Semi-parametric Estimation of Biomarker Age Trends with Endogenous Medication Use in Longitudinal Data

Andrew J. Spieker
Department of Biostatistics
Vanderbilt University Medical Center
Nashville, TN 37203, USA
andrew.spieker@vumc.org

Joseph A.C. Delaney
College of Pharmacy
University of Manitoba
Winnipeg, MB R3E 0T5, CA
joseph.delaney@umanitoba.ca

Robyn L. McClelland
Department of Biostatistics
University of Washington
Seattle, Washington 98195, USA
rmcclell@u.washington.edu

Abstract
In cohort studies, non-random medication use can pose barriers to estimation of the natural history trend in a mean biomarker value—namely, the association between a predictor of interest and a biomarker outcome that would be observed in the total absence of biomarker-specific treatment. Common causes of treatment and outcomes are often unmeasured, obscuring our ability to easily account for medication use with assumptions commonly invoked in causal inference such as conditional ignorability. Further, without a high degree of confidence in the availability of a variable satisfying the exclusion restriction, use of instrumental variable approaches may be difficult to justify. Heckman’s hybrid model with structural shift (sometimes referred to less specifically as the treatment effects model) can be used to correct endogeneity bias via a homogeneity assumption (i.e., that average treatment effects do not vary across covariates) and parametric specification of a joint model for the outcome and treatment. In recent work, we relaxed the homogeneity assumption by allowing observed covariates to serve as treatment effect modifiers. While this method has been shown to be reasonably robust in settings of cross-sectional data, application of this methodology to settings of longitudinal data remains unexplored. We demonstrate how the assumptions of the treatment effects model can be extended to accommodate clustered data arising from longitudinal studies. Our proposed approach is semi-parametric in nature in that valid inference can be obtained without the need to specify any component of the longitudinal correlation structure. As an illustrative example, we use data from the Multi-Ethnic Study of Atherosclerosis to evaluate trends in low-density lipoprotein by age and gender. Results from a collection of simulation studies, as well as our illustrative example, confirm that our generalization of the treatment effects model can serve as a useful tool to uncover natural history trends in longitudinal data that are obscured by endogenous treatment.

Keywords: Biomarker, Cohort study, Endogenous, Longitudinal data; Multi-Ethnic Study of Atherosclerosis

© Andrew J. Spieker, Joseph A.C. Delaney, and Robyn L. McClelland.
1. Introduction

Often in epidemiologic cohort studies involving cardiovascular biomarkers, a large segment of study participants are on one or more medications intended to specifically alter those biomarker values. For such subjects, the treated biomarker value is perhaps of greatest relevance when seeking to understand subject-specific risk of subsequent cardiovascular events. However, when seeking to uncover associations between certain deterministic or time-stable exposures (e.g., age or gender) and a biomarker outcome, the natural history of the biomarker that would have occurred in the absence of treatment is more informative than the value under observed treatment. For treated subjects, the biomarker’s natural history is contaminated by endogeneity in that participants differing in their medication use status tend to differ in their underlying biomarker values due to unmeasured confounding by indication. Since the goal of treatment is to shift biomarker values toward a healthier range, the association between an exposure of interest and the observed biomarker value cannot be expected to provide an adequate representation of the natural history association that would have occurred had treatment not been a factor.

For observational studies in which patients enter the cohort already on medication, pre-treatment values are typically unavailable; in this setting, endogeneity cannot be overcome by standard approaches such as regression adjustment or inverse probability weighting, in which it is assumed that all confounders are measured and properly accounted for. Heckman’s hybrid model with structural shift (henceforth referred to as the “treatment effects model”) was proposed as a method to estimate treatment effects that are subject to unmeasured confounding (Heckman, 1978). In prior work, we examined this approach specifically through the likelihood-based framework of Maddala (1983) in order to correct endogeneity bias for estimation of the natural history association between a predictor and a biomarker in cross-sectional data; from this vantage point, we view treatment as a nuisance to be accounted for rather than a parameter of interest (Spieker et al., 2015). Identification of the natural history association is obtained through (1) parametric specification of a joint model for the outcome and treatment, and (2) a treatment effect homogeneity assumption; these assumptions allow us to bypass both the conditional ignorability assumption typically associated with inverse probability of treatment weighting methods (Robins et al., 2000) and the exclusion restriction assumption typically associated with instrumental variable methods (Imbens and Angrist, 1994). Recently, we developed an extension of the treatment effects model to allow treatment effects to vary across covariates, thereby relaxing the homogeneity assumption (Spieker et al., 2018). Further, our prior exploration of this approach has demonstrated that this modeling framework exhibits fairly robust behavior for estimation of the natural history association under violations to a range of assumptions, including misspecified error distributions, variable omission, and non-differential exposure mismeasurement.

Many longitudinal studies over the last two decades have sought to characterize trends in mean LDL across age and/or by gender in sub-populations of interest (Carroll et al., 2005; Duncan et al., 2019; Gupta et al., 2016; Russo et al., 2015; Singh et al., 2012; Zitnanova et al., 2020). Conclusions from these studies were largely derived on the basis of trends in observed LDL values, and were not based on methods that attempt to correct for endogenous medication use. Although the framework of Heckman’s treatment effects model
Biomarker Trends with Endogenous Treatment

can be used to correct endogeneity bias in the setting of cross-sectional observational data, generalizations to longitudinal studies in which observations are correlated within subjects over time remain unexplored. In this manuscript, we seek to address this methodological gap by developing and justifying a longitudinal endogeneity model (LEM).

The remainder of this manuscript is organized as follows. In Section 2, we provide background on the assumptions of Heckman’s treatment effects model for estimation of natural history associations in cross-sectional data, including our extension to relax the assumption of treatment effect heterogeneity. We further provide information on computational strategies to aid the implementation of this approach. In Section 3, we present our proposed longitudinal endogeneity model, focusing specifically on the assumptions necessary for identification. In Section 4, we conduct a variety of simulation studies to evaluate the finite sample properties of the LEM. In Section 5, we illustrate the use of our approach for estimation of LDL trends by age and gender using data from the Multi-Ethnic Study of Atherosclerosis. Finally, we conclude in Section 6 with a discussion of our findings, conclusions, and possible directions for further research.

2. Background

In this section, we summarize how Heckman’s treatment effects model can be used to achieve the goal of estimating what we refer to as the natural history association between a predictor of interest and a biomarker outcome in cross-sectional data, including our prior extension to accommodate covariate-specific treatment effects (Spieker et al., 2018).

2.1 Notation, assumptions, and definitions

Let \( i = 1, \ldots, N \) index independently sampled study participants, each having observed treatment status \( A_i \) (assumed binary). We follow the potential outcomes notation of Rubin (2005). We let \( Y_a^i \) denote the potential biomarker value that would be observed under treatment \( A_i = a \), and \( Y_i \) the observed biomarker outcome. Each subject has total vector of exogenous covariates \( O_i \) having length \( J \) that can be subsetted in the following useful ways: we let \( X_i \subseteq O_i \) denote predictors of the outcome, \( Z_i \subseteq O_i \) denote predictors of treatment, and \( W_i \subseteq O_i \) denote covariates that modify the association between treatment and the outcome. Importantly, covariates appearing in \( X_i \), \( Z_i \), and \( W_i \) need not be mutually exclusive. A directed acyclic graph (DAG) is depicted in Figure 1 in order to illustrate the presumed relationship between variables.

Our goal is to estimate the association between (at least one predictor appearing in) \( X \) and \( Y^0 \), the biomarker value that would be observed in the absence of medication use. Under a linearity assumption, a parameter of this sort can be expressed succinctly as:

\[
    \beta_j = E[Y^0 | X = (x_1, \ldots, x_j-1, x_j + 1, x_{j+1}, \ldots, x_JX)] - E[Y^0 | X = (x_1, \ldots, x_j-1, x_j, x_{j+1}, \ldots, x_{j+1}, \ldots, x_JX)],
\]

where \( J_X \) denotes the number of covariates appearing in \( X \). Since \( Y^0 \) is only observed for those subjects who are not treated, standard regression-based approaches for estimation of \( \beta_j \) are not valid. Importantly, a traditional causal interpretation for \( \beta_j \) may not be appropriate, as \( X_j \) need not be manipulable (e.g., age). As previously discussed in Section
Figure 1: DAG depicting the relationship between variables. The total covariate vector of exogenous covariates, $O$, contains variables that can be associated with either $A$ alone, $Y$ alone, or both; for ease of presentation, we do not partition $O$ in this graphical depiction. The parameter of interest can be described as the association between any possible subset of the covariates in $O$ and the off-treatment potential outcome, $Y^0$ (shown in bold). Importantly, $U$ contains a collection of unobserved confounders (e.g., a pre-treatment biomarker value), such that $A$ is endogenous.

1, the population average causal effect of the endogenous treatment variable, $A$, on the outcome, $Y$, is not itself of interest the setting we are describing in this work; instead, treatment serves as a nuisance in the goal of estimating $\beta$. Still, certain conditions must hold regarding the relationship between these variables in order for $\beta$ to be identified. To that end, we make the following initial identifying assumptions:

1. No interference: $Y_i^a \perp \perp A_{i'}, 1 \leq i \neq i' \leq N$.

2. Consistency: $Y_i = Y_i^{A_i}$.


The assumption of no interference states that the treatment status of one individual does not influence the potential outcome of another individual. Taken together with the assumption of no interference, consistency ensures that potential outcomes are well defined and that for each subject, the observed outcome corresponds to the potential outcome under the observed treatment. The conditional homogeneity assumption is an extension of the treatment effects model that allows any arbitrary subset of observed covariates to serve as possible modifiers of the treatment effect (Spieker et al., 2018). As is implied by the notation, all covariates in $O$ that serve as effect modifiers must be included in $W$.

The common assumption of conditional ignorability (namely, that $Y^a \perp A|O$) is notably not one of the identifying assumptions we will plan to invoke in this work. Specifically, we wish to allow for the possibility of unmeasured confounding of the treatment-outcome relationship. Further, we do not propose an exclusion restriction or monotonicity assumption.
Identification of parameters of interest will instead depend upon assumptions specific to the estimation procedure, discussed in Section 2.2.

2.2 Estimation in cross-sectional data

For ease of notation, let $Y = (Y_1, \ldots, Y_N)^T$ denote the outcome vector, and let $X$, $Z$, and $W$ denote design matrices for each of the respective covariate groupings described in Section 2.1. The estimation approach described by Spieker et al. (2018) uses maximum likelihood to estimate parameters from the following system of simultaneous structural equations:

$$
Y = X\beta + W\eta \circ A + \epsilon;
A^* = Z\alpha + \gamma.
$$

Here, $\circ$ denotes element-wise multiplication, and $\epsilon$ and $\gamma$ denote vectors of error terms that are each independent and identically distributed across subjects, but are possibly themselves correlated. This error correlation is designed to reflect unmeasured common causes of $A$ and $Y$. Note that $A^*$ denotes a continuous latent underlying treatment variable, from which the dichotomous treatment $A = 1(A^* > 0)$ is observed. This parameterization is equivalent to a probit model for the probability of treatment, though the error representation presented above is convenient in forming the likelihood estimator: if $\theta$ denotes all model parameters, a general form for the log-likelihood can be expressed as follows:

$$
\ell(\theta; O, A, Y) = \sum_{i=1}^{N} \left[ \log p_\theta(Y_i|O_i) + \log \int_{A_i} p_\theta(A_i^*|O_i) dA_i^* \right],
$$

where $A_i = (-\infty, 0)$ if $A = 0$ and $(0, \infty)$ if $A = 1$. One notable example in which identifiability of $\beta$ has been shown to be achieved is the setting under which the error terms are presumed to follow a common bivariate normal distribution with outcome error variance $\sigma_Y^2$ and correlation $\rho$. As the latent treatment error variance, $\sigma_A^2$ is not itself identifiable, the usual procedure it to set $\sigma_A^2 = 1$, from which weak identifiability of $\rho$ and $\alpha$ are achieved through the equivalence between $\rho$ and the point-biserial correlation (Freedman and Sekhon, 2010; Tate, 1954). Under the bivariate normal specification, the log-likelihood can be expressed for $\theta = (\alpha, \beta, \eta, \sigma_Y, \rho)$ as follows:

$$
\ell(\theta; O, A, Y) = \sum_{i=1}^{N} \left[ \log \phi \left( \frac{Y_i - X_i^T \beta - W_i^T \eta \times A_i}{\sigma_Y} \right) - \log \sigma_Y 
+ \log \Phi \left( \frac{(-1)^{-A_i} Z_i^T \alpha + \rho(Y_i - X_i^T \beta - W_i^T \eta \times A_i) / \sigma_Y}{\sqrt{1 - \rho^2}} \right) \right].
$$

Note that $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and cumulative distribution functions, respectively. Building off of identifiability justifications provided by Heckman (1978) and Maddala (1983), our prior work shows that $\eta$ is identifiable (Spieker et al., 2018). Further note that the parametric specification of a probit model for medication use implies the usual positivity assumption (namely, that $P(A = 1|Z = z) > 0 \forall z$ such that $f_Z(z) > 0$). For the purposes of this work, in which our primary focus is on $\beta$, we will not give $\eta$ much attention and will consider it as a nuisance parameter.
2.3 Likelihood evaluation, software, and robustness

We briefly comment on evaluation of the log-likelihood of Equation (3). First, we note that in order for the negative log-likelihood to possess a global maximum on the interior of the parameter space, we require the easily evaluable condition of outcome overlap, meaning that \( \bigcap_{a \in \{0,1\}} (\min_{A_i=a} Y, \max_{A_i=a} Y) \neq \emptyset \). Second, the resulting score equations do not possess a closed-form analytic solution; numerical techniques are required to evaluate the likelihood.

The correlation parameter, \( \rho \), must satisfy the condition that \(-1 < \rho < 1\); to that end, one may wish to re-parameterize via either the relationship \( \rho = 2 \arctan(\varphi)/\pi \) or \( \rho = 2/(1+\exp(-\varphi)) - 1 \). Since each of these mappings is bijective, one can optimize over \( \varphi \) to obtain the maximum likelihood estimate of \( \rho \), thereby bypassing computational challenges associated with constraining the optimization. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm can be used to find a numeric solution to the likelihood (Fletcher, 1987).

We have made an R package available on CRAN titled endogenous. The function hybrid within this package can be used to maximize the likelihood of Equation (3). Parameters are initialized based on standard regression approaches, and the BFGS algorithm is used to iterate toward the maximum likelihood solution based on the gradient, included to provide more computationally efficient and algorithmically stable estimation as compared to quasi-Newton based techniques. Standard likelihood theory allows the use of the asymptotically efficient inverse Fisher information matrix for settings in which the parametric assumptions are thought to hold. The hybrid function also contains a robust sandwich variance estimator in the spirit of White (1980) to provide valid standard errors under heteroscedasticity. Either can be used to formulate confidence intervals and conduct robust Wald-based hypothesis tests, described in further detail by Spieker et al. (2018).

Despite what are sometimes viewed as stringent parametric assumptions, prior work has demonstrated that this modeling framework exhibits fairly robust behavior for estimation of \( \beta \) in a variety of settings, including skewed or heavy-tailed errors, various forms of variable omission, and non-differential exposure mismeasurement (Spieker et al., 2015).

3. The longitudinal endogeneity model

In this section, we propose a longitudinal endogeneity model (LEM) to estimate marginal natural history associations in longitudinal data.

3.1 Expanding notation, assumptions, and definitions

We expand the notation of Section 2.1 as follows, following the potential outcomes notation of Robins (1986) for longitudinal data. Suppose each subject has \( T \) observation times (we will later discuss variable observation times). We let \( O_{it}, A_{it}, \) and \( Y_{it} \) denote the covariate vector, treatment status, and biomarker outcome, respectively, for subject \( i \) at time \( t \). We use overbar notation in defining variable history (e.g., \( \overline{A}_{it} = (A_{i1}, \ldots, A_{it}) \), treatment history up through time \( t \)). Analogously, let \( \overline{A}_i = (A_{i1}, \ldots, A_{iT}) \) denote treatment from time \( t \) onward. For convenience of notation, we let \( \overline{A}_{it} = \overline{A}_{iT} \) denote the entire treatment history. Potential outcomes are in turn defined by hypothetical treatment history, \( Y_{it}^\pi \). We expand our goal of Section 2.1 to estimation of the population-average natural history
association between \( X \) and \( Y \) that would be observed in the absence of medication use:

\[
\beta_j = \mathbb{E}[Y_{i,t}^\pi=0 | \mathbf{X}_t = (x_{t1}, \ldots, x_{t(j-1)}, x_{tj} + 1, x_{t(j+1)} \ldots, x_{tJ})] \\
- \mathbb{E}[Y_{i,t}^\pi=0 | \mathbf{X}_t = (x_{t1}, \ldots, x_{t(j-1)}, x_{tj}, x_{t(j+1)} \ldots, x_{tJ})].
\]

(4)

We update the identifying assumptions of Section 2.1 to generalize to the longitudinal
setting as follows:

1. No interference: \( Y_{i,t}^\pi \perp \perp A_{j,t}', 1 \leq i \neq i' \leq N, 1 \leq t \neq t' \leq T. \)

2. Consistency: \( Y_{i,t} = Y_{i,t}^{A_{i,t}}. \)

3. Conditional homogeneity: \( \mathbb{E}[Y_{i,t}^\pi_{t-1:0,\pi_{t+1}^{a_t+1}} - Y_{i,t}^\pi_{t-1:0,\pi_{t+1}^{a_t+1}} | \mathcal{O}_t] = W_t^T \eta \forall (\pi_{t-1}, \pi_{t+1}). \)

The generalization of Assumptions 1 and 2 to the setting of repeated measures is straight-
forward. Conditional homogeneity implies that, conditional on \( W_t \), the causal effect of
\( t \) is homogeneous with respect to all other covariates in \( \mathcal{O}_t \).

3.2 Semi-parametric estimation

Given the joint nature of the modeling approach (i.e., between the treatment \( A \) and the
outcome \( Y \)), there are a resulting four major classes of correlation that must be considered
when both \( A \) and \( Y \) are measured over time. Figure 2 illustrates these correlation classes
in the simple setting of two observations. The first form of correlation is one we have
already considered, and that is the correlation between the error terms of the treatment
and outcome models at concurrent times: \( \rho = \text{Cov}(\epsilon_{i,t}, \gamma_{i,t}). \) The other types of correlation
are longitudinal in nature, and we make no assumption about their values or functional
structures.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Illustration of the different classes of correlation under consideration. Of note,
\( \rho_A, \rho_Y, \) and \( \rho_{AY} \) may be nonzero, though they are neither modeled nor estimated
in our proposed LEM.}
\end{figure}
Let \( \ell_{it}(\theta; O_{it}, A_{it}, Y_{it}) \) denote the contribution of subject \( i \) to the log-likelihood for \( \theta \) at observation \( t \). We propose the following semi-parametric estimating equations for \( \theta \) based on the total of \( \sum_{i=1}^{N} T \) observations:

\[
U_N(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{\partial \ell_{it}(\theta; O_{it}, A_{it}, Y_{it})}{\partial \theta} = 0.
\] (5)

In contrast to the likelihood proposed for cross-sectional data, these equations are not fully parametric score equations as they do not involve the correlation within subjects over time. Although such correlations are all but certain to be nonzero in most practical settings, they need not be estimated. As the cross-sectional score equations have expectation zero, it follows that \( E[U_N(\theta)] = 0 \), and hence these estimating equations can be solved for \( \theta \) to obtain an estimator \( \hat{\theta} \). A robust cluster-based covariance estimator can be used to form standard errors and conduct inference on \( \beta \subseteq \theta \). Denoting the indices of \( \beta \) as a subset of \( \theta \) as 1:(\( J_X + 1 \)) without loss of generality,

\[
\hat{\text{Var}}(\beta) = \left[ \left( \frac{\partial U(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right)^T \left[ \sum_{i=1}^{N} U_i(\hat{\theta})U_i(\hat{\theta})^T \right] \left( \frac{\partial U(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right)^{-1} \right]_{((1:(J_X+1)),(1:(J_X+1)))}.
\] (6)

The hybrid function in the R package endogenous accommodates repeated measures on subjects via a subject identifier argument. The robust cluster-based covariance estimator is the only one available estimator in this setting, although a nonparametric bootstrap procedure could be manually programmed in its place (Davison and Hinkley, 1997).

3.3 Variable follow-up times and informative missingness

The proposed approach can be extended to accommodate variation in the number of observations per subject. Assumptions 1 through 3 generalize in the expected way, and the inner sum of the estimating equations is taken over the number of observation times for the particular subject indexed in the outer sum. Consistency of \( \hat{\beta} \) for \( \beta \) will not generally hold if cluster size is informed by the latent error terms. This is analogous to assumptions made by approaches such as generalized estimating equations (GEE) with a working independence correlation structure (Hoffman et al., 2001).

4. Simulation study

We present the results of a simulation study to elucidate the finite-sample properties of the proposed semi-parametric estimation procedure of Section 3.2. In particular, we focus on bias, standard errors, and coverage probability for estimation of \( \beta \). All simulation studies were conducted using R, version 4.0.2 (2020).

4.1 Correct specification with panel data

Consider a study of \( N \) independent observations, each with \( T = 3 \) observations. Let \( O_{it} \) denote a total covariate vector of length \( J = 7 \), with \( X_{it} = (1, O_{it1}, O_{it4}, O_{it5}, O_{it7}) \), \( Z_{it} = (1, O_{it2}, O_{it4}, O_{it6}, O_{it7}) \), and \( W_{it} = (1, O_{it3}, O_{it5}, O_{it6}, O_{it7}) \); that is, four of the seven
Biomarker Trends with Endogenous Treatment

covariates are related to each the outcome, four are related to treatment, and four related
to effect size. The purpose of this formulation is so that there is a predictor appearing in
each of the three subsets alone, a predictor appearing in each of the three possible pairs
of subsets, and a predictor that appears in all three; of note, we are also allowing an in-
tercept in this formulation. We generate \( O_i \) from a multivariate normal distribution
with the following correlations: (1) a common correlation of 0.20 between covariates measured
at the same time, (2) a common correlation of 0.30 within each variable over time, and (3)
a common correlation of 0.10 between different covariates measured at different times. Fur-
ther, suppose the treatments \( A_{it} \) and outcomes \( Y_{it} \) are generated according to the following
mechanism:

\[
\begin{bmatrix}
Y_i \\
A_i
\end{bmatrix} = \begin{bmatrix}
X_i \beta + W_i \eta \\
Z_i \alpha
\end{bmatrix} + \begin{bmatrix}
\epsilon_i \\
\gamma_i
\end{bmatrix}; \quad \begin{bmatrix}
\epsilon_i \\
\gamma_i
\end{bmatrix} \sim \mathcal{N} \left( 0, \begin{bmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{T12} & \Sigma_{22}
\end{bmatrix} \right). \tag{7}
\]

Here, \( \Sigma_{11} \) is an exchangeable covariance matrix having outcome error variance of \( \sigma_Y^2 = 1.0 \)
and a correlation of \( \rho_Y = 0.60 \) between outcome errors over time; \( \Sigma_{12} \) includes a correlation
of \( \rho = 0.50 \) between the latent treatment error \( \gamma_{it} \) and the concurrent outcome error \( \epsilon_{it} \)
on the diagonal entries and a correlation of \( \rho_{AY} = 0.20 \) between the two errors at different
times on off-diagonal entries; \( \Sigma_{22} \) contains diagonal entries of \( \sigma_A^2 = 1 \) for the treatment
error variance, off-diagonal entries of \( \rho_A = 0.50 \), signifying the correlation in treatment
errors over time. We set \( \beta = \alpha = (0, 1, 1, 1, 1)^T \), and we set \( \eta = (0, 0.20, 0.20, 0.20, 0.20)^T \)
in this simulation.

As an initial comparator method, we use GEE with a working independence corre-
lation structure, adjusted for treatment (in Section 4.5, we will compare the LEM to an
instrumental variable approach). We conduct one-thousand simulation replicates under this
setup, extracting the corresponding point estimates, robust cluster-based standard errors,
and indicator of coverage based on symmetric 95% Wald-based confidence intervals at each
replicate. Tables 1 and 2 depict the results of this study under respective sample sizes of
\( N = 50 \) and \( N = 500 \).

Table 1: Simulation study results under constant cluster size (panel data) with \( N = 50 \).
Presented are the mean estimates across simulation replicates, the empirical standard
error (ESE), the average estimated standard error (\( \hat{SE} \)), and the coverage probability (CP) for each of \( \beta_0 \) through \( \beta_4 \).

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>LEM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 = 0 )</td>
<td>0.01</td>
<td>0.018</td>
</tr>
<tr>
<td>( \beta_1 = 1 )</td>
<td>1.00</td>
<td>0.097</td>
</tr>
<tr>
<td>( \beta_2 = 1 )</td>
<td>1.00</td>
<td>0.121</td>
</tr>
<tr>
<td>( \beta_3 = 1 )</td>
<td>1.00</td>
<td>0.131</td>
</tr>
<tr>
<td>( \beta_4 = 1 )</td>
<td>1.00</td>
<td>0.160</td>
</tr>
</tbody>
</table>

Based on the results of this study, we observe that the LEM achieves low bias under this
setup under both a smaller and larger sample size. The estimated standard errors appear
Table 2: Simulation study results under constant cluster size (panel data) with $N = 500$. Presented are the mean estimates across simulation replicates, the empirical standard error (ESE), the average estimated standard error ($\hat{SE}$), and the coverage probability (CP) for each of $\beta_0$ through $\beta_4$.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>LEM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>ESE</td>
</tr>
<tr>
<td>$\beta_0 = 0$</td>
<td>0.00</td>
<td>0.067</td>
</tr>
<tr>
<td>$\beta_1 = 1$</td>
<td>1.00</td>
<td>0.030</td>
</tr>
<tr>
<td>$\beta_2 = 1$</td>
<td>1.00</td>
<td>0.036</td>
</tr>
<tr>
<td>$\beta_3 = 1$</td>
<td>1.00</td>
<td>0.041</td>
</tr>
<tr>
<td>$\beta_4 = 1$</td>
<td>1.00</td>
<td>0.048</td>
</tr>
</tbody>
</table>

to be smaller than the empirical standard errors under a smaller sample size; this finding is not surprising given the large number of parameters estimated by the LEM. Under a larger sample size, the standard errors adequately represent the true repeat-sample variability of the estimates across Monte Carlo iterations, with coverage close to the desired 95%. Estimates produced by the GEE approach are not, in general, unbiased for $\beta$, though the robust standard errors appear to capture the repeat-sample variability well. We make note that GEE achieves low bias for estimation of $\beta_1$. Recall that the variable $O_1$ is associated with neither the treatment assignment nor the effect of the treatment, and so endogeneity does not pose a challenge to estimation of $\beta_1$ using standard approaches. The relative loss of efficiency associated with the LEM is not surprising given that it involves estimation of a substantially larger number of parameters.

4.2 Variable cluster size

4.2.1 Completely random cluster size

We conduct a simulation study under the same general setup as that of Section 4.1, with completely random variation in subject-specific observation times. In particular, we introduced a $1/3$ probability that a subject’s second observation was missing, and a $1/2$ probability a subject’s third observation was missing. Conclusions from this simulation study are analogous to those of Section 4.1, and so we do not supply them in detail.

4.2.2 Exposure-dependent cluster size

We conduct a simulation study under the same general setup as that of Section 4.1, with another example of variable observation times, $T_i$. The indicator of missingness for subject $i$ at time $t$ was generated to depend upon covariates as follows:

$$M_{it} \sim \text{Bernoulli} \left( p = \expit \left( -1 + 0.2 \times \sum_{j=1}^{7} O_{ij} \right) \right).$$  (8)
This results in an overall missingness rate of approximately 30%. As expected, conclusions from this simulation study are analogous to those of Section 4.1, and so we do not supply them in detail.

### 4.2.3 Outcome-dependent cluster size

Of interest is to evaluate the performance of the longitudinal endogeneity model under a missingness mechanism that depends upon the outcome. We conduct a simulation study under the same general setup as that of Section 4.1, with the indicator of missingness or subject $i$ at time $t$ generated as:

$$M_{it} \sim \text{Bernoulli}(p = 0.1 \times 1(Y_{it} \leq -1) + 0.4 \times 1(-1 < Y_{it} \leq 2) + 0.7 \times 1(Y_{it} \leq 2)).$$  \hspace{1cm} (9)

Results from this simulation study are depicted in Table 3. As expected, the LEM produces estimates that are biased; in settings of informative cluster size, the LEM cannot be expected to produce consistent estimates of $\beta$.

Table 3: Simulation study results under outcome-dependent cluster size. Presented are the mean estimates across simulation replicates, the empirical standard error (ESE), the average estimated standard error ($\hat{SE}$), and the coverage probability (CP) for each of $\beta_0$ through $\beta_4$.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>LEM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0 = 0$</td>
<td>-0.14</td>
<td>-0.51</td>
</tr>
<tr>
<td>$\beta_1 = 1$</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>$\beta_2 = 1$</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>$\beta_3 = 1$</td>
<td>0.97</td>
<td>0.89</td>
</tr>
<tr>
<td>$\beta_4 = 1$</td>
<td>0.97</td>
<td>0.82</td>
</tr>
</tbody>
</table>

### 4.3 Performance under a misspecified error distribution

In this section, we illustrate the performance of the LEM under a moderate perturbation to the assumption of bivariate normal errors. We conduct a simulation study under the same general setup as that of Section 4.1, instead with skewed and heavy-tailed error terms:

$$\begin{bmatrix} \epsilon_{it}^* \\ \gamma_{it}^* \end{bmatrix} \sim \mathcal{N} \left( 0, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \right) + \begin{bmatrix} \epsilon_{it} \\ \gamma_{it} \end{bmatrix},$$  \hspace{1cm} (10)

where $\epsilon_{it}^*, \gamma_{it}^* \sim \text{Exponential}(\lambda = 1/3) + t(df = 3) - 3$ is represented by the sum of an exponential distribution and a $t$-distribution with three degrees of freedom.

The results of this study are presented in Table 4. Though the LEM appears to show low bias in this setting, the standard errors do not appear, on average, to adequately reflect the true repeat-sample variability for either $X_4$ or $X_7$, both of which are also predictors of treatment. We further found that the nonparametric bootstrap produced standard errors that matched the empirical variability more closely.
Table 4: Simulation study results under a misspecified error distribution. Presented are the mean estimates across simulation replicates, the empirical standard error (ESE), the average estimated standard error (\(\hat{SE}\)), and the coverage probability (CP) for each of \(\beta_0\) through \(\beta_4\).

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>ESE</th>
<th>(\hat{SE})</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_0 = 0)</td>
<td>0.26</td>
<td>0.530</td>
<td>0.396</td>
<td>0.91</td>
</tr>
<tr>
<td>(\beta_1 = 1)</td>
<td>1.00</td>
<td>0.099</td>
<td>0.010</td>
<td>0.95</td>
</tr>
<tr>
<td>(\beta_2 = 1)</td>
<td>1.00</td>
<td>0.173</td>
<td>0.145</td>
<td>0.93</td>
</tr>
<tr>
<td>(\beta_3 = 1)</td>
<td>1.00</td>
<td>0.133</td>
<td>0.134</td>
<td>0.93</td>
</tr>
<tr>
<td>(\beta_4 = 1)</td>
<td>1.00</td>
<td>0.195</td>
<td>0.172</td>
<td>0.95</td>
</tr>
</tbody>
</table>

4.4 Performance in the absence of an exclusion restriction

In this section, we demonstrate that the LEM has proper finite-sample performance in the absence of an exclusion restriction (i.e., when there is no instrument). We therefore conduct a simulation study under the same general setup as that of Section 4.1, with the exception that \(\alpha_1\) is set to zero (recall that \(O_2\) had served the role of what would traditionally be described as an instrumental variable, although the existence of such a variable is not one of our identification assumptions).

Table 5: Simulation study results in the absence of an exclusion restriction. Presented are the mean estimates across simulation replicates, the empirical standard error (ESE), the average estimated standard error (\(\hat{SE}\)), and the coverage probability (CP) for each of \(\beta_0\) through \(\beta_4\).

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>ESE</th>
<th>(\hat{SE})</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_0 = 0)</td>
<td>0.00</td>
<td>0.076</td>
<td>0.073</td>
<td>0.94</td>
</tr>
<tr>
<td>(\beta_1 = 1)</td>
<td>1.00</td>
<td>0.030</td>
<td>0.030</td>
<td>0.95</td>
</tr>
<tr>
<td>(\beta_2 = 1)</td>
<td>1.00</td>
<td>0.040</td>
<td>0.039</td>
<td>0.95</td>
</tr>
<tr>
<td>(\beta_3 = 1)</td>
<td>1.00</td>
<td>0.041</td>
<td>0.040</td>
<td>0.95</td>
</tr>
<tr>
<td>(\beta_4 = 1)</td>
<td>1.00</td>
<td>0.051</td>
<td>0.049</td>
<td>0.94</td>
</tr>
</tbody>
</table>

The results of this study are depicted in Table 5. As expected, the LEM achieves the desired finite sample properties, similar to those depicted in Section 4.1, even in the absence of an exclusion restriction.

4.5 Bias and efficiency relative to instrumental variable approach

In this section, we seek to compare the LEM to a longitudinal instrumental variables model. In the interest of a fair comparison, we consider the setting in which treatment effect homogeneity is satisfied. In turn, we let \(X_{it} = (1, O_{it1}, O_{it2}, O_{it3})\) and \(Z_{it} = (1, O_{it2}, O_{it3})\), with \(\beta = (0, 1, 1, \beta_3)^T\) (for \(\beta_3\) varying between zero and one), and \(\alpha = (0, 1, 1)^T\). When \(\beta_3 = 0\), the exclusion restriction is met, with \(O_{it3}\) serving the role of the instrument.
When $\beta_3 \neq 0$, the exclusion restriction is violated, but the identifying assumptions of the LEM are still met. Letting $X_i = (1^{T}, \overbar{\Gamma}_1^{T}, \overbar{\Gamma}_2^{T}, \overbar{\Gamma}_3^{T})$ and $P_i = (1^{T}, \overbar{\Gamma}_1^{T}, \overbar{\Gamma}_2^{T}, \overbar{\Gamma}_3^{T})$, the just-identified longitudinal instrumental variable estimator is the solution to the following estimating equation:

$$\sum_{i=1}^{N} X_i^{T} (Y_i - P_i \beta) = 0,$$

and has the closed-form expression (Wooldridge, 1983):

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^{N} X_i^{T} P_i\right)^{-1} \left(\sum_{i=1}^{N} X_i^{T} Y_i\right).$$

The results of this simulation are depicted in Figure 3. We observe in this study that when the exclusion restriction is met, the instrumental variable method can provide efficiency gains over the LEM. The level of bias increases sharply with departures from the exclusion restriction (most notably for the variable associated with treatment), while the bias of the LEM remains low. The level of variability associated with the LEM remains constant across variations in $\beta_3$, whereas it appears to grow for the instrumental variable method.

5. Age and LDL: The Multi-Ethnic Study of Atherosclerosis

The Multi-Ethnic Study of Atherosclerosis (MESA) is a longitudinal cohort study of 6,814 men and women from six U.S. communities. At the time of first observation, participants ranged from 45 to 84 years of age. The demographic breakdown of the study is as follows: 47% male, 38% white, 28% African-American, 22% Hispanic, and 12% Chinese-American. This study was designed to provide insights into the prevalence and progression of subclinical cardiovascular disease. We use data from the first five exams of MESA, collected over ten years. All subjects provided written informed consent. Further details regarding sampling, recruitment, and data collection are reported elsewhere (Bild et al., 2002). Of note, 16.1% of study participants reported use of at least one lipid-lowering drug at the time of their first observation. The study prevalence of lipid-lowering medication use increased during the period of time over which subjects were observed, with 39.0% of the 4,464 subjects observed for a fifth visit having reported use of at least one lipid-lowering drug at the time of their fifth observation. Participants were aged 45 to 84 years at the baseline observation; by the time of the fifth observation, the maximum age was 94 years.

We use these data to illustrate the utility of the LEM in order to estimate and compare natural history age trends in LDL by gender. Treatment was considered as the use of any lipid-lowering medications. In the outcome model of the LEM, we included gender, a natural cubic spline for age with knots at 55, 70, and 85 years, and an interaction between the two. The medication use (probit) model included age category, gender, race category, diabetes status, health insurance status, and the Framingham risk score. Treatment effect magnitude was permitted to vary with continuous age, gender, and lipid drug type, categorized as follows: 0 = statin; 1 = other (e.g., fibrates, niacin, bile acid resins).
Figure 3: Depiction of bias and variability of the LEM and instrumental variable (IV) approach for estimation of $\beta_1$ (corresponding to a variable associated with outcome only) and $\beta_2$ (corresponding to a variable associated with both the outcome and the treatment).

We include results from three comparator models in this illustration. In the first, we apply the treatment effects model (with covariate-specific treatment effects) to data from the first observation—i.e., so that we may apply the treatment effects model suitable for cross-sectional data. We further estimate the parameters of two separate GEE models with a working independence correlation structure: one that adjusts for medication use (0 = none; 1 = statins, 2 = other), and one that excludes all observations in which a subject was treated at the time of measurement.

Figure 4 presents the predicted age trends by gender and corresponding point-wise 95% confidence intervals for each of the four approaches. In comparing the results of the cross-sectional and longitudinal models, we note little difference in the overall age-trends by
Unsurprisingly, the confidence bands for the LEM are narrower as compared to the cross-sectional counterpart. This is consistent with the idea that both models are targeting the same parameter (namely the natural history trend that would be observed in the absence of medication use). The LEM has an advantage over the cross-sectional method in being able to estimate the natural history trend more efficiently by making use of additional observations, as evidenced by narrower confidence bands. That the cross-sectional model displays very wide confidence intervals for larger values of age is a product of the fact that there were fewer data points at older ages; of note, we only show the predicted results through 84 years in the cross-sectional model, the maximum baseline age. Each model
correcting for endogeneity displays graphical evidence of a decrease in off-treatment LDL at older ages. This trend would be consistent with poorer liver function associated with lower age.

The GEE models represent trends in the observed biomarker. The overall trends based on the GEE model suggest a much sharper downward trend as compared to the models that account for endogeneity. In the exclusion-based model, this is easily explained by a problem of selection bias, in which those with a higher underlying LDL value are systematically excluded from the analysis. That treatment adjustment does little to correct this is attributable to endogeneity of treatment; the adjusted model conditions on observed treatment status but does not handle the correlation between treatment and the error term. The GEE models also tend to suggest much stronger evidence of a difference in mean observed LDL between genders as compared to the models for the natural history trend. Interestingly, this result is consistent with a prior study that concluded that mean LDL shows substantives decreases with age (Ferrara et al., 1997), in which the study investigators adjusted for medication use in both cross-sectional and longitudinal regression models.

6. Discussion

Endogenous medication use acts as a contaminant when seeking to estimate marginal associations between predictors of interest and biomarkers in longitudinal data. A subject’s natural history is distorted by the effects of the medication on the biomarker. Since medication users differ from non-users in their expected underlying off-medication biomarker values, and since the predictor of interest is often associated with higher medication user prevalence, naïve approaches to account for this distortion are not appropriate. In this manuscript, we have derived and presented an approach for estimation of trends in the natural history of a biomarker using longitudinal data. The primary contribution of the proposed approach is that while it preserves the parametric components of its cross-sectional counterpart, no further parametric structure is introduced when applied to the longitudinal setting. The R package, endogenous, is freely available on CRAN and will accommodate clustered data. We have also added to the body of literature comparing parametric approaches to their more commonly used alternatives including instrumental variable methods. Our findings illustrate that in the presence of a valid instrument, efficiency gains can be made over the LEM, though our findings are also consistent with previous critiques of instrumental variable approaches in observational data (Hernán and Robins, 2006).

Utilizing a working independence model, as we have proposed, is a means of extending the cross-sectional treatment effects model to accommodate clustering while bypassing computational difficulties associated with full specification of a longitudinal covariance matrix. In our application, we were able to demonstrate substantial efficiency gains by making use of repeated measures on subjects. In addition to computational convenience, the working independence approach as proposed in our work offers an additional advantage over full covariance specification. In particular, non-independence approaches are known to suffer from the challenge of requiring the “full-covariate conditional mean” assumption to hold, in which the mean model for an outcome at a particular time must hold conditional on the entire covariate history rather than merely the concurrent covariates (Diggle et al., 2002; Pepe and Anderson, 1994). Although this assumption may hold for certain primordial ex-
posures that are time-stable and deterministic in nature, factors that influence medication use are typically iterative and time-dependent in nature, such that this assumption could not be guaranteed. Of note, our proposed methodology can be employed when seeking to evaluate associations between biomarkers and predictors such as race/ethnicity, genetic exposures, or chronic illnesses; in such settings, more structured proposals for longitudinal correlation structures may be desirable, although further work would be required to determine conditions under which longitudinal correlation parameters are identifiable.

We point out that we are not offering a blanket criticism of GEE for estimation of associations. Our purpose in using GEE as a comparator method in simulations and in our application to MESA was to underscore the idea that simple regression models, commonly applied in this setting, do not provide consistent estimates of the natural history trend in the specific setting of endogenous medication use.

There are several possible directions for future work. For settings in which cluster size may be informative, Hoffman et al. (2001) propose a within-cluster resampling approach in GEE models; it would be of interest to evaluate whether this approach can be generalized to the LEM in order to reduce bias arising from informative missingness. Some recent studies of biomarker trends have focused on latent class growth curves via finite mixture modeling (Loucks et al., 2011; de Groot et al., 2014; Allen et al., 2014). In settings of endogenous treatment, it is all but certain that finite mixture modeling approaches will fail to capture latent underlying natural history curves; therefore, a generalization of our framework to this setting could also be of interest.

Acknowledgments

This work was supported by R01-HL-103729-01A1. MESA is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169 and CTSA UL1-RR-024156. The authors thank the other investigators, the staff, and MESA participants for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.
References


