Homework 3

STAT 448 - Advanced Data Analysis

Due: Thursday, March 1 at 5:00 pm

The data sets are provided in the **HW3Data.sas** file in the Homework 3 folder on the course website. For exercise 1, use the `cardata` data, for exercise 2 and 3, use the `housing` data. The `cardata` data is a subset of the sashelp.cars data and contains the number of cylinders, the location of origin, the type of car and the highway fuel efficiency for a subset of the cars. The `housing` data set is a subset of `housing.data` from the UCI Machine Learning Repository. The data is for suburbs of Boston in 1978 and the original variables are described on the UCI Machine Learning Repository website referenced at: `https://archive.ics.uci.edu/ml/machine-learning-databases/housing`. 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, 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)

1. Use the `cardata` and answer the following questions.

   (a) For **mpg_highway** (highway miles per gallon), create a cross-tabulation of the mean, standard deviation and counts by **cylinders**, **origin**, and **type**. Comment on any interesting features (e.g. apparent differences in fuel efficiency across groups and balanced-ness of the data).

   (b) Start with a three-way main effects ANOVA and choose the best main effects ANOVA model for **mpg_highway** as a function of **cylinders**, **origin**, and **type** for the cars in this set. Comment on which terms should be kept in a model for **mpg_highway** and why. For the model with just the predictors you decide to keep, comment on the significance of the model and of the terms in the model and comment on how much variation in highway fuel efficiency the model describes.

   (c) Starting with main effects chosen in part (b), find your best ANOVA model by adding in any additional interaction terms that will significantly improve the model. For your final model, comment on significance of the model and the individual terms in the model, variation explained by the model, and any significant group differences. What does this tell us about fuel efficiency differences across **cylinders**, **origin**, or **type** groups?

   Note: For interpretations of differences for the main effects, state quantitative interpretations of the significantly different groups (e.g. difference estimates and 95% confidence intervals of the differences, and what the difference tells us about **mpg_highway**). For interaction term, identify significant interactions, but no need to interpret it quantitatively.

2. 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?

3. In addition to average **age**, let's consider data on proportion of non-retail business acres (**indus**), nitric oxides concentration (**nox**), and average rooms per house (**rm**) information as well.

   (a) Fit a linear regression model for **logmedv** as a function of **age**, **indus**, **nox**, and **rm** for suburbs with less than 1 crime per capita. Analyze the diagnostics, describe any issues that need to be remedied, but no need to make adjustments for undue influence. (i.e., no need to re-fit the model)

   (b) Comment on how much variation in **logmedv** is described by the model. Comment on the relationship between **logmedv** and the remaining significant predictor(s). Check multicollinearity issues among predictors. Explain to the medical director whether this is a good model based on variation explained and diagnostics and diagnostics plots.

4. Now we perform model selection.

   (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).