University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2008 *Paper* 26

On the designation of the patterned associations for longitudinal Bernoulli data: weight matrix versus true correlation structure?

Hanjoo Kim* Joseph M. Hilbe†
Justine Shults‡

*University of Pennsylvania School of Medicine, hanjoo@mail.med.upenn.edu
†Arizona State University, j.m.hilbe@gmail.com
‡University of Pennsylvania School of Medicine, jshults@cceb.med.upenn.edu

On the designation of the patterned associations for longitudinal Bernoulli data: weight matrix versus true correlation structure?

Hanjoo Kim, Joseph M. Hilbe, and Justine Shults

Abstract

Due to potential violation of standard constraints for the correlation for binary data, it has been argued recently that the working correlation matrix should be viewed as a weight matrix that should not be confused with the true correlation structure. We propose two arguments to support our view to the contrary for the first-order autoregressive AR(1) correlation matrix. First, we prove that the standard constraints are not unduly restrictive for the AR(1) structure that is plausible for longitudinal data; furthermore, for the logit link function the upper boundary value only depends on the regression parameter and the change in covariate values between successive measurements. In addition, for given marginal means and parameter α, we provide a general proof that satisfaction of the standard constraints for consecutive marginal means will guarantee the existence of a compatible multivariate distribution with an AR(1) structure. The relative laxity of the standard constraints for the AR(1) structure coupled with the existence of a simple model that yields data with an AR(1) structure bolsters our view that for the AR(1) structure at least, it is appropriate to view this model as a correlation structure versus a weight matrix.

, *pp*. 1–12

©

On the designation of the patterned associations for longitudinal Bernoulli data: weight matrix versus true correlation structure?

BY HANJOO KIM

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine,
Philadelphia, Pennsylvania 19104, U.S.A.

hanjoo@mail.med.upenn.edu

JOSEPH M. HILBE

School of Social and Family Dynamics, Arizona State University, Tempe, AZ 85287 and
Department of Nutrition, Arizona State University Polytechnic, Mesa, Arizona 85212, U.S.A.

Hilbe@asu.edu

AND JUSTINE SHULTS

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine,
Philadelphia, Pennsylvania 19104, U.S.A.

jshults@mail.med.upenn.edu

SUMMARY

Due to potential violation of standard constraints for the correlation for binary data, it has been argued recently that the working correlation matrix should be viewed as a weight matrix that should not be confused with the true correlation structure. We propose two arguments to support our view to the contrary for the first-order autoregressive AR(1) correlation matrix. First, we prove that the standard constraints are not unduly restrictive for the AR(1) structure that is plausible for longitudinal data; furthermore, for the logit link function the upper boundary value only depends on the regression parameter and the change in covariate values between successive measurements. In addition, for given marginal means and parameter α, we provide a general proof that satisfaction of the standard constraints for consecutive marginal means will guarantee the existence of a compatible multivariate distribution with an AR(1) structure. The relative laxity of the standard constraints for the AR(1) structure coupled with the existence of a simple model that yields data with an AR(1) structure bolsters our view that for the AR(1) structure at least, it is appropriate to view this model as a correlation structure versus a weight matrix.

Some key words: Bernoulli data; correlated binary data; first-order autoregressive AR(1) structure.

1. INTRODUCTION

Correlated binary data that occur in many settings. For example, in a study in which repeated blood pressure measurements are collected on subjects, the binary variables Y_{ij} that take value 1 if subject i has high blood pressure and 0 otherwise, will be expected to be correlated within subjects.

We consider longitudinal binary measurements $Y_{i1}, \ldots, Y_{in_i}$ with expected values $E(Y_{ij}) = \text{pr}(Y_{ij} = 1) = P_{ij}$, where $Q_{ij} = \text{pr}(Y_{ij} = 0) = 1 - P_{ij}$, and the correlation between measurements Y_{ij} and Y_{ik} is given by $\text{corr}(Y_{ij}, Y_{ik}) = C_{ijk}$. An important feature of correlated binary data is that the P_{ij}, Q_{ij} and C_{ijk} completely determine the bivariate distribution of Y_{ij} and Y_{ik}

because the pair-wise probabilities pr$(Y_{ij} = y_{ij}, Y_{ik} = y_{ik}) =$ pr(y_{ij}, y_{ik}) can be expressed as

$$\text{pr}(y_{ij}, y_{ik}) = P_{ij}^{y_{ij}} Q_{ij}^{1-y_{ij}} P_{ik}^{y_{ik}} Q_{ik}^{1-y_{ik}} \left\{ 1 + C_{ijk} \frac{(y_{ij} - P_{ij})(y_{ik} - P_{ik})}{(P_{ij}\, P_{ik}\, Q_{ij}\, Q_{ik})^{1/2}} \right\} \quad (1)$$

as noted in Prentice (1988).

Prentice (1988) pointed out that the probabilities in (1) will be non-negative, i.e. pr$(y_{ij}, y_{ik}) \geq 0$, only if the correlations satisfy the following constraints that depend on the marginal means:

$$L_i(j,k) \leq \text{corr}(Y_{ij}, Y_{ik}) \leq U_i(j,k) \quad (2)$$

where

$$L_i(j,k) = \max\left\{ -(w_{ij}w_{ik})^{1/2}, -(w_{ij}w_{ik})^{-1/2} \right\},$$
$$U_i(j,k) = \min\left\{ (w_{ij}/w_{ik})^{1/2}, (w_{ij}/w_{ik})^{-1/2} \right\},$$

and $w_{ij} = P_{ij}Q_{ij}^{-1}$ for $i = 1, \dots, m$, $j = 1, \dots, n_i$, and $k = 1, \dots, n_i$.

Chaganty & Joe (2004) noted that the standard bounds (2) can be extremely tight for modeling correlated binary data with a vector of covariates x_{ij}, under the constant correlation assumption between all measurement pairs, i.e. corr$(Y_{ij}, Y_{ik}) = \alpha$ for all i, and $j \neq k$. As a result, they suggest the working correlation matrix should be viewed as a *weight matrix* that is not to be confused with the true correlation matrix of the binary measurements. They also propose simple rules for analysis with either an exchangeable weight matrix, or a first-order autoregressive AR(1) weight matrix, for which corr$(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$. The AR(1) structure is often applied in longitudinal studies because it forces the correlation to decrease with increasing separation in measurement occasion; this is plausible for many biological outcomes.

In this note we argue that it is appropriate and beneficial to view the AR(1) structure as the true correlation structure versus a weight matrix. First, in §2 we prove that the constraints (2) are not necessarily severe for the AR(1) structure. Our proof consists of showing that satisfaction of (2) for consecutive P_{ij} and P_{ij+1} implies the satisfaction of the constraints for all other P_{ij} and P_{ik}. Further, we simplify (2) for the logit link function and prove that for this link function, the upper boundary value for α only depends on the regression parameter β and the change in *consecutive* covariate values on each subject. In many situations the consecutive changes will not be large. For example, in many clinical studies, the temporal spacing of measurements is relatively small, so that time varying covariates should not change much from one measurement to the next. In addition, for studies that only contain cluster level covariates, e.g. sex and treatment group indicators, the consecutive changes will be zero, in which case the upper value of the constraints for α will be 1. In all situations, the lower boundary value for α is negative.

In general, as discussed in Chaganty & Joe (2006), satisfaction of (2) for given marginal means P_{ij} and parameter α is necessary but not sufficient to guarantee so-called *compatibility* which refers to the existence of a multivariate binary distribution with the given marginal means P_{ij} and patterned correlation structure with parameter α. We therefore next prove in §3 that for the AR(1) structure, satisfaction of the constraints (2) does guarantee the existence of a compatible multivariate distribution; this distribution is based on a Markovian model that was discussed by Liu & Liang (1997) and Jung & Ahn (2005).

Considered jointly, our proofs show (i) the standard bounds (2) are not necessarily unduly restrictive and (ii) that there exists a relatively simple compatible distribution with an AR(1) structure for a given set of marginal means and parameter α; these results suggest that is appropriate to view the AR(1) structure as a correlation structure versus a weight matrix. In §4 we argue that in addition to being appropriate, there are important benefits to be gained by viewing the AR(1) structure as the true correlation structure, versus a weight matrix.

2. STANDARD CONSTRAINTS FOR THE AR(1) STRUCTURE

For the AR(1) structure the correlation between consecutive measurements Y_{ij} and Y_{ij+1} is α. Theorem 1 establishes that if the standard constraints (2) are satisfied for α, and all consecutive marginal means P_{ij} and P_{ij+1}, i.e.

$$L_i(j, j+1) \leq \alpha \leq U_i(j, j+1) \tag{3}$$

for all $i = 1, \ldots, m$, and $j = 1, \ldots, n_i - 1$, then (2) will be satisfied for the correlation $\alpha^{|j-k|}$ between any Y_{ij} and Y_{ik}.

THEOREM 1 (CONSECUTIVE BOUNDS FOR AR(1) STRUCTURE). *Suppose* $L_i(j, j+1) \leq \alpha \leq U_i(j, j+1)$ *for all* $i = 1, \ldots, m$, *and* $j = 1, \ldots, n_i - 1$. *Then for* $j, k = 1, \ldots, n_i$ *such that* $|j - k| \geq 2$,

$$L_i(j, k) \leq \alpha^{|j-k|} \leq U_i(j, k)$$

for all $i = 1, \ldots, m$.

Proof. Without loss of generality assume that $j < k$. We first consider the upper bounds. Let

$$\begin{aligned} m &= \min\{U_i(j, j+1), U_i(j+1, j+2), \ldots, U_i(k-1, k)\} \\ &= \min\left\{v_{ij}, (v_{ij})^{-1}, v_{ij+1}, (v_{ij+1})^{-1}, \ldots, v_{ik-1}, (v_{ik-1})^{-1}\right\}, \end{aligned}$$

where $v_{ij} = (w_{ij}/w_{ij+1})^{1/2}$. Then (3) implies that $\alpha^{k-j} \leq m^{k-j}$ where

$$\begin{aligned} m^{k-j} &\leq v_{ij}v_{ij+1}\ldots v_{ik-1} = (w_{ij}/w_{ik})^{1/2}, \text{ and} \\ m^{k-j} &\leq (v_{ij})^{-1}(v_{ij+1})^{-1}\ldots(v_{ik-1})^{-1} = (w_{ij}/w_{ik})^{-1/2}. \end{aligned}$$

Therefore $\alpha^{k-j} \leq \min\{(w_{ij}/w_{ik})^{1/2}, (w_{ij}/w_{ik})^{-1/2}\} = U_i(j, k)$. Next, we consider the lower bounds. The lower bounds will be satisfied when $k - j$ is even because in this case $\alpha^{k-j} > 0$

and the lower bound is always negative. We therefore only need to consider the case that $k-j$ is odd. Let

$$\begin{aligned} s &= \max\{L_i(j,j+1), L_i(j+1,j+2), \ldots, L_i(k-1,k)\} \\ &= = \max\left\{-z_{ij}, -(z_{ij})^{-1}, -z_{ij+1}, -(z_{ij+1})^{-1}, \ldots, -z_{ik-1}, -(z_{ik-1})^{-1}\right\}, \end{aligned}$$

where $z_{ij} = (w_{ij}w_{ij+1})^{1/2}$. Then (3) and the fact that $k-j$ is odd implies that $\alpha^{k-j} \geq s^{k-j}$ where

$$\begin{aligned} s^{k-j} &\geq -z_{ij}(z_{ij+1})^{-1}z_{ij+2}\ldots(z_{ik-2})^{-1}z_{ik-1} = -(w_{ij}w_{ik})^{1/2}, \text{ and} \\ s^{k-j} &\geq -(z_{ij})^{-1}z_{ij+1}(z_{ij+2})^{-1}\ldots z_{ik-2}(z_{ik-1})^{-1} = -(w_{ij}w_{ik})^{-1/2}. \end{aligned}$$

Therefore $\alpha^{k-j} \geq \max\{-(w_{ij}w_{ik})^{1/2}, -(w_{ij}w_{ik})^{-1/2}\} = L_i(j,k)$. Since our results do not depend on the particular choice of i, j, and k, the result of the theorem follows. □

Theorem 1 is useful in establishing the boundary constraints for AR(1) structure for multivariate binary data with the logit link function that is widely used in practice. Specifically let $\log(w_{ij}) = x'_{ij}\beta$ where x_{ij} is $1 \times n_i$ vector of covariates and β is $1 \times p$ vector of regression coefficients. Then it follows that $w_{ij}w_{ij+1} = \exp\{(x_{ij} + x_{ij+1})'\beta\}$ and $w_{ij}/w_{ij+1} = \exp\{(x_{ij} - x_{ij+1})'\beta\}$. Substituting theses two expressions into (3) yields the following constraints for α:

$$\max_{i,j}\left[-\exp\left\{-\frac{|(x_{ij}+x_{ij+1})'\beta|}{2}\right\}\right] \leq \alpha \leq \min_{i,j}\left[\exp\left\{-\frac{|(x_{ij}-x_{ij+1})'\beta|}{2}\right\}\right]. \qquad (4)$$

Therefore, any α which satisfy the constraints (4) must also satisfy the standard bound (2) for correlated binary data with an AR(1) structure and logit link function.

In all practical applications for which α is thought to be positive and only positive estimates of α will be considered, the restriction on the lower bound no longer applies since it is always negative. Moreover, the upper bound in (4) depends only on the changes in consecutive time-varying covariates values and their corresponding regression parameters. Consequently, if the multivariate binary data do not contain any time-varying covariates, the upper bound becomes 1. Furthermore, even for data with time-varying covariates, the standard bounds need not necessarily be too tight because the upper bounds depend only on the change in consecutive values.

3. Compatible multivariate binary distribution with AR(1) structure

Here we prove that assumption of constant correlation α between consecutive Y_{ij}, Y_{ij+1} and satisfaction of the constraints (2) for consecutive marginal means P_{ij}, P_{ij+1} will guarantee the existence of a compatible multivariate binary distribution with AR(1) structure parameterized by α; this result was shown for $n \leq 14$ in §7 of Chaganty & Joe (2006) through a numerical approach.

As a first step, we consider the following Markovian model described by Liu & Liang (1997) and Jung & Ahn (2005) for the probability of a particular realization of the random variables Y_{ij} on subject i:

$$\text{pr}(y_{i1}, \ldots, y_{in_i}) = \text{pr}(Y_{i1} = y_{i1}) \prod_{j=2}^{n_i} \text{pr}(Y_{ij} = y_{ij} | Y_{ij-1} = y_{ij-1}) \tag{5}$$

where $\text{pr}(y_{i1}, \ldots, y_{in_i}) = \text{pr}(Y_{i1} = y_{i1}, \ldots, Y_{in} = y_{in_i})$.

Next, we prove the following theorem. A shorter proof for the following result is available in the technical report by Shults et al. (Shults, J., Sun, W. & Tu, X. 2006. On the violation of bounds for the correlation in generalized estimating equation analyses of binary data from longitudinal trials. *UPenn Biostatistics Working Papers. Working Paper 8*, http://biostats.bepress.com/upennbiostat/papers/art8.)

THEOREM 2 (CONSTANT α AND MARKOVIAN MODEL YIELDS AR(1) STRUCTURE). *Consider the Markovian model in (5) with the constant correlation between any two consecutive* *observation on a subject, i.e. corr*$(Y_{ij}, Y_{ij+1}) = \alpha$ *for all* $i = 1, \ldots, m$ *and* $j = 1, \ldots, n_i - 1$. *Then the correlation structure of the measurements for this model is given by AR(1), i.e. corr*$(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$ *for all* $i = 1, \ldots, m$ *and* $j, k = 1, \ldots, n_i$.

A proof is given in the Appendix. Next, for a given set of marginal means P_{ij} and parameter values α, satisfaction of (3) for all consecutive P_{ij-1} and P_{ij} guarantees that all bivariate distributions for Y_{ij-1} and Y_{ij} will be valid. Consequently, a compatible joint distribution with the given marginal means and corr$(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$ does exist by taking products of the Markovian model in (5) over all $i = 1, \ldots, m$, i.e. $\prod_{i=1}^{m} \text{pr}(y_{i1}, \ldots, y_{in_i}) = \prod_{i=1}^{m} \text{pr}(Y_{i1} = y_{i1}) \prod_{j=2}^{n_i} \text{pr}(Y_{ij} = y_{ij} | Y_{ij-1} = y_{ij-1})$. Therefore, we have

$$\text{pr}(y_{11}, \ldots, y_{mn_i}) = \prod_{i=1}^{m} \text{pr}(Y_{i1} = y_{i1}) \prod_{j=2}^{n_i} \frac{\text{pr}(y_{ij-1}, y_{ij})}{\text{pr}(Y_{ij-1} = y_{ij-1})} \tag{6}$$

where pr(y_{ij-1}, y_{ij}) is defined using(1) in §1 evaluated at $C_{ij-1j} = \alpha$.

Lastly, note that if a compatible distribution exists with given means P_{ij} and correlations corr$(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$, then the constraints in (2) will be satisfied because the bivariate probabilities pr(y_{ij}, y_{ik}) will clearly be non-negative, since they can be obtained by summing the appropriate probabilities in the multivariate distribution. However, Theorem 1 was helpful to establish that even if data are not distributed according to the multivariate Markovian model (6), then satisfaction of the constraints (2) for consecutive marginal means will guarantee satisfaction of (2) for all marginal means.

4. DISCUSSION

In this note we have proven (Theorem 1) that the standard constraints on α are not necessarily severe for the AR(1) structure. In addition, we have proven that for given marginal means and parameter α, there will exist a relatively simple compatible multivariate distribution with an

AR(1) structure. Our goal was to argue in favor of designating the AR(1) structure as a true correlation structure versus a weight matrix, for longitudinal binary data. We suggest that this designation may be appropriate for other structures as well, although more work may be needed to establish this conclusion.

Admitting that our working structure plays the role of a correlation structure could encourage us to think more carefully about our choice of structure and the implications of improper selection. For example, this note was motivated by the authors recent experience with submission to an applied statistics journal of a paper that discussed methods for choosing the correct correlation structure for binary data; one reviewer mentioned in three places, the recent publication that promotes the view that the AR(1) structure should be viewed as a weight matrix versus a correlation structure. These multiple citations implied the question "Why work on choosing a correlation structure for binary data when the working structure should be viewed as a weight matrix that is not to be confused with the true correlation structure of the data?"

More generally, we note that wording can have an important and potentially detrimental impact on research. For example, we suspect that the early designation of the correlation parameters as nuisance parameters for GEE, discouraged efforts to implement the same wide variety of patterned correlation matrices that have been applied for maximum likelihood analysis of normally distributed data, due to the fact that this designation encouraged researchers to dismiss the correlation parameters as unimportant.

There are also potential practical drawbacks to viewing working structures as weight matrices, because lack of a defined role for the structure could result in ambiguity with respect to selection of the estimated value of the weight parameter α. For example, the proponents of designating α as a weight parameter (Chaganty & Joe, 2004) suggest choosing the value of α for the AR(1) weight matrix from within an interval of potential values. For data that suggest strong dependence they suggest choosing a value for α that is the midpoint of the estimated bounds, or that is in the interval $(0.70, 0.90)$. That their approach requires the analyst to choose from an infinite number of possible values for α suggests that their weight-based approach could be difficult to implement in practice.

As Hardin & Hilbe (2002) pointed out, priority factors for choosing an appropriate generalized estimating equations (GEE) (Liang & Zeger, 1986) model are the scientific questions of interest, the size and nature of the clusters, and the nature of the covariates. Therefore, viewing a patterned matrix as a correlation structure as opposed to merely treating it as a weight matrix should lead to searching for a limited number of candidate structures that are plausible in the context of the investigation. Careful modeling of the correlations is admittedly difficult due to complicated boundary conditions for α; however, promising research is being done in this area. For example, Chaganty & Deng (2007) have derived constraints for α for familial correlation structures that are useful in analysis of genetic data. Furthermore, Qaqish (2003) has described a wide class of multivariate distributions that can be used to yield Bernoulli data with particular patterned correlation structures.

The theory of optimal estimating functions of Godambe (1960, 1991) can be used to show that we will suffer a loss in efficiency in estimation of the regression parameter if the correlation structure is misspecified. Further study of the impact of incorrect specification for new correlation structures, in addition to further development of methods for choosing between and implementing those structures in GEE analysis of binary data should therefore be beneficial.

APPENDIX

Proof

Proof of Theorem 2. The proof is by induction. First note that for binary data, $E(Y_{ij}Y_{ij+1}) = \text{pr}(Y_{ij} = 1, Y_{ij+1} = 1)$ so that

$$\text{corr}(Y_{ij}, Y_{ij+1}) = \{\text{pr}(Y_{ij} = 1, Y_{ij+1} = 1) - P_{ij}P_{ij+1}\} / (P_{ij}P_{ij+1}Q_{ij}Q_{ij+1})^{1/2}. \tag{A1}$$

Since $\text{corr}(Y_{ij}, Y_{ij+1}) = \alpha$ by the assumption in Theorem 2, rearranging (A1) gives

$$\text{pr}(Y_{ij} = 1, Y_{ij+1} = 1) = P_{ij}P_{ij+1} + \alpha(P_{ij}P_{ij+1}Q_{ij}Q_{ij+1})^{1/2}. \tag{A2}$$

Without loss of generality assume that $k > j$. For the first step in the induction argument, let $k = j + 1$. We need to show that the Markovian model (5) coupled with $\text{corr}(Y_{ij}, Y_{ij+1}) = \alpha$ implies that $\text{corr}(Y_{ij}, Y_{ij+2}) = \alpha^2$, which is equivalent to showing that

$$\text{pr}(Y_{ij} = 1, Y_{ij+2} = 1) = P_{ij}P_{ij+2} + \alpha^2(P_{ij}P_{ij+2}Q_{ij}Q_{ij+2})^{1/2}.$$

Thus,

$$\begin{aligned}
\text{pr}(Y_{ij} = 1, Y_{ij+2} = 1) &= \sum_{y\in\{0,1\}} \text{pr}(Y_{ij} = 1, Y_{ij+1} = y, Y_{ij+2} = 1) \\
&= \sum_{y\in\{0,1\}} \text{pr}(Y_{ij+2} = 1|Y_{ij+1} = y)\text{pr}(Y_{ij+1} = y|Y_{ij} = 1)P_{ij} \\
&= \sum_{y\in\{0,1\}} \frac{\text{pr}(Y_{ij+2} = 1, Y_{ij+1} = y)}{\text{pr}(Y_{ij+1} = y)}\text{pr}(Y_{ij} = 1, Y_{ij+1} = y) \\
&= \frac{\text{pr}(Y_{ij+2} = 1, Y_{ij+1} = 1)}{P_{ij+1}}\text{pr}(Y_{ij} = 1, Y_{ij+1} = 1) \\
&\quad + \frac{P_{ij+2} - \text{pr}(Y_{ij+1} = 1, Y_{ij+2} = 1)}{Q_{ij+1}}\left\{P_{ij} - \text{pr}(Y_{ij} = 1, Y_{ij+1} = 1)\right\} \\
&= P_{ij}P_{ij+2} + \alpha^2(P_{ij}P_{ij+2}Q_{ij}Q_{ij+2})^{1/2}
\end{aligned}$$

where the second equality follows from the Markovian model (5) and the last expression is obtained from (A2), and some algebra. Next, assume that $\text{corr}(Y_{ij}, Y_{ij+k}) = \alpha^k$ is true for some $k > j + 1$. Then we can use almost identical calculations as for $k = j + 1$ to show that

$$\text{pr}(Y_{ij} = 1, Y_{ij+k+1} = 1) = \sum_{y\in\{0,1\}} \text{pr}(Y_{ij} = 1, Y_{ij+k} = y, Y_{ij+k+1} = 1)$$

$$= P_{ij}P_{ij+k+1} + \alpha^{k+1}(P_{ij}P_{ij+k+1}Q_{ij}Q_{ij+k+1}).$$

Therefore, $\text{corr}(Y_{ij}, Y_{ij+k+1}) = \alpha^{k+1}$. □

Acknowledgments

Work on this report was supported by the NIH funded Longitudinal Analysis for Diverse Populations project (R01CA096885).

References

Chaganty N.R. & Deng Y. (2007). Ranges of measures of association for familial binary variables. *Comm. statist. theo. meth.* **36**, 587-598.

Chaganty, N.R. & Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *J. R. Statist. Soc. B* **66**, 851-860.

Chaganty, N.R. & Joe, H. (2006). Range of correlation matrices for dependent bernoulli random variables. *Biometrika* **93**, 197-206.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. math. statist.* **31**, 1208-1211.

Godambe, V.P. (1991). *Estimating functions*. Oxford: University Press.

Hardin, J.W. & Hilbe, J.M. (2002). *Generalized estimating equations*. Florida: Chapman & Hall/CRC Press.

Jung S.H. & Ahn W.W. (2005). Sample size for a two-group comparison of repeated binary measurements using GEE. *Statist. med.* **24**, 2583-2596.

Liang K.Y. & Zeger, S.L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13-22.

Liu G. & Liang K.Y. (1997). Sample size calculation for studies with correlated observations. *Biometrics* **53**, 937-947.

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**, 455-463.

[*Received*]