605-HW12-Regression-WorldHealth

Michael Y.

November 17, 2019

Contents

HW11 - Regression2 - World Health

WHO datset

Load the dataset

```r
setwd(mydir)
who.df <- read.csv('who.csv')
attach(who.df)
```

The attached who.csv dataset contains real-world data from 2008. The variables included follow:

VariableName	Description
Country	name of the country
LifeExp	average life expectancy for the country in years
InfantSurvival	proportion of those surviving to one year or more
Under5Survival	proportion of those surviving to five years or more
TBFree	proportion of the population without TB.
PropMD	proportion of the population who are MDs
PropRN	proportion of the population who are RNs
PersExp	mean personal expenditures on healthcare in US dollars at average exchange rate
GovtExp	mean government expenditures per capita on healthcare, US dollars at average exchange rate
TotExp	sum of personal and government expenditures.

EDA

```r
# Dimension of the dataset
dim(who.df)
```

```
## [1] 190  10
```

```r
# structure of the dataset
str(who.df)
```

```
## 'data.frame':    190 obs. of  10 variables:
##  $ Country       : Factor w/ 190 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ LifeExp       : int  42 71 71 82 41 73 75 69 82 80 ...
##  $ InfantSurvival: num  0.835 0.985 0.967 0.997 0.846 0.99 0.986 0.979 0.995 0.996 ...
##  $ Under5Survival: num  0.743 0.983 0.962 0.996 0.74 0.989 0.983 0.976 0.994 0.996 ...
##  $ TBFree        : num  0.998 1 0.999 1 0.997 ...
##  $ PropMD        : num  0.0002288 0.0011431 0.0010605 0.0032973 0.0000704 ...
##  $ PropRN        : num  0.000572 0.004614 0.002091 0.0035 0.001146 ...
##  $ PersExp       : int  20 169 108 2589 36 503 484 88 3181 3788 ...
##  $ GovtExp       : int  92 3128 5184 169725 1620 12543 19170 1856 187616 189354 ...
##  $ TotExp        : int  112 3297 5292 172314 1656 13046 19654 1944 190797 193142 ...
```

```r
# summary of the dataset
summary(who.df)
```

```
##                Country       LifeExp        InfantSurvival     Under5Survival        TBFree             PropMD                PropRN
##  Afghanistan        :  1   Min.   :40.0000   Min.   :0.835000   Min.   :0.731000   Min.   :0.987000   Min.   :0.000019600   Min.   :0.0
##  Albania            :  1   1st Qu.:61.2500   1st Qu.:0.943250   1st Qu.:0.925250   1st Qu.:0.996905   1st Qu.:0.000244355   1st Qu.:0.0
##  Algeria            :  1   Median :70.0000   Median :0.978500   Median :0.974500   Median :0.999215   Median :0.001047359   Median :0.0
##  Andorra            :  1   Mean   :67.3789   Mean   :0.962447   Mean   :0.945942   Mean   :0.998038   Mean   :0.001795380   Mean   :0.0
##  Angola             :  1   3rd Qu.:75.0000   3rd Qu.:0.991000   3rd Qu.:0.990000   3rd Qu.:0.999760   3rd Qu.:0.002458363   3rd Qu.:0.0
##  Antigua and Barbuda:  1   Max.   :83.0000   Max.   :0.998000   Max.   :0.997000   Max.   :0.999980   Max.   :0.035129032   Max.   :0.0
##  (Other)            :184
##     PersExp           GovtExp             TotExp
##  Min.   :   3.00   Min.   :    10.0   Min.   :    13.0
##  1st Qu.:  36.25   1st Qu.:   559.5   1st Qu.:   584.0
##  Median : 199.50   Median :  5385.0   Median :  5541.0
##  Mean   : 742.00   Mean   : 40953.5   Mean   : 41695.5
##  3rd Qu.: 515.25   3rd Qu.: 25680.2   3rd Qu.: 26331.0
##  Max.   :6350.00   Max.   :476420.0   Max.   :482750.0
##
```

The dataset contains 190 observations of 10 variables (where each country is an obseervation.)

1. Provide a scatterplot of LifeExp~TotExp, and run simple linear regression.

Do not transform the variables.

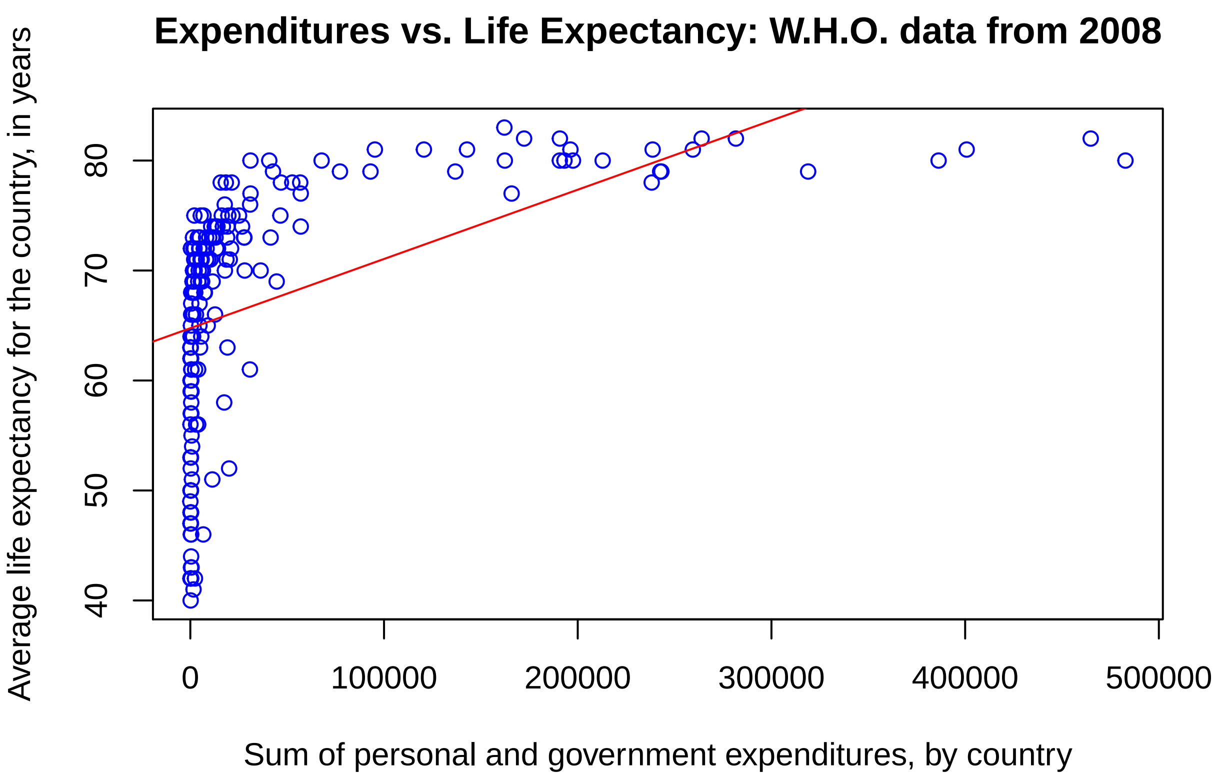
Linear Regression

```r
# Simple linear regression model
Model1 <- lm(LifeExp ~ TotExp, data=who.df)
Summ1 <- summary(Model1)
Fstat1 <- Summ1$fstatistic[1]
Tstat1_0 <- Summ1$coefficients["(Intercept)","t value"]
Tstat1_1 <- Summ1$coefficients["TotExp","t value"]
Pval1_0 <- Summ1$coefficients["(Intercept)","Pr(>|t|)"]
Pval1_1 <- Summ1$coefficients["TotExp","Pr(>|t|)"]
Rsq1 <- Summ1$r.squared
AdjRsq1 <- Summ1$adj.r.squared
Correlation1 <- cor(who.df$LifeExp, who.df$TotExp)
print(Summ1)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = who.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -24.76421  -4.77817   3.15437   7.11620  13.29178
##
## Coefficients:
##                   Estimate     Std. Error  t value               Pr(>|t|)
## (Intercept) 64.75337453357  0.75353661143 85.93262 < 0.0000000000000000222 ***
## TotExp       0.00006297019  0.00000779467  8.07863    0.00000000000007714 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.37103 on 188 degrees of freedom
## Multiple R-squared:  0.257692,   Adjusted R-squared:  0.253744
## F-statistic: 65.2642 on 1 and 188 DF,  p-value: 0.0000000000000771399
```

Make a scatterplot, with regression line

```r
plot(LifeExp~TotExp,
     xlab="Sum of personal and government expenditures, by country",
     ylab="Average life expectancy for the country, in years",
     main="Expenditures vs. Life Expectancy: W.H.O. data from 2008",
     col="blue")
abline(Model1,col="red")
```



Provide and interpret the F statistics, R^2, standard error,and p-values only.

The **F-statistic**, 65.26419817 is large, indicating significance. (The critical value for the F-test is 3.891398098).

Because there is only 1 degree of freedom in the numerator, the F-statistic equals the **square of the t-value**, 8.078626008, on the coefficient on the independent variable, `TotExp`.

The R^2 (0.25769216) and the **adjusted-R^2** (0.253743714) values are not very strong, indicating a poor fit. As they are about $\frac{1}{4}$, this indicates that the correlation between the independent and dependent variables is about $\frac{1}{2}$.

(We have confirmed that the actual correlation is 0.507633883 .)

The `Null Hypothesis` is H_0: The regression coefficients are ***not*** significantly different from zero,
indicating no relationship between the dependent and independent variables.

The `Alternative` is H_A: The coefficients ***are*** significantly different from zero.

The results indicate that the coefficients are significant, as the p-values are close to zero.

Regression Assumptions

Discuss whether the assumptions of simple linear regression met.

There are four assumptions of simple linear regression:

1. Linearity of the data. The relationship between the predictor (x) and the outcome (y) is assumed to be linear.

2. Normality of residuals. The residual errors are assumed to be normally distributed.

3. Homogeneity of residuals variance. The residuals are assumed to have a constant variance (homoscedasticity)

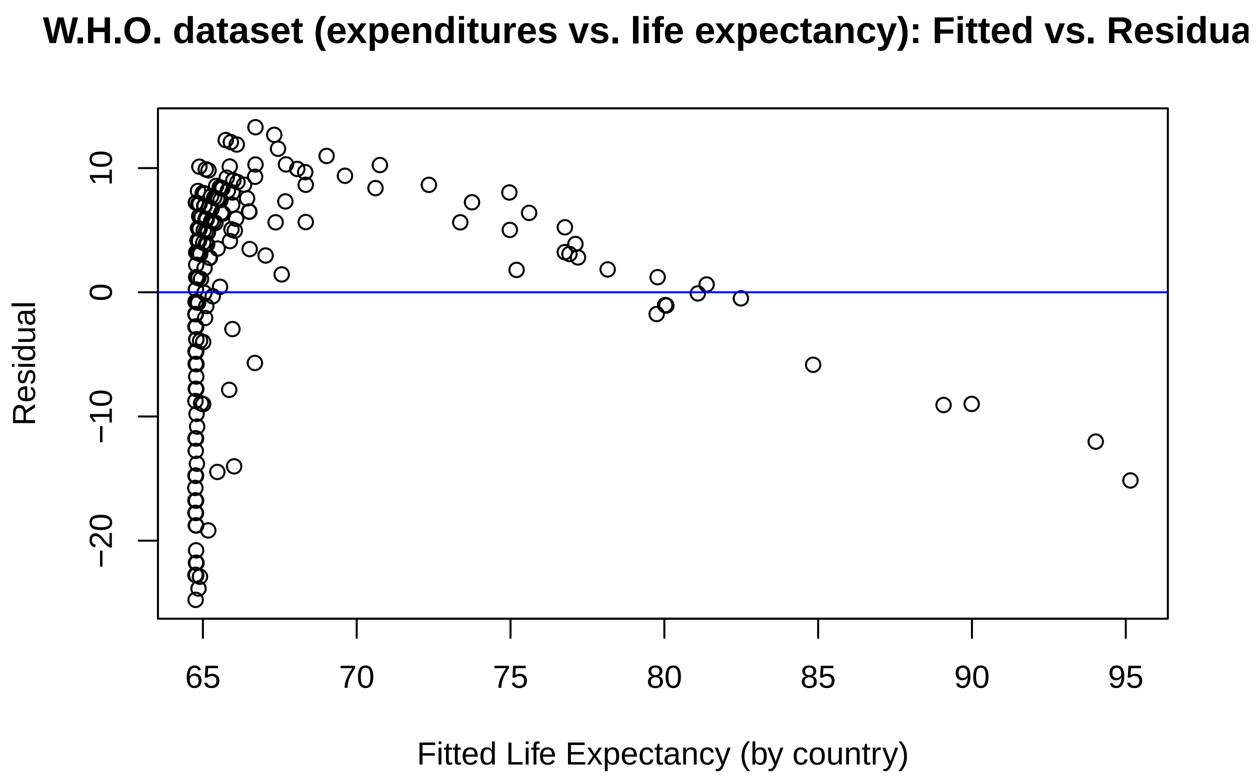
4. Independence of residuals error terms.

We will find that in this example, these assumptions are not met.

1. Linearity of the data

Clearly from the plot above, the data is not linear. We could resolve this through a transformation, such as taking logs, but have been asked not to. Additionally, the plot of the residuals has the following pattern:

```r
Residual = resid(Model1)
Fitted = fitted(Model1)
plot(Fitted,Residual,
     main="W.H.O. dataset (expenditures vs. life expectancy): Fitted vs. Residuals",
     xlab="Fitted Life Expectancy (by country)")
abline(h=0, col="blue")
```



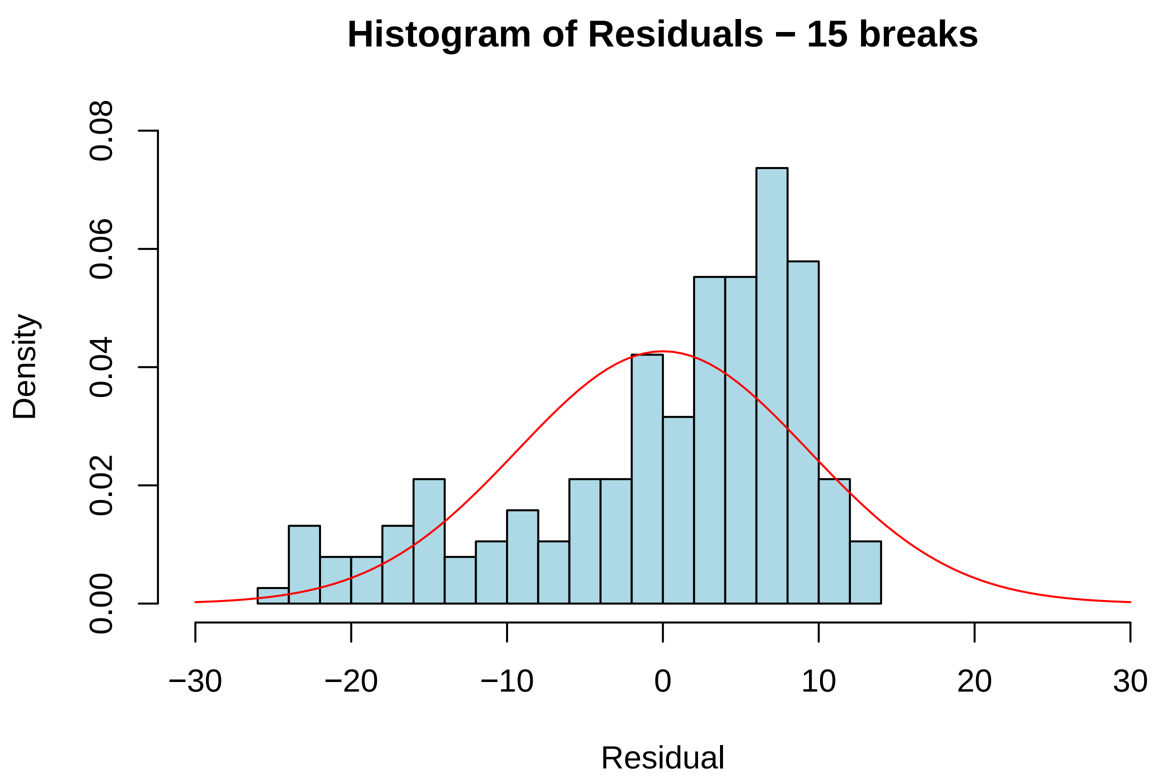
This does not look good. . .

2. Normality of residuals.

Histogram:

We plot a histogram of the residuals:

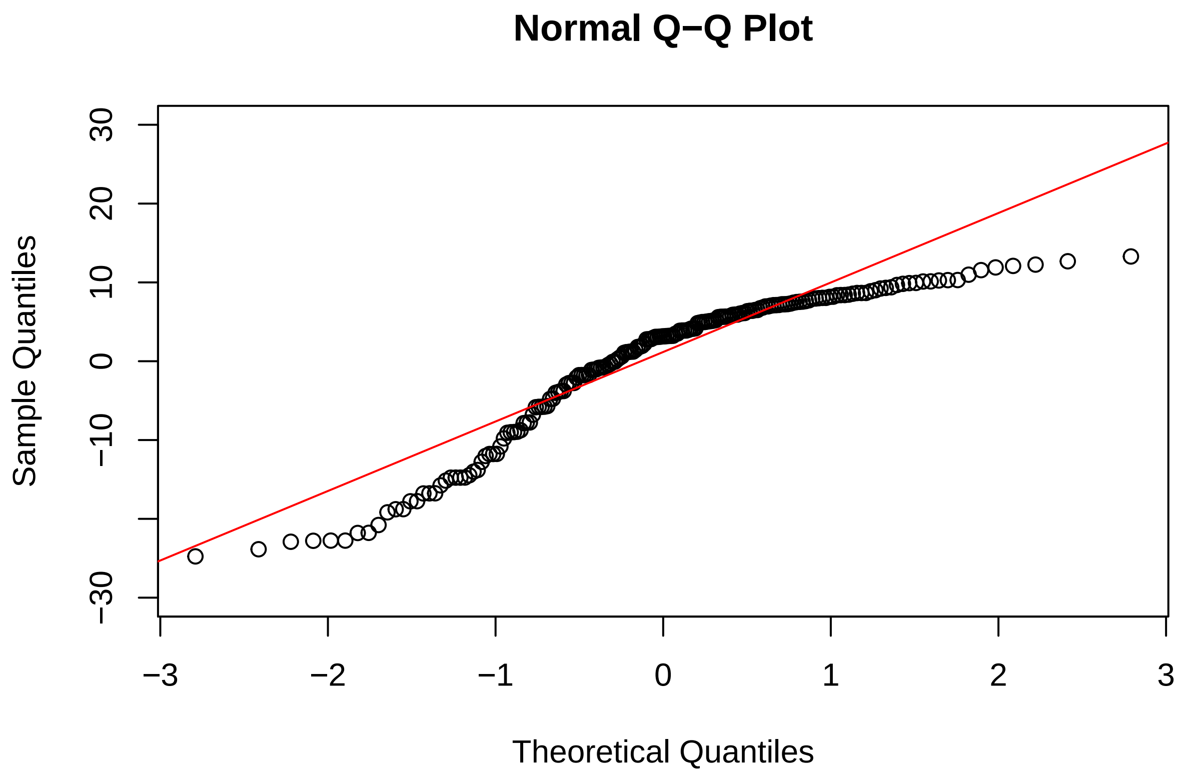
```
Residual = resid(Model1)
hist(Residual, main = "Histogram of Residuals - 15 breaks", ylab = "Density",
     ylim = c(0, 0.08),
     xlim = c(-30,30),
     prob = TRUE,breaks=15, col="lightblue")
curve(dnorm(x, mean = mean(Residual), sd = sd(Residual)), col="red", add=TRUE)
```



Clearly the above distribution is highly skewed and does not resemble the required normal distribution, which is superimposed in red.

QQ-Plot:

```
qqnorm(Residual, ylim=c(-30,30))
qqline(Residual, col="red")
```



Clearly the residuals fail the QQplot test.

We can run several standard tests of normality:

```
library(olsrr)
```

```
##
```

```
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
ols_test_normality(Model1)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.8915         0.0000
## Kolmogorov-Smirnov        0.1587         0.0001
## Cramer-von Mises         17.1443         0.0000
## Anderson-Darling          7.2968         0.0000
## -----------------------------------------------
```

Every test failed.

3. Homogeneity of residuals variance.

Clearly looking at the graphs above suggests that this test should fail.

We can test using the `lmSupport` package:

```r
library(lmSupport)
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```r
modelAssumptions(Model1,"LINEAR")
```

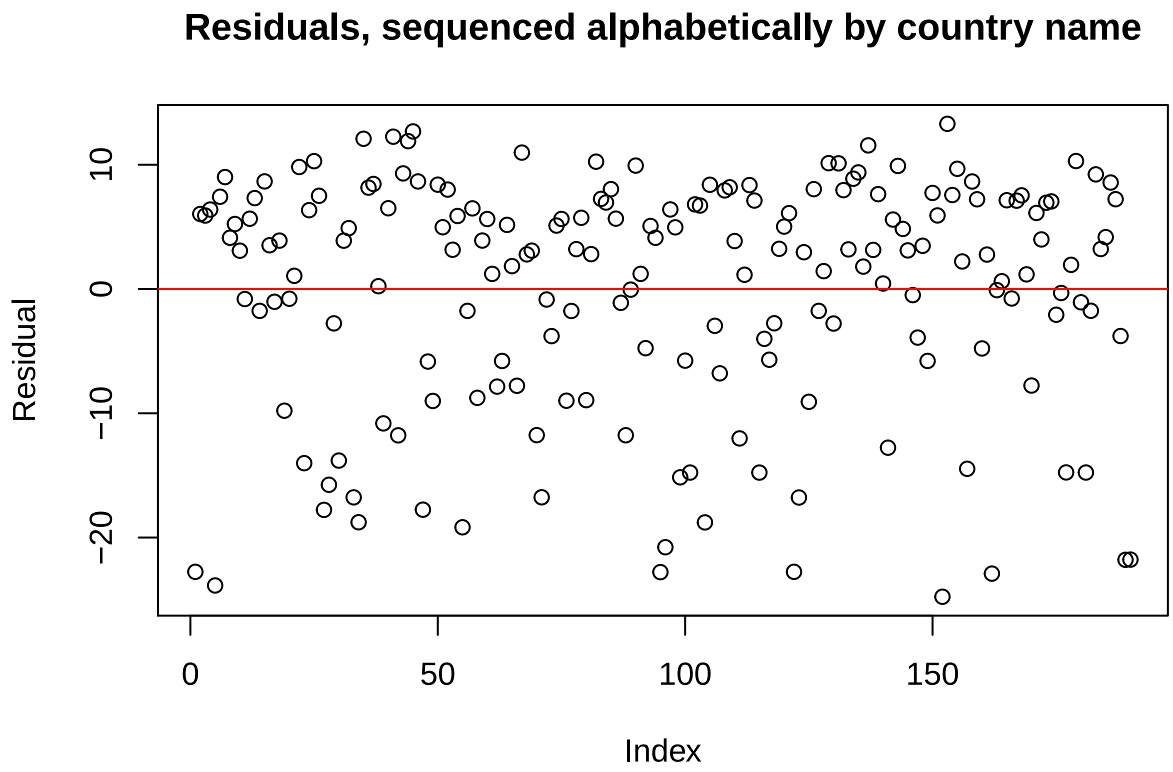
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = who.df)
##
## Coefficients:
##   (Intercept)         TotExp
## 64.7533745336   0.0000629702
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = Model)
##
##                           Value            p-value                   Decision
## Global Stat        56.73701065 0.0000000000140474 Assumptions NOT satisfied!
## Skewness           30.53275694 0.0000000328276598 Assumptions NOT satisfied!
## Kurtosis            0.00280358 0.9577726303075544    Assumptions acceptable.
## Link Function      26.07470333 0.0000003284593019 Assumptions NOT satisfied!
## Heteroscedasticity  0.12674679 0.7218292148467906    Assumptions acceptable.
```

While various tests do fail, according to the above diagnostic, the heteroscedasticity test does not fail.

4. Independence of residuals

While there is clearly a relationship between the independent variable vs. the residual, the residuals themselves are actually independent of each other, as each represents the result from an individual country. Plotting the residuals sequentially (where the sequence happens to be alphabetical by country name) doesn't yield any discernable pattern:

```r
plot(Residual, main="Residuals, sequenced alphabetically by country name")
abline(h=0, col="red")
```



Having failed several of the assumptions of Linear Regression, the simple linear regression model is ***NOT SUITABLE*** for this data.

---

2. Raise life expectancy to the 4.6 power (i.e., $LifeExp^{4.6}$).

Raise total expenditures to the 0.06 power (nearly a log transform, $TotExp^{.06}$).

```
who.df$xformLifeExp <- who.df$LifeExp^(4.6)
who.df$xformTotExp <- who.df$TotExp^(0.06)
```

Plot $LifeExp^{4.6}$ as a function of $TotExp^{.06}$, and r re-run the simple regression model using the transformed variables.

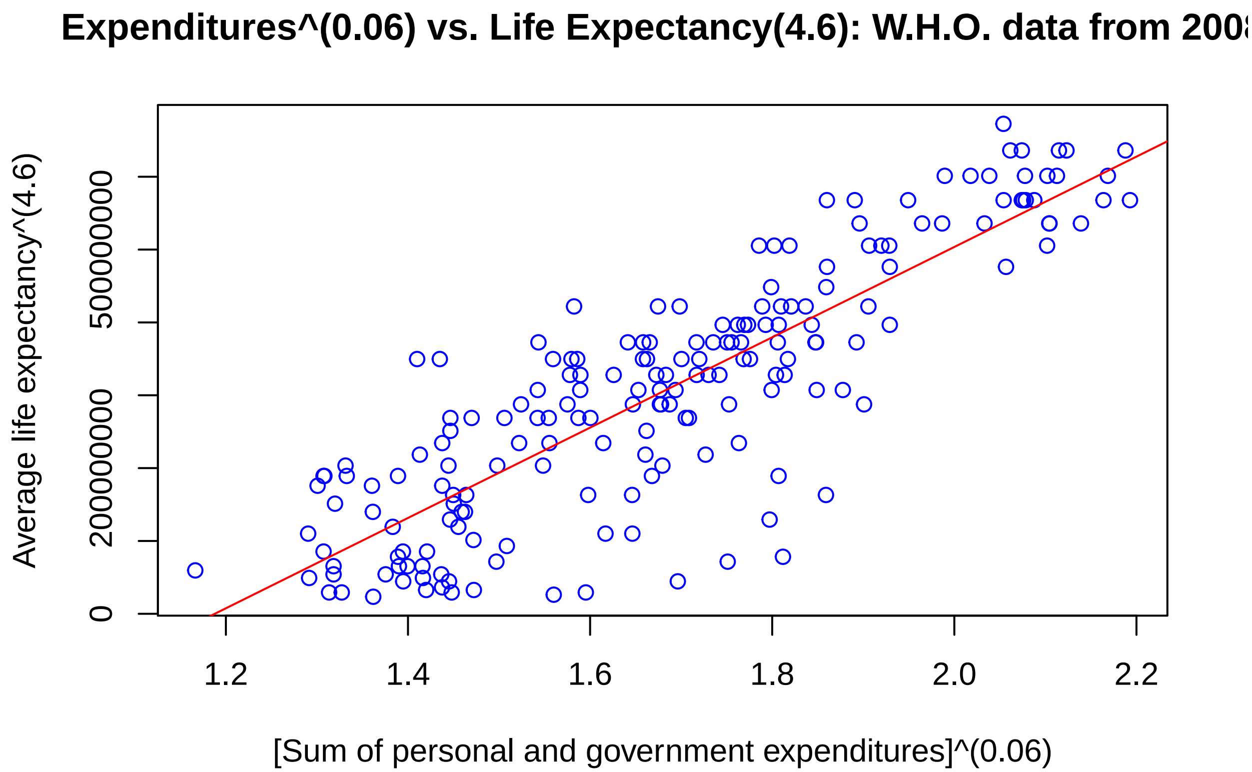
```
# Simple linear regression model, with transformed variables
Model2 <- lm(xformLifeExp ~ xformTotExp, data=who.df)
Summ2 <- summary(Model2)
Fstat2 <- Summ2$fstatistic[1]
Tstat2_0 <- Summ2$coefficients["(Intercept)","t value"]
Tstat2_1 <- Summ2$coefficients["xformTotExp","t value"]
Pval2_0 <- Summ2$coefficients["(Intercept)","Pr(>|t|)"]
Pval2_1 <- Summ2$coefficients["xformTotExp","Pr(>|t|)"]
Rsq2 <- Summ2$r.squared
AdjRsq2 <- Summ2$adj.r.squared
Correlation2 <- cor(who.df$xformLifeExp, who.df$xformTotExp)
print(Summ2)
```

```
##
## Call:
## lm(formula = xformLifeExp ~ xformTotExp, data = who.df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##               Estimate Std. Error  t value               Pr(>|t|)
## (Intercept) -736527910   46817945 -15.7317 < 0.0000000000000000222 ***
## xformTotExp  620060216   27518940  22.5321 < 0.0000000000000000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 90492400 on 188 degrees of freedom
## Multiple R-squared:  0.729767,   Adjusted R-squared:  0.72833
## F-statistic: 507.697 on 1 and 188 DF,  p-value: < 0.0000000000000000222
```

Make a scatterplot, with regression line

```r
plot(xformLifeExp~xformTotExp, data=who.df,
     xlab="[Sum of personal and government expenditures]^(0.06)",
     ylab="Average life expectancy^(4.6)",
     main="Expenditures^(0.06) vs. Life Expectancy(4.6): W.H.O. data from 2008",
     col="blue")
abline(Model2,col="red")
```



Provide and interpret the F statistics, R^2, standard error, and p-values.

The **F-statistic**, 507.696705395 is large, indicating significance. (The critical value for the F-test is 3.891398098).

Because there is only 1 degree of freedom in the numerator, the F-statistic equals the **square of the t-value**, 22.532126074, on the coefficient on the independent variable, `TotExp`.

The R^2 (0.729767299) and the **adjusted-**R^2 (0.728329891) values are strong, much better than those in the previous model. This indicating a much better fit. As they are about $\frac{3}{4}$, this indicates that the correlation between the independent and dependent variables is about $\frac{\sqrt{3}}{2} = 0.866025404$.

(We have confirmed that the actual correlation is 0.854264186 .)

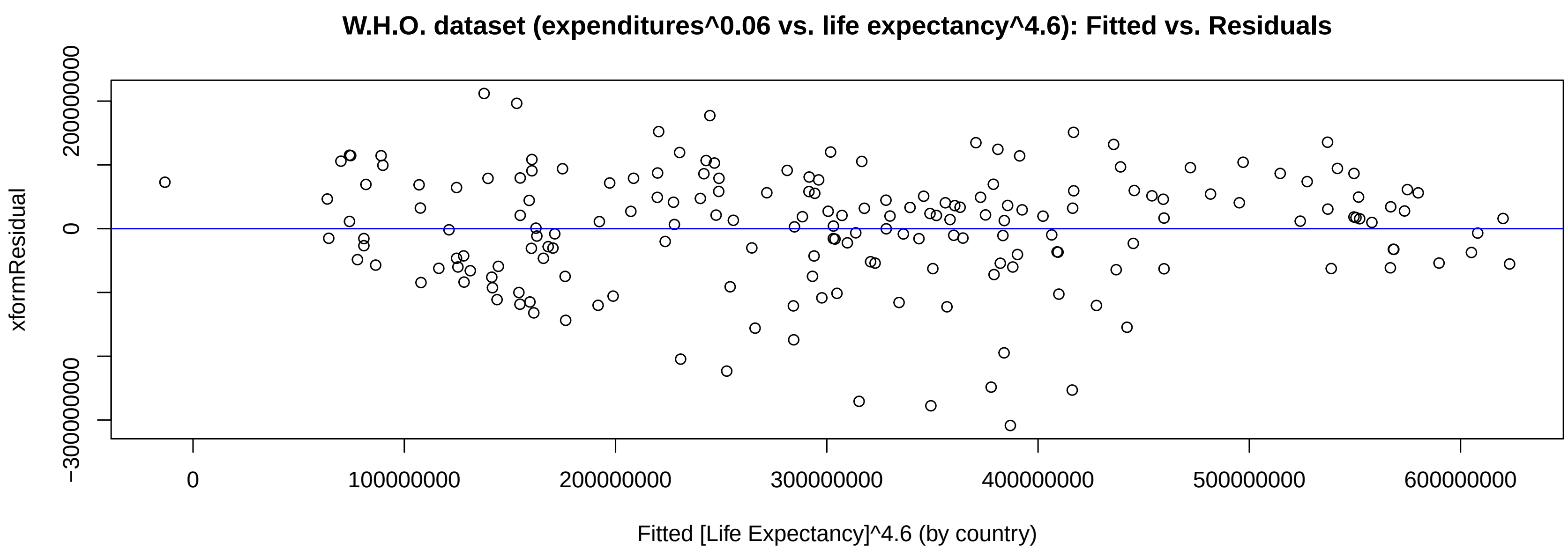
The `Null Hypothesis` is H_0: The regression coefficients are ***not*** significantly different from zero, indicating no relationship between the dependent and independent variables.

The `Alternative` is H_A: The coefficients ***are*** significantly different from zero.

The results indicate that the coefficients are significant, as the p-values are close to zero.

Linearity of the (transformed) data

```
xformResidual = resid(Model2)
xformFitted = fitted(Model2)
plot(xformFitted,xformResidual,
     main="W.H.O. dataset (expenditures^0.06 vs. life expectancy^4.6): Fitted vs. Residuals",
     xlab="Fitted [Life Expectancy]^4.6 (by country)")
abline(h=0, col="blue")
```

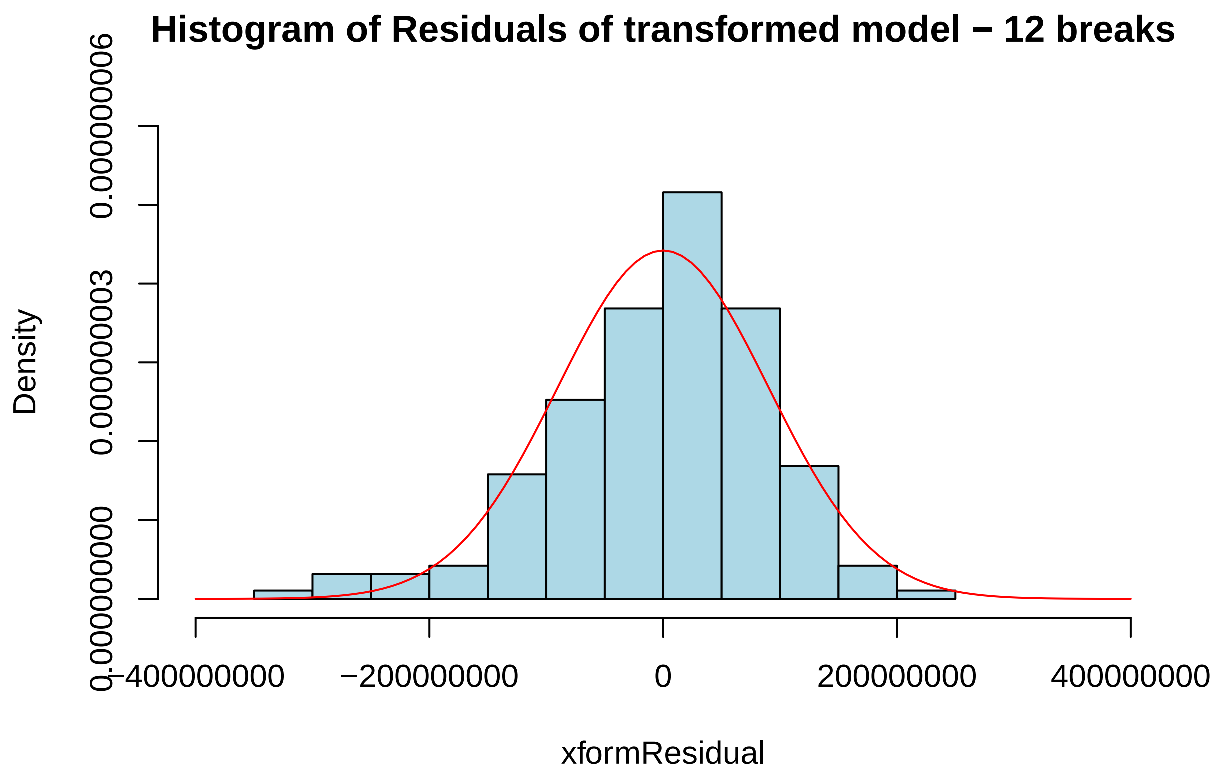


2. Normality of residuals of transformed model.

Histogram of transformed model:

We plot a histogram of the residuals:

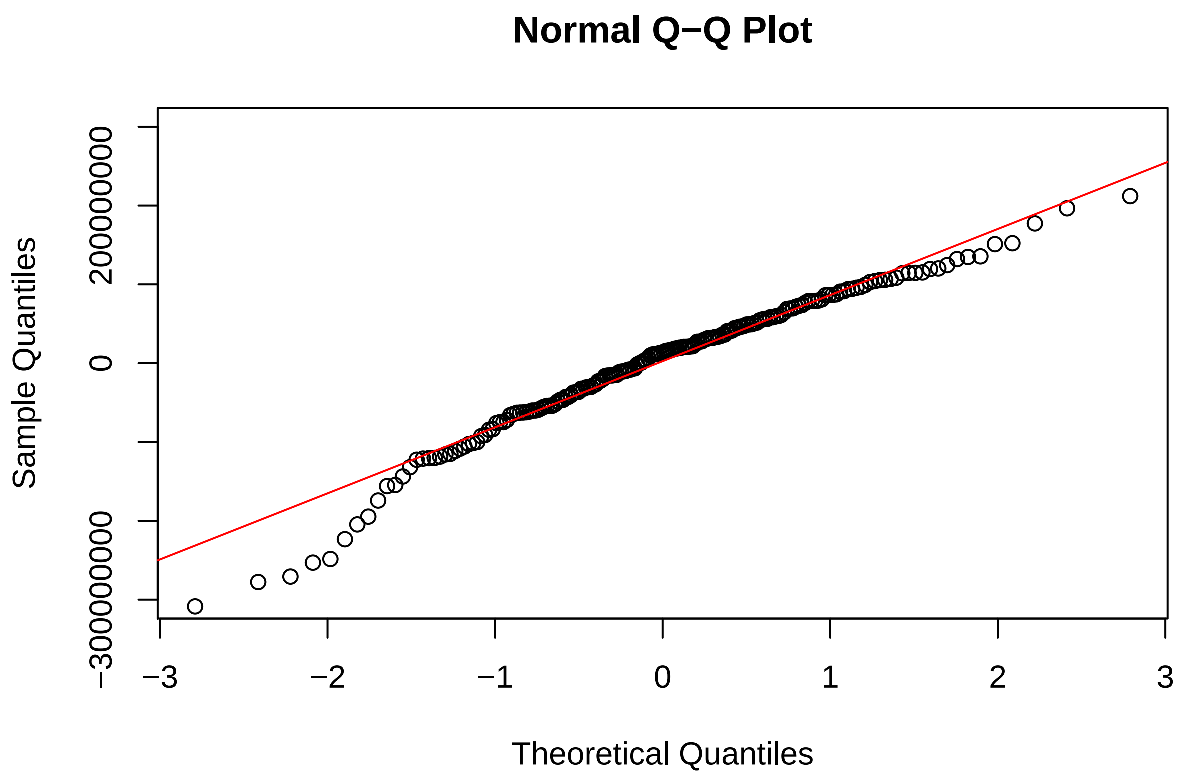
```
xformResidual = resid(Model2)
hist(xformResidual, main = "Histogram of Residuals of transformed model - 12 breaks", ylab = "Density",
     ylim = c(0, 0.000000006),
     xlim = c(-400000000,400000000),
     prob = TRUE,breaks=12, col="lightblue")
curve(dnorm(x, mean = mean(xformResidual), sd = sd(xformResidual)), col="red", add=TRUE)
```



The above distribution much more closely resembles the normal distribution, which is superimposed in red.

QQ-Plot on transformed model:

```
qqnorm(xformResidual, ylim=c(-300000000,300000000))
qqline(xformResidual, col="red")
```



The quantiles of the residuals on the transformed model much more closely match those of the Normal distribution, though there are is a thicker lower tail.

We can run several standard tests of normality:

```r
#library(olsrr)
ols_test_normality(Model2)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.967          0.0002
## Kolmogorov-Smirnov        0.0753         0.2322
## Cramer-von Mises         16.3596         0.0000
## Anderson-Darling          1.4003         0.0012
## -----------------------------------------------
```

While the Kolmogorov-Smirnov test no longer fails, every other test still fails, despite the transformations.

3. Homogeneity of residuals variance on transformed model.

Looking at the improvement of the graphs above suggests that this test should pass.

We can test using the `lmSupport` package:

```r
#library(lmSupport)
modelAssumptions(Model2,"LINEAR")
```

```
##
## Call:
## lm(formula = xformLifeExp ~ xformTotExp, data = who.df)
##
## Coefficients:
## (Intercept)  xformTotExp
##  -736527909    620060216
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
```

```
##  gvlma(x = Model)
##
##                         Value      p-value                   Decision
## Global Stat        27.811724 0.0000136181 Assumptions NOT satisfied!
## Skewness           17.137169 0.0000347751 Assumptions NOT satisfied!
## Kurtosis            7.458088 0.0063152042 Assumptions NOT satisfied!
## Link Function       2.986551 0.0839588386    Assumptions acceptable.
## Heteroscedasticity  0.229917 0.6315853166    Assumptions acceptable.
```

Despite the improvement in the model, various tests do fail according to the above diagnostic.
However, the heteroscedasticity test does pass.

4. Independence of residuals on transformed model

While there is clearly a relationship between the independent variable vs. the residual, the residuals themselves are actually independent of each other, as each represents the result from an individual country. Plotting the residuals sequentially (where the sequence happens to be alphabetical by country name) doesn't yield any discernable pattern:

```r
plot(xformResidual, main="Residuals of the transformed model")
abline(h=0, col="red")
```

Which model is "better?"

Despite the fact that the second model still fails several tests, the graphs indicate that it is much better than the first model.

---

3. Using the results from 3, forecast life expectancy when $TotExp^{.06} = 1.5$.

Then forecast life expectancy when $TotExp^{.06} = 2.5$.

```
# Create a dataframe with the values 1.5 and 2.5 for the x-value (here, renamed "xformTotExp")
predictor_values <-  data.frame(xformTotExp=c(1.5,2.5))

# call the predict function to obtain the prediction.
# Because the initial life expectancy values have been raised to power 4.6, we need to reverse this to obtain the desired result
prediction_results <- predict(object=Model2,predictor_values,interval = 'predict')^(1/4.6)
prediction_results
```

```
##          fit        lwr        upr
## 1 63.3115334 35.9354497 73.0079291
## 2 86.5064485 81.8064334 90.4341379
```

```
res <- round(prediction_results,2)
res
```

```
##     fit   lwr   upr
## 1 63.31 35.94 73.01
## 2 86.51 81.81 90.43
```

The life expectancy when $TotExp^{.06} = 1.5$ is 63.31 with a confidence interval of (35.94,73.01) .

The life expectancy when $TotExp^{.06} = 2.5$ is 86.506448484 with a confidence interval of (81.806433445,90.434137942) .

---

4. Build the following multiple regression model and interpret the F Statistics, R^2, standard error, and p-values.

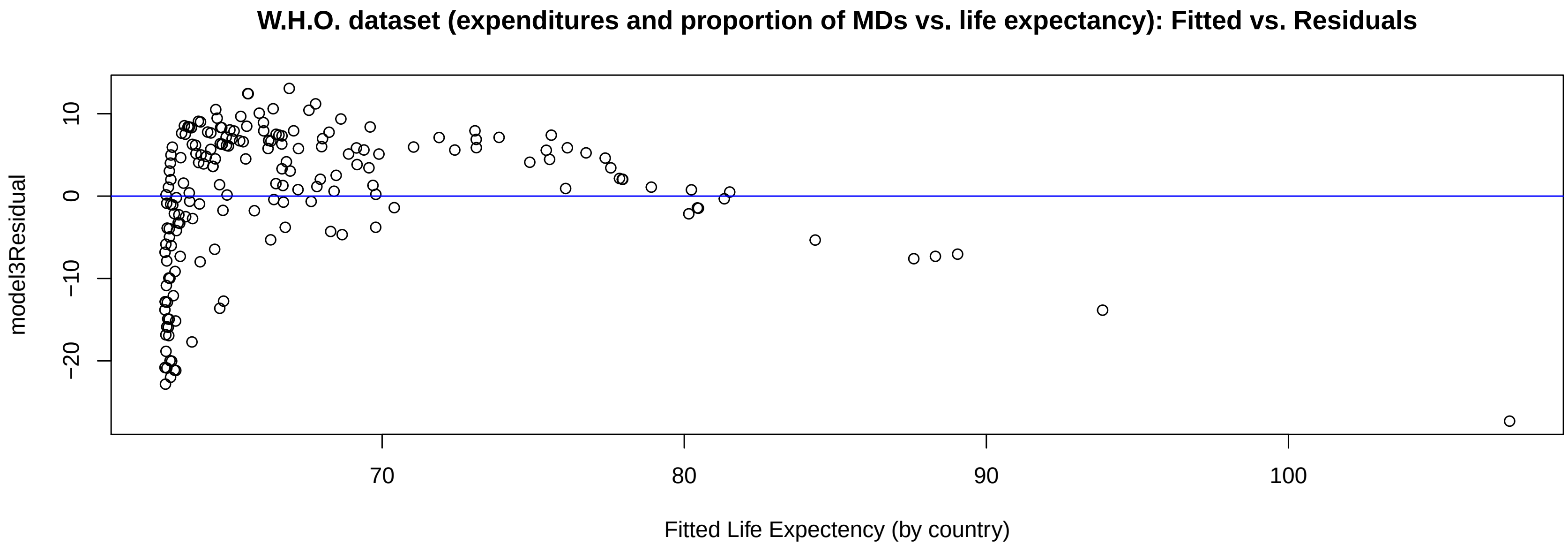
$$LifeExp = b_0 + b_1 \cdot PropMd + b_2 \cdot TotExp + b_3 \cdot PropMD \cdot TotExp$$

```r
Model3 <- lm(LifeExp ~ PropMD + TotExp + (PropMD*TotExp),data=who.df)
summary(Model3)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + (PropMD * TotExp), data = who.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -27.32028  -4.13191   2.09766   6.53970  13.07385
##
## Coefficients:
##                      Estimate       Std. Error  t value               Pr(>|t|)
## (Intercept)     62.77270325541    0.79560523844 78.89931 < 0.0000000000000000222 ***
## PropMD        1497.49395251893  278.81687965214  5.37089   0.000000232060277382 ***
## TotExp           0.00007233324    0.00000898193  8.05320   0.000000000000093863 ***
## PropMD:TotExp   -0.00602568644    0.00147235740 -4.09254   0.000063527329494147 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.76549 on 186 degrees of freedom
## Multiple R-squared:  0.357435,   Adjusted R-squared:  0.347071
## F-statistic: 34.4883 on 3 and 186 DF,  p-value: < 0.0000000000000000222
```

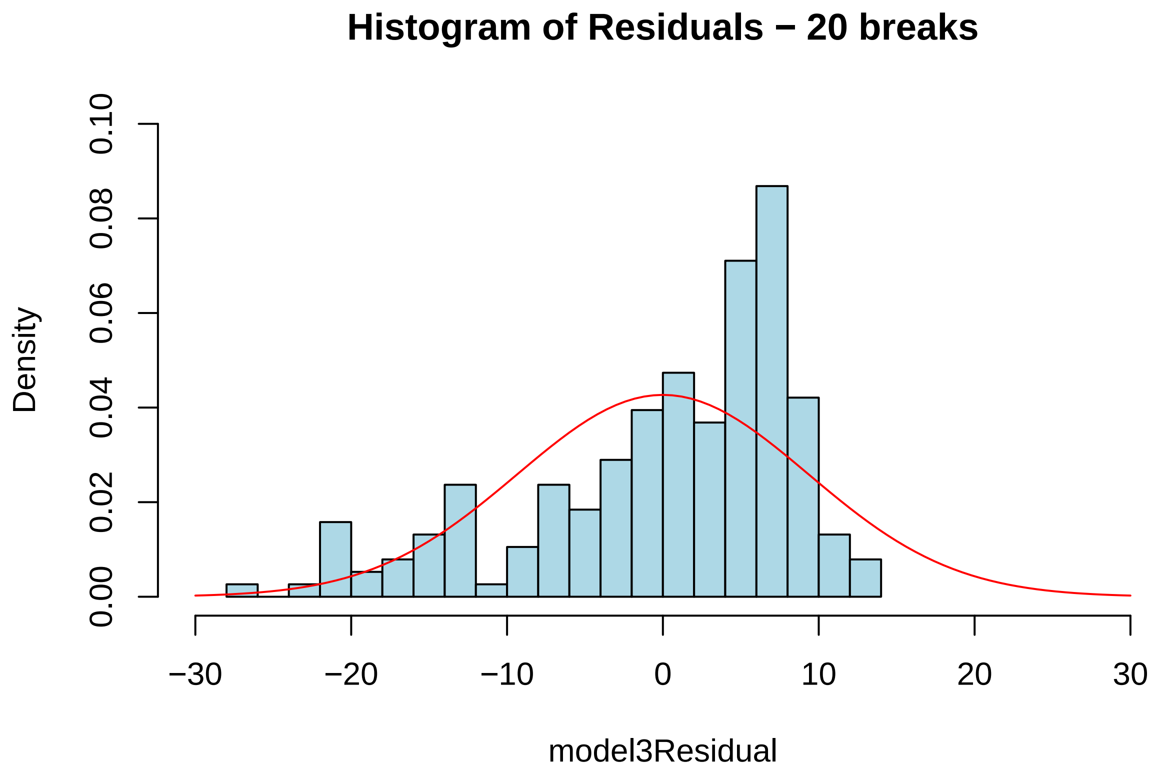
Linearity of the model with interaction term

```r
model3Residual = resid(Model3)
model3Fitted = fitted(Model3)
plot(model3Fitted,model3Residual,
     main="W.H.O. dataset (expenditures and proportion of MDs vs. life expectancy): Fitted vs. Residuals",
     xlab="Fitted Life Expectency (by country)")
abline(h=0, col="blue")
```



The model does not show much improvement over the initial model.

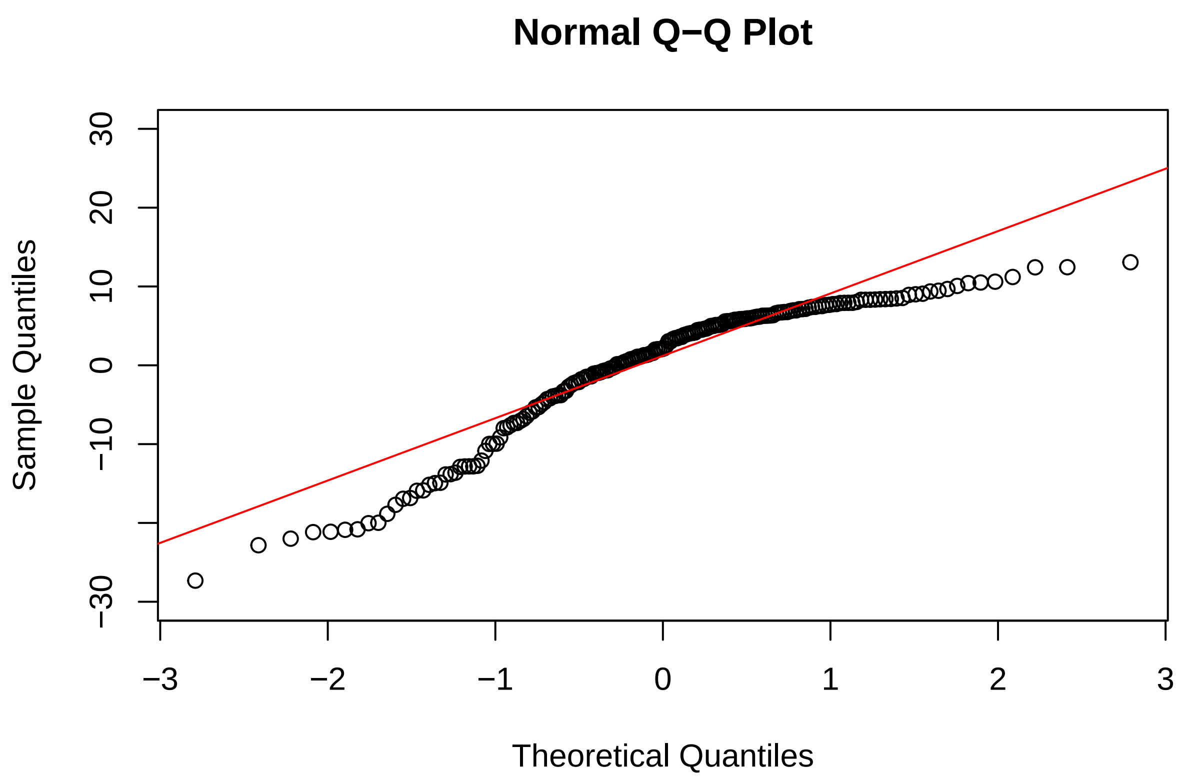
```
model3Residual = resid(Model3)
hist(model3Residual, main = "Histogram of Residuals - 20 breaks", ylab = "Density",
     ylim = c(0, 0.10),
     xlim = c(-30,30),
     prob = TRUE,breaks=20, col="lightblue")
curve(dnorm(x, mean = mean(Residual), sd = sd(Residual)), col="red", add=TRUE)
```



The above histogram does not closely resemble the corresponding Normal distribution.

QQPlot:

```
qqnorm(model3Residual, ylim=c(-30,30))
qqline(model3Residual, col="red")
```



Not a good result.

Tests of Normality:

```
#library(olsrr)
ols_test_normality(Model3)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
```

```
## Shapiro-Wilk              0.8991         0.0000
## Kolmogorov-Smirnov        0.1274         0.0042
## Cramer-von Mises         15.7378         0.0000
## Anderson-Darling          6.5834         0.0000
## -----------------------------------------------
```

All tests fail.

Homogeneity of residuals variance:

Clearly looking at the graphs above suggests that this test should fail.

```
#library(lmSupport)
modelAssumptions(Model3,"Normal")
```

```
## Descriptive Statistics for Studentized Residuals
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + (PropMD * TotExp), data = who.df)
##
## Coefficients:
##     (Intercept)           PropMD           TotExp    PropMD:TotExp
##   62.7727032554  1497.4939525189     0.0000723332    -0.0060256864
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = Model)
##
##                       Value              p-value                   Decision
## Global Stat        87.070286 0.000000000000000000 Assumptions NOT satisfied!
## Skewness           33.421944 0.000000007418221926 Assumptions NOT satisfied!
## Kurtosis            0.560019 0.454252478434432505    Assumptions acceptable.
## Link Function      52.728414 0.000000000000383027 Assumptions NOT satisfied!
## Heteroscedasticity  0.359909 0.548556883534016215    Assumptions acceptable.
```

The model does pass the heteroscedasticity test.

How good is the model?

While the model has a higher R^2 than the initial model, it still fails numerous tests.

5. Forecast LifeExp when $PropMD = .03$ and $TotExp = 14$.

```
# Create a dataframe with the PropMD = .03 and TotExp = 14 for the x-values
predictor_values3 <-  data.frame(PropMD=0.03, TotExp=14)

# call the predict function to obtain the prediction.

prediction_results3 <- predict(object=Model3,predictor_values3,interval = 'predict')
prediction_results3
```

```
##          fit       lwr      upr
## 1 107.696004 84.2479069 131.144101
```

```
res3 <- round(prediction_results3,2)
res3
```

```
##     fit   lwr    upr
## 1 107.7 84.25 131.14
```

Under this model, the life expectancy when is 107.7 with a confidence interval of (84.25,131.14) .

Does this forecast seem realistic?

No, it does not seem realistic, as this represents an expected lifespan in years.

Why or why not?

If we look at the input data, the highest life expectancy is 83.

Furthermore, there are only two countries (Cyprus and San Marino) with a proportion of MDs greater than 3 percent of the population (indeed, greater than 1 percent):

```
many_MDs <- who.df[who.df$PropMD > 0.03,]
(many_MDs[,1:10]) %>% kable() %>% kable_styling(c("striped", "bordered"))
```

	Country	LifeExp	InfantSurvival	Under5Survival	TBFree	PropMD	PropRN	PersExp	GovtExp	TotExp
45	Cyprus	80	0.997	0.996	0.99994	0.033228132	0.003972813	1350	39399	40749
146	San Marino	82	0.997	0.997	0.99995	0.035129032	0.070838710	3490	278163	281653

These are both relatively wealthy (and, small) European countries with long life expectancies (80 and 82).

There is only one country (Burundi) with a `TotExp` less than 50:

```
small_expenditure <- who.df[who.df$TotExp < 50,]
small_expenditure[,1:10] %>% kable() %>% kable_styling(c("striped", "bordered"))
```

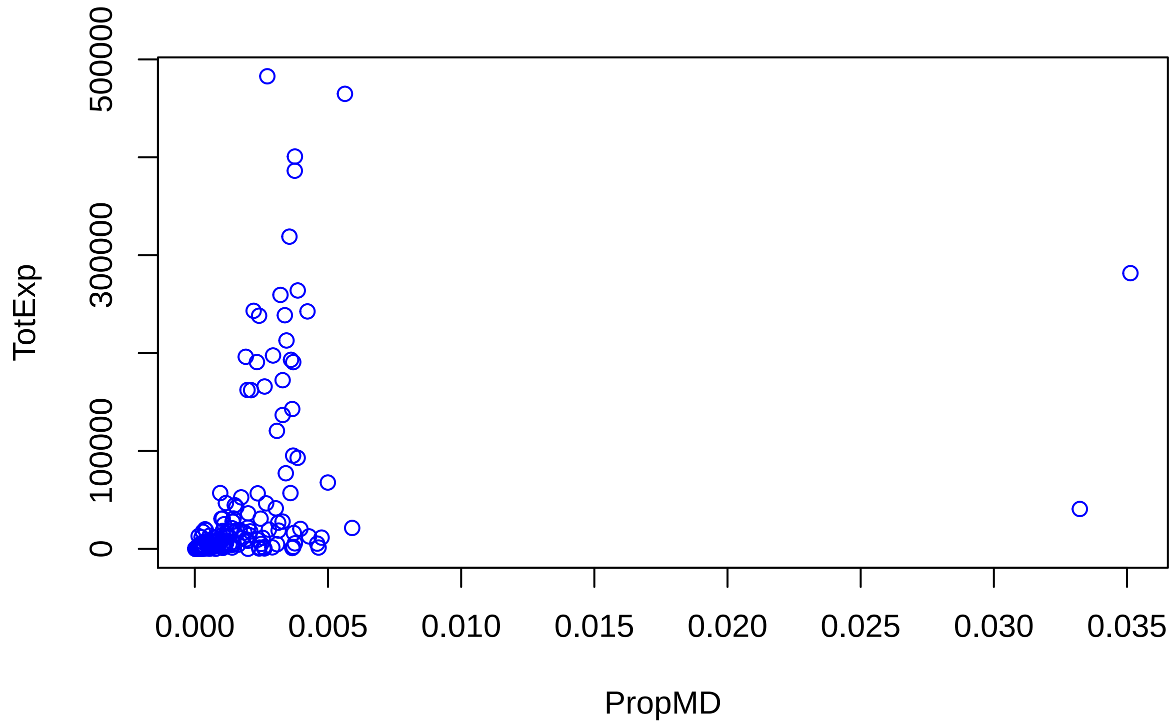
	Country	LifeExp	InfantSurvival	Under5Survival	TBFree	PropMD	PropRN	PersExp	GovtExp	TotExp
28	Burundi	49	0.891	0.819	0.99286	0.0000245	0.000164933	3	10	13

Here the life expectancy of 48 is among the lowest such expectancies, and the proportion of MDs in the country is miniscule.

So, there are no countries similar to that requested for the prediction – with a very high proportion of MDs in the population but extremely low expenditures on healthcare.

If we plot the relationship between PropMD and TotExp we see the following:

```
plot(PropMD,TotExp, col="blue")
```



If there were any country which had PropMD and TotExp similar to those requested for prediction, it would be located at the very bottom (immediately above the 0.030 tickmark.) WHile there is a country which appears to be relatively nearby (Cyprus, in the lower right) the key difference is that Cyprus is a rather small country which which spends a sizable amount on healthcare and has high life expectancy.

Accordingly, the data doesn't support forecasting the life expectancy for a country where 3 percent of the population are doctors but the average healthcare spending is only $14 because there are no countries like this.
