Inferring causal effects from longitudinal repeated measures data has high relevance to a number of areas of research, including economics, social sciences and epidemiology. In observational studies in particular, the treatment receipt mechanism is typically not under the control of the investigator; it can depend on various factors, including the outcome of interest. This results in differential selection into treatment levels, and can lead to selection bias when standard routines such as least squares regression are used to estimate causal effects.

Interestingly, both the characterization of and methodology for handling selection bias can differ substantially by disciplinary tradition. In social sciences and economics, instrumental variables (IV) is the standard method for estimating linear and nonlinear models in which the error term may be correlated with an observed covariate. When such correlation is not ruled out, the covariate is called endogenous and least squares estimates of the covariate effect are typically biased. The availability of an instrumental variable can be used to reduce or eliminate the bias.

In public health and clinical medicine (e.g., epidemiology and biostatistics), selection bias is typically viewed in terms of confounders, and the prevailing methods are geared toward making proper adjustments via explicit use of observed confounders (e.g., stratification, standardization). A class of methods known as inverse probability weighting (IPW) estimators, which relies on modeling selection in terms of confounders, is gaining in popularity for making such adjustments.

Our objective is to review and compare IPW and IV for estimating causal treatment effects from longitudinal data, where the treatment may vary with time. We accomplish this by defining the causal estimands in terms of a linear stochastic model of potential outcomes (counterfactuals). Our comparison includes a review of terminology typically used in discussions of causal inference (e.g., confounding, endogeneity); a review of assumptions required to identify causal effects and their implications for estimation and interpretation; description of estimation via inverse weighting and instrumental variables; and a comparative analysis of data from a longitudinal cohort study of HIV-infected women. In our discussion of assumptions and estimation routines, we try to emphasize sufficient conditions needed to implement relatively standard analyses that can essentially be formulated as regression models. In that sense this review is geared toward the quantitative practitioner.

The objective of the data analysis is to estimate the causal (therapeutic) effect of receiving combination antiviral therapy on longitudinal CD4 cell counts, where receipt of therapy varies with time and depends on CD4 count and other covariates. Assumptions are reviewed in context, and resulting inferences are compared. The analysis illustrates the importance of considering the existence of unmeasured confounding and of checking for ‘weak instruments.’ It also suggests that IV methodology may have a role in longitudinal cohort studies where potential instrumental variables are available.
1 Introduction

1.1 The causal inference conundrum

The idea of drawing causal inferences has a long and vibrant history in several fields, including economics, philosophy, statistics, computer science and epidemiology. Causal estimands can sometimes be deceptively simple to define; for example, in the evaluation of an intervention such as worker retraining, the analyst may be interested in the difference in wages for a population that received the intervention versus the same population under the scenario that the intervention was not received. Posing the problem in this way makes it clear that we are interested in the effect of job training on wages, all other things being held equal. In biomedical applications, it is frequently of interest to determine if a specific therapy or drug is efficacious for treatment or prevention of disease. In studies of HIV and AIDS, CD4 cell count is an important marker of immune system function and disease stage. To evaluate a new antiviral therapy regimen, researchers may be interested to know the difference in CD4 cell count for an individual who receives the new therapy, versus if that same individual did not receive therapy. This is a well-defined, if unobservable, quantity.

The potential outcomes (counterfactuals) framework, which can be traced to Neyman\(^1\) and has been formalized by Rubin\(^2,3\) and Robins\(^4,5\) provides a coherent structure for defining causal effects, characterizing selection biases, and articulating assumptions that allow for identification of causal effects from observed data. (See Pearl\(^6,7\), Dawid\(^8\) and Holland\(^9\) and references therein, for wide-ranging discussion of other frameworks for causal inferences.)

As a simple illustration in the context of HIV, let \(a \in \{0, 1\}\) index receipt of an antiviral treatment (1 if received, 0 if not); then each person has two potential CD4 counts, \(Y(1)\) under receipt of treatment, and \(Y(0)\) under nonreceipt. The causal effect of receiving treatment on CD4 count is \(Y(1) - Y(0)\), a within-individual contrast. The average causal effect is \(E[Y(1) - Y(0)]\), the average of causal effects across the population. Writing the potential outcomes in terms of a linear model, we have

\[
E[Y(a)] = \alpha + \beta a
\]

where \(\beta\) is the average causal effect of receiving treatment (i.e., of changing \(a\) from 0 to 1). Model (1) is a simple type of marginal structural model\(^10,11\) ‘marginal’ because it represents the marginal mean of \(Y(a)\), and ‘structural’ because it is written in terms of potential outcomes, as opposed to observed data.

The difficulty with inferring \(\beta\) from observed data is that only one of the potential outcomes can be observed. In particular, for binary treatment, the observed data on an individual consists of \((A, Y)\), where \(A \in \{0, 1\}\) is the realized treatment and \(Y = AY(1) + (1 - A)Y(0)\) is the observed response. Clearly, only one potential outcome can be observed (the other is ‘counterfactual’), and therefore inferences for structural models such as (1) can be viewed as missing data problems.

Structural models must be distinguished from regression (or association) models such as

\[
E(Y|A) = \alpha^* + \beta^* A
\]
in which the right-hand side is the conditional expectation or regression function. The regression parameter $\beta$ is not, in general, equal to the causal parameter $\beta$ because selection to receipt of treatment is typically nonrandom.

An immediate consequence is that estimation of $\beta$ via ordinary least squares (OLS) under the assumed model (2) will generally yield inconsistent estimates of the causal parameter $\beta$. In this paper, we review two moment-based estimation methods, inverse probability weighting (IPW) and two-stage least squares using instrumental variables (IV). IPW is motivated by an epidemiologic approach to nonrandom selection based on adjustment for confounding variables that may explain the treatment selection process. The IV approach assumes there may be unmeasured or unobserved confounders, and relies on the use of an instrument or randomizer that can be used to identify causal treatment effects. Both approaches can be formulated in terms of weighted least squares estimation and, under some very specific assumptions, each can be used to derive consistent estimation of causal parameters, even for longitudinal data with time-varying treatments.

### 1.2 Motivating example: a longitudinal HIV natural history study

The HIV Epidemiologic Research Study (HERS) is an observational cohort study of the natural history of HIV in women, in which 871 HIV-infected women and another 439 HIV-negative women were followed prospectively for up to seven years. Data are collected at four sites, one each in Baltimore, New York, Detroit and Providence; two of these (Detroit and Providence) also provide primary HIV care to the patients enrolled in the study. Study visits occurred every six months, at which time a number of clinical, behavioral and demographic variables were collected.

Enrolment to HERS began in 1993 and continued through 1995, during which time the standard antiviral treatment regimen consisted of a single medication such as protease inhibitor. Subsequently, multiple treatment regimens known as highly active antiviral therapy (HAART) were found to be much more effective, and gained widespread use around 1996. HAART is now the recommended therapy for many with HIV, in particular for those with CD4 cell count below 500.

For the analysis in Section 6, we use both IPW and IV to estimate the causal effect of exposure to HAART, which varies with time, on longitudinal CD4 cell count, an important surrogate marker for disease progression. As virtually no HIV-infected persons had access to HAART prior to 1996, we use data on the 357 women whose CD4 was less than 500 at visit 7 (three years post enrolment), and estimate treatment effects for the two-year period from visits 8 to 11.

A host of potential covariates is available, including clinical markers such as plasma viral load (HIV-1 RNA levels) and HIV symptomatology, demographic variables such as age and race, behavioral characteristics such as intravenous drug use history, and other personal characteristics such as the time since becoming aware of one’s HIV infection. Two variables – an indicator of whether primary care is provided at the study site, and the lag time from opening of the study to individual enrolment – will be used as instruments. Justification is provided in the data analysis.
The use of HAART is dictated primarily by HIV disease stage as measured by CD4 and viral load; in particular, even among those with CD4 less than 500, prevalence of HAART usage is much greater among those with CD4 less than 200. Furthermore, a recent analysis of these data suggests that the causal effect of HAART is heterogeneous across initial CD4 cell count. Accordingly, our analysis in Section 6 is stratified by CD4 count at visit 7 (>200 versus ≤200). Table 1 depicts three important processes: mean CD4 cell count among women still in follow-up, percentage on HAART, and percentage still in follow-up at each visit. Clearly, the percentage on HAART is increasing with study visit, which reflects the steady penetration of HAART into standard clinical treatment of HIV-infected individuals. Figure 1 shows the joint CD4/HAART process for 20 randomly selected subjects, illustrating that HAART use has considerable variation both within and between subjects.

1.3 Structural models for longitudinal data

To address the longitudinal nature of the HER Study, we must generalize the potential outcome framework. Continuing with the case where treatment is binary, and assuming responses are recorded at common time points \( t = 1, \ldots, T \), the set of potential treatments is now defined in terms of treatment histories \( \tilde{a}_t = \{a_1, \ldots, a_t\} \), where \( a_t \in A_t \) and \( A_t = \{0, 1\}^{\otimes t} \) is the set of all possible \( t \)-sequences of 0s and 1s. Model (1) can now be expanded as

\[
E[Y_t(\tilde{a}_t)] = \alpha_t + \beta g(\tilde{a}_t)
\]

where \( g \) is a known function of treatment history, such as the current exposure or cumulative exposure. If the objective is to model causal effect conditional on covariates \( V_t = (V_{1t}, \ldots, V_{pt}) \) (these may or may not overlap with potential confounders), then the covariates can be included via \( \alpha_t = V_t \theta \), where \( \theta = (\theta_1, \ldots, \theta_p)^T \) is a vector of coefficients. In general we will suppress reference to covariates.

The implied observed data model is

\[
E(Y_t | \tilde{A}_t) = \alpha_t^* + \beta^* g(\tilde{A}_t)
\]

Table 1 Summary statistics by visit

<table>
<thead>
<tr>
<th>Visit 7 CD4</th>
<th>Statistic</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–200</td>
<td>n</td>
<td>124</td>
<td>116</td>
<td>108</td>
<td>101</td>
</tr>
<tr>
<td></td>
<td>% on HAART</td>
<td>0.35</td>
<td>0.39</td>
<td>0.45</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>CD4 mean (SD)</td>
<td>133 (99)</td>
<td>147 (103)</td>
<td>149 (121)</td>
<td>169 (143)</td>
</tr>
<tr>
<td>200–500</td>
<td>n</td>
<td>233</td>
<td>222</td>
<td>209</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>% on HAART</td>
<td>0.29</td>
<td>0.35</td>
<td>0.42</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>CD4 mean (SD)</td>
<td>359 (124)</td>
<td>367 (157)</td>
<td>376 (182)</td>
<td>370 (196)</td>
</tr>
</tbody>
</table>

SD, standard deviation.
where $\bar{A}_t = \{A_1, \ldots, A_t\}$ is the observed treatment history at $t$ and

$$Y_t = \sum_{\bar{a}_t \in A_t} I(\bar{A}_t = \bar{a}_t)Y_t(\bar{a}_t)$$

As in the cross-sectional case, the parameter $\beta^*$ does not correspond to the causal effect $\beta$. The longitudinal case is more complicated because selection to treatment may be a function of time-dependent stochastic covariates, including previous responses $Y_1, \ldots, Y_{t-1}$.

### 1.4 Perspectives on estimating a causal effect

Regardless of how one characterizes the problem, it is well understood across disciplines that standard regression methods applied to observed treatment and response data cannot generally be used to infer causal effects, particularly for longitudinal data. The first method we describe, IPW, is designed to adjust for confounders that can be
observed. Other approaches to adjusting for observed confounders include popular methods such as direct and indirect standardization, stratification procedures such as Mantel–Haenszel estimators, matching at the design stage, and propensity score methods. IPW estimation of structural models distinguishes itself because it has obvious generalizations to longitudinal data. The IPW approach views confounding as a mechanism leading to nonrandom selection from the population of potential outcomes, which provides a natural motivation for the use of inverse weighting – commonly used in sample surveys – as a method to correct selection bias.

An economist or social scientist may view methods such as IPW with suspicion because the assumption that all confounders have been observed may seem unrealistic; however, many epidemiologic studies differ from those in social sciences in that the collection of candidate confounders is an integral part of study design. By contrast, important research questions in economics and the social sciences are usually addressed by analysing data that have been collected or are maintained by government agencies or survey organizations (e.g., Current Population Survey, Medicaid data, etc.). The databases serve as important resources for investigating a wide variety of issues, but the variables are not typically selected for a specific research agenda. Consequently, econometric methods for causal inference are predicated on the existence of at least one and possibly several unmeasured confounding variables; therefore, confounding is essentially viewed as an omitted variables problem that leads to correlation between errors and covariates (endogeneity).

Under some specific assumptions, IV methods can be used to mitigate this problem; succintly, an instrument is a variable that is associated with receiving treatment, but conditional on treatment is uncorrelated with outcome. Hence any marginal association between instrument and outcome is wholly explained by receiving treatment. Random assignment to active treatment in a clinical trial is a variable that meets these assumptions, and is an instrument by definition. In observational studies, however, it may be less obvious and must be justified on substantive rather than empirical grounds. Causal inference via IV can be carried out using a variety of techniques; to maintain the theme of moment-based estimators, we focus in Section 5 on two-stage least squares (2SLS), which can be viewed more generally as generalized method of moment (GMM) estimators.

Just as economists may view the ‘no unmeasured confounders’ assumption as unrealistic, statisticians and epidemiologists tend to be wary of methods that purport to adjust for variables that cannot be directly observed. Indeed, both IPW and IV estimation rely on assumptions that cannot be empirically verified. For IPW we require that all confounders are observed, and in IV the instrument must be uncorrelated with an unobservable error term.

1.5 Organization

The objective of this paper is to describe, apply and compare IPW and IV estimation of the linear structural model (3) for longitudinal data with time-varying treatment. Working with a single underlying model allows us to describe assumptions from each method in terms of a common notation and underlying structure.

In Sections 2 and 3, we provide a detailed description of our working structural model (3) and its implied observed-data association model. Section 2 sets out the
structural model in detail (for purposes of clarity, our focus is on average causal effects on an additive scale, but we fully recognize the existence of more general characterizations of potential outcomes distributions; see references 22 and 23 for example). Section 3 draws distinctions between structural models and regression models, reviews the concept of ignorability, which is crucial to framing assumptions for identification of structural models, and describes the connection between confounding and endogeneity.

Sections 4 and 5, respectively, describe identification and estimation of the causal parameters in (3) via IPW and IV; Section 4 concentrates on assumptions regarding observed confounders, and Section 5 is primarily concerned with assumptions related to the instrumental variables. We show that both methods can be represented as weighted least squares estimators, with different choices of weights. In Section 6, we apply each method to the HIV data, characterizing assumptions in context and providing comparisons of estimated effects. Finally, Section 7 concludes with a summary and brief discussion.

In the course of describing each methodology and applying them to real data, we attempt to clarify, to the extent possible, terms and concepts that may be inconsistently understood across disciplines; these include instrumental variable, confounding, endogeneity, and causal effect. We also hope to make a case that carefully applied econometric methods may have important applications in public health research, even in epidemiologic studies.

2 Linear potential outcomes model for causal effect of a binary time-varying treatment

This section elaborates the construction and interpretation of (3), a linear structural model of causal effect of a time-varying treatment, wherein treatment history acts additively on potential outcomes and each potential outcome has its own associated error term. Versions of this model have been studied in a variety of contexts. Various causal parameters can be defined in terms of this model, including average causal effect (ACE), average causal effect among the treated, average causal effect among compliers, and local average treatment effect (see Heckman and Chapter 18 of Wooldridge for accessible reviews). As our analysis of the HIV data restricts attention to women with initial CD4 counts less than 500, and because those who meet this criterion are potentially eligible to receive HAART, our focus throughout is on ACE.

2.1 Potential outcomes

We define potential outcomes as random variables corresponding to outcomes under different treatments applied to the same individual; naturally it is possible to observe only one outcome, so the observed outcome is one element from the set of potential outcomes. We confine attention to binary (yes/no) treatments and their accumulated histories, although many of the ideas discussed here are readily generalized to allow multilevel treatments.
Elaborating the notation of Section 1, we assume that \( n \) individuals are scheduled to have measurements taken at a common set of time points \( t = 1, \ldots, T \). For \( i = 1, \ldots, n \), let \( A_{it} \) denote treatment received at \( t \), such that

\[
A_{it} = \begin{cases} 
1 & \text{if treatment received} \\
0 & \text{otherwise}
\end{cases}
\]

Define the history of treatment received as \( \bar{A}_{it} = \{A_{i1}, \ldots, A_{iT}\} \), and let \( \mathcal{A}_i \) denote the set of all possible treatment combinations. Now let \( a_t \in [0, 1] \) denote a possible realization of \( A_{it} \), and \( \bar{a}_t \in \mathcal{A}_t \) a possible realization of \( \bar{A}_{it} \). Then, at each measurement occasion, every subject has \( 2^T \) potential outcomes, denoted by the set \( \{Y_{it}(\bar{a}_t): \bar{a}_t \in \mathcal{A}_t\} \). Each element in this set corresponds to a unique treatment history up to and including time \( t \). The only potential outcome that can be observed corresponds to the actual treatment history \( \bar{A}_{it} \),

\[
Y_{it} = \sum_{\bar{a}_t \in \mathcal{A}_t} I(\bar{A}_{it} = \bar{a}_t)Y_{it}(\bar{a}_t),
\]

where \( I(\mathcal{E}) = 1 \) if event \( \mathcal{E} \) occurs and equals zero otherwise. Equation (5) connects potential outcomes to the observed data, and is known as the consistency relation.\(^{26,37}\) It also captures implicitly the stable unit treatment value assumption (SUTVA) because we are assuming responses of individual \( i \) can be affected only by treatment received by individual \( i \).\(^{2,21,38}\)

### 2.2 Linear structural model

For cross-sectional sampling, average causal effect is the average difference, across the population, of individual-specific differences in potential outcomes under \( a = 1 \) and \( a = 0 \), \( E(Y_i(1) - Y_i(0)) \). In longitudinal data, average causal effects can be defined as contrasts in the mean of potential outcomes for different treatment histories, such as \( E(Y_{it}(\bar{a}_t') - Y_{it}(\bar{a}_t)) \), for \( \bar{a}_t' \neq \bar{a}_t \).

Causal effects can be represented using a structural model defined in terms of potential outcomes, as follows. Let \( X_{it}(\bar{a}_t) \) denote a \( 1 \times L \) design matrix for modeling treatment effects, let \( \theta \) represent an \( L \times 1 \) vector of parameters, and assume \( Y_{it}(\bar{a}_t) \) is related to \( \bar{a}_t \) via

\[
Y_{it}(\bar{a}_t) = X_{it}(\bar{a}_t)\theta + \varepsilon_{it}(\bar{a}_t)
\]

where \( E(\varepsilon_{it}(\bar{a}_t)) = 0 \) for all \( i, t, \) and \( \bar{a}_t \). Pairs \( \{\varepsilon_{it}(\bar{a}_t), \varepsilon_{it}(\bar{a}_t')\} \) (for \( \bar{a}_t \neq \bar{a}_t' \)) and \( \{\varepsilon_{it}(\bar{a}_t), \varepsilon_{it'}(\bar{a}_t')\} \) (for \( t \neq t' \)) may be dependent, thereby allowing within-subject correlation between potential outcomes at fixed times and across times. No distributional assumptions (e.g. normality) about the error terms are made. Model (6) elaborates (3) by including the error terms and generalizes it slightly by using \( X_{it}(\bar{a}_t) \).
Model (6) can be appreciated by considering specialized cases. The simplest is a cross-sectional version \( T^\hat{1} \) and hence \( \cdot a \) and \( t^\hat{a} \), wherein the causal effect is a contrast of means; that is, \( X_i(a) = [1, a] \) and \( \theta = (\alpha, \beta)^T \). In this setting, (6) becomes

\[
Y_i(a) = \alpha + \beta a + \epsilon_i(a) \tag{7}
\]

where \( \beta \) is the average causal effect of receiving treatment. Under (7), where each potential outcome has an associated mean-zero error term, causal effects are heterogeneous across individuals, such that the individual-specific causal effect is

\[
Y_i(1) - Y_i(0) = \beta + [\epsilon_i(1) - \epsilon_i(0)], \tag{8}
\]

the average causal effect plus a mean-zero individual-specific effect.

Models such as (6) and (7) are frequently termed \textit{structural} because they derive from a theory of how the world might work if we were allowed to vary \( a \) arbitrarily (i.e. they are written in terms of \( a \) and not \( A_i \)). Importantly, neither is a regression model: referring to (7), if we were to replace \( a \) with \( A_i \), there is no assumption that \( E(\epsilon_i | A_i) = 0 \), from which it follows in general that \( E(Y_i | A_i) \neq \alpha + \beta A_i \). From this point of view, the error terms in (6) and (7) contain variables that have an effect on the observed response \( Y_i \) but whose effects are either ignored or held constant in the theorizing that leads to (7). These variables, some of which may be unobservable, can be viewed as confounders in an epidemiologic context.

We note also that (7) can be viewed as a special case of the ‘Rubin Causal Model’, which is more general because it assumes potential treatment outcomes \( \{A_i(r): r \in \mathcal{R}\} \), where \( r \) is the realization of a randomly assigned variable with sample space \( \mathcal{R} \). \cite{angrist96} provide a detailed study of IV estimators under this model.

Turning back to the longitudinal case, for the purposes of comparing methods and interpreting data analyses, we assume that cumulative treatment history can be summarized using a scalar via the function \( g: A_t \rightarrow \mathbb{R} \), and that no covariates other than visit are of interest. This yields our working model

\[
Y_{it}(\tilde{a}_t) = \alpha_t + \beta g(\tilde{a}_t) + \varepsilon_{it}(a_t) \tag{9}
\]

We reiterate that the key feature of (9) that is most germane to comparisons between IV and IPW is the lack of distributional assumptions on the potential outcomes errors \( \varepsilon_{it}(a_t) \). Identification and estimation of model parameters for both procedures requires unverifiable assumptions about these error terms; Sections 4 and 5 are concerned primarily with describing these.

### 3 Implications for observed data

Data on an individual selected randomly from the population of interest consist of

\[
\{(Y_{it}, A_{it}, W_{it}): t = 1, \ldots, T\}
\]
where, for measurement occasion \( t \), \( A_{it} \) is treatment received, \( Y_{it} \) is the observed outcome (5) – the potential outcome corresponding to treatment history \( \tilde{A}_{it} \) – and \( W_{it} = (W_{1it}, \ldots, W_{Qit}) \) is a \( 1 \times Q \) vector of covariates that may include confounders, instrumental variables, or both (definitions for these terms are given in subsequent sections). Note that \( X_{it} \) from the structural model is a subset of the available covariates \( W_{it} \).

### 3.1 Structural models versus regression models

To fix ideas, we return to the cross-sectional structural model (7). Contrast (7) with the regression model

\[
E(Y_i | A_i) = \alpha^* + \beta^* A_i
\]

written in terms of observables. When the observed data are generated by the structural model (7), then

\[
Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0)
\]

\[
= (\alpha + \varepsilon_i(0)) + A_i(\beta + \varepsilon_i(1) - \varepsilon_i(0))
\]

and

\[
\beta^* = E(Y_i | A_i = 1) - E(Y_i | A_i = 0)
\]

\[
= \beta + E(\varepsilon_i(0) | A_i = 1) - E(\varepsilon_i(0) | A_i = 0) + E(\varepsilon_i(1) - \varepsilon_i(0) | A_i = 1)
\]

The regression slope, \( \beta^* \), differs from the average causal effect \( \beta \) by the generally nonzero collection of expectations on the right-hand side. From a data analysis point of view, the immediate consequence is that consistent or unbiased estimates of \( \beta^* \) will in general be biased for the causal parameter \( \beta \), by an amount equal to the added terms in (12). Hence for the cross-sectional case at least, standard estimators such as a contrast in sample means between those who receive and do not receive treatment, or equivalently the OLS estimator of the coefficient of \( A_i \) in a regression of \( Y_i \) on \( A_i \) – which are in fact unbiased estimators of \( \beta^* \) – are biased estimators of the causal contrast \( \beta \). The bias is attributable to differential selection for receipt of treatment based on the errors \( \varepsilon_i(0) \) and \( \varepsilon_i(1) \); for example,

i) selection bias attributable to dependence between \( A_i \) and \( \varepsilon_i(0) \), characterized (e.g.) by sicker patients preferentially receiving treatment, manifest as \( E(\varepsilon_i(0) | A_i = 1) \neq E(\varepsilon_i(0) | A_i = 0) \); and

ii) selection bias attributable to dependence between \( A_i \) and the individual-specific gain \( \varepsilon_i(1) - \varepsilon_i(0) \), characterized (e.g.) by good responders preferentially receiving treatment, and manifest as \( E(\varepsilon_i(1) - \varepsilon_i(0) | A_i) \neq 0 \). In general, the bias will be nonzero when \( A_i \) depends either on \( \varepsilon_i(0) \) or \( \varepsilon_i(1) \).

### 3.2 Ignorability and sequential ignorability

Identification of causal effects from observed data is not possible without unverifiable assumptions. Primary among these is ignorability of treatment receipt.\(^{3,19}\) In what
follows, we argue that although the causes of selection bias may take different characterizations (e.g., endogeneity, confounding), they all trace back to violations of the following assumption.

**Assumption 1 (Ignorability of treatment receipt)** Treatment receipt is ignorable if $A_i$ is marginally independent of both $e_i(0)$ and $e_i(1)$, i.e., if $A_i \perp e_i(0)$ and $A_i \perp e_i(1)$.

(It is possible to distinguish weak ignorability (our assumption here) from strong ignorability, under which $A_i$ is jointly independent of the errors, i.e., $A_i \perp \{e_i(0), e_i(1)\}$. In the present paper, our use of the term ‘ignorability’ implicitly refers to weak ignorability.)

If data are generated by (11), and if treatment receipt is ignorable, then $\beta^*$ is equivalent to the causal parameter $\beta$; this follows directly by applying Assumption 1 to (12) (in the linear model, a sufficient condition for $\beta = \beta^*$ is independence between $A_i$ and $e_i(0) + A_i(e_i(1) - e_i(0))$; see Heckman and Angrist et al. for further discussion).

The notion of ignorability can be extended to the longitudinal data setting. Consider our working model (9) with $g(\tilde{a}_t) = a_t$; the implied observed-data model is

$$ Y_{it} = A_{it} Y_{it}(1) + (1 - A_{it}) Y_{it}(0) $$

$$ = \{x_t + e_{it}(0)\} + A_{it} \{\beta + e_{it}(1) - e_{it}(0)\} $$

$$ = x_t + \beta A_{it} + v_{it} $$ (13)

where $v_{it} = e_{it}(0) + A_{it} \{e_{it}(1) - e_{it}(0)\}$. Contrasting means based on treatment receipt at $t$ yields

$$ E(Y_{it} | A_{it} = 1) - E(Y_{it} | A_{it} = 0) = \beta + E[e_{it}(0) | A_{it} = 1] - E[e_{it}(0) | A_{it} = 0] $$

$$ + E[e_{it}(1) - e_{it}(0) | A_{it} = 1] $$

Similar to the cross-sectional case, the contrast in means between those receiving and not receiving treatment at time $t$ does not in general correspond to the causal parameter $\beta$, motivating a sequential version of the ignorability assumption that amounts to ignorability at each time point.

**Assumption 2 (Sequential ignorability)** Receipt of treatment at time $t$ is sequentially ignorable if, for $t = 1, \ldots, T$ and for each $\tilde{a}_t \in A$, $A_{it} \perp e_{it}(\tilde{a}_t) | A_{i,t-1}$

Assumption 2 also can be interpreted to mean that treatment is ‘sequentially randomized’ because it implies treatment received at $t$ is independent of the potential outcomes, given treatment history.

In contrast to the cross-sectional case, Assumption 2 is unlikely to hold even in a randomized trial, except in the unusual circumstance that each participant fully complies with assigned treatment for the duration of the study. In longitudinal studies, decisions about whether to administer treatment at time $t$ typically depend on responses and covariates observed prior to $t$, which in turn may be correlated with the potential
outcomes at $t$. For example, suppose treatment is antiviral therapy, and outcome is CD4 cell count. Physicians prescribe therapy for time period $t$ based on their clinical assessment of the patient at time $t-1$. It is not difficult to imagine that physicians are more likely to prescribe therapy to untreated patients who are observed to be more adversely affected by HIV (based on $W_{it}$). To the extent that $W_{it}$ is correlated with $e_{it}(0)$, the differential assignment in treatment results in dependence between $A_{it}$ and $e_{it}(0)$. If physicians are treating sicker patients on average, then $E\{e_{it}(0) \mid A_{it} = 1\} < E\{e_{it}(0) \mid A_{it} = 0\}$, thus potentially violating Assumption 2. Under this scenario, (time-dependent) variables excluded from the model but which contribute to the error term (whether measured or unmeasured) lead to treatment endogeneity, and can be viewed as time-dependent confounders.$^{29,41}$

### 3.3 Endogeneity and confounding

Scientists in public health and epidemiologic research tend to characterize violations of ignorability in terms of confounding, whereas those in social sciences and econometrics refer to endogenous treatments. Making this distinction in viewpoints is important to the extent that they motivate prevailing modes of data analysis in the primary and related fields. Very frequently, epidemiologic studies are designed so that relevant confounders will be collected in advance, and prevailing data analytic methods operate under the assumption that these confounders are available for use in a model or analysis. Methods such as matching, stratification, standardization, and regression adjustments are all designed to make use of confounding variables, usually (but not always) by conditioning on them. By contrast, quantitative studies in econometrics, social sciences, and policy research rely on datasets whose construction may not have had a specific research purpose; hence, the analyst frequently does not have the luxury of having relevant confounders in hand. Prevailing data-analytic methods (consisting primarily of IV methods) are geared toward addressing endogeneity without the benefit of all measured confounders.

**Endogeneity** A covariate $A$ in the linear model $Y = \alpha + \beta A + \nu$ is said to be endogenous if $\text{cov}(A, \nu) \neq 0$, and exogenous otherwise.$^{34}$ Endogeneity can arise from a number of sources, including differential selection (e.g., violations of the ignorability assumption) and covariate measurement error. The observed-data model (11) provides an example: although $Y_i$ is linear in the causal parameters $\alpha$ and $\beta$, that is

$$Y_i = \alpha + \beta A_i + [e_i(0) + A_i(e_i(1) - e_i(0))]$$

$$= \alpha + \beta A_i + v_i$$

(14)

the treatment covariate $A_i$ is endogenous unless it is independent of the (complex) error term $v_i$. One such condition that induces independence is ignorability (Assumption 1). Hence, when data are generated by additive potential outcomes models, ignorability of treatment receipt implies that observed treatment $A_i$ is exogenous in the corresponding observed-data model. For the longitudinal case, sequential ignorability (Assumption 2) implies exogeneity.

**Confounding.** Confounding is a term whose meaning seems to be understood in general terms, but whose precise definition is rather elusive. A comprehensive and
general discussion of this issue can be found in Greenland et al., Stone, and most recently Geng et al. The definition of ‘confounder’ found in many epidemiology texts makes reference to causal effects but not explicitly to potential outcomes. This passage from Kelsey et al. is representative:

A confounder is a variable that is (a) causally related to the disease or condition under study, [conditional on] the exposure of primary interest . . . , and (b) is associated with the exposure under study in the population, but is not a consequence of this exposure.

It turns out that this condition is sufficient but not necessary for confounding to exist. Because most studies typically have a number of candidate confounders, it is somewhat more intuitive to define sufficient control for confounding. For individual i, let $C_i$ denote a set of candidate confounders. Confounding can be sufficiently controlled if treatment receipt is conditionally ignorable, given $C_i$:

**Assumption 3 (Ignorability given confounders)** There exists a set of confounders $C_i$ such that for all $i$, both $A_i \perp e_i(0) \mid C_i$ and $A_i \perp e_i(1) \mid C_i$.

If all components of $C_i$ are observed, then Assumption 3 is equivalent to the ‘no unmeasured confounders’ assumption.

When data are longitudinal and treatments may be time-varying, the set of candidate confounders may include any (possibly unobserved) covariate process histories up to the instant prior to time $t$, and in particular may include the histories $Y_{i, t-1} = \{Y_{ij}: j = 1, \ldots, t - 1\}$ of the outcomes process and $A_{i, t-1} = \{A_{ij}: j = 1, \ldots, t - 1\}$ of the treatment process. Denote this time-varying set of potential confounders as $C_{it}$. The longitudinal generalization of Assumption 3 is given in Assumption 4.

**Assumption 4 (Sequential ignorability given confounders)** For $i = 1, \ldots, n$ and for $t = 1, \ldots, T$, there exists a set $C_{it}$ of time-varying confounders such that $A_{it} \perp e_{it}(\tilde{a}_t) \mid (A_{it-1}, C_{it})$ for each $\tilde{a}_t \in A_t$.

Maintaining this assumption means that for individuals with a common history $(C_{it}, A_{i,t-1})$, receipt of treatment is sequentially randomized; when $C_{it}$ is fully observed, it corresponds to the longitudinal version of ‘no unmeasured confounders’.

Naturally there may exist more than one confounder, and Assumptions 3 and 4 may hold for more than one set of potential confounders; see reference 23 for a discussion of minimally sufficient sets of confounders.

**Controlling the effects of observed confounders.** Depending on the objectives of inference, there exist many methods for making adjustments for confounding. The inferential objectives can be divided into conditional, where interest is in causal effects conditional on $C_{it}$, and marginal, where interest is in the causal effect averaged over $C_{it}$. Commonly used methods for estimating conditional effects include stratification on confounders (e.g., Mantel–Haenszel estimator) and regression model adjustment, for example including confounders as covariates in an observed data model such as (11), and treating it as a regression. An important consideration when using regression adjustment is proper specification of the model, which can be difficult when there exist
several potential confounders; one solution is to summarize covariate information using a propensity score (Rosenbaum and Rubin\textsuperscript{19}), which then can be used as a single covariate (propensity scores can also be used to estimate unconditional effects\textsuperscript{19,45}).

When interest is in the marginal effect of treatment, the methodologies can become computationally more complex, particularly for longitudinal data. Recent developments for longitudinal data include the G-computation algorithm\textsuperscript{44} wherein treatment effects are estimated within levels of the confounder under Assumption 4 and then averaged over the distribution of the confounder, and inverse probability weighted estimates of marginal structural models such as (9), wherein confounders are used to build a treatment selection model and individual observations are weighted by the inverse probability of realizing their observed treatment history $A_n$.\textsuperscript{10,11} Propensity scores can be used to find marginal causal effects,\textsuperscript{18,19,45} but the generalization to time-varying treatments is not obvious. Readers are referred to Robins\textsuperscript{46} for a review of several methods in the context of equivalence trials with noncompliance.

4 Inverse probability weighting

The IPW method is a logical extension of inverse weighting methods used in survey sampling\textsuperscript{47,48} and in missing data problems.\textsuperscript{49–51} In the sample survey context, interest lies in drawing inference about a population $\mathcal{P}$. The sample used to draw inferences is not a simple random sample (SRS), but rather one in which members of certain subpopulations of $\mathcal{P}$ are over- or under-sampled (e.g., ethnic subpopulations). In using standard analyses such as regression modeling, each unit is weighted relative to its inverse probability of being sampled (in an SRS of size $n$, the sampling probability is $1/n$, so the relative weights are equal to 1). The weights themselves can be interpreted to quantify the number of nonsampled members of the population that are being represented by the sampled unit; for example, if the weight for an observed unit is $1/4$, then this unit’s data represents information from four members of the population $\mathcal{P}$.

The problem of estimating parameters from a population model of counterfactuals can be cast in much the same way. The hypothetical sample of counterfactuals is assumed to be an SRS from the population of interest, but the observed portion of these is realized nonrandomly. To illustrate, consider cross-sectional sampling with binary treatment. The (hypothetical) SRS of counterfactuals is the set of pairs

$$\mathcal{Y} = \{Y_i(0), Y_i(1): i = 1, \ldots, n\}$$

Selection to treatment or no treatment causes only one member of each pair to be observed; the selection mechanism maps the SRS $\mathcal{Y}$ to the observed data

$$\mathcal{Y} = \{Y_i: i = 1, \ldots, n\} = \{Y_i(A_i): i = 1, \ldots, n\}$$

hence, $\mathcal{Y}$ can be viewed as a nonrandom sample from $\mathcal{Y}$. This representation motivates a strategy for parameter estimation based on weighting observed-data contributions $I(A_i = a)Y_i(a)$ inversely by $\text{pr}(A_i = a)$. 
Referring now to our working model (9), the objective is to infer the parameters from observed longitudinal response data. IPW estimation is typically described in the context of marginal structural models (MSM), which technically are specified only in terms of the marginal mean $E[Y_{it}(\tilde{A}_i)]$ of a structural model such as (9). As the errors in (9) have mean zero, the associated MSM is

$$E[Y_{it}(\tilde{A}_i)] = \alpha_t + \beta g(\tilde{A}_i)$$

(15)

To motivate the estimation procedure, consider first the problem of estimating parameters from (15) from a hypothetical sample that is the longitudinal analogue of $Y$. As the errors have zero mean, and because measurement times are exogenous (in our example they are fixed by design), parameters from (15) would be consistently estimated by OLS as follows. Denote the $T + 1$ parameters in (15) by $\theta = (\alpha_1, \ldots, \alpha_T, \beta)^T$, let $X_{it}(\tilde{A}_i)$ represent the corresponding $1 \times (T + 1)$ covariate vector for occasion $t$ (i.e. the design matrix), and let $\mu_{it}(\tilde{A}_i, \theta) = E[Y_{it}(\tilde{A}_i)] = X_{it}(\tilde{A}_i)\theta$. Then the $(T + 1) \times 1$ system of (OLS) estimating equations $U_0(\theta) = 0$, where

$$U_0(\theta) = \sum_{i=1}^n \sum_{t=1}^T \sum_{\tilde{A}_i \in A_i} X_{it}(\tilde{A}_i)^T [Y_{it}(\tilde{A}_i) - \mu_{it}(\tilde{A}_i, \theta)]$$

(16)

is unbiased, which implies that its solution is a consistent estimator of $\theta$.

Because $Y'$ is a (nonrandom) subsample of $Y$, it is natural to consider weighted estimating equations for inferring model (15) from the observables. Consider the estimating equations $U_1(\theta) = 0$, where

$$U_1(\theta) = \sum_{i=1}^n \sum_{t=1}^T \sum_{\tilde{A}_i \in A_i} I(\tilde{A}_{it} = \tilde{A}_i) D_{it} X_{it}(\tilde{A}_i)^T [Y_{it}(\tilde{A}_i) - \mu_{it}(\tilde{A}_i, \theta)]$$

(17)

or equivalently,

$$U_1(\theta) = \sum_{i=1}^n \sum_{t=1}^T D_{it} X_{it}(\tilde{A}_{it})^T [Y_{it} - X_{it}(\tilde{A}_{it})\theta]$$

(18)

where $D_{it}$ is a scalar weight. When $D_{it} = 1$, solving $U_1(\theta) = 0$ is simply OLS applied to the observed data marginal model $E(Y_{it} | A_{it}) = \alpha_t + \beta g(A_{it})$, evident from (18). The more general form (17), written explicitly in terms of potential outcomes, is useful for motivating IPW estimation of the causal parameters under various assumptions about ignorability.

Consider first the simple case where $D_{it} = 1$ for all $i$ and $t$ (OLS applied to observed data). Under sequential ignorability (Assumption 2), the solution to $U_1(\theta) = 0$ is consistent for the causal parameters $\theta$ because each contribution to $U_1(\theta)$ has mean zero at the true value of $\theta$ (see Hernán et al. for further details).

If, as is more likely in applied problems, sequential ignorability does not hold, then using $D_{it} = 1$ and $U_1(\theta) = 0$ is no longer unbiased at the true $\theta$. Suppose, however, that
sufficient information from confounding variables is available such that Assumption 4 (sequential ignorability conditional on confounders) can be made. Further suppose that at each time \( t \), each subject has nonzero probability of receiving each treatment history \( \tilde{A}_t \in A_t \), and that the selection probabilities

\[
\lambda_{it}(a_t) = \Pr(A_{it} = a_t | \tilde{C}_{it}, \tilde{A}_{i,t-1}), \quad t = 1, \ldots, T
\]

are a known function of the confounders.

Following the logic of sample survey methodology, the IPW strategy is to weight each contribution to (17) by its inverse probability of being observed, or equivalently, by the inverse of

\[
\pi_{it}^{*}(\tilde{a}_t) = \Pr(\tilde{A}_{it} = \tilde{a}_t | \tilde{C}_{it}, \tilde{A}_{i,t-1})
\]

\[
= \prod_{j=1}^{t} A_{it} \lambda_{it} + (1 - A_{it})(1 - \lambda_{it})
\]

If \( D_{it} = \{\pi_{it}^{*}(\tilde{a}_t)\}^{-1} \) in (17), then it can be shown that \( U_i(\theta) = 0 \) is an unbiased estimating equation for the causal parameters \( \theta \), and hence its solution is consistent for \( \theta \).

In practice, of course, it is unusual that \( \pi_{it}^{*}(\tilde{a}_t) \) will be known, so it must be estimated. A reasonably straightforward approach is to model the probabilities \( \lambda_{it} \) using a regression for binary data, for example, logistic regression or probit regression. As parameter estimation and not standard error estimation is the chief objective, the \( A_{it} \)'s can be treated as independent within subject. If the model is correctly specified and \( \hat{\lambda}_{it} \) consistently estimates \( \lambda_{it} \), then replacing \( \lambda_{it} \) by \( \hat{\lambda}_{it} \) (20) yields consistent estimates of the sampling probabilities. Robins et al. show that the estimating equations remain unbiased with probability approaching one when \( \pi_{it}^{*}(\tilde{a}_t) \) is replaced with an \( n^{1/4} \)-consistent estimator, and in fact the resulting estimates of \( \theta \) are more efficient compared to using known \( \pi_{it}^{*} \). It is also possible to use stabilized weights to further increase efficiency.

To summarize, IPW is motivated by viewing the observed outcomes as a nonrandom sample from the set of potential outcomes, within an individual. If Assumption 4 holds and the probability of selection to observed treatment history \( \tilde{A}_{it} \) is known or can be consistently estimated as a known function of confounders, then the inverse of these probabilities can be used to reweight contributions to the OLS estimating equations (17). The solution to the weighted estimating equations is consistent for the causal parameters in (9).

5 Instrumental variables

Although the use of IV dates back at least to Riersol, research on IV methods actively continues. Important reviews can be found in Angrist et al., Heckman, Heckman and Robb, and Wooldridge. Rather than providing a detailed treatment of the many assumptions and scenarios under which IVs can be used to infer causal effects, our purpose here is to describe a set of reasonable (but certainly not minimal) assumptions that lead to estimators that relate directly to regression techniques such as OLS. The key consideration in application of IV methods is to evaluate whether the assumptions
about an instrument will hold for a given application; particularly when treatment effects may vary across individuals, this evaluation is necessarily qualitative rather than quantitative, because the key assumptions are not verifiable using observed data.

In this section, as in the previous one, we assume that the objective is to infer causal parameters from the structural model (9). To fix ideas, we begin by considering the cross-sectional case with binary treatment, and generalize to the longitudinal case.

To parallel the IPW approach, we consider moment-based estimators, and in particular two-stage least squares (2SLS).

5.1 Cross-sectional data with binary treatment

We begin again with the cross-sectional structural model (7). As illustrated in Section 3.3, replacing treatment index \( a \) by observed treatment \( A_i \) gives rise to the model (14) in which \( A_i \) is an endogenous covariate because of its potential correlation with the error term \( e_i(0) + A_i[e_i(1) - e_i(0)] \).

To fix key ideas, we first consider the widely used specialization of (7) where \( e_i(0) = e_i(1) = e_i \), known as the endogenous treatment model; its defining characteristic is that treatment effects are assumed homogeneous across the population. We then consider the more general (and perhaps more realistic) heterogeneous treatment model for cross-sectional data, and finally the heterogeneous treatment model for longitudinal data.

Causal inference using IVs relies on a system of equations in which parameters of the first equation correspond to those in the underlying structural model; the second equation characterizes probability of receiving treatment. The key in forming the system of equations is to identify an instrumental variable (or variables) \( R_i \), which has the property of being associated with treatment but conditionally on treatment, not with the potential outcomes (we formalize this below). Hence an instrument is a variable whose effect on observables \( Y \) can be exerted only through treatment.

Homogeneous treatment effects (endogenous treatment model). Consider the structural model

\[
Y_i(a) = \alpha + \beta a + e_i \tag{21}
\]

where \( E(e_i) = 0 \). Suppose the observed data comprise \( \{Y_i, A_i, R_i: i = 1, \ldots, n\} \), where \( Y_i \) and \( A_i \) are as before, and \( R_i \) is an instrumental variable that satisfies the following condition.

**Definition 1** With respect to Model (21), \( R_i \) is an instrumental variable if

i) \( R_i \) is independent of \( e_i \); 
ii) \( E(A_i|R_i = r) \) is a nonconstant function of \( r \).

Under the assumption that \( R_i \) is an instrument, the causal parameter \( \beta \) can be consistently estimated from the following system of equations that characterize – on the inessential hypothesis that the dependence of treatment on instrument satisfies a linear model – the mechanism giving rise to observed data,

\[
Y_i = \alpha + \beta A_i + e_i \\
A_i = \gamma_0 + \gamma_1 R_i + \delta_i \tag{22}
\]
Here, $\delta_i$ is a mean-zero error term and the instrument $R_i$ is exogenous with respect to the entire system (i.e., it is independent of both $\varepsilon_i$ and $\delta_i$). The error term for the second equation is sometimes written as $\delta_i = \gamma_2 \varepsilon_i + \xi_i$, where $\xi_i$ is a mean-zero error term, to emphasize its possible dependence on $\varepsilon_i$.

Substituting the right-hand side of the second equation for $A_i$ in the first yields the reduced form of (22),

$$
Y_i = (x + \beta \gamma_0) + \beta \gamma_1 R_i + [(1 + \beta \gamma_2) \varepsilon_i + \beta \xi_i]
$$

$$
A_i = \gamma_0 + \gamma_1 R_i + \gamma_2 \varepsilon_i + \xi_i
$$

If $R_i$ satisfies Definition 1 and so is uncorrelated with both $\varepsilon_i$ and $\delta_i$ then least squares applied in (23) to each equation separately will give consistent estimates of both $\beta \gamma_1$ and $\gamma_1$, and it follows that a consistent estimate of average causal effect $\beta$ can be formed from their ratio. Such a calculation is an instrumental variable (IV) estimator of the average causal effect. The validity of this procedure clearly requires that $\gamma_1$ is not zero, which is the reason for part (ii) of Definition 1. When $R_i$ is scalar the explicit formula is

$$
\hat{\beta}^{IV} = \frac{\text{cov}(Y_i, R_i)}{\text{cov}(A_i, R_i)}
$$

where $\text{cov}$ denotes sample covariance.\textsuperscript{21} Importantly, consistency of $\hat{\beta}^{IV}$ for this simple model does not require distributional assumptions about the error terms, but it does require that the means are correctly specified in (22). For cases where $A$ is continuous, the estimator $\hat{\beta}^{IV}$ is the ML estimator of $\beta$ under the assumption that the error terms $(1 + \beta \gamma_2) \varepsilon_i + \beta \varepsilon_i$ and $\gamma_2 \varepsilon_i + \varepsilon_i$ are bivariate normal (for binary and discrete treatments, such as we consider here, probit regression formulations are possible for the second equation; for a recent example, see Chib\textsuperscript{55}). Under ignorable treatment (Assumption 1), $A_i \perp \varepsilon_i$ and $A_i$ can therefore serve as its own instrument (it is necessarily correlated with itself), an additional $R_i$ is unnecessary, and $\hat{\beta}^{IV}$ will then reduce to the least squares estimator. The form of $\hat{\beta}^{IV}$ as a ratio of covariances highlights the importance of having an IV that is associated nontrivially with receipt of treatment ($\gamma_1 \neq 0$), because otherwise its denominator will be zero. An instrument that is weakly correlated with treatment is highly susceptible to bias attributable to small departures from part (i) of Definition 1.\textsuperscript{56,57}

As illustrated in our analysis in Section 6, the partial correlation coefficient (partial $R^2$) from the treatment model in (22) provides a useful measure of this association.\textsuperscript{56}

The estimator $\hat{\beta}^{IV}$ can be thought of in several different ways. One is that it derives from doing two least squares calculations in sequence. In the first, least squares is used to get the predicted value of $A_i$ given $R_i$ and in the second this predicted value of $A_i$ is substituted for $A_i$ in the first equation and the estimator of $\beta$ is found by doing least squares regression of $Y_i$ on these predicted values. On this interpretation $\hat{\beta}^{IV}$ is called the two-stage least squares (2SLS) estimator. This point of view suggests a generalization to the case where there are many instruments $- R_i$ is vector valued – possibly together with additional potential measured confounders. First, using OLS, regress (binary) $A_i$ on $R_i$, and obtain its linear predictor $\hat{A}_i$; second, regress $Y_i$ on $\hat{A}_i$ using OLS. The coefficient in the second regression is a consistent estimator of $\beta$, so long as $R_i$ is a valid instrument, and the linear relationships are correctly specified.\textsuperscript{34} This procedure also can be...
represented as a generalized method of moments estimator, which reduces to weighted least squares for a particular choice of weights (see Wooldridge,\textsuperscript{34} Section 8.2); we elaborate this point in our discussion of longitudinal data in Section 5.2.

Another interpretation is to show that the existence of a valid instrument permits the construction of an unbiased estimating equation for $\beta$. Multiply the first (structural) equation in (22) by $R_i - \bar{R}$ and sum. As the product of this term with $e_i$ is zero in expectation, the result is an unbiased estimating equation for $\beta$. Finally, as we have already remarked, $\beta^{IV}$ is a maximum likelihood estimator under bivariate normality for the reduced form errors and this suggests another generalization, different in general from 2SLS, to the case of many instruments. It also suggests a Bayesian treatment of causal inference using instrumental variables.\textsuperscript{25,55,58}

**Heterogeneous treatment effects (switching regression model).** With heterogeneous treatment effects, $e_i(0) \neq e_i(1)$, and the observed data model follows (14), which is sometimes referred to as the ‘switching regression model’\textsuperscript{59} because the error terms may differ depending on whether $A_i$ is zero or one.\textsuperscript{24,60} The apparently simple generalization to heterogeneous treatment effects has important implications for the definition of an IV.

**Definition 2** With respect to model (14), the variable $R_i$ is an instrument if

1. $R_i \perp [e_i(0) + A_i[e_i(1) - e_i(0)]]$,
2. $E(A_i|R_i = r)$ is nonconstant in $r$.

When $R_i$ meets this definition, $\beta$ can be estimated using the same 2SLS outlined above (mean independence is sufficient in part (i) of Definition 2 when the underlying linear stochastic model is additive, but is not sufficient in general\textsuperscript{22,40}). Part (i) of this definition is potentially unnatural to conceptualize in light of part (ii); namely, we need to assume $R_i$ is independent of a term that includes $A_i$, but we also must assume that $R_i$ is correlated with $A_i$. Wooldridge\textsuperscript{34} points out that part (i) is most easily understood to mean $\text{cov}[A_i, e_i(1) - e_i(0)]$ is constant across levels of the instrumental variable.

The validity of IV estimators such as 2SLS for estimating $\beta$ under heterogeneous treatment effects depends solely on the assumptions one is willing to make regarding $R_i$; the estimator itself does not depend on the assumptions, but the interpretation does. From an inferential viewpoint, this lack of distinction occurs because without making distributional assumptions, the data provide no direct information to distinguish homogeneous from heterogeneous treatment effects; in principle, this distinction is possible in a likelihood-based formulation when $e_i(0)$ and $e_i(1)$ are assumed to follow known distributions (e.g., normal).

### 5.2 Heterogeneous treatment model for longitudinal data

We turn now to extensions that allow identification and estimation of the causal parameter $\beta$ in the longitudinal linear stochastic model (9). The observed-data model associated with (9) can be written

$$Y_i = \sum_{a_i \in \mathcal{A}_i} I(\tilde{A}_{it} = \tilde{a}_t)[x_t + g(\tilde{a}_t)\beta + \varepsilon_{it}(\tilde{a}_t)]$$

$$= x_t + g(\tilde{A}_{it})\beta + \sum_{a_i \in \mathcal{A}_i} I(\tilde{A}_{it} = \tilde{a}_t)\varepsilon_{it}(\tilde{a}_t)$$

(24)
where the last term (summation over $A_t$) is the error term. The covariate $g(\tilde{A}_{it})$ is endogenous if it is correlated with the error term (or more generally if $\tilde{A}_{it}$ correlated with the error term). Moment-based IV estimation requires a more general definition of the instrument.

**Definition 3** With respect to model (24), $R_i$ is an instrumental variable if, for all $t = 1, \ldots, T$,

1. $R_i \perp \sum_{\tilde{a}_i \in A_i} I(A_{it} = \tilde{a}_i)\epsilon_{it}(\tilde{a}_i)$,
2. $\mathbb{E}[g(\tilde{A}_{it})|R_i = r]$ is a nontrivial function of $r$.

Definition 3 requires that $R_i$ is an IV for every time period where a causal effect is to be estimated. In principle, this suggests a generalization of 2SLS to repeated measurements.

To describe this generalization, which – like the IPW estimator – amounts to weighted least squares, we follow the development given by Wooldridge, 34 pp. 190 ff. Note first that (9) can be written in matrix form. Let $x_i(\tilde{a}) = [I(t = 1), \ldots, I(t = T), g(\tilde{a})]$ denote the $1 \times L$ covariate vector needed to parameterize (9); here $L = T + 1$. Then

$$Y_i(\tilde{a}) = X_i(\tilde{a})\theta + \epsilon_i(\tilde{a})$$

where

$$Y_i(\tilde{a})_{(T \times 1)} = [Y_{i1}(\tilde{a}_1), \ldots, Y_{iT}(\tilde{a}_T)]^T$$

$$X_i(\tilde{a})_{(T \times L)} = \begin{bmatrix} x_{i1}(\tilde{a}_1) \\ \vdots \\ x_{iT}(\tilde{a}_T) \end{bmatrix}$$

$$\epsilon_i(\tilde{a})_{(T \times 1)} = [\epsilon_{i1}(\tilde{a}_1), \ldots, \epsilon_{iT}(\tilde{a}_T)]^T$$

$$\theta_{(L \times 1)} = (\alpha_1, \ldots, \alpha_T, \beta)^T$$

Denote the error term in the observed-data model (24) as $\nu_{it}$, and define the $T \times 1$ vector $v_i = (v_{i1}, \ldots, v_{iT})^T$. Then the observed-data model (24) can be written as

$$Y_i = X_i\theta + v_i$$

where $Y_i = Y_i(\tilde{A}_i) = [Y_{i1}(A_{i1}), \ldots, Y_{iT}(A_{iT})]^T$ and $X_i = X_i(\tilde{A}_i)$ are obtained by replacing components of the treatment indices $a_1, \ldots, a_T$ from the structural model with their observed values $A_{i1}, \ldots, A_{iT}$.

Now assume the existence of up to $S$ instrumental variables $R_{i1}, \ldots, R_{iS}$ that meet Definition 3. These are collected in a matrix $Z_i$ as follows. Let $x^*_{it}$ represent the exogenous covariates in $x_{it}$ (i.e., excluding those related to treatment), and let $z_{it} = [x^*_{it}, R_{i1}, \ldots, R_{iS}]$, a $1 \times M$ vector. Then

$$Z_{i(T \times M)} = \begin{bmatrix} z_{i1} \\ \vdots \\ z_{iT} \end{bmatrix}$$
We now make two assumptions about the instrumental variables in terms of \( Z_i \); under the form of (9), where treatment effects act linearly, these will allow identification of a unique estimator of \( \beta \) that takes the form of a weighted least squares estimator.

**Assumption 5**  
\[ E(Z_i^T \nu_i) = 0. \]

**Assumption 6**  
Rank \( \{E(Z_i^T X_i)\} = L. \)

When \( M = L \) (i.e., when the number of IVs is equal to the number of treatment parameters), then Assumption 5 implies that \( Z_i \nu_i = Z_i (Y_i - X_i \theta) \) has mean zero, motivating use of the \( L \times 1 \) system of unbiased estimating equations \( V_0(\theta) = 0 \), where

\[
V_0(\theta) = \sum_{i=1}^{n} Z_i^T (Y_i - X_i \theta)
\]

When \( M > L \), the parameters are over-identified, and the solution to \( V_0 \) is not unique; however, 2SLS for the multivariate model (24) can be cast as a generalized method of moments (GMM) procedure (cf. reference 20). For a positive definite weight matrix \( G \), a GMM estimator minimizes a quadratic form

\[
\left\{ \sum_{i=1}^{n} Z_i(Y_i - X_i \beta) \right\}^T G \left\{ \sum_{i=1}^{n} Z_i(Y_i - X_i \beta) \right\}
\]

over \( \beta \); a solution that gives a consistent estimator is a WLS estimator with weight matrix \( Z_i^T G Z_i \). The 2SLS procedure is obtained by taking

\[
\hat{G}_{(T \times T)} = \left\{ \frac{1}{n} \sum_{i=1}^{n} Z_i^T Z_i \right\}^{-1}
\]

leading to (weighted) estimating equations \( V_1(\theta) = 0 \), where

\[
V_1(\theta) = \sum_{i=1}^{n} X_i^T Z_i^T \hat{G} Z_i (Y_i - X_i \theta)
\]

(this is precisely the estimator implemented in the ‘ivreg’ procedure in Stata Version 7 (reference 61, pp. 136–7)). Thus the weights are given by \( Z_i^T \hat{G} Z_i \), which projects the covariates from the outcomes model (including treatment) onto the covariate space of the treatment model (which includes the instrumental variables). By comparison, the IPW estimator solves

\[
U_1(\theta) = \sum_{i=1}^{n} X_i^T D_i(Y_i - X_i \theta) = 0
\]
where \( D_i \) is a \( T \times T \) diagonal matrix with entries \( (\pi_{it}(\tilde{A}_{it}))^{-1} \) along the diagonal; thus both 2SLS and IPW take the form of weighted least squares estimators, although under very different assumptions.

### 6 Analysis of HIV observational study

#### 6.1 Summaries and exploratory analyses

As indicated in Section 1.2, our analysis uses data from 357 women in the HERS who were still in follow-up at their seventh visit (three years from enrolment), and whose CD4 cell count was less than 500 at that visit. Visit 7 was considered ‘baseline’, and our interest is in estimating the causal effect of HAART during the two-year period spanned by visits 8 through 11. Table 1 summarizes the mean and variation of CD4 cell counts, the proportion of women with data at visit 7 who are still in follow-up, and the proportion on HAART. Both the proportion on HAART and the overall mean CD4 count are increasing over this period. HAART use varies considerably within individual, as seen in Figure 1, which shows individual-level data on CD4 and HAART for 20 randomly selected women.

Our structural model relating potential CD4 count to treatment history \( \tilde{a}_t \) is

\[
Y_{it}(\tilde{a}_t) = \alpha_t + \beta g(\tilde{a}_t) + \varepsilon_{it}(\tilde{a}_t)
\]

for \( t = 8, \ldots, 11 \). The model will be estimated using OLS, IPW and IV (2SLS) under three parameterizations of \( g(\cdot) \):

\[
\begin{align*}
g_1(\tilde{a}_t) & = a_t, \\
g_2(\tilde{a}_t) & = a_{t-1} + a_t, \\
g_3(\tilde{a}_t) & = \sum_{j=7}^{t} a_j
\end{align*}
\]

Under \( g_1 \), only the most recent treatment affects CD4; \( g_2 \) implies that exposure over the past year has an additive effect, regardless of the timing of treatment; and \( g_3 \) allows exposure to be cumulative from visit 7 onwards. Certainly other parameterizations are possible. Under parameterization \( g_1(\tilde{a}_t) = a_t \), the coefficient \( \beta \) is the mean difference in potential CD4 outcomes at time \( t \), \( E\{Y_{it}(1) - Y_{it}(0)\} \), regardless of past treatment history. Using \( g_3(\tilde{a}_t) \), \( \beta \) is the mean difference between potential CD4 outcomes that differ by one unit (one six-month period) on cumulative exposure to HAART. This parameterization captures a cumulative effect of being treated over time, but irrespective of when the exposures occur. Finally, \( g_2(\tilde{a}_t) \) is similar to \( g_3 \) but assumes cumulative effect of HAART depends only on exposures in the past 12 months.

For subject \( i \), the observed data consist of

\[
\{(Y_{it}, A_{it}, C_{it}, R_i): t = 8, \ldots, T_i\}
\]

with \( Y_{it} \) and \( A_{it} \) respectively denoting observed CD4 count and HAART status at time \( t \), \( C_{it} \) consisting of a vector of candidate confounders, some of which may be time-varying, and \( R_i \) consisting of a vector of potential instrumental variables, all of which
are time-stationary. The specific components of $C_t$ are listed in Table 2; the two instrumental variables are $R_{i1}$, a binary indicator of whether the clinician-investigators at a particular site also provided primary HIV care to the women enrolled at that site, and $R_{i2}$, the lag time from opening of enrolment to actual enrolment of woman $i$. Rationale for both IVs is given in Section 6.4.

Drop-out occurs throughout the HER Study, which is why observed data series have length $T_i$. For purposes of this analysis, we assume drop-out is a missing-completely-at-random (MCAR) process in the sense that drop-out at $t$ does not depend on $Y_{it}$. An illustration of the use of inverse weighting for handling drop-out using moment-based estimators of structural models can be found in Hernán et al.; more formal justification related to semiparametric regression is found in Rotnitzky et al. and Scharfstein et al.

### 6.2 Estimation by ordinary least squares

For purposes of comparison with IPW and IV analyses, and to provide a starting point, we first estimate (25) using OLS, by solving $U_0(\theta) = 0$. To adjust for residual correlation between repeated CD4 measures, robust standard errors for all models were computed using the ‘sandwich estimator’ (computations for OLS and IPW estimation were carried out using PROC GENMOD in Version 8.0 of SAS). Provided that (25) is correctly specified (an assumption we maintain throughout), OLS will consistently estimate $\beta$ only under Assumption 2, that HAART is sequentially randomized at each visit.

OLS estimates and associated standard errors for each parameterization of $g$ appear in the third column of Table 3. Under $g_1$ and $g_3$, which emphasize recent receipt of HAART, the estimates suggest a beneficial effect on CD4 count, with slightly more benefit to those with lower CD4 at visit 7.

### 6.3 Estimation by inverse weighting

Parameter estimation under IPW proceeds under Assumption 4, that receipt of treatment is sequentially ignorable given observed confounders. The confounders

---

### Table 2 Covariates used in analysis of HERS data

<table>
<thead>
<tr>
<th>Type</th>
<th>Covariate</th>
<th>Unit of category</th>
<th>Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confounders</td>
<td>HAART status</td>
<td>Yes/no</td>
<td>$t - 1, t - 2$</td>
</tr>
<tr>
<td>ART status</td>
<td>Yes/no</td>
<td></td>
<td>$t - 1$</td>
</tr>
<tr>
<td>AIDS status</td>
<td>Yes/no</td>
<td></td>
<td>$t - 1$</td>
</tr>
<tr>
<td>HIV symptom scale</td>
<td>0–10</td>
<td></td>
<td>$t - 1$</td>
</tr>
<tr>
<td>log CD4</td>
<td>log cells/mm$^3$</td>
<td></td>
<td>$t - 1$</td>
</tr>
<tr>
<td>log HIV RNA</td>
<td>log copies/mm$^3$</td>
<td></td>
<td>$t - 1$</td>
</tr>
<tr>
<td>log CD4 $\times$ log HIV RNA</td>
<td></td>
<td></td>
<td>$t - 1$</td>
</tr>
<tr>
<td>HAART $\times$ log CD4</td>
<td></td>
<td></td>
<td>$t - 1$</td>
</tr>
<tr>
<td>HAART $\times$ log HIV RNA</td>
<td></td>
<td></td>
<td>$t - 1$</td>
</tr>
<tr>
<td>Recent IV drug use</td>
<td>Yes/no</td>
<td></td>
<td>Enrolment only</td>
</tr>
<tr>
<td>Lifetime IV drug use</td>
<td>Yes/no</td>
<td></td>
<td>Enrolment only</td>
</tr>
<tr>
<td>Race</td>
<td>Black, white, other</td>
<td></td>
<td>Enrolment only</td>
</tr>
<tr>
<td>Years aware of HIV status</td>
<td>0, 1–5, 6+ years</td>
<td></td>
<td>Enrolment only</td>
</tr>
<tr>
<td>Instruments</td>
<td>Primary HIV care site</td>
<td>Yes/No</td>
<td>Enrolment only</td>
</tr>
<tr>
<td>Lag time to enrolment</td>
<td>Years</td>
<td></td>
<td>Enrolment only</td>
</tr>
</tbody>
</table>
used in our analysis are drawn from Table 2; our inferences depend on the subjective assumption that some subset of these variables is sufficient to ensure Assumption 4 holds. In words, we need to know that at each visit, within each distinct level of the set of confounders, receipt of treatment is independent of the error terms \( \epsilon_{it}(\tilde{a}_t) \) corresponding to all possible histories of HAART up to \( t \). Another way to conceptualize this assumption is that within every distinct level of the confounder set, and given past treatment history \( A_{it-1} \), treatment is randomly assigned at each visit.

The potential CD4 outcomes \( \{ Y_{it}(\tilde{a}_t) \} \) cannot, of course, be observed by the clinician prior to deciding whether to administer treatment; however, the decision about whether to treat (i.e., whether to set \( A_{it} = 0 \) or \( A_{it} = 1 \)) clearly will be based on both measured and unmeasured individual-level characteristics that are related to the potential CD4 outcomes. For example, a patient’s CD4 count, viral load, AIDS status, and current treatment at visit \( t - 1 \) all will factor into whether treatment is prescribed for the interval ending with visit \( t \). In deciding whether a set of confounders is sufficient for controlling selection bias, the analyst must decide whether the information in the confounders captures all of the systematic variation in the treatment assignment process.

The treatment selection model is assumed to follow a logistic regression that is linear in those confounders thought to be sufficient for control of selection bias (Assumption 2). The time-varying confounders are CD4, viral load, HIV symptom scale, previous HAART status, previous ART status, progression to AIDS (yes/no), and the following interactions: CD4/viral load, CD4/HAART, and viral load/HAART. The main effects are included because treatment guidelines are based on the clinical markers CD4 and plasma viral load (HIV RNA), and because both are highly likely to be correlated with the potential CD4 outcomes at the next visit; the interactions are included because treatment decisions can depend on variations in one of the markers at fixed levels of the other. The time-stationary confounders are CD4 and viral load at enrolment, ART status at enrolment, years aware of HIV status at enrolment, race, and drug use history.

To compute the weights, we first assume that for \( t = 6, \ldots, T_i \),

\[
\logit(\lambda_{it}) = h(\tilde{C}_{it}, \tilde{A}_{i,t-1}; \psi)
\]

where \( \lambda_{it} \) is the selection probability defined in (19) and \( h \) is a known function that linearly relates the covariates to regression parameters \( \psi \). Because a small proportion of women received HAART prior to visit 6, we fixed \( \tilde{A}_{it} = A_{it} \) for \( t = 1, \ldots, 5 \) (this gives weight 1 for visits prior to visit 6). Sampling probabilities are computed using (20),

<table>
<thead>
<tr>
<th>CD4</th>
<th>( g(\tilde{a}_t) )</th>
<th>OLS</th>
<th>IPW</th>
<th>IV1-2SLS</th>
<th>IV2-2SLS</th>
<th>ptl. ( R^2 )</th>
<th>ptl. ( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–200</td>
<td>( a_t )</td>
<td>48 (15)</td>
<td>47 (15)</td>
<td>125 (52)</td>
<td>0.13</td>
<td>131 (52)</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>( \sum_j a_j )</td>
<td>11 (7)</td>
<td>11 (8)</td>
<td>40 (19)</td>
<td>0.14</td>
<td>25 (17)</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>( a_{t-1} + a_t )</td>
<td>33 (10)</td>
<td>33 (11)</td>
<td>65 (28)</td>
<td>0.16</td>
<td>66 (28)</td>
<td>0.16</td>
</tr>
<tr>
<td>200–500</td>
<td>( a_t )</td>
<td>39 (18)</td>
<td>63 (22)</td>
<td>180 (248)</td>
<td>0.01</td>
<td>-14 (133)</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>( \sum_j a_j )</td>
<td>11 (6)</td>
<td>20 (8)</td>
<td>269 (911)</td>
<td>0.0004</td>
<td>-17 (28)</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>( a_{t-1} + a_t )</td>
<td>24 (11)</td>
<td>42 (14)</td>
<td>147 (220)</td>
<td>0.004</td>
<td>-24 (74)</td>
<td>0.03</td>
</tr>
</tbody>
</table>
taking the product from $j = 5$ to $T_i$; weights were computed as $\{\hat{\pi}_{ii}(\tilde{A}_{it})\}^{-1}$. Lack of fit is assessed using Hosmer and Lemeshow’s deciles of risk statistic, $6.5$ which gives 8.0, with $P = 0.56$ relative to $\chi^2$ distribution.

The column labeled ‘IPW’ in Table 3 shows IPW estimates. An upward adjustment of estimated causal effect relative to OLS is evident in the cohort with CD4 200–500, but not among those with CD4 $< 200$. One possible explanation is that treatment with HAART is relatively uniform for less healthy patients, but is more dependent on factors other than CD4 for those with CD4 $\geq 200$. The upward adjustment among those with higher initial CD4 suggests that in this cohort, the effect of confounding is that sicker patients tend preferentially to receive HAART, thereby attenuating estimates that do not adjust for confounding.

One concern in implementing IPW estimation is that the weight distribution may be heavily skewed, thereby assigning undue influence to a small number of observations. The use of stabilized weights $^{31}$ mitigates this problem to a large degree. Figure 2 shows the distribution of stabilized weights at each visit, and does not suggest outsized influence from any individual observation.

### 6.4 Estimation by instrumental variables

Our IV analysis is carried out using the GMM procedure described in Section 5 with two instrumental variables. The first, denoted by $R_{1i}$, is an indicator of whether the

---

**Figure 2** Boxplots of inverse weights from IPW estimation, by visit. Left box for CD4 $< 200$, right box for CD4 200–500
study site also provides primary HIV care. Participants at Providence and Baltimore received their HIV care from HERS clinical investigators, and are more likely to receive HAART; one possible explanation is that the caregivers, in their dual role as HIV researchers, had earlier awareness of the potential benefits of HAART for their HIV-infected patients, for example, due to their attendance at scientific meetings and connection with evolving scientific literature. Whether the study investigators also were caregivers is purely a consequence of geographic location and not of individual patient characteristics, which for this demonstration analysis is assumed to be sufficient to meet part (i) of Definition 3.

The second instrumental variable $R_{2i}$ is lag time from the opening of HERS enrolment to a participant’s enrolment time, and exploits two key features of HERS: i) enrolment accrued over two years (1993–1995), giving rise to staggered entry times relative to the first enrolment; and ii) HAART did not become widely available until 1996. Thus, HAART is available to the earliest enrollees (1993) around their 5th and 6th visits, and for the latest ones (1995) around their 2nd and 3rd visits. For $R_{2i}$ to be a valid instrument, we need Assumption 5 to hold, which means lag time $R_{2i}$ is independent of potential CD4 counts, conditional on visit number and on treatment received; this is equivalent to meeting part (i) of Definition 3 (conditionally on visit number).

**Treatment model.** Although the parameter estimates for (25) are determined by solving $V^2(\hat{y}) = 0$ for the appropriate choice of $Z_i$, it is instructive to examine parameters from the treatment model

$$E(A_{it}) = \gamma_1 R_{1i} + \gamma_2 R_{2i}$$

(26)

because the solution to $V^2(\theta) = 0$ corresponds to the 2SLS solution. (Here, $x_{it}$ consists of the exogenous visit indicators.) The partial $R^2$ associated with the instrumental variables is useful for diagnosing a possibly weak instrument.56

Table 4 summarizes estimates and standard errors for two versions of (26) across the three parameterizations of treatment exposure; the first version includes only care site (hence $\gamma_2 = 0$ by assumption), and the second includes both instruments. The first row ($CD4 < 200$ and $g(A_t) = A_{it}$) indicates that percentage of participant visits with HAART use differs by 35% between ‘care sites’ and non ‘care sites,’ and the standard error

<table>
<thead>
<tr>
<th>Visit 7 CD4</th>
<th>Treatment</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\gamma}_1$ (s.e.)</td>
<td>Partial $R^2$</td>
<td>$\hat{\gamma}_1$ (s.e.)</td>
</tr>
<tr>
<td>0–200</td>
<td>$A_{it}$</td>
<td>0.35 (0.04)</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>$\sum_j A_{ij}$</td>
<td>1.09 (0.13)</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>$A_{it} + A_{it}$</td>
<td>0.66 (0.07)</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>$A_{it}$</td>
<td>0.08 (0.03)</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>$\sum_j A_{ij}$</td>
<td>0.06 (0.10)</td>
<td>0.0004</td>
</tr>
<tr>
<td>200–500</td>
<td>$A_{it}$</td>
<td>0.10 (0.06)</td>
<td>0.004</td>
</tr>
</tbody>
</table>
suggests the difference is highly statistically significant. The partial $R^2$ indicates that ‘care site’ explains about 13% of the variation in HAART use; adding lag time as an instrument does not give a significant effect, nor does it increase variation explained (partial $R^2 = 0.14$). For parameterization $g_2$, however, lag time is significant in Model 2, but the partial $R^2$ is increased only incrementally compared to Model 1 (0.17 versus 0.14).

The treatment model for CD4 200–500 under the same treatment parameterization (row 4 of the Table 4) is an interesting contrast: HAART use at care sites is only about 8% higher (see $\hat{\gamma}_1$); the effect is statistically significant, but partial $R^2 = 0.01$, signaling a very weak instrument. In the second model, both instruments are significantly associated with HAART, but still only weakly explain its variation because the partial $R^2 = 0.02$. The same pattern emerges for the other treatment parameterizations, namely that the first instrument (care site) can be statistically significant without yielding a high $R^2$, and that adding a second instrument that also exhibits significant association may not lead to substantial improvements.

**Structural model.** GMM (or 2SLS) estimates of the structural model based on one and two instruments appear in columns 5–6 and 7–8 of Table 3, respectively. The partial $R^2$ values from corresponding treatment models are transcribed next to the estimated effects. A striking feature is that standard error estimates are closely tied to $R^2$ values; in the CD4 200–500 cohort, standard errors are so high as to render the effect estimates virtually meaningless (cf. row 5 of Table 3). Bias also is a concern: in general, for a single instrument,

$$\hat{\beta}_{2SLS} \rightarrow \beta + \frac{\text{cov}(R, v)}{\text{cov}(R, A)} \text{cov}(R, n)$$

If $R$ is an instrumental variable according to Definition 1, then $\text{cov}(R, v)$ is exactly zero, and the 2SLS estimator is consistent (in theory at least), regardless of whether $\text{cov}(R, A)$ is small, so long as it is nonzero. Weak instruments, however, lead to small values of $\text{cov}(R, A)$, and therefore small departures from part (i) of Definition 1 can lead to large asymptotic bias. This is evidently the case for 2SLS estimators of the HAART effect for those with CD4 200–500.

Turning to the cohort with CD4 < 200, the instruments explain more treatment variability and standard error estimates are far more reasonable. The 2SLS point estimates are greater than their IPW counterparts, suggesting the possibility of unmeasured confounding, an entirely reasonable possibility because i) from a clinical standpoint, anyone with CD4 < 200 is a candidate for HAART, regardless of other observable clinical indicators, and therefore these indicators are not likely to be confounders; ii) in the early stages of HAART availability, before clinical data on efficacy and side effects was widely available to patients and physicians, nonclinical factors may have played a key role in treatment decisions (e.g., patients or physicians may want to hold off on a new treatment for fear of unexpected side effects); therefore patient and/or physician preference may be an important unmeasured confounder. Despite differences in estimated effect, the point estimates from 2SLS are within two standard errors of the IPW estimates, and both analyses suggest that HAART has a beneficial (if imprecisely estimated) therapeutic effect for those with CD4 < 200.
Summary and discussion

This paper has reviewed two moment-based methods for estimating causal treatment effects from longitudinal observational data. Inverse probability weighting is closely aligned with traditional epidemiologic approaches in that it assumes all relevant confounders have been recorded. Instrumental variables is more prevalent in social sciences, where the existence of unmeasured confounding is a standard assumption. We have reviewed the key assumptions underlying each approach, and provided a comparative data analysis for illustration.

The key difference between the methods we have reviewed here is that IPW builds selection weights using the observed confounders, whereas with IV, the need to identify confounders – whether measured or unmeasured – is largely circumvented if an appropriate instrumental variable exists. Inverse probability weighting (IPW) uses observed confounders to estimate treatment selection probabilities corresponding to observed treatment histories $A_{it}$, and the inverses of these are used as observation weights. IPW operates under the ‘no unmeasured confounders’ assumption (Assumption 4), and hence cannot be used directly to handle unmeasured confounding. The method of instrumental variables (IV) exploits the existence of one or more instruments, variables that are associated with receipt of treatment but otherwise uncorrelated with the potential outcomes (in the sense of Definition 3). IV can be used to adjust for unmeasured confounding, but as with the ‘no unmeasured confounders’ assumption required for IPW, the validity of an instrumental variable cannot be empirically verified and must be defended on subject-matter grounds.

Although the analysis of the HIV data is meant to illustrate use and interpretation of the methods, several lessons can be drawn. First, each of the methods is relatively easy to apply. The IPW estimation was carried out using SAS Proc Genmod (see Hernán for an example of relevant code), and the IV analysis was implemented in Stata Version 7 using the `ivreg` command. Secondly, the analysis illustrates that IV can be applied in epidemiologic studies. In our analysis, we exploited the incident nature of HAART to derive an instrumental variable (time to enrolment), and focused our analysis on the two-year period where HAART use increased rapidly in the population. Thirdly, it underscores the importance of evaluating assumptions and their implications in context. For example, among those with CD4 < 200, IPW and OLS estimators of treatment effect are roughly the same, while the IV estimates are quite different. This suggests that observed clinical variables explain very little of the variation in treatment receipt for this subsample of the cohort; however, the IV estimator may be picking up effects of unmeasured confounding, for example due to reluctance of doctors and patients to use a new therapy (HAART) before it has gained widespread popularity. On the other hand, the IPW estimators of treatment effect for those with CD4 200–500 are quite different from the OLS estimator, suggesting that clinical variables beyond CD4 explain variations in treatment. The IV estimators are less useful in this subsample because the instrumental variables are quite weak, making the estimates highly susceptible to biases from violations of part (i) in Definition 3, and leading to very large standard errors.

The ‘no unmeasured confounders’ assumption, although unrealistic on its face, is not un-reasonable in many epidemiologic studies because the collection of confounder data
is usually an integral part of study design. Our HIV example is representative because many if not all of the concrete clinical factors that inform treatment decisions have been recorded. Naturally there exist nonclinical, unmeasured factors such as clinician’s professional judgment and patient preferences, but it is assumed that these effects are negligible compared to the measured clinical variables. Because it is not possible to empirically critique ‘no unmeasured confounders,’ vigorous discussion between the quantitative and clinical members of the scientific team should lead either to rejection of this assumption or a careful justification on qualitative grounds.

Instrumental variables methods such as 2SLS rest on the existence of a valid instrument that explains a substantial portion of variability in treatment. These methods have traditionally been applied to analyse data in settings where the data were not collected for a specific purpose; for example, studies of health care utilization are typically based on databases from the Heath Care Finance Association (HCFA) (see reference 66, for example). It is highly likely that, for a particular research question of interest, even some obvious confounding variables are not included, and the ‘no unmeasured confounders’ assumption is prima facie untenable. Thus, a valid instrument (in the sense that one can make a valid qualitative case) can be a very powerful tool for inference. As illustrated by the HERS analysis, the instrument needs to be both qualitatively plausible and highly correlated with treatment; failure to meet the latter criterion can lead to inconsistent and imprecise estimators.

The reliance on empirically untestable assumptions, inherent in any inference of causal effects, suggests the importance of sensitivity analysis. For IPW estimation, Robins has proposed methods that capture sensitivity to unmeasured confounding in a selection parameter, which is assumed known at a range of plausible values. See Ko et al. for a detailed application. Related to IV methods, Little discusses violations of important structural assumptions such as the linear relationship between error terms, while Bound et al. and Staiger and Stock provide some guidelines for diagnosing and handling weak instruments.

Other approaches to inference about the effects of time-varying treatments include the G-computation algorithm, which like IPW relies on all confounders being measured, and likelihood-based approaches to instrumental variables estimation; see also Hogan and Daniels for a related approach in the context of longitudinal clinical trials with varying duration of treatment. Frangakis and colleagues have recently developed a new method called principal stratification for time-varying treatments; this approach is similar in spirit to IV but unlike (22), does not require additivity. Another important and difficult area of research concerns dynamic treatment regimes, wherein the regime specifies a priori whether and how treatment at t should be modified in response to stochastic variables observed just prior to t. Robins and Murphy et al. provide methods for estimating the causal effect of a dynamic regime, and most recently Murphy develops methodology for finding an optimal dynamic strategy.

**Acknowledgements**

We are grateful to Lytt Gardner and the Centers for Disease Control for allowing use of the HERS data, to Christina Korkontzelou for assistance with data analysis, to Sharon-Lise Normand, Jason Roy, Charles Carpenter, and an anonymous referee for helpful
comments on the manuscript, and to members of the Biostatistics Working Group in HIV/AIDS at Brown University for useful discussions and feedback. We also thank the Editors for the opportunity to contribute to this special issue on longitudinal data. Primary support provided through NIH Grant R01-AI50505, with additional support for the first author from the Lifespan-Tufts-Brown Center for AIDS Research (NIH Grant P30-AI42853). Data from HERS were collected under CDC cooperative agreements U64-CCU106795, U64-CCU206798, U64-CCU306802, and U64-CCU506831.

References


Yau, LHY, Little RJ. Inference for the compiler-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association* 2001; 96: 1232–44.


