STA 303H1S / STA 1002HS: Loglinear Models / Poisson Regression Practice Problems

1. Consider the elephant mating example from lecture.
   (a) Both the binomial and the Poisson distributions provide probability models for counts. Is the binomial distribution appropriate for the number of successful matings of the male African elephants?
   (b) For the model fit in lecture, interpret the coefficient of age.
   (c) Consider the plot of the number of matings versus age. The spread of the responses is larger for larger values of the mean response. Should we be concerned?
   (d) From the estimated log-linear regression of the elephants' successful matings on age, what are the mean and variance of counts of successful matings (in the 8 years of the study) for the elephants who are aged 25 years at the beginning of the observation period? What are the mean and variance for elephants who are aged 45 years?
   (e) While it is hypothesized that the number of matings increases with age, there may be an optimal age for matings where, for older elephants the number of matings starts to decline. One way to investigate this is to add a quadratic term for age into the model to allow the log of the mean number of matings to reach a peak. Does the inclusion in the model of age^2 improve the fit?
2. What is the difference between a log-linear model and a linear model after the log transformation of the response?
3. Why are ordinary residuals $(y_i - \hat{\mu}_i)$ not particularly useful for Poisson regression?
4. Consider the deviance goodness-of-fit test.
   (a) Under what conditions is it valid for Poisson regression?
   (b) When it is valid, what possibilities are suggested by a small p-value?
   (c) When it is valid, what possibilities are suggested by a large p-value?
5. Poisson regression fits the model
$$\log(\mu_i) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$$
   where the μ_i's are the means of the Poisson distributions with observed counts y_i, $i = 1, \ldots, n$. Write down the log likelihood function used for maximum likelihood estimation of the β's.
6. Consider a table that categorizes 1000 subjects into 5 rows and 10 columns.
   (a) If Poisson log-linear regression is used to analyze the data, what is the sample size? (That is, how many Poisson counts are there?)
   (b) How would one test for independence of row and column factors?
7. The Physician's Health Study is a famous experiment in which male physicians between 40 and 84 years old were randomly assigned to take an aspirin or placebo every day. They were then followed and the number of myocardial infarctions (heart attacks) in each group was recorded. The data are summarized in the table below.

	Myocardial Infarction	
Group	Yes	No
Placebo	189	10,845
Aspirin	104	10,933

(a) Assuming that the number of physicians in each treatment group is fixed, carry out a test to see if the probability of a myocardial infarction in each treatment group is the same.

(b) SAS output is given on the practice problem website. In the output, several numbers have been replaced by `xxxxxx`. Fill in these numbers by using other numbers in the output or an appropriate statistical table.

(c) How do the conclusions from the output from proc genmod agree with your conclusions from the test in part (a)?

(d) For the first model fit using proc genmod (the independence model), do you trust the estimates from the model? For the second model fit using proc genmod (the saturated model), do you trust the p-values and confidence intervals?

8. In this question, you will show why the multinomial and Poisson models for the distribution of the counts in a 2×2 contingency table are equivalent.

   The sum of independent Poisson random variables has a Poisson distribution with mean equal to the sum of the means. That is, if the four counts, Y_{ij}, $i = 1, 2$, $j = 1, 2$, are independent Poisson random variables with means μ_{ij} then $\sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij} \sim \text{Poisson}(\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij})$.

   Recall that the conditional probability mass function of a random variable U given another random variable V is the probability mass function of the joint distribution of U and V divided by the probability mass function of the marginal distribution of V.

   For the model where $Y_{ij} \sim \text{Poisson}(\mu_{ij})$, show that the conditional distribution of the Y_{ij}'s given that the total count is n is multinomial with $\pi_{ij} = \frac{\mu_{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}}$.

9. Consider the drug survey example example from lecture where students were asked whether they used alcohol, cigarettes, or marijuana.

   (a) For the (AM,CM) model, calculated the estimated count of the number of students who use all of alcohol, cigarettes, and marijuana.

   (b) For the (AM,CM) model, estimate the odds of using marijuana for students who use both alcohol and cigarettes.

   (c) For the (AM,CM) model, calculate the estimated ratio between the odds of using marijuana for cigarette users and non-users. How do the estimated odds ratios differ depending on alcohol use?

   (d) For the (AC,AM,CM) model, conduct a likelihood ratio test to test whether the coefficient is 0 for the interaction term between alcohol and marijuana use. Is your answer consistent with the Wald test?

   (e) Suppose we want to know whether students who use alcohol or cigarettes are more likely to use marijuana. Then we will treat marijuana use as an outcome variable. Output from

4 logistic regression models is given on the practice problems website. If you compare this output to the output from the loglinear models considered in class, you'll find that the deviance for each of the logistic models is identical to the deviance for one of the loglinear models.

i. Match the corresponding models.

ii. Explain why the corresponding loglinear and logistic models are consistent in the way they model the alcohol, cigarette, and marijuana use relationship.

10. For a loglinear model for a 3-way contingency table, show that the deviance is

$$2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}\log\left(\frac{y_{ijk}}{\hat{\mu}_{ijk}}\right)$$