Bios 765: Homework # 2, Fall 2020
due September 23, 2020

1. (**Applied, 15 pts.**) The outcome of interest from Homework # 1, problem #2 is the number of roots produced by 270 micropropagated shoots of the columnar apple cultivar *Trajan*. During the rooting period, all shoots were maintained under identical conditions, but the shoots themselves were cultured on media containing different concentrations of the cytokinin BAP concentration (2.2, 4.4, 8.8, 17.6) in growth cabinets with an 8 or 16 hour photoperiod. Define the follow explanatory variables for analysis: $x_{i1} = 1$ if the $i-$th shoot was assigned to the 16-hour photoperiod and $x_{i1} = 0$ if assigned to the 8-hour photoperiod; and $x_{i2} = \ln(BAP) - \ln(2.2) = \ln(BAP/2.2)$.

   (a) Report the percentage of shoots that gave zero roots for each photoperiod.

   (b) Write down a model expression for the zero-inflated Poisson (ZIP) regression model for number of roots as the outcome that includes x_{i1}, x_{i2} and their interaction as explanatory variables in the distribution part of the model and only x_{i1} as a covariate in the zero-inflation model part. Also, fit the corresponding zero-inflated negative binomial (ZINB) regression model. Create a single table with regression parameter estimates, standard errors and p-values for both models.

   (c) Provide a test of the dispersion parameter from the ZINB model in part b) and interpret the result with respect whether ZIP or ZINB is the preferred model.

   (d) From your selected "best" model from part c), estimate and interpret the exponentiated regression coefficients from both model parts (excluding the intercepts) in language that a non-statistician can understand. Report 95% confidence intervals along with these estimates.

   (e) Fit a marginalized zero-inflated regression model as an alternative to the "best" model chosen in part c). Write down an expression for the model equation. Additionally, write down the fitted model. Interpret the estimated exponentiated regression coefficients from both model parts (excluding the intercepts) in language that a non-statistician can understand. Report 95% confidence intervals along with these estimates. Refer to "notes5_RegTwoYear-DMFS.sas" on website for SAS code example.

   (f) Based on the marginalized model from part (e), estimate the incidence density ratio (and its 95% confidence interval) of the comparison of the 16 hour photoperiod versus the 8 hour photoperiod for each of the four BAP levels. Interpret the results.

2. (**Applied 15 pts.**) Table A gives the number of mentally ill persons from King County between 1993 and 1998 that have been captured by one, two or all three systems, where systems are J=Jail, M=Medicaid, and C=County Mental Health Services. We assume that these individuals represent a certain population from King County which we will call the poor mentally ill; this population would not include those mentally ill persons who are able to pay for private medical care. Also assume persons' responses can be described by identical independent multinomial distributions with $n = 1$ characterized by $2^3 = 8$ cells; the first seven cells correspond to the seven possibilities of being captured (see Table A); the last cell, the outcome of not being captured by any system is unobserved.

Table A. Frequency Table of the Poor Mentally Ill as captured by three systems

SYSTEM	count
J only	77,269
M only	138,578
C only	13,391
J and M	25,631
J and C	3,079
M and C	23,876
J,M and C	8,528

(a) Fit the model of no three factor interaction (JM,JC,MC). Estimate $\boldsymbol{\pi}_0$, the conditional probabilities of falling into the categories represented by the rows of Table A. Create a table like Table A but with two more columns, for the observed and model predicted conditional probabilities. Additionally, fit the best 2 two-factor interaction model, and add a column to your table reporting its estimated conditional probabilities. Discuss the goodness of fit of this model relative to model (JM,JC,MC).

(b) Let $\pi_{ijk}, i = 1, 2; j = 1, 2; k = 1, 2$ be the unconditional probability from the full 8 cell multinomial, such that π_{222} is the probability of not being captured by any of the three systems. Give the X matrix for model (JM,JC,MC) including the deleted row, x'_{222}, that resulted in X_0 used in part a. Compute $\hat{\pi}_{222}$ and $\hat{\gamma}$, where $\gamma = \pi_{222}/(1 - \pi_{222})$.

(c) Determine $\hat{n}_{222}$ and $\hat{N}$ where n_{222} is the number of the poor mentally ill not captured, and N is the total population size of the poor mentally ill in King County.

(d) Using SAS IML, compute an estimate of var $\begin{bmatrix} \hat{\boldsymbol{\pi}}_0 \\ \hat{\gamma} \end{bmatrix}$.

(e) Determine $se(\hat{n}_{222})$ and $se(\hat{N})$.

(f) Briefly state the assumptions of your analysis.

3. (**Applied 10 pts.**) In Bios 765 Notes 8, a loglinear model was fit to selected variables from the diabetes self-management study, a cross-sectional survey of older adults in rural North Carolina (the study is described in Quandt *et al.*, *Ethnicity and Disease*, 2005, 15:656-663). Consider the data set of seven categorical variables: E (ethnicity: 1-White, 2-African American, 3-Native American), D (diet), X (exercise), G (glucose), I (Insulin use), F (foot care), and C (medical care). Except for ethnicity, all variables are coded as 1=yes, and -1=no to indicate whether a particular strategy is used to manage diabetes. Be complete and precise when asked to provide interpretations. Make sure you properly handle sampling zeros.

   (a) Fit the hierarchical loglinear model containing all two-factor interactions as well as the three-factor interaction EDC (shortcut notation in SAS CATMOD is "loglin E|D|X|G|I|F|C@2 E*D*C;"). Report the likelihood ratio test. Is it appropriate to use this test to justify the goodness-of-fit of the model to this data?

   (b) Fit a reduced hierarchical loglinear model by deleting two-factor interaction terms from the model in (a) that are not statistically significant ($p > 0.10$). Provide a chi-square test of goodness-of-fit for the reduced model based upon comparison of deviance to the model in (a).

   (c) Using the model results from (b), give the estimate (and 95% large sample confidence interval) for the odds ratio of diet and Medical Care, for each ethnic group. Provide interpretations.

   (d) Using the model results from (b), provide odds ratio estimates, 95% confidence intervals, and interpretations corresponding to each two-factor interaction term that does not involve ethnicity.

4. (**Theory, 5 pts.**)

Assume $Y_1, \ldots, Y_s$ have independent Poisson distributions with respective means $\mu_1, \ldots, \mu_s$, and define the loglinear model

$$\log(\mu_i) = \alpha + x_i'\boldsymbol{\beta}$$

where x_i is a $t \times 1$ covariate vector for the i-th observation and $\boldsymbol{\beta}$ is the corresponding parameter vector. Note that exposure $N_i = 1$ for all observations.

(a) Show that the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\text{var}(\hat{\boldsymbol{\beta}}) = (1/n)[X'(D_{\boldsymbol{\pi}} - \boldsymbol{\pi}\boldsymbol{\pi}')X]^{-1}$$

where X is the $s \times t$ design matrix with rows x_i', $n = \sum_{i=1}^{s} \mu_i$, and the ith element of $\boldsymbol{\pi}$ is $\pi_i = \mu_i/n$.

(b) Given that the score equations for $\boldsymbol{\beta}$ are identical for the Poisson loglinear model and the Multinomial loglinear model based on single multinomial sampling (assuming the same covariate structure X in each), what are the implications of the result in **a.** for statistical inference. In particular, explain why it is permissable (or not permissable, if that is the case) to use the Poisson loglinear model for analyzing contingency table data generated from single multinomial sampling. Are results based on likelihood ratio tests for the Poisson loglinear model applicable to data generated from single multinomial sampling? Please explain.