Biometry 711 - Categorical Data Analysis
Summer 2015
Homework 3
Due Wednesday 29 July 2015, 6pm

1. Read Sections 14.3.1 and 14.3.2 in your text. In class we showed that a gamma mixture of Poisson distributions results in the negative binomial distribution, a distribution that accommodates overdispersed count data. A similar approach can be used to model overdispersed binomial data by constructing a beta mixture of binomial distributions (see Section 14.3 of Agresti). Accordingly, let $Y|\pi \sim \text{Binomial}(n, \pi)$ and $\pi \sim \text{Beta}(\alpha_1, \alpha_2)$.

   (a) Show that the marginal distribution of Y is the beta-binomial distribution given by

   $$p(y; \alpha_1, \alpha_2) = \binom{n}{y} \frac{B(\alpha_1 + y, n + \alpha_2 - y)}{B(\alpha_1, \alpha_2)}, \; y = 0, 1, \ldots, n$$

   where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$.

   (b) Show that $E(Y) = n\mu$ where $\mu = \frac{\alpha_1}{\alpha_1 + \alpha_2}$.

   (c) Show that $Var(Y) = n\mu(1 - \mu)(1 + (n - 1)\rho)$ where $\rho = 1/(\alpha_1 + \alpha_2 + 1)$. The factor $1 + (n - 1)\rho$ is the variance inflation factor (VIF), the relative inflation in variance due to overdispersion, and ρ is the intra-cluster correlation (ICC) which measures the correlation between the n Bernoulli random variables that sum to Y. Give two different conditions under which the VIF equals 1 and so no overdispersion is present.

2. Read Sections 14.3.2 - 14.3.4 of your text.

   (a) Using the data shown in Table 4.7 of your text (page 151) and available as either `teratology.csv` or `teratology.sas7bdat` on the class website, fit the logistic model described in Section 14.3.4 on page 550 of your text assuming no overdispersion. Provide estimates and corresponding SEs for α, β_2, β_3 and β_4.

   (b) Based on the output, report the value of X^2, the Pearson GOF statistic, and its corresponding degrees of freedom. Recall that the scale parameter, ϕ, is equal to 1 for the Binomial distribution but $\phi > 1$ indicates overdispersion. Explain how this output supports a conclusion of overdispersion, and provide an estimate of the scale parameter.

   (c) In class, we discussed how to use a Quasi-likelihood (QL) approach to inflate the variance for overdispersed count data. A similar approach can be used to adjust the variance estimates for overdispersed binomial data, and QL for overdispersed binomial data is presented in Section 14.3.3 of your text. Use the `scale = pearson` model option in `SAS PROC LOGISTIC` to estimate α, β_2, β_3, β_4 and their SEs. Explain why the point estimates of the model parameters remain unchanged under QL, but the SEs are larger.

   (d) As noted in Section 14.3.2 of your text, the beta-binomial distribution is not a member of the scaled exponential family and therefore corresponding regression models are not in the class of generalized linear models. `SAS PROC FMM` (which stands for finite mixture model) does allow you to fit a beta-binomial regression model with *nearly* the same syntax as in `PROC LOGISTIC`. Review the help file for this procedure, and fit the same model as in question 2a but this time using a beta-binomial distribution. Report the parameter estimates and corresponding SEs. The estimated

'scale parameter' in the parameter estimates table is equivalent to $(1-\rho)/\rho$, where ρ is the ICC defined in question 1c. Based on the reported scale parameter, estimate the ICC, and the variance inflation for litters of 10 pups.

3. The dataset `homicides.csv` (or `homicides.sas7bdat`) on the class website contains responses from 1308 subjects to the question: within the past 12 months, how many people have you known personally that were victims of homicide?

   (a) Let y_i denote the response for subject i and let $x_i = 1$ for blacks and $x_i = 0$ for whites. Fit the Poisson GLM $\log \mu_i = \beta_0 + \beta x_i$ and interpret $\hat{\beta}$.

   (b) Describe factors of heterogeneity such that a Poisson GLM may be inadequate. Fit the corresponding negative binomial GLM. Report the estimated dispersion parameter (reported in the parameter estimates table of the output, and defined on page 553 of your text), and subsequently write an explicit expression demonstrating how the variance depends on the mean. What evidence does this model fit provide that the Poisson GLM had overdispersion?

   (c) Show that the Wald 95% confidence interval for the incidence rate ratio for blacks and whites is (4.2, 7.5) for the Poisson GLM but (3.5, 9.0) for the negative binomial GLM. Which do you think better reflects the uncertainty in the estimated rate ratio? Why?

4. CDA 8.2.

5. CDA 8.12(a).