Dr. Sudhakar Raju

BUS 6309 – LINEAR & MULTIVARIATE MODELS

SPRING 2016

**ANSWERS TO ASSIGNMENT 6**

**1. Customer arrivals at a bank are random and independent – the probability of arrival in any one minute period is the same as the probability of arrival in any other one minute period. Suppose the mean number of arrivals in any one minute period is 3.**

**a.) What is the probability of exactly 3 arrivals in a one minute period?**

The Poisson model is given by:

P (x) = 

µ = 3. In this case, x is also 3. Thus:

P (x=3) = 

= 

= .2240

**b.) What is the probability of at least 3 arrivals in a one minute period?**

ANS: At least 3 arrivals implies that X≤3. Applying the formula above, results in:

P(X=0) → .0499

P(X=1) → .1494

P(X=2) → .2240

P(X=3) → .2240

P (X≤3) = .0499 + .1494 + .2240 + .2240 = .6473 or 64.73%

**2. Phone calls arrive at the rate of 48 per hour at the reservation desk of Regional Airways.**

**a.) What is the probability of receiving exactly 3 calls in a 5 minute period?**

µ = 48 calls in 60 minutes. This means that in a 5 minute period you get 4 calls. Thus, the mean value for a 5 minute window is 4 calls.

a.) P (x=3) = 

= 

= .1953 or 19.53%

**b.) What is the probability of receiving no calls in a 5 minute period?**

ANS: b.

P (x=0) = 

= 

= .0183 or 1.83%

3. Use the attached dataset (credit\_default\_data) which pertains to a sample of 10,000 credit card holders. The dependent variable is default/no default and the independent variables are student /not a student, credit card balance and income.

a. Run a Logistic Regression on this data. What proportion of the credit card holders default?

> table(default)

default

No Yes

9667 333

Thus, only [333/10,000] or 3.33% default.

> summary(LOGIT\_MODEL)

Call:

glm(formula = default ~ student + balance + income, family = binomial(),

data = credit\_default\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.4691 -0.1418 -0.0557 -0.0203 3.7383

Coefficients:

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

(Intercept) -10.869045196 0.492255516 -22.080 < 0.0000000000000002 \*\*\*

student -0.646775807 0.236252529 -2.738 0.00619 \*\*

balance 0.005736505 0.000231895 24.738 < 0.0000000000000002 \*\*\*

income 0.000003033 0.000008203 0.370 0.71152

---

Something to note in the results above is the fact that since balance and income are both measured in the same units (dollars), they are comparable. The coefficient values imply that credit card balance (.0057) has a greater impact on the probability of default than income (.000003).

**b. What is the predicted probability of default for a student with a credit card balance of $1500 and income of $40,000?**

ln(Odds) = -10.8690 - .6468 student + .00574 balance + .000003 income

ln (Odds) = -10.8690 - .6468 (1) + .00574 ($1500) + .000003 ($40,000)

= -10.8690 -.6468 + 8.61 + .12

= -2.7858

Odds = e-2.7858 ≈ .0617

Prob. of Default = Odds / (1+ Odds)

= (.0617) / (1.0617)

= .0581 or 5.81%

**What is the predicted probability of default for a non-student with a credit card balance of $1500 and income of $40,000?**

c. ln (Odds) = -10.8690 - .6468 (0) + .00574 ($1500) + .000003 ($40,000)

= -10.8690 - 0 + 8.61 + .12

= -2.1390

Odds = e-2.1390 ≈ .1178

Prob. of Default = Odds / (1+ Odds)

= (.1178) / (1.1178)

= .1054 or 10.54%

The results above imply that for the same balance and income values, non-students have a higher probability of default. Note also that in R, default is coded as: Yes = 1; No default = 0. If R had coded this as: Yes = 0; No = 1, the interpretation above would be “Probability of No Default”. You can always confirm the way R has coded a factor by using: contrasts (as.factor(default))

**d. What is the highest credit card balance in the dataset? What is the predicted probability of default for this individual? Does this result make sense?**

d. The highest credit card balance in the dataset is for id 8496. You can use Excel’s “Find” function to quickly find the values corresponding to a balance of 2654.32. ID 8496 has the following values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id=8496 | default=Yes | student=1 | balance =2654.32 | income=21930.39 |

ln(Odds) = -10.8690 - .6468 student + .00574 balance + .000003 income

= -10.8690 - .6468 (1) + .00574 (2654.32) + .000003 (21930.9)

= -10.8690 - .6468 + 15.24 + .0658

= 3.79

Odds = e3.79 ≈ 44.26

Prob. of Default = Odds / (1+ Odds)

= (44.26) / (1 + 44.26)

≈.98 or 98%

This make sense- with a large outstanding balance, the probability of default is very high. You can also check the fitted values of probability for id 8496. The exact probability is 97.76%.

**e. What is the highest income in the dataset? What is the predicted probability of default for this individual? Does this result make sense?**

The highest income in the dataset is for id: 5371. You can use Excel’s “Find” function to quickly find the values corresponding to an income of $73,554. ID 5371 has the following values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id:5371 | default=No | student =0 | balance=1593.433 | income=73554.23 |

ln(Odds) = -10.8690 - .6468 student + .00574 balance + .000003 income

= -10.8690 - .6468 (0) + .00574 (1593) + .000003 (73554)

= -10.8690 – 0 + 9.14 + .2207

= -1.5293

Odds = e-1.5293 ≈ .2167

Prob. of Default = Odds / (1+ Odds)

= (.2167) / (1 + .2167)

≈.18 or 18%

This make sense - with a large income, the probability of default is low. You can also check the fitted values of probability for id: 5371. The exact probability is 18.17%.

**4. Use the same credit\_default\_data in question 3 above. Instead of a Logit Model, run an LDA (Linear Discriminant Model) on the data. What are the Fisher Classification functions for the different groups?**

The Fisher Classification functions for the default and no default groups are given by:

Fisher Function = -12.08 + 12.14 student + .0038 balance + .000521 income

(No Default Group)

Fisher Function = -21.56 + 11.77 student + .0085 balance + .000529 income

(Default Group)

**b. What is the predicted Fisher Score for a student with a credit card balance of $1500 and income of $40,000? Is this person in the default or no default group according to LDA? Compare the LDA result to Logit result. Are they in agreement?**

b. Fisher Function = -12.08 + 12.14 (1) + .0038 ($1500) + .000521 ($40,000)

(No Default Group) = -12.08 + 12.14 + 5.70 + 20.84

= 26.60

Fisher Function = -21.56 + 11.77 (1) + .0085 ($1500) + .000529 ($40,000)

(Default Group) = -21.56 + 11.77 + 12.75 + 21.16

= 24.12

Since the **Fisher score is higher for the No Default group, this individual would be classified as No Default**. The Logit Model implies that for students with a balance of $1500 and income of $40,000 the default probability is only 5.81%. The models are clearly in agreement.

Note that the **advantage of the Logit Model** is that it explicitly generates a probability value while LDA simply tells you whether you are part of a particular group.

**C. What is the predicted LDA scores for the individual with the highest income?**

The highest income in the dataset is for id: 5371. ID 5371 has the following values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Id:5371 | default=No | student =0 | balance=1593 | income=73,554 |

Fisher Function = -12.08 + 12.14 (0) + .0038 ($1593) + .000521 ($73,554)

(No Default Group) = -12.08 + 0 + 6.05 + 38.32

= 32.29

Fisher Function = -21.56 + 11.77 (0) + .0085 ($1593) + .000529 ($73,554)

(Default Group) = -21.56 + 0 + 13.54 + 38.91

= 30.89

Since the Fisher score is higher for the No Default group, this individual is predicted to belong to the No Default group.

**d. What is the predicted LDA scores for a non-student with a balance of $1530 and income of $30,003? What is this individual predicted class?**

d.

$functions

No Yes

constant -12.083941 -21.563343

student 12.140185 11.773927

balance 0.003804 0.008508

income 0.000521 0.000529

Fisher Function = -12.08 + 12.14 (0) + .0038 ($1530) + .000521 ($30,003)

(No Default Group) = -12.08 + 0 + 5.81 + 15.63

= 9.36

Fisher Function = -21.56 + 11.77 (0) + .0085 ($1530) + .000529 ($30,003)

(Default Group) = -21.56 + 0 + 13.01 + 15.87

= 7.32

Since the Fisher score is higher for the No Default group, this individual is predicted to belong to the No Default group.

**e. The data above pertains to ID: 264. What is this individual’s actual class? What is his predicted class?**

The predicted class is “No Default”. The Actual Class, is however “Default” Clearly, this individual is misclassified.

**f. Interpret the Confusion Matrix for this dataset.**

$confusion

predicted

original No Yes

No 9645 22

Yes 254 79

$error\_rate

[1] 0.0276

Of the 10,000 individuals, 9645 + 79 = 9724 are classified correctly. The remaining 276 individuals are classified incorrectly (error rate = 276 / 10,000 = 2.76%). Of this incorrectly classified group, 22 were predicted to default but did not actually default. The remaining 254 were not predicted to default, but actually ended up defaulting.

5. See the R code. Cluster analysis does not do as effective a job of identifying the default/no default groups as LDA.

R CODE

library(psych)

library(DiscriMiner)

library(plyr)

library(outliers)

options(scipen=999)

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

dim(credit\_default\_data)

names(credit\_default\_data)

str(credit\_default\_data)

attach(credit\_default\_data)

describe(credit\_default\_data)

table(default)

outlier(balance)

outlier(income)

####################LOGIT MODEL############################

LOGIT\_MODEL<-glm(default~student+balance+income, data=credit\_default\_data,

family=binomial())

summary(LOGIT\_MODEL)

predict(LOGIT\_MODEL)

ODDS<-exp(predict(LOGIT\_MODEL))

ODDS

PROB<-(ODDS)/(1+ODDS)

PROBABILITY<-data.frame(PROB\*100)

describe(PROBABILITY)

#####################LDA MODEL########################

library(DiscriMiner)

contrasts(as.factor(default))

LDA\_MODEL<-linDA(credit\_default\_data[,3:5], credit\_default\_data[,2])

LDA\_MODEL

LDA\_MODEL$scores

LDA\_MODEL$classification

COMPARISON<-data.frame(credit\_default\_data$default,

LDA\_MODEL$classification)

COMPARISON

table(COMPARISON)

**5. Using the credit\_default\_data apply the k-means algorithm.**

a. To apply the k-means algorithm, clean the data. Eliminate the columns with id and default leaving you with purely numeric data.

b. Use the 3 variables (student, balance and income) to create k = 2 clusters.

c. Compare the LDA actual and predicted classification with the k means clusters. How well does cluster analysis perform compared to LDA?

#########################K-MEANS ############################

NUMERIC\_DATA<-credit\_default\_data[-1:-2]

NUMERIC\_DATA

SCALED\_DATA<-scale(NUMERIC\_DATA)

SCALED\_DATA

CLUSTERS<-kmeans(SCALED\_DATA, 2)

CLUSTERS

CLUSTERS$cluster

COMPARE\_METHODS<-data.frame(credit\_default\_data$default,

LDA\_MODEL$classification,

CLUSTERS$cluster)