**BSTA 6651** **Midterm One** **Fall, 2011**

**Name:** Solution

**Coverage:** CDA 2nd ed., Chapters 1-4 **Time:** 100 minutes

**Instruction:** close book but 6 pages of help sheets are allowed; no computer access

**Question 1**

Consider a cohort study of comparing the death rates between therapies A and B. It was concluded that "the death rate of therapy A is two times that of therapy B." Answer the following parts.

(C) (a) Which of the following statement is correct?

A. The relative risk is 2.0
B. The odds ratio is 2.0
C. The death rate of therapy A is 2/3 while that of therapy B is 1/3
D. None of the above

Answer: A

(5) (b) [10 pts] Further assume that the odds of death were 0.5 for therapy A. Which of the following statement is correct?

A. The odds of death for therapy B is 0.25
B. The odds ratio is less than 2.0
C. The death rate of therapy A is 1/3
D. None of the above

Answer: C

**Question 2**

It is interested to study how the smoking habit of a student depends on the smoking habit of his/her parents. All students in a school district at that year were surveyed and answered questions. *The researcher did not know or preset the total number of students in advance.* Here is the data.

| | Student smokes | Student does not smoke |
|---|---|---|
| Both parents smoke | 410 | 1370 |
| One parent smokes | 416 | 1823 |
| Neither parent smokes | 188 | 1168 |

(5) (a) Which of the following sampling method was adopted here?

A. Poisson sampling
B. single multinomial sampling
C. independent multinomial sampling
D. None of the above

Answer: A

(10) (b) Does this table support that the smoking habit of a student was significantly associated with that of his/her parents? Identify your test and its p-value from the SAS output.

| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 2 | 37.5663 | <.0001 |
| Likelihood Ratio Chi-Square | 2 | 38.3658 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 37.4148 | <.0001 |

Answer: Yes. The test I choose is L-R $\chi^2$ test and its p-value is <.0001

(15) (c) Partition the 3x2 table into independent 2x2 subtables. CALCULATE and INTERPRET their odds ratios.

Answer: Y = Student smoking = {Yes, No};
X = # of parents who smoke = 0, 1, 2

(1)

| X \ Y | Yes | No |
|---|---|---|
| 2 | 410 | 1370 |
| 1 | 416 | 1823 |

$\theta_{XY} = \frac{410 \times 1823}{416 \times 1370} = 1.31$

(Consider those students whose parents smoke) the odds of smoking for students whose parents both smoke is 1.31 times that for those whose parent smokes.

(2)

| X \ Y | Yes | No |
|---|---|---|
| 1 or 2 | 826 | 3193 |
| 0 | 188 | 1168 |

$\theta_{XY} = \frac{826 \times 1168}{188 \times 3193} = 1.61$

The odds of smoking for students at least one of whose parents smoke is 1.61 times that for those both whose parents don't smoke.

(d). response = smoking. binary response ⇒ logit model.

$\text{logit}(\pi) = \alpha + \beta_1(\text{Race}) + \beta_2(\text{Gender})$ ; $Y_i = \text{smoking} \overset{\text{indept.}}{\sim} \text{Ber}(\pi_i)$, $i = 1, \cdots, n = 462$

link component; systematic component; random component.

**Question 3**

Part of data in a national study of 15-16 year old adolescents is described as below. The event of interest is ever smoking.

| | Race | Gender | Smoking | |
|---|---|---|---|---|
| | | | Yes | No |
| 0 = | White | Male (0) | 43 | 134 |
| | | Female (1) | 26 | 149 |
| 1 = | Black | Male (0) | 29 | 23 |
| | | Female (1) | 22 | 36 |

(6) (a) Calculate the conditional odds ratios between Gender (G) and Smoking (S).

(6) (b) Interpret these conditional odds ratios in context of the situation.

(6) (c) Based on the odds ratios found in part (a) and allowing for sampling error, is it reasonable to say the (conditional) association between Smoking and Gender is homogeneous? Defense your answer.

(12) (d) Identify the response variable here. What are the three components (random, systematic, link) of the corresponding generalized linear model (GLM)? Use the canonical link.

Answer:

(a). For whites:

| Gender | Smoking: Yes | Smoking: No |
|---|---|---|
| M | 43 | 134 |
| F | 26 | 149 |

$\theta_{GS(W)} = \frac{43 \times 149}{26 \times 134} = 1.84$

For Blacks:

| Gender | Smoking: Yes | Smoking: No |
|---|---|---|
| M | 29 | 23 |
| F | 22 | 36 |

$\theta_{GS(B)} = \frac{29 \times 36}{22 \times 23} = 2.06$

(b). The odds of smoking for male whites is 1.84 times the odds for female whites

The odds of smoking for male blacks is 2.06 times the odds for female blacks

(c) Yes, we can assume homogeneous association between Smoking & Gender because $1.84 \approx 2.06$.

**Question 4**

Let $X \sim Poisson(\mu)$. The mean and variance are $E(X) = \text{var}(X) = \mu$. The likelihood is $l(\mu) = e^{-\mu}\frac{\mu^x}{x!}$, for a non-negative observation $x$.

(6) (a) Prove that log function is the canonical link for Poisson random component in GLM.

Answer: $\theta = \mu$ here,

$$e^{-\mu}\frac{\mu^x}{x!} = \frac{1}{x!} \times e^{-\mu} \times e^{(\log \mu^x)}$$

$$= \underbrace{\frac{1}{x!}}_{b(x)} \times \underbrace{e^{-\mu}}_{a(\theta)} \times \exp[\underbrace{\log \mu}_{c(\theta)}], \text{ so the canonical link is } \log(\mu).$$

Parts (b) to (e) to derive the Wald CI for Poisson data:

(6) (b) Find the score function $S(\mu) = \frac{\partial \log l}{\partial \mu}$.

$$\log l = -\mu + x \log \mu - \log x!$$

$$\frac{\partial \log l}{\partial \mu} = -1 + \frac{x}{\mu}(-0) = \boxed{\frac{x}{\mu} - 1}$$

(6) (c) Find the MLE $\hat{\mu}$ by solving $S(\mu) = 0$.

$$S(\mu) = 0 \Rightarrow \frac{x}{\mu} = 1 \Rightarrow \hat{\mu} = \boxed{x}$$

(6) (d) Find the Fisher information $I(\mu) = -E\left\{\frac{\partial^2}{\partial \mu^2}\log l\right\}$.

$$\frac{\partial^2 \log l}{\partial \mu^2} = -\frac{x}{\mu^2} \Rightarrow I(\mu) = \frac{E(x)}{\mu^2} = \frac{\mu}{\mu^2} = \boxed{\frac{1}{\mu}}$$

(6) (e) Formulate the standard error of $\hat{\mu}$, and the Wald CI.

Wald C.I. is $\hat{\mu} \pm z_{\alpha/2}\, S.E(\hat{\mu})$

where $S.E.(\hat{\mu}) = \sqrt{I^{-1}(\hat{\mu})} = \sqrt{\hat{\mu}} = \boxed{\sqrt{x}}$

So Wald C.I. for $\mu$ is $\boxed{x \pm z_{\alpha/2}\sqrt{x}}$