Regression - Final project

Victoire de Termont and Sarah Jallot

12/3/2019

I. Introduction

We are given a dataset, containing variables that describle almost comprehensively every aspect of residential homes in Ames, Iowa. For a certain number of homes (the number of lines of our dataframe), the associated price is given. The goal of this analysis is to select the good criteria and the most adapted regression model in order to be able to dertermine the price of other homes in Ames, based on this information.
Firt, we have to look at the data, and get a first intuition of an efficient model. Indeed, before starting our full analysis, we have to understand what we are given, check that there are no missing values, and than R classifies correctly each regressor.

II. Exploratory data analysis / initial modeling

Now that our dataset is correctly formated, we can try to extract trends and relationships among the regressors. Indeed, histograms are a good way to understand the behavior of each regressor. Comparing them is also essential, for instance using pairplots. As the goal is to have an efficient model, we don't want to keep variables that are not correlated with the output, or that are very correlated among them as it leads to redundance. Thus, we could get a first intuition of variables to keep using a correlarion matrix. We will have to deal string regressors before doing so.
As our intuition is that the location of the flat has a big impact, we could try to mathematically test this hypothesis.

III. Modeling and diagnostics

We wikll start by a linear model, and then add a Lasso penalization as we have a lot of regressors. Using those two models, Anova methods will enable us to select the appropriate regressors.
Then, we will compare our models using criteria seen in class. We will also test the gaussian assumption of our model.

IV. Final models

Parameters of our model:
Error:
Confidence interval:
p-value:
Estimate of the generalization error of our final model:

V. Discuttion
Final conclusions on our model:
key learnings:
Improvements for future projects: