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

MT 5019
SOLUTIONS
August 21 2025

Solutions for Examination Categorical Data Analysis, August 21, 2025

Problem 1

a. The linear logistic regression model has

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}. \tag{1}$$

b. We want to find a confidence interval for $\pi(1.5)$. We first look at $\text{logit}\,[\pi(1.5)] = \alpha + 1.5\beta$, whose point estimate is

$$\text{logit}\,[\hat{\pi}(1.5)] = \hat{\alpha} + 1.5\hat{\beta} = -2.2 + 1.5 \cdot 0.8 = -1. \tag{2}$$

Since

$$\begin{array}{rcl} \text{Var}\,[\text{logit}(\hat{\pi}(1.5))] & = & \text{Var}(\hat{\alpha}) + 2 \cdot 1.5 \cdot \text{Cov}(\hat{\alpha}, \hat{\beta}) + 1.5^2 \cdot \text{Var}(\hat{\beta}) \\ & = & \text{Var}(\hat{\alpha}) + 3 \cdot \text{Cov}(\hat{\alpha}, \hat{\beta}) + 2.25 \cdot \text{Var}(\hat{\beta}), \end{array}$$

this gives a standard error for the estimate in (2) that equals

$$\begin{array}{rcl} \text{SE} & = & \sqrt{\widehat{\text{Var}}(\hat{\alpha}) + 3 \cdot \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}) + 2.25 \cdot \widehat{\text{Var}}(\hat{\beta})} \\ & = & \sqrt{0.02 + 3 \cdot (-0.005) + 2.25 \cdot 0.01} \\ & = & 0.1658 \end{array}$$

and a Wald type confidence interval

$$(-1 - 1.96 \cdot \text{SE}, -1 + 1.96 \cdot \text{SE}) = (-1.3250, -0.6750)$$

for $\text{logit}[\pi(1.5)]$ with approximate coverage probability 95%, since $z_{0.025} = 1.96$ is the 97.5% quantile of a standard normal distribution. The corresponding confidence interval for $\pi(1.5)$, with approximate coverage probability 95%, is

$$\left(\frac{\exp(-1.3250)}{1 + \exp(-1.3250)}, \frac{\exp(-0.6750)}{1 + \exp(-0.6750)} \right) = (0.210, 0.337).$$

c. Suppose Ben's smoking consumption is x pack years. Then Adam's is $x + 10/20 = x + 0.5$ pack years, since 10 cigarettes corresponds to half a pack. The odds ratio asked for is

$$\mathrm{OR} = \frac{\pi(x+0.5)/(1-\pi(x+0.5))}{\pi(x)/(1-\pi(x))} = \frac{\exp(\alpha+\beta(x+0.5))}{\exp(\alpha+\beta x)} = \exp(0.5\beta).$$

By a similar argument as in a), we first compute a confidence interval

$$\begin{aligned}\left(\hat{\beta} - 1.96\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}, \hat{\beta} + 1.96\sqrt{\widehat{\mathrm{Var}}(\hat{\beta})}\right) &= (0.8 - 1.96 \cdot \sqrt{0.01}, 0.8 + 1.96 \cdot \sqrt{0.01}) \\ &= (0.6040, 0.9960)\end{aligned}$$

for β with approximate coverage probability 95%. The corresponding confidence interval for the odds ratio is

$$(\exp(0.5 \cdot 0.6040), \exp(0.5 \cdot 0.9960)) = (1.353, 1.645).$$

Problem 2

a. Let $\pi(x)$ be the probability in (1) of having lung cancer for a *randomly drawn* person with cumulative exposure x to smoking. For a *sampled* person with the same cumulative exposure x, the corresponding probability is

$$\begin{aligned}\pi^*(x) &= \rho\pi(x)/\left[(1-\pi(x)) + \rho\pi(x)\right] \\ &= \rho\exp(\alpha+\beta x)/\left[1+\rho\exp(\alpha+\beta x)\right] \\ &= \exp(\alpha^*+\beta^* x)/\left[1+\exp(\alpha^*+\beta^* x)\right],\end{aligned} \tag{3}$$

with

$$\begin{aligned}\alpha^* &= \alpha + \log(\rho), \\ \beta^* &= \beta.\end{aligned} \tag{4}$$

We used that $1-\pi(x) = 1/\left[1+\exp(\alpha+\beta x)\right]$, and therefore $1+\exp(\alpha+\beta x)$ cancelled out in the second step of (3). We recognize the right hand side of (3) as a logistic regression model, with a different intercept $\alpha^* \neq \alpha$ but the same slope $\beta^* = \beta$ as in Problem 1.

b. Viewing $\hat{\alpha}$ and $\hat{\beta}$ of Problem 1 as estimates of α^* and β^* in (4), we obtain the corresponding estimates

$$\begin{aligned}\tilde{\alpha} &= \hat{\alpha} - \log(\rho) = -2.2 - \log(10) = -4.5026, \\ \tilde{\beta} &= \hat{\beta} = 0.8\end{aligned} \tag{5}$$

of α and β. Inserting (5) into the formula for $\pi(1.5)$, we obtain a corrected estimate

$$\tilde{\pi}(1.5) = \frac{\exp(\tilde{\alpha}+1.5\tilde{\beta})}{1+\exp(\tilde{\alpha}+1.5\tilde{\beta})} = \frac{\exp(-4.5026+1.5\cdot 0.8)}{1+\exp(-4.5026+1.5\cdot 0.8)} = 0.0355 \tag{6}$$

of the probability of developing lung cancer for a person with a cumulative amount of smoking 1.5 pack years. Since this corrected estimate 0.0355 is almost six times smaller than the lower end point 0.21 of the confidence interval for $\pi(1.5)$ in Problem 1b, this shows that this interval is severely biased upwards.

Alternatively, we can compute the corrected estimate $\tilde{\pi}(1.5)$ of $\pi(1.5)$ directly from the estimates $\hat{\alpha}$ and $\hat{\beta}$ of Problem 1. Since lung cancer cases are oversampled by a factor $\rho = 10$ among persons with cumulative exposure 1.5, the odds of $\tilde{\pi}(1.5)$ is 10 times too large. Therefore,

$$\tilde{\pi}(1.5) = \frac{0.1 \cdot \exp(\hat{\alpha} + 1.5\hat{\beta})}{1 + 0.1 \cdot \exp(\hat{\alpha} + 1.5\hat{\beta})} = \frac{0.1 \cdot \exp(-2.2 + 1.5 \cdot 0.8)}{1 + 0.1 \cdot \exp(-2.2 + 1.5 \cdot 0.8)} = 0.0355.$$

c. Replacing ρ by $\hat{\rho}$ in (5), we find that

$$\begin{array}{rcl} \tilde{\alpha} & = & \hat{\alpha} - \log(\hat{\rho}) \\ & = & \alpha^* - \log(\rho) + (\hat{\alpha} - \alpha^*) - [\log(\hat{\rho}) - \log(\rho)] \\ & = & \alpha + (\hat{\alpha} - \alpha^*) - [\log(\hat{\rho}) - \log(\rho)] \\ & \approx & \alpha + (\hat{\alpha} - \alpha^*) - \frac{\hat{\rho}-\rho}{\rho}, \end{array} \tag{7}$$

where in the fourth step we used a first order Taylor expansion of the logarithmic function around ρ. Since α, α^* and ρ are constants, it follows from (7) that

$$\text{Var}(\tilde{\alpha}) \approx \text{Var}(\hat{\alpha}) - \frac{2}{\rho}\text{Cov}(\hat{\alpha}, \hat{\rho}) + \frac{1}{\rho^2}\text{Var}(\hat{\rho}).$$

The corresponding estimate of this variance is

$$\begin{array}{rcl} \widehat{\text{Var}}(\tilde{\alpha}) & = & \widehat{\text{Var}}(\hat{\alpha}) - \frac{2}{\hat{\rho}}\widehat{\text{Cov}}(\hat{\alpha}, \hat{\rho}) + \frac{1}{\hat{\rho}^2}\widehat{\text{Var}}(\hat{\rho}) \\ & = & 0.02 - \frac{2}{10} \cdot 0.005 + \frac{1}{10^2}0.5 \\ & = & 0.02 - 0.001 + 0.005 \\ & = & 0.024. \end{array}$$

Problem 3

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

$$\mu_{ijk} = \exp(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}) \tag{8}$$

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

$$\boldsymbol{\beta} = (\lambda, \lambda_1^X, \lambda_1^Y, \lambda_1^Z, \lambda_{11}^{XY}). \tag{9}$$

b. It follows from (8) 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

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

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, which does not require the loglinear parametrization from a), 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 $n = n_{+++}$. Since the total number of observations of the two partial tables are $n_{++0} = 148$ and $n_{++1} = 154$, and the total number of observations is $n = 302$, we get

$$\hat{\mu}_{000} = \frac{n_{00+}n_{++0}}{n} = \frac{115 \cdot 148}{302} = 56.36$$

for cell $(0, 0, 0)$. A similar calculation of all other $\hat{\mu}_{ijk}$ gives the following result:

Values of $\hat{\mu}_{ij0}$:

	$j = 0$	$j = 1$
$i = 0$	56.36	19.11
$i = 1$	41.66	30.87

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

	$j = 0$	$j = 1$
$i = 0$	58.64	19.89
$i = 1$	43.34	32.13

d. The chisquare statistic for testing the null hypothesis $H_0 : (XY, Z)$ against the alternative hypothesis $H_a : (XYZ)$ but not (XY, Z), is

$$\begin{aligned} X^2(XY, Z) &= \textstyle\sum_{ijk} \frac{(n_{ijk}-\hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}} \\ &= \frac{(60-56.36)^2}{56.36} + \ldots + \frac{(35-32.13)^2}{32.13} \\ &= 3.248 \\ &< \chi^2_3(0.05) = 7.81. \end{aligned} \tag{10}$$

Hence we cannot reject H_0, that gene B has no effect on Alzheimer's disease, at level 5%. Let $p(M)$ be the number of parameters of model M. In the last step of (10) we used that the number of degrees of freedom och the chisquare distribution is

$$\text{df} = p(XYZ) - p(XY, Z) = 8 - 5 = 3,$$

since the saturated model has one parameter for each cell, and there are $2 \times 2 \times 2 = 8$ cells in the table. From (9) we also know that the parameter vector $\boldsymbol{\beta}$ of (XY, Z) contains 5 parameters.

Problem 4

a. Submodel $(A*Z+A*W)$ has one intercept, three types of main effects (A, Z and W) and two types of second order interactions (AZ and AW). It follows that

$$P(Y=1|A=i,Z=k,W=h)=\frac{\exp(\alpha+\beta_i^A+\beta_k^Z+\beta_h^W+\beta_{ik}^{AZ}+\beta_{ih}^{AW})}{1+\exp(\alpha+\beta_i^A+\beta_k^Z+\beta_h^W+\beta_{ik}^{AZ}+\beta_{ih}^{AW})}.$$

b. The number of parameters of $(A*W+A*W)$ is

$$p=1+(4-1)+(4-1)+(4-1)+(4-1)(4-1)+(4-1)(4-1)=28,$$

where the first term corresponds to an intercept, each main effect contributes with $4-1=3$ parameters (one per level; excluding the baseline level), and each second order interaction adds $(4-1)(4-1)=9$ parameters (one for each pair of levels, none of which is a baseline level).

c. Reasoning as in b), each main effect and second order interaction adds $4-1=3$ and $(4-1)^2=9$ parameters respectively. Since each model is balanced, we know how many main effects and second order interactions it includes. This gives the following completion of the third column of the given table:

M	$L(M)$	$p(M)$	$-2L(M)+2p(M)$
$(A*Z+A*W)$	-5000.0	$1+3\cdot 3+2\cdot 9=28$	10056
$(Z+A*W)$	-5005.5	$1+3\cdot 3+9=19$	10049
$(W+A*Z)$	-5008.0	$1+3\cdot 3+9=19$	10054
$(A+Z+W)$	-5025.0	$1+3\cdot 3=10$	10070
None	-5200.0	1	10402

d. Akaike's information criterion is

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

It has been evaluated in the fourth column of the above table. The chosen model, with lowest AIC(M), is therefore $(Z+A*W)$.

e. In backward elimination (BE), we first select the largest model $(A*Z+A*W)$ among those that are being tested. Then we test each one of $(Z+A*W)$ and $(W+A*Z)$, in which one type of second order interaction has been removed from $(A*Z+A*W)$, against $(A*Z+A*W)$. The corresponding two likelihood ratio statistics are

$$\begin{aligned}G^2(Z+A*W|A*Z+A*W) &= 2\left[L(A*Z+A*W)-L(Z+A*W)\right]\\ &= 2\left[-5000-(-5005.5)\right]=11\\ &< \chi^2_{28-19}(0.05)=16.62,\end{aligned}$$

and

$$\begin{aligned}G^2(W+A*Z|A*Z+A*W) &= 2\left[L(A*Z+A*W)-L(W+A*Z)\right]\\ &= 2\left[-5000-(-5008)\right]=16\\ &< \chi^2_{28-19}(0.05)=16.62,\end{aligned}$$

respectively. We notice that the null hypothesis (the smaller model) is not rejected at level 5% in any of these two tests, but since the LR statistic is smaller when $Z + A * W$ is being tested against $A * Z + A * W$, we select $Z + A * W$ after the first step.

Then, in the second step of the BE procedure, we check whether removal of interaction $A * W$ degrades model fit significantly. That is, we only test one model $A + Z + W$ against $Z + A * W$. This gives an LR statistic

$$\begin{aligned} G^2(A + Z + W | Z + A * W) &= 2\,[L(Z + A * W) - L(A + Z + W)] \\ &= 2\,[-5005.5 - (-5025)] = 39 \\ &> \chi^2_{19-10}(0.05) = 16.62. \end{aligned}$$

Since the null hypothesis $A + Z + W$ is rejected at level 5%, the BE scheme stops and model $Z + A * W$ is selected. Therefore, backward elimination and Akaike's criterion in d) give the same result.