Bios 765: Homework # 1, Fall 2020
due September 9, 2020

1. (**Applied, 15 pts.**) Consider the data from Notes 3 "Distribution of discovered Down Syndrome and of total live births by maternal age and birth order, Michigan, 1950-1964". Conduct the following steps in your statistical analysis of these data.

   (a) Construct an alternative analysis of deviance table to the one presented in Notes 3, slide 31. Your table should include the following four models: [0] Saturated model, [2] m_cat, b_cat, mb, [3*] m_cat, b_cat, and [4] m_cat, b. Note that models [0], [2], and [4] correspond to the models fitted in Notes 3, whereas model [3*] is a new model that is a main effects model treating both maternal age and birth order as having a nominal scale. Your table should follow the same format as in Notes 3, slide 31 by including likelihood ratio tests and their corresponding p-values for first-order differences (i.e., comparison of deviances for pairs of models from consecutive rows).

   (b) From your results in part 1, present an argument for the case that model [4] is the "best" model. Next, construct the likelihood ratio test for the second-order comparison of model [2] versus model [4]. Interpret the result.

   (c) Using β-notation, write down a model equation for model [4] using "Under 20" as the reference group for maternal age. Define all notation. Additionally, write down the fitted model equation for model [4]. Create a table containing the regression parameter estimates, their estimated standard errors, z-scores and p-values.

   (d) Compute 95% confidence intervals (CI) for incidence density ratios describing the effect on Down Syndrome of each maternal age interval relative to "Under 20", and for a one unit increase in birth order. Provide an interpretation for each CI.

   (e) Strengthen your argument in favor of choosing model [4] as the "best" model by determining the observation with the largest (in absolute terms) standardized Pearson residual among all the observations. What is the largest standardized Pearson residual for model [4] and how does it compare to the largest one from model [3] from Notes 3?

2. (**Applied, 15 pts.**) The outcome of interest in the Table 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.

Table. *Frequency distributions of the number of roots produced by 270 shoots of the apple cultivar Trajan, classified by the experimental conditions (BAP concentration and photoperiod) under which the shoots were reared; shown are the numbers of shoots that produced 0,1,...,12 roots; counts that exceed 12 are shown individually.*

	Photoperiod							
	8-hour				16-hour			
BAP(μM)	2.2	4.4	8.8	17.6	2.2	4.4	8.8	17.6
Number of roots								
0	0	0	0	2	15	16	12	19
1	3	0	0	0	0	2	3	2
2	2	3	1	0	2	1	2	2
3	3	0	2	2	2	1	1	4
4	6	1	4	2	1	2	2	3
5	3	0	4	5	2	1	2	1
6	2	3	4	5	1	2	3	4
7	2	7	4	4	0	0	1	3
8	3	3	7	8	1	1	0	0
9	1	5	5	3	3	0	2	2
10	2	3	4	4	1	3	0	0
11	1	4	1	4	1	0	1	0
12	0	0	2	0	1	1	1	0
> 12	13,17	13	14,14	14				
Total number of roots	175	233	300	286	98	82	94	98
Number of shoots	30	30	40	40	30	30	30	40

(a) Compute the mean and the variance of the number of roots and report them along with sample size (number of shoots) for each combination of photoperiod and BAP. Comment on the relationship between the variance and the mean.

(b) Fit the Poisson regression model to the n=270 count observations where the number of roots is the outcome, and explanatory variables are an indicator variable for photoperiod=16, the natural log of BAP, and their interaction. Report the estimates and their standard errors and p-values. Is the Pearson chi-squared goodness-of-fit test valid in this situation? Explain.

(c) For the data set with eight observations corresponding to the combinations of photoperiod and bap, fit the same Poisson regression model as in part b) but with total number of roots as the outcome and the natural log of the number of shoots as the offset. Considering the sample size, is this model valid? What can you conclude about the fit of the model, if anything, from the Pearson chi-square goodness-of-fit test? Explain. Comment on how the parameter estimates and standard errors compare to those model results in part b).

(d) Using the original data of 270 observations (or shoots), find the "best" negative binomial regression model when the number of roots is the outcome, photoperiod and log-BAP are explanatory variables. Include their interaction in the model if it is statistically significant at the two-sided 0.05 level; report the Wald test p-value for the interaction term. Write down an expression for your model using β-notation. Write down the fitted model equation. Create a table with regression parameter estimates, standard errors, z-scores (or chi-square values) and p-values.

(e) Provide a test of the hypothesis that the dispersion parameter in your negative binomial regression model in part d) is equal to zero, $H_0 : \alpha = 0$ versus $H_1 : \alpha > 0$. Interpret the result.

(f) From the model in part d), estimate the incidence density ratio and its 95% confidence interval for the effects of photoperiod on the mean number of roots and provide an interpretation that non-statisticians would understand.

3. (**Theory, 10 pts.**) Determine the asymptotic variance matrix of the survival estimates from the duodenal ulcer data in Koch, Atkinson and Stokes (1986). Specifically, do the following:

(a) Applying the general matrix method for the propagation of variances, derive a matrix expression for the asymptotic variance of the predicted survival rates from the piecewise exponential regression model. Give a formula for an estimator of this asymptotic variance matrix.

(b) Using PROC IML (or some other matrix computer language), determine the numerical value of the estimated asymptotic variance matrix of predicted survival rates for the duodental ulcer data. From this matrix, determine the standard error estimates of the maximum likelihood model predicted survival rates.

4. (**Theory, 10 pts.**) Suppose $y_{11}, y_{12}, \ldots, y_{1r}, \ldots, y_{s1}, y_{s2}, \ldots, y_{sr}$ are independent random variables which have Poisson distributions with means μ_{ij} for $i = 1, 2, \ldots, s$ and $j = 1, 2, \ldots, r$.

   (a) Show that the $n_{i.} = \sum_{j=1}^{r} y_{ij}$ for $i = 1, 2, \ldots, s$ are independent Poisson random variables with means $\mu_{i.} = \sum_{j=1}^{r} \mu_{ij}$.

   (b) Show that $(y_{11}, y_{12}, \ldots, y_{1r}, \ldots, y_{s1}, y_{s2}, \ldots, y_{sr})$ given $(n_{1.}, n_{2.}, \ldots, n_{s.})$ conditionally has the product multinomial distribution

$$\phi = \prod_{i=1}^{s} \frac{n_{i.}! \prod_{j=1}^{r} \pi_{ij}^{y_{ij}}}{\prod_{j=1}^{r} y_{ij}!} \text{ with } \pi_{ij} = \frac{\mu_{ij}}{\mu_{i.}}.$$