MINI PROJECT LINEAR REGRESSION

Project Summary

Humberto Carvalho – March 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 saved 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).

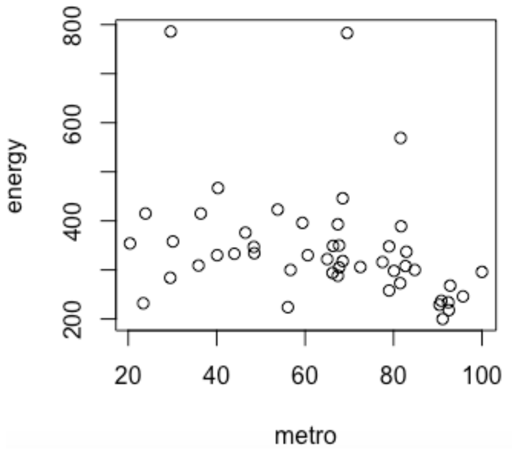
The subset is named **sts.energy.metro**.

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 – linear regression

```
> 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
```

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.

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

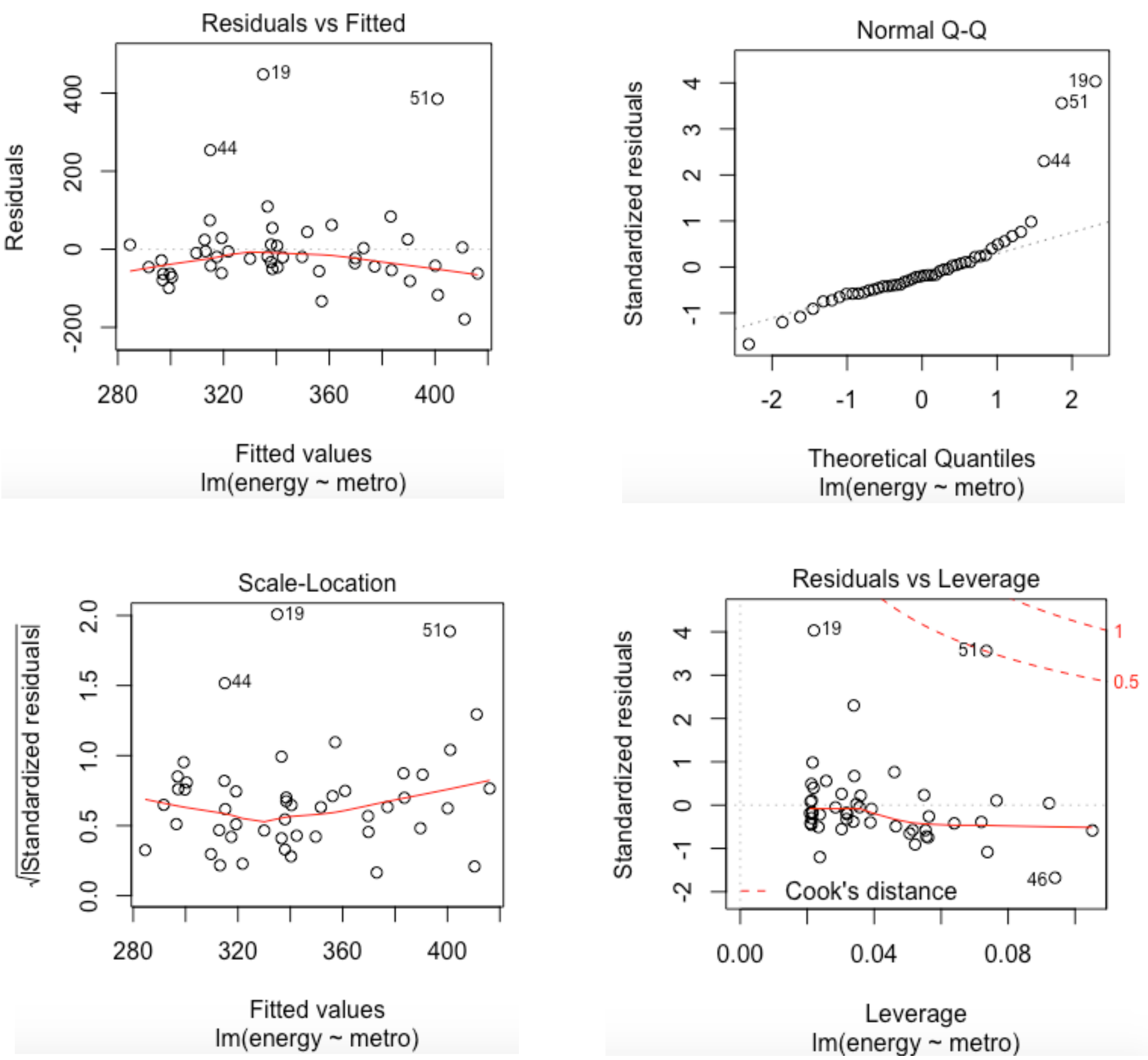
The R-squared statistic (R^2) parameter is 0.09714, it means that only about 9.71% of the variance found in the dependent variable energy consumed per capita can be explained by the percentage of people living in metropolitan areas, i.e., the independent variable (predictor).

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 other variables.

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

Analysing the below diagnostics plots we reach some additional conclusions as follows.



The Residuals vs Fitted plot shows if residuals have non-linear patterns. The residuals are not spread equally around the red line, meaning some lacking of linear relationship.

The Normal Q-Q plot does not show a straight line evidencing some deviations from it. Therefore, it means that the residuals are not normally distributed.

Regarding the Scale-Location plot, it shows a no horizontal line with an angle step and the residuals spreading on a unequally way.

Finally, the last plot, the Residuals vs Leverage, can help to find some cases that could be influent for the linear regression analysis. The cases outside of the Cook's distance are influential to the regression results.

Based on those outcomes, we have to go back and rethink the model.

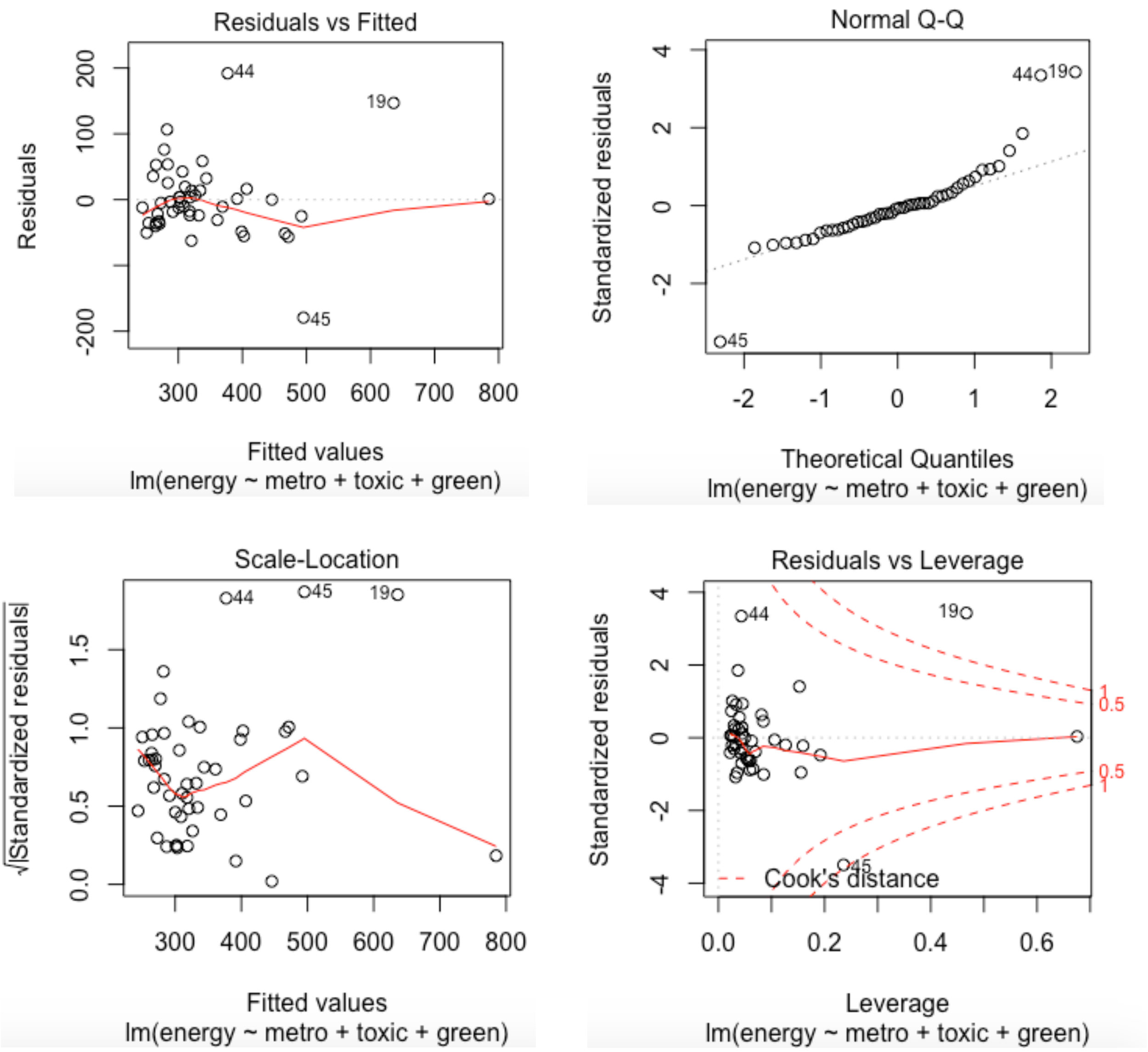
4. Improving the model with new predictors

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:

- 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:

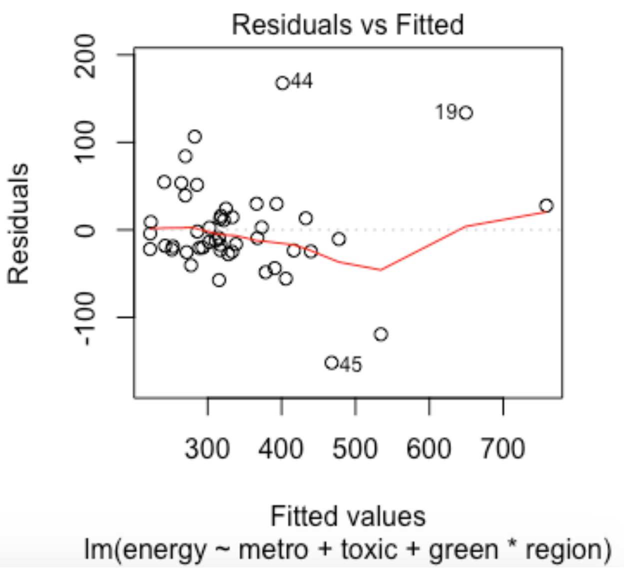
- the distribution of residuals is now much more symmetrical than the previous version;
- the Residual Standard Error is 58.67, it means a percentage error about 36.5%;
- the R^2 is 0.7644 (76.4%);
- the F-statistics indicator is now 47.58.

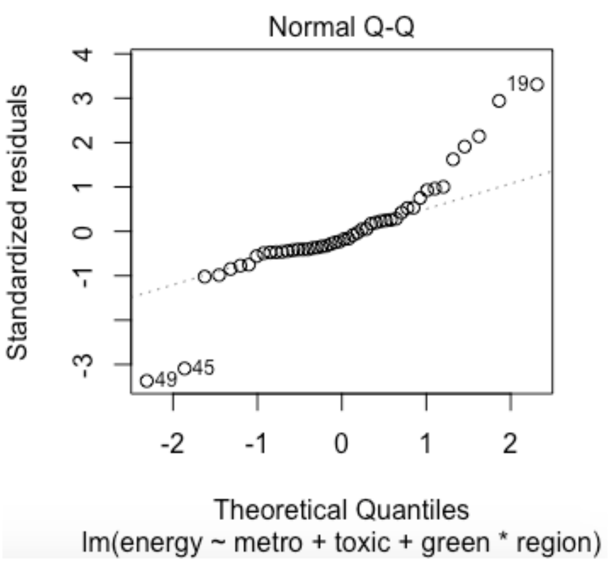
2nd Exercise - interactions and factors

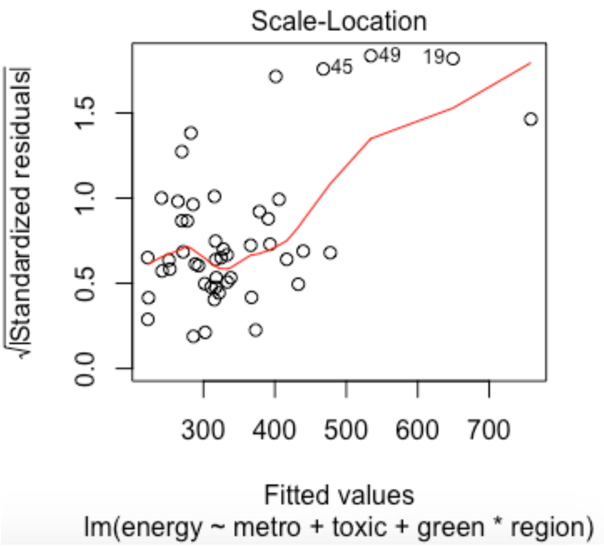
1. Add on to the regression equation that you created in exercise 1 by generating an interaction term and testing the interaction

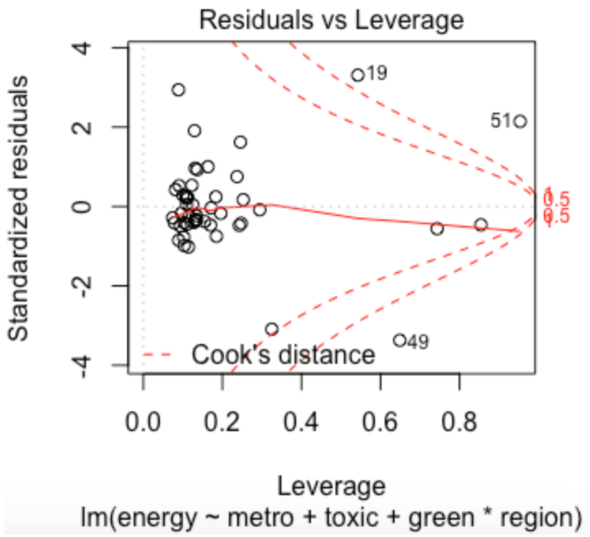
We add one interaction term, the categorical variable region, to the previous model.

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

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

Residuals:
    Min      1Q  Median      3Q     Max 
-151.78  -23.23   -9.94   18.14  167.71 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         157.2399    44.2931   3.550 0.001046 ** 
metro                 0.3557     0.4417   0.805 0.425738    
toxic                 2.2669     0.5301   4.276 0.000123 ***
green                 4.6566     0.6916   6.733 5.66e-08 ***
regionN. East       -94.6585   119.9182  -0.789 0.434800    
regionSouth         -19.7798    52.5118  -0.377 0.708512    
regionMidwest         8.3120    50.3941   0.165 0.869866    
green:regionN. East   5.1051     7.6193   0.670 0.506896    
green:regionSouth     1.9070     1.8391   1.037 0.306313    
green:regionMidwest  -0.2074     1.5488  -0.134 0.894173    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.77 on 38 degrees of freedom
Multiple R-squared:  0.7888,	Adjusted R-squared:  0.7388 
F-statistic: 15.77 on 9 and 38 DF,  p-value: 2.553e-10
```









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 the N. East region varies with all variables together through an opposing way when compared with the South and Midwest regions.

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

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

Residuals:
    Min      1Q  Median      3Q     Max 
-158.30  -23.39  -12.53   17.00  172.54 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   153.6718    42.3815   3.626 0.000788 ***
metro           0.2914     0.4301   0.678 0.501816    
toxic           2.4238     0.5010   4.838 1.89e-05 ***
green           4.7999     0.5988   8.016 6.31e-10 ***
regionN. East -12.3014    28.2791  -0.435 0.665843    
regionSouth    28.4084    23.2879   1.220 0.229482    
regionMidwest   3.7223    24.7604   0.150 0.881239    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.71 on 41 degrees of freedom
Multiple R-squared:  0.7802,	Adjusted R-squared:  0.748 
F-statistic: 24.26 on 6 and 41 DF,  p-value: 4.909e-12
```