DATA 605 Assignment Week 12

CUNY Spring 2021

Philip Tanofsky

24 April 2021

Prompt

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.

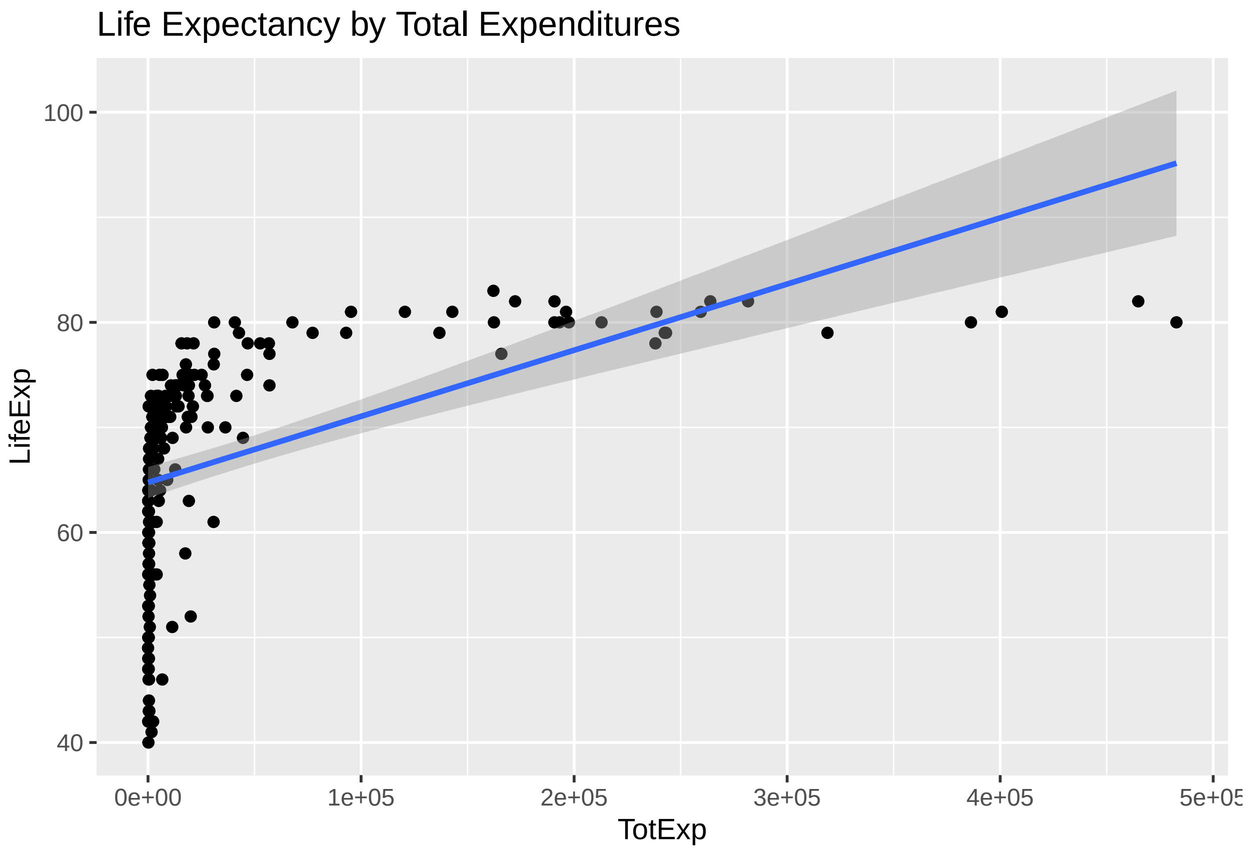
```r
# load libraries
library(tidyverse)

# Read in data
who <- read_csv("who.csv")

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
```

Part 1

Provide a scatterplot of LifeExp~TotExp, and run simple linear regression. Do not transform the variables. Provide and interpret the F statistics, Rˆ2, standard error,and p-values only. Discuss whether the assumptions of simple linear regression met.

```
# Scatterplot
who %>%
  ggplot(aes(x=TotExp, y=LifeExp)) +
  geom_point() +
  labs(title = 'Life Expectancy by Total Expenditures') + geom_smooth(method='lm', formula= y~x)
```



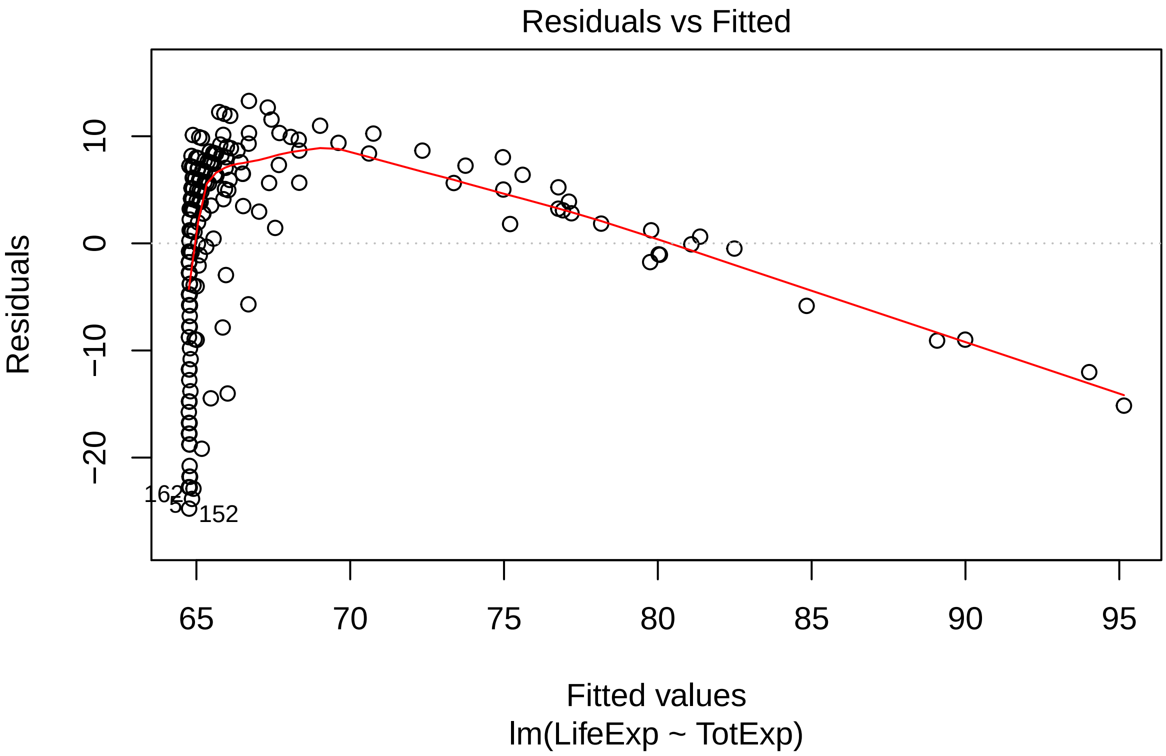
```
# Simple linear regression
simple_lm <- lm(LifeExp ~ TotExp, data=who)

summary(simple_lm)
```

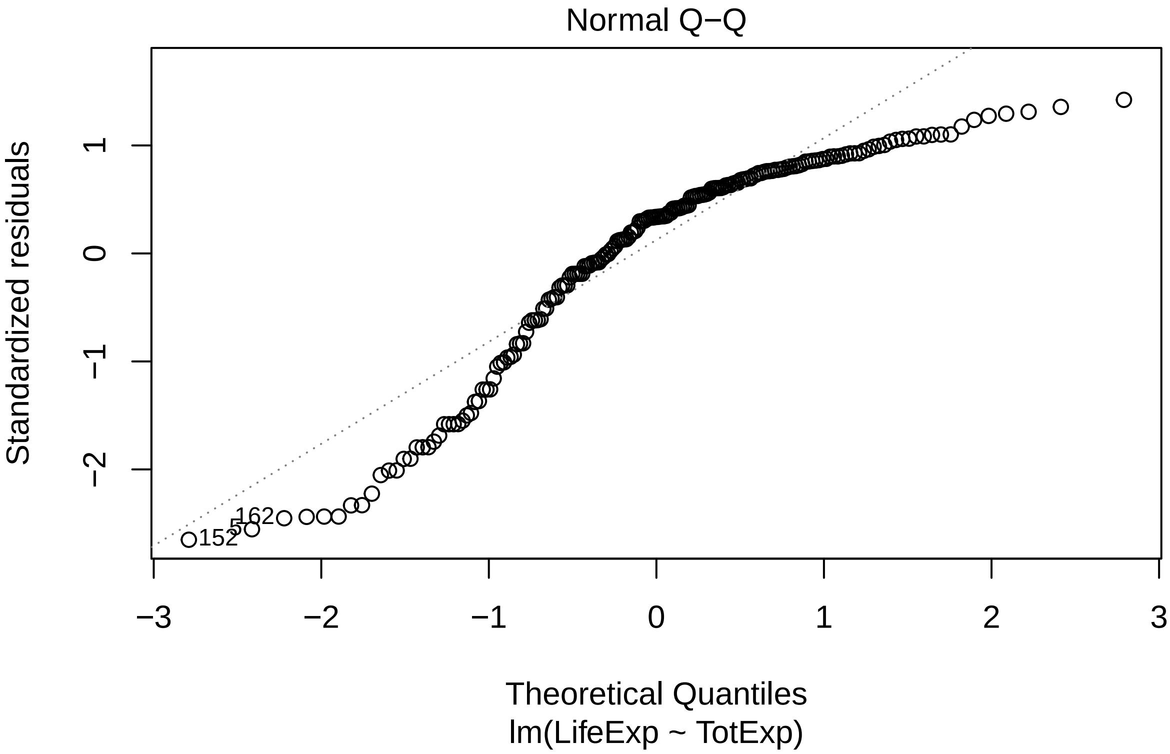
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = who)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
## TotExp      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
```

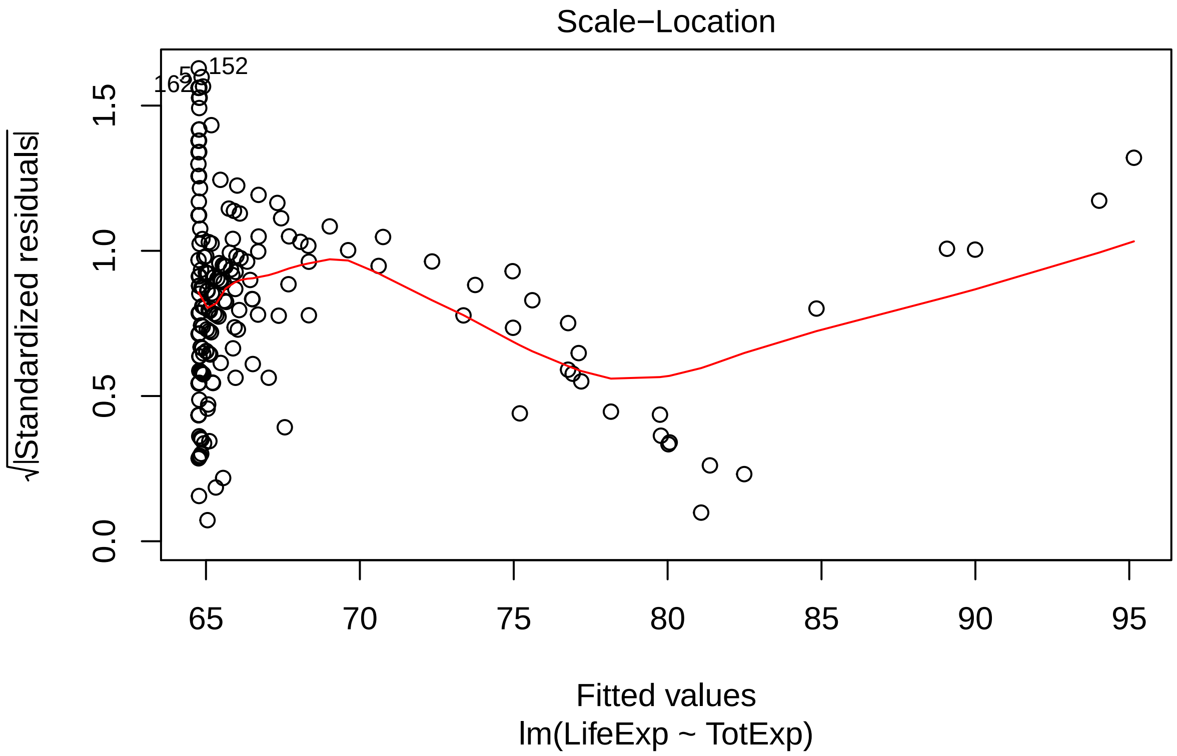
- F statistic: Not useful given this is a single parameter model
- Rˆ2: Multiple R-squared is 0.2577, indicating the model explains 25.77% of the data's variation … not good. Adjusted R-squared not important because of the single variable regression model.
- Standard error: Using the t-values to interpret the standard errors, the standard error for the intercept is quite small indicating little variability in the intercept value. The t-value for Total Expenditure is over 8, which according to the textbook is good. But, as I've plotted the Residuals vs Fitted, I know these values are misleading.
- p-values: Both p-values are quite small, indicating the probability of the coefficients as not relevant to the model as quite low. As mentioned in Standard Error, according to the textbook, these values represent a good model, yet the model plots and R-squared value indicate otherwise.

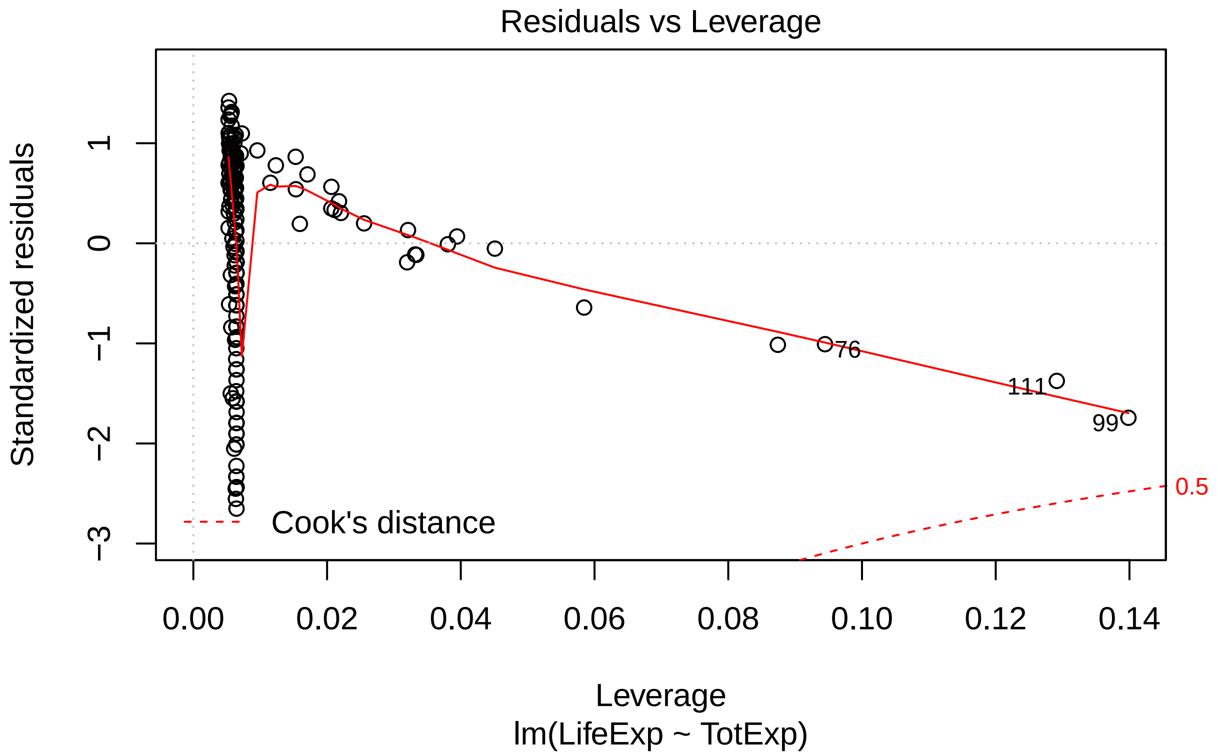
Based on the plots: Assumptions of linear regression not met.

```
plot(simple_lm)
```


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)

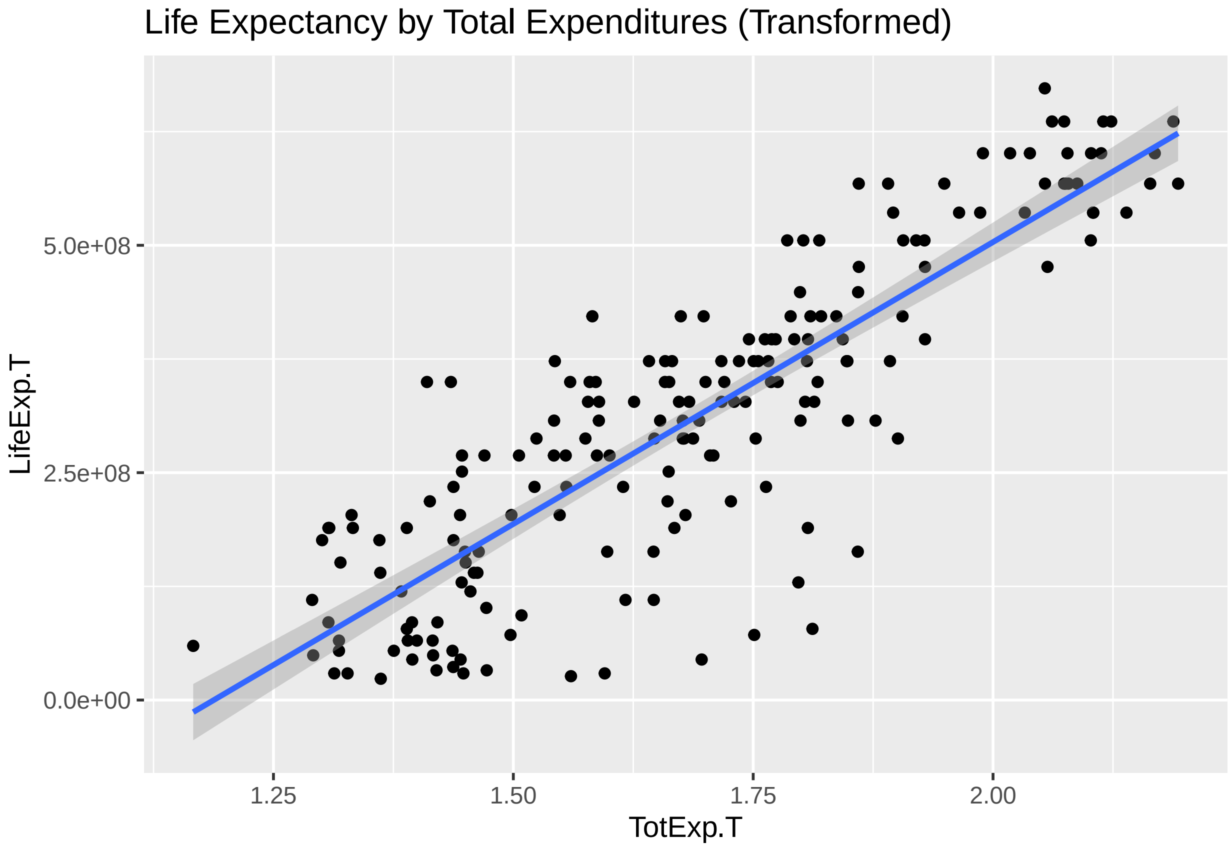

Scale−Location
162
5
152
1.5
1.0
0.5
0.0
√|Standardized residuals|
65
70
75
80
85
90
95
Fitted values
lm(LifeExp ~ TotExp)



Part 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?"

```
who$LifeExp.T <- who$LifeExp^4.6
who$TotExp.T <- who$TotExp^.06

who %>%
  ggplot(aes(x=TotExp.T, y=LifeExp.T)) +
  geom_point() +
  labs(title = 'Life Expectancy by Total Expenditures (Transformed)') + geom_smooth(method='lm', formula
```



```
simple_t_lm <- lm(LifeExp.T ~ TotExp.T, data=who)

summary(simple_t_lm)
```

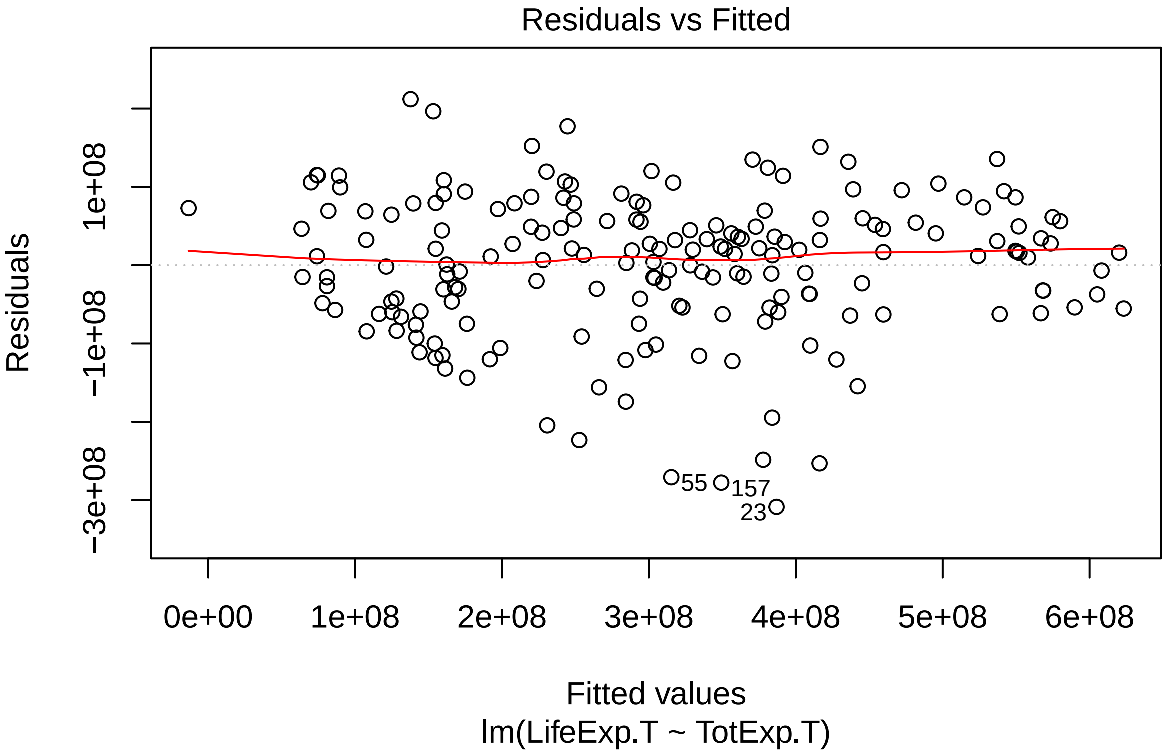
```
##
## Call:
## lm(formula = LifeExp.T ~ TotExp.T, data = who)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -736527910  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

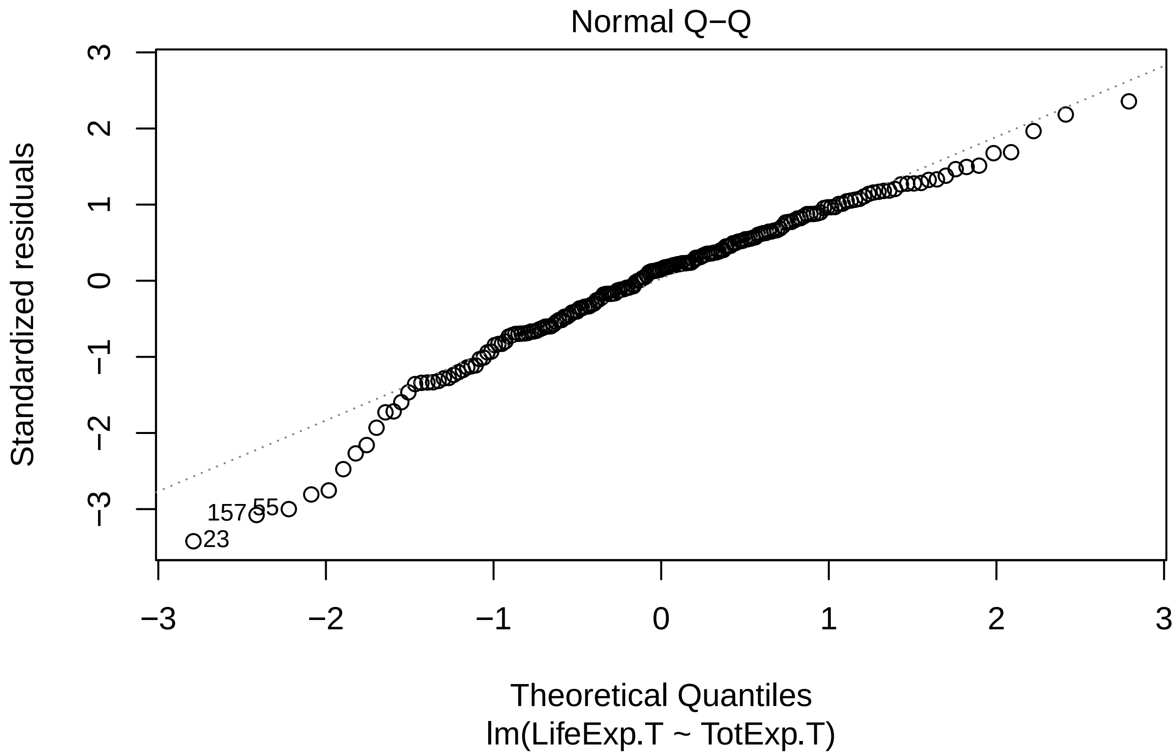
- F statistic: Again, not useful given this is a single parameter model

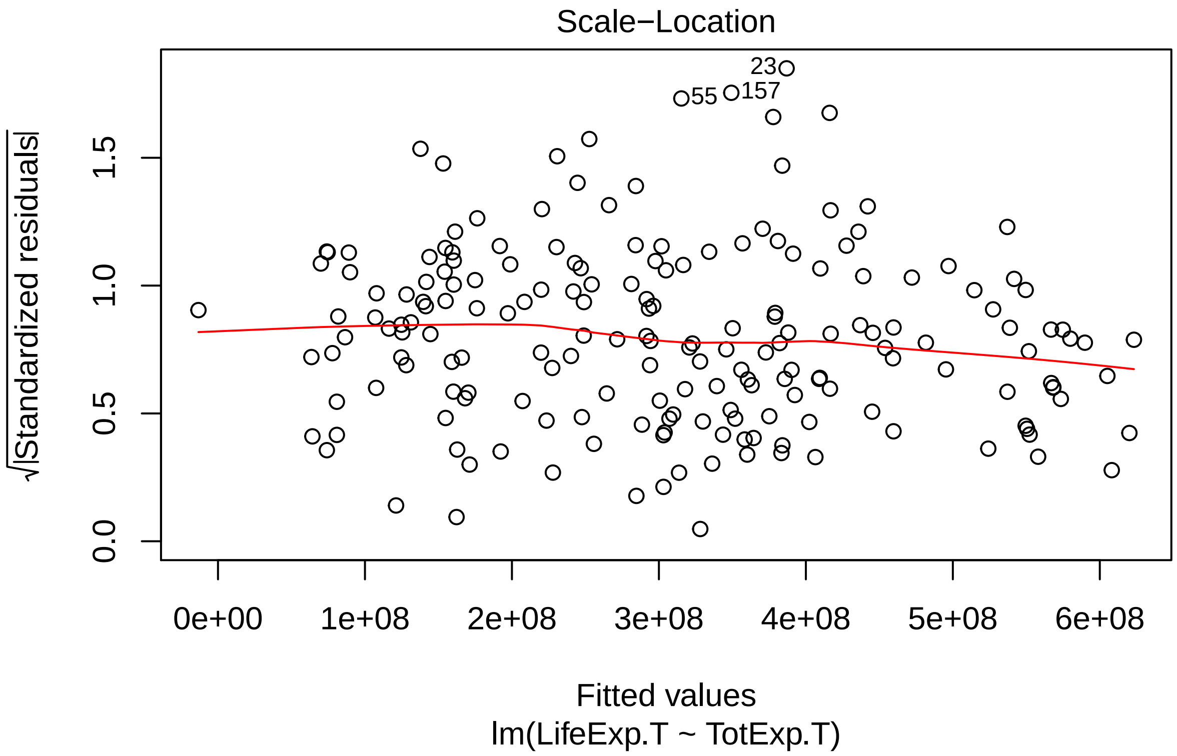
- Rˆ2: Multiple R-squared is 0.7298, indicating the model explains 72.98% of the data's variation … quite an improvement compared to the first model. Adjusted R-squared not important because of the single variable regression model.
- Standard error: Using the t-values to interpret the standard errors, the standard error for the intercept shows a ratio greater than 15 when comparing the standard error the coefficient, indicating little variability in the intercept value. The t-value for Total Expenditure is over 22, which according to the textbook is good. Both standard error values as compared to the coefficient values are indicative of a good model because both denote little variability in the coefficient values.
- p-values: The p-value for the intercept and Total Expenditure coefficients are very small which means the chance the coefficient value is not relevant to the model is very, very infrequent. For both values, good sign the model is valid for the dataset.

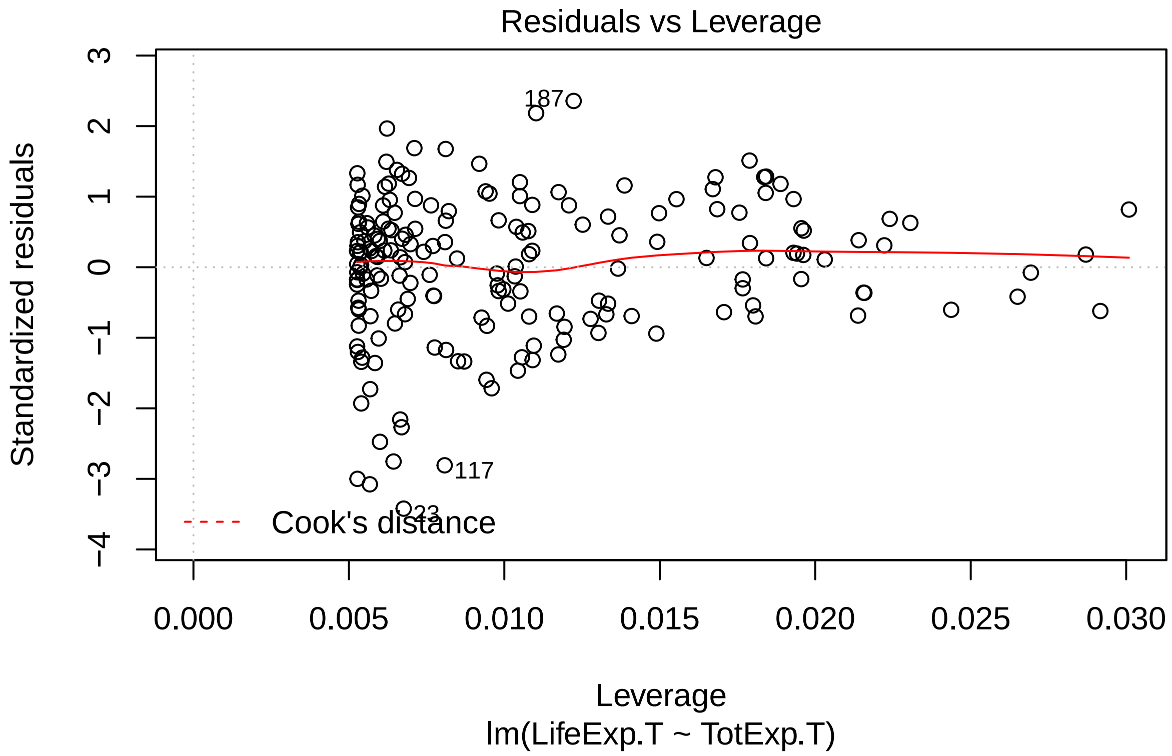
The model with transformed values (second model) is better. Scatterplot above and Residuals vs Fitted plot below confirm this assessment.

```
plot(simple_t_lm)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
157
55
23
Theoretical Quantiles
lm(LifeExp.T ~ TotExp.T)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0e+00
1e+08
2e+08
3e+08
4e+08
5e+08
6e+08
Fitted values
lm(LifeExp.T ~ TotExp.T)
23
55
157



Part 3

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

```
result_3.1 <- simple_t_lm$coefficients[[1]] + (1.5 * simple_t_lm$coefficients[[2]])

result_3.1 <- result_3.1^(1/4.6)

result_3.1
```

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

Life expectancy when TotExp^.06=1.5 is 63.3115334.

```
result_3.2 <- simple_t_lm$coefficients[[1]] + (2.5 * simple_t_lm$coefficients[[2]])

result_3.2 <- result_3.2^(1/4.6)

result_3.2
```

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

Life expectancy when TotExp^.06=2.5 is 86.5064485.

Part 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

```r
mult_lm <- lm(LifeExp ~ PropMD + TotExp + (PropMD * TotExp), data=who)

summary(mult_lm)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + (PropMD * TotExp), 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
```

- F statistic: The model results in a value of 34.49 on 3 and 186 degrees of freedom. According to the textbook, "the F-test compares the current model to a model with one fewer predictor. If the current model is better than the reduced model, the p-value will be small." Given the p-value is quite small, this model would be better than a reduced model.
- R^2: The multiple R-squared value is 0.3574, and the adjusted R-squared is 0.3471. In assessing the model, these two values are not indicative of a good model. These results show the model accounts for only 35% of the variation in the data. As points of comparison to the first two models, this multiple regression model performs better than the first simple linear regression model, but much worse than the transformed simple linear regression model.
- Standard error: Interpreting the coefficient standard error values through the t-values, the intercept appears to denote a good model. Seeing as the other t-values have absolute values of less than 10, the coefficients for the three predictor variables may not represent a good model.
- p-values: The p-values for all 4 coefficients are quite small, typically a sign of a good model.

How good is the model? Based on the R-squared values, I would not recommend this model.

Part 5

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

```
result_5.1 <- mult_lm$coefficients[[1]] + (0.03 * mult_lm$coefficients[[2]]) + (14 * mult_lm$coefficien

result_5.1
```

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

Life Expectancy when PropMD=.03 and TotExp=14 is 107.6960037. This forecast does not seem realistic, a Life Expectancy of 107 is literally off the charts, as the maximum value from the initial dataset is 83. Also, trying to put the values into context given the raw data makes the combination unlikely. The PropMD of .03 indicates 3% of the population is medical doctors, which would be one of the higher values compared to the provided instances. The Total Expenditure of 14 is quite low compared to the original dataset as the minimum is 13 and the mean is 41696. These values indicate a country with one of the highest proportions of medical doctors and yet one of the poorest countries in personal and government expenditures. I would imagine, even with a relatively large proportion of doctors, the facilities for these doctors to treat patients would dictate a larger expenditure value, or else, these numbers indicate many doctors working in extremely poor conditions/facilities. Also, given the makeup of the regression model, the PropMD variable has quite a bit of influence on this result given the small value of TotExp … another reason the result is not realistic.