Department of Statistics

COURSE STATS 330/762

Assignment 4, 2011

Instructions: Hand in your completed assignment to the Student Resource Centre by **4pm October 3rd.**

The data set for this assignment is in the file **pima.txt** which is available on the course web page. There are no names in the file but the variables are in the order below.

The data set contains measurements on 768 Pima Indian women aged 21 or more. The response variable test records whether or not the subject showed signs of diabetes during routine medical care. The population lives near Phoenix, Arizona, USA. The variables are

pregnant:	Number of times pregnant
glucose:	Plasma glucose concentration after 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 m)^2)
diabetes:	Diabetes pedigree function (a measure of genetic risk)
age:	Age (years)
test:	0=no diabetes, 1=diabetes

The aim of the data analysis is to create a prediction equation that will help predict if a subject has diabetes.

1. Read the data into R, and make a data frame containing the variables. Check for any obvious outliers. Print out the first 16 lines. [5 marks]
2. Make some plots that will let you see how the various explanatory variables affect the response. What are these effects, if any? [5 marks]
3. Fit a model to the data, discarding variables and making transformations as appropriate. [10 marks]
4. From your model, develop a rule that will allow you to predict if a female Pima Indian over 21 will be have diabetes. [10 marks]
5. Evaluate the performance of your rule. [10 marks]

Note that the data for this assignment has **not** been corrupted. However, you should still check for any large outliers when you read the data in. Total marks are 40.