STOCKHOLM UNIVERSITY
DEPT OF MATHEMATICS
Div. of Mathematical statistics
Ola Hössjer

MT 5006
SOLUTIONS
January 13 2023

# Solutions for Examination Categorical Data Analysis, January 13, 2023

## Problem 1

a. Since the total number of students is random, a reasonable model is Poisson sampling, where $N_{ij} \sim \text{Po}(\mu_{ij})$ are independent and Poisson distributed random variables for $0 \le i, j \le 2$, with $N_{ij}$ the number of students providing alternatives $i$ and $j$ to the first two questions. Then $\pi_{ij} = \mu_{ij}/\mu_{++}$ is the probability that a randomly chosen student belongs to cell $(i, j)$. The null hypothesis of independence between fantasy and sports watching habits can be phrased as

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \Longleftrightarrow \mu_{ij} = \frac{\mu_{i+}\mu_{+j}}{\mu_{++}} \tag{1}$$

for all $i, j$, where in the second step we used that $\pi_{i+} = \mu_{i+}/\mu_{++}$.

b. Let $n_{ij}$ be the observed value of $N_{ij}$, and $n = n_{++} = 102$ the total number of students. The maximum likelihood estimate of $\mu_{ij}$ under $H_0$ is

$$\hat{\mu}_{ij} = \frac{\hat{\mu}_{i+}\hat{\mu}_{+j}}{\hat{\mu}_{++}} = \frac{\frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n}}{n} = \frac{n_{i+}n_{+j}}{n}.$$

Inserting all values of $n_{ij}$, we get the following table of fitted expected values $\hat{\mu}_{ij}$:

| | Sports ($j$) | | | |
|---|---|---|---|---|
| Fantasy ($i$) | 0 | 1 | 2 | Total |
| 0 | 5.167 | 8.167 | 3.667 | 17 |
| 1 | 8.814 | 13.931 | 6.255 | 29 |
| 2 | 17.020 | 26.902 | 12.078 | 56 |
| Total | 31 | 49 | 22 | 102 |

c. The chisquare test statistic is

$$\begin{aligned} X^2 &= \textstyle\sum_{i,j=0}^{2} \frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \\ &= \frac{(3-5.167)^2}{5.167} + \frac{(2-8.167)^2}{8.167} + \frac{(12-3.667)^2}{3.667} + \ldots + \frac{(3-12.078)^2}{12.078} \\ &= 33.40 \\ &> \chi^2_4(0.05) = 9.49, \end{aligned} \tag{2}$$

and from this it follows that independence between watching sports and fantacy movies can be rejected at level 0.05. In the last step of (2) we used that the number of degrees of freedom is df $= 9-5 = 4$, since there are $3\times 3 = 9$ parameters $\mu_{ij}$ of the full model, and 5 parameters (e.g. $\mu_{++}$, $\pi_{1+}$, $\pi_{2+}$, $\pi_{+1}$ and $\pi_{+2}$) for the independence model. We can also make use of df $= (3-1)(3-1) = 4$.

d. The number of concordant and discordant pairs are

$$\begin{aligned} C &= 3(4+17+33+3)+2(7+3)+8(33+3)+14\cdot 3 = 521, \\ D &= 2(8+20)+12(8+14+20+33)+14\cdot 20+7(20+33) = 1607 \end{aligned}$$

respectively. Therefore, an estimator of the difference between the fraction of all concordant/discordant pairs that are concordant and discordant, is

$$\hat{\gamma} = \frac{521-1607}{521+1607} = -0.5103.$$

This indicates a negative association between watching fantacy and sports movies.

# Problem 2

a. We merge categories 0 and 1 of fantacy and sports into a new level 1. This gives a condensed $2\times 2$ table with the following cell counts $\tilde{n}_{ij}$:

| | Sports ($j$) | | |
|---|---|---|---|
| Fantasy ($i$) | 1 | 2 | Total |
| 1 | 27 | 19 | 46 |
| 2 | 53 | 3 | 56 |
| Total | 80 | 22 | 102 |

The estimator of the odds ratio

$$\theta = \frac{\tilde{\mu}_{11}\tilde{\mu}_{22}}{\tilde{\mu}_{12}\tilde{\mu}_{21}} \tag{3}$$

is

$$\hat{\theta} = \frac{\tilde{n}_{11}\tilde{n}_{22}}{\tilde{n}_{12}\tilde{n}_{21}} = \frac{27\cdot 3}{19\cdot 53} = 0.0804, \tag{4}$$

indicating quite strongly that watching fantacy and sports movies are negatively correlated.

b. Equations (3)-(4), and a first order Taylor expansion of the logarithmic function around the expected cell counts $\tilde{\mu}_{ij}$ gives

$$\begin{aligned} \log(\hat{\theta}) &= \log\tilde{N}_{11}+\log\tilde{N}_{22}-\log\tilde{N}_{12}-\log\tilde{N}_{21} \\ &\approx \left[\log\tilde{\mu}_{11}+\tfrac{\tilde{N}_{11}-\tilde{\mu}_{11}}{\tilde{\mu}_{11}}\right]+\left[\log\tilde{\mu}_{22}+\tfrac{\tilde{N}_{22}-\tilde{\mu}_{22}}{\tilde{\mu}_{22}}\right]-\left[\log\tilde{\mu}_{12}+\tfrac{\tilde{N}_{12}-\tilde{\mu}_{12}}{\tilde{\mu}_{12}}\right]-\left[\log\tilde{\mu}_{21}+\tfrac{\tilde{N}_{21}-\tilde{\mu}_{21}}{\tilde{\mu}_{21}}\right] \\ &\overset{(3)}{=} \log\theta+\tfrac{\tilde{N}_{11}-\tilde{\mu}_{11}}{\tilde{\mu}_{11}}+\tfrac{\tilde{N}_{22}-\tilde{\mu}_{22}}{\tilde{\mu}_{22}}-\tfrac{\tilde{N}_{12}-\tilde{\mu}_{12}}{\tilde{\mu}_{12}}-\tfrac{\tilde{N}_{21}-\tilde{\mu}_{21}}{\tilde{\mu}_{21}}. \end{aligned}$$

Since $\tilde{N}_{ij}$ are independent and Poisson distributed with $E(\tilde{N}_{ij}) = \mathrm{Var}(\tilde{N}_{ij}) = \tilde{\mu}_{ij}$ we find that approximately,

$$\mathrm{Var}[\log(\hat{\theta})] = \sum_{i,j=1}^{2} \frac{\mathrm{Var}(\tilde{N}_{ij})}{\tilde{\mu}_{ij}^2} = \sum_{i,j=1}^{2} \frac{1}{\tilde{\mu}_{ij}}. \tag{5}$$

c. The standard error

$$\begin{aligned} \mathrm{SE} &= \sqrt{\widehat{\mathrm{Var}}[\log(\hat{\theta})]} \\ &= \sqrt{\tfrac{1}{\tilde{n}_{11}} + \tfrac{1}{\tilde{n}_{12}} + \tfrac{1}{\tilde{n}_{21}} + \tfrac{1}{\tilde{n}_{22}}} \\ &= \sqrt{\tfrac{1}{27} + \tfrac{1}{19} + \tfrac{1}{53} + \tfrac{1}{3}} \\ &= 0.6647 \end{aligned}$$

of $\log(\hat{\theta})$ is obtained by first replacing all $\tilde{\mu}_{ij}$ by estimates $\tilde{n}_{ij}$ in (5), and then taking the square root. An approximate 95% confidence interval for $\theta$ is

$$I = \Big(\exp[\log(\hat{\theta}) - 1.96 \cdot \mathrm{SE}], \exp[\log(\hat{\theta}) + 1.96 \cdot \mathrm{SE}]\Big) = (0.0219, 0.2960). \tag{6}$$

The negative association between fantacy and sports watching is significant, since $1 \notin I$.

d. The accuracy of (6) is quite poor, since it relies on a large sample approximation, and there are only $\tilde{n}_{22} = 3$ observations in cell $(2, 2)$. But a more exact analysis is unlikely to change the conclusion $1 \neq I$, since the association between wathing fantacy and sports movies is strong.

# Problem 3

a. By adding the two partial contingency tables for $Z = 0$ and $Z = 1$, we get the following marginal $2 \times 2$ contingency table $\{n_{ij+}\}$ for $X$ and $Y$:

| Father's aff status | Son's aff status | |
|---|---|---|
| | $Y = 0$ | $Y = 1$ |
| $X = 0$ | 868 | 49 |
| $X = 1$ | 50 | 33 |

From the marginal and the two partial tables we obtain the following estimated marginal and conditional odds ratios:

$$\begin{aligned} \hat{\theta}_{XY} &= (868 \cdot 33)/(49 \cdot 50) = 11.691, \\ \hat{\theta}_{XY(0)} &= (841 \cdot 4)/(27 \cdot 30) = 4.153, \\ \hat{\theta}_{XY(1)} &= (27 \cdot 29)/(22 \cdot 20) = 1.779. \end{aligned}$$

The fact that $\hat{\theta}_{XY}$ is much larger than the two partial odds ratios indicate strongly that $Z$ is a common risk factor for fathers and sons. Since $\hat{\theta}_{XY(0)}$ and $\hat{\theta}_{XY(1)}$ are both larger than 1, this indicates (more weakly) other possible common (genetic or shared environmental) risk factors. Since $\hat{\theta}_{XY(0)}$ is larger than $\hat{\theta}_{XY(1)}$, there is possibly a third order interaction between $X$, $Y$ and $Z$.

b. The loglinear parametrization of $(XZ, YZ)$ is

$$\mu_{ijk} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}) \tag{7}$$

for $0 \leq i, j, k \leq 1$. Assume that $X = 0$, $Y = 0$ and $Z = 0$ are chosen as baseline levels. Then those loglinear parameters are put to zero for which at least one index $i$, $j$ or $k$ equals 0. The remaining parameters are

$$\boldsymbol{\beta} = (\lambda, \lambda_1^X, \lambda_1^Y, \lambda_1^Z, \lambda_{11}^{XZ}, \lambda_{11}^{YZ}). \tag{8}$$

c. It follows from (7) that

$$\mu_{ijk} = A_k B_{ik} C_{jk}, \tag{9}$$

with $A_k = \exp(\lambda + \lambda_k^Z)$, $B_{ik} = \exp(\lambda_i^X + \lambda_{ik}^{XZ})$ and $C_{jk} = \exp(\lambda_j^Y + \lambda_{jk}^{YZ})$. Then, summing over one of $i$ or $j$, or over both indeces simultaneously in (9), we find that

$$\begin{aligned} \mu_{i+k} &= A_k B_{ik} C_{+k}, \\ \mu_{+jk} &= A_k B_{+k} C_{jk}, \\ \mu_{++k} &= A_k B_{+k} C_{+k}. \end{aligned}$$

Consequently,

$$\frac{\mu_{i+k}\mu_{+jk}}{\mu_{++k}} = \frac{A_k B_{ik} C_{+k} \cdot A_k B_{+k} C_{jk}}{A_k B_{+k} C_{+k}} = A_k B_{ik} C_{jk} = \mu_{ijk}.$$

Alternatively, we may work directly with the cell probabilities $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$. Since $X$ and $Y$ are conditionally independent given $Z$ for model $(XZ, YZ)$, it follows that

$$\pi_{ijk} = \pi_{++k}\pi_{ij|k} = \pi_{++k}\pi_{i+|k}\pi_{+j|k} = \pi_{++k} \cdot \frac{\pi_{i+k}}{\pi_{++k}} \cdot \frac{\pi_{+jk}}{\pi_{++k}} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}},$$

and hence

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

d. The maximum likelihood estimates

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

of the expected cell counts are obtained by replacing $\mu_{i+k}$, $\mu_{+jk}$ and $\mu_{++k}$ by estimates $n_{i+k}$, $n_{+jk}$ and $n_{++k}$. From the given marginals of the two partial tables we can read off all $n_{i+k}$, $n_{+jk}$ and $n_{++k}$, for instance

$$\hat{\mu}_{000} = \frac{n_{0+0}n_{+00}}{n_{++0}} = \frac{868 \cdot 871}{902} = 838.2.$$

Continuing in this way for the other cells $(i, j, k)$, we get the following predicted expected cell counts $\hat{\mu}_{ijk}$:

Genetic variant $Z = 0$:

| Father's aff status | Son's aff status $Y=0$ | Son's aff status $Y=1$ | Sum |
|---|---|---|---|
| $X = 0$ | 838.2 | 29.8 | 868 |
| $X = 1$ | 32.8 | 1.17 | 34 |
| Sum | 871 | 31 | 902 |

Genetic variant $Z = 1$:

| Father's aff status | Son's aff status $Y=0$ | Son's aff status $Y=1$ | Sum |
|---|---|---|---|
| $X = 0$ | 23.5 | 25.5 | 49 |
| $X = 1$ | 23.5 | 25.5 | 49 |
| Sum | 47 | 51 | 98 |

e. The log likelihood ratio statistic for testing $(XZ, YZ)$ against the saturated model $(XYZ)$, is

$$\begin{aligned} G^2 &= 2\sum_{ijk} n_{ijk} \log \tfrac{n_{ijk}}{\hat{\mu}_{ijk}} \\ &= 2\left(841 \cdot \log \tfrac{841}{838.2} + \ldots + 29 \cdot \log \tfrac{29}{25.5}\right) \\ &= 6.731 \\ &> \chi^2_2(0.05) = 5.99, \end{aligned}$$

where in the last step we used that $\text{df} = 8 - 6 = 2$, since the saturated model has $2 \times 2 \times 2 = 8$ parameters, and the conditional independence model $(XZ, YZ)$ has 6 parameters according to (8). We thus reject conditional independence between $X$ and $Y$ given $Z$ at level 5%, indicating that there might be other common risk factors for fathers and sons.

# Problem 4

a. The loglinear parametrization for $(XY, XZ, YZ)$ requires addition of an $XY$-interaction term compared to (7). This gives

$$\mu_{ijk} = \exp(\lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{ij} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{jk}). \tag{10}$$

b. Let $\pi_{ijk} = \mu_{ijk}/\mu_{+++} = P(X = i, Y = j, Z = k)$, so that $\pi_{i+k} = P(X = i, Z = k)$. Using (10) we find that

$$\begin{aligned} \text{logit}[P(Y = 1|X = i, Z = k)] &= \log[P(Y = 1|X = i, Z = k)/P(Y = 0|X = i, Z = k)] \\ &= \log[(\pi_{i1k}/\pi_{i+k})/(\pi_{i0k}/\pi_{i+k})] \\ &= \log(\pi_{i1k}/\pi_{i0k}) \\ &= \log(\mu_{i1k}/\mu_{i0k}) \\ &= (\lambda + \lambda^X_i + \lambda^Y_1 + \lambda^Z_k + \lambda^{XY}_{i1} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{1k}) \\ &- (\lambda + \lambda^X_i + \lambda^Y_0 + \lambda^Z_k + \lambda^{XY}_{i0} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{0k}) \\ &= \alpha + \beta^X_i + \beta^Z_k, \end{aligned}$$

where in the last step we used that

$$\begin{aligned} \alpha &= \lambda^Y_1 - \lambda^Y_0, \\ \beta^X_i &= \lambda^{XY}_{i1} - \lambda^{XY}_{i0}, \\ \beta^Z_k &= \lambda^{YZ}_{1k} - \lambda^{YZ}_{0k}. \end{aligned}$$

If $X = 0$ and $Z = 0$ are chosen as baseline levels, then any loglinear parameter with $i = 0$ or $k = 0$ among it indeces is zero, which implies $\beta^X_0 = \beta^Z_0 = 0$. The only remaining parameters are $(\alpha, \beta^X_1, \beta^Z_1)$.

c. Since there is no third order interaction $XYZ$ in the model, the conditional odds ratio between $X$ and $Y$ does not depend on the level $k$ of the conditioning variable $Z$. (In contrast, the conditional odds ratios between $X$ and $Y$ of the saturated model, that are estimated in Problem 3a, depend on the level of $Z$.) We find that

$$\begin{aligned}\log(\theta_{XY}) &= \text{logit}[P(Y=1|X=1,Z=k)] - \text{logit}[P(Y=1|X=0,Z=k)] \\ &= \alpha + \beta_1^X + \beta_k^Z - (\alpha + \beta_0^X + \beta_k^Z) \\ &= \beta_1^X - \beta_0^X \\ &= \beta_1^X.\end{aligned}$$

A Wald type approximate 95% confidence interval for $\log(\theta_{XY})$ is

$$\begin{aligned}&(\hat{\beta}_1^X - 1.96\sqrt{\widehat{\text{Var}}(\hat{\beta}_1^X)}, \hat{\beta}_1^X + 1.96\sqrt{\widehat{\text{Var}}(\hat{\beta}_1^X)}) \\ =\ &(0.8347 - 1.96\sqrt{0.1255}, 0.8347 + 1.96\sqrt{0.1255}) \\ =\ &(0.1404, 1.5290),\end{aligned}$$

and the one for $\theta_{XY}$ is

$$I = (\exp(0.1404), \exp(1.5290)) = (1.15, 4.61).$$

Since $1 \notin I$, this indicates (weakly) that there are additional common risk factors for the father and son apart from $Z$.

d. Since

$$\text{logit}[\pi(0,1)] = \text{logit}[P(Y=1|Z=0,X=1)] = \alpha + \beta_1^X,$$

we first compute a standard error

$$\begin{aligned}\text{SE} &= \sqrt{\widehat{\text{Var}}(\hat{\alpha} + \hat{\beta}_1^X)} \\ &= \sqrt{\widehat{\text{Var}}(\hat{\alpha}) + 2\widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}_1^X) + \widehat{\text{Var}}(\hat{\beta}_1^X)} \\ &= \sqrt{0.0342 - 2 \cdot 0.0096 + 0.1255} \\ &= \sqrt{0.1405} \\ &= 0.3748,\end{aligned}$$

in order to find a Wald type 95% confidence interval

$$(\hat{\alpha} + \hat{\beta}_1^X - 1.96 \cdot \text{SE}, \hat{\alpha} + \hat{\beta}_1^X + 1.96 \cdot \text{SE}) = (-3.2816, -1.8125)$$

for $\text{logit}[\pi(0,1)]$, which we transform to find the confidence interval

$$\left(\frac{\exp(-3.2816)}{1+\exp(-3.2816)}, \frac{\exp(-1.8215)}{1+\exp(-1.8125)}\right) = (0.036, 0.140)$$

for $\pi(0,1)$.

# Problem 5

(a) The likelihood of data $\{n_{ik};\, 0 \le i,k \le 1\}$ is

$$l = \prod_{i,k=0}^{1} \exp(-\mu_{ik}) \frac{\mu_{ik}^{n_{ik}}}{n_{ik}!},$$

and the log likelihood

$$L = \log(l) = \sum_{i,k=0}^{1} \left[-\mu_{ik} + n_{ik} \log(\mu_{ik})\right] + C, \tag{11}$$

where $C = -\sum_{i,k} \log(n_{ik}!)$ is a constant not depending on the parameters.

(b) There are five parameters $\lambda, \lambda_0^X, \lambda_1^X, \lambda_0^Z, \lambda_1^Z$ in the given formula for all $\mu_{ik}$, but in order to avoid overparametrization we can only have one marginal parameter for $X$ and one for $Z$. If $X = 0$ and $Z = 0$ are both baseline levels, then $\lambda_0^X = \lambda_0^Z = 0$, and three parameters of $\boldsymbol{\beta} = (\lambda, \lambda_1^X, \lambda_1^Z)$ remain.

(c) Using (11) and the hint, score function component for $\lambda$ is

$$u_1(\boldsymbol{\beta}) = \frac{\partial L}{\partial \lambda} = (n_{00} - \mu_{00}) + (n_{01} - \mu_{01}) + (n_{10} - \mu_{10}) + (n_{11} - \mu_{11}). \tag{12}$$

In order to find the score function components for the other two parameters, we notice that $\partial \mu_{ik}/\partial \lambda_1^X = \mu_{ik}$ and $\partial \log(\mu_{ik})/\partial \lambda_1^X = 1$ if $(i,k) = (1,0)$ or $(1,1)$, whereas both of these partial derivatives are 0 if $(i,k) = (0,0)$ or $(0,1)$, so that

$$u_2(\boldsymbol{\beta}) = \frac{\partial L}{\partial \lambda_1^X} = (n_{10} - \mu_{10}) + (n_{11} - \mu_{11}). \tag{13}$$

Similarly one obtains

$$u_3(\boldsymbol{\beta}) = \frac{\partial L}{\partial \lambda_1^Z} = (n_{01} - \mu_{01}) + (n_{11} - \mu_{11}). \tag{14}$$

The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is found by solving the nonlinear system of equations

$$\begin{array}{rcl} u_1(\boldsymbol{\beta}) & = & 0, \\ u_2(\boldsymbol{\beta}) & = & 0, \\ u_3(\boldsymbol{\beta}) & = & 0 \end{array} \tag{15}$$

iteratively with respect to $\boldsymbol{\beta} = (\lambda, \lambda_1^X, \lambda_1^Z)$, using the fact that all $\mu_{ik} = \mu_{ik}(\boldsymbol{\beta})$ depend on the parameter vector in (12)-(14). The times of exposure $t_{ik}$ seem to be absent in (12)-(14), but they enter in $\mu_{ik}$. In conclusion, (15) are the three likelihood equations.

(d) The annual premium for a young driver that lives in a rural area, is

$$\begin{array}{rcl} \hat{P}_{10} & = & 110 \cdot \exp\left(\hat{\lambda} + \hat{\lambda}_1^X + \hat{\lambda}_0^Z\right) \\ & = & 110 \cdot \exp\left(\hat{\lambda} + \hat{\lambda}_1^X\right) \\ & = & 110 \cdot \exp\left(-3.10 + 0.25\right) \\ & = & 6.36, \end{array}$$

or 6 360 Swedish crowns.

(e) The elements of the Fisher information matrix are

$$J_{ab}(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_a \partial\beta_b}\right)$$

for $1 \leq a, b \leq 3$. Focusing on $a = b = 3$, i.e. the diagonal element of $\beta_3 = \lambda_1^Z$, it follows by differentiating (14) that

$$\frac{\partial^2 L}{\partial^2 \lambda_1^Z} = -\frac{\partial \mu_{01}}{\partial \lambda_1^Z} - \frac{\partial \mu_{11}}{\partial \lambda_1^Z} = -\mu_{01} - \mu_{11} = E\left(\frac{\partial^2 L}{\partial^2 \lambda_1^Z}\right) \Longrightarrow J_{33}(\boldsymbol{\beta}) = \mu_{01} + \mu_{11},$$

since the second derivative does not depend on data $\{n_{ik}\}$. Therefore it is constant and equal to its expected value.