Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Simulating longer vectors of correlated binary random variables via multinomial sampling

Justine Shults

Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, 423 Guardian Drive, 635 Blockley Hall, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Received 22 June 2016
Received in revised form 25 March 2017
Accepted 6 April 2017
Available online xxxx

Keywords:
Binary random variables
Generalized estimating equations
Multinomial sampling
Simulation

ABSTRACT

The ability to simulate correlated binary data is important for sample size calculation and comparison of methods for analyzing clustered and longitudinal data with dichotomous outcomes. One available approach for simulating vectors of length n of dichotomous random variables is to sample them from multinomial distribution of all possible length n permutations of zeros and ones. However, the multinomial sampling method has only been implemented in a general form (without making the initial restrictive assumptions) for vectors of length 2 and 3 because constructing multinomial distribution is very challenging for longer vectors. This difficulty can be overcome by presenting an algorithm for simulating correlated binary data via multinomial sampling that can be easily used for directly computing the multinomial distribution for any value of n. To demonstrate the approach, vectors of length 4 and 8 are simulated for assessing the power during the planning phase of a study and for evaluating the choice of working correlation structure in an analysis with generalized estimating equations.

1. Introduction

Methods to simulate realizations of dependent variables with specified marginal means and pairwise correlations are useful to assess semi-parametric approaches such as generalized estimating equations (GEE) (Liang and Zeger, 1986), which only require models for the first two moments of the distribution of the outcome variable. Continuous variables can be simulated via multivariate normal distribution that is determined by its mean and covariance matrix. In contrast, dependent Bernoulli random variables present a greater simulation challenge due to the lack of an equally general and flexible equivalent of the normal distribution for discrete data.

Several useful methods have been proposed; however, the best way to simulate correlated binary data remains an active area of research in statistical literature. Bahadur (1961) developed an elegant representation of a correlated Bernoulli distribution with specified marginal means and second- and higher-order correlations. However, complex constraints on the correlations limit the use of Bahadur's representation for the simulation of shorter vectors (Fitzmaurice and Molenberghs, 2009), or after simplifying the assumptions such as setting all third and higher order correlations to zero. Farrell and Rogers-Stewart (2008) reviewed methods to simulate correlated binary data, including an approach by Qaqish (2003) that allows for unstructured correlations and non-stationary data. Emrich and Piedmonte (1991) proposed a flexible method for simulation that is slightly complex because it involves solving non-linear equations via numerical integration. Al Osh and Lee (2001) proposed an approach that relies on the association among random variables resulting from the sharing of some common

E-mail address: jshults@mail.med.upenn.edu.

http://dx.doi.org/10.1016/j.csda.2017.04.002
0167-9473/© 2017 Published by Elsevier B.V.

components that induce correlation. Recently, Preisser and Qaqish (2014) compared the approach proposed by Qaqish (2003) with a method based on multivariate probit distribution.

This manuscript proposes the simulation of Bernoulli random variables with specified marginal means and pairwise correlations via the multinomial sampling approach that considers "k-variate binary data as a multinomial distribution with 2^k possible outcomes" (Kang and Jung, 2001). For paired data, Kang and Jung's multinomial sampling approach is a special case of the simulation method proposed for bivariate binomial data by Hamdan and Nasro (1986, p. 751). Recently, Haynes et al. (2015) showed that multinomial sampling was comparable with the method proposed by Emrich and Piedmonte (1991), which Haynes et al. (2015) referred to as the gold-standard approach. (However, Haynes et al., 2015 also acknowledged that they did not compare it with the approach proposed by Qaqish, 2003.)

Because the multinomial sampling approach involves complete specification of the underlying distribution of all possible permutations of zeros and ones, its use ensures that a valid multivariate parent distribution exists that is compatible with specified marginal means and covariance matrix. The lack of a compatible parent is not typically a concern for continuous variables because multivariate normal distribution is a possible valid parent even if it does not fit the data well. However, for discrete random variables there is no guarantee that a valid distribution exists for a given marginal mean and covariance pair (Chaganty and Joe, 2006). As discussed in Molenberghs (2010), although not fully specified, the parent distribution provides an estimation framework and probabilistic basis for semi-parametric methods such as GEE or the quasi-least squares approach (Chaganty, 1997; Shults and Chaganty, 1998; Chaganty and Shults, 1999, QLS) in the framework of GEE.

From a practical perspective, Rochon (1998) cautioned that the additional constraints necessary to ensure a valid parent distribution should be evaluated during the planning phase of a study. For example, consider sample size calculation with a standard formula, such as the one provided in Diggle et al. (2002, p. 167). If means and correlations with no valid parent distribution are used, the formula will provide results; however, the results will be invalid and no warning will be provided. Assessing power using a method that simulates from a compatible parent distribution will ensure that the results are appropriate. Shults et al. (2009) suggested that a *severe* violation of constraints during the analysis could be considered as a rule-out criterion for the selection of a working correlation structure to describe the pattern of association in the data.

Although the multinomial sampling approach is useful for assessing the parent distribution, it has not been implemented for vectors of length four or more, without first simplifying assumptions such as the first-order Markov property (Shults et al., 2006, p. 13) or the exchangeability condition (Kang and Jung, 2001, Section 5). As explained by Haynes et al. (2015) (who simulated vectors of length 2 and 3), "the CDF for establishing decision rules becomes complicated for cases of four or more repeated measures. While not impossible, constructing higher order joint probabilities can be computationally challenging". To overcome the difficulty described by Haynes et al. (2015), this paper presents an easy to implement algorithm for constructing higher order joint probabilities. The method is described in Section 2 and demonstrated in Section 3 for assessment of power and selection of a working correlation structure for GEE.

2. Methods

2.1. Notation and assumptions

Let $\underline{Y}_n = (Y_1, \ldots, Y_n)'$ be an $n \times 1$ vector of Bernoulli random variables Y_j with marginal means $E(Y_j) = P(Y_j = 1) = p_j$ and variances $\text{Var}(Y_j) = p_j q_j$, where $q_j = 1 - p_j$ $(j = 1, \ldots, n)$. Let $R_n = \text{Corr}(\underline{Y}_n)$ be the $n \times n$ correlation matrix of $\underline{Y}_n$, with (j, k)th entry as follows:

$$R_n[j, k] = \rho_{jk} = \frac{p_{jk} - p_j p_k}{\sqrt{p_j p_k q_j q_k}},$$

where $p_{jk} = E(Y_j Y_k) = P(Y_j = 1, Y_k = 1)$.

Let $mod(x, y)$ represent the modulus of x with respect to y, such that

$$mod(x, y) = x - y\,\text{floor}(x/y),$$

where $\text{floor}(x/y)$ is an unique integer n such that $n \le x/y < n + 1$. Let $B_n(i)$ be the length n binary representation of $i - 1$ that is expressed as an $n \times 1$ vector $(i = 1, \ldots, 2^n)$. For example, $B_3(2) = \left(1, 0, 0\right)'$ is the length 3 binary representation of 1 because $1 = 1 \times 2^0 + 0 \times 2^1 + 0 \times 2^2$. Furthermore, let $P_n(i) = P(\underline{Y}_n = B_n(i))$ $(i = 1, \ldots, 2^n)$.

2.2. Construction of multinomial distribution for Bernoulli vectors of length n

Kang and Jung (2001) proposed an algorithm for simulating dependent Bernoulli random variables $\underline{Y}_n$ via multinomial sampling, which can be described as follows with slightly different notation. To simulate a sample of size m, alternate m times between the following two steps. *Step One:* Simulate a value U from a uniform $(0, 1)$ distribution. *Step Two*: Select sequence $B_n(i)$ if $Z_n(i-1) \le U < Z_n(i)$, where $Z_n(0) = 0$ and $Z_n(i) = \sum_{j=0}^{i} P_n(j)$ for $i = 1, \ldots, 2^n$. The probability that $B_n(i)$ is selected in Step Two is equal to the length of the interval $[Z_n(i-1), Z_n(i)\rangle = Z_n(i) - Z_n(i-1) = P_n(i)$ $(i = 1, \ldots, 2^n)$.

This algorithm is easy to implement for $n = 2$ because the marginal means p_1, p_2 and the pairwise correlation ρ_{12} can be used to easily construct the well-known bivariate Bernoulli distribution, for which $P_2(1) = q_1 q_2 + \rho_{12}\sqrt{p_1 q_1 p_2 q_2}$; $P_2(2) = p_1 q_2 - \rho_{12}\sqrt{p_1 q_1 p_2 q_2}$; $P_2(3) = q_1 p_2 - \rho_{12}\sqrt{p_1 q_1 p_2 q_2}$; and $P_2(4) = p_1 p_2 + \rho_{12}\sqrt{p_1 q_1 p_2 q_2}$. Kang and Jung (2001)

Please cite this article in press as: Shults, J., Simulating longer vectors of correlated binary random variables via multinomial sampling. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.04.002