Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER I
SESSION 2014/2015

COURSE NAME	:	CATEGORICAL DATA ANALYSIS
COURSE CODE	:	BWB 31703
PROGRAMME	:	3 BWQ
EXAMINATION DATE	:	DECEMBER 2014/JANUARY 2015
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ANY FOUR (4)** QUESTIONS

THIS QUESTION PAPER CONSISTS OF **TWELVE (12)** PAGES

Q1 (a) Fill in the **blanks (A to J)** in the following questions separately.

(i) A more formal way of analysing the kinds of two-way tables would be to use ------- (**A**). In a log-linear model, we try to model the ------- (**B**) of each cell directly. The simplest log-linear model is the ------- (**C**).

(3 marks)

(ii) It is unlikely that the null model will give a very good fit in most cases, because it does not even take account of the different ------- (**D**) of the variables. The next model would be the ------- (**E**) which accounts for marginal distributions, but assumes ------- (**F**).

(3 marks)

(iii) Log-linear models are a form of ------- (**G**). The ------- (**H**) is the log because this extends counts over the full number line. We use the ------- (**I**) because we are dealing with the accumulation of events.

(3 marks)

(iv) To fit a log-linear model one uses the ------- (**J**) as the dependent variable and the row and column variables are entered using the traditional dummy coding. This is entered into a generalised linear model with a log link and a poisson error distribution.

(1 mark)

(b) The following Table **Q1 (b)(i)** was taken from the 1991 General Social Survey and the `R output` (Table **Q1(b)(ii)**) for given data.

Table Q1(b)(i) : General Social Survey

	Party Identification			
Race	Democrat	Independent	Republican	Total
White	341	105	405	851
Black	103	15	11	129
Total	444	120	416	980

Table Q1(b)(ii) : R-Output

```
> 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
> fit <- glm(Count ~ Racew+ PartyD + RacewPartyD +
RacewPartyI,family=poisson(link="log"))
```

---continue next page---

```
> summary(fit)
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
```

Let X and Y denote the race and party respectively. The 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. Based on the `R output` :

(i) Write down the loglinear regression model. (2 marks)

(ii) Compute all the estimated cell accounts. (12 marks)

(iii) State if the loglinear model fit the data well. (1 mark)

Q2 (a) State either the following questions **True** or **False**:

(i) For General Social Survey Data on Y=political ideology (categories liberal, moderate, conservative), X_1=gender (1=female, 0=male) and X_2 = political party (1=Democrat, 0=Republican), the Maximum Likelihood (ML) fit of the cumulative logit model is

$$\text{logit}\left[\hat{P}(Y \le j)\right] = \hat{\alpha}_j + 0.12x_1 + 0.96x_2$$

For each gender, according to this model fit the estimated odds that a Democrat's response is liberal rather than moderate or conservative and the estimated odds that a Democrat's response is

liberal or moderate rather than conservative, is $e^{0.96} = 2.6$ times the corresponding estimate odds for a Republican's response. This odds ratio estimate indicates that in this sample Democrats tended to be more liberal than Republicans.

(1 mark)

(ii) Subject suffering from mental depression is measured after 1 week of treatment, 2 weeks of treatment and 4 weeks of treatment in terms of a (normal, abnormal) response outcome. Covariates are severity of condition at original diagnosis (1=severe, 0=mild) and treatment used (1=new, 0=standard). Since each subject contributes three observations to the analysis, we can use the Generalised Estimating Equations (GEE) method to fit the model.

(1 mark)

(iii) A difference between logit and loglinear models is that the logit model is a generalised linear model assuming a binomial random component whereas the loglinear model is a generalised linear model assuming a Poisson random component. If both are fitted to a contingency table having 50 cells, the logit model treats the cell counts as 25 binomial observations whereas the loglinear model treats the cell counts as 50 Poisson observations.

(1 mark)

(iv) The cumulative logit model assumes that the response variable Y is ordinal; it should not be used with nominal variables. By contrast, the baseline-category logit model treats Y as nominal. It can be used with ordinal Y, but it then ignores the ordering information.

(1 mark)

(v) The cumulative logit model for J response categories corresponds to a logistic regression model holding for each of the $J - 1$ cumulative probabilities, such that the curves for each cumulative probability have exactly the same shape (e.g : the same β parameter); that is they increase or decrease at the same rate, so one can use $\hat{\beta}$ to describe effects that apply to all $J - 1$ of the cumulative probabilities.

(1 mark)

(vi) For a sample of retired subjects in Italy, a contingency table is used to relate X=cholesterol (8 ordered levels) to Y=whether the subject has symptoms of heart disease (yes=1, no=0). For the linear logit model $\text{logit}[P(Y=1)] = \alpha + \beta x$ fitted to the 8 binomials in the 8 x 2 contingency table by assigning scores to the 8 cholesterol levels,

the deviance statistic equals 6.0. Thus, this model provides a poor fit to the data.

(1 mark)

(b) 1319 schoolchildren were questioned on the prevalence of symptoms of severe cold at the age of 12 and again at the age of 14 years. At age 12, 356 (27%) children were reported to have severe colds in the past 12 months compared to 468 (35.5%) at age 14 as in the Table **Q2(b)(i)**. By using the `R output` given in the Table **Q2(b)(ii)**, test if there is significant increase of the prevalence of severe cold.

Table Q2(b)(i) : Prevalence of Symptoms Severe Cold

Severe cold at age 12	Severe colds at age 14		Total
	Yes	No	
Yes	212	144	356
No	256	707	963
Total	468	851	1319

Table Q2(b)(ii) : R-Output

```
> data <- matrix(c(212,256,144,707),
+ nrow=2,dimnames = list("Severe colds at age 12"=c("Yes","No"),
+ "Severe colds at age 14"=c("Yes","No")))
> data
                    Severe colds at age 14
Severe colds at age 12 Yes  No
                   Yes      212 144
                   No       256 707
> mcnemar.test(data)

        McNemar's Chi-squared test with continuity correction

data:  data
McNemar's chi-squared = 30.8025, df = 1, p-value = 2.857e-08
```

(6 marks)

(c) A model fit predicting preference fo President (Democrat, Republican, Independent) using x=annual income (in \$10,000 dollars) is $\log\left(\hat{\pi}_D/\hat{\pi}_I\right)=3.3-0.2x$ and $\log\left(\hat{\pi}_R/\hat{\pi}_I\right)=1.0+0.3x$

(i) State the prediction equation for $\log\left(\hat{\pi}_R/\hat{\pi}_D\right)$. Interpret the slope.

(3 marks)

(ii) Find the range of x for which $\hat{\pi}_R > \hat{\pi}_D$.

(2 marks)

(iii) State the prediction equation for $\hat{\pi}_I$.

(2 marks)

(d) The `teenagers` dataset divided the students' responses into `Rural`, `Suburban` and `Urban` playing areas. A two-way table for student criteria and school area appears in the Table **Q2(d)(i).** By using the `R output` given in the Table **Q2(d)(ii)**, test if there are any association between the type of playing areas and the students' choice of good skills, sports ability or popularity as most important.

Table Q2(d)(i) : Student's Playing Area

	Playing Area		
Criteria	Rural	Suburban	Urban
Skills	57	87	24
Sports	50	42	6
Popularity	42	22	5

Table Q2(d)(ii) : R-Output

```
> rural <- c(57,50,42)
> suburban <- c(87,42,22)
> urban <- c(24,6,5)
> teenagers <- data.frame(rural,suburban,urban)
> teenagers
rural suburban urban
1    57       87    24
2    50       42     6
3    42       22     5
> chisq.test(teenagers)
     Pearson's Chi-squared test
data:  goal
X-squared = 18.5642, df = 4, p-value = 0.000957
```

(6 marks)

Q3 (a) State either the following questions **True** or **False**:

(i) A simple linear regression model with explanatory variable x and outcome variable y, we have these summary statistics for sample means and standard deviations; $\bar{x} = 10$, $s_x = 3$, $\bar{y} = 20$ and $s_y = 5$. For a new data point with x = 13, it is possible that the predicted value $\hat{y} = 26$.

(1 mark)

(ii) A standard multiple regression model with quantitative predictors, x_1 and x_2, a factor predictor T with four levels, an interaction

between x_1 and T, and an intercept has for its model coefficients an 11 x 1 vector β.

(1 mark)

(iii) In a standard multiple regression model, if a plot of residuals versus fitted values shows a fan-shaped pattern with residuals becoming more spread out as fitted values increase, a log transformation of the response variable may result in data more consistent with model assumptions.

(1 mark)

(iv) If the outcome variable is quantitative and all explanatory variables take values 0 or 1, a logistics regression model is most appropriate.

(1 mark)

(v) In a greenhouse experiment with several predictors, the response variable is the number of seeds that germinate out of 60 planted with each treatment combination. A Poisson regression model is most appropriate for this data.

(1 mark)

(vi) In a greenhouse experiment with several predictors, the response variable is the number of seeds produced for each plant with a sample size of 60 plants. A Poisson regression model is most appropriate for this data.

(1 mark)

(vii) The same data is fit with two models using exactly the same predictors. The first model uses standard logistic regression (with `glm(…, family=binomial)`) while the second model accounts for overdispersion (with `glm (…, family=quasibinomial)`). The estimated coefficients for the predictors in the two models will be identical.

(1 mark)

(viii) When there is a single categorical response variable, logistic models are more appropriate than loglinear models.

(1 mark)

(ix) When you want to model the association and interaction structure among several categorical response variables, logistic models are more appropriate than loglinear models.

(1 mark)

(x) A difference between logistic and loglinear models is that the logistic model is a GLM assuming a binomial random component

whereas the loglinear model is a GLM assuming a Poisson random component. If both are fitted to a contingency table having 50 cells with a binary response, the logistic model treats the cell counts as 25 binomial observations whereas the loglinear model treats the cell counts as 50 Poisson observations.

(1 mark)

(b) A group of dairy scientists is interested in studying the effect that the percentage of crude protein (CP) in the diet and the stage of lactation (Early, Mid and Late) have on the amount of nitrogen excreted in urine. Dairy cows typically produce milk for at least 300 days after calving. The amount milk they produce per day changes over time. In the study, cows in the early lactation stage averaged 123 days in milk (DIM), mid stage average 175 DIM and the late stage averaged 221 DIM. There are four treatment diets with CP percentage taking values 15%, 17%, 19% and 21% of diet dry matter. CP was treating as quantitative variable in the model. The outcome variable, urea urine nitrogen (UUN) is measured in grams per day. The scientists conduct an experiment in which 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. The two models, the first with CP and Lactation as inputs, the second with an interaction term included where the results were summarised in the Table **Q3(b)**.

Table Q3(b) : R-Output

```
Model 1
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 2
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) Each model describes the relationship between UUN and CP with a line for each lactation group. Fill in the Table **Q3(b)(i)** indicating the slope and intercepts for each model for each group.

Table Q3(b)(i) : Model 1 & Model 2

	Model 1		Model 2	
Lactation	Slope	Intercept	Slope	Intercept
Early	A	B	C	D
Mid	E	F	G	H
Late	I	J	K	L

(12 marks)

(ii) For each model, predict the UUN for a cow in the mid lactation stage with a diet CP of 20%.

(2 marks)

(iii) For Model 1, provide an interpretation of slope.

(1 mark)

Q4 (a) A study by a group of biologist interested in adaptive evolution studies the charismatic threespine stickleback. These fish colonised lakes in coastal British Columbia about 10,000 years ago when glaciers receded. In each lake, the fish appear as *ecomorph pairs* where each of two types of fish nests, feeds and breeds in a different ecological niche. The *Benthic type* feeds near the edges of lakes whereas the *Limnetic type* feeds in open water. There are independently evolved species pairs in each of three lakes. As part of a study on sexual selection, the biologist conducted 753 mating trials with a single male and single female fish. For each mating trial, it is recorded whether or not the fish spawn or not within a 30 minute time period. In each trial, the fish can be from the same lake or different lakes, a factor we will call `Lakes` with levels `Same` and `Different`. Also, for each trial, the pair of potential mates are either both Benthic (BB), both Limnetic (LL), or one of each (BL) which we record in a factor `Types`. Consider two logistic regression models, one with `Lakes` and `Types` as inputs and one with these inputs plus an interaction. The estimated coefficients from these models are displayed in the Table **Q4(a)(i)** below.

Table Q4(a)(i) : R-Output

Model 1

	coef.est	coef.se
(Intercept)	-0.36	0.15
LakesSame	-0.02	0.16
TypesBL	-1.15	0.21
TypesLL	0.13	0.18

Model 2

	coef.est	coef.se
(Intercept)	-0.38	0.17
LakesSame	0.02	0.26
TypesBL	-0.94	0.27
TypesLL	0.00	0.26
LakesSame:TypesBL	-0.51	0.43
LakesSame:TypesLL	0.22	0.37

(i) Briefly explain why logistic regression is an appropriate model for this data.

(1 mark)

(ii) Calculate the probability of spawning for various mating trial classes under each model indicated in the Table **Q4(a)(ii)** below using the inverse logit function : $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$.

Table Q4(a)(ii) : Probability Of Spawning For Various Mating Trial

Types *Lakes*	Class 1 LL *Same*	Class 2 LL *Different*	Class 3 BL *Same*
Model 1	A	B	C
Model 2	D	E	F

(6 marks)

(iii) Using Model 1, compare the spawning probabilities between Class 1 and Class 2, also between Class 1 and Class 3. Does it appear that the types of fish or the lakes of origin is more important predictor of spawning probability ?

(2 marks)

(b) In the Austria, the estimated annual probability that a woman over the age of 35 dies of lung cancer equals 0.001304 for current smokers and 0.000121 for non-smokers.

(i) Calculate and interpret the difference of proportions and the relative risk.

(6 marks)

(ii) Calculate and interpret the odds ratio. Explain why the relative risk and the odds ratio take similar values.

(4 marks)

(c) The data from a case-control study in Thailand Health Office stratified by the gender of the people studied. The two different tables, Table **Q5(c)(i)** and Table **Q5(c)(ii)** given the information on the relationship between exposure (either to a certain drug or smoking cigarettes) and having a particular disease for women and men. Refer to the `R output` given in the Table **Q5(c)(iii)**, explain the function Cochran-Mantel-Haenszel Chi-squared in this study and make an appropriate conclusions based on the output.

Table Q5(c)(i) : Women

Data for Women		
Goals	Disease	Control
Exposure	4	5
Unexpected	5	103

Table Q5(c)(ii) : Men

Data for Men		
Goals	Disease	Control
Exposure	10	3
Unexpected	5	43

Table Q5(c)(iii) : R-Output

```
> library(lawstat)
> mywomen <- matrix(c(4,5,5,103),nrow=2,byrow=TRUE)
> colnames(mywomen) <- c("Disease","Control")
> rownames(mywomen) <- c("Exposure","Unexposed")
> print(mywomen)
           Disease Control
Exposure         4       5
Unexposed        5     103
> mymen <- matrix(c(10,3,5,43),nrow=2,byrow=TRUE)
> colnames(mymen) <- c("Disease","Control")
> rownames(mymen) <- c("Exposure","Unexposed")
> print(mymen)
           Disease Control
Exposure        10       3
Unexposed        5      43
> myarray <- array(c(mywomen,mymen),dim=c(2,2,2))
> cmh.test(myarray)
        Cochran-Mantel-Haenszel Chi-square Test
data:  myarray
CMH statistic = 40.512, df = 1.000, p-value = 0.000, MH Estimate =
23.001,Pooled Odd Ratio = 25.550, Odd Ratio of level 1 = 16.480, Odd
Ratio of level 2 = 28.667
```

(6 marks)

Q5 (a) State the appropriate scale of measurements for the following variables.

(i) Infant condition (Good, Fair, Critical).

(1 mark)

(ii) Political party affiliation (Republican, Democrat, unaffiliated).

(1 mark)

(iii) Highest degree achieved (Certificate, Diploma, Bachelors, Masters, Doctorate).

(1 mark)

(iv) Favorite drink (Milo, Juice, Milk, Soft drink).

(1 mark)

(v) School location (Urban, Suburban, Rural).

(1 mark)

(b) The analysis of ------- **(i)** is concerned with more than one variable, ------- **(ii)** are employed. These tables provide a foundation for statistical inference, where ------- **(iii)** question the relationship between the variables on the basis of the data observed. The ------- **(iv)** provides a method for testing the association between the row and column variables in a two-way table. The ------- **(v)** assumes that there is no association between the

variables (one variable does not vary according to the other variable), while the ------- **(vi)** claims that some association does exist.

(6 marks)

(c) The two-way table for children intelligence level, reading ability and thinking ability by grade (1 to 3) given in the Table **Q5(c)**. Calculate the expected values for each cell.

Table Q5(c) : Children Intelligence Level (Grade)

	Grade		
Goals	1	2	3
Intelligence	38	47	64
Reading	23	56	38
Thinking	18	29	23

(9 marks)

(d) (i) For the generalised linear models (GLM), logistic regressions and also Poisson regressions, they all have variants that include an overdispersion parameter. Why we not suggested to model overdispersion in a standard multiple regression models with a normally distributed response?

(2 marks)

(ii) Given that the fitted values and goodness-of-fit measures do not change when transforming predictors by centering (e.g:subtracting the mean), what is the benefit of making such a transformation?

(1 mark)

(iii) After fitting a multiple regression models, briefly explain how we could detect that the linearity assumption of the model below is violated.

$$E[y_i] = X_i\beta = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

(1 mark)

(iv) Explain when logistics regression is an appropriate in modelling the data.

(1 mark)

- END OF QUESTION -