

# MA318 computer based R test answer template

– Postgraduate Version

Please read the following instructions carefully before start answering the questions:

- Please answer ALL TWO questions and Upload your Rmarkdown and output files to FASER.
- There are 100 marks in total
- Please write your code and comments in the R markdown template (.rmd) and save it as “MA318\_registrationnum”. For example, if your registration number is 2000999, then save your file as MA318\_2000999.rmd. After you have done all your questions, you should knit your R markdown to produce either .html or .doc or .pdf file, the name should be the same as your rmarkdown file.

## Q1 Solution:

- 1) Read the "result.csv" file in R, show the first 6 entries of the dataset.

```
library(boot)

# Q1 1)
result<-read.csv("result.csv")
head(result)

##   numresit department pass
## 1         0   Psychology   1
## 2         0   Sociology   1
## 3         0   Psychology   1
## 4         0   Psychology   0
## 5         0   Psychology   0
## 6         0   Sociology   1
```



Your comments: There are 420 observations and 3 variables.

- 2) Conduct a cross-validation on the data set by doing the following steps:
- (i) Using the last three digits of your registration number as the random seed, split the data set into training and test set. Your training set should contain 300 randomly selected observations from the data set.

```
# Q1 2) (i)
set.seed(511)
train.index=sample(x=420,size=300)
result.train=result[train.index, ]
result.test=result[-train.index, ]
```



- (ii) Taking "pass" as the response variable and all other columns as features, fit a logistic regression model on the training set. Comments on which variables are statistically significant associated with "pass" under 5 percent significance level (0.05) and explain the meaning of coefficient of "numresit" in terms of odds ratio.

```
# Q1 2) (ii)
fit1<-glm(pass~.,family=binomial(link="logit"),data=result.train)
summary(fit1)

##
## Call:
## glm(formula = pass ~ ., family = binomial(link = "logit"), data =
## result.train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3802   0.3483   0.5709   0.7167   1.3251
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.7318    0.2500   6.927 4.31e-12 ***
## numresit       -0.6908    0.1517  -4.553 5.30e-06 ***
```

```
## departmentPsychology  -0.5036      0.3457  -1.457   0.1451
## departmentSociology   1.0403      0.5671   1.834   0.0666 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 318.66  on 299  degrees of freedom
## Residual deviance: 280.87  on 296  degrees of freedom
## AIC: 288.87
##
## Number of Fisher Scoring iterations: 5

b.logit=coefficients(fit1)
b.logit

##           (Intercept)                numresit departmentPsychology
departmentSociology
##           1.7317619             -0.6908060             -0.5036352
1.0402666

est=exp(b.logit)
est

##           (Intercept)                numresit departmentPsychology
departmentSociology
##           5.6506011             0.5011720             0.6043298
2.8299715
```

Your comments: The numresit is the only variable that statistically significant under 5 percent significance level. Every unit increase of a number the student had taken resit the exam is associated with reducing 0.501 times odds of passing.

- (iii) Define a cost function to compute the classification error rate for the cross-validation pipeline. Use 0.5 as the prediction threshold, i.e. classify as “pass” if the predicted probability is > 0.5. Use the defined cost function to compute the 10-fold cross-validation error of your model on the training set. Report the cross-validation error on training set.

```
# Q1 2) (iii)
#define cost function
error.rate.cost = function(y,prediction)
{
  predict.tf=prediction>0.5
  error.rate=mean(predict.tf!=y)
  return(error.rate)
}
#calculate the error
cv.error=cv.glm(data=result.train,glmfit=fit1,cost=error.rate.cost,K=10)
error=cv.error$delta[1]
error
```

```
## [1] 0.22
```

Your comments: The error is 0.22.

- 3) Using the fitted model in 2), make predictions using 0.5 threshold respectively on the test set and compute and report the test classification error rate. Note that the “pass” column takes directly values of 0 and 1.

```
# Q1 3)
predict.test=predict(object=fit1,newdata=result.test,type='response')
tf.test=rep(0,length(predict.test))
tf.test[predict.test>0.5]=1
rate=mean(result.test$pass!=tf.test)
rate

## [1] 0.2416667
```

Your comments: The test classification error rate is 0.242.

- 4) Using the predictions in 3) and table() function, create the confusion matrix on the test set. Report the number of true negative and false positive.

```
# Q1 4)
table(result.test$pass,tf.test)

##      tf.test
##      0  1
## 0   5 25
## 1   4 86
```

Your comments: The true negative is 5 and the false positive is 25.

## Q2 Solution

- 1) Read the data “survey.csv” into R and report the sample size and the number of male interviewers and the number of female interviewers.

```
#Q2 1)
survey<-read.csv("survey.csv")
dim(survey)

## [1] 500  3

head(survey)

##   num size gender
## 1    0    L     M
## 2    1    M     M
## 3    0    L     M
## 4    0    L     M
## 5    0    L     M
## 6    1    M     M


table(survey$gender)
```

```
##
##   F   M
##  87 413
```

Your comments: There are 500 observations. There are 413 male interviewers and 87 female interviewers.

- 2) Fit the Poisson regression model to investigate the associations of number of surveys with the size of the family and the gender of the interviewers. Print the summary of your model and identify which covariate(s) is(are) statistically significantly associated with number of accidents under 5 percent significance level (0.05).

```
#Q2 2)
fit2<-glm(num~.,family=poisson(link="log"),data=survey)
summary(fit2)
```



```
##
## Call:
## glm(formula = num ~ ., family = poisson(link = "log"), data = survey)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.9732  -0.6846  -0.3572   0.2751   2.2462
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.07582    0.18802  -5.722 1.05e-08 ***
## sizeM         1.59641    0.19367   8.243  < 2e-16 ***
## sizeS         2.38648    0.18002  13.257  < 2e-16 ***
## genderM       -0.37516    0.07411  -5.062 4.15e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 694.89  on 499  degrees of freedom
## Residual deviance: 201.69  on 496  degrees of freedom
## AIC: 1225.6
##
## Number of Fisher Scoring iterations: 6
```

Your comments: The variables representing family sizeM, sizeS, and male interviewer are statistically significant under 5 percent significance level.

- 3) Calculate the expected average number of surveys for
  - i) small size family interviewed by female interviewer;
  - ii) middle size family interviewed by female interviewer;
  - iii) large size family interviewed by female interviewer;

- iv) small size family interviewed by male interviewer;
- v) middle size family interviewed by male interviewer;
- vi) large size family interviewed by male interviewer;

and report on average which combination of size and gender will give the smallest expected number of surveys.

```
#Q2 3)
b.pois=coefficients(fit2)
b.pois

## (Intercept)      sizeM      sizeS      genderM
## -1.0758210    1.5964099    2.3864797   -0.3751603

#i)
exp(b.pois[1]+b.pois[3])

## (Intercept)
##    3.708616

#ii)
exp(b.pois[1]+b.pois[2])

## (Intercept)
##    1.683019

#iii)
exp(b.pois[1])

## (Intercept)
##    0.3410177

#iv)
exp(b.pois[1]+b.pois[3]+b.pois[4])

## (Intercept)
##    2.548483

#v)
exp(b.pois[1]+b.pois[2]+b.pois[4])

## (Intercept)
##    1.156535

#vi)
exp(b.pois[1]+b.pois[4])

## (Intercept)
##    0.2343402
```

Your comments: The large size family interviewed by male interviewer gives the lowest expected number of surveys.

- 4) Comment on if the fitted model in part 2) is significantly better than the NULL model under 5 percent significance level (0.05), based on the deviance value of the models.

```
#Q2 4)
#calculate model deviance
694.89-201.69

## [1] 493.2

#calculate difference in degree of freedom
499-496

## [1] 3

#calculate p-value
1-pchisq(493.2,3)

## [1] 0
```



Your comments: The fitted model is better than the null model under 5 percent significance level.