Homework 3

Due: Friday March 17 at noon

See general homework tips and submit your files via the course website.

For all exercises, use the **Auto** data set defined in the **HW3Data.sas** file. The **Auto** data is based on the **automobile** data[1] from the UCI Machine Learning Repository.

Auto data contains two measures for miles per gallon (mpg) fuel efficiency, three categorical predictor variables, and five continuous predictor variables. The groups of variables are as follows:
Mileage variables:

- **highwaympg**: miles per gallon on the highway
- **citympg**: miles per gallon in the city

Categorical predictors:

- **weight_cat**: light or heavy
- **type**: luxury brand or non-luxury brand (type=1 if luxury brand, or type=0 if non-luxury brand)
- **ndoors**: number of doors (four, two)

Continuous predictors:

- **weight**: weight of the car
- **height**: height of the car
- **horsepower**: horsepower of the engine
- **enginesize**: the size of the engine
- **price**: price of the car

Exercise 1

a) For **highway mpg**, create a cross-tabulation of the mean, standard deviation and counts by **weight_cat, type** and **ndoors**. Comment on any interesting features (e.g. apparent differences between groups).

b) Start with a three-way ANOVA (be sure to use the correct SAS procedure) and choose the best ANOVA model for **highway mpg** as a function of **weight_cat, type** and **ndoors**. Comment on which terms should be kept in a model for **highway mpg** and based on type III SS.

c) Use GLMSELECT and stepwise selection procedure, retain only the statistically significant main predictors. Which predictors did you choose? Investigate the interaction term(s) among the significant main predictors, should you retain the interaction term(s) based on significance level?

d) For the model chosen in part **c**, comment on significance of the model and the individual terms in the model, and variation explained by the model. For each significant main effect, rank the levels from the highest highway mpg to the lowest highway mpg based on LS mean estimates. For each significant main effect, highlight (on SAS output table) the pairs of levels that are significantly different in highway mpg. For the significantly different pairs, highlight the differences of LS means, 95% CI of the differences and p-values. Obtain model diagnostics to validate your assumptions. Check the normality assumption for the residuals.

Exercise 2

a) Fit an ANCOVA model, with **highwaympg** as the response, and the following predictors: **weight_cat, type, ndoors,** and covariate **horsepower**. If you include **weight_cat** as a categorical predictor in your model, should you also include **weight** as a continuous predictor in this model?

b) Use GLMSELECT and stepwise selection procedure, retain only the statistically significant predictors. Which predictors did you choose for the final model? How are these predictors different from the model obtained in exercise 1?

c) Comment on the quality of the final model, significance of the parameters, and variation explained by the model. Compare the R-square of this ANCOVA model with the ANOVA model obtained from Exercise 1. Which model is better in terms of % variation explained? Obtain model diagnostics to validate your assumptions. You don't need to check the equal slope assumption. Check the normality assumption for the residuals.

Exercise 3

The director at the auto company wants to build a model that can better predict a car's city mpg. Consider modeling **citympg** as a function of all continuous variables in the data.

a) Fit a linear regression model for **citympg** as a function of all continuous variables in the data. Comment on significance of the model and the individual terms in the model, and variation explained by the model. Comment on which predictor(s) could be removed from the model based on significance level. However, do not remove any statistically insignificant terms and do not remove any outliers yet.

b) Use stepwise selection method and keep only the statistically significant predictors. Which predictor(s) are removed by the stepwise selection?

c) Using the model from b), remove highly correlated predictors one at a time until multicollinearity problems are resolved. If there's no predictor with high VIF, state so in your report. State the predictors you are keeping in your model.

d) Using the model you selected in part c), identify any influential observations, remove any points you deem unduly influential, and refit the model if necessary. Consider the Cook's distance cutoff to be 1. If no observation has high Cook's distance, state so in your report. Comment on the quality of the final model, significance of the parameters, variation explained by the model, and any remaining issues noted in the diagnostics. Check the normality assumption of the residuals. Comment on the relationship between the outcome variable and the remaining statistical significant predictors (eg, how does one unit increase in predictor X impacts city's mpg?). Explain to the company's director whether this is a good model based on variation explained and diagnostics.

Exercise 4

a) Now let's get back to highway mpg. Fit an ANCOVA model, with **highway** as the response, and the statistically significant main effects that you chose in your final model from Exercise 2 as predictors. Now investigate the equal slope assumption by examining the interaction terms of the covariate with all categorical predictors. For example, suppose covariate X and categorical predictors A and B are in your final model from Exercise 2, investigate the interactions of X*A, X*B, and X*A*B. Remove the highest order insignificant interaction term one at a time. State whether the equal slop assumption is valid or not, and which interaction terms should be kept in the model. Compare the R-square of this ANCOVA model with the final ANCOVA model you obtained from Exercise 2. Check the normality assumption for the residuals.

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