A random effects model for simulating clustered binary data

Michael Parzen
Emory University

Submitted to Statistics and Computing
Summary

In this paper, we propose a new random effects model for clustered binary data. This model adds the random effect to the marginal probability of success for each binary response in the cluster. We show that the beta-binomial is a special case of our model. Our random effects model may be particularly useful for simulating correlated binary data, in which the marginal model is a logistic regression model.

Key words: correlated binary data, marginal model
1 Introduction

Repeated measures studies have become very popular in applied statistics. In repeated measures studies, the basic sampling unit is a cluster of subjects; a measurement is made on each subject within the cluster. The observations within the cluster, one from each subject, constitute the “repeated measures” on the cluster. For example, in a toxicological study, the cluster is a litter, and the fetuses or newborns are the subjects within the cluster; in a genetic study, the cluster is the family, and the members of the same family are the subjects within the cluster. In repeated measures studies, we expect the repeated measurements on the cluster to be correlated. In this paper, the response is binary for each subject within the cluster. For example, consider the data in Table 1, which is based on data from an Eastern Cooperative Oncology Group liver cancer clinical trial (Falkson et al., 1994), where, for confidentiality, the real data cannot be released. The clinical trial compared three doses (0=low, 1=medium, 2=high) of a chemotherapy for the treatment of advanced liver cancer with respect to the binary outcome ‘alive at one year from randomization’ (yes or no); another covariate of interest is cardiac status at baseline (normal, abnormal). In Table 1, there are a total of 21 patients from 6 hospitals. In this setting the hospital can be thought of as clusters and patients as units within a cluster.

Several random effects models have been proposed for analyzing clustered binary data. In the usual random effects logistic regression model (Stiratelli et al, 1984; Anderson and Aitken, 1985), for each subject in the cluster, a random cluster effect is added to the linear part (covariates times the fixed parameter effects) of the logistic regression model. In our model proposed here, for a subject in a cluster, we add a random cluster effect to the marginal probability of success. In particular, the random cluster effect is not
added to the linear part of the logistic regression, but instead is added to the marginal probability of success for the individual in the cluster. This type of model is called a ‘marginal model’ because the marginal expectation of each binary outcome follows a logistic regression; in the usual random effects logistic regression model of Stiratelli et al, (1984), the marginal expectation of each binary outcome does not follow a logistic regression. We note that McDonald (1994) has proposed a different random effects model which also adds the random effect to the probability of success instead of the linear part of the logistic regression.

Our random effects model may be particularly useful for for simulating correlated binary data. In particular, there has been a great deal of research in the last 15 years on semi-parametric methods for estimating parameters for ‘marginal models’ for correlated binary data. Papers describing the methods include Liang and Zeger (1986), Prentice (1988), Zhao and Prentice (1990), Lipsitz, et al. (1991), Liang, et al. (1992), Carey, et al. (1993), Fitzmaurice and Laird (1993), Molenberghs and Lesaffre (1994), Fitzmaurice and Lipsitz (1995), and Lipsitz, et al. (1995). To determine the finite sample properties of all of these estimators, simple methods to simulate correlated data with logistic regression margins are needed. Thus, one can simulate from our random effects model to determine the finite sample properties of these estimators. Section 2 defines the notation. Section 3 defines the constraints on the correlation for our model and illustrates the simulation algorithm with the cancer clinical trial example given above. Section 4 discusses extensions.
2 Notation

Suppose, in the $i^{th}$ cluster ($i = 1, \ldots, N$), there are $n_i$ subjects ($j = 1, \ldots, n_i$). For the $j^{th}$ subject in cluster $i$, the response is binary, i.e., we let $Y_{ij} = 1$ if the $j^{th}$ subject in cluster $i$ has response 1 (say, success), and $Y_{ij} = 0$ otherwise. Subject $j$ in cluster $i$ also has a $(K \times 1)$ covariate vector $x_{ij}$. The $n_i$ responses of the subjects from the same cluster may be correlated, but the clusters are independent. In a random effects model, the correlation among members of a cluster is induced by adding a common random effect to the model for $Y_{ij}$ for observations in the same cluster.

In the usual random effects logistic regression model (Stiratelli et al, 1984; Anderson and Aitken, 1985), the conditional mean of $Y_{ij}$, given the cluster, is

$$p_{ij} = E(Y_{ij}|x_{ij}, a_i) = \Pr(Y_{ij} = 1|x_{ij}, a_i) = \frac{e^{a_i + x_{ij}'\beta}}{1 + e^{a_i + x_{ij}'\beta}},$$

(1)

where $a_i$ is a random cluster effect with mean 0, often assumed to be $N(0, \sigma_a^2)$, and $\beta$ is the vector of fixed covariate effects. In this random effects model, given $a_i$ (and thus $p_{ij}$), the $Y_{ij}$’s in the cluster are independent Bernoulli random variables, i.e.,

$$Y_{ij}|a_i \sim \text{Bern}(p_{ij}).$$

An element of the parameter vector $\beta$ represents the log-odds ratio for a one unit change in the covariate, given $a_i$ and the other covariates. Note, given $a_i$, this log-odds ratio is the same for all subjects in all clusters. However, the marginal probability of success,

$$\Pr(Y_{ij} = 1|x_{ij}) = \int \Pr(Y_{ij} = 1|x_{ij}, a_i) f(a_i) da_i = \int \left( \frac{e^{a_i + x_{ij}'\beta}}{1 + e^{a_i + x_{ij}'\beta}} \right) f(a_i) da_i,$$

is not of logistic form, where $f(a_i)$ is the distribution of $a_i$. Thus, this random effects logistic regression model has a nice conditional (on $a_i$) interpretation, but a complicated marginal interpretation.
We propose a random effects model that uses similar logic to the above random effects logistic regression, except that the formulation of $p_{ij}$ is different than in (1). In our random effects model, we let

$$p_{ij} = \Pr(Y_{ij} = 1|x_{ij}, a^*_i) = \pi_{ij} + a^*_i \sqrt{\pi_{ij}(1 - \pi_{ij})},$$

(2)

where

$$\pi_{ij} = \frac{e^{x_{ij}'\beta^*}}{1 + e^{x_{ij}'\beta^*}};$$

(3)

and $a^*_i$ is a random cluster effect with mean $E(a^*_i) = 0$ and variance $Var(a^*_i) = \rho$ (for $0 \leq \rho \leq 1$). As before, in this random effects model, given $a^*_i$, the $Y_{ij}$’s in a cluster are independent Bernoulli random variables with probabilities $p_{ij}$. Also, as shown in Appendix 1, $\rho$ is the marginal correlation between $Y_{ij}$ and $Y_{i\ell}$ (integrating over $a^*_i$) i.e.,

$$\rho = \text{Corr}(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}).$$

Further, the marginal mean of $Y_{ij}$ is $\pi_{ij}$,

$$E(Y_{ij}|x_{ij}) = E_{a^*_i|x_{ij}}[E_{Y_{ij}|x_{ij},a^*_i}(Y_{ij})]$$

$$= E_{a^*_i|x_{ij}}[p_{ij}]$$

$$= E_{a^*_i|x_{ij}}[\pi_{ij} + a^*_i \sqrt{\pi_{ij}(1 - \pi_{ij})}]$$

$$= \pi_{ij} + \sqrt{\pi_{ij}(1 - \pi_{ij})} E_{a^*_i|x_{ij}}[a^*_i]$$

$$= \pi_{ij}.\quad(4)$$

(Note, $a^*_i$ is assumed to be independent of $x_{ij}$ so that the conditional distribution of $a^*_i$ given $x_{ij}$ is the same as the marginal distribution of $a^*_i$). Also, as shown in Appendix 1, the marginal variance of $Y_{ij}$ is $\pi_{ij}(1 - \pi_{ij})$ (this is to be expected since $Y_{ij}$ is Bernoulli).

An element of the parameter vector $\beta^*$ represents the log-odds ratio for a one unit change in the corresponding covariate given the other covariates, for any subject in the population (integrating over $a^*_i$). This random effects model has a complicated conditional interpretation, but a simple marginal interpretation.
3 Constraints on $a_i^*$ and $\rho$

In order for $p_{ij}$ in (2) to be between 0 and 1, the maximum and minimum possible values of $a_i^*$ (and thus the maximum value of $\rho$) are constrained. In particular, we have the constraint

$$0 < \pi_{ij} + a_i^* \sqrt{\pi_{ij}(1 - \pi_{ij})} < 1$$

for all $j$ in cluster $i$, which, after some algebra, implies

$$-\sqrt{\frac{\pi_{ij}}{(1 - \pi_{ij})}} < a_i^* < \sqrt{\frac{1 - \pi_{ij}}{\pi_{ij}}}.$$ 

Thus, the sample space of $a_i^*$ is bounded. If we let $\pi_{i,min}$ be the minimum value of $\pi_{ij}$ in cluster $i$ and $\pi_{i,max}$ be the maximum value of $\pi_{ij}$ in cluster $i$, then, for cluster $i$, the constraint is

$$-\sqrt{\frac{\pi_{i,min}}{1 - \pi_{i,min}}} < a_i^* < \sqrt{\frac{1 - \pi_{i,max}}{\pi_{i,max}}}. \quad (5)$$

As the upper and lower bounds in (5) approach 0, the maximum possible value of $\rho = Var(a_i^*) = E(a_i^{*2})$ is also lowered. The bound on $\rho$ depends on the posed distribution of $a_i^*$. Here, we consider three possible distributions for $a_i^*$: dichotomous (discrete with two levels), uniform (continuous) or transformed beta (continuous). Table 2 gives values for $\pi_{i,min}$ and $\pi_{i,max}$ and the maximum possible value for $\rho$ for these three distributions. Next, we explain these distributions in a little more detail.

As shown in Appendix 2, the following dichotomous random variable has mean 0 and variance $\rho$,

$$a_i^* = \begin{cases} m_i & \text{with probability } \frac{\rho}{(\rho + m_i^2)} \\ -\rho/m_i & \text{with probability } \frac{m_i^2}{(\rho + m_i^2)} \end{cases}, \quad (6)$$

for any given $-1/m_i$ and $m_i$ in the interval

$$\left[ -\sqrt{\frac{\pi_{i,min}}{(1 - \pi_{i,min})}}, \sqrt{\frac{1 - \pi_{i,max}}{\pi_{i,max}}} \right].$$
In order for the bounds in (5) to hold, for this distribution, we must have (Appendix 2)

\[ 0 \leq \rho \leq \sqrt{\frac{\pi_{i,\min}(1 - \pi_{i,\max})}{\pi_{i,\max}(1 - \pi_{i,\min})}}. \] (7)

From Table 2, the maximum possible value of \( \rho \) depends on the absolute value of the difference between \( \pi_{i,\min} \) and \( \pi_{i,\max} \), but not the value of \( \pi_{i,\min} \) or \( \pi_{i,\max} \). As \(|\pi_{i,\max} - \pi_{i,\min}|\) gets larger, the upper bound on \( \rho \) gets smaller.

Another possible distribution for \( a_i^* \) is the following uniform distribution,

\[ f(a_i^*) = \frac{1}{2\sqrt{3}\rho} \text{ for } -\sqrt{3}\rho \leq a_i^* \leq \sqrt{3}\rho. \]

The mean of a uniform distribution is the average of the endpoints of the range of \( a_i^* \), which is easily shown to equal 0 here. Also, the variance is the square of the range of \( a_i^* \), divided by 12, which is easily shown to equal \( \rho \). For this uniform distribution, we must have

\[ 0 \leq \rho \leq \min \left\{ \frac{\pi_{i,\min}}{3(1 - \pi_{i,\min})}, \frac{1 - \pi_{i,\max}}{3\pi_{i,\max}} \right\}. \]

As we see from Table 2, in terms of the possible values of \( \rho \), this uniform distribution is much more restrictive than the previous distribution. Because the uniform is a symmetric distribution (and it has to be symmetric about 0 in this case for \( E(a_i^*) = 0 \)), \( a_i^* \) cannot take on the full range of values in (5). Further, in Table 2, we also see that the maximum value \( \rho \) can take on depends on both \( \pi_{i,\min} \) and \( \pi_{i,\max} \).

The last distribution we discuss is a transformed beta distribution. Suppose \( b_i \) is a beta random variable,

\[ f(b_i) = b_i^{\nu-1}(1 - b_i)^{\omega-1}/B(\nu, \omega) \text{ for } 0 \leq b_i \leq 1, \] (8)

where \( B(\nu, \omega) \) is the beta function. Then, we form the linear transformation

\[ a_i^* = \left[ \sqrt{\frac{\pi_{i,\min}}{(1 - \pi_{i,\min})}} + \sqrt{\frac{1 - \pi_{i,\max}}{\pi_{i,\max}}} \right] b_i - \sqrt{\frac{\pi_{i,\min}}{(1 - \pi_{i,\min})}}, \] (9)
which can have maximum value (when \( b_i = 1 \))

\[
U_i = \sqrt{\frac{1 - \pi_{i,\text{max}}}{\pi_{i,\text{max}}}}
\]

and minimum value (when \( b_i = 0 \))

\[
L_i = -\sqrt{\frac{\pi_{i,\text{min}}}{(1 - \pi_{i,\text{min}})}}.
\]

In terms of \( U_i \) and \( L_i \), we can rewrite (9) as

\[
a_i^* = (U_i - L_i)b_i + L_i.
\]

Under the restriction that \( E(a_i^*) = 0 \) and \( \text{Var}(a_i^*) = \rho \), it can be shown \( \nu \) and \( \omega \) from the beta distribution satisfy

\[
\nu = \frac{U_i(-U_iL_i - \rho)}{(U_i - L_i)\rho}
\]

and

\[
\omega = \frac{-L_i(-U_iL_i - \rho)}{(U_i - L_i)\rho}.
\]

Since \( \nu \) and \( \omega \) must both be positive in the beta distribution, and, by definition \( U_i \geq L_i \), both equations (11) and (12) imply that

\[-U_iL_i - \rho \geq 0.
\]

After a little algebra, this further implies that, for the beta distribution, \( \rho \) has the same constraints as in (7), which are also the constraints for the dichotomous distribution given in (6). Again, the constraints on \( \rho \) depend on the difference between \( \pi_{i,\text{min}} \) and \( \pi_{i,\text{max}} \), and not the value of \( \pi_{i,\text{min}} \) or \( \pi_{i,\text{max}} \). Thus, this transformed beta gives us much more flexibility than the uniform. Even though it has the same range as the dichotomous distribution given in (6), this transformed beta allows \( a_i^* \) to be continuous instead of taking on just two values.
Since the distribution of $a_i^*$ depends on $\pi_{i,min}$ and $\pi_{i,max}$ in both (6) and (9), the marginal likelihood,

$$f(y_{i1}, \ldots, y_{ni}|\beta, \rho) = \int \left[ \prod_{j=1}^{n_i} p_{ij}^{y_{ij}}(1 - p_{ij})^{(1-y_{ij})} \right] f(a_i^*) da_i^*, \quad (13)$$

is complicated. Further, because the distribution of $a_i^*$ depends on $\pi_{i,min}$ and $\pi_{i,max}$ the maximum likelihood estimate will not have the nice properties of the usual maximum likelihood estimate, i.e., the asymptotic distribution of the maximum likelihood estimate is very difficult to derive, so that the inverse of the second derivative of the negative log-likelihood will not be the asymptotic variance. The one exception to this occurs when $\pi_{ij} = \pi_i$ is the same for all subjects in a cluster, so that $\pi_i = \pi_{i,max} = \pi_{i,min}$. For example, when $a_i^*$ has the distribution in (9), then the marginal distribution of $\sum_{i=1}^{n_i} Y_{ij}$ (integrating over $a_i^*$) follows a beta-binomial distribution, which has the usual properties of a maximum likelihood estimate.

Instead of using maximum likelihood to estimate the parameters $(\beta, \rho)$, we suggest using generalized estimating equations (Liang and Zeger, 1986), which is the semi-parametric efficient estimate. Further, unlike maximum likelihood for this model, the asymptotic properties of generalized estimating equations are well understood (Liang and Zeger, 1986). Using generalized estimating equations, one only needs to specify the mean vector and the covariance matrix of the binary outcomes in a cluster. Although the joint distribution in (13) is complicated, the marginal mean, the marginal variance, and covariance are easy to calculate, i.e.,

$$E(Y_{ij}|x_{ij}) = \pi_{ij} = \frac{e^{x_{ij}'\beta}}{1 + e^{x_{ij}'\beta}}; \quad (14)$$

$$V ar(Y_{ij}|x_{ij}) = \pi_{ij}(1 - \pi_{ij}); \quad (15)$$

and

$$\rho = Corr(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}). \quad (16)$$
This is the usual marginal model for which the generalized estimating equations are used.

4 Simulation based on Liver Cancer Example

Our example deals with the clustered binary data given in Table 1, which is based on a liver cancer study comparing three doses (0=low, 1=medium, 2=high) of a chemotherapy with respect to survival. In table 1, a total of 21 patients from $N = 6$ hospitals are displayed, with an average of 3.5 patients ($n_i$) per hospital. The outcome for the $j^{th}$ patient in institution $i$ is the binary outcome $Y_{ij}$ (1 if survived at least one year from randomization, 0 if survived less than one year from randomization).

In determining dose differences ($DOSE_{ij}$ equals 0 if low, 1 if medium, 2 if high), another possibly important predictor is baseline cardiac status ($heart_{ij}$ equals 1 if normal, 0 if abnormal). For the $j^{th}$ patient in institution $i$, we let the marginal logistic model be

\[
\text{logit} [\pi_{ij}] = \text{logit} [\text{pr}\{Y_{ij} = 1 | x_{ij}\}] = -0.25 - 0.25DOSE_{ij} + 0.35heart_{ij},
\]

and the correlation be

\[
\rho = 0.4.
\]

We assume that the conditional (on $a_i^*$) model is

\[
p_{ij} = \text{pr}\{Y_{ij} = 1 | x_{ij}, a_i^*\} = \pi_{ij} + a_i^* \sqrt{\pi_{ij}(1-\pi_{ij})},
\]

where the random effect $a_i^*$ is given in (10), with $b_i$ following the beta distribution given in (8). To simulate $b_i$, we use the probability integral transform (Hoel, et. al., 1971). Applying the probability integral transform, $F(b_i)$ has a uniform (0,1) distribution, where $F(b_i)$ is the cumulative beta distribution function of $b_i$. Letting $d_i$ be a uniform (0,1) random variable, it then follows that $b_i = F^{-1}(d_i)$ has a beta distribution, where $F^{-1}(\cdot)$
is the inverse cumulative distribution function of a beta random variable. Table 3 gives the SAS commands and selected output for a simulation with the patients in Table 1. We see from Table 3 that the maximum value \((\max \rho)\) that \(\rho\) can take on is 0.42741. We note that the program can easily be written in other software such as S-Plus.

5 Extensions and Discussion

To extend the model proposed here to general correlation patterns among responses for subjects in the same cluster (such as autoregressive in a longitudinal study), we must extend the notation a little. Suppose we form the \((n_i \times 1)\) vectors

\[
Y_i = [Y_{i1}, ..., Y_{in_i}]', \\
\pi_i = [\pi_{i1}, ..., \pi_{in_i}]', \\
p_i = [p_{i1}, ..., p_{in_i}]',
\]

and form the \((n_i \times n_i)\) diagonal matrix

\[
W_i = \text{Diag}[\pi_{i1}(1 - \pi_{i1}), ..., \pi_{in_i}(1 - \pi_{in_i})],
\]

which contains the marginal variances \(\text{Var}(Y_{ij}|x_{ij}) = \pi_{ij}(1 - \pi_{ij})\) on the main diagonal. Then, we let

\[
p_i = \pi_i + W_i^{1/2}a_i,
\]

where the elements of \(W_i^{1/2}\) are the square roots of the elements of \(W_i\), and \(a_i\) is an \(n_i \times 1\) random vector with mean 0 and covariance matrix \(\Gamma\). In this random effects model, given \(a_i\) (and thus the \(p_{ij}\)'s), the \(Y_{ij}\)'s in the cluster are independent bernoulli random variables with probabilities \(p_{ij}\). Here, it is easily shown that \(\Gamma\) is the marginal correlation matrix of the elements of \(Y_i\) (integrated over \(a_i^*\)).
For the random effect model discussed here, the joint distribution of the $Y_{ij}$'s from a cluster (integrated over the $a_i^*$'s) is multinomial, but finding the probabilities of this joint distribution involve complicated integrations (or sums, depending on the distribution of the random effect). Thus, to estimate the parameters $\beta^*$ and $\rho$, we suggest the use of generalized estimating equations, which gives the semi-parametric efficient estimate. The most important use of the random effects models proposed here is for simulating correlated binary data, in which we specify a given correlation between pairs of observations in a cluster. A popular model for simulating correlated binary data in which (14), (15) and (16) hold, is the Bahadur model (1961). However, using the Bahadur model, besides the correlation coefficient, one also has to specify higher order ‘correlation coefficients’; with $n_i$ observations per cluster, one has to specify $[2^{n_i} - n_i(n_i + 1)/2]$ ‘higher order correlations’ per cluster. These higher order correlations are nuisance parameters which we would rather not have to specify. With the random effects models proposed here, we do not have to specify these higher order correlations.
References


Table 1. Data from the Liver Cancer study

<table>
<thead>
<tr>
<th>PATIENT</th>
<th>HOSPITAL</th>
<th>≥ 1 year</th>
<th>DOSE</th>
<th>ABNORMAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>YES</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>NO</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>YES</td>
<td>2</td>
<td>NO</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>NO</td>
<td>0</td>
<td>YES</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>NO</td>
<td>2</td>
<td>YES</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>NO</td>
<td>1</td>
<td>NO</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>NO</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>NO</td>
<td>2</td>
<td>YES</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>YES</td>
<td>1</td>
<td>NO</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>NO</td>
<td>1</td>
<td>NO</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>NO</td>
<td>0</td>
<td>YES</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>YES</td>
<td>1</td>
<td>NO</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>YES</td>
<td>2</td>
<td>YES</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>NO</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>YES</td>
<td>2</td>
<td>NO</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>YES</td>
<td>1</td>
<td>YES</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>YES</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>NO</td>
<td>2</td>
<td>NO</td>
</tr>
<tr>
<td>19</td>
<td>6</td>
<td>YES</td>
<td>0</td>
<td>NO</td>
</tr>
<tr>
<td>20</td>
<td>6</td>
<td>NO</td>
<td>2</td>
<td>YES</td>
</tr>
<tr>
<td>21</td>
<td>6</td>
<td>YES</td>
<td>1</td>
<td>YES</td>
</tr>
</tbody>
</table>
Table 2. Maximum value of $\rho$ for various values of $\pi_{i,min}$ and $\pi_{i,max}$

<table>
<thead>
<tr>
<th>$\pi_{i,min}$</th>
<th>$\pi_{i,max}$</th>
<th>Maximum value of $\rho$</th>
<th>transformed beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.9</td>
<td>0.11</td>
<td>0.04</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9</td>
<td>0.33</td>
<td>0.04</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5</td>
<td>0.33</td>
<td>0.04</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7</td>
<td>0.65</td>
<td>0.14</td>
</tr>
<tr>
<td>0.3</td>
<td>0.5</td>
<td>0.65</td>
<td>0.14</td>
</tr>
<tr>
<td>0.8</td>
<td>0.9</td>
<td>0.67</td>
<td>0.08</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.67</td>
<td>0.08</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>1.00</td>
<td>0.33</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>1.00</td>
<td>0.04</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
<td>1.00</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table 3. SAS commands to simulate data for Table 1.

data liver;
  input INST SURV DOSE heart;
  xb = -.25 - .25*DOSE + .35*heart;
  pi_ij = exp(xb)/(1+exp(xb));
datalines;
  1 1 1 1
  1 1 2 0
  1 0 0 1
  2 0 2 1
  2 0 1 0
  3 0 2 1
  3 1 1 0
  3 0 1 0
  3 0 0 1
  4 1 1 0
  4 1 2 1
  4 0 0 0
  4 1 2 0
  5 1 1 1
  5 1 0 0
  5 0 2 0
  6 1 0 0
  6 0 2 1
  6 1 1 1
;
proc means noprint data=liver;
  var pi_ij;
  by inst;
  output out=min_max(drop=_type_ _freq_) min=min_pi max=max_pi;
run;
data min_max;
  set min_max;
  max_rho = min_pi*( 1 - max_pi )/max_pi/( 1 - min_pi )
run;
proc print data=min_max;
run;
/* MIN_MAX DATASET */
Obs INST min_pi max_pi max_rho
  1 1 0.32082 0.52498 0.42741
  2 2 0.37754 0.43782 0.77880
  3 3 0.37754 0.52498 0.54881
  4 4 0.32082 0.43782 0.60653
  5 5 0.32082 0.46257 0.54881
  6 6 0.40131 0.46257 0.77880
*/
Table 3. SAS commands to simulate data for Table 1 (continued).

data min_max(keep = inst a_i) ;
set min_max;
rho=.4;
U_i = sqrt( (1 - max_pi)/max_pi );
L_i = -sqrt( min_pi/(1 - min_pi) );
nu = ( U_i*(-U_i*L_i - rho )/( (U_i - L_i)*rho ) ;
omega = ( -L_i*(-U_i*L_i - rho )/( (U_i - L_i)*rho ) ;
d_i = ranuni(0); /* uniform (0,1) random variate */
b_i = BETAINV(d_i,nu,omega); /* probability integral transform */
a_i = (U_i - L_i)*b_i + L_i ;
run;

data liver;
merge liver min_max;
by inst;
p_ij = pi_ij + a_i* sqrt ( pi_ij*(1-pi_ij) );
y_ij = ranbin(0,1,p_ij);
run;

proc print;
var inst DOSE heart pi_ij a_i p_ij y_ij;
run;

/* Random Sample

Obs INST DOSE heart pi_ij a_i p_ij y_ij
1 1 1 1 0.46257 0.94400 0.93325 1
2 1 0 0 0.43782 0.94400 0.90616 1
3 1 2 0 0.32082 0.94400 0.76147 1
4 1 0 1 0.52498 0.94400 0.76147 1
5 2 2 1 0.40131 -0.66070 0.07746 0
6 2 1 0 0.37754 -0.66070 0.05725 0
7 2 0 0 0.43782 -0.66070 0.11004 1
8 3 2 1 0.40131 0.55450 0.67311 0
9 3 1 0 0.37754 0.55450 0.64635 1
10 3 1 0 0.37754 0.55450 0.64635 1
11 3 0 1 0.52498 0.55450 0.80188 1
12 4 1 0 0.37754 -0.16177 0.29912 1
13 4 2 1 0.40131 -0.16177 0.32202 1
14 4 0 0 0.43782 -0.16177 0.35757 0
15 4 2 0 0.32082 -0.16177 0.24531 1
16 5 1 1 0.46257 0.88113 0.90190 1
17 5 0 0 0.43782 0.88113 0.87497 1
18 5 2 0 0.32082 0.88113 0.73212 1
19 6 0 0 0.43782 0.78458 0.82707 1
20 6 2 1 0.40131 0.78458 0.78588 1
21 6 1 1 0.46257 0.78458 0.85376 1

*/
Appendix 1

First, we show that \( \text{Var}(Y_{ij}|x_{ij}) = \pi_{ij}(1-\pi_{ij}) \). We showed in (4) that \( E(Y_{ij}|x_{ij}) = \pi_{ij} \). Since \( Y_{ij} \) is binary, \( Y^2_{ij} = Y_{ij} \), and

\[
E(Y^2_{ij}|x_{ij}) = E(Y_{ij}|x_{ij})
\]

Then,

\[
\text{Var}(Y_{ij}|x_{ij}) = E(Y^2_{ij}|x_{ij}) - [E(Y_{ij}|x_{ij})]^2 = E(Y_{ij}|x_{ij}) - [E(Y_{ij}|x_{ij})]^2 = \pi_{ij}(1-\pi_{ij}).
\]

Next, we show that

\[
\rho = \text{Corr}(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}).
\]

Now,

\[
\text{Corr}(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}) = \frac{\text{Cov}(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell})}{\sqrt{\text{Var}(Y_{ij}|x_{ij})\text{Var}(Y_{i\ell}|x_{i\ell})}} = \frac{\text{Cov}(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell})}{\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{i\ell}(1-\pi_{i\ell})}}
\]

Thus, we need to calculate the covariance,

\[
\text{Cov}(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}) = E(Y_{ij}Y_{i\ell}|x_{ij}, x_{i\ell}) - \pi_{ij}\pi_{i\ell}.
\]

Next, we apply the double expectation formula,

\[
E(Y_{ij}Y_{i\ell}|x_{ij}, x_{i\ell}) = Ea^*_i|x_{ij},x_{i\ell}[EY_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}, a^*_i(Y_{ij}Y_{i\ell})].
\]

Note, \( a^*_i \) is independent of \( x_{ij} \), so that (20) can be rewritten as

\[
Ea^*_i[EY_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}, a^*_i(Y_{ij}Y_{i\ell})].
\]

Since, conditional on \( a^*_i, Y_{ij} \) and \( Y_{i\ell} \) are independent,

\[
EY_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}, a^*_i(Y_{ij}Y_{i\ell}) = [EY_{ij}|x_{ij}, a^*_i(Y_{ij})][EY_{i\ell}|x_{i\ell}, a^*_i(Y_{i\ell})]
\]

\[
= [\pi_{ij} + a^*_i\sqrt{\pi_{ij}(1-\pi_{ij})}][\pi_{i\ell} + a^*_i\sqrt{\pi_{i\ell}(1-\pi_{i\ell})}]
\]

\[
= \pi_{ij}\pi_{i\ell} + a^*_i[\pi_{i\ell}\sqrt{\pi_{ij}(1-\pi_{ij})} + \pi_{ij}\sqrt{\pi_{i\ell}(1-\pi_{i\ell})}]
\]

\[
+ a^*_i^2 \sqrt{\pi_{ij}(1-\pi_{ij})\pi_{i\ell}(1-\pi_{i\ell})}.
\]

Then, taking expectations of (21) with respect to \( a^*_i \) gives

\[
E(Y_{ij}Y_{i\ell}|x_{ij}, x_{i\ell}) = \pi_{ij}\pi_{i\ell} + \rho\sqrt{\pi_{ij}(1-\pi_{ij})\pi_{i\ell}(1-\pi_{i\ell})}
\]

18
since $E_{a_i^*}(a_i^*) = 0$ and $E_{a_i^*}(a_i^{*2}) = Var_{a_i^*}(a_i^{*2}) = \rho$.

Substituting (22) in (19), we obtain

\[
Cov(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}) = \rho \sqrt{\pi_{ij}(1 - \pi_{ij})\pi_{i\ell}(1 - \pi_{i\ell})},
\]

and thus

\[
Corr(Y_{ij}, Y_{i\ell}|x_{ij}, x_{i\ell}) = \rho.
\]
Appendix 2

First, let $a_i^*$ be a dichotomous random variable. Suppose we let

$$m_i \in \left[ 0, \sqrt{\frac{1 - \pi_{i,\text{max}}}{\pi_{i,\text{max}}}} \right].$$

be one of two possible values that $a_i$ can take on, with $\alpha_i = \text{pr}(a_i^* = m_i)$. We denote the other possible value that $a_i^*$ can take on by $M_i$. Since $m_i$ is specified, to fully specify the distribution, we need to find a value for $M_i$ and $\alpha_i$. Note, the mean and variance of $a_i^*$ are

$$\mu_i = E(a_i^*) = \alpha_i m_i + (1 - \alpha_i) M_i$$  \hspace{1cm} (23)

and

$$\text{Var}(a_i^*) = \alpha_i (m_i - \mu_i)^2 + (1 - \alpha_i) (M_i - \mu_i)^2.$$  \hspace{1cm} (24)

Under the assumptions of the model, we have that $\mu_i = 0$ and $\text{Var}(a_i^*) = \rho$. After substituting $\mu = 0$ and $\text{Var}(a_i^*) = \rho$ in equations (23) and (24) and working through some algebra, we obtain

$$\alpha_i = \frac{\rho}{\rho + m_i^2}$$

and

$$M_i = -\frac{\rho}{m_i}.$$  \hspace{1cm} (25)

This gives the following discrete distribution for $a_i^*$,

$$a_i^* = \begin{cases} m_i & \text{with probability } \rho/(\rho + m_i^2) \\ -\rho/m_i & \text{with probability } m_i^2/(\rho + m_i^2) \end{cases}.$$  \hspace{1cm} (25)

Recall, the bounds on $a_i^*$ in (5) are such that $M_i$ must be in the interval

$$\left[ -\sqrt{\frac{\pi_{i,\text{min}}}{(1 - \pi_{i,\text{min})}}}, \sqrt{\frac{1 - \pi_{i,\text{max}}}{\pi_{i,\text{max}}}} \right].$$

In particular,

$$-\sqrt{\frac{\pi_{i,\text{min}}}{(1 - \pi_{i,\text{min})}}} \leq M_i$$

or, substituting $M_i = -\rho/m_i$, we have

$$-\sqrt{\frac{\pi_{i,\text{min}}}{(1 - \pi_{i,\text{min})}}} \leq -\frac{\rho}{m_i},$$

or, equivalently,

$$\rho \leq m_i \sqrt{\frac{\pi_{i,\text{min}}}{(1 - \pi_{i,\text{min})}}}.$$
However, since, using the bounds in (5),

\[ m_i \leq \sqrt{\frac{1 - \pi_{i,\text{max}}}{\pi_{i,\text{max}}}}; \]

which implies that

\[ \rho \leq \sqrt{\frac{\pi_{i,\text{min}}(1 - \pi_{i,\text{max}})}{\pi_{i,\text{max}}(1 - \pi_{i,\text{min}})}}. \]