 1, lifedata) # null model with no predictors
step(mod.null, scope=formula(life.lm), direction='forward', trace=FALSE)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + Adult.Mortality +
##     polio.vacc + Status + HIV.AIDS + GDP, data = lifedata)
##
## Coefficients:
##      (Intercept)         Schooling   Adult.Mortality  polio.vaccmedium
##        6.121e+01         1.193e+00        -3.601e-02        -6.163e-01
##   polio.vacchigh  StatusDeveloping    HIV.AIDSmedium      HIV.AIDShigh
##        2.067e+00        -1.388e+00        -2.275e+00        -2.367e+00
##              GDP
##        3.452e-05
```

- The best model using Forward AIC is Y = β_0 + $\beta_1 X_1$ + $\beta_2 X_2$ + $\beta_4 X_4$ + $\beta_7 X_7$ + $\beta_8 X_8$ + $\beta_9 X_9$ + $\beta_{10} X_{10}$

Backward AIC

```
step(life.lm, direction='backward', trace=FALSE)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     GDP + polio.vacc + HIV.AIDS + Status, data = lifedata)
##
## Coefficients:
##      (Intercept)   Adult.Mortality         Schooling               GDP
##        6.121e+01        -3.601e-02         1.193e+00         3.452e-05
## polio.vaccmedium    polio.vacchigh    HIV.AIDSmedium      HIV.AIDShigh
##       -6.163e-01         2.067e+00        -2.275e+00        -2.367e+00
## StatusDeveloping
##       -1.388e+00
```

- The best model using Backward AIC is Y = β_0 + $\beta_1 X_1$ + $\beta_2 X_2$ + $\beta_4 X_4$ + $\beta_7 X_7$ + $\beta_8 X_8$ + $\beta_9 X_9$ + $\beta_{10} X_{10}$

Stepwise AIC

```
step(life.lm, direction='both', trace=FALSE)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     GDP + polio.vacc + HIV.AIDS + Status, data = lifedata)
##
## Coefficients:
##      (Intercept)   Adult.Mortality         Schooling               GDP
##        6.121e+01        -3.601e-02         1.193e+00         3.452e-05
## polio.vaccmedium    polio.vacchigh    HIV.AIDSmedium      HIV.AIDShigh
##       -6.163e-01         2.067e+00        -2.275e+00        -2.367e+00
## StatusDeveloping
##       -1.388e+00
```

- The best model using Stepwise AIC is Y = β_0 + $\beta_1 X_1$ + $\beta_2 X_2$ + $\beta_4 X_4$ + $\beta_7 X_7$ + $\beta_8 X_8$ + $\beta_9 X_9$ + $\beta_{10} X_{10}$

All three methods produced the same model but it seems overfit. This process will be repeated using stepwise elimination with BIC, which has a strong penalty associated with addin too many terms.

Forward BIC

```
mod.null <- lm(Life.expectancy ~ 1, lifedata) # null model with no predictors
step(mod.null, scope=formula(life.lm), direction='forward', trace=FALSE, k=log(nrow(lifedata)))
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Schooling + Adult.Mortality +
##     polio.vacc, data = lifedata)
##
## Coefficients:
##      (Intercept)         Schooling   Adult.Mortality  polio.vaccmedium
##         58.54392           1.34239          -0.03991          -0.31961
##   polio.vacchigh
##          2.33516
```

- The best model using Forward BIC is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7 + \beta_8 X_8$

Backward BIC

```
step(life.lm, direction='backward', trace=FALSE, k=log(nrow(lifedata)))
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     polio.vacc, data = lifedata)
##
## Coefficients:
##      (Intercept)   Adult.Mortality         Schooling  polio.vaccmedium
##         58.54392          -0.03991           1.34239          -0.31961
##   polio.vacchigh
##          2.33516
```

- The best model using Backward BIC is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7 + \beta_8 X_8$

Stepwise BIC

```
step(life.lm, direction='both', trace=FALSE, k=log(nrow(lifedata)))
```

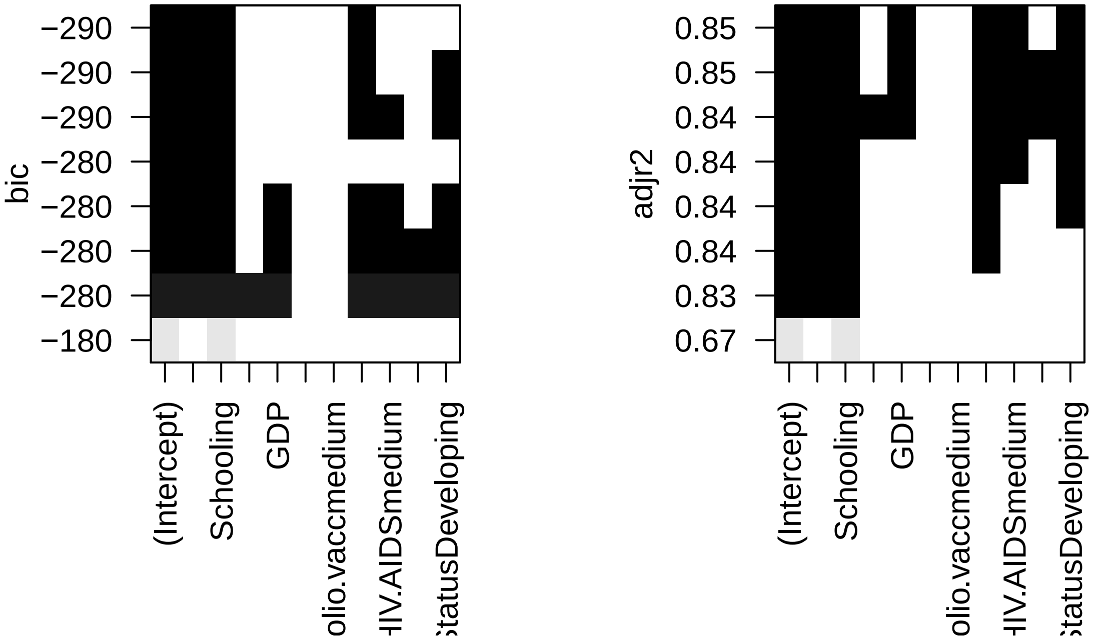
```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     polio.vacc, data = lifedata)
##
## Coefficients:
##      (Intercept)   Adult.Mortality         Schooling  polio.vaccmedium
##         58.54392          -0.03991           1.34239          -0.31961
##   polio.vacchigh
##          2.33516
```

- The best model using Stepwise BIC is Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7 + \beta_8 X_8$

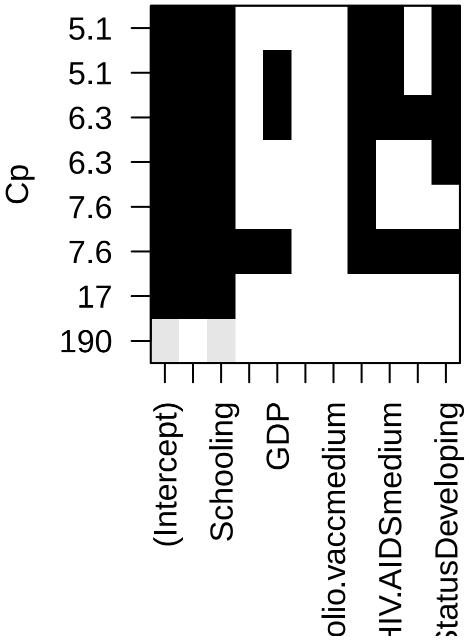
Again, all three methods produced the same model but with fewer predictors/variables using the BIC criteria.

Best subsets model selection

```
par(mfrow=c(1,2))
plot(life.rs)
plot(life.rs, scale='adjr2')
```

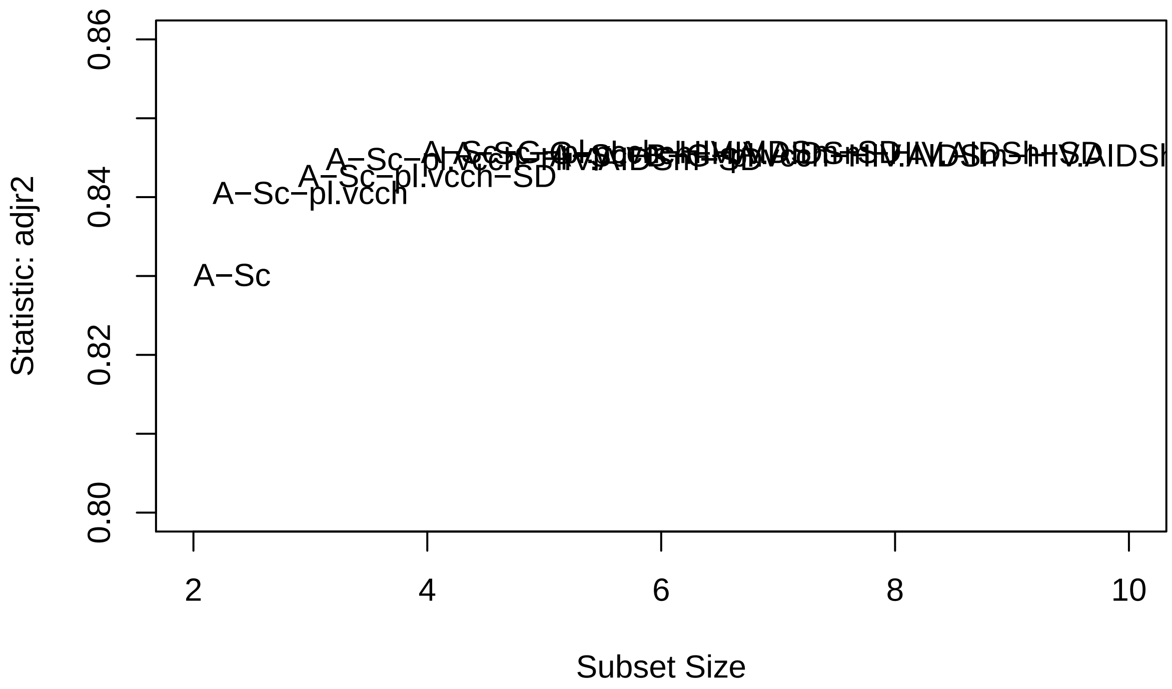


```
plot(life.rs, scale='Cp')
```



- The model with lowest BIC: Y = β_0 + $\beta_1 X_1$ + $\beta_2 X_2$ + $\beta_7 X_7$
- The model with highest R$^2 adj$: Y = β_0 + $\beta_1 X_1$ + $\beta_2 X_2$ + $\beta_4 X_4$ + $\beta_7 X_7$ + $\beta_8 X_8$ + $\beta_{10} X_{10}$
- The model with lowest C{p}: Y = β_0 + $\beta_1 X_1$ + $\beta_2 X_2$ + $\beta_7 X_7$ + $\beta_8 X_8$ + $\beta_{10} X_{10}$

```
subsets(life.rs, statistic='adjr2',  ylim=c(.8,.86), min.size=2, legend=FALSE)
```

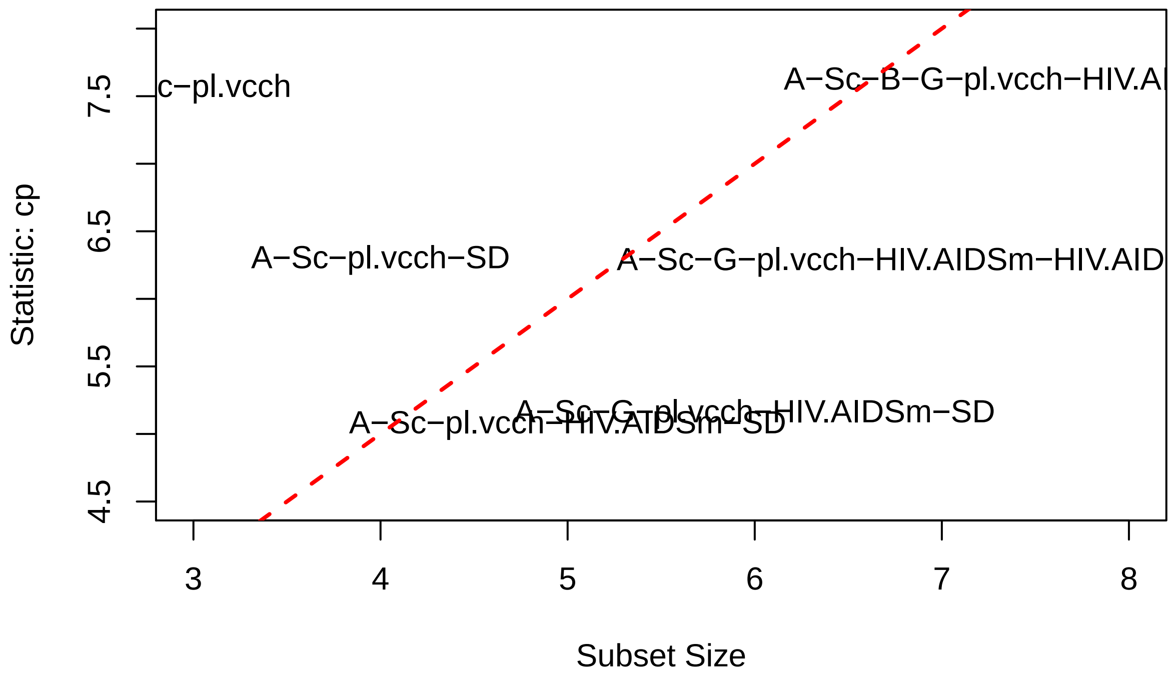


```
##                      Abbreviation
## Adult.Mortality                 A
## Schooling                      Sc
## BMI                             B
## GDP                             G
## infant.deaths                   i
## polio.vaccmedium          pl.vccm
## polio.vacchigh            pl.vcch
## HIV.AIDSmedium          HIV.AIDSm
## HIV.AIDShigh            HIV.AIDSh
## StatusDeveloping               SD
```

```
subsets(life.rs, statistic='cp', xlim=c(3,8), ylim=c(4.5,8), min.size=2, legend=FALSE)
```

```
##                      Abbreviation
## Adult.Mortality                 A
## Schooling                      Sc
## BMI                             B
## GDP                             G
## infant.deaths                   i
## polio.vaccmedium          pl.vccm
## polio.vacchigh            pl.vcch
## HIV.AIDSmedium          HIV.AIDSm
## HIV.AIDShigh            HIV.AIDSh
## StatusDeveloping               SD
```

```
abline(a=1, b=1, col='red', lty='dashed', lwd=2)
```



## Summary: Subsets

The $\mathrm{R}^2 adj$ plot is difficult to interpret, since many of the models have $\mathrm{R}^2 adj$ from 0.83-0.85, as seen in the *plot* for $\mathrm{R}^2 adj$.

From the cp plot, the first model which crosses below the reference line is also the model with the lowest C{p}: Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7 + \beta_8 X_8 + \beta_{10} X_{10}$

Variable selection procedures

Lowest Cp Model

Using the model with the lowest C{p}:

```
life_cp <- lm(Life.expectancy ~
            Adult.Mortality
              + Schooling
              + polio.vacc
              + HIV.AIDS
              + Status, lifedata)

summary(life_cp)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling +
##     polio.vacc + HIV.AIDS + Status, data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7403 -1.7833  0.1144  1.7692  7.1251
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61.138088   2.509620  24.361   <2e-16 ***
## Adult.Mortality  -0.036629   0.003573 -10.252   <2e-16 ***
## Schooling         1.230118   0.130963   9.393   <2e-16 ***
## polio.vaccmedium -0.552757   1.265909  -0.437   0.6630
## polio.vacchigh    2.139281   0.987223   2.167   0.0317 *
## HIV.AIDSmedium   -2.208781   1.153961  -1.914   0.0574 .
## HIV.AIDShigh     -2.248463   2.420154  -0.929   0.3543
## StatusDeveloping -1.556737   0.781547  -1.992   0.0481 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.119 on 160 degrees of freedom
## Multiple R-squared:  0.8501, Adjusted R-squared:  0.8435
## F-statistic: 129.6 on 7 and 160 DF,  p-value: < 2.2e-16
```

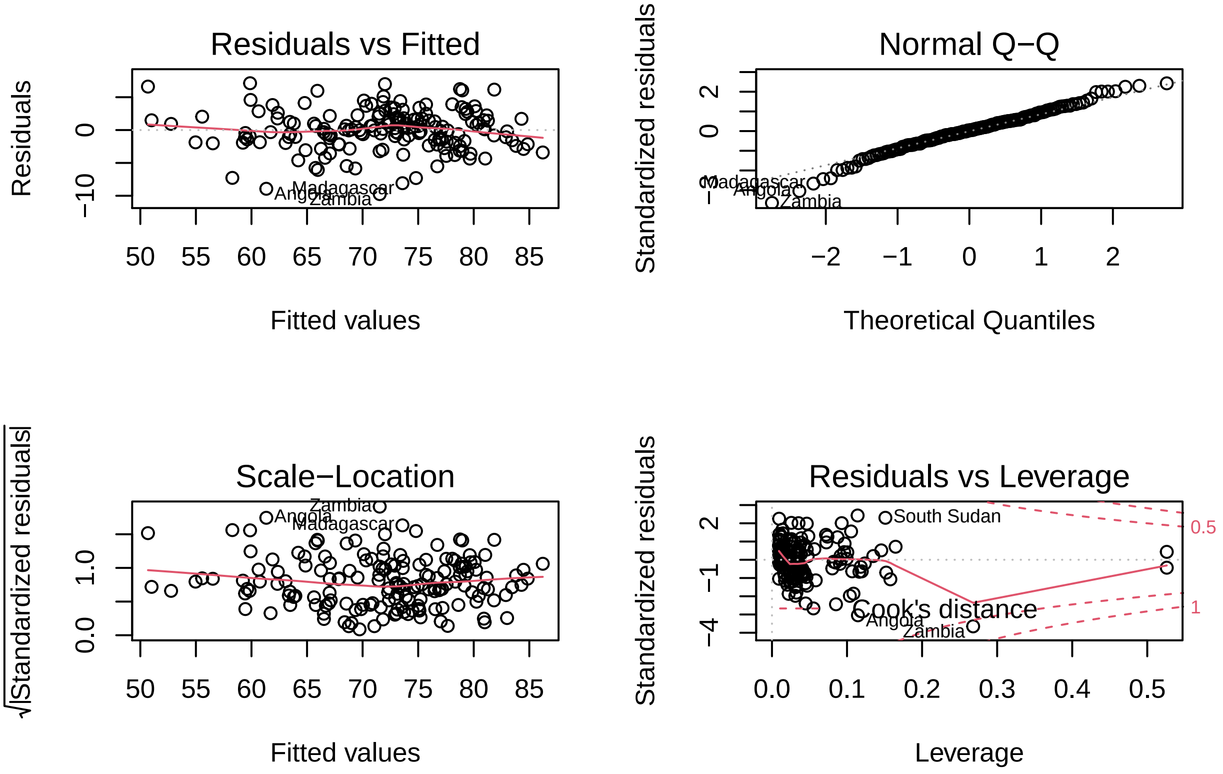
```
summary(life_cp)$sigma
```

```
## [1] 3.119192
```

```
vif(life_cp)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## Adult.Mortality 2.014561  1        1.419352
## Schooling       2.380462  1        1.542874
## polio.vacc      1.409288  2        1.089557
## HIV.AIDS        1.478605  2        1.102714
## Status          1.506363  1        1.227340
```

```
par(mfrow=c(2,2))
plot(life_cp)
```



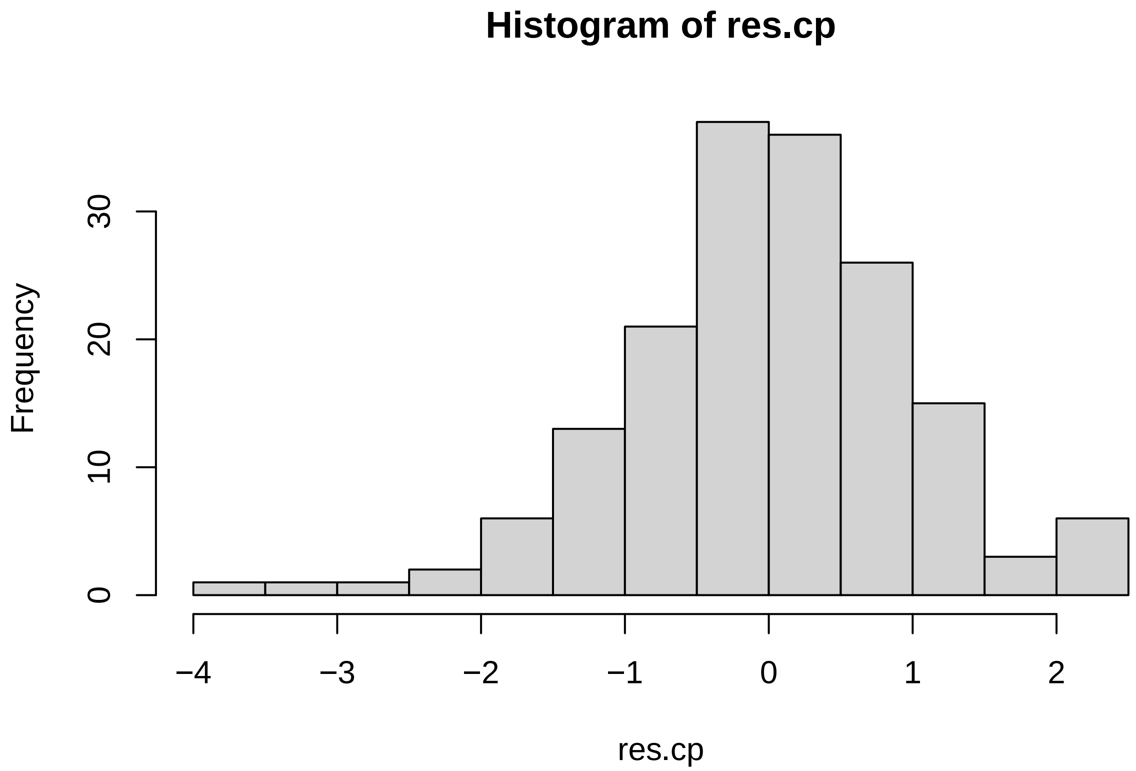
```
res.cp <- rstandard(life_cp)
shapiro.test(res.cp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.cp
## W = 0.98279, p-value = 0.03551
```

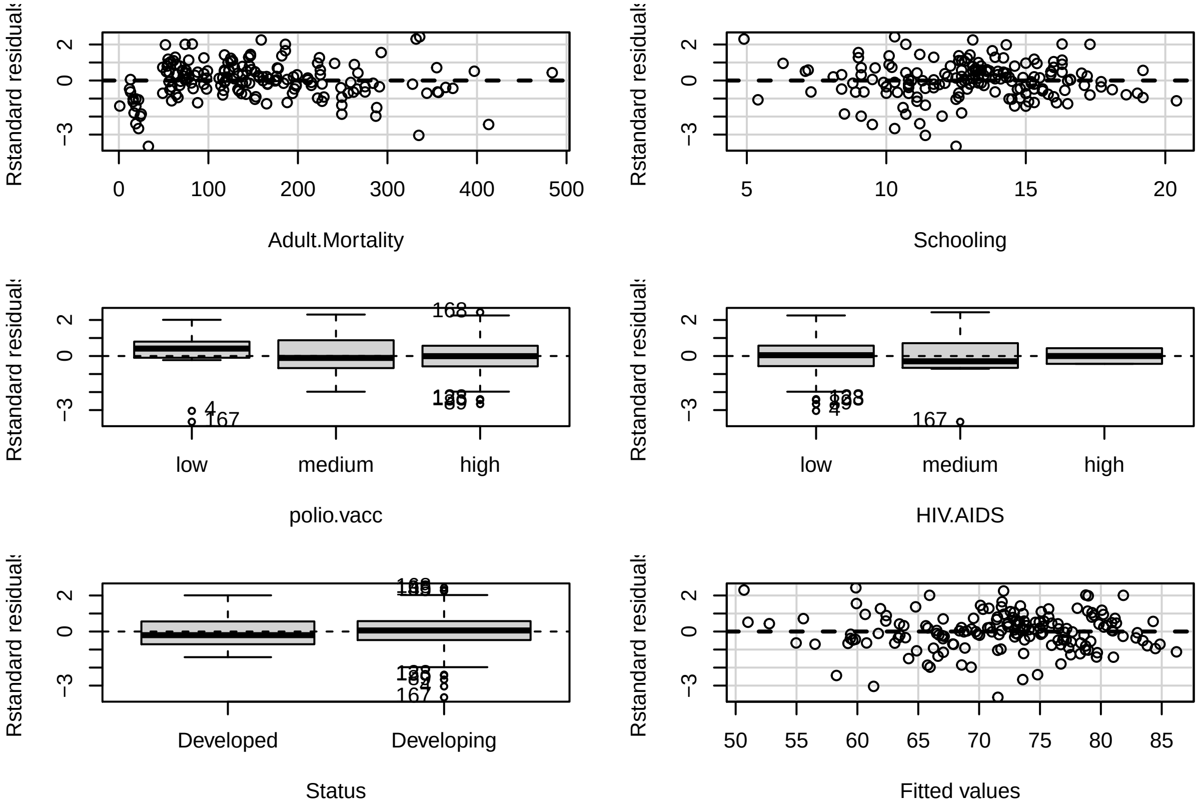
```
car::ncvTest(life_cp)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.707706, Df = 1, p = 0.099865
```

```
par(mfrow=c(1,1))
hist(res.cp)
```



```
residualPlots(life_cp, type='rstandard', quadratic=FALSE)
```



```
##                  Test stat Pr(>|Test stat|)
## Adult.Mortality   -4.5639        9.993e-06 ***
## Schooling         -0.3714           0.7108
## polio.vacc
## HIV.AIDS
## Status
## Tukey test         0.4298           0.6674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary: Lowest Cp Model

The summary statistics look good, with most predictors being influential and a R$^2 adj$ of 0.8435 and sigma of 3.27.There are no issues with collinearity as all VIFs are less than 10.

The plots do not indiate any issues with normality, lineairtiy or homoscedasticity. There is one influential point (Zambia).

The Breusch-Pagan test indicates there is an issue with heteroscadisity that was not seen in the plots. The histogram of residuals is unfortunately left-skewed, with some high outliers.

Lowest BIC Model

Next we will try using the model (tied) with lowest BIC: Y = $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_7 X_7 + \beta_{10} X_{10}$

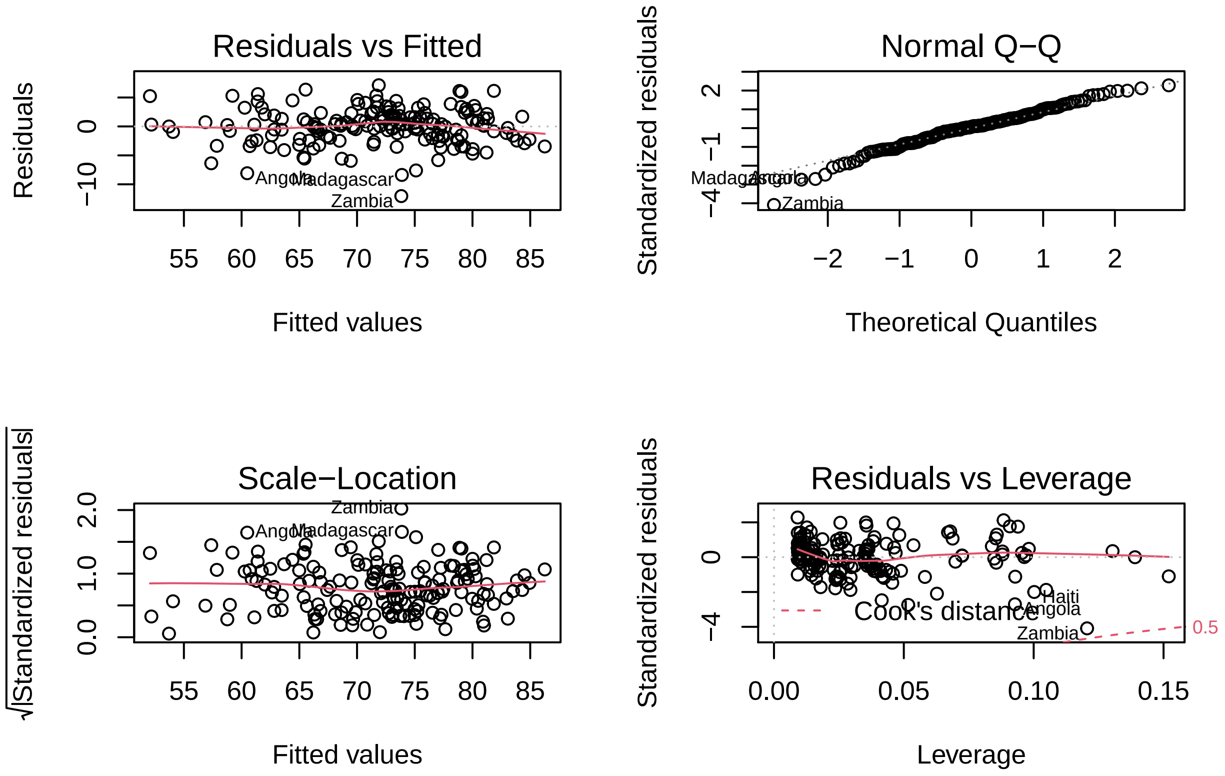
```
life_BIC <- lm(Life.expectancy ~
               Adult.Mortality
              + Schooling
              + polio.vacc
              + Status, lifedata)

summary(life_BIC)
```

```
## 
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Schooling + 
##     polio.vacc + Status, data = lifedata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.0334  -1.8226   0.2118   1.7904   7.1177 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      61.219316   2.518172  24.311   <2e-16 ***
## Adult.Mortality  -0.039717   0.003185 -12.471   <2e-16 ***
## Schooling         1.228873   0.131078   9.375   <2e-16 ***
## polio.vaccmedium -0.557400   1.266114  -0.440   0.6603    
## polio.vacchigh    2.315900   0.985490   2.350   0.0200 *  
## StatusDeveloping -1.436163   0.784017  -1.832   0.0688 .  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 3.138 on 162 degrees of freedom
## Multiple R-squared:  0.8463, Adjusted R-squared:  0.8416 
## F-statistic: 178.4 on 5 and 162 DF,  p-value: < 2.2e-16
```

```
summary(life_BIC)$sigma
```

```
## [1] 3.138474
```

```
par(mfrow=c(2,2))
plot(life_BIC)
```



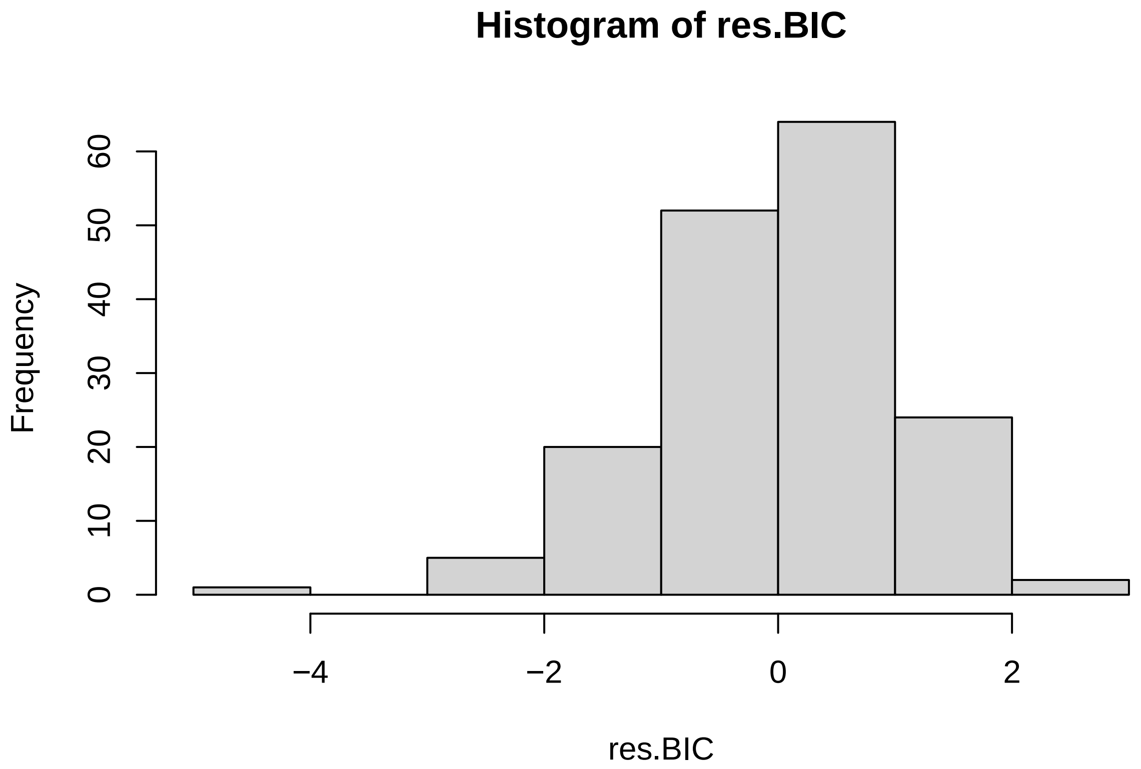
```r
res.BIC <- rstandard(life_BIC)
shapiro.test(res.BIC)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.BIC
## W = 0.97848, p-value = 0.0103
```

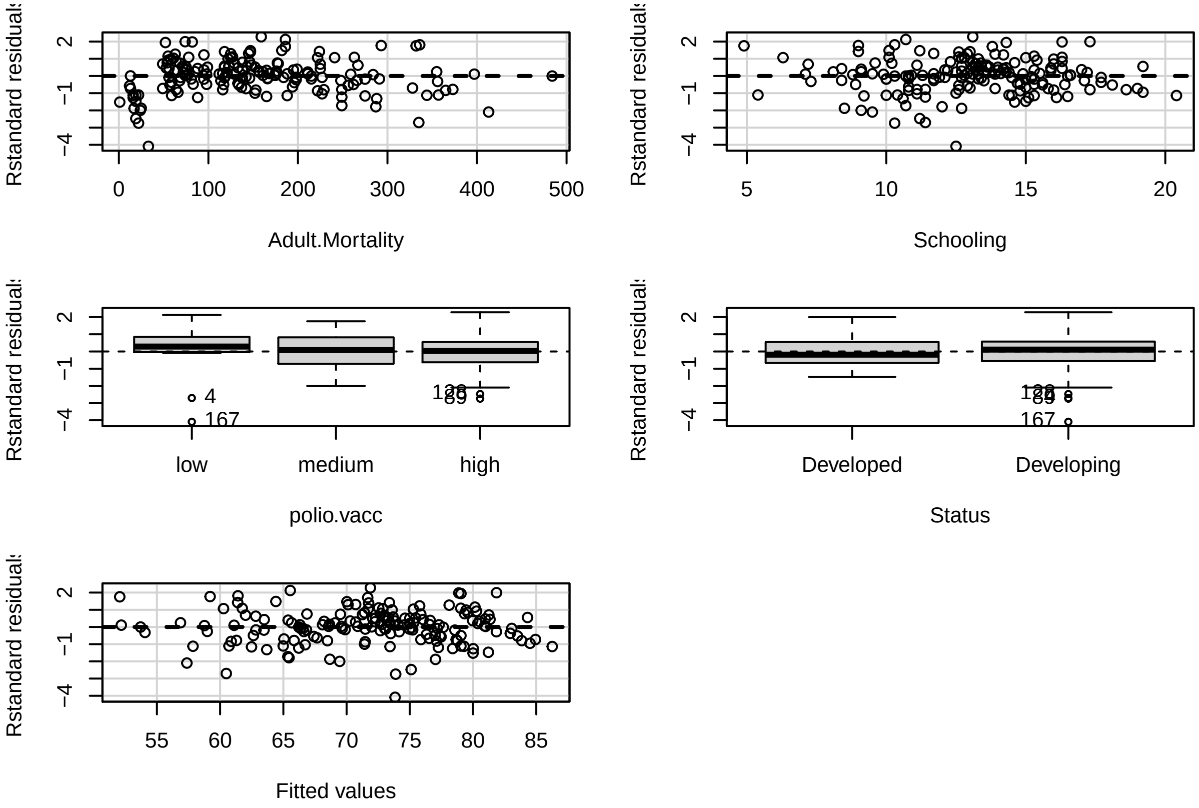
```r
car::ncvTest(life_BIC)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8300842, Df = 1, p = 0.36225
```

```r
par(mfrow=c(1,1))
hist(res.BIC)
```



```r
residualPlots(life_BIC, type='rstandard', quadratic=FALSE)
```



```
##                   Test stat Pr(>|Test stat|)
## Adult.Mortality     -4.5897        8.89e-06 ***
## Schooling           -0.5319          0.5955
## polio.vacc
## Status
## Tukey test          -0.7796          0.4356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary: Lowest BIC Model

The summary statistics look good, with most predictors being influential and a R$^2 adj$ of 0.8416 and sigma of 3.99.

The Breusch-Pagan test improved significantly, as did the influntial point (Zambia), which has moved back under a Cook's distance of 0.5.

The histogram of residuals has improved with fewer outliers, but still appears slightly left-skewed.

Lowest BIC Model with Quadratic Transformation

Next we will try using the same model with the lowest BIC, but with a quadratic predictor: Y = β_0 + $\beta_1 X_1^2$ + $\beta_2 X_2$ + $\beta_7 X_7$ + $\beta_{10} X_{10}$

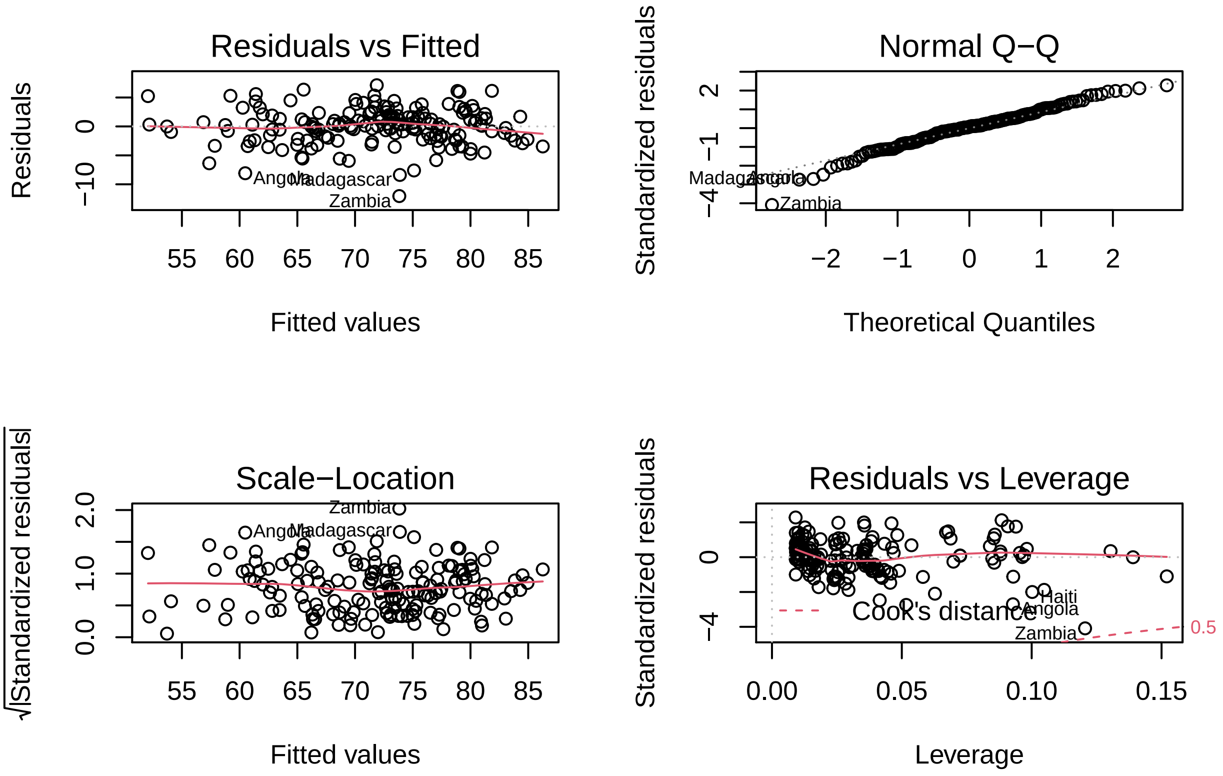
```
life_BIC2 <- lm(Life.expectancy ~
               Adult.Mortality^2
              + Schooling
              + polio.vacc
              + Status, lifedata)

summary(life_BIC2)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality^2 + Schooling +
##     polio.vacc + Status, data = lifedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0334  -1.8226   0.2118   1.7904   7.1177
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61.219316   2.518172  24.311   <2e-16 ***
## Adult.Mortality  -0.039717   0.003185 -12.471   <2e-16 ***
## Schooling         1.228873   0.131078   9.375   <2e-16 ***
## polio.vaccmedium -0.557400   1.266114  -0.440   0.6603
## polio.vacchigh    2.315900   0.985490   2.350   0.0200 *
## StatusDeveloping -1.436163   0.784017  -1.832   0.0688 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.138 on 162 degrees of freedom
## Multiple R-squared:  0.8463, Adjusted R-squared:  0.8416
## F-statistic: 178.4 on 5 and 162 DF,  p-value: < 2.2e-16
```

```
summary(life_BIC2)$sigma
```

```
## [1] 3.138474
```

```
par(mfrow=c(2,2))
plot(life_BIC2)
```



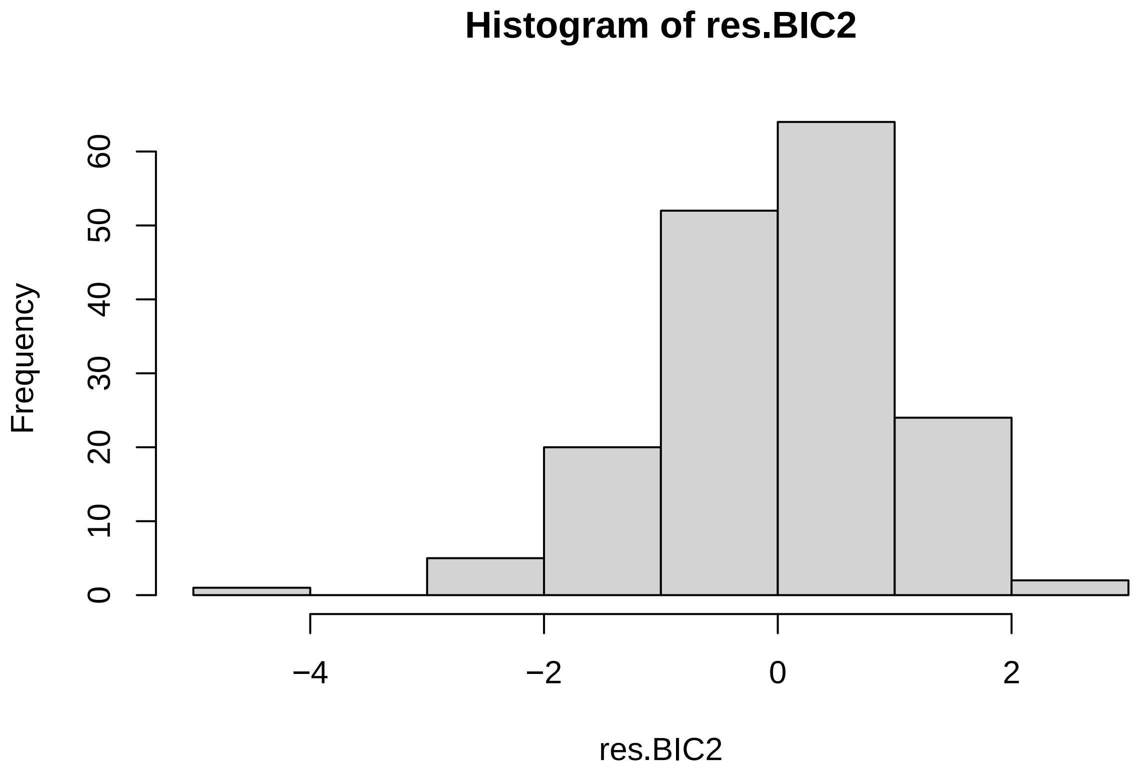
```
res.BIC2 <- rstandard(life_BIC2)
shapiro.test(res.BIC2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.BIC2
## W = 0.97848, p-value = 0.0103
```

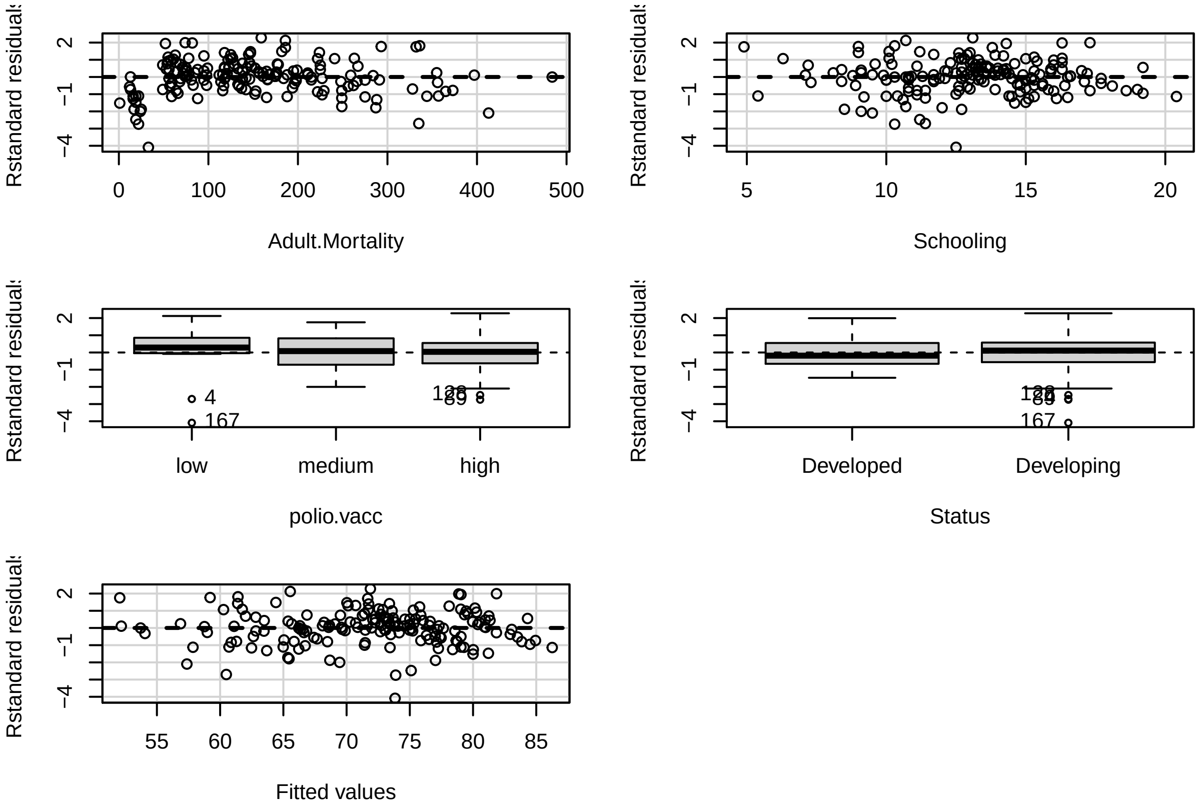
```
car::ncvTest(life_BIC2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8300842, Df = 1, p = 0.36225
```

```
par(mfrow=c(1,1))
hist(res.BIC2)
```



```
residualPlots(life_BIC2, type='rstandard', quadratic=FALSE)
```



```
##                  Test stat Pr(>|Test stat|)
## Adult.Mortality    -4.5897        8.89e-06 ***
## Schooling          -0.5319          0.5955
## polio.vacc
## Status
## Tukey test         -0.7796          0.4356
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary: Lowest BIC Model with Quadratic Transformation

The residuals have improved a bit but there are still some significant outliers in the histogram of residuals (>|3|). Sigma has improved a bit to 3.35. Adding in some interaction terms like in Week 4 could help improve the model even further.

Final model

Using the lowest BIC model coefficients, along with transformations and interactions found previously to further explore the "best" model for this data.

```
life.final <- lm(Life.expectancy ~
I(Adult.Mortality^2)
+ Schooling
```

```r
+ I(Adult.Mortality^2):Status
+ I(Adult.Mortality^2):HIV.AIDS
+ Schooling:polio.vacc
+ Schooling:Status
+ Schooling:HIV.AIDS,
data = lifedata)

summary(life.final)
```

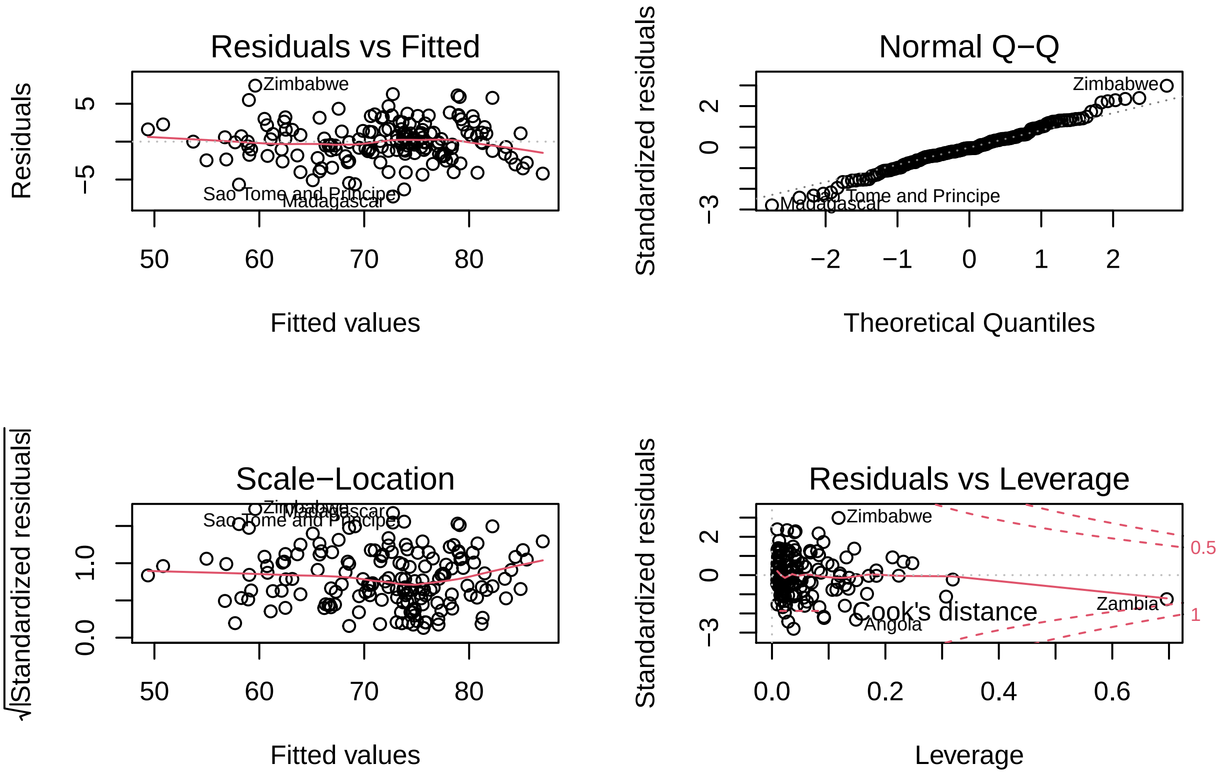
```
##
## Call:
## lm(formula = Life.expectancy ~ I(Adult.Mortality^2) + Schooling +
##     I(Adult.Mortality^2):Status + I(Adult.Mortality^2):HIV.AIDS +
##     Schooling:polio.vacc + Schooling:Status + Schooling:HIV.AIDS,
##     data = lifedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2427 -1.4298 -0.0901  1.3582  7.3874
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            6.083e+01  1.560e+00  38.997  < 2e-16
## I(Adult.Mortality^2)                  -3.240e-04  6.935e-05  -4.672 6.37e-06
## Schooling                              1.232e+00  1.287e-01   9.572  < 2e-16
## I(Adult.Mortality^2):StatusDeveloping  1.921e-04  6.950e-05   2.765  0.00638
## I(Adult.Mortality^2):HIV.AIDSmedium    9.055e-05  1.769e-05   5.119 8.87e-07
## I(Adult.Mortality^2):HIV.AIDShigh      9.165e-05  3.653e-05   2.509  0.01314
## Schooling:polio.vaccmedium            -1.360e-01  1.050e-01  -1.295  0.19713
## Schooling:polio.vacchigh               1.075e-01  7.637e-02   1.407  0.16131
## Schooling:StatusDeveloping            -1.762e-01  5.070e-02  -3.475  0.00066
## Schooling:HIV.AIDSmedium              -8.282e-01  1.824e-01  -4.541 1.11e-05
## Schooling:HIV.AIDShigh                -8.414e-01  6.165e-01  -1.365  0.17428
##
## (Intercept)                           ***
## I(Adult.Mortality^2)                  ***
## Schooling                             ***
## I(Adult.Mortality^2):StatusDeveloping **
## I(Adult.Mortality^2):HIV.AIDSmedium   ***
## I(Adult.Mortality^2):HIV.AIDShigh     *
## Schooling:polio.vaccmedium
## Schooling:polio.vacchigh
## Schooling:StatusDeveloping            ***
## Schooling:HIV.AIDSmedium              ***
## Schooling:HIV.AIDShigh
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.637 on 157 degrees of freedom
## Multiple R-squared:  0.8949, Adjusted R-squared:  0.8882
## F-statistic: 133.6 on 10 and 157 DF,  p-value: < 2.2e-16
```

```r
summary(life.final)$sigma
```

```
## [1] 2.636859
```

```r
par(mfrow=c(2,2))
plot(life.final)
```

```
## Warning: not plotting observations with leverage one:
##   84, 144
```



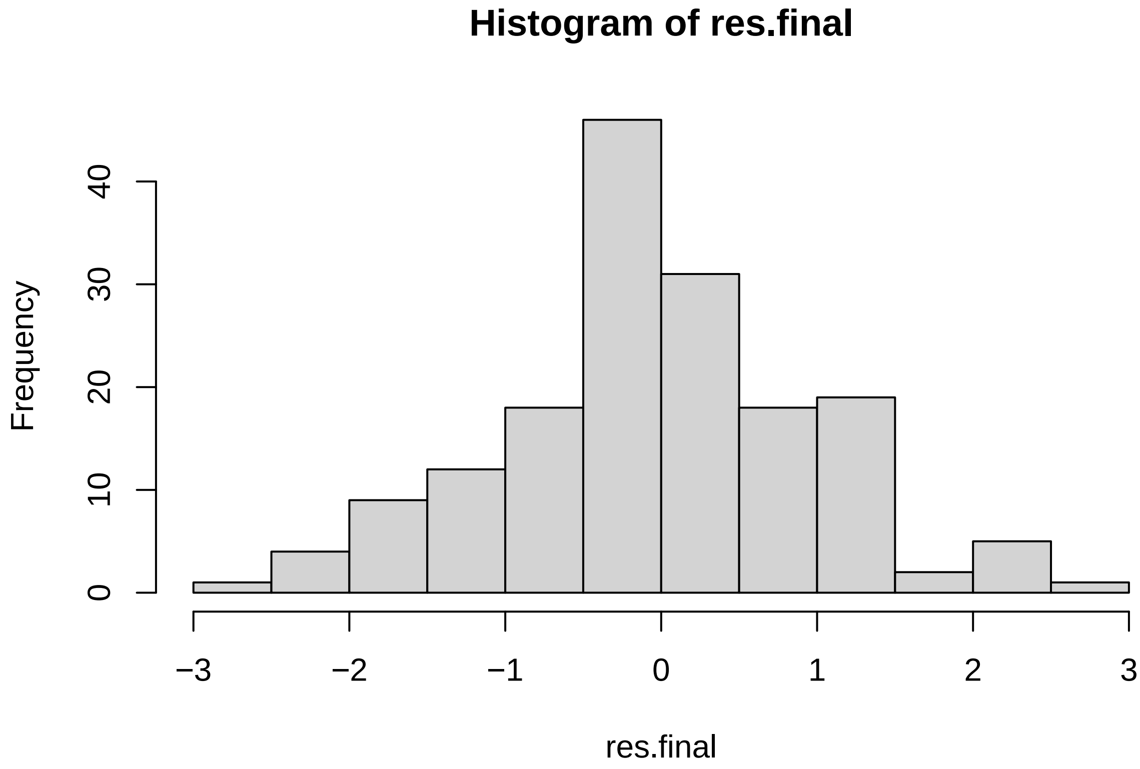
```r
res.final <- rstandard(life.final)
shapiro.test(res.final)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.final
## W = 0.99195, p-value = 0.4801
```

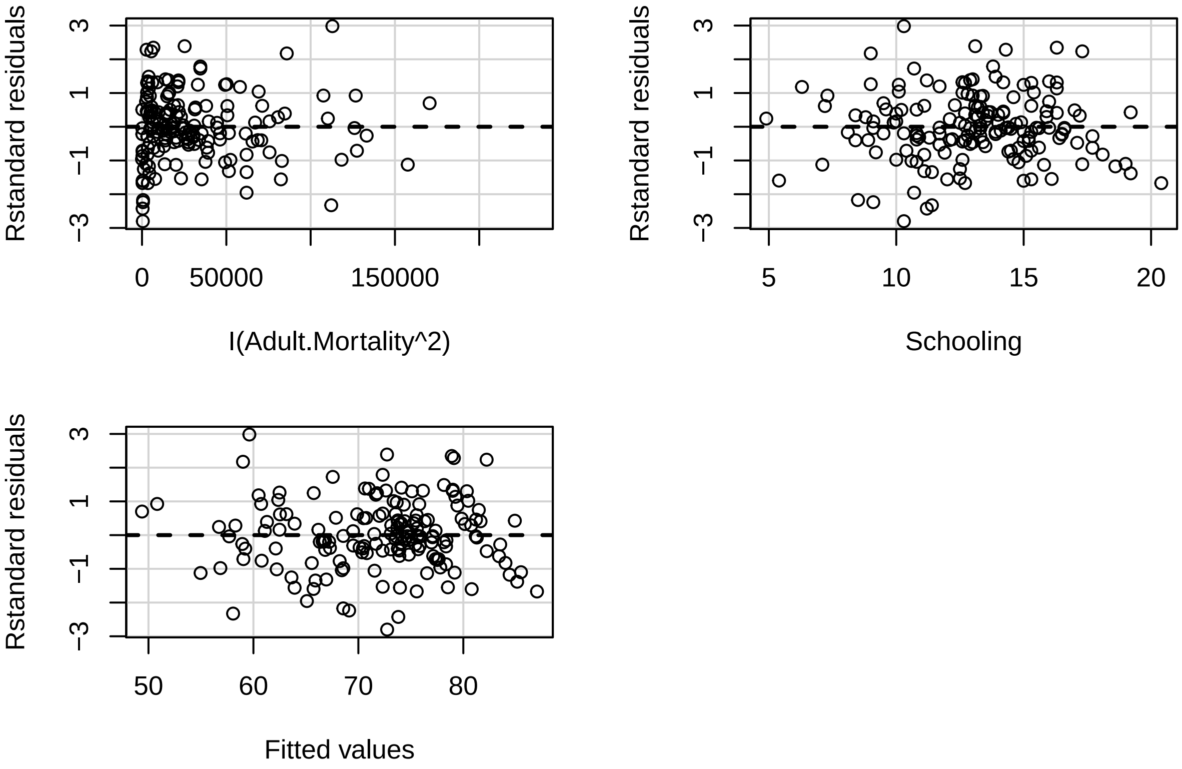
```r
car::ncvTest(life.final)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.04413981, Df = 1, p = 0.83359
```

```r
par(mfrow=c(1,1))
hist(res.final)
```

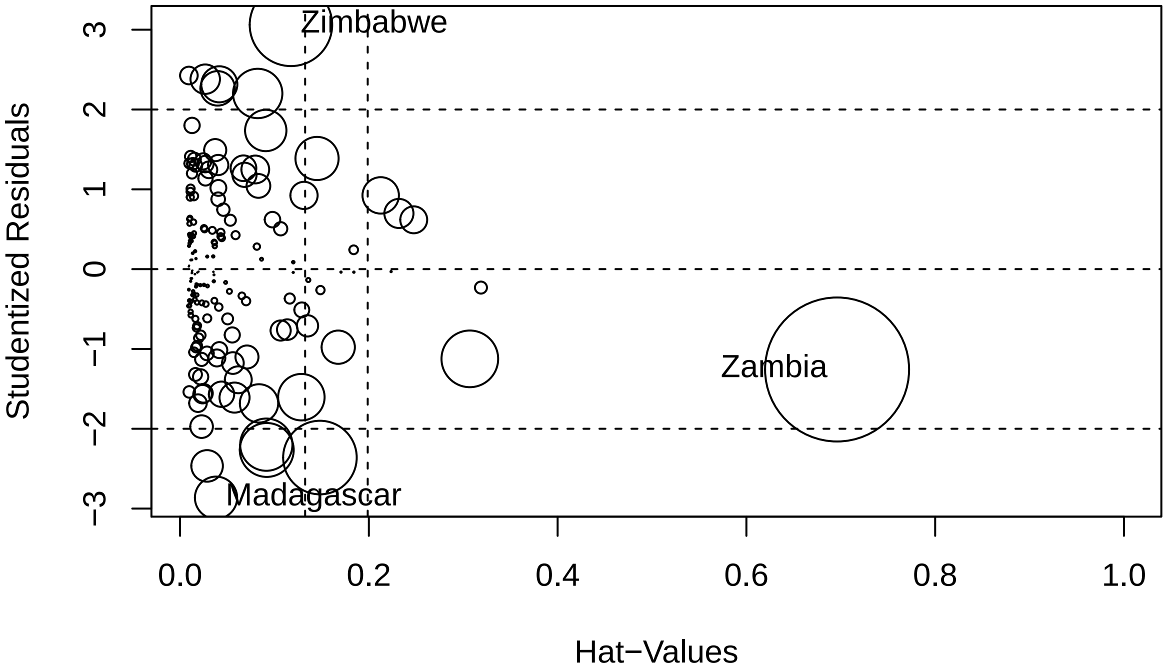


```r
residualPlots(life.final, type='rstandard', quadratic=FALSE)
```



```
##                       Test stat Pr(>|Test stat|)
## I(Adult.Mortality^2)    -0.4938          0.62214
## Schooling               -1.9593          0.05186 .
## Tukey test              -0.0192          0.98470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
influencePlot(life.final)
```



```
##                StudRes        Hat      CookD
## Lesotho            NaN 1.00000000        NaN
## Madagascar -2.864116 0.03810002 0.02824242
## Swaziland          NaN 1.00000000        NaN
## Zambia     -1.257220 0.69608153 0.32789146
## Zimbabwe    3.060659 0.11747728 0.10762527
```

Summary: Final Model

Y = β_0 + $\beta_1 X_1^2$ + $\beta_2 X_2$ + $\beta_1 X_1^2{:}\beta_{10} X_{10}$ + $\beta_1 X_1^2{:}\beta_8 X_8$ + $\beta_1 X_1^2{:}\beta_9 X_9$ + $\beta_2 X_2 \beta_6 X_7$ + $\beta_2 X_2 \beta_7 X_7$ + $\beta_2 X_2 \beta_{10} X_{10}$ + $\beta_2 X_2 \beta_8 X_8$ + $\beta_2 X_2 {*} \beta_9 X_9$

The final model using the lowest BIC as well as the transformations and interactions from Week 5 produced the best results yet. It produced the highest R^2*adj* value yet: 0.8882, and lowest sigma yet: 2.78.

Most interestingly is that the model includes the variables most commonly found in the best p-variable models observed during the variable selection procedures. The variables that occurred most often in the models (X_2 *Schooling*, X_1 *Adult Mortality*, X_7 *Polio Vaccine - High*, and X_{10} *Status - Developing*) are all predictors used in the final model.

The assumptions for the residuals in the final model are also the best found in the 7 weeks of the course. The Breusch-Pagan test indicates homoscedasticity is met, and the normality assumption is met using the Shapiro-Wilk test. The linearity assumption is met; the only influential point is “Zambia” but it does not appear to be an outlier