

# R Programming Exercise: Analysis of Air Quality Data

Bob Gotwals  
Intro to Computational Science, NCSSM Online

September 19, 2013

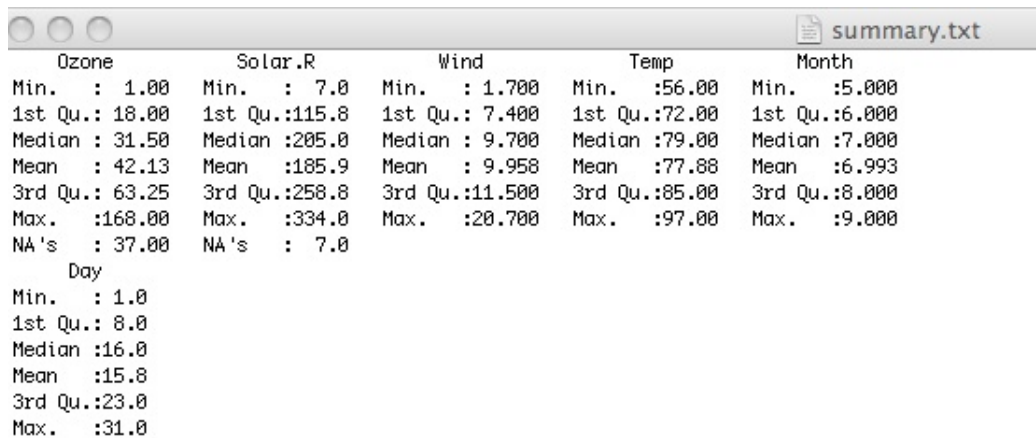
## 1 Part 1

### 1.1 Programming Specs

Your task is to develop a *fully documented* script in R that analyzes data in the dataset "airquality". The data is available from BrainHoney. Download it, put it in a folder on your desktop and load it with the "read.csv" command.

Your program should do the following:

1. Find and print the names (to an output file called "names.txt") of all of the items contained in this dataset
2. Calculate and print (to an output file called "summary.txt") the basic descriptive statistics of the dataset. The file should look like Figure 1:



Ozone	Solar.R	Wind	Temp	Month
Min. : 1.00	Min. : 7.0	Min. : 1.700	Min. :56.00	Min. :5.000
1st Qu.: 18.00	1st Qu.:115.8	1st Qu.: 7.400	1st Qu.:72.00	1st Qu.:6.000
Median : 31.50	Median :205.0	Median : 9.700	Median :79.00	Median :7.000
Mean : 42.13	Mean :185.9	Mean : 9.958	Mean :77.88	Mean :6.993
3rd Qu.: 63.25	3rd Qu.:258.8	3rd Qu.:11.500	3rd Qu.:85.00	3rd Qu.:8.000
Max. :168.00	Max. :334.0	Max. :20.700	Max. :97.00	Max. :9.000
NA's : 37.00	NA's : 7.0			
Day				
Min. : 1.0				
1st Qu.: 8.0				
Median :16.0				
Mean :15.8				
3rd Qu.:23.0				
Max. :31.0				

Figure 1: Summary output file

3. Perform a multiple linear regression between wind and temperature (x-values) and ozone (y value).

4. Output a summary of the regression results to a file ("multireg.txt"). The file should look like Figure 2 (I got some strange characters in my file, yours may be cleaner):

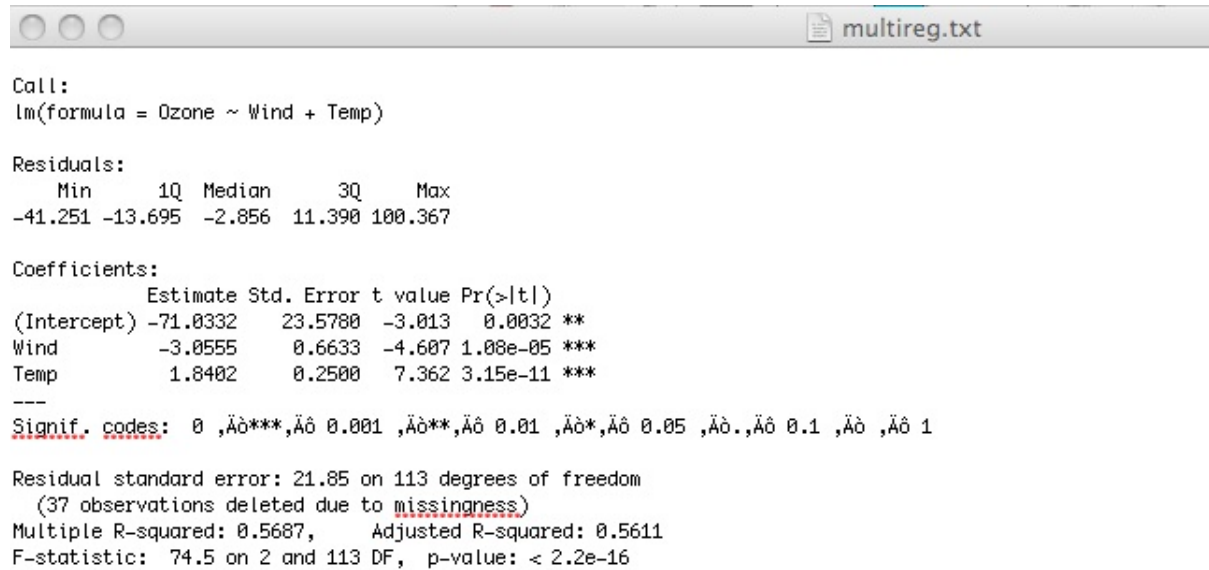


Figure 2: Multiple linear regression summary output file

5. Calculate a linear regression between temperature (x value) and ozone (y value). Plot this data. Then, plot the linear regression line on top of the data. The plot should look like Figure 3 and be saved as a file "regressionplot.jpg":
6. Plot all four variables (wind, temp, solar radiation, and ozone) as horizontal boxplots (with appropriate titles) in a 2x2 matrix. Your plots should look like Figure 4 and be saved as a file "boxplots.jpg":

Your code should be HEAVILY documented, including your name, date, name of the program, and brief descriptor of what the program does. The R script and all of the output files should be located in ONE folder, and the appropriate "setwd()" command should be at the top of the script. I should be able to change the working directory to my folder (for example, "/Users/gotwals/BobStuff/NCSSM Stuff/NCSSM Online/CompBioCourse/RFolder/airquality") and your script should run and produce the correct output files (text files and jpeg files).

## 2 Part 2: Bayesian Information Criterion (BIC) Programming Project

This should be done in your EXISTING script for air quality that was completed in Part 1. NEW CODE SHOULD BE WELL DOCUMENTED!

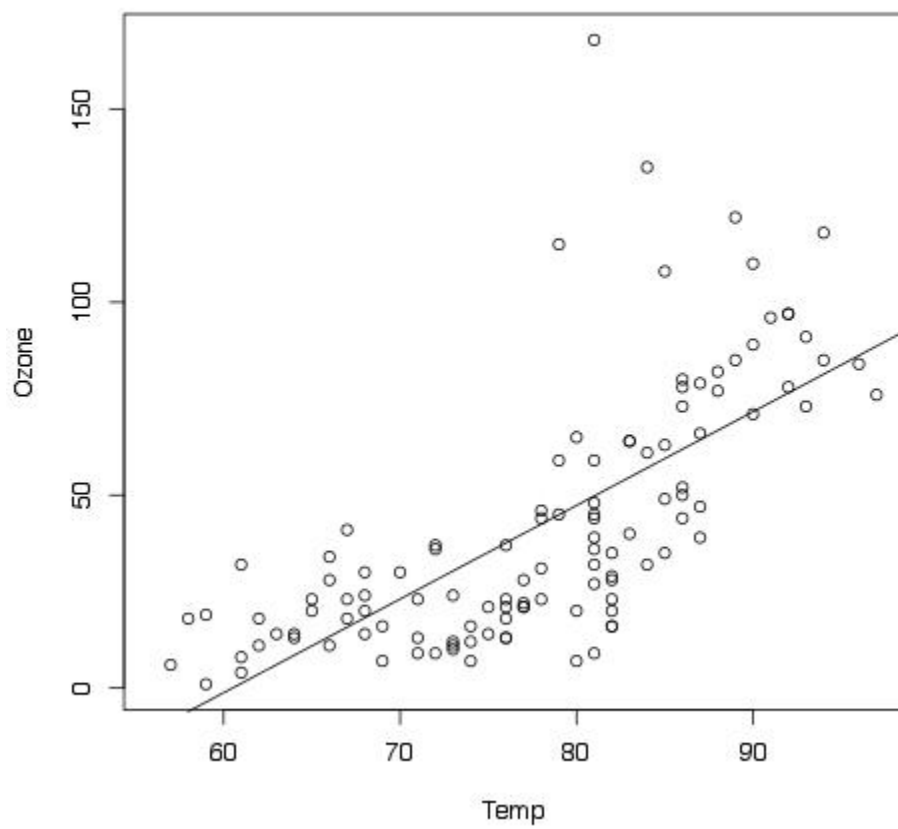


Figure 3: Linear regression plot

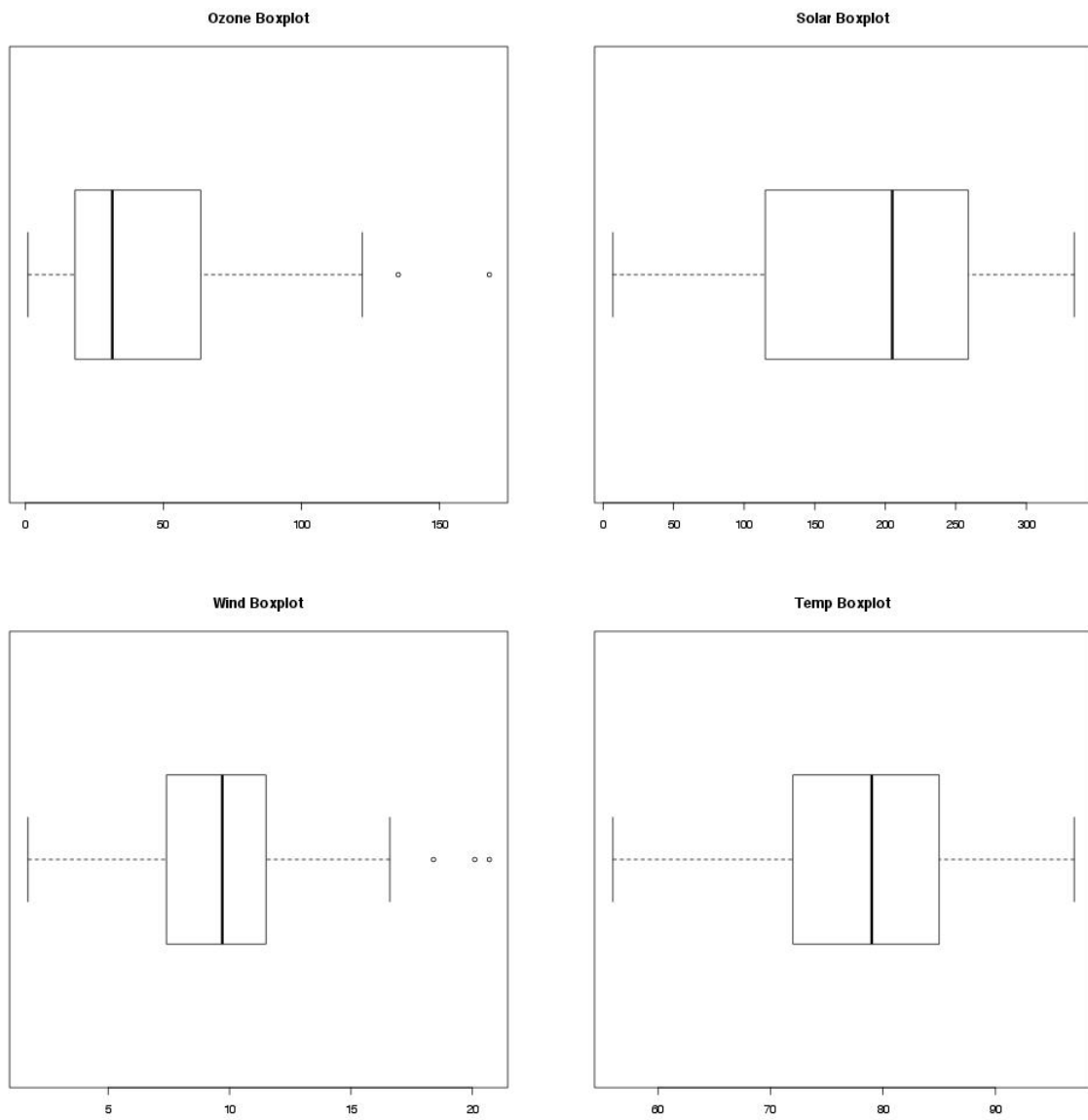


Figure 4: Boxplots

1. Build BIC models for EACH of the models shown in the graphic below. Your results should be output to a file called "bicscores.txt", which (upon changing the directory from YOUR directory to MY directory) should have a PRINT statement showing the model, and then the BIC score. For example: "Temp → Ozone" 1075.967
2. Models to be included:

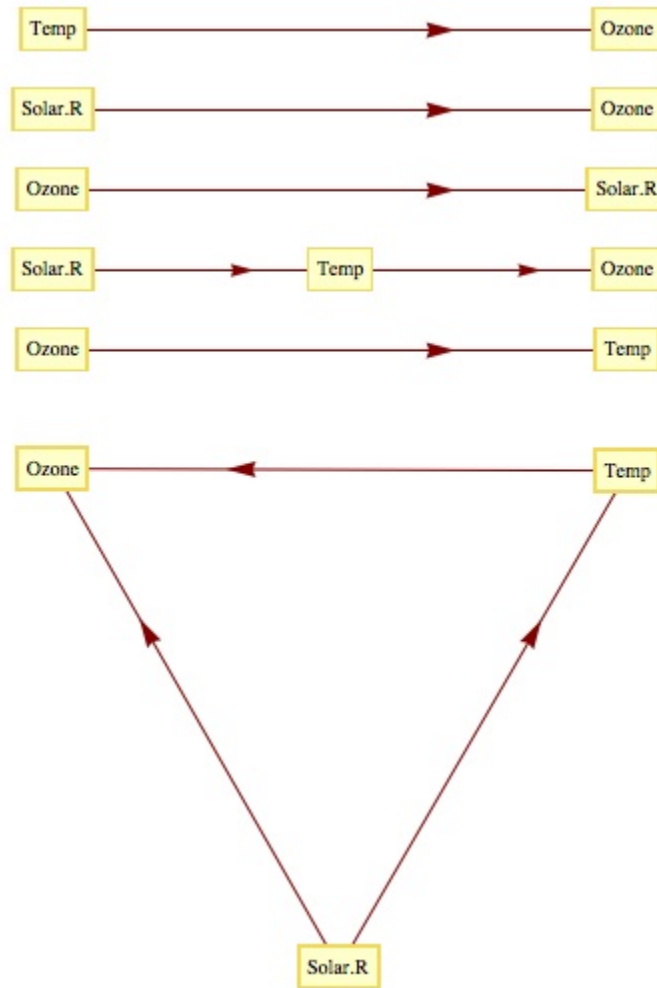


Figure 5: BIC Models

3. Your analysis should include a brief description of what you think these scores are telling you.