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

MT 5006
SOLUTIONS
January 10 2019

Solutions for Examination Categorical Data Analysis, January 10, 2019

Problem 1

a. We will first find a 95% confidence interval for $\mathrm{logit}[\pi(2)] = \beta_0 + 2\beta_1 + 4\beta_2$. A point estimate of this quantity is

$$\begin{aligned}\mathrm{logit}[\hat{\pi}(2)] &= \hat{\beta}_0 + 2\hat{\beta}_1 + 4\hat{\beta}_2 \\ &= -6.0 + 2\cdot 1.0 + 4\cdot 0.5 \\ &= -2.0.\end{aligned}$$

Since

$$\begin{aligned}\mathrm{Var}[\mathrm{logit}(\hat{\pi}(2))] &= \mathrm{Var}(\hat{\beta}_0) + 4\mathrm{Var}(\hat{\beta}_1) + 16\mathrm{Var}(\hat{\beta}_2) \\ &+ 4\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1) + 8\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_2) + 16\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2),\end{aligned}$$

the squared standard error of $\mathrm{logit}(\hat{\pi}(2))$ is

$$\begin{aligned}\widehat{\mathrm{Var}}[\mathrm{logit}(\hat{\pi}(2))] &= \widehat{\mathrm{Var}}(\hat{\beta}_0) + 4\widehat{\mathrm{Var}}(\hat{\beta}_1) + 16\widehat{\mathrm{Var}}(\hat{\beta}_2) \\ &+ 4\widehat{\mathrm{Cov}}(\hat{\beta}_0,\hat{\beta}_1) + 8\widehat{\mathrm{Cov}}(\hat{\beta}_0,\hat{\beta}_2) + 16\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2) \\ &= 1\cdot 0.01 + 4\cdot 0.02 + 16\cdot 0.02 - 4\cdot 0.01 - 8\cdot 0.01 - 16\cdot 0.01 \\ &= 0.13.\end{aligned}$$

Using the normal quantile $z_{0.025} = \sqrt{\chi_1^2(0.05)} = \sqrt{3.8415} = 1.96$, this gives an approximate 95% confidence interval

$$(-2.0 - 1.96\sqrt{0.13}, -2.0 + 1.96\sqrt{0.13}) = (-2.7067, -1.2933) \tag{1}$$

for $\mathrm{logit}[\pi(2)]$. The corresponding approximate 95% confidence interval for $\pi(2)$ is obtained by transforming the left and right end points of (1) by the inverse of the logit transformation, i.e.

$$\left(\frac{\exp(-2.7067)}{1+\exp(-2.7067)}, \frac{\exp(-1.2933)}{1+\exp(-1.2933)}\right) = (0.063, 0.215).$$

b. The odds of dying, for a person with blood concentration x mmHg of the gas, is

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\beta_0 + x\beta_1 + x^2\beta_2).$$

Taking the ratio of this expression for $x = 2$ and $x = 1$ we obtain the odds ratio

$$\mathrm{OR} = \frac{\pi(2)/(1-\pi(2))}{\pi(1)/(1-\pi(1))} = \frac{\exp(\beta_0 + 2\beta_1 + 4\beta_2)}{\exp(\beta_0 + \beta_1 + \beta_2)} = \exp(\beta_1 + 3\beta_2)$$

of dying between two persons with concentrations 2 and 1 mmHg. The sought for log odds ratio is therefore

$$\log \mathrm{OR} = \beta_1 + 3\beta_2. \tag{2}$$

c. We will first compute an approximate 95% confidence interval for the log odds ratio in (2). We estimate this quantity by

$$\log \widehat{\mathrm{OR}} = \hat{\beta}_1 + 3\hat{\beta}_2 = 1.0 + 3 \cdot 0.5 = 2.5,$$

and then find the variance

$$\mathrm{Var}(\log \widehat{\mathrm{OR}}) = \mathrm{Var}(\hat{\beta}_1) + 9\mathrm{Var}(\hat{\beta}_2) + 6\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) \tag{3}$$

of this estimate. Plugging in the estimated variances and covariances into the last expression, we obtain the squared standard error

$$\begin{aligned} \widehat{\mathrm{Var}}(\log \widehat{\mathrm{OR}}) &= \widehat{\mathrm{Var}}(\hat{\beta}_1) + 9\widehat{\mathrm{Var}}(\hat{\beta}_2) + 6\widehat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2) \\ &= 0.02 + 9 \cdot 0.02 - 6 \cdot 0.01 \\ &= 0.14 \end{aligned}$$

of $\log \widehat{\mathrm{OR}}$. This gives an approximate 95% confidence interval

$$(2.5 - 1.96\sqrt{0.14}, 2.5 + 1.96\sqrt{0.14}) = (1.7666, 3.2334)$$

for $\log \mathrm{OR}$, and a corresponding approximate 95% confidence interval

$$(\exp(1.7666), \exp(3.2334)) = (5.85, 25.36)$$

for OR.

Problem 2

a. Let n_{ik} be the number of observations in cell (i,k), which is an observation of the random variable N_{ik}. The joint distribution of all cell counts is multinomial

$$\boldsymbol{N} = (N_{ik})_{i,k=0}^{2} \sim \mathrm{Mult}(500, (\pi_{ik})_{i,k=0}^{2}).$$

Since the cell probabilities sum to 1 ($\sum_{i,k=0}^{2} \pi_{ik} = 1$), there are 8 free parameters, for instance

$$\boldsymbol{\theta} = (\pi_{00}, \pi_{01}, \pi_{02}, \pi_{10}, \pi_{11}, \pi_{12}, \pi_{20}, \pi_{21}).$$

This gives a likelihood

$$\begin{aligned} l(\boldsymbol{\theta}) &= \frac{500!}{\prod_{i,k=0}^{2} n_{ik}!} \prod_{(i,k)\neq(2,2)} \pi_{ik}^{n_{ik}} \cdot \left(1 - \sum_{(i,k)\neq(2,2)} \pi_{ik}\right)^{n_{22}} \\ &= \frac{500!}{30!60!28!47!120!89!16!60!50!} \pi_{00}^{30}\pi_{01}^{60}\pi_{02}^{28}\pi_{10}^{47}\pi_{11}^{120}\pi_{12}^{89}\pi_{20}^{16}\pi_{21}^{60}\left(1 - \sum_{(i,k)\neq(2,2)} \pi_{ik}\right)^{50}. \end{aligned}$$

b. The expected cell counts equal $\mu_{ik} = E(N_{ik}) = n_{++}\pi_{i+}\pi_{+k}$ under H_0, which we estimate by

$$\hat{\mu}_{ik} = n_{++} \cdot \frac{n_{i+}}{n_{++}} \cdot \frac{n_{+k}}{n_{++}} = \frac{n_{i+}n_{+k}}{n_{++}},$$

for instance

$$\hat{\mu}_{00} = \frac{118 \cdot 93}{500} = 21.95$$

for cell $(0,0)$. Continuing in this way for the other 8 cells we obtain the following values of $\hat{\mu}_{ik}$:

Values of $\hat{\mu}_{ik}$ under H_0:				
	k			
i	0	1	2	Sum
0	21.95	56.64	39.41	118
1	47.62	122.88	85.50	256
2	23.44	60.48	42.08	126
Sum	93	240	167	500

This gives a X^2-statistic

$$X^2 = \sum_{i,k=0}^{2} \frac{(n_{ik} - \hat{\mu}_{ik})^2}{\hat{\mu}_{ik}} = \frac{(30 - 21.95)^2}{21.95} + \ldots \frac{(50 - 42.08)^2}{42.08} = 10.53.$$

Since the saturated model has $8 - 4 = 4$ more parameters than the independence model, and $X^2 > \chi_4^2(0.05) = 9.49$, we reject H_0 at level 5%.

c. In order to estimate p, we notice that there are 1000 copies of gene I, two for each individual. Under H_0' we have that each copy of gene I is either A with probability p, or a with probability $1 - p$, independently between gene copies. Since there are $N_{1+} + 2N_{2+}$ gene copies that equal A it follows that $N_{1+} + 2N_{2+} \sim \text{Bin}(1000, p)$. Therefore, the maximum likelihood estimate of p is

$$\hat{p} = \frac{n_{1+} + 2n_{2+}}{1000} = \frac{256 + 2 \cdot 126}{1000} = 0.508. \tag{4}$$

In a similar way we find a maximum likelihood estimate

$$\hat{q} = \frac{n_{+1} + 2n_{+2}}{1000} = \frac{240 + 2 \cdot 167}{1000} = 0.574 \tag{5}$$

of q. Since the expected cell counts under H_0' are

$$\mu_{ik} = 500 \cdot \binom{2}{i}(1 - p)^{2-i}p^i \cdot \binom{2}{k}(1 - q)^{2-k}q^k, \tag{6}$$

we simply plug (4) and (5) into (6), and find that

$$\hat{\mu}_{ik} = 500 \cdot \binom{2}{i}(1 - \hat{p})^{2-i}\hat{p}^i \cdot \binom{2}{k}(1 - \hat{q})^{2-k}\hat{q}^k,$$

for all $i, k \in \{0, 1, 2\}$. For instance, cell $(0,0)$ has

$$\hat{\mu}_{00} = 500(1 - \hat{p})^2(1 - \hat{q})^2 = 500(1 - 0.508)^2(1 - 0.574)^2 = 21.96.$$

Continuing in this way for the other 8 cells, we obtain the following values of $\hat{\mu}_{ik}$:

Values of $\hat{\mu}_{ik}$ under H_0':				
	k			
i	0	1	2	Sum
0	21.96	59.19	39.88	121.03
1	45.36	122.23	82.35	249.94
2	23.42	63.10	42.51	129.03
Sum	90.74	244.52	164.74	500

This gives a X^2-statistic that equals

$$\sum_{i,k=0}^{2} \frac{(n_{ik} - \hat{\mu}_{ik})^2}{\hat{\mu}_{ik}} = \frac{(30 - 21.96)^2}{21.96} + \dots \frac{(50 - 42.51)^2}{42.51} = 10.95.$$

There are only 2 parameters p and q under H_0', and therefore the saturated model has $8 - 2 = 6$ more parameters. Since $X^2 < \chi^2_6(0.05) = 12.59$, we do not reject H_0' at level 5%.

Problem 3

a. For the loglinear model $M = (GI, GL, GS, IL, IS, LS)$, we have that

$$\mu_{gils} = \exp(\lambda + \lambda_g^G + \lambda_i^I + \lambda_l^L + \lambda_s^S + \lambda_{gi}^{GI} + \lambda_{gl}^{GL} + \lambda_{gs}^{GS} + \lambda_{il}^{IL} + \lambda_{is}^{IS} + \lambda_{ls}^{LS}),$$

for all cells $g, i, l, s \in \{0, 1\}$. If $g = i = l = s = 0$ are chosen as baseline levels, then all loglinear parameters equal 0 for which at least index is 0. This gives a parameter vector with the remaining nonzero loglinear parameters

$$\boldsymbol{\beta} = (\lambda, \lambda_1^G, \lambda_1^I, \lambda_1^L, \lambda_1^S, \lambda_{11}^{GI}, \lambda_{11}^{GL}, \lambda_{11}^{GS}, \lambda_{11}^{IL}, \lambda_{11}^{IS}, \lambda_{11}^{LS}).$$

The number of parameters is $p(M) = 11$.

b. All of the listed models in the tables are balanced, and all four categorical variables are binary. Therefore, each model has 1 baseline parameter, 4 main effect parameters (1 per main effect), $1 = (2-1) \cdot (2-1)$ parameter per second order association, and $1 = (2-1) \cdot (2-1) \cdot (2-1)$ parameter per third order association. Adding the number of baseline, main effect, second order, and third order association parameters, we find the total number of parameters

$$\begin{aligned}
p(G, I, L, S) &= 1 + 4 + 0 + 0 = 5, \\
p(GI, GL, GS, IL, IS, LS) &= 1 + 4 + 6 + 0 = 11, \\
p(GIL, GS, IS, LS) &= 1 + 4 + 6 + 1 = 12, \\
p(GIS, GL, IL, LS) &= 1 + 4 + 6 + 1 = 12, \\
p(GLS, GI, IL, IS) &= 1 + 4 + 6 + 1 = 12, \\
p(ILS, GI, GL, GS) &= 1 + 4 + 6 + 1 = 12, \\
p(GIL, GIS, GLS, ILS) &= 1 + 4 + 6 + 4 = 15
\end{aligned}$$

of all models.

c. Let $M_1 = (GILS)$ refer to the saturated model, with $2^4 = 16$ parameters. Akaike's Information Criterion of model M is

$$\begin{aligned}\mathrm{AIC}(M) &= -2L(M) + 2p(M) \\ &= -2[L(M) - L(M_1)] + 2p(M) - 2L(M_1) \\ &= G^2(M) + 2p(M) - 2L(M_1),\end{aligned}$$

where $L(M)$ and $G^2(M)$ is the log likelihood and deviance of model M. We select the best model, according to the AIC-criterion, by minimizing $\mathrm{AIC}(M)$, which is equivalent to minimizing $G^2(M) + 2p(M)$. We found the number of parameters $p(M)$ of all models in b). This makes it possible to fill in the second column of the given table, and then add a third column:

Model M	$G^2(M)$	$p(M)$	$G^2(M) + 2p(M)$
(G, I, L, S)	2792.8	5	2802.8
(GI, GL, GS, IL, IS, LS)	23.4	11	45.4
(GIL, GS, IS, LS)	18.6	12	42.6
(GIS, GL, IL, LS)	22.8	12	46.8
(GLS, GI, IL, IS)	7.5	12	31.5
(ILS, GI, GL, GS)	20.6	12	44.6
(GIL, GIS, GLS, ILS)	1.33	15	31.3

Since $M = (GIL, GIS, GLS, ILS)$ minimizes $G^2(M) + 2p(M)$, this is the model chosen by the AIC-criterion.

d. In the first step of backward elimination (BE), the largest model among those listed in the table, $M_1' = (GIL, GIS, GLS, ILS)$, is tested against each one of the four models M for which three second order associations have been removed from M_1', by means of a likelihood ratio test. The log likelihood ratios of these four tests are

$$\begin{aligned}G^2(M|M_1') &= -2[L(M) - L(M_1')] \\ &= G^2(M) - G^2(M_1') \\ &= \begin{cases} 18.6 - 1.33 = 17.27, & M = (GIL, GS, IS, LS), \\ 22.8 - 1.33 = 21.47, & M = (GIS, GL, IL, LS), \\ 7.5 - 1.33 = 6.17, & M = (GLS, GI, IL, IS), \\ 20.6 - 1.33 = 19.27, & M = (ILS, GI, GL, GS), \end{cases}\end{aligned}$$

respectively. In all of these tests, the null hypothesis

$$H_0 : \text{model } M \text{ holds}$$

is rejected if $G^2(M|M_1') > \chi^2_3(0.05) = 7.81$, where $3 = 15 - 12$ is the number of parameters being tested. We find that H_0 is *not* rejected for model $M = (GLS, GI, IL, IS)$, whereas H_0 is rejected for the other three models with one third order association. Therefore (GLS, GI, IL, IS) is selected in the first step of the BE-scheme. In the second step of the BE-scheme we test

$$H_0 : M_0 = (GI, GL, GS, IL, IS, LS)$$

against the alternative that $M = (GLS, GI, IL, IS)$ holds but not M_0. This gives a log likelihood ratio

$$\begin{aligned} G^2(M_0|M) &= G^2(M_0) - G^2(M) \\ &= 23.4 - 7.5 \\ &= 15.9 \\ &> \chi_1^2(0.05) \\ &= 3.84, \end{aligned}$$

since $1 = 12 - 11$ parameter is tested. The null hypothesis is rejected in this second step, and therefore the BE-scheme stops, with (GLS, GI, IL, IS) as the chosen model.

Problem 4

a. Let $\pi_{gils} = \mu_{gils}/\mu_{++++}$ be the probability of cell (g, i, l, s) for multinomial sampling when we condition on the total number of observations of the Poisson model (GI, GL, GS, IL, IS, LS). Regarding I as the outcome variable and G, L, S as predictor variables of this mutinomial model, we find that $I|G, L, S$ is an ANOVA type logistic regression model, since

$$\begin{aligned} &\text{logit} P(I = 1|G = g, L = l, S = s) \\ &\quad = \log[P(I = 1|G = g, L = l, S = s)/P(I = 0|G = g, L = l, S = s)] \\ &\quad = \log[(\pi_{g1ls}/\pi_{g+ls})/(\pi_{g0ls}/\pi_{g+ls})] \\ &\quad = \log(\pi_{g1ls}/\pi_{g0ls}) \\ &\quad = \log(\mu_{g1ls}/\mu_{g0ls}) \\ &\quad = \log(\mu_{g1ls}) - \log(\mu_{g0ls}) \\ &\quad = \lambda + \lambda_g^G + \lambda_1^I + \lambda_l^L + \lambda_s^S + \lambda_{g1}^{GI} + \lambda_{gl}^{GL} + \lambda_{gs}^{GS} + \lambda_{1l}^{IL} + \lambda_{1s}^{IS} + \lambda_{ls}^{LS} \\ &\quad -(\lambda + \lambda_g^G + \lambda_0^I + \lambda_l^L + \lambda_s^S + \lambda_{g0}^{GI} + \lambda_{gl}^{GL} + \lambda_{gs}^{GS} + \lambda_{0l}^{IL} + \lambda_{0s}^{IS} + \lambda_{ls}^{LS}) \\ &\quad = \alpha + \beta_g^G + \beta_l^L + \beta_s^S, \end{aligned} \tag{7}$$

with

$$\begin{aligned} \alpha &= \lambda_1^I - \lambda_0^I = \lambda_1^I, \\ \beta_g^G &= \lambda_{g1}^{GI} - \lambda_{g0}^{GI} = \lambda_{g1}^{GI}, \\ \beta_l^L &= \lambda_{1l}^{IL} - \lambda_{0l}^{IL} = \lambda_{1l}^{IL}, \\ \beta_s^S &= \lambda_{1s}^{IS} - \lambda_{0s}^{IS} = \lambda_{1s}^{IS}. \end{aligned}$$

In the last step we assumed that $g = i = l = s = 0$ are baseline levels, putting to zero all loglinear parameters with at least one 0 index. Then all effect parameters $\beta_0^G = \beta_0^L = \beta_0^S = 0$ vanish, and the remaining four nonzero parameters of the logistic regression model, are

$$\boldsymbol{\beta} = (\alpha, \beta_1^G, \beta_1^L, \beta_1^S).$$

b. The conditional odds ratio of injury between those that use safety belt and those that do not, conditional on gender and location, is

$$\theta_{IS(gl)} = \frac{P(I = 1|S = 1, G = g, L = l)/P(I = 0|S = 1, G = g, L = l)}{P(I = 1|S = 0, G = g, L = l)/P(I = 0|S = 0, G = g, L = l)}. \tag{8}$$

It follows from (7) that

$$\begin{aligned}
\log \theta_{IS(gl)} &= \text{logit} P(I=1|S=1,G=g,L=l) - \text{logit} P(I=1|S=0,G=g,L=l) \\
&= \alpha + \beta_g^G + \beta_l^L + \beta_1^S - (\alpha + \beta_g^G + \beta_l^L + \beta_0^S) \\
&= \beta_1^S - \beta_0^S \\
&= \beta_1^S \\
&= \lambda_{11}^{IS}
\end{aligned}$$

when $i = s = 0$ are chosen as baseline levels of injury and safety belt use. Equivalently,

$$\theta_{IS(gl)} = \exp(\lambda_{11}^{IS}). \tag{9}$$

c. There is homogeneous association between injury I and safety belt use S if the conditional odds ratio $\theta_{IS(gl)}$ does not depend on the levels g and l of gender G and location L. It follows from (9) that model (GI, GL, GS, IL, IS, LS) has homogeneous association, since the right hand side of this equation does not depend on g or l. Similarly, one shows that all loglinear models M for which I and S are not involved in any third order association, have homogeneous association between I and S. Hence, among the loglinear models listed in the table of Problem 3, the ones with homogeneous association between injury and safety belt use, are (G, I, L, S), (GI, GL, GS, IL, IS, LS), (GIL, GS, IS, LS), and (GLS, GI, IL, IS).

d. For the loglinear model $M_0 = (IS, IGL)$ we have that S and G, L are conditionally independent given I. In conjunction with Bayes' Theorem, this gives

$$\begin{aligned}
P(I=i|S=s,G=g,L=l) &= \frac{P(S=s|I=i,G=g,L=l)P(I=i|G=g,L=l)}{P(S=s|G=g,L=l)} \\
&= \frac{P(S=s|I=i)P(I=i|G=g,L=l)}{P(S=s|G=g,L=l)}.
\end{aligned} \tag{10}$$

Insertion of (10) into the definition (8) of the conditional odds ratio gives

$$\theta_{IS(gl)} = \frac{P(S=1|I=1)P(S=0|I=0)}{P(S=0|I=1)P(S=1|I=0)}, \tag{11}$$

since all terms $P(I=i|G=g,L=l)$ and $P(S=s|G=g,L=l)$ appear twice, in the numerator and denominator, and hence cancel out. A second application of Bayes' Theorem gives $P(S=s|I=i) = P(I=i|S=s)P(S=s)/P(I=i)$. Inserting this expression into (11), we find that

$$\theta_{IS(gl)} = \frac{P(I=1|S=1)P(I=0|S=0)}{P(I=0|S=1)P(I=1|S=0)} = \theta_{IS},$$

since all terms $P(I=i)$ and $P(S=s)$ appear twice, in the numerator and denominator, and hence cancel out. From this it follows that the conditional odds ratio $\theta_{IS(gl)}$ of having an injury between those that use seat belt and those that don't, for model $M_0 = (IS, IGL)$, equals the corresponding marginal odds ratio θ_{IS}. There is also an alternative way of showing this (without using Bayes' Theorem). We start by noticing that

$$\mu_{gils} = A_{is} B_{gil} \tag{12}$$

for model M_0, with $A_{is} = \exp(\lambda + \lambda_i^I + \lambda_s^s + \lambda_{is}^{IS})$ and $B_{gil} = \exp(\lambda_g^G + \lambda_l^L + \lambda_{gi}^{GI} + \lambda_{il}^{IL} + \lambda_{gl}^{GL} + \lambda_{gil}^{GIL})$. From the calculations in (7) and (12) we find that the conditional odds ratio between injury and seat belt use can be expressed as

$$\theta_{IS(gl)} = \frac{\mu_{g0l0}\mu_{g1l1}}{\mu_{g0l1}\mu_{g1l0}} = \frac{A_{00}A_{11}}{A_{01}A_{10}},$$

since all the B_{gil}-terms cancel out. Similarly, we find that the marginal odds ratio between injury and seat belt use equals

$$\theta_{IS} = \frac{\mu_{+0+0}\mu_{+1+1}}{\mu_{+0+1}\mu_{+1+0}} = \frac{A_{00}A_{11}}{A_{01}A_{10}},$$

since $\mu_{+i+s} = A_{is}B_{+i+}$, and all the B_{+i+}-terms cancel out. From the last two displayed equations, it follows that $\theta_{IS(gl)} = \theta_{IS}$.

We finally estimate the marginal odds ratio from the data set of Problem 3, as

$$\begin{aligned}
\hat{\theta}_{IS} &= \tfrac{n_{+1+1}n_{+0+0}}{n_{+1+0}n_{+0+1}} \\
&= \tfrac{(759+757+380+513)\cdot(7287+3246+10381+6123)}{(996+973+812+1084)\cdot(11587+6134+10969+6693)} \\
&= \tfrac{2409\cdot 27037}{3865\cdot 35383} \\
&= 0.4763,
\end{aligned}$$

which is slightly higher than the estimated conditional odds ratio $\hat{\theta}_{IS(gl)} = 0.44$ between injury and seat belt use for model $M = (GI, GL, GS, IL, IS, LS)$.

Problem 5

a. A 2×3 contingency table has three 2×2 subtables:

$$\begin{aligned}
\mathrm{I} &= \{11, 12, 21, 22\}, \\
\mathrm{II} &= \{12, 13, 22, 23\}, \\
\mathrm{III} &= \{11, 13, 21, 23\}.
\end{aligned}$$

The corresponding estimated odds ratios for (a higher degree of) job-satisfaction, between middle-aged and young people, are

$$\begin{aligned}
\hat{\theta}_{\mathrm{I}} &= (n_{11}n_{22})/(n_{12}n_{21}) = (34 \cdot 174)/(53 \cdot 80) = 1.395, \\
\hat{\theta}_{\mathrm{II}} &= (n_{12}n_{23})/(n_{13}n_{22}) = (53 \cdot 304)/(88 \cdot 174) = 1.052, \\
\hat{\theta}_{\mathrm{III}} &= (n_{11}n_{23})/(n_{13}n_{21}) = (34 \cdot 304)/(88 \cdot 80) = 1.468,
\end{aligned}$$

where n_{ij} is the number of observations with $X = i$ and $Y = j$.

b. Since subtables I and II have adjacent columns, their odds ratios are local, whereas the odds ratio of subtable III is not local. We have that

$$\hat{\theta}_{\mathrm{III}} = \frac{n_{11}n_{23}}{n_{13}n_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \cdot \frac{n_{12}n_{23}}{n_{13}n_{22}} = \hat{\theta}_{\mathrm{I}} \cdot \hat{\theta}_{\mathrm{II}}. \tag{13}$$

c. The number of concordant and discordant pairs are

$$\begin{aligned} C &= n_{11}(n_{22}+n_{23}) + n_{12}n_{23} = 34(174+304) + 53 \cdot 304 = 32364, \\ D &= n_{12}n_{21} + n_{13}(n_{21}+n_{22}) = 53 \cdot 80 + 88(80+174) = 26592, \end{aligned}$$

and consequently

$$\hat{\gamma} = \frac{C-D}{C+D} = \frac{32364 - 26592}{32364 + 26592} = 0.0979. \tag{14}$$

d. Notice first that

$$\frac{C}{D} = \frac{n_{11}n_{22} + n_{12}n_{23} + n_{11}n_{23}}{n_{12}n_{21} + n_{13}n_{22} + n_{13}n_{21}} = w_{\mathrm{I}}\hat{\theta}_{\mathrm{I}} + w_{\mathrm{II}}\hat{\theta}_{\mathrm{II}} + w_{\mathrm{III}}\hat{\theta}_{\mathrm{III}}, \tag{15}$$

where the weights

$$\begin{aligned} w_{\mathrm{I}} &= n_{12}n_{21}/(n_{12}n_{21} + n_{13}n_{22} + n_{13}n_{21}), \\ w_{\mathrm{II}} &= n_{13}n_{22}/(n_{12}n_{21} + n_{13}n_{22} + n_{13}n_{21}), \\ w_{\mathrm{III}} &= n_{13}n_{21}/(n_{12}n_{21} + n_{13}n_{22} + n_{13}n_{21}) \end{aligned}$$

are non-negative and sum to 1. If the two local odds ratios are larger than 1 ($\hat{\theta}_{\mathrm{I}} > 1$, $\hat{\theta}_{\mathrm{II}} > 1$), it follows from (13) that the third non-local odds ratio is larger than 1 as well ($\hat{\theta}_{\mathrm{III}} = \hat{\theta}_{\mathrm{I}} \cdot \hat{\theta}_{\mathrm{II}} > 1$). Since C/D is a weighted average (15) of the three odds ratios, all of which are greater than 1, we find that $C/D > 1$. Dividing the numerator and denominator of (14) by D we finally obtain

$$\hat{\gamma} = \frac{C/D - 1}{C/D + 1} > 0.$$