Indian Statistical Institute
Chennai Centre
M-Stat (NB Stream)Semester II : 2016-18
Categorical Data Analysis

Mid-Semester Examination

February 22, 2017] [9:30 am - 12:30 noon

- The paper is of **100** marks.
- Answer **All** the questions.

1. (a) The table below presents cross-classified data on race (white, black and others) of randomly selected individuals from four counties (C1, C2, C3, C4).

	Counties			
Race	C1	C2	C3	C4
White	95	72	32	8
Black	66	129	116	13
Others	31	101	133	82

Compute the Goodman-Kruskal **symmetric** measure of association and interpret it.

(b) The table below displays the diagnosis of carcinoma of 106 patients based on two different pathological evaluations.

	Evaluation B	
Evaluation A	Malignant	benign
Malignant	38	21
Benign	33	14

Formally justify whether there is any agreement between the two evaluations. [20]

2. (a) Briefly state what is Simpson's paradox.

(b) Analysis of a $2 \times 2 \times 2$ table for Gender×Medium×Performance gave the following

```
> names(dat)
[1] "Gender"      "Medium"      "Performance"
> oddsratio(xtabs(~Gender+Medium+Performance,dat))
      Bad      Good
-4.576833 -3.734559
> oddsratio(xtabs(~Gender+Performance+Medium,dat))
 Bengali  English
1.449116 2.335157
> oddsratio(xtabs(~Medium+Performance+Gender,dat))
    Female       Male
0.06252036 0.84255142
> oddsratio(xtabs(~Gender+Medium,dat))
[1] -4.465384
> oddsratio(xtabs(~Gender+Performance,dat))
[1] 1.704546
> oddsratio(xtabs(~Medium+Performance,dat))
[1] -0.3689405
```

Does this output show Simpson's paradox? Justify your answer. [10]

3. (a) Briefly explain the Latent Variable Approach to model ordinal categorical response variable with possible covariates. Suppose the underlying distribution of latent variables is exponential.

   i. Introduce the ordinal response variable Y.

   ii. Suppose a cumulative logit proportional odds model is to be fitted to the data with a single covariate. Show how the cumulative probabilities of the ordinal response variables Y_i will be linked to the linear predictor $\alpha + \beta x_i$ using the information on the underlying latent distribution.

   iii. Write the appropriate expressions for the cumulative logits using (ii) above.

   iv. Derive the log-likelihood equations to get maximum likelihood estimates of the parameters of interest.

   (b) Can you detect over dispersion when the response variables are bernouli? Justify your answer.

   (c) The data on a production process records classification of 651 units in 5 ordinal categories, very bad to very good, represented by $1, 2, \cdots, 5$. There are two covariates: i)Quality of material, recorded in three groups A, B and C and ii) Weight of unit, a continuous variable. A proportional odds logistic regression model was fitted to the data and the analysis is presented below.

```
Response= Level ~ Quality + weight

Coefficients:
            Value     Std. Error    t value
Quality B   -3.352     0.4287       -7.819
Quality C   -1.710     0.3715       -4.603
Weight       0.616     0.2606        2.363

Intercepts:
         Value  Std. Error t value
1|2     -0.9078  0.2833    -3.2017
2|3      0.0443  0.2646     0.1673
3|4      2.1763   0.7671      2.8370
4|5      4.2716   0.7922      5.3924

Residual Deviance: 711.3479
AIC: 736.0657
```

   i. Interpret the outcome.

   ii. State the estimated model.

   iii. What is the estimated probability that a unit of weight 5 gms and quality B will be bad?

   iv. Can this output be used to estimate odds of baseline category logit model? Justify your answer. If yes, illustrate by estimating one baseline category odds of your choice.

   (d) A data on 121 patients recorded number of times each patient experienced total black-out in 1 month, apart from their age, gender, and whether they are on a certain medication or not. The number of black-outs are between 0 and 6 with an average of 2.6. It is verified that the Poisson regression model with selected covariates and the choice of link function fits the data well.

   If the observed variance is 4, state with justification what will be your conclusion and what will be your next step of analysis. [40]

4. (a) Suppose $Y_1, \cdots, Y_n$ be independent random variables with covariates $x_1, \cdots, x_n$ and $\mathrm{E}(Y_i) = \beta_1 + \ln(\beta_2 + \beta_3 x_i)$ where $Y_i,\ i = 1, \cdots, n$ follow Normal distribution with common variance σ^2. Can this set up be modeled as a generalized linear model? Justify your answer.

(b) The table below gives observations on independent random variables $Y_i,\ i = 1, \cdots, 5$ which follow exponential distribution with mean $\mu_i,\ i = 1, \cdots, 5$ respectively. Suppose x is the covariate and $\mu_i = \exp(\beta_1 + \beta_2 x_i),\ i = 1, \cdots, 5$

Observation no	1	2	3	4	5
y_i	2.5	5.9	44.6	52.8	262.2
x_i	0	1	2	3	4

i. Does the data support choice of the link function? Justify.

ii. Write the log-likelihood function and derive likelihood equations to get maximum likelihood estimates of β_1 and β_2.

iii. The likelihood equations need to be solved iteratively. Suggest a method which can easily give sensible initial values of β_1 and β_2 and find the values.

iv. The log-likelihood function at the MLEs is given by $l(\hat{\beta}_1, \hat{\beta}_2) = -21.68$. Can the hypotheses $H_o : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$, be tested? Justify your answer. If possible, test the hypotheses. [30]
