DATA 605 - Homework #12

Zach Alexander

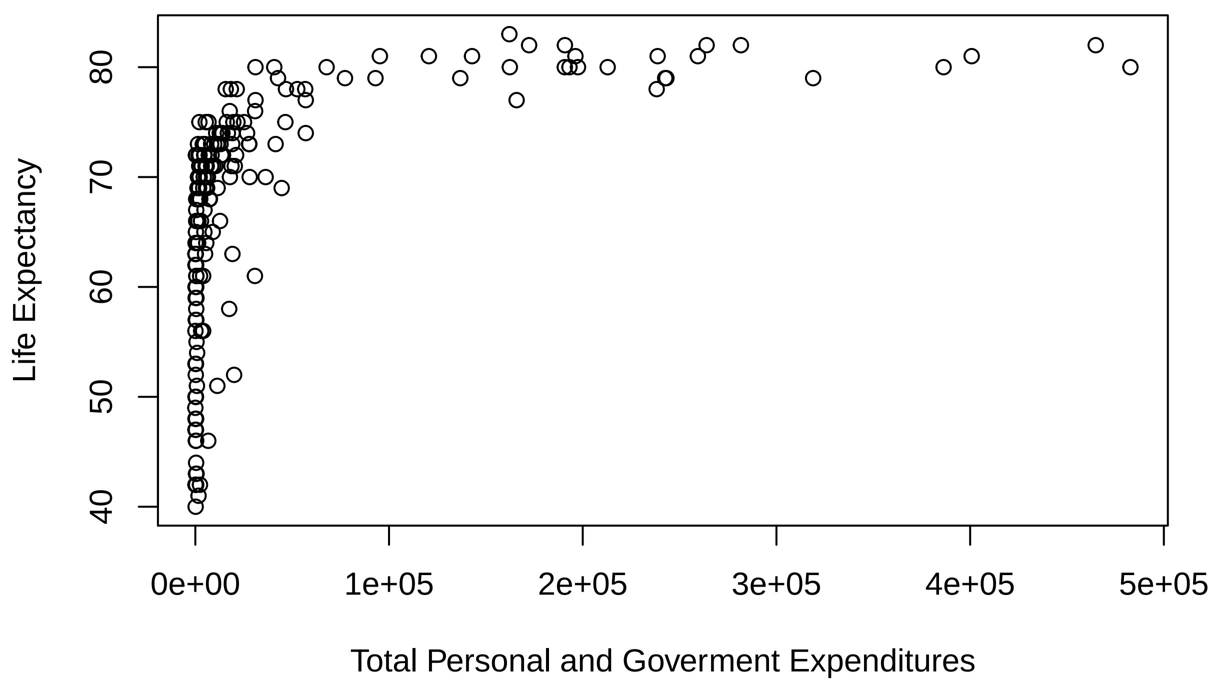
4/25/2020

Downloading the data With the dataframe ready to go, let's start:

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.

To set this up, we can use the following R functions:

```
plot(LifeExp ~ TotExp, xlab = 'Total Personal and Goverment Expenditures',
     ylab = 'Life Expectancy')
```

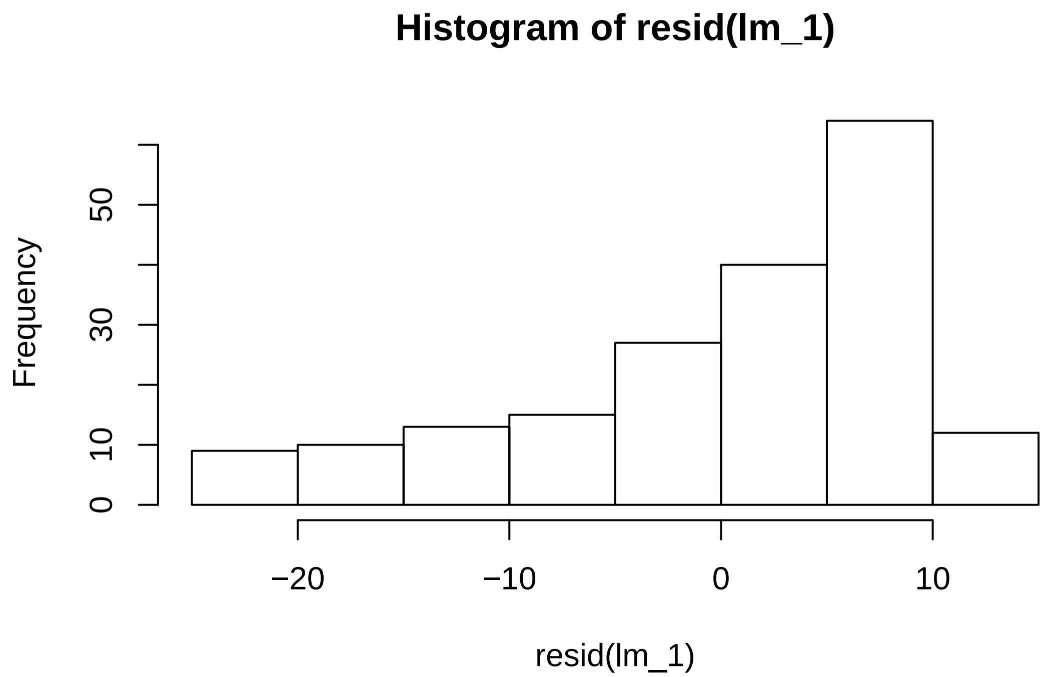


```r
lm_1 <- lm(LifeExp ~ TotExp)
summary(lm_1)
```

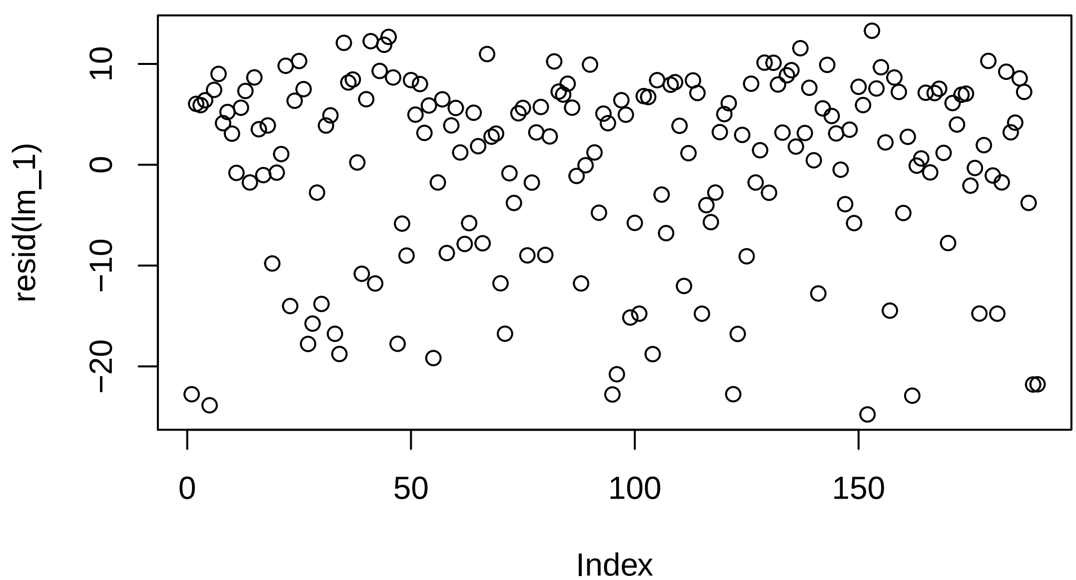
```
##
## Call:
## lm(formula = LifeExp ~ TotExp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
## TotExp      6.297e-05  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
```

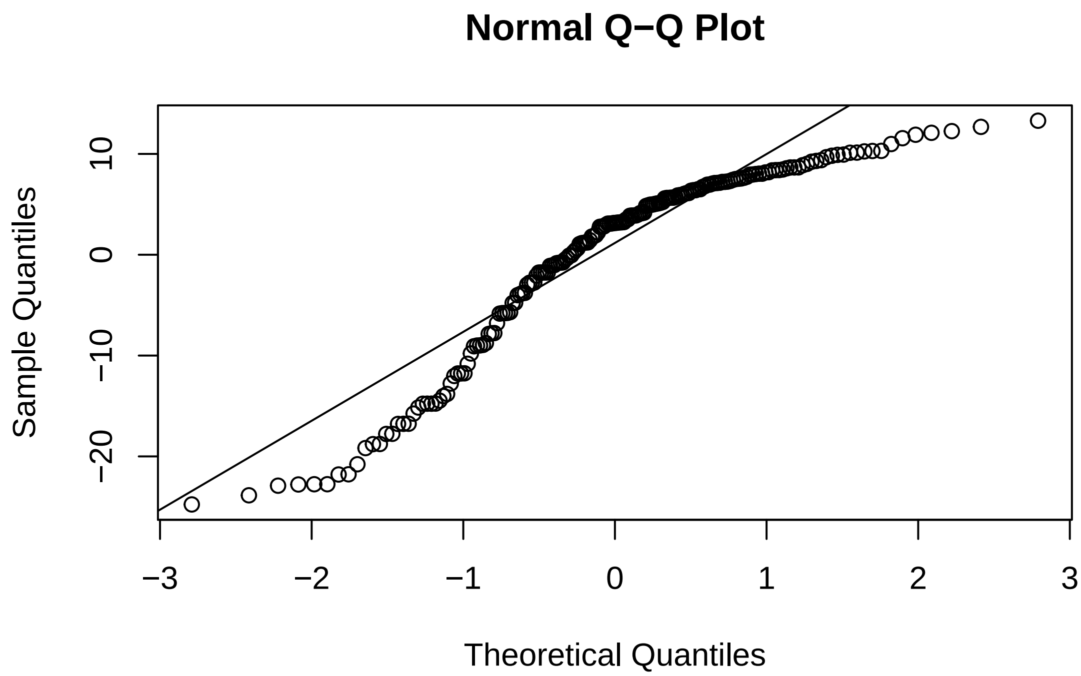
After running the linear regression, I've found the following:

- The F-statistic is 65.26, which isn't particularly valuable for this model given that this value compares the current model to a model with one fewer parameters. Since this one-factor model already has only a single parameter, it's not really useful at this point, but will become more important when we add additional factors later.
- The multiple R^2 value is 0.2577, which is quite low since this value is between 0 and 1. It measures how well the model describes the measured data. We can say that the sum of personal and goverment expenditures explains about 25% of the variation in average life expectancy of a country in years.
- The residual standard error that we see here is 9.371, which is the total variation of the residual values. The residuals seem to be pretty normally distributed (1Q and 3Q of residuals are roughly the same), but the residual standard error is not 1.5 times less these values for 1Q and 3Q, which isn't a good sign. This indicates that the residuals may not be normally distributed. We can confirm this below from a histogram of the residuals, which shows that the distribution is left-skewed.
- The p-values of the coefficients are both less than 0.001, which indicates that the probability that the intercept or `TotExp` are not relevant in the model are quite small. This shows that they may be good predictors of life expectancy.



Residual Analysis After testing the residuals, we can see that they are indeed uniformly scattered above and below zero, although they reach past -20 and only reach slightly above 10, which indicates that it may not be a very normal distribution. This is proven when we use the q-q plot, and confirmed with the histogram earlier.





From this, we can say that the assumptions for a linear regression are not completely met given the non-normal distribution of the residuals.

---

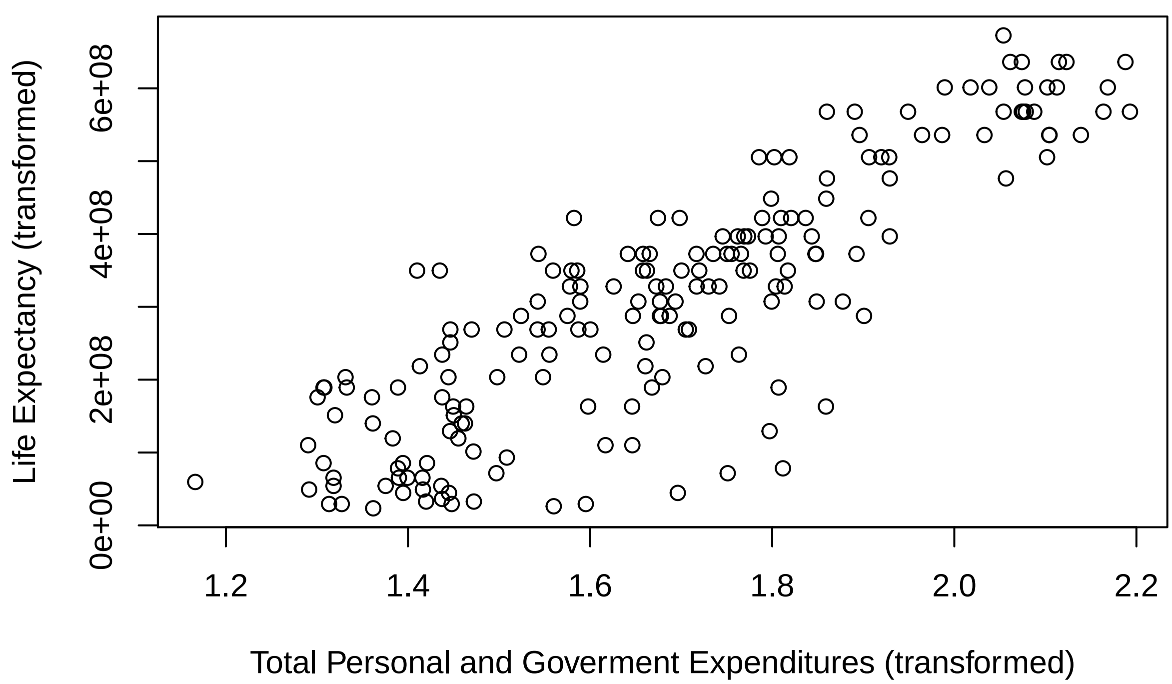
2. Raise life expectancy to the 4.6 power (i.e., LifeExp^4.6). Raise total expenditures to the 0.06 power (nearly a log transform, TotExp^.06). Plot LifeExp^4.6 as a function of TotExp^.06, and re-run the simple regression model using the transformed variables. Provide and interpret the F statistics, R^2, standard error, and p-values. Which model is "better?"

First, I'm transforming the coefficients:

```
lifeexp_tran <- LifeExp^4.6
totexp_tran <- TotExp^.06
```

Now, I'll plot the transformed variables:

```
plot(lifeexp_tran ~ totexp_tran, xlab = 'Total Personal and Goverment Expenditures (transformed)',
     ylab = 'Life Expectancy (transformed)')
```



And to rerun the model with the transformed variables:

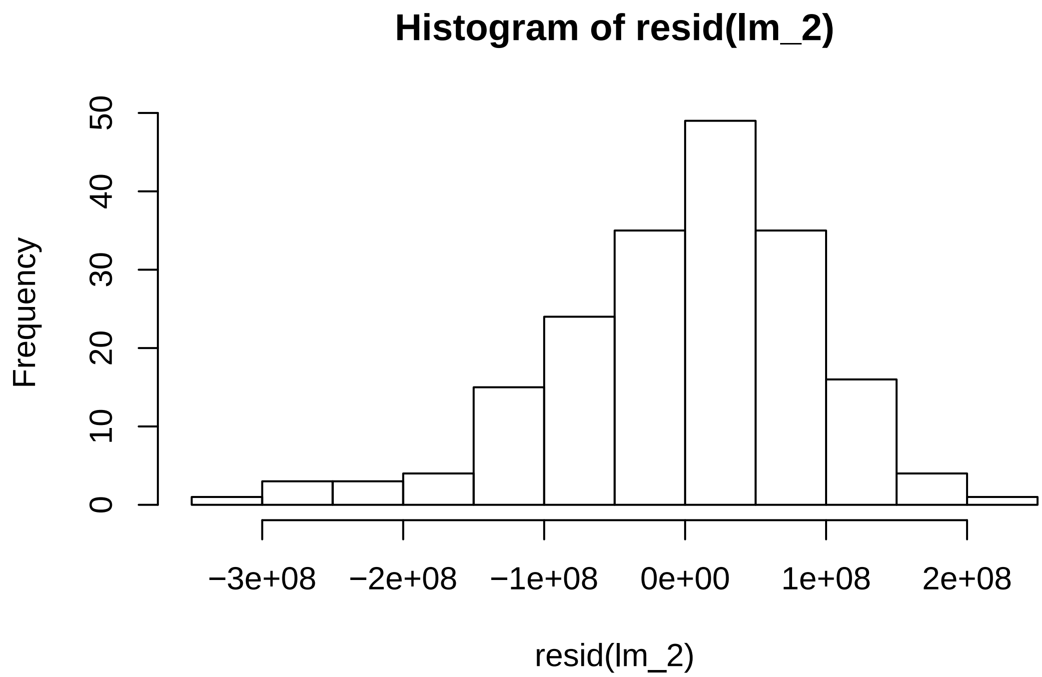
```
lm_2 <- lm(lifeexp_tran ~ totexp_tran)
summary(lm_2)
```

```
##
## Call:
## lm(formula = lifeexp_tran ~ totexp_tran)
##
## Residuals:
```

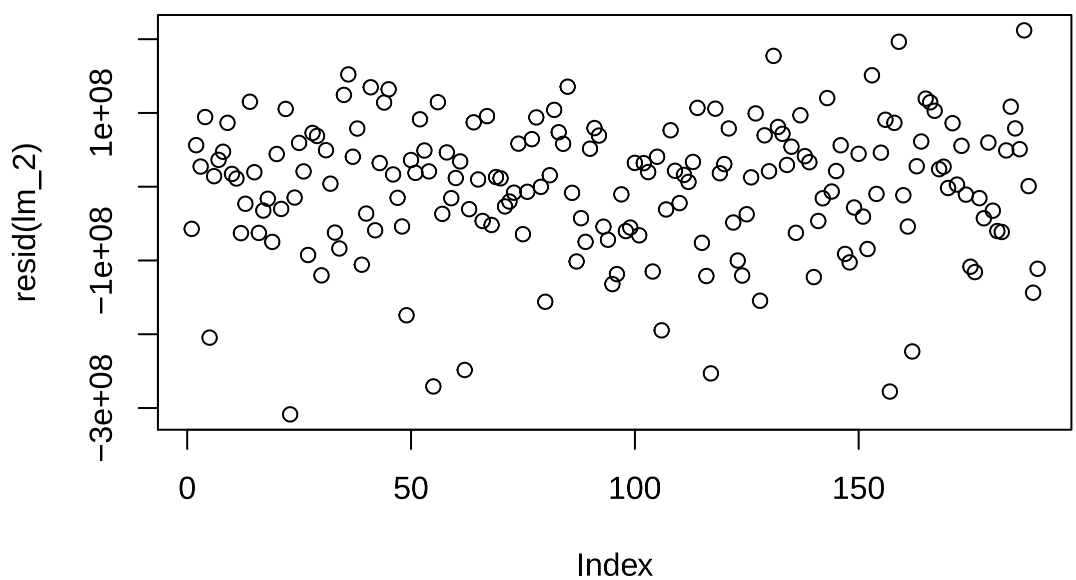
```
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -736527910   46817945  -15.73   <2e-16 ***
## totexp_tran  620060216   27518940   22.53   <2e-16 ***
## ---
## 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
```

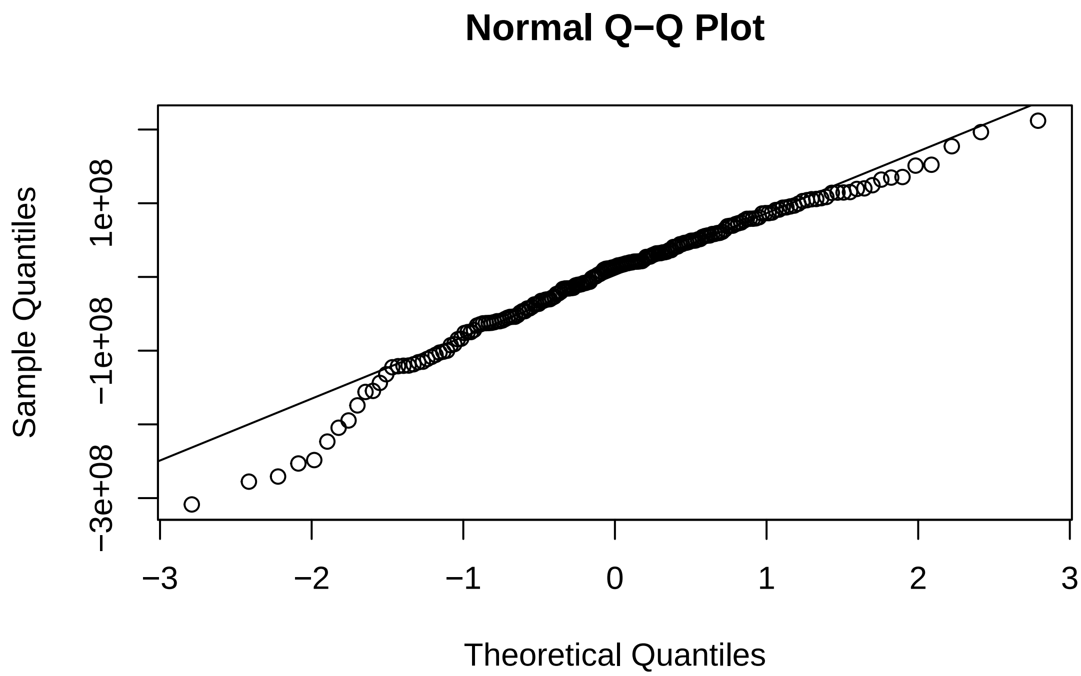
After running the linear regression, I've found the following:

- The F-statistic is 507.7, which, again isn't particularly valuable for this model given that this value compares the current model to a model with one fewer parameters. Since this one-factor model already has only a single parameter, it's not really useful at this point, but will become more important when we add additional factors later.
- The multiple R^2 value is 0.7283, which is a great deal higher than the model without transformed variables. It measures how well the model describes the measured data. We can say that the sum of personal and goverment expenditures explains about 73% of the variation in average life expectancy of a country in years.
- The residual standard error that we see here is 90490000, which is the total variation of the residual values (after being transformed). The residuals seem to be pretty normally distributed (1Q and 3Q of residuals are roughly the same magnitude), which is a good sign.
- Similar to the non-transformed model, the p-values of the coefficients are both less than 0.001, which indicates that the probability that the intercept or `TotExp`(transformed) are not relevant in the model are quite small. This shows that they may be good predictors of life expectancy.



Residual Analysis After testing the residuals for our transformed model, we can see that they are indeed pretty uniformly scattered above and below zero, and from the histogram above, follows a fairly normal distribution. This is proven when we use the q-q plot.





From this, we can say that the assumptions for a linear regression are met for this model, given uniform scattering of residuals, normally distributed residuals, and independent observations.

Furthermore, the “better” model is definitely the transformed model in question #2, given that the R-squared value is much higher than the model in question #1 and the residuals are much more normally distributed in this second model.

---

3. Using the results from 2, forecast life expectancy when TotExp^.06 =1.5. Then forecast life expectancy when TotExp^.06=2.5.

Given that our linear regression outputs the following equation:

- $LifeExp(transformed) = -736527909 + 620060216 \times TotExp(transformed)$

And we want to find out what LifeExp(transformed) will be when TotExp(transformed) is equal to 1.5, we can solve the equation:

- $LifeExp(transformed) = -736527909 + 620060216 \times 1.5$
- $LifeExp(transformed) = 193562415$

However, since this is the transformed value, we need to convert this back to normal units:

- $LifeExp^{4.6} = 193562415$
- $LifeExp = 193562415^{\frac{1}{4.6}}$
- $LifeExp = 63.312$ years

The life expectancy is about 63.31 years when the `TotExp`(transformed) is equal to 1.5.

Similarly, we can use the same equation again to solve the second part of the question:

- $LifeExp(transformed) = -736527909 + 620060216 \times TotExp(transformed)$

And we want to find out what LifeExp(transformed) will be when TotExp(transformed) is equal to 2.5, we can solve the equation:

- $LifeExp(transformed) = -736527909 + 620060216 \times 2.5$
- $LifeExp(transformed) = 813622631$

However, since this is the transformed value, we need to convert this back to normal units:

- $LifeExp^{4.6} = 813622631$
- $LifeExp = 813622631^{\frac{1}{4.6}}$
- $LifeExp = 86.506$ years

The life expectancy is about 86.51 years when the `TotExp`(transformed) is equal to 2.5.

---

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

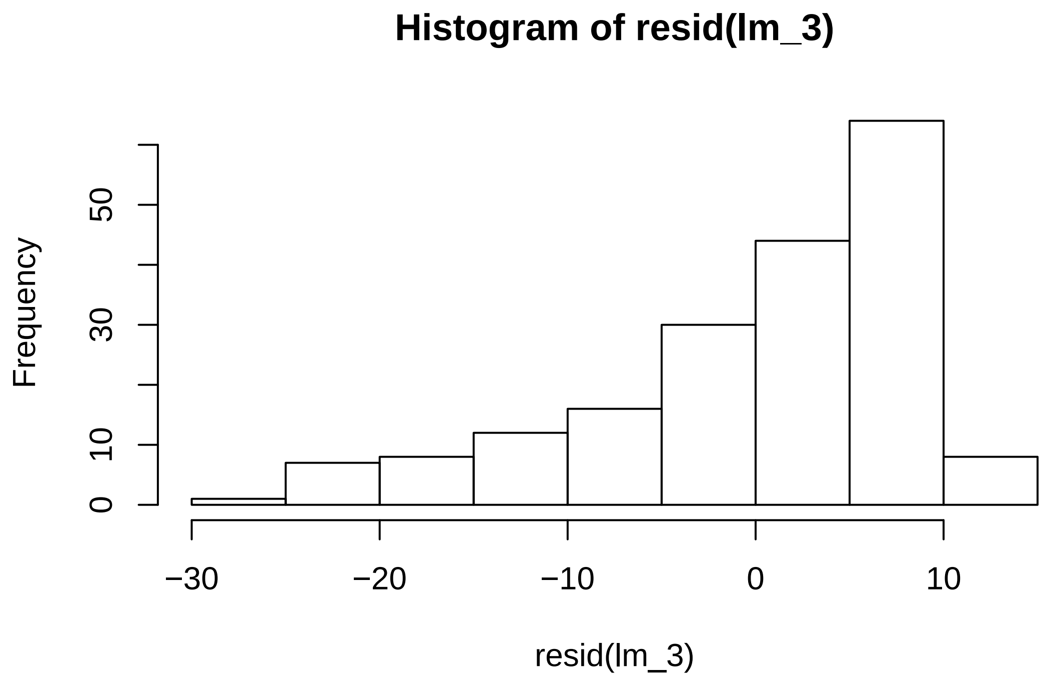
$$LifeExp = b0 + b1 \times PropMd + b2 \times TotExp + b3 \times PropMD \times TotExp$$

```r
lm_3 <- lm(LifeExp ~ PropMD + TotExp + PropMD*TotExp)
summary(lm_3)
```

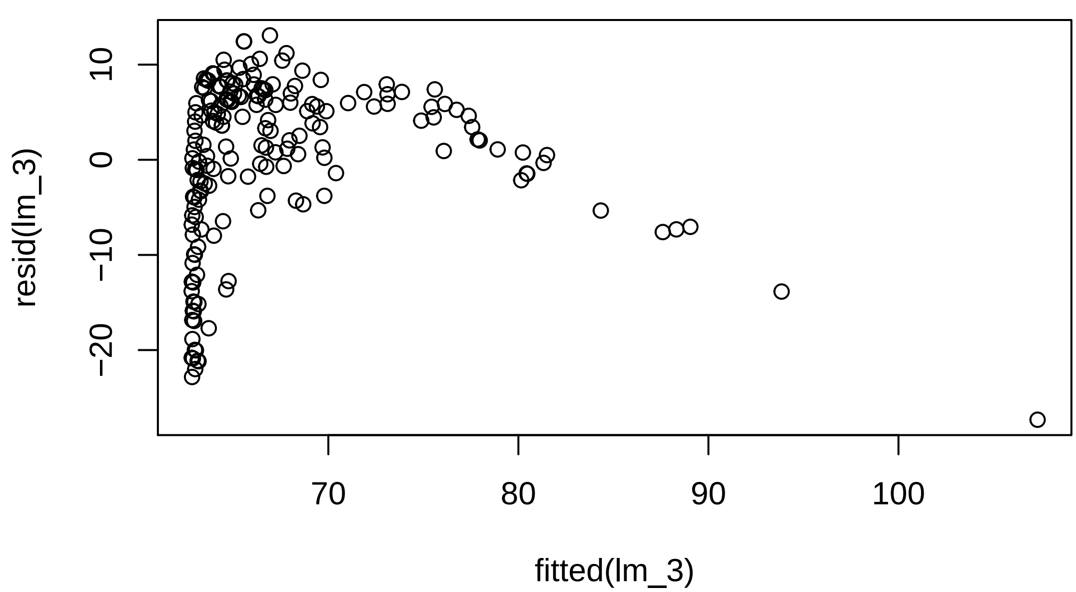
```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMD * TotExp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.277e+01  7.956e-01  78.899  < 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
```

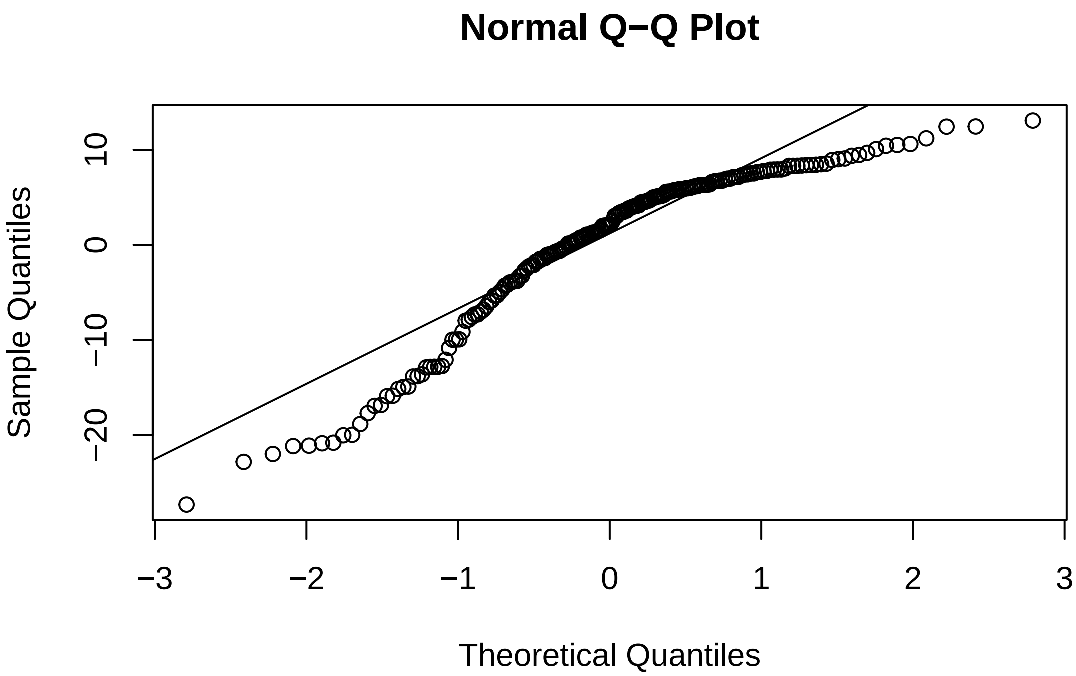
After running the multiple regression, I've found the following:

- The F-statistic is 34.49, which is less than when we just had `TotExp` in our model, so this is a good sign. However, the decrease is only slight, so we can't read into this too much.
- The multiple R^2 value is 0.3574, which is a little better than our model in question #1, but still only gives us an indication that about 36% of the variation in average life expectancy of a country in years can be explained by these three variables – `PropMD`, `TotExp` and the interaction variable of `PropMD:TotExp`.
- The residual standard error that we see here is 8.765, which is the total variation of the residual values. The residuals seemed to be balanced along the quartiles (1Q and 3Q of residuals are roughly the same magnitude), which is a good sign. However, the maginitude between Min and Max is quite different, which indicates that the residual distribution may be skewed.
- Similar to the non-transformed model in question #1, all p-values of the coefficients are less than 0.001, which indicates that the probability that the intercept, `PropMD`, `TotExp`, and the interaction of `PropMD` $\times$ `TotExp` are not relevant in the model are quite small.



Residual Analysis After testing the residuals for our multiple regression model, we can see that they are not uniformly scattered above and below zero, and from the histogram above, we can see that the residuals are fairly left skewed. This is proven when we use the q-q plot.





In the end, the residual analysis shows that the model doesn't meet many of the criteria needed to make accurate assumptions or predictions – it can be deemed as invalid. Additionally, the R-squared value is quite low, which means that the model isn't very good.

---

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

Using the equation we generated from our linear model, we can forecast the following:

- $LifeExp = 62.277 + (1497 \times PropMD) + (0.00007233 \times TotExp) + (-0.006026 \times PropMD : TotExp)$

With values for PropMD and TotExp, we can calculate and forecast:

- $LifeExp = 62.277 + (1497 \times 0.03) + (0.00007233 \times 14) + (-0.006026 \times (0.03 \times 14))$
- $LifeExp = 107.185$ years

The average life expectancy when `PropMD` is equal to 0.03 and `TotExp` is equal to 14 is about 107.19 years. This is not a realistic forecast. Given the average life expectancy for humans is much lower, in the year 2008 when this data was released it was about 68 years, an estimation of an average life expectancy for a country of 107 is very unreasonable.