UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER II
SESSION 2023/2024

COURSE NAME	:	CATEGORICAL DATA ANALYSIS
COURSE CODE	:	BWB 21703
PROGRAMME CODE	:	BWQ
EXAMINATION DATE	:	JULY 2024
DURATION	:	2 HOURS 30 MINUTES
INSTRUCTIONS	:	1. ANSWER ALL QUESTIONS 2. THIS FINAL EXAMINATION IS CONDUCTED VIA □ Open book ☒ Closed book 3. STUDENTS ARE **PROHIBITED** TO CONSULT THEIR OWN MATERIAL OR ANY EXTERNAL RESOURCES DURING THE EXAMINATION CONDUCTED VIA

THIS QUESTION PAPER CONSISTS OF **SEVEN (7)** PAGES

Q1 (a) (i) Explain **THREE (3)** assumptions for the generalised linear model (GLM).

(3 marks)

(ii) Explain **THREE (3)** components of GLM appropriately.

(3 marks)

(iii) Explain deviance appropriately.

(2 marks)

(b) In a study that involve 500 pregnant women has the dependent variable known as the occurrence of preterm birth and the related explanatory variables of preterm birth are the age of the woman, socio-economic status, body mass index, smoking status, bleeding during pregnancy, serum level and several dietary factors. Formulate the situation by forming an appropriate GLM modelling framework.

(4 marks)

(c) (i) Describe **FOUR (4)** characteristics of log-linear model.

(4 marks)

(ii) Identify **TWO (2)** advantages of using the log-linear model.

(4 marks)

Q2 In 2017, a group of researchers conducted a cross-sectional survey using dataset titled 'Balance's Affairs', which included 600 respondents. The study examined nine variables: frequency of affairs in past years, age, gender, education level, years of married, presence of children (yes/no), religiosity (rated on 5-point scale from 1=not religious to 5=very religious), occupation (categorised into 7-point scale), and a self-assessed marital happiness (rated from 1=very unhappy to 5=very happy). The researchers applied logistic regression to analyse the data and the results are presented in **Table APPENDIX A.1** generated by R software.

(a) Discuss the coefficient estimations and its significant values based on Model A.

(5 marks)

(b) Discuss the coefficient estimations and its significant values based on Model B.

(5 marks)

(c) Explain the use of analysis of variance (ANOVA) and its result as shown in **Table APPENDIX A.1.**

(4 marks)

(d) Interpret the parameters of Model B.

(4 marks)

(e) Explain overdispersion appropriately.

(2 marks)

TERBUKA

Q3 A study tracked the success of students in gaining admission to Universiti Kebangsaan Malaysia (UKM) via an admissions test, which was measured by a binary outcome variable (`success`). The explanatory variables analysed are the student's numeracy test scores (`numeracy`) and their anxiety test scores (`anxiety`). The results of this analysis are displayed in **Table APPENDIX A.2** as generated by the R software.

(a) Define the response variable.

(2 marks)

(b) Identify the values of mean for `numeracy` and `anxiety`.

(2 marks)

(c) Identify the value of odds ratio for `numeracy`.

(1 mark)

(d) Determine the default link function for a binary outcome variable.

(1 mark)

(e) Discuss the coefficient estimations and its significant values for Model C.

(6 marks)

(f) Discuss the value of null deviance and residual deviance for Model C.

(5 marks)

(g) Demonstrate the relationship between the response variable and two explanatory variables.

(3 marks)

Q4 (a) A two-way contingency table is presented in **Table APPENDIX A.3** illustrating the frequency scale of students' achievement across three objectives: academic grades, athletic ability and popularity in sport, with a frequency range set from four to six.

(i) Construct a contingency table for marginal distribution.

(3 marks)

(ii) Compute the expected counts for each of the cell.

(9 marks)

(b) **Table APPENDIX A.4** presents the results from R software's for chi-square test of independence, which was conducted to determine whether there is an association between the type of school areas (A, B and C) and the scores in three subjects at their respective schools.

(i) Construct a two-way contingency table.

(3 marks)

(ii) Conduct the hypothesis testing and interpret the result.

(5 marks)

- END OF QUESTIONS –

APPENDIX A

Table APPENDIX A.1

```
> ModelA = glm(ynaffair ~ gender + age + yearsmarried + children + religiousness
+ education + occupation +rating,data=Affairs, family=binomial())
> summary(ModelA)

Call:
glm(formula = ynaffair ~ gender + age + yearsmarried + children +
religiousness + education + occupation + rating, family = binomial(),
data = Affairs)

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
gendermale     0.28029    0.23909   1.172 0.241083
age           -0.04426    0.01825  -2.425 0.015301 *
yearsmarried   0.09477    0.03221   2.942 0.003262 **
childrenyes    0.39767    0.29151   1.364 0.172508
religiousness -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
rating        -0.46845    0.09091  -5.153 2.56e-07 ***
---
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 599  degrees of freedom
Residual deviance: 609.51  on 591  degrees of freedom
AIC: 627.51

Number of Fisher Scoring iterations: 4
```

Table APPENDIX A.1 (continue)

```
> ModelB = glm(ynaffair ~ age + yearsmarried + religiousness + rating,
+ data=Affairs, family=binomial())
> summary(ModelB)

Call:
glm(formula = ynaffair ~ age + yearsmarried + religiousness +
    rating, family = binomial(), data = Affairs)

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 *
yearsmarried  0.10062    0.02921   3.445 0.000571 ***
religiousness -0.32902    0.08945  -3.678 0.000235 ***
rating       -0.46136    0.08884  -5.193 2.06e-07 ***
---
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 599  degrees of freedom
Residual deviance: 615.36  on 595  degrees of freedom
AIC: 625.36

Number of Fisher Scoring iterations: 4



> anova(ModelA, ModelB,test="Chisq")
Analysis of Deviance Table

Model 1: ynaffair ~ gender + age + yearsmarried + children + religiousness +
    education + occupation + rating
Model 2: ynaffair ~ age + yearsmarried + religiousness + rating
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       591     609.51
2       595     615.36 -4  -5.8474   0.4509
```

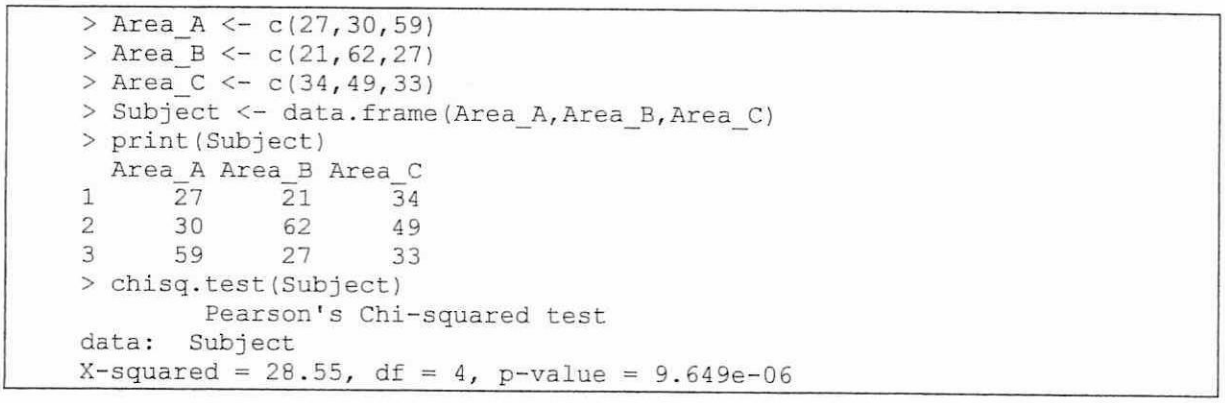
Table APPENDIX A.2

```
> A <- structure(list(numeracy = c(6.6, 7.1, 7.3, 7.5, 7.9, 7.9, 8,
+ 8.2, 8.3, 8.3, 8.4, 8.4, 8.6, 8.7, 8.8, 8.8, 9.1, 9.1, 9.1, 9.3,
+ 9.5, 9.8, 10.1, 10.5, 10.6, 10.6, 10.6, 10.7, 10.8, 11, 11.1,
+ 11.2, 11.3, 12, 12.3, 12.4, 12.8, 12.8, 12.9, 13.4, 13.5, 13.6,
+ 13.8, 14.2, 14.3, 14.5, 14.6, 15, 15.1, 15.7), anxiety = c(13.8,
+ 14.6, 17.4, 14.9, 13.4, 13.5, 13.8, 16.6, 13.5, 15.7, 13.6, 14,
+ 16.1, 10.5, 16.9, 17.4, 13.9, 15.8, 16.4, 14.7, 15, 13.3, 10.9,
+ 12.4, 12.9, 16.6, 16.9, 15.4, 13.1, 17.3, 13.1, 14, 17.7, 10.6,
+ 14.7, 10.1, 11.6, 14.2, 12.1, 13.9, 11.4, 15.1, 13, 11.3, 11.4,
+ 10.4, 14.4, 11, 14, 13.4), success = c(0L, 0L, 0L, 1L, 0L, 1L,
+ 0L, 0L, 1L, 0L, 1L, 1L, 0L, 1L, 0L, 0L, 0L, 0L, 0L, 1L, 0L, 0L,
+ 1L, 1L, 1L, 0L, 0L, 0L, 1L, 0L, 1L, 0L, 0L, 1L, 1L, 1L, 1L, 1L,
+ 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L, 1L)), .Names = c("numeracy",
+ "anxiety", "success"), row.names = c(NA, -50L), class = "data.frame")
> names(A)
[1] "numeracy" "anxiety"  "success"
> dim(A)
[1] 50  3
> head(A)
  numeracy anxiety success
1      6.6    13.8       0
2      7.1    14.6       0
3      7.3    17.4       0
4      7.5    14.9       1
5      7.9    13.4       0
6      7.9    13.5       1
> ModelC <- glm(success ~ numeracy * anxiety, family=binomial)
> summary(ModelC)

Call:
glm(formula = success ~ numeracy * anxiety, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.85712  -0.33055   0.02531   0.34931   2.01048

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.87883   46.45256   0.019    0.985
numeracy          1.94556    4.78250   0.407    0.684
anxiety          -0.44580    3.25151  -0.137    0.891
numeracy:anxiety -0.09581    0.33322  -0.288    0.774

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.029  on 49  degrees of freedom
Residual deviance: 28.201  on 46  degrees of freedom
AIC: 36.201

Number of Fisher Scoring iterations: 7
> mean(numeracy)
[1] 10.722
> mean(anxiety)
[1] 13.954
> exp(1.94556)
[1] 6.997549
```

Table APPENDIX A.3

	Frequency Scale		
Goals	4	5	6
Grades score	49	50	69
Athletic ability	24	36	38
Sports popularity	19	22	28

Table APPENDIX A.4



```
    > Area_A <- c(27,30,59)
    > Area_B <- c(21,62,27)
    > Area_C <- c(34,49,33)
    > Subject <- data.frame(Area_A,Area_B,Area_C)
    > print(Subject)
      Area_A Area_B Area_C
    1     27     21     34
    2     30     62     49
    3     59     27     33
    > chisq.test(Subject)
            Pearson's Chi-squared test
    data:  Subject
    X-squared = 28.55, df = 4, p-value = 9.649e-06
```