Eric_Hirsch_605_Assignment_12

Eric Hirsch

11/13/2021

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.5
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.0.5
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Country Comparison Data

We will perform multiple regression analysis on country level data.

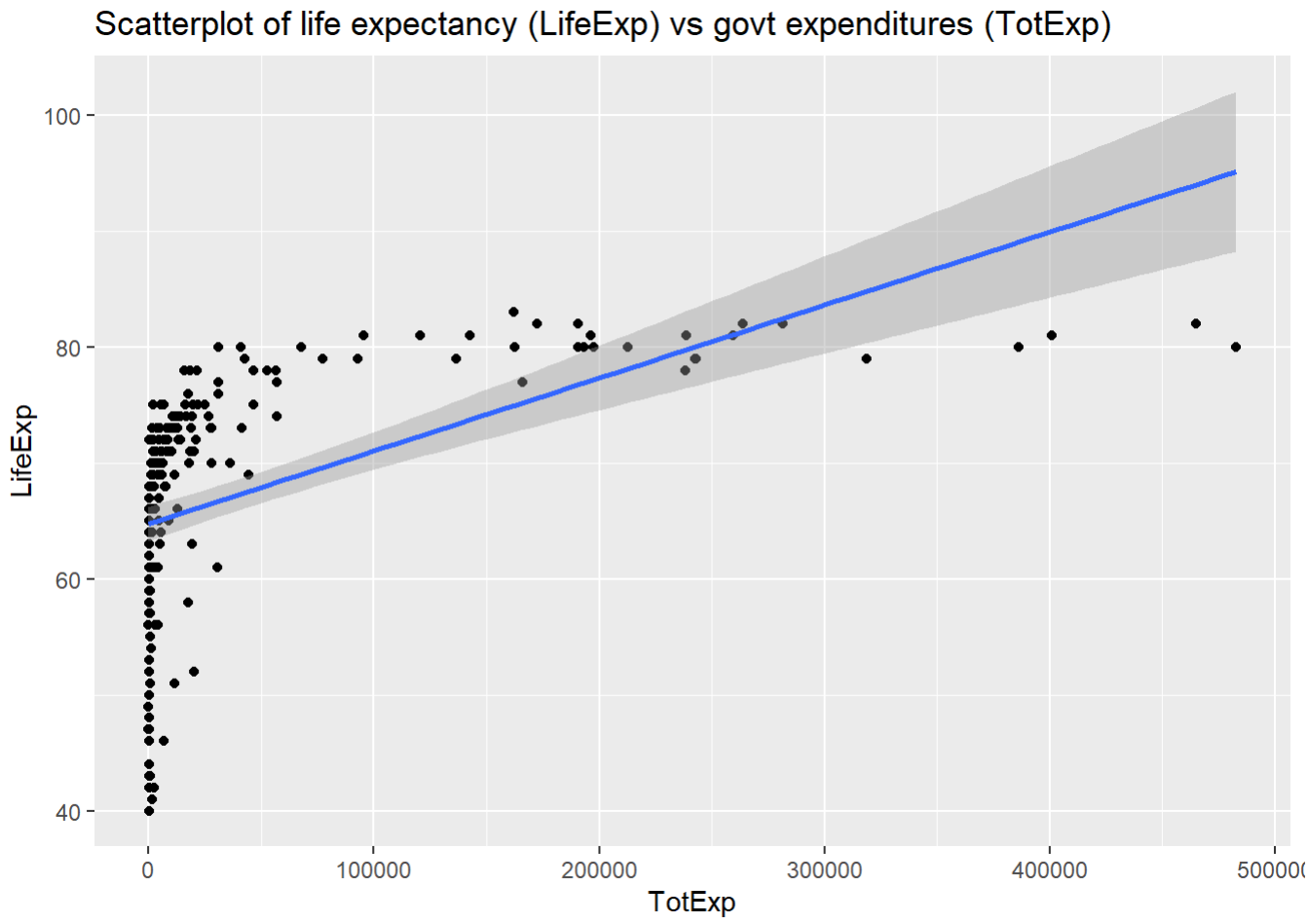
Get the data

```
dfCountry <- read.csv("D:\\RStudio\\CUNY_605\\12\\who.csv", header = TRUE)
```

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.

```
options(scipen = 999)

ggplot(dfCountry, aes(TotExp, LifeExp)) +
  geom_point() +
  stat_smooth(method = "lm") +
  ggtitle("Scatterplot of life expectancy (LifeExp) vs govt expenditures (TotExp)")
```

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




```
m1 <- lm(LifeExp ~ TotExp, data = dfCountry)
summary(m1)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = dfCountry)
##
## 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) 64.753374534  0.753536611  85.933 < 0.0000000000000002 ***
## TotExp       0.000062970  0.000007795   8.079   0.0000000000000771 ***
## ---
## 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: 0.00000000000007714
```

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

- F statistic: The F statistic tells you something when there is more than one independent variable, so here it has little meaning.
- R^2: 25% of the variation in life expectancy is explained by govt. spending.
- p-values: the p values are near 0, suggesting that our independent variable is significant.

Here are our assumptions:

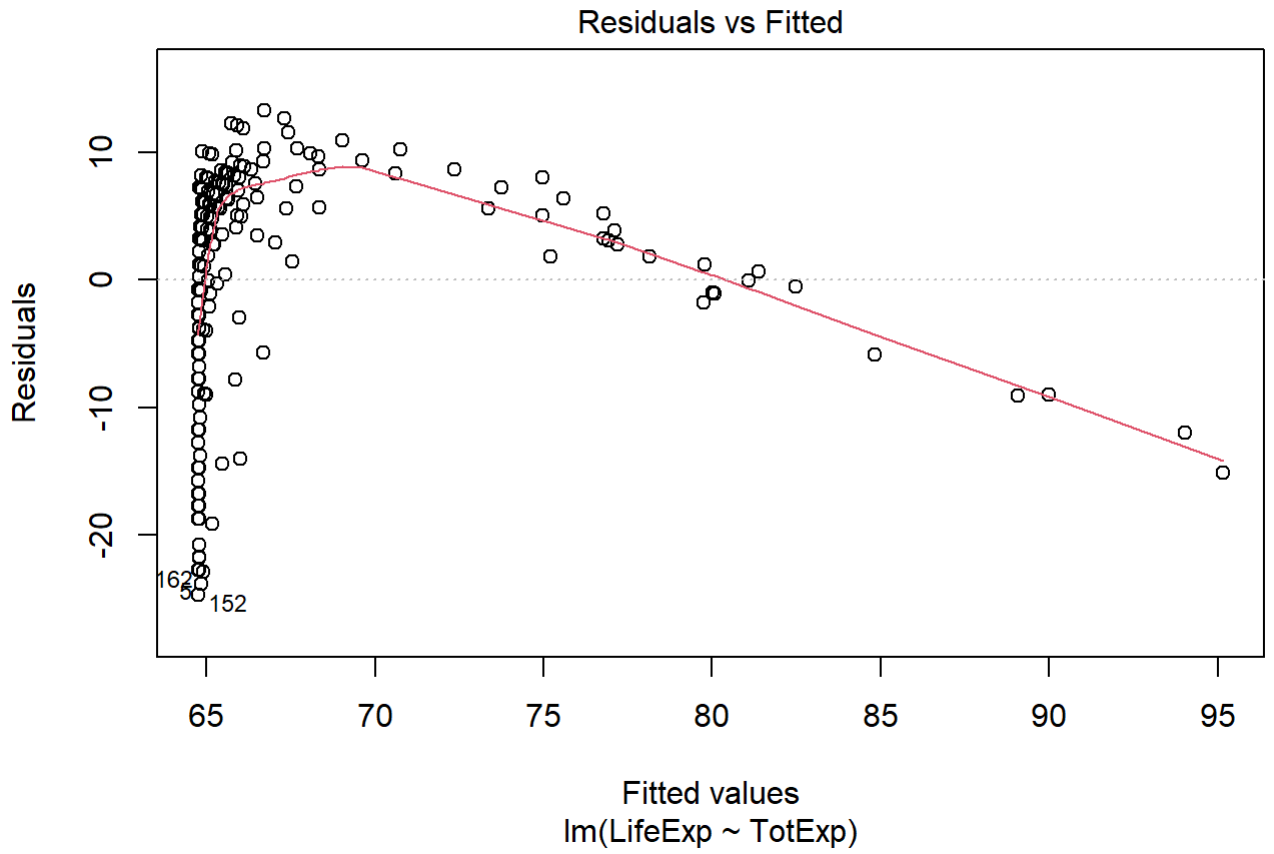
1. Linear relationship:
No, the relationship is clearly strong but not linear. At low levels of expenditure, life expectancy varies widely. At higher levels, life expectancy levels out around 80 and stays there. The regression analysis apparently picks up the correlation between the extremes.

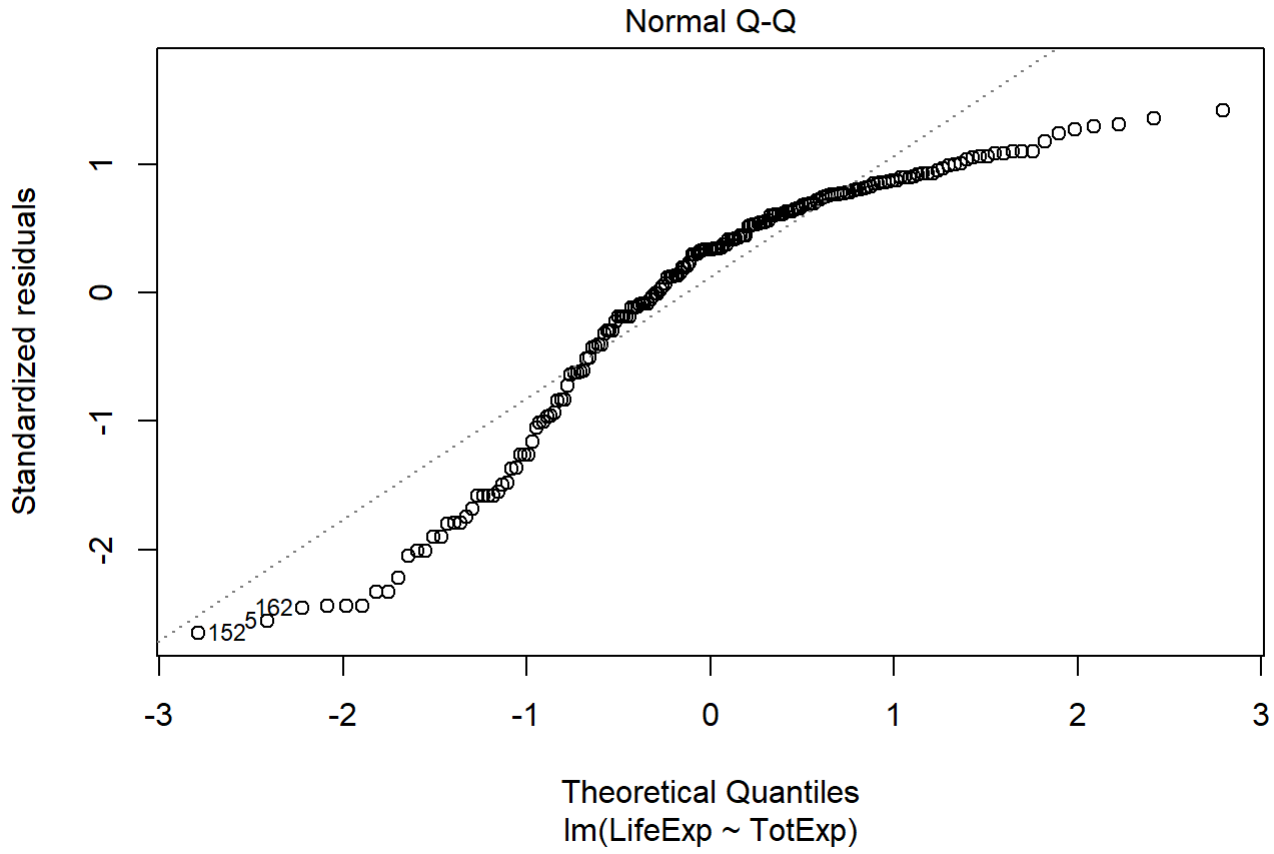
2. Independence:

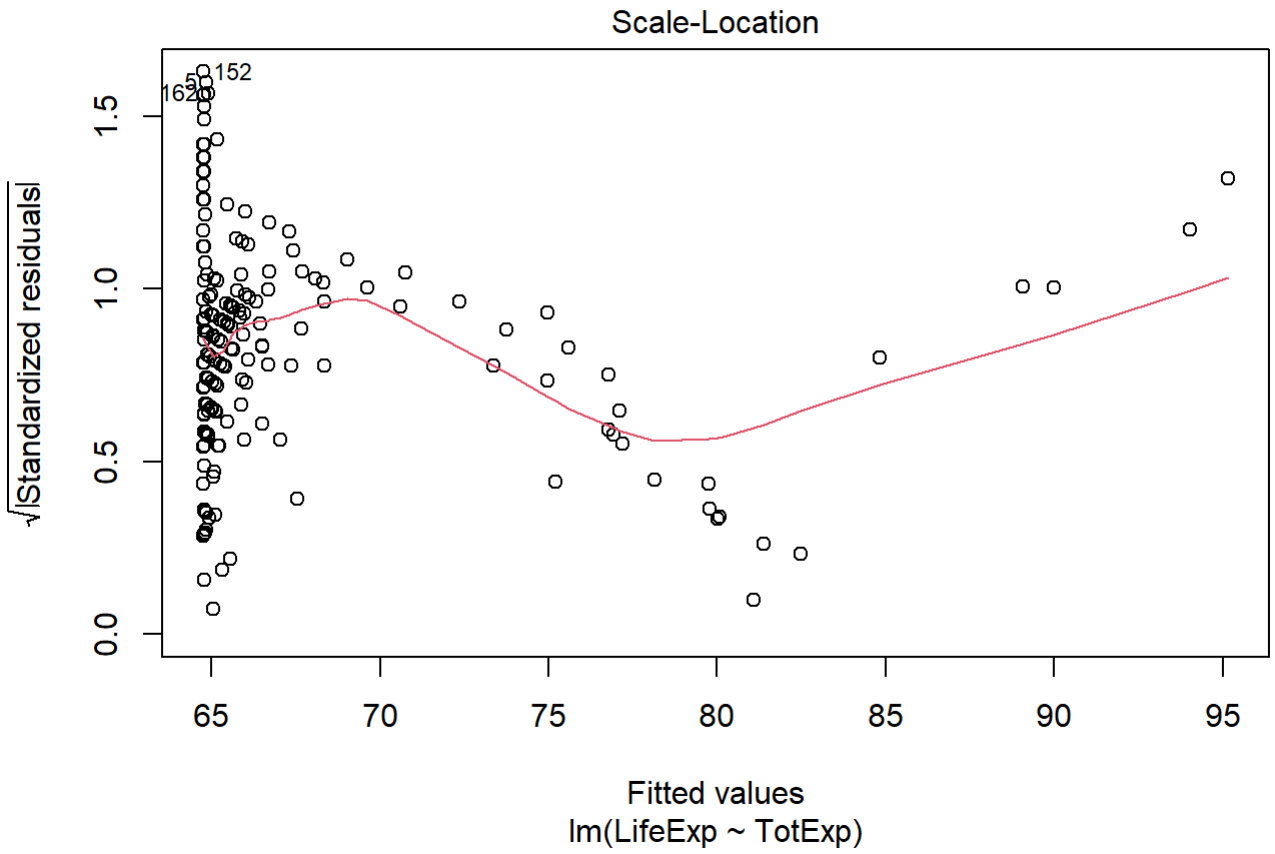
We have no reason to expect otherwise.

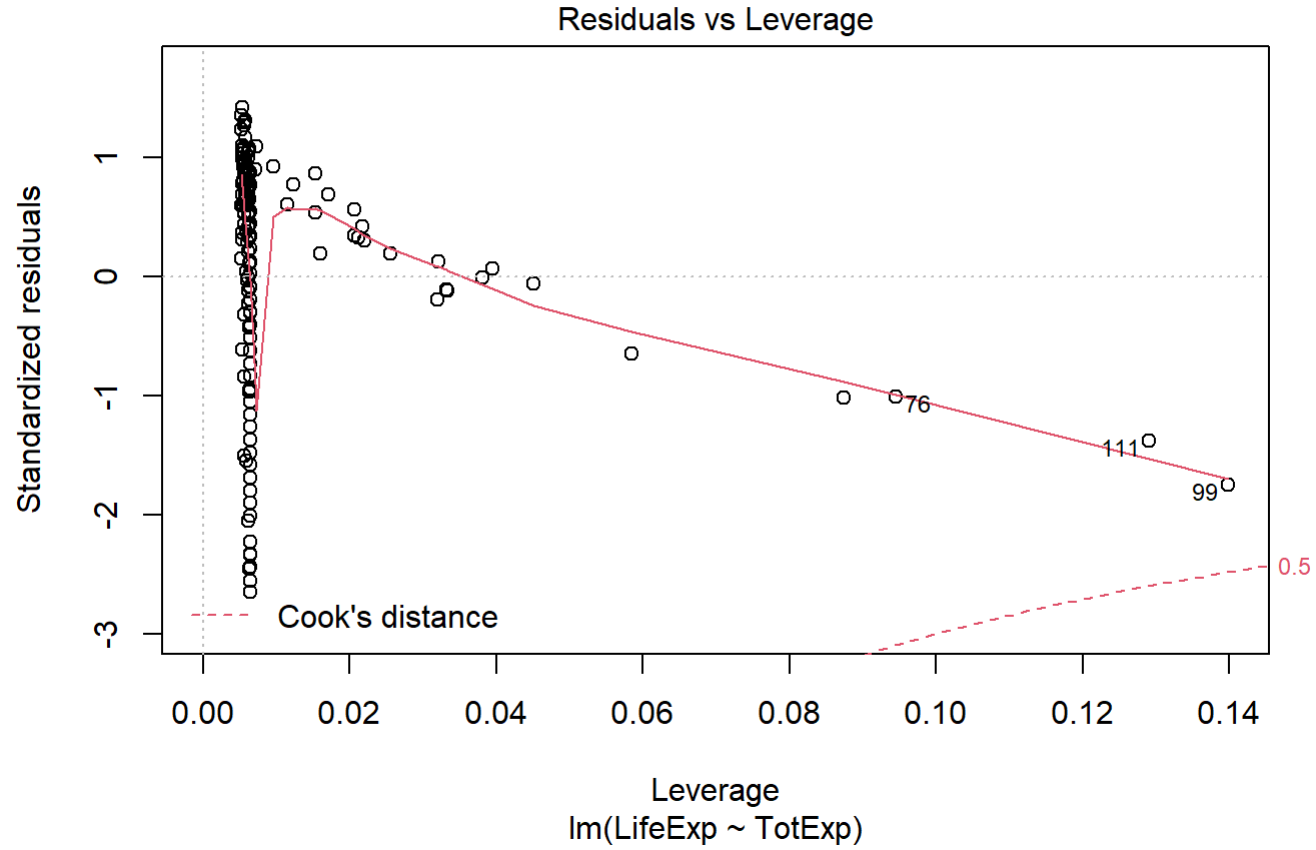
3. Homoscedasticity:

```
plot(m1)
```


Residuals vs Fitted
Residuals
10
0
-10
-20
162
5
152
65
70
75
80
85
90
95
Fitted values
lm(LifeExp ~ TotExp)


Normal Q-Q
Standardized residuals
1
0
-1
-2
162
5
152
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(LifeExp ~ TotExp)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
65
70
75
80
85
90
95
Fitted values
lm(LifeExp ~ TotExp)


Residuals vs Leverage
Standardized residuals
-3
-2
-1
0
1
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Leverage
lm(LifeExp ~ TotExp)
Cook's distance
0.5

```
bptest(m1)
```

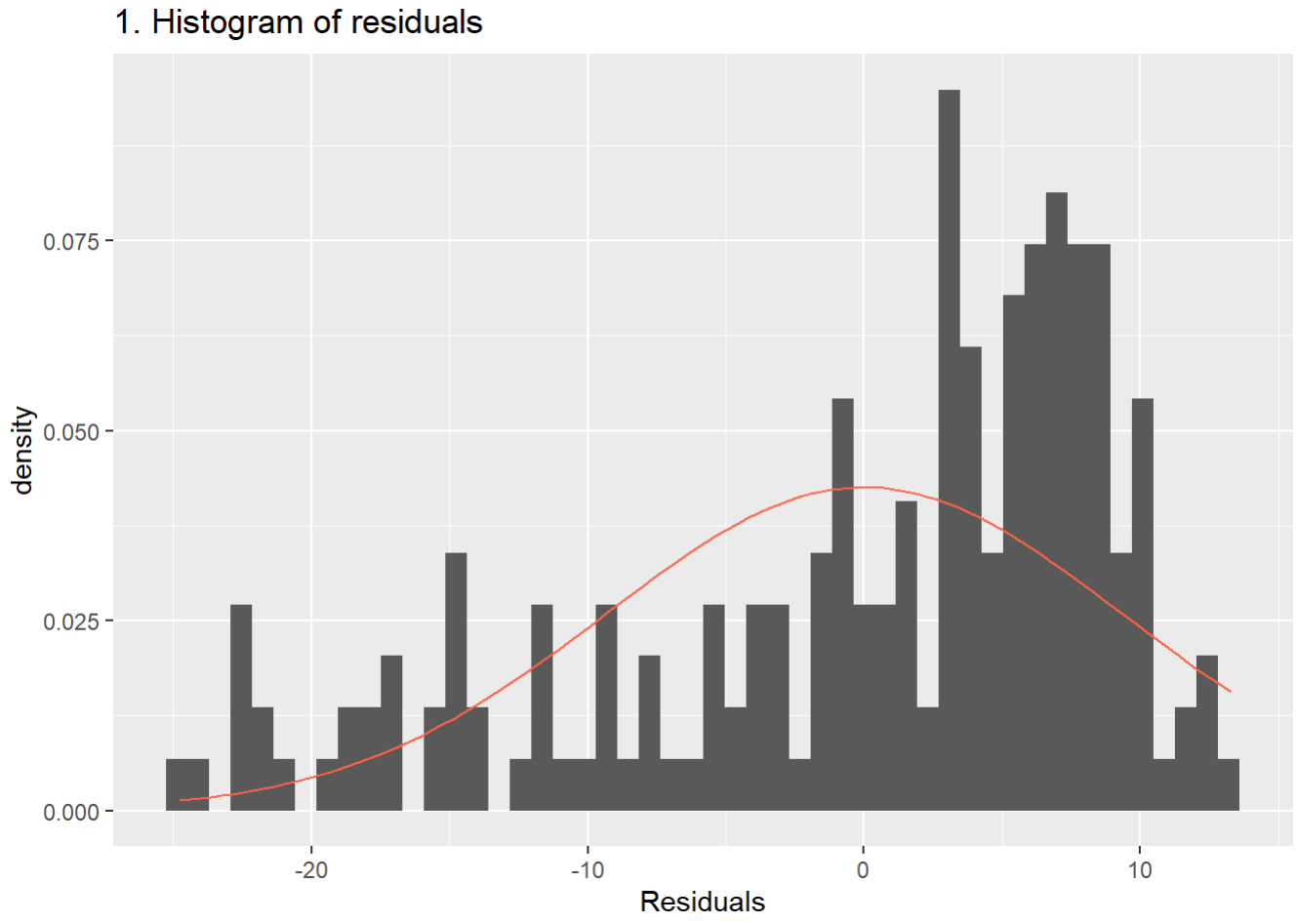
```
##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 2.6239, df = 1, p-value = 0.1053
```

The variance is very high at low values, despite the bp test. We also see that the residuals are not evenly distributed around 0 - they have a distinctly non-normal distribution as the residuals of the fitted values rise and then plunge.

This pattern is already evident in the original scatterplot.

4. Normality:

```
dmean <- 0
dse <- summary(m1)$sigma

ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(aes(y = ..density..), bins = 50) +
  xlab("Residuals") +
  ggtitle("1. Histogram of residuals") +
  stat_function(fun = dnorm, args = c(mean = dmean, sd = dse), col = "tomato")
```

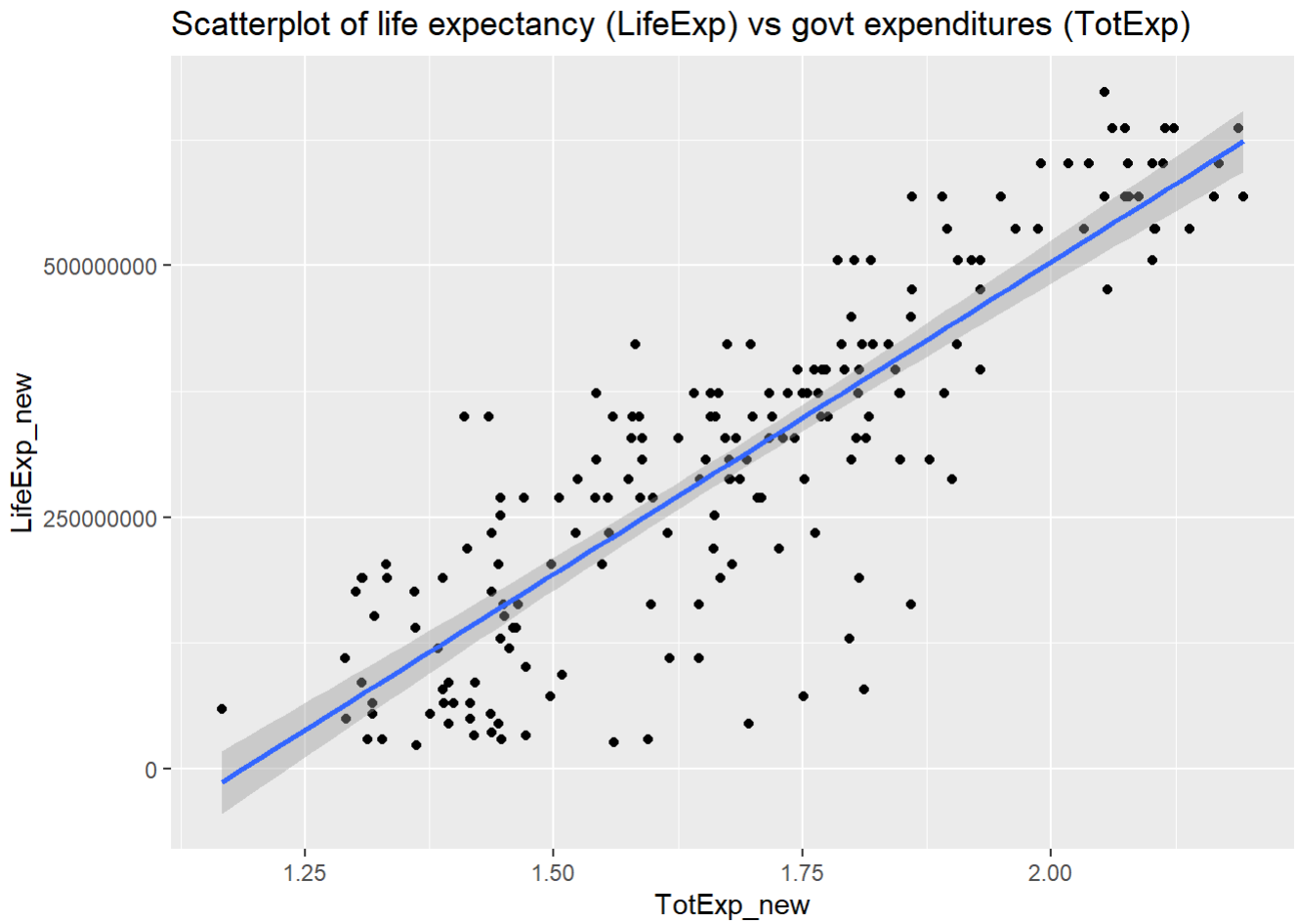


The qq-plot and histogram show that the residuals are not normally distributed but tend to have a number of small underpredictions and a smaller but larger number of overpredicitons.

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?"

```
options(scipen = 999)

dfCountry1 <- dfCountry %>%
  mutate(LifeExp_new = LifeExp^4.6) %>%
  mutate(TotExp_new = TotExp^.06)

ggplot(dfCountry1, aes(TotExp_new, LifeExp_new)) +
  geom_point() +
  stat_smooth(method = "lm") +
  ggtitle("Scatterplot of life expectancy (LifeExp) vs govt expenditures (TotExp)")
```

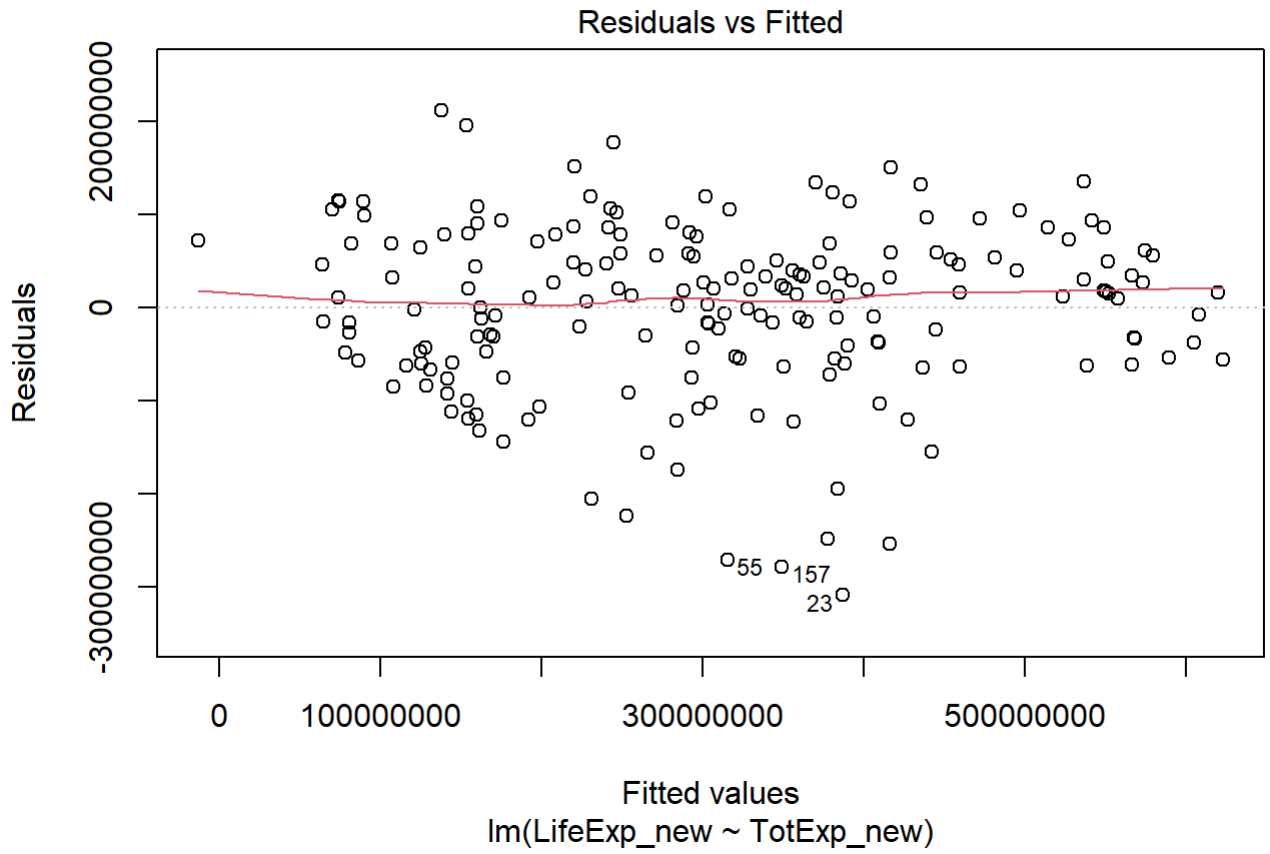
```
## `geom_smooth()` using formula 'y ~ x'
```

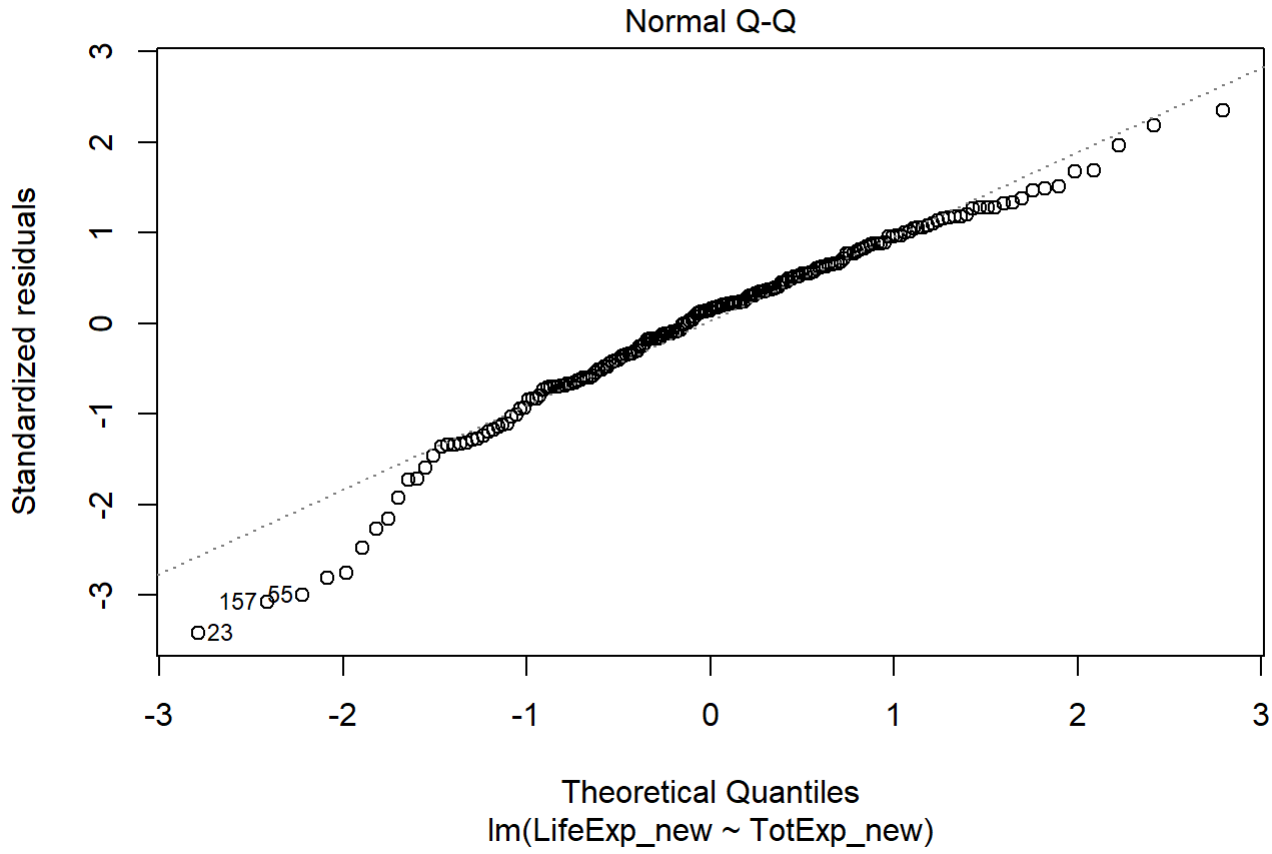


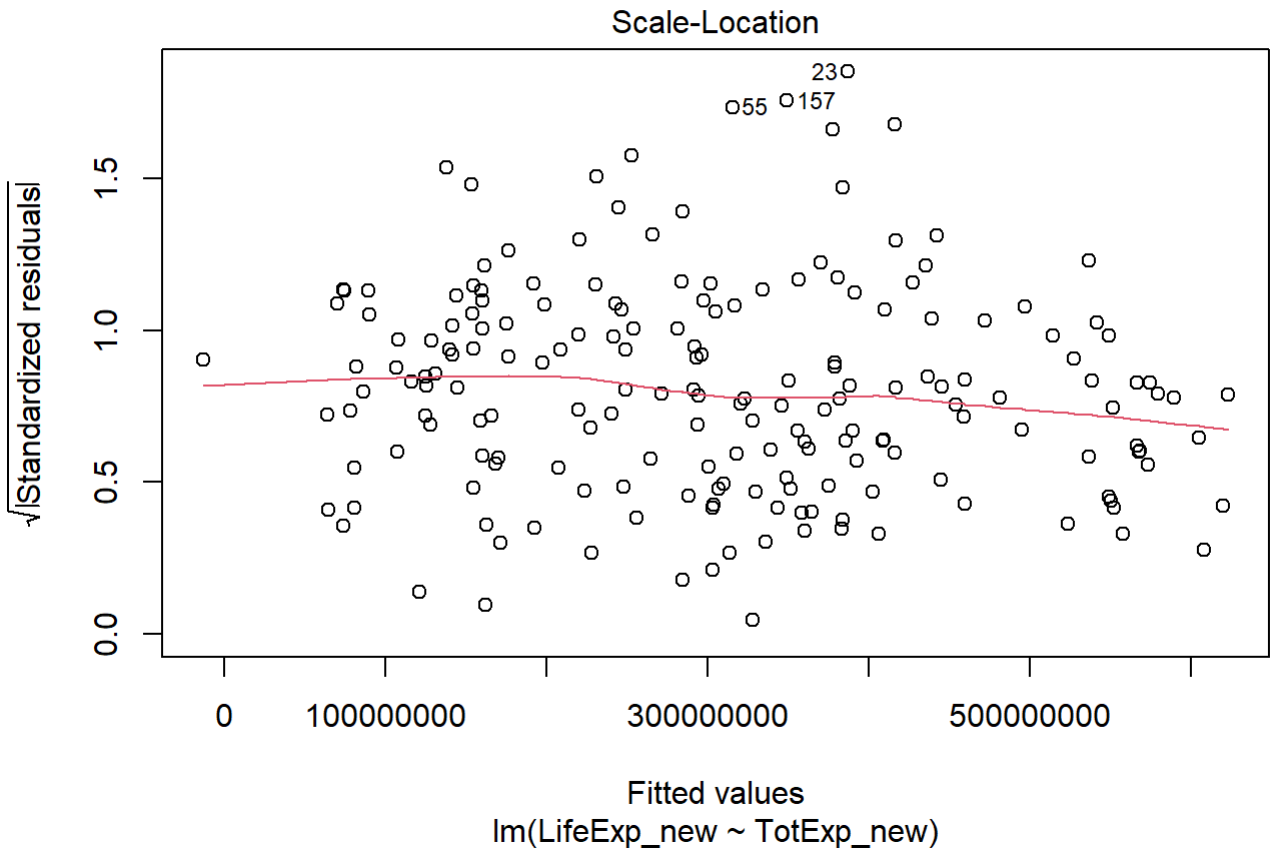

```
m2 <- lm(LifeExp_new ~ TotExp_new, data = dfCountry1)
summary(m2)
```

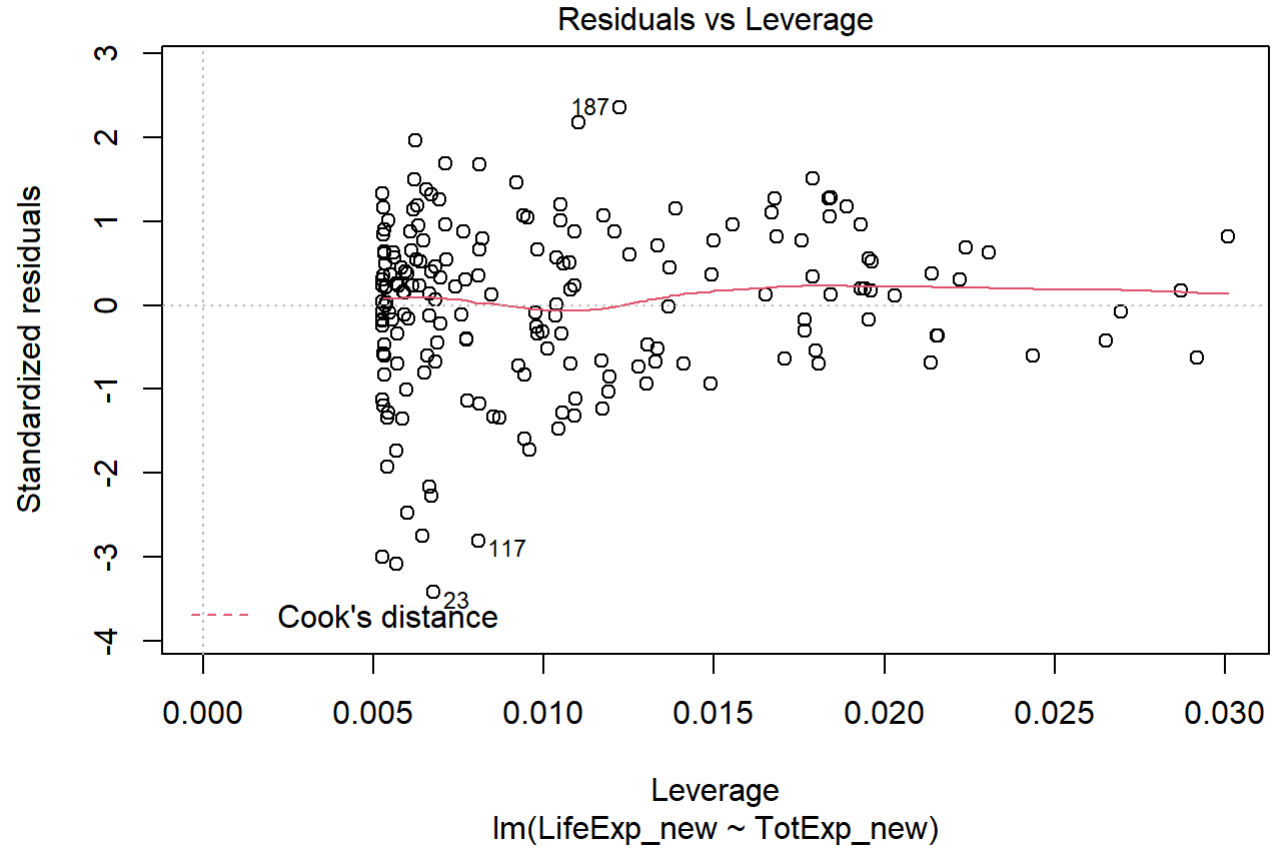
```
##
## Call:
## lm(formula = LifeExp_new ~ TotExp_new, data = dfCountry1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -736527910   46817945  -15.73 <0.0000000000000002 ***
## TotExp_new   620060216   27518940   22.53 <0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

```
plot(m2)
```


Residuals vs Fitted
Residuals
200000000
0
-300000000
55
157
23
0
100000000
300000000
500000000
Fitted values
lm(LifeExp_new ~ TotExp_new)


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
157
55
23
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(LifeExp_new ~ TotExp_new)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
23
55
157
0
100000000
300000000
500000000
Fitted values
lm(LifeExp_new ~ TotExp_new)


Residuals vs Leverage
Standardized residuals
-4
-3
-2
-1
0
1
2
3
187
117
23
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(LifeExp_new ~ TotExp_new)

```
bptest(m2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m2
## BP = 0.28802, df = 1, p-value = 0.5915
```

- F statistic: The F statistic tells you something when there is more than one independent variable, so here it has little meaning.
- R^2: 72% of the variation in life expectancy is explained by govt. spending.
- p-values: the p values are near 0, suggesting that our independent variable is significant.

The model does a much better job of predicting the dependent variable, and the transformation brings the model much more into alignment with the necessary assumptions for running a regression. The model still tends to skew a bit and there is more variability in the middle of the plot.

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

```
(-736527910 + 1.5*620060216)^(1/4.6)
```

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

```
(-736527910 + 2.5*620060216)^(1/4.6)
```

```
## [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?

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

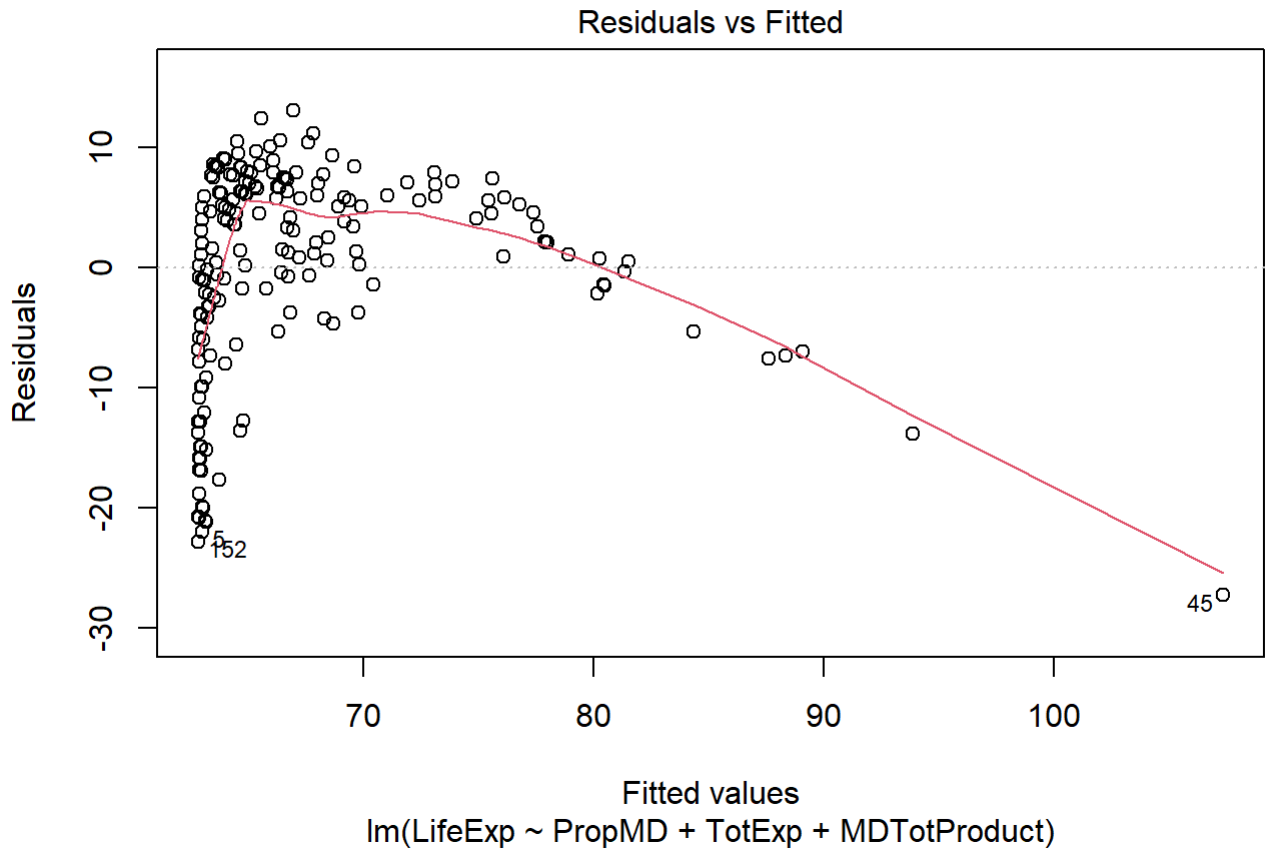
```
dfCountry2 <- dfCountry %>%
  mutate(MDTotProduct = PropMD*TotExp)

m3 <- lm(LifeExp ~ PropMD + TotExp + MDTotProduct, data = dfCountry2)
summary(m3)
```

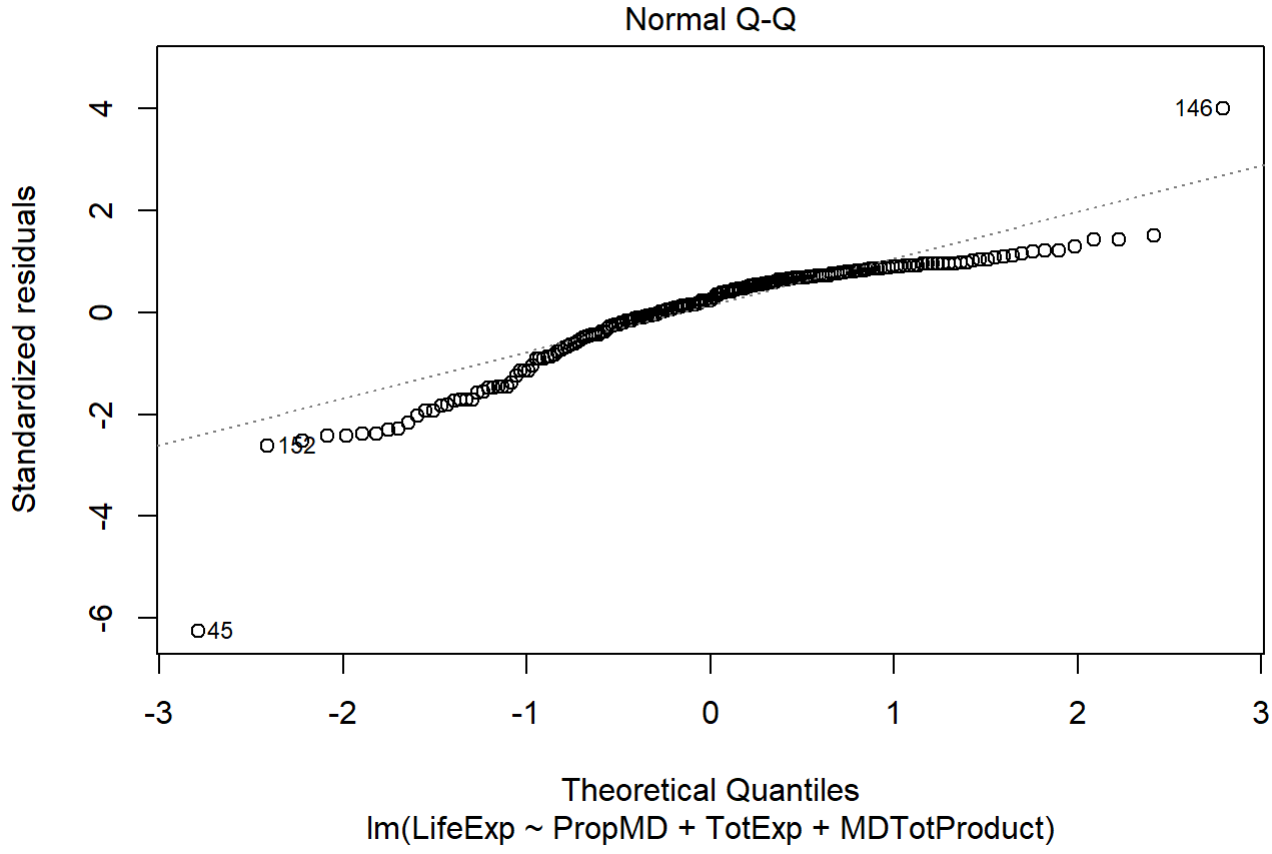
```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + MDTotProduct, data = dfCountry2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                    Estimate     Std. Error t value             Pr(>|t|)
## (Intercept)    62.772703255    0.795605238  78.899 < 0.0000000000000002 ***
## PropMD       1497.493952519  278.816879652   5.371   0.0000002320602774 ***
## TotExp          0.000072333    0.000008982   8.053   0.0000000000000939 ***
## MDTotProduct   -0.006025686    0.001472357  -4.093   0.0000635273294941 ***
## ---
## 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: < 0.00000000000000022
```

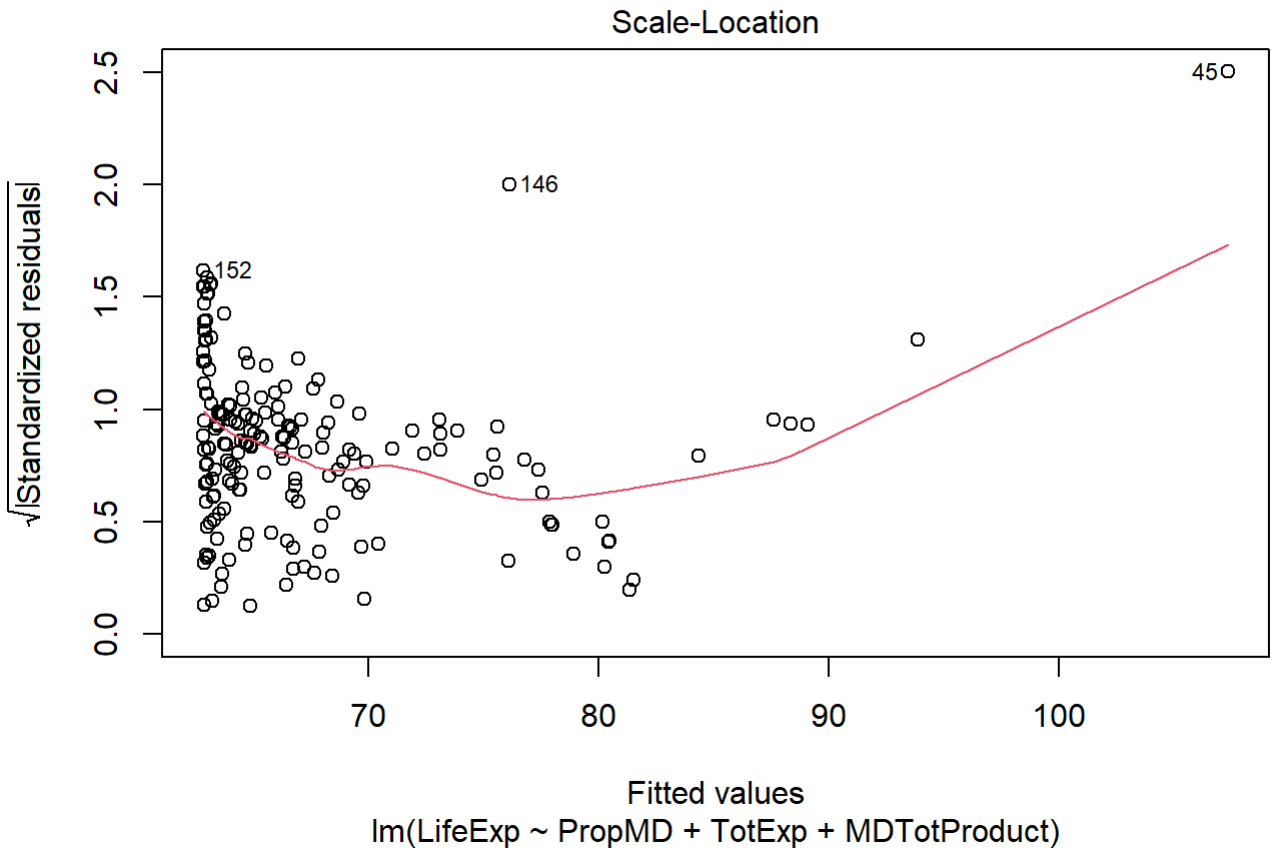
```
m4 <- lm(LifeExp ~ PropMD, data = dfCountry2)
summary(m4)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD, data = dfCountry2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.450  -5.347   3.004   7.065  15.274
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)    65.42       0.82  79.774 < 0.0000000000000002 ***
## PropMD       1092.15     203.03   5.379          0.000000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 188 degrees of freedom
## Multiple R-squared:  0.1334, Adjusted R-squared:  0.1288
## F-statistic: 28.94 on 1 and 188 DF,  p-value: 0.0000002206
```

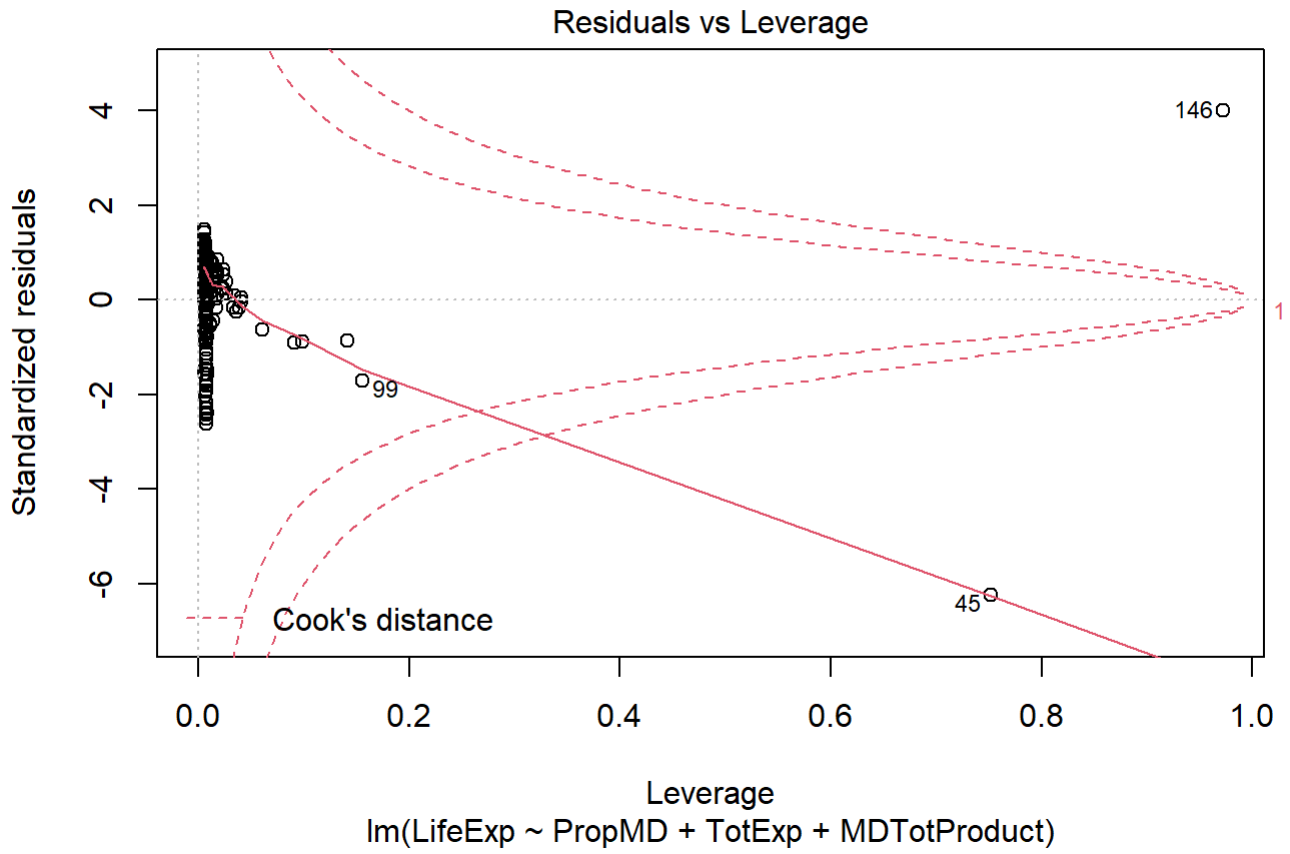
```
plot(m3)
```


Residuals vs Fitted
Residuals
10
0
-10
-20
-30
5
152
45
70
80
90
100
Fitted values
lm(LifeExp ~ PropMD + TotExp + MDTotProduct)


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
-6
146
152
45
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(LifeExp ~ PropMD + TotExp + MDTotProduct)




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

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




```
bptest(m3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m3
## BP = 12.005, df = 3, p-value = 0.007366
```

- F statistic: The F statistic and p values are significant.
- R^2: 35% of the variation in life expectancy is explained by the independent variables.
- coefficient p-values: the p values are near 0, suggesting that our independent variables are significant.

The model has a higher R2 than the original model. However, the issues with normality and heteroskedasticity remain. The residuals show that the model is still clearly not linear. What's more, the coefficients are very, very small for the TotEXp variables and won't affect the result much.

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

```
62.772703255 + .03*1497.493952519  + 14*0.000072333 + .03*14*-0.006025686
```

```
## [1] 107.696
```

107 years old is unreasonable, but so is a propMD of .03. More reasonable is .003:

```
62.772703255 + .003*1497.493952519  + 14*0.000072333 + .003*14*-0.006025686
```

```
## [1] 67.26594
```

67 years old is reasonable.