Dr. Sudhakar Raju

BIA 6309 – LINEAR & MULTIVARIATE MODELS

SUMMER 2018

**ANSWERS FOR ASSIGNMENT 6**

I. a.) A Binomial distribution is one in which a variable has only two possible outcomes – guilty/not guilty, heads/tails, malignant/benign, etc. The outcome variable here (YNAFFAIRS) has only two possible outcomes - no affairs/had affairs. Hence, this is a binomial (binary) outcome and must be estimated using logistic regression.

Standard linear regression is inappropriate when estimating binomial outcomes because it can result in estimated values that are greater than 1 and lesser than zero. Since probabilities cannot be more than 1 (100%) or negative, linear regression is usually inappropriate when there are binomial outcomes. Whenever the outcome variable (dependent or left hand side variable) is binomial, a LOGISTIC REGRESSION or LOGIT MODEL must be applied.

b.)

> describe(affairs\_data)

vars n mean sd median trimmed mad min max range skew kurtosis se

id 1 601 301.00 173.64 301 301.00 222.39 1.00 601 600.00 0.00 -1.21 7.08

YNAFFAIRS 2 601 0.25 0.43 0 0.19 0.00 0.00 1 1.00 1.15 -0.67 0.02

number\_affairs 3 601 1.46 3.30 0 0.55 0.00 0.00 12 12.00 2.34 4.19 0.13

male 4 601 0.48 0.50 0 0.47 0.00 0.00 1 1.00 0.10 -1.99 0.02

age 5 601 32.49 9.29 32 31.37 7.41 17.50 57 39.50 0.88 0.21 0.38

years\_married 6 601 8.18 5.57 7 8.26 8.15 0.12 15 14.88 0.08 -1.57 0.23

children 7 601 0.72 0.45 1 0.77 0.00 0.00 1 1.00 -0.95 -1.09 0.02

religious 8 601 3.12 1.17 3 3.12 1.48 1.00 5 4.00 -0.09 -1.02 0.05

education 9 601 16.17 2.40 16 16.21 2.97 9.00 20 11.00 -0.25 -0.32 0.10

occupation 10 601 4.19 1.82 5 4.34 1.48 1.00 7 6.00 -0.74 -0.79 0.07

marriage\_rating 11 601 3.93 1.10 4 4.07 1.48 1.00 5 4.00 -0.83 -0.22 0.04

The mean number of affairs is 1.46. This value is misleading since the data is skewed by a few outliers – 38 participants in the survey had 12 affairs. A better indicator in the presence of skewed data is the median. The median number of affairs is zero which indicates that at least 50% had no affairs.

c.) > table(number\_affairs)

number\_affairs

0 1 2 3 7 12

451 34 17 19 42 38

|  |
| --- |
| > PROBABILITY\_TABLE  number\_affairs  0 1 2 3 7 12  0.75041597 0.05657238 0.02828619 0.03161398 0.06988353 0.06322795 |

Notice that 75% of the participants had no affairs.

d.)

Call:

glm(formula = YNAFFAIRS ~ male + age + years\_married + children +

religious + education + occupation + marriage\_rating, family = binomial(),

data = affairs\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.5713 -0.7499 -0.5690 -0.2539 2.5191

Coefficients:

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

(Intercept) 1.37726 0.88776 1.551 0.120807

male 0.28029 0.23909 1.172 0.241083

age -0.04426 0.01825 -2.425 0.015301 \*

years\_married 0.09477 0.03221 2.942 0.003262 \*\*

children 0.39767 0.29151 1.364 0.172508

religious -0.32472 0.08975 -3.618 0.000297 \*\*\*

education 0.02105 0.05051 0.417 0.676851

occupation 0.03092 0.07178 0.431 0.666630

marriage\_rating -0.46845 0.09091 -5.153 0.000000256 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 675.38 on 600 degrees of freedom

Residual deviance: 609.51 on 592 degrees of freedom

AIC: 627.51

Number of Fisher Scoring iterations: 4

The coefficients are not easily interpretable since they are in the form of “logarithmic odds”. The variables that best explain the logarithmic odds of having an affair is age, years\_married, religious, and marriage\_rating. Notice that age, religious and marriage\_rating have negative effects. Thus, older participants, more religious persons (higher religious rating implies more religious) and those who self-rated their marriage as happy have lesser logarithmic odds of having extra marital affairs.

Curiously, gender, number of children, education and occupation have no statistically significant effect in explaining marital infidelity.

e.) > summary(REDUCED\_MODEL)

Call:

glm(formula = YNAFFAIRS ~ age + years\_married + religious + marriage\_rating,

family = binomial(), data = affairs\_data)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6278 -0.7550 -0.5701 -0.2624 2.3998

Coefficients:

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

(Intercept) 1.93083 0.61032 3.164 0.001558 \*\*

age -0.03527 0.01736 -2.032 0.042127 \*

years\_married 0.10062 0.02921 3.445 0.000571 \*\*\*

religious -0.32902 0.08945 -3.678 0.000235 \*\*\*

marriage\_rating -0.46136 0.08884 -5.193 0.000000206 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 675.38 on 600 degrees of freedom

Residual deviance: 615.36 on 596 degrees of freedom

AIC: 625.36

Number of Fisher Scoring iterations: 4

The Reduced Model has all statistically significant coefficients. We can demonstrate that this Reduced Model is a better fit by using a Chi\_Square test.

f.)

|  |
| --- |
| > coef(REDUCED\_MODEL)  (Intercept) age years\_married religious marriage\_rating  1.93083017 -0.03527112 0.10062274 -0.32902386 -0.46136144 |

It is important to recognize that in Logit Models, the coefficient estimates are expressed in terms of the logarithm of odds (“logits”). Since logarithms are nothing but exponents, the coef values above are actually exponents to the base e.

To express them in terms of Odds, do the following:

e1.9308 = 6.90

e-.0353 = .9653

e.1006 = 1.106

e-.3290 = .7196

e-.46714 = .63

It is easy to do the above in R:

exp(coef(REDUCED\_MODEL))

(Intercept) age years\_married religious marriage\_rating

6.8952321 0.9653437 1.1058594 0.7196258 0.6304248

The above implies that the odds of an extramarital affair increases by 1.106 for every additional year of marriage. Conversely, the odds of an extramarital affair decrease with age, being more religious and higher happiness marriage rating (the odds for all these 3 variables is less than 1).

g.)

|  |
| --- |
| > describe(PROBABILITY)  vars n mean sd median trimmed mad min max range skew kurtosis se  predict.REDUCED\_MODEL. 1 601 0.25 0.14 0.21 0.23 0.13 0.03 0.74 0.71 0.99 0.55 0.01 |

The minimum and maximum values for probability is 3% and 74%. The probability values make sense since they are between 0% and 100%. Any violation here is a definitive indicator that the model is not formulated correctly.

h.)

> CONFUSION\_MATRIX

predict.REDUCED\_MODEL.

ACTUAL\_STATE FAITHFUL UNFAITHFUL

faithful 0.71880200 0.03161398

unfaithful 0.21297837 0.03660566

**Accuracy Rate = .7188 + .0366 = .7554**

**Error Rate = .2130 + .0316 = .2446**

i.) > LDA\_MODEL$functions

faithful unfaithful

constant -20.4514621 -18.3367814

age 0.6672121 0.6320848

years\_married -0.5469150 -0.4470490

religious 1.9627125 1.6396917

marriage\_rating 4.1032111 3.5867551

j. ) Suppose we fit the above equations for observation 1. Observation 1 implies that:

age = 37; years\_married = 10; religious = 3; marriage\_rating = 4.

Thus:

LDA score for “faithful” = -20.45 + .6672 (37) -.5469 (10) + 1.9627 (3) + 4.1032 (4)

= - 20.45 + 24.69 – 5.47 + 5.89 + 16.41

= 21.07

LDA score for “unfaithful” = -18.34 + .6321 (37) -.4471 (10) + 1.6397 (3) + 3.5868 (4)

= - 18.34 + 23.39 – 4.47 + 4.92 + 14.35

= 19.85

Since the score for “faithful” is higher than the score for “unfaithful”, this person would be classified as “faithful” according to the LDA model. The scores and classification are given by the R commands: LDA\_MODEL$functions; LDA\_MODEL$scores; LDA\_MODEL$classification

k.) > LDA\_MODEL$confusion

predicted

original faithful unfaithful

faithful 430 21

unfaithful 125 25

> LDA\_CONFUSION\_MATRIX

predicted

original faithful unfaithful

faithful 0.71547421 0.03494176

unfaithful 0.20798669 0.04159734

Accuracy Rate = [.7155 + .0416] = .7571 or about 75.71%.

The accuracy rate of the Logistic Regression model is 75.54% while the LDA has an accuracy rate of 75.71%. Essentially, the performance of both models is almost identical.

**R CODE FOR AFFAIRS**

attach(affairs\_data)

names(affairs\_data)

str(affairs\_data)

options(scipen=999)

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

library(psych)

describe(affairs\_data)

FREQUENCY\_TABLE<-table(number\_affairs)

FREQUENCY\_TABLE

PROBABILITY\_TABLE<-prop.table(FREQUENCY\_TABLE)

PROBABILITY\_TABLE

##################LOGISTIC REGRESSION MODEL######

FULL\_MODEL<-glm(YNAFFAIRS~male+age+years\_married+children+

religious+education+occupation+marriage\_rating,

data=affairs\_data, family=binomial())

summary(FULL\_MODEL)

REDUCED\_MODEL<-glm(YNAFFAIRS~age+years\_married+religious+marriage\_rating,

data=affairs\_data, family=binomial())

summary(REDUCED\_MODEL)

coef(REDUCED\_MODEL)

exp(coef(REDUCED\_MODEL)) ####This tranforms regression coefficients into Odds###

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

LOGITS<-data.frame(predict(REDUCED\_MODEL))

LOGITS

ODDS<-data.frame(exp(LOGITS))

ODDS

PROBABILITY<-data.frame(ODDS/(1+ODDS))

PROBABILITY

min(PROBABILITY)

max(PROBABILITY)

describe(PROBABILITY)

############CONFUSION MATRIX###########

ACTUAL\_STATE<- ifelse(number\_affairs>0, "unfaithful", "faithful")

ACTUAL\_STATE

PREDICTED\_STATE<-ifelse(PROBABILITY>.50, "UNFAITHFUL", "FAITHFUL")

PREDICTED\_STATE

COMPARISON<-data.frame(ACTUAL\_STATE, PREDICTED\_STATE)

TABLE<-table(COMPARISON)

TABLE

CONFUSION\_MATRIX<-prop.table(TABLE)

CONFUSION\_MATRIX

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

library(DiscriMiner)

LDA\_DATASET<- data.frame(ACTUAL\_STATE, age, years\_married,religious,

marriage\_rating)

LDA\_MODEL<-linDA((LDA\_DATASET) [,2:5],ACTUAL\_STATE)

summary(LDA\_MODEL)

LDA\_MODEL$functions

LDA\_MODEL$scores

LDA\_MODEL$classification

LDA\_MODEL$confusion

LDA\_TABLE<-LDA\_MODEL$confusion

LDA\_CONFUSION\_MATRIX<-prop.table(LDA\_TABLE)

LDA\_CONFUSION\_MATRIX

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

II.

a.

> 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. 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%

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. 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. The highest income in the dataset is for id: 5371. 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%.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

a. The LDA functions (also called “Fisher Classification functions”) for the default and no default groups are given by:

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

(No Default Group)

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

(Default Group)

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

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

= 26.60

LDA 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 LDA 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. 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 |

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

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

= 32.29

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

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

= 30.89

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

d.

$functions

No Yes

constant -12.083941 -21.563343

student 12.140185 11.773927

balance 0.003804 0.008508

income 0.000521 0.000529

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

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

= 9.36

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

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

= 7.32

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

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

f. $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.

**R CODE FOR CREDIT DEFAULT DATASET**

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)

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