Stats 539 Homework 6: Due Tuesday Apr. 4 by 10:50am

1. In murder trials in 20 Florida counties during 1976 and 1977, the death penalty was given in 19 out of 151 cases in which a white killed a white, in 0 out of 9 cases in which a white killed a black, in 11 out of 63 cases in which a black killed a white, and in 6 out of 103 cases in which a black killed a black (M. Radelet, *Am. Sociol. Rev.*, 46: 918-927, 1981). We will model these data using both log-linear models for contingency tables and logistic regression.

   (a) Enter these data into R as a `data.frame` with four columns: `Defendant` (race of defendant with levels "B" and "W"), `Victim` (race of victim with levels "B" and "W"), `Penalty` (whether or not death penalty was invoked with levels "N" and "Y"), and `Freq` (cell frequency). Your data.frame should have eight rows (not including variable names). Display this data.frame in R.

   (b) Use the R function `xtabs` with the `data.frame` format of the data to construct the partial tables needed to study the conditional association between defendant's race and the death penalty verdict, conditional on victim's race. Find and interpret the sample conditional odds ratios, adding 0.5 to each cell to reduce the impact of the zero cell count.

   (c) Find and interpret the sample marginal odds ratio between defendant's race and the death penalty verdict. Do these data exhibit Simpson's paradox? Explain.

   (d) Fit a logistic regression model which allows you to study the marginal association between defendant's race and the death penalty verdict. *Use your fitted logistic regression model* to answer the following questions.

      i. What is the estimated marginal odds ratio between defendant's race and the death penalty verdict? Show your work. (*Hint:* You should get the same answer as in part (c).)

      ii. Calculate and interpret an approximate 95% confidence interval for the marginal odds ratio of the death penalty comparing white defendants to black defendants.

      iii. Is there significant statistical evidence to suggest that the defendant's race has an effect on the odds of the death penalty? Justify your answer.

   (e) Fit a logistic regression model which allows you to study the conditional association between defendant's race and the death penalty verdict (conditional on victim's race). Assume homogeneous association between defendant's race and the death penalty, conditional on victim's race. *Use your fitted logistic regression model* to answer the following questions.

      i. What is the estimated conditional odds ratio between defendant's race and the death penalty for black victims? for white victims? Show your work. (Note that these answers will *not* match the answers in part (b).)

ii. Calculate and interpret an approximate 95% confidence interval for the conditional odds ratio of the death penalty comparing white defendants to black defendants conditioned on victim's race.

iii. Is there significant statistical evidence to suggest that the defendant's race has an effect on the odds of the death penalty after we control for the victim's race? Justify your answer.

(f) Fit the following four loglinear models (using the same notation as in the book), where D = Defendant, V = Victim, and P = Penalty: (D, V, P), (DV, VP), (DV, VP, DP), (DVP). You may use either R function `glm` or `loglm`. For each of the four fitted models, report the fitted conditional odds ratio(s) between Defendant and Penalty, conditioned on Victim.

(g) Assess the goodness of fit for each of the four models fit in part (f). Which model would you choose and why?

(h) For your chosen model from part (g), calculate an approximate 95% confidence interval for the conditional odds ratio(s) between Victim and Penalty, conditioned on Defendant. Show your work. How does this interval compare to the interval calculated in part (e)ii.

2. Assume that λ has a Gamma distribution with mean μ and shape parameter $k > 0$. That is, the pdf of λ is:

$$f(\lambda) = \frac{(k/\mu)^k}{\Gamma(k)} \lambda^{k-1} e^{-k\lambda/\mu}$$

and that $Y|\lambda \sim Pois(\lambda)$. Show that the marginal probability mass function of Y (given μ and k)

(a) is equal to the negative binomial pmf

$$p(y|\mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k, \quad y = 0, 1, 2, \ldots,$$

and

(b) is a member of the exponential dispersion family.

3. In this problem we will use data obtained from a sexually transmitted disease (STD) clinic to determine factors that are associated with the rate of reinfection of STDs. The dataset contains aggregated (grouped) data including the total number of reinfections per group, the total amount of time each group was followed for, and various other covariates. The data can be found on the course website at http://www.math.montana.edu/shancock/courses/stat539/data/stdgrp.txt. The following variables are available in the dataset:

Variable	Description
`white`	Indicator of white race (1=white, 0=black)
`married`	Marital status (1=single, 2=married, 3=divorced/separated)
`agegrp`	Age group (categorized as [13,19], (19,22], and (22,48])
`edugrp`	Years of schooling (categorized as [6,11.9], (11.9,12.9], (12.9,18])
`inftype`	Initial infection (1= gonorrhea, 2=chlamydia, 3=both)
`npartnr`	Number of sexual partners (categorized as [0,1], (1,2], (2,3], and (3,19])
`condom`	Condom use (1=always, 0=sometimes/never)
`n.reinfect`	Total number of reinfections observed in the group
`yrsfu`	Total number of years individuals in the group were followed for

(a) If we model the total number of reinfections observed in a group (`n.reinfect`) as a Poisson random variable, what variable in the data set should serve as the offset term?

(b) Write down the Poisson regression model that shows how the rate of reinfection varies with race, education group, initial infection and condom use (without interaction terms). Explain, using your model, how the rate is related to the actual count of reinfections.

(c) Fit the model in part (b) (show R code and output) and choose two estimated coefficients to interpret in context of the problem.

(d) Create an appropriate plot to examine the data for overdispersion. Refit your model accounting for overdispersion using a quasi-Poisson family. How do your conclusions change (if at all) after accounting for overdispersion?

(e) Do the grouped reinfection counts appear to follow a Poisson distribution? Why or why not?

(f) If we control for race and education, does condom use have a significant effect on the rate of re-infection? Conduct the appropriate likelihood ratio test (accounting for overdispersion) to address this question and write a conclusion of the test in context of the problem.