DATA 605 - Homework 12

Omer Ozeren

Table of Contents


```
library(tidyverse)
library(knitr)
library(kableExtra)
library(gvlma)
library(gridExtra)
```

Objective

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.
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?"
3. Using the results from 3, forecast life expectancy when TotExp^.06 =1.5. Then forecast life expectancy when TotExp^.06=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 x PropMd + b2 x TotExp +b3 x PropMD x TotExp
5. Forecast LifeExp when PropMD=.03 and TotExp = 14. Does this forecast seem realistic? Why or why not?

Question 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.

Scatterplot

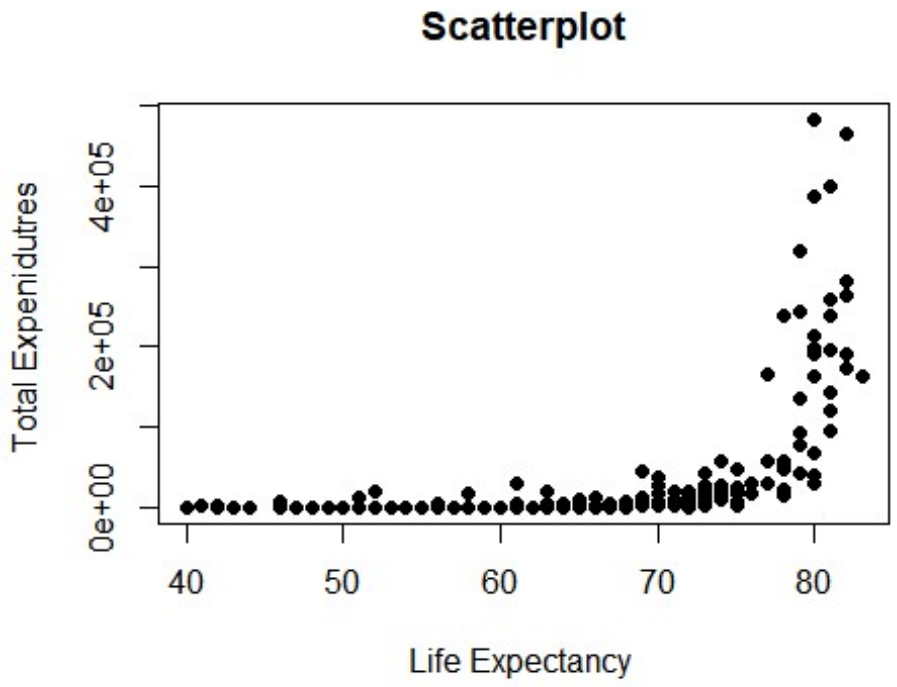
```
library(knitr)
url <- "C:/Users/OMERO/Desktop/who.csv"
who_df  <- read.csv(file = url, header = T, stringsAsFactors = F)
summary(who_df)
##    Country             LifeExp      InfantSurvival   Under5Survival
##  Length:190         Min.   :40.00   Min.   :0.8350   Min.   :0.7310
##  Class :character   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
##  Min.   :0.9870   Min.   :0.0000196   Min.   :0.0000883
##  1st Qu.:0.9969   1st Qu.:0.0002444   1st Qu.:0.0008455
##  Median :0.9992   Median :0.0010474   Median :0.0027584
##  Mean   :0.9980   Mean   :0.0017954   Mean   :0.0041336
##  3rd Qu.:0.9998   3rd Qu.:0.0024584   3rd Qu.:0.0057164
##  Max.   :1.0000   Max.   :0.0351290   Max.   :0.0708387
##     PersExp           GovtExp             TotExp
##  Min.   :   3.00   Min.   :    10.0   Min.   :    13
##  1st Qu.:  36.25   1st Qu.:   559.5   1st Qu.:   584
##  Median : 199.50   Median :  5385.0   Median :  5541
##  Mean   : 742.00   Mean   : 40953.5   Mean   : 41696
##  3rd Qu.: 515.25   3rd Qu.: 25680.2   3rd Qu.: 26331
##  Max.   :6350.00   Max.   :476420.0   Max.   :482750
kable(head(who_df))
```

Country
LifeExp
InfantSurvival
Under5Survival
TBFree
PropMD
PropRN
PersExp
GovtExp
TotExp
Afghanistan
42
0.835

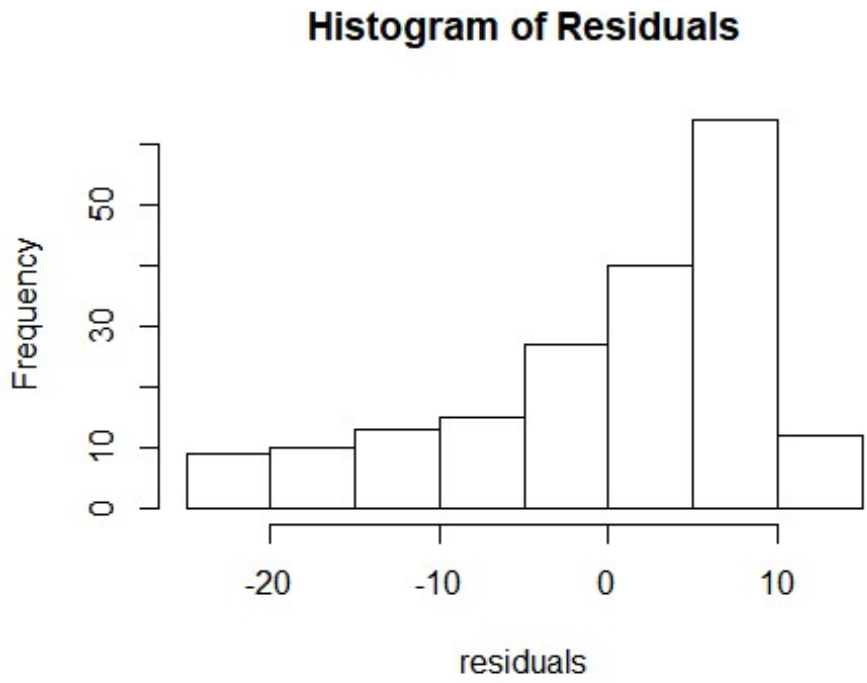
0.743
0.99769
0.0002288
0.0005723
20
92
112
Albania
71
0.985
0.983
0.99974
0.0011431
0.0046144
169
3128
3297
Algeria
71
0.967
0.962
0.99944
0.0010605
0.0020914
108
5184
5292
Andorra
82
0.997
0.996
0.99983
0.0032973
0.0035000
2589
169725
172314
Angola
41
0.846

0.740
0.99656
0.0000704
0.0011462
36
1620
1656
Antigua and Barbuda
73
0.990
0.989
0.99991
0.0001429
0.0027738
503
12543
13046

```r
#scatter plot
plot(who_df$LifeExp, who_df$TotExp, main="Scatterplot",
    xlab="Life Expectancy ", ylab="Total Expenidutres ", pch=19)
#simple linear regression
lm_who_df <- lm(who_df$LifeExp ~ who_df$TotExp)
abline(who_df, col = "red")
```




```r
#residuals
hist(resid(lm_who_df), main = "Histogram of Residuals", xlab = "residuals")
```

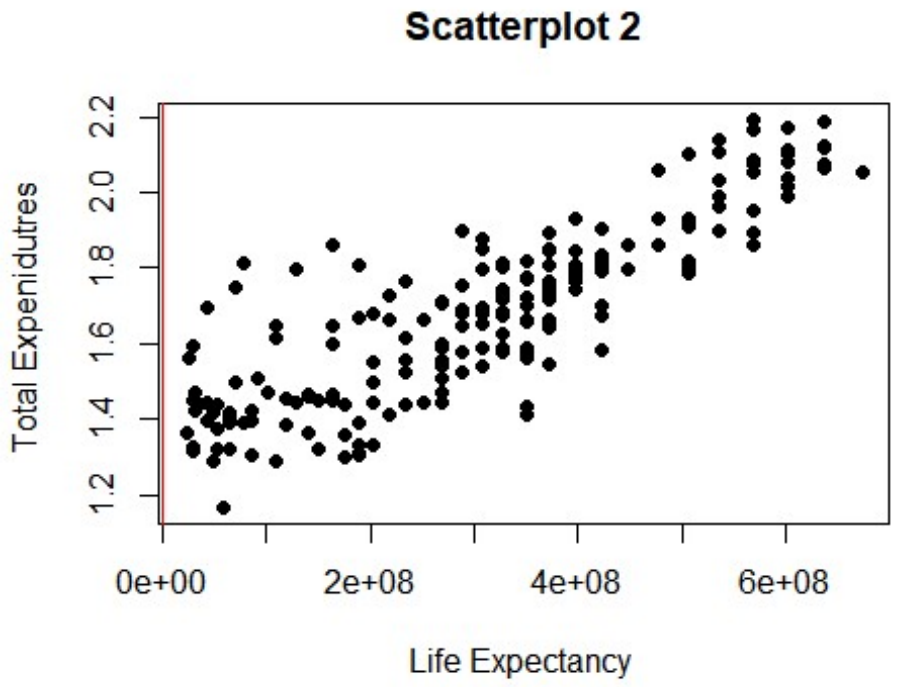


```r
#summary
summary(lm_who_df)
##
## Call:
## lm(formula = who_df$LifeExp ~ who_df$TotExp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.475e+01  7.535e-01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

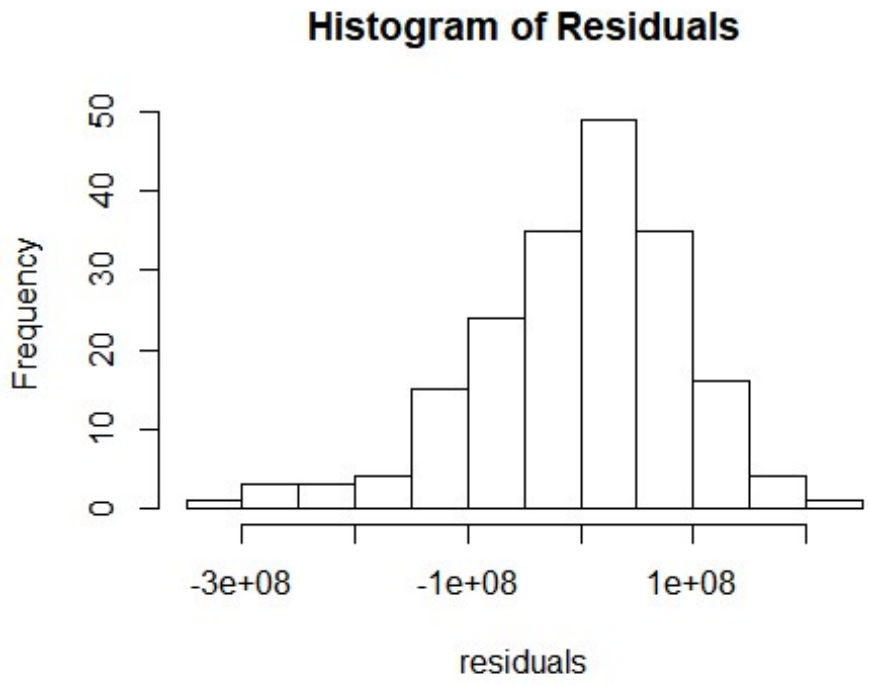
Question 2.

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?"

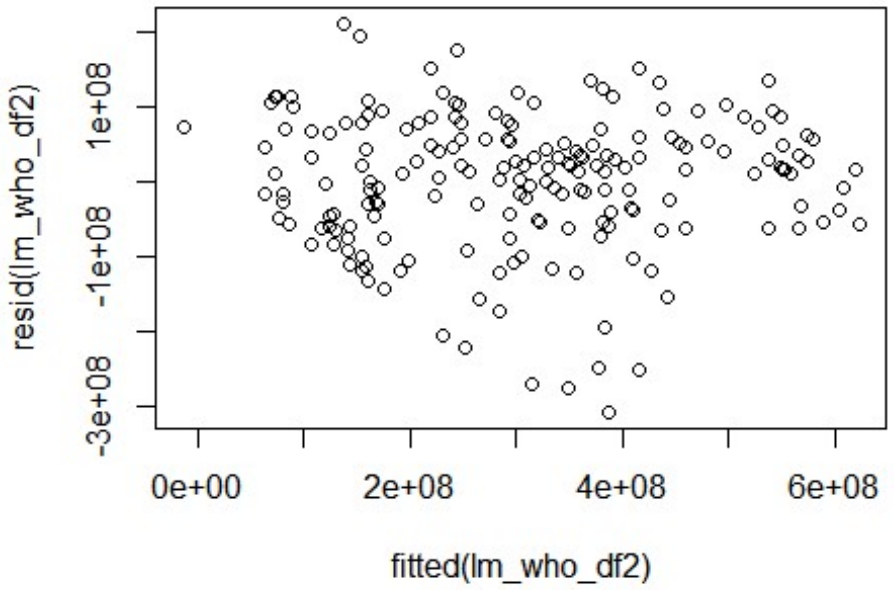
```r
#simple linear regression
x <- who_df$LifeExp^4.6
y <- who_df$TotExp^0.06
lm_who_df2 <- lm(x ~ y)
#scatter plot
plot(x, y, main="Scatterplot 2",
    xlab="Life Expectancy ", ylab="Total Expenidutres ", pch=19)
abline(lm_who_df2, col = "red")
```




```r
#residuals
hist(resid(lm_who_df2), main = "Histogram of Residuals", xlab = "residuals")
```



```r
#summary
summary(lm_who_df2)
##
## Call:
## lm(formula = x ~ y)
##
## 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 ***
## y            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
plot(fitted(lm_who_df2), resid(lm_who_df2))
plot(fitted(lm_who_df2), resid(lm_who_df2))
```



Model2 is highly different and better compared to Model1. Adjusted Rsquare is 72% whereas Model1 is only 25%. There seems to be a good correlation. p-value is less in

Model2 compared to Model1. F-stat is 507 in model2 whereas only 65 in Model1. Residual standard error is high in Model2 and normally distributed in Model2.

Question 3.

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

$$y = -736527910 + 620060216x$$

$$lifeexpectancy = y^(1/4.6)$$

```
le <- function(fc)
{   y <- -736527910 + 620060216 * (fc)
    y <- y^(1/4.6)
    print(y)
}
#Life expectancy when TotExp^.06 =1.5
le(1.5)
## [1] 63.31153
#Life expectancy when TotExp^.06 =2.5
le(2.5)
## [1] 86.50645
```

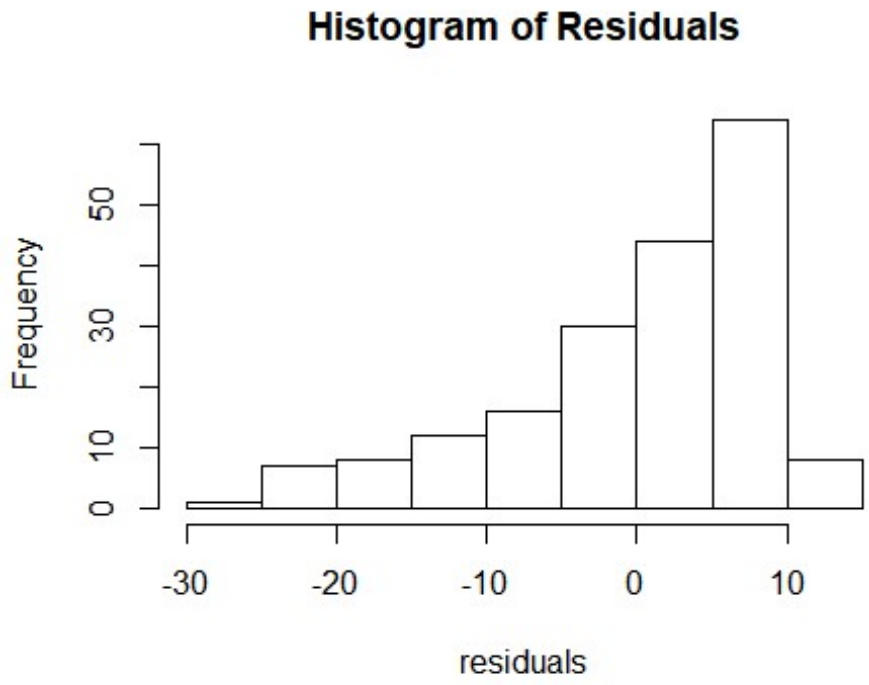
Question 4.

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

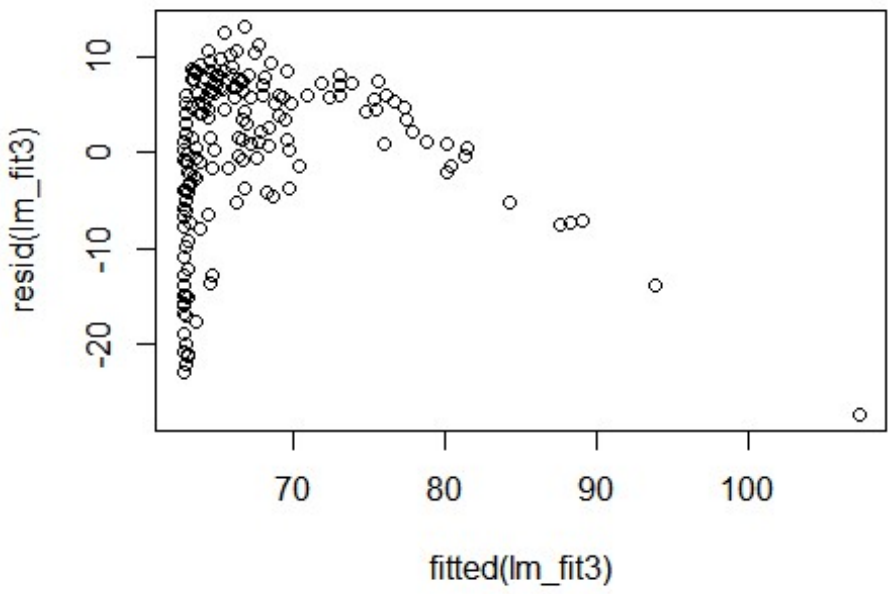
LifeExp = b0+b1 x PropMd + b2 x TotExp +b3 x PropMD x TotExp

```
lm_fit3 <- lm(who_df$LifeExp ~ who_df$PropMD + who_df$TotExp +
who_df$PropMD*who_df$TotExp)
summary(lm_fit3)
##
## Call:
## lm(formula = who_df$LifeExp ~ who_df$PropMD + who_df$TotExp +
##     who_df$PropMD * who_df$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 ***
## who_df$PropMD                  1.497e+03  2.788e+02   5.371 2.32e-07 ***
## who_df$TotExp                  7.233e-05  8.982e-06   8.053 9.39e-14 ***
## who_df$PropMD:who_df$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
hist(resid(lm_fit3), main = "Histogram of Residuals", xlab = "residuals")
```




```
plot(fitted(lm_fit3), resid(lm_fit3))
```



p-value is less than .05. model is statistically significant. F-statistic is 34.49 by adding 3 variables. Based on Rsquare only 35% of the variability can be explained by 3 variables. Correlation is moderate in this case. Residuals is right skewed. So, linear model is not valid.

Question 5.

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

$$LifeExp = 6.277 * 10^1 + 1.497 * 10^3 * PropMD + 7.233 * 10^-5 TotExp - 6.026 * 10^-3 * PropMD * TotExp$$

```
LE <- ( (6.277*10^1) + (1.497*10^3)*.03 + (7.233*10^(-5))*14 - ((6.026*10^(-3))*0.03*14) )
LE
## [1] 107.6785
```

This prediction does not seem realistic, since the total personal and government expenditure is near the minimum, yet life expentancy exceeds that of any country in the dataset.Hence, The forecast age 107.6 is an outlier and seems to be unrealistic. The expenditure is also low.