Homework 4

Due: Wednesday March 29 at 11:59pm

See general homework tips and submit your files via the course website. Code for creating the data set is in **HW4Data.sas** in the Homework 4 folder on the course website.

The data set is based on the **imports-85.data**[1] file from the UCI Machine Learning Repository[2] and defined as the **autos** data set in **HW4Data.sas**. The data is for cars in use in 1985 and the original variables are described on the UCI Machine Learning Repository website referenced in the endnote at the end of this assignment. The **autos** data set for this homework contains observations for 4, 6 and 8 cylinder vehicles which did not have 4 wheel drive.

The variables in **autos** are the following:

- **cityover30mpg** – indicator for city fuel efficiency over 30 mpg (1 for city mpg over 30, 0 otherwise)
- **price1k** – price in thousands of dollars
- **rpm** – peak rpm for the motor
- **fuel** – fuel source (gas or diesel)
- **drive** – front (fwd) or rear (rwd) wheel drive
- **enginesize** – engine size (presumably in units of cc)
- **hwaympg** – highway mpg rounded to the nearest mile per gallon
- **hp** – horsepower of the engine
- **cylinders**– categorical variable for number of cylinders ('four', 'six', or 'eight')

Note that for logistic regression models the Hosmer-Lemeshow test (see the **lackfit** option) can be used to test for lack of fit for a model. Rejection of the Hosmer-Lemeshow test indicates there is a lack of fit (e.g. the model does not fit the data well). When a lack of fit is determined, this could be an indication that the model does not fit well in particular segments of the data or it could mean that the model does not fit well in general.

Exercise 1

We want to consider logistic models for having city fuel efficiency over 30 mpg. In this exercise, we consider possible categorical predictors.

a) There are three possible categorical predictors in the data: **fuel**, **drive**, and **cylinders**. For each of these categorical variables, create a frequency table of that variable against **cityover30mpg**. Based on the frequency tables, which of the categorical variables might be useful for predicting the odds that a vehicle has city fuel efficiency over 30 mpg?

b) Perform model selection to find your best logistic model for having greater than 30 mpg city fuel efficiency based on these 3 possible categorical predictors. State how you obtained your final model and what terms remain in the final model.

c) For the final model in b, comment on the significance of parameter estimates, what Hosmer-Lemeshow's test tells us about possible lack of fit in the model, and the significance of odds ratios. Interpret what the model tells us about relationships between the categorical predictors and the expected odds of a vehicle having city fuel efficiency over 30 mpg.

Exercise 2

Now we consider possible relationships between continuous predictors (exclude **hwaympg**) and having city fuel efficiency over 30 mpg.

a) Fit a logistic regression model for **cityover30mpg** as a function of **price1k**, **rpm**, **enginesize**, and **hp**. Comment on what the AIC suggests about this model as compared to a constant only model, comment on what the global tests suggest about the presence of non-zero parameters, and comment on what the parameter estimates and odds ratios tell us about continuous predictors we may want to keep in the model.

b) Perform model selection to find your best logistic model for having greater than 30 mpg city fuel efficiency based on these 4 possible continuous predictors. State how you obtained your final model and what terms remain in the final model.

c) For the final model in b, comment on the significance of parameter estimates, what Hosmer-Lemeshow's test tells us about possible lack of fit in the model, and the significance of odds ratios. Interpret what the model tells us about relationships between the continuous predictors and the expected odds of a vehicle having city fuel efficiency over 30 mpg.

Exercise 3

We now consider a generalized linear model for highway miles per gallon as a function of possible predictors in the data. Do not include **cityover30mpg** as a possible predictor. Since **hwaympg** is rounded, we will consider a log-linear Poisson model (a Poisson generalized linear model with log link).

a) Determine the best subset of predictors for a log-linear Poisson model of **hwaympg**, and account for overdispersion if necessary. Explain how you got to your final model, and comment on what type 1 and type 3 analyses tell us about predictors you may want to keep or remove from the model as you select terms for your final model.

b) For your final model, comment on the significance of parameter estimates and interpret what the parameter estimates tell us about the expected relationship between changes in the predictors and highway mpg.

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

[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.