BIOSTAT/STAT 570: Midterm Take-Home Exam

To be submitted to the course canvas site by 11:59pm Monday 15th November, 2021. This is an exam, so no collaboration.

In this exam you will investigate the modeling of binary data, accounting for potential overdispersion.

Consider the situation in which we have a sequence of binomial trials of size N_i, with Y_i being the number of events of interest, $\boldsymbol{x}_i$ being covariates associated with trial i, and π_i being the proportion of the events of interest, for $i = 1, \ldots, n$.

The simplest (appropriate) approach to analyzing such data is to assume the *Binomial Model*:

$$\begin{aligned} Y_i|\pi_i &\sim \text{Binomial}(N_i, \pi_i) \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{x}_i\boldsymbol{\beta}, \qquad i = 1, \ldots, n. \end{aligned}$$

If the data do not exhibit binomial variance, we can consider the *Quasi-Binomial Model:*

$$\begin{aligned} \text{E}[Y_i] &= N_i\pi_i \\ \text{var}(Y_i) &= N_i\pi_i(1-p_i) \times \kappa \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{x}_i\boldsymbol{\beta}, \qquad i = 1, \ldots, n. \end{aligned}$$

An alternative for modeling overdispersion is the *Beta-Binomial Model:*

$$\begin{aligned} Y_i|P_i &\sim \text{Binomial}(N_i, P_i) \\ P_i|\pi_i, \tau_i^2 &\sim \text{Beta}(a_i, b_i), \qquad \text{with} \qquad \pi_i = \frac{a_i}{a_i + b_i}, \quad \tau_i^2 = \frac{1}{a_i + b_i + 1} \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{x}_i\boldsymbol{\beta}, \qquad i = 1, \ldots, n. \end{aligned}$$

1. **Methods**

   (a) **4 Points** For the Beta-binomial model, write $\text{E}[P_i]$ and $\text{var}(P_i)$ in terms of π_i, τ_i^2. Hence, derive the marginal mean and variance $\text{E}[Y_i]$, $\text{var}(Y_i)$ in terms of π_i and τ_i^2.

   (b) **4 Points** Suppose one fits the binomial model and forms Pearson residuals:

$$r_i = \frac{y_i - N_i\widehat{\pi}_i}{\sqrt{N_i\widehat{\pi}_i(1-\widehat{\pi}_i)}}.$$

By comparing the forms of the variances for the Beta-Binomial and Quasi-Binomial models, explain why a plot of r_i versus N_i might help to choose between these two models. Hence, explain how one might choose between these models.

(c) **3 Points** For the Beta-binomial model, derive the marginal distribution of the data, $\Pr(Y_i|a_i, b_i)$.

(d) **4 Points** Is the Beta-Binomial model a member of the exponential family? What are the implications for inference?

2. **Simulation Study** In practice, the Beta-Binomial model is often used with $\tau_i^2 = \tau^2$.

The logistic form in the simulation is based on $\beta_0 = -2$, $\beta_1 = \log 4$ with x uniformly spaced on $[0, 2]$, i.e., $x_i = 2 \times (i-1)/(n-1)$. The simulation experiments will have $n = 50, 100, 200, 500, 1000$ trials with $N_i \sim_{iid}$ Poisson(30), $i = 1, \ldots, n$. Hence, for example, in the $n = 50$ experiment there will be denominators, N_i, $i = 1, \ldots, 50$.

Simulate under the following scenarios:

- *No excess-binomial variation:* Simulate Y_i from a Binomial distribution.
- *Moderate excess-binomial variation:* Simulate Y_i from a Beta-Binomial distribution with $a_i + b_i = 10$.
- *Strong excess-binomial variation:* Simulate Y_i from a Beta-Binomial distribution with $a_i + b_i = 4$.
  [Hint: In the second and third cases, solve for a_i, b_i, given π_i and $a_i + b_i$, in order to perform the simulation, for $i = 1, \ldots, n$.]

For each of the three scenarios, fit Binomial, Quasi-Binomial and Beta-Binomial models.

[Hint: for the Binomial and Quasi-Binomial use the `glm()` function, and for the Beta-Binomial use the `betabin()` function in the `aod` library.]

(a) **20 Points** For each of $\widehat{\beta}_0$ and $\widehat{\beta}_1$, report both the bias and the coverage of 90% asymptotic confidence intervals (use Wald intervals), giving your results in graphical form.

(b) **6 Points** Summarize and explain the results.

3. **Frequentist Data Analysis**

Toxicology Data: Weil (1970) and Williams (1975) describe a toxicological experiments in which there were two randomized groups (placebo and chemical treatment), each containing 16 pregnant rats (so $n = 32$). For each of the pregnant rats, the total litter size was recorded (N_i) , along with the number of the litter who were alive at 21 days

(Y_i). These data may be found in the `VGAM` library, under the name `prats`.

Seeds Data: Crowder (1978) reports $n = 21$ binomial experiments in which two factors (each with 2 levels) were varied, seed type and the type of root extract. The total number of seeds (N_i) and the number that germinated (Y_i) were recorded. These data may be found in the `BradleyTerry2` library, under the name `seeds`.

(a) **6 Points** For the Toxicology data, fit the Binomial, Quasi-Likelihood and Beta-Binomial models with a logistic regression model with treatment as the covariate. On the basis of the estimated levels of overdispersion, examination of Pearson residuals versus N_i, or otherwise, decide on the most appropriate analysis. Summarize the association between survival and treatment, based on your favored model.

(b) **6 Points** For the Seeds data, fit the Binomial, Quasi-Likelihood and Beta-Binomial models with a logistic regression model with seed types and extract as covariates. On the basis of the estimated levels of overdispersion, examination of Pearson residuals versus N_i, or otherwise, decide on the most appropriate analysis. Summarize the association between germination of seeds and the two covariates, based on your favored model.

4. **Bayesian Data Analysis**

(a) **10 Points** For each of the Toxicology and Seeds data, carry out a Bayesian analysis, fitting the Beta-Binomial model in `INLA`. Produce tables that compare the posterior means and posterior standard deviations with the MLEs and standard errors found in the Frequentist Data Analysis part, and comment.

5. **Extensions**

(a) **6 Points** Suppose we assume only:

$$\begin{aligned} \mathsf{E}[Y_i] &= N_i\pi_i \\ \mathsf{var}(Y_i) &= N_i\pi_i(1-\pi_i)[1+(N_i-1)\tau^2], \end{aligned}$$

with $\pi_i = \pi(\boldsymbol{\beta}) = \mathsf{expit}(\boldsymbol{x}_i\boldsymbol{\beta})$. What philosophical approach to inference could be used? Sketch out an algorithm for estimation of $\boldsymbol{\beta}$ and τ^2.