Weighting Condom Use Data to Account for Nonignorable Cluster Size

JOHN M. WILLIAMSON, MSc, ScD, HAE-YOUNG KIM, MSc, AND LEE WARNER, MPH, PhD

PURPOSE: We examined the impact of weighting the generalized estimating equation (GEE) by the inverse of the number of sex acts on the magnitude of association for factors predictive of recent condom use.
METHODS: Data were analyzed from a cross-sectional survey on condom use reported during vaginal intercourse during the past year among male students attending two Georgia universities. The usual GEE model was fit to the data predicting the binary act-specific response indicating whether a condom was used. A second cluster-weighted GEE model (i.e., weighting the GEE score equation by the inverse of the number of sex acts) was also fit to predict condom use.
RESULTS: Study participants who engaged in a greater frequency of sex acts were less likely to report condom use, resulting in nonignorable cluster-size data. The GEE analysis weighted by sex act (usual GEE) and the GEE analysis weighted by study subject (cluster-weighted GEE) produced different estimates of the association between the covariates and condom use in last year. For example, the cluster-weighted GEE analysis resulted in a marginally significant relationship between age and condom use (odds ratio of 0.49 with 95% confidence interval (0.23–1.03) for older versus younger participants) versus a nonsignificant relationship with the usual GEE model (odds ratio of 0.67 with a 95% confidence interval of 0.28–1.60).
CONCLUSIONS: The two ways of weighting the GEE score equation, by the sex act or by the respondent, may produce different results and a different interpretation of the parameters in the presence of nonignorable cluster size.
Ann Epidemiol 2007;17:603–607. © 2007 Elsevier Inc. All rights reserved.

KEY WORDS: Condom use, Generalized Estimating Equations, HIV Infections, Informative Cluster Size, Sex Behavior, Sexually Transmitted Diseases.

INTRODUCTION

Approximately 19 million cases of sexually transmitted diseases (STD) occur in the United States each year (1). For persons who are sexually active, male latex condoms remain a critical component of public health strategies for prevention of STD (2). When used consistently and correctly, condom use also has been associated with reduced risk of human immunodeficiency virus (HIV) as well as many other STDs (3–7). Although levels of condom use have increased in recent years (8–10), overall use remains suboptimal for effective prevention of STD. Thus, there is continued interest in identifying demographic and behavioral characteristics of persons who report using condoms for STD prevention to aid development of interventions.

One potential difficulty in identifying predictors of condom use is the effect of clustering of correlated data on sex acts within individuals. Such clusters of correlated observations often arise in public health or biomedical studies (e.g., longitudinal data, familial data, ophthalmology data) (11). Similarly, studies of sexual behavior are subject to many of the same issues involving correlated data. For example, in studies of condom use, data are correlated as the result of subjects' use of condoms on multiple sex acts (12). In this case, the cluster would be the subject and the individual sex act would be the observation (subunit) within the cluster.

Generalized estimating equations (GEE) (13, 14) are a common method for valid estimation of the marginal parameters while taking into account this correlation within clusters and has been used to analyze such data on condom use (12, 15). GEE involves specification of a "working" correlation matrix for the observations within a cluster to allow estimation of the marginal parameters. Then, the parameter estimate standard errors are "empirically-corrected" through the use of the sandwich estimator allowing valid inference. Choosing an independence working correlation matrix for the GEE analysis results in an equal weight for each observation and clusters of greater size have proportionately greater weight in the parameter estimation.

It is implicitly assumed with the usual GEE analysis that the response is independent of the cluster size (the number of observations in the cluster). However, in a number of applications the response among cluster members may be related to the cluster size (known as nonignorable or

From the National Center for Infectious Diseases, Division of Parasitic Diseases, Centers for Disease Control and Prevention (CDC), Atlanta, GA (J.M.W.); Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC (H.-Y.K.); and National Center for Chronic Disease Prevention and Health Promotion, Division of Reproductive Health, Centers for Disease Control and Prevention (CDC), Atlanta, GA (L.W.).

Address correspondence to: John M. Williamson, ScD, Centers for Disease Control and Prevention, National Center for Infectio, Atlanta, GA 30341. E-mail: jow5@cdc.gov.

Received September 29, 2006; accepted March 26, 2007.

360 Park Avenue South, New York, NY 10010

1047-2797/07/$–see front matter
doi:10.1016/j.annepidem.2007.03.008

Selected Abbreviations and Acronyms

STD = sexually transmitted disease
HIV = human immunodeficiency virus
GEE = generalized estimating equations
WCR = within-cluster resampling
CWGEE = cluster-weighted generalized estimating equations
OR = odds ratio
CI = confidence interval

informative cluster size) (16). For example, dental studies may result in nonignorable cluster size data. Assume that the response is the health of each tooth. Persons with a dental disease may already have lost some teeth as a result of the disease and, therefore, the number of teeth (cluster size) in a person's mouth is related to the outcome. Hoffman et al. proposed a within-cluster resampling (WCR) method that remains valid for the analysis of clustered data when cluster size is nonignorable (16). Inversely weighting the GEE score equation by cluster size with an independence working correlation matrix (cluster-weighted generalized estimating equations, CWGEE) has been shown to be asymptotically equivalent to WCR (17, 18).

Here, we illustrate the impact of using a CWGEE analysis versus the usual unweighted GEE analysis by applying both analyses to a cross-sectional study on condom use that was conducted on a sample of male university students attending two Georgia universities (12). The objective of this article was to compare how weighting the GEE analysis impacts the magnitude of association for factors predicting condom use.

MODELS

GEE

Let Y_{ij} denote the response of the j^{th} subunit in the i^{th} cluster, $i = 1,\ldots, N$. For our example, the cluster is the male study participant and the subunit is the vaginal sex act. The response Y_{ij} may denote either a binary, count or continuous random variable. Let $\boldsymbol{Y_i} = [Y_{i1},Y_{i2},\ldots Y_{in_i}]'$ denote the response vector for the i^{th} respondent with $\mathbf{E[Y}_i\mathbf{]} = \boldsymbol{\mu}_i$, where n_i is the number of subunits in the i^{th} cluster (cluster size). The $n_i \times (P+1)$ design matrix for the i^{th} subject is denoted by $\boldsymbol{X}_i = [\mathbf{1},\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\ldots,\boldsymbol{x}_{ip}]$, where $\mathbf{1}$ is a $n_i \times 1$ vector of ones. The covariates may be either cluster-specific (related to the study participant) or subunit-specific (related to the specific sex act). The usual GEE modeling has the following setting:

$$g(\mu_{ij}) = x_{ij}\boldsymbol{\beta}, \qquad (1)$$

where x_{ij} is the covariate vector for the j^{th} subunit of the i^{th} cluster, $\beta = [\beta_0,\beta_1,\ldots,\beta_P]'$ is the marginal parameter vector of interest, and $g(\cdot)$ is a link function relating the mean response with the covariates and parameters (e.g., the logit link function for a binary response). Estimation of β is obtained by solving the generalized estimating equations:

$$\nu(\hat{\beta}) = \sum_{i=1}^{N} \hat{D}_i'\hat{V}_i^{-1}(Y_i - \mu_i(\hat{\beta})) = 0, \qquad (2)$$

where $\boldsymbol{D}_i = \boldsymbol{D}_i\ (\beta) = d\boldsymbol{\mu}i(\beta)/d\beta$, $\mathbf{V}_i = \mathbf{V}_i(\beta,\alpha) = \mathbf{A}_i^{1/2}\mathbf{R}_i\mathbf{A}_i^{1/2} \approx \mathrm{var}(\mathbf{Y}_i)$, $\mathbf{A}_i = diag(\mathrm{var}(y_{i1}),\ldots, \mathrm{var}(y_{in_i}))$, and $\mathbf{R}_i$ is the working correlation matrix for $\mathbf{Y}_i$ (13, 14). The empirical estimate of variance can be consistently estimated by

$$\hat{V}_\beta = N\left[\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\hat{D}_i\right]^{-1}\left[\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}(Y_i - \hat{\pi}_i) \times (Y_i - \hat{\pi}_i)'\hat{V}_i^{-1}\hat{D}_i\right]\left[\sum_{i=1}^{N}\hat{D}_i'\hat{V}_i^{-1}\hat{D}_i\right]^{-1}$$

(13, 14). This "robust" variance estimator allows valid inference on the parameters even when the within-cluster dependence structure is misspecified. Each cluster of observations is weighted inversely to its variance matrix $\mathbf{V}_i$, which is a function of the working correlation matrix $\boldsymbol{R}_i$ (11). Choosing an independence working correlation matrix for a GEE analysis results in an equal weight for each observation and clusters with greater size have proportionately greater weight in the parameter estimation.

Analysis of clustered data focusing on inference of the marginal distribution may be problematic when cluster size is nonignorable. WCR is a method for analyzing such data which still results in valid parameter estimation when estimation of marginal effects weighted at the cluster level is of interest (16). WCR is based on resampling replicate data sets each containing one observation from each cluster. Each resampled data set is analyzed with an appropriate marginal model (e.g., Poisson regression for a count response since the observations are now independent). The resulting WCR parameter estimate is the average of the parameter estimates from the analyses of each of the resampled data sets. The variance of the WCR estimator is estimated as the average of the variances of the parameter estimates from the replicated data sets minus the variance-covariance matrix of parameter estimates from the replicate data sets. See Hoffman et al. for details (16).

CWGEE

Weighting the GEE score equation by the inverse of the cluster size (n_i) while using an identity working correlation matrix ($\mathbf{V}_i = \boldsymbol{I}_n$):