Multiple Regression Analysis

Jeff Shamp

4/21/2020

Problem 1

```r
who<-read_csv("who.csv")
```

```
## Parsed with column specification:
## cols(
##   Country = col_character(),
##   LifeExp = col_double(),
##   InfantSurvival = col_double(),
##   Under5Survival = col_double(),
##   TBFree = col_double(),
##   PropMD = col_double(),
##   PropRN = col_double(),
##   PersExp = col_double(),
##   GovtExp = col_double(),
##   TotExp = col_double()
## )
```

```r
head(who)
```

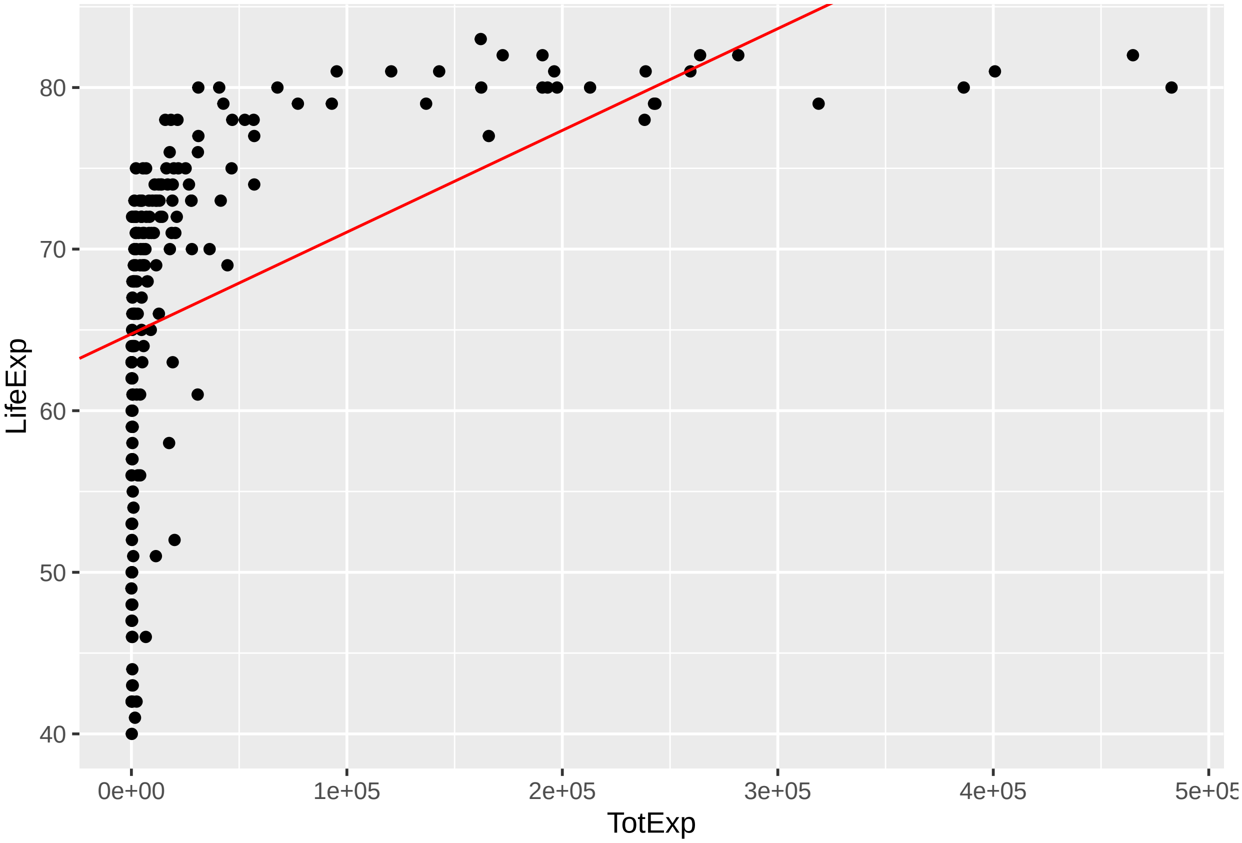
```
## # A tibble: 6 x 10
##   Country LifeExp InfantSurvival Under5Survival TBFree  PropMD  PropRN PersExp
##   <chr>     <dbl>          <dbl>          <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 Afghan~      42          0.835          0.743  0.998 2.29e-4 5.72e-4      20
## 2 Albania      71          0.985          0.983  1.00  1.14e-3 4.61e-3     169
## 3 Algeria      71          0.967          0.962  0.999 1.06e-3 2.09e-3     108
## 4 Andorra      82          0.997          0.996  1.00  3.30e-3 3.50e-3    2589
## 5 Angola       41          0.846          0.74   0.997 7.04e-5 1.15e-3      36
## 6 Antigu~      73          0.99           0.989  1.00  1.43e-4 2.77e-3     503
## # ... with 2 more variables: GovtExp <dbl>, TotExp <dbl>
```

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.

```r
lm_1<-lm(LifeExp ~ TotExp, data=who)
summary(lm_1)
```

```
##
## 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
```

```
lm_1 %>%
  ggplot(aes(x=TotExp, y=LifeExp)) +
  geom_point() +
  geom_abline(slope = lm_1$coefficients[2],
              intercept = lm_1$coefficients[1],
              color="red")
```



The R squared is terrible, but standard error, and p-values are super low and the F-stat is large enought to safely reject that the slope is zero. A simple scatter plot with an abline from the linear regression shows that a linear model is bad idea here. Adding a line to the data, in this case is not a good idea, even if the metrics for rejecting a zero slope is met. Not the right tool for the job.

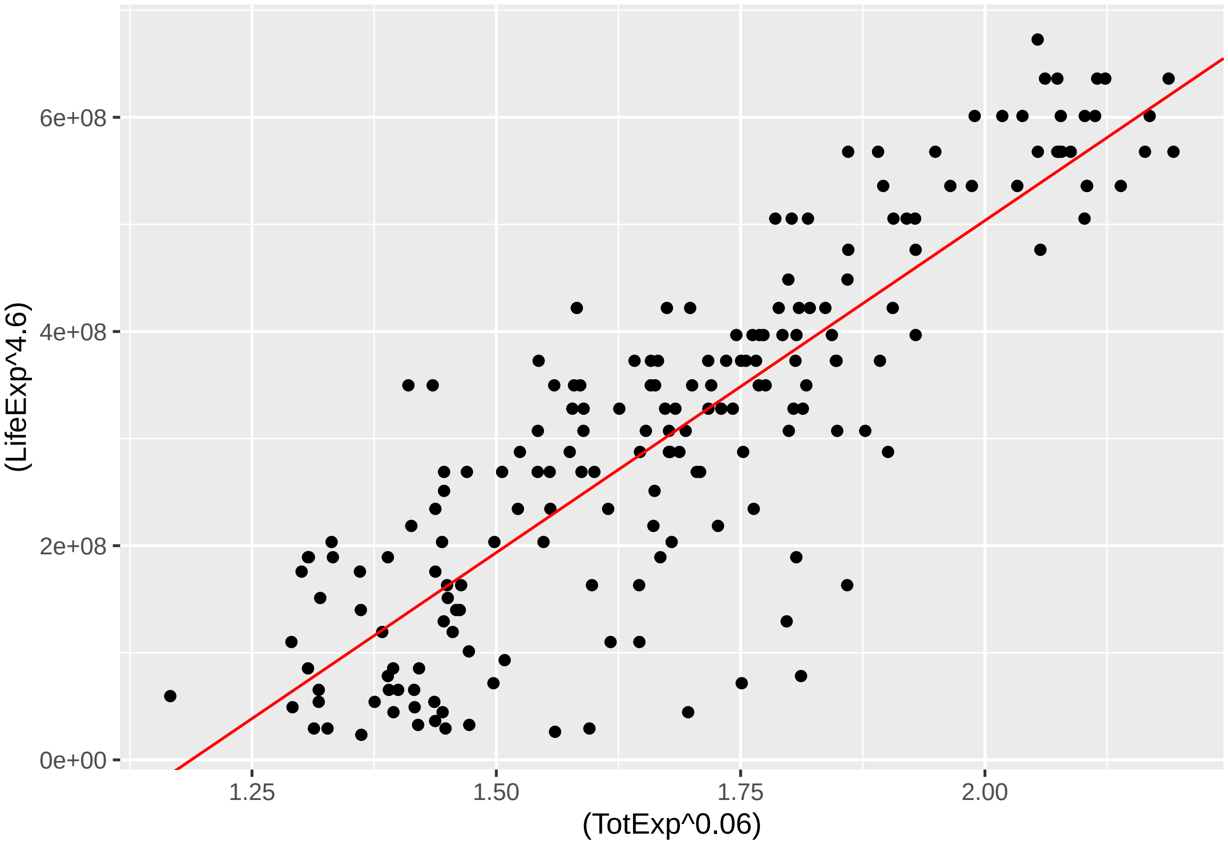
Problem 2

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?"

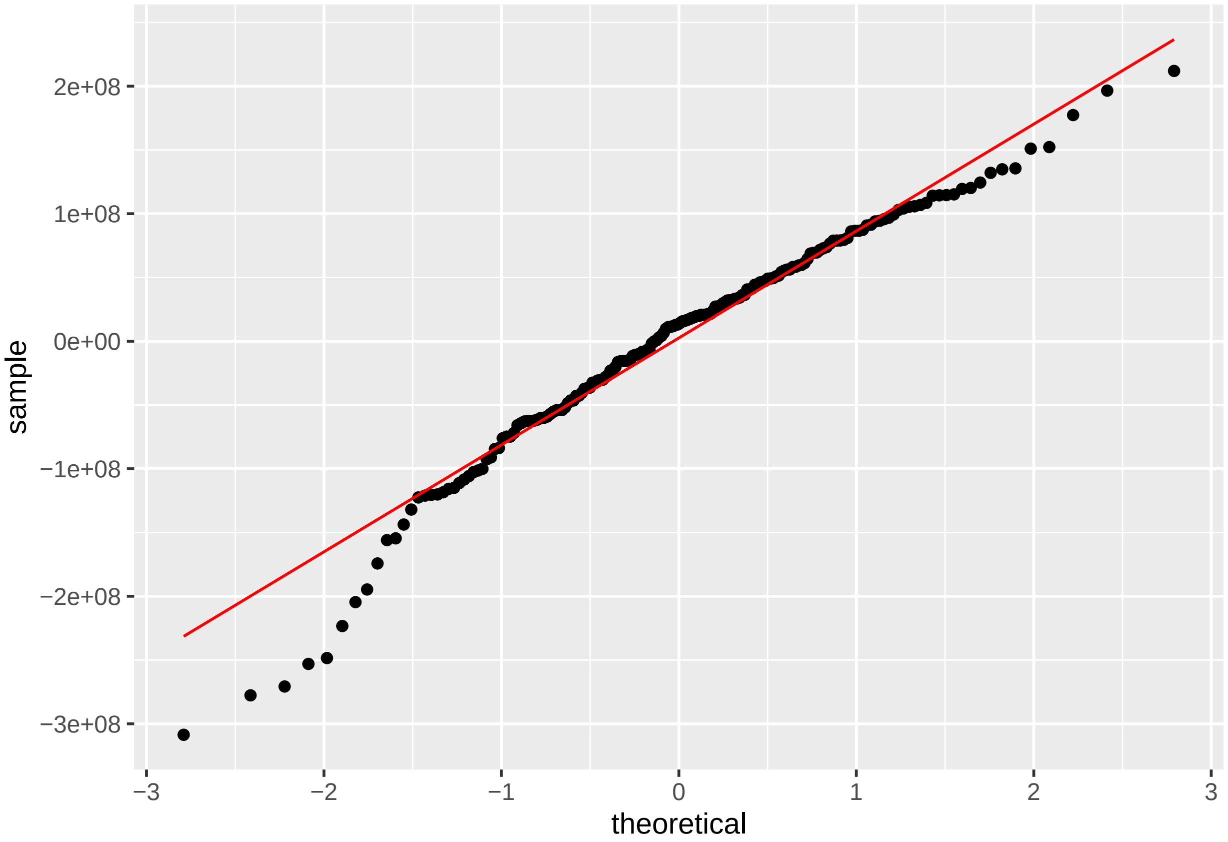
```
lm_2<-lm((LifeExp^4.6) ~ I(TotExp^.06), data=who)
summary(lm_2)
```

```
##
## 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
```

```
who %>%
  ggplot(aes(x=(TotExp^.06), y=(LifeExp^4.6))) +
  geom_point() +
  geom_abline(slope = lm_2$coefficients[2],
              intercept = lm_2$coefficients[1],
              color="red")
```



```
lm_2 %>%
  ggplot(aes(sample = resid(lm_2))) +
  stat_qq() +
  stat_qq_line(color="red")
```



This actually looks linear, which is great. The F-statistic is huge and the it's p-value is very small, so we can confident that slope is non-zero. This is also true the slope and intercepts from the regression. The residuals are huge, but that might be sue to the transformation of the numbers. The R square and adjusted R squared are much better in that they are 0.72, which means the model can explain 72% of the variance in the data.The residual lower tail is very heavy, but otherwise ok.

Problem 3

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

```
life_exp<- (lm_2$coefficients[2]*(1.5)+lm_2$coefficients[1])^(1/4.6)
paste("Life Expectancy when Total Expenses^0.6 = 1.5 =",life_exp, "years")
```

```
## [1] "Life Expectancy when Total Expenses^0.6 = 1.5 = 63.3115334469743 years"
```

```
life_exp<- (lm_2$coefficients[2]*(2.5)+lm_2$coefficients[1])^(1/4.6)
paste("Life Expectancy when Total Expenses^0.6 = 2.5 =",life_exp, "years")
```

```
## [1] "Life Expectancy when Total Expenses^0.6 = 2.5 = 86.5064484844719 years"
```

Spend tax money, live longer.

Problem 4

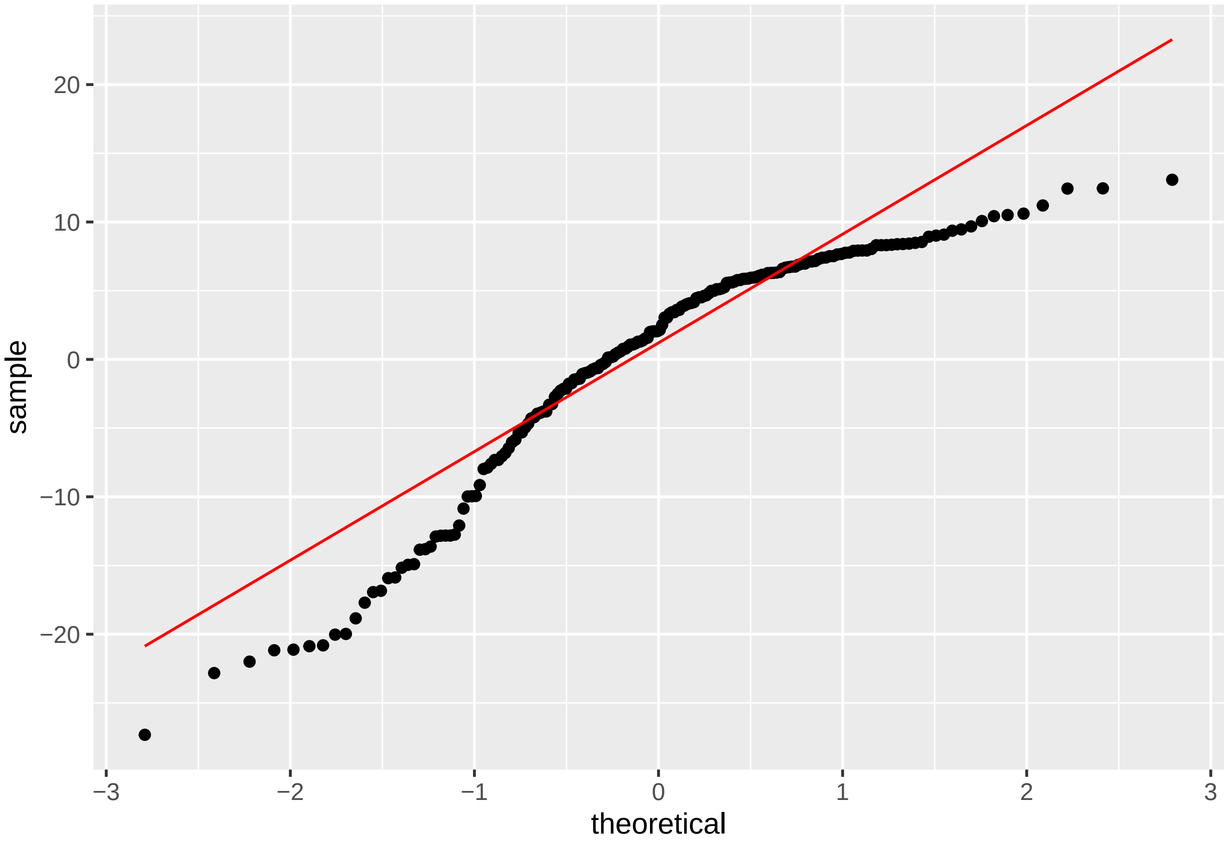
Build the following multiple regression model and interpret the F Statistics, R^2, standard error, and p-values. How good is the model?

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

```r
lm_3<-lm(LifeExp ~ PropMD + TotExp + (PropMD*TotExp), data=who)
summary(lm_3)
```

```
##
## 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
```

```r
lm_3 %>%
  ggplot(aes(sample = resid(lm_3))) +
  stat_qq() +
  stat_qq_line(color="red")
```



This is a better model than the first, but not great. The std error, p-value, F stat all indicate the the slope of the line is non-zero, but the R squared values cannot account more much variance in the data. Further the residuals don't seem to be nearly normal, but makes it hard to provide convinencing evidence that these variables should be used for a linear model.

Problem 5

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

```
lm_3$coefficients
```

```
##   (Intercept)        PropMD        TotExp PropMD:TotExp
##  6.277270e+01  1.497494e+03  7.233324e-05 -6.025686e-03
```

```
#LifeExp ~ PropMD + TotExp + (PropMD * TotExp)
life_exp<- lm_3$coefficients[1]+
          (.03)*lm_3$coefficients[2]+
          (14)*lm_3$coefficients[3]+
          (.03*14)*lm_3$coefficients[4]

paste("Life Expectancy when Total Expenses = 14 and Proportion MD = .03 =",life_exp, "years")
```

```
## [1] "Life Expectancy when Total Expenses = 14 and Proportion MD = .03 = 107.696003708063 years"
```

This is what happens when you over extrapolate the data.