UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Business Statistics

Assignment 1

The file ames.xls is a dataset describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. It was compiled by Dean De Cock, and documented in an article appeared in the Journal of Statistics Education, Volume 19, Number, 3 (2011).

The data set contains 1460 observations on 80 variables. The last column is the Sale Price, which is the output variable we aim at predicting. There are 34 quantitative house characteristics relating to the dimension of the house and its age. There are also a few categorical variables associated with this data set with a varying number of categories. The full description of the variables is in the file data_description.pdf

1 Reading the data and preliminary operations

Some of the variables contain a large number of missing data. The following statements import the data and delete the columns with missing values, returning a complete dataset. You should copy and paste them as they are.

```
ames = read.csv("ames.csv", stringsAsFactors = F)
# drop variables with missing obs
ames = ames[sapply(ames, function(x) !any(is.na(x)))]
sapply(ames, function(x) sum(is.na(x)))
attach(ames)
```

2 Your task

You should address the following points:

a. Explore the functions `hist()` and `boxplot()`. Plot the histogram of `SalePrice`, choosing the number of classes as you deem appropriate. Using the `boxplot()` function, display its conditional distribution by `BldgType`.

b. Using the function `pairs` produce a matrix scatterplot of the variables `SalePrice, x1stFlrSF,x2ndFlrSF, TotalBsmtSF`. What potential problems do you see?

b. Compare the results of the following regression models:

- Regress `SalePrice` on each of the explanatory variables `x1stFlrSF,x2ndFlrSF, TotalBsmtSF` separately (simple regressions).

- Regress `SalePrice` on all the explanatory variables `x1stFlrSF,x2ndFlrSF, TotalBsmtSF` (multiple regressions).
- Regress `SalePrice` on the single explanatory variable `xsum = x1stFlrSF+ x2ndFlrSF+TotalBsmtSF` (you will need to define it first).

Which specification do you prefer? For the preferred specification, do you think that you obtained a good fit? Can we conclude that the value of an additional square feet in the basement (TotalBsmtSF) is the same as that of a square feet in the first floor?

This assignment is due by 10:00 p.m. of September 29. You should upload in

https://www.dropbox.com/request/zJcaaTpfKcgmqcuJsHK2

a unique pdf file with the results and your comments. The filename must be YourSurname_YourName_Assignment1.pdf.

3 Getting started

To get the work started, recall the following steps:

1. Download the file in a working directory that you can trace back.
2. Open RStudio
3. Set the working directory from the menu Session → Set Working Directory → Choose Directory
4. Open a new script that contains your R statements. From the menu File → New → R Script
5. Copy and paste the following lines:

```
ames = read.csv("ames.csv", stringsAsFactors = F)
# drop variables with missing obs
ames = ames[sapply(ames, function(x) !any(is.na(x)))]
attach(ames)
```

6. Enjoy working with the data.
7. Remember to save the script.