Statistics 191
Introduction to Regression Analysis and Applied Statistics
Data analysis project
Due March 14, 2014

Prof. J. Taylor

This project is a full analysis of a (moderately) large dataset using the tools we have learnt throughout the course. The project should be **individually**.

The data set is based on real estate sales in Ames, Iowa in the years 2006-2010 [1]. A description of the data can be found here

http://www.stanford.edu/class/stats191/data/amesdoc.txt

I have created a subsample of 2000 of the cases, reserving a separate 500 for testing. The training data for the assignment is available here

http://www.stanford.edu/class/stats191/data/ames2000.csv

Your task is to build a model to predict `SalePrice` based on the remaining variables. After you submit your written report, you will be asked to test your model on a separate subset of 500 cases in the data set. Part of your grade will be based on the accuracy of the final model.

The final report should be no more than 10 pages.

Beware: the data set is large enough so simple stepwise model building procedures may be very slow.

The project should have the following parts

§1) **The study:** In this section, you should give a description of the study underlying their dataset. Possible questions to be answered are the following:

(a) What field does the data come from?

(b) What are the goals of the study? Are there any effects of particular interest?

(c) How might these goals be answered, i.e. tests / confidence intervals?

In this data set, it is of particular interest to estimate the effect of the number of bedrooms on the sale price, some measure of the overall price per square foot, and the added value of a pool. Your final model should include Bonferroni corrected confidence intervals for these parameters.

§2) **The data:** In this section, you should describe the data set and possibly do some exploratory data analysis. For instance:

(a) How are the predictor variables spread out? Are there any noteworthy features to their spread that could be highly influential observations?

(b) Are any of the predictor variables highly correlated?

§3) **The models:** In this section, you should develop a model for the data that will allow them to answer some of the specific goals of the study. Possible questions to be addressed here are the following:

(a) Which predictor variables, if any, should be included in the model *a priori*?

(b) Are there any interactions that should be considered for inclusion in the model?

(c) Are there any second order interactions that should be considered?

(d) Are there any interactions that should NOT be considered?

§4) **Results:** In this section, you should report their results obtained by fitting the proposed models in the previous section. Emphasis should be placed on clarity, as if the report were a statistical consultant's report for a non-statistician. For instance, loads of *R* output would, in general, not be acceptable. Plots and well-organized tables are good things to have in this section. Possible questions to be addressed here are the following?

(a) What is the final regression model for the data?

(b) How was this model obtained, i.e. forward stepwise search (to be seen in class)?

(c) Using the standard diagnostic tests, does the model appear to fit the data well?

(d) What are the final confidence intervals for the effects of interest mentioned in the study section? Do these intervals seem very sensitive to the choice of model (i.e. do they vary widely for different choices of variables in the model)?

(e) Report a 5-fold cross-validated estimate of error for predicting sale price.

§5) **Appendix (*R* Code):** In this section, you should attach a final, editied, copy of the *R* code used in the analysis. Ideally, there will be comments in the file, i.e. lines beginning with "#" to clarify what each part of the code is doing.

§6) **Acknowledgements:** If you consult outside sources that refer to this data set, you should cite these as references, and describe what you used from each source. Sources include material found on the internet, journal articles and books.

Note: There are no right or wrong answers for many of these questions. The goal of the project is to try to mimic the analysis of a real data set that you might come across in your own field of application.

References

[1] Dean De Cock. Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistical Education*, 19(3), 2011.