UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Business Statistics

Assignment 2

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.

1. As your training sample, select randomly $N = 1000$ observations from the original dataset, using the following statements (copy and paste in your R script). The `seed` for the sample selection (which is the argument of the `set.seed()` function), must be set equal to the eight digits number making up the your birth date (DDMMYYYY), e.g. 29021996; if the date starts with a 0, drop it and use a 7 digits number:

   ```
   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)
   ##########################################################
   Ntot = nrow(ames)
   set.seed(29021996); # please, set seed as specified below
   N = 1000;
   s = sample(1:Ntot, N);
   ames.train = ames[s,];  # training sample of size N
   ames.test = ames[-s,];  # validation sample of size Ntot-N
   summary(ames.train);
   ```

   The test set contains the remaining observations.

2. Consider the following candididate input variables: `x1stFlrSF`,`x2ndFlrSF`, `LotArea`, `FullBath`, `PoolArea`, `GarageCars`, `TotRmsAbvGrd`, `KitchenAbvGr`, `GrLivArea`, `FullBath`, `BedroomAbvGr`, `YearRemodAdd`, `YearBuilt`, `OverallCond`, `OverallQual` for predicting the output variable `SalePrice`.

- Estimate the full model (containing all the input variables) and comment about the results. What effects are not significant at the 5% level?
- Perform variable selection using forward and backward stepwise selection. Are there differences between the two selection methods? Consider the variables that are selected by the backward stepwise method and compare the estimated coefficients with the full model specification. Is there any surprising result?
- Would the model selected by forward stepwise change if you use BIC as your estimate of the test error?
- Estimate the test error of the rival specifications using the test set `ames.test`. Which one delivers the smallest test error? To get the out of sample prediction use the function `predict(model, newdata = ames.test)`, where `model` is the specification under investigation.

3. Compute the eigenvalues and eigenvectors of the correlation matrix P of the input variables in the data frame X below:

   ```
   X = data.frame(x1stFlrSF,x2ndFlrSF,  LotArea, FullBath, PoolArea, GarageCars,
                  TotRmsAbvGrd, KitchenAbvGr,  GrLivArea, FullBath, BedroomAbvGr,
                  YearRemodAdd, YearBuilt, OverallCond, OverallQual)
   P = cor(X)
   ```

   How many eigenvalues are larger than 1? What is the sum of the eigenvalues? What is the share of the total variability accounted for by the first principal component? Consider the first eigenvector. Can we say that `OverallCond` behaves somewhat idiosyncratically?

This assignment is due by 10:00 p.m. of October 9. You should upload in

https://www.dropbox.com/request/EOUpt8gfJCiE7QY91g43

a unique pdf file with the results and your comments. The filename must be YourSurname_YourName_Assignment2.pdf.