University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2009 *Paper* 251

Nonparametric population average models: deriving the form of approximate population average models estimated using generalized estimating equations

Alan E. Hubbard[*] Mark J. van der Laan[†]

[*] Division of Biostatistics, UC Berkeley, hubbard@berkeley.edu

[†] University of California - Berkeley, laan@berkeley.edu

Nonparametric population average models: deriving the form of approximate population average models estimated using generalized estimating equations

Alan E. Hubbard and Mark J. van der Laan

Abstract

For estimating regressions for repeated measures outcome data, a popular choice is the population average models estimated by generalized estimating equations (GEE). We review in this report the derivation of the robust inference (sandwich-type estimator of the standard error). In addition, we present formally how the approximation of a misspecified working population average model relates to the true model and in turn how to interpret the results of such a misspecified model.

1 Introduction

First, introduce a simple repeated measures data structure. Let Y_{ij} be the outcome for sub-unit *i* within unit *j*, and associated covariates $X_{ij},$ For this paper, the parameter of interest I the so-called population average model or $E(Y_{ij}|X_{ij}),$ which is modeled as:

$$E(Y_{ij} \mid X_{ij}) = m(X_{ij} \mid \beta) = g[\mu(X_{ij} \mid \beta)] \quad (1)$$

g is the link function depends on the regression (e.g., linear: $g^{-1}(u)=u$, $log{:}g^{-1}(u)=log(u)$, logistic: $g^{-1}(u)=log[u/(1-u)]$), β are the regression coefficients *u* is a linear function of basis functions constructed from the vector of covariates, X_{ij}. Typically, a generalized estimation equation approach is used to estimated the β (Liang and Zeger, 1986).

A reasonably parameter one could estimate is an informative approximation of the true population average model, $g[\mu(X_{ij} \mid \beta)]$. With this approach, one could avoid any model specification (i.e., the model is nonparametric) and the parameter of interest would be some projection of $E[Y_{ij}|X_{ij}]$ onto the proposed class of estimating models (e.g., linear with only main effects); one can define explicitly the parameter of interest as a function of the distribution of the observed data (X,Y)

Below, we first review the generalized estimating equation approach and somewhat informally discuss the derivation of the robust inference. Then, we derive parameter of interest as some projection of the working model onto the true population average model, where the projection is a function of the so-called working correlation model.

2 Generalized Estimating Equations

The general form of this estimating function for regression problems is (see van der Laan and Robins, 2003):

$$D_V(X_{.j},Y_{.j} \mid \beta) = \frac{\partial m(X_{.j} \mid \beta)}{\partial \beta} V^{-1}\left\{\varepsilon_{.j}(\beta)\right\}\varepsilon_{.j}(\beta)^T \qquad (2)$$

where $X_{.j}$ represents the design matrix for all observations in neighborhood j, $Y_{.j}$ represents the vector of outcomes observations corresponding to the rows of $X_{.j}$, $m(X_{.j}|\beta)=\{m(X_{ij}|\beta),\ i=1,\ldots,n_j\}$ a vector of regression functions corresponding again to each row of $X_{.j}$, $\varepsilon_{.j}(\beta)=Y_{.j} - m(X_{.j}\ |\beta) = \{Y_{ij}\text{-}m(X_{ij}|\beta),\ i=1,\ldots,n_j\}$, β a $px1$ vector of coefficients and $V\{\varepsilon_{.j}(\beta)\}$ is the working n_j by n_j variance-covariance of the residuals, consisting of the entries, $cov(\varepsilon_{ij},\ \varepsilon_{i'j})$, based on the working correlation model. One can show that this is a consistent estimating function, that is, it has mean 0 if the true β's are entered. The estimator of β is defined by setting the average of these estimating functions for each unit (e.g., neighborhood) to 0 and solving for β:

$$0 = \sum_{i=1}^{m} D_{\hat{V}}(X_{.j},Y_{.j} \mid \beta). \qquad (3)$$

Instead, robust or sandwich inference is typically provided (Liang and Zeger, 1986). Without going into the technical details, one can show that these estimators are asymptotically linear, which means that:

$$\beta_m - \beta = \frac{1}{m}\sum_{j=1}^{m} IC(X_{.j},Y_{.j} \mid \beta) + o_p\left(\frac{1}{\sqrt{m}}\right) \qquad (4)$$

where β_n is the estimate based on *m* observations (neighborhoods); *IC* is the so-called influence curve, is the same dimension as β and is derived from the estimating function (2); and the last term, $o_p(1/\sqrt{m})$, is a second order term that becomes negligible as *m* gets large. Thus, one can derive the variance estimate of β_n and the resulting standard errors by simply looking at the empirical variance-covariance of the components of the *IC*. In this case, the *IC,* a standardized version of the estimating function, is:

$$IC(X_{\cdot j},Y_{\cdot j} \mid \beta) = E\left\{h(X_{.j})\frac{\partial m(X_{\cdot j} \mid \beta)}{\partial \beta^T}\right\}^{-1} h(X_{.j})\varepsilon_{\cdot j}(\beta)^T$$

where $h(X_{.j}) = \dfrac{\partial m}{\partial \beta}(X_{\cdot j} \mid \beta)^T_{pxn_j} V^{-1}\left\{\varepsilon_{\cdot j}(\beta)\right\}_{n_j x\, n_j}$.

3 Definition of Projections Related to GEE regression models

The motivation for defining the parameter of interest as some an approximation of a population average model borrows extensively from Neugebauer and van der Laan, 2007). The true population average conditional mean is $\mu^{*}=E(Y_{ij}|X_{ij})$. Define $\mu : \Re^p \rightarrow \mathrm{M}$, where $\mu(X \mid \beta) \subset \mathrm{M}$ is the working model (informative approximation) and denote M as the set of possible working models (μ is the proposed approximation of μ^{*}). Note that μ^{*} is not necessarily an element of M (the class of working models does not necessarily contain the true model); we will refer to $\mu(X|\beta)$ as the model approximation. The parameter of interest in the context of misspecification of μ^{*} is: $\beta(P_{X,Y} \mid A)$, where $P_{X,Y}$ represents the distribution of the data and is a function of some weight matrix, $A = V$ ^{1} in which the variance is derived from the mean according to some guessed distribution (e.g., $var(Y_{ij}|X_{ij})=E(Y_{ij}|X_{ij})$ if model assumed is log-linear, Poisson model; see Liang and

Zeger, 1986 for details). The estimation function (2) can be derived from solving the objective function:

$$\beta_A = \beta(P_{X,Y} \mid A) \equiv \underset{\beta \in \Re^k}{\arg\min} E_X\left[\varepsilon^T(\beta) A \varepsilon(\beta)\right], \quad \varepsilon_{.j}(\beta) = Y_{.j} - \mu(X_{.j} \mid \beta).$$

This can be shown to be equivalent to:

$$\beta_A = \underset{\beta \in \Re^k}{\arg\min} E_X\left[\varepsilon^{*T}(\beta) A \varepsilon^*(\beta)\right], \quad \varepsilon^*_{.j}(\beta) = u^*(X_{.j}) - \mu(X_{.j} \mid \beta), (5)$$

or the β_A that minimizes the weighted Euclidean distance between the model and the true mean. This can be seen by differentiation product rule, so estimation function derived from (5) is (dropping the vector *j* subscript):

$$\frac{d}{d\beta} E_X\left[\varepsilon^{*T}(\beta) A \varepsilon^*(\beta)\right] = 0 \to E_X\left\{\frac{\partial \mu(X \mid \beta)^T}{\partial \beta} A[\mu^*(X) - \mu(X \mid \beta)]\right\} = 0.$$

But, by the law of iterated expectation, one can exchange *Y* for $\mu^*(X)$ this is the same as solving:

$$E_X\left\{\frac{\partial \mu(X \mid \beta)^T}{\partial \beta} A[Y - \mu(X \mid \beta)]\right\} = E_X E\left\{\frac{\partial \mu(X \mid \beta)^T}{\partial \beta} A[Y - \mu(X \mid \beta)] \middle| X\right\} =$$
$$E_X\left\{\frac{\partial \mu(X \mid \beta)^T}{\partial \beta} A[\mu^*(X) - \mu(X \mid \beta)]\right\}.$$

Thus, using the GEE estimating equations to estimate the coefficients, β_A, as part of a misspecified model, $\mu(X \mid \beta)$, is asymptotically equivalent to the β_A that minimize the weighted Euclidean distance of the model to the true mean, where different weight matrices A (working model for the inverse of the variance-covariance matrix of the vector of repeated measures, Y) can result in different coefficients. Thus, the approximation model varies as a function of the working correlation structure chosen by the user (e.g., independence, auto-correlation, unstructured, etc), so that the estimate will represent estimates of different parameters for different working correlation structures Under misspecification of the population average model, the typical GEE algorithm will provide estimation and proper robust inference for β_A. Thus, in this case, we have defined the coefficient parameters that GEE estimates as explicit nonparametric functions of the distribution of the data, $P_{X,Y}$.

References

Liang K-Y, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1):13-22.

Neugebauer R, van der Laan MJ. 2007. Nonparametric Causal Effects based on marginal structural models, *Journal of Statistical Planning and Inference Journal of Statistical Planning and Inference* **137**:419–434.

van der Laan, MJ, Robins, JM. 2003. *Unified Methods for Censored Longitudinal Data and Causality*. Springer-Verlag, New York.