# DATA605: Fundamentals of Computational Mathematics

Assignment 12

Donald Butler

04/24/2022

The attached who.csv dataset contains real-world data from 2008. The variables included follow.

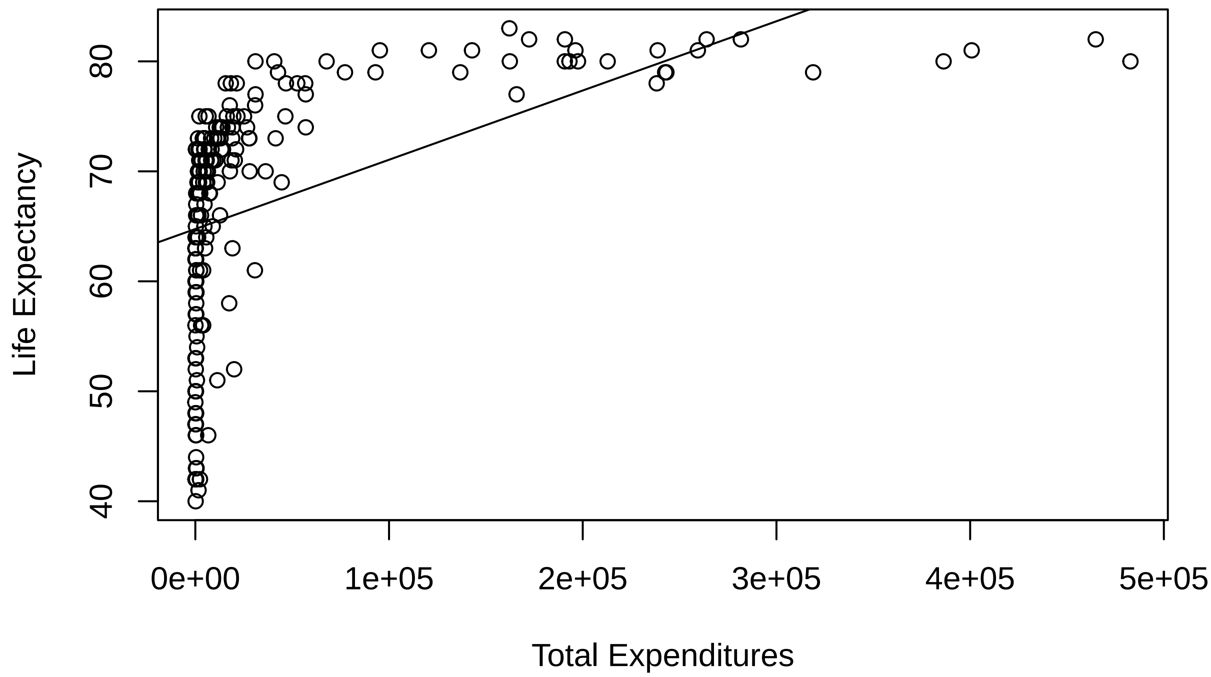
```
Country: name of the country
LifeExp: average life expectancy for the country in years
InfantSurvival: proportion of those surviving to one year or more
Under5Survival: proportion of those surviving to five years or more
TBFree: proportion of the population without TB
PropMD: proportion of the population who are MDs
PersExp: mean personal expenditures on healthcare in US dollars at average exchange rate
GovtExp: mean government expenditures per capital on healthcare, US dollars at average exchange rate
TotExp: sum of personal and government expenditures
```

```r
who_data <- read.csv('who.csv')
```

### Problem 1

Provide a scatterplot of LifeExp ~ TotExp, and run a 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 are met.

```r
who.lm <- lm(formula = LifeExp ~ TotExp, data = who_data)
plot(formula = LifeExp ~ TotExp, data = who_data, xlab = 'Total Expenditures', ylab = 'Life Expectancy')
abline(who.lm)
```




```
summary(who.lm)
```

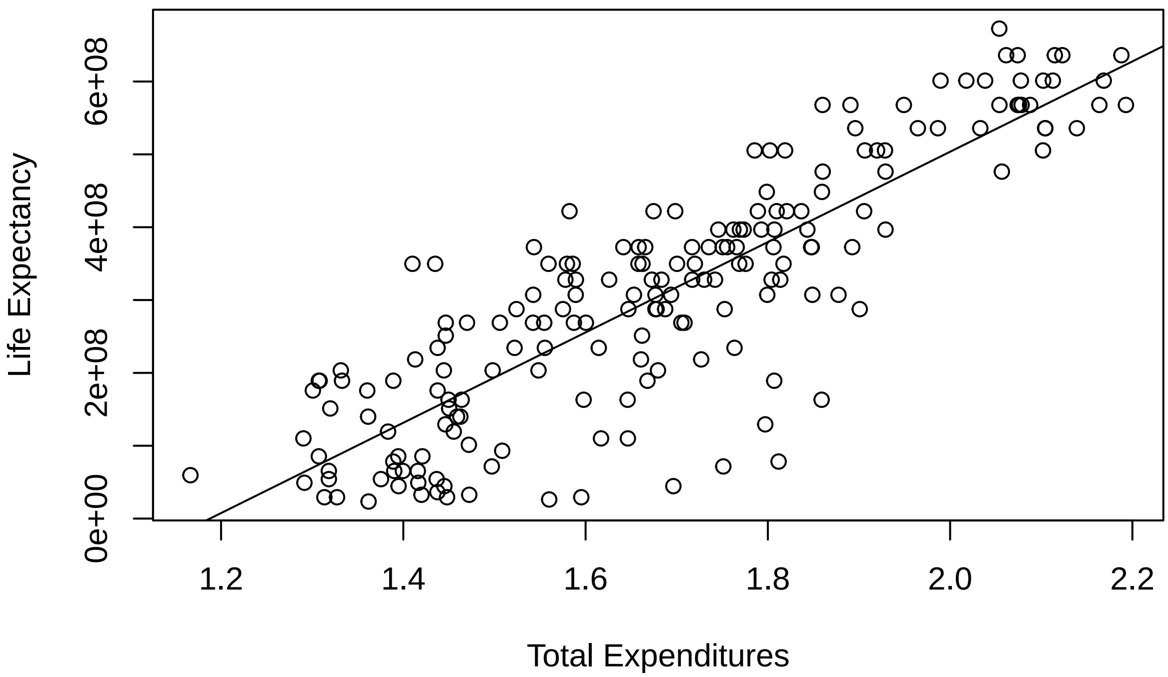
```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = who_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.764  -4.778   3.154   7.116  13.292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.475e+01  7.535e-01  85.933  < 2e-16 ***
##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

With a p-value less than .05, we would reject the null hypotheses and conclude that Life Expectancy is dependent on the total healthcare expenditures. The $R^2$ value indicates that only 25% of the variation in life expectancy is attributed to total expenditures. The standard error indicates that observed values fall $\pm 9.4$ years from the linear regression line.

## 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 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_data$LifeExp2 <- who_data$LifeExp ^ 4.6
who_data$TotExp2 <- who_data$TotExp ^ .06
who.lm2 <- lm(formula = LifeExp2 ~ TotExp2, data = who_data)
plot(formula = LifeExp2 ~ TotExp2, data = who_data, xlab = 'Total Expenditures', ylab = 'Life Expectanc
abline(who.lm2)
```




```
summary(who.lm2)
```

```
##
## Call:
## lm(formula = LifeExp2 ~ TotExp2, data = who_data)
##
## 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 ***
## TotExp2      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
```

The p-value is again less than .05, so we would again reject the null hypotheses and conclude that Life Expectancy is dependent on the total healthcare expenditures. The $R^2$ value indicates that 73% of the variation in $LifeExp^{4.6}$ is attributed to $TotExp^{.06}$. The standard error indicates that observed values fall within 9.05e7 $years^{4.6}$ from the linear regression line.

The second model is better because the p-value is less and the $R^2$ value is higher than in the original model.

## Problem 3

Using the results from 2, forecast life expectancy when $TotExp^{.06} = 1.5$. Then forecast life expectancy when $TotExp^{.06} = 2.5$.

```
(coef(summary(who.lm2))[1,1] + coef(summary(who.lm2))[2,1] * 1.5)^(1/4.6)
```

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

The model predicts a life expectancy of 63.3 years when $TotExp^{.06} = 1.5$.

```
(coef(summary(who.lm2))[1,1] + coef(summary(who.lm2))[2,1] * 2.5)^(1/4.6)
```

```
## [1] 86.50645
```

The model predicts a life expectancy of 86.5 years when $TotExp^{.06} = 2.5$.

## 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?

$$LifeExp = b_0 + b_1 * PropMD + b_2 * TotExp + b_3 * PropMD * TotExp$$

```
who.lm4 <- lm(data = who_data, formula = LifeExp ~ PropMD + TotExp + PropMD * TotExp)
summary(who.lm4)
```

```
##
## Call:
## lm(formula = LifeExp ~ PropMD + TotExp + PropMD * TotExp, data = who_data)
##
## 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 F-statistic and p-values indicate that the variables are providing a statistically significant influence on the life expectancy. The $R^2$ value of 34% is better than the first model, but still leaves 66% of the variation in life expectancy to other factors.

## Problem 5

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

```r
coef(summary(who.lm4))[1,1] + coef(summary(who.lm4))[2,1] * .03 + coef(summary(who.lm4))[3,1] * 14 + coe
```

```
## [1] 107.696
```

The model forecasts the life expectancy to be 107 years, which is about the maximum age a person can live. What is more unrealistic though, is the inputs of 3% of the population are doctors, yet total healthcare expenditures are only $14.