Data 605 - Assignment 12

Leticia Salazar

4/11/2022

Contents

Regression Analysis

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

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.

```
# Import libraries
library(tidyverse)
library(expm)
```

```
# Load Data
who <- read.csv('https://raw.githubusercontent.com/letisalba/Data-605/main/Week-12/who.csv', header = T
head(who)
```

```
##               Country LifeExp InfantSurvival Under5Survival  TBFree      PropMD
## 1         Afghanistan      42          0.835          0.743 0.99769 0.000228841
## 2             Albania      71          0.985          0.983 0.99974 0.001143127
## 3             Algeria      71          0.967          0.962 0.99944 0.001060478
## 4             Andorra      82          0.997          0.996 0.99983 0.003297297
## 5              Angola      41          0.846          0.740 0.99656 0.000070400
## 6 Antigua and Barbuda      73          0.990          0.989 0.99991 0.000142857
##        PropRN PersExp GovtExp TotExp
## 1 0.000572294      20      92    112
## 2 0.004614439     169    3128   3297
## 3 0.002091362     108    5184   5292
## 4 0.003500000    2589  169725 172314
## 5 0.001146162      36    1620   1656
## 6 0.002773810     503   12543  13046
```

```
# glimpse of data
glimpse(who)
```

```
## Rows: 190
## Columns: 10
## $ Country        <chr> "Afghanistan", "Albania", "Algeria", "Andorra", "Angola~
## $ LifeExp        <int> 42, 71, 71, 82, 41, 73, 75, 69, 82, 80, 64, 74, 75, 63,~
## $ InfantSurvival <dbl> 0.835, 0.985, 0.967, 0.997, 0.846, 0.990, 0.986, 0.979,~
## $ Under5Survival <dbl> 0.743, 0.983, 0.962, 0.996, 0.740, 0.989, 0.983, 0.976,~
## $ TBFree         <dbl> 0.99769, 0.99974, 0.99944, 0.99983, 0.99656, 0.99991, 0~
## $ PropMD         <dbl> 0.000228841, 0.001143127, 0.001060478, 0.003297297, 0.0~
## $ PropRN         <dbl> 0.000572294, 0.004614439, 0.002091362, 0.003500000, 0.0~
## $ PersExp        <int> 20, 169, 108, 2589, 36, 503, 484, 88, 3181, 3788, 62, 1~
## $ GovtExp        <int> 92, 3128, 5184, 169725, 1620, 12543, 19170, 1856, 18761~
## $ TotExp         <int> 112, 3297, 5292, 172314, 1656, 13046, 19654, 1944, 1907~
```

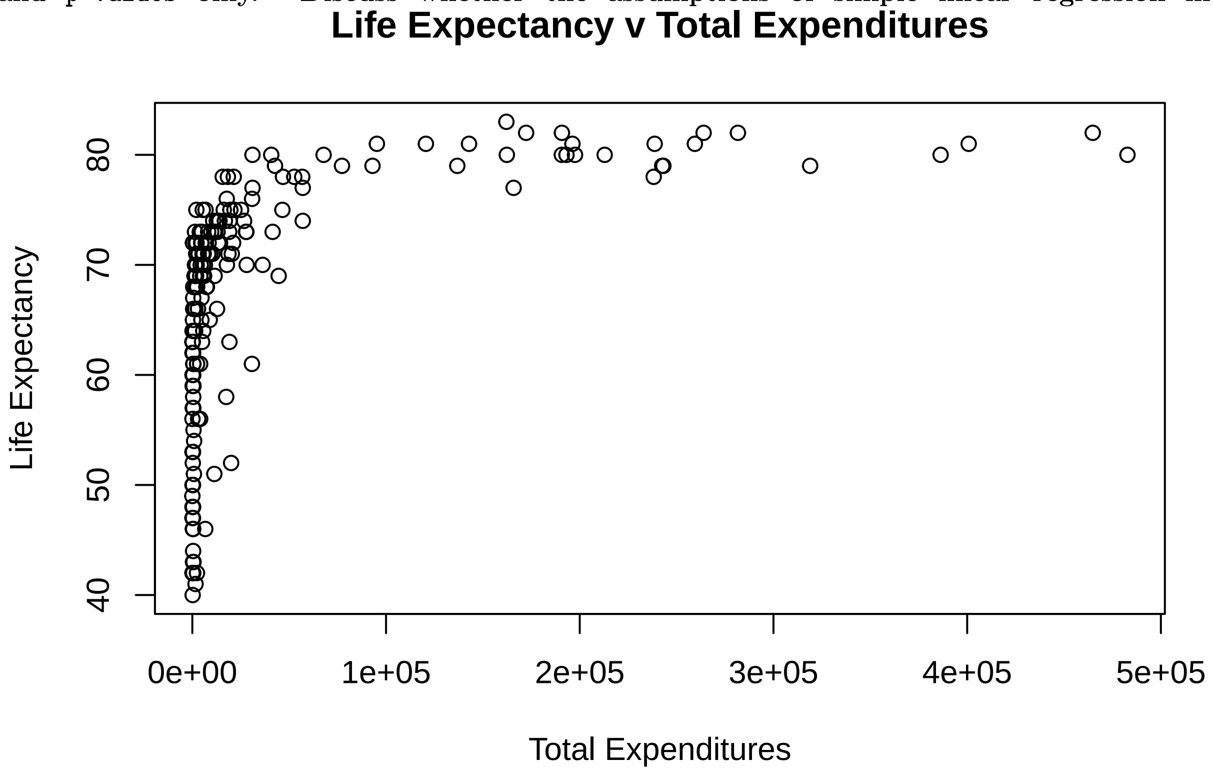
```
# summary of data
summary(who)
```

```
##    Country             LifeExp      InfantSurvival   Under5Survival
##  Length:190         Min.   :40.00   Min.   :0.8350   Min.   :0.7310
##  Class :character   1st Qu.:61.25   1st Qu.:0.9433   1st Qu.:0.9253
##  Mode  :character   Median :70.00   Median :0.9785   Median :0.9745
##                     Mean   :67.38   Mean   :0.9624   Mean   :0.9459
##                     3rd Qu.:75.00   3rd Qu.:0.9910   3rd Qu.:0.9900
##                     Max.   :83.00   Max.   :0.9980   Max.   :0.9970
##      TBFree            PropMD              PropRN              PersExp
##  Min.   :0.9870   Min.   :0.0000196   Min.   :0.0000883   Min.   :   3.00
##  1st Qu.:0.9969   1st Qu.:0.0002444   1st Qu.:0.0008455   1st Qu.:  36.25
##  Median :0.9992   Median :0.0010474   Median :0.0027584   Median : 199.50
##  Mean   :0.9980   Mean   :0.0017954   Mean   :0.0041336   Mean   : 742.00
##  3rd Qu.:0.9998   3rd Qu.:0.0024584   3rd Qu.:0.0057164   3rd Qu.: 515.25
##  Max.   :1.0000   Max.   :0.0351290   Max.   :0.0708387   Max.   :6350.00
##     GovtExp             TotExp
##  Min.   :    10.0   Min.   :    13
##  1st Qu.:   559.5   1st Qu.:   584
##  Median :  5385.0   Median :  5541
##  Mean   : 40953.5   Mean   : 41696
##  3rd Qu.: 25680.2   3rd Qu.: 26331
##  Max.   :476420.0   Max.   :482750
```

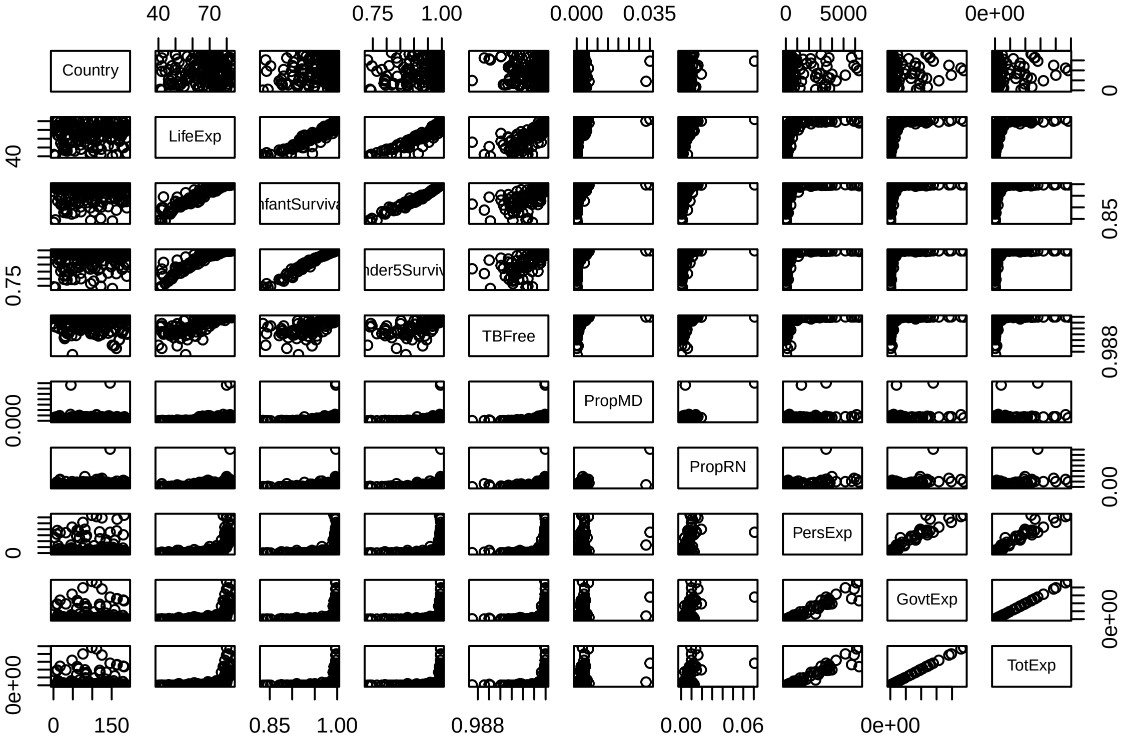
```r
# Linear Model
my_lm <- lm(LifeExp ~ TotExp, who)

# Scatter plot
plot(LifeExp ~ TotExp, who,
     xlab = "Total Expenditures", ylab = "Life Expectancy",
     main = "Life Expectancy v Total Expenditures")
```

1. Provide a scatterplot of `LifeExp ~ TotExp`, and run simple linear regression. Do not transform the variables. Provide and interpret the `F statistics`, `R^2`, `standard error`, and `p-values` only. Discuss whether the assumptions of simple linear regression met.



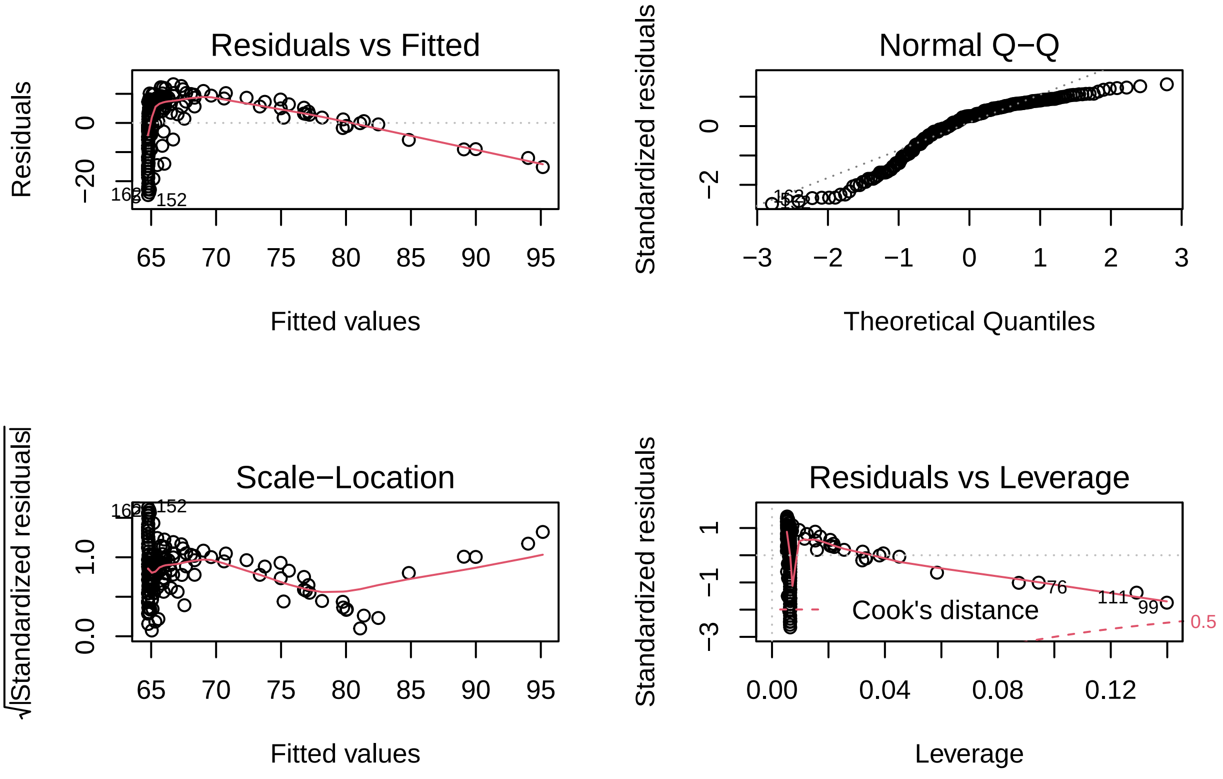
```r
par(mfrow = c(2,2))
plot(who)
```



```r
# Summary of linear model
summary(my_lm)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
## TotExp      6.297e-05  7.795e-06   8.079 7.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.371 on 188 degrees of freedom
## Multiple R-squared:  0.2577, Adjusted R-squared:  0.2537
## F-statistic: 65.26 on 1 and 188 DF,  p-value: 7.714e-14
```

```r
par(mfrow = c(2,2))
plot(my_lm)
```



F-Statistic and P-Value: This test if any of the independent variables are related to the Y outcome. If the p-value associated is $\geq$ 0.05 then there is no relationship and if $\leq$ 0.05 then there is at least 1 independent variable related to Y. From the summary we can see that the F-statistic is 65.26 and the p-value is 7.714e-14 which is less than 0.05, meaning there is at least one possible independent variable related to Y. Being that the p-value is relatively small, we can reject the null hypothesis and accept the alternative that the linear model is a better fit for the data.

R^2**: This measures how well the model describes our data. With a 0.2577** R^2 **value, then 25.77% explains the variance in our data set.**

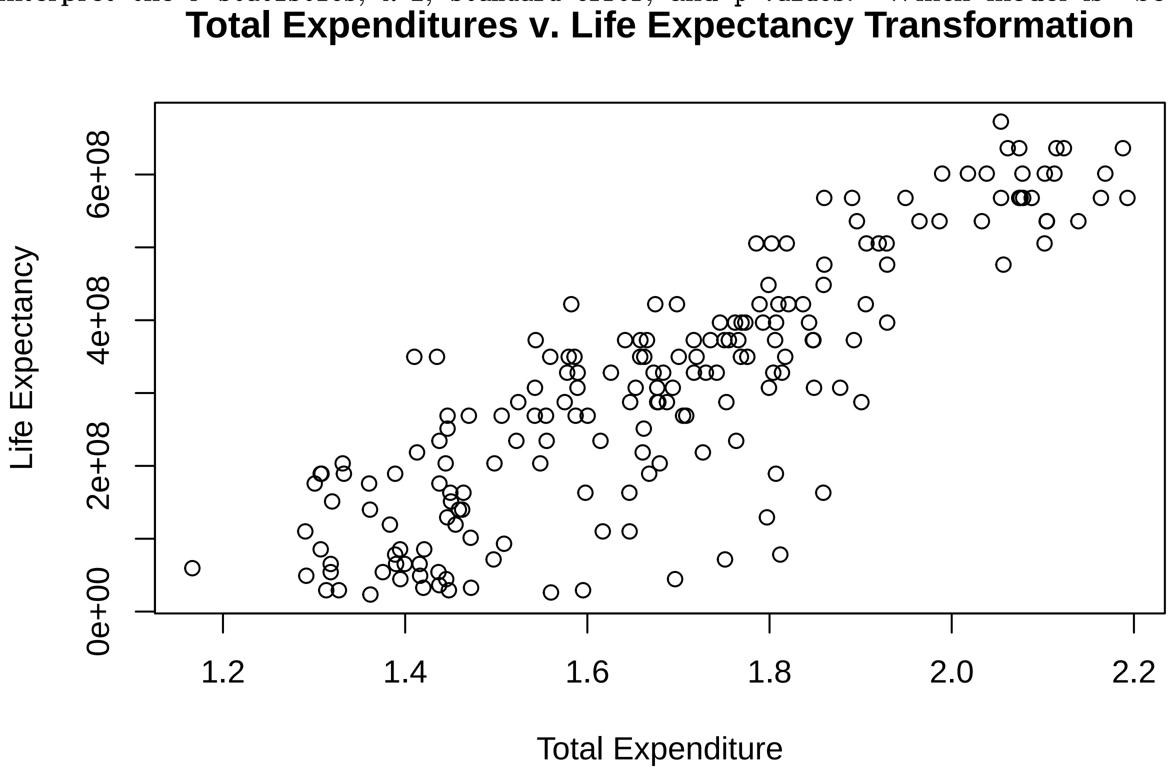
Standard Error: When looking at the standard error you are looking for the variation in the residuals. For this data set the standard error is 9.371 on 188 degrees of freedom.

Based on this we cannot asssume that linear regression is met because it doesn't seem to fully follow a linear trend and there a very low variance (R^2**) with the data, so there may be other factors that come to play.**

```
LifeExp46 <- (who$LifeExp ** (4.6))
TotExp06 <- (who$TotExp ** (.06))

plot(TotExp06, LifeExp46,
     xlab = "Total Expenditure",
     ylab = "Life Expectancy",
     main = "Total Expenditures v. Life Expectancy Transformation")
```

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, $TotExp^{.06}$). Plot $LifeExp^{4.6}$ as a function of $TotExp^{.06}$, and r re-run the simple regression model using the transformed variables. Provide and interpret the `F` statistics, `R^2`, `standard error`, and `p-values`. Which model is "better"?

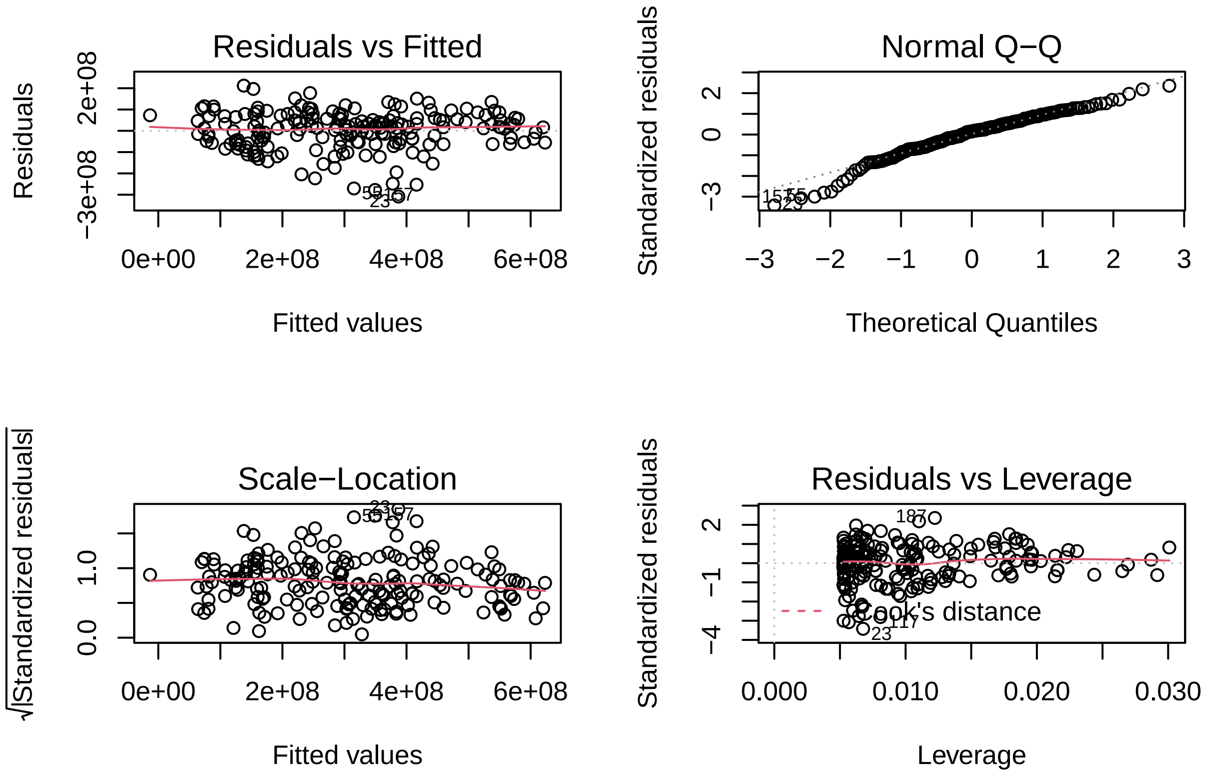


```
my_lm2 <- lm(LifeExp46 ~ TotExp06, who)
summary(my_lm2)
```

```
##
## Call:
## lm(formula = LifeExp46 ~ TotExp06, data = who)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -736527910   46817945  -15.73   <2e-16 ***
## TotExp06     620060216   27518940   22.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 90490000 on 188 degrees of freedom
## Multiple R-squared:  0.7298, Adjusted R-squared:  0.7283
## F-statistic: 507.7 on 1 and 188 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(my_lm2)
```



F-Statistic and P-Value: Similarly to the first model, the F-statistic is 507.7 and the P-value is $<$ 2.2e-16, being that the P-value is small we know this model fits the data well.

R^2: With a 0.7298 R^2 value, then 72.98% explains the variance in our data set which is much higher than the first model.

Standard Error: For this model the standard error is 90490000 on 188 degrees of freedom.

Based on this we can asssume that linear regression is met because the plot looks more linear, the R^2 value is much higher than the first model and the P-value is still less than 0.05.

```r
# Key
a <- -736527910
b <- 620060216

# Forecasting life expectancy when TotExp^.06 = 1.5
LifeExp_46 <- a + b * 1.5
LifeExp15 <- exp(log(LifeExp_46) / 4.6)
LifeExp15
```

3. Using the results from 3, forecast `life expectancy when TotExp^.06 = 1.5`. Then forecast `life expectancy when TotExp^.06 = 2.5`.

```
## [1] 63.31153
```

```r
# Forecasting life expectancy when TotExp^.06 = 2.5
LifeExp2_46 <- a + b * 2.5
LifeExp25 <- exp(log(LifeExp2_46) / 4.6)
LifeExp25
```

```
## [1] 86.50645
```

4. Build the following multiple regression model and interpret the `F Statistics, $R^2$, standard error, and p-values`. How good is the model?

```r


my_lm3 <- lm(LifeExp ~ PropMD + TotExp + TotExp:PropMD, who)
summary(my_lm3)
```

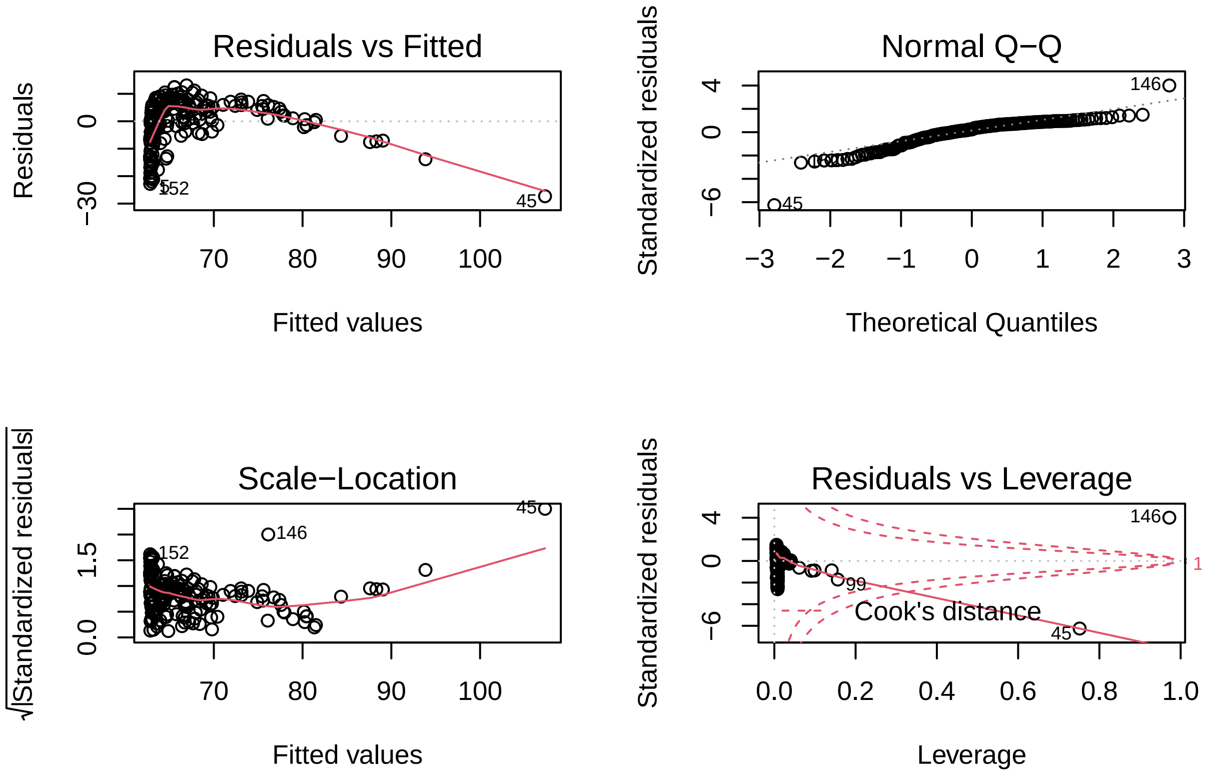
`LifeExp = b0 + b1 x PropMd + b2 x TotExp +b3 x PropMD x TotExp`

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + TotExp:PropMD, data = who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.277e+01  7.956e-01  78.899  < 2e-16 ***
## PropMD         1.497e+03  2.788e+02   5.371 2.32e-07 ***
## TotExp         7.233e-05  8.982e-06   8.053 9.39e-14 ***
## PropMD:TotExp -6.026e-03  1.472e-03  -4.093 6.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.765 on 186 degrees of freedom
## Multiple R-squared:  0.3574, Adjusted R-squared:  0.3471
## F-statistic: 34.49 on 3 and 186 DF,  p-value: < 2.2e-16
```

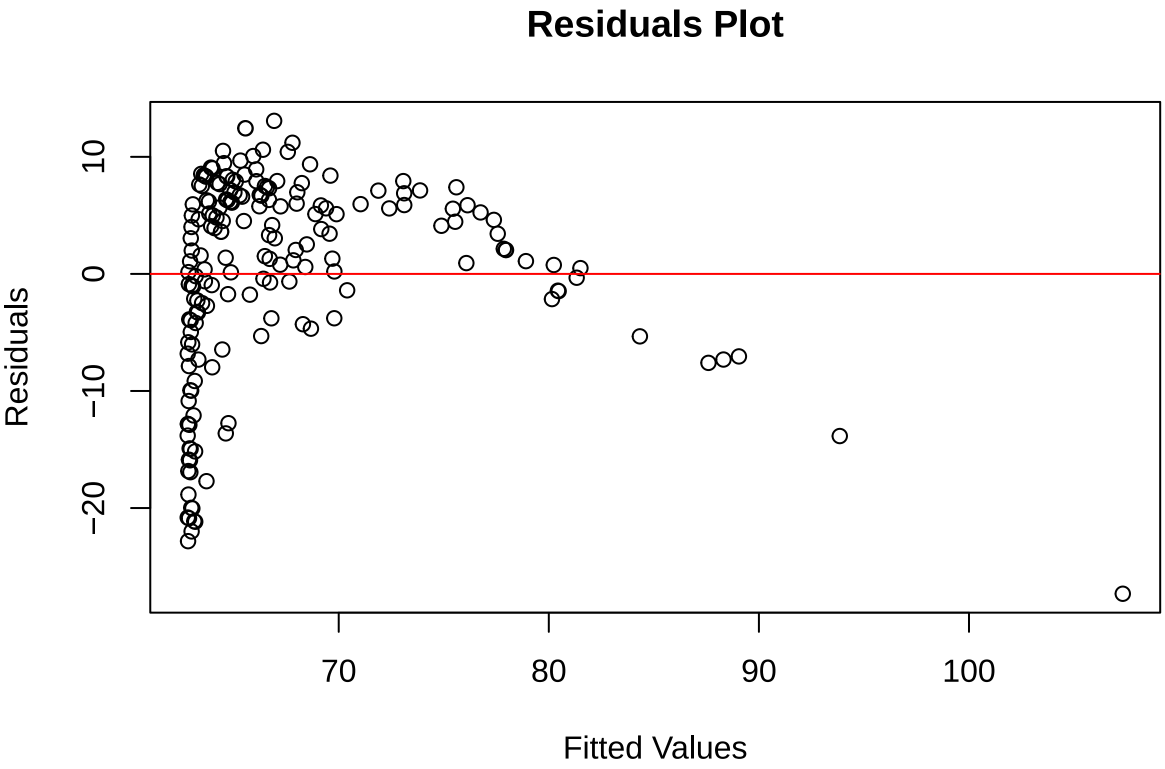
```
par(mfrow = c(2,2))
plot(my_lm3)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



```
plot(my_lm3$fitted.values, my_lm3$residuals,
     xlab="Fitted Values", ylab="Residuals",
     main="Residuals Plot")
abline(0,0, col = 'red')
```



F-Statistic and P-Value: the F-statistic and p-value are still relatively low so we know the model fits the data well.

R^2: Comparing it with the second model, the R^2 decreased to 35.74% of variance in the data.

Standard Error: For this model the standard error is 8.765 on 186 degrees of freedom.

Based on the model and plot above, this doesn't look like it's normally distributed and the R^2 value only accounts for a low amount of variance compared to the second model. Therefore, this model is not a good fit.

```
# Key
inter <-  62.8
co2 <- 0.00007233
co3 <- 1497
PropMD <- .03
TotExp <- 14

pred_5 <- inter + co2 * TotExp + co3 * PropMD + .006 * 14 * PropMD
pred_5
```

5. `Forecast LifeExp when PropMD = .03 and TotExp = 14.` Does this forecast seem realistic? Why or why not?

```
## [1] 107.7135
```

This forecast doesn't seem realistic because it is such a long time for a person's life expectancy.