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

SOLUTIONS

1. (a) No. The binomial count is the count of events in a fixed number of trials with a definite upper limit.

   (b) The estimated mean number of matings increases by a factor of $e^{0.0687} = 1.071$ (that is, by about 7%) for a one year increase in age.

   (c) No. In the Poisson model we expect the variance to be equal to the mean and since the estimated mean varies with the predictor variable, so should the variance.

   (d) For 25-year-old elephants, mean$= \exp(-1.582 + 0.0687 age) = 1.15$, variance=1.15. For 45-year-old elephants, mean=4.53, variance=4.53.

   (e) No. From the Wald test, there is no evidence against the null hypothesis that the coefficient of age^2 is zero. (Moreover, because of the correlation between age and age^2, the coefficient for neither term is significant when both are in the model.) We can also look at the likelihood ratio test to compare the models with and without the quadratic term. The null hypothesis is that the coefficient of age^2 is zero (so that the two models are equivalent). The test has test statistic 0.1854 (the difference in the deviances, or 2 times the difference in the log likelihoods). From the chi-square distribution with 1 degree of freedom, the p-value is 0.67 (from tables we can say the p-value is between 0.1 and 0.9). So there is no evidence that the coefficient of age^2 is different from 0.

2. In a log-linear model, the mean of Y is μ and the model is $\log(\mu) = \beta_0 + \beta_1 X_1$. Y is not transformed. If a simple linear regression is used after a log transformation, the model is expressed in terms of the mean of the logarithm of Y. Moreover, the model assumptions are not the same.

3. The residuals with larger means will have larger variances. So if an observation has a large residual it is difficult to know whether it is an outlier or an observation from a distribution with larger variance than the others. Residuals that are studentized so that they have the same variance are more useful for identifying outliers.

4. (a) Since it is an asymptotic test (only approximate except in the limit where the sample size goes to infinity), we need large Poisson counts (expected cell counts at least 5 for contingency tables is one rule-of-thumb for "large").

   (b) The Poisson distribution is an inadequate model (for example, there may be extra-Poisson variation), the explanatory variables are inadequate (need more explanatory variables or a different form of the explanatory variables than you have in the model), or there are some outliers.

   (c) Either the model is correct, or there is insufficient data to detect any inadequacies.

5. The likelihood function is

$$\prod_{i=1}^{n} \frac{\mu_i^{y_i}\, e^{-\mu_i}}{y_i!} = \frac{e^{-\sum \mu_i} \prod \mu_i^{y_i}}{\prod y_i!}$$

and the log likelihood function is

$$\log\left(L(\beta_0,\ldots,\beta_p)\right) = -\sum_{i=1}^{n}\mu_i + \sum_{i=1}^{n} y_i\log(\mu_i) - \sum_{i=1}^{n} log(y_i!)$$

where $\mu_i = \exp\left(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}\right)$.

6. (a) 50

   (b) Test for the significance of the 36 interaction terms in the log-linear regression with row, column, and row-column interaction effects. This may be accomplished by fitting the model without the interaction terms and comparing the deviance to a chi-square distribution with 36 degrees of freedom.

7. (a) Testing $H_0: \; p_{aspirin} = p_{placebo}$ versus $H_a: \; p_{aspirin} \neq p_{placebo}$ where $p_{aspirin}$ and $p_{placebo}$ are the probabilities of an MI in the aspirin and placebo groups, respectively.
   The test statistic is

$$z_{obs} = (\hat{p}_{placebo} - \hat{p}_{aspirin}) \Bigg/ \sqrt{\hat{p}_{pooled}(1-\hat{p}_{pooled})\left(\frac{1}{189+10845}+\frac{1}{104+10933}\right)} = 5.0014$$

   where $\hat{p}_{placebo} = \frac{189}{189+10845}$ and $\hat{p}_{aspirin} = \frac{104}{104+10933}$ and $\hat{p}_{pooled} = \frac{189+104}{189+10845+104+10933}$.
   From tables for the standard normal distribution, $p < 0.0004$.
   So we have strong evidence that the probability of an MI is different for the two treatment groups.

   (b) See the complete SAS output on the website.

   (c) From the proc genmod output for the saturated model, there is strong evidence that the coefficient of the interaction term is not zero (Wald test, $p < 0.0001$), indicating that having an MI and treatment taken are not independent. This is consistent with the conclusion in part (a) which was that the probability of having an MI depends on the treatment group.

   We could also look at the output for the first model fit in proc genmod which is the complete independence model. This model does not adequately fit the data. The test with null hypothesis that this model fits as well as the saturated model (equivalent to the coefficient of the additional term (the interaction term) in the saturated model being 0) has test statistic 25.3720 (the deviance). From the chi-square distribution with 1 degree of freedom, the p-value for this test is < 0.005 (from the table). So we have strong evidence that the independence model does not fit as well as the saturated model. So modeling treatment group and occurrence of an MI as independent is not adequate. So we conclude that there is a relationship between treatment group and MI status, that is, the probability of having an MI depends on the treatment group.

   (It's worth noting that the chance of having an MI was higher for the placebo group.)

   (d) The independence model does not fit the data well. As noted above, we have strong evidence from the deviance goodness of fit test that the saturated model fits the data better. It is clear that the problem is that the model doesn't fit well and not just a case of extra-Poisson variation since the residuals are large. Thus the estimates from the independence model are not trustworthy since it is the wrong model.

The saturated model fits the data perfectly, so we have no concerns about model fit. For the inferences to be valid, we need large enough counts for the likelihood ratio tests and Wald tests and confidence intervals to be approximately correct. One rule-of-thumb is that all estimated counts should be at least 5, which is the case.

There is no concern here about independent observations since there is no reason to believe that the physicians are related in any way.

8. The conditional distribution of the Y_{ij}'s (assuming they are independent) given the total number of observations is

$$\prod_{i,j}\left(\frac{e^{-\mu_{ij}}\mu_{ij}^{y_{ij}}}{y_{ij}!}\right)\Bigg/\frac{\exp(-\sum_{i,j}\mu_{ij})(\sum_{i,j}\mu_{ij})^{\sum_{i,j}y_{ij}}}{(\sum_{i,j}y_{ij})!}$$
$$=\frac{(\sum_{i,j}y_{ij})!}{y_{11}!y_{12}!y_{21}!y_{22}!}\left(\frac{\mu_{11}}{\sum_{i,j}\mu_{ij}}\right)^{y_{11}}\left(\frac{\mu_{12}}{\sum_{i,j}\mu_{ij}}\right)^{y_{12}}\left(\frac{\mu_{21}}{\sum_{i,j}\mu_{ij}}\right)^{y_{21}}\left(\frac{\mu_{22}}{\sum_{i,j}\mu_{ij}}\right)^{y_{22}}$$

9. (a) The estimated count is $\exp(5.1921+1.1272-0.2351-6.6209+4.1251+3.2243)=909.3$.

(b) The estimated odds of using marijuana for students who use both alcohol and cigarettes are $\exp(-6.6209+4.1251+3.2243)=2.07$. That is, students who use both alcohol and cigarettes are twice as likely to use marijuana as they are not to use marijuana.

(c) Since we aren't given the estimated counts, we must use the fitted coefficients to find the estimated probabilities.

When A is 1 and C is 1, the estimated odds of using marijuana are $\exp(-6.6209+4.1251+3.2243)=2.07$. When A is 1 and C is 2, the estimated odds of using marijuana are $\exp(-6.6209+4.1251)=0.082$. So for the odds of marijuana use, the odds ratio for cigarette users to non-users is $2.07/0.082=25.1$ for alcohol users. That is, for alcohol users, cigarette users have 25 times the odds of using marijuana as students who do not use cigarettes.

When A is 2 and C is 1, the estimated odds of using marijuana are $\exp(-6.6209+3.2243)=0.033$. When A is 2 and C is 2, the estimated odds of using marijuana are $\exp(-6.6209)=0.0013$. So for the odds of marijuana use, the odds ratio for cigarette users to non-users is $0.033/0.0013=25.1$ for alcohol non-users. That is for students who do not use alcohol, cigarette users have 25 times the odds of using marijuana as students who do not use cigarettes.

Note that we get the same answer whether or not the student uses alcohol. This is to be expected since for the (AM,CM) model, alcohol use and cigarette use are conditionally independent given marijuana use.

(d) We need to conduct a likelihood ratio test to see if the (AC,AM,CM) fits better than the (AC,CM) model. The null hypothesis is that the two models fit equally well, which is equivalent to the extra term in the (AC,AM,CM) (the interaction between alcohol and marijuana use) having coefficient 0. The test statistic is the difference in the deviances (or 2 times the difference in the log likelihoods) for the models which is 91.6444. If the null hypothesis is true, this is an observation from a chi-square distribution with 1 degree of freedom. From the chi-square table, the p-value is < 0.005 and we conclude that we have strong evidence that the coefficient of the alcohol-marijuana interaction term is not zero. Although the test statistic is different for the Wald test, the conclusion is the same as the p-value for the Wald test (with null hypothesis that the coefficient of the AM interaction term in the (AC,AM,CM) model is 0) is < 0.0001.

(e) i.

	Loglinear model	Logistic model	Deviance
1	(AC,M)	null	843.8266
2	(AC,AM)	A	497.3693
3	(AC,CM)	C	92.0184
4	(AC,AM,CM)	(A,C)	0.3740

ii. For model 1, the form of the loglinear model says that there is no relationship between both alcohol and cigarette use and marijuana use, i.e. alcohol and cigarette use are independent of marijuana use. This is consistent with the logistic regression model in which neither alcohol nor cigarette use are used as predictors of the probability of marijuana use.

For model 2, the form of the loglinear model says that there is no relationship between cigarette use and marijuana use, i.e. cigarette use is independent of marijuana use. This is consistent with the logistic regression model in which cigarette use is not used as a predictor of the probability of marijuana use.

For model 3, the form of the loglinear model says that there is no relationship between alcohol use and marijuana use, i.e. alcohol use is independent of marijuana use. This is consistent with the logistic regression model in which alcohol use is not used as a predictor of the probability of marijuana use.

For model 4, the form of the loglinear model says that each pair of variables is related, but the relationship between any pair of variables does not vary with the value of the third variable. This is consistent with the logistic regression model in which there is no interaction term for alcohol and cigarette use so we are not modeling the relationships between alcohol and marijuana use and between cigarette and marijuana use as being different depending on whether or not the students use cigarettes or alcohol, respectively.

10. The deviance is 2 times the difference between the log-likelihood for the saturated model and the log-likelihood for the fitted model. Using the result from question 5, the log-likelihood function is

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(-\mu_{ijk} + y_{ijk}\log(\mu_{ijk}) - \log(y_{ijk}!)\right).$$

Let $\hat{\mu}_{ijk}$ be the estimated value of μ_{ijk}. For the saturated model, the estimated value of μ_{ijk} is y_{ijk}. So the deviance is

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

and since $\sum_i\sum_j\sum_k \hat{\mu}_{ijk} = \sum_i\sum_j\sum_k y_{ijk} =$ the total number of counts, the deviance is the formula given.