Dr. Sudhakar Raju

BIA 6309 – LINEAR & MULTIVARIATE MODELS

SUMMER 2018

**QUESTIONS FOR ASSIGNMENT 6**

I. The attached csv file called “affairs\_data” relates to a cross sectional survey that was published in *“Psychology Today”* and analyzed in a classic paper “*A Theory of Extramarital* *Affairs”,*  Ray Fair, Journal of Political Economy, Vol.86, 1978, pp. 45-61*.* The data relates to 601 participants and includes information on the number of affairs, age, gender (Male=1, Female = 0), number of children (No Children = 0, Children = 1), religiousness on a 5 point scale (1 = Anti Religious to 5 = Religious), occupation (“Hollingshead” scale 1 to 7) and self–rating of marriage (1 = Very Unhappy, 5 = Very happy). YNAFFAIRS is a binary variable created from number of affairs. If a person had at least one extramarital affair, YNAFFAIR is coded as 1, otherwise it is coded as 0.

a.) Why does standard linear regression not work when estimating binary (binomial) outcomes?

b.) Describe the characteristics of the affairs\_data set. What is the mean and median value of the number\_affairs variable?

c.) Create a frequency table showing the distribution of affairs among the survey participants. (In other word, how many participants had no affairs, 1 affair, 2 affairs, etc.). Create a probability table – what percentage of the participants had no affairs, 1 affair, etc.

d.) Run a Logit Model on the data. What are the significant coefficients? Explain the results of the model.

e.) Drop all the insignificant variables and run a reduced model. Does this model fit better?

f.) Interpret the model parameters of the REDUCED LOGIT MODEL.

g.) Using the Reduced Logit Model, create data frames for Logits, Odds and Probabilities. What is the range for probabilities?

h.) How well does the Reduced Model fit the data? In other words, create a Confusion Matrix. What is the accuracy of the model?

i.) Next, contrast the Logistic Regression model with the Linear Discriminant model. The Linear Discriminant is a variant of linear regression that creates separate linear equations (also called “Fisher Classification Functions”) for the binary states. Note that the LDA model requires the outcome variable to be in the form of a categorical variable such as “yes/no”, “faithful/unfaithful”, “churn/no churn”, etc. Thus, use the categorical variable ACTUAL\_STATE as the outcome variable.

What are the linear discriminant equations for “faithful” and “unfaithful”?

j). Fit observation 1 to the linear discriminant equations. According to the LDA equation, what is the predicted class of observation 1?

k.) Create a Confusion Matrix for the LDA model. Contrast the performance of the LDA model with the Logistic Regression model?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

II. Use the attached dataset (credit\_default\_data) which pertains to a sample of 10,000 credit card holders. The dependent variable is default/no default and the independent variables are student /not a student, credit card balance and income.

a.) Run a Logistic Regression on this data. What proportion of the credit card holders default?

b.) What is the predicted probability of default for a student with a credit card balance of $1500 and income of $40,000?

c.) What is the predicted probability of default for a non-student with a credit card balance of $1500 and income of $40,000?

d.) What is the highest credit card balance in the dataset? What is the predicted probability of default for this individual? Does this result make sense?

e.) What is the highest income in the dataset? What is the predicted probability of default for this individual? Does this result make sense?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

III. Use the same credit\_default\_data above. Instead of a Logit Model, run an LDA (Linear Discriminant Model) on the data.

a.) What are the LDA equations (i.e. Fisher Classification functions) for the different groups?

b.) What is the predicted LDA score for a student with a credit card balance of $1500 and income of $40,000? Is this person in the default or no default group according to LDA? Compare the LDA result to Logit result. Are they in agreement?

c.) What is the predicted LDA scores for the individual with the highest income?

d.) What is the predicted LDA scores for a non-student with a balance of $1530 and income of $30,003? What is this individual predicted class?

e.) The data above pertains to ID: 264. What is this individual’s actual class? What is his predicted class?

f.) Interpret the Confusion Matrix for this dataset.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_