Homework 4

STAT 448 - Advanced Data Analysis

Due: Friday, March 30 at 5:00 pm

To complete this assignment, you will need to access the data sets in **Program_HW4_Data_Spring2018.sas** in the Homework 4 link on Compass. For all exercises, use the sleep data set which originates from `http://lib.stat.cmu.edu/datasets/sleep`. The **nondreamingsleep** and **dreamingsleep** variables have been removed, and a variable called **maxlife10** has been added. **Maxlife10** is 1 if the species maximum life span is less than 10 years and 0 if its maximum life span is greater than or equal to 10 years.

For logistic regression models, we can use the **Cbar** *measure in SAS as an analogue of Cook's distance to check for pointwise influence, and the Hosmer-Lemeshow test -* **lackfit** *option - to test goodness of fit for a model. Rejection of the Hosmer-Lemeshow test indicates there is a lack of fit. 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.*

1. Consider finding the best logistic model for predicting the probability that a species' maximum lifespan will be at least 10 years. Consider all 7 variables as candidates and three index variables of them are categorical in nature. Treat those variables as nominal categorical variables (e.g. ignore the fact that they are ordinal). [1]

   (a) First determine and specify the best set of predictors through stepwise selection with criteria .05. Comment on any unduly influential points on the final model. Use cut-off as 1. If any extremely unduly influential points exist, remove them for the final model.

   (b) Comment on the significance of parameter estimates, what Hosmer-Lemeshow's test tells us about goodness of fit, and point out any remaining issues with diagnostics.

   (c) Comment on the significance of odds ratios and interpret what the model tells us about relationships between the predictors and the odds of a species' maximum lifespan being at least 10 years.

[1] The data is file is based on the sleep data set from statlib located and described on `http://lib.stat.cmu.edu/datasets/sleep`.

2. The index variables in the data set are ordinal, meaning they are categorical and they have a natural ordering. If we treat an index variable as a continuous variable, this will imply a linear change as the index changes. Repeat Exercise 1 by following the parts below and treating the index variables as continuous variables in SAS.

   (a) Determine and specify the best set of predictors through stepwise selection with criteria .05. Comment on any unduly influential points on the final model. Use cut-off as 1. If any extremely unduly influential points exist, remove them for the final model.

   (b) Comment on the significance of parameter estimates, what Hosmer-Lemeshow's test tells us about goodness of fit, and point out any remaining issues with diagnostics.

   (c) Comment on the significance of odds ratios and interpret what the model tells us about relationships between the predictors and the odds of a species' maximum lifespan being at least 10 years.

3. Gestation time is the species' expected number of days in a full-term pregnancy. It might be reasonable for size of an animal to have a relationship with gestation time. Risk of being killed by other species might also be related, for instance from an evolutionary standpoint there may be benefits to a species having a longer or shorter gestation time. Consider a Poisson log-linear model for gestation time with body weight, brain weight, predation index, and sleep exposure index as possible predictors.

   (a) Before determining the best set of predictors for the model (accounting for overdispersion if necessary) comment on any unduly influential points for the full model with all 4 variables. Use cut-off as 1. If any extremely unduly influential points exist, remove them and refit before proceeding with model selection. Repeat this until all observations have Cook' distance less than 1. After removing all influential points, determine predictors you may want to keep or remove based on type 1 and type 3 analyses.

   (b) Based on the results from part (a), determine what terms should be retained for the final model. After choosing the final terms, remove again any points that are still unduly influential (for the reduced model) and refit. Leave no points with Cook's distance greater than 1 in your data.

   (c) Comment on the significance of parameter estimates and point out any remaining issues with diagnostics in your final model. Interpret what the parameter estimates tell us about how the predictors are related to expected gestation time.