Some remarks about the project

TA: Nhan Huynh

5/24/2017

Data analysis

- It's a crucial step to have prior knowledge before building any linear models.
- Some common methods:
  - statistical description (min,max,mean,range,...) $\rightarrow$ learn the distribution of variables and detect any unsual patterns; should we treat predictors as numerical or categorical in the model?
  - visualization such as histograms and scatterplots $\rightarrow$ correlation between variables (response vs. predictor, predictor vs. predictor). For example: if there exists a strong positive correlation between the response and predictor but the linear slope is negative, what should we do in this case?
  - scatterplot to determine if we should add interactions in the model (please look at lab8.R for more details)

Oscam's Razor

"The model that fits observations sufficiently well in the least complex way should be preferred."

That says if you have two models: a simple and a complex one, choose the simpler one when both models do an equivalent job at fitting the data.

Interaction terms

- If interaction between the predictors has a significant effect on the repsonse variable but excluded in the model, the results might be misleading.
- There is a trade-off between interpretation and goodness of fit when adding interactions in the model.
- In lab8, we see the presence of interaction between categorical and continuous predictors. It is also possible to have interaction between continuous variables.