Harvard University
Harvard University Biostatistics Working Paper Series

Year 2012 *Paper* 151

Locally Efficient Estimation of Marginal Treatment Effects when Outcomes are Correlated: Is the Prize Worth the Chase?

Alisa J. Stephens* Eric J. Tchetgen Tchetgen†
Victor De Gruttola‡

*Harvard University, alisa.j.stephens@gmail.com
†Harvard University, etchetge@hsph.harvard.edu
‡Harvard University, degrut@hsph.harvard.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/harvardbiostat/paper151

Copyright ©2012 by the authors.

Locally efficient estimation of marginal treatment effects when outcomes are correlated: is the prize worth the chase?

Alisa J. Stephens*, Eric J. Tchetgen Tchetgen, and Victor De Gruttola

Department of Biostatistics, Harvard University, Boston MA 02115, USA

Abstract

Semiparametric methods have been developed to increase efficiency of inferences in randomized trials by incorporating baseline covariates. Locally efficient estimators of marginal treatment effects, which achieve minimum variance under an assumed model, are available for settings in which outcomes are independent. The value of the pursuit of locally efficient estimators in other settings, such as when outcomes are multivariate, is often debated. We derive and evaluate semiparametric locally efficient estimators of marginal mean treatment effects when outcomes are correlated; such outcomes occur in randomized studies with clustered or repeated-measures responses. The resulting estimating equations modify existing generalized estimating equations (GEE) by identifying the efficient score under a mean model for marginal effects when data contain baseline covariates. Locally efficient estimators are implemented for longitudinal data with continuous outcomes and clustered data with binary outcomes. Methods are illustrated through application to AIDS Clinical Trial Group Study 398, a longitudinal randomized clinical trial that compared the effects of various protease inhibitors in HIV-positive subjects who had experienced antiretroviral therapy failure. In addition, extensive simulation studies characterize settings in which locally efficient estimators result in efficiency gains over suboptimal estimators and assess their feasibility in practice.

CORA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Keywords: Clinical trials; Correlated data; Covariate adjustment; Semiparametric efficiency

1 Introduction

Semiparametric estimators are appealing because of their robustness to distributional assumptions and model misspecification. In the analysis of randomized trials, semiparametric theory has been used to develop estimators of treatment effects that improve efficiency of inferences by incorporating baseline covariates, where 'baseline' describes data measured prior to randomization. In this paper, we present a semiparametric locally efficient estimator to improve efficiency of inferences in randomized experiments with correlated outcomes when baseline covariates are available. We begin with a review of current estimators for multivariate outcomes and then introduce our locally efficient estimator.

Correlated outcomes are often observed in medical research studies, such as those that randomize clusters of subjects or that randomize individual subjects but collect repeated measures of the response of interest. We denote the outcome for the i^{th} independent randomized unit, $i = 1, ..., m$, in such studies is denoted by the n_i-dimensional response vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})^{\mathrm{T}}$, which may represent longitudinal measurements from a single subject or a set of responses from subjects within a cluster, defined by a family, hospital, or class. Considering the substantial costs incurred by such studies, it is of interest to maximize efficiency in the estimation of treatment effects by using all available data.

In general, studies collect data on i.i.d. observations $O_i = (\mathbf{Y}_i, A_i, \mathbf{X}_i)$, where A_i denotes a scalar treatment assignment to 1 of K possible treatments, and $\mathbf{X}_i$ is a matrix of baseline covariates. Throughout we allow n_i to be fixed or random and assume ignorability when n_i is random. Longitudinal data also include a time variable $t_i = (t_{i1}, t_{i2}, ..., t_{in_i})^{\mathrm{T}}$ denoting time points at which outcomes are measured. As in the case of unit size n_i, we allow t_i to be either fixed or random but ignorable. When repeated measures are taken on the same subject, baseline covariates are measured

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

at $t_{ij} = 0$; thus $X_{ij} = X_i$ for all $j = 1, 2, ..., n_i$, resulting in a single level of baseline covariate information. Clustered data, however, may include pre-treatment covariates at the level of the group or the individual, creating two layers of auxiliary data. In the longitudinal context, we refer to the vector $\mathbf{Y}_i$ as the subject, or independent unit and Y_{ij} as observation- or measurement-level data. For clustered data, we refer to $\mathbf{Y}_i$ as cluster-level and Y_{ij} as individual-level observations.

Semiparametric estimation often involves specifying a restricted mean model. When estimating marginal treatment effects, a model for the expected outcomes given treatment assignment is usually assumed. Consequently, only data on the treatment and outcome are used in estimation. For example, in longitudinal studies, the marginal effect of treatment over time may be measured by assuming the restricted mean model

$$E(Y_{ij}|A_i, t_{ij}) = g\{\beta_0 + \beta_A A_i + \beta_t^{\mathrm{T}} f_1(t_{ij}) + \beta_{A,t}^{\mathrm{T}} A_i f_1(t_{ij})\}, \tag{1}$$

where $f_1(t_{ij})$ is a function of t_i. The main effect β_A, which measures imbalance in $E(Y_{ij}|A_i, t_{ij})$ at baseline, is expected to be zero when randomization successfully balances covariate profiles across treatment arms. The post-baseline effect of treatment is measured by $\beta_{A,t}$. Parameters β_t and $\beta_{A,t}$ may be vector-valued, as the function describing the effect of time on expected outcomes may be of some polynomial form. Similarly, for clustered data, the semiparametric model

$$E(Y_{ij}|A_i) = g(\beta_0 + \beta_1 A_i) \tag{2}$$

may be assumed, with treatment effects determined by inference on β_1.

Estimating equations are determined by geometric arguments that distinguish parameters of interest, such as the treatment-outcome association (β) in the context of randomized studies, from other components needed to fully specify the data-generating distribution, which are represented by η. For parameters of interest, we aim to derive regular asymptotically linear (RAL) estimators,

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

where an asymptotically linear estimator $\hat{\beta}$ is one for which there exists a function $\psi(O_i)$ such that

$$\sqrt{n}(\hat{\beta} - \beta) = n^{-1} \sum_{i=1}^{n} \psi(O_i) + o_p(1). \tag{3}$$

Regularity conditions ensure that variance bounds are well-defined and exclude superefficient estimators that have undesirable properties under local alternatives (Newey (1990)). The function $\psi(O_i)$ is called the influence function of $\hat{\beta}$ and determines its limiting distribution. As (3) suggests, any RAL estimator may be obtained by solving an influence function equation. To derive the class of estimating functions under an assumed model $\mathcal{M}$, one first defines the nuisance scores $\partial log(\mathcal{L}_\eta)/\partial\eta$ for the data-generating distribution $\mathcal{L}_\eta$; one then determines the subspace defined by the closed linear span of all scores of smooth parametric submodels $\mathcal{L}_\eta$ in model $\mathcal{M}$. This nuisance tangent space is denoted by Λ_{nuis} (Bickel et al. (1993)). The orthogonal complement of Λ_{nuis}, $\Lambda_{nuis}^{\perp}$ defines the set $\{\psi(h) : h\}$, indexed by h, which contains the set of influence functions of all regular asymptotically linear estimators (Bickel et al. (1993); van der Vaart (1998)). For correlated outcomes, the geometric arguments of semiparametric theory may be viewed as a generalization of the quasilikelihood approach of Liang and Zeger (1986) in deriving generalized estimating equations (GEE). We denote as $\mathcal{M}_1$ the set of distributions of $W_i = (\mathbf{Y}_i, A_i)$ with known treatment process satisfying (1). Under model $\mathcal{M}_1$, $\Lambda_{nuis_1}^{\perp}$ is shown to be

$$\sum_{i=1}^{m} \psi(W_i; h, \beta) = \sum_{i=1}^{m} h(A_i, t_i)\{\mathbf{Y}_i - \mathbf{g}(A_i, t_i; \beta)\} = \mathbf{0}, \tag{4}$$

for estimating the $p-$dimensional vector β. The index or weight $h(A_i, t_i)$ is a $p \times n_i$ function of a random treatment variable A_i and time t_i, and $\mathbf{g}(A_i, t_i; \beta) = \{g(A_i, t_0; \beta), g(A_i, t_1; \beta), ..., g(A_i, t_{n_i}; \beta)\}^{\mathrm{T}}$. We use bold $\mathbf{g}(A_i, t_i; \beta)$ to denote the vector-valued mean function and $g(A_i, t_{ij}; \beta)$ to represent its scalar components.

A locally efficient estimator of a semiparametric model is defined as an estimator that achieves the semiparametric efficiency bound (minimum asymptotic variance among all RAL estimators) at

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

a given submodel for the data-generating law, but remains consistent outside of the data-generating submodel (Bickel et al. (1993)). A locally efficient estimator is determined by finding the optimal estimating function, referred to as the efficient score, which for GEE requires finding the optimal $h(\cdot)$. When no baseline covariates are observed, Chamberlain (1986) showed that the efficient score of β, is obtained by setting $h(A_i, t_i) = \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}$, where $\mathbf{V}_i$ is the $n_i \times n_i$ variance-covariance matrix of $\mathbf{Y}_i$, and $\mathbf{D}_i = \frac{\partial \mathbf{g}(A_i,t_i;\beta)}{\partial \beta}$. The estimator remains consistent, however, when a working covariance other than the true covariance is substituted into the estimating equations, thereby demonstrating that consistency is achievable outside of the data-generating law.

For model $\mathcal{M}_2$, defined by observations O_i, marginal model (1), and known treatment process, the set of influence functions was derived by Robins et al. (1994) and arises by augmenting the influence functions of β under model $\mathcal{M}_1$. Augmented estimators are constructed by subtracting the orthogonal projection of the standard estimating function onto the span of the scores of the treatment mechanism from the standard estimating function (Robins et al. (1994), Robins (1999)). For correlated outcomes, $\Lambda^{\perp}_{nuis_2} = \{\psi(O_i, h, \gamma, \beta) : h, \gamma\}$, and augmented GEE are

$$\sum_{i=1}^{m} \psi_i(O_i; \beta, h, \gamma) = \sum_{i=1}^{m} \left[\overbrace{h(A_i, t)\{\mathbf{Y}_i - \mathbf{g}(A_i, t_i; \beta)\}}^{Standard\ GEE} - \overbrace{\sum_{a=0}^{K-1} \{I(A_i = a) - \pi_a\}\gamma_a(\mathbf{X}_i)}^{arbitrary\ score\ of\ [A|X]} \right] = \mathbf{0}, \qquad (5)$$

where for K-level treatment A_i, $P(A_i = a) = \pi_a$. Fixing $h(A_i, t_i)$, the most efficient estimating function sets $\gamma_k(\mathbf{X}_i) = \gamma_{k_{opt}}(\mathbf{X}_i) = h(k, t)\{E(Y|A_i = a, \mathbf{X}_i, t) - \mathbf{g}(k, t; \beta)\}$ (Robins et al. (1994), Robins (2000); van der Laan and Robins (2003); Zhang et al. (2008)). The augmentation therefore involves estimation of the conditional mean outcome regression model $E(\mathbf{Y}_i|\mathbf{X}_i, A_i)$. When baseline covariates are predictive of the outcome augmentation reduces variability in estimated treatment effects, irrespective of the outcome distribution. For the longitudinal marginal model (1), if outcomes Y_{ij} are restricted to post-baseline measurements, the baseline measurement Y_{i0} may be utilized as a baseline covariate and included in $\mathbf{X}_i$. The interpretation of model parameters then changes,

Hosted by The Berkeley Electronic Press

with the effect of treatment over time evaluated through β_A and $\beta_{A,t}$. In contrast to the previous interpretation, β_A now measures a constant shift in $g^{-1}\{E(Y_{ij}|A_{ij}, t_{ij})\}$ due to treatment, while nonzero $\beta_{A,t}$ indicates a change in the impact of treatment on $g^{-1}\{E(Y_{ij}|A_{ij}, t_{ij})\}$ over time.

Locally efficient estimators of parameters in restricted mean models of marginal treatment effects have been implemented for univariate data in the presence of baseline covariates by Robins (2000); Bang and Robins (2005); van der Laan and Rubin (2006); Tsiatis et al. (2008); Zhang et al. (2008); Moore and van der Laan (2009b) and Moore and van der Laan (2009a). In these developments, the choice of $h(\cdot)$ has no impact on the resulting asymptotic variance and is therefore not considered for deriving efficient estimators. For a univariate outcome, the model $g_s(A_i; \beta)$ defined by a unique parameter for each treatment level is saturated and the choice of $h(\cdot)$ is inconsequential. When $\mathbf{Y}_i$ is multivariate, $g_s(A_i; \beta)$ is not saturated because a single parameter β is shared across components of the vector $\mathbf{g}_s(A_i; \beta)$. As a result, $\Lambda^{\perp}_{nuis_2}$ provides a larger set of estimating functions than $\Lambda^{\perp}_{nuis_1}$, where each element in $\Lambda^{\perp}_{nuis_2}$ is indexed by $h(\cdot)$. The choice of $h(\cdot)$ impacts efficiency, and the optimal $h(\cdot)$ must be found to achieve minimum variance.

The efficient score in model $\mathcal{M}_2$ does not generally have the same optimal index $h(A_i, t_i)$ as the efficient score in model $\mathcal{M}_1$. When incorporating auxiliary covariates in the estimation of marginal treatment effects via augmented GEE, the choice $h(A_i, t_i) = \mathbf{D}_i^{\mathrm{T}}\mathbf{V}_i^{-1}$, while resulting in a consistent estimator, is therefore no longer optimal in model $\mathcal{M}_2$. The efficient score is determined by optimizing over all $p \times n_i$ index functions $h(A_i, t_i)$ (Robins et al. (1994); Robins (1999); van der Laan and Robins (2003)). Robins (1999) established general theory for deriving the efficient score of treatment effects in marginal structural models (MSMs) of time-dependent exposures, including the case of multivariate outcomes. Application of the Robins (1999) theory to establish locally efficient estimators in specific settings, such as for randomized trials with correlated, requires further derivation. Additionally, the locally efficient estimators of Robins (1999) were not implemented nor evaluated for practical use. Models (1) and (2) may be viewed as examples of MSMs for a point exposure; the Robins (1999) theory therefore equally applies. Although the efficient score

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

may be obtained theoretically, it is often computationally intensive to calculate. Consequently, inefficient estimators are typically used. The suboptimal estimator based on augmenting GEE with $h(A_i, t_i) = \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1}$ was shown to improve efficiency by Stephens et al. (2011). In subsequent text, we refer to unaugmented GEE (4) under model $\mathcal{M}_1$ with the index function $h(A_i, t_i) = \mathbf{D}_i^{\mathrm{T}} \mathbf{V}_i^{-1}$ as standard GEE, and the suboptimal estimator obtained by augmenting standard GEE is referred to as simple augmented GEE. Here we show how to further improve on simple augmented GEE by deriving the corresponding semiparametric locally efficient estimator for model $\mathcal{M}_2$. We then evaluate the feasibility of achieving such improvement in practice.

The following section presents the efficient score and derives a locally efficient estimator of marginal treatment effects in randomized trials with correlated outcomes when auxiliary data are available as in model $\mathcal{M}_2$. We also discuss an implementation procedure detailing how to appropriately estimate each component of the efficient score. In Sections 3 and 4 we compare the derived semiparametric locally efficient estimator to standard and simple augmented GEE through simulations and application to the AIDS Clinical Trial Group study 398, a randomized longitudinal HIV intervention trial.

2 Methods

2.1 The Efficient Score

We consider the setting of longitudinal data and note that results follow analogously for clustered data by omitting t_i. Before presenting the main result, some additional notation is required. Conditioning on t_i, the matrix $h(A_i, t_i)$ takes K possible values, which may be denoted by K $p \times n_i$ constant matrices $h_0(t_i),\ h_1(t_i), ...,\ h_{K-1}(t_i)$. For binary treatment, we have $\mathbf{h}_1 = h_1(t_i)$ and $\mathbf{h}_0 = h_0(t_i)$, which denote the index functions under treatment ($A = 1$) and control ($A = 0$), respectively. Let $\Delta_{a_i}(X) = E(\mathbf{Y}_i | A_i = a, \mathbf{X}_i, t_i) - \mathbf{g}(a, t_i; \beta)$, the n_i-dimensional vector of the difference in the conditional and marginal mean outcomes given time. Using this construction, let $\mathbf{h} = [\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{K-1}]$,

Hosted by The Berkeley Electronic Press

the complete index matrix of dimension $p \times Kn_i$. Using a result from Newey and McFadden (1994), we show in the supplementary material that the optimal index $h_{opt}(A,t)$ and resulting efficient score may be determined by solving a generalized information equality. Here we present our main result:

Proposition 1. *The efficient score for model $\mathcal{M}_2$ is*

$$\mathbf{h_{opt}} = \left[\pi_0 \frac{\partial \mathbf{g}(0,t;\beta)}{\partial \beta^{\mathrm{T}}}, \pi_1 \frac{\partial \mathbf{g}(1,t;\beta)}{\partial \beta^{\mathrm{T}}}, ..., \pi_{K-1} \frac{\partial \mathbf{g}(K-1,t;\beta)}{\partial \beta^{\mathrm{T}}}\right]^{\mathrm{T}} \mathbf{C}^{-1}, \tag{6}$$

$\mathbf{C} = \mathbf{C}_1 - \mathbf{C}_2$, where

$$\mathbf{C}_1 = \begin{bmatrix} \pi_0 V(\mathbf{Y}|A=0) & 0 & \cdots & 0 \\ 0 & \pi_1 V(\mathbf{Y}|A=1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_{K-1} V(\mathbf{Y}|A=K-1) \end{bmatrix},$$

and

$$\mathbf{C_2} = \begin{bmatrix} \pi_0(1-\pi_0) E_X\left[\Delta_0(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X})\right] & \cdots & -\pi_0\pi_{K-1} E_X\left[\Delta_0(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X})\right] \\ -\pi_1\pi_0 E_X\left[\Delta_1(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X})\right] & \ddots & -\pi_1\pi_{K-1} E_X\left[\Delta_1(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X})\right] \\ \vdots & \ddots & \vdots \\ -\pi_{K-1}\pi_0 E_X\left[\Delta_{K-1}(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X})\right] & \cdots & \pi_{K-1}(1-\pi_{K-1}) E_X\left[\Delta_{K-1}(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X})\right] \end{bmatrix}.$$

As shown above, $\mathbf{C}$ is of dimension $Kn_i \times Kn_i$ and may be decomposed into the difference $\mathbf{C} = \mathbf{C}_1 - \mathbf{C}_2$, where $\mathbf{C}_1$ is a block diagonal matrix with diagonal components $\pi_a V(Y|A=a,t)$. The block diagonal of $\mathbf{C}_2$ contains the matrices $\pi_a(1-\pi_a)E_X\left[\Delta_a(X)\Delta_a^{\mathrm{T}}(X)\right]$, and off-diagonal block components are determined by $-\pi_a\pi_{a'}E_X\left[\Delta_a(X)\Delta_{a'}^{\mathrm{T}}(X)\right]$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

When treatment is binary, $\mathbf{C}$ simplifies to

$$\mathbf{C} = \begin{bmatrix} \pi_1 V(Y|A=1,t) - \pi_1\pi_0 E_X\left[\Delta_1(X)\Delta_1^{\mathrm{T}}(X)\right] & \pi_1\pi_0 E_X\left[\Delta_1(X)\Delta_0^{\mathrm{T}}(X)\right] \\ \pi_1\pi_0 E_X\left[\Delta_0(X)\Delta_1^{\mathrm{T}}(X)\right] & \pi_0 V(Y|A=0,t) - \pi_1\pi_0 E_X\left[\Delta_0(X)\Delta_0^{\mathrm{T}}(X)\right] \end{bmatrix},$$

where $\pi_0 = 1 - \pi_1$. Inverting $\mathbf{C}$ analytically and letting $\zeta^{a,a'} = E_X\left[\Delta_a(X)\Delta_{a'}^{\mathrm{T}}(X)\right]$,

$$h_{opt}(A) = \left\{D^{\mathrm{T}}(A) - \pi_1^{1-A}(1-\pi_1)^A D^{\mathrm{T}}(1-A)\left[V(1-A) - \pi_1^A(1-\pi_1)^{1-A}\zeta^{1-A,1-A}\right]^{\mathrm{T}^{-1}}\zeta^{1-A,A}\right\} \times \left\{V(A) - \pi_1^{1-A}(1-\pi_1)^A\left(\zeta^{A,A} + \zeta^{A,1-A}\left[\frac{V(1-A)}{\pi_1^A(1-\pi_1)^{1-A}} - \zeta^{1-A,1-A}\right]^{-1}\zeta^{A,1-A^{\mathrm{T}}}\right)\right\}^{-1}. \quad (7)$$

Expressing the optimal index as in (7) demonstrates that $\mathbf{h_{opt}}$ incorporates information on the treatment assignment and auxiliary covariates $\mathbf{X}$ through $\zeta^{a,a'}$, while the standard index $h_{std} = D^{\mathrm{T}}(A)V(A)^{-1}$, does not. The matrix $\zeta^{a,a'}$ is by definition the covariance of $E(\mathbf{Y}_i|\mathbf{X}_i, a, t_i)$ and $E(\mathbf{Y}_i|\mathbf{X}_i, a', t_i)$, the expected outcomes given baseline covariates and treatment assignment to a and a', respectively. The optimal index $\mathbf{h_{opt}}$ therefore boosts efficiency by incorporating information on the covariance in expected outcomes when weighting the residuals $\mathbf{Y}_i - \mathbf{g}(A_i;\beta)$ in the marginal model estimating equations. To implement locally efficient GEE for model $\mathcal{M}_2$, estimates of $V(\mathbf{Y}_i|A_i, t_i)$, $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$, and $\zeta^{a,a'}$ for all unique pairs of treatment levels $\{a, a'\}$, including $a = a'$, are needed. The next section details an estimation procedure for each component of $\mathbf{h_{opt}}$ when $\mathbf{Y}_i$ is continuous and $g(\cdot)$ is the identity link, or $\mathbf{Y}_i$ is binary and $g(\cdot)$ is the inverse logit link.

2.2 Estimation of $\mathbf{h_{opt}}$

The semiparametric locally efficient estimator requires estimates of 3 additional parameters:

1. $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$
2. $\zeta^{a,a'} = Cov\{E(\mathbf{Y}_i|\mathbf{X}_i, A_i = a, t_i), E(\mathbf{Y}_i|\mathbf{X}_i, A_i = a', t_i)|A_i, t_i\}$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

3. $V(\mathbf{Y}_i|A_i, t_i)$.

These quantities may be linked by the law of total variance, $V(\mathbf{Y}_i|A_i, t_i) = E[V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i)|A_i, t_i] + V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]|A_i, t_i)$. For the i^{th} independent unit, the n_i-dimensional vector $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$ determines the $n_i \times n_i$ matrix $V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]|A_i, t_i)$ and ultimately impacts the form of the marginal variance matrix $V(\mathbf{Y}_i|A_i, t_i)$. Observing the relationship among each of these parameters provides guidance for estimation. For example, the working marginal covariance selected must be compatible with the working model chosen for $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$. More generally, the models for each component of $\mathbf{h_{opt}}$ must be specified so that the model selected for one component does not preclude the choice of model chosen for another. One approach that ensures compatibility is to start by estimating $E(Y_{ij}|X_{ij}, A_i, t_{ij})$ through an appropriate regression technique to provide an estimate $\hat{E}(\mathbf{Y}_i|\mathbf{X}_i, A_i = a, t_i)$. The conditional mean outcome may be modeled by

$$E[Y_{ij}|X_{ij}, A_i, t_{ij}] = g\{\eta_0 + \eta_A A_i + \eta_t^{\mathrm{T}} f(t_{ij}) + \eta_{A,t}^{\mathrm{T}} A_i f(t_{ij}) + \eta_X^{\mathrm{T}} X_{ij} + \eta_{X,t}^{\mathrm{T}} X_{ij} f(t_{ij}) + \eta_{A,X}^{\mathrm{T}} A_i X_{ij}\}, \quad (8)$$

where X_{ij} represents the collection of covariates for the j^{th} measurement in the i^{th} unit. The next step is to estimate the conditional expectation by noting how the model of $E(Y_{ij}|X_{ij}, A_i = a, t_i)$ impacts the form of the matrix $\zeta^{a,a'}$. The final step is estimation of $V(\mathbf{Y}_i|A_i, t_i)$ by summing the estimates of $E[V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i)|A_i, t_i]$ and $V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]|A_i, t_i)$.

2.2.1 General estimation of $\zeta^{a,a'}$ and $V(Y|A)$

Since $\zeta^{a,a'}$ is a covariance matrix, it may generally be estimated in a similar fashion to estimating the correlation parameters in standard GEE. Let $\zeta^{a,a'} = \mathbf{R}^{1/2}\mathbf{S}\mathbf{R}^{1/2}$, where $\mathbf{R}$ is a $n_i \times n_i$ diagonal matrix with the j^{th} diagonal component $R_{j,j} = Cov(E[Y_{ij}|X_{ij}, A_i = a, t_{ij}], E[Y_{ij}|X_{ij}, A_i = a', t_{ij}]|A_i, t_{ij}) = \nu_j^{a,a'}$, the covariance of the predicted outcomes of element j under treatments a and a', and $\mathbf{S}$ is a $n_i \times n_i$ correlation matrix with $S_{j,j} = 1$ and $S_{j,j'} = f(\tau^{a,a'})$, denoting the correlation in the predicted outcomes of element j under treatment a and element j' under treatment a'. Parameters $\tau^{a,a'}$, which

http://biostats.bepress.com/harvardbiostat/paper151

may be a vector, and $\nu^{a,a'} = (\nu_1^{a,a'}, \nu_2^{a,a'}, ..., \nu_{n_i}^{a,a'})$ characterize the covariance in conditional mean outcomes under treatments a and a'. Letting $\hat{\Delta}_{a_{ij}} = \hat{E}(Y_{ij}|X_{ij}, A_i = a, t_{ij}) - g(a, t_{ij}; \hat{\beta}_{init})$, where $\hat{\beta}_{init}$ is an initial estimate of β, $\nu_j^{a,a'}$ may be estimated by

$$\hat{\nu}_j^{a,a'} = \frac{1}{m - p_\eta} \sum_{i=1}^{m} \hat{\Delta}_{a_{ij}} \hat{\Delta}_{a'_{ij}},$$

where p_η is the dimension of the outcome regression parameter η. The correlation parameter $\tau^{a,a'}$ is then estimated by the moment equations

$$\sum_{i=1}^{m} \sum_{j<j'} \left\{ \frac{\hat{\Delta}_{a_{ij}}}{\sqrt{\hat{\nu}_j^{a,a}}} \frac{\hat{\Delta}_{a'_{ij'}}}{\sqrt{\hat{\nu}_{j'}^{a',a'}}} - f(\tau^{a,a'}) \right\} = 0.$$

For $a = a'$, we obtain an estimate of $\zeta^{a,a} = V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i = a, t_i]|A_i, t_i)$.

As an alternative approach, one may also derive an expression of $\zeta_{j,j'}^{a,a'}$, the j, j' element of $\zeta^{a,a'}$, that depends on $\eta = (\eta_0, \eta_A, \eta_t^{\mathrm{T}}, \eta_{A,t}^{\mathrm{T}}, \eta_X^{\mathrm{T}}, \eta_{X,t}^{\mathrm{T}}, \eta_{A,X}^{\mathrm{T}})^{\mathrm{T}}$ and the covariance in baseline covariates. An empirical estimate of $Cov(\mathbf{X}_i)$ may then be substituted into this expression.

After estimating $\zeta^{a,a}$, the conditional variance of $\mathbf{Y}_i$, $V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i)$, may be estimated using the correlation parameters from GEE based on the conditional mean model (8). Under homoscedasticity $V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i) = \lambda$ for all i. To ensure compatibility of all parameters, the marginal variance $V(\mathbf{Y}_i|A_i, t_i)$ is then estimated by $\hat{V}(\mathbf{Y}_i|A_i, t_i) = \hat{\zeta}^{a,a} + \hat{\lambda}$, where $\hat{\zeta}^{a,a}$ and $\hat{\lambda}$ are estimates of $\zeta^{a,a}$ and λ, respectively.

2.2.2 Estimation of $\zeta^{a,a'}$ for clustered data or longitudinal data with $\eta_{X,t} = \mathbf{0}$

For clustered data and longitudinal data with $\eta_{X,t} = \mathbf{0}$ in (8), calculating $\zeta^{a,a'}$ is straightforward. When data are clustered, $\eta_t = \eta_{A,t} = \eta_{X,t} = \mathbf{0}$, leaving $E[Y_{ij}|X_{ij}, A_i] = g(\eta_0 + \eta_A A_i + \eta_X^{\mathrm{T}} X_{ij} + \eta_{A,X}^{\mathrm{T}} A_i X_{ij})$. In this setting, $\zeta_{j,j'}^{a,a'}$ is calculated as $\zeta_{j,j'}^{a,a'} = Cov_X\{g(\eta_0 + \eta_A a + \eta_X^{\mathrm{T}} X_{ij} + \eta_{A,X}^{\mathrm{T}} a X_{ij}),\ g(\eta_0 + \eta_A a' + \eta_X^{\mathrm{T}} X_{ij'} + \eta_{A,X}^{\mathrm{T}} a' X_{ij'})\}$. If auxiliary covariates $X_{ij}, X_{ij'}$ are equally correlated among subjects

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

within a cluster $\zeta_{j,j'}^{a,a'} = \rho^{a,a'}$ for all j, j'. This holds for all link functions $g(\cdot)$. For longitudinal data when $\eta_{X,t} = \mathbf{0}$ (i.e. the effects of baseline covariates on the conditional mean outcome do not vary over time) $\zeta_{j,j'}^{a,a'} = Cov_X\{g(\eta_0 + \eta_A a + \eta_t^{\mathrm{T}} f(t_{ij}) + \eta_{A,t}^{\mathrm{T}} a f(t_{ij}) + \eta_X^{\mathrm{T}} X_i + \eta_{A,X}^{\mathrm{T}} a X_i),\ g(\eta_0 + \eta_A a' + \eta_t^{\mathrm{T}} t_{ij'} + \eta_{A,t}^{\mathrm{T}} a' f(t_{ij'}) + \eta_X^{\mathrm{T}} X_i + \eta_{A,X}^{\mathrm{T}} a' X_i\}$. If $g(\cdot)$ is the identity link, this reduces to $\zeta_{j,j'}^{a,a'} = Cov(\eta_X^{\mathrm{T}} X_i + \eta_{A,X}^{\mathrm{T}} a X_i,\ \eta_X^{\mathrm{T}} X_i + \eta_{A,X}^{\mathrm{T}} a' X_i) = \rho^{a,a'}$ for all j, j'. For clustered data $\rho^{a,a'}$ is a constant that depends on $Cov(X_{ij}, X_{ij'})$, η_X^{T} and $\eta_{A,X}$, whereas for longitudinal data, $Cov(X_{ij}, X_{ij'})$ is replaced by $Var(X_i)$ since $X_{ij} = X_i$ for all j.

2.2.3 Estimation of $V(Y_i|A_i = a)$ under a compatible standard form

In some special cases where summing $E[V(\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i)|A_i, t_i]$ and $V(E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]|A_i, t_i)$ results in a marginal covariance matrix $V(\mathbf{Y}_i|A_i, t_i)$ with a standard form, e.g., exchangeable, $V(\mathbf{Y}_i|A_i, t_i)$ may be estimated directly while maintaining compatibility with $E[\mathbf{Y}_i|\mathbf{X}_i, A_i, t_i]$ and $\zeta^{a,a}$. As stated above, if individual-level covariates X_{ij} are equally correlated among subjects within the i^{th} cluster, the model $E[Y_{ij}|X_{ij}, A_i = a]$ imposes compound symmetry on $\zeta^{a,a'}$, where diagonal components depend on $Var(X_{ij})$ and off-diagonal components are determined by $Cov(X_{ij}, X_{ij'})$. If the conditional variance $V(\mathbf{Y}_i|\mathbf{X}_i, A_i)$ is also exchangeable, $V(\mathbf{Y}_i|A, t_i)$ is the sum of two exchangeable matrices and therefore also has an exchangeable structure. The optimal index $\mathbf{h_{opt}}$ may then be calculated by estimating $V(\mathbf{Y}_i|A_i, t_i)$ directly as in standard or simple augmented GEE and using the above procedure to estimate $\zeta^{a,a'}$.

A consistent estimator of the asymptotic variance of $\hat{\beta}_{opt}$, the solution to the augmented estimating equations (5) evaluated under (6), may be calculated using the sandwich variance formula of Huber (1964).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

3 Simulation Study

Semiparametric locally efficient GEE for model $\mathcal{M}_2$ were compared to standard and simple augmented GEE through a simulation study. Simulations were completed for clustered data with continuous and binary outcomes and longitudinal data with continuous outcomes. Results are based on 1,000 Monte Carlo datasets.

3.1 Continuous outcomes

3.1.1 Clustered Data

Data for $m = 500$ clusters were generated, with n_i=2,4,6,8,10,12 with equal probability for the first set of simulations and n_i=10,20,30,40,50 in the second set. Auxiliary covariates X_{ij1}, X_{ij2}, and X_{ij3} were each generated from a multivariate normal distribution with $Var(X_{ij1})$=2, $Var(X_{ij2})$=6, and $Var(X_{ij3})$=5. Correlation was induced among individual-level covariates within the same cluster by setting $Cov(X_{ij1}, X_{ij'1}) = \varsigma_{X_1}$, $Cov(X_{ij2}, X_{ij'2}) = \varsigma_{X_2}$, and $Cov(X_{ij3}, X_{ij'3})$=1. Covariance terms ς_{X_1} and ς_{X_2} were varied from 0.5 to 2 and 1.5 to 6, respectively, to evaluate the effect of auxiliary covariate correlation on the performance of locally efficient augmented GEE. At ς_{X_1}=0.5 and ς_{X_2}=1.5 covariates were weakly correlated among individuals in the same cluster, while at ς_{X_1}=5 and $\varsigma_{X_2,4}$=6, covariates were perfectly correlated, thereby becoming cluster-level. The exact values considered for ς_{X_1} and ς_{X_2} were (0.5, 1, 1.5, 2) and (1.5, 3, 4.5, 6), for simulation sets 1-4 at each set of cluster sizes. Within the j^{th} individual in the i^{th} cluster, auxiliary covariates were independent. The treatment variable A_i was drawn from the Bernoulli distribution with p=1/2. Clustered responses were generated from the following model, with individual-level error terms $\varepsilon_{ij} \sim N(0, 40)$ and cluster-level effects $b_i \sim N(0, \sigma_b^2)$: $Y_{ij}|A_i, X_{ij}, b_i = 1.0 + 1.1X_{ij1}^2 + 0.9X_{ij2} + 0.5A_i + b_i + \varepsilon_{ij}$. The proportion of variability in Y_{ij} explained by auxiliary covariates X_{ij} was held fixed at roughly 25%. Simulations were completed with $\sigma_b^2 = 0$ and $\sigma_b^2 = 6$, representing the case in which covariates account for all between-cluster heterogeneity ($V(Y|X, A)$ independent) and the alternative of some

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

intracluster correlation caused by an unmeasured variable ($V(Y|X,A)$ exchangeable), respectively.

For each dataset, the marginal effect of treatment was estimated by fitting model (2) through standard, simple augmented, and locally efficient GEE for $\mathcal{M}_2$. The impact of misspecification on the locally efficient estimator and its efficiency relative to simple augmented and standard GEE was evaluated by fitting various models to estimate $E(\mathbf{Y}|\mathbf{X},A)$. The correct model for $E(\mathbf{Y}|\mathbf{X},A)$, denoted by 'C' in tables and figures, was $E(Y_{ij}|X_{ij},A_i) = \eta_0 + \eta_1 X_{ij1}^2 + \eta_2 X_{ij2} + \eta_3 A_i$, and two incorrect models were Wrong 1, 'W1'=$E(Y_{ij}|X_{ij},A_i) = \eta_0 + \eta_1 X_{ij1} + \eta_2 X_{ij2} + \eta_3 A_i$ and Wrong 2, 'W2'=$E(Y_{ij}|X_{ij},A_i) = \eta_0 + \eta_1 X_{ij1}^2 + \eta_2 X_{ij2} + \eta_3 X_{ij3} + \eta_4 A_i$. 'Wrong 1' evaluated the impact of misspecifying the functional form of X_{ij1}, while 'Wrong 2' examined the effect of adding noise to the outcome regression model. All working covariance matrices were fit under exchangeable structure.

Efficiency comparisons relative to standard GEE are summarized in Figures 1a-1b, while the Monte Carlo Relative Efficiency (MCRE) of the locally efficient estimator for $\mathcal{M}_2$ to simple augmented GEE may be found in Table 1. Small cluster figures are included in the supplementary material. Across all levels of correlation, augmented estimators resulted in increased efficiency compared to the unaugmented estimator (MCRE 1.25-11.6). For low correlation among X_{ij} simple augmented and locally efficient augmented estimators performed similarly. Simple augmented GEE and locally efficient GEE for $\mathcal{M}_2$ also resulted in similar efficiency when the conditional mean model did not include the data-generating quadratic term X_{ij1}^2 or the true conditional variance was exchangeable (MCRE locally efficient to simple augmented GEE 0.99-1.01). When correlation was increased among X_{ij} within a cluster, the assumed conditional mean model included all important covariates in the correct functional form, and baseline covariates accounted for all within-subject correlation, locally efficient GEE for $\mathcal{M}_2$ gained in efficiency over the simple augmented GEE (MCRE locally efficient to simple augmented GEE 1.04-1.22). Increased covariance among auxiliary covariates also resulted in greater efficiency gains for any augmented GEE relative to the standard estimator. Trends were more pronounced for large average cluster size (average n_i=30 vs. average n_i=7).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

3.1.2 Longitudinal Responses

For each Monte Carlo dataset, m=500 longitudinal response vectors $\mathbf{Y}_i$ were generated from the model $Y_{ij} = 1.5 + 1.1X_{i1}^2 + 0.9X_{i2} + 1.0t_{ij} + 1.0A_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, and $Cov(\varepsilon_{ij}, \varepsilon_{ij'})$ had an AR-1 structure with correlation parameter $\alpha = 0.1, 0.3$, or 0.5 for different sets of simulations. Covariates X_{i1} and X_{i2} were normally distributed with mean 0 and variance σ_{X1}^2 and σ_{X2}^2, respectively. Variance parameters σ_ε^2, σ_{X1}^2, and σ_{X2}^2 were varied so that baseline covariates accounted for 10-60% of the variability in $\mathbf{Y}|A$ in increments of 10%. Subjects were randomly assigned to treatment (A_i=1) with probability 1/2. For each subject $t_i = (t_{i1} = 1, t_{i2} = 2, ..., t_{in_i} = n_i)$, where n_i varied from 1 to 8, as might be the case in a longitudinal study with staggered entry.

Standard GEE, simple augmented GEE, and locally efficient GEE for $\mathcal{M}_2$ were applied to each Monte Carlo dataset to estimate marginal treatment effects. All GEE were fit based on the marginal mean model $E(Y_{ij}|A_i) = \beta_0 + \beta_1 A_i + \beta_2 t_{ij}$ with inferences on the treatment effect completed through β_1. Standard and simple augmented GEE were applied to each Monte Carlo dataset with AR-1, exchangeable, and true working covariance structures, with the true structure under the marginal model being a summation of AR-1 and exchangeable matrices as described in section 2. Locally efficient GEE for $\mathcal{M}_2$ were fit under the true covariance structure and a misspecified marginal AR-1 working covariance. Baseline covariates were incorporated fitting several outcome regression models. We use 'C' to denote the correct model $E(Y_{ij}|X_{ij}, A_i, t_{ij}) = \eta_0 + \eta_1 A_i + \eta_2 t_{ij} + \eta_3 X_{i1}^2 + \eta_4 X_{i2}$, which corresponds to the true data generating mechanism; 'W1' indicates the model $E(Y_{ij}|X_i, A_i, t_{ij}) = \eta_0 + \eta_1 A_i + \eta_2 t_{ij} + \eta_3 X_{i1} + \eta_4 X_{i2}$, omitting the exponent on X_{i1}; and 'W2' is the model that includes a noisy covariate X_{i3}, such that $E(Y_{ij}|X_i, A_i, t_{ij}) = \eta_0 + \eta_1 A_i + \eta_2 t_{ij} + \eta_3 X_{i1}^2 + \eta_4 X_{i2} + \eta_5 X_{i3}$.

Efficiency comparisons are summarized in Figure 2 and Table 2. Additional figures may be found in the supplementary material. For well-specified variance components and conditional mean models, the locally efficient GEE for $\mathcal{M}_2$ was more efficient than the simple augmented GEE, with the difference in efficiency increasing with the percent variability explained by $\mathbf{X}_i$ (MCRE of locally efficient to simple augmented GEE 1.0-1.27). Similarly, all augmented estimators were more

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

efficient than standard GEE, with efficiency gains from augmenting increasing with correlation in $\mathbf{Y}$ and $\mathbf{X}$ (MCRE of Augmented GEE to Standard GEE 1.36-5.28). For poorly specified conditional mean models, locally efficient GEE for $\mathcal{M}_2$ and simple augmented GEE were nearly equally efficient (MCRE of locally efficient to simple augmented GEE 0.97-1.0), but when the marginal variance was also misspecified locally efficient GEE were less efficient than simple augmented GEE (MCRE 0.88-0.99). This demonstrates that the locally efficient efficient GEE for $\mathcal{M}_2$ is a bit more sensitive to working marginal covariance misspecification than simple augmented GEE. Among the simple augmented estimators, the estimator with the incorrect marginal AR-1 working covariance resulted in the β_1 estimate with the lowest variability. This illustrates an important distinction between locally efficient and suboptimal estimating functions. Considering estimators using a suboptimal index, misspecified models for parameters in the index may result in more efficient inferences than correctly specified models. For the locally efficient estimator, asymptotic efficiency is achieved only in the absence of model misspecification for all parameters in $h_{opt}(\cdot)$.

3.2 Clustered Binary Data

As for continuous outcomes, data for m=500 clusters of variable size were generated with n_i=2,4,6,8,10,12 for small cluster settings and n_i=10,20,30,40,50 for the large cluster scenario. The binary treatment variable A_i was simulated from the Bernoulli(1/2) distribution. Individual-level covariates X_{ij1}, X_{ij2}, and X_{ij3} were each generated from a multivariate normal distribution with $\mu_{X_{ijk}} = 0$, $\sigma^2_{X_{ij1}} = \sigma^2_{X_{ij3}}$=2, $\sigma^2_{X_{ij2}} = 5$, and $Cov(X_{ijk}, X_{ij'k}) = \varsigma_{X_k}$, inducing marginal correlation among individuals within the same cluster. Covariance parameters ς_{X_k} were varied to evaluate the impact of covariance in auxiliary covariates on the performance of augmented estimators, with $\varsigma_{X_1} = \varsigma_{X_3} = 0.5, 1.0, 1.5, 2.0$ and $\varsigma_{X_3} = 1.25, 2.5, 3.75, 5.0$ for different sets of simulations. For low levels of ς_{X_k}, covariates were weakly correlated, while for $\varsigma_{X_k} = \sigma^2_{X_{ijk}}$, covariates were cluster-level. Binary outcomes were simulated from the model $logit[E(Y_{ij}|X_{ij}, A_i, b_i)] = 0.7{X_{ij1}}^2 + 0.4X_{ij2} - 0.5A_i + b_i$, where b_i was drawn from the bridge distribution for the logit link (Wang and Louis (2003)) with scale pa-

http://biostats.bepress.com/harvardbiostat/paper151

rameter θ. Simulations were completed with two values of the bridge distribution scale parameter, $\theta = 1$ and $\theta = 0.8$, representing settings in which all sources of between-cluster heterogeneity are measured through auxiliary covariates, or when unmeasured sources of between-cluster heterogeneity are present. A total of 16 sets of simulations were done, varying cluster size, correlation in $\mathbf{X}$, and θ.

Standard, simple augmented, and locally efficient GEE for $\mathcal{M}_2$ were applied to each dataset and compared for efficiency. For each estimator, the restricted mean model of interest was model (2) with $g(\cdot)$ the inverse logit link and β_1 measuring the marginal effect of treatment. Among augmented estimators, four outcome regression models were considered: 1) 'C'-Correct, $E(Y_{ij}|X_{ij}, A_i) = g(\eta_0 + \eta_1 X_{ij1}^2 + \eta_2 X_{ij2} + \eta_3 A_i)$; 2) 'W1'-Wrong 1, $E(Y_{ij}|X_{ij}, A_i) = g(\eta_0 + \eta_1 X_{ij1} + \eta_2 X_{ij2} + \eta_3 A_i)$; 3) 'W2'-Wrong 2, $E(Y_{ij}|X_{ij}, A_i) = g(\eta_0 + \eta_1 X_{ij1}^2 + \eta_2 X_{ij2} + \eta_3 X_{ij3} + \eta_4 A_i)$; and 4)'W1 OLS'-Wrong 1 OLS, $E(Y_{ij}|X_{ij}, A_i) = \eta_0 + \eta_1 X_{ij1} + \eta_2 X_{ij2} + \eta_3 X_{ij3} + \eta_4 A_i$. With the exception of model 4, which was fit through ordinary least squares (OLS), all outcome regression models were fit by logistic regression. All estimators were fit with exchangeable working covariances.

Large cluster results are shown in figures 3a-3b and Table 3, while small cluster results are included in the supplementary material. Conclusions are similar to those obtained for continuous outcomes. Efficiency improvement with augmented estimators relative to standard GEE increased with correlation in auxiliary covariates (MCRE 1.10-10.54), as did the additional efficiency gains for the locally efficient GEE for $\mathcal{M}_2$ over simple augmented GEE (MCRE 1.0-1.23). Simple and locally efficient augmented estimators were equally efficient for $\theta = 0.8$ or when conditional mean models left out important transformations, but differences in efficiency favoring the optimal estimator were observed for $\theta = 1$ and well-specified covariate-adjusted models.

3.3 Simulation Study Summary

Results from the various simulation settings provide insight into the performance of the locally efficient GEE for model $\mathcal{M}_2$ and its practical value. The locally efficient estimator theoretically

Hosted by The Berkeley Electronic Press

achieves minimum asymptotic variance when all components of $h_{opt}(\cdot)$ and the augmentation are correctly specified. The results show that achieving the efficiency bound is not robust to model misspecification for working covariances and conditional means; the locally efficient GEE for $\mathcal{M}_2$ was only more efficient than simple augmented GEE when all mean models included important covariates in the correct polynomial form, and the correct structure was specified for working covariances. Even under well-specified models, the locally efficient GEE only improved over the simple augmented GEE when the data-generating mechanism was such that the underlying conditional variance, $V(\mathbf{Y}|\mathbf{X}, A)$ had a sparse structure, such as AR1 or independence. The difficulty of correctly specifying models for nuisance parameters, particularly covariances, as well as measuring all sources of correlation so that $V(\mathbf{Y}|\mathbf{X}, A)$ is sparse present challenges for successfully implementing locally efficient estimators in real-world analysis. This challenge is further illustrated in the following section with application to AIDS Clinical Trial Group Study 398.

4 Application

The semiparametric locally efficient estimator of marginal treatment effects for correlated outcomes was applied to data from AIDS Clinical Trial Group Study 398 (ACTG 398) (Hammer et al. (2002)). ACTG 398 was a multicenter, double-blind trial, in which 481 HIV-infected patients were randomized to one of four arms, A) saquinavir, B) indinavir, C) nelfinavir, or D) placebo based on their past protease inhibitor (PI) treatment. Patients were only randomized to drugs to which they had no prior exposure. Randomized treatments were added to a common antiviral regimen for all subjects. Subjects' CD4 counts were measured at weeks 0 (baseline), 4, 8, and every 8 weeks thereafter until 48 weeks or dropout. Here, we apply the GEE estimators to compare the nelfinavir and placebo arms among patients who were eligible for both according to the stratified randomization scheme. Additional baseline covariates were age, sex, past PI use, past non-nucleoside reverse transcriptase inhibitor (NNRTI) exposure, weight, Karnofsky score, intravenous drug use,

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

and race/ethnicity. Weeks 4-32 of followup were included for analysis, with CD4 measurements at week 4 and beyond included as outcomes and week 0 CD4 included as a baseline covariate. Data were approximately 90% complete through week 32. In evaluating the effect of treatment on CD4, the best fitting marginal model was $E(Y_{ij}|A_i) = \beta_0 + \beta_1 A_i + \beta_2 t_{ij}$, where t_{ij} indicates the week of the j^{th} measurement on the i^{th} individual, and A_i was an indicator for the placebo arm. Since only follow-up measurements were modeled as outcomes and no interaction was detected between treatment and time, the effect of treatment was captured by β_1.

Standard, simple augmented, and locally efficient GEE for $\mathcal{M}_2$ were applied to estimate β_1. Several candidate outcome regression models for augmented GEE were identified through model selection procedures. Cross validation was used to select the final model, $E(Y_{ij}|A_i, X_i, t_i) = \eta_0 + \eta_1 A_i + \eta_2 t_{ij} + \eta_3 CD4_{0_i} + \eta_4 Sex_i$, where CD4_0 is baseline CD4. The QIC goodness-of-fit statistic (Pan (2001)) was compared among GEE fit to unaugmented marginal and conditional models to guide the choice of working covariance structures. To enforce compatibility of the marginal variance, conditional variance, and outcome regression in fitting locally efficient augmented GEE, the additive estimate of the marginal covariance was used. The working conditional variance was chosen by selecting the covariance structure resulting in the lowest QIC when fitting GEE on the conditional mean model. Simple augmented GEE were computed under all possible working marginal covariance structures, including the additive estimator motivated by the locally efficient GEE.

Results are shown in Table 4. Regarding covariance selection, unstructured working covariance resulted in the lowest QIC for the conditional model (supplementary material), suggesting the locally efficient estimator should be fit assuming an unstructured form of $V(\mathbf{Y}_i|\mathbf{X}_i, A_i)$. Several other covariance structures were also implemented for the locally efficient estimator to explore variance misspecification. Among simple augmented estimators, the additive marginal covariance obtained by summing the unstructured $V(\mathbf{Y}|\mathbf{X}, A, t)$ and $V(E[\mathbf{Y}|\mathbf{X}, A, t]|A, t)$ induced by the chosen conditional mean model resulted in lower variability than estimators using standard marginal covariance

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

structures. Comparing standard GEE with different working covariance models, the estimated difference in average CD4 for the placebo arm versus nelfinavir ranged from 9.9 to 20.17. The direction of the effect was reversed for estimators that incorporated baseline covariates, with average CD4 on the placebo arm 0.07 to 8.11 units lower than the nelfinavir arm. Treatment did not have a significant impact on CD4 at the 0.05 level for any of the estimators considered.

Estimators that incorporated baseline covariates greatly increased efficiency, with $SE(\hat{\beta}_1) \approx 20$ for standard GEE and $SE(\hat{\beta}_1) \approx 9$ among augmented estimators (Relative efficiency augmented to standard GEE ≈ 5.0). Simple augmented and locally efficient GEE for $\mathcal{M}_2$ resulted in similar efficiency-a result that may be explained by several factors: 1) Subjects had the same number of follow-up visits. For GEE, the index impacts efficiency most when the number of observations per unit is variable, 2) The unstructured conditional variance is not sparse, and 3) The components of $\mathbf{h_{opt}}$ may be misspecified. As a benchmark for efficiency, we also fit unaugmented GEE assuming the conditional mean model $E(Y_{ij}|A_i, X_i, t_i) = \beta_0 + \beta_1 A_i + \beta_2 t_{ij} + \beta_3 CD4_{0_i} + \beta_4 Sex_i$ with an unstructured working covariance. This estimator represents the most efficient estimator of β_1 that may be obtained using $\mathbf{X}_i$, which requires assuming that the more restrictive conditional mean model is correct. From this estimator, we can determine that for this particular case, there is little additional efficiency to be gained by locally efficient GEE if simple augmented GEE are fit under the best working covariance (Table 4).

5 Discussion

We derived and implemented a closed-form expression for the efficient score and a locally efficient estimator in model $\mathcal{M}_2$ for correlated outcomes. Through simulation, we demonstrated that the locally efficient estimator is more efficient than corresponding suboptimal estimators in certain settings, particularly when randomized units vary in size, baseline covariates account for a large portion of the within-unit correlation, and baseline covariates are at least moderately predictive of

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

the outcome. In longitudinal studies, variable size may occur when studies have staggered entry or as subjects are lost to follow-up. The estimator derived is only semiparametric locally efficient in the first case, as the locally efficient estimator for incomplete data incorporates information on the missingness process. Accounting for correlation through measured covariates and correctly specifying the form of correlation are challenges to achieving local efficiency. Such challenges may be addressed through use of scientific knowledge and covariance structure diagnostic tools, but are still likely to make local efficiency unachievable in most settings, rendering the simple augmented GEE the more useful option. Although theoretically possible, the prize of implementing local efficiency is usually not worth the chase for estimating treatment effects with baseline covariates when outcomes are correlated. Nonetheless, large efficiency gains were shown for longitudinal analysis when the baseline level of the outcome was incorporated in estimation as an auxiliary covariate. Baseline levels of outcomes can be highly predictive of followup levels, suggesting that in the analysis of data from longitudinal studies, failing to incorporate baseline covariates in analysis can be highly inefficient. These results suggest the value of incorporating baseline covariates in both interim and final analyses of data from randomized clinical trials.

Acknowledgements

This research was supported by NIH grants AIR0151164 and T32AI007358.

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–972.

Bickel, P. J., Klassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press, Baltimore.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics* **32,** 189–218.

Hammer, S. M., Vaida, F., Bennett, K., Holohan, M. K., Sheiner, L., Eron, J., Wheat, L. J., Mitsuyasu, R. T., Gulick, R. M., Valentine, F. T., Aberg, J. A., Rogers, M. D., Karol, C. N., Saah, A. J., Lewis, R. H., Bessen, L. J., Brosgart, C., De Gruttola, V., and Mellors, J. W. (2002). Dual vs. single protease inhibitor therapy following antiretroviral treatment failure. *Journal of the American Medical Association* **288,** 169–180.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35,** 73–101.

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

Moore, K. L. and van der Laan, M. J. (2009a). Application of time-to-event methods in the assessment of safety in clinical trials. In Peace, K. E., editor, *Design, Summarization, Analysis & Interpretation of Clinical Trials with Time-to-Event Endpoints*. Chapman & Hall.

Moore, K. L. and van der Laan, M. J. (2009b). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine* **28,** 39–64.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* **5,** 99–135.

Newey, W. K. and McFadden, D. (1994). Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pages 2111 – 2245. Elsevier.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57,** 120–125.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/harvardbiostat/paper151

Robins, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, pages 6–10.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In Berry, D. and Halloran, M. E., editors, *Statistical Models in Epidemiology: The Environment and Clinical Trials*, volume 116, pages 95–134. NY: Springer-Verlag.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Stephens, A. J., Tchetgen Tchetgen, E. J., and De Gruttola, V. (2011). Augmented gee for improving efficiency of inferences in cluster randomized trials by leveraging cluster and individual-level covariates. *Statistics in Medicine* In press.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons for randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine* **27,** 4658–4677.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. NY: Springer-Verlag.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2,** 1–40.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wang, A. and Louis, T. A. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika* **90,** 765–775.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

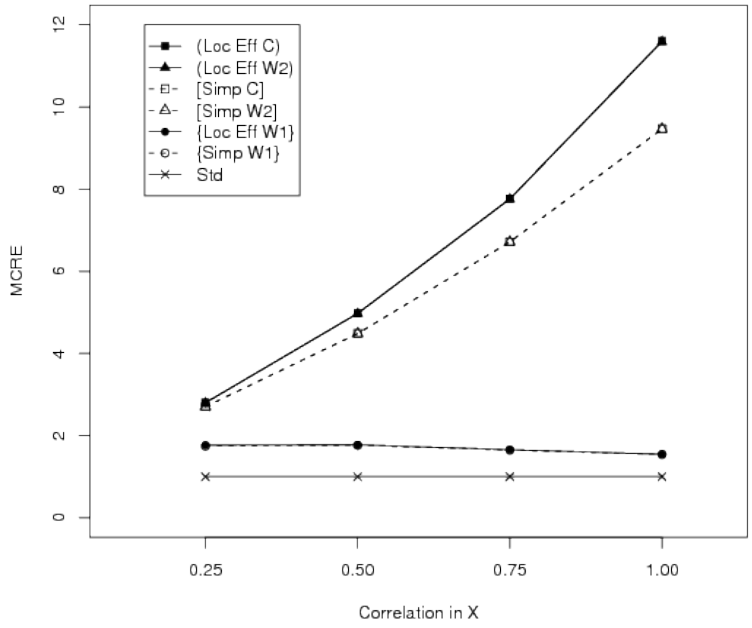
Hosted by The Berkeley Electronic Press

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in clinical randomized trials using auxiliary covariates. *Biometrics* **64,** 707–715.

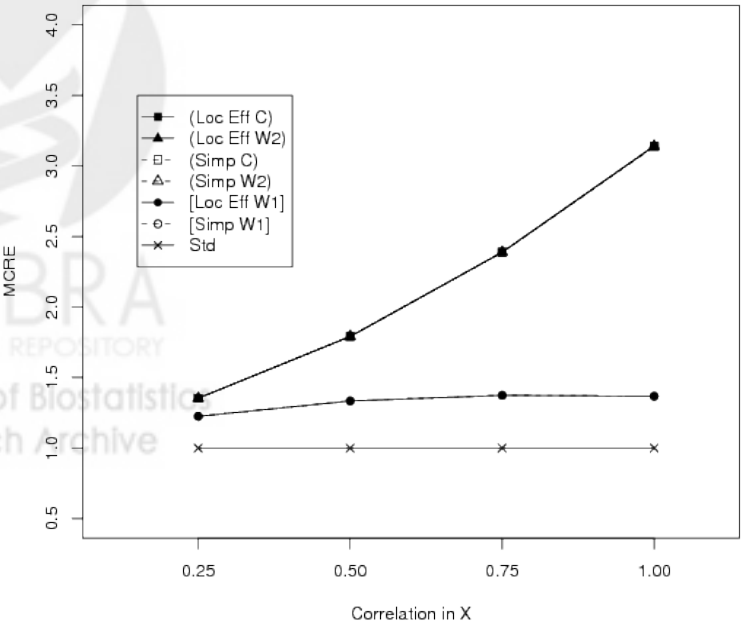
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Figure 1: MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE. Continuous clustered outcomes. Estimators corresponding to each curve are denoted by 'Estimator-Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; C-Correct, W1-Wrong 1, W2-Wrong 2. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$. The order of curves in the legend follows the order of curves on the figure, with sets of superimposed curves denoted by '()', '[]', or'{}'.

(a) n_i=(10,20,30,40,50), $\sigma_b^2 = 0$



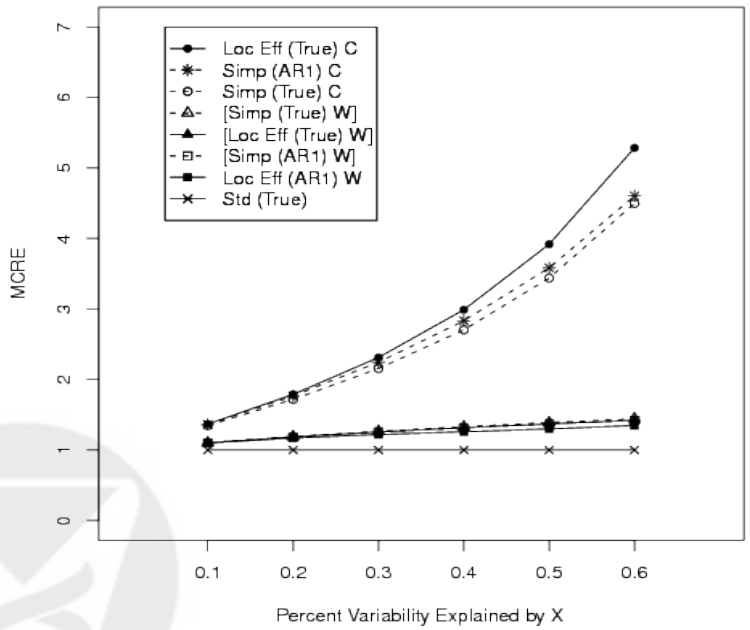
(b) n_i=(10,20,30,40,50), $\sigma_b^2 = 6$



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

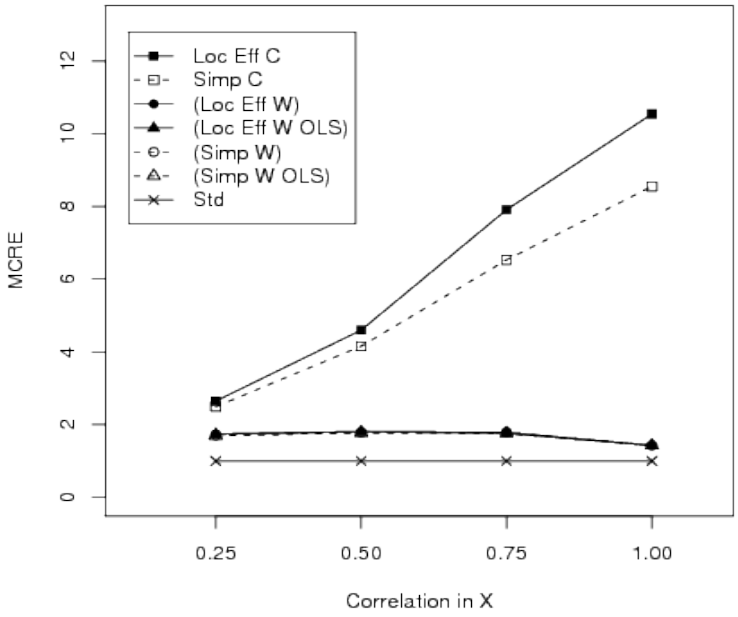
Figure 2: MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE. Continuous longitudinal outcomes. Estimators corresponding to each curve are denoted by 'Estimator (Marginal Working Covariance) Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; AR1-Autoregressive(1) $V(Y|A)$, True-Exchangeable/AR1 for $V\{E(Y|X,A)\}$ and $V(Y|X,A)$, respectively; C-Correct, W1-Wrong 1, W2-Wrong 2;α=0.3. The order of curves in the legend follows the order of curves on the figure, with the set of superimposed curves denoted by '[]'.



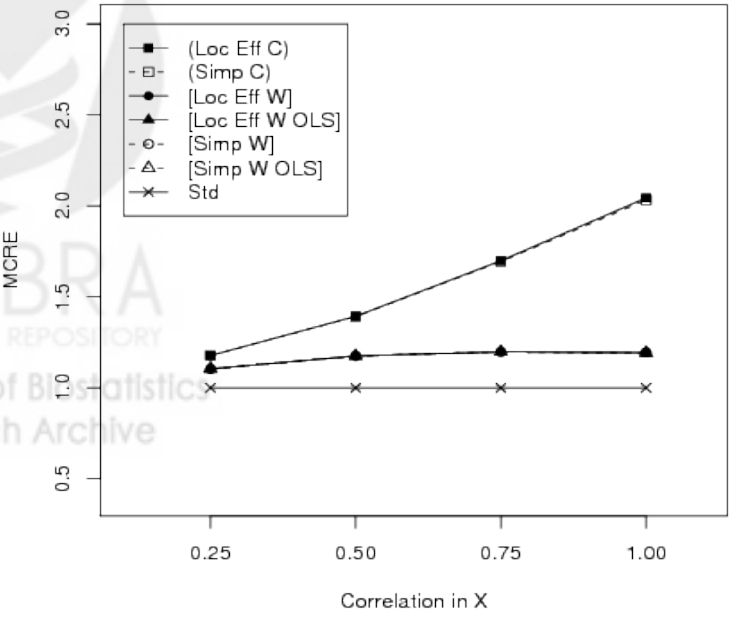
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Figure 3: MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE. Binary clustered outcomes. Estimators corresponding to each curve are denoted by 'Estimator-Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; C-Correct, W1-Wrong 1, W2-Wrong 2. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$. The order of curves in the legend follows the order of curves on the figure, with sets of superimposed curves denoted by '()'and '[]'.

(a) n_i=(10,20,30,40,50), $\theta = 1$



(b) n_i=(10,20,30,40,50), $\theta = 0.8$



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Table 1: **Monte Carlo Relative Efficiency of Locally Efficient Augmented GEE to Sub-optimal Augmented GEE:** Continuous clustered outcomes. Working Marginal Covariance (WM-Cov): Exchangeable (Exch). Outcome Regression (OR): Correct (C), Wrong 1(W1), Wrong 2 (W2). First entry $\sigma_b^2 = 0$, second entry $\sigma_b^2 = 6$. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$.

Cluster Size = 2,4,6,8,10,12

	Correlation among X_{ij}			
WMCov/OR	0.25	0.50	0.75	1.00
Exch/C	1.0115	1.0450	1.0907	1.1464
	1.0036	0.9991	1.0010	1.0085
Exch/W1	1.0062	1.0089	1.0064	1.0038
	1.0006	1.0008	1.0018	1.0019
Exch/W2	1.0114	1.0448	1.0905	1.1462
	1.0036	0.9990	1.0009	1.0083

Cluster Size =10,20,30,40,50

	Correlation among X_{ij}			
Cov/OR	0.25	0.50	0.75	1.00
Exch C	1.0356	1.1096	1.1563	1.2259
	1.0005	0.9999	1.0002	1.0011
Exch W1	1.0126	1.0081	1.0050	1.0032
	1.0000	1.0000	1.0001	1.0003
Exch W2	1.0352	1.1090	1.1556	1.2247
	1.0006	0.9998	1.0001	1.0009

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Table 2: **Monte Carlo Relative Efficiency of Locally Efficient Augmented GEE to Suboptimal Augmented GEE:** Continuous longitudinal outcomes. Working Marginal Covariance (WMCov): 1) True, exchangeable for $V(E(Y|X,A)|A)$ and AR1 for $V(Y|X,A)$ 2) AR1 for V(Y—A). Outcome Regression (OR): Correct (C), Wrong 1(W1), Wrong 2 (W2). First entry $\alpha = 0.1$, second entry $\alpha = 0.3$, third entry $\alpha = 0.5$.

	Correlation between Y and X					
WMCov/OR	10	20	30	40	50	60
True/C	1.0281	1.0700	1.1168	1.1662	1.2175	1.2702
	1.0166	1.0425	1.0728	1.1055	1.1398	1.1752
	1.0090	1.0234	1.0409	1.0603	1.0811	1.1028
True/W1	0.9995	0.9929	0.9851	0.9783	0.9735	0.9717
	1.0006	0.9974	0.9930	0.9887	0.9854	0.9837
	1.0009	0.9999	0.9982	0.9961	0.9943	0.9931
True/W2	1.0284	1.0703	1.1171	1.1664	1.2176	1.2701
	1.0168	1.0428	1.0731	1.1058	1.1401	1.1754
	1.0092	1.0237	1.0412	1.0606	1.0814	1.1031
AR1/W1	0.9916	0.9645	0.9300	0.8902	0.8832	0.8887
	0.9972	0.9858	0.9707	0.9567	0.9481	0.9481
	0.9996	0.9958	0.9903	0.9849	0.9811	0.9802

Table 3: **Monte Carlo Relative Efficiency of Locally Efficient Augmented GEE to Suboptimal Augmented GEE:** Binary clustered outcomes. Working Marginal Covariance (WMCov): Exch-Exchangeable. Outcome Regression (OR): Correct (C), Wrong 1 (W1), Wrong 2 (W2), Wrong 1 Linear Model (W1-LM). First entry $\theta = 1$, second entry $\theta = 0.8$. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$.

	Correlation between Y and X			
WMCov/OR	0.25	0.50	0.75	1.00
Exch/C	1.0624	1.1068	1.2113	1.2329
	0.9996	1.0009	1.0025	1.0057
Exch/W1	1.0247	1.0179	1.0025	1.0015
	1.0001	1.0003	1.0002	1.0001
Exch/W2	1.0630	1.1072	1.2080	1.2353
	0.9995	1.0009	1.0024	1.0056
Exch/W1-LM	1.0238	1.0171	1.0016	1.0008
	1.0001	1.0003	1.0001	1.0000

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Table 4: **Application of Standard, Simple Augmented, and Locally Efficient Augmented GEE to AIDS Clinical Trial Group Study 398.** Estimator (Working Marginal Covariance). Estimator: Unaugmented GEE (Standard), Simple Augmented GEE (Simple Aug. GEE), Locally Efficient Augmented GEE (Loc. Eff.). Working Marginal Covariance: Independence (Ind), Exchangeable (Exch), Autoregressive(1) (AR1), Unstructured (Un), Exchangeable for $V(E(Y|X,A)|A)$ and Unstructured for $V(Y|X,A)(Exch/Un)$, Exchangeable for $V(E(Y|X,A)|A)$ and AR1 for $V(Y|X,A)(Exch/AR1)$. Sandwich Standard Error (SE). Relative Efficiency (RE)

Estimator	$\hat{\beta}_1$	SE	RE
Standard (Ind)	9.971	20.772	0.942
Standard (Exch)	14.182	20.593	0.958
Standard (AR1)	16.977	20.222	0.993
Standard (Un)	20.173	20.156	1.000
Standard (Exch/Un)	14.615	20.347	0.981
Simple Aug. (Ind)	-8.110	9.203	4.797
Simple Aug. (Exch)	-6.385	8.904	5.124
Simple Aug. (AR1)	-3.059	9.244	4.754
Simple Aug. (Un)	-0.079	9.411	4.587
Simple Aug. (Exch/Un)	-5.972	8.571	5.530
Simple Aug. (Exch/AR1)	-5.048	8.920	5.106
Loc. Eff. (Ind)	-8.110	9.203	4.797
Loc. Eff. (Exch)	-6.821	8.953	5.068
Loc. Eff. (Exch/AR1)	-5.715	9.073	4.936
Loc. Eff. (Exch/Un)	-6.277	8.601	5.492
Adjusted (Un)	-6.649	8.621	5.467

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Supplementary Material

Appendices and additional figures referenced in Sections 2 & 3 are shown below.

Appendix: Deriving the Efficient Score

Let $O_i = (\mathbf{Y}_i, A_i, \mathbf{X}_i)$, where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, ...Y_{in_i})^{\mathrm{T}}$ is the n_i-dimensional response vector for the i^{th} independent unit, $i = 1, ..., m$, A_i is a scalar treatment assignment, and $\mathbf{X}_i$ is a matrix of auxiliary covariates. For data O_i augmented estimating functions $\psi_{i_{aug}}(O_i, t; \beta, h, \gamma)$ are constructed by (5). The optimal index $h_{opt}(A, t)$ is determined by solving the generalized information equality

$$-E\left[\frac{\partial\psi\{\mathbf{Y}, A, \mathbf{X}, t; \beta, \gamma, h(\cdot)\}}{\partial\beta^{\mathrm{T}}}\bigg|_{\beta=\beta_0}\right] = E\left[\psi\{\mathbf{Y}, A, \mathbf{X}, t; \beta, \gamma, h(\cdot)\}\psi^{\mathrm{T}}\{\mathbf{Y}, A, \mathbf{X}, t; \beta, \gamma, h_{opt}(\cdot)\}\bigg|_{\beta=\beta_0}\right],$$

for h_{opt}, where $h(\cdot)$ is any $p \times n_i$ function such that $E[\psi^{\mathrm{T}}\psi] < \infty$.

Conditioning on t, $h(A, t)$ takes up to K different matrix values, $h_0(t), h_1(t), ..., h_{K-1}(t)$, which may be denoted by K $p \times n_i$ constant matrices $\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{K-1}$. Similarly, we define $\Delta_k(\mathbf{X}) = E(\mathbf{Y}|A = k, \mathbf{X}, t) - \mathbf{g}(k, t; \beta)$, the n_i-dimensional vector of the difference in the conditional and marginal mean outcomes under treatment k, where $k = 0, 1, ..., K-1$. Using this construction, let $\mathbf{h} = [\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{K-1}]$ and $\Delta_K(\mathbf{X}) = \{\Delta_0^{\mathrm{T}}(\mathbf{X}), ..., \Delta_{K-1}^{\mathrm{T}}(\mathbf{X})\}^{\mathrm{T}}$. The complete index matrix $\mathbf{h}$ is therefore of dimension $p \times Kn_i$, while Δ_K is a Kn_i-dimensional vector. Estimating functions are then expressed concisely through defining two auxiliary matrix functions of A. Let $\mathbf{A}$ be the $Kn_i \times n_i$ matrix $\mathbf{A} = [I(A = 1)\mathbf{I}_n, ..., I(A = K)\mathbf{I}_n]^{\mathrm{T}}$ and $\mathbf{A}_\pi$ be the $Kn_i \times Kn_i$ block diagonal matrix composed of the diagonal matrices $\{I(A = k) - \pi_k\}\mathbf{I}_n$, where $\mathbf{I}_n$ denotes the $n_i \times n_i$ identity matrix.

Rewriting (5) using this notation, we obtain

$$\sum_{i=1}^{m} \mathbf{h}_i\mathbf{A}_i\{\mathbf{Y_i} - \mathbf{g}(A_i, t; \beta)\} - \mathbf{h}\mathbf{A}_{\pi_{\mathbf{i}i}}\Delta_i(\mathbf{X}_i) = \mathbf{0}. \tag{9}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Substituting this expression into Newey's equations we have

$$E\left[\mathbf{h}\mathbf{A}\frac{\partial \mathbf{g}(A,t;\beta)}{\partial \beta^{\mathrm{T}}}\right] = E\big[\{\mathbf{h}\mathbf{A}(\mathbf{Y}-\mathbf{g}(A,t;\beta)) - \mathbf{h}\mathbf{A}_{\pi}\Delta_K(\mathbf{X})\}\times \{(\mathbf{Y}-\mathbf{g}(A,t;\beta))^{\mathrm{T}}\mathbf{A}\mathbf{h}_{\mathbf{opt}}^{\mathrm{T}} - \Delta_K^{\mathrm{T}}(\mathbf{X})\mathbf{A}_{\pi}\mathbf{h}_{\mathbf{opt}}^{\mathrm{T}}\}\big]$$

We first note that since $\mathbf{h}$ and $\mathbf{h_{opt}}$ are constant, we can extract them from the expectation, leaving

$$\mathbf{h}^{\mathrm{T}}E\left[\mathbf{A}\frac{\partial \mathbf{g}(A,t;\beta)}{\partial \beta^{\mathrm{T}}}\right] = \mathbf{h}^{\mathrm{T}}E\big[\{\mathbf{A}(\mathbf{Y}-\mathbf{g}(A,t;\beta)) - \mathbf{A}_{\pi}\Delta_K(\mathbf{X})\}\times \{(\mathbf{Y}-\mathbf{g}(A,t;\beta))^{\mathrm{T}}\mathbf{A} - \Delta_K^{\mathrm{T}}(\mathbf{X})\mathbf{A}_{\pi}\}\big]\mathbf{h}_{\mathbf{opt}}^{\mathrm{T}}$$

Since $\mathbf{h}$ is nonzero, it must hold that

$$E\left[\mathbf{A}\frac{\partial \mathbf{g}(A,t;\beta)}{\partial \beta^{\mathrm{T}}}\right] = E\left[\{\mathbf{A}(\mathbf{Y}-\mathbf{g}(A,t;\beta)) - \mathbf{A}_{\pi}\Delta_K(\mathbf{X})\}\{(\mathbf{Y}-\mathbf{g}(A,t;\beta))^{\mathrm{T}}\mathbf{A} - \Delta_K^{\mathrm{T}}(\mathbf{X})\mathbf{A}_{\pi}\}\right]\mathbf{h}_{\mathbf{opt}}^{\mathrm{T}} \tag{10}$$

Evaluating the left hand side of the equation, we have

$$E\left\{\begin{bmatrix} A_0\mathbf{I}_n \\ A_2\mathbf{I}_n \\ \vdots \\ A_{K-1}\mathbf{I}_n \end{bmatrix}\frac{\partial \mathbf{g}(A,t;\beta)}{\partial \beta^{\mathrm{T}}}\right\} = \begin{bmatrix} \pi_0\frac{\partial \mathbf{g}(0,t;\beta)}{\partial \beta^{\mathrm{T}}} \\ \pi_1\frac{\partial \mathbf{g}(1,t;\beta)}{\partial \beta^{\mathrm{T}}} \\ \vdots \\ \pi_{K-1}\frac{\partial \mathbf{g}(K-1,t;\beta)}{\partial \beta^{\mathrm{T}}} \end{bmatrix} \tag{$\mathbf{D}^*$}$$

Evaluating the right hand side, we note that we have an expression of the form $E[A-B][A_{opt}-B_{opt}]^{\mathrm{T}}$. Interpreting the augmented estimating function as a residual, we note that $A-B \perp B_{opt}$. We can therefore evaluate $E[A-B][A_{opt}-B_{opt}]^{\mathrm{T}} = E[A-B][A_{opt}]^{\mathrm{T}}$. In (10), this becomes

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

$$E[\mathbf{A}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}\{\mathbf{Y}-g(A,t;\beta)\}^{\mathrm{T}}\mathbf{A}]-E[\mathbf{A}_{\pi}\Delta(\mathbf{X})\{\mathbf{Y}-g(A,t;\beta)\}^{\mathrm{T}}\mathbf{A}] \tag{11}$$

Regarding the first term in (11), we have

$$E[\mathbf{A}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}\{\mathbf{Y}-g(A,t;\beta)^{\mathrm{T}}\}\mathbf{A}]=$$

$$E\begin{bmatrix} A_0A_0\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & A_0A_1\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & \cdots & A_0A_{K-1}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} \\ A_1A_0\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & A_1A_1\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & \cdots & A_1A_{K-1}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{K-1}A_0\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & A_{K-1}A_1\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} & \cdots & A_{K-1}A_{K-1}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2} \end{bmatrix}, \tag{12}$$

where $U^{\otimes 2}$=UU^{T}. Since each individual is only assigned to one treatment, only one of A_0, A_1,...,A_{K-1} is nonzero. The non diagonal blocks of (12) are identically 0. The diagonal blocks contain terms of the form $E[A_kA_k\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2}]=E[A_k\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\otimes 2}]=\pi_kV(\mathbf{Y}|A=k)$. Matrix (12) is written as

$$\begin{bmatrix} \pi_0V(\mathbf{Y}|A=0) & 0 & \cdots & 0 \\ 0 & \pi_1V(\mathbf{Y}|A=1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \pi_{K-1}V(\mathbf{Y}|A=K-1) \end{bmatrix} \tag{$\mathbf{C}_1$}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Evaluating the second term of (11), we have

$$
\begin{aligned}
&E[\mathbf{A}_\pi \Delta_K(\mathbf{X})\{\mathbf{Y} - g(A,t,\beta)\}^{\mathrm{T}}\mathbf{A}] = \\
&E\Bigg\{ \begin{bmatrix} (A_0-\pi_0)\mathbf{I}_n & \dots & 0 \\ \vdots & (A_1-\pi_1)\mathbf{I}_n & \vdots \\ 0 & \cdots & (A_{K-1}-\pi_{K-1})\mathbf{I}_n \end{bmatrix} \begin{bmatrix} \Delta_0(\mathbf{X}) \\ \Delta_1(\mathbf{X}) \\ \vdots \\ \Delta_{K-1}(\mathbf{X}) \end{bmatrix} \times \\
&\{\mathbf{Y} - \mathbf{g}(A,t;\beta)\}^{\mathrm{T}} \begin{bmatrix} A_0\mathbf{I}_n & A_1\mathbf{I}_n & \cdots & A_{K-1}\mathbf{I}_n \end{bmatrix} \Bigg\} \\
=&E\left\{ \begin{bmatrix} (A_0-\pi_0)A_0\Delta_0(\mathbf{X})\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\mathrm{T}} & \cdots & (A_0-\pi_0)A_{K-1}\Delta_0(\mathbf{X})\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\mathrm{T}} \\ (A_1-\pi_1)A_0\Delta_1(\mathbf{X})\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\mathrm{T}} & \ddots & (A_1-\pi_1)A_{K-1}\Delta_1(\mathbf{X})\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\mathrm{T}} \\ \vdots & \ddots & \vdots \\ (A_{K-1}-\pi_{K-1})A_0\Delta_{K-1}(\mathbf{X})\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\mathrm{T}} & \cdots & (A_{K-1}-\pi_{K-1})A_{K-1}\{\mathbf{Y}-\mathbf{g}(A,t;\beta)\}^{\mathrm{T}} \end{bmatrix} \right\} \\
=&E\left\{ \begin{bmatrix} (A_0-\pi_0)A_0\Delta_0(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) & \cdots & (A_0-\pi_0)A_{K-1}\Delta_0(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) \\ (A_1-\pi_1)A_0\Delta_1(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) & \ddots & (A_1-\pi_1)A_{K-1}\Delta_1(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) \\ \vdots & \ddots & \vdots \\ (A_{K-1}-\pi_{K-1})A_0\Delta_{K-1}(\mathbf{X})\Delta_A^{\mathrm{T}}(\mathbf{X}) & \cdots & (A_{K-1}-\pi_{K-1})A_{K-1}\Delta_A^{\mathrm{T}}(\mathbf{X}) \end{bmatrix} \right\} \\
=&\begin{bmatrix} \pi_0(1-\pi_0)\Delta_0(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X}) & \cdots & -\pi_0\pi_{K-1}\Delta_0(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X}) \\ -\pi_1\pi_0\Delta_1(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X}) & \ddots & -\pi_1\pi_{K-1}\Delta_1(\mathbf{X})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X}) \\ \vdots & \ddots & \vdots \\ -\pi_{K-1}\pi_0\Delta_{K-1}(\mathbf{X})\Delta_0^{\mathrm{T}}(\mathbf{X}) & \cdots & \pi_{K-1}(1-\pi_{K-1})\Delta_{K-1}^{\mathrm{T}}(\mathbf{X}) \end{bmatrix} \qquad (\mathbf{C_2})
\end{aligned}
$$

From ($\mathbf{C_2}$), we see that generally, the second term of (11) contains block diagonal terms $\pi_k(1-\pi_k)E_{\mathbf{X}}\left\{\Delta_k^{\otimes 2}(\mathbf{X})\right\}$, and block off-diagonal terms $-\pi_j\pi_k E_{\mathbf{X}}\{\Delta_j(\mathbf{X})\Delta_k^{\mathrm{T}}(\mathbf{X})\}$.

Referring back to (10), we see that $\mathbf{h_{opt}} = [\mathbf{C}_1 - \mathbf{C}_2]^{-1}\mathbf{D}^*$, as labeled above.

CODRA
Collection of Biostatistics
Research Archive

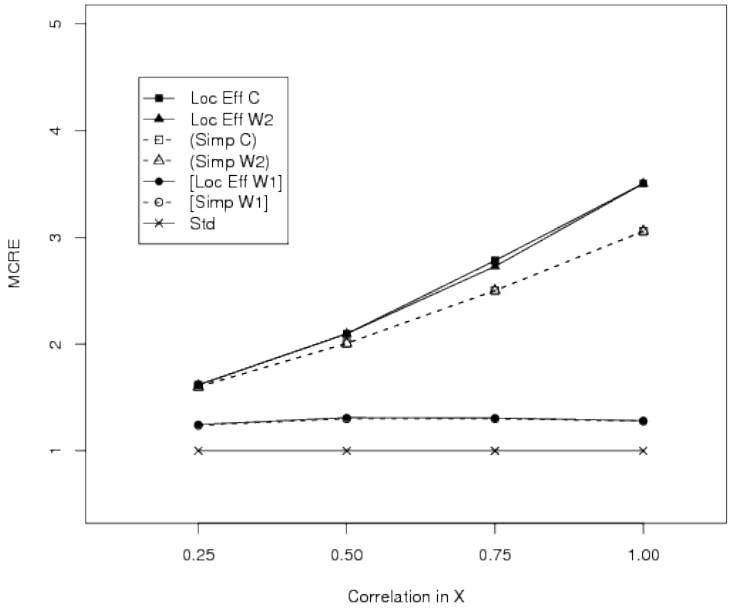
Supplementary Figures: Simulation Results

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

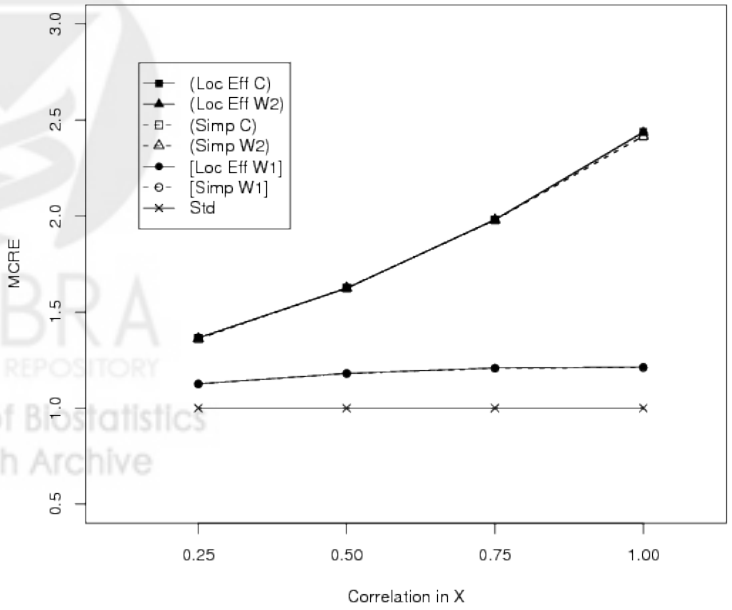
Hosted by The Berkeley Electronic Press

Figure 4: MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE. Continuous clustered outcomes. Estimators corresponding to each curve are denoted by 'Estimator-Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; C-Correct, W1-Wrong 1, W2-Wrong 2. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$. The order of curves in the legend follows the order of curves on the figure, with sets of superimposed curves denoted by '()' and '[]'.

(a) n_i=(2,4,6,8,10,12), $\sigma_b^2 = 0$



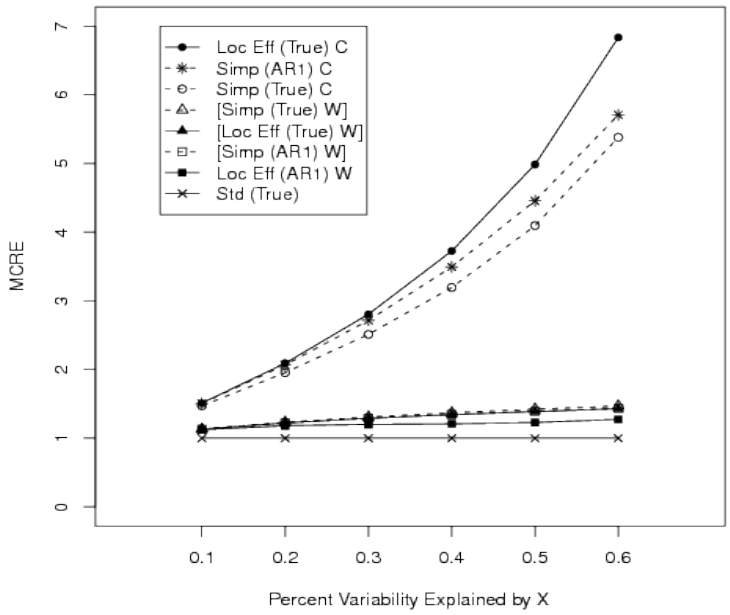
(b) n_i=(2,4,6,8,10,12), $\sigma_b^2 = 6$



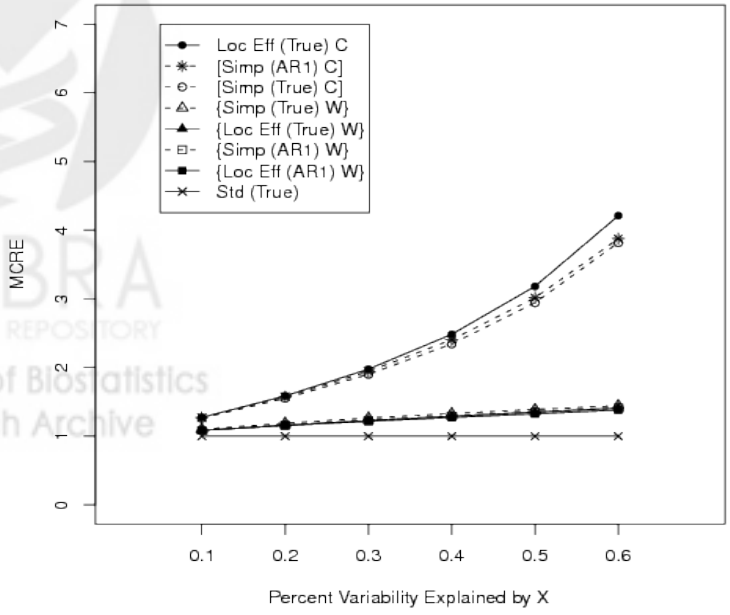
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Figure 5: MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE. Continuous longitudinal outcomes. Estimators corresponding to each curve are denoted by 'Estimator (Marginal Working Covariance) Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; AR1-Autoregressive(1) $V(Y|A)$, True-Exchangeable/AR1 for $V\{E(Y|X,A)\}$ and $V(Y|X,A)$, respectively; C-Correct, W1-Wrong 1, W2-Wrong 2;α=0.3. The order of curves in the legend follows the order of curves on the figure, with the set of superimposed curves denoted by '[]' and '{}'.

(a) α=0.1



(b) α=0.5

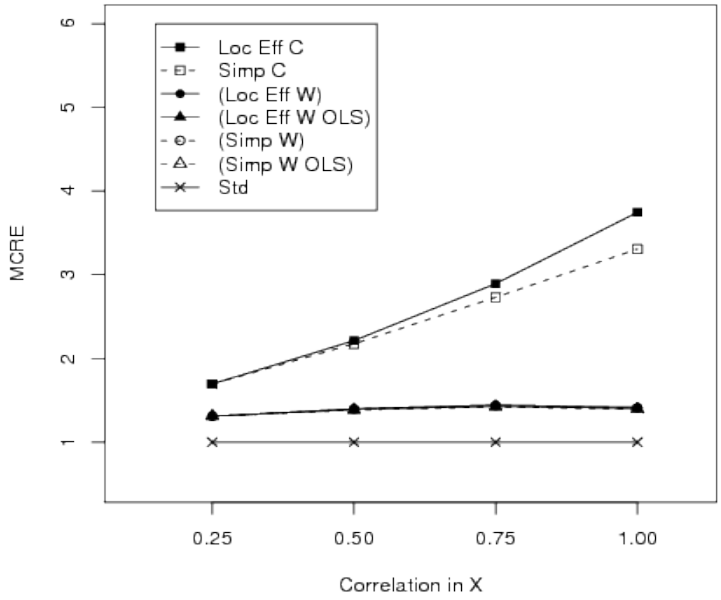


COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

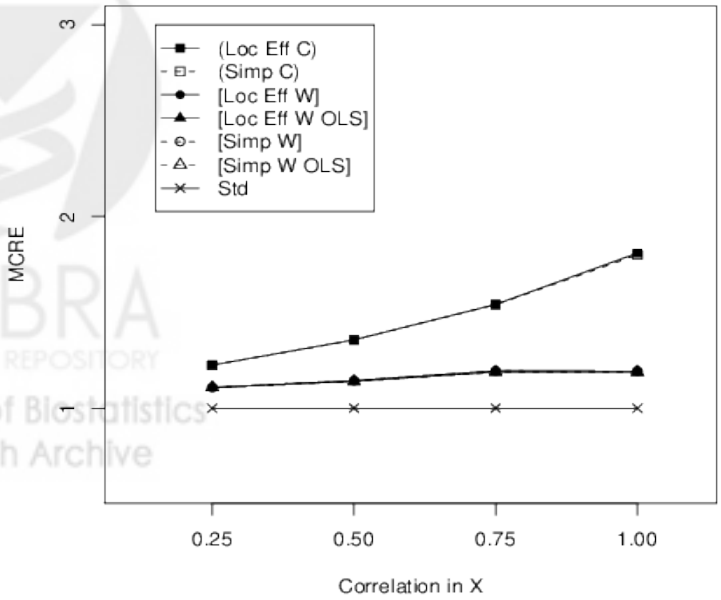
Hosted by The Berkeley Electronic Press

Figure 6: MCRE of Locally Efficient and Simple Augmented GEE Relative to Standard (Unaugmented) GEE. Binary clustered outcomes. Estimators corresponding to each curve are denoted by 'Estimator-Outcome Regression' using the abbreviations: Loc Eff-Locally Efficient, Simp-Simple Augmented, Std-Standard; C-Correct, W1-Wrong 1, W2-Wrong 2. All estimators use exchangable working covariance for $V(Y|A)$ and $V\{E(Y|X,A)\}$. The order of curves in the legend follows the order of curves on the figure, with sets of superimposed curves denoted by '()'.

(a) n_i=(2,4,6,8,10,12), $\theta = 1$



(b) n_i=(2,4,6,8,10,12), $\theta = 0.8$



COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Supplementary Table: QIC for selecting working covariance structures

Table 5: QIC for selecting working covariance structures. Conditional model: $E(CD4_{ij}|Trt_i, Week_{ij}, \mathbf{X}_i) = \eta_0 + \eta_1 A_i + \eta_2 Week_{ij} + \eta_3 Sex_i + \eta_4 CD4_{0_i}$. Marginal model: $E(CD4_{ij}|Trt_i, Week_{ij}) = \eta_0 + \eta_1 Trt_i + \eta_2 Week_{ij}$

Conditional Model

Working Covariance Structure	QIC
Independence	1053.44
Exchangeable	1051.9
AR1	1052.29
Unstructured	1049.72

Marginal Model

Working Covariance Structure	QIC
Independence	1047.59
Exchangeable	1047.1
AR1	1046.56
Unstructured	1049.35

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press