Homework #3
Logistic Regression

ENGM 182
Data Analytics
Due Wednesday April 24, 2019

1 NBA predictions revisited

Consider the nbaspread data from Class 6. In class, we fitted a logistic model of favwin as a function of spread. Examine other possible models using additional variables. Note that favwin, favscr, undscr are linearly related (e.g., $favwin = 1$ if $favscr > undscr$, else 0) so be careful which combination of variables you include. One way to select your model is by using the AIC criteria; for background see:

`https://en.wikipedia.org/wiki/Akaike_information_criterion`

1. Test the effect of regional rivalry and whether the favorite team is playing at home on predicting whether the favorite wins.

2 Graduate school admissions

Consider graduate admissions data available on Canvas as "admissions.csv." This dataset has a binary response variable called admit. There are three predictor variables: gre, gpa, and rank. Treat the variables gre and gpa as continuous. The variable rank takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.

1. Characterize the data by calculating means and standard deviations.

2. For the categorial variables of admit and rank, how many observations are in each of the $2 \times 4 = 8$ cells? Recall that the xtabs function can do this.

3. Predict admit as a function of gre score, gpa, and undergraduate school rank. Note that rank should be recoded as a categorical variable with 4 factors. Which of the predictor variables is significant?

4. What does the coefficient on rank = 3 mean? In other words, interpret the number and explain how it changes the probability of admission.

5. What is a 95% confidence interval around each parameter estimate? Hint: use the confint function.

6. Calculate the predicted probability of admission at each value of rank, holding gre and gpa at their mean value.

   - Predict admission probabilities using the logistic function you created earlier in the term.
   - Predict admission probabilities using the predict function we covered previously.

7. Perform a chi-square test on the difference in deviance between the null model and the model with gre, gpa, and rank. Is the fitted model a significant improvement? For background, see the solutions from the Titanic survival analysis. The chi square test is done after question number 10.

8. Optional: create a table that shows how admissions probabilities vary at each rank level across gre scores from 200 to 800 (increment gre by 10) while holding gpa at its mean.

3 Credit Card analysis

Consider a data set of 1000 bank customers. Banks must make decisions regarding whether to approve loans or not. If an applicant is a good credit risk then the cost of not approving a loan is the loss of potentially profitable business to the bank. If the applicant is a bad credit risk, then approving a loan exposes the bank to a significant default risk.

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants. Your task is to create a predictive model using this data to help make decisions about whether to approve loans to potential borrowers dependent upon the variables in the data set.

Load the GermanCredit.csv file from Canvas. Using the “caret” package from the Titanic exercise, create a training data set with 60% of the data and a testing data set with the remaining data. Use logistic regression to predict the “Class” variable outcome (good or bad) as a function of the remaining variables. Try to find a model that best fits the data while also minimizing the AIC criterion.

Once you have specified your model, use it to create predictions of the “Class” outcome and see how accurate your predictions are as compared to the actual outcomes. What do you conclude?

4 Optional question - Titanic revisited

Consider the Titanic analysis. Create 100 training and testing samples. Use “survived” is the variable to split the dataset using the partition function in the “caret” package. Determine the model accuracy of each, calculate the mean of the accuracies, the standard deviation, and plot as a histogram.