Math 263 – Excel Assignment 3

Sections 001 and 003

Purpose

In this assignment you will use the same data as in Excel Assignment 2. You will perform an exploratory data analysis using R. You shall reproduce a series of graphs included for reference in this document.

Software

Excel with the RExcel plugin installed.

Review - Instructions to use the RExcel plugin

1. Open Excel.
2. Go to Add-Ins tab; you should see now the RExcel tab.
3. Choose “Start R” from the RExcel tab.
4. After R started after a few seconds, choose “RCommander > With Excel Menus”.
5. You should see a menu of RExcel commands which you will use.

Create an R data frame

1. Select all data in the spreadsheet, including the labels (A10:E664).
2. Right-click to see the context menu.
3. Select “Put R Dataframe”. Name the dataframe “FEV”.
4. Now your data are available for analysis.

Change the “sex” and “smoke” variables to categorical

In the original spreadsheed, there are columns “sex” and “smoke”. The data in these columns indicates, of course, the gender of the child, and whether the child smokes or not. However, the data are “0” and “1”, which is confusing, because these are numbers. In the dataframe you have created, variables “sex” and “smoke” are numerical. You need to convert them to categorical by using the following steps:

1. Select “Data > Manage variables in active data set > Recode variables.”
2. Choose variable “sex”.
3. Specify “sex” as new variable name (overwrite variable “sex”).
4. Type in the bottom window: 0=”F”, 1=”M” (that is, “F” means “girl”, “M” means “boy”)
5. Press OK. Confirm overwriting of the “sex” variable.

6. Repeat steps 1-5 with the “smoke” variable, but in step 4 use: 0=”S”, 1=”N” (that is, “S” means “Smoker” and “N” means “Non-Smoker”).

Now type “FEV” in the top RCommander window, and press the “Submit” button. In the bottom window you should see the data:

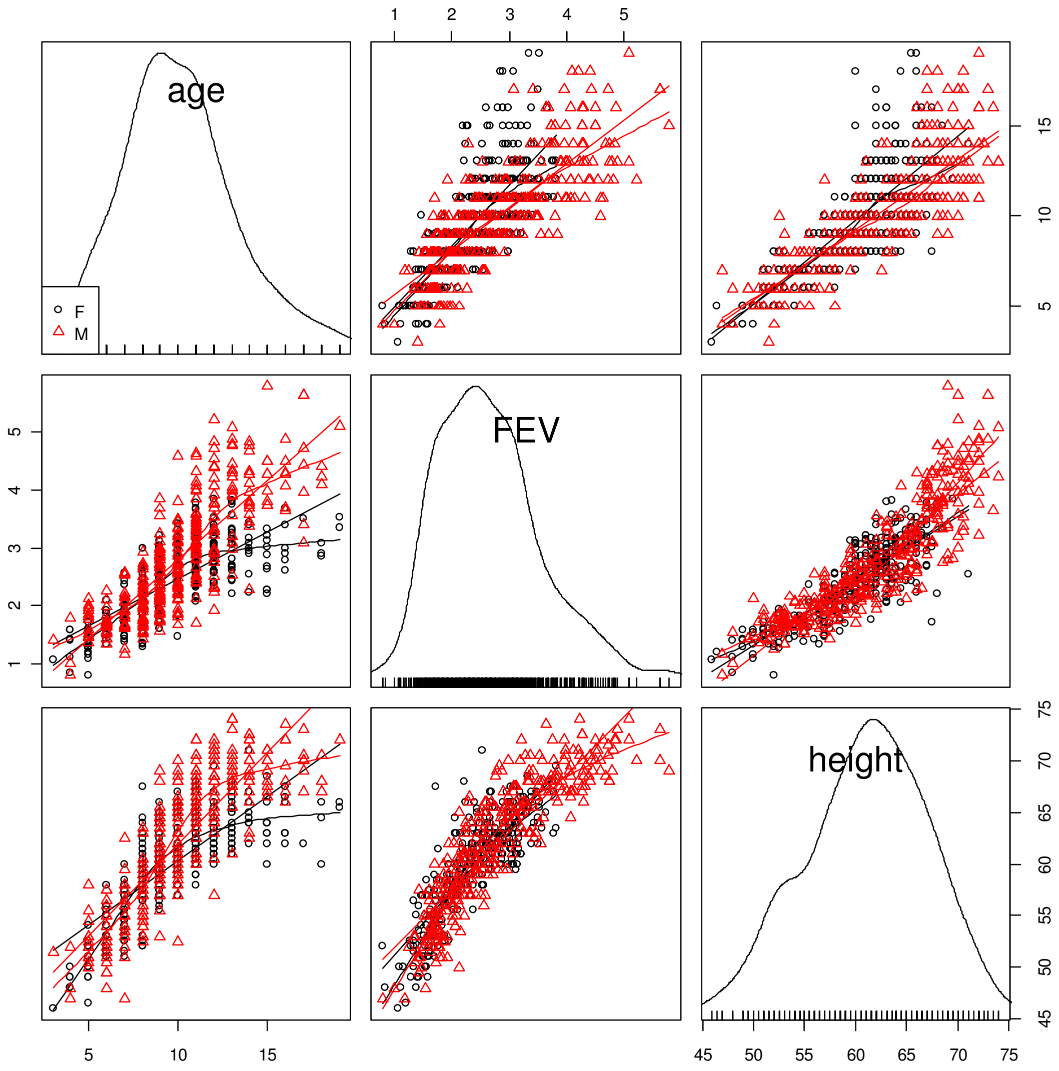
```
> FEV
    age   FEV height sex smoke
1     9 1.708   57.0   F     S
2     8 1.724   67.5   F     S
3     7 1.720   54.5   F     S
4     9 1.558   53.0   M     S
….
```

That is, the variables “sex” and “smoke” were “recoded” to use new symbols (letters) to denote categories, rather than numbers. Also, now R understands that “sex” and “smoke” are categorical variables. This fact is reflected in plots, as will be seen below.

Use matrix plot function to quickly explore the data

1. Choose “Graphs > Scatterplot Matrix” from the RExcel menus.
2. When the option window appears, accept all 3 variables.
3. Click on the “Plot by groups” button and select “sex” to group points by gender (girls and boys will be plotted with different markers, circles and triangles)
4. Once the graph appears, choose “File > Copy to clipboard > As Metafile”.
5. Paste in your document.
6. The result is visible below.

The scatterplot matrix is a very useful kind of plot, as you can quickly see the relationships, their directions, and the smoothed histograms for all variables involved. You have also divided the data using “sex” as factor.



Conduct regression of FEV on age

1. Select “Statistics > Fit model > Linear Regression”
2. Copy the results from either the RCommander bottom window or transfer the output to Excel first by using the “Get R Output” command in the context menu (right-click while pointing at some empty block of cells).
3. Please keep all information, even the pieces that you don’t yet understand.
4. You will notice familiar slope and intercept of the regression line, and R-squared value.

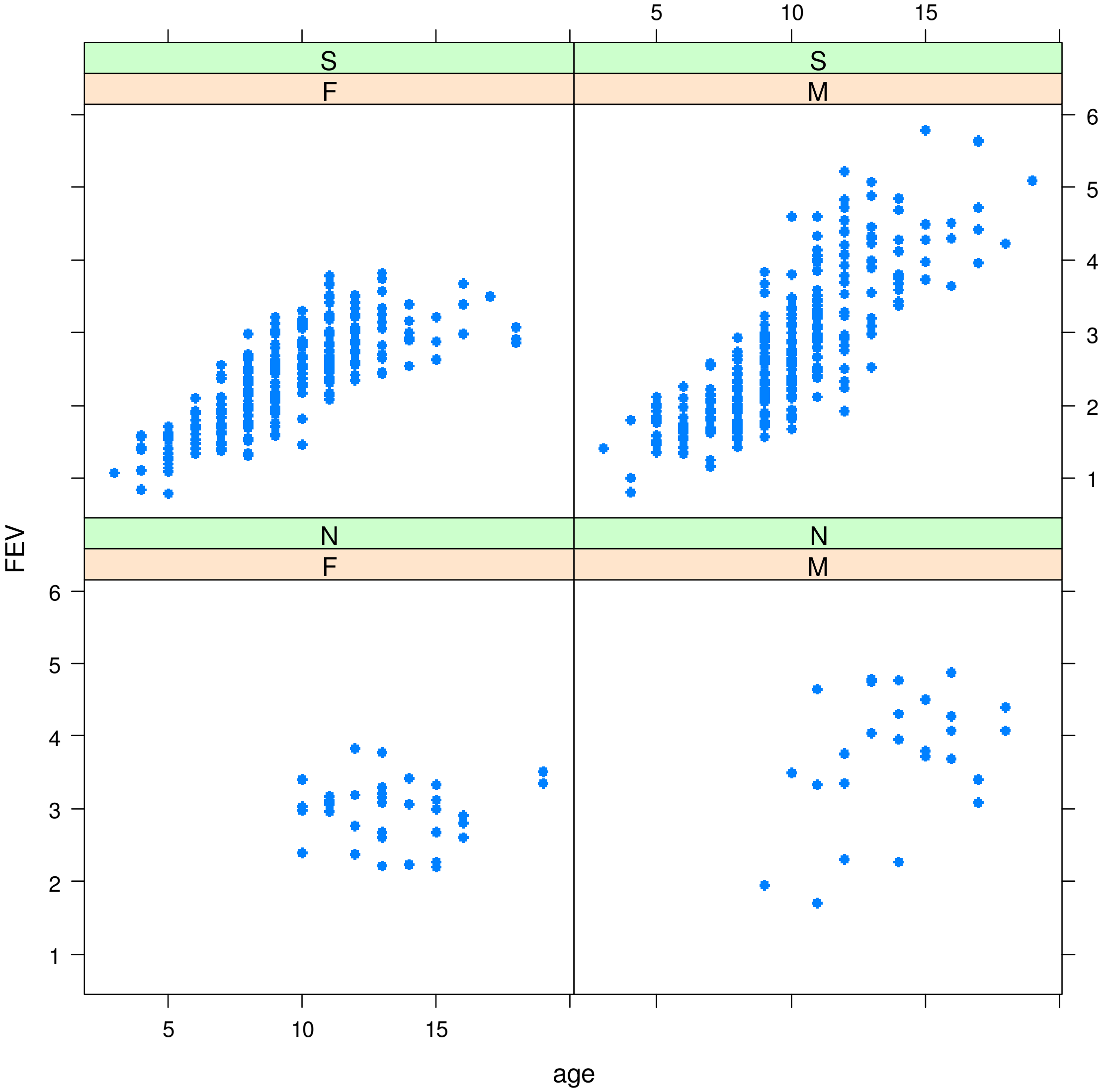
```
Call:
lm(formula = age ~ FEV, data = FEV)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9675 -1.2974 -0.2618  1.0462  7.5116 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.13585    0.24220   12.95   <2e-16 ***
FEV          2.57714    0.08726   29.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.933 on 652 degrees of freedom
Multiple R-squared: 0.5722,   Adjusted R-squared: 0.5716 
F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

The XY conditioning plot

The “conditioning” plot uses categorical variables to provide additional information about the data (you should have no trouble figuring out how it works). Here is how it is created:

1. Choose “Graphs > XY Conditioning” plot
2. Where the dialog pops up, choose the “Group Variables” to be both “sex” and “smoke”.

The result is below for reference.


S
F
S
M
N
F
N
M
FEV
age
5
10
15
1
2
3
4
5
6