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

MT 5006
SOLUTIONS
January 13, 2021

Solutions for Examination
Categorical Data Analysis, January 13, 2021

Problem 1

a. Let $\pi(x) = P(Y = 1|X = x)$. The simple linear logistic regression model asserts that

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + x\beta \Longleftrightarrow \pi(x) = \frac{e^{\alpha+x\beta}}{1+e^{\alpha+x\beta}}, \tag{1}$$

where α is the intercept and β the effect parameter.

b. Let

$$\theta = \frac{\pi(65)/[1-\pi(65)]}{\pi(45)/[1-\pi(45)]} = \frac{\exp(\alpha+65\beta)}{\exp(\alpha+45\beta)} = \exp(20\beta)$$

be the odds ratio of having experienced CHD symptoms between persons of age 65 and 45. The esimate of this odds ratio is

$$\hat{\theta} = \exp(20\hat{\beta}) = \exp(20 \cdot 0.111) = 9.21.$$

c. We use the delta method with a logarithmic transformation. Therefore, we start by calculating a Wald-type 95% confidence interval for $\log(\theta) = 20\beta$. It is given by

$$\begin{aligned} I_{\log(\theta)} &= 20\hat{\beta} \pm 1.96 \cdot \sqrt{\widehat{\text{Var}}(20\hat{\beta})} \\ &= 20 \cdot 0.111 \pm 1.96 \cdot 20 \cdot \sqrt{0.0006} \\ &= (1.260, 3.180) \end{aligned}$$

The corresponding 95% confidence interval for θ is

$$I_\theta = (\exp(1.260), \exp(3.180)) = (3.52, 24.1).$$

Since 1 is not included in this interval, we conclude that the data set is sufficiently large to support the conclusion that the odds of having experienced CHD symptoms is significantly higher (at level 5%) at age 65 compared to age 45.

d. We want to test

$$\begin{aligned} H_0: &\quad \text{Model (1) holds for age groups } x_1, x_2, x_3, x_4, \\ H_a: &\quad \text{Model (1) does not hold for all age groups } x_1, x_2, x_3, x_4. \end{aligned}$$

To this end we use the chisquare goodness-of-fit statistic

$$\begin{aligned} X^2 &= \sum_{a=1}^{4} \left[\frac{(n_{a0} - n_a(1-\hat{\pi}_a))^2}{n_a(1-\hat{\pi}_a)} + \frac{(n_{a1} - n_a\hat{\pi}_a)^2}{n_a\hat{\pi}_a} \right] \\ &= \sum_{a=1}^{4} \frac{(n_{a1} - n_a\hat{\pi}_a)^2}{n_a\hat{\pi}_a(1-\hat{\pi}_a)}. \end{aligned}$$

Insertion of numerical values from the table gives

$$X^2 = \frac{(3-2.49)^2}{2.49(1-0.100)} + \frac{(8-7.97)^2}{7.97(1-0.295)} + \frac{(11-11.75)^2}{11.75(1-0.560)} + \frac{(21-21.44)^2}{21.44(1-0.794)} = 0.269.$$

Under H_0, the X^2 statistic has a X_2^2-distribution, since the number of degrees of freedom for the test is $d = 4 - 2$. This follows since there are 4 parameters of the saturated independent binomial rows model, whereas the null model (the logistic regression model) has 2 parameters α and β. Since $0.269 < \chi_2^2(0.05) = 5.99$, we do not reject H_0 at significance level 5%.

Problem 2

a. The likelihood function is

$$l = \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}}. \tag{2}$$

Taking the logarithm of (2) we find that the log likelihood function equals

$$L = \log\left(\frac{n!}{\prod_{i,j} n_{ij}!} \right) + \sum_{i,j} n_{ij} \log(\pi_{ij}). \tag{3}$$

Since the cell probabilities π_{ij} sum to 1, there are only $IJ - 1$ free parameters. We may therefore parametrize the (log) likelihood by $\boldsymbol{\theta} = (\pi_{11}, \pi_{12}, \ldots, \pi_{IJ} - 1)$ and substitute $\pi_{IJ} = 1 - \sum_{(i,j)\neq(I,J)} \pi_{ij}$ into (2) and (3).

b. When testing the null hypothesis H_0 of independence between X and Y against the alternative hypothesis H_a of non-independence, we formulate this as

$$\begin{aligned} H_0: &\quad \pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all } i, j, \\ H_a &\quad \pi_{ij} \neq \pi_{i+}\pi_{+j} \text{ for at least one } i, j. \end{aligned}$$

c. The estimated proportional reduction in entropy equals

$$\hat{U} = \frac{\sum_{i,j} \hat{\pi}_{ij} \log(\hat{\pi}_{ij}/(\hat{\pi}_{i+}\hat{\pi}_{+j}))}{-\sum_{j=1}^{J} \hat{\pi}_{+j} \log(\hat{\pi}_{+j})}.$$

From this it follows that

$$\begin{aligned} -2n\hat{U} \sum_{j=1}^{J} \hat{\pi}_{+j} \log(\hat{\pi}_{+j}) &= 2n \sum_{i,j} \hat{\pi}_{ij} \log(\hat{\pi}_{ij}/(\hat{\pi}_{i+}\hat{\pi}_{+j})) \\ &= 2 \sum_{ij} n_{ij} \log(n_{ij}/\hat{\mu}_{ij}) \\ &= G^2 \\ &\overset{H_0}{\sim} \chi_d^2, \end{aligned} \tag{4}$$

where

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

are the fitted expected cell counts under the independence model H_0. The last step of (4) follows since G^2 is the likelihood ratio statistic for testing H_0 against H_a, which under the null hypothesis (and large samples) is chisquare distributed with

$$d = (IJ - 1) - (I + J - 2) = (I - 1)(J - 1)$$

degrees of freedom. This follows since the full multinomial model, according to 3a), has $IJ - 1$ parameters, whereas the independence model has $I + J - 2$ parameters ($I - 1$ marginal probabilities for X and $J - 1$ marginal probabilities for Y).

d. Inserting $n = 1009$, $\hat{U} = 0.052$ and the values of $\hat{\pi}_{+j}$ from the table, we find that

$$\begin{aligned}
&-2n\hat{U}\textstyle\sum_{j=1}^{J}\hat{\pi}_{+j}\log(\hat{\pi}_{+j}) \\
&\quad = -2 \cdot 1009 \cdot 0.0052 \cdot [0.142\log(0.142) + 0.299\log(0.299) + 0.559\log(0.559)] \\
&\quad = 10.11 \\
&\quad > \chi^2_4(0.05) = 9.49,
\end{aligned}$$

where in the last step we used that $I = J = 3$ and consequently $d = (3-1)(3-1) = 4$. We conclude that independence between age and job satisfaction is rejected at significance level 5%.

Problem 3

a. The data set is a threeway contingency table, where X has $I = 2$ levels, Y has $J = 2$ levels and Z has $K = 3$ levels. The four loglinear models that we will compare are nested. In the table below. we have denoted them M_0, M_1, M_2, M_3, and also computed the number of parameters $p(M)$ of each model. All four models share one intercept parameter λ and $(I - 1) + (J - 1) + (K - 1) = 4$ marginal parameters. the two rightmost terms of the $p(M)$ column. For M_1, M_2, M_3, we also added the relevant number of secdond order interaction parameters

M	$p(M)$
$M_3 = (XY, XZ, YZ)$	$10 = (J-1)\cdot(K-1) + (I-1)\cdot(K-1) + (I-1)\cdot(J-1) + 4 + 1$
$M_2 = (XY, XZ)$	$8 = (I-1)\cdot(K-1) + (I-1)\cdot(J-1) + 4 + 1$
$M_1 = (XY, Z)$	$6 = (I-1)\cdot(J-1) + 4 + 1$
$M_0 = (X, Y, Z)$	$5 = 4 + 1$

In the first step of the Forward Inclusion (FI) scheme we test

$$\begin{aligned}
H_0 &: \ M_0, \\
H_a &: \ M_1 \setminus M_0.
\end{aligned}$$

We use the likelihood ratio (LR) statistic

$$\begin{aligned}
G^2(M_0|M_1) &= G^2(M_0) - G^2(M_1) = 147.0 - 25.42 = 121.58 \\
&> \chi^2_{p(M_1)-p(M_0)}(0.05) = \chi^2_1(0.05) = 3.84.
\end{aligned}$$

Since H_0 is rejected at significance level 5% we proceed to the second step of the FI scheme and test

$$\begin{aligned} H_0: &\quad M_1, \\ H_a: &\quad M_2 \setminus M_1, \end{aligned}$$

using the LR statistic

$$\begin{aligned} G^2(M_1|M_2) &= G^2(M_1) - G^2(M_2) = 25.42 - 2.69 = 22.73 \\ &> \chi^2_{p(M_2)-p_(M_1)}(0.05) = \chi^2_2(0.05) = 5.99. \end{aligned}$$

Since H_0 is rejected also in the second step of the FI scheme, we proceed to the third step and test

$$\begin{aligned} H_0: &\quad M_2, \\ H_a: &\quad M_3 \setminus M_2, \end{aligned}$$

using the LR statistic

$$\begin{aligned} G^2(M_2|M_3) &= G^2(M_2) - G^2(M_3) = 2.69 - 1.67 = 1.02 \\ &< \chi^2_{p(M_3)-p_(M_2)}(0.05) = \chi^2_2(0.05) = 5.99. \end{aligned}$$

Since H_0 is not rejected in the third step of the FI scheme, $M_2 = (XY, XZ)$ is the selected model.

b. For model $M_1 = (XY, Z)$ we have that

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{++k}}{\mu_{+++}} \Longrightarrow \hat{\mu}_{ijk} = \frac{n_{ij+}n_{++k}}{n_{+++}}.$$

In particular,

$$\hat{\mu}_{122} = \frac{n_{12+}n_{++2}}{n_{+++}} = \frac{(25+39+13)(41664+4291+39+26)}{105633} = \frac{77 \cdot 46020}{105636} = 33.54.$$

For model $M_2 = (XY, XZ)$ we have that

$$\mu_{ijk} = \frac{\mu_{ij+}\mu_{i+k}}{\mu_{i++}} \Longrightarrow \hat{\mu}_{ijk} = \frac{n_{ij+}n_{i+k}}{n_{i++}}.$$

In particular,

$$\hat{\mu}_{122} = \frac{n_{12+}n_{1+2}}{n_{1++}} = \frac{(25+39+13)(41664+39)}{39160+41664+15163+25+39+13} = \frac{77 \cdot 41703}{96064} = 33.42.$$

c. Since X and Y are jointly independent of Z for model M_1, it follows that

$$P(Y=j|X=i, Z=k) = P(Y=j|X=i) = \mu_{ij+}/\mu_{i++}.$$

for all i, j, k. From this it follows that the conditional odds ratio

$$\theta^{XY}_{(k)} = \frac{P(Y=2|X=2)/P(Y=1|X=2)}{P(Y=2|X=1)/P(Y=1|X=1)} = \theta^{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

collapses to the marginal odds ratio between X and Y. Therefore the maximum likelihood estimate of $\theta^{XY}_{(k)} = \theta^{XY}$ is

$$\hat{\theta}^{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}} = \frac{(39160 + 41664 + 15163)(26 + 26 + 10)}{(25 + 39 + 13)(3882 + 4291 + 1337)} = \frac{95987 \cdot 62}{77 \cdot 9510} = 8.13.$$

Consequently, according to model M_1, the estimated odds of dying in leukemia is 8.13 times higher among those with a high radiation dose, compared the estimated odds among those with a low radiation dose, regardless of age.

Problem 4

a. The expected cell counts μ_{ijk} of the loglinear model $M_2 = (XY, XZ)$ are parametrized as

$$\log(\mu_{ijk}) = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{ij} + \lambda^{XZ}_{ik}, \quad 1 \le i, j \le 2, 1 \le k \le 3.$$

If we choose the lowest level ($i = j = k = 1$) of each variable as baseline, any parameter with at least one index at its lowest level is put to zero in order to avoid overparametrization. The remaining eight freely variable parameters are

$$\boldsymbol{\theta} = (\lambda, \lambda^X_2, \lambda^Y_2, \lambda^Z_2, \lambda^Z_3, \lambda^{XY}_{22}, \lambda^{XZ}_{22}, \lambda^{XZ}_{23}).$$

Similarly, for model $M_3 = (XY, XZ, YZ)$ we have that

$$\log(\mu_{ijk}) = \lambda + \lambda^X_i + \lambda^Y_j + \lambda^Z_k + \lambda^{XY}_{ij} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{jk}, \quad 1 \le i, j \le 2, 1 \le k \le 3.$$

If the lowest level of each variable is chosen as baseline, we get ten freely variable parameters

$$\boldsymbol{\theta} = (\lambda, \lambda^X_2, \lambda^Y_2, \lambda^Z_2, \lambda^Z_3, \lambda^{XY}_{22}, \lambda^{XZ}_{22}, \lambda^{XZ}_{23}, \lambda^{YZ}_{22}, \lambda^{YZ}_{23}).$$

b. Write $\pi_{ijk} = \mu_{ijk}/\mu_{+++}$ for the cell probabilities of model M_3 under multinomial sampling. Then

$$\begin{aligned}
&\text{logit}P(Y = 2|X = i, Z = k) \\
&\quad = \log P(Y = 2|X = i, Z = k) - \log P(Y = 1|X = i, Z = k) \\
&\quad = \log(\pi_{i2k}/\pi_{i+k}) - \log(\pi_{i1k}/\pi_{i+k}) \\
&\quad = \log(\pi_{i2k}) - \log(\pi_{i1k}) \\
&\quad = \log(\mu_{i2k}/\mu_{+++}) - \log(\mu_{i1k}/\mu_{+++}) \\
&\quad = \log(\mu_{i2k}) - \log(\mu_{i1k}) \qquad (5) \\
&\quad = (\lambda + \lambda^X_i + \lambda^Y_2 + \lambda^Z_k + \lambda^{XY}_{i2} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{2k}) \\
&\quad -(\lambda + \lambda^X_i + \lambda^Y_1 + \lambda^Z_k + \lambda^{XY}_{i1} + \lambda^{XZ}_{ik} + \lambda^{YZ}_{1k}) \\
&\quad = (\lambda^Y_2 - \lambda^Y_1) + (\lambda^{XY}_{i2} - \lambda^{XY}_{i1}) + (\lambda^{YZ}_{2k} - \lambda^{YZ}_{1k}) \\
&\quad = \lambda^Y_2 + \lambda^{XY}_{i2} + \lambda^{YZ}_{2k} \\
&\quad =: \alpha + \beta^X_i + \beta^Z_k,
\end{aligned}$$

if we use the parameter constraints of M_3 from a). It follows from (5) that $Y|X, Z$ is a logistic regression model with four nonzero parameters $\alpha = \lambda^Y_2$, $\beta^X_2 = \lambda^{XY}_{22}$, $\beta^Z_2 = \lambda^{YZ}_{22}$ and $\beta^Z_3 = \lambda^{YZ}_{23}$.

c. The likelihood of the loglinear models M_2 and M_3, with parameter vector $\boldsymbol{\theta}$, can be factorized into two parts;

$$\begin{array}{rcl} l(\boldsymbol{\theta}) & = & \prod_{i,j,k} P(N_{ijk} = n_{ijk}) \\ & = & \prod_{i,k} P(N_{i+k} = n_{i+k}) \\ & & \cdot \prod_{i,k} \binom{n_{i+k}}{n_{i2k}} P(Y = 2|X = i, Z = k)^{n_{i2k}} P(Y = 1|X = i, Z = k)^{n_{i1k}}, \end{array} \tag{6}$$

where the first term on the right hand side of (6) corresponds to the likelihood of the saturated loglinear model (XZ) for the two predictor variables X and Z, whereas the second term corresponds to the likelihood of the logistic regression models derived from M_2 and M_3 respectively. Taking the logarithm of both sides of (6) we find that the log likelihood of M_2 and M_3 is given by

$$\begin{array}{rcl} L(\boldsymbol{\theta}) & = & \sum_{i,k} \log[P(N_{i+k} = n_{i+k})] \\ & + & \sum_{i,k} \log \left[\binom{n_{i+k}}{n_{i2k}} P(Y = 2|X = i, Z = k)^{n_{i2k}} P(Y = 1|X = i, Z = k)^{n_{i1k}} \right], \end{array} \tag{7}$$

i.e. the sum of the log likelihood for (XZ) and the log likelihood of the logistic regression model derived from M_2 and M_3 respectively.

Denote the logistic regression models obtained from $M_2 = (XY, XZ)$ by (X), whereas the logistic regression model obtained from $M_3 = (XY, XZ, YZ)$ is written as $(X + Z)$. Let $L(M_2)$ and $L(M_3)$ be the maximized the log likelihoods of M_2 and M_3, obtained by maximizing (7) with respect to the parameter vector $\boldsymbol{\theta}$. Since the saturated model (XZ) is used for the log likelihood of the first term in (7), it follows that the two terms in (7) can be maximized separately, Therefore, the maximized log likelihoods of M_2 and M_3 satisfy

$$\begin{array}{rcl} L(M_2) & = & L(XZ) + L(X), \\ L(M_3) & = & L(XZ) + L(X + Z) \end{array}$$

respectively, where $L(XZ)$ is the maximized log likelihood of the loglinear model (XZ), whereas $L(X)$ and $L(X + Z)$ are the maximized log likelihoods of the two logistic regression models. Since the deviances of M_2 and M_3 are provided in Problem 3, it follows that

$$2L(X + Z) - 2L(X) = 2L(M_3) - 2L(M_2) = G^2(M_2) - G^2(M_3) = 2.69 - 1.67 = 1.02.$$

We know from 4b) that $(X + Z)$ has four parameters. In the same way it can be shown that (X) has two parameters $\alpha = \lambda_2^Y$ and $\beta_2^X = \lambda_{22}^{XY}$. Therefore

$$\begin{array}{rcl} \mathrm{AIC}(X) & = & -2L(X) + 2 \cdot 2 = (1.02 + 2 \cdot 2 - 2 \cdot 4) - 2L(X + Z) + 2 \cdot 4 \\ & = & -2.98 + \mathrm{AIC}(X + Z) < \mathrm{AIC}(X + Z). \end{array}$$

From this we conclude that the smaller logistic regression model X (derived from M_2), is chosen by the AIC criterion.

Problem 5

a. Recall that z_S is approximately standard normal under the null hypothesis $H_0 : \beta = 0$. Since the alternative hypothesis $H_a : \beta > 0$ is one-sided, we reject the null hypothesis at level 0.1 %, if $z_S > z_{0.001} = 3.09$. Evaluation of the score statistic yields

$$z_S = \frac{300.5}{\sqrt{0.43(1-0.43)\cdot 14475}} = 5.045 > 3.09.$$

Consequently, the dataset is large enough in order to reject H_0 at significance level 0.1 %.

b. We have that

$$\begin{array}{rcl} L(\alpha,\beta) & = & \sum_{a=1}^{4}\left[\log\binom{n_a}{n_{a1}} + n_{a1}\log\frac{\pi_a}{1-\pi_a} + n_a\log(1-\pi_a)\right] \\ & = & \sum_{a=1}^{4}\left\{\log\binom{n_a}{n_{a1}} + n_{a1}(\alpha+\beta x_a) - n_a\log[1+\exp(\alpha+\beta x_a)]\right\}. \end{array} \tag{8}$$

c. Differentiating (8) with respect to α and β we find that

$$\begin{array}{rcl} u_\alpha(\alpha,\beta) & = & \sum_a (n_{a1} - n_a\pi_a), \\ u_\beta(\alpha,\beta) & = & \sum_a x_a(n_{a1} - n_a\pi_a). \end{array} \tag{9}$$

d. It is assumed that the sampling scheme of the age grouped data set has independent binomial rows sampling, with $N_{a1} \sim \mathrm{Bin}(n_a, \pi_a)$. From this and (9) it follows that

$$\begin{array}{rcl} J_{\alpha\alpha}(\alpha,\beta) & = & \mathrm{Var}(u_{\alpha\alpha}(\alpha,\beta)) = \sum_a \mathrm{Var}(N_{a1}) = \sum_a n_a\pi_a(1-\pi_a), \\ J_{\alpha\beta}(\alpha,\beta) & = & \mathrm{Cov}(u_{\alpha\alpha}(\alpha,\beta), u_{\beta\beta}(\alpha,\beta)) = \sum_a x_a\mathrm{Var}(N_{a1}) = \sum_a x_a n_a\pi_a(1-\pi_a), \\ J_{\beta\beta}(\alpha,\beta) & = & \mathrm{Var}(u_{\beta\beta}(\alpha,\beta)) = \sum_a x_a^2\mathrm{Var}(N_{a1}) = \sum_a x_a^2 n_a\pi_a(1-\pi_a). \end{array} \tag{10}$$

e. It follows from the second equation of (9) and the first part of the hint that the numerator of the score statistic equals

$$u_\beta(\hat\alpha(0),0) = \sum_a x_a(n_{a1} - n_a p) = \sum_a (x_a - \bar x)(n_{a1} - n_a p) = \sum_a (x_a - \bar x)n_{a1}, \tag{11}$$

where in the second and third steps we used the definitions of p and $\bar x$ respectively. As for the denominator of the score statistic, we combine (10) with the first part of the hint. This gives

$$\begin{array}{rcl} J_{\alpha\alpha}(\hat\alpha(0),0) & = & p(1-p)\sum_a n_a, \\ J_{\alpha\beta}(\hat\alpha(0),0) & = & p(1-p)\sum_a x_a n_a, \\ J_{\beta\beta}(\hat\alpha(0),0) & = & p(1-p)\sum_a x_a^2 n_a. \end{array} \tag{12}$$

Then we use the second part of the hint, together with (12), and deduce

$$\begin{array}{rcl} \mathrm{Var}[u_\beta(\hat\alpha(0),0)] & = & p(1-p)\sum_a x_a^2 n_a - \left[p(1-p)\sum_a x_a n_a\right]^2 / \left[p(1-p)\sum_a n_a\right] \\ & = & p(1-p)\left[\sum_a x_a^2 n_a - \left(\sum_a x_a n_a\right)^2 / \sum_a n_a\right] \\ & = & p(1-p)\sum_a (x_a - \bar x)^2 n_a. \end{array} \tag{13}$$

Dividing (11) by the square root of (13), we finally arrive at the sought for expression

$$z_S = \frac{\sum_a (x_a - \bar{x}) n_{a1}}{\sqrt{p(1-p) \sum_a (x_a - \bar{x})^2 n_a}}$$

of the score statistic.