PH245
Introduction to Multivariate Statistics
Homework Set 3

Due date: November 6, Monday

Problems:

1. The Cleveland Heart Disease dataset contains 303 patients with 13 attributes plus a diagnosis of heart disease. The goal is to distinguish presence (values 1,2,3,4) from absence (value 0) of heart disease given the 13 attributes, which are [1] age, [2] gender, [3] chest pain type (1-4), [4] resting blood pressure, [5] serum cholestoral, [6] fasting blood sugar (1=true,0=false), [7] resting electrocardiographic results, [8] maximum heart rate achieved, [9] exercise induced angina (1=yes; 0=no), [10] ST depression induced by exercise, [11] the slope of the peak exercise ST segment, [12] number of major vessels (0-3) colored by flourosopy, [13] thal: 3 = normal; 6 = fixed defect; 7 = reversable defect. The data can be found in "Data-HW3-CHeartDisease.dat". Remove 6 patients with missing values from the data analysis.

   (a) Understand your data first. Among the 297 remaining patients, how many had heart disease and how many had none? Which predictors are numerical, which are categorical, and which are unclear?

   (b) Fit a logistic regression with the binary response of presence/absence of heart disease, and the 13 predictors. Present the summary of the logistic regression fit.

   (c) Recognizing that variable [3] chest pain type, and variable [13] thal are actually categorical variables, create an appropriate set of dummy variables to encode the chest pain type and the thal type. (You may find the R package "dummies" useful. We still treat variables [7] and [11] as numerical variables in our analysis.) Then refit the logistic regression model with the dummy variables you create, plus the other 11 predictors. Present the summary of the new logistic regression fit.

   (d) Based on the logistic regression fit from (c), interpret the coefficient estimate in front of the predictor [5] serum cholestoral. If one wishes to test the null hypothesis that this coefficient equals zero, what is the p-value of this test? If the significance level is set at 0.05, what is your conclusion of this hypothesis test?

(e) Based on the logistic regression fit from (c), interpret the coefficient estimate in front of the predictor [3] chest pain type 4. If one wishes to test the null hypothesis that this coefficient equals zero, what is the p-value of this test? If the significance level is set at 0.05, what is your conclusion of this hypothesis test?

(f) Based on the logistic regression fit from (c), classify the 297 patients into the presence class or absence class, using the cutoff probability 0.5. What is the misclassification rate for this logistic regression fit?

Policy: You must do the homework on your own. Please ask the Instructor or the GSI if you have any question.

Hint: You may find the following R code helpful to process the data.

```
data<-read.table(file="Data_CHeartDisease.dat", header=FALSE,
      quote="", sep=",")
n.all<-nrow(data)
id.ms<-sort(c(seq(1,n.all)[data[,12]=='?'], seq(1,n.all)[data[,13]=='?']))
data2<-data[-id.ms,]
data2[,12]<-as.numeric(data2[,12]) - 2
data2[,13]<-as.numeric(data2[,13]) - 1
X<-data.matrix(data2[,1:13])
Y<-data2[,14]; Y[Y > 0]<-1
colnames(X)<-c("age", "gender", "chestpain", "bldpressure", "chol",
      "bldsugar", "electrocardio", "heartrate", "angina", "STdepression",
      "STslope", "vessel", "thal")
```