Homework12

Jonathan Hernandez

November 11, 2018

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?

- Let's start by loading the data 'who.csv'

```
country_stats_2008 <- read.csv("who.csv")
head(country_stats_2008)
```

```
##               Country LifeExp InfantSurvival Under5Survival  TBFree
## 1         Afghanistan      42          0.835          0.743 0.99769
## 2             Albania      71          0.985          0.983 0.99974
## 3             Algeria      71          0.967          0.962 0.99944
## 4             Andorra      82          0.997          0.996 0.99983
## 5              Angola      41          0.846          0.740 0.99656
## 6 Antigua and Barbuda      73          0.990          0.989 0.99991
##        PropMD      PropRN PersExp GovtExp TotExp
## 1 0.000228841 0.000572294      20      92    112
## 2 0.001143127 0.004614439     169    3128   3297
```

```
## 3 0.001060478 0.002091362     108    5184   5292
## 4 0.003297297 0.003500000    2589  169725 172314
## 5 0.000070400 0.001146162      36    1620   1656
## 6 0.000142857 0.002773810     503   12543  13046
```

```
str(country_stats_2008)
```

```
## 'data.frame':    190 obs. of  10 variables:
##  $ Country       : Factor w/ 190 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ LifeExp       : int  42 71 71 82 41 73 75 69 82 80 ...
##  $ InfantSurvival: num  0.835 0.985 0.967 0.997 0.846 0.99 0.986 0.979 0.995 0.996 ...
##  $ Under5Survival: num  0.743 0.983 0.962 0.996 0.74 0.989 0.983 0.976 0.994 0.996 ...
##  $ TBFree        : num  0.998 1 0.999 1 0.997 ...
##  $ PropMD        : num  2.29e-04 1.14e-03 1.06e-03 3.30e-03 7.04e-05 ...
##  $ PropRN        : num  0.000572 0.004614 0.002091 0.0035 0.001146 ...
##  $ PersExp       : int  20 169 108 2589 36 503 484 88 3181 3788 ...
##  $ GovtExp       : int  92 3128 5184 169725 1620 12543 19170 1856 187616 189354 ...
##  $ TotExp        : int  112 3297 5292 172314 1656 13046 19654 1944 190797 193142 ...
```

```
dim(country_stats_2008)
```

```
## [1] 190  10
```

For 1

```
lm_lifeexp_totexp <- lm(LifeExp ~ TotExp, data = country_stats_2008)
summary(lm_lifeexp_totexp)
```

```
##
## Call:
## lm(formula = LifeExp ~ TotExp, data = country_stats_2008)
##
## 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
```

- F-statistic: The F-statistic is a measure of how statistically significant or how R^2 is significant or not. It's the ratio of the sum of squared residuals (SSR) and sum of squared errors (SSE) that is (SSR/SSE). If the p-value is low for a F-statistic, it means adding more features contributes to a better fit modeling rather than having a intercept only model. The F-statistic is a bit high so our variable TotExp helps to improve the model.

- The R^2 value tells us how close the data are to the fitted regression line. It is bound from 0 to 1. The closer to 1, the more the model's regression line fit the data. Here the R^2 is pretty low and thus our model's regression line doesn't fit the data well.
- The Residual Standard error is the standard deviation of the residuals. The residual standard error is about 9.371 and means that it is quite far off from the data points. The Standard error in the variables are the standard deviation estimate of the variable coefficient. The coefficient standard errors are small and the coefficients are good to use in the model.
- The only condition that is not met would be the R^2 squared value and Residual Standard error. The R^2 is low and the residual standard error is too off from the data points so simple linear regression will not work.

For 2.

```
LifeExp2 <- country_stats_2008$LifeExp^4.6
TotalExp2 <- country_stats_2008$TotExp^.06
lm_lifeexp_totalexp2 <- lm(LifeExp2 ~ TotalExp2, data = country_stats_2008)
summary(lm_lifeexp_totalexp2)
```

```
##
## Call:
## lm(formula = LifeExp2 ~ TotalExp2, data = country_stats_2008)
##
## 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 ***
## TotalExp2    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
```

- F-statistic is higher than before with a p-value for it small so we can say that having TotExp raised to the 4.6th power is good to have in this model.
- The R^2 value is higher than before and more closer to 1 so the fitted regression line constitues to a good amount of the data.
- The residual standard error is much higher than before but that is expected as we scaled the explanatory variable exponentially.
- The p-values in the coefficients are also very low so that means that each coefficient is a good fit for the regression model.
- This model does better describe the data with its R^2 value but the values in the coefficients and standard error can be a bit misleading. Overall this model is preferred than the first one.

For 3)

- Let's use the predict function with our model from 2)

```
life_expectancy_predict <- predict(lm_lifeexp_totalexp2,data.frame(TotalExp2 = 1.5))
life_expectancy_predict
```

```
##         1
## 193562414
```

```
life_expectancy_predict <- predict(lm_lifeexp_totalexp2,data.frame(TotalExp2 = 2.5))
life_expectancy_predict
```

```
##         1
## 813622630
```

For 4)

```
lm_lifeexp_multreg <- lm(LifeExp ~ (PropMD + TotExp + (PropMD*TotExp)),
                         data = country_stats_2008)
summary(lm_lifeexp_multreg)
```

```
##
## Call:
## lm(formula = LifeExp ~ (PropMD + TotExp + (PropMD * TotExp)),
##     data = country_stats_2008)
##
## 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
```

- F-statistic: 34.49 is smaller than before as we have more attributes and the p-value is small so the terms used are better than doing the intercept only model.
- $R^2_{adj} = 0.3574$ is small so the fitted line doesn't contribute to a good amount of the data. (only about 35% of the fitted line estimates the data well)
- The Standard Error of coefficients
  - Intercept: 0.7956 standard deviations from the true intercept value
  - TotExp: 8.982e-06 standard deviations from the true value of the coefficient TotExp

  - PropMD: 278.8 standard deviations from the true value of the coefficient PropMD
  - PropMD*TotExp: 1.472e-03 standard deviations from the true value of the coefficient

- The PropMD coefficient estimate is rather large compared to the others.
- p-values of coefficients:
  - Intercept, TotExp, PropMD, PropMD*TotExp are all below 0.001 which means that we can reject the null hypothesis (coefficients are zero) and accept the alternative hypothesis (coefficients !=0) This tells us that the features are to be included in our model.

For 5)

```
life_expectancy_predict <- predict(lm_lifeexp_multreg,
                                   data.frame(TotExp = 14, PropMD = 0.03))
life_expectancy_predict
```

```
##       1
## 107.696
```

- This value is not realistic as it is not common for someone to be expected to live over 107 years. Few people live to be that long and the prediction doesn't account other countries and other factors.