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

MT 5006
SOLUTIONS
February 16 2022

Solutions for Examination
Categorical Data Analysis, February 16, 2022

Problem 1

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

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

where α is the intercept and β the effect parameter.

b. We obtain $\hat{\pi}(60)$ by plugging the ML estimates of α and β into (1), i.e.

$$\hat{\pi}(60) = \frac{e^{\alpha+60\cdot\hat{\beta}}}{1+e^{\alpha+60\cdot\hat{\beta}}} = \frac{e^{-5.31+60\cdot 0.111}}{1+e^{-5.31+60\cdot 0.111}} = \frac{e^{1.350}}{1+e^{1.350}} = 0.794.$$

c. We have that

$$\begin{aligned}\text{Var}(\text{logit}(\hat{\pi}(60)) &= \text{Var}\left(\hat{\alpha} + 60\cdot\hat{\beta}\right)\\ &= \text{Var}(\hat{\alpha}) + 2\cdot 60\cdot\text{Cov}(\hat{\alpha},\hat{\beta}) + 60^2\cdot\text{Var}(\hat{\beta}),\end{aligned}$$

which gives a standard error

$$\begin{aligned}\text{SE} &= \sqrt{\widehat{\text{Var}}\left(\text{logit}(\hat{\pi}(60))\right)}\\ &= \sqrt{1.2852 - 120\cdot 0.0267 + 3600\cdot 0.0006}\\ &= 0.4911.\end{aligned}$$

a 95% Wald type confidence interval

$$(1.350 \pm 1.96\cdot 0.4911) = (0.3874, 2.3126)$$

for $\text{logit}(\pi(60))$, and

$$\left(\frac{e^{0.3874}}{1+e^{0.3874}}, \frac{e^{2.3126}}{1+e^{2.3126}}\right) = (0.596, 0.910)$$

for $\pi(60)$.

Problem 2

a. The Wald test statistic for testing $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ is

$$z_S = \frac{\hat{\beta}}{\sqrt{\widehat{\text{Var}}(\hat{\beta})}} = \frac{0.111}{\sqrt{0.0006}} = 4.53,$$

which exceeds $z_{0.025} = 1.96$, the 0.975 quantile of a standard normal distribution. Hence we conclude that age has a significant effect.

b. We can equivalently formulate the two hypotheses as

$$\begin{array}{rl} H_0 : & M_0 \text{ holds}, \\ H_a : & M_1 \text{ holds but not } M_0. \end{array}$$

The deviance of model M is $G^2(M) = -2(L(M) - L(M_{\text{sat}}))$, where $L(M)$ and $L(M_{\text{sat}})$ are the log likelihoods for M and the saturated model M_{sat}. Therefore, the likelihood ratio test statistic is

$$G^2(M_0|M_1) = -2(L(M_0) - L(M_1)) = G^2(M_0) - G^2(M_1) = 136.66 - 107.35 = 29.31,$$

which approximately has a χ_1^2-distribution under H_0, since M_1 has $2 - 1 = 1$ additional parameter compared to M_0. By the table in Appendix A, $G^2(M_0|M_1)$ exceeds $\chi_1^2(0.05) = 3.84$, and therefore H_0 is rejected.

c. The log likelihood is

$$\begin{aligned} L(\alpha, \beta) &= \textstyle\sum_{i=1}^{100} \log\left(\pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}\right) \\ &= \textstyle\sum_{i=1}^{100} \left(y_i \log(\pi(x_i)/(1 - \pi(x_i))) + \log(1 - \pi(x_i))\right) \\ &= \textstyle\sum_{i=1}^{100} \left(y_i(\alpha + \beta x_i) - \log(1 + e^{\alpha+\beta x_i})\right). \end{aligned}$$

Differentiating with respect to α and β we find $\hat{\alpha}$ and $\hat{\beta}$ as the solution of the two likelihood equations

$$\begin{aligned} 0 &= \partial L(\alpha, \beta)/\partial\alpha = \textstyle\sum_{i=1}^{100} (y_i - \pi(x_i)), \\ 0 &= \partial L(\alpha, \beta)/\partial\beta = \textstyle\sum_{i=1}^{100} x_i (y_i - \pi(x_i)), \end{aligned} \tag{2}$$

using that $\partial \log(1 + e^{\alpha+\beta x_i})/\partial\alpha = \pi(x_i)$ and $\partial \log(1 + e^{\alpha+\beta x_i})/\partial\beta = x_i\pi(x_i)$. The dependence on α and β is implicit in (2) through all $\pi(x_i)$.

d. Data can be summarized in a $I \times 2$ table, with cell (i, j) containing N_{ij}, the number of patients of age x^i and CHD status j for $i = 1, \ldots, I$ and $j = 0, 1$. Then M_1 corresponds to a loglinear model where all N_{ij} are independent and Poisson distributed;

$$N_{ij} \sim \text{Po}\left(\exp(\lambda_i^X + \alpha 1_{\{j=1\}} + \beta x^i 1_{\{j=1\}})\right),$$

and $1_{\{j=1\}}$ is the indicator function for $j = 1$. This model has I main effect parameters $\lambda_1^X, \ldots, \lambda_I^X$ for X, one main effect parameter α for Y and one single interaction parameter β between X and Y.

Problem 3

a. The relative risk ratio is $r = \pi_1/\pi_2$.

b. The likelihood function $l(\pi_1, \pi_2)$ is maximized by

$$\begin{aligned} \hat{\pi}_1 &= n_{11}/n_1, \\ \hat{\pi}_2 &= n_{21}/n_2, \end{aligned}$$

We could equivalently write the two parameters as (for instance) π_2 and $r = \pi_1/\pi_2$, and since the maximum likelihood estimates transform in the same way as the parameters, it follows that the ML estimate of r is

$$\hat{r} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{n_{11}n_2}{n_{21}n_1}. \tag{3}$$

c. We estimate $\mathrm{Var}(\log(\hat{r}))$ by replacing π_1 and π_2 with $\hat{\pi}_1$ and $\hat{\pi}_2$ respectively. This gives

$$\mathrm{SE} = \sqrt{\widehat{\mathrm{Var}}(\log(\hat{r}))} = \sqrt{\frac{1-\hat{\pi}_1}{n_1\hat{\pi}_1} + \frac{1-\hat{\pi}_2}{n_2\hat{\pi}_2}} = \sqrt{\frac{n_{12}}{n_1 n_{11}} + \frac{n_{22}}{n_2 n_{21}}}. \tag{4}$$

d. We find from (3) and (4) that

$$\hat{r} = \frac{200 \cdot 40000}{360 \cdot 20000} = 1.11 \Longleftrightarrow \log(\hat{r}) = 0.1054,$$

and

$$\mathrm{SE} = \sqrt{\frac{19800}{20000 \cdot 200} + \frac{39640}{40000 \cdot 360}} = 0.0878.$$

This gives a 95% Wald confidence interval

$$(0.1054 \pm 1.96 \cdot 0.0878) = (-0.0667, 0.2775)$$

for $\log(r)$ and

$$(e^{-0.0667}, e^{0.2775}) = (0.935, 1.320) \tag{5}$$

for r. Since (5) includes 1, there is no significant difference in accident rates between the two regions. Even though the data set is large, the smallness of the accident rates makes the confidence intervals wide.

Problem 4

a. All models have one baseline parameter λ. Since S, E and I are binary variables, each type of main effect (e.g. λ^S), second order interaction effect (e.g. λ^{SE}) and third order interaction effect (λ^{SEI}) contributes with one parameter. Adding the number of parameters of different interaction orders, we find:

Model	p	df
(S,E,I)	1+3+0+0=4	4
(SE,I)	1+3+1+0=5	3
(SI,E)	1+3+1+0=5	3
(S,EI)	1+3+1+0=5	3
(SE,EI)	1+3+2+0=6	2
(SE,SI)	1+3+2+0=6	2
(SE,SI,EI)	1+3+3+0=7	1
(SEI)	1+3+3+1=8	0

Since the saturated model has $2 \times 2 \times 2 = 8$ parameters, in the last column, a model with p parameters has $\text{df} = 8 - p$.

b. Let $L(M)$ and $p(M)$ be the log likelihood and number of parameters for model M. The deviance equals $G^2(M) = -2(L(M) - L(\text{SEI}))$, with (SEI) the saturated model. Hence

$$\text{AIC}(M) = -2L(M) + 2p(M) = G^2(M) + 2p(M) - 2L(\text{SEI}). \tag{6}$$

Since the last term on the right hand side of (6) is the same for all models M, it is equivalent to minimize $\text{AIC}(M)$ and $G^2(M) + 2p(M)$. The values of $G^2(M)$ were given, and those for $p(M)$ were found in a). From this we find that (SEI) minimizes AIC.

c. The likelihood ratio test statistic for testing

$$\begin{aligned} H_0 &: \ \text{(SE,SI,EI)}, \\ H_a &: \ \text{(SEI) but not (SE,SI,EI)}, \end{aligned}$$

is

$$\begin{aligned} G^2(\text{SE,SI,EI}|\text{SEI}) &= -2(L(\text{SE,SI,EI}) - L(\text{SEI})) = G^2(\text{SE,SI,EI}) \\ &= 2.85 < \chi^2_{8-7}(0.05) = 3.8415, \end{aligned}$$

so that (SE,SI,EI) is not rejected at level 0.05.

d. The cell counts of model (SE,SI,EI) have a Poisson distribution

$$N_{sei} \sim \text{Po}\left(\exp(\lambda + \lambda_s^S + \lambda_e^E + \lambda_i^I + \lambda_{se}^{SE} + \lambda_{si}^{SI} + \lambda_{ei}^{EI})\right),$$

with $s, e, i \in \{1, 2\}$. If level 2 of each variable is used as baseline, any parameter with at least one of its indeces s, e or i equal to 2 is put to zero. This gives a parameter vector

$$(\lambda, \lambda_1^S, \lambda_1^E, \lambda_1^I, \lambda_{11}^{SE}, \lambda_{11}^{SI}, \lambda_{11}^{EI}).$$

Problem 5

a. Since N_{ij} are independent Poisson variables with means μ_{ij}, we find that

$$P(N_{11} = n_{11}, \ldots, N_{IJ} = n_{IJ}) = \prod_{i,j} e^{-\mu_{ij}} \frac{\mu_{ij}^{n_{ij}}}{n_{ij}!}. \tag{7}$$

b. The cell counts $\{N_{ij}\}$ have a multinomial distribution with n trials and success probabilities $\{\pi_{ij}\}$, i.e.

$$P(N_{11} = n_{11}, \ldots, N_{IJ} = n_{IJ}) = \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}}. \qquad (8)$$

c. The cohort study typically has multinomial sampling. In contrast, if the rows and columns correspond to different levels of the explanatory and response variables, the clinical trial has independent multinomial rows, whereas the case-control study has independent multinomial columns.

d. By additivity properties of independent Poisson variables, it follows that the total cell count has distribution

$$N = \sum_{ij} N_{ij} \sim \text{Po}\left(\sum_{ij} \mu_{ij}\right) = \text{Po}(\mu)$$

under Poisson sampling. Together with (7), this implies

$$\begin{aligned} P(N_{11} = n_{11}, \ldots, N_{IJ} = n_{IJ} | N = n) &= P(N_{11} = n_{11}, \ldots, N_{IJ} = n_{IJ}, N = n)/P(N = n) \\ &= P(N_{11} = n_{11}, \ldots, N_{IJ} = n_{IJ})/P(N = n) \\ &= \prod_{i,j} e^{-\mu_{ij}} \frac{\mu_{ij}^{n_{ij}}}{n_{ij}!} / \left(e^{-\mu} \frac{\mu^n}{n!}\right) \\ &= \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{ij} \left(\frac{\mu_{ij}}{\mu}\right)^{n_{ij}}, \end{aligned}$$

which agrees with (8), since $\pi_{ij} = \mu_{ij}/\mu$.