Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2017/2018

COURSE NAME	:	CATEGORICAL DATA ANALYSIS
COURSE CODE	:	BWB 31703
PROGRAMME	:	BWQ
EXAMINATION DATE	:	DECEMBER 2017/JANUARY 2018
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ALL** QUESTIONS

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

Q1 (a) Explain **THREE (3)** assumptions of generalised linear model (GLM).

(6 marks)

(b) Describes **THREE (3)** components of GLM.

(6 marks)

(c) Discuss **ONE (1)** situation to apply binary logistic in fitting GLM.

(3 marks)

(d) Define the meaning of deviance and state **TWO (2)** forms of deviance.

(4 marks)

(e) **Table Q1(e)** shows the summary model of GLM by using `mtcars` dataset, where variable `vs` indicates, if a car has a V-engine or a straight-engine. One vehicle company decided to develop a model to predict the probability of a vehicle either having a V-engine or a straight-engine, given a weight (`wt`) of 4.8kg and engine displacement (`disp`) of 210 cubic inches. Briefly explain each estimates of coefficients at 0.05 significant level.

(6 marks)

Table Q1(e): Summary Model of `mtcars`

```
> model <- glm(formula= vs ~ wt + disp, data=mtcars, family=binomial)
> summary(model)
Call:
glm(formula = vs ~ wt + disp, family = binomial, data = mtcars)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.67506  -0.28444  -0.08401   0.57281   2.08234
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.60859    2.43903   0.660    0.510
wt           1.62635    1.49068   1.091    0.275
disp        -0.03443    0.01536  -2.241    0.025 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 43.86  on 31  degrees of freedom
Residual deviance: 21.40  on 29  degrees of freedom
AIC: 27.4
Number of Fisher Scoring iterations: 6
```

Q2 Dataset in **Table Q2** consists test scores for numeracy and anxiety for predictors and success as binary outcome variable. This study aims to determine whether the students pass to the admission of University of Rotterdam, Netherland through test scores. A GLM analysis has been applied to identify success on the numeracy and anxiety scores.

Table Q2: Test Scores

```
> 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(0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0,
+ 0, 0, 0, 0, 1, 0, 0, + 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1,
+ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, + 1, 1)),Names = c("numeracy",
+ "anxiety", "success"), row.names = c(NA, -50L), + class =
+ "data.frame")
> names(A)
[1] "numeracy" "anxiety"  "success"
> 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
> model <- glm(success ~ numeracy * anxiety, family=binomial)
> summary(model)
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
```

(a) Identify the response variable based on **Table Q2**.

(6 marks)

(b) State the default link function for a binomial GLM outcome variable.

(2 marks)

(c) Explain each of estimates coefficients with their significant value.

(8 marks)

(d) Discuss the value of null and residual deviance in the model.

(6 marks)

(e) State the equation based on **Table Q2** to represent the model.

(3 marks)

Q3 (a) Suppose that log transformation applies to response variable. Explain possible differences between log-linear model and simple linear regression.

(4 marks)

(b) What does ordinal or nominal mean?

(4 marks)

(c) State the most common model for nominal outcomes.

(3 marks)

(d) Milk Dairy Company in Johor investigated an effect of percentage toward crude protein (`CP`) with three stages of `Lactation` which are `Early`, `Mid` and `Late`. The cows in the `Early` lactation stage averaged 123 days in milk, `Mid` stage averaged 175 days in milk and the `Late` stage averaged 221 days in milk. Given that four treatment diets which are CP percentage with values of 15%, 17%, 19% and 21% of diet dry matter and CP was treating as quantitative variable in the model. The outcome variable, urea urine nitrogen (`UUN`) is measured in grams per day. The experiment conducting to four cows at each stage of lactation are given each diet for one week with `UUN` measurements being taken at the end of this period, a total of $n = 4 \times 3 \times 4 = 48$ measurements. Finally, two models produce (Model A and Model B), first with `CP` and `Lactation` as inputs and second with an interaction term included. The assumption are the dairy cows produce milk for at least 180 days after calving and the amount of milk produced per day changes over time. The output of both models summarises in **Table Q3(d)(i)**.

Table Q3(d)(i): Model A and Model B

Model A

```
lm(formula = UUN ~ CP + Lactation)
                coef.est coef.se
(Intercept)      -419.3    29.4
CP                 33.8     1.6
LactationMid        4.6     8.7
LactationLate     -11.6     8.7

n=48, k=4
residual sd=24.73, R-Squared=0.91
```

Model B

```
lm(formula = UUN ~ CP * Lactation)
                    coef.est coef.se
Intercept)          -427.6     46.6
CP                    34.2      2.6
LactationMid         -80.9     65.8
LactationLate         98.9     65.8
CP:LactationMid        4.8      3.6
CP:LactationLate      -6.1      3.6

n=48, k=6
residual sd=22.96, R-Squared=0.93
```

(i) Complete the value of **A** to **L** in **Table Q3(d)(ii)** which indicate the slopes and intercepts for each of Model A and Model B.

(12 marks)

Table Q3(c)(ii): Model A & Model B

	Model A		Model B	
Lactation	Slope	Intercept	Slope	Intercept
Early	**A**	**D**	**G**	**J**
Mid	**B**	**E**	**H**	**K**
Late	**C**	**F**	**I**	**L**

(ii) Predict the value of UUN for Model B in the late lactation stage given that the diet CP is 30%.

(2 marks)

Q4 Early result of general survey election process in the United States for 2015 could be seen in **Table Q4(i)** and the summary of model as in **Table Q4(ii)**. Suppose X and Y denote the Race and Party respectively, then given 95th percentiles of χ^2 distribution with 1, 2, 3, 4, 5, 6 degrees of freedom are 3.841, 5.99, 7.81, 9.49, 11.07 and 12.59 respectively.

Table Q4(i): General Survey Election of United States in 2015

	Party Identification		
Race	Democrat	Independent	Republic
White	341	105	405
Black	103	15	11

Table Q4(ii): Summary Model

```
> Racew <- c(1,1,1,0,0,0) #white=1; black=0
> PartyD <- c(1,0,0,1,0,0) #democrat=1;others=0
> PartyI <- c(0,1,0,0,1,0) #independent=1;others=0
> Count <- c(341,105,405,103,15,11)
> RacewPartyD <- Racew*PartyD
> RacewPartyI <- Racew*PartyI
> model <- glm(Count ~ Racew+ PartyD + RacewPartyD +
RacewPartyI,family=poisson(link="log"))
> summary(model)
Call:
glm(formula = Count ~ Racew + PartyD + RacewPartyD + RacewPartyI,
family = poisson(link = "log"))

Deviance Residuals:
      1       2       3       4       5       6
 0.0000  0.0000  0.0000  0.0000  0.5413 -0.5699

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.5649     0.1961  13.079   <2e-16 ***
Racew         3.4389     0.2023  16.998   <2e-16 ***
PartyD        2.0698     0.2195   9.430   <2e-16 ***
RacewPartyD  -2.2418     0.2315  -9.686   <2e-16 ***
RacewPartyI  -1.3499     0.1095 -12.327   <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 918.81894  on 5  degrees of freedom
Residual deviance: 0.61784  on 1  degrees of freedom
AIC: 47.906

Number of Fisher Scoring iterations: 4
```

(a) Define the log-linear analysis.

(3 marks)

(b) Explain **THREE (3)** characteristics of log-linear model.

(6 marks)

(c) Distinguish **TWO (2)** major difference between logistics models and log-linear models.

(4 marks)

(d) Calculate all the expected cell counts of log-linear model.

(12 marks)

- END OF QUESTIONS -