Homework 3

Due: Wednesday March 1 at 11:59pm

See general homework tips and submit your files via the course website. Code for creating the data set is in **HW3Data.sas** in the Homework 3 folder on the course website.

The data set is based on the **housing.data**[1] file from the UCI Machine Learning Repository[2] and defined as the **housing** data set in **HW3Data.sas**. The data is for suburbs of Boston in 1978 and the original variables are described on the UCI Machine Learning Repository website referenced in the endnote below. The **housing** data set for this homework contains observations where **medv** is less than 50 as it appears the 50 values may be censored (50 seems to be the recorded value for values of 50 or greater).

The variables remaining in **housing** are the following:

- **age** – proportion of houses built before 1940
- **crim** – crime rate per capita
- **indus** – proportion of non-retail business acres
- **nox** – nitric oxides concentration
- **ptratio** – pupil-teacher ratio
- **rm** – average number of rooms per house
- **over25kSqFt**– categorical variable for whether there is residential land zoned for lots over 25,000 square feet (**'none'** if the original **zn** variable is 0 and **'some'** if it is greater than 0)
- **ptlevel** – categorical variable for pupil-teacher ratio (**'lower'** for ratios less than 15, **'higher'** for ratios greater than 20, and **'medium'** for ratios in between)
- **taxlevel**– categorical variable for property tax per $10,000 (**'lower'** for taxes below 500 and **'higher'** for taxes of 500 or greater)
- **logmedv** – log of median value of owner-occupied homes in $1000's (log of the **medv** variable from the original data)

The logs of salaries and asset values (such as home prices) are often modeled rather than the actual value because they often behave more normally than the actual values.

Exercise 1

a) Obtain the best main effects only model of **logmedv** using **over25kSqFt**, **ptlevel** and **taxlevel** as possible main effects. Explain how you chose the predictors in the final model, and comment on the significance of the model, the amount of variation in the log of median home values described by the model, and the significance of the terms in the model.

b) Identify and interpret the significant differences in expected log home values across groups and comment on what that tells us about suburb features that had a greater and lesser relationship with home values in 1978 Boston.

Exercise 2

Repeat **Exercise 1** now considering possible interactions as well. Start with the main effects chosen in Exercise 1 and repeat parts **a** and **b** by adding any significant interaction terms to get your best model including significant interactions.

Exercise 3

A family in Boston in 1978 is interested in the relationship between average age of homes in a suburb and the median home value in that suburb. They are also concerned about crime and will only consider suburbs with less than 1 crime per capita.

a) Fit a simple linear regression model for **logmedv** as a function of **age** for suburbs with less than 1 crime per capita. Analyze the diagnostics, note any issues that need to be remedied, make any necessary adjustments for undue influence, and re-fit the model if necessary. For Cook's distances, do not leave any points in the final model that have Cook's distance greater than 4 times the cutoff line in the plot.

b) For your final model, interpret what the model tells us about the relationship between average home age and the median home value (Note: the interpretation should be in terms of the median value, not the log of the median value). Comment on how much variation in log of median home value is described by the model, and note any remaining issues in the diagnostics and state how you might remedy them (you do not need to actually remedy the issues and re-fit). Based on these diagnostics and results, how useful would this model based on age alone be for estimating median home value?

Exercise 4

In addition to average **age**, the family also has data on proportion of non-retail business acres (**indus**), nitric oxides concentration (**nox**), and average rooms per house (**rm**) information as well.

a) Perform model selection starting with these 4 predictors and obtain your best linear regression model for **logmedv** for suburbs with less than 1 crime per capita. Analyze the diagnostics, note any issues that need to be remedied, make any necessary adjustments for undue influence, and re-fit the model if necessary. For Cook's distances, do not leave any points in the final model that have Cook's distance greater than 4 times the cutoff line in the plot.

b) For the final model from part a), note any remaining diagnostic issues, comment on significance of the model and how much variation in log median home value is described by crime rate for towns in this subset of the data, and interpret what the model tells us about the relationship between the chosen predictors and median home value (As before, interpret in terms of home value and not log home value).

---

[1] http://archive.ics.uci.edu/ml/datasets/housing

[2] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.