

STA 138: Fall 2006

Homework 4 Solutions

Problems: 3.30, 3.32, 8.1, 8.5(a,c), 8.7, 8.19

3.30 For testing $H_0 : \pi_1 = \pi_2$ using independent binomial variates y_1 and y_2 with n_1 and n_2 trials, the score statistic is

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1 - \hat{\pi})(1/n_1 + 1/n_2)}},$$

where $\hat{\pi} = (y_1 + y_2)/(n_1 + n_2)$ is the pooled estimate of $\pi_1 = \pi_2$ under H_0 . Show that $z^2 = X^2$.

Solution:

Begin with X^2 and show that it can be written as the score statistic z^2 .

$$\begin{aligned} X^2 &= \frac{(y_1 - n_1\hat{\pi})^2}{n_1\hat{\pi}} + \frac{(n_1 - y_1 - n_1(1 - \hat{\pi}))^2}{n_1(1 - \hat{\pi})} + \frac{(y_2 - n_2\hat{\pi})^2}{n_2\hat{\pi}} + \frac{(n_2 - y_2 - n_2(1 - \hat{\pi}))^2}{n_2(1 - \hat{\pi})} \\ &= \frac{(y_1 - n_1\hat{\pi})^2}{n_1\hat{\pi}} + \frac{(n_1\hat{\pi} - y_1)^2}{n_1(1 - \hat{\pi})} + \frac{(y_2 - n_2\hat{\pi})^2}{n_2\hat{\pi}} + \frac{(n_2\hat{\pi} - y_2)^2}{n_2(1 - \hat{\pi})} \\ &= \frac{1}{\hat{\pi}(1 - \hat{\pi})} \left(\frac{(y_1 - n_1\hat{\pi})^2}{n_1} + \frac{(y_2 - n_2\hat{\pi})^2}{n_2} \right) \\ &= \frac{1}{\hat{\pi}(1 - \hat{\pi})} \left[n_1 \left(\frac{y_1}{n_1} - \frac{y_1 + y_2}{n_1 + n_2} \right)^2 + n_2 \left(\frac{y_2}{n_2} - \frac{y_1 + y_2}{n_1 + n_2} \right)^2 \right] \\ &= \frac{1}{\hat{\pi}(1 - \hat{\pi})} \left[n_1 \left(\frac{y_1 n_2 - y_2 n_1}{n_1(n_1 + n_2)} \right)^2 + n_2 \left(\frac{y_2 n_1 - y_1 n_2}{n_2(n_1 + n_2)} \right)^2 \right] \\ &= \frac{1}{\hat{\pi}(1 - \hat{\pi})} \frac{(y_1 n_2 - y_2 n_1)^2}{(n_1 + n_2)^2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\hat{\pi}(1-\hat{\pi})} \frac{(n_1\hat{\pi}_1n_2 - n_2\hat{\pi}_2n_1)^2}{(n_1+n_2)^2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \\
&= \frac{1}{\hat{\pi}(1-\hat{\pi})} (\hat{\pi}_1 - \hat{\pi}_2)^2 \frac{(n_1n_2)^2}{(n_1+n_2)^2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \\
&= \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}(1-\hat{\pi})} \frac{n_1n_2}{n_1+n_2} \\
&= \frac{(\hat{\pi}_1 - \hat{\pi}_2)^2}{\hat{\pi}(1-\hat{\pi}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\
&= z^2
\end{aligned}$$

3.32 For a 2×2 table, show that:

- The four Pearson residuals may take different values.
- All four standardized Pearson residuals have the same absolute value. (This is sensible, since $df = 1$.)
- The square of each standardized Pearson residual equals X^2 . [Note: $X^2 = n(n_{11}n_{22} - n_{12}n_{21})^2 / (n_{1+}n_{2+}n_{+1}n_{+2})$ for 2×2 tables.]

Solution:

- Pearson residuals $e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$. Consider the following 2×2 table:

		Total
	1 3	4
	2 4	6
Total	3 7	10

Then, $e_{11} = -0.8$, $e_{12} = 0.2$, $e_{21} = 0.5$, $e_{22} = -0.08$. So, they take different values.

- For a 2×2 table, the standardized Pearson residuals are $ste_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}$.

Then,

$$\begin{aligned}
ste_{11}^2 &= \frac{(n_{11} - \frac{n_{1+}n_{+1}}{n})^2}{\frac{n_{1+}n_{+1}}{n} (1 - \frac{n_{1+}}{n})(1 - \frac{n_{+1}}{n})} \\
&= \frac{n(nn_{11} - n_{1+}n_{+1})^2}{n_{1+}n_{+1}(n - n_{1+})(n - n_{+1})} \\
&= \frac{n(nn_{11} - n_{1+}n_{+1})^2}{n_{1+}n_{+1}n_{2+}n_{+2}}
\end{aligned}$$

Similarly, $ste_{12}^2 = \frac{n(nn_{12} - n_{1+}n_{+2})^2}{n_{1+}n_{+1}n_{2+}n_{+2}}$. Note that $nn_{12} - n_{1+}n_{+2} = n(n_{1+} - n_{11}) - n_{1+}(n - n_{+1}) = -nn_{11} + n_{1+}n_{+1}$. So, $ste_{11}^2 = ste_{12}^2$. Similarly, all four squared standardized residuals are equal: $ste_{11}^2 = ste_{12}^2 = ste_{21}^2 = ste_{22}^2$. Thus, all four standardized Pearson residuals have the same absolute value.

c. Note that

$$\begin{aligned} nn_{11} - n_{1+}n_{+1} &= (n_{11} + n_{12} + n_{21} + n_{22})n_{11} - (n_{11} + n_{12})(n_{11} + n_{21}) \\ &= n_{11}n_{22} - n_{12}n_{21} \end{aligned}$$

So, from (b) we have

$$\begin{aligned} ste_{ij}^2 &= ste_{11}^2 \\ &= \frac{n(nn_{11} - n_{1+}n_{+1})^2}{n_{1+}n_{+1}n_{2+}n_{+2}} \\ &= \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{+1}n_{2+}n_{+2}} \\ &= X^2. \end{aligned}$$

8.1 The 1988 General Social Survey compiled by the National Opinion Research Center asked: “Do you support or oppose the following measures to deal with AIDS? (1) Have the government pay all of the health care costs of AIDS patients; (2) Develop a government information program to promote safe sex practices, such as the use of condoms.” Table 8.16 summarizes opinions about health care costs (H) and the information program (I), classified also by the respondent’s gender (G).

- a. Fit loglinear models (GH,GI), (GH,HI), (GI,HI), and (GH,GI,HI). Show that models that lack the HI term fit poorly.
- b. For model (GH,GI,HI), show that 95% Wald confidence intervals equal (0.55,1.10) for the GH conditional odds ratio and (0.99,2.55) for the GI conditional odds ratio. Interpret. Is it plausible that gender has no effect on opinion for these issues?

Solution:

- a. The likelihood ratio chi-squared test results are as follows:

Model	G^2	df	P-value
(GH,GI)	11.67	2	0.0029
(GH,HI)	4.13	2	0.13
(GI,HI)	2.38	2	0.30
(GH,GI,HI)	0.30	1	0.58

The model (GH,GI) without the HI term has the worst fit (smallest P-value). Also, the residuals are generally larger than those from the other model. All of the other models fit well.

- b. The estimated GH conditional log odds ratio is $\log(\theta_{GH}) = -0.2516$, with standard error $se(\log(\theta_{GH})) = 0.1749$. Hence, the 95% Wald confidence interval for the GH conditional log odds ratio is $-0.2516 \pm 1.96 \times 0.1749 = [-0.594, 0.0912]$. So, the 95% Wald confidence interval for the GH conditional odds ratio is $[e^{-0.594}, e^{0.0912}] = [0.55, 1.10]$. Since this interval covers 1, we can not reject the null hypothesis of no interaction between gender and health care opinion. Similarly, the estimate of the GI conditional log odds ratio is $\log(\theta_{GI}) = 0.4636$, with standard error 0.2406. The 95% Wald confidence interval for the GI conditional log odds ratio is $0.4636 \pm 1.96 \times 0.2406 = [-0.008, 0.935]$. So, 95% Wald confidence interval for the GI conditional odds ratio is $[0.99, 2.55]$. Again, the interval covers 1, indicating that we can not reject the null hypothesis of no interaction between gender and information opinion.

8.5 Table 8.18 refers to automobile accident records in Florida in 1988.

- a. Find a loglinear model that describes the data well. Interpret associations.
 c. Since n is large, goodness-of-fit statistics are large unless the model fits very well. Calculate the dissimilarity index for the model in part (1), and interpret.

Solution:

- a. The model (SE,SI,EI) with all three interaction terms has $G^2 = 2.85$ with P-value 0.091, suggesting that this model fits the data well. There are associations between all variables.

For the model (SI,EI), $G^2 = 7134$, suggesting a conditional association between seat belt use and ejection when injury is fixed. The conditional odds ratio for S and E is 0.091, so seat belt users are less likely to be ejected at either level of injury. The model (SE,EI) has $G^2 = 1146$, indicating that there is a conditional association between seat belt use and injury when ejection is fixed. The conditional odds ratio for S and I is 5.57, so seat belt users are more likely to have a

nonfatal injury than a fatal injury at either level of ejection. The model (ES,SI) has $G^2 = 1680$, suggesting a conditional association between ejection and injury when seat belt use is fixed. The conditional odds ratio between E and I is 0.061, so being ejected is less likely with nonfatal injuries at either level of seat belt use.

- c. The dissimilarity index is $\hat{\Delta} = \sum_i |n_i - \hat{\mu}_i|/2n = 0.000048$. Since this is close to zero, the model fits well. The predicted values are close to the observed values.

8.7 Refer to the loglinear models for Table 8.8.

- Explain why the fitted odds ratios in Table 8.10 for model (GI,GL,GS,IL,IS,LS) suggest that the most likely accident case for injury is females not wearing seat-belts in rural locations.
- Fit model (GLS,GI,IL,IS). Using model parameter estimates, show that the fitted IS conditional odds ratio equals 0.44. Show that for each injury level, the estimated conditional LS odds ratio is 1.17 for (G=female) and 1.03 for (G=male). How can you get these using the model parameter estimates?

Solution:

- Injury has estimated conditional odds ratios of 0.58 with gender, 2.13 with location, and 0.44 with seat belt use. The variables are coded so that category 1 is “No” for injury, “female” for gender, “urban” for location, and “No” for seat belt. Thus, the conditional odds of no injury for females are 0.58 times the odds for males. The conditional odds of no injury for urban locations are 2.13 times the odds for rural locations. The conditional odds of no injury for no seat belt use are 0.44 times the odds for seat belt use. Since there are no three-way interactions in this model, the overall highest odds for no injury are for males in urban locations wearing seat belts. Thus, the most likely category for injury is females not wearing seat belts in rural locations.
- First, consider computing conditional odds ratios using the fitted values for each category (Table 8.8). For the IS conditional odds ratio, we can work with any of the 2×2 tables, since they all give the same answer. For example, females in urban areas, the odds ratio is $(7273.2 \times 713.4)/(11632.6 \times 1009.8) = 0.44$. Since there is a GLS three-way interaction, the LS conditional odds ratio can differ between males and females. For females, we can use either the injury or the non-injury predicted values. For example, using the non-injury values, the LS conditional odds ratio is $(7273.2 \times 6093.5)/(11632.6 \times 3254.7) = 1.17$. Similarly, using non-injury values for males, the LS conditional odds ratio is $(10358.9 \times 6697.6)/(10959.2 \times 6150.2) = 1.03$.
Alternatively, these conditional odds ratios could be computed directly from model parameter estimates. The estimated conditional IS odds ratio is simply the exponential of the parameter estimate for the IS term: $0.44 = e^{-0.82}$. The estimated conditional LS odds ratio in females is the exponential of the parameter

estimate for the LS term: $1.17 = e^{0.157}$. In males, we must also add the GLS parameter estimate: $1.03 = e^{0.157-0.127}$.

8.19 For three categorical variables X, Y, Z:

- When Y is jointly independent of X and Z, show that X and Y are conditionally independent, given Z.
- Prove that mutual independence of X, Y, and Z implies that X and Y are both marginally and conditionally independent.
- When X is independent of Y and Y is independent of Z, does it follow that X is independent of Z? Explain.
- When any pair of variables is conditionally independent, explain why there is no three-factor interaction.

Solution:

Let $\pi_{ijk} = P(X = i, Y = j, Z = k)$ and $\pi_{ij|k} = P(X = i, Y = j|Z = k)$ for all i, j, k .

- If Y is jointly independent of X and Z, then $\pi_{ijk} = \pi_{i+k}\pi_{j+}$ for all i, j, k . And $\pi_{+j|k} = \pi_{j+}$. Now show that X and Y are conditionally independent, given Z:

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{++k}} = \frac{\pi_{i+k}\pi_{j+}}{\pi_{++k}} = \frac{\pi_{i+k}}{\pi_{++k}}\pi_{j+} = \pi_{i+|k}\pi_{+j|k}.$$

- Mutual independence of X, Y, and Z means $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ for all i, j, k . Also, $\pi_{i+|k} = \pi_{i++}$ and $\pi_{+j|k} = \pi_{+j+}$. First, show that this implies marginal independence of X and Y:

$$\pi_{ij+} = \sum_k \pi_{ijk} = \sum_k \pi_{i++}\pi_{+j+}\pi_{++k} = \pi_{i++}\pi_{+j+} \sum_k \pi_{++k} = \pi_{i++}\pi_{+j+}.$$

Next, show conditional independence of X and Y given Z:

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{++k}} = \frac{\pi_{i++}\pi_{+j+}\pi_{++k}}{\pi_{++k}} = \pi_{i++}\pi_{+j+} = \pi_{i+|k}\pi_{+j|k}.$$

- Independence of Y from both X and Z does not necessarily imply independence of X and Z. Consider the loglinear model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$$

which has an XZ interaction whenever $\lambda_{ik}^{XZ} \neq 0$, but no interaction between Y and either X or Z. For example, if $\lambda_{ik}^{XZ} > 0$, then $\pi_{i+k} > \pi_{i++}\pi_{++k}$ so that X and Z can not be independent.

- d. Suppose X and Y are conditionally independent given Z. Then, within any level of Z there is no XY interaction. This implies that the conditional odds ratios that describe the XY association are all equal to 1. Hence, there can not be a three factor interaction λ_{ijk}^{XYZ} in the loglinear model, because this would mean that the XY association differed between levels of Z.