Dr. Sudhakar Raju

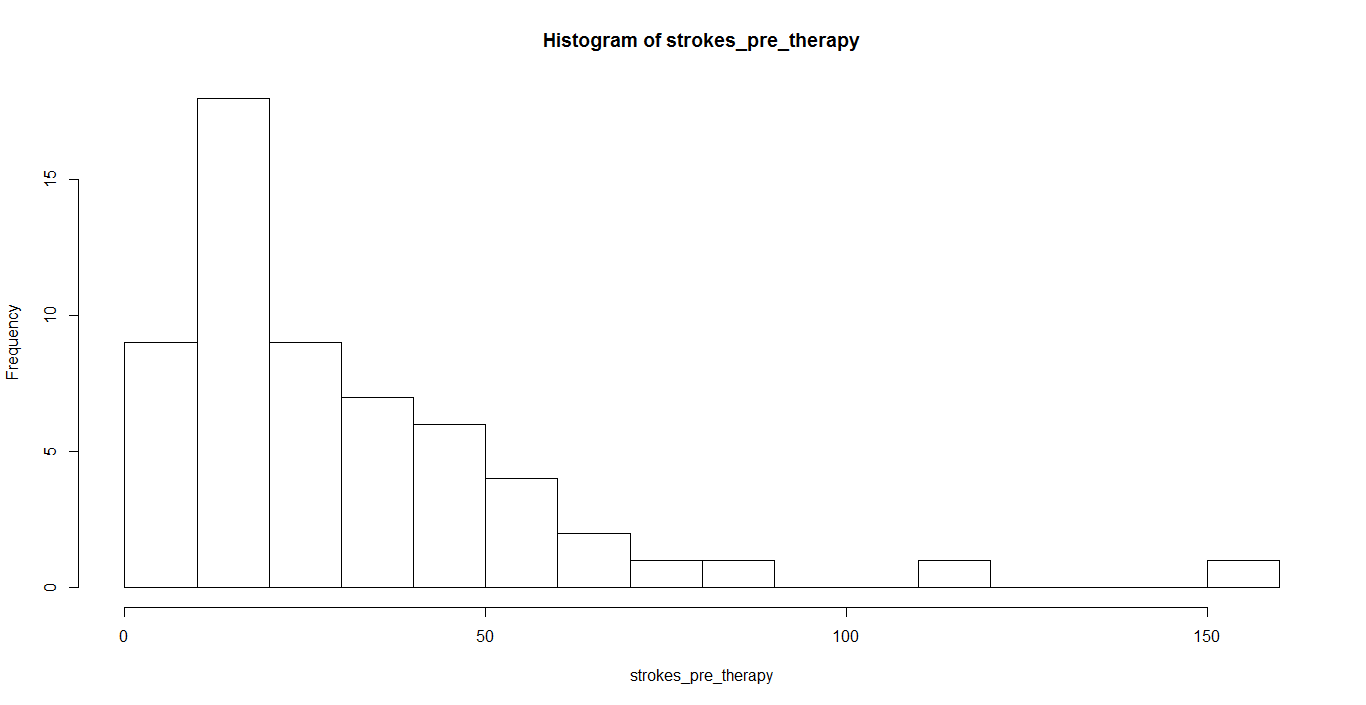
BUS 6309 – LINEAR & MULTIVARIATE MODELS

SPRING 2016

**ANSWERS TO 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.**

1. **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?**



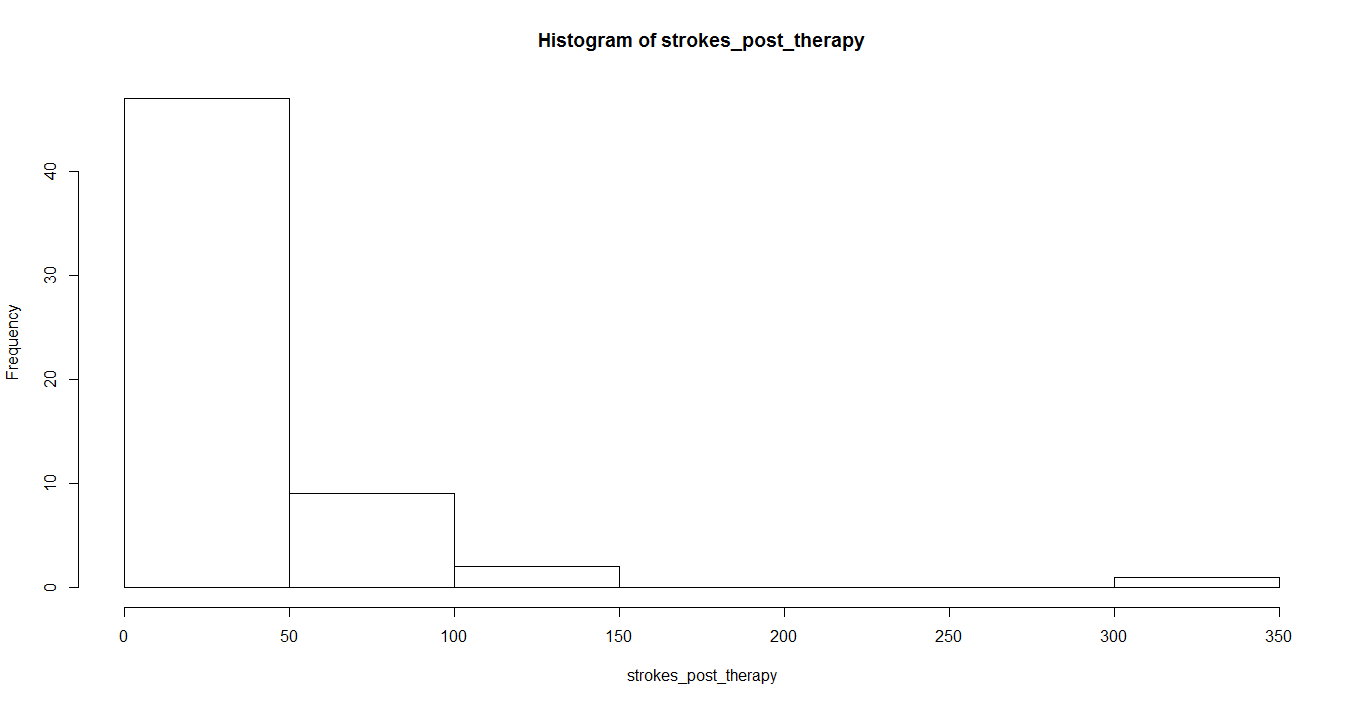
summary(strokes\_pre\_therapy)

Min. 1st Qu. Median Mean 3rd Qu. Max.

6.00 12.00 22.00 31.22 41.00 151.00

(Clearly the data is not Normally distributed – Poisson Distribution to be applied)

hist(strokes\_post\_therapy)



summary(strokes\_post\_therapy)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.00 11.50 16.00 33.05 36.00 302.00

(Clearly the data is not Normally distributed – Poisson Distribution to be applied)

|  |
| --- |
| describe(DATA)  vars n mean sd median trimmed mad min max range skew kurtosis  patient\_id 1 59 30.00 17.18 30 30.00 22.24 1 59 58 0.00 -1.26  strokes\_post\_therapy 2 59 33.05 45.58 16 24.37 14.83 0 302 302 3.86 18.50  treatment\* 3 59 1.47 0.50 1 1.47 0.00 1 2 1 0.10 -2.02  age 4 59 28.34 6.30 28 28.10 7.41 18 42 24 0.29 -0.91  strokes\_pre\_therapy 5 59 31.22 26.88 22 26.80 16.31 6 151 145 2.13 5.73  se  patient\_id 2.24  strokes\_post\_therapy 5.93  treatment\* 0.07  age 0.82  strokes\_pre\_therapy 3.50 |
|  |
| |  | | --- | |  | |

The mean number of strokes pre and post therapy is almost the same. In fact, the mean number of strokes post therapy increased marginally. The SD of strokes post therapy increases significantly from 27 to 46. The most significant effect is on the range. Notice that post treatment the maximum number of strokes increases substantially from 151 to 302.

**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?**

Call:

lm(formula = strokes\_post\_therapy ~ treatment + age + strokes\_pre\_therapy,

data = strokes\_data)

Residuals:

Min 1Q Median 3Q Max

-54.335 -8.803 -0.854 8.078 102.037

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -34.4909 16.9548 -2.034 0.0468 \*

treatmentplacebo 2.8109 6.6825 0.421 0.6757

age 0.7454 0.5438 1.371 0.1761

strokes\_pre\_therapy 1.4441 0.1269 11.382 4.42e-16 \*\*\*

**CODE:**

*REG <- lm(strokes\_post\_therapy~strokes\_pre\_therapy+treatment+age)*

*summary(REG)*

Liner Model Equation:

strokes\_post\_therapy = -34.4 + 1.4441 strokes\_pre\_therapy + treatmentplacebo 2.8109 + 0.7454 age

**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?**

predict(REG)

10.41681264 STROKES

Strokes\_post\_therapy = -34.49 + 2.81 (Placebo = 0, Treatment = 1) + .7454 (37) + 1.44 (12)

= 10.42

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?

11.45702862 STROKES

Strokes\_post\_therapy = -34.49 + 2.81 (Placebo = 0, Treatment = 1) + .7454 (23) + 1.44 (18)

= 11.46

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

Call:

glm(formula = strokes\_post\_therapy ~ treatment + age + strokes\_pre\_therapy,

family = poisson(), data = strokes\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-6.0569 -2.0433 -0.9397 0.7929 11.0061

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.7961250 0.1252994 14.335 < 2e-16 \*\*\*

treatmentplacebo 0.1527009 0.0478051 3.194 0.0014 \*\*

age 0.0227401 0.0040240 5.651 1.59e-08 \*\*\*

strokes\_pre\_therapy 0.0226517 0.0005093 44.476 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2122.73 on 58 degrees of freedom

Residual deviance: 559.44 on 55 degrees of freedom

AIC: 850.71

Number of Fisher Scoring iterations: 5

strokes\_post\_therapy = 1.7961+ 0.0227 strokes\_pre\_therapy + 0.1527 treatmentplacebo + 0.0227 age

**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?**

**Strokes\_post\_therapy = 1.80 + .1527 (placebo = 1, Treatment =0) + .0227 Age + .0226 strokes-pre\_therapy**

**Strokes\_post\_therapy = 1.80 + .1527 (Treatment =0) + .0227 \*37 + .0226 \*12**

**= 2.91**

Note that what is actually estimated is:

As the predicted values are logits – the value to be raised to the power of EXP

**Ln (µ of strokes\_post\_therapy) = 1.80 + .1527 (placebo = 1, Treatment =0) + .0227 Age + .0226 strokes-pre\_therapy**

= 2.91

To solve for µ, exponentiate. Thus:

e2.91 = 18.36 strokes

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?

Similarly, for a 23 year old patient who is treated with the placebo and experienced 18 strokes per-therapy, the estimated number of strokes post therapy according to the Poisson model is:

**Ln (µ of strokes\_post\_therapy) = 1.80 + .1527 (placebo = 1, Treatment =0) + .0227 Age + .0226 strokes-pre\_therapy**

**= 1.80 + .1527 (Placebo = 1) + .0227 \*23 + .0226 \*18**

**= 2.88**

To solve for µ, exponentiate. Thus:

e2.88 = 17.81 strokes

2**. 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** |

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

a.)

There are clearly two clusters here.

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

|  |  |  |
| --- | --- | --- |
| ID | INCOME | EDUCATION |
| A | 100 | 25 |
| B | 90 | 20 |
| C | 50 | 12 |
| D | 40 | 10 |
| **MEAN** | **70.00** | **16.75** |
| **SAMPLE SD** | **29.44** | **6.99** |
|  |  |  |

|  |  |
| --- | --- |
| **ZINCOME** | **ZEDUCATION** |
| 1.0190 | 1.1796 |
| 0.6794 | 0.4647 |
| -0.6794 | -0.6791 |
| -1.0190 | -0.9651 |
| MEAN=0.00 | MEAN =0.00 |
| SAMPLE SD = 1.00 | SAMPLE SD = 1.00 |

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **EUCLIDEAN MATRIX** | | |
|  | **A** | **B** | **C** | **D** |
| **A** | 0.0000 | 0.7915 | 2.5178 | 2.9586 |
| **B** | 0.7915 | 0.0000 | 1.7761 | 2.2201 |
| **C** | 2.5178 | 1.7761 | 0.0000 | 0.4440 |
| **D** | 2.9586 | 2.2201 | 0.4440 | 0.0000 |

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

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

|  |  |  |  |
| --- | --- | --- | --- |
| **CLUSTER 1** | | | |
|  | 1.0190 | 1.1796 |  |
|  | 0.6794 | 0.4647 |  |
| **MEAN (CENTROID)** | 0.8492 | 0.8221 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **CLUSTER 2** | | | |
|  | -0.6794 | -0.6791 |  |
|  | -1.0190 | -0.9651 |  |
| **MEAN (CENTROID)** | -0.8492 | -0.8221 |  |

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

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

> CLUSTER\_DATA1

Country CLUSTER

1 Albania 1

2 Austria 3

3 Belgium 2

4 Bulgaria 1

5 Czechoslovakia 3

6 Denmark 3

7 E Germany 3

8 Finland 1

9 France 2

10 Greece 1

11 Hungary 3

12 Ireland 2

13 Italy 1

14 Netherlands 3

15 Norway 1

16 Poland 3

17 Portugal 1

18 Romania 1

19 Spain 1

20 Sweden 1

21 Switzerland 2

22 UK 2

23 USSR 1

24 W Germany 3

25 Yugoslavia 1

b.)

> CLUSTER\_DATA2

Country CLUSTER2.cluster

1 Albania 4

2 Austria 3

3 Belgium 2

4 Bulgaria 5

5 Czechoslovakia 3

6 Denmark 3

7 E Germany 3

8 Finland 4

9 France 7

10 Greece 4

11 Hungary 1

12 Ireland 2

13 Italy 4

14 Netherlands 3

15 Norway 4

16 Poland 1

17 Portugal 6

18 Romania 6

19 Spain 6

20 Sweden 5

21 Switzerland 2

22 UK 7

23 USSR 4

24 W Germany 3

25 Yugoslavia 6

**R CODE FOR CLUSTER ANALYSIS**

attach(protein\_data)

str(protein\_data)

protein\_data$Country<-as.character(protein\_data$Country)

names(protein\_data)

MEAT<-protein\_data[,c("RedMeat", "WhiteMeat")]

meat<-data.frame(MEAT)

################################

set.seed(1)

CLUSTER1<-kmeans(meat, 3)

CLUSTER1

CLUSTER\_DATA1<-data.frame(Country, CLUSTER1$cluster)

CLUSTER\_DATA1

####################################

set.seed(1)

CLUSTER2<-kmeans(meat, 7)

CLUSTER2

CLUSTER\_DATA2<-data.frame(Country, CLUSTER2$cluster)

CLUSTER\_DATA2

#################################