DATA 605 - Homework 11

Manolis Manoli

```
library(ggplot2)
library(psych)
library(dplyr)
library(knitr)
library(tidyr)
library(GGally)
```

World Health Organisation

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.

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?

Solution

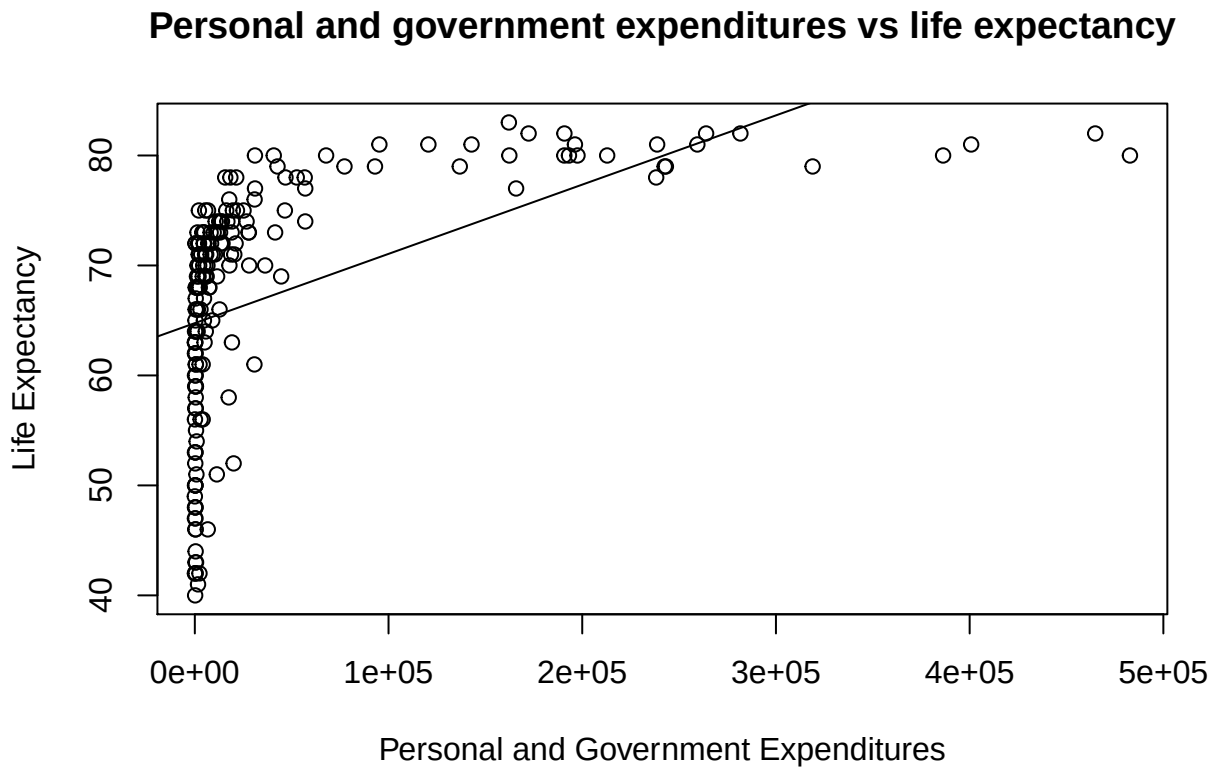
```
url = "https://raw.githubusercontent.com/chilleundso/Data605_CompMath/master/Homework12/who.csv"
df = read.csv(url)
```

Question 1

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.

```
expendVSexpect.lm <- lm(df$LifeExp ~ df$TotExp)
```

```
plot(df$LifeExp ~ df$TotExp,main='Personal and government expenditures vs life expectancy', xlab = 'Pers
abline(expendVSexpect.lm)
```




```
summary(expendVSexpect.lm)
```

```
##
## Call:
## lm(formula = df$LifeExp ~ df$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 ***
## 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
```

Since the p-Value is quite small this model significantly better describes the data than the null hypothesis (average of dependant variable). However looking at the R-squared of 25% we can see that only a quarter of the datas variance is described by the model which is not very satisfying.

Question 2

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?"

```
expendVSexpectRAISED.lm <- lm(df$LifeExp^4.6 ~ I(df$TotExp^0.06))
summary(expendVSexpectRAISED.lm)
```

```
##
## Call:
## lm(formula = df$LifeExp^4.6 ~ I(df$TotExp^0.06))
##
## 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(df$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
```

We can see tt the R-Squared has gone up to 73%, now explaining a far larger amount of the variability and the F-statistic has gone from 65 to 508, showing that the model has increased in significance versus using the average observations.

Question 3

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

```
coef(expendVSexpectRAISED.lm)
```

```
##       (Intercept) I(df$TotExp^0.06)
##        -736527909         620060216
```

```
intercept=coef(expendVSexpectRAISED.lm)[1]
slope=coef(expendVSexpectRAISED.lm)[2]
```

life expectancy when TotExp^.06 =1.5

```
(intercept + slope * 1.5)^(1/4.6)
```

```
## (Intercept)
##    63.31153
```

life expectancy when TotExp^.06 =2.5

```
(intercept + slope * 2.5)^(1/4.6)
```

```
## (Intercept)
##    86.50645
```

Question 4

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

```
multi.lm <- lm(df$LifeExp ~ df$PropMD + df$TotExp + df$PropMD * df$TotExp)
summary(multi.lm)
```

```
##
## Call:
## lm(formula = df$LifeExp ~ df$PropMD + df$TotExp + df$PropMD *
##     df$TotExp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.320  -4.132   2.098   6.540  13.074
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.277e+01  7.956e-01  78.899  < 2e-16 ***
## df$PropMD            1.497e+03  2.788e+02   5.371 2.32e-07 ***
```

```
## df$TotExp            7.233e-05  8.982e-06   8.053 9.39e-14 ***
## df$PropMD: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
```

R-squared is less the previous model and with ~35% not very satisfying. The p-value for each variable is below the significance threashold and the model as a whol has a relatively high significance with a p-value: < 2.2e-16.

Question 5

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

```
coef(multi.lm)
```

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

```
a=coef(multi.lm)[1]
b1=coef(multi.lm)[2]
b2=coef(multi.lm)[3]
b3=coef(multi.lm)[4]
PropMD=.03
TotExp = 14
```

```
a + (b1 * PropMD) + (b2 * TotExp) + (b3 * PropMD * TotExp)
```

```
## (Intercept)
##     107.696
```

Clearly the model is not a great predicter given that an average life expecancy of 108 years is very unrealistic (currently).

https://github.com/chilleundso/Data605_CompMath/tree/master/Homework12