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

MT 5006
SOLUTIONS
February 23 2023

Solutions for Examination
Categorical Data Analysis, February 23, 2023

Problem 1

a. The logistic regression with one single predictor x has

$$\pi(x) = P(Y=1|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}. \tag{1}$$

b. To test if the medicine has any preventive effect we formulate null and alternative hypotheses

$$\begin{array}{ll} H_0: & \beta = 0, \\ H_a: & \beta < 0. \end{array}$$

The Wald test statistic is

$$z_W = \frac{\hat{\beta}}{\sqrt{\widehat{\text{Var}}(\hat{\beta})}} = \frac{-0.32}{\sqrt{0.0225}} = -2.133 < -z_{0.05} = -1.645.$$

We can therefore reject the null hypothesis that the medicine has no preventive effect at level 0.05.

c. Plugging in parameter estimates and $x = 10$ into (1), we find that

$$\text{logit}[\hat{\pi}(10)] = \hat{\alpha} + 10\hat{\beta} = -3.1 - 10 \cdot 0.32 = -6.30,$$

so that the predicted probability of suffering from a heart attack within one year is

$$\hat{\pi}(10) = \frac{e^{-6.30}}{1+e^{-6.30}} = 0.0018 = 0.18\%,$$

for a patient with daily dose of 10 mg.

d. We use the delta method, so that a confidence interval for logit$[\pi(10)]$ is constructed at first. We have that

$$\begin{aligned} \widehat{\text{Var}}(\hat{\alpha} + 10\hat{\beta}) &= \widehat{\text{Var}}(\hat{\alpha}) + 2 \cdot 10 \cdot \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}) + 10^2 \cdot \widehat{\text{Var}}(\hat{\beta}) \\ &= 1.1 - 20 \cdot 0.06 + 100 \cdot 0.0225 \\ &= 2.15, \end{aligned}$$

which gives an approximate 95% confidence interval

$$\left(-6.30 - 1.96 \cdot \sqrt{2.15}, -6.30 + 1.96 \cdot \sqrt{2.15}\right) = (-9.1739, -3.4261)$$

for logit$[\pi(10)]$, and

$$\left(\frac{e^{-9.1739}}{1+e^{-9.1739}}, \frac{e^{-3.4261}}{1+e^{-3.4261}}\right) = (0.000104, 0.0313)$$

for $\pi(10)$.

Problem 2

a. Let

$$\begin{aligned}
\theta_I &= \mu_{11}\mu_{22}/(\mu_{12}\mu_{21}), \\
\theta_{II} &= \mu_{11}\mu_{33}/(\mu_{13}\mu_{31}), \\
\theta_{III} &= \mu_{22}\mu_{33}/(\mu_{23}\mu_{32}),
\end{aligned}$$

be the odds ratios of subtables I, II and III. They are estimated by

$$\begin{aligned}
\hat{\theta}_I &= n_{11}n_{22}/(n_{12}n_{21}) = 2.45, \\
\hat{\theta}_{II} &= n_{11}n_{33}/(n_{13}n_{31}) = 42, \\
\hat{\theta}_{III} &= n_{22}n_{33}/(n_{23}n_{32}) = 10.5.
\end{aligned}$$

These estimates suggest that degree of injury is strongly associated with health one year later, if the severe injury and bad health levels are included, as for subtables II and III. The association between no/mild injury and good/fair health is weaker, and possibly not significant for this rather small data set.

b. Since this data set has Poisson sampling, the null hypothesis of independence between the rows and columns of subtable I is $H_0 : \mu_{11}\mu_{22} = \mu_{12}\mu_{21}$, or equivalently $H_0 : \theta_I = 1$.

c. Fisher's exact test uses a hypergeometric distribution

$$P_{H_0}(N_{11} = n_{11}|n_{1+}, n_{2+}, n_{+1}, n_{+2}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}} = \frac{\binom{11}{n_{11}}\binom{12}{12-n_{11}}}{\binom{23}{12}}.$$

d. The one-sided alternative is $H_a : \theta_I > 1$. Since $n_{11} = 7$, we get a

$$\begin{aligned}
P - \text{value} &= \textstyle\sum_{k=7}^{11} P(N_{11} = k|n_{1+}, n_{2+}, n_{+1}, n_{+2}) \\
&= 0.1933 + 0.0604 + 0.0089 + 0.0005 + 0.0000 \\
&= 0.2632 \\
&> 0.05.
\end{aligned}$$

The association of subtable I is therefore not significant at level 0.05.

Problem 3

a. We regard (n_{11}, n_{21}) as data, since they determine uniquely the number of observations in the other two cells of subtable I. Since N_{11} and N_{21} are independent and binomially distributed with success probabilities π_1 and π_2, the likelihood is

$$
\begin{aligned}
l(\pi_1,\pi_2) &= P(N_{11}=n_{11}, N_{21}=n_{21}|\pi_1,\pi_2) \\
&= \binom{n_{1+}}{n_{11}}\pi_1^{n_{11}}(1-\pi_1)^{n_{1+}-n_{11}} \cdot \binom{n_{2+}}{n_{21}}\pi_2^{n_{21}}(1-\pi_2)^{n_{2+}-n_{21}} \\
&= \binom{n_{1+}}{n_{11}}\pi_1^{n_{11}}(1-\pi_1)^{n_{12}} \cdot \binom{n_{2+}}{n_{21}}\pi_2^{n_{21}}(1-\pi_2)^{n_{22}} \\
&= \binom{11}{7}\pi_1^{7}(1-\pi_1)^{4} \cdot \binom{12}{5}\pi_2^{5}(1-\pi_2)^{7} \\
&= 261360 \cdot \pi_1^7(1-\pi_1)^4\pi_2^5(1-\pi_2)^7.
\end{aligned}
$$

b. The relative risk is $r = \pi_1/\pi_2$. The twosided test that mild injury has no effect on health status, is based on null and alternative hypotheses

$$
\begin{aligned}
H_0:&\quad r = 1, \\
H_a:&\quad r \neq 1.
\end{aligned}
$$

c. Let

$$
\begin{aligned}
L(\pi_1,\pi_2) &= \log[l(\pi_1,\pi_2)] \\
&= n_{11}\log(\pi_1) + n_{12}\log(1-\pi_1) + n_{21}\log(\pi_2) + n_{22}\log(1-\pi_2) + \text{constant}
\end{aligned}
$$

be the log likelihood, with a constant not depending on the parameters. Since $r = 1 \Leftrightarrow \pi_1 = \pi_2 = \pi$ under H_0, the null likelihood $L(\pi,\pi)$ is the same as for one binomial experiment with $n = n_{++}$ trials, success probability π and n_{+1} successes. Maximizing the corresponding log likelihood, we find that

$$
\begin{aligned}
L_0 &= \max_\pi L(\pi,\pi) \\
&= \max_\pi[n_{+1}\log(\pi) + n_{+2}\log(1-\pi) + \text{constant}] \\
&= L(\hat\pi,\hat\pi) \\
&= n_{11}\log(\tfrac{n_{+1}}{n}) + n_{12}\log(\tfrac{n_{+2}}{n}) + n_{21}\log(\tfrac{n_{+1}}{n}) + n_{22}\log(\tfrac{n_{+2}}{n}) + \text{constant},
\end{aligned}
$$

with $\hat\pi = n_{+1}/n$ the ML estimate of π. For the full model we maximize the log likelihoods for each row separately with respect to π_1 and π_2. This give a maximized log likelihood

$$
\begin{aligned}
L_1 &= \max_{\pi_1,\pi_2} L(\pi_1,\pi_2) \\
&= L(\hat\pi_1,\hat\pi_2) \\
&= n_{11}\log\tfrac{n_{11}}{n_{+1}} + n_{12}\log\tfrac{n_{12}}{n_{+1}} + n_{21}\log\tfrac{n_{21}}{n_{2+}} + n_{22}\log\tfrac{n_{22}}{n_{2+}} + \text{constant}
\end{aligned}
$$

for both rows combined. From this it follows that the likelihood ratio statistic is

$$
\begin{aligned}
G^2 &= -2(L_0 - L_1) \\
&= 2\left(n_{11}\log\tfrac{n_{11}/n_{1+}}{n_{+1}/n} + n_{12}\log\tfrac{n_{12}/n_{1+}}{n_{+2}/n} + n_{21}\log\tfrac{n_{21}/n_{2+}}{n_{+1}/n} + n_{22}\log\tfrac{n_{22}/n_{2+}}{n_{+2}/n}\right).
\end{aligned}
\tag{2}
$$

d. Insertion of the observed cell counts of subtable I into (2) gives

$$\begin{aligned} G^2 &= 2\left(7 \cdot \log \tfrac{7\cdot 23}{11\cdot 12} + 4 \cdot \log \tfrac{4\cdot 23}{11\cdot 11} + 5 \cdot \log \tfrac{5\cdot 23}{12\cdot 12} + 7 \cdot \log \tfrac{7\cdot 23}{12\cdot 11}\right) \\ &= 1.12 \\ &< \chi_1^2(0.05) = 3.84, \end{aligned}$$

where in the last step, the degrees of freedom is

$$\mathrm{df} = 2 - 1 = 1,$$

since the full model has 2 parameters (π_1 and π_2) and the null model only 1 (π). Therefore, we cannot conclude from this data set (at level 0.05) that a mild injury impacts health one year later.

Problem 4

a. The loglinear parametrization of (XY, Z) is

$$\mu_{ijk} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}) \tag{3}$$

for $1 \leq i, j, k \leq 2$. Assume that $X = 2$, $Y = 2$ and $Z = 2$ are chosen as baseline levels. Then all loglinear parameters are put to zero for which at least one index i, j or k equals 2. The remaining parameters are

$$\boldsymbol{\beta} = (\lambda, \lambda_1^X, \lambda_1^Y, \lambda_1^Z, \lambda_{11}^{XY}). \tag{4}$$

b. It follows from (3) that

$$\mu_{ijk} = A_{ij} B_k,$$

with $A_{ij} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY})$ and $B_k = \exp(\lambda_k^Z)$. Then

$$\begin{aligned} \mu_{ij+} &= A_{ij} B_+, \\ \mu_{++k} &= A_{++} B_k, \\ \mu_{+++} &= A_{++} B_+. \end{aligned}$$

Consequently,

$$\frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}} = \frac{A_{ij}B_+ \cdot A_{++}B_k}{A_{++}B_+} = A_{ij}B_k = \mu_{ijk}.$$

An alternative solution uses cell probabilities

$$\pi_{ijk} = \frac{\mu_{ijk}}{\mu_{+++}}$$

of the multinomial model, obtained by conditioning the Poisson model on the total cell count n_{+++}. Since Z is independent of X, Y, we have that

$$\mu_{ijk} = \mu_{+++} \cdot \pi_{ijk} = \mu_{+++} \cdot \pi_{ij+}\pi_{++k} = \mu_{+++} \cdot \frac{\mu_{ij+}}{\mu_{+++}} \cdot \frac{\mu_{++k}}{\mu_{+++}} = \frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}},$$

as was to be proved.

c. The ML-estimates

$$\hat{\mu}_{ijk} = \frac{n_{ij+}n_{++k}}{n}$$

of all expected cell counts of model (XY, Z) are found by replacing μ_{ij+}, μ_{++k} and μ_{+++} in the definition of μ_{ijk} by their corresponding observed values n_{ij+}, n_{++k} and $n = n_{+++}$. By summing data from the two partial tables we get the following marginal table for X and Y:

Values of n_{ij+}

	$i=1$	$i=2$
$j=1$	72	119
$j=2$	32	239

Since the total number of observations of the two partial tables are $n_{++1} = 168$ and $n_{++2} = 294$, and the total number of observations is $n = 168 + 294 = 462$, we get

$$\hat{\mu}_{111} = \frac{n_{11+}n_{++1}}{n} = \frac{72 \cdot 168}{462} = 26.18,$$

for cell $(1, 1, 1)$. A similar calculation of all other $\hat{\mu}_{ijk}$ gives the following result:

Values of $\hat{\mu}_{ij1}$:

	$j=1$	$j=2$
$i=1$	26.18	43.27
$i=2$	11.64	86.91

Values of $\hat{\mu}_{ij2}$:

	$j=1$	$j=2$
$i=1$	45.82	75.73
$i=2$	20.36	152.09

d. With Akaike's information criterion one chooses the model M that minimizes

$$\text{AIC}(M) = -2L(M) + 2p(M),$$

where $L(M)$ is the maximum log likelihood of M. We can use the log likelihood ratio statistic G^2 between (XY, Z) and (XYZ) for AIC-based selection between these two models, since

$$\begin{aligned} G^2 &= 2[L(XYZ) - L(XY, Z)] \\ &= 2\textstyle\sum_{ijk} n_{ijk} \log \frac{n_{ijk}}{\hat{\mu}_{ijk}} \\ &= 2\left(25 \cdot \log \frac{25}{26.18} + \ldots + 146 \cdot \log \frac{146}{152.09}\right) \\ &= 1.796 \\ &< 2[p(XYZ) - p(XY, Z)] = 2(8 - 5) = 6. \end{aligned}$$

In the last step we used that the saturated model has $p(XYZ) = 2\times2\times2 = 8$ parameters, and that the joint independence model between XY and Z has $p(XY, Z) = 5$ parameters according to (4). Since $\text{AIC}(XY, Z) < \text{AIC}(XYZ)$, we select the joint independence model.

Problem 5

a. The likelihood of Problem 3b can be written as

$$
\begin{aligned}
l(\alpha,\beta) &= \binom{n_1}{n_{11}} \left(\tfrac{\exp(\alpha)}{1+\exp(\alpha)}\right)^{n_{11}} \left(\tfrac{1}{1+\exp(\alpha)}\right)^{n_{12}} \\
&\cdot \binom{n_2}{n_{21}} \left(\tfrac{\exp(\alpha+\beta)}{1+\exp(\alpha+\beta)}\right)^{n_{21}} \left(\tfrac{1}{1+\exp(\alpha+\beta)}\right)^{n_{22}},
\end{aligned}
\tag{5}
$$

where $n_1 = n_{1+}$ and $n_2 = n_{2+}$. By taking the logarithm of (5) we get a log likelihood

$$
\begin{aligned}
L(\alpha,\beta) &= n_{11}\alpha - n_1 \log[1+\exp(\alpha)] \\
&+ n_{21}(\alpha+\beta) - n_2 \log[1+\exp(\alpha+\beta)] + C,
\end{aligned}
\tag{6}
$$

where $C = \log\binom{n_1}{n_{11}} + \log\binom{n_2}{n_{21}}$ is a constant that does not depend on the parameter vector (α,β).

b. Let $J_{ij} = J_{ij}(\alpha,\beta)$ denote element i,j of the Fisher information matrix. We have that

$$
J_{11} = -E\left(\frac{\partial^2 L(\alpha,\beta)}{\partial^2\alpha}\right),\ J_{12} = J_{21} = -E\left(\frac{\partial^2 L(\alpha,\beta)}{\partial\alpha\partial\beta}\right),\ J_{22} = -E\left(\frac{\partial^2 L(\alpha,\beta)}{\partial^2\beta}\right).
\tag{7}
$$

c. The score vector components are obtained from (6) as

$$
\begin{aligned}
\tfrac{\partial L(\alpha,\beta)}{\partial\alpha} &= n_{11} - n_1\left(1 - \tfrac{1}{1+\exp(\alpha)}\right) + n_{21} - n_2\left(1 - \tfrac{1}{1+\exp(\alpha+\beta)}\right), \\
\tfrac{\partial L(\alpha,\beta)}{\partial\beta} &= n_{21} - n_2\left(1 - \tfrac{1}{1+\exp(\alpha+\beta)}\right).
\end{aligned}
\tag{8}
$$

By differentiating (8) we find that the second order partial derivatives of L only depend on n_1 and n_2, which are fixed, not on the cell counts n_{ij}. Since the second order partial derivatives are constant they equal their expected values, and therefore (7) implies

$$
\begin{aligned}
J_{11} &= -\tfrac{\partial^2 L(\alpha,\beta)}{\partial\alpha^2} = n_1 \tfrac{\exp(\alpha)}{(1+\exp(\alpha))^2} + n_2 \tfrac{\exp(\alpha+\beta)}{(1+\exp(\alpha+\beta))^2} \\
&= n_1\pi_1(1-\pi_1) + n_2\pi_2(1-\pi_2), \\
J_{12} = J_{21} &= -\tfrac{\partial^2 L(\alpha,\beta)}{\partial\alpha\partial\beta} = n_2 \tfrac{\exp(\alpha+\beta)}{(1+\exp(\alpha+\beta))^2} = n_2\pi_2(1-\pi_2), \\
J_{22} &= -\tfrac{\partial^2 L(\alpha,\beta)}{\partial^2\beta} = n_2 \tfrac{\exp(\alpha+\beta)}{(1+\exp(\alpha+\beta))^2} = n_2\pi_2(1-\pi_2).
\end{aligned}
\tag{9}
$$

d. Replacing π_1 and π_2 by their estimates $\hat{\pi}_1 = n_{11}/n_1$ and $\hat{\pi}_2 = n_{21}/n_2$ in (9), we find that the observed Fisher information matrix

$$
\hat{\boldsymbol{J}} = \begin{pmatrix} \hat{J}_{11} & \hat{J}_{12} \\ \hat{J}_{21} & \hat{J}_{22} \end{pmatrix}
$$

has elements

$$
\begin{aligned}
\hat{J}_{11} &= n_{11}n_{12}/n_1 + n_{21}n_{22}/n_2, \\
\hat{J}_{12} = \hat{J}_{21} = \hat{J}_{22} &= n_{21}n_{22}/n_2.
\end{aligned}
\tag{10}
$$

Since the estimated covariance matrix of the parameter estimates is the inverse of the observed Fisher information matrix, we use the (10) and the hint to conclude that

$$\hat{J}_{11}\hat{J}_{22} - \hat{J}_{12}\hat{J}_{21} = \frac{n_{11}n_{12}}{n_1} \cdot \frac{n_{21}n_{22}}{n_2}$$

and

$$\begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\alpha}) & \widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}) \\ \widehat{\mathrm{Cov}}(\hat{\beta}, \hat{\alpha}) & \widehat{\mathrm{Var}}(\hat{\beta}) \end{pmatrix} = \begin{pmatrix} \hat{J}_{11} & \hat{J}_{12} \\ \hat{J}_{21} & \hat{J}_{22} \end{pmatrix}^{-1} = \frac{n_1}{n_{11}n_{12}} \cdot \frac{n_2}{n_{21}n_{22}} \begin{pmatrix} \hat{J}_{22} & -\hat{J}_{12} \\ -\hat{J}_{21} & \hat{J}_{11} \end{pmatrix},$$

which, in view of (10), simplifies to the expression given in Problem 5d, since $n_1 = n_{11} + n_{12}$ and $n_2 = n_{21} + n_{22}$.