Stat 420 Project

Statistical Analysis of the Life Expectancy Dataset

Introduction

An Examination of Life Expectancy Across Various Countries The data is a combination of many different data sets from the World Health Organization website. The combined data allows us to get a better idea of the economic and health situations within each country. The data contains information about different countries' average life expectancy along with other macroeconomic predictors such as GDP, adult mortality rate, Income composition of resources, etc. This is concerning because countries experience rapid economic growth along with technological advances, but does any of this lead towards increasing the lifespan of humans?

```r
life_expectancy = read.csv("Life Expectancy Data.csv")
life_expectancy = filter(life_expectancy, Year == 2014)
life_expectancy = na.omit(life_expectancy)

#Plotting helper functions and assumption testing for later
loocv_rmse <- function(model) {
  sqrt(mean(resid(model) / (1 - hatvalues(model))) ^ 2)
}

plot_fitted_resid <- function(model, title, pointcol = "dodgerblue", linecol
= "darkorange") {
  plot(fitted(model), resid(model), col = pointcol, pch = 19, cex = 0.75,
xlab = "Fitted", ylab = "Residuals", main = title)
  abline(h = 0, col = linecol, lwd = 2)
}

plot_qq <- function(model, pointcol = "dodgerblue", linecol = "darkorange") {
  qqnorm(resid(model), col = pointcol, pch = 19, cex = 0.75)
  qqline(resid(mode), col = linecol, lwd = 2)
}
```

Dataset Explanation

The dataset *life_expectancy* was condensed down to 131 observations. The goal of this study is to predict lifetime expectancy using information on the following:

1. Country: Name of country
2. Year: Year in which data for that country was taken
3. Status: Dummy variable indicating 'developed' or 'undeveloped'
4. Life.expectancy: Number of years a person lives on average in specified country
5. Adult.mortality: Probability of dying between ages 15-60 per 1000 people

6. infant.deaths: Number of infant deaths per 1000 people
7. Alcohol: Recorded (15+) consumption of alcohol per capita
8. percentage.expenditure: Expenditure on health care as a % of GDP
9. Hepatitis.B: Percentage of HepB immunization among 1-year olds
10. Measles: Number of reported measles cases per 1000 people
11. BMI: Average Body Mass Index of entire population
12. under-five-deaths: Number of deaths under the age of 5 per 1000 people
13. Polio: Percentage of Polio (Pol3) immunizations among 1-year olds
14. Total.Expenditure: General government expenditure on healthcare as a percentage of total government expenditure
15. Diptheria: Diptheria immunization coverage percentage among 1-year olds
16. HIV/AIDS: Percentage of deaths per 1000 live births from 0-4 years old
17. GDP: Gross Domestic Product per capita in USD
18. Population: Size of the country
19. thinness.1.19.years: Prevalence of thinness among children and adolescents from age 10-19 (%)
20. thinness.5.9.years: Prevalence of thinness among children age 5-9 (%)
21. Income.composition.of.resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
22. Schooling: Number of years of schooling on average

Methods

Life Expectancy Data Analysis

```
avgLifetime = mean(life_expectancy$Life.expectancy)
max(life_expectancy$Life.expectancy)

## [1] 89

avgLifetime

## [1] 70.51985
```

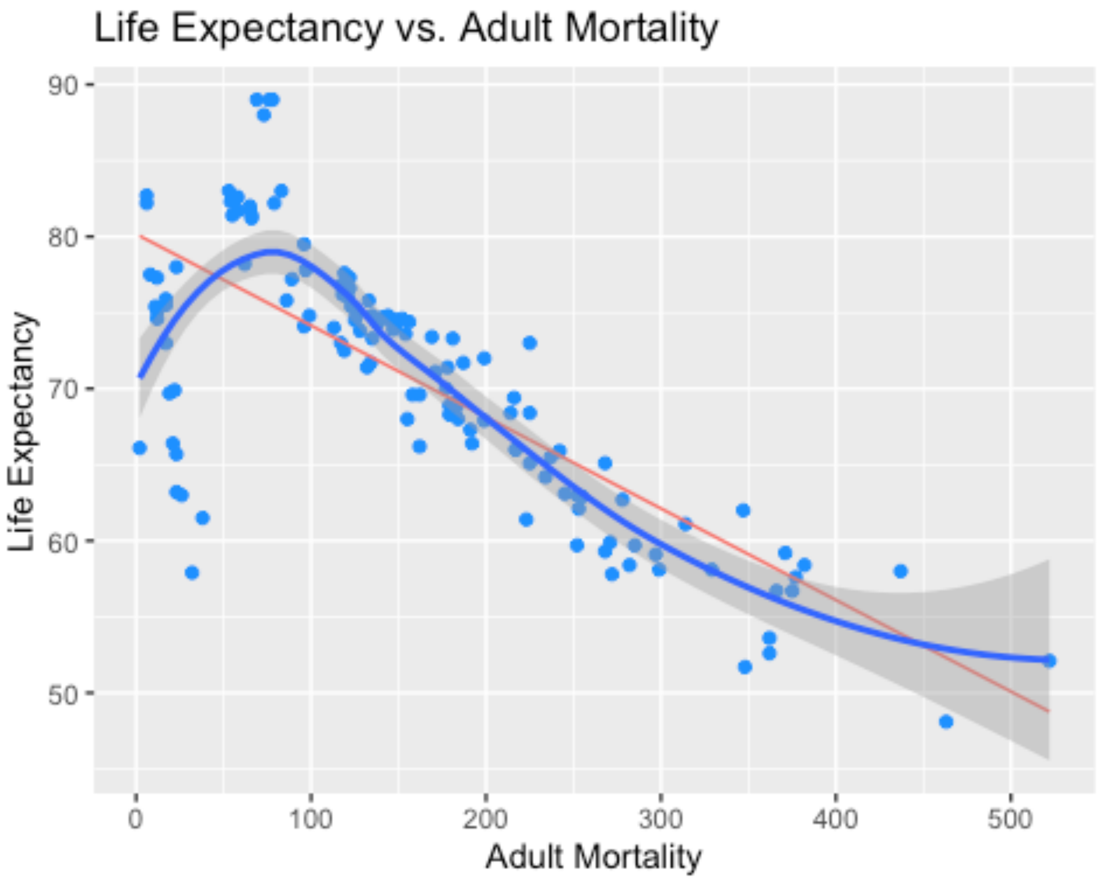
The average lifetime across all 131 countries is *70* and a half years, while Portugal has the highest life expectancy with *89* years.

Scatterplots

Adult Mortality

```
adultMortalityData = as.data.frame(cbind(life_expectancy$Adult.Mortality,
life_expectancy$Life.expectancy))
adultMortalityData = adultMortalityData[complete.cases(adultMortalityData),]
colnames(adultMortalityData) = c("Adult Mortality", "Life Expectancy")
pred.adultMortality <- predict(lm(life_expectancy$Life.expectancy ~
life_expectancy$Adult.Mortality, data = adultMortalityData))
```

```
adultPlot <- ggplot(adultMortalityData, aes(x =
life_expectancy$Adult.Mortality, y = life_expectancy$Life.expectancy))
adultPlot + geom_point(col = "dodgerblue", ) + geom_line(aes(y =
pred.adultMortality, col = "darkorange")) + geom_smooth() + ggtitle("Life
Expectancy vs. Adult Mortality") + xlab("Adult Mortality") + ylab("Life
Expectancy") +  theme(legend.position = "none")

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

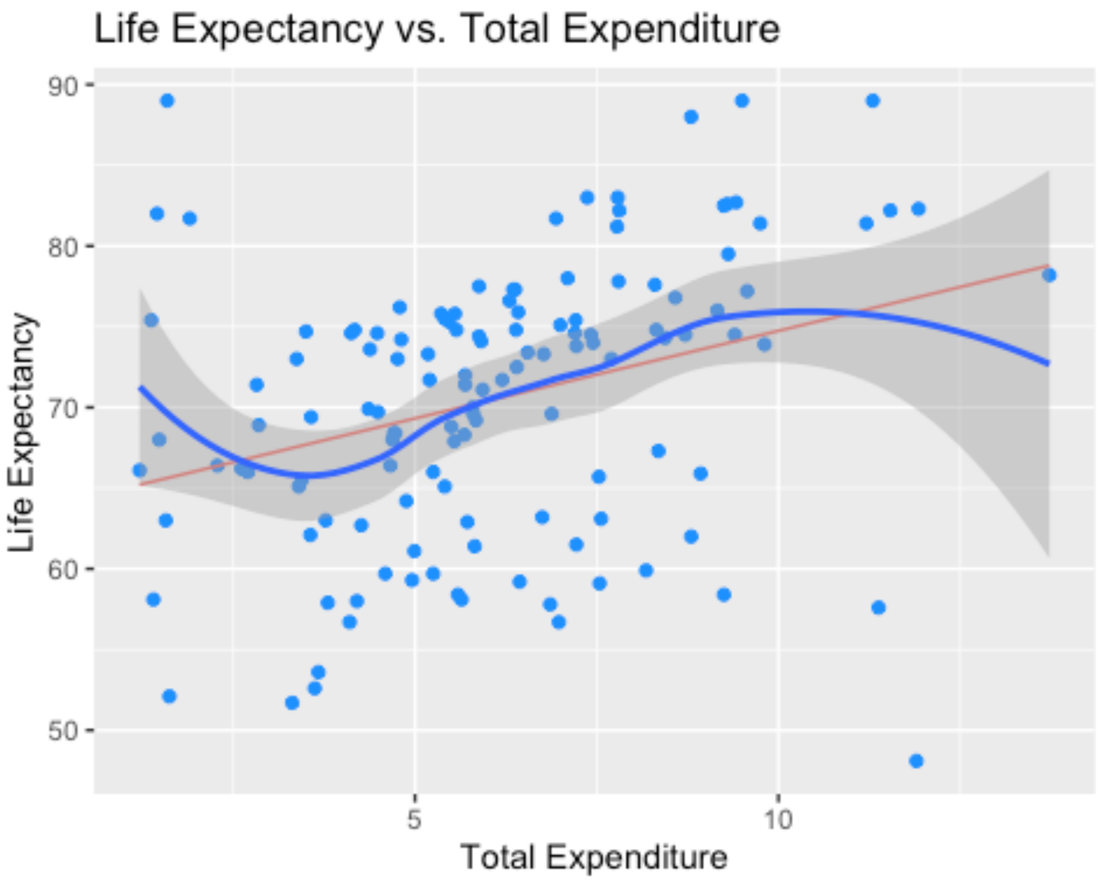


This plot shows the effect of adult mortality on life expectancy. There is a clear negative relationship between the two variables and variance increases with Adult Mortality levels above 400.

Total Expenditure

```
expendData = as.data.frame(cbind(life_expectancy$Total.expenditure,
life_expectancy$Life.expectancy))
expendData = expendData[complete.cases(expendData),]
colnames(expendData) = c("Total Expenditure", "Life Expectancy")
pred.expend <- predict(lm(life_expectancy$Life.expectancy ~
life_expectancy$Total.expenditure, data = expendData))

expendPlot <- ggplot(expendData, aes(x = life_expectancy$Total.expenditure, y
= life_expectancy$Life.expectancy))
```

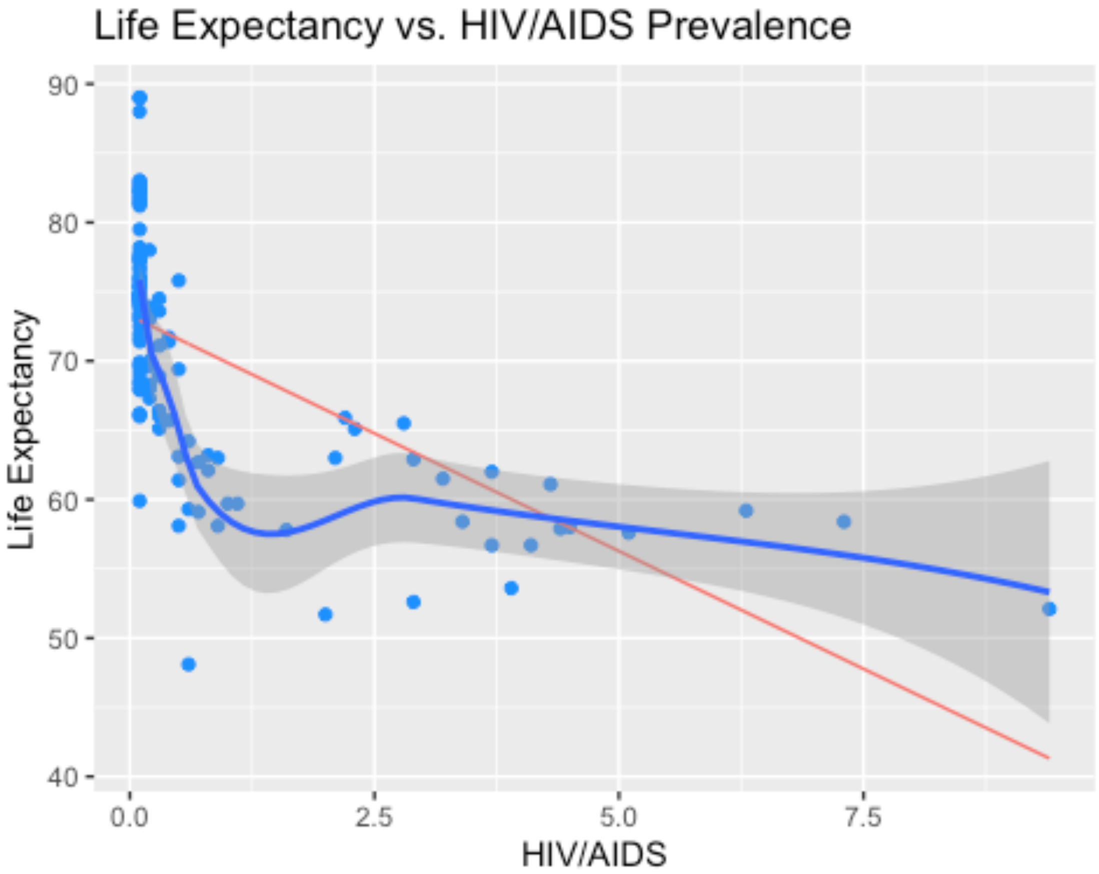
```
expendPlot + geom_point(col = "dodgerblue", ) + geom_line(aes(y =
pred.expend, col = "darkorange")) + geom_smooth() + ggtitle("Life Expectancy
vs. Total Expenditure") + xlab("Total Expenditure") + ylab("Life Expectancy")
+  theme(legend.position = "none")

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



This plot shows the effect of a country's total expenditure on life expectancy. There appears to be a slightly positive relationship between the two variables with variance increasing near the ends of the data.

HIV/AIDS

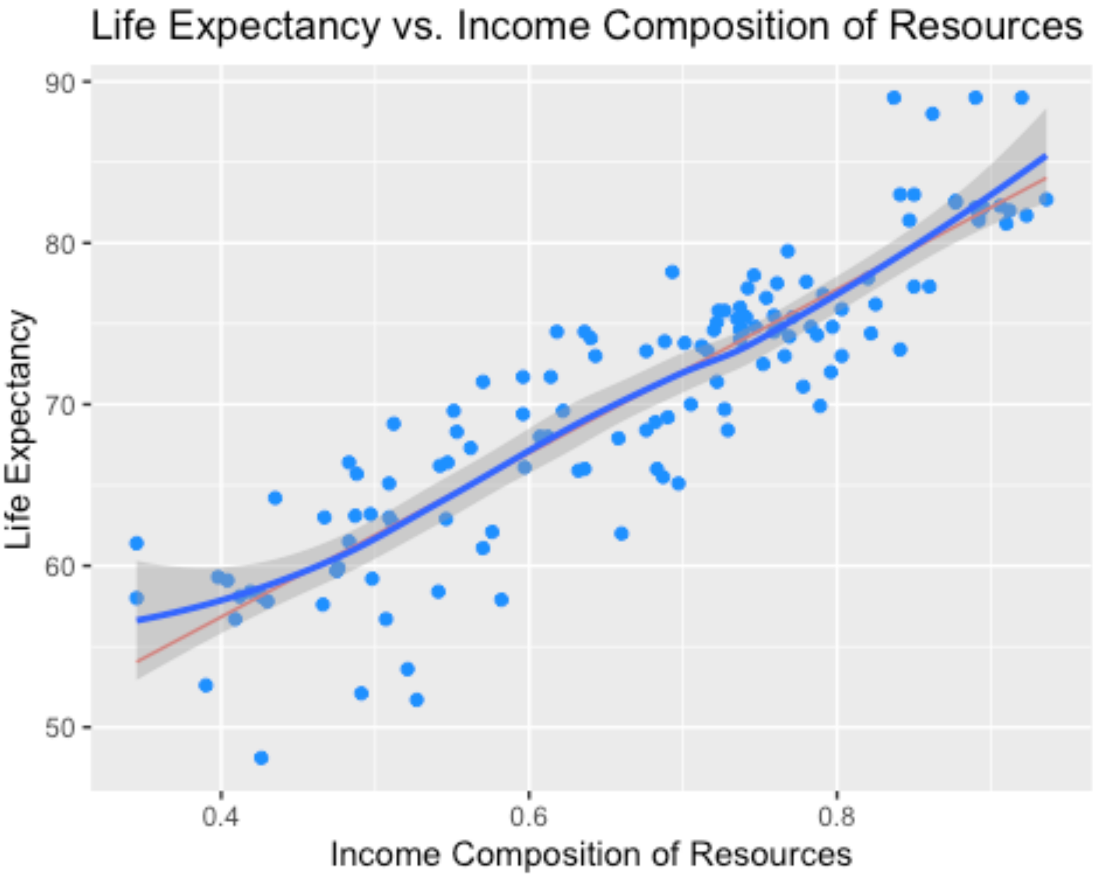
```
HIVData = as.data.frame(cbind(life_expectancy$HIV.AIDS,
life_expectancy$Life.expectancy))
HIVData = HIVData[complete.cases(HIVData),]
colnames(HIVData) = c("HIV/AIDS", "Life Expectancy")
pred.HIV <- predict(lm(life_expectancy$Life.expectancy ~
life_expectancy$HIV.AIDS, data = HIVData))
HIVPlot <- ggplot(HIVData, aes(x = life_expectancy$HIV.AIDS, y =
life_expectancy$Life.expectancy))
HIVPlot + geom_point(col = "dodgerblue", ) + geom_line(aes(y = pred.HIV, col
= "darkorange")) + geom_smooth() + ggtitle("Life Expectancy vs. HIV/AIDS
Prevalence") + xlab("HIV/AIDS") + ylab("Life Expectancy") +
theme(legend.position = "none")

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



This plot shows the effect of the prevalence of HIV/AIDS within a country on life expectancy. There appears to be a negative relationship between the two variables with variance increasing as HIV/AIDS increases.

Income Composition of Resources

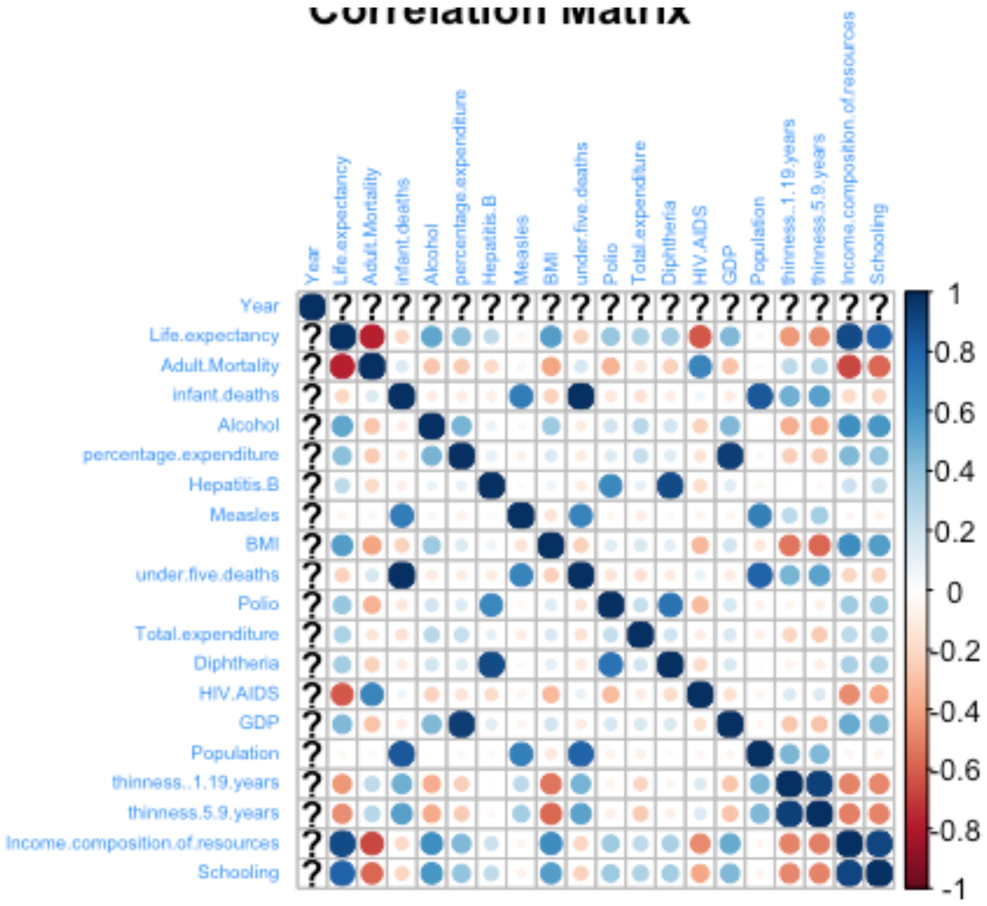
```
incomeData =
as.data.frame(cbind(life_expectancy$Income.composition.of.resources,
life_expectancy$Life.expectancy))
incomeData = incomeData[complete.cases(incomeData),]
colnames(incomeData) = c("Income Comp", "Life Expectancy")
pred.income <- predict(lm(life_expectancy$Life.expectancy ~
life_expectancy$Income.composition.of.resources, data = incomeData))

incomePlot <- ggplot(incomeData, aes(x =
life_expectancy$Income.composition.of.resources, y =
life_expectancy$Life.expectancy))
incomePlot + geom_point(col = "dodgerblue", ) + geom_line(aes(y =
pred.income, col = "darkorange")) + geom_smooth() + ggtitle("Life Expectancy
vs. Income Composition of Resources") + xlab("Income Composition of
Resources") + ylab("Life Expectancy") +  theme(legend.position = "none")

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



This plot shows the relationship between income composition of resources and life expectancy. There is clearly a very positive relationship with variance remaining constant.

Now we can look at the correlation matrix and see if we have any evidence of multicollinearity.

```
corrplot(cor(life_expectancy[sapply(life_expectancy, is.numeric)]), title =
"Correlation Matrix", tl.cex = 0.5, tl.pos = "lt", tl.col = "dodgerblue")

## Warning in cor(life_expectancy[sapply(life_expectancy, is.numeric)]): the
## standard deviation is zero
```



Model Selection

1. Full Model

```
full_Model = lm(Life.expectancy~. - Country, data = life_expectancy)
summary(full_Model)

##
## Call:
## lm(formula = Life.expectancy ~ . - Country, data = life_expectancy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4098  -1.7264  -0.0392   1.7715   8.3880
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.122e+01  3.314e+00  15.457  < 2e-16 ***
## Year                                    NA         NA      NA       NA
## StatusDeveloping                -1.170e+00  1.035e+00  -1.130 0.261006
## Adult.Mortality                 -1.724e-02  4.148e-03  -4.157 6.36e-05 ***
## infant.deaths                    8.287e-02  5.619e-02   1.475 0.143057
## Alcohol                          5.674e-03  9.749e-02   0.058 0.953689
## percentage.expenditure           4.627e-04  4.639e-04   0.997 0.320716
## Hepatitis.B                      1.205e-02  2.808e-02   0.429 0.668582    
## Measles                         -3.361e-05  4.823e-05  -0.697 0.487345    
## BMI                             -7.576e-03  2.000e-02  -0.379 0.705531    
## under.five.deaths               -6.014e-02  3.838e-02  -1.567 0.119989    
## Polio                           -8.746e-03  2.117e-02  -0.413 0.680327    
## Total.expenditure                2.878e-01  1.274e-01   2.259 0.025833 *  
## Diphtheria                       7.644e-03  3.445e-02   0.222 0.824805    
## HIV.AIDS                        -8.363e-01  2.470e-01  -3.385 0.000984 ***
## GDP                             -5.980e-05  6.656e-05  -0.898 0.370911    
## Population                      -1.729e-09  6.804e-09  -0.254 0.799816    
## thinness..1.19.years            -1.300e-01  2.267e-01  -0.574 0.567462    
## thinness.5.9.years               5.458e-03  2.227e-01   0.025 0.980489    
## Income.composition.of.resources  3.597e+01  6.228e+00   5.775 7.11e-08 ***
## Schooling                       -1.617e-01  2.740e-01  -0.590 0.556279    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.186 on 111 degrees of freedom
## Multiple R-squared:  0.8829, Adjusted R-squared:  0.8629 
## F-statistic: 44.06 on 19 and 111 DF,  p-value: < 2.2e-16
```

2. Reduced additive model with significant predictors

```
reduced_model = lm(Life.expectancy ~ Adult.Mortality + Total.expenditure + 
HIV.AIDS + Income.composition.of.resources, life_expectancy)
summary(reduced_model)
```

```
## 
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Total.expenditure + 
##     HIV.AIDS + Income.composition.of.resources, data = life_expectancy)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.3808  -1.6174  -0.0501   1.6143   9.9760 
## 
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     47.614113   2.047692  23.253  < 2e-16 ***
## Adult.Mortality                 -0.017949   0.003855  -4.656 8.04e-06 ***
## Total.expenditure                0.355162   0.111458   3.187 0.001816 ** 
## HIV.AIDS                        -0.844894   0.230049  -3.673 0.000353 ***
## Income.composition.of.resources 36.285086   2.488491  14.581  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.102 on 126 degrees of freedom
## Multiple R-squared:  0.8741, Adjusted R-squared:  0.8701 
## F-statistic: 218.7 on 4 and 126 DF,  p-value: < 2.2e-16
```

3. Full Interaction Model

```
full_interaction_model = lm(Life.expectancy ~
Adult.Mortality*Total.expenditure*HIV.AIDS*Income.composition.of.resources,
life_expectancy)
summary(full_interaction_model)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality * Total.expenditure *
##     HIV.AIDS * Income.composition.of.resources, data = life_expectancy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4637 -1.5656 -0.1007  1.5709 10.2628
##
## Coefficients:
##
Estimate
## (Intercept)
40.656947
## Adult.Mortality
0.041776
## Total.expenditure
2.050656
## HIV.AIDS
-5.193188
## Income.composition.of.resources
44.491764
## Adult.Mortality:Total.expenditure
-0.011089
## Adult.Mortality:HIV.AIDS
0.012179
## Total.expenditure:HIV.AIDS
-0.193318
## Adult.Mortality:Income.composition.of.resources
-0.076117
## Total.expenditure:Income.composition.of.resources
-1.911898
## HIV.AIDS:Income.composition.of.resources
3.847932
## Adult.Mortality:Total.expenditure:HIV.AIDS
0.001557
## Adult.Mortality:Total.expenditure:Income.composition.of.resources
0.013078
## Adult.Mortality:HIV.AIDS:Income.composition.of.resources
-0.018158
## Total.expenditure:HIV.AIDS:Income.composition.of.resources
0.559848
## Adult.Mortality:Total.expenditure:HIV.AIDS:Income.composition.of.resources
-0.002375
##
```

```
Std. Error
## (Intercept)
6.634816
## Adult.Mortality
0.037920
## Total.expenditure
1.018573
## HIV.AIDS
16.103344
## Income.composition.of.resources
9.396611
## Adult.Mortality:Total.expenditure
0.005058
## Adult.Mortality:HIV.AIDS
0.043469
## Total.expenditure:HIV.AIDS
3.074968
## Adult.Mortality:Income.composition.of.resources
0.068598
## Total.expenditure:Income.composition.of.resources
1.380268
## HIV.AIDS:Income.composition.of.resources
29.791079
## Adult.Mortality:Total.expenditure:HIV.AIDS
0.008619
## Adult.Mortality:Total.expenditure:Income.composition.of.resources
0.009690
## Adult.Mortality:HIV.AIDS:Income.composition.of.resources
0.083418
## Total.expenditure:HIV.AIDS:Income.composition.of.resources
5.889307
## Adult.Mortality:Total.expenditure:HIV.AIDS:Income.composition.of.resources
0.016781
##
t value
## (Intercept)
6.128
## Adult.Mortality
1.102
## Total.expenditure
2.013
## HIV.AIDS
-0.322
## Income.composition.of.resources
4.735
## Adult.Mortality:Total.expenditure
-2.192
## Adult.Mortality:HIV.AIDS
0.280
## Total.expenditure:HIV.AIDS
```

```
-0.063
## Adult.Mortality:Income.composition.of.resources
-1.110
## Total.expenditure:Income.composition.of.resources
-1.385
## HIV.AIDS:Income.composition.of.resources
0.129
## Adult.Mortality:Total.expenditure:HIV.AIDS
0.181
## Adult.Mortality:Total.expenditure:Income.composition.of.resources
1.350
## Adult.Mortality:HIV.AIDS:Income.composition.of.resources
-0.218
## Total.expenditure:HIV.AIDS:Income.composition.of.resources
0.095
## Adult.Mortality:Total.expenditure:HIV.AIDS:Income.composition.of.resources
-0.142
##
Pr(>|t|)
## (Intercept)
1.28e-08
## Adult.Mortality
0.2729
## Total.expenditure
0.0464
## HIV.AIDS
0.7477
## Income.composition.of.resources
6.30e-06
## Adult.Mortality:Total.expenditure
0.0304
## Adult.Mortality:HIV.AIDS
0.7798
## Total.expenditure:HIV.AIDS
0.9500
## Adult.Mortality:Income.composition.of.resources
0.2695
## Total.expenditure:Income.composition.of.resources
0.1687
## HIV.AIDS:Income.composition.of.resources
0.8975
## Adult.Mortality:Total.expenditure:HIV.AIDS
0.8570
## Adult.Mortality:Total.expenditure:Income.composition.of.resources
0.1798
## Adult.Mortality:HIV.AIDS:Income.composition.of.resources
0.8281
## Total.expenditure:HIV.AIDS:Income.composition.of.resources
0.9244
## Adult.Mortality:Total.expenditure:HIV.AIDS:Income.composition.of.resources
```

```
0.8877
##
## (Intercept)
***
## Adult.Mortality
## Total.expenditure
*
## HIV.AIDS
## Income.composition.of.resources
***
## Adult.Mortality:Total.expenditure
*
## Adult.Mortality:HIV.AIDS
## Total.expenditure:HIV.AIDS
## Adult.Mortality:Income.composition.of.resources
## Total.expenditure:Income.composition.of.resources
## HIV.AIDS:Income.composition.of.resources
## Adult.Mortality:Total.expenditure:HIV.AIDS
## Adult.Mortality:Total.expenditure:Income.composition.of.resources
## Adult.Mortality:HIV.AIDS:Income.composition.of.resources
## Total.expenditure:HIV.AIDS:Income.composition.of.resources
## Adult.Mortality:Total.expenditure:HIV.AIDS:Income.composition.of.resources
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.834 on 115 degrees of freedom
## Multiple R-squared:  0.904,  Adjusted R-squared:  0.8915
## F-statistic: 72.23 on 15 and 115 DF,  p-value: < 2.2e-16
```

4. Reduced Interaction Model

```
#Based on significance we can drop all interactions except
Adult.Mortality:Total.expenditure

reduced_interaction_model = lm(Life.expectancy ~ Total.expenditure + HIV.AIDS
+ Income.composition.of.resources + Adult.Mortality:Total.expenditure,
life_expectancy)

#We can also drop Adult.Mortality based on its insignificant t value

summary(reduced_interaction_model)

##
## Call:
## lm(formula = Life.expectancy ~ Total.expenditure + HIV.AIDS +
##     Income.composition.of.resources + Adult.Mortality:Total.expenditure,
##     data = life_expectancy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7271 -1.8870 -0.1355  1.7972 11.3208
```

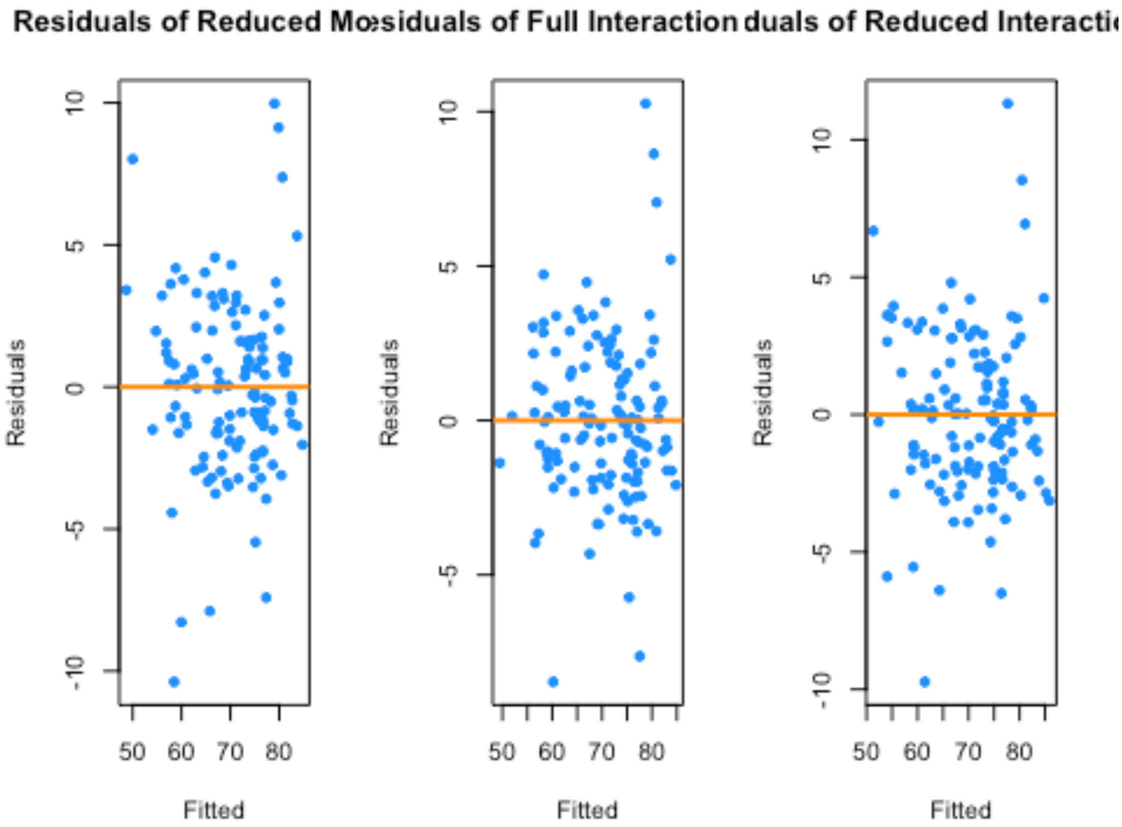
```
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       45.2043035  1.6211321  27.884  < 2e-16
***
## Total.expenditure                  0.8095382  0.1359027   5.957 2.40e-08
***
## HIV.AIDS                          -0.9822090  0.2041526  -4.811 4.21e-06
***
## Income.composition.of.resources   35.5319886  2.3685925  15.001  < 2e-16
***
## Total.expenditure:Adult.Mortality -0.0027935  0.0004898  -5.704 7.92e-08
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.994 on 126 degrees of freedom
## Multiple R-squared:  0.8827, Adjusted R-squared:  0.879
## F-statistic:   237 on 4 and 126 DF,  p-value: < 2.2e-16
```

Model Assumptions

Now, between the reduced_model and the full_interaction_model

1. Linearity and Constant Variance

```r
par(mfrow=c(1,3))
plot_fitted_resid(reduced_model, title = "Residuals of Reduced Model")
plot_fitted_resid(full_interaction_model, title = "Residuals of Full
Interaction Model")
plot_fitted_resid(reduced_interaction_model, title = "Residuals of Reduced
Interaction Model")
```

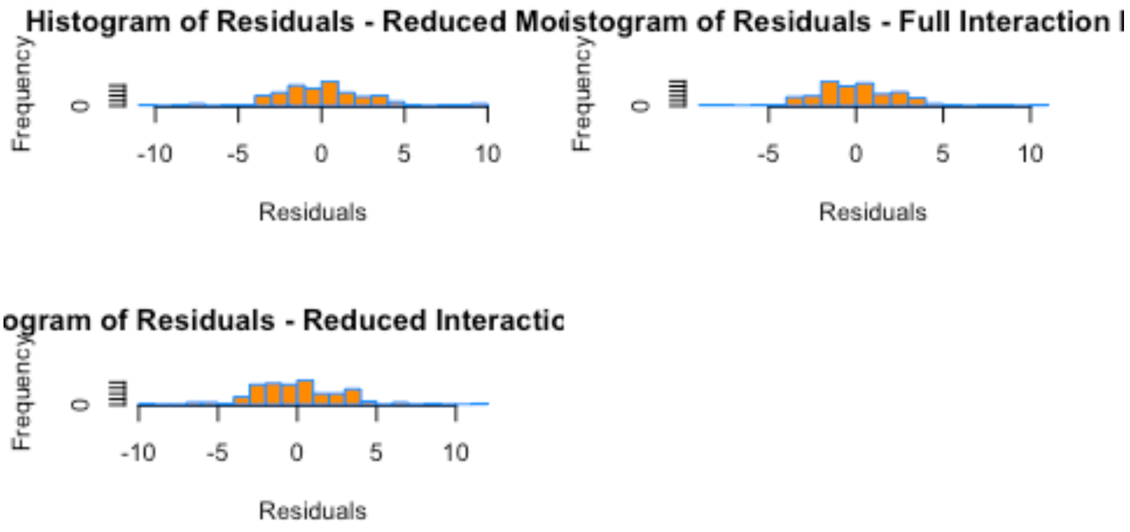


```
bptest(reduced_model)

##
##  studentized Breusch-Pagan test
##
## data:  reduced_model
## BP = 1.7498, df = 4, p-value = 0.7817

bptest(full_interaction_model)

##
##  studentized Breusch-Pagan test
##
## data:  full_interaction_model
## BP = 13.363, df = 15, p-value = 0.5743

bptest(reduced_interaction_model)

##
##  studentized Breusch-Pagan test
##
## data:  reduced_interaction_model
## BP = 5.2247, df = 4, p-value = 0.265
```

We can observe that the residuals are closely related to 0 with some outliers, however the Breusch-Pagan test confirms our residuals are homoskedastic in nature. The best model, is the reduced interaction model in terms of homoskedasticity.

2. Normality of Errors

```
par(mfrow=c(3,2))
hist(resid(reduced_model), xlab = "Residuals", main = "Histogram of Residuals
- Reduced Model", col = "darkorange", border = "dodgerblue", breaks = 20)


hist(resid(full_interaction_model), xlab = "Residuals", main = "Histogram of
Residuals - Full Interaction Model", col = "darkorange", border =
"dodgerblue", breaks = 20)


hist(resid(reduced_interaction_model), xlab = "Residuals", main = "Histogram
of Residuals - Reduced Interaction Model", col = "darkorange", border =
"dodgerblue", breaks = 20)

shapiro.test(resid(reduced_model))

##
##  Shapiro-Wilk normality test
##
## data:  resid(reduced_model)
## W = 0.96648, p-value = 0.002531

shapiro.test(resid(full_interaction_model))

##
##  Shapiro-Wilk normality test
##
## data:  resid(full_interaction_model)
## W = 0.96229, p-value = 0.001075

shapiro.test(resid(reduced_interaction_model))

##
##  Shapiro-Wilk normality test
##
## data:  resid(reduced_interaction_model)
## W = 0.97224, p-value = 0.008679
```



3. Outliers/Influential Points

```
#Outliers
length(rstandard(reduced_model)[abs(rstandard(reduced_model)) > 2])

## [1] 8

length(rstandard(full_interaction_model)[abs(rstandard(full_interaction_model
)) > 2])

## [1] 7

length(rstandard(reduced_interaction_model)[abs(rstandard(reduced_interaction
_model)) > 2])

## [1] 8

#Influential Points
length(cooks.distance(reduced_model)[cooks.distance(reduced_model) > 4 /
length(cooks.distance(reduced_model))])

## [1] 9

length(cooks.distance(full_interaction_model)[cooks.distance(full_interaction
_model) > 4 / length(cooks.distance(full_interaction_model))])
```

```
## [1] 11
```

```r
length(cooks.distance(reduced_interaction_model)[cooks.distance(reduced_inter
action_model) > 4 / length(cooks.distance(reduced_interaction_model))])
```

```
## [1] 10
```

As expected, there were outliers and influential points present. This is very common in large, aggregated data sets. Fortunately, in this case the points are not too influential on the results of our regression.

Model Evaluations

```r
summary(reduced_model)$adj.r.squared
```

```
## [1] 0.8700815
```

```r
summary(full_interaction_model)$adj.r.squared
```

```
## [1] 0.891531
```

```r
summary(reduced_interaction_model)$adj.r.squared
```

```
## [1] 0.8789724
```

```r
loocv_rmse(reduced_model)
```

```
## [1] 0.01557687
```

```r
loocv_rmse(full_interaction_model)
```

```
## [1] 0.05187392
```

```r
loocv_rmse(reduced_interaction_model)
```

```
## [1] 0.0114956
```

```r
# AIC
extractAIC(reduced_model)
```

```
## [1]   5.0000 301.4718
```

```r
extractAIC(full_interaction_model)
```

```
## [1]  16.000 287.867
```

```r
extractAIC(reduced_interaction_model)
```

```
## [1]   5.0000 292.1854
```

```r
#BIC
extractAIC(reduced_model, k = log(nrow(life_expectancy)))
```

```
## [1]   5.0000 315.8478
```

```r
extractAIC(full_interaction_model, k = log(nrow(life_expectancy)))
```

```
## [1]  16.0000 333.8701
```

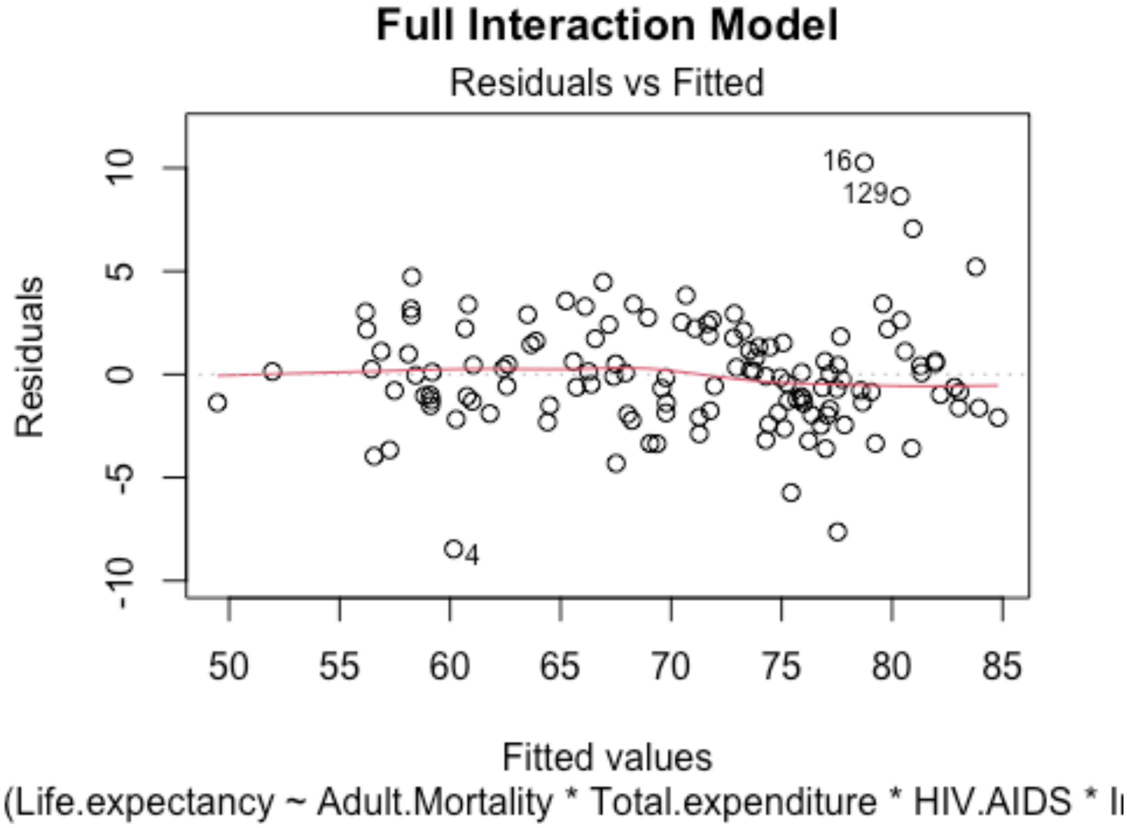
```r
extractAIC(reduced_interaction_model, k = log(nrow(life_expectancy)))
```

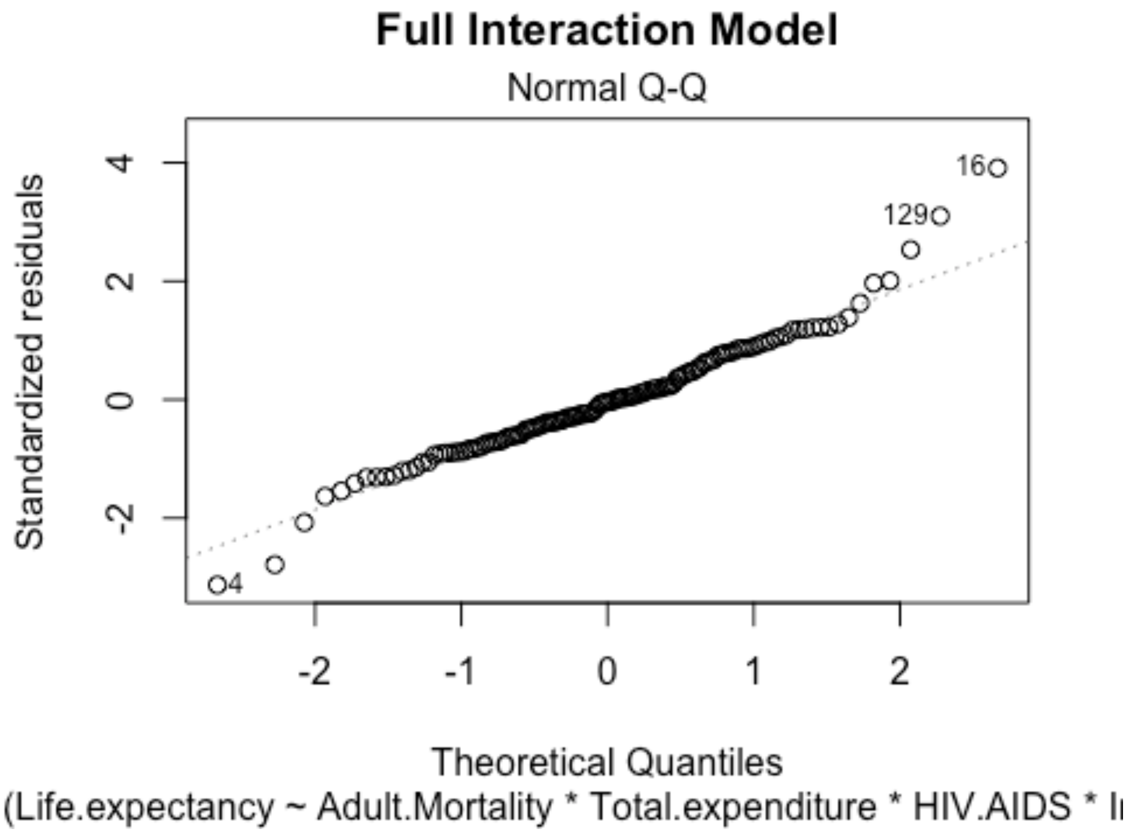
```
## [1]   5.0000 306.5614
```

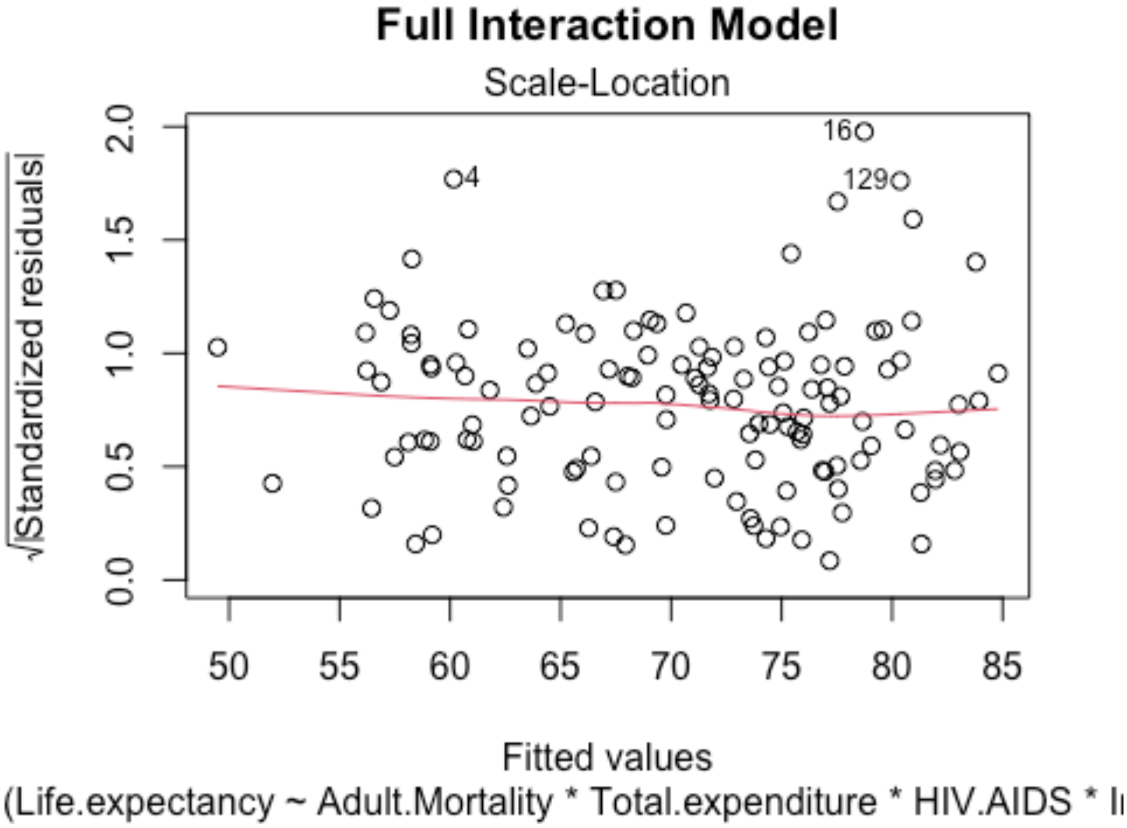
Results

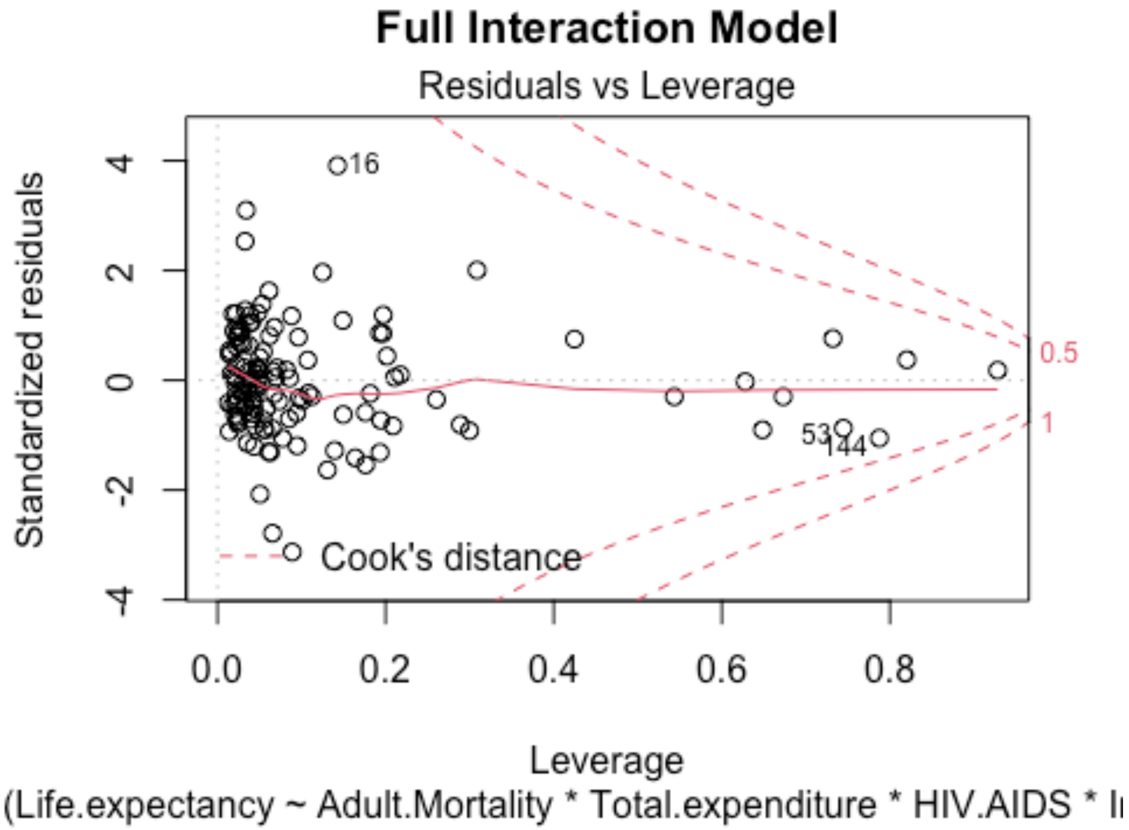
We ended up choosing the 'Full Interaction Model' as our best model, mainly because of the values we received for the Adjusted R Squared. Since we know it passed the assumptions, we can take a look at how it performed individually.

```r
plot(full_interaction_model, main = "Full Interaction Model")
```


Full Interaction Model
Residuals vs Fitted
Residuals
10
5
0
-5
-10
16
129
4
50
55
60
65
70
75
80
85
Fitted values
(Life.expectancy ~ Adult.Mortality * Total.expenditure * HIV.AIDS * I


Full Interaction Model
Normal Q-Q
Standardized residuals
4
2
0
-2
-2
-1
0
1
2
16
129
4
Theoretical Quantiles
(Life.expectancy ~ Adult.Mortality * Total.expenditure * HIV.AIDS * I


Full Interaction Model
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
16
129
4
50
55
60
65
70
75
80
85
Fitted values
(Life.expectancy ~ Adult.Mortality * Total.expenditure * HIV.AIDS * Ir



```
#Analysis
summary(full_interaction_model)$adj.r.squared

## [1] 0.891531

loocv_rmse(full_interaction_model)

## [1] 0.05187392

#VIF
sum(vif(full_interaction_model) > 5) / length(coef(full_interaction_model))

## [1] 0.9375

#AIC
extractAIC(full_interaction_model)

## [1]  16.000 287.867

#BIC
extractAIC(full_interaction_model, k = log(nrow(life_expectancy)))

## [1]  16.0000 333.8701
```

This model seems to outperform all the other models generated and examined previously.

Discussion

Predictor Significance

```
sum(summary(full_interaction_model)$coefficients[ ,4] < 0.05) /
length(coef(full_interaction_model))

## [1] 0.25

summary(full_interaction_model)$coefficients[summary(full_interaction_model)$
coefficients[ ,4] < 0.05,]

##                                      Estimate  Std. Error   t value
## (Intercept)                       40.65694678 6.634815537  6.127819
## Total.expenditure                  2.05065563 1.018573290  2.013263
## Income.composition.of.resources   44.49176390 9.396610691  4.734874
## Adult.Mortality:Total.expenditure -0.01108882 0.005058191 -2.192250
##                                       Pr(>|t|)
## (Intercept)                       1.283004e-08
## Total.expenditure                 4.642238e-02
## Income.composition.of.resources   6.301794e-06
## Adult.Mortality:Total.expenditure 3.037589e-02
```

Significant R^2 and F − statistic The final model that we chose *full_interaction_model* achieved an R^2 value of .904 which means that we were able to explain 90.4% of the variation in our depedent variable, life expectancy. This model also had a statistically significant F-stat of 72.2 on 15 and 115 degrees of freedom which suggests that the model is valid.

anova test

```
anovaTest = anova(full_interaction_model, reduced_interaction_model, test =
"F")
anovaTest$`Pr(>F)`[2]

## [1] 0.01279788
```

We tested the full interaction model against the reduced interaction model using an anova test and found that the p-value was .0128. This suggests we should reject the null at the 5% level and conclude that the full interaction model should be used at the 5% significance level. We also tested the full interaction model against the additive reduced model and found that there was little evidence to suggest that the terms in the additive model were significant enough.

```
anovaTest2 = anova(full_interaction_model, reduced_model, test = "F")
anovaTest2$`Pr(>F)`[2]

## [1] 0.0006479437
```

Testing Assumptions The first assumption that we tested was constant variance. We created residual plots and ran Breusch Pagan tests for the reduced model, full interaction model and reduced interaction model. The residual plots seemed to indicate a few outliers, but the BP test confirmed that each model's residuals were homoskedastic.

The next assumption that we tested was normality of errors. We created a histogram for the reduced model, full interaction model and reduced interaction model. The histograms weren't perfect, but the distributions appeared to be relatively normal. We then conducted a Shapiro-Wilk test on each of the models which confirmed that all 3 had normally distributed errors.

Appendix

After omitting empty entries, we were left with still enough data to generate a report on. However, this resulted in not being able to use the USA as a data point. Unfortunately, factors such as GDP were left blank so we had to choose between having the USA as a data point or GDP and we figured GDP would be a better metric than one more observation. In the end, GDP ended up not being significant enough to make it past the first round of tests to determine the models.