ALTERNATIVE ESTIMATION TECHNIQUES FOR CORRELATED DISCRETE DATA

William J. Price Ph.D.
bprice@uidaho.edu

Bahman Shafii Ph.D.
University of Idaho, bshafii@uidaho.edu

Alternative Estimation Techniques for Correlated Discrete Data

William J. Price and Bahman Shafii

Statistical Programs

University of Idaho

Moscow, Idaho 83843 USA

Binary or multinomial data often occur in agricultural and biological research. Advancements in measurement and video technologies now allow such data to be sequentially recorded through time or space. These data sets, however, can exhibit a serial correlation structure, which in turn, can bias and influence point estimates as well as inferences made regarding the data. Statistical methods using generalized mixed models and probability distributions such as the beta-binomial and correlated binomial have been proposed as potential solutions for estimating the parameters of interest in these cases. In this paper, we will explore the properties of these techniques through simulation studies and demonstrate each scenario using real data related to olfactometer choice tests of a seed eating weevil.

I. Introduction

Discrete binary or multinomial data are common in agricultural and biological research. Plant studies, for example, may explore the occurrence of discrete events, such as plant emergence, the formation of plant structures, or plant survival/mortality. Animal research, on the other hand, may also involve discrete responses such as the occurrence and frequency of behavioral changes in individual animals. Recent advances in measurement techniques such as radio tagging, video technology, and geolocation allow discrete data events to be sequentially recorded in time or space at relatively high resolutions and accuracy. This can result in a serially correlated data structure and lead to biased and inconsistent parameter estimates if not properly accounted for. While this inherent correlation structure can be problematic, it can also be modeled, thereby allowing necessary adjustments for subsequent parameter estimation and inference.

Modeling techniques may include the use of generalized mixed models assuming predefined serial correlation structures. This method accounts for the inherent discrete nature of the data while simultaneously weighting the variance-covariance structure with a specified correlation structure.

As an alternative, the correlation structure can be explicitly accounted for by merging an underlying data model with an ancillary distribution for the correlation. A common model following this technique, the Beta-Binomial, includes a correlation parameter applied through a Beta distribution in conjunction with a Binomial distribution assumed for the discrete response.

Yet, a more direct and processed-based approach involves deriving a likelihood function from a time sequence of discrete binary events. This technique which is developed based on the data generation process, directly models the inherent and underlying correlation structure.

The purpose of this paper is to explore the properties of the aforementioned techniques using simulation studies, as well as demonstrating the results utilizing data from an olfactometer choice test on a gall forming weevil.

II. Methods

Experimental Design and Data Description

This work was motivated by insect host plant choice tests. The research was designed to assess host plant preference cues of insects based on their response to visual or olfactory stimuli, as well as the combination of both. These were tested using a Y-tube olfactometer, where individual insects were presented with two targets, one at each branch end of a Y shaped tube arena. Various targets were considered including: live host plant material from several species, volatized chemical(s) from the host plants, and colored targets spectrally similar to the host plants. In each case, one or more of these positive selection choices was paired with a paper control target having no cue present. For both the positive and control targets, actual contact with the target material was prevented by a mesh screen. To begin the test, a single individual was released into an olfactometer at the base of the Y shaped tube. Traditionally, the insect's choice would then be recorded after a predetermined period of time. After several runs of N individuals, the resulting data could be represented as binomial (N, p), where parameter p represents the preference for the positive target. In the current experiment, however, the insect location in the olfactometer was determined via video, and recorded every minute for 60 minutes. While the proportion of times an individual spent on a given target could be computed over the 60 minutes, these observations over time would not be independent and, hence, the data process for a given individual would no longer be considered as a binomial process.

Statistical Models and Estimation

Four potential models were considered for statistical analyses of the data. These included a standard binomial ignoring the serial correlation, a generalized linear mixed model (GLMM) assuming a conditionally binomial response with an autoregressive correlation structure, a beta-binomial model accounting for the correlation between observations using a beta distribution in conjunction with a binomial distribution for the response frequency, and finally, a process-derived likelihood model developed based on the serial structure inherent in the data. The details of each statistical model are described below.

Binomial Model

$$\boldsymbol{y_{ij}} \sim \textbf{bin}\ (\textbf{T}, \boldsymbol{\theta_i}) = \binom{T}{y_{ij}} \boldsymbol{\theta_i}^{y_{ij}} (\mathbf{1} - \boldsymbol{\theta_i})^{T - y_{ij}} \qquad (1)$$

where $\boldsymbol{y_{ij}}$ is the number of successes from j=1 to T time points (trials) for the $\boldsymbol{i^{th}}$ weevil. The time points are assumed to be independent, ignoring potential auto-correlation. Estimation can be carried out through least squares or maximum likelihood. The final estimate for host preference across all weevils is computed as avg ($\boldsymbol{\theta_i}$).

Generalized Linear Mixed Model

If x_{ij} are individual [0, 1] measurements of success for the jth time in the ith weevil, then

$x_{ij} \mid s_{ij}$ ~ binary (θ) and s_{ij} is a random effect assumed i.i.d. N (0,σ^2). A linear Predictor: $\eta_{ij} = \alpha_i + s_{ij}$ can then be formed where η_{ij} is a logit link function, i.e.: logit (θ). An autocorrelation structure such as the autoregressive AR(1) may then be imposed on the variance-covariance structure, Var (x_{ij} | s_{ij}) with parameter ρ. Estimation is then completed using maximum likelihood techniques (Stroup 2012). The estimate of θ can be considered as a measurement of host preference adjusted for the serial correlation structure.

Beta-Binomial Model

If y_i is the number of successes for the ith weevil over M time points, then y_i is distributed as:

$$\boldsymbol{y_i \sim \textbf{beta-bin}\ (\mathrm{M}, \theta_i', \tau) = \binom{M}{y_i}\frac{B(y_i+\frac{\theta_i'}{\tau},\ M-y_i+\frac{1-\theta_i'}{\tau})}{B(\frac{\theta_i'}{\tau},\ \frac{1-\theta_i'}{\tau})}} \qquad (2)$$

where the usual binomial parameter, θ, is now a random variate for the ith weevil, and $\boldsymbol{\theta}_i'$, is assumed to follow a beta distribution. The term B (a,b) represents a beta function with parameters (a) and (b), written as functions of the response y_i, the probability of success, $\boldsymbol{\theta}_i'$, and τ, a measure of over-dispersion.

Parameters of Eq.2 can be estimated through maximum likelihood procedures (Diniz, et al. 2010; Martinez et al,. 2010). The host preference probability is then estimated as the

avg ($\boldsymbol{\theta}_i'$).

Process Derived Likelihood

Given a binary time series of successes (1s representing time on the positive olfactometer target) and failures (0s, time spent off target), e.g.:

0 **1 1 1 1 1** 0 0 0 **1 1 1** 0 0 0 0 **1 1 1 1**,

we note that the strings of 1s are of interest and can be characterized by three phases:

1) Initiation from an off target value of 0, to an on target value, 1,
2) Continuation of 1s with $\boldsymbol{n_{il}}$ successes for the lth series in the ith weevil, and
3) Termination of the sequence with a return to an off target value of 0.

The probability of each phase can then be assigned as:

1) p (initiation) = r' ,
2) p (continuation) = $r^{\eta_{il}-1}$, where r is the probability of choosing a value of 1, and
3) p (termination) = 1 – r.

The probability of the lth series, S_l, is then:

$$p(S_l) = \mathrm{r}' \cdot r^{\eta_{il}-1} \cdot (1 - \mathrm{r}\,) \qquad (3)$$

and the probability of a sequence of series, S_l , in the ith weevil, l = 1, 2, 3, …, N, is:

$$\prod_{l=1}^{N} \mathrm{r}' \cdot r^{\eta_{il}-1} \cdot (1 - \mathrm{r}\,) \qquad (4)$$

Equation (4) is then used as the basis for a likelihood function across all weevils and r is the estimate of host preference. Estimation can be carried out through standard maximum likelihood techniques.

The estimation process can also be extended to incorporate the random effect of weevil by redefining r as:

$$\mathrm{Logit}\ (\rho') = \ln\ (r/(1\text{-}r)) + \phi_i \ ;\ \phi_i \sim N\ (0, \sigma^2)\ . \qquad (5)$$

The E [ρ'| ϕ_i] is then used as an estimator for host preference.

Computations

All statistical computations were carried out using SAS 9.4 64 bit (SAS 2012):

- Binomial Model; least squares; Proc Means
- GLMM Model; LaPlace optimization; Proc Glimmix
- Beta-Binomial and Process Likelihood Models; maximum likelihood; Proc Nlmixed

SAS Codes for all techniques are available at: http://webpages.uidaho.edu/cals-statprog/

III. Simulation Results and Demonstration

Simulation Process

The simulation process was designed to mimic the available experimental data, where ten weevils per target type were individually observed in the olfactometer tests every minute for 1 hour. The probability of initiation, r′, was set to 0.50, with the expectation that initially there would be no preference for either the positive target or the control. The probability of continuing, r, was then varied from 0.50 to 0.90 that is, ranging from no preference to high positive preference. Each setting of r was then simulated for ten weevils and the entire process repeated B = 1000 times, resulting in a total of 3 million data points. Each model estimation technique was then fitted to these simulation data, recording the estimated host preference for each of the simulated data iteration. Bias, computed as the difference between the estimated host preference and the known simulated value of r was also recorded.

Binomial Simulation Results

Figure 1 presents the simulation results for the standard binomial model, ignoring any serial correlation. As might be expected, the simulation with no preference, r = 0.50, shows a good match between the estimated and the simulated value, 0.5. Higher values of r, however, show a systematic deviation from the one-to-one relationship between estimated and simulated values. The resulting estimates of host preference at high values of r consistently underestimate the true value.

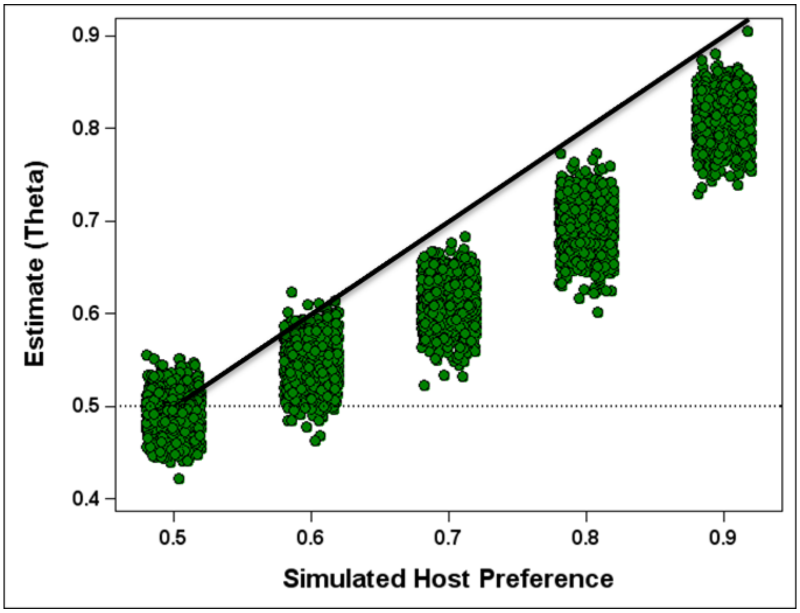

Figure 1. Estimated host versus simulated preference assuming the uncorrelated binomial model for 1000 simulated data sets. Dashed line represents no preference; solid line represents a theoretical one-to-one relationship.

A summary of the bias for these estimates is shown in Figure 2. It can be seen that ignoring the serial correlation in the data when host preference is present can lead to serious negative bias in the estimates, even at moderate levels of host preference.

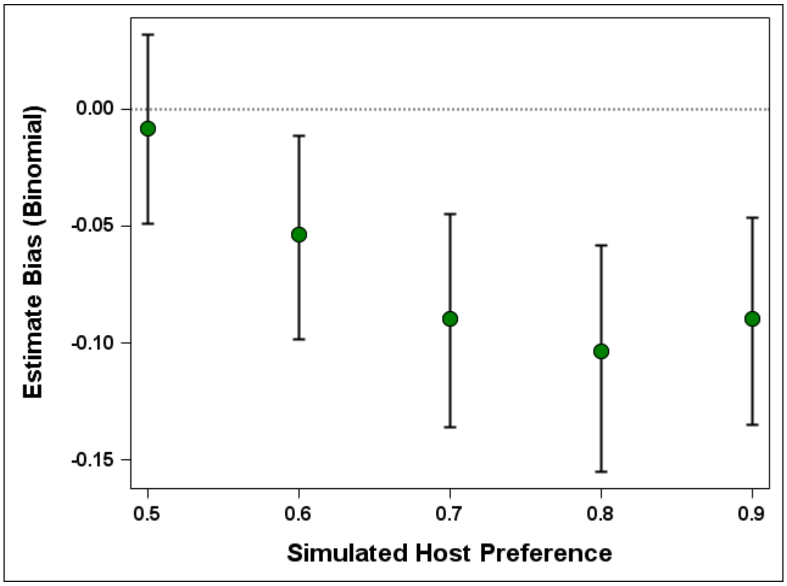

Figure 2. Mean binomial model estimate bias for host preference versus simulated values. Error bars represent the lower and upper 95% percentiles of the 1000 simulation runs. The dashed line represents no bias.

GLMM Simulation Results

The GLMM technique, assuming an AR (1) correlation structure, performed better than the binomial model and showed smaller bias, which was not different from zero based on percentile intervals (Figure 3). Some bias was evident, however, at very high values for r, i.e. r=0.9, where the estimated host preference was consistently estimated at or very near to 1.0. This behavior may be due to problems in the estimation algorithms with high levels of correlation in an AR (1). process. While the details of this issue were not considered here, they may present problems in applying this technique to actual data and should be investigated further.

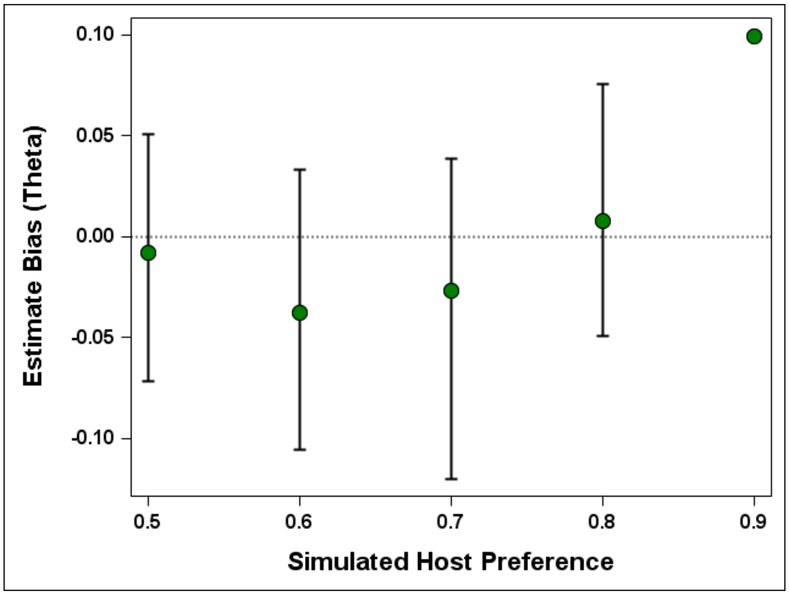

Figure 3. Mean GLMM estimate bias for host preference versus simulated values. Error bars represent the lower and upper 95% percentiles of the 1000 simulation runs. The dashed line represents no bias.

Beta-Binomial Simulation Results

The bias of the beta-binomial model was similar to that of the standard binomial model (Figure 4). As simulated host preference increased, the model consistently underestimated the simulated value. Only the case of no host preference, r = 0.50, showed no bias. It is noted that, while this model considers correlated binary data, it does so in a manner more consistent with a constant spatial or temporal correlation (not serial correlation). This may account for the lack of performance in the beta-binomial case.

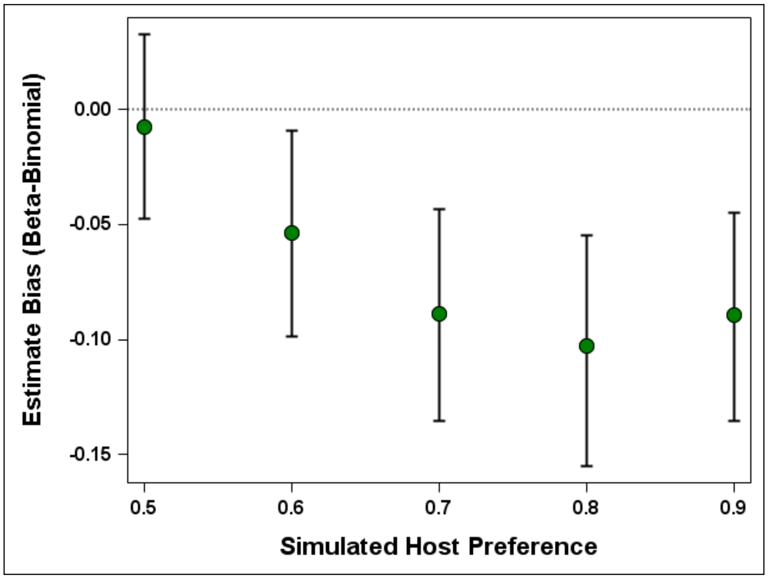

Figure 4. Mean estimate bias of the Beta-Binomial model for host preference versus simulated values. Error bars represent the lower and upper 95% percentiles of the1000 simulation runs. The dashed line represents no bias.

Process Likelihood Simulation Results

The process likelihood model had good performance over all simulated values of host preference (Figure 5). There was little bias and all 95% percentile intervals covered the zero bias region.

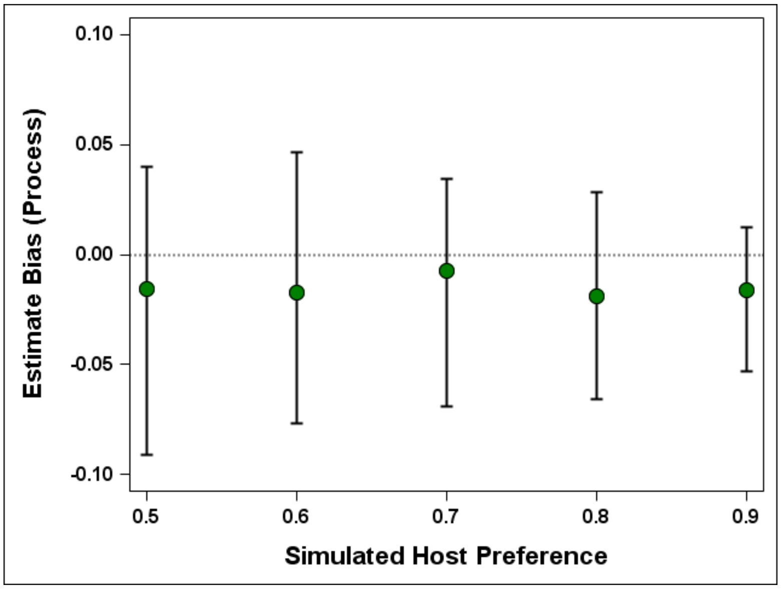

Figure 5. Mean estimate bias of the process likelihood model for host preference versus simulated values. Error bars represent the lower and upper 95% percentiles of the 1000 simulation runs. The dashed line represents no bias.

This performance might be expected, however, because the process used to generate the simulation data, as well as the method used to derive the process likelihood were very similar.

Demonstration

Olfactometer tests were carried out to assess the factors influencing the biological control of the weed species whitetop (Lepidium draba). The potential control agent was the gall forming weevil Ceutorhynchus cardariae. The experimental objectives were to determine what stimulus cues attract the weevil to the plant. This was assessed using plant material or plant volatiles in combination with blank paper controls. Ten to fifteen weevils were separately tested and measured as to their position in the olfactometer every minute for 60 minutes. For the purpose of this demonstration, the positions were recorded as either on the positive target or off the target (control target or neither target). Only the process-derived likelihood model and GLMM are considered for this demonstration.

Table 1 shows the estimated host preferences for two positive target choices (L. draba plant material, L. draba volitiles only) under each specified model type. Both models show very high preference for each target type, although the process likelihood has a slightly smaller value for the volatiles at 0.86.

A comparison of targets can be set up by redefining r in (5) with indicator variables for the target types:

$$r = I_{Plant} \cdot r_{Plant} + \ I_{Volitile} \cdot r_{Volitile} \qquad (6)$$

where I_{Plant} and $I_{Volitile}$ are [0,1] indicators for the plant material and volatile treatments, respectively with corresponding preferences, r_{Plant} and $r_{Volitile}$. Under this full model, a contrast of host preferences for the process likelihood model gives an approximate p-value of 0.2181, while a similar comparison for the GLMM has a p-value of 0.0024. Although the process likelihood model provided a larger difference in estimates, it did so with less precision. The GLMM model, however, has a tendency to overestimate at high levels of r and may exhibit over precision and bias in this case. A bootstrap simulation or other nonparametric techniques may be more appropriate for estimation and inference under these circumstances.

Target Type	Process Model	GLMM
Plant Material	r = 0.9487	r = 0.9953
Volatiles only	r = 0.8672	r = 0.9858

Table 1. Estimated host preference for two target types using the likelihood process model and GLMM.

IV. Concluding Remarks

Observation of binary data over time presents estimation issues in choice tests due to serial correlation. Ignoring this correlation leads to biased estimates and inaccurate inference. Alternative statistical models may help in this regard, but not all methods are successful. GLMM, for example, adjusts for the correlation well, but may have computational problems when target preferences are high. The beta-binomial (or similar distributions) may not model the appropriate serial process. Alternatively, a likelihood developed based on the data generation process will perform well in terms of the estimated bias, but may lack precision. The performance of this method could potentially be improved if the model incorporated negative preferences (host avoidance) in addition to positive host preferences, as given here. The GLMM alternative, for example, considers preferences that can be either positive or negative and, hence, may be a more robust estimation technique covering a wider array of practical scenarios. Inference and precision for both the process likelihood and GLMM methods, however, may not perform well at extreme values of the host preference, r. Such cases could require estimation methods other than maximum likelihood as well as a re-specification of the underlying correlation structure. While the GLMM and process likelihood models currently show some limitations, these can be anticipated and potentially corrected for. In summary, methods that inherently account for and model serial correlation are preferable to those that lack that characteristic. Failing to do so will result in biased estimates and incorrect inference.

References

Diniza, C. A. R., M. H. Tutiab and J. G. Leitea. 2010. Bayesian analysis of a correlated binomial model. Brazilian Journal of Probability and Statistics. Vol. 24, No. 1, 68–77.

Martinez, E. Z., J. A. Achcar, and D. C. Aragon. 2010. Parameter estimation of the beta-binomial distribution: an application using the SAS software. Ciência e Natura, Santa Maria, v. 37 n. 4, 2015, p. 12–19.

SAS/Stat Version 9.4 Copyright © 2012, SAS Institute Inc., Cary, NC, USA

Stroup, W. W. 2012. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. Chapman & Hall/CRC Texts in Statistical Science, pp 555.