University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2013 *Paper* 35

On the Simulation of Longitudinal Discrete Data with Specified Marginal Means and First-Order Antedependence

Matthew Guerra* Justine Shults†

*Division of Biometrics III, OB, CDER, FDA, Silver Spring, MD 20993., guerra.m.w@gmail.com

†Univeristy of Pennsylvania Department of Biostatistics, jshults@mail.med.upenn.edu

On the Simulation of Longitudinal Discrete Data with Specified Marginal Means and First-Order Antedependence

Matthew Guerra and Justine Shults

Abstract

We propose a straightforward approach for simulation of discrete random variables with overdispersion, specified marginal means, and product correlations that are plausible for longitudinal data with equal, or unequal, temporal spacings. The method stems from results we prove for variables with first-order antedependence and linearity of the conditional expectations. The proposed approach will be especially useful for assessment of methods such as generalized estimating equations, which specify separate models for the marginal means and correlation structure of measurements on a subject.

On the Simulation of Longitudinal Discrete Data with Specified Marginal Means and First-Order Antedependence

Matthew W. Guerra and Justine Shults

SUMMARY. We propose a straightforward approach for simulation of discrete random variables with overdispersion, specified marginal means, and product correlations that are plausible for longitudinal data with equal, or unequal, temporal spacings. The method stems from results we prove for variables with first-order antedependence and linearity of the conditional expectations. The proposed approach will be especially useful for assessment of methods such as generalized estimating equations, which specify separate models for the marginal means and correlation structure of measurements on a subject.

KEY WORDS: Antedependence models; Correlated discrete data; General-

---

Matthew W. Guerra is in the Division of Biometrics III, OB, CDER, FDA, Silver Spring, MD 20993. Justine Shults is Associate Professor, Department of Biostatistics, University of Pennsylvania, Philadelphia, PA 19034 (E-mail: matthew.guerra@fda.hhs.gov and jshults@mail.med.upenn.edu). Disclaimer: This manuscript reflects the views of the authors and should not be construed to represent the FDA's views or policies. Matthew Guerra began this work while he was a PhD student in the Department of Biostatistics at the University of Pennsylvania, and was supported by a grant from the National Cancer Institute (T32 CA93283). With permission from the journal, we note that an earlier version of this paper was submitted to *The American Statistician* on May 2, 2011, and a shorter version was tentatively accepted on May 28, 2013. The authors are also very grateful to their friend and mentor Professor Thomas Ten Have for suggesting that they work on this manuscript and dedicate this paper to his memory.

ized estimating equations; Longitudinal data; Markovian dependence of order one; Overdispersion; Product correlations; Simulation

1. INTRODUCTION

Longitudinal discrete data are commonly encountered in research. For example, a study might record the monthly number of kidney transplants performed in each of a large number of centers, along with the portion that were from live donors.

Semi-parametric approaches, such as generalized estimating equations (Liang and Zeger 1986), are especially attractive for the analysis of discrete data, as the likelihoods of discrete random variables for a likelihood based approach can be very complex. However, construction of the underlying distribution is useful to evaluate methods, such as generalized estimating equations, if the likelihoods can be used to simulate realizations of random variables with the same features that were specified by the semi-parametric approach.

Quite a few methods have been proposed for the simulation of correlated binary variables, including approaches by Emrich and Piedmonte (1991), Qaqish (2003), and those reviewed by Farrell and Rogers-Stewart (2008). However, fewer authors considered correlated discrete random variables (and in particular, count variables) that are not Bernoulli. Gange (1995) used iterative proportional fitting (IPF) to simulate correlated categorical variables. Schulman et al. (1996) described how the linear programming (LP) method of Lee (1993) for simulation of dichotomous variables could be generalized for the multi-category case, but also cautioned that neither the IPF method or the LP method is satisfactory for simulation of a large number of random variables. Other methods are described in Devroye (1986).

We propose an approach that previously was unavailable, for the simulation of discrete variables with specified marginal means, overdispersion that is a common feature of discrete data (Efron 1992), and product correlations that are plausible for longitudinal trials (Nuñez-Antón and Woodworth 1994). The proposed approach is straightforward for simulation of categorical or count variables, and its ease of implementation does not necessarily lessen with an increase in the number of simulated variables.

2. SIMULATION APPROACH

2.1 Results

The following results will be used to construct likelihoods that allow for simulation of random variables $Y_1, \ldots, Y_n$ with specified marginal means, overdispersion, and product correlations.

Theorem 2.1. *Let $E(Y_j \mid Y_{j-1})$ be linear in Y_{j-1}, so that $E(Y_j \mid Y_{j-1}) = a_j + b_j Y_{j-1}$ $(j = 2, \ldots, n)$. Then,*

$$E\left(Y_j \mid Y_{j-1}\right) = \mu_j + C_{j-1,j}\sigma_j/\sigma_{j-1}\left(Y_{j-1} - \mu_{j-1}\right), \qquad (2.1)$$

where $\mu_j = E(Y_j)$; $C_{j-1,j} = corr(Y_{j-1}, Y_j)$; and ${\sigma_j}^2 = var(Y_j)$; furthermore,

$${\sigma_j}^2 \;=\; 1/(1 - C^2_{j-1,j})E\{var(Y_j \mid Y_{j-1})\} \quad (j = 2, \ldots, n). \qquad (2.2)$$

Proof. Utilizing results from Christensen (1997), the conditional expectation $\mathrm{E}\left(Y_j \mid Y_{j-1}\right)$ is the function of Y_{j-1} that minimizes the squared-error loss, $\mathrm{E}\left(Y_j - f(Y_{j-1})\right)^2$, while the best *linear* predictor of Y_j based on Y_{j-1} is the *linear* function of Y_{j-1} that minimizes the squared-error loss. If the conditional expectation is linear, it will also be the best linear predictor and can then be expressed as in Equation (2.1), which was obtained using the expression for the best linear predictor (Christensen 1997, p.108). The result can also be shown directly. Next, as a consequence of (2.1), the marginal means $\mathrm{E}(Y_j) = \mathrm{E}\{E\left(Y_j \mid Y_{j-1}\right)\} = \mu_j$. Furthermore, from the variance decomposition formula and (2.1)

$$\begin{aligned}
{\sigma_j}^2 &= \mathrm{E}\{\mathrm{var}(Y_j \mid Y_{j-1})\} + \mathrm{var}\{\mathrm{E}(Y_j \mid Y_{j-1})\} \\
&= \mathrm{E}\{\mathrm{var}(Y_j \mid Y_{j-1})\} + \mathrm{var}\{\mu_j + C_{j-1,j}\sigma_j/\sigma_{j-1}(Y_{j-1} - \mu_{j-1})\} \\
&= \mathrm{E}\{\mathrm{var}(Y_j \mid Y_{j-1})\} + C^2_{j-1,j}{\sigma_j}^2. \qquad (2.3)
\end{aligned}$$

Solving (2.3) for ${\sigma_j}^2$ then yields (2.2), so that the proof is complete.

Theorem 2.2. *Consider random variables $Y_1, \ldots, Y_n$ with first order antedependence, so that each Y_j given the immediate antecedent Y_{j-1}, is independent of all further preceding variables (Gabriel 1962). Then, if $E(Y_j \mid Y_{j-1})$ $(j = 2, \ldots, n)$ have linear form (2.1), $corr(Y_j, Y_{j+t}) = C_{j,j+t}$ is a product of the adjacent correlations, so that*

$$corr(Y_j, Y_{j+t}) = \prod_{w=j}^{j+t-1} C_{w,w+1} \quad (j = 1, \ldots, n-1; t = 1, \ldots, n-j). \qquad (2.4)$$

Proof. We use induction to prove this result. For the first step,

$$\begin{aligned}
\mathrm{E}(Y_j Y_{j+2}) &= \mathrm{E}\left\{\mathrm{E}\left(Y_j, Y_{j+2} \mid Y_1, \ldots, Y_{j+1}\right)\right\} \\
&= \mathrm{E}\left\{Y_j \,\mathrm{E}\left(Y_{j+2} \mid Y_1, \ldots, Y_{j+1}\right)\right\} \\
&= \mathrm{E}\left\{Y_j \left(\mu_{j+2} + C_{j+1,j+2}\sigma_{j+2}/\sigma_{j+1}\left(Y_{j+1} - \mu_{j+1}\right)\right)\right\}.
\end{aligned}$$

Hence, $\text{cov}(Y_j, Y_{j+2}) = C_{j+1,j+2}\sigma_{j+2}/\sigma_{j+1}\text{cov}(Y_j, Y_{j+1})$, so that $\text{corr}(Y_j, Y_{j+2}) = C_{j,j+1}C_{j+1,j+2}$. Next, we assume that $\text{corr}(Y_j, Y_{j+k}) = \prod_{w=j}^{j+k-1} C_{w,w+1}$. Using a very similar argument as for the first step, it is straightforward to show that $\text{cov}(Y_j, Y_{j+k+1}) = C_{j+k,j+k+1}\sigma_{j+k+1}/\sigma_{j+k}\text{cov}(Y_j, Y_{j+k})$, so that $\text{corr}(Y_j, Y_{j+k+1}) = \prod_{w=j}^{j+k} C_{w,w+1}$ and the proof is complete.

It is also interesting to note that if the conditional expectations are linear and the correlations have product form (2.4), then the conditional expectations can be expressed as in (2.1). A proof is provided in Appendix A.1.

Different parameterizations $C_{w,w+1} = \alpha^{\theta_w}$ in (2.4) yield structures that were implemented by Nuñez-Antón and Woodworth (1994), Shults and Chaganty (1998), and Zimmerman and Nuñez-Antón (2010): $\theta_w = 1$ yields a first-order autoregressive structure that was also implemented for binary variables by Zeger et al. (1985) and Qaqish (2003); $\theta_w = t_{w+1} - t_w$ (where t_w is the timing of Y_w) yields a Markov structure; and $\theta_w = ({t_{w+1}}^{\gamma} - {t_w}^{\gamma})/\gamma$ yields a generalized Markov structure. Letting $C_{w,w+1} = \alpha_k$ yields an unstructured product correlation matrix that, in addition to the first-order autoregressive and Markov structures, was implemented for simulation and maximum likelihood based analysis of longitudinal Bernoulli data by Guerra et al. (2012). To achieve positive-definite matrices, the following restrictions must be satisfied: $-1 < \alpha < 1$ for the AR(1); $0 < \alpha < 1$ and $t_{k+1} - t_k \geq 1$ $(k = 1, \cdots, n-1)$ for the Markov; $0 < \alpha < 1$ and $\gamma > 0$ for the generalized Markov; and $0 < \alpha_k < 1$ $(k = 1, \cdots, n-1)$ for the AD(1) structure.

2.2 Constructed likelihoods

We construct joint distributions of $Y_1, \ldots, Y_n$ for specified marginal means $\mu_1, \ldots, \mu_n$ and adjacent correlations $C_{1,2}, \ldots, C_{n-1,n}$, assuming first-order antedependence, linearity of the conditional expectations, and the same distribution for Y_1 and for Y_j *given* Y_{j-1} $(j = 2, \ldots, n)$. The details for each distribution are provided in the Appendix.

Conditional Binomial: Specify the distribution of Y_1 as binomial with $\mu_1 = N_1 p_1$, so that ${\sigma_1}^2 = N_1 p_1 q_1$, where $q_1 = 1 - p_1$. Then, the *conditional* distribution of Y_j given Y_{j-1} is specified as binomial with mean given by (2), with $\mu_j = N_j p_j$, and ${\sigma_j}^2$ as defined in Equation (2) (j = 2,...,n). For this distribution,

$$ {\sigma_j}^2 = N_j p_j q_j / \{1 + C^2_{j-1,j} \, (1 - N_j)/N_j\} \quad (j = 2, \ldots, n), \tag{2.5} $$

where $q_j = 1 - p_j$; the Y_j are therefore over-dispersed relative to the binomial distribution if $N_j > 1$, and $C_{j-1,j} \neq 0$, because in this case ${\sigma_j}^2 = \phi_j N_j p_j q_j$, where $\phi_j > 1$. Also, note that ${\sigma_j}^2 > 0$ if $-1 < C_{j-1,j} < 1$ in (2.5). The constructed distribution will be valid if N_j, p_j and $C_{j-1,j}$ satisfy the following: N_j is an integer ≥ 1; $0 < p_j < 1$ $(j = 1, \ldots, n)$;

$$ 0 < N_j p_j + C_{j-1,j} \, N_{j-1} \, q_{j-1} \, \sigma_j / \sigma_{j-1} < N_j \quad (j = 2, \ldots, n); \tag{2.6} $$

$$ 0 < N_j p_j - C_{j-1,j} \, N_{j-1} \, p_{j-1} \, \sigma_j / \sigma_{j-1} < N_j \quad (j = 2, \ldots, n); \tag{2.7} $$

and $C_{j-1,j}$ $(j = 2, \ldots, n)$ satisfy the constraints required to achieve a positive definite correlation matrix.

For the *conditional Bernoulli* distribution $(N_j = N_{j-1} = 1; j = 2, \ldots, n)$, there is no overdispersion, and (2.6) and (2.7) reduce to the constraints for the bivariate Bernoulli distribution (Prentice 1988, p. 1046) .

Conditional Poisson: The distribution of Y_1 is specified as Poisson with $\mu_1 = \lambda_1$ and ${\sigma_1}^2 = \lambda_1$. Then, the *conditional* distribution of Y_j given Y_{j-1} is specified as Poisson with conditional mean given by (2.1), and ${\sigma_j}^2$ as defined in Equation (2) (j = 2,...,n). For this distribution,

$$ {\sigma_j}^2 = \lambda_j/(1 - C^2_{j-1,j}) \quad (j = 2, \dots, n); \tag{2.8} $$

the Y_j are therefore overdispersed relative to the Poisson distribution if $C_{j-1,j} \neq 0$, because in this case ${\sigma_j}^2 = \phi_j \lambda_j$, where $\phi_j > 1$. The constructed distribution will be valid if $\lambda_j \geq 0$ $(j = 1, \dots, n)$;

$$ \lambda_j - \lambda_{j-1} C_{j-1,j} \sigma_j / \sigma_{j-1} > 0 \quad (j = 2, \dots, n); \tag{2.9} $$

and $C_{j-1,j}$ $(j = 2, \dots, n)$ satisfy the constraints required to achieve a positive definite correlation matrix.

2.3 Simulation approach

The following algorithm can be easily applied to simulate realizations $y_1, \dots, y_n$ of $Y_1, \dots, Y_n$ with a joint distribution of the type described in Section 2.2.

Step One: Specify a particular distribution for Y_1 and for Y_j given Y_{j-1} $(j = 2, \dots, n)$. *Step Two:* Specify marginal means $\mu_1, \dots, \mu_n$ and adjacent correlations $C_{1,2}, \dots, C_{n-1,n}$. As shown in Theorem 2.2, different choices for the adjacent correlations $C_{j-l,j}$ in (2.1) will induce different product correlation structures. *Step Three:* Check that the specified marginal means and adjacent correlations satisfy the necessary constraints for the assumed distributions. If not, change the values of the marginal means, or choose correlations that are closer to zero. *Step Four:* Simulate a realization from Y_1 from the specified distribution for Y_1 and then from Y_j given Y_{j-1} ($j = 2, \ldots, n$).

To obtain longitudinal data that comprise repeated measurements on each of m independent subjects, the algorithm can be applied successively to obtain n_i measurements on subject i ($i = 1, \ldots, m$). Covariates can also be incorporated in the definition of the marginal means. For example, for $\mu_j = N_j p_j$ (conditional Binomial), we might specify a logistic model with $\text{logit}(p_j) = x_j^{'}\beta$ for covariates x_j and corresponding regression parameter β. Or, for $\mu_j = \lambda_j$ (conditional Poisson), we might specify $\lambda_j = \exp(x_j^{'}\beta)$.

2.4 An Example of a Constructed Distribution

The simulation approach *does not* require the enumeration of all possible realizations of the random variables and the probability of each realization. However, it is instructive to demonstrate the construction of one joint distribution. We construct the joint distribution of Y_1, Y_2, Y_3, assuming the *conditional binomial* distribution, with marginal means $\mu_1 = 2.4$ (for $N_1 = 3$ and $p_1 = 0.8$); $\mu_2 = 0.4$ (for $N_2 = 1$ and $p_2 = 0.4$); and $\mu_3 = 0.6$ (for $N_3 = 2$ and $p_3 = 0.3$). In addition, the AD(1) structure is specified, with adjacent correlations $C_{1,2}$=0.2 and $C_{2,3} = 0.3$. These values satisfy the constraints provided in (2.6) and (2.7). Then, since the assumed distribution of Y_1 is binomial, ${\sigma_1}^2 = N_1 p_1 q_1 = 0.48$. Next, using (2.5), ${\sigma_2}^2 = 0.24$ and ${\sigma_3}^2 = .43979058$. Next, Y_j given Y_{j-1} are assumed to be binomial with $\text{E}(Y_j|Y_{j-1}) = N_j {p_j}^*$ calculated using (2.1), so that ${p_j}^* = 1/N_j\,[\mu_j + C_{j-1,j}\sigma_j/\sigma_{j-1}(Y_{j-1} - \mu_{j-1})]$ for $j = 2, 3$. Table E1 provided in Appendix A.4 lists all possible realizations of (Y_1, Y_2, Y_3) and the associated probabilties $pr(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) =$

$pr(y_1, y_2, y_3) =$

$$\binom{N_1}{y_1} {p_1}^{y_1} {q_1}^{N_1 - y_1} \binom{N_2}{y_2} {p_2}^{*y_2} {q_2}^{*N_2 - y_2} \binom{N_3}{y_3} {p_3}^{*y_3} {q_3}^{*N_3 - y_3}. \tag{2.10}$$

In Appendix A.4, we also verify that this constructed distribution is valid; furthermore, by summing over the appropriate functions of $pr(y_1, y_2, y_3)$, we do indeed obtain the assumed values for the marginal means and adjacent correlations, in addition to the values of ${\sigma_j}^2$ (for $j = 1, 2, 3$) and $\mathrm{corr}(Y_j, Y_k)$ (for $j = 1, 2, 3$ and $k = 1, 2, 3$) that we expect based on Theorem 2.1 and Proposition 2.2, respectively.

3. DEMONSTRATION

We now demonstrate the proposed approach to estimate the power to detect a difference between two treatment groups over time. Our earlier notation is readily generalized for longitudinal data that comprise realizations y_{ij} of ordered discrete random variables Y_{ij} on subject i ($j = 1, \ldots, n_i$). We assume the marginal means $\mathrm{E}(Y_{ij}) = \mu_{ij}$ are a function of $x'_{ij}\beta = \eta_{ij}$, where

$$\eta_{ij} = \beta_0 x_{ij1} + \beta_1 x_{ij2} + \beta_2 x_{ij3} + \beta_3 x_{ij4}, \tag{3.1}$$

where $x'_{ij} = (x_{ij1},\, x_{ij2},\, x_{ij3},\, x_{ij4})$; $x_{ij1} = 1$; x_{ij2} is an indicator variable for treatment group, which equals 1 for subjects treated with a treatment A and 0 for treatment B; x_{ij3} represents time, which will vary for different examples; and x_{ij4} is the time by treatment interaction that represents the product of x_{ij2} and x_{ij3}. The interaction term β_3 is of primary interest, because if it differs significantly from zero then this indicates that the change over time in the marginal means differs significantly between the two treatment groups.

We consider the following data types, true correlation structures, and specified values for time: (i) *Conditional Poisson,* with $\mu_{ij} = exp(\eta_{ij})$, an AR(1) structure, and $x_{ij3} = j$ for $j = 1, \ldots 6$; (ii) *Conditional Binomial with all* $N_{ij} = 1$, with $logit(\mu_{ij}) = \eta_{ij}$, a Markov structure, and $x_{ij3} = j$ for $j = 1, 2, 3$ and $x_{ij3} = (j-2) \times 3$ for $j = 4, 5, 6$; (iii) *Conditional Binomial,* with $logit(\mu_{ij}/N_{ij}) = \eta_{ij}$ and $N_{ij} = 4$, an AD(1) structure, and the same timings used for simulation of Bernoulli data. We specified identical timings for the Markov and AD(1) structures, so that the Markov structure is a special case of the AD(1) structure, and is a correctly specified working structure when the true structure is AD(1).

For each simulation scenario, we simulated 10000 data sets using our software in R and Stata to compare quasi-least squares (QLS), a method in the framework of GEE that allows for easy implementation of the Markov structure (Shults and Chaganty 1998; Chaganty and Shults 1999), with application of GEE when the working structure is an identity matrix but the standard errors are adjusted for misspecification of the correlation structure via application of a "sandwich" covariance matrix for estimation of the covariance matrix of $\widehat{\beta}$. GEE was implemented using geepack in R (Halekoh, Hjsgaard, and Yan 2006) and using xtgee in Stata, while QLS was implemented using the qlspack package in R and xtqls in Stata (Shults, Ratcliffe, and Leonard 2007).

There were no simulation runs that resulted in a failure to converge for either approach. Therefore, the power to test the hypothesis $\beta_3 = 0$ with type-one error of 0.05 was estimated as the proportion of 10000 simulation runs that resulted in a p-value less than 0.05 (based on Wald's test as implemented in each software package). Simulations were duplicated in both Stata and R, with the exception of the conditional binomial example for which QLS and GEE were only implemented in Stata, owing to the inability of qlspack and geepack to fit a binomial model with $N_j > 1$. Assessment of power for these two approaches allows us to compare correct specification of the marginal means and correlation structure with ignoring the correlations, but adjusting for misspecification of the correlation structure via application of a sandwich covariance matrix. We specified a sandwich covariance matrix for each approach, and also correctly specified the mean and link functions that relate the mean and variance for each distribution, with one important exception- we ignored the overdispersion that is present for all data types except the Bernoulli. As described in Efron (1992), overdispersion is a common feature of count data; therefore, simulating data with overdispersion is useful for assessing power under conditions that are likely to be encountered in practice.

Table 1 displays the results for two conditions, when β_3 differs from zero, and when it is identically zero; the latter set of simulation results are important to assess departures from a level of 0.05 for the test. Table 1 indicates that correctly modeling the correlation structure with QLS yields a small gain in power (that decreases as the sample size increases) over fitting GEE with an identity working structure, but with adjusted standard errors. For example, for group sizes of 20, the power for QLS versus GEE was 65.4 % versus 60.5 %, respectively; however, for group sizes of 80, the power was almost identical for QLS versus GEE (99.7 % versus 99.4 %, respectively). This suggests that for smaller samples it can be important to correctly model the correlation structure, because even a small improvement in power that allows us to reduce the sample size by a several subjects, can yield considerable savings over the course of a clinical trial that involves expensive tests and monitoring of the participants. The upper constraint for α displayed directly beneath Table 1 were obtained using a grid search and (2.6) and (2.7) for the conditional binomial and conditional Bernoulli, and a grid search and (2.9) for the conditional Poisson. Other results (including estimation of percentage bias and mean-square error of the regression and correlation parameters) are available on request.

4. DISCUSSION

The proposed algorithm for simulating overdispersed random variables with specified marginal means and product correlations is straightforward to implement, even for an increasingly large number of random variables. The method constructs a likelihood for $Y_1, \ldots, Y_n$ based on assumptions of first-order antedependence, the same distribution for Y_1 and for Y_j given Y_{j-1}, and linearity of the conditional expectations $E(Y_j|Y_{j-1})$. The key is to select a conditional distribution for Y_j given Y_{j-1} whose conditional expectation coincides with the best linear predictor (Christensen 1997, p.108) of Y_j given Y_{j-1} $(j = 2, \ldots, n)$.

The algorithm requires specification of the marginal means and adjacent intra-subject correlations $C_{j-1j}(\alpha)$, which induces in the discrete random variables a decaying-product correlation structure that has been thoroughly studied for continuous outcomes (Zimmerman and Nuñez-Antón 2010). The

Table 1

Estimated power for testing the hypothesis $\beta_3 = 0$ and for several data types, true correlation structures, and group sizes (m/2), for the model defined in (3.1) when $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)' = (1, 0.1, -0.1, -0.1)'$ and $n_i = 6$ for i=1, ..., m. To estimate power when $\beta_3 = 0$, we also considered $\beta = (1, 0.1, -0.1, 0.0)'$. The data types considered are conditional Poisson, conditional binomial, and Bernoulli. The true correlation structures are AR(1) (with $\alpha = 0.65$) for the conditional Poisson, Markov (with $\alpha = 0.55$) for the conditional binomial, and AD(1) (with $\alpha = (0.70, 0.70, 0.70, 0.343, 0.343, 0.343)'$ for the Bernoulli. The simulated AD(1) structure is identical to a Markov structure with $\alpha = 0.7$ for this example. The working correlation structures were correctly specified for QLS in the qlspack package in R and the xtqls command in Stata, respectively. GEE with an identity working structure was implemented in the xtgee command in Stata and in the geepack package in R.

		Overdispersed Poisson Data[a]		Bernoulli Data[b]		Overdispersed Binomial Data[c]	
β_3	$m/2$	GEE-IND	QLS-AR1	GEE-IND	QLS-MARK	GEE-IND	QLS-MARK
-0.1	20	0.252	0.272	0.220	0.227	0.605	0.654
-0.1	30	0.328	0.362	0.289	0.301	0.773	0.808
-0.1	50	0.483	0.526	0.424	0.446	0.939	0.957
-0.1	80	0.665	0.719	0.624	0.652	0.994	0.997
-0.1	120	0.830	0.875	0.795	0.820	0.999	1.000
0	20	0.067	0.064	0.060	0.062	0.058	0.062
0	30	0.061	0.062	0.053	0.054	0.053	0.056
0	50	0.058	0.057	0.051	0.054	0.057	0.058
0	80	0.056	0.056	0.051	0.052	0.051	0.049
0	120	0.049	0.050	0.052	0.053	0.050	0.051

[a]The largest value for α that will yield a valid distribution for the assumed marginal means is 0.6709.
[b]The largest value for α that will yield a valid distribution for the assumed marginal means is 0.5959.
[c]The largest value for α that will yield a valid distribution for the assumed marginal means is 0.7408.

decaying-product structure includes several structures as special cases that are plausible for the analysis of longitudinal data, including the AR(1), Markov, generalized Markov, and AD(1). However, in contrast to other available methods for simulation of binary data (Emrich and Piedmonte 1991; Qaqish 2003), our approach cannot be used to simulate data with a correlation structure that differs from the decaying-product form, including the equicorrelated structure that has been recommended for cross-sectional studies with binary "clustered" data (Chaganty and Joe 2004, p.858).

It is also interesting to note that the algorithm in Section 2.3 has a long history for the special case of Bernoulli data and an induced AR(1) structure. Zeger et al. (1985) implemented a maximum likelihood approach for estimation of the parameters for the *Conditional Binomial* distribution, for $C_{j-1j}(\alpha) = \alpha$; all $N_j = 1$; a logistic model for the marginal means; and time-independent covariates, so that $p_j = p$ within a subject. Zeger et al. (1985) did not mention that their assumed likelihood induces data with an AR(1) structure; however, Liang and Zeger (1986) noted that they made use of a Markov chain of order one with first lag autocorrelation α to simulate binary data for Table 2 of Liang and Zeger (1986), and therefore presumably implemented the algorithm in Section 2.3 to simulate binary data with an AR(1) structure. Qaqish (2003) did not discuss a general correlation structure with form (2.4), but did consider the AR(1) structure and obtained the conditional mean in (6) of Qaqish (2003) that determines the same likelihood (but with time-varying covariates) that was considered by Zeger et al. (1985). Jung and Ahn (2005) proposed a simple method for simulation of data with an AR(1) structure that also follows from the likelihood assumed by Zeger et al. (1985). In addition, as noted earlier, if we start with an assumed product correlation structure and assumed conditional expectations that are also the best linear predictors (Christensen 1997), then the conditional expectations will be of form (2.1).

Our approach is also similar to the method of Azzalini (1994) that assumes first-order antedependence and can be applied to generate realizations of Bernoulli random variables with specified marginal means and association parameters. Heagerty (2002) extended the approach of Azzalini (1994) to allow for higher-order antedependence. However, Azzalini (1994) and Heagerty (2002) modeled association via pairwise odds-ratios, while we model the association via correlations, which allows for simulation of data with decaying product correlations and has a more natural extension for discrete data that are not binary.

Future work might focus on constructing additional likelihoods under assumptions of the first order Markov property and linearity of the expectations of the conditional distributions. Plans are also underway to implement the proposed likelihoods for analysis of longitudinal discrete data.

References

Azzalini, A. (1994), "Logistic regression for autocorrelated data with application to repeated measures," *Biometrika,* 84, 767–775.

Chaganty, N.R., and Joe, H. (2004), *Efficiency of generalized estimating equations for binary responses. Journal of the Royal Statistical Society, Series B,* 66, 851–860.

Chaganty, N.R., and Shults, J. (1999), "On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter," *Journal of Statistical Planning and Inference,* 76, 145–161.

Christensen, R. (1987), *Plane Answers to Complex Questions: the Theory of Linear Models* New York: Springer-Verlag.

Christensen, R. (1997), *Linear Models for Multivariate, Time Series, and Spatial Data.* New York: Springer-Verlag.

Devroye, L. (1986), *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

Efron, B. (1992), "Poisson overdispersion estimates based on the method of asymmetric maximum likelihood," *Journal of the American Statistical Association,* 87, 98–107.

Emrich, L.J., and Piedmonte, M.R. (1991), "A method for generating high-dimensional multivariate binary variates," *The American Statistician,* 45, 302–304.

Farrell, P.J., and Rogers-Stewart, K. (2008), "Methods for generating longitudinally correlated binary data," *International Statistical Review,* 76, 28–38.

Gabriel, K. R. (1962), Ante-dependence Analysis of an Ordered Set of Variables. *Annals of Mathematical Statistics,* 33, 201–212.

Gange, S. J. (1995), Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician,* 49, 134–138.

Guerra, M. W., Shults, J., Amsterdam, J., and Ten-Have, T. (2012), The analysis of binary longitudinal data with time-dependent covariates. *Statistics in Medicine,* 31, 1097–0258.

Halekoh, U., Hjsgaard, S. and Yan, J. (2006), "The R package geepack for generalized estimating equations," *Journal of Statistical Software,* 15, 1–11.

Heagerty, P. H. (2002), "Marginalized transition models and likelihood inference for longitudinal categorical data Ante-dependence Analysis of an Ordered Set of Variables," *Biometrics,* 58, 342–351.

Jung, S.H., and Ahn, W.W. (2005), "Sample size for a two-group comparison of repeated binary measurements using GEE," *Statistics in Medicine,* 24, 2583–2596.

Lee, A.J. (1993), "Generating random binary deviates having fixed marginal distributions and specified degrees of association," *The American Statistician,* 47 (3), 209–215.

Liang, K.Y., and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika,* 73, 13–22.

Nuñez-Antón, V. and Woodworth, G.G. (1994), "Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors," *Biometrics,* 50 (2), 445–456

Prentice, R.L. (1988), "Correlated binary regression with covariates specific to each binary observation," *Biometrics,* 44, 1033–1048.

Qaqish, B.F. (2003), "A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations," *Biometrika,* 90, 455–63.

Schulman, A., Gange, S.J., Ormsby, C.G., Johnston, R.G., Lienhard, C. W., and Jolliffe, I. T. (1996), "Letter to the Editor," *The American Statistician,* 50, 280–282.

Shults, J., and Chaganty, N.R. (1998), "Analysis of serially correlated data using quasi-least squares," *Biometrics,* 54, 1622–1630.

Shults, J., Ratcliffe, S., and Leonard, M. (2007), "Improved generalized estimating equation analysis via xtqls for implementation of quasi-least squares in Stata," *Stata Journal,* 7, 147–166.

Zeger, S.L., Liang, K.Y., and Self, S.G. (1985), "The analysis of binary longitudinal data with time-independent covariates," *Biometrika,* 72, 31–38.

Zimmerman, D.L. & Nuñez-Antón, V.A. (2010), *Antedependence Models for Longitudinal Data.* Boca Raton: Chapman and Hall/CRC Press.

A. APPENDIX

A.1 *Assumed Linear Expectations and Product Correlations*

Assume product correlations (2.4) and linear conditional expectations

$$\mathrm{E}\,(\,Y_j \mid H_{j-1}\,) = \mu_j + \sum_{k=1}^{j-1} b_{jk}\,(\,Y_k - \mu_k\,)\,, \tag{A.1}$$

where $H_{j-1} = (Y_1, \ldots, Y_{j-1})'$. Then, using results from the discussion of best linear prediction (Christensen 1987, Chapter 6) presented on p. 108 of Christensen (1997),

$$\Sigma[1 : j-1,\ 1 : j-1]B_j = \Sigma[1 : j-1,\ j], \tag{A.2}$$

where Σ is the assumed covariance matrix for $(Y_1, \ldots, Y_n)$ and $B_j = (b_{j1}, \ldots, b_{j\,j-1})'$.

Qaqish (2003) used the Choleski decomposition of $\Sigma[1 : j-1,\ 1 : j-1]$ to solve for B_j in (A.2), in order to construct multivariate distributions for binary variables. We utilize the simple tri-diagonal form (Zimmerman and Nuñez-Antón 2010) of the product covariance structure, to directly obtain

$$b_j = \Sigma[1 : j-1,\ 1 : j-1]^{-1}\Sigma[1 : j-1,\ j]. \tag{A.3}$$

The elements of $\Sigma^{-1}[1 : j-1\ 1 : j-1]$ are given by $\Sigma^{-1}[1,1] = 1/(\sigma_1^2(1 - C_{1,2}^2)$; $\Sigma^{-1}[k,k] = (1 - C_{k-1,k}^2\, C_{k,k+1}^2)/(\sigma_k^2(1 - C_{k-1,k}^2)(1 - C_{k,k+1}^2))$ for k $= 2, \ldots, j-2$; $\Sigma^{-1}[k, k+1] = -C_{k,k+1}/(\sigma_k\, \sigma_{k+1}(1 - C_{k,k+1}^2)$ for k $= 1, \ldots, j-2$; $\Sigma^{-1}[j-1, j-1] = 1/(\sigma_{j-1}{}^2(1 - C_{j-2,j-1}^2)$; and $\Sigma^{-1}[k, k'] = 0$ for $|k - k'| > 0$. In addition, $\Sigma[1 : j-1,\ j] = \Big(\sigma_1\sigma_j \prod_{k=1}^{j-1} C_{k,k+1},\ \ \sigma_2\sigma_j \prod_{k=2}^{j-1} C_{k,k+1},\ \ldots,$ $\sigma_{j-1}\sigma_j C_{j-1,j}\Big)'$. Substitution for $\Sigma[1 : j-1,\ 1 : j-1]^{-1}$ and $\Sigma[1 : j-1,\ j]$ in (A.3) and some algebra then yields $b_j = (0, \ldots, 0, b_{jj-1})'$ where $b_{jj-1} = \sigma_j/\sigma_{j-1}C_{j-1,j}$. Substituting b_j into (A.1) then yields $E\,(Y_j|H_j)$ with form (2.1), so that we have the result.

A.2 *Conditional Binomial*

We specify the distribution of Y_1 as binomial with $\mu_1 = N_1p_1$ and $\sigma_1{}^2 = N_1p_1q_1$, where $q_1 = 1 - p_1$. Then, the conditional distribution of Y_j given Y_{j-1} is specified as binomial with mean given by (2.1), with $\mu_j = N_jp_j$ and $\sigma_j{}^2$ $(j = 2, \ldots, n)$ as obtained using (2.2), as follows. First, $\text{var}(Y_j \mid Y_{j-1}) =$

$N_j p_j^*(1-p_j^*)$, where $p_j^* = p_j + b_j^*(Y_{j-1} - N_{j-1}p_{j-1})$ for $b_j^* = C_{j-1,j}\sigma_j/(\sigma_{j-1}N_j)$. We can then directly obtain $E\{\text{var}(Y_j \mid Y_{j-1})\}$, substitute its value into (2.2), and solve the resultant equation for ${\sigma_j}^2$ to obtain (2.5).

Next, in order for $E(Y_j \mid Y_{j-1})$ to be valid for the conditional binomial distribution, they must satisfy $0 < E(Y_j \mid Y_{j-1}) < N_j$ for $Y_{j-1} \in \{0, \ldots, N_{j-1}\}$. For $C_{j-1,j} > 0$ the maximum value of $E(Y_j \mid Y_{j-1})$ is obtained at $Y_{j-1} = N_{j-1}$ and the minimum value is obtained at $Y_{j-1} = 0$. For $C_{j-1,j} < 0$ the *minimum* value of $E(Y_j \mid Y_{j-1})$ is obtained at $Y_{j-1} = N_{j-1}$ and the *maximum* value is obtained at $Y_{j-1} = 0$. Since $0 < E(Y_j \mid Y_{j-1}) < N_j$ as long as $\min\{E(Y_j \mid Y_{j-1})\} > 0$ and $\max\{E(Y_j \mid Y_{j-1})\} < N_j$, we can easily check whether the constraints are satisfied for a particular set of parameter values by first calculating $E(Y_j \mid Y_{j-1} = N_{j-1})$ and $E(Y_j \mid Y_{j-1} = 0)$, which are provided in (2.6) and (2.7), respectively. We can then check whether (2.6) and (2.7) both take value between 0 and N_j recursively $(j = 2, \ldots, n)$.

A.3 *Conditional Poisson*

Here the distribution of Y_1 is specified as Poisson with $E(Y_1) = \mu_1 = \lambda_1$. Then, the conditional distribution of Y_j given Y_{j-1} is specified as Poisson with $\mu_j = \lambda_j$ and conditional mean given by (2.1) $(j = 2, \ldots, n)$. Then, since the mean and variance are identical for the Poisson distribution, $\text{E}\{\text{var}(Y_j \mid Y_{j-1})\} = \text{E}\{\text{E}(Y_j \mid Y_{j-1})\} = \lambda_j$; substitution into (2.2) then yields ${\sigma_j}^2$ in (2.8).

In order for the conditional expectations $E(Y_j \mid Y_{j-1})$ to be valid for the conditional Poisson distribution, they must satisfy $E(Y_j \mid Y_{j-1}) > 0$ for $Y_{j-1} \geq 0$. In order for this inequality to be satisfied for all $Y_{j-1} \geq 0$ we must specify $C_{j-1,j} \geq 0$; then the minimum value of $E(Y_j \mid Y_{j-1})$ is obtained at $Y_{j-1} = 0$. Since $E(Y_j \mid Y_{j-1}) > 0$ as long as $\min\{E(Y_j \mid Y_{j-1})\} > 0$, substituting $Y_{j-1} = 0$ yields the constraints (2.9) that must be satisfied in order for the conditional Poisson distributions to be valid.

A.4 *Example of a Constructed Distribution*

[Table 1 about here.]

Using the probabilities displayed in Table E1, it is straightforward to verify that

$$\sum_{y_1}\sum_{y_2}\sum_{y_3} pr(y_1, y_2, y_3) = 1,$$

so that the constructed distribution is valid. We can then show that

$$\sum_{y_1}\sum_{y_2}\sum_{y_3} y_j\, pr(y_1, y_2, y_3) \quad = \quad \mu_j \quad (j = 1,\, 2,\, 3),$$

where $\mu_1 = 2.4$, $\mu_2 = 0.4$, and $\mu_3 = 0.6$. Furthermore,

$$\sum_{y_1}\sum_{y_2}\sum_{y_3} {y_j}^2\, pr(y_1, y_2, y_3) - {\mu_j}^2 \quad = \quad {\sigma_j}^2 \quad (j = 1,\, 2,\, 3),$$

where ${\sigma_1}^2 = 0.48$, ${\sigma_2}^2 = 0.24$, and ${\sigma_3}^2 = 0.43979058$. Finally, we can verify that

$$\left(\sum_{y_1}\sum_{y_2}\sum_{y_3} y_j y_k\, pr(y_1, y_2, y_3) - \mu_j\, \mu_k\right) / \left(\sigma_j\, \sigma_k\right) \quad = \quad corr(Y_j,\, Y_k),$$

where $corr(Y_1,\, Y_2) = 0.20 = C_{1,\,2}$; $corr(Y_2,\, Y_3) = 0.30 = C_{2,\,3}$; and $corr(Y_1,\, Y_3) = 0.06 = C_{1,\,2}\, C_{2,\,3}$. The constructed distribution therefore has the expected properties, based on Theorem 2.1 and Proposition 2.2, respectively.

Table AE1

Example of a Constructed Distribution of Y_1, Y_2, Y_3.

y_1	y_2	y_3	$pr(y_1, y_2, y_3)$
0	0	0	0.00458663
0	0	1	0.00256895
0	0	2	0.00035971
0	1	0	0.00016203
0	1	1	0.00023643
0	1	2	0.00008625
1	0	0	0.04675376
1	0	1	0.02618654
1	0	2	0.00366674
1	1	0	0.00648266
1	1	1	0.00945949
1	1	2	0.00345082
2	0	0	0.15387185
2	0	1	0.08618283
2	0	2	0.01206764
2	1	0	0.04408390
2	1	1	0.06432721
2	1	2	0.02346656
3	0	0	0.16097156
3	0	1	0.09015935
3	0	2	0.01262445
3	1	0	0.08298290
3	1	1	0.12108862
3	1	2	0.04417312