Dr. Sudhakar Raju

BIA 6309 – LINEAR & MULTIVARIATE MODELS

SPRING 2017

**ANSWERS FOR ASSIGNMENT 4**

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 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 variables that best explain the number of extramarital affairs 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 probability 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.

**R CODE FOR AFFAIRS**

attach(affairs\_data)

library(psych)

names(affairs\_data)

options(scipen=999)

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

describe(affairs\_data)

FREQUENCY\_TABLE<-table(number\_affairs)

FREQUENCY\_TABLE

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

PROBABILITY\_TABLE

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))

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

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)

II.

a. > FREQ\_TABLE<-table(default)

> FREQ\_TABLE

default

No Yes

9667 333

> PROB\_TABLE<-prop.table(FREQ\_TABLE)

> PROB\_TABLE

default

No Yes

0.9667 0.0333

About 3.33% of credit card holders default.

b. > MODEL1<-glm(default~ balance, data=Default, family=binomial())

> summary(MODEL1)

Call:

glm(formula = default ~ balance, family = binomial(), data = Default)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2697 -0.1465 -0.0589 -0.0221 3.7589

Coefficients:

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

(Intercept) -10.6513306 0.3611574 -29.49 <0.0000000000000002 \*\*\*

balance 0.0054989 0.0002204 24.95 <0.0000000000000002 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom

Residual deviance: 1596.5 on 9998 degrees of freedom

AIC: 1600.5

Number of Fisher Scoring iterations: 8

c. It is important to be clear about how R is defining categorical variables such as default probability. [Default = 0 & No Default = 1] is one possibility. It is also possible that [Default = 1 & No Default = 0]. To verify the manner in which R is defining default/no default use the following command:

contrasts(as.factor(default))

The result in R is:

|  |
| --- |
| > contrasts(as.factor(default))  Yes  No 0  Yes 1 |

Thus, No Default is coded in R as 0 and Yes Default is coded as 1.

The Logistic Regression above implies that:

Ln(Odds) = -10.6513 + .0055 balance

The fitted value at a balance of $2000 is:

Ln(Odds) = -10.6513 + .0055 ($2000)

= .3487

This looks like a probability but it is not. The above is actually the natural logarithm of Odds. To solve for Odds, exponentiate:

e.3487 = 1.417

Next, recognize that:

Probability = Odds / (1 + Odds) = 1.417 / (1 + 1.417) = .5863 or 58.63%

d.

> contrasts(as.factor(student))

Yes

No 0

Yes 1

> MODEL2<-glm(default~ student, data=Default, family=binomial())

> summary(MODEL2)

Call:

glm(formula = default ~ student, family = binomial(), data = Default)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.2970 -0.2970 -0.2434 -0.2434 2.6585

Coefficients:

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

(Intercept) -3.50413 0.07071 -49.55 < 0.0000000000000002 \*\*\*

studentYes 0.40489 0.11502 3.52 0.000431 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom

Residual deviance: 2908.7 on 9998 degrees of freedom

AIC: 2912.7

Number of Fisher Scoring iterations: 6

e. The Logistic Regression above implies that:

Ln(Odds) = -3.5041 + .4049 student[Yes]

If the credit card holder is a student (student = 1), the fitted value is:

Ln(Odds) = -3.5041 + .4049 (1)

= -3.10

To solve for Odds, exponentiate:

e-3.10 = .0451

Thus:

Probability of Default for Students = [Odds / (1 + Odds)] = [.0451 / (1 + .0451)]

= .0432 or 4.32%

Similarly, the probability of default for non-students is given by:

Ln(Odds) = -3.5041 + .4049 (0)

= -3.5041

To solve for Odds, exponentiate:

e-3.5041 = .0301

Thus:

Probability of Default for Non-Students = [Odds / (1 + Odds)] = [.0301 / (1 + .0301)]

= .0292 or 2.92%

The results indicate that the probability of default for students is higher than that of non-students.

**R CODE FOR CREDIT CARD DEFAULTS**

library(psych)

attach(creditdefault\_data)

names(creditdefault\_data)

dim(creditdefault\_data)

str(creditdefault\_data)

options(scipen=999)

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

describe(creditdefault\_data)

FREQ\_TABLE<-table(default)

FREQ\_TABLE

PROB\_TABLE<-prop.table(FREQ\_TABLE)

PROB\_TABLE

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

contrasts(as.factor(default))

MODEL1<-glm(default~ balance, data=creditdefault\_data, family=binomial())

summary(MODEL1)

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

contrasts(as.factor(student))

MODEL2<-glm(default~ student, data=Default, family=binomial())

summary(MODEL2)

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