MINI PROJECT LINEAR REGRESSION

Project Summary

Humberto Carvalho – February 2017

1ST EXERCISE - LEAST SQUARES REGRESSION

0. INTRODUCTION

The 1st Exercise, Least Squares Regression, has the goal to fit a model predicting the energy consumed per capita (energy) versus the percentage of residents living in the metropolitan areas (metro).

For this purpose, we use the **states.rds** data set.

First of all, we clean the data, removing the null data. The cleaned data is save in a new data set called **states.data_clean**.

1. EXAMINE/PLOT THE DATA BEFORE FITTING THE MODEL

A subset from states.data.clean is created with all rows for the variables:

- independent variable metro (percentage of residents living in metropolitan areas), and
- dependent variable, energy (energy consumed per capita).

The subset is named **sts.energy.metro**.

In order to analyse the data, we summarize both variables using the summary() and cor() functions (see results below).

From summary(), we get the information that the minimum percentage of residents in metropolitan areas is about 20%, and the variation of the energy consumed per capita is from 200 btu until about 786 btu. The mean is about 340 btu.

Based on the result coming from cor() function, the correlation between energy and metro is closer 0 than 1. Hence, it means that the linear relationship between these two variables is poor.

```
> summary(sts.energy.metro)
     metro            energy     
 Min.   : 20.40   Min.   :200.0  
 1st Qu.: 47.92   1st Qu.:287.0  
 Median : 67.55   Median :320.0  
 Mean   : 64.31   Mean   :343.6  
 3rd Qu.: 81.62   3rd Qu.:362.5  
 Max.   :100.00   Max.   :786.0  
> #
> # Correlating between metro and energy
> cor(sts.energy.metro)
            metro     energy
metro   1.0000000 -0.3116753
energy -0.3116753  1.0000000
```

2. Print and interpret the model 'summary'

From the residuals, its distribution is not symmetrical, meaning that the model predicts some points that are far away from the actual observed points.

The expected value of energy consumed per capita, when we consider the average percentage of people living in metropolitan areas, is about 449 btu. The energy consumed by 1% of people in metropolitan areas is almost 1.7 units and it decreases when the population increases.

```
> summary(energy.metro.model)

Call:
lm(formula = energy ~ metro, data = states.data_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-179.17  -54.21  -21.64   15.07  448.02 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 449.8382    50.4472   8.917 1.37e-11 ***
metro        -1.6526     0.7428  -2.225    0.031 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 112.3 on 46 degrees of freedom
Multiple R-squared:  0.09714,	Adjusted R-squared:  0.07751 
F-statistic: 4.949 on 1 and 46 DF,  p-value: 0.03105
```

The Residual Standard Error is 112.3, it means a percentage error about 25%. Therefore, any prediction would be off by 25%.

The R-squared statistic (R^2) parameter is 0.09714, it means that only about 9.71% of the variance found in the dependent variable energy consumed per capita can be explained by the percentage of people living in metropolitan areas, i.e., the independent variable (predictor).

Concerning the F-statistic indicator (4.949), it is not far from 1, also meaning that the relationship between energy and metro is poor.

Based on above results, we can conclude that the studied model is not featured by a good linear relationship between its variables. We have to improve it, adding more variables.

3. `PLOT' THE MODEL TO LOOK FOR DEVIATIONS FROM MODELLING ASSUMPTIONS

The obtained plots reinforce the above conclusions regarding the need to improve the model. Identifying the available data in the dataset **states.dta**, we have decided adding some predictors to the model as follows:

- pop (1990 population),
- toxic (Per capita toxics released, lbs), and
- green (Per capita greenhouse gas, tons).

```
> summary(sts.energy.metro.toxic.green)
     metro            toxic              green            energy     
 Min.   : 20.40   Min.   :  1.810   Min.   : 11.76   Min.   :200.0  
 1st Qu.: 47.92   1st Qu.:  7.232   1st Qu.: 16.98   1st Qu.:287.0  
 Median : 67.55   Median : 11.705   Median : 21.38   Median :320.0  
 Mean   : 64.31   Mean   : 17.544   Mean   : 25.11   Mean   :343.6  
 3rd Qu.: 81.62   3rd Qu.: 21.363   3rd Qu.: 26.34   3rd Qu.:362.5  
 Max.   :100.00   Max.   :101.280   Max.   :114.40   Max.   :786.0  
> cor(sts.energy.metro.toxic.green)
            metro      toxic      green     energy
metro   1.0000000 -0.1848052 -0.4111107 -0.3116753
toxic  -0.1848052  1.0000000  0.2622973  0.5985974
green  -0.4111107  0.2622973  1.0000000  0.7706181
energy -0.3116753  0.5985974  0.7706181  1.0000000
```

```
> summary(energy.metro.toxic.green.model)

Call:
lm(formula = energy ~ metro + toxic + green, data = states.data_clean)

Residuals:
     Min       1Q   Median       3Q      Max 
-179.311  -31.415   -4.114   17.108  191.943 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 160.5506    37.4912   4.282 9.87e-05 ***
metro         0.2437     0.4273   0.570    0.571    
toxic         2.6691     0.4730   5.643 1.13e-06 ***
green         4.7992     0.5819   8.247 1.79e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.67 on 44 degrees of freedom
Multiple R-squared:  0.7644,	Adjusted R-squared:  0.7483 
F-statistic: 47.58 on 3 and 44 DF,  p-value: 7.305e-14
```

With this combination, we have found out a better model:

- the distribution of residuals is now much more symmetrical than the previous version;
- the Residual Standard Error is 58.67, it means a percentage error about 36.5%;
- the R^2 is 0.7644 (76.4%);
- the F-statistics indicator is now 47.58.

2nd Exercise - interactions and factors

1. ADD ON TO THE REGRESSION EQUATION THAT YOU CREATED IN EXERCISE 1 BY GENERATING AN INTERACTION TERM AND TESTING THE INTERACTION

We add two interaction terms, toxic and green. Based on some previous tests, we conclude that these two variables could significantly be the most relevant for the consumed energy.

```
> summary(energy.metro.by.toxic.green.model)

Call:
lm(formula = energy ~ metro * toxic * green, data = states.data_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-76.959 -24.842  -2.734  24.212 184.494 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        285.004138  78.498396   3.631 0.000794 ***
metro               -1.374970   1.261691  -1.090 0.282331    
toxic                1.541243   3.728558   0.413 0.681549    
green               -0.232294   3.346150  -0.069 0.945000    
metro:toxic         -0.076904   0.061354  -1.253 0.217317    
metro:green          0.086708   0.064150   1.352 0.184085    
toxic:green          0.023237   0.128579   0.181 0.857498    
metro:toxic:green    0.002302   0.002150   1.071 0.290654    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.72 on 40 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8399,	Adjusted R-squared:  0.8119 
F-statistic: 29.99 on 7 and 40 DF,  p-value: 5.342e-14
```

2. TRY ADDING REGION TO THE MODEL. ARE THERE SIGNIFICANT DIFFERENCES ACROSS THE FOUR REGIONS?

In fact, the four regions have important differences across themselves. For instance, the consumed energy in South region varies with all variables together through an opposing way when compared with the N. East and Midwest regions.

```
> summary(energy.metro.by.toxic.green.region.model)

Call:
lm(formula = energy ~ metro * toxic * green + region, data = states.data_clean)

Residuals:
    Min      1Q  Median      3Q     Max 
-75.803 -27.276  -1.792  20.394 172.312 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       291.051239  77.537244   3.754 0.000598 ***
metro              -1.018579   1.297758  -0.785 0.437523    
toxic               0.927627   3.704099   0.250 0.803638    
green               0.333805   3.325149   0.100 0.920578    
regionN. East     -32.442871  25.517875  -1.271 0.211529    
regionSouth        18.570268  20.708755   0.897 0.375658    
regionMidwest     -10.783173  23.387708  -0.461 0.647453    
metro:toxic        -0.081172   0.061355  -1.323 0.193956    
metro:green         0.063207   0.066772   0.947 0.349973    
toxic:green         0.017170   0.129426   0.133 0.895178    
metro:toxic:green   0.002724   0.002186   1.246 0.220514    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 49.38 on 37 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8596,	Adjusted R-squared:  0.8217 
F-statistic: 22.66 on 10 and 37 DF,  p-value: 7.618e-13
```