Data Analysis Project

MAP 553 - Regression

11/4/2019

General Instructions

This project is about the implementation on a real data set of the statistical methods we reviewed in this course. You *should* use the statistical software R to perform your analysis.

- You are required to submit your work electronically on moodle. Deadline for the project is December 20th at noon. no late work is accepted except in exceptional circumstances. Supporting documents will be required.
- Your write your report in Rmarkdown format. Your submission should contain two files: the report in pdf format and the Rmarkdown file containing all your code that you compiled to obtain the pdf report.
- Your work will not be graded if you do not respect any of the above instructions.

Report instructions

- Your code should be clearly commented so that it is clear which parts of your code corresponds to which parts of your report.
- *8 page lenngth limit including figures and tables.* Fonts should be no smaller than 10 points, margins should be reasonable.
- Your report should have the following sections:
  1. **Introduction.** Write a short introduction describing the research problem. Clearly state the research hypothesis at the end.
  2. **Exploratory Data Analysis/Inital Modeling.** Provide graphical displays ornumerical summaries for all variables and pairs of variables. Describe your results.
  3. **Modeling and Diagnostics.** Start by building any multiple linear regression models you think are appropriate. Create diagnostic plots to determine the appropriateness of your model. Discuss whether the assumptions are met. If not, take any actions you think are justified such as: transformations, removing outliers, removing variables, etc… Explain what decisions you make and explain why you made them. Check again your diagnostics for the new models
  4. **Final Models.** Summarize your final models: report the parameter estimates, standard errors, confidence intervals, and p-values. Interpret the fitted models in the context of the problem. If you have several models, then compare them. Note that there are several ways to compare different regression models: (i) partial F-tests, (ii) residuals and diagnostics and (iii) cross-validation (or other measures of prediction error). Use the test sample to get an estimate of the generalization error of your final model.
  5. **Discussion.** What are your final conclusions? Mention any limitations of your analysis, or possible future directions of research.

The Data

The data set is taken from Kaggle competition **House Prices: Advanced Regression Techniques**

Data description. The data records the selling price of about 1500 houses along with 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. The file *data_description.txt* contains a full description of each column.

Goal: This competition challenges you to predict the final price of each home.

Note that the original kaggle data sets contain some missing values. For the sake of simplicity, I have already imputed these missing values using elementary techniques. I deleted some variables with too many missing values from the Kaggle data set. If you are interested in this aspect of data analysis as well, feel free to download the original data and propose your own treatment of missing values.

You will use the following commade line to read the data:

```r
train <- readr::read_csv("train.csv")
```