STK 320

Assignment 2: Logistic Regression

Question sheet

Submission: Before 05 September 2016, 12:30

Total marks: 60

PLEASE NOTE: make use of a seed value of 390 when completing this assignment

Question 1 (Data Preprocessing):

Data: Description: Predict whether a patient will have a congenital heart defect (CHD), given health factors such as smoking (tobacco), cholesterol (ldl), diabetes (typea), weight (obesity), alcohol consumption and age.

File: Heart_data_stk320
Output variable: chd
Input variables: tobacco, ldl, adiposity, typea, obesity, alcohol, age

1.1 Split the dataset into a 80% training and 20% test set.

a) Copy and paste the part of your code indicating this step [4]

Question 2 (Estimation/Training):

2.1 Perform logistic regression from first principles on the training set.

a) Copy and paste the part of your code indicating this step. [3]

b) Give the estimations for βs [3]

2.2 Perform logistic regression using proc logistic.

a) Give the estimations for βs [3]

Question 3 (Prediction/Testing):

Using the estimation output from proc logistic (question 2.2), get the following counts (cell NOT grey) of the confusion matrix **on the test set**:

Table 3.1 [9]

		Predicted positive	Predicted negative	Total
conditions	Condition positive	True positive (TP)	False Negative (FP)	
	Condition negative	False positive (FP)	True negative (TN)	
	Total			

3.2 From the table in 3.1 get the following: [5]

a) model accuracy (also count R^2)

b) recall for positive and negative (hint: for positive: $recall = \frac{true\ positive}{condition\ positive}$)

c) precision for positive and negative (hint: for positive: $precision = \frac{true\ positive}{predicted\ positive}$

3.3 Display the ROC on the test set [3]

Question 4 [21]

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The outcome variable, admit/don't admit, is binary.

Perform logistic regression on this dataset, using 'proc logistic' in SAS.

File: admit_data.xlsx
Data preprocessing
a) Which predictor variable (s) in the data set is categorical? Which category of this variable will you admit in the analysis? [2]

Part 1

b) Is the overall model statistically significant? Motivate your answer? [2]
c) Which of the predictor variables are statistically significant? Motivate your answer [6]

Part 2

d) Comment on the significance of the variables given the output of the 'Analysis of MLE'. How does it differ from question 4.3? [4]
e) Given your answer in 4.3, would you consider making changes to your predictor variables? Motivate your answer. [2]

Part 3

f) Give an interpretation for the odds ratio of each predictor variable in the model. [5]

Question 5 [9]

Given the output of 'Analysis of MLE' – remove the least significant predictor variable from the dataset. Answer the following questions:

a) Which predictor variable did you remove and why? [1]
b) Is the overall model statistically significant? Motivate your answer? [1]
c) Which of the predictor variables are statistically significant? Motivate your answer [2]
d) Give an interpretation for the odds ratio of each predictor variable in the model. [3]

e) Which one of the two models (question 4 vs question 5) is the best model? Motivate your answer [2].