Data 605 - Assignment 11

Jai Jeffryes

4/26/2020

Assignment

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

Variables

- 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.

Load

```
dataset <- read.csv("Dataset_Assignment12.csv")
```

Workflow

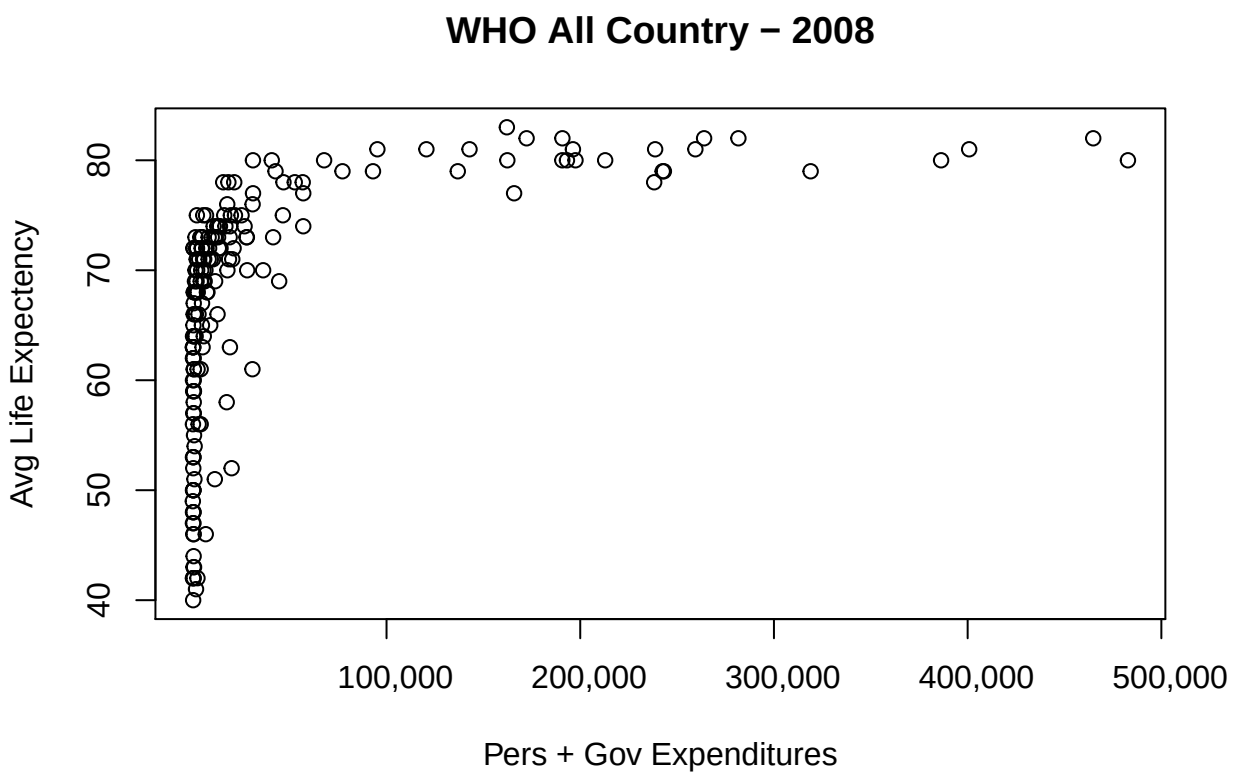
Wrapper for repeated tasks.

```
#function
regressit=function(model,data,level) {
    myreg <- lm(model, data, y=TRUE) #runregression
    r0 <- myreg
    r1 <- summary(myreg) #regression summary
    par(mfrow = c(2,2))
  mydiagplot <- plot(myreg)
    mylist <- list(r0, r1) #,r2,r3,r4,r5,r6,r7,r7a,r9,all)
    return(mylist)
}
```

1. Simple linear regression, first model

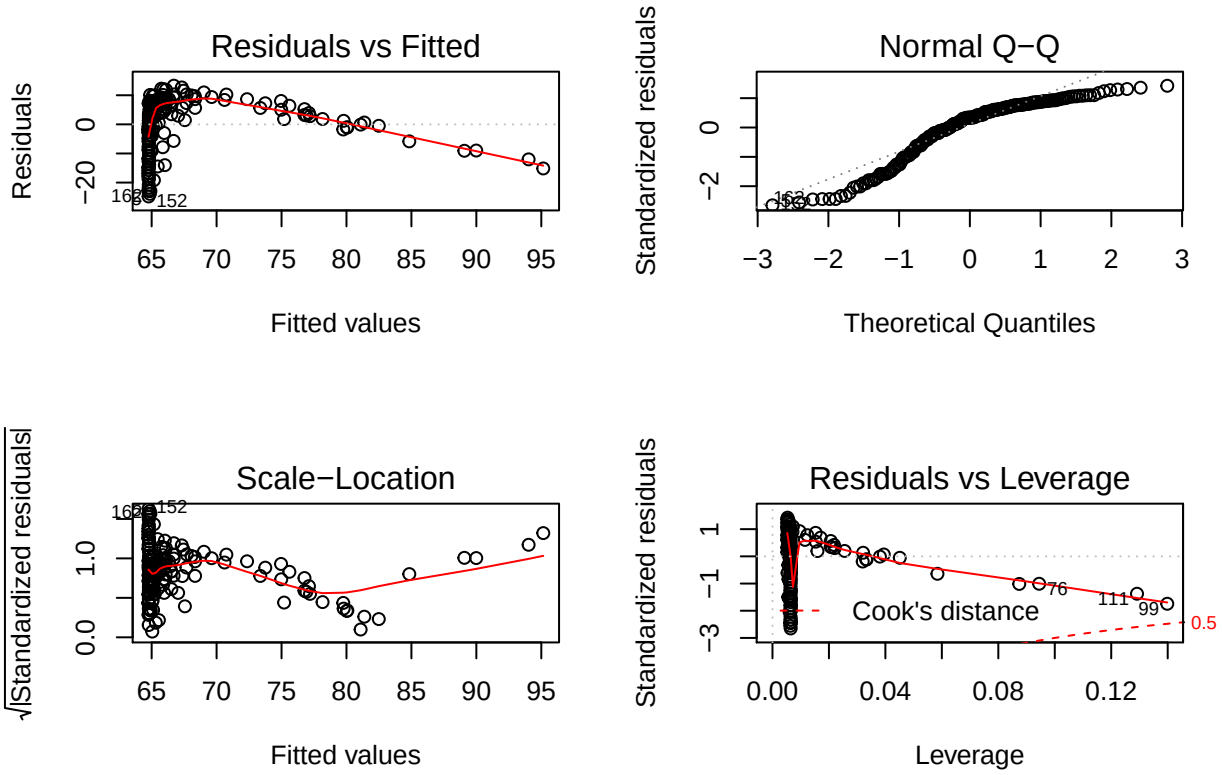
Provide a scatterplot of LifeExp~TotExp, and run simple linear regression. Do not transform the variables. Provide and interpret the F statistics, R^2, standard error,and p-values only. Discuss whether the assumptions of simple linear regression met.

```r
plot(dataset$TotExp, dataset$LifeExp,
     xlab = "Pers + Gov Expenditures",
     ylab = "Avg Life Expectency",
     #type = "h",
     main = "WHO All Country - 2008",
     xaxt = "n")
axis(1, at = seq(100000, 1000000, 100000),
     labels = format(seq(100000,1000000,100000), big.mark=",", scientific=FALSE))
```




```r
#specify model, data, and level
model <- LifeExp~TotExp
data <- dataset
level <- 0.95

fit1 <- regressit(model, data, level)
```




```
fit1[[2]]
```

```
##
## Call:
## lm(formula = model, data = data, y = TRUE)
##
## 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
```

Analysis

- F statistic. The value (65.26) is far from 0 and its p-value is near zero. We reject the null hypothesis, which states the variability in the data is due to chance.

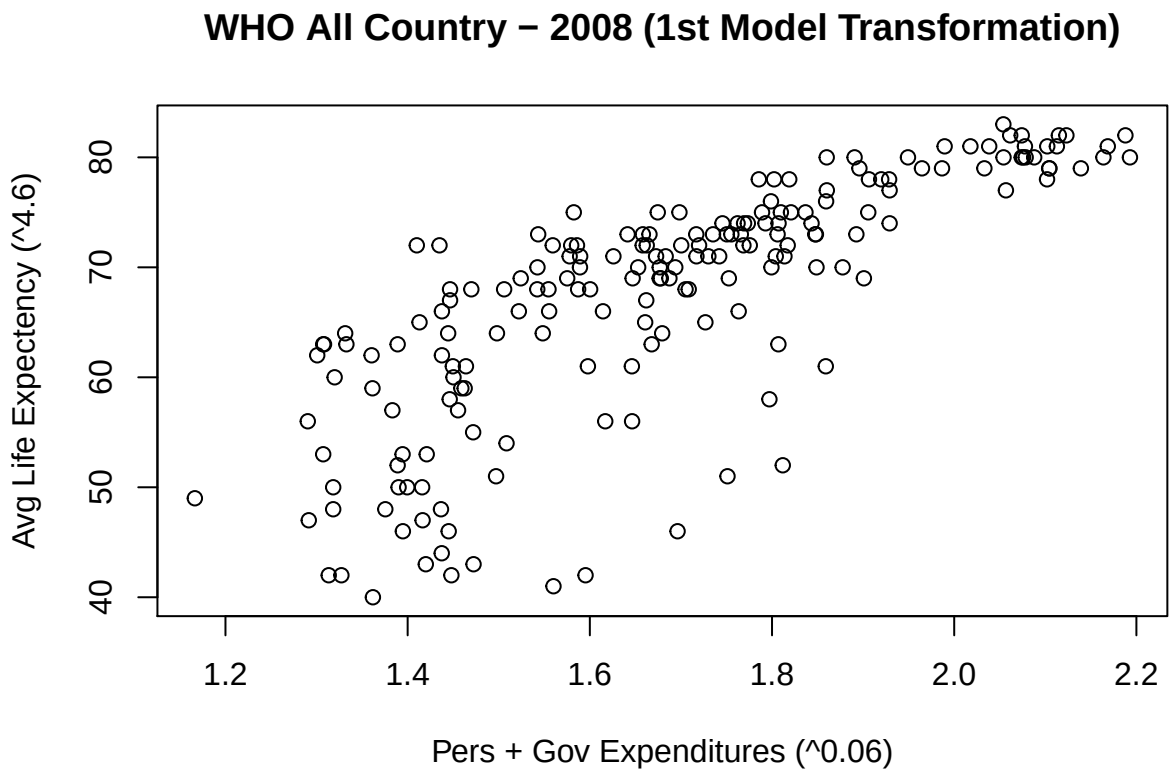
- R^2. The value of 0.2577 means that the model explains 26% of the variability in the data.
- Standard error.
  - The residual standard error indicates the model is off, on average by 9 years.
  - For the coefficient, in order to have reasonable variability of the residuals we want a value of at least 5-10 for the ratio of the coefficient for `TotExp` to standard error. The ratio is 8.08, so there is some variability of the residuals, but it is neither low nor high.
  - The ratio for the intercept is 85.9323159, indicating it is very stable.
- The data do not meet the assumptions of linear regression as the scatterplot and residuals vs. fitted plot reveal high non-linearity.

2. Transformations

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?"

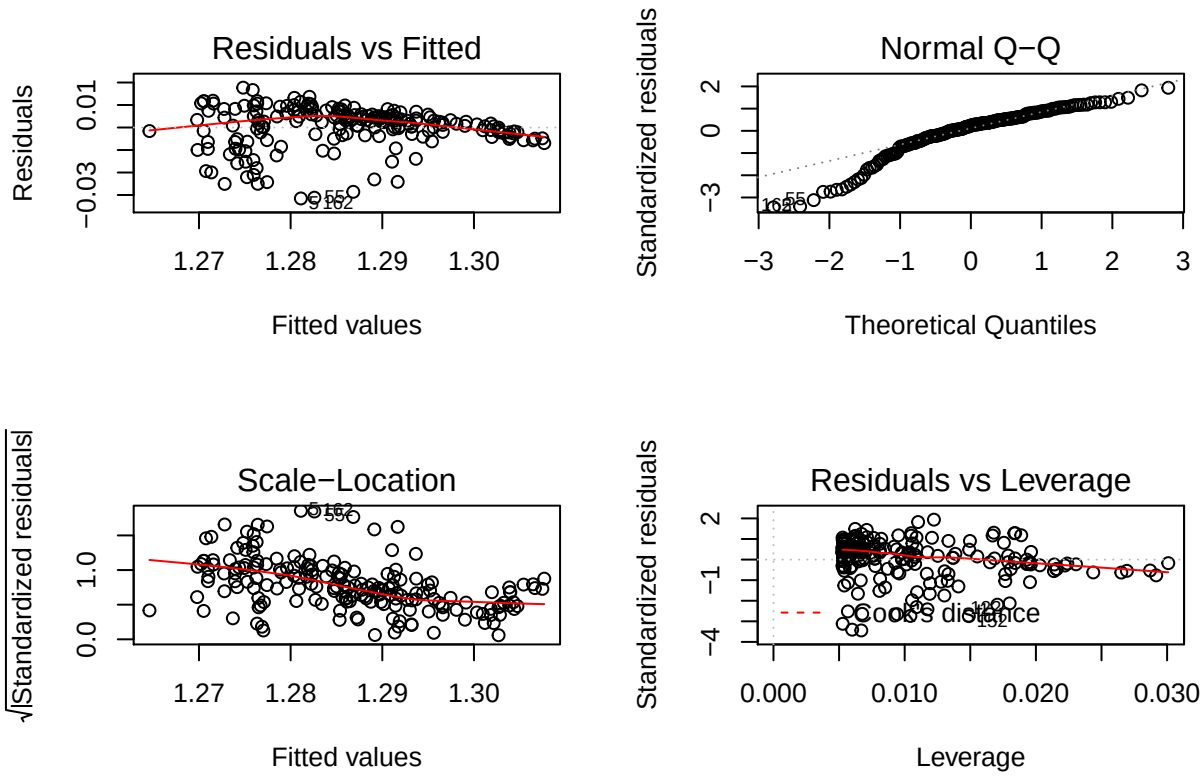
```
dataset2 <- dataset
dataset2$TotExp <- dataset2$TotExp^0.06
dataset2$LifeExp <- dataset2$LifeExp^0.06

plot(dataset2$TotExp, dataset$LifeExp,
     xlab = "Pers + Gov Expenditures (^0.06)",
     ylab = "Avg Life Expectency (^4.6)",
     #type = "h",
     main = "WHO All Country - 2008 (1st Model Transformation)")
```




```
#specify model, data, and level
model <- LifeExp~TotExp
data <- dataset2
level <- 0.95

fit2 <- regressit(model, data, level)
```




```
fit2[[2]]
```

```
##
## Call:
## lm(formula = model, data = data, y = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.031530 -0.003478  0.002165  0.005598  0.017704
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.215667   0.004754  255.73   <2e-16 ***
## TotExp      0.041954   0.002794   15.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.009188 on 188 degrees of freedom
## Multiple R-squared:  0.5453, Adjusted R-squared:  0.5429
## F-statistic: 225.4 on 1 and 188 DF,  p-value: < 2.2e-16
```

Analysis

- F statistic. The value is much further from 0, from 65.26 in the first model to 225.4, and the p-value is

smaller than can be recorded by machine precision. We reject the null hypothesis again, but this time on a stronger basis.
- R^2. We have more than doubled the value, now 0.5453. The model explains 55% of the variability in the data.
- Standard error.
  - The residual standard error is down to 0.01.
  - The ratios of intercept and `TotExp` to standard error are 255.71 and 15.02. Residual error is reduced in this model.
- This model meets assumptions of linear regression. The scatterplot appears linear. The plot of residuals vs. fitted is nearly flat, indicating strong linearity. The Q-Q plot shows strong normal distribution of residuals, except in the region below 1 standard deviation, where non-linearity is still evident.

3. Forecast

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

Note that we report prediction intervals, suitable for point predictions, with a confidence of 0.95, rather than confidence intervals, which are suitable for an average and typically narrower.

```
predict(fit2[[1]],data.frame(TotExp=c(1.5, 2.5)), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 1.278598 1.260397 1.296800
## 2 1.320552 1.301831 1.339273
```

4. Regression model

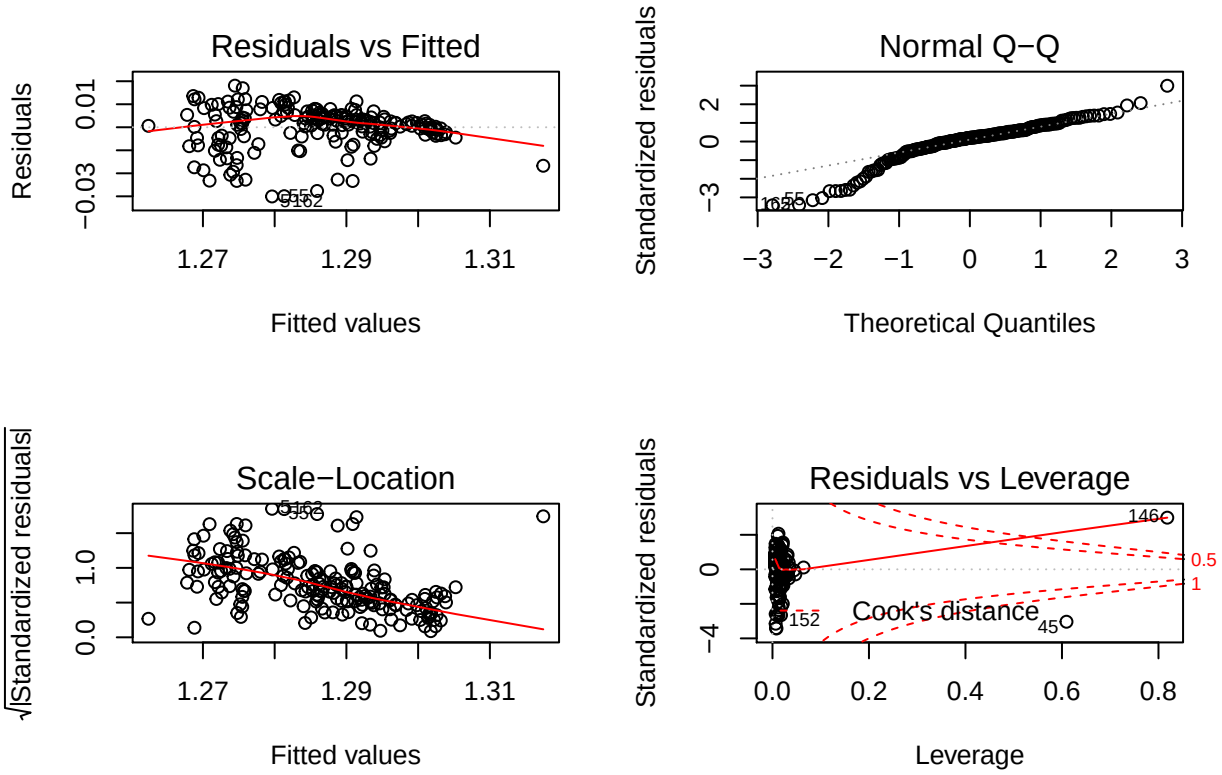
Build the following multiple regression model and interpret the F Statistics, Rˆ2, standard error, and p-values. How good is the model?

$LifeExp = b0 + b1 \times PropMd + b2 \times TotExp + b3 \times PropMD \times TotExp$

Assumption: The intent of the question is to use the transformed data from the previous step, rather than the original data set.

```
#specify model, data, and level
model <- LifeExp~PropMD*TotExp
data <- dataset2
level <- 0.95

fit3 <- regressit(model, data, level)
```




```
fit3[[2]]
```

```
##
## Call:
## lm(formula = model, data = data, y = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.030048 -0.003198  0.001731  0.005011  0.018071
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.211443   0.005059 239.468  < 2e-16 ***
## PropMD         9.393782   2.247219   4.180 4.48e-05 ***
## TotExp         0.043611   0.003009  14.493  < 2e-16 ***
## PropMD:TotExp -4.593767   1.125567  -4.081 6.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.008811 on 186 degrees of freedom
## Multiple R-squared:  0.5863, Adjusted R-squared:  0.5796
## F-statistic: 87.86 on 3 and 186 DF,  p-value: < 2.2e-16
```

Analysis

- F statistic. The F statistic still supports rejection of the null hypothesis. However, its value has moved the wrong direction, from 225.4 in the prior model to 87.86.
- R^2. We have improved the value slightly, from 0.5453 to 0.5863. The model now explains 59% of the variability in the data.
- Standard error.
  - The residual standard error is still 0.01.
  - All of the coefficients have low p values, indicating their influence is not due to chance. The ratios of their values to their standard errors are:
    - `Intercept`: 239.46
    - `PropMD`: 4.18
    - `TotExp`: 14.49
    - `PropMD:TotExp`: -4.08
  - The ratios for `PropMD` and the interaction term `PropMD:TotExp` lie below the desirable minimum range of 5-10. The estimates for these coefficients in this model can vary significantly.
- This model meets assumptions of linear regression. The plot of residuals vs. fitted is not quite as flat as the prior plot. The Q-Q plot appears unchanged, still showing a strong normal distribution of residuals, except in the region below 1 standard deviation, where non-linearity is still evident. The plot for residuals vs. leverage now reveals an outlier, observation with index 146, lying beyond Cook's distance, indicating that it exerts influence on the regression line. That can signal dirty data and merits examination.

5. Forecast

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

```
predict(fit3[[1]], data.frame(PropMD = 0.03, TotExp = 14), interval = "prediction")
```

```
##         fit        lwr       upr
## 1 0.1744331 -0.6103905 0.9592566
```

The prediction does not seem reasonable, as the prediction interval includes the value 0. This is due to the model including in its range, total expenditures near 0. `TotExp` was transformed. Conversion back to the original scale (the 0.6^{th} root) yields a very small expenditure, 81.32. The age predicted and converted to its original scale (the 4.6^{th} root) is 0.68. Models need boundaries on the inputs. Linearity does not extend infinitely in the real world.