UPPSALA UNIVERSITY
Department of Mathematics
Shaobo Jin

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

Time: 8.00-13.00. Limits for the credits 3, 4, 5 are 18, 25 and 32 points, 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?
   Solution: Yes, since any odds ratio is a series of products of local odds ratios.

   (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?
   Solution: No, unless the collapsibility condition is fulfilled.

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). Let $\pi_{ik} = P\left(S = 1 \mid C = i, A = k\right)$.

   (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.
   Solution: The model with conditional independence is

$$\log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = \alpha + \beta_k^A.$$

   This is because the log of the conditional odds ratio is

$$\begin{aligned}\log\left(\frac{\pi_{ik}/\left(1-\pi_{ik}\right)}{\pi_{i+1,k}/\left(1-\pi_{i+1,k}\right)}\right) &= \left(\alpha + \beta_k^A\right) - \left(\alpha + \beta_k^A\right) \\ &= 0.\end{aligned}$$

   We cannot have

$$\log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = \alpha,$$

   since it means that S and C are marginally independent.

   (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?
   Solution: No, it is not enough to claim independence of S and C. We can only claim that S and C are conditionally independent given A. But it does not imply marginal independence.

   (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?
Solution: The fitted model is

$$\log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) \quad = \quad \alpha + \beta_i^C + \beta_k^A.$$

(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.
Solution: Note that the residual deviance is a lot larger than the degrees of freedom 1. Hence, the residual deviance is larger than the 95% quantile of χ_1^2. Hence, the model does not fit the data well. The model that we fitted imply homogeneous association. Hence, we cannot claim we have homogeneous SC association.

(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?
Solution: By AIC, we will choose the cloglog link. The cloglog model fits the data well. The distribution assumption behind the loglog link is the Gumbel distribution. If a random variable X is Gumbel, then the cdf of $-X$ yields the cloglog link.

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, data = GLM)
##
## Coefficients:
## (Intercept)           C1           S2           S3           S4           S5
##  -7.715e-01    1.543e+00    1.249e+00   -1.249e+00   -5.999e-15   -1.249e+00
##          S6           S7           S8           S9          S10          S11
##  -5.746e-15   -1.249e+00   -1.249e+00    2.003e+01   -2.003e+01   -2.003e+01
##         S12          S13          S14          S15          S16          S17
##  -5.952e-15   -1.249e+00   -1.249e+00   -6.281e-15    1.249e+00   -2.003e+01
##         S18          S19          S20          S21          S22          S23
##  -6.060e-15   -1.249e+00   -1.249e+00   -2.003e+01    1.249e+00   -6.280e-15
##         S24          S25          S26          S27          S28          S29
##  -1.249e+00   -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?
Solution: The conditional odds ratio is $\exp(1.543)$.

(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
```

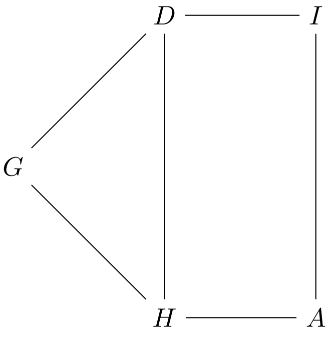
Which analysis is more reasonable, the first or the second statistician? Motivate your answer.
Solution: We have a sparse data. The number of observations in each partial table is fixed to 4. Hence, $n = 4K$. If we increase n, then K also increases, as well as the parameter λ_j^S. Hence, the approach by the second statistician is more reasonable.

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 $\log \mu$).
Solution: The model equation is

$$\log \mu_{ijkls} \;=\; \lambda + \lambda_i^D + \lambda_j^H + \lambda_k^G + \lambda_l^I + \lambda_s^A + \lambda_{ij}^{DH} + \lambda_{ik}^{DG} + \lambda_{jk}^{HG} + \lambda_{il}^{DI} + \lambda_{js}^{HA} + \lambda_{is}^{IA} + \lambda_{ijk}^{DHG}.$$

(b) (2p) Draw its conditional independence graph.
Solution: The CIG of this model is



(c) (2p) Is G conditionally independent of I given D and H?
Solution: Yes, since $\{D, H\}$ separates G and I.

(d) (2p) Does the model have homogeneous DH association?
Solution: Note that

$$\begin{aligned}\log \frac{\mu_{ijkls}\mu_{i+1,j+1,kls}}{\mu_{i+1,jkls}\mu_{i,j+1,kls}} &= \lambda_{ij}^{DH} - \lambda_{i+1,j}^{DH} - \lambda_{ij+1}^{DH} + \lambda_{i+1,j+1}^{DH} \\ &\quad + \lambda_{ijk}^{DHG} - \lambda_{i+1,jk}^{DHG} - \lambda_{i,j+1,k}^{DHG} + \lambda_{i+1,j+1,k}^{DHG},\end{aligned}$$

where the DHG interation remains. Hence, it does not have homogeneous DH association.

(e) (2p) If we collapse A, will it change DH association?
Solution: $\{I, H\}$ separates $\{D, G\}$ and A. Hence, $(D, G) \perp A \mid (I, H)$. If we collapse A, the DH association will not change.

(f) (2p) Is there any other loglinear model with the same conditional independence graph?
Solution: Yes, the model (DH, DG, HG, DI, HA, IA) has 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, data = Ship)
##
## Coefficients:
## (Intercept)           S2           G2           C2           A2        S2:A2
##      2.5068      -1.0402      -2.3686       0.8188       2.2757       0.2829
##       S2:C2        S2:G2        C2:A2        G2:C2
##     -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?
Solution: Note that the residual deviance is a lot larger than the degrees of freedom 10. Hence, we can speculate that the residual deviance is larger than the 95% quantile of χ^2_{10}. Hence, the model does not fit the data well.

(b) (1p) Find the conditional SC odds ratio given $A = 1$, $G = 2$.
Solution: The fitted loglinear model is

$$\log \mu_{ijkl} = \lambda + \lambda_i^C + \lambda_j^S + \lambda_k^A + \lambda_l^G + \lambda_{ij}^{CS} + \lambda_{ik}^{CA} + \lambda_{il}^{CG} + \lambda_{jk}^{SA} + \lambda_{jl}^{SG}.$$

The conditional SC log odds ratio given $A = 1$, $G = 2$ is

$$\log\left(\frac{\mu_{1112}\mu_{2212}}{\mu_{1212}\mu_{2112}}\right) = \lambda_{11}^{CS} - \lambda_{12}^{CS} - \lambda_{21}^{CS} + \lambda_{22}^{CS} = -1.0371.$$

So the conditional odds ratio is $\exp(-1.0371)$.

(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 $\hat{\mu}_{2+++} = n_{2+++}$?
Solution: We need to derive the minimal sufficient statistics first. The kernel of the likelihood is

$$f = \exp\left\{\sum_{ijkl} n_{ijkl} \log(\mu_{ijkl})\right\}.$$

If we have a ghost sample n^*_{ijkl}, then

$$\begin{aligned}
\log\left[\frac{f(\boldsymbol{n})}{f(\boldsymbol{n}^*)}\right] &= n\lambda + \sum_i n_{i+++}\lambda^C_i + \sum_j n_{+j++}\lambda^S_j + \sum_k n_{++k+}\lambda^A_k + \sum_l n_{+++l}\lambda^G_l + \sum_{ij} n_{ij++}\lambda^{CS}_{ij} \\
&\quad + \sum_{ik} n_{i+k+}\lambda^{CA}_{ik} + \sum_{il} n_{i++l}\lambda^{CG}_{il} + \sum_{jk} n_{+jk+}\lambda^{SA}_{jk} + \sum_{jl} n_{+j+l}\lambda^{SG}_{jl} \\
&\quad - \left[n^*\lambda + \sum_i n^*_{i+++}\lambda^C_i + \sum_j n^*_{+j++}\lambda^S_j + \sum_k n^*_{++k+}\lambda^A_k + \sum_l n^*_{+++l}\lambda^G_l\right] \\
&\quad - \left[\sum_{ij} n^*_{ij++}\lambda^{CS}_{ij} + \sum_{ik} n^*_{i+k+}\lambda^{CA}_{ik} + \sum_{il} n^*_{i++l}\lambda^{CG}_{il} + \sum_{jk} n^*_{+jk+}\lambda^{SA}_{jk} + \sum_{jl} n^*_{+j+l}\lambda^{SG}_{jl}\right]
\end{aligned}$$

does not depend on the parameters if and only if $n_{ij++} = n^*_{ij++}$, $n_{i+k+} = n^*_{i+k+}$, $n_{i++l} = n^*_{i++l}$, $n_{+jk+} = n^*_{+jk+}$, $n_{+j+l} = n^*_{+j+l}$. Hence, they will be our minimal sufficient statistics. Hence, we will have for example, $\hat{\mu}_{ij++} = n_{ij++}$, which means that we will have $\hat{\mu}_{1+++} = n_{1+++}$ 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

$$\log\left(\frac{P(S = 1 \mid C = i, A = k, G = l)}{1 - P(S = 1 \mid C = i, A = k, G = l)}\right) = \alpha + \beta^C_i + \beta^A_k + \beta^G_l.$$

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

Solution: Note that

$$\begin{aligned}
\log\left(\frac{P(S = 1 \mid i, k, l)}{1 - P(S = 1 \mid i, k, l)}\right) &= \lambda + \lambda^C_i + \lambda^S_1 + \lambda^A_k + \lambda^G_l + \lambda^{CS}_{i1} + \lambda^{CA}_{ik} + \lambda^{CG}_{il} + \lambda^{SA}_{1k} + \lambda^{SG}_{1l} \\
&\quad - \left(\lambda + \lambda^C_i + \lambda^S_2 + \lambda^A_k + \lambda^G_l + \lambda^{CS}_{i2} + \lambda^{CA}_{ik} + \lambda^{CG}_{il} + \lambda^{SA}_{2k} + \lambda^{SG}_{2l}\right) \\
&= \left(\lambda^S_1 - \lambda^S_2\right) + \left(\lambda^{CS}_{i1} - \lambda^{CS}_{i2}\right) + \left(\lambda^{SA}_{1k} - \lambda^{SA}_{2k}\right) + \left(\lambda^{SG}_{1l} - \lambda^{SG}_{2l}\right)
\end{aligned}$$

Hence, we will have $\alpha = 1.0402$, $\beta^C_1 = 0$, $\beta^C_2 = 1.0371$, $\beta^A_1 = 0$, $\beta^A_2 = -0.2829$, $\beta^G_1 = 0$, and $\beta^G_2 = -3.2294$.

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

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

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.

Solution: The joint distribution is

$$\begin{aligned}
f &= \exp\left\{\sum_{k=1}^K \left[\sum_{i=1}^n y_{ik}\log(\mu_{ik}) - \sum_{i=1}^n \mu_{ik} + \sum_{i=1}^n \log(y_{ik}!)\right]\right\} \\
&= \exp\left\{\sum_{k=1}^K \left(\sum_{i=1}^n y_{ik}\right)\lambda_k + \beta\left(\sum_{k=1}^K\sum_{i=1}^n y_{ik}x_i\right) - \sum_{k=1}^K\sum_{i=1}^n \mu_{ik} - \sum_{k=1}^K\sum_{i=1}^n \log(y_{ik}!)\right\}.
\end{aligned}$$

Hence, the sufficient statistics are $\{\sum_{i=1}^n y_{ik}\}$ and $\sum_{k=1}^K\sum_{i=1}^n y_{ik}x_i$.We conditional on the sufficient statistics for $\{\lambda_k\}$. Let

$$S = \left\{(y^*_{11}, \cdots, y^*_{NK});\ \sum_{i=1}^n y^*_{ik} = t_k,\ k = 1, 2, ..., K\right\}.$$

Then,

$$\begin{aligned}
& P\left(Y_{11} = y_{11}, \cdots, Y_{NK} = y_{NK} \mid \sum_{i=1}^{n} y_{ik} = t_k, \forall k\right) \\
= & \frac{P\left(Y_{11} = y_{11}, \cdots, Y_{NK} = y_{NK}, \sum_{i=1}^{n} y_{ik} = t_k, \forall k\right)}{P\left(\sum_{i=1}^{n} y_{ik} = t_0, \forall k\right)} \\
= & \frac{P\left(Y_{11} = y_{11}, \cdots, Y_{NK} = y_{NK}\right)}{P\left(\sum_{i=1}^{n} y_{ik} = t_k, \forall k\right)} \\
= & \frac{\exp\left\{\sum_{k=1}^{K}\left(\sum_{i=1}^{n} y_{ik}\right)\lambda_k + \beta\left(\sum_{k=1}^{K}\sum_{i=1}^{n} y_{ik}x_i\right) - \sum_{k=1}^{K}\sum_{i=1}^{n}\mu_{ik} - \sum_{k=1}^{K}\sum_{i=1}^{n}\log\left(y_{ik}!\right)\right\}}{\sum_S \exp\left\{\sum_{k=1}^{K}\left(\sum_{i=1}^{n} y_{ik}^*\right)\lambda_k + \beta\left(\sum_{k=1}^{K}\sum_{i=1}^{n} y_{ik}^*x_i\right) - \sum_{k=1}^{K}\sum_{i=1}^{n}\mu_{ik} - \sum_{k=1}^{K}\sum_{i=1}^{n}\log\left(y_{ik}!\right)\right\}} \\
= & \frac{\exp\left\{\sum_{k=1}^{K} t_k\lambda_k + \beta\left(\sum_{k=1}^{K}\sum_{i=1}^{n} y_{ik}x_i\right) - \sum_{k=1}^{K}\sum_{i=1}^{n}\mu_{ik} - \sum_{k=1}^{K}\sum_{i=1}^{n}\log\left(y_{ik}!\right)\right\}}{\sum_S \exp\left\{\sum_{k=1}^{K} t_k\lambda_k + \beta\left(\sum_{k=1}^{K}\sum_{i=1}^{n} y_{ik}^*x_i\right) - \sum_{k=1}^{K}\sum_{i=1}^{n}\mu_{ik} - \sum_{k=1}^{K}\sum_{i=1}^{n}\log\left(y_{ik}^*!\right)\right\}} \\
= & \frac{\exp\left\{\beta\left(\sum_{k=1}^{K}\sum_{i=1}^{n} y_{ik}x_i\right) - \sum_{k=1}^{K}\sum_{i=1}^{n}\log\left(y_{ik}!\right)\right\}}{\sum_S \exp\left\{\beta\left(\sum_{k=1}^{K}\sum_{i=1}^{n} y_{ik}^*x_i\right) - \sum_{k=1}^{K}\sum_{i=1}^{n}\log\left(y_{ik}^*!\right)\right\}},
\end{aligned}$$

which is a function of β.

7. (3p) Consider the loglinear model (XZ, YZ). Find the MLE of μ_{ijk}.
Solution: We need to derive the minimal sufficient statistics first. The kernel of the likelihood is

$$f = \exp\left\{\sum_{ijk} n_{ijk}\log\left(\mu_{ijk}\right)\right\}.$$

If we have a ghost sample n_{ijk}^*, then

$$\begin{aligned}
\log\left[\frac{f\left(\boldsymbol{n}\right)}{f\left(\boldsymbol{n}^*\right)}\right] = & \; n\lambda + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z + \sum_{ij} n_{ij+}\lambda_{ij}^{XY} + \sum_{jk} n_{+jk}\lambda_{jk}^{YZ} \\
& - \left[n^*\lambda + \sum_i n_{i++}^*\lambda_i^X + \sum_j n_{+j+}^*\lambda_j^Y + \sum_k n_{++k}^*\lambda_k^Z + \sum_{ij} n_{ij+}^*\lambda_{ij}^{XY} + \sum_{jk} n_{+jk}^*\lambda_{jk}^{YZ}\right]
\end{aligned}$$

does not depend on the parameters if and only if $n_{ij+} = n_{ij+}^*$, $n_{+jk} = n_{+jk}^*$. Hence, we will have $\hat{\mu}_{ik+} = n_{ij+}$ and $\hat{\mu}_{+jk} = n_{+jk}$. Note that the model implies $X \perp Y \mid Z$. Then,

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}$$

which implies that

$$\hat{\mu}_{ijk} = n\frac{\hat{\pi}_{i+k}\hat{\pi}_{+jk}}{\hat{\pi}_{++k}} = \frac{\hat{\mu}_{i+k}\hat{\mu}_{+jk}}{\hat{\mu}_{++k}} = \frac{n_{i+k}n_{+jk}}{n_{++k}}.$$