Department of Statistics

COURSE STATS 330/762

Assignment 5, 2010

Instructions: Hand in your completed assignment to the Student Resource Centre by **4pm October 14th.**

The data set for this assignment is in the file **acath.txt** which is available on the course web page.

These data are from a medical study of individuals who had exhibited symptoms of coronary artery disease. The response variable **`tvdlm`** is an indicator of whether or not that patent, when examined, actually has severe coronary artery disease (Three-vessel or left main disease). It has value 1 if severe disease is found, and 0 otherwise.

The other variables in the data set are

sex:	gender , scored as 0=male, 1=female
age:	Age in years
cad.dur:	Duration of the symptoms (weeks)
choleste:	Blood cholesterol reading

1. Read the data into R, and make a data frame. Check for gross errors. (Note that, due to the size of this data set, I have not reproduced it at the end of the assignment. I have not deliberately introduced any errors into the data.) Print out the first 16 lines. [5 marks]

2. Perform a graphical analysis of the data, without fitting any model, that will let you see how the risk factors sex, age, cad.dur and choleste affect the probability of having severe coronary artery disease . [10 marks]

3. Fit a logistic regression model to the data, diagnosing any major problems with the fit. Interpret the coefficients. Does your interpretation confirm the conclusions you reached in question 2? [15 marks]

4. Produce a graph that will illustrate how the fitted probabilities for your model change with age and sex. (You can fix the values of cad.dur and choleste at their averages.) Is the risk higher for males? [10 marks]

5. Extra question for 762 only: Do you think the model you have chosen has any value as a predictive tool?