Journal of Data Science **2**(2004), 245-257

A GEE Approach for Estimating Correlation Coefficients Involving Left-censored Variables

Jingli Song, Huiman X. Barnhart and Robert H. Lyles
Emory University

Abstract: HIV (Human Immunodeficiency Virus) researchers are often concerned with the correlation between HIV viral load measurements and CD4+ lymphocyte counts. Due to the lower limits of detection (LOD) of the available assays, HIV viral load measurements are subject to left-censoring. Motivated by these considerations, the maximum likelihood (ML) method under normality assumptions was recently proposed for estimating the correlation between two continuous variables that are subject to left-censoring. In this paper, we propose a generalized estimating equations (GEE) approach as an alternative to estimate such a correlation coefficient. We investigate the robustness to the normality assumption of the ML and the GEE approaches via simulations. An actual HIV data example is used for illustration.

Key words: Correlation, generalized estimating equations, left-censored variables, missing data.

1. Introduction

Viral load assessment via quantification of plasma viral RNA (Ribonucleic Acid) plays an important role in current HIV (Human Immunodeficiency Virus) research. It has provided valuable insights into the pathogenesis of HIV disease and the activity of anti-viral drugs. However, inherent limits of detection (LOD) in existing HIV RNA assays lead to the possibility of left-censored (also termed missing) data. Such left-censored data is also characteristic of many other types of bioassay studies (Lynn, 2001). As both CD4+ cell counts and plasma HIV RNA are bio-markers of the progression of HIV disease, the study of the correlation between the CD4 + cell count and the HIV viral load is often among HIV researchers' interests. Therefore, there is a need to estimate the correlation between two variables, where one of them may be left-censored.

An *ad hoc* but convenient approach is to replace the censored values by, e.g., 1, 1/2 or $1/\sqrt{2}$ times the LOD, depending on the assumed shape of the left tail of the distribution (Hornung and Reed, 1990; Olson, 1993). This method can cause bias and even misinterpretation, especially when the censoring rate is high (Lyles

et al., 2001). Two available parametric methods for the case of one left-censored variable are multiple imputation (MI) and maximum likelihood estimation (ML) assuming an underlying bivariate normal distribution. Comparing the two approaches, Lynn (2001) favored the ML approach in terms of bias. Lyles (2001) found the two methods comparable for point estimates and argued for an MI paradigm on the basis of confidence interval performance. However, the ML approach makes a full normality assumption and the performance of the ML method with non-normal data has not been studied.

In this paper, we propose an alternative approach using generalized estimating equations (GEE) to estimate correlations involving left-censored data, and we compare it with the ML approach via simulated normal and non-normal data. The GEE approach has been widely applied in many statistical applications due to its less stringent distributional assumptions and robustness properties (Liang and Zeger, 1986; Zeger and Liang, 1986). It has also been used to estimate correlation coefficients when data are not censored (Qu *et al*, 1992; Catalano and Ryan, 1992; Barnhart and Williamson, 2001).

In section 2, we present GEE models for estimating correlation coefficients when one variable is subject to left-censoring. In section 3, we report several simulation studies aimed at comparing the GEE and the ML approaches for both normal and non-normal data. In section 4, we apply the proposed method to a real data set stemming from clinical trials conducted in Bangkok, Thailand. In this example, we investigate the correlation between cervicovaginal HIV viral load measurements and CD4 + lymphocyte counts from HIV positive women. We also include an example in this section to illustrate the potential advantage of the proposed GEE approach when modeling covariates that may impact the correlation coefficient. A discussion is presented in section 5.

2. GEE Approach

2.1 Method

Let X and Y be continuous random variables. Let X_L be a left-censored variable corresponding to X with LOD (cut point) L_x. Specifically, we define X_L as follows:

$$X_L = \begin{cases} x & \text{if } X = x \geq L_x \\ x_0 & \text{if } X = x < L_x, \end{cases}$$

where x_0 can be any fixed constant that is less than or equals to L_x (see section 2.3 regarding the choice of x_0). We assume that $(X, Y)'$ has a mean of $(\mu_x, \mu_y)'$

and a covariance matrix of

$$\begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Our main interest is to estimate the correlation ρ between X and Y, where X is subject to left-censoring. For simplicity, we do not introduce covariates initially; an extension to incorporate covariates will be presented later. Although we make use of the bivariate normal distribution when computing expectations in the development of the methodology specified below, the full bivariate normality assumption is not theoretically necessary. Rather, the GEE approach requires only that the moment specifications are approximately correct.

We present two sets of generalized estimating equations for estimating the correlation coefficient ρ, given paired observations of the left-censored variable (X_L) and the complete variable (Y). Let $\{x_{Li}, y_i\}$, $i = 1, \cdots, N$, be random realizations of X_L and Y. In order to estimate the correlation coefficient, one first needs to estimate the parameter $\boldsymbol{\theta} = (\mu_x, \mu_y, \sigma_x, \sigma_y)'$. Letting $\mathbf{Z_i} = (x_{Li}, y_i, x_{Li}^2, y_i^2)'$, $i = 1, \cdots, N$, we propose the first GEE equation to estimate $\boldsymbol{\theta}$ by modeling the marginal mean of $\mathbf{Z_i}$ with $E(\mathbf{Z_i}) = \mathbf{U}(\boldsymbol{\theta})$ as follows:

$$\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V_i}^{-1} (\mathbf{Z}_i - \mathbf{U}(\boldsymbol{\theta})) = \mathbf{0} \tag{2.1}$$

where $\mathbf{U}(\boldsymbol{\theta}) = (\mathbf{U}[1], \mu_y, \mathbf{U}[3], \sigma_y^2 + \mu_y^2)$. Computing moments as if X and Y are distributed as bivariate normal, then, $\mathbf{U}[1] = x_0\Phi(\tau_x) + \mu_x\Phi(-\tau_x) + \sigma_x\phi(\tau_x)$ and $\mathbf{U}[3] = x_0^2\Phi(\tau_x) + (\mu_x^2 + \sigma_x^2)\Phi(-\tau_x) + (L_x + \mu_x)\sigma_x\phi(\tau_x)$, where $\tau_x = (L_x - \mu_x)/\sigma_x$, $\Phi(x)$ denotes the standard univariate normal cumulative density function and $\phi(x)$ refers to the standard normal density function. $\mathbf{D}_i = \partial\mathbf{U}/\partial\boldsymbol{\theta}$ is a 4×4 matrix (see Appendix) and $\mathbf{V}_i$ is the working covariance matrix. The GEE approach uses empirical covariance estimates to adjust for a mis-specified covariance structure without loss of much efficiency (Liang and Zeger, 1986; Zeger and Liang, 1986). For convenience, we take $\mathbf{V}_i$ as a diagonal matrix with diagonal entries as the variances of X_L, Y, X^2 and Y^2 that apply under normality. Thus, we have $\mathbf{V}_i$ $= diag(\mathbf{U}[3] - \mathbf{U}[1]^2,\ \sigma_y^2,\ 2\sigma_x^4 + 4\mu_x^2\sigma_x^2,\ 2\sigma_y^4 + 4\mu_y^2\sigma_y^2)$, where $diag(A)$ denotes the diagonal matrix with vector A as the diagonal elements.

We propose to estimate ρ by modeling the mean of the conditional distribution of $X_{Li}|Y_i$. Note that under normality, $X|Y$ is distributed as $N(\mu_x + \rho\sigma_x(y -$

$\mu_y)/\sigma_y, (1-\rho^2)\sigma_x^2)$. Let $\nu_y(Y) = (Y-\mu_y)/\sigma_y$. Then,

$$\begin{aligned}\gamma_i(\rho,\boldsymbol{\theta}) &= E(X_{Li}|Y_i)\\ &= \int_{L_x}^{\infty} x f(x|y_i)dx + x_0\int_{-\infty}^{L_x} f(x|y_i)dx\\ &= x_0\Phi(\omega_{xy}(y_i)) + (\mu_x + \rho\sigma_x\nu_y(y_i))\Phi(-\omega_{xy}(y_i))\\ &\qquad + \sigma_x\sqrt{1-\rho^2}\phi(\omega_{xy}(y_i)),\end{aligned}$$

where

$$\omega_{xy}(Y) = \tau_x/\sqrt{1-\rho^2} - \rho\nu_y(Y)/\sqrt{1-\rho^2}.$$

We solve for ρ using the following second estimating equation:

$$\sum_{i=1}^{N} C_i' W_i^{-1}(x_{Li} - \gamma_i(\rho,\boldsymbol{\theta})) = 0 \tag{2.2}$$

where

$$\begin{aligned}C_i &= \frac{\partial\gamma_i}{\partial\rho}\\ &= \sigma_x\nu_y(y_i)\Phi(-\omega_{xy}(y_i)) - [\frac{\partial\omega_{xy}(y_i)}{\partial\rho}](L_x - x_0)\phi(\omega_{xy}(y_i))\\ &\qquad - \frac{\rho}{\sqrt{1-\rho^2}}\sigma_x\phi(\omega_{xy}(y_i)),\end{aligned}$$

and

$$\frac{\partial\omega_{xy}(y_i)}{\partial\rho} = \frac{\rho}{(1-\rho^2)^{3/2}}\tau_x - \frac{1}{(1-\rho^2)^{3/2}}\nu_y(y_i).$$

Here, W_i is the working variance of $X_{Li}|Y_i$. In practice, we use $W_i = var(X|Y) = (1-\rho^2)\sigma_x^2$, as obtained under normality. To obtain the point estimate of ρ, a modified Fisher-scoring iterative procedure is used. Specifically, we obtain the estimate of $\boldsymbol{\theta}$, by the iteration process,

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} + \left(\sum_{i=1}^{N}(\hat{\mathbf{D}}_i^{(t)})'(\hat{\mathbf{V}}_i^{(t)})^{-1}(\hat{\mathbf{D}}_i^{(t)})\right)^{-1}\sum_{i=1}^{N}(\hat{\mathbf{D}}_i^{(t)})'(\hat{\mathbf{V}}_i^{(t)})^{-1}(\hat{\mathbf{Z}}_i - \hat{\mathbf{U}}(\hat{\boldsymbol{\theta}}^{(t)})).$$

By replacing $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$ in equation (2.2), the estimate of ρ is obtained by the iteration process,

$$\hat{\rho}^{(t+1)} = \hat{\rho}^{(t)} + \left(\sum_{i=1}^{N}(\hat{C}_i^{(t)})'(\hat{W}_i^{(t)})^{-1}\hat{C}_i\right)^{-1}\sum_{i=1}^{N}(\hat{C}_i^{(t)})'(\hat{W}_i^{(t)})^{-1}(x_{Li} - \gamma_i(\hat{\rho}^{(t)},\hat{\boldsymbol{\theta}})).$$

Following prior arguments regarding generalized estimating equations (Liang, Zeger and Qaqish, 1992), we can readily show that the parameter estimates are consistent provided that $\mathbf{U}(\boldsymbol{\theta})$ and $\gamma_i(\rho, \boldsymbol{\theta})$ are correctly specified. This is true regardless of whether the working covariance matrices in the two sets of equations are correctly specified. To obtain the standard error of $\hat{\rho}$, we follow the similar procedures as in Prentice (1988) and Barnhart and Williamson (2001).

2.2 Evaluating impact of covariates on the correlation

A particular advantage of the proposed GEE approach is that it can easily be extended to investigate covariates' impact on the correlation. Let $\mathbf{Q_1}$ denote the design matrix formed by covariates. First, we model $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = \mathbf{Q_1}\boldsymbol{\beta}$, where the parameter estimates of $\boldsymbol{\beta}$ can be obtained by modifying equation (2.1) as

$$\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V_i}^{-1} (\mathbf{Z}_i - \mathbf{U}_i(\boldsymbol{\beta})) = \mathbf{0} \tag{2.3}$$

where $\mathbf{D}_i = \partial \mathbf{U}_i / \partial \boldsymbol{\beta}$. Second, we use Fisher's Z-transformation to model the correlation coefficient as

$$\frac{1}{2} \log \frac{1+\rho}{1-\rho} = \mathbf{Q_2}\boldsymbol{\alpha}.$$

We use Fisher's Z-transformation here because it ranges from $-\infty$ to ∞ and its quantity is more stable than ρ. $\mathbf{Q_2}$ is a design matrix of covariates impacting ρ that may or may not be the same as $\mathbf{Q_1}$. The estimates of $\boldsymbol{\alpha}$ can be solved by modifying equation (2.2) as

$$\sum_{i=1}^{N} \mathbf{C}_i' W_i^{-1} (x_{Li} - \gamma_i(\boldsymbol{\alpha}, \boldsymbol{\beta})) = 0, \tag{2.4}$$

with $\gamma_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = E(X_{Li}|Y_i)$, a function of α and β, $\mathbf{C}_i = \partial \gamma_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) / \partial \boldsymbol{\alpha}$.

2.3 Remark on the choice of x_0

The censored observations of X are assigned to a fixed value, x_0 (less than or equal to the LOD). In general, the consistency of the proposed GEE estimator holds regardless of the choice of x_0. We have performed simulations to evaluate different choices of x_0 (not shown) and found minimal bias in the GEE estimate in each case. However, defining x_0 based on the appealing condition that $E(X) = E(X_L)$ tends to provide improved efficiency and 95% coverage. This implies that

$$x_0 = \mu_x - \sigma_x \frac{\phi(\tau_x)}{\Phi(\tau_x)}. \tag{2.5}$$

Because μ_x and σ_x are unknown, we employ the following procedure in solving equation (2.1). First, $x_0 = L_x/2$ is used as the initial value, then x_0 is updated with the updated $\hat{\boldsymbol{\theta}}$. Alternatively, one can insert ML estimates (e.g., Cohen, 1959) of μ_x and σ_x assuming a univariate left-censored normal model to get reasonable estimate for x_0.

3. Simulations

We summarize two simulation studies in this section. The first is designed to assess the performance of the GEE estimates for ρ, and to compare them with the MLEs based on normal data. For ease of comparison, simulations were performed using the same true parameter settings as in Lyles *et al.* (2001). Without loss of generality, we set $\boldsymbol{\theta} = (0, 0, 1, 1)'$. The second simulation study is conducted to assess the robustness to the normality assumption using both the proposed GEE approach and the ML approach of Lyles *et al.* (2001).

We first generated continuous responses $\{x_i, y_i\}, i = 1, \cdots, N$, from the standard bivariate normal distribution. Then, $\{x_{Li}\}$ was generated from $\{x_i\}$ according to L_x. The true value of L_x is determined by the censoring rate and the true parameter settings. Simulations were conducted for several combinations of censoring rate (20%, 60%), the true value of ρ (0.25, 0.5, or 0.75) and sample size N (50 or 100), with a total of 12 simulation settings. A total of 1000 simulated data sets were generated for each combination. Table 1 summarizes these simulation results. We report the means of $\hat{\rho}$, the means of the empirically corrected standard error of $\hat{\rho}$, the standard deviation of the 1000 $\hat{\rho}$'s and the 95% coverage based on the estimated standard error of $\hat{\rho}$. Corresponding results using ML (Lyles *et al.*, 2001a) are also included for comparison. We observe that the GEE approach performed similarly to ML, although there is a tendency toward underestimation of the standard error of $\hat{\rho}$ for sample size $N = 50$. As suggested by Barnhart and Williamson (2001), an adjustment to the estimated standard error, such as multiplying by $\sqrt{N/(N-1)}$ or $\sqrt{(N/(N-2)}$, may be useful when N is less than or equal to 50.

To explore the robustness to non-normality of the GEE and ML approaches, we considered two kinds of non-normally distributed data: correlated uniform data in the interval $(-1, 1)$, and correlated chi-square data (with degrees of freedom 10, moderately skewed). These correlated non-normal data were generated by a method similar to that of Saucier (2000). Although data were generated from non-normal distributions, the mean specifications in the two sets of estimating equations and the full likelihood for the ML method were based on the normality assumption and thus misspecified, motivating the robustness study.

Simulations were performed for each combination of the censoring rate (20%, 60%) and true ρ (0.25 or 0.5). We used a sample size of 100 and a total of

Table 1: Simulation Study for the Case of One Left-censored Variable

True ρ	Sample Size	Method	Mean $\hat{\rho}$	Mean Std.err	Empirical Std. Dev.	95% Coverage
			X is 20% *left-censored,* $L_x = -0.842$			
0.25	50	MLE	0.256	–	0.135	0.930
		GEE	0.249	0.130	0.137	0.928
	100	MLE	0.248	–	0.099	0.933
		GEE	0.249	0.094	0.097	0.928
0.50	50	MLE	0.499	–	0.111	0.927
		GEE	0.498	0.106	0.112	0.924
	100	MLE	0.500	–	0.080	0.926
		GEE	0.498	0.077	0.079	0.924
0.75	50	MLE	0.749	–	0.069	0.930
		GEE	0.749	0.065	0.070	0.919
	100	MLE	0.747	–	0.047	0.918
		GEE	0.749	0.047	0.049	0.924
			X is 60% *left-censored,* $L_x = 0.253$			
0.25	50	MLE	0.248	–	0.160	0.915
		GEE	0.250	0.152	0.162	0.915
	100	MLE	0.254	–	0.110	0.936
		GEE	0.249	0.110	0.114	0.927
0.50	50	MLE	0.498	–	0.130	0.923
		GEE	0.499	0.130	0.141	0.900
	100	MLE	0.500	–	0.091	0.932
		GEE	0.499	0.094	0.096	0.927
0.75	50	MLE	0.751	–	0.081	0.903
		GEE	0.750	0.084	0.092	0.908
	100	MLE	0.750	–	0.055	0.928
		GEE	0.749	0.060	0.063	0.930

1000 simulated data sets in each case. As shown in Table 2, both the GEE and ML approaches continue to perform reasonably well for correlated uniform data and correlated chi-square data in the context of point estimates and 95% coverage. The point estimate for ρ from the GEE approach shows minimal bias for correlated uniform data with a low or moderate censoring rate, and a slight tendency toward underestimation for moderately skewed correlated chi-square data. Compared with the GEE estimates, the MLEs of ρ tend to be slightly larger in the latter case. For correlated uniform data, the mean standard error based on the GEE method is closer to the empirical standard deviation than the one based on the ML method, where the empirical standard deviation is calculated from 1000 estimates obtained from the 1000 simulated data sets. The disparity between the two based on ML translates into the overly conservative 95% coverage observed in Table 2.

In summary, the proposed GEE method performs very consistently with ML

Table 2: Simulation Study for Investigating Robustness of Normality Assumption ($N = 100$)

True ρ	Censored Rate	Method	Mean $\hat{\rho}$	Mean Std.err	Empirical Std.dev	95% Coverage
			Correlated Uniform Data			
0.25	20%	GEE	0.254	0.077	0.080	0.932
		MLE	0.252	0.094	0.079	0.970
	60%	GEE	0.237	0.102	0.099	0.952
		MLE	0.249	0.111	0.098	0.967
0.50	20%	GEE	0.500	0.061	0.063	0.928
		MLE	0.498	0.075	0.059	0.980
	60%	GEE	0.477	0.089	0.089	0.945
		MLE	0.510	0.089	0.077	0.968
			Correlated Chi-square Data			
0.25	20%	GEE	0.248	0.093	0.095	0.936
		MLE	0.245	0.095	0.096	0.944
	60%	GEE	0.239	0.102	0.100	0.945
		MLE	0.258	0.106	0.101	0.947
0.50	20%	GEE	0.488	0.077	0.081	0.924
		MLE	0.505	0.075	0.082	0.916
	60%	GEE	0.464	0.090	0.094	0.922
		MLE	0.515	0.084	0.086	0.921

for point estimation under bivariate normality. For heavy-tailed and moderately skewed non-normal data, both the GEE and ML methods remain relatively robust, with the GEE method marginally better than ML with respect to closeness of the estimated standard errors to the empirical standard deviations based on simulated data sets .

4. Examples

Ever since the emergence of HIV viral load as a new virologic marker for describing HIV/AIDS progression, researchers have been interested in the correlation of the HIV viral load measurements with the immunologic marker, the CD4+ cell count, which was traditionally the standard benchmark for prognosis of HIV patients (Mulder, *et al.*, 1994; Mellors *et al.*, 1995). Because the assay for HIV viral load has a lower limit of detection, it is important to account for this in estimating the correlation between the HIV viral load and CD4+ cell counts. We apply the proposed methods to estimate this correlation from a HIV study from a Centers for Disease Control and Prevention (CDC)-sponsored clinical trial, conducted in Bangkok, Thailand. Data were collected from 154 women in a Zidovudine-treated subgroup of the study. Available data are CD4+ cell

counts (Y) at 36 weeks' gestation and HIV viral load from cervicovaginal lavage (CVL, X_L) at 38 weeks' gestation. Details of the study and the data were reported elsewhere (Shaffer *et. al.*, 1999). A total of 120 out of 154 women had non-detectable CVL measurements, implying a 77.9% left censoring rate. As suggested by Lyles *et al.* (2001), a log transformation was applied to the CVL measurements. As the lower limit of detection of the assay is 400 copies/ML, we have $L_x = log_{10}(400) = 2.60206$ in this example. Results of the analysis are presented in Table 3. We also include results based on 1000 simulated data sets, where the true parameters are set to the GEE estimates $\hat{\boldsymbol{\theta}} = (1.572, 444.63, 1.335, 201.78)'$ and the sample size is the same as in the data set. The corresponding results from ML and the *ad hoc* method that replaces censored values by LOD/2 are also included for comparison. Both the GEE and ML methods give similar results, while the *ad hoc* approach is strongly biased. These results suggest that the correlation between CVL measurements and CD4+ counts for HIV-infected women in the zidovudine group is negative, with an estimated magnitude of 0.24.

Table 3: Correlation between CD4+ cell counts and log-transformed HIV RNA in CVL

	Analysis Results		Results based on simulations		
Method	$\hat{\rho}$	$95\%CI$	Mean $\hat{\rho}$	Std.dev	Coverage
MLE¶	-0.2430	$(-0.4610, -0.0250)$	-0.2420	0.1040	0.939
GEE†	-0.2385	$(-0.4774, 0.0004)$	-0.2386	0.1102	0.933
Ad hoc§	-0.1420	–	-0.1800	0.0780	–

¶. Simulations are based on parameters setting $\boldsymbol{\theta} = (1.558, 444.63, 1.348, 201.78)'$, and $\rho = -0.2430$

†. Simulations are based on parameters setting $\boldsymbol{\theta} = (1.572, 444.63, 1.335, 201.78)'$, and $\rho = -0.2385$

§. Values of CVL below LOD are set to LOD/2.

To illustrate the GEE approach for data with covariates, we use additional data from the same study as in the first example by including an additional 155 women who received a placebo instead of Zidovudine. A total of 74 women in the placebo group had non-detectable measurements of HIV RNA in CVL, implying a 47.7% censoring rate, as compared to the 77.9% censoring rate among women in Zidovudine group. We seek to determine whether the correlation between CD4+ counts and HIV RNA measurements in CVL for women in the Zidovudine group is significantly different from that in the placebo group. We model $\boldsymbol{\theta}$ by $\boldsymbol{\theta} = \mathbf{Q_1}\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x$ and $(1/2)log((1+\rho)/(1-\rho)) = \mathbf{Q_2}\boldsymbol{\alpha} = \alpha_0 + \alpha_1 x$, where x is the indicator variable for the Zidovudine group. The estimate of

$\boldsymbol{\alpha} = (\alpha_0, \alpha_1)'$ is $(-0.323, 0.090)'$ with standard error of $(0.189, 0.274)$. Thus the estimated correlations are -0.23 and -0.31 for the Zidovudine and placebo groups, respectively. This GEE result immediately implies a p-value of 0.37 for testing $\alpha_1 = 0$, indicating that the correlation between CD4+ cell counts and HIV RNA measurements in CVL for HIV-infected women in the Zidovudine group is not significantly different from that in the placebo group.

5. Discussion

We have proposed a GEE approach for estimating the correlation between two continuous variables, where one variable is subject to left censoring. We evaluated the method by simulation studies and illustrated it with a clinical example. We also investigated the robustness to the normality assumption for both the ML approach and the GEE approach via simulations. As an alternative to ML, the GEE estimates for ρ performed comparably in both simulation studies and on real data sets. In addition to providing a unified framework for estimation of and inference about ρ, GEE may be easier to use than ML with respect to modeling covariates' impact on the correlation. Although we make use of the bivariate normality assumption in computing expectations, the full normal distribution is not required as long as the means are correctly specified in the two GEE sets. Based on the simulation study for non-normal data with misspecified means, we found that the GEE method performed similarly to or marginally better than the ML method for uniform and moderately skewed data.

Although the ML approach may be used as a standard method for estimating the correlation between left-censored variables, we tend to favor use of the GEE approach because of its theoretical advantage, because it performs similarly to ML without requiring optimization routines, and because of its convenience for modeling covariates' impact on correlation. When one or both variables under study are very strongly skewed, we recommend use of a log or other transformation before applying either approach (as seen in the real data examples). Additional simulation studies (not detailed here) for highly right-skewed bivariate log-normal data indicated potentially poor confidence interval coverage without a transformation.

We have employed a two-stage estimating equations approach. Specifically, we estimate $\boldsymbol{\theta}$ in the first set of equations and then plug into the second set of equations to estimate ρ. Another alternative (not explored here) would be to use a three-stage approach, i.e, to estimate $(\mu_y, \sigma_y)'$ in the first equation, estimate $(\mu_x, \sigma_x)'$ in the second equation, and then estimate ρ in the third equation. We believe that the 2-stage approach is more efficient than the 3-stage approach in estimating $\boldsymbol{\theta}$ (Liang, *et al.*, 1992).

An alternative way to estimate ρ is to model the mean of the product $(E(X_L Y))$,

instead of the conditional mean ($E(X_L|Y)$). However, $E(X_L Y)$ involves double integration and is thus computationally intensive. Furthermore, according to Carey *et al.* (1993), modeling the conditional mean has reasonable efficiency as compared to modeling the mean of the product. Therefore, we prefer our approach to estimate ρ by modeling the conditional mean.

The computer programs used in this research were written in Splus (Windows 2000) and R (Version 1.5.1, 2002). All programs are available from the first author upon request.

Acknowledgments

The data used in the examples were provided by the Bangkok Collaborative Perinatal HIV Transmission Study Group (Chuachoowong *et al.*, 2000; Shaffer *et al.*, 1999), and we sincerely thank the participants and investigators. The full list of the Bangkok study group could be found in Shaffer *et al.* (1999). This research is partially supported by the Emory Quadrangle Research Grant (for Jingli Song and Huiman X. Barnhart). We sincerely thank the editor and the reviewer for their valuable comments that led to the improvement of the manuscript.

Appendix

Referring to equation (2.1), we have

$$\mathbf{D}_i = \frac{\partial \mathbf{U}_i}{\partial \boldsymbol{\theta}} = \begin{pmatrix} d_{11} & 0 & d_{13} & 0 \\ 0 & 1 & 0 & 0 \\ d_{31} & 0 & d_{33} & 0 \\ 0 & 2\mu_y & 0 & 2\sigma_y \end{pmatrix},$$

where
$d_{11} = \Phi(-\tau_x) + (L_x - x_0)\phi(\tau_x)/\sigma_x$,
$d_{13} = \phi(\tau_x)[\tau_x(L_x - x_0)/\sigma_x + 1]$,
$d_{31} = 2\mu_x\Phi(-\tau_x) + \phi(\tau_x)(L_x^2 + 2\sigma_x^2 - x_0^2)/\sigma_x$, and
$d_{33} = 2\sigma_x\Phi(-\tau_x) + \phi(\tau_x)[(L_x + \mu_x)(1 + \tau_x^2) + \tau_x(\mu_x^2 + \sigma_x^2 - x_0^2)/\sigma_x]$.

References

Barnhart, H. X. and Williamson, J. M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics* **57**, 931-940.

Carey, V., Zeger, S. L. and Diggle, P. (1993). Modeling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517-526.

Catalano, P. J. and Ryan, M. L. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of American Statistical Association* **87**, 651-658.

Chuachoowong, R., Shaffer, N., Siriwasin, W., Chaisilwattana, P., Young, N.L., Mock, P.A., Chearskul, S., Waranawat, N., Chaowanachan, T., Karon, J., Simonds, R.J., Mastro, T.D. (2000). Short-course antenatal zidovudine reduces both cervicovaginal human immunodeficiency virus type I RNA levels and risk for perinatal transmission. *Journal of Infectious Disease 2000* **181**, 99-106.

Cohen, A. C. (1959) Simplified estimators for normal distribution when samples are singly censored or truncated. *Technometrics* **1**, 217-237.

Hornung, R. W. and Reed, L. D. (1990). Estimation of average concentration in the presence of nondetectable value. *Applied Occupational and Environmental Hygiene* **4**, 46-51.

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

Liang, K. Y, Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54**, 3-24.

Lyles, R. H., Fan, D., and Chuachoowong, R. (2001) Correlation coefficient estimation involving a left censored laboratory assay variable. *Statistics in Medicine* **20**, 2921-2933.

Lynn, H. S. (2001) Maximum likelihood inference for left-censored HIV RNA data. *Statistical in Medicine* **20**, 33-45.

Mellors, J. W., Kingsley, L. A., Rinaldo, C. R. Jr., Todd, J. A., Hoo, B. S., Kokka, R. P. *et al.* (1995) Quantization of HIV-1 RNA in plasma predicts outcome after seroconversion. *Annals Internal Medicine* **122**, 573-579.

Mulder, J., McKinney, N., Christopherson, C., Sninsky. J., Greenfield, L., and Kwod, S. (1994). Rapid and simple PCR assay for quantification of human immunodeficiency virus type 1 RNA in plasma: Application to acute retroviral infection. *Journal of Clinical Microbiology* **32**, 292-300.

Olson, D. R. (1993). A simple method for estimation when there is a detection limit. *Joint Statistical Meeting of American Statistical Society and Biometric Society* San Francisco. (abstract)

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

Qu, Y., Williams, W., Beck, J. and Medendorp, S. V. (1992). Latent variable models for clustered dichotomous data with multiple subclusters. *Biometrics* **48**, 1095-1102.

R (2002). *The R Development Core Team*, http://cran.r-project.org, **Version 1.5.1**.

Saucier, R. (2000). Computer Generation of Statistical Distribution. *Army Research Laboratory* ARL-TR-2168.

Shaffer, N., Chuachoowong, R., Mock, P.A., *et al.* (1999). Short-course zidovudine for perinatal HIV-1 transmission in Bangkok, Thailand: A randomized controlled trial. *Lancet* **353**, 773-780.

Splus 2000 (1999). *S-PLUS 2000 Guide to Statistics, Volume 2.* Data Analysis Products Division, MathSoft.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Received April 20, 2003; accepted June 17, 2003.

Jingli Song
DC 6134
Eli Lilly and Company
Lilly Corporate Center
Indianapolis, IN 46285, USA
songji@lilly.com

Huiman X Barnhart
Department of Biostatistics and Bioinformatics
and Duke Clinical Research Institute
Duke University
P O BOX 17969
Durham, NC 27715, USA
huiman.barnhart@duke.edu

Robert H. Lyles
Department of Biostatistics
The Rollins School of Public Health
Emory University
1518 Clifton Road, NE
Atlanta GA 30322, U.S.A