UPPSALA UNIVERSITY
Department of Mathematics

Shaobo Jin

Exam in Mathematical Statistics
Analysis of Categorical Data 1MS370
Categorical Data Analysis 2ST121
2024-01-09

Time: 8.00-13.00. For **1MS370**, limits for the grades 3, 4, 5 are 18, 25 and 32 points, respectively. For **2ST121**, the limits for grades G and VG are 20 and 32, respectively. The solutions should be well motivated.

Permitted aids: A sheet of your own notes (A4 paper, two-sided). Pocket calculator. Dictionary. No electronic device with internet connection is allowed.

1. (4p) Consider an $I \times J \times K$ contingency table.
   (a) (2p) Let $I = 3$ and $J = 3$. Suppose that all local odds ratio in a partial table are 1. Can we claim all pairs of odds ratio in that partial table are 1?
   (b) (2p) Let $I = 2$ and $J = 2$. Suppose that the contingency table has homogeneous association. Can we claim the marginal local odds ratio is the same as the conditional odds ratio?
2. (8p) We have a data set about the survival status of the ship Titanic. The variables that we have are survival (S, 2 levels), class of cabin (C, 2 levels), age group (A, 2 levels). The variable S is coded as "Alive" and "Dead". Let $\pi_{ik} = P(S = \text{Alive} \mid C = i, A = k)$.
   (a) (2p) A statistician wants to fit the model such that S and C are conditionally independent given A, but not necessarily that S and C are marginally independent. Write down a model that satisfy such hypothesis.
   (b) (1p) The statistician wants to test independence of S and C. If he finds out that S and C are conditionally independent at any level of A, can he claim that S is independent of C?
   (c) (1p) The statistician fitted a logit model in R. The results are presented below.

```
##
## Call:
## glm(formula = cbind(Alive, Dead) ~ C + A, family = binomial(),
##     data = Data)
##
## Deviance Residuals:
##       1        3        7        9
##  1.4758  -1.4285  -0.5590   0.8948
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4488     0.3364  -1.334 0.182139
## C2            1.3381     0.2707   4.943 7.68e-07 ***
## A2            1.2670     0.3414   3.712 0.000206 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.6502  on 3  degrees of freedom
## Residual deviance:  5.3317  on 1  degrees of freedom
## AIC: 28.487
##
## Number of Fisher Scoring iterations: 4
```

Which model has been fitted?

(d) (2p) Can we claim that we have homogeneous SC association from the above output? You may need the following quantiles. The 95% quantiles of a chi-square distribution with 1, 2, 3 degrees of freedom are 3.84, 5.99, 7.81, respectively.

(e) (2p) Another statistician played around with more link functions in R. The results are presented below.

```
##
## Call:  glm(formula = cbind(Alive, Dead) ~ C + A, family = binomial("probit"),
##     data = Data)
##
## Coefficients:
## (Intercept)           C2           A2
##     -0.2607       0.7823       0.7636
##
## Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
## Null Deviance:     43.65
## Residual Deviance: 4.477  AIC: 27.63
##
## Call:  glm(formula = cbind(Alive, Dead) ~ C + A, family = binomial("cloglog"),
##     data = Data)
##
## Coefficients:
## (Intercept)           C2           A2
##     -0.6308       0.7204       0.7806
##
## Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
## Null Deviance:     43.65
## Residual Deviance: 2.725  AIC: 25.88
##
```

```
## Call:  glm(formula = cbind(Alive, Dead) ~ C + A, family = binomial("cauchit"),
##     data = Data)
##
## Coefficients:
## (Intercept)           C2           A2
##     -0.3451       1.5104       1.0570
##
## Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
## Null Deviance:     43.65
## Residual Deviance: 11.76  AIC: 34.92
```

Which link function do you prefer? What is the distribution assumption behind the link function of your choice?

3. (3p) The effect of a new curriculum is going to be studied. 30 schools are included in the study. Within each school, two students are assigned to the old curriculum, and two students are assigned to the new curriculum. The data set includes the following variables: which school the student is from (S), whether the student has the new or the old curriculum (C), and student's satisfaction after 1 year of study (R). The result R is coded as "Y" (satisfied) and "N" (not satisfied). Two statisticians try to analyze the data set by estimating the odds ratio between C and R.

   (a) (2p) The result of the first statistician is

```
##
## Call:  glm(formula = cbind(Y, N) ~ C + S, family = binomial)
##
## Coefficients:
## (Intercept)            C           S2           S3           S4
##  -7.715e-01    1.543e+00    1.249e+00   -1.249e+00   -5.999e-15
##          S5           S6           S7           S8           S9
##  -1.249e+00   -5.746e-15   -1.249e+00   -1.249e+00    2.003e+01
##         S10          S11          S12          S13          S14
##  -2.003e+01   -2.003e+01   -5.952e-15   -1.249e+00   -1.249e+00
##         S15          S16          S17          S18          S19
##  -6.281e-15    1.249e+00   -2.003e+01   -6.060e-15   -1.249e+00
##         S20          S21          S22          S23          S24
##  -1.249e+00   -2.003e+01    1.249e+00   -6.280e-15   -1.249e+00
##         S25          S26          S27          S28          S29
##  -1.249e+00   -6.201e-15   -6.680e-15   -6.381e-15   -2.003e+01
##         S30
##  -5.403e-15
##
## Degrees of Freedom: 59 Total (i.e. Null);  29 Residual
## Null Deviance:     89.46
## Residual Deviance: 38.97  AIC: 135.6
```

What is the estimated odds ratio between C and R? Is your estimate an conditional odds ratio or a marginal odds ratio?

(b) (1p) The result of the second statistician is

```
##
##  Mantel-Haenszel chi-squared test with continuity correction
##
## data:  Data
## Mantel-Haenszel X-squared = 7.0843, df = 1, p-value = 0.007776
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.370549 7.207004
## sample estimates:
## common odds ratio
##          3.142857
```

Which statistician's analysis is more reasonable, the first or the second statistician? Motivate your answer.

4. (12p) Consider a contingency table with the following variables: diet (D), health (H), gender (G), income (I), age (A). Consider the loglinear model (DHG, DI, HA, IA).

   (a) (2p) Write down the model equation (i.e., the expression of $\log \mu$).

   (b) (2p) Draw its conditional independence graph.

   (c) (2p) Is G conditionally independent of I given D and H?

   (d) (2p) Does the model have homogeneous DH association?

   (e) (2p) If we collapse A, will it change DH association?

   (f) (2p) Is there any other loglinear model with the same conditional independence graph?

5. (6p) Consider again the Titanic data in Task 1. An additional variable Gender (G) is also collected.

   (a) (1p) A statistician fitted a loglinear model to the data set. The following outputs are obtained.

```
##
## Call:  glm(formula = count ~ S + G + C + A + S:A + S:C + S:G + C:A +
##     C:G, family = poisson)
##
## Coefficients:
## (Intercept)           S2           G2           C2           A2
##      2.5068      -1.0402      -2.3686       0.8188       2.2757
##       S2:A2        S2:C2        S2:G2        C2:A2        G2:C2
##      0.2829      -1.0371       3.2294      -0.6632       0.3699
```

```
##
## Degrees of Freedom: 15 Total (i.e. Null);  6 Residual
## Null Deviance:     867.6
## Residual Deviance: 74.95  AIC: 170.6
```

Does the model fit the data well?

(b) (1p) Find the conditional SC odds ratio given $A = 1$, $G = 2$.

(c) (2p) It is known that the number of tickets for each class is fixed, that is, n_{1+++} and n_{2+++} are fixed by design. Will the model fitted above satisfy $\hat{\mu}_{1+++} = n_{1+++}$ and $\hat{\mu}_{2+++} = n_{2+++}$?

(d) (2p) Suppose that we want to build a logit model for $P(S = 1 \mid C = i, A = k, G = l)$ such as

$$\log\left(\frac{P(S=1 \mid C=i, A=k, G=l)}{1 - P(S=1 \mid C=i, A=k, G=l)}\right) \quad = \quad \alpha + \beta_i^C + \beta_k^A + \beta_l^G.$$

Can you obtain its estimated parameters from the above loglinear model? If so, present the estimates. Otherwise, state the reason.

6. (4p) Consider the loglinear model for independent $Y_{ik} \sim \text{Poisson}(\mu_{ik})$, where

$$\log \mu_{ik} \quad = \quad \lambda_k + \beta x_i, \quad i = 1, 2, ..., n, \ k = 1, 2, ..., K,$$

where $x_i = 1$ or 0. Suppose that only β is the focus parameter and all $\{\lambda_k\}$ are nuisance parameters. Derive the conditional likelihood of β.

7. (3p) Consider the loglinear model (XZ, YZ). Find the MLE of μ_{ijk}.