Department of Statistics

COURSE STATS 330

Assignment 4, 2006

Instructions: Hand in your completed assignment to the Student Resource Centre by 4pm on Thursday 28 September.

Question 1

The file **`diabetes.txt`** contains data collected by the U.S. National Institute of Diabetes and Digestive and Kidney Diseases who conducted a study on 768 adult female Pima Indians living near Phoenix, Arizona. The study investigated risk and prognostic factors for diabetes. The dependent variable is **`test`**, which is a binary variable recording if the patient has diabetes as determined by a standard test. The other variables in the data set are described in the list below.

`pregnant:`	Number of times pregnant
`glucose:`	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
`diastolic:`	Diastolic blood pressure (mm Hg)
`triceps:`	Triceps skin fold thickness (mm)
`insulin:`	2-Hour serum insulin (mu U/ml)
`bmi:`	Body mass index (weight in kg/(height in metres squared))
`diabetes :`	Diabetes pedigree function
`age:`	Age (years)

(a) Read the data into R and make a pairs plot. These data have an unusual feature, which you should diagnose and correct before proceeding further. [10 marks]

(b) Fit a logistic regression model to the data, using test as the response and the other variables as explanatory variables. Do any of the explanatory variables need transforming? Are there any influential points? Make any adjustments to your model you feel necessary. [10 marks]

(c) Develop a formula that will allow you to evaluate the probability that a randomly chosen Pima woman will have diabetes, given the values of the explanatory variables above.
Make up an example (i.e. a set of values of the above variables) to illustrate the use of your formula, using suitable R code to evaluate your formula. [5 marks]

Question 2

Make a plot of Pearson residuals versus fitted values for the data of Question 1. Explain the peculiar pattern that you see. Repeat for deviance residuals. [15 marks]

NB. We don't list the data in this assignment, due to its size, but you may assume that there are no typographical errors in the data set.