DATA 605 : Week 12 - Regression Analysis in R 2

Ramnivas Singh

11/14/2021

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.

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. Read in the data from github

```
who <- read.csv(url("https://raw.githubusercontent.com/vindication09/DATA-605/master/who.csv"), header =
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
```

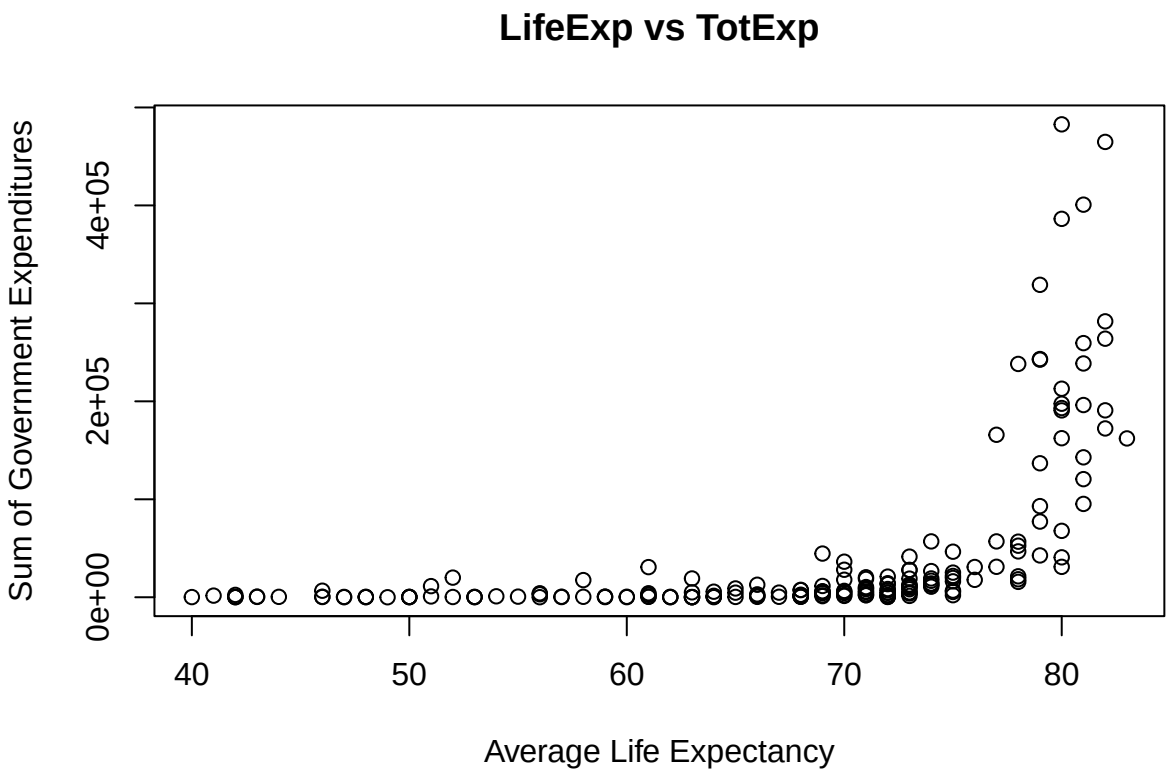
```
summary(who)
```

```
##    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
```

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.

Lets visualize

```
plot(who$LifeExp, who$TotExp, xlab='Average Life Expectancy', ylab='Sum of Government Expenditures',
     main='LifeExp vs TotExp')
```



Run Simple Regression Model

```
mod <- lm(LifeExp~TotExp, data=who)
summary(mod)
```

```
##
## 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
```

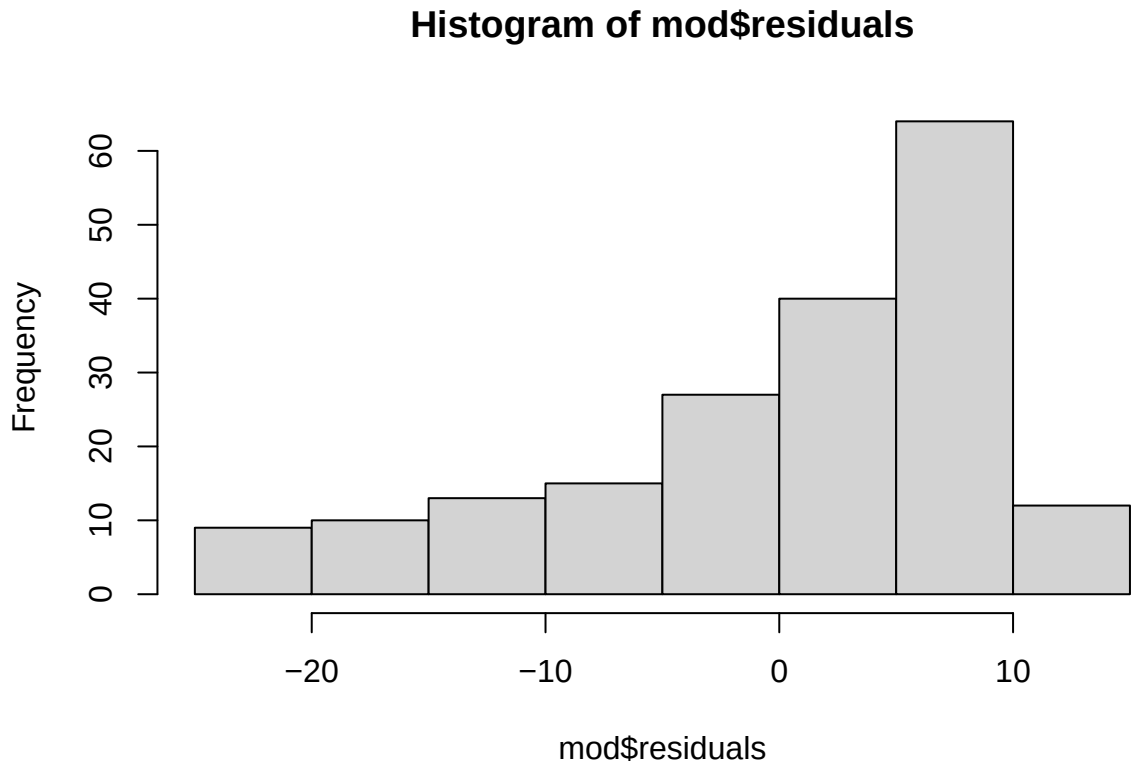
What can we learn about our model from the output? The F-Statistic is 65.26 on 1 and 188 degrees of freedom. A statistics greater than 1 could indicate there is a relationship between response and predictor

but how large exactly depends on the number of data points. The F test is actually testing the model against the null model. Based on the p-value, the model is not equal to the null model. If the null hypothesis is that the model is equal to the null model, then we can reject.

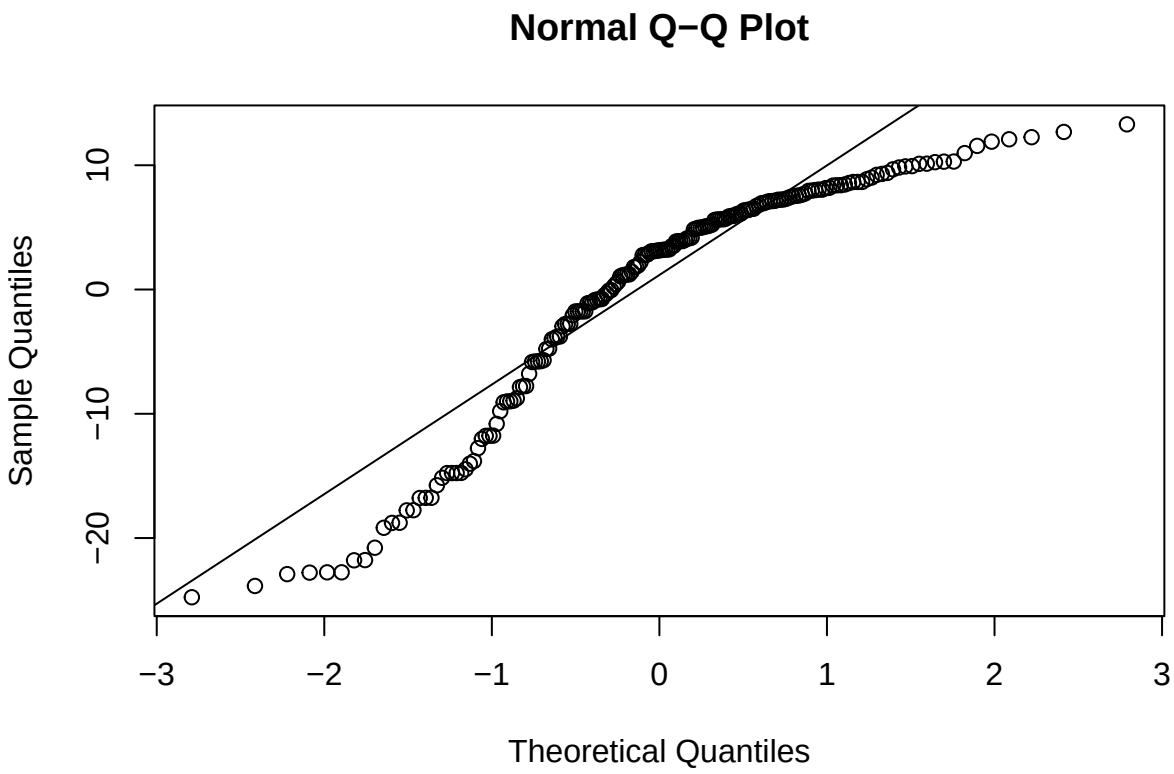
The adjusted R squared is .2, meaning roughly 20% of the variability in the data is accounted for. The standard error is about 9%,which is larger than what we would want it to be.

Assumptions of regression

```
hist(mod$residuals);
```




```
qqnorm(mod$residuals);
qqline(mod$residuals)
```



The residuals are clearly not normal or close to normal. This assumption is not met.

```
library(olsrr)
```

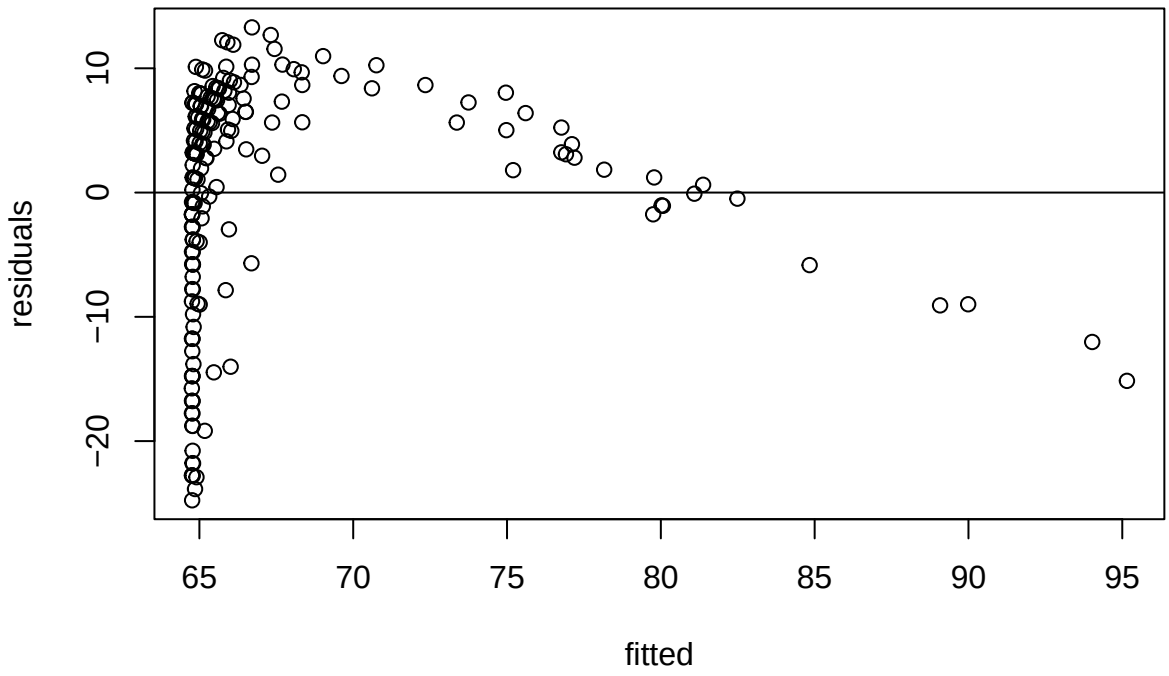
```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_test_breusch_pagan(mod);
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                Data
##  ----------------------------------
##  Response : LifeExp
##  Variables: fitted values of LifeExp
##
##         Test Summary
```

```
##  ----------------------------
##  DF            =     1
##  Chi2          =     2.599177
##  Prob > Chi2   =     0.1069193
```

```
plot(fitted(mod), residuals(mod), xlab="fitted", ylab="residuals")
abline(h=0)
```



Constant variance condition also fails. Observe the high p value for the Breusch Pagan Test for Heteroskedasticity.

There is an abundance of information to indicate that the model is not a good fit at all.

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?"

```
mod2 <- lm((LifeExp^4.6)~I(TotExp^.06), data=who)
summary(mod2)
```

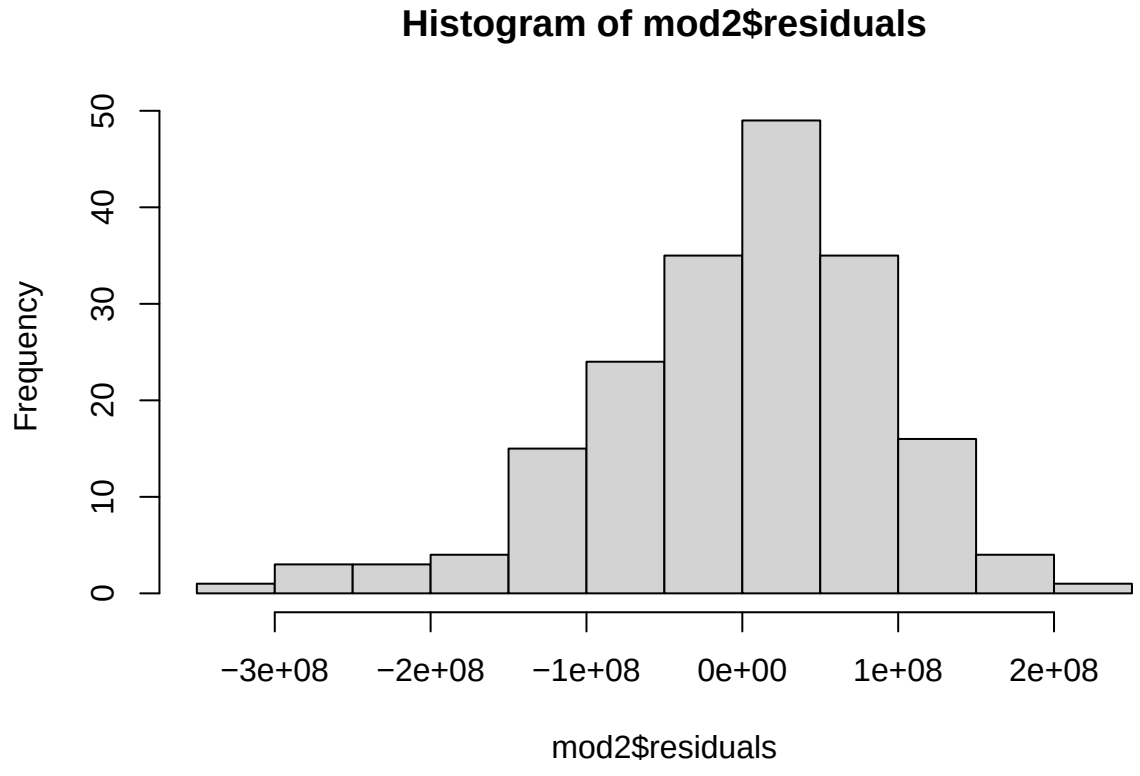
```
##
## Call:
## lm(formula = (LifeExp^4.6) ~ I(TotExp^0.06), data = who)
##
```

```
## 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 ***
## I(TotExp^0.06)  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
```

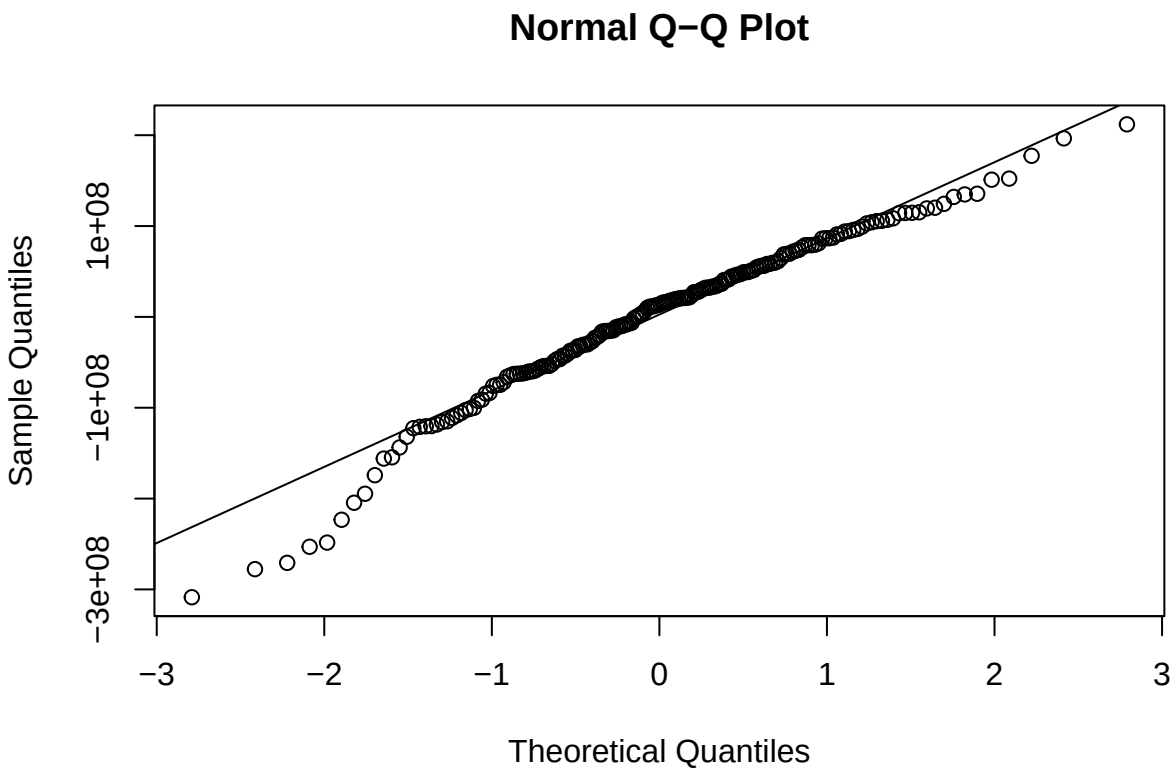
There already is a massive improvement in the adjusted r squared. 72 percent of the variability in the data is accounted for.Our F statistic is much larger (over 500) indicating a strong relationship between predictor and response.

Residuals

```
hist(mod2$residuals);
```




```
qqnorm(mod2$residuals);
qqline(mod2$residuals)
```

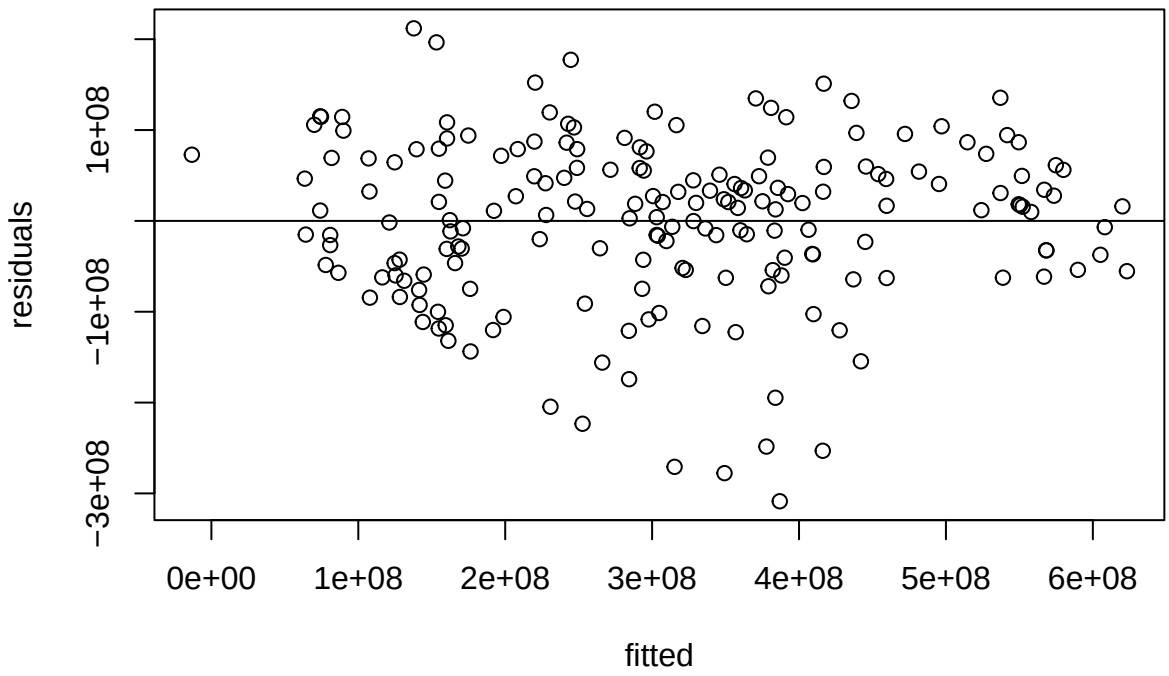


Residuals are much closer to the normal distribution than the previous model.

```
ols_test_breusch_pagan(mod2);
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                  Data
##  -----------------------------------------
##  Response : (LifeExp^4.6)
##  Variables: fitted values of (LifeExp^4.6)
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    0.4278017
##  Prob > Chi2   =    0.5130696
```

```
plot(fitted(mod2), residuals(mod2), xlab="fitted", ylab="residuals")
abline(h=0)
```



With 90% confidence , we can say the variance is constant.

The transformed model is much better than the original model. It should be noted that the residual standard error in model 2 is much much larger.

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

This problem is asking us to find the value of y given x. Lets make a function using the coefficients from model 2

```
mod2_compute <- function(x)
  {   y <- -736527910 + 620060216 * (x)
      y <- yˆ(1/4.6)
    print(y)
  }
```

Compute

```
mod2_compute(1.5)
```

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

Compute other

```
mod2_compute(2.5)
```

```
## [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

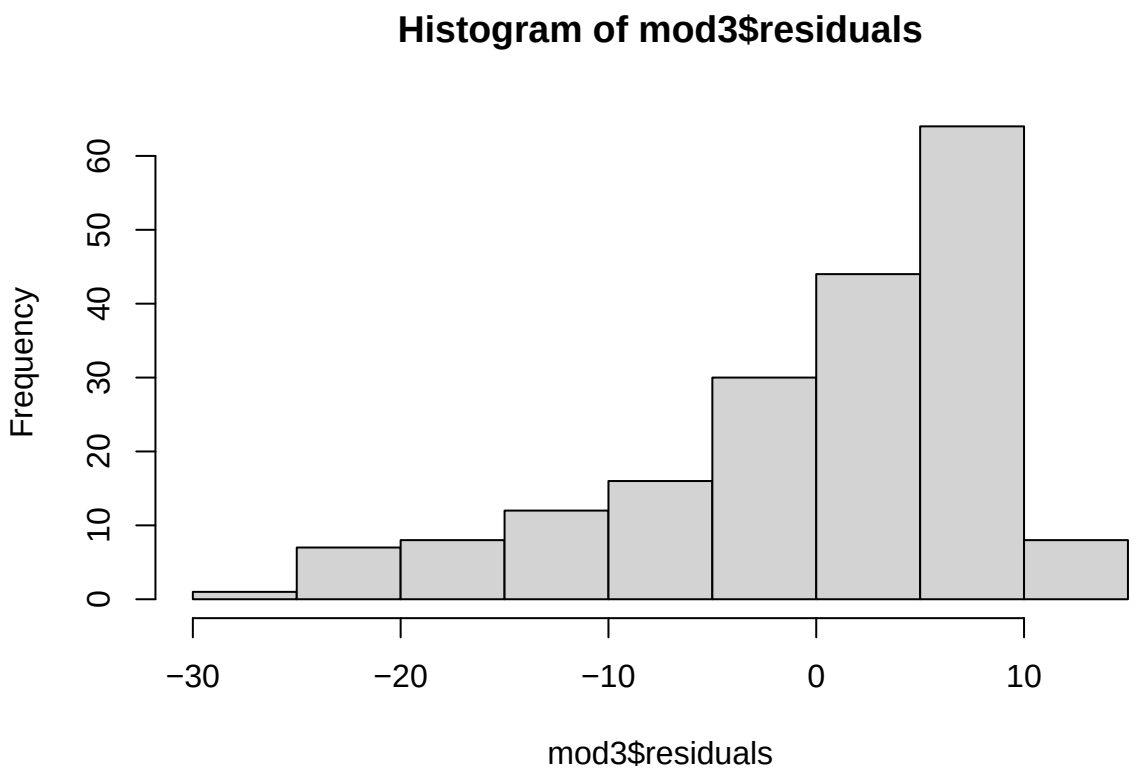
```
mod3 <- lm(LifeExp ~ PropMD+TotExp+(PropMD*TotExp), data=who)
summary(mod3)
```

```
##
## 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
```

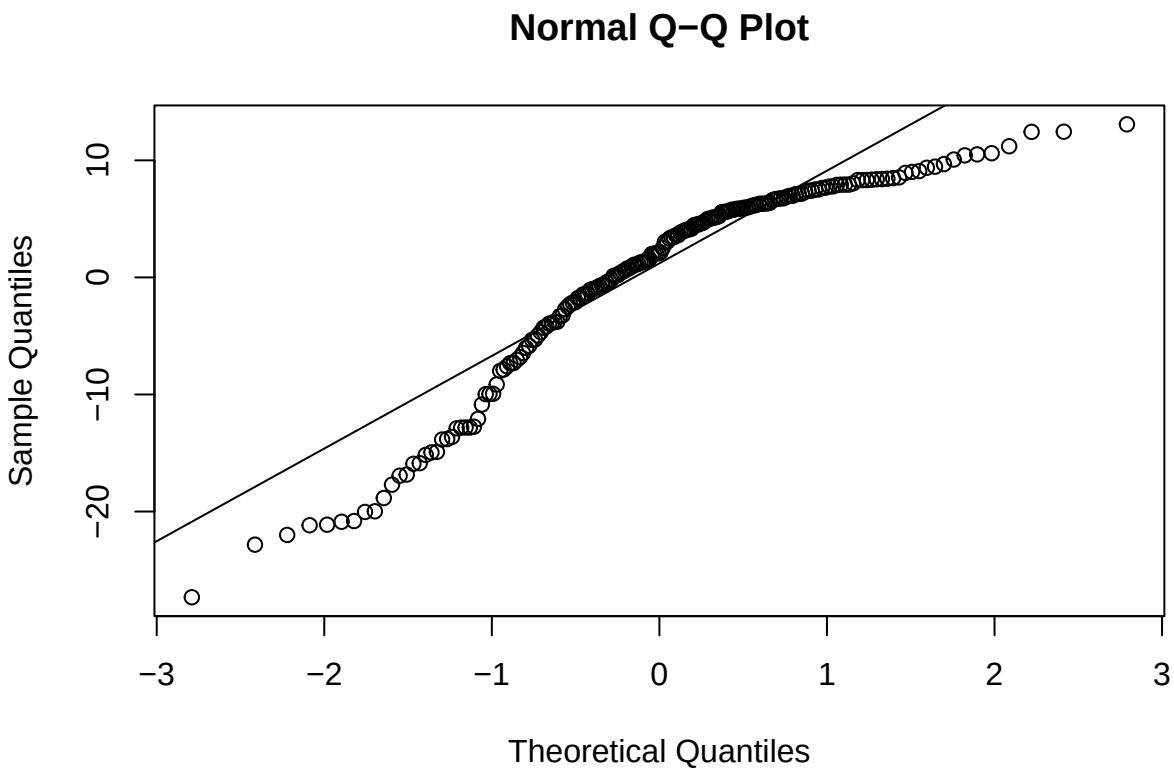
The model with additional predictors and interaction terms is better than the original model (mod1). The adjusted r squared is higher. The residual error is slightly smaller. What can we learn from the residuals?

Residuals

```
hist(mod3$residuals);
```




```
qqnorm(mod3$residuals);
qqline(mod3$residuals)
```

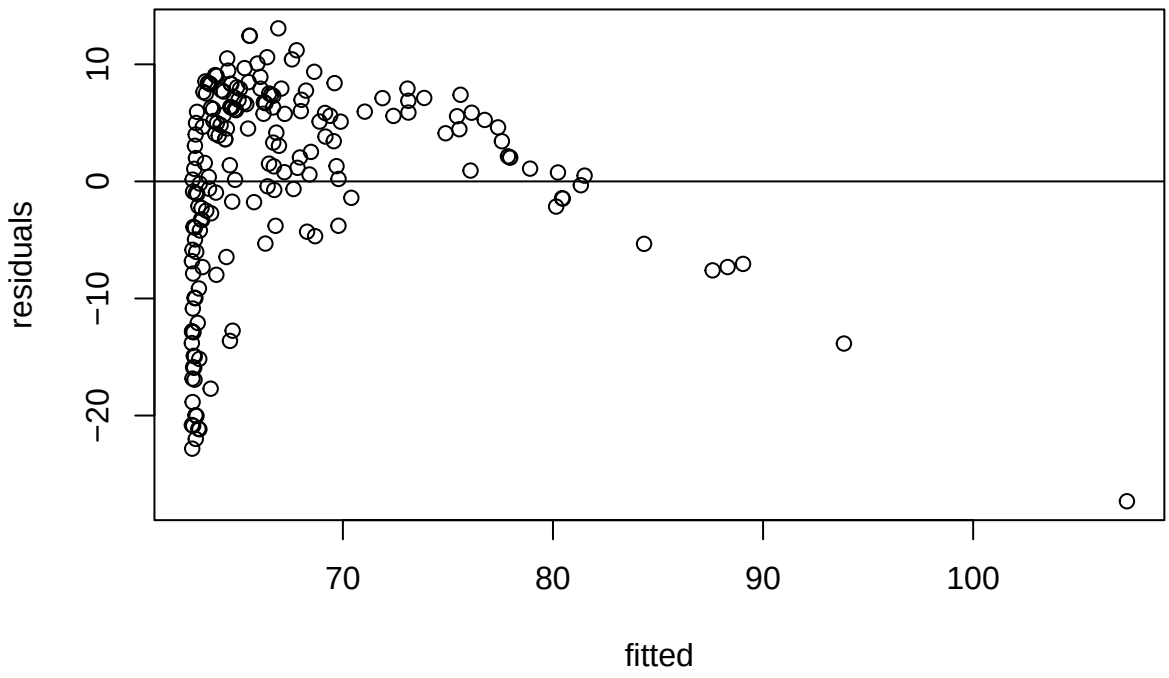


Residuals do not appear normal. There is a heavy right skew.

```
ols_test_breusch_pagan(mod3);
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                Data
##  ----------------------------------
##  Response : LifeExp
##  Variables: fitted values of LifeExp
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    0.0031467
##  Prob > Chi2   =    0.9552658
```

```
plot(fitted(mod3), residuals(mod3), xlab="fitted", ylab="residuals")
abline(h=0)
```



We do not have constant variance. Our third model with interaction term and additional predictors is not a good model and does not satisfy the assumptions of regression.

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

```
#remove scientific notation with
options(scipen=999)
coef(mod3)
```

```
##      (Intercept)           PropMD           TotExp    PropMD:TotExp
##   62.77270325541 1497.49395251893    0.00007233324   -0.00602568644
```

```
mod3_compute <- function(x,y)
  {
  z <- 62.77270325541+1497.49395251893*(x)+(0.00007233324*(x*y))
  return(z)
  }
```

calculate when PropMD=.03 and TotExp = 14

```
mod3_compute(0.03,14)
```

```
## [1] 107.6976
```

Our predicted life exp is not realistic. The max life exp is around the 80's.
