DATA 605 Multiple Regression

Warner Alexis

2024-04-14

Data Analysis

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

> 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.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(cowplot)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
data <- read.csv('who.csv', stringsAsFactors = F)
head(data)
```

```
##               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
```

```r
str(data)
```

```
## 'data.frame':    190 obs. of  10 variables:
##  $ Country       : chr  "Afghanistan" "Albania" "Algeria" "Andorra" ...
##  $ 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  2.29e-04 1.14e-03 1.06e-03 3.30e-03 7.04e-05 ...
##  $ 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 ...
```

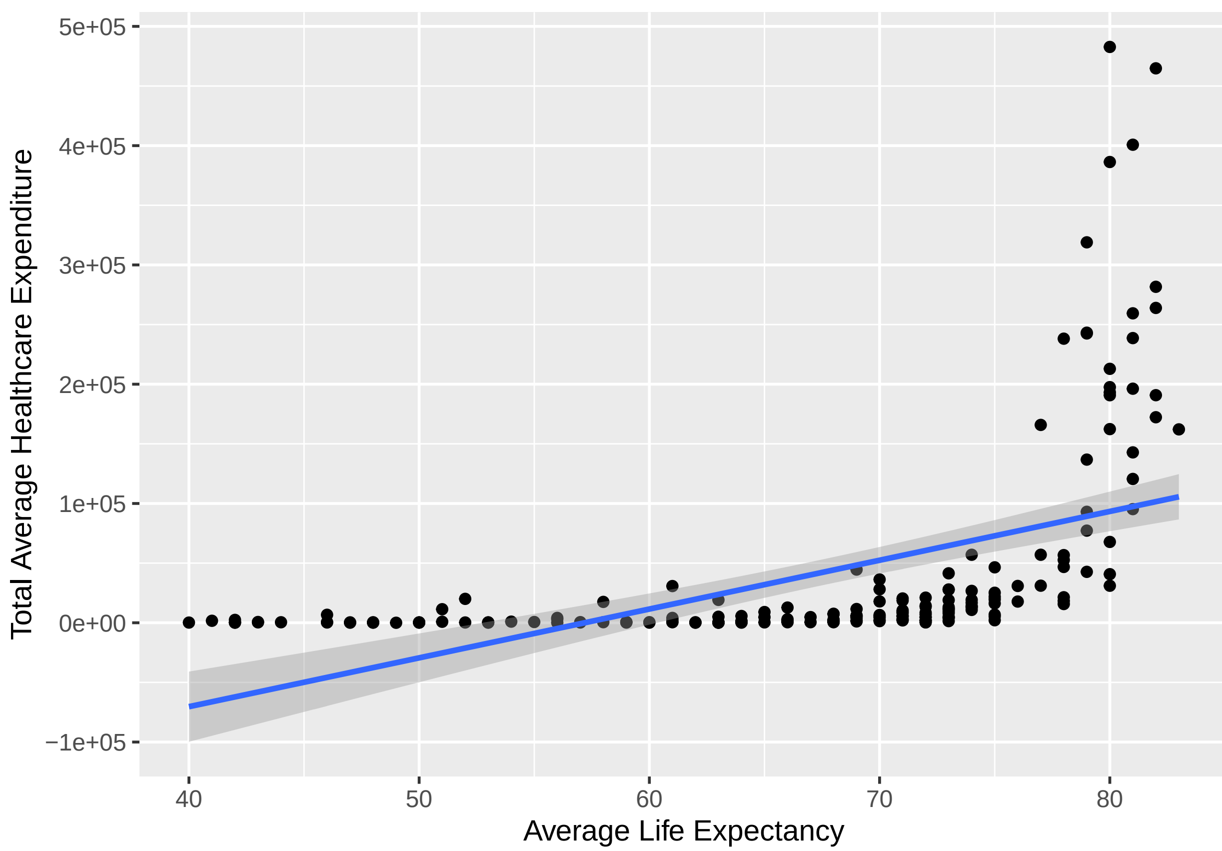
```r
summary(data)
```

```
##    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
ggplot(data, aes(x = LifeExp ,y = TotExp)) +
  geom_point() +
  labs(x = "Average Life Expectancy", y = "Total Average Healthcare Expenditure") +
  geom_smooth(method=lm)
```

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




```r
set.seed(123)
linear_lm  <- lm(LifeExp ~ TotExp, data= data)
summary(linear_lm)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = data)
##
```

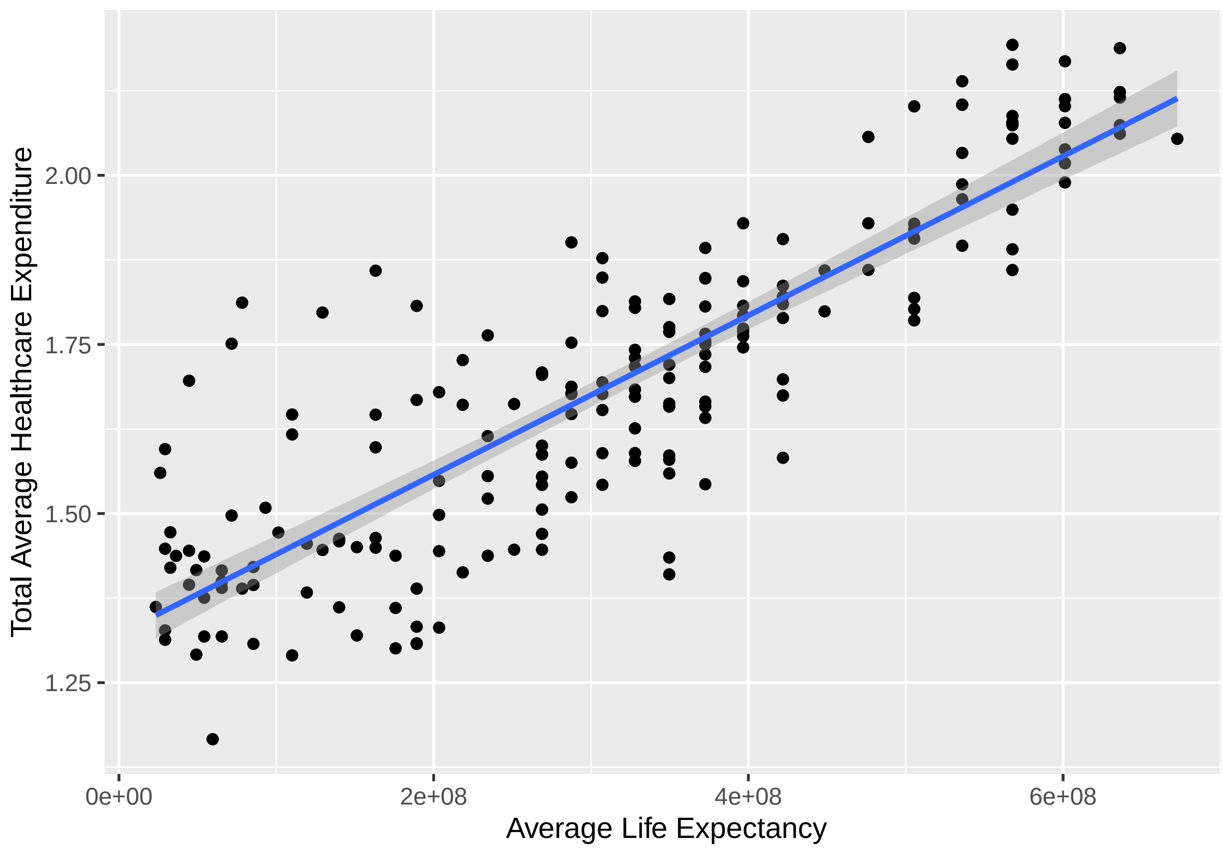
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
## TotExp      6.297e-05  7.795e-06   8.079 7.71e-14 ***
## ---
## 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
```

pvalue for TotExp is very small which is lower than 0.05 indicates that it is significant for the prefiction of the LifeExp. the adjusted R-squared 0.2537 is too low which show us that the model need a lot of more works. We can assume there is a linear relationship betwen the feature and the target variables, but not a strong one since thye model has low pvalue.

```r
x_e = 4.6
y_e = 0.06
df <- data %>%
  mutate(LifeExpT = LifeExp^x_e,
         TotExpT = TotExp^y_e)


ggplot(df, aes(x = LifeExpT ,y = TotExpT)) +
  geom_point() +
  labs(x = "Average Life Expectancy", y = "Total Average Healthcare Expenditure") +
  geom_smooth(method=lm)
```

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

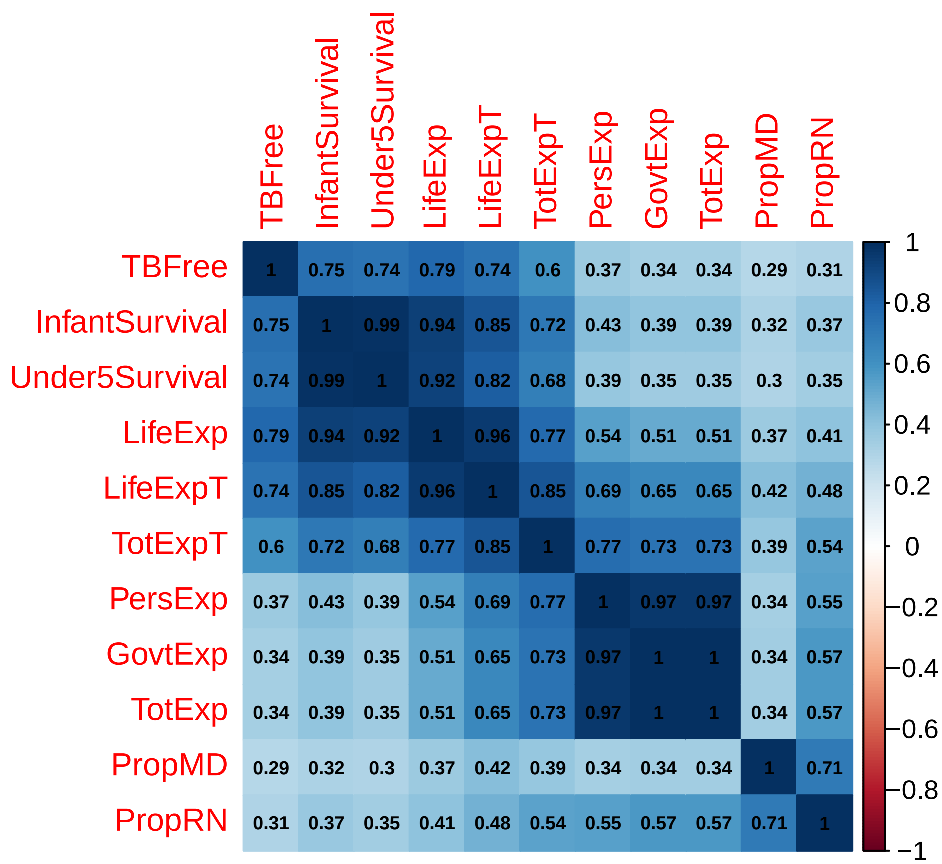



```r
set.seed(123)
linear_lm2 <- lm(LifeExpT ~ TotExpT, data= df)
summary(linear_lm2)
```

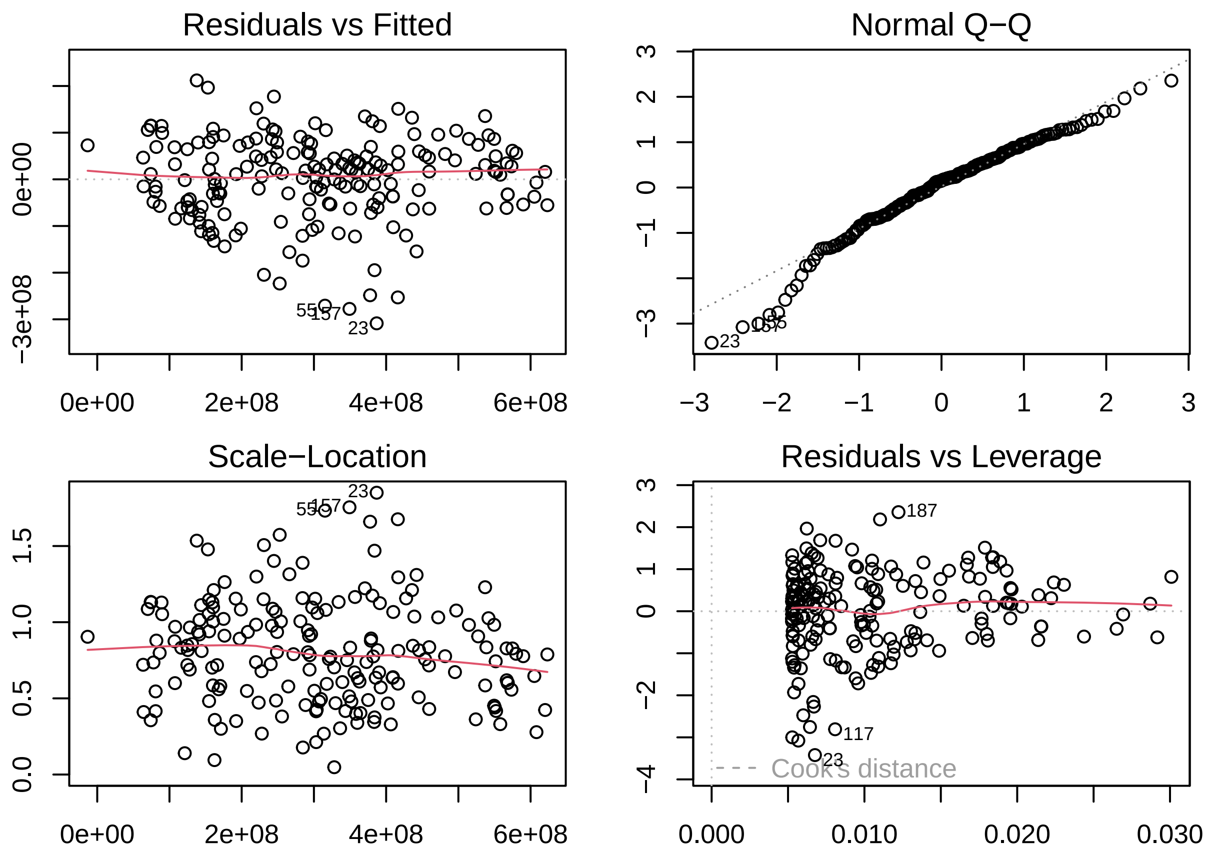
```
##
## Call:
## lm(formula = LifeExpT ~ TotExpT, data = df)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -736527910   46817945  -15.73   <2e-16 ***
## TotExpT      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
```

pvalue is significantly low which means TotExpT is a statistically significant predictor of LifeExpT. The Rsquare value is close to 1 which means the model did very good and we can use it predict LifeExpT. The feature variable and target variables (LifeExpT ~ TotExpT) has a correlation of 0.85 which means they are related. The residual from y axis are scatted randomly. We can see most fitted value has good correlated with a residual.

```
ctrd <- cor(df[, sapply(df, is.numeric)])
corrplot(ctrd
         , method = 'color' # I also like pie and ellipse
         , order = 'hclust' # Orders the variables so that ones that behave similarly are placed next t
         , addCoef.col = 'black'
         , number.cex = .6 # Lower values decrease the size of the numbers in the cells
)
```




```
par(mfrow = c(2, 2), mar = c(2,2,2,2))
plot(linear_lm2)
```




```r
prediction1 <- predict(linear_lm2, newdata = data.frame(TotExpT = 1.5))^(1/4.6)
prediction2 <- predict(linear_lm2, newdata = data.frame(TotExpT = 2.5))^(1/4.6)
cat(
  'Prediction with 1.5: ',
  scales::comma(prediction1),
  '\nPrediction with 2.5: ',
  scales::comma(prediction2),
  sep = ''
)
```

```
## Prediction with 1.5: 63
## Prediction with 2.5: 87
```

```r
set.seed(123)
multiple_lm <- lm(LifeExp ~ PropMD + TotExp + PropMD:TotExp, data= data)
summary(multiple_lm)
```

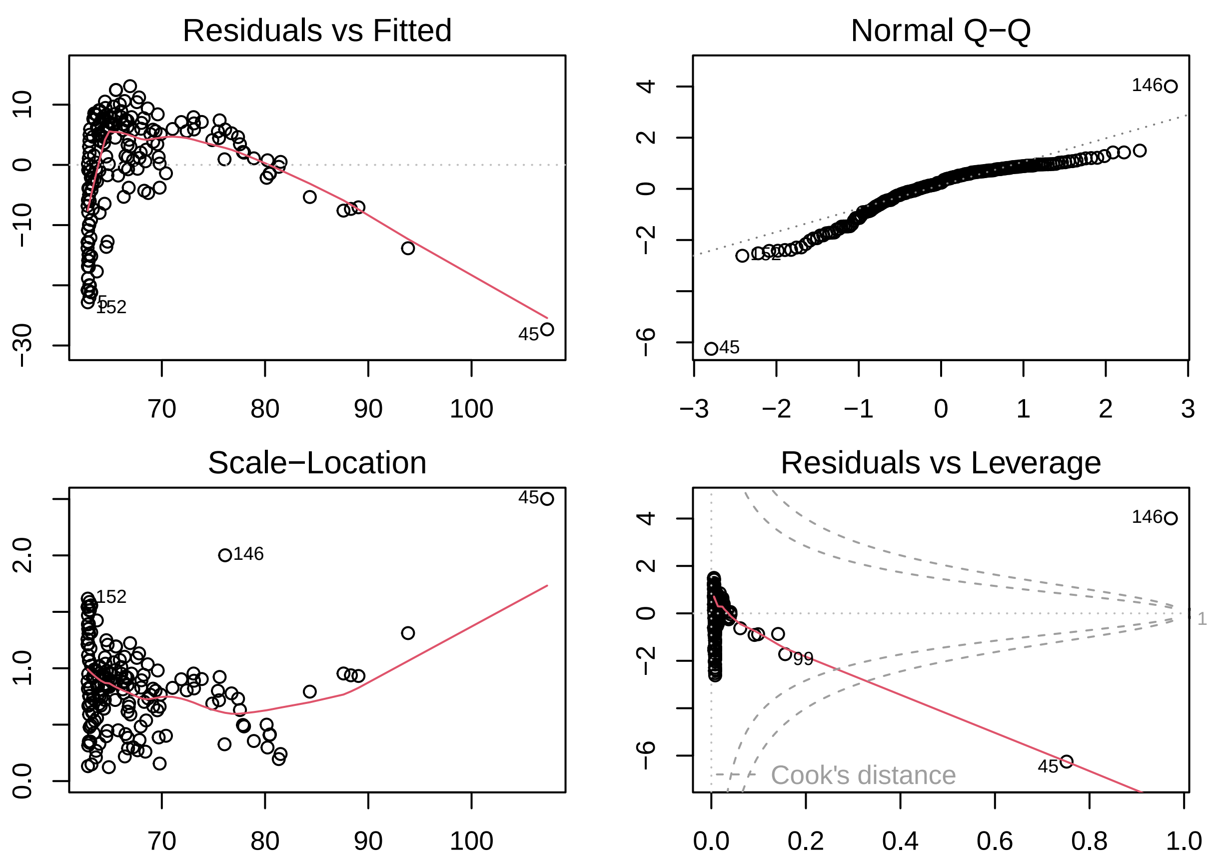
```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMD:TotExp, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
```

```
## 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
```

```r
par(mfrow = c(2, 2), mar = c(2,2,2,2))
plot(multiple_lm)
```

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

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



the Rsquare is 0.3471 which means the model didn't perfom better that the previous model. we need more work to evaluate the model and do hyperprameters tiunning on the features. The residuals are clustered on the y-axis which means the model didn't predict the value corerctly.

```r
newdata = data.frame(PropMD = 0.03, TotExp = 14)
predict(multiple_lm, newdata = newdata)
```

```
##       1
## 107.696
```

Based on the life expextancy recorded in the databased, the presiction seems to be irrelevant because the model didnt do a good job. The predicted life expancy is way higher than the one in the dataset