Term Project: STA 108

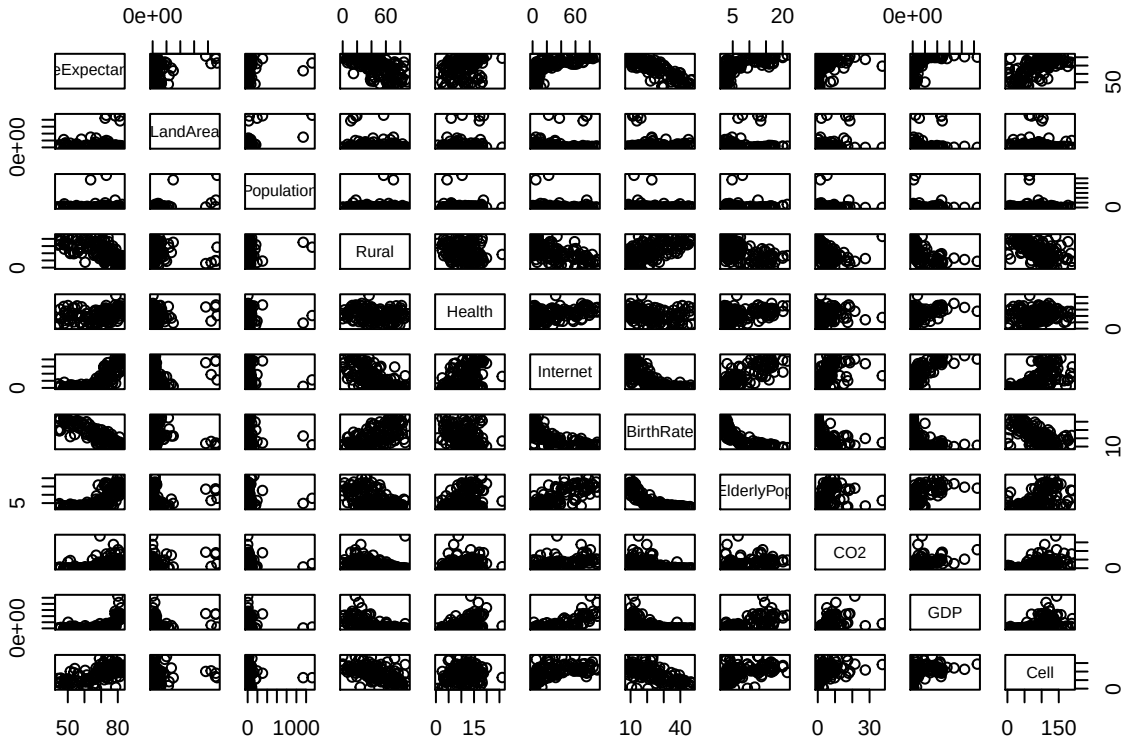
Riley Adams

12/13/2020

```
## [1] 186  13
```

```
## [1] 149  13
```

```
#scatterplots (initial)
pairs(cbind(LifeExpectancy, LandArea, Population, Rural, Health, Internet, BirthRate, ElderlyPop, CO2, G
```



Life Expectancy appears to have no correlation with LandArea, Population.

There appears to be a negative linear trend with Rural, BirthRate.

There appears to be positive linear trend with Health, Cell.

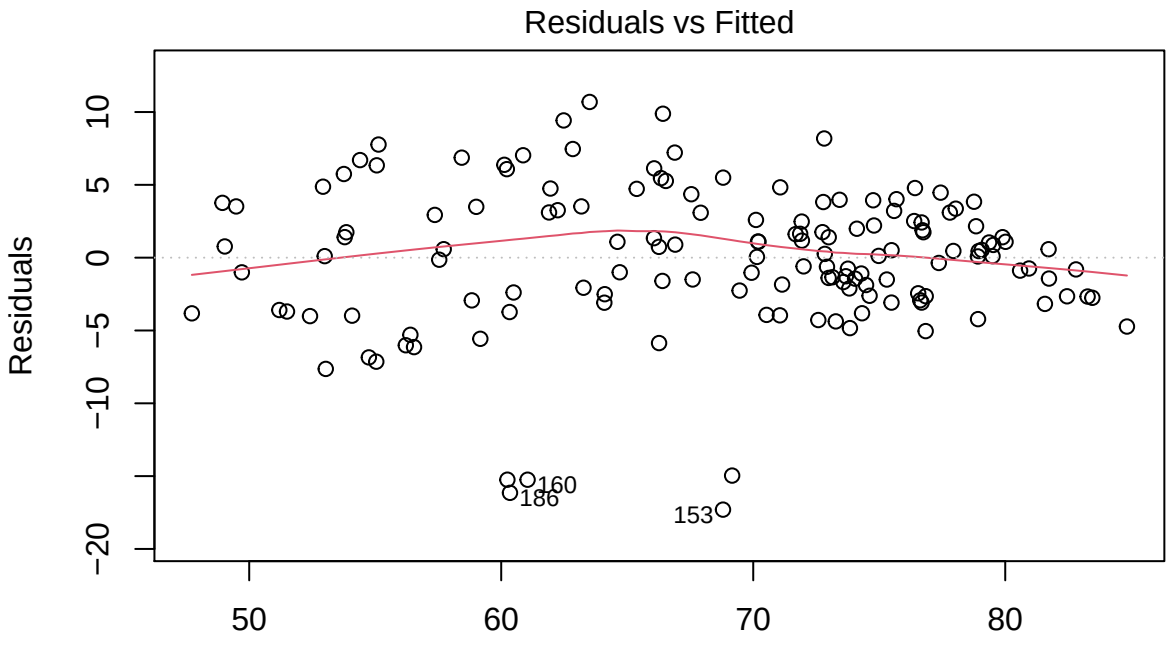
With internet, elderlypopulation, CO2 and GDP there is som kind of positive relationship that is not linear.

```r
## fit model and summary output [1] -----
fit1 <- lm(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet + BirthRate + ElderlyPop +
summary(fit1)
```

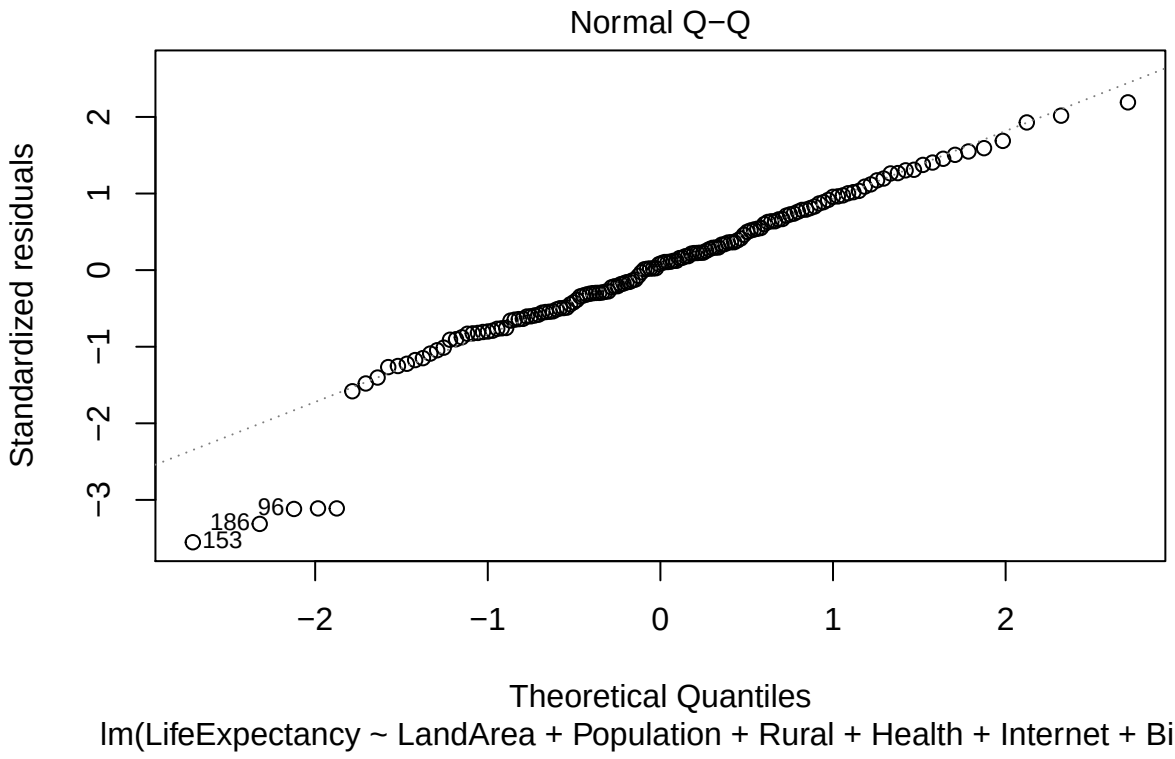
```
##
## Call:
## lm(formula = LifeExpectancy ~ LandArea + Population + Rural +
##     Health + Internet + BirthRate + ElderlyPop + CO2 + GDP +
##     Cell, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3006  -2.6467   0.3558   3.1298  10.6910
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.464e+01  3.643e+00  23.236  < 2e-16 ***
## LandArea    -1.655e-07  3.308e-07  -0.500   0.6176
## Population   4.588e-04  3.519e-03   0.130   0.8965
## Rural       -5.196e-02  2.670e-02  -1.947   0.0536 .
## Health       1.942e-01  1.046e-01   1.857   0.0655 .
## Internet     5.611e-02  3.525e-02   1.592   0.1137
## BirthRate   -7.180e-01  7.847e-02  -9.150 7.07e-16 ***
## ElderlyPop  -4.482e-01  1.716e-01  -2.612   0.0100 *
## CO2         -6.835e-02  1.031e-01  -0.663   0.5086
## GDP          7.031e-05  4.262e-05   1.650   0.1013
## Cell         1.736e-02  1.412e-02   1.230   0.2209
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.959 on 137 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7899, Adjusted R-squared:  0.7745
## F-statistic:  51.5 on 10 and 137 DF,  p-value: < 2.2e-16
```

Initial model above.

```r
# Residual analysis [1] ---

# Residual plots
plot(fit1, which = 1)
```




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




```
# Model Selection ----
library(leaps)
```

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

```
## Forward selection
fit1_forward <- regsubsets(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet + BirthRa
                           data = country80, method = "forward")
cbind(summary(fit1_forward)$which, "adjusted r^2" = summary(fit1_forward)$adjr2)
```

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

```
## 6   0   1    0    0.7767840
## 7   0   1    1    0.7777896
## 8   1   1    1    0.7773005
```

```
## Backward elimination
fit1_backward <- regsubsets(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet + BirthRa
                            data = country80, method = "backward")
cbind(summary(fit1_backward)$which, "adjusted r^2" = summary(fit1_backward)$adjr2)
```

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

Backward Elimination suggests best model is: LifeExpectancy ~ Rural, Health, Internet, BirthRate, ElderlyPop, GDP, Cell

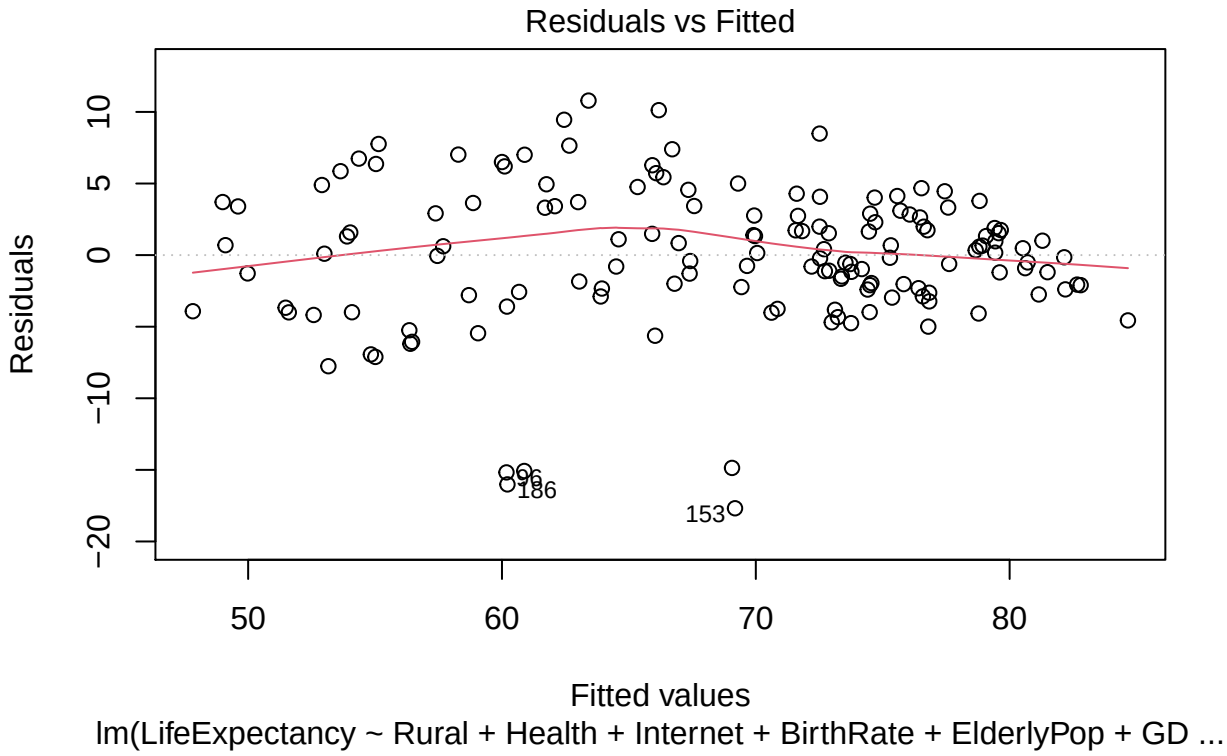
I'll fit that as the new model, fit2.

```
## fit model and summary output [2] -----
fit2 <- lm(LifeExpectancy ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cell, data = cou
summary(fit2)
```

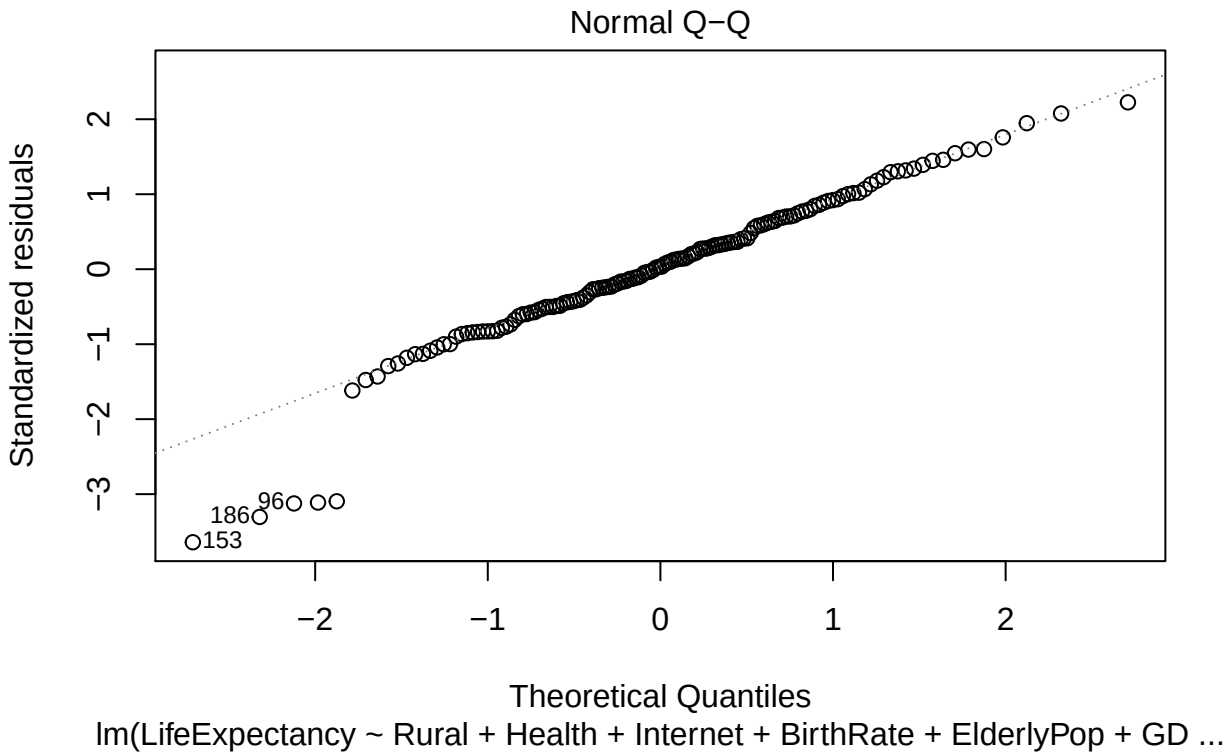
```
##
## Call:
## lm(formula = LifeExpectancy ~ Rural + Health + Internet + BirthRate +
##     ElderlyPop + GDP + Cell, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.6813  -2.4483   0.1573   3.1532  10.7907
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.376e+01  3.469e+00  24.147   <2e-16 ***
## Rural       -5.063e-02  2.560e-02  -1.977   0.0500 *
## Health       1.990e-01  1.017e-01   1.956   0.0524 .
## Internet     5.164e-02  3.450e-02   1.497   0.1367
## BirthRate   -7.031e-01  7.487e-02  -9.391   <2e-16 ***
```

```
## ElderlyPop  -4.078e-01  1.653e-01  -2.467   0.0148 *
## GDP          5.893e-05  4.019e-05   1.466   0.1449
## Cell         1.722e-02  1.346e-02   1.280   0.2027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.923 on 140 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7884, Adjusted R-squared:  0.7778
## F-statistic: 74.51 on 7 and 140 DF,  p-value: < 2.2e-16
```

```r
# Residual analysis [2] ---

# Residual plots
plot(fit2, which = 1)
```




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



Residual Plot:

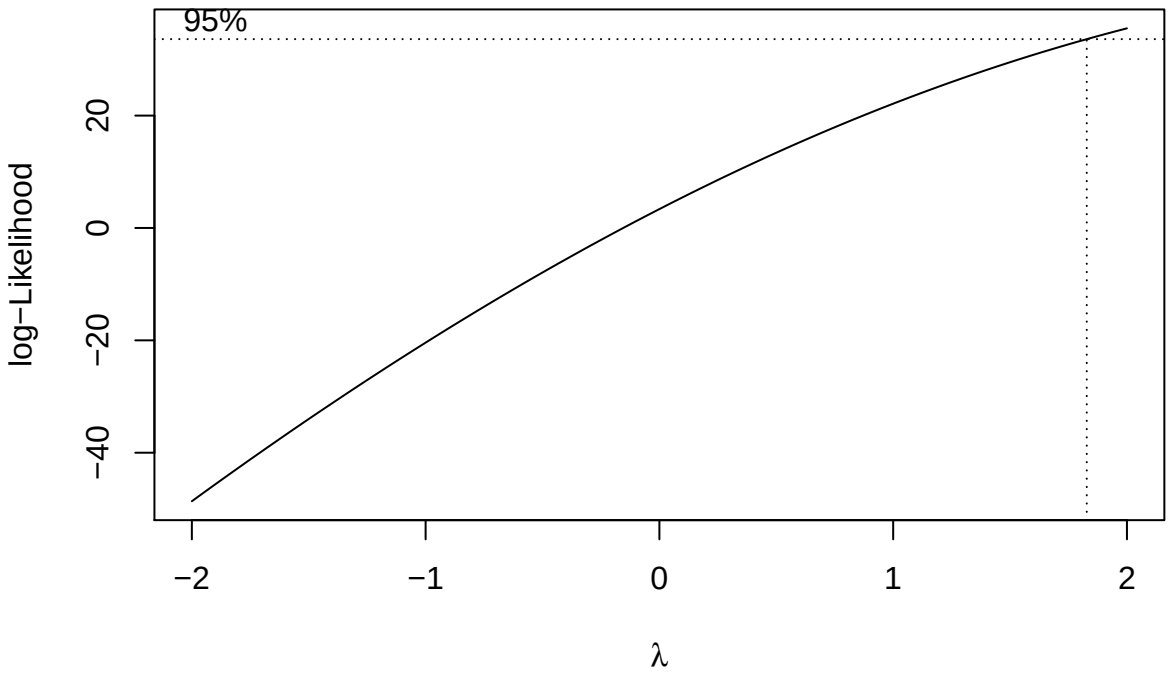
- Linearity: Residuals a little high in the middle. somewhat nonlinear.
- Variance: non-constant. Lower variance at high end.

QQ Plot:

- Fairly good. Values at lowest end are too low.

Good candidate for boxcox.

```
library(MASS)

# Boxcox for fit2
boxcox(fit2)
```



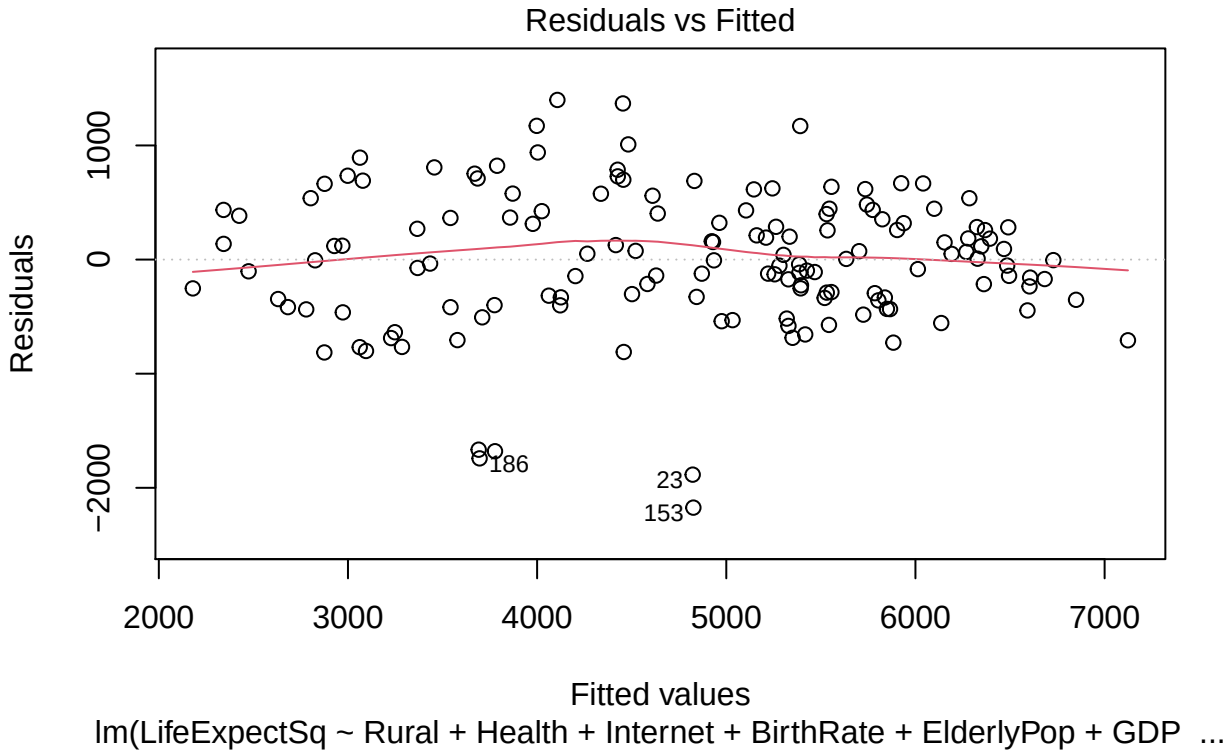
Box Cox suggests we raise Y to approx 1.5 or 2. Closer to 2. So we will try that.

```
## fit model and summary output [3] -----
LifeExpectSq <- (LifeExpectancy)^2
fit3 <- lm(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cell, data = counti
summary(fit3)
```

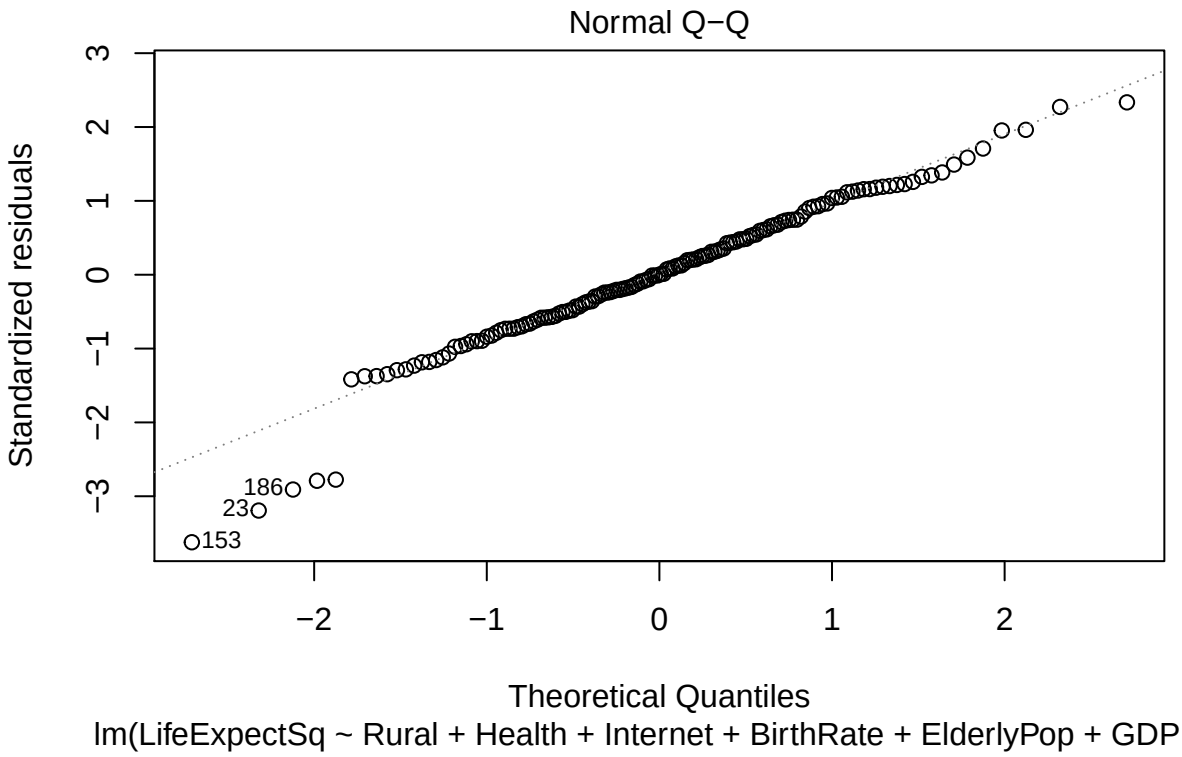
```
##
## Call:
## lm(formula = LifeExpectSq ~ Rural + Health + Internet + BirthRate +
##     ElderlyPop + GDP + Cell, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2173.61  -338.56     0.37   399.77  1398.60
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.520e+03  4.286e+02  15.213  < 2e-16 ***
## Rural       -6.914e+00  3.164e+00  -2.185   0.0305 *
## Health       2.682e+01  1.257e+01   2.134   0.0346 *
## Internet     7.567e+00  4.262e+00   1.775   0.0780 .
## BirthRate   -8.416e+01  9.252e+00  -9.096 8.32e-16 ***
## ElderlyPop  -4.355e+01  2.043e+01  -2.132   0.0347 *
## GDP          9.770e-03  4.967e-03   1.967   0.0512 .
## Cell         2.035e+00  1.663e+00   1.224   0.2230
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 608.3 on 140 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8075, Adjusted R-squared:  0.7979
## F-statistic: 83.92 on 7 and 140 DF,  p-value: < 2.2e-16
```

```r
# Residual analysis [3] ---

# Residual plots
plot(fit3, which = 1)
```




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



Residuals are looking better. Variance has tightened up and linearity as well. Lower end of QQ plot tucked in a little and the rest almost perfectly normal.

We will run another model selection process to see if any predictors have become insignificant.

```
# Model Selection [3] ----
## Forward selection
fit3_forward <- regsubsets(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cel
                           data = country80, method = "forward")
cbind(summary(fit3_forward)$which, "adjusted r^2" = summary(fit3_forward)$adjr2)
```

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

```
## Backward elimination
fit3_backward <- regsubsets(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Ce
                            data = country80, method = "backward")
cbind(summary(fit3_backward)$which, "adjusted r^2" = summary(fit3_backward)$adjr2)
```

```
##   (Intercept) Rural Health Internet BirthRate ElderlyPop GDP Cell adjusted r^2
```

```
## 1           1     0      0        0         1          0   0    0    0.7364894
## 2           1     0      0        1         1          0   0    0    0.7756859
## 3           1     1      0        1         1          0   0    0    0.7852103
## 4           1     1      1        1         1          0   0    0    0.7891219
## 5           1     1      1        1         1          1   0    0    0.7932579
## 6           1     1      1        1         1          1   1    0    0.7972095
## 7           1     1      1        1         1          1   1    1    0.7979232
```

```
# Mallow's Cp
fit3_subset <- regsubsets(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP + Cel
                          data = country80, method = "exhaustive")
cbind(summary(fit3_subset)$which, "Mallows' Cp" = summary(fit3_subset)$cp)
```

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

Step-wise method says to keep Cell, but not by a lot.

Mallows' Cp seems to fit the model better without cell.

Furthermore t-test for cell suggests it is not significant after other predictors accounted for.

Try building model without it and see if residuals tighten up.

```
## fit model and summary output [4] -----
fit4 <- lm(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop + GDP, data = country80)
summary(fit4)
```

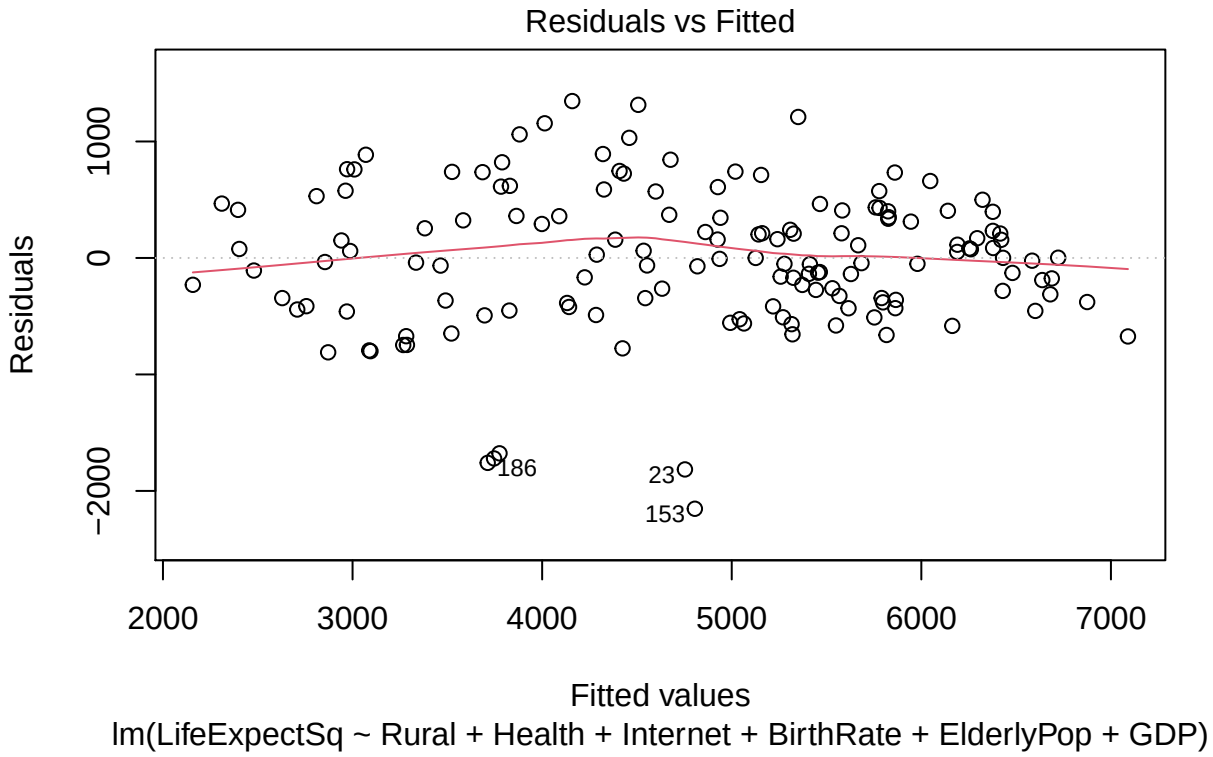
```
##
## Call:
## lm(formula = LifeExpectSq ~ Rural + Health + Internet + BirthRate +
##     ElderlyPop + GDP, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2153.15  -379.01     0.57   397.42  1346.88
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.848e+03  3.352e+02  20.434   <2e-16 ***
## Rural       -7.796e+00  3.086e+00  -2.526   0.0126 *
## Health       2.631e+01  1.259e+01   2.090   0.0384 *
## Internet     8.367e+00  4.219e+00   1.983   0.0493 *
## BirthRate   -8.841e+01  8.590e+00 -10.292   <2e-16 ***
## ElderlyPop  -4.715e+01  2.025e+01  -2.328   0.0213 *
## GDP          9.655e-03  4.975e-03   1.941   0.0543 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 609.4 on 141 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8055, Adjusted R-squared:  0.7972
## F-statistic: 97.31 on 6 and 141 DF,  p-value: < 2.2e-16
```

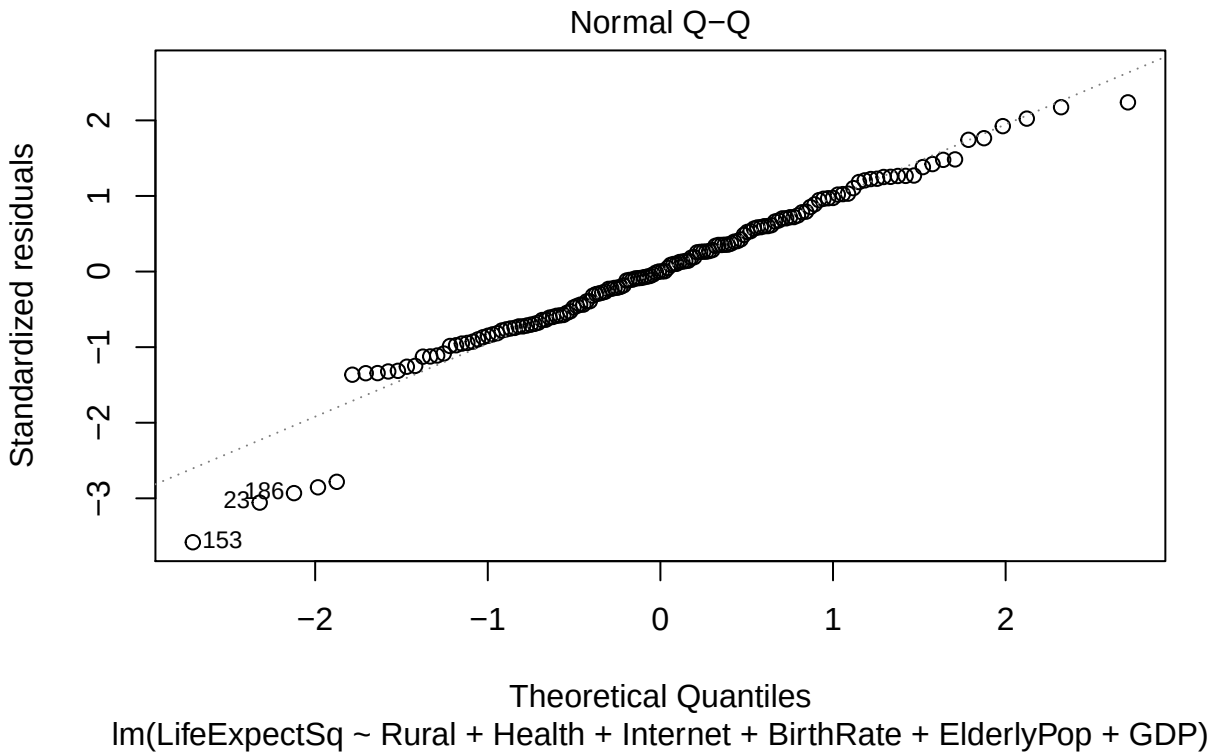
adjR^2 dropped just a little bit. But t-tests for all variables are now significant.

Let's carry out residual analysis, and see if it is any better.

```
# Residual analysis [4] ---

# Residual plots
plot(fit4, which = 1)
```



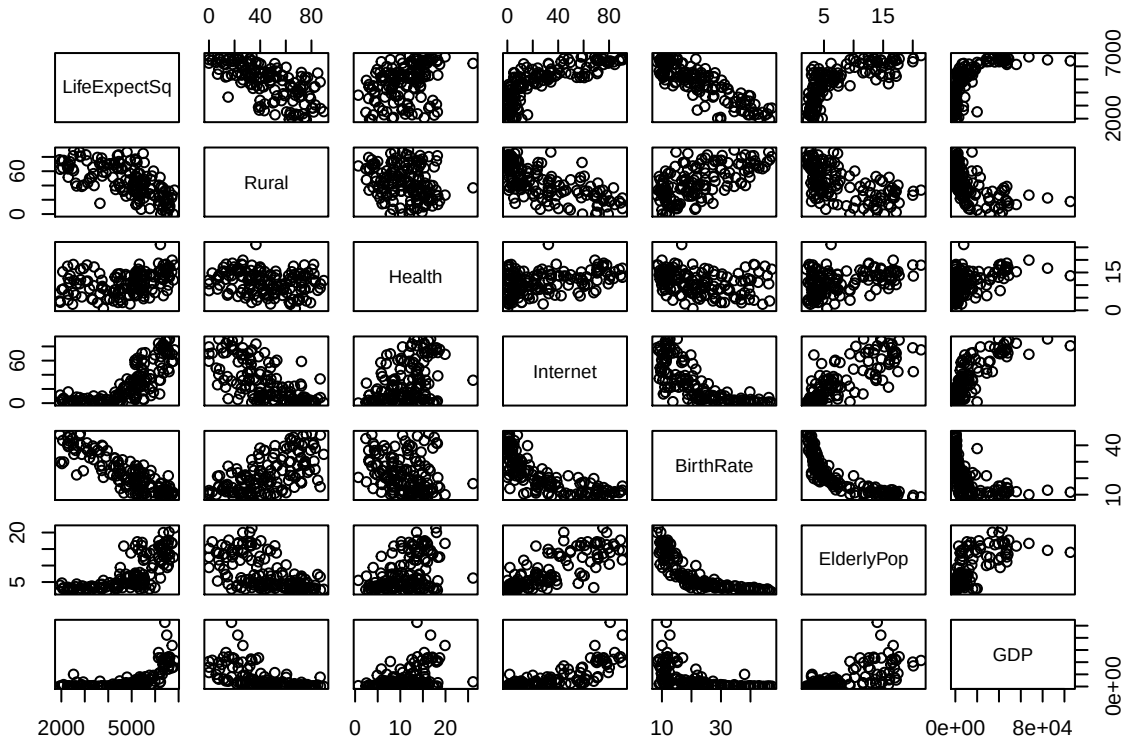

```
#Normal Probability Plot
plot(fit4, which = 2)
```



hmmm QQ plot got worse. Let's roll with this for now though. Try some X transformations.

We'll plot pairs() to see what could be more linear.

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




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

```
##
##  Pearson's product-moment correlation
##
## data:  Rural and BirthRate
## t = 9.4013, df = 146, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5025229 0.7055204
## sample estimates:
##       cor
## 0.6140779
```

I'll try transforming:

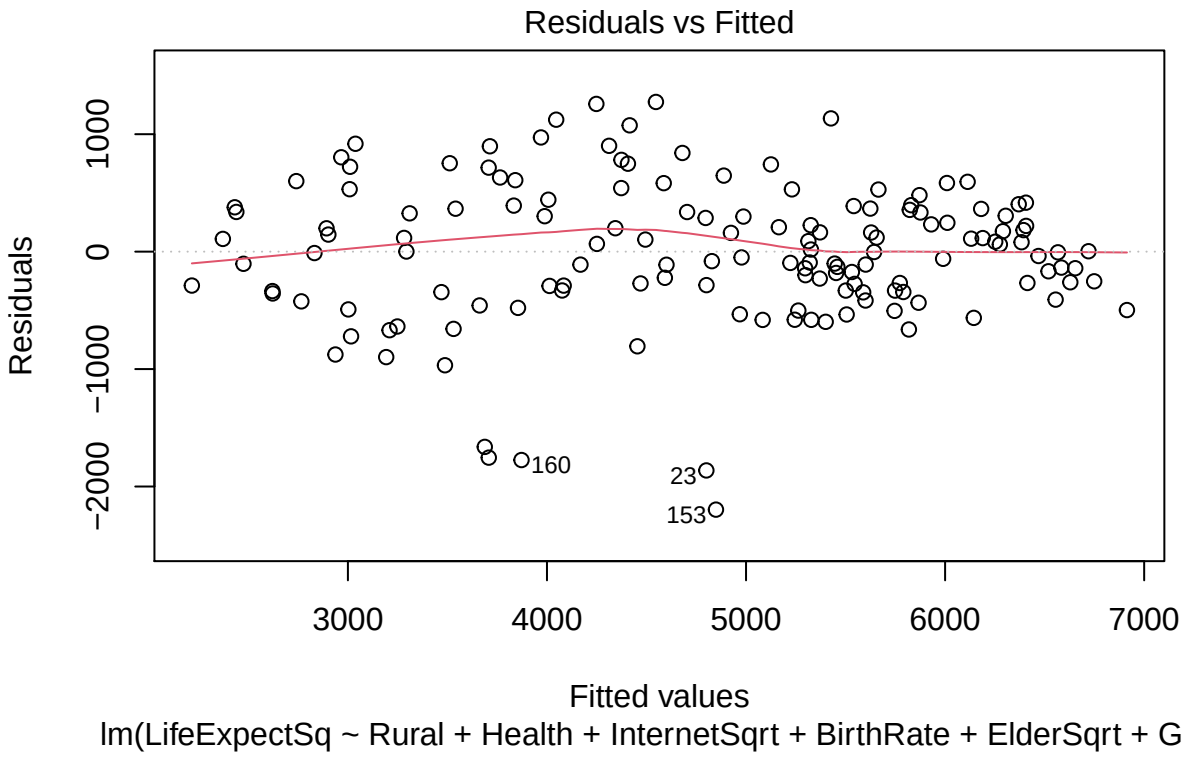
internet -> internet^(1/2) elderlypop -> elderlypop^(1/2) GDP -> GDP^(1/2)

```
InternetSqrt <- Internet^(.5)
ElderSqrt <- ElderlyPop^(.5)
GDPsqrt <- GDP^(.5)

## fit model and summary output [5] -----
fit5 <- lm(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate + ElderSqrt + GDPsqrt, data = counti
summary(fit5)
```

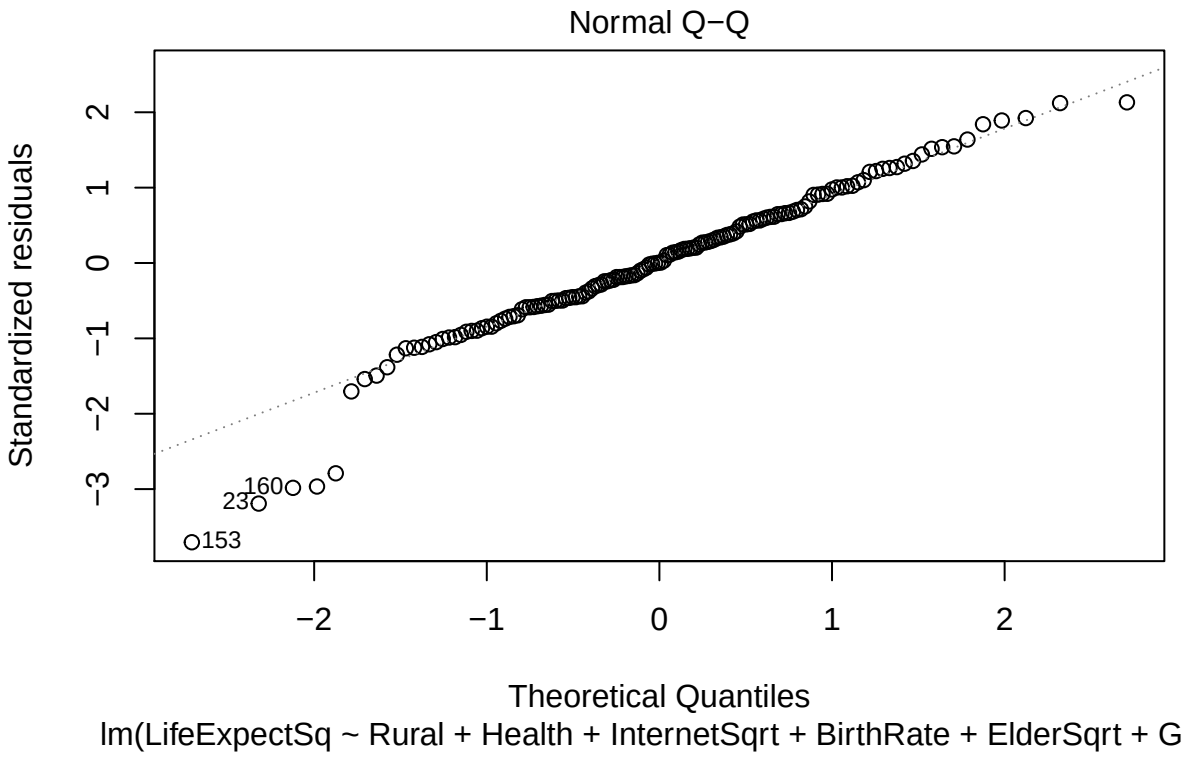
```
##
## Call:
## lm(formula = LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate +
##     ElderSqrt + GDPsqrt, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2196.67  -331.61     2.18   365.71  1274.34
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6699.619    509.300  13.155  < 2e-16 ***
## Rural          -5.916      3.205  -1.846   0.0670 .
## Health         24.473     12.526   1.954   0.0527 .
## InternetSqrt   89.764     43.224   2.077   0.0396 *
## BirthRate     -83.520      9.657  -8.649 1.05e-14 ***
## ElderSqrt    -299.805    126.530  -2.369   0.0192 *
## GDPsqrt         3.873      1.510   2.565   0.0113 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 602.5 on 141 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8099, Adjusted R-squared:  0.8018
## F-statistic: 100.1 on 6 and 141 DF,  p-value: < 2.2e-16
```

adjR2 remains the same as fit4. predictors all still significant.

```
# Residual analysis [5] ---

# Residual plots
plot(fit5, which = 1)
```




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



Best looking QQ so far. Still have the same problem with non-constant variance.

Try model selection again.

```
# Model Selection [5] ----

## Forward selection
fit5_forward <- regsubsets(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate + ElderSqrt + GDPsq
                           data = country80, method = "forward")
cbind(summary(fit5_forward)$which, "adjusted r^2" = summary(fit5_forward)$adjr2)
```

```
##   (Intercept) Rural Health InternetSqrt BirthRate ElderSqrt GDPsqrt
## 1           1     0      0            0         1         0       0
## 2           1     0      0            0         1         0       1
## 3           1     0      0            1         1         0       1
## 4           1     0      0            1         1         1       1
## 5           1     0      1            1         1         1       1
## 6           1     1      1            1         1         1       1
##   adjusted r^2
## 1    0.7364894
## 2    0.7842098
## 3    0.7903413
## 4    0.7954793
## 5    0.7984503
## 6    0.8018114
```

```r
## Backward elimination
fit5_backward <- regsubsets(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate + ElderSqrt + GDPs
                            data = country80, method = "backward")
cbind(summary(fit5_backward)$which, "adjusted r^2" = summary(fit5_backward)$adjr2)
```

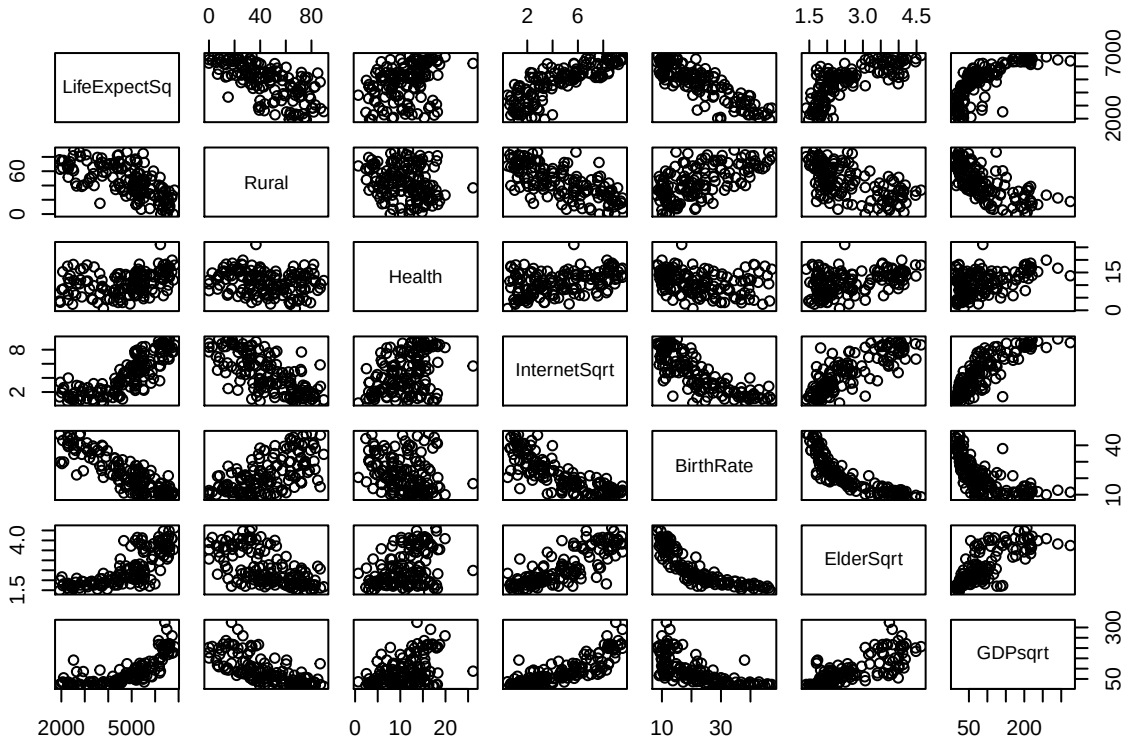
```
##   (Intercept) Rural Health InternetSqrt BirthRate ElderSqrt GDPsqrt
## 1           1     0      0            0         1         0       0
## 2           1     0      0            0         1         0       1
## 3           1     0      0            1         1         0       1
## 4           1     0      0            1         1         1       1
## 5           1     0      1            1         1         1       1
## 6           1     1      1            1         1         1       1
##   adjusted r^2
## 1    0.7364894
## 2    0.7842098
## 3    0.7903413
## 4    0.7954793
## 5    0.7984503
## 6    0.8018114
```

```r
# Mallow's Cp
fit5_subset <- regsubsets(LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate + ElderSqrt + GDPsqr
                          data = country80, method = "exhaustive")
cbind(summary(fit5_subset)$which, "Mallows' Cp" = summary(fit5_subset)$cp)
```

```
##   (Intercept) Rural Health InternetSqrt BirthRate ElderSqrt GDPsqrt Mallows' Cp
## 1           1     0      0            0         1         0       0   50.120852
## 2           1     0      0            0         1         0       1   15.877756
## 3           1     0      0            1         1         0       1   12.333948
## 4           1     0      0            1         1         1       1    9.568842
## 5           1     0      1            1         1         1       1    8.408165
## 6           1     1      1            1         1         1       1    7.000000
```

All model selection methods confirm current model is best.

```r
pairs(cbind(LifeExpectSq, Rural, Health, InternetSqrt, BirthRate, ElderSqrt, GDPsqrt))
```



InternetSqrt looks much more linear now than internet did agains LifeExpectency.

Eldersqrt saw a bit of improvement.

GDP sqrt got closer to linear as well, but still not as much as the other two.

Rural and BirthRate are quite collinear, yet BirthRate has a stronger linear relationship with LifeExpectancy.

Drop Rural, build new model. Compare results.

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

```
##
## Call:
## lm(formula = LifeExpectSq ~ Health + InternetSqrt + BirthRate +
##     ElderSqrt + GDPsqrt, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2140.58  -333.73   -20.19   372.70  1295.78
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6390.883    485.122  13.174  < 2e-16 ***
## Health         22.156     12.568   1.763  0.08006 .
## InternetSqrt  105.793     42.701   2.478  0.01440 *
```

```
## BirthRate     -85.596      9.672  -8.850 3.16e-15 ***
## ElderSqrt    -313.841    127.368  -2.464  0.01493 *
## GDPsqrt         4.721      1.450   3.255  0.00142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 607.6 on 142 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8053, Adjusted R-squared:  0.7985
## F-statistic: 117.5 on 5 and 142 DF,  p-value: < 2.2e-16
```

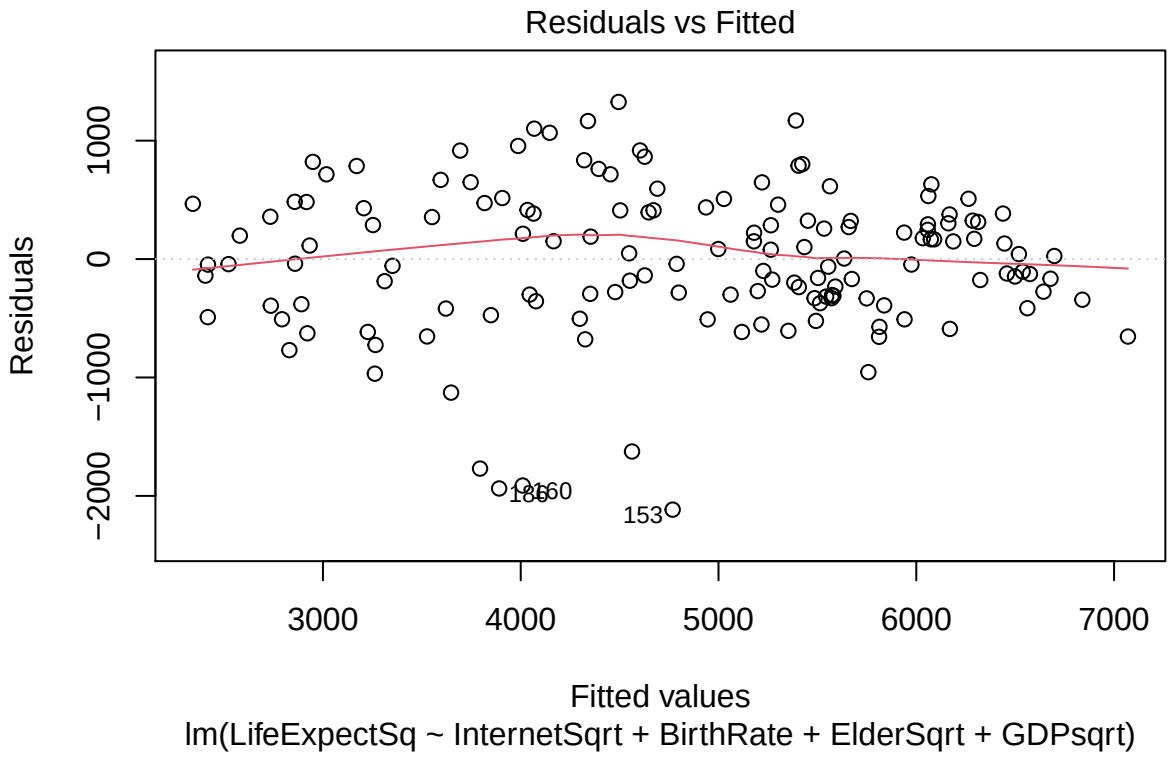
All predictors are significant except Health. Drop health. New model

```
## fit model and summary output [6] -----
fit7 <- lm(LifeExpectSq ~ InternetSqrt + BirthRate + ElderSqrt + GDPsqrt, data = country80)
summary(fit7)
```

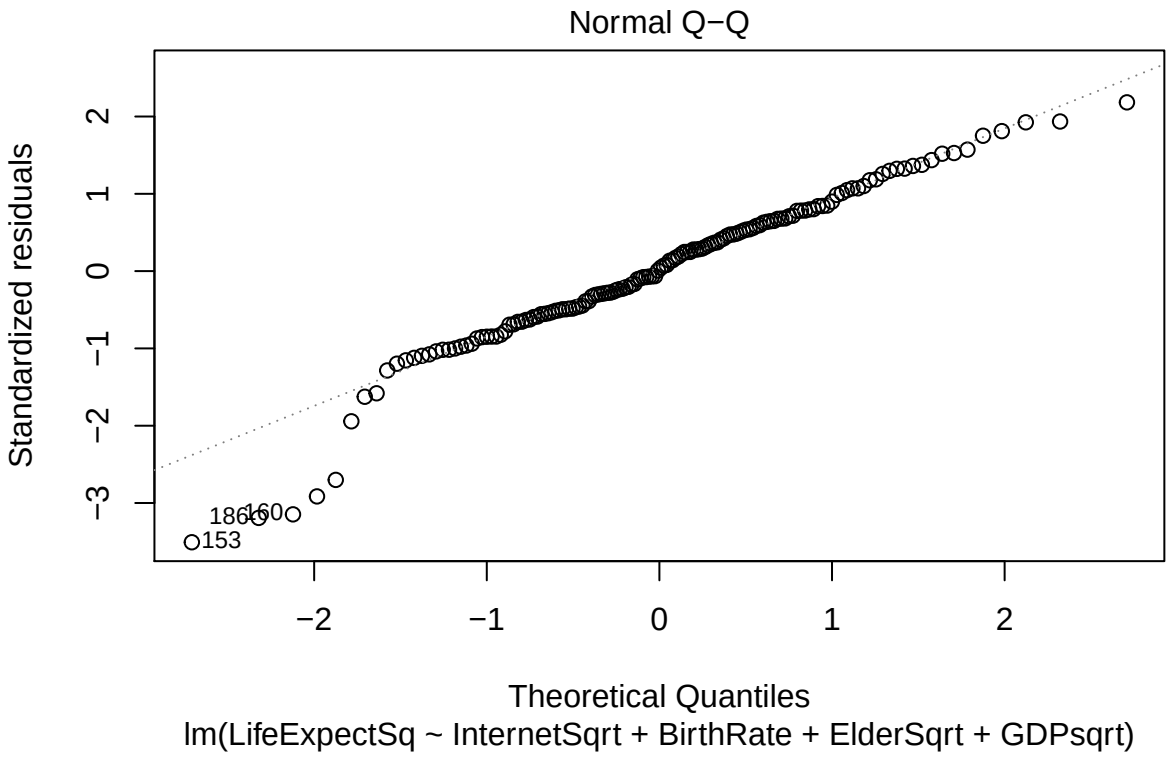
```
##
## Call:
## lm(formula = LifeExpectSq ~ InternetSqrt + BirthRate + ElderSqrt +
##     GDPsqrt, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2115.95  -330.82    15.31   398.55  1326.78
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6389.493    488.684  13.075  < 2e-16 ***
## InternetSqrt  113.700     42.776   2.658 0.008756 **
## BirthRate     -82.371      9.567  -8.610 1.21e-14 ***
## ElderSqrt    -270.526    125.893  -2.149 0.033330 *
## GDPsqrt         4.981      1.453   3.427 0.000796 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 612 on 143 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.801,  Adjusted R-squared:  0.7955
## F-statistic: 143.9 on 4 and 143 DF,  p-value: < 2.2e-16
```

All predictors significant. AdjR2 has increased. Conduct model residual analysis.

```
# Residual analysis [7] ---

# Residual plots
plot(fit7, which = 1)
```




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




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

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit7)
## W = 0.9584, p-value = 0.0001952
```

qqplot sucks. try untransforming X vars.

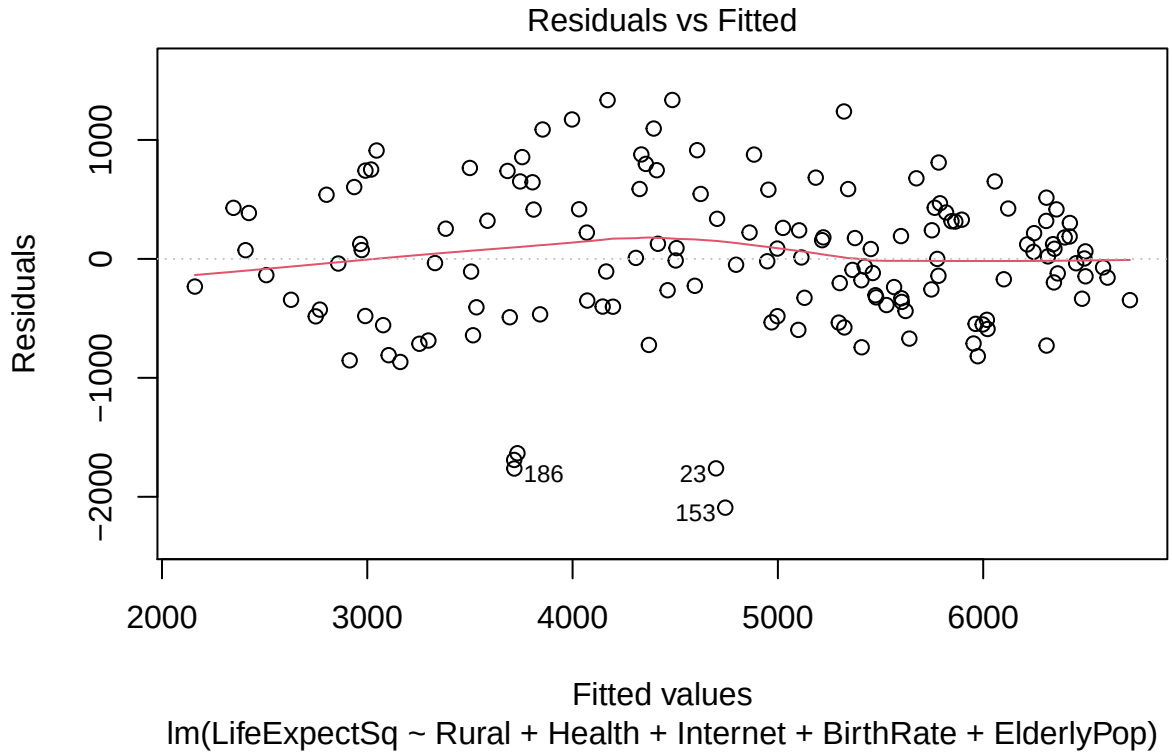
Back up to fit4. notice that GDP had t-test pval $> .05$

```
## fit model and summary output [8] -----
fit8 <- lm(LifeExpectSq ~ Rural + Health + Internet + BirthRate + ElderlyPop, data = country80)
summary(fit8)
```

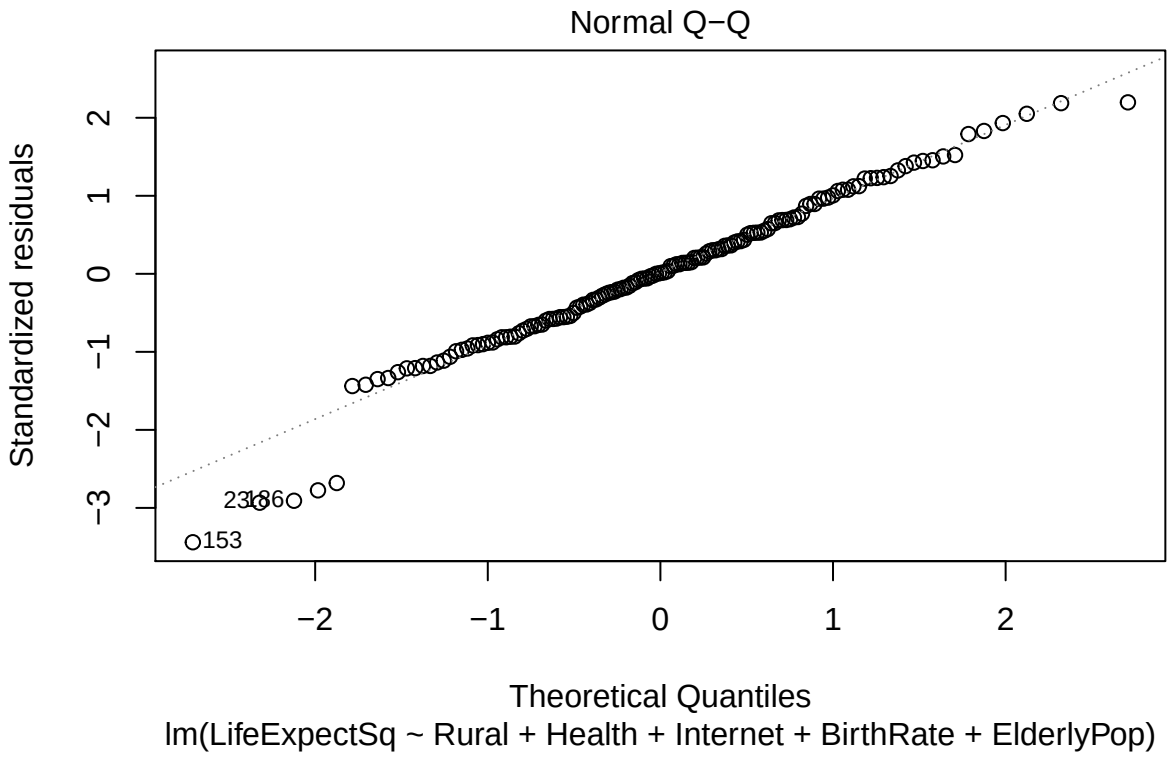
```
##
## Call:
## lm(formula = LifeExpectSq ~ Rural + Health + Internet + BirthRate +
##     ElderlyPop, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2091.47  -368.27     5.91   396.47  1335.88
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6709.962    330.653  20.293  < 2e-16 ***
## Rural         -8.714      3.079  -2.830  0.00533 **
## Health        28.800     12.642   2.278  0.02421 *
## Internet      12.664      3.627   3.492  0.00064 ***
## BirthRate    -84.497      8.432 -10.021  < 2e-16 ***
## ElderlyPop   -39.381     20.042  -1.965  0.05138 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 615.3 on 142 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.8003, Adjusted R-squared:  0.7933
## F-statistic: 113.8 on 5 and 142 DF,  p-value: < 2.2e-16
```

```r
# Residual analysis [8] ---

# Residual plots
plot(fit8, which = 1)
```




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




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

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit8)
## W = 0.97112, p-value = 0.003265
```

model looking pretty good, however, elderlypop has pval $> .05$ and collinear with internet.

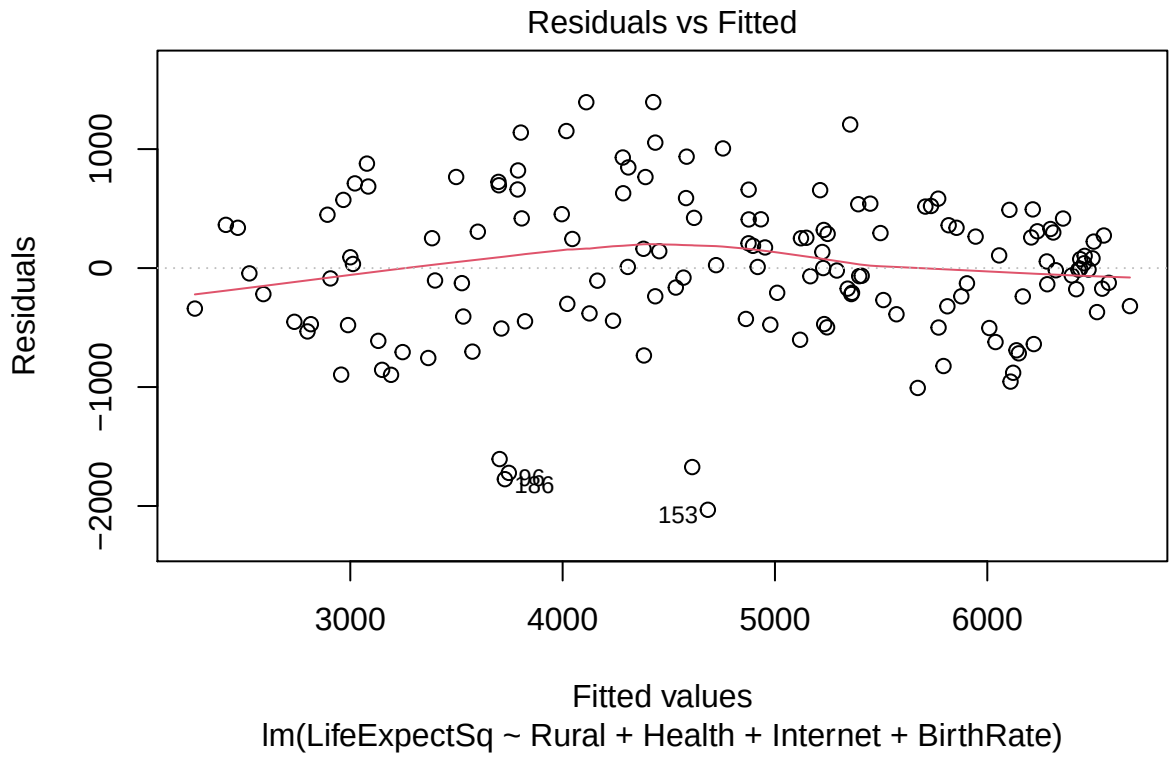
drop elderlypop.

```
## fit model and summary output [9] -----
fit9 <- lm(LifeExpectSq ~ Rural + Health + Internet + BirthRate, data = country80)
summary(fit9)
```

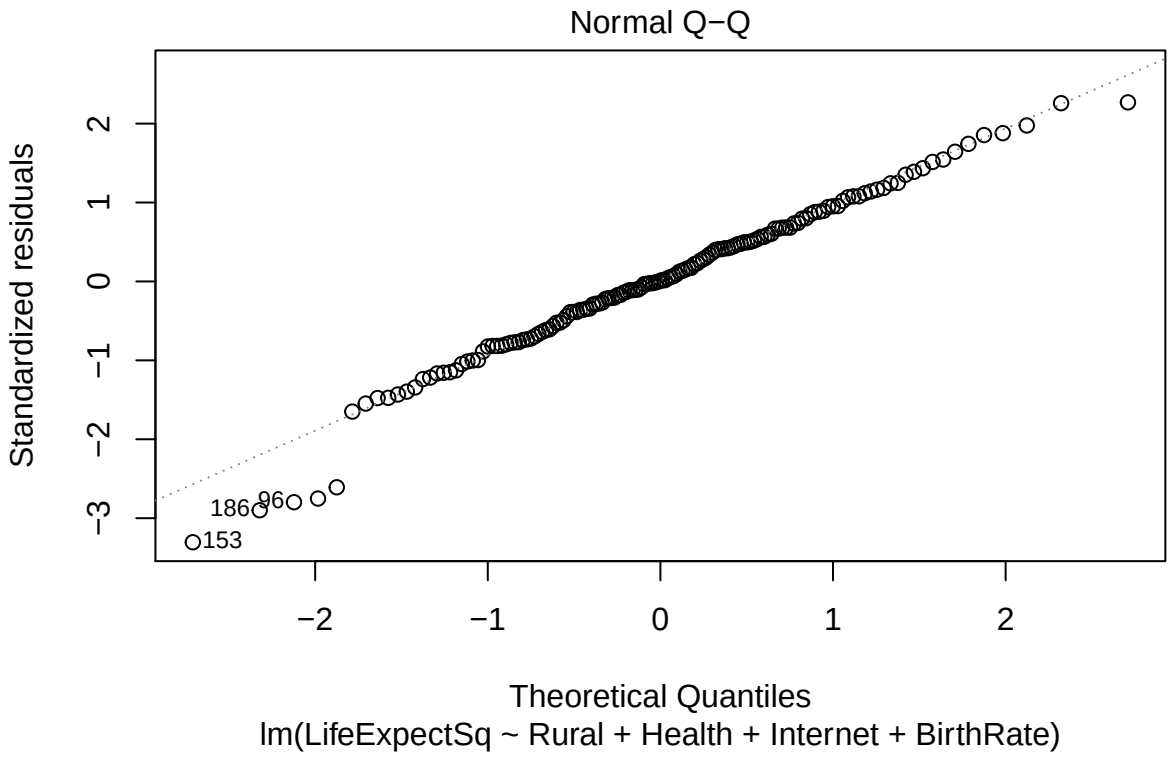
```
##
## Call:
## lm(formula = LifeExpectSq ~ Rural + Health + Internet + BirthRate,
##     data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2031.67  -382.93     4.78   409.11  1394.98
##
```

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

```r
# Residual analysis [9] ---

# Residual plots
plot(fit9, which = 1)
```



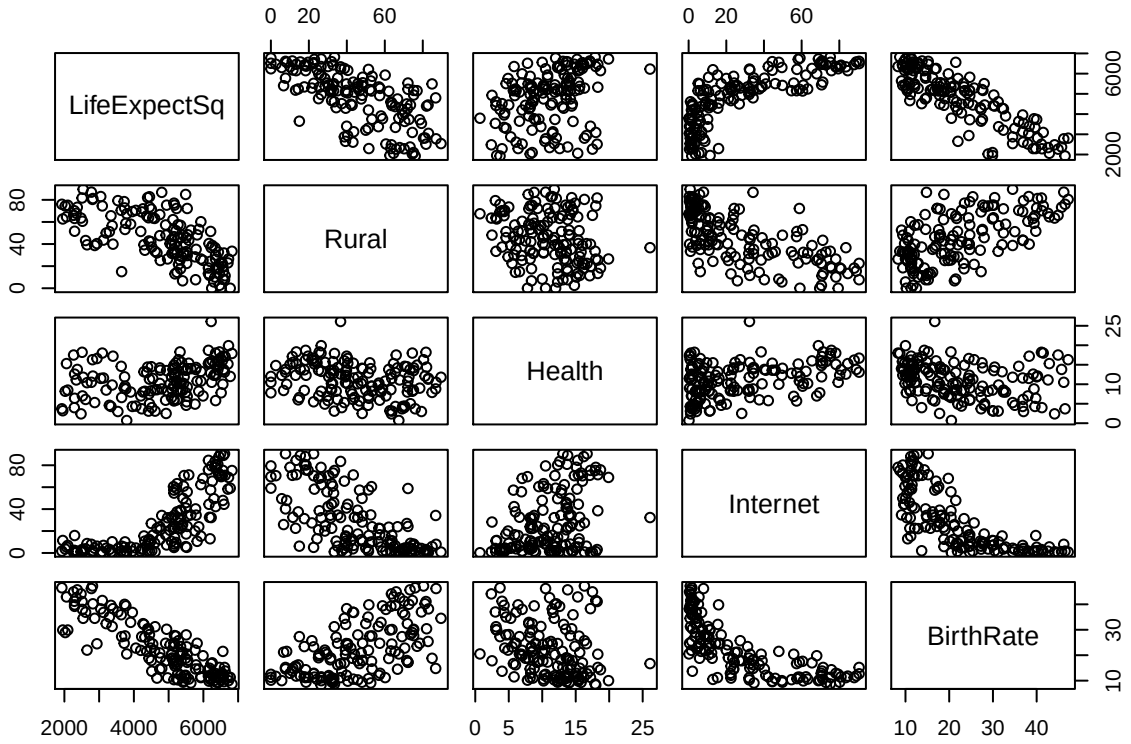

```r
#Normal Probability Plot
plot(fit9, which = 2)
```




```
shapiro.test(resid(fit9))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit9)
## W = 0.97807, p-value = 0.01806
```

```
pairs(cbind(LifeExpectSq, Rural, Health, Internet, BirthRate))
```



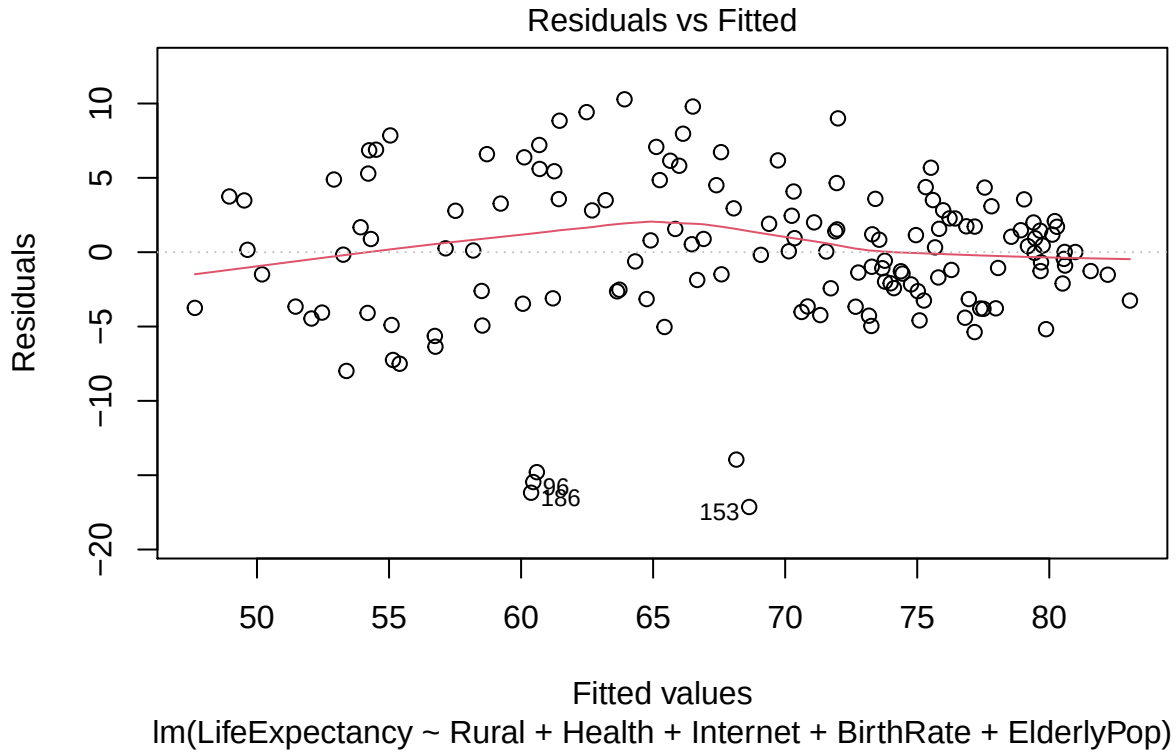
looks quite good. see about untransforming Y, for comparison to above model.

```
## fit model and summary output [10] -----
fit10 <- lm(LifeExpectancy ~ Rural + Health + Internet + BirthRate + ElderlyPop, data = country80)
summary(fit10)
```

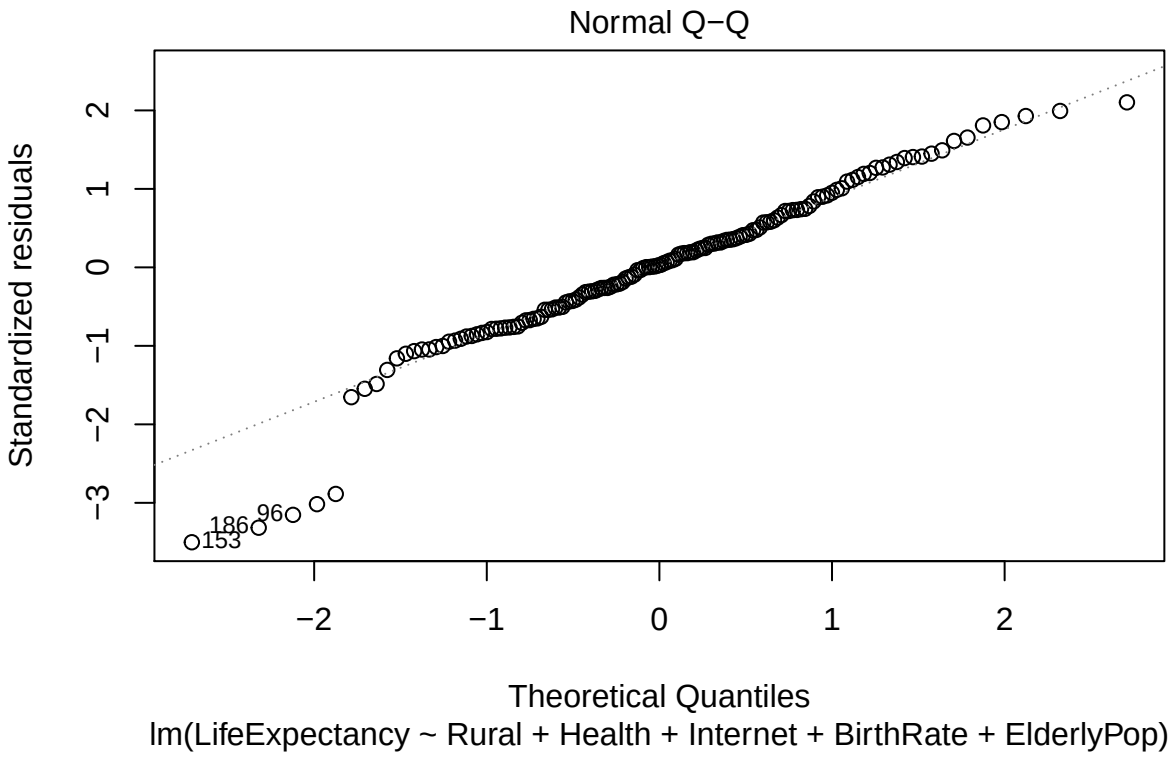
```
##
## Call:
## lm(formula = LifeExpectancy ~ Rural + Health + Internet + BirthRate +
##     ElderlyPop, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1379  -2.7523   0.1361   2.9832  10.2826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 85.70067    2.66149  32.200  < 2e-16 ***
## Rural       -0.06360    0.02478  -2.566  0.01132 *
## Health       0.20962    0.10175   2.060  0.04122 *
## Internet     0.08420    0.02919   2.884  0.00453 **
## BirthRate   -0.71564    0.06787 -10.545  < 2e-16 ***
## ElderlyPop  -0.39165    0.16133  -2.428  0.01645 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 4.953 on 142 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7828, Adjusted R-squared:  0.7751
## F-statistic: 102.3 on 5 and 142 DF,  p-value: < 2.2e-16
```

```r
# Residual analysis [10] ---

# Residual plots
plot(fit10, which = 1)
```




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




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

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit10)
## W = 0.95109, p-value = 4.55e-05
```

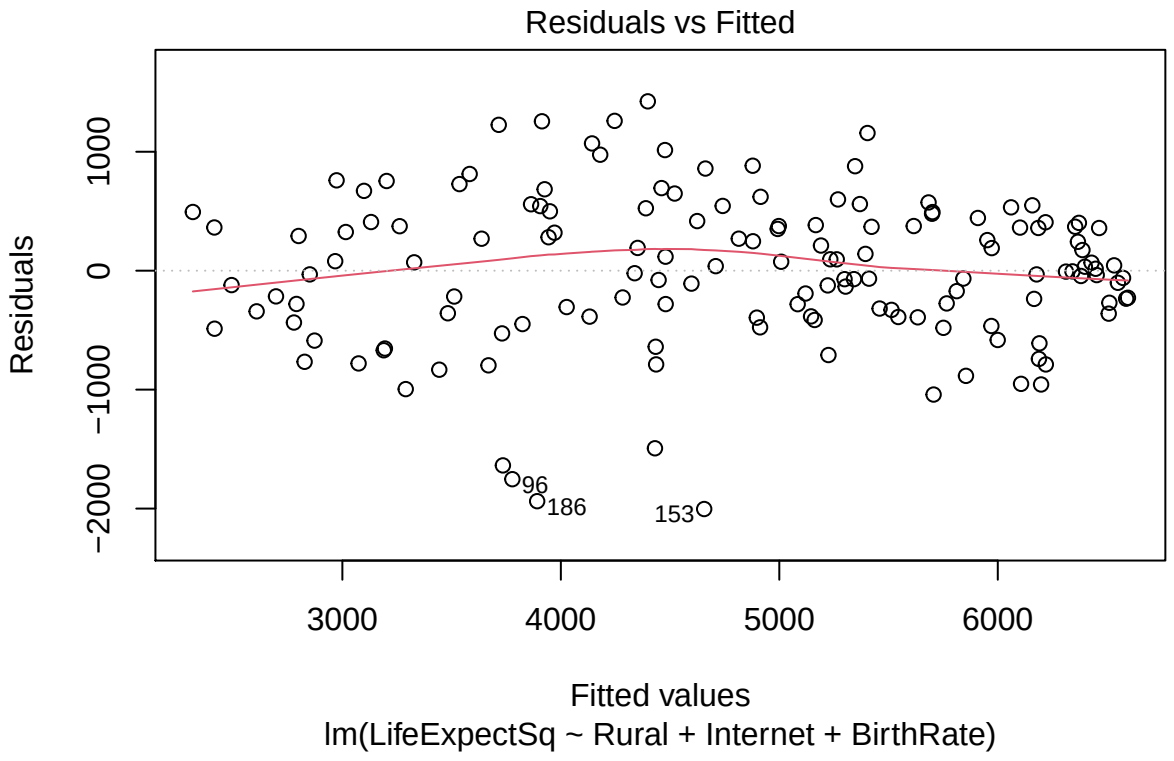
Yea fit9 is best. but health still has borderline pval. try dropping to compare fit11 with fit9 and make judgement call about significance of Health predictor.

```
## fit model and summary output [11] -----
fit11 <- lm(LifeExpectSq ~ Rural + Internet + BirthRate, data = country80)
summary(fit11)
```

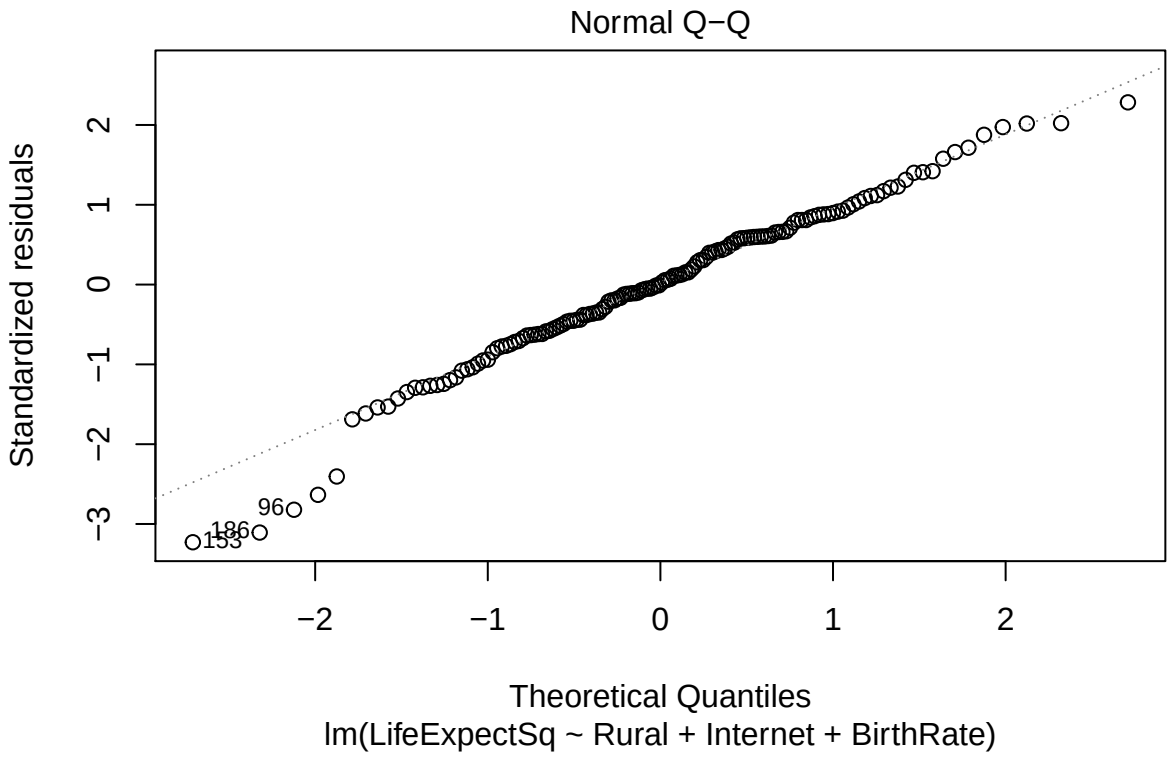
```
##
## Call:
## lm(formula = LifeExpectSq ~ Rural + Internet + BirthRate, data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2003.80  -366.99     5.13   403.27  1423.65
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6555.016    274.218  23.904  < 2e-16 ***
## Rural         -8.537      3.132  -2.726 0.007212 **
## Internet      11.327      3.091   3.664 0.000348 ***
## BirthRate    -75.494      7.407 -10.192  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 627.2 on 144 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7896, Adjusted R-squared:  0.7852
## F-statistic: 180.1 on 3 and 144 DF,  p-value: < 2.2e-16
```

```r
# Residual analysis [11] ---

# Residual plots
plot(fit11, which = 1)
```



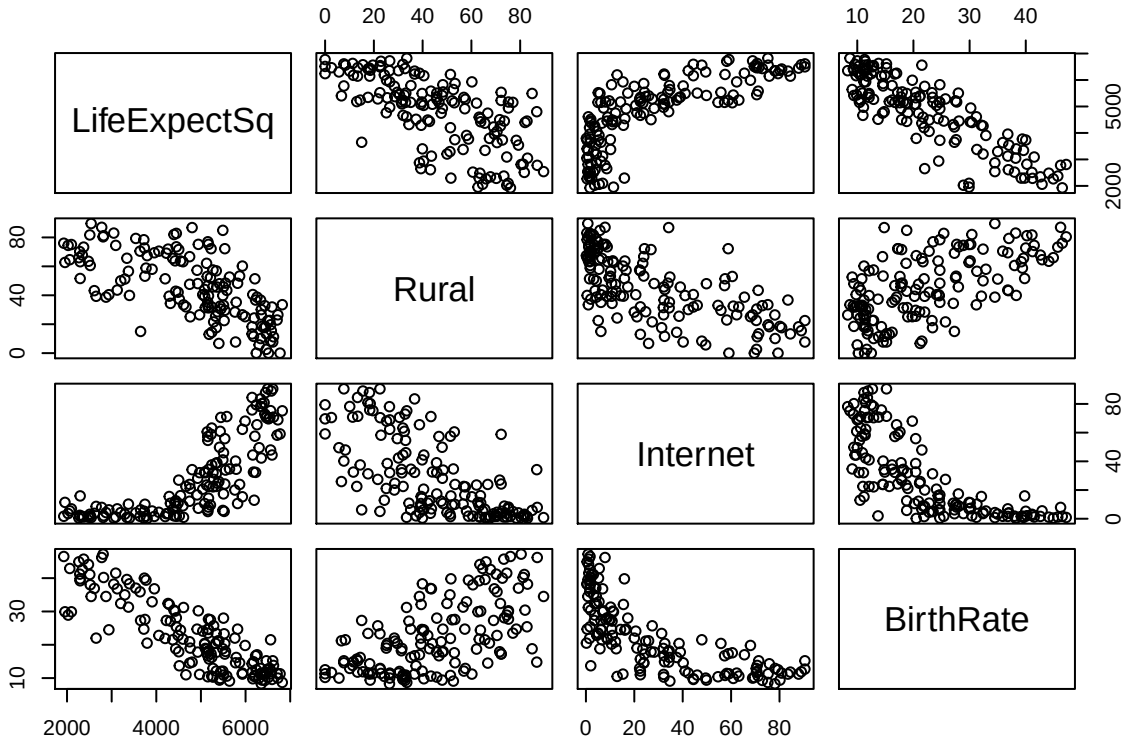

```r
#Normal Probability Plot
plot(fit11, which = 2)
```




```
shapiro.test(resid(fit11))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit11)
## W = 0.97794, p-value = 0.01743
```

```
pairs(cbind(LifeExpectSq, Rural, Internet, BirthRate))
```



WORSE.

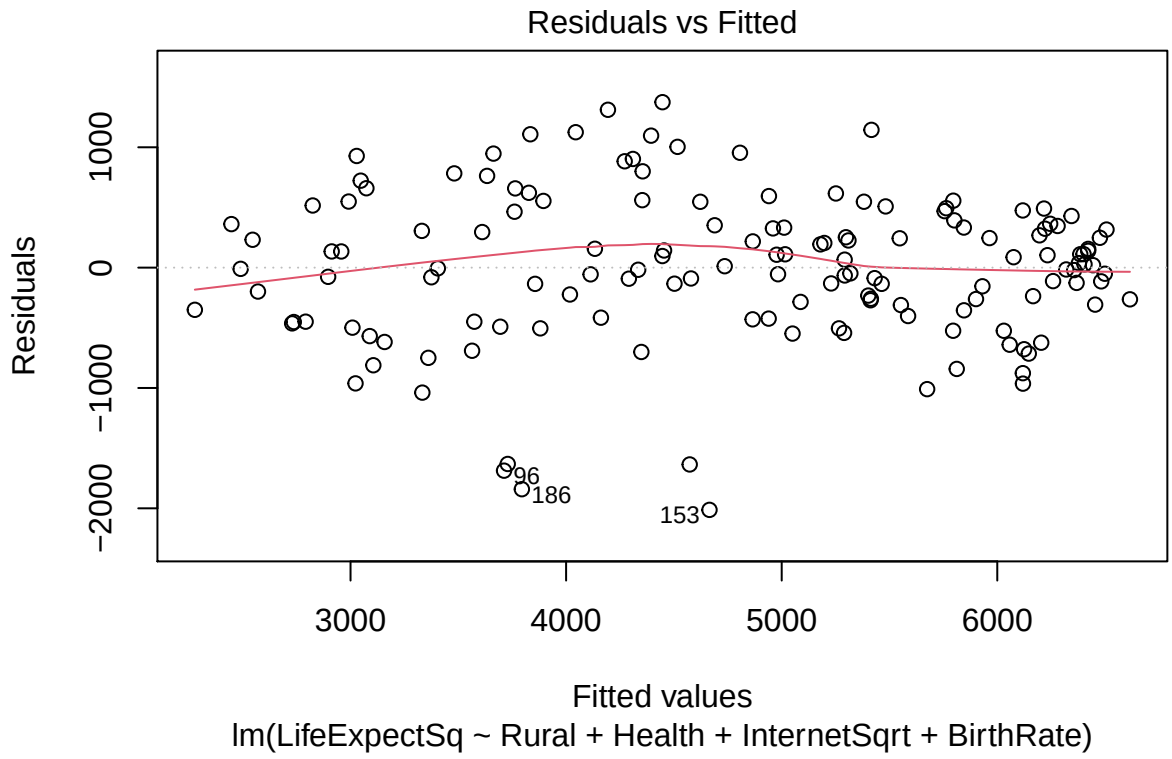
Back to fit9 try one more with internet transformed to sqrt(internet)

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

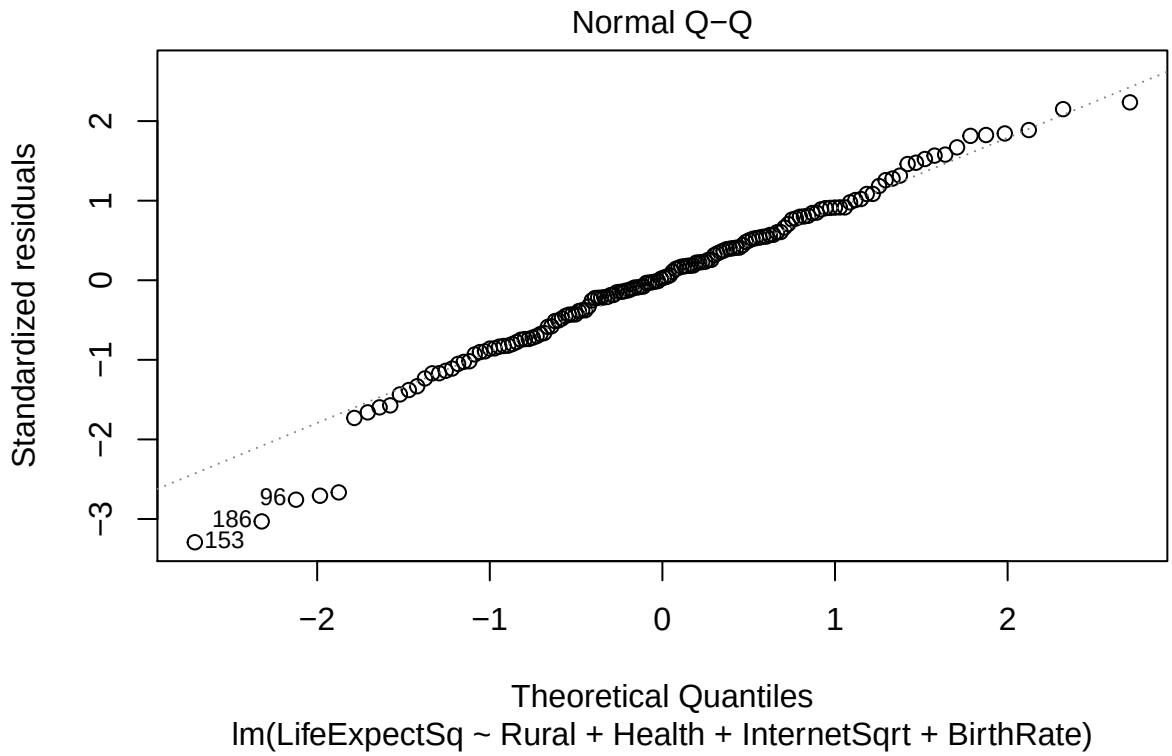
```
##
## Call:
## lm(formula = LifeExpectSq ~ Rural + Health + InternetSqrt + BirthRate,
##     data = country80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2012.86  -366.63    15.86   362.24  1374.63
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5956.840    375.762  15.853  < 2e-16 ***
## Rural          -8.263      3.130  -2.640  0.00921 **
## Health         23.601     12.394   1.904  0.05888 .
## InternetSqrt  117.630     37.245   3.158  0.00194 **
## BirthRate     -70.791      8.131  -8.706 6.95e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 618.1 on 143 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7971, Adjusted R-squared:  0.7914
## F-statistic: 140.4 on 4 and 143 DF,  p-value: < 2.2e-16
```

```r
# Residual analysis [12] ---

# Residual plots
plot(fit12, which = 1)
```



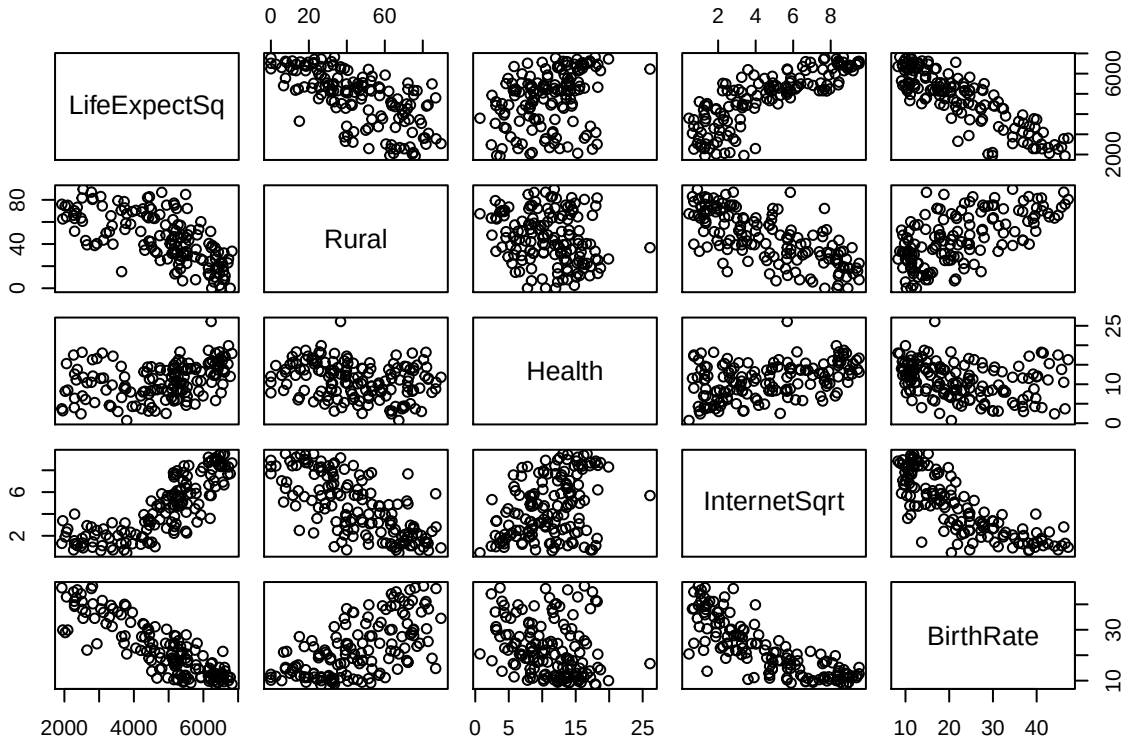

```r
#Normal Probability Plot
plot(fit12, which = 2)
```




```
shapiro.test(resid(fit12))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit12)
## W = 0.97583, p-value = 0.01026
```

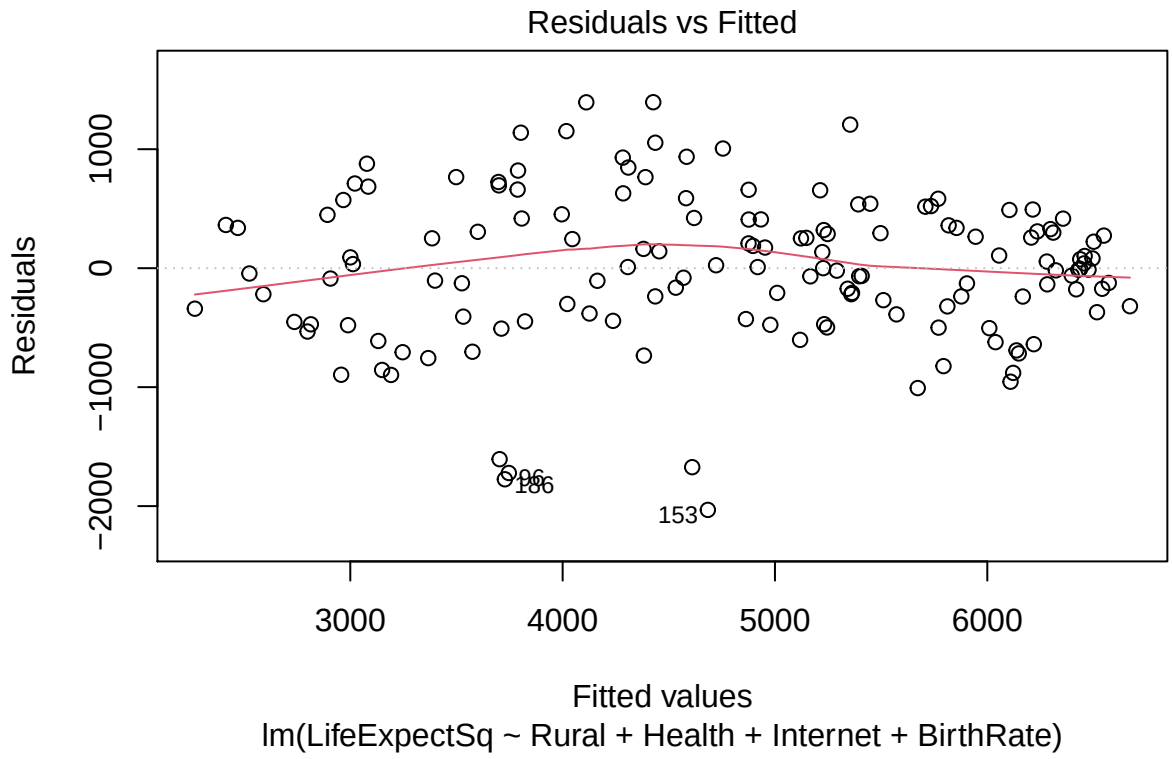
```
pairs(cbind(LifeExpectSq, Rural, Health, InternetSqrt, BirthRate))
```



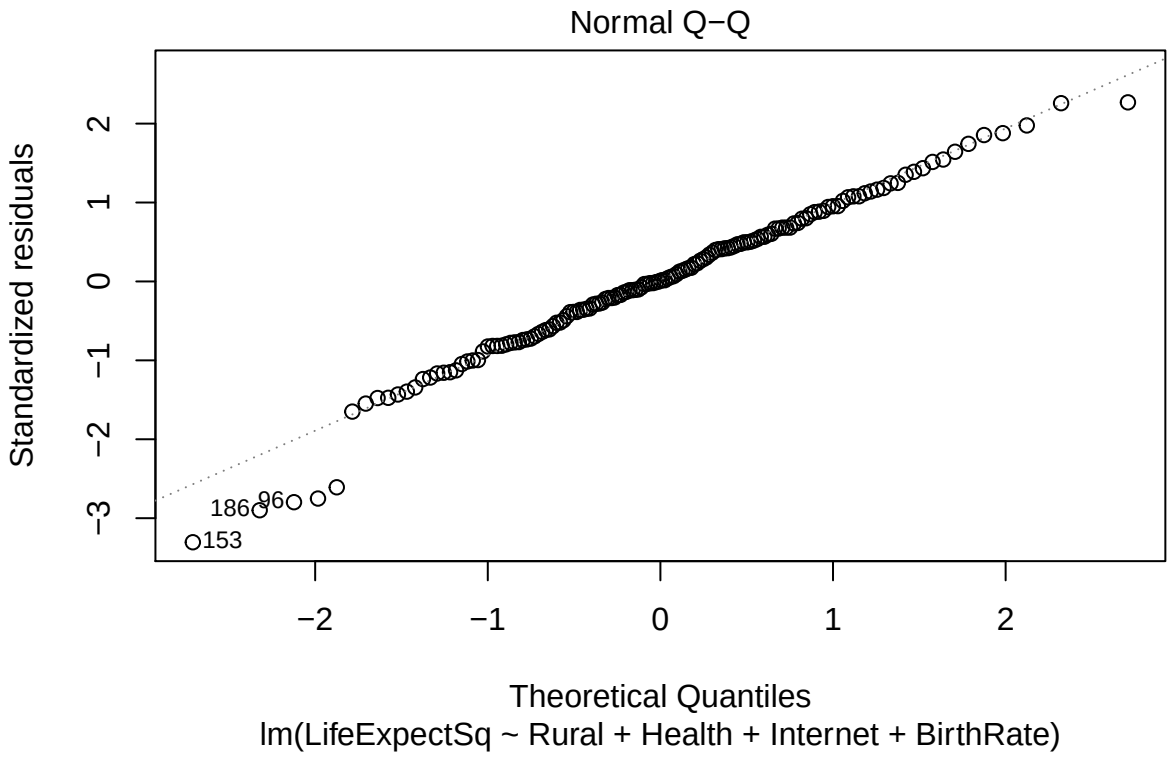

```
## fit 9 for comparison side-by-side --------
summary(fit9)
```

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

```r
# Residual analysis [9] ---

# Residual plots
plot(fit9, which = 1)
```



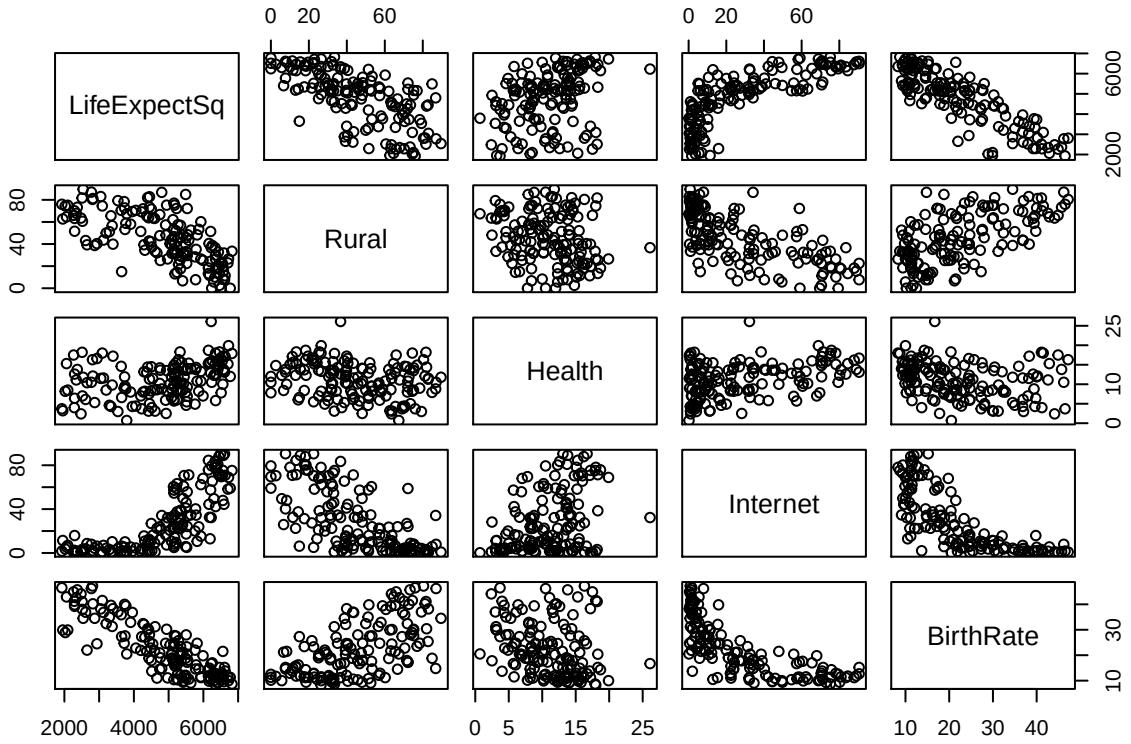

```r
#Normal Probability Plot
plot(fit9, which = 2)
```




```
shapiro.test(resid(fit9))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit9)
## W = 0.97807, p-value = 0.01806
```

```
pairs(cbind(LifeExpectSq, Rural, Health, Internet, BirthRate))
```



Although fit12 has higher adjR2 and all predictors highly significant, fit9 has what could be considered normality in qq plot. best of all models, the only variable of concern is health, but this is borderline and seems to be important in model.

fit9 is best model.