Dr. Sudhakar Raju

BUS 6309 – LINEAR & MULTIVARIATE MODELS

SPRING 2016

**QUESTIONS FOR ASSIGNMENT 5**

I. The attached csv file called “strokes\_data” relates to data collected on patients suffering from strokes before (strokes\_pre\_therapy) and strokes after (strokes\_post\_therapy) treatment with a new drug therapy. The age of the patients and whether the patients were actually placed on the drug or received a placebo is also reported in the data set.

a.) Create a histogram of the number of strokes pre and post therapy. What can be said about the distribution of strokes pre and post therapy?

b.) Formulate a model of the form:

strokes\_post\_therapy = f( treatment, age, strokes\_pre\_therapy)

Run a linear regression. What are the estimated coefficients of this Linear Regression?

c.) What is the estimated number of strokes\_post\_therapy for a 37 year old patient who is treated with the drug and experienced 12 strokes prior to therapy according to Linear Regression?

What is the estimated number of strokes\_post\_therapy for a 23 year old patient who is treated with the placebo and experienced 18 strokes prior to therapy according to Linear Regression?

d.) Run a Poisson regression. What are the estimated coefficients of this Poisson Regression?

e.) What is the estimated number of strokes\_post\_therapy for a 37 year old patient who is treated with the drug and experienced 12 strokes prior to therapy? What is the estimated number of strokes\_post\_therapy for a 23 year old patient who is treated with the placebo and experienced 18 strokes prior to therapy?

II. Consider the following very simple dataset consisting of 4 individuals. (Do this problem in Excel or by hand. The focus here is on understanding the logic of cluster analysis).

|  |  |
| --- | --- |
| INCOME | EDUCATION (Years) |
| $100 | 25 |
| $90 | 20 |
| $50 | 12 |
| $40 | 10 |

a.) Plot the points on a graph. How many clusters are there?

b.) What is the mean and sample SD of the clusters? What are the scaled (normalized) values of income and education?

c.) What are the Euclidean distances of the individual points in the clusters? Depict these distances in the form of a matrix.

d.) Show the scaled (normalized) values of income and education on a graph.

e.) What is the centroid (mean) value of each cluster?

III. The attached data relates to protein consumption (protein\_csv) in a group of European countries.

a.) Run a cluster analysis using only Red Meat and White Meat. Use 3 clusters. What countries fall under each cluster?

b.) Run a cluster analysis using only Red Meat and White Meat but use 7 clusters. What countries fall under each cluster?

**NOTE: R may read Country as a categorical variable and label is as 1,2,3,4,…etc. You may want to use the following command to let R know that Country is a “character” variable and not a categorical variable.**

**protein$Country<-as.character(protein$Country)**