BIOSTAT/STAT 570: Midterm Takehome Exam Key

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:*

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

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) 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.
   Moments of prior:
   $$\text{E}[P_i] = \pi_i, \qquad \text{var}(P_i) = \pi_i(1-\pi_i)\tau_i^2.$$

Marginal mean:

$$\mathsf{E}[Y_i] = \mathsf{E}[\mathsf{E}[Y_i|P_i]] = \mathsf{E}[N_i P_i] = N_i \pi_i$$

Marginal variance:

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

(b) 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.

If the Beta-Binomial is appropriate, the residual plot will show increasing spread as a function of N_i.

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

$$\begin{aligned} \Pr(Y_i|a_i, b_i) &= \int \Pr(Y_i|P_i) \times p(P_i|a_i, b_i)\, dP_i = ... \\ &= \binom{N_i}{Y_i} \frac{\Gamma(a_i + b_i)}{\Gamma(a_i)\Gamma(b_i)} \frac{\Gamma(Y_i + a_i)\Gamma(N_i - Y_i + b_i)}{\Gamma(N_i + a_i + b_i)} \end{aligned}$$

(d) Is the Beta-binomial model a member of the exponential family? What are the implications for inference?

No, not a member of the exponential family. If the model is not specified correctly, no guarantees of consistency for $\boldsymbol{\beta}$.

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

Take $N_i = 11$ in each binomial trial, and $\beta_0 = -2$, $\beta_1 = \log 4$ with x_i uniformly spaced on $[0, 2]$, i.e. $x_i = 2 \times (i-1)/(n-1)$. We have $i = 1, \ldots, n$, and carry out simulation experiments with $n = 10, 100, 500, 1000$.

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 (for the Binomial and Quasi-Binomial use the `glm` function, and 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.

We plot the empirical bias and the coverage as a function of n for each true data generating mechanism and each estimation strategy. The bias and the coverage are each calculated over 1000 simulated datasets. The results are given in Figures 1 and 2.

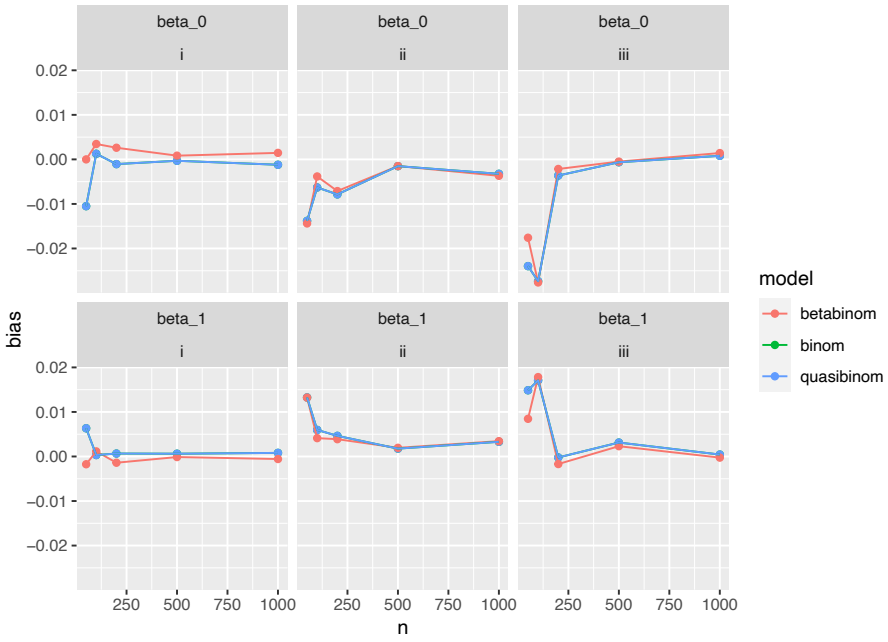

Figure 1: Empirical bias of confidence intervals, computed using 1,000 simulated datasets. Note that the binomial model results are hidden under the quasibinomial results, as the two estimation strategies yield identical point estimates (and thus the same bias). The columns of the plot relate to the different true data generating mechanisms.

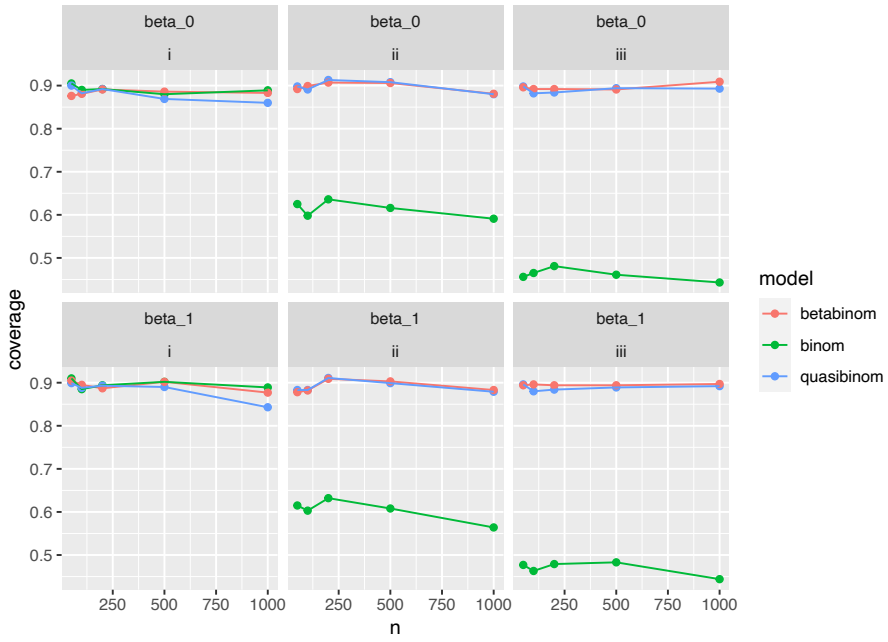

Figure 2: Empirical coverage of confidence intervals, computed using 1,000 simulated datasets. The columns correspond to the data generating mechanisms.

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

As n increases, the bias of all three estimation strategies shrinks to 0 under all three data generating mechanisms. This is because the mean model is correctly specified in all three data generating mechanisms for all three estimation strategies.

All three methods seem to achieve approximately correct nominal coverage under data generating mechanism (i) where there is no true overdispersion. Interestingly, the quasibinomial model seems to have slightly incorrect coverage for large n here. The quasibinomial model should be correct with $\kappa = 1$, but perhaps the model is accidentally attributing some of the true signal to overdispersion and is thus estimating $\kappa \neq 1$. One takeaway from this is that if there is no true overdispersion it is best to just use a binomial model. This under-coverage seen for quasibinomial may simply be Monte Carlo error, and so we would want to explore this idea more with more iterations before trusting this conclusion too much.

The binomial model has terrible coverage under mechanisms 2 and 3. This is because the binomial model does not allow for overdispersion, and mechanisms 2 and 3 have a lot of overdispersion. Coverage is even worse for model 3, as the overdispersion gets worse. The quasibinomial model does not correctly specify the variance for mechanisms 2 and 3, because it assumes that the variance is

a linear function of the mean and in reality here we have beta-binomial data and this have a more complex mean-variance relationship. Nonetheless, quasibinomial model seems to get pretty close to 90% nominal coverage for all three mechanisms. As the beta-binomial model is a correctly specified likelihood model, it achieves correct nominal coverage. For both bias and coverage, the beta binomial model sometimes performs slightly better than the quasibinomial model for smaller n. This is because a correctly specified likelihood model should be most efficient, whereas more flexible models require larger sample sizes.

Some students lost points for suggesting that both the quasibinomial model and the beta binomial model are correctly specified in the case of overdispersion. The quasibinomial model does not have the correct form here; it just turned out empirically to perform well. For full credit, students needed to make this point.

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.

In each model we fit, let

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(treatment_i).$$

Table 1 shows the estimates and corresponding standard errors for β_0 and β_1 in each model.

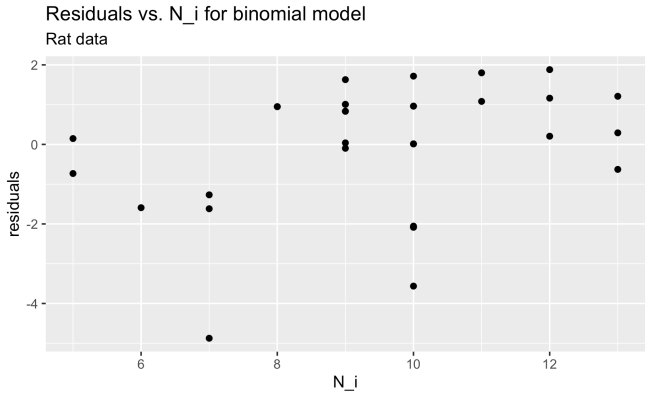

Figure 3: Residuals vs. N_i for the binomial model for the rats data.

Figure 3 shows the Pearson residuals verses the N_i for the binomial model. We do see a trend in the residuals; we seem to have mostly positive residuals when N_i is large and mostly negative residuals when N_i is small. However, with the exception of a few points when $N_i = 10$, the spread in the residuals seems to be constant across N_i. This suggests that the pattern could be due to an issue with the mean model, not the variance model.

Figure 4 shows the Pearson residuals verses the N_i for the beta binomial; these are different residual than before because we use our new model and we now divide by the beta-binomial variance. In terms of trends that we see in the plot, we still see a few strange low outliers at certain values of N_i. But it is hard to say for sure if "constant spread over N_i" has been achieved by switching to the beta binomial model.

Justify a choice of either the quasibinomial or the Beta Binomial. After justifying choice, be sure to interpret the coefficients from that model. Using the Beta Binomial model, we estimate that $\beta_0 = 1.86$ and $\beta_1 = -0.66$. This means that, on average, rats with treatment have their children's odds of survival multiplied by $exp(-0.66) = 0.52$. It seems surprising that the treatment decreases their odds of survival. The odds of a baby's survival for a rat without treatment is $exp(1.86) = 6.42$.

(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.

We carry out essentially the same workflow as in the rats dataset, but now our

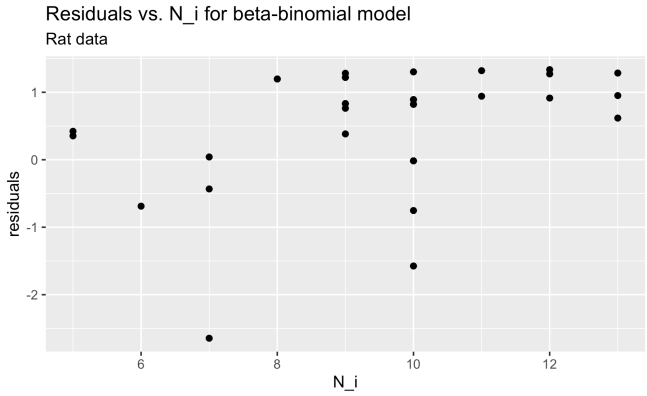

Figure 4: Residuals vs. N_i for the beta-binomial model for the rats data.

model is

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(seed) + \beta_2(extract).$$

Some students may have also included an interaction term between the two treatments.

For the model that I chose (no interaction terms), Figure 5 does not seem to show a pattern of the spread of the Pearson residuals increasing with N_i. Figure 6 shows the Pearson residuals for the beta-binomial model; it does not necessarily seem as though the beta-binomial model is a better choice. The quasibinomial model estimates an overdispersion parameter of 2, which suggests that perhaps there really is some overdispersion in the data, and so maybe the quasibinomial model is the best choice.

The fitted beta-binomial model is given by

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = -0.43 - 0.27(seed) + 1.06(extract).$$

Compared to experiments with seed type = 0, experiments with seed type = 1 are estimated to have odds-of-germination multiplied by $exp(-0.27) = 0.76$. Compared to experiments with root extract type = 0, experiments with root extract type = 1 are estimated to have their odds of germination multiplied by $exp(1.01) = 2.88$. The odds of germination for experiments with seed type = 0 and root type = 0 are $exp(-0.39) = 0.65$.

4. **Bayesian Data Analysis**

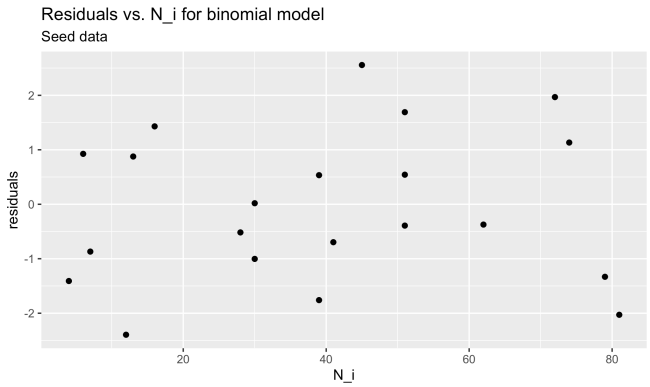

Figure 5: Residuals vs. N_i for the binomial model for the seeds data.

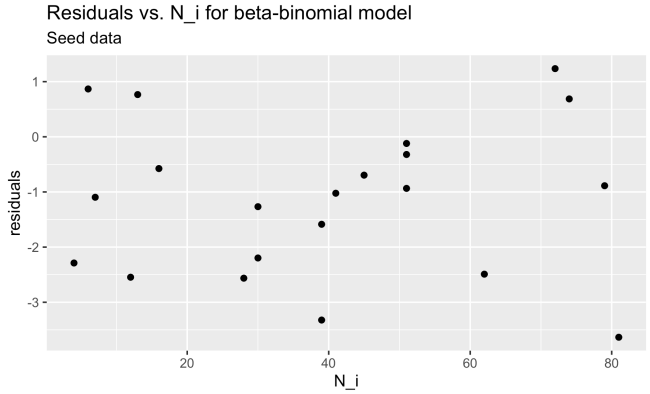

Figure 6: Residuals vs. N_i for the beta-binomial model for the seed data.

Model	β_0 estimate	β_0 SE	β_1 estimate	β_1 SE
Binomial	2.18	0.26	-0.96	0.33
Quasibinomial	2.18	0.43	-0.96	0.54
Beta Binomial	1.86	0.36	-0.66	0.46
Bayesian	1.81	0.37	-0.65	0.46

Table 1: Table summarizing coefficient estimates and standard deviations for β_0 and β_1 for the rats dataset. For the bayesian analysis we are reporting the posterior mean and posterior standard deviations.

Model	β_0 estimate	β_0 SE	β_1 estimate	β_1 SE	β_2 estimate	β_2 SE
Binomial	0.43	0.11	-0.27	0.15	1.06	0.14
Quasibinomial	-0.43	0.17	-0.27	0.23	1.06	0.21
Beta Binomial	-0.39	0.16	-0.34	0.21	1.01	0.20
Bayesian	-0.35	0.21	-0.39	0.26	0.98	0.25

Table 2: Table summarizing coefficient estimates and standard deviations for β_0 and β_1 for the seeds dataset. For the bayesian analysis we are reporting the posterior mean and posterior standard deviations.

(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.

For both the Toxicology and the seeds data, I fit a bayesian model in INLA with `family = "betabinomial"`. I used the default INLA priors for this family. According to the INLA documentation for the betabinomial family, this corresponds to putting a $N(0, 0.4)$ prior on θ, where

$$\tau^2 = \frac{1}{a_i + b_i + 1} = \frac{e^\theta}{1 + e^\theta}.$$

As Tables 1 and 2 show, the bayesian version of the beta binomial model gives very similar results to the frequentist beta binomial model for both datasets. The beta binomial MLE is quite close to the posterior mean, and the beta binomial standard error is quite close to the posterior standard error. This suggests that the INLA default priors are not particularly informative. The bayesian results differ from the frequentist results more for the seeds data than for the rats data; this may be due to the smaller sample size (21 observations as opposed to 32 rats). For the seeds dataset, the Bayesian model has slightly larger posterior standard deviations than the MLE standard errors for the frequentist version.

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.

The approach and the algorithm are essentially given in Section 2.5.2 of the Wakefield regression textbook; simply replace α in those equations with parameter τ.

We are able to fit this model using the theory of estimating equations. The form of the estimating equation is given in (2.38) in the textbook.

Fitting this model is more difficult than fitting a quasi-likelihood model because the parameter τ^2 appears in the estimating equation for the mean model. We need to carry out estimation using the same two-step algorithm that was suggested in Section 2.5.2 of the textbook. We need to start out with an initial guess for $\hat{\tau}$ called $\hat{\tau}^t$ for $t = 0$. Once we have this guess, we can solve the estimating equation to get $\hat{\beta}^{t+1}$. We then use the residuals from the current guess for the mean model to estimate $Var(Y_i)$ and thus obtain a new estimate for the overdispersion parameter $\hat{\tau}^{t+1}$. We increment t and continue in this manner until convergence.

As the assumption for the variance is used in the estimating equation that estimates β, the estimates of β are only consistent if both the mean and the variance are correctly specified.

Note that using this estimating equation approach, we have a way to get asymptotically valid standard errors to do inference on the parameters because we know the asymptotic distribution. This allows us to do inference. Some students lost points for suggesting vague estimation strategies (such as optimization or least squares) without explaining how they could do inference.