HW12

Gabe Abreu

11/15/2020

Data 605 #12

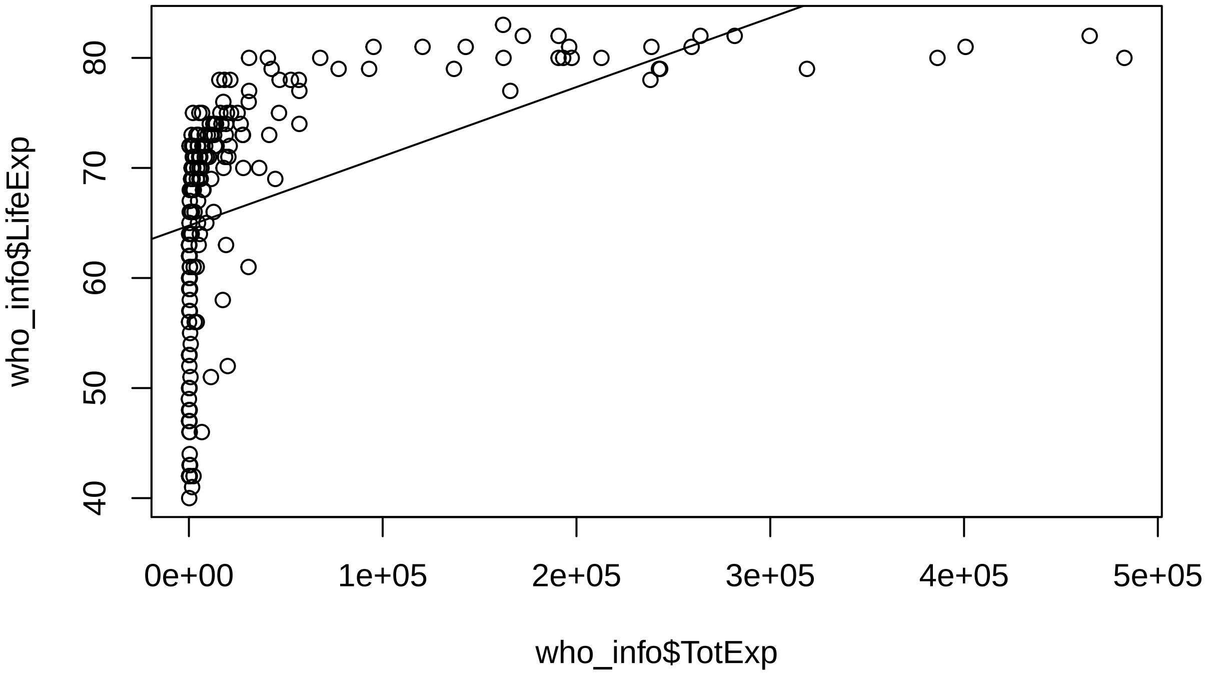
1. Provide a scatterplot of LifeExp~TotExp, and run simple linear regression. Do not transform the variables. Provide and interpret the F statistics, R^2, standard error,and p-values only. Discuss whether the assumptions of simple linear regression met.

```
who_info <- read.csv('who.csv')

head(who_info)
```

```
##               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
```

```
simpleRegression <- lm(LifeExp ~ TotExp, data = who_info)
plot(who_info$TotExp, who_info$LifeExp)
abline(simpleRegression)
```

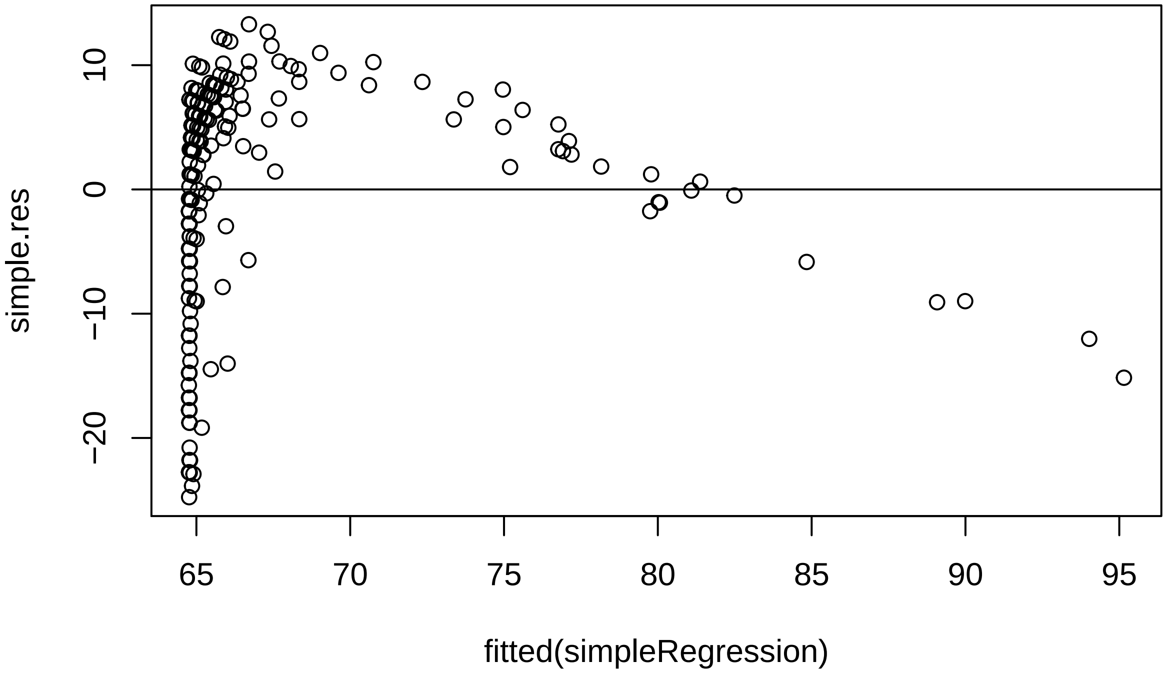


```r
summary(simpleRegression)
```

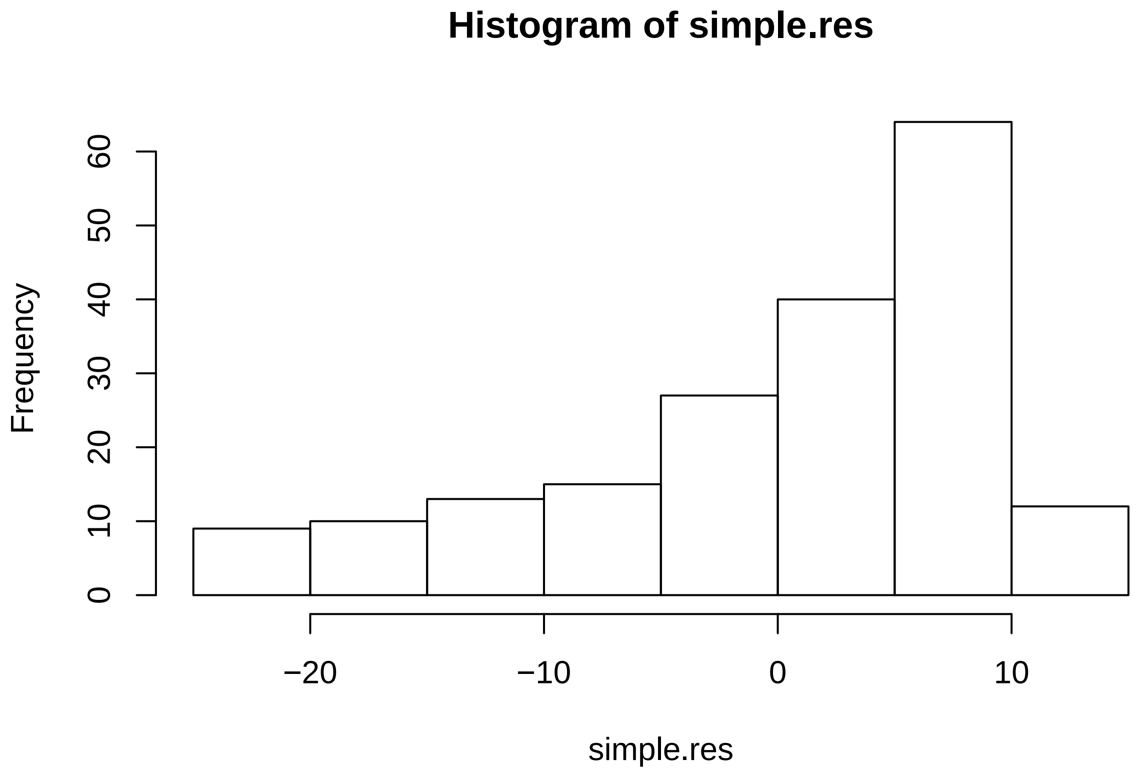
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = who_info)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
##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

```r
#Residuals
simple.res = resid(simpleRegression)

 plot(fitted(simpleRegression), simple.res)
 abline(0,0)
```

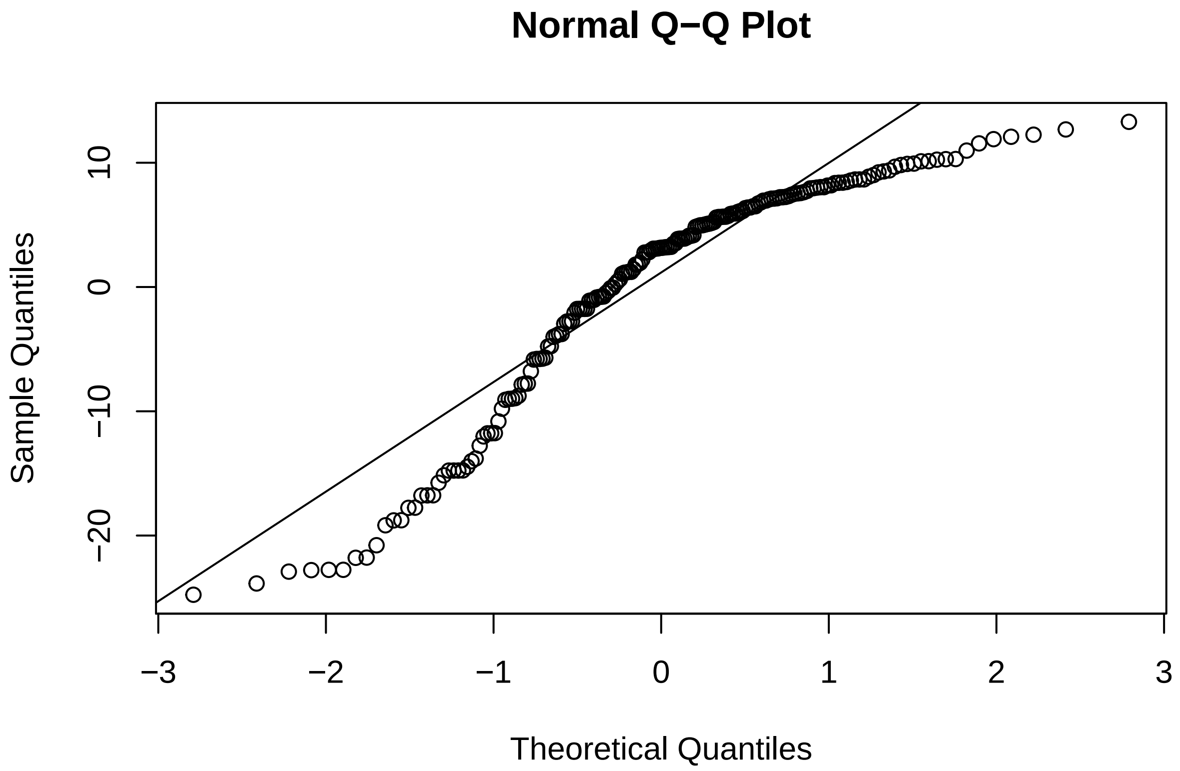


```
hist(simple.res)
```



Q-Q Plots

```
qqnorm(resid(simpleRegression))
qqline(resid(simpleRegression))
```



The F-Statistics suggests there is a relationship between the indepdent and dependent variables. The value of the F statistics is 65.26.

The R^2 represents how well the data fits the model which in this case the value of R-square (0.2577) is slow, indicating the data doesn't fit the model.

The residual standard error is the standard deviation of the residuals, the smaller the value the better. In in our model the residual standard error is high at 9.371

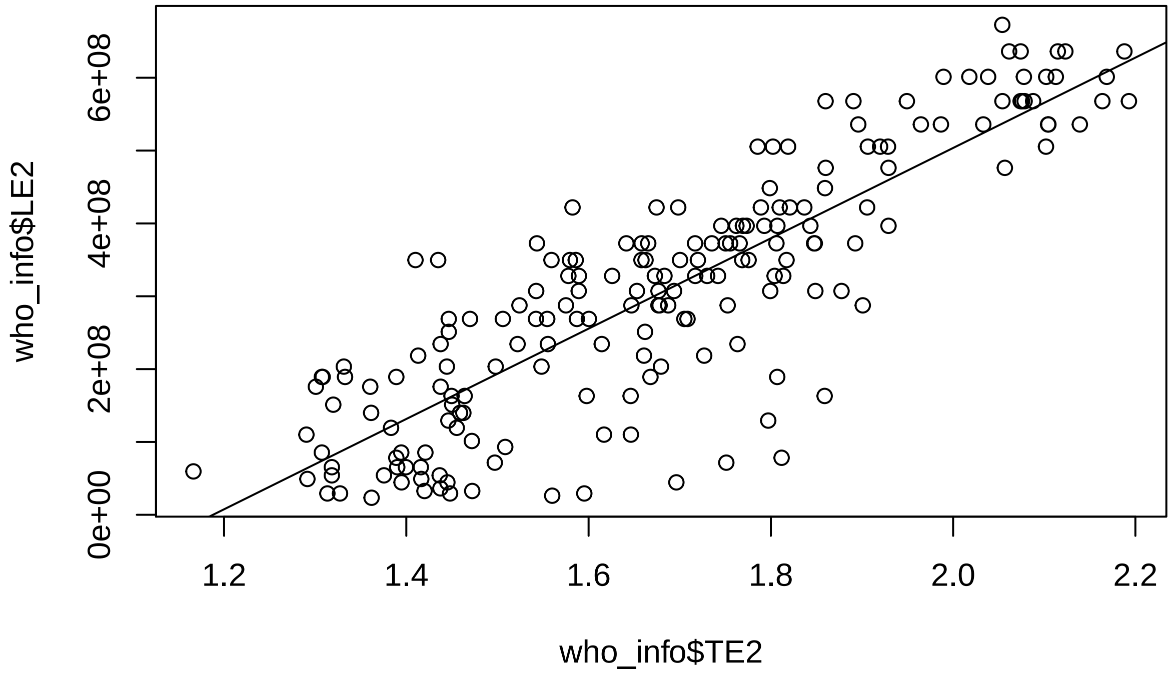
The p-value shows statistical significance at 7.71e-14.

As it stands, this is not a good model.

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_info$LE2 <- who_info$LifeExp^4.6
who_info$TE2 <- who_info$TotExp^0.06

simpleRegression2 <- lm(who_info$LE2 ~ who_info$TE2, data = who_info)
plot(who_info$TE2, who_info$LE2)
abline(simpleRegression2)
```

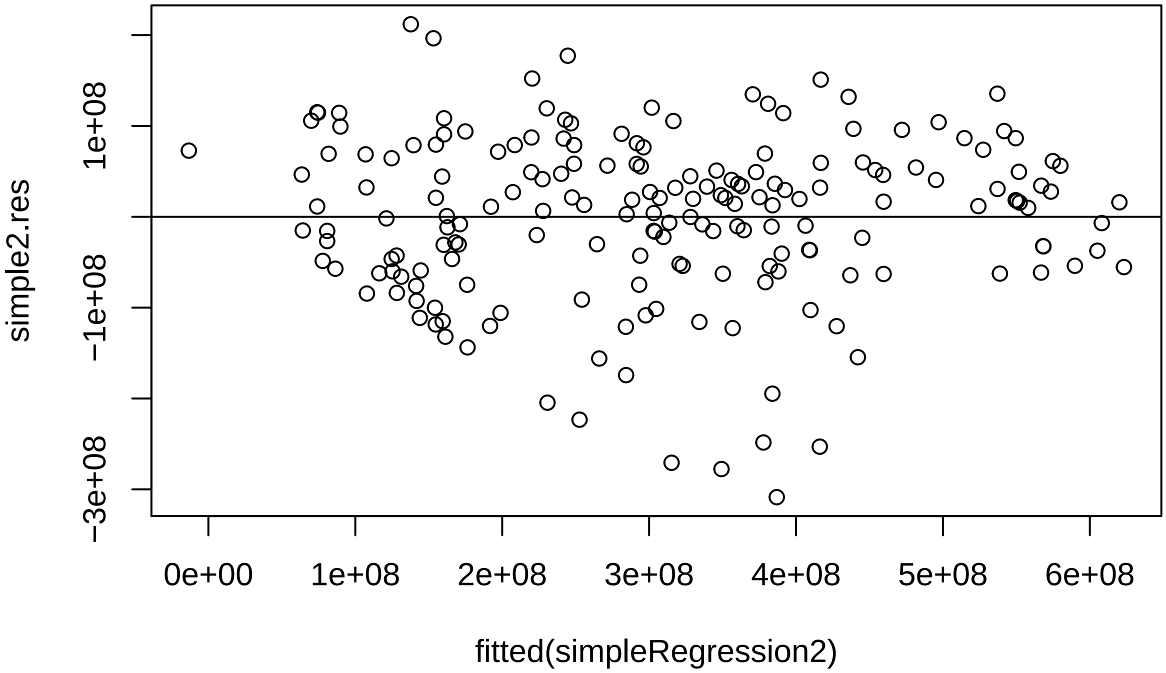


```
summary(simpleRegression2)
```

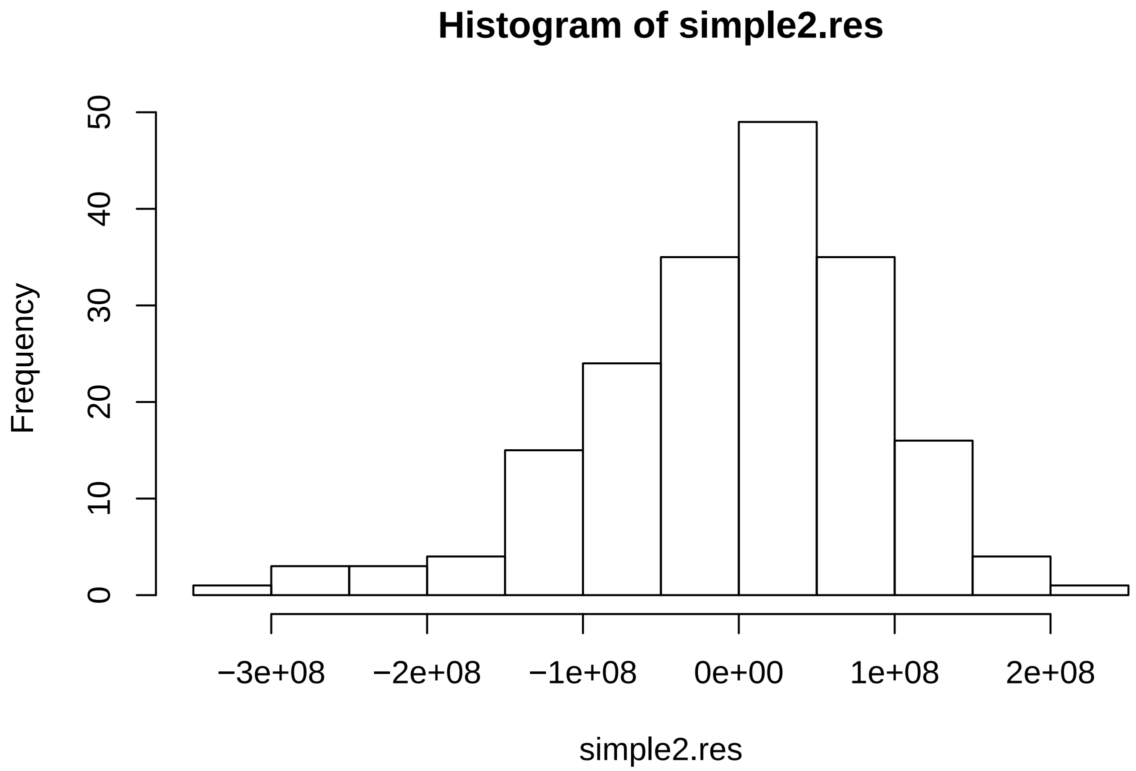
```
##
## Call:
## lm(formula = who_info$LE2 ~ who_info$TE2, data = who_info)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -308616089  -53978977   13697187   59139231  211951764
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -736527910   46817945  -15.73   <2e-16 ***
## who_info$TE2  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
```

```
#Residuals
simple2.res = resid(simpleRegression2)

 plot(fitted(simpleRegression2), simple2.res)
 abline(0,0)
```

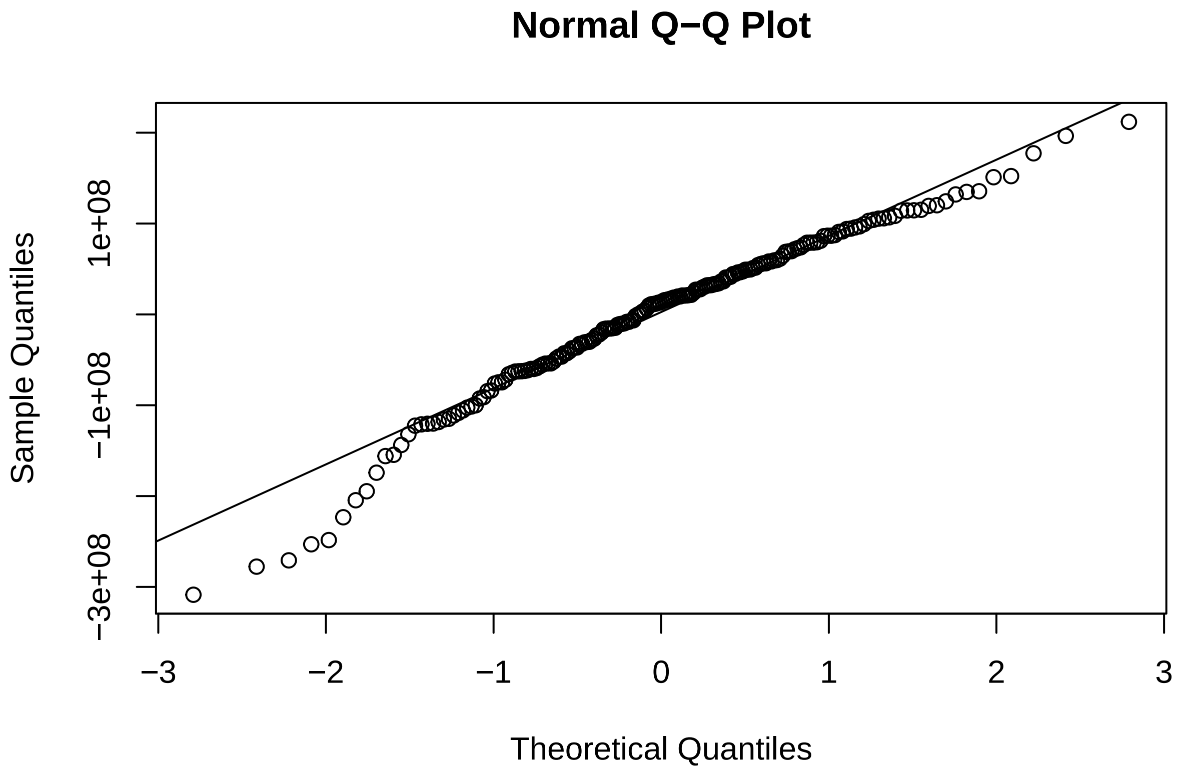


```
hist(simple2.res)
```



Q-Q Plots

```
qqnorm(resid(simpleRegression2))
qqline(resid(simpleRegression2))
```



The model is better based on the criteria used to judge linear models.

The F statistics is much higher and p-value is still significant. The Rˆ2 valueu is greatly improved. The residuals show normality and linearity.

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

The formula is y = -736527910 + 620060216 * x

```
predictor <- function(x) {
    y <- -736527910 + 620060216 * x
    y <- y^(1/4.6)

    return(y)
}
```

totExpˆ0.6 = 1.5

```
predictor(1.5)
```

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

totExpˆ0.6 = 2.5

```r
predictor(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

```r
multiRegression <- lm(LifeExp ~ PropMD + TotExp + PropMD*TotExp, data = who_info)
summary(multiRegression)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMD * TotExp, data = who_info)
##
## 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 doesn't work as well as the second model. The residual standard error is high at 8.765. The R^2 value is low at 0.3574 and the F-Statistic is 34.49. The p-value is still significant being well below 0.05.

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

```r
expdata <- data.frame(PropMD=0.03, TotExp=14)
round(predict(multiRegression, expdata))
```

```
##   1
## 108
```

The value is much too high at 107. Doesn't seem realistic given the rest of the data.