Lecture 2: Covariate Adjustment with Time-Varying Policy Variables

Bryan S. Graham, UC - Berkeley & NBER

September 4, 2015

Many econometric evaluation problems involve treatments which may vary over time. Consider a researcher interested in the relationship between measured teacher 'quality' in elementary school and long run life outcomes (e.g., adult earnings). Each year a student is assigned to a new teacher. Different sequences of teacher quality during elementary school may influence adult outcomes differently. For example teacher assignments in kindergarten may be more consequential, than those in a later grades. It may be that consecutive years of assignment to low quality teachers is especially detrimental, while intermittent exposure to poor teachers is not. Assessing hypotheses likes these requires methods for evaluating the effects of *sequences* of treatments. Developing such methods raises new conceptual issues relative to those encountered in the more familiar 'point in time', or static, econometric evaluation problem.

In dynamic settings the econometrician may observe intermediate variables which are (i) influenced by *past* treatment assignments, (ii) influence *future* treatment assignment and (iii) predict the final outcome of interest. Continuing with our example, the econometrician may observe an end-of-grade test score for each student. A student's test score likely varies with the quality of her past teachers, likewise her current performance may influence which type of teacher she is assigned to in the future. Test scores may also help to predict the final outcome of interest. Test score is both a confounder and an intermediate outcome (since it is influenced by prior teacher assignments).

Our intuitions from the static program evaluation problem do not provide directly useful guidance on how to incorporate test scores into an analysis of the effects of different sequences of teacher qualities for, say, adult earnings. A test score has characteristics of a confounder, suggesting that adjustment for it is warranted (Rubin, 1977), as well as of a concomitant variable, suggesting that such adjustment is not advisable (Rosenbaum, 1984). In what

follows I will call variables like test scores time varying confounders. Properly controlling for time varying confounders is not straightforward.

Robins (1986) extends the basic theory of covariate adjustment for static program evaluation problems to dynamic settings. The development in this note will be based on Robins (1997) and Robins (2000). Dynamic methods of covariate adjustment are used with some (albeit modest) frequency in biostatistics; their application within economics is unusual (e.g., Lechner, 2009). Imbens and Wooldridge's (2009) recent survey of program evaluation methods devotes only a few paragraphs to dynamic program evaluation problems (while also noting that the area is understudied within econometrics). The growing availability for research use of administrative educational databases, population registers, and other large datasets with longitudinal structure, in conjunction with the intrinsically dynamic nature of many economic policies, suggests that economists could benefit from a better understanding of how to undertake covariate adjustment in longitudinal settings.

Average structural function

Assume the availability of a random sample of N units from the study population of interest. In each of $t = 0, 1, \ldots, T$ periods we observe a unit's treatment assignment $X_t \in \mathbb{X}_t$, and a vector of time-varying confounders, $W_t \in \mathbb{W}_t$. Also observed are a vector of baseline covariates $V \in \mathbb{V}$ and a final outcome of interest $Y \in \mathbb{Y}$. In what follows, for the purposes of exposition, I will often assume that X_t is a measure of teacher quality in grade t, W_t an end-of-school year test score in grade $t-1$ (with W_0 equaling measured achievement prior to kindergarten entry), V a vector of background controls (such as parental education, race and gender), and Y an adult outcome of interest.

A unit's treatment history from baseline through period t is denoted by $X_0^t = (X_0', X_1', \ldots, X_t')'$, with the convention that $X_0^{-1} = \oslash$. The history of time-varying confounders is denoted by $W_0^t = (W_0', W_1', \ldots, W_t')'$ (with $W_0^{-1} = \oslash$).

Let $\mathbb{X}_0^t = \mathbb{X}_0 \times \mathbb{X}_1 \times \ldots \times \mathbb{X}_t$ so that the number of possible treatment sequences coincides with the cardinality of the set $\mathbb{X}_0^T$, which we will take to be finite. Even with a simple binary treatment there will be 2^{T+1} possible treatment sequences; a fact which complicates identification.

For each unit we posit the existence of a potential response function

$$Y\left(x_0^T\right) = m\left(x_0^T, U\right). \tag{1}$$

The function $Y\left(x_0^t\right)$ gives the outcome that an individual would have experienced if assigned

to, possibly contrary to fact, treatment sequence $x_0^T \in \mathbb{X}_0^T$. The expression to the right of the equality in (1) provides a structural equation representation of the potential outcome response. The equality in (1) is without loss of generality: we are free to conceptualize U as a (potentially very high dimensional) vector of all attributes which generate individual-level variation in treatment response. In my own work I have sometimes found the potential outcome representation most convenient, while at other times the structural function representation more so. We will work mostly with the structural function in what follows.

Each individual's observed outcome is given by their potential response function (1) evaluated at their actual treatment sequence, a so-called 'consistency' condition,

$$Y = Y\left(X_0^T\right) = m\left(X_0^T, U\right). \tag{2}$$

The average structural function (ASF) (cf., Blundell and Powell, 2003; Wooldridge, 2005) is given by

$$\mathbb{E}\left[Y\left(x_0^T\right)\right] = \mathbb{E}\left[m\left(x_0^T, U\right)\right] \overset{def}{\equiv} m^{\text{ASF}}\left(x_0^T\right). \tag{3}$$

The ASF is an average of $m\left(x_0^T, U\right)$ over the marginal distribution of U; it coincides with the expected response of a randomly sampled individual to treatment regime x_0^T. Our goal will be to identify the ASF, at different logically feasible treatment sequences, from the joint distribution of the observed data $Z = \left(V, W_0^T, X_0^T, Y\right)$.

The econometrician faces a selection problem because the average outcome among those actually assigned to protocol $X_0^T = x_0^T$

$$\mathbb{E}\left[Y \middle| X_0^T = x_0^T\right] = \mathbb{E}\left[m\left(x_0^T, U\right) \middle| X_0^T = x_0^T\right],$$

does not, in general, equal $m^{\text{ASF}}\left(x_0^T\right)$. This is because the distribution of U among individuals experiencing treatment regime $X_0^T = x_0^T$ will typically differ from that in the population as a whole. For example the unobserved (time-invariant) determinants of adult earnings among individuals who attended schools staffed by high quality teachers will generally differ from those who did not attend such schools.

Assumptions

Robins (1986) extends the standard notion of exogeneity (Rubin, 1977) to the dynamic setting.

Assumption 1. (Sequential Exogeneity) *For* $t = 0, 1, \ldots, T$

$$U \perp X_t \big| X_0^{t-1}, W_0^t, V$$

for all $X_0^{t-1} \in \mathbb{X}_0^{t-1}$, $W_0^t \in \mathbb{W}_0^t$ *and* $V \in \mathbb{V}$.

Assumption 1 implies that – conditional on past treatment assignments, X_0^{t-1}, current and past time-varying controls, W_0^t, and baseline characteristics, V – current period treatment assignment varies independently of U (or equivalently potential outcomes). Assumption 1 implies that the econometrician observes all *joint* predictors of the treatment and outcome at each point in time. To understand the implications of this assumption it is helpful to consider an alternative identification condition. Specifically, instead of maintaining Assumption 1 we might have instead conceptualized the dynamic treatment regime as a static, but multi-valued, one. In that case the appropriate extension of Rubin's (1977) exogeneity condition was shown by Imbens (2000) to be

$$U \perp X_0^T \big| V \tag{4}$$

for all $X_0^T \in \mathbb{X}_0^T$ and $V \in \mathbb{V}$. Condition (4) asserts 'as if' random assignment of the *entire* treatment sequence conditional on baseline characteristics alone. This is much stronger than what is implied by Assumption 1, which asserts 'as if' *sequential* random assignment, conditional on an information set which grows with time.

Under Assumption 1 we have the following joint density factorization

$$\begin{aligned}
f\left(u, v, w_0, \ldots, w_T, x_0, \ldots, x_T\right) &= f\left(x_T \middle| u, v, w_0^T, x_0^{T-1}\right) f\left(u, v, w_0^T, x_0^{T-1}\right) \\
&= f\left(x_T \middle| v, w_0^T, x_0^{T-1}\right) f\left(w_T \middle| u, v, w_0^{T-1}, x_0^{T-1}\right) \\
&\quad \times f\left(u, v, w_0^{T-1}, x_0^{T-1}\right) \\
&= \left\{\prod_{t=1}^{T} f\left(x_t \middle| v, w_0^t, x_0^{t-1}\right) f\left(w_t \middle| u, v, w_0^{t-1}, x_0^{t-1}\right)\right\} \\
&\quad \times f\left(u, v, w_0, x_0\right) \\
&= \left\{\prod_{t=1}^{T} f\left(x_t \middle| v, w_0^t, x_0^{t-1}\right) f\left(w_t \middle| u, v, w_0^{t-1}, x_0^{t-1}\right)\right\} \\
&\quad \times f\left(x_0 \middle| v, w_0\right) f\left(u, v, w_0\right),
\end{aligned}$$

with the second and fourth equalities following from Assumption 1.

A key implication of Assumption 1 is that the density of W_t

$$f\left(w_t \middle| u, v, w_0^{t-1}, x_0^{t-1}\right) \tag{5}$$

may vary with u and past treatment assignment x_0^{t-1}.

Assumption 1 implies that conditional on background, V, prior achievement, W_0^t, and prior teacher quality, X_0^{t-1}, current teacher quality, X_t, is 'as good as' randomly assigned. The condition *does* allow, for example, students with low quality teachers in period $t-1$ to be systematically assigned to higher equality teachers in period t (on average). Likewise it allows for current period teacher assignments to depend on past test scores. Furthermore, by (5) above, it allows test scores, W_t, to depend on unobserved ability, U, past achievement, W_0^{t-1}, and past teacher assignments, X_0^t. Test scores are both a confounder and an intermediate outcome (or concomitant variable).

Condition (4), in contrast to Assumption 1, requires that the entire sequence of teacher assignments is as good as random given baseline characteristics (V). Specifically it does not allow period t assignments to be influenced by time-varying prognostic variables. In the static program evaluation setting, conditioning on intermediate outcomes is known to generate bias (e.g., Rosenbaum, 1984). A key contribution of Robins (1986, 1997, 2000) is to show how to control for such confounders.

Identification of the ASF also requires a support condition. To state this condition consider the set of values for the baseline attribute, V, and time-varying confounders through period t, W_0^t that are observed in the subpopulation of units receiving treatment sequence $X_0^{t-1} = x_0^{t-1}$ from baseline to period $t-1$:

$$\mathbb{S}\left(x_0^{t-1}\right) = \left\{v, w_0^t \,:\, f\left(v, w_0^t, x_0^{t-1}\right) > 0\right\}. \tag{6}$$

Let x_0^t be some sequence of treatments from baseline to period t. Set (6) defines a subpopulation of units that has followed protocol x_0^t through period $t-1$ and thus could, in principle, continue on with the protocol in period t.

Assumption 2. (Overlap) *Given the specific treatment regime $x_0^T \in \mathbb{X}_0^T$,*

$$f\left(x_t | \, v, w_0^t, x_0^{t-1}\right) \geq \kappa > 0$$

for all $v, w_0^t \in \mathbb{S}\left(x_0^{t-1}\right)$ and $t = 0, 1, \ldots, T$.

Assumption 2 implies that in order to learn about the distribution of potential outcomes at $X_0^T = x_0^T$ is must be the case that, for any set of units following the x_0^T protocol for $t-1$ periods, a positive fraction will continue to follow the protocol in period t

Begin with $t = 0$. Assumption 2 implies that the probability of the initial assignment $X_0 = x_0$ is bounded away from zero for all values of V and W_0 observed in the population.

Consequently at least some units experience treatment $X_0 = x_0$ within all subpopulations defined in terms of W_0 and V. In the next period, for all subpopulations defined in terms of V, W_0, W_1 that were also previously assigned $X_0 = x_0$, a positive fraction will receive assignment $X_1 = x_1$. Hence among any subpopulation of units that can logically continue from assignment $X_0 = x_0$ to $X_1 = x_1$ at least some positive mass will do so. For a generic t the probability of the assignment $X_t = x_t$ is bounded away from zero for all values of V and W_0^t observed in the subpopulation that was previously assigned treatments $X_0^{t-1} = x_0^{t-1}$. Assumption 2 is thus a sequential generalization of the usual overlap condition from the static program evaluation problem.

If the cardinality of $\mathbb{X}_0^T$ is large we might expect Assumption 2 to hold for only a subset of logically feasible treatment regimes. Without further assumptions we will only be able to identify $m^{\text{ASF}}\left(x_0^T\right)$ at sequences x_0^T which satisfy Assumption 2.

To understand the implications of the assumption for empirical work it is helpful to consider some examples. Consider the average impact of living in a high poverty neighborhood in all $t = 0, 1, \ldots, T$ years of childhood. If there is some subpopulation of units defined in terms of W_0 and V that are never assigned to a high poverty neighborhood in period 0, then clearly we will be unable to learn about the population average effect of living in high poverty neighborhood in all years of childhood. Identification will also fail if, conditional on initial assignment to a high poverty neighborhood, the transition rate out in subsequent periods is "nearly" one hundred period for some sub-group. For example, even if some children whose parents' have completed a graduate degree begin their lives in high poverty environments, if they all eventually move out of such neighborhoods, then the average structural function at the sequence "high poverty neighborhood in all years of childhood" is not identified.

G-Computation Formula

Robins (1986) develops a method of covariate adjustment for environments characterized by Assumptions 1 and 2. Central to his method is the so-called G-Computation formula. To understand this formula it is helpful to recall some basic results for the static program evaluation problem under exogeneity (e.g., Imbens, 2004). In that problem the structural function is, setting $T = 0$,

$$Y(x_0) = m(x_0, U)$$

with x_0 scalar-valued. With $T = 0$ Assumption 1 simplifies to to the point-in-time exogeneity condition

$$U \perp X_0 | V, W_0$$

and Assumption 2 to the requirement that

$$f\left(x_0 \middle| v, w_0\right) \geq \kappa > 0$$

for all $v \in \mathbb{V}$ and $w_0 \in \mathbb{W}_0$.

Let $q\left(v, w_0, x_0\right) = \mathbb{E}\left[Y \middle| V = v, W_0 = w_0, X_0 = x_0\right]$ be the proxy variable regression (PVR) function. In the static case ($T = 0$), using the law of iterated expectations, we can express the ASF at $X_0 = x_0$ as

$$\begin{aligned}
\mathbb{E}\left[Y\left(x_0\right)\right] &= \mathbb{E}\left[m\left(x_0, U\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[m\left(x_0, U\right) \middle| W_0, V\right]\right] \\
&= \int\int \mathbb{E}\left[m\left(x_0, U\right) \middle| V = v, W_0 = w_0\right] f\left(v, w_0\right) \mathrm{d}v\mathrm{d}w_0 \\
&= \int\int \mathbb{E}\left[m\left(x_0, U\right) \middle| V = v, W_0 = w_0, X_0 = x_0\right] f\left(v, w_0\right) \mathrm{d}v\mathrm{d}w_0 \\
&= \int\int q\left(v, w_0, x_0\right) f\left(v, w_0\right) \mathrm{d}v\mathrm{d}w_0 \\
&= \mathbb{E}\left[q\left(V, W_0, x_0\right)\right]
\end{aligned}$$

The second equality follows from the law of iterated expectations and the third by re-writing the outer expectation in integral form. The fourth equality is an implication of conditional independence of U and X_0 given (V, W_0). Since the distribution of U given (V, W_0) does not vary with treatment assignment, we are free to additionally condition on $X_0 = x_0$ in the integrand. The fifth equality follows because the observed outcome for units assigned to treatment $X_0 = x_0$, coincides with $m\left(x_0, U\right)$.

In the static case the ASF is identified by a partial mean: compute the mean outcome given covariates and treatment and then average this over the marginal distribution of covariates. Under the overlap condition $f\left(x_0 \middle| v, w_0\right) \geq \kappa > 0$ for all $v \in \mathbb{V}$ and $w_0 \in \mathbb{W}_0$ this outer average is well-defined (cf., Newey, 1994).

It is tempting to naively apply the same logic to evaluate a sequence of treatments under Assumptions 1 and 2. Specific, in the dynamic case, we can let

$$q\left(v, w_0^T, x_0^T\right) = \mathbb{E}\left[Y \middle| V = v, W_0^T = w_0^T, X_0^T = x_0^T\right] \tag{7}$$

be the proxy variable regression function. We then might hope that its marginal mean

$$\int_v \int_{w_0} \cdots \int_{w_T} q\left(v, w_0^T, x_0^T\right) f\left(v, w_0^T\right) \mathrm{d}w_T \cdots \mathrm{d}w_0 \mathrm{d}v$$

identifies the ASF. An implication of Robins (1986) is that this is not the case. Robins (1986) instead shows that, in order to appropriately adjust for bias cause by time-varying confounders, one must sequentially average the proxy variable regression function vis-a-vis a particular sequence of conditional distributions for the confounders.

Lemma 1. (G-Computation Formula) *Under (2) and Assumption 1*

$$m^{\text{ASF}}\left(x_0^T\right) = \int_v \int_{w_0} \cdots \int_{w_T} q\left(v, w_0^T, x_0^T\right) \prod_{t=0}^{T} f\left(w_t \middle| v, w_0^{t-1}, x_0^{t-1}\right) f\left(v\right) \mathrm{d}w_T \cdots \mathrm{d}w_0 \mathrm{d}v \tag{8}$$

for all treatment sequences x_0^T satisfying Assumption 2.

Proof. An outline of the argument is as follows. Under random sampling $q\left(v, w_0^T, x_0^T\right)$ is non-parametrically identified at all points in the joint support $\mathbb{V} \times \mathbb{W}_0^T \times \mathbb{X}_0^T$. For the average (8) to be computable we require that $\prod_{t=0}^{T} f\left(w_t \middle| v, w_0^{t-1}, x_0^{t-1}\right) f\left(v\right)$ is non-zero only at values of v and w_0^T contained in the conditional support of $\mathbb{V} \times \mathbb{W}_0^T$ given $X_0^T = x_0^T$. Assumption 2 ensures this condition. To see this consider the factorization

$$\begin{aligned}
\prod_{t=0}^{T} f\left(w_t \middle| v, w_0^{t-1}, x_0^{t-1}\right) f\left(v\right) &= \prod_{t=0}^{T} \frac{f\left(v, w_0^t, x_0^{t-1}\right)}{f\left(v, w_0^{t-1}, x_0^{t-1}\right)} f\left(v\right) \\
&= \frac{f\left(v, w_0^T, x_0^{T-1}\right)}{f\left(v, w_0^{T-1}, x_0^{T-1}\right)} \frac{f\left(v, w_0^{T-1}, x_0^{T-2}\right)}{f\left(v, w_0^{T-2}, x_0^{T-2}\right)} \times \\
&\quad \cdots \times \frac{f\left(v, w_0^1, x_0\right)}{f\left(v, w_0, x_0\right)} \frac{f\left(v, w_0\right)}{f\left(v\right)} f\left(v\right) \\
&= f\left(v, w_0^T, x_0^{T-1}\right) \frac{1}{f\left(x_{T-1} \middle| v, w_0^{T-1}, x_0^{T-2}\right)} \\
&\quad \times \frac{1}{f\left(x_{T-2} \middle| v, w_0^{T-2}, x_0^{T-3}\right)} \times \cdots \times \frac{1}{f\left(x_0 \middle| v, w_0\right)}.
\end{aligned}$$

Under Assumption 2 $q\left(v, w_0^T, x_0^T\right)$ is identified at all points $v, w_0^T \in \mathbb{S}\left(x_0^{T-1}\right)$ and, hence, whenever the numerator in the ratio to the right of the last equality above is non-zero. We also have the denominator is non-zero and positive. To see this note that if $v, w_0^{T-1} \in \mathbb{S}\left(x_0^{T-1}\right)$ it is also an element of $\mathbb{S}\left(x_0^{T-2}\right)$. This follows because any sub-group that follows treatment $X_0^{T-2} = x_0^{T-2}$ through periods $t = 0, 1, \ldots, T-2$ continues on to treatment $X_{T-1} = x_{T-1}$ in period $T-1$ with positive probability under Assumption 2. Hence any group defined in terms of period $T-1$ observables (V, W_0^{T-1}) contained in $\mathbb{S}\left(x_0^{T-2}\right)$ will also be in $\mathbb{S}\left(x_0^{T-1}\right)$. This gives positivity of the denominator. Consequently the right-hand-side of (8) is identified.

We can then use Assumptions 1 to show that the stated equality holds. Manipulating the right-hand-side of (8):

$$\begin{aligned}
&\int_v \int_{w_0} \cdots \int_{w_T} \left\{ q\left(v, w_0^T, x_0^T\right) \right. \\
&\qquad \left. \prod_{t=0}^{T} f\left(w_t \middle| v, w_0^{t-1}, x_0^{t-1}\right) f\left(v\right) \right\} \mathrm{d}w_T \cdots \mathrm{d}w_0 \mathrm{d}v \quad = \\
&\int_v \int_{w_0} \cdots \int_{w_T} \left\{ \left[\int_u m\left(x_0^T, u\right) f\left(u \middle| v, w_0^T, x_0^T\right) \mathrm{d}u \right] \right. \\
&\qquad \left. \prod_{t=0}^{T} f\left(w_t \middle| v, w_0^{t-1}, x_0^{t-1}\right) f\left(v\right) \mathrm{d}w_T \cdots \mathrm{d}w_0 \mathrm{d}v \right\} \quad = \\
&\int_u \int_v \int_{w_0} \cdots \int_{w_T} \left\{ m\left(x_0^T, u\right) f\left(u \middle| v, w_0^T, x_0^{T-1}\right) f\left(w_T \middle| v, w_0^{T-1}, x_0^{T-1}\right) \right. \\
&\qquad \left. \prod_{t=0}^{T-1} f\left(w_t \middle| v, w_0^{t-1}, x_0^{t-1}\right) f\left(v\right) \right\} \mathrm{d}w_T \cdots \mathrm{d}w_0 \mathrm{d}v \mathrm{d}u \quad = \\
&\int_v \int_{w_0} \cdots \int_{w_T} \int_u \left\{ m\left(x_0^T, u\right) f\left(u, w_T \middle| v, w_0^{T-1}, x_0^{T-1}\right) \right. \\
&\qquad \left. \prod_{t=0}^{T-1} f\left(w_t \middle| v, w_0^{t-1}, x_0^{t-1}\right) f\left(v\right) \right\} \mathrm{d}w_T \cdots \mathrm{d}w_0 \mathrm{d}v \mathrm{d}u \quad = \\
&\int_v \int_{w_0} \cdots \int_{w_{T-1}} \int_u \left\{ m\left(x_0^T, u\right) f\left(u \middle| v, w_0^{T-1}, x_0^{T-1}\right) \right. \\
&\qquad \left. \prod_{t=0}^{T-1} f\left(w_t \middle| v, w_0^{t-1}, x_0^{t-1}\right) f\left(v\right) \right\} \mathrm{d}w_{T-1} \cdots \mathrm{d}w_0 \mathrm{d}v \mathrm{d}u \quad = \\
&\qquad \vdots \\
&\int_v \int_u m\left(x_0^T, u\right) f\left(u \middle| v\right) f\left(v\right) \mathrm{d}v \mathrm{d}u \quad = \\
&\qquad m^{\mathrm{ASF}}\left(x_0^T\right) \quad =
\end{aligned}$$

where the first equality follows from (2) and the definition of the conditional expectation function and the second equality follows from sequential exogeneity. The third and fourth equalities involve density factorization and integration over w_T. These last two steps are repeated for $t = T-1, T-2, \ldots\ldots, 0$ to get the final equality. □

When $T = 0$ equation (8) specializes to

$$m^{\mathrm{ASF}}\left(x_0\right) = \int_v \int_{w_0} q\left(v, w_0, x_0\right) f\left(v, w_0\right) \mathrm{d}w_0 \mathrm{d}v,$$

which is familiar from the static problem. Lemma 1 is therefore a dynamic generalization of the familiar approach to covariate adjustment in static models by marginal integration (e.g., Newey, 1994; Blundell and Powell, 2003; Wooldridge, 2005).

Recursive representation

Bang and Robins (2005) discuss a backwards recursion formulation of G-Computation. This representation, which helps to build intuition for Lemma 1, is given by

$$\begin{aligned}
\tilde{m}_{T+1}\left(V, W_0^T, x_0^T\right) &= \mathbb{E}\left[Y \middle| V, W_0^T, X_0^T = x_0^T\right] = q\left(V, W_0^T, x_0^T\right) \\
\tilde{m}_T\left(V, W_0^{T-1}, x_0^T\right) &= \mathbb{E}\left[\tilde{m}_{T+1}\left(V, W_0^T, x_0^T\right) \middle| V, W_0^{T-1}, X_0^{T-1} = x_0^{T-1}\right] \\
&\vdots \\
\tilde{m}_t\left(V, W_0^{t-1}, x_0^T\right) &= \mathbb{E}\left[\tilde{m}_{t+1}\left(V, W_0^t, x_0^T\right) \middle| V, W_0^{t-1}, X_0^{t-1} = x_0^{t-1}\right] \\
&\vdots \\
\tilde{m}_2\left(V, W_0^1, x_0^T\right) &= \mathbb{E}\left[\tilde{m}_3\left(V, W_0^2, x_0^T\right) \middle| V, W_0^1, X_0^1 = x_0^1\right] \\
\tilde{m}_1\left(V, W_0, x_0^T\right) &= \mathbb{E}\left[\tilde{m}_2\left(V, W_0^1, x_0^T\right) \middle| V, W_0, X_0 = x_0\right] \\
\tilde{m}_0\left(x_0^T\right) &= \mathbb{E}\left[\tilde{m}_1\left(V, W_0, x_0^T\right)\right] = m^{\text{ASF}}\left(x_0^T\right).
\end{aligned}$$

Each line above corresponds to one of the integrals in the G-Computation formula. Integration occurs sequentially, beginning with the inmost integral and working outwards.

Covariate imbalance in dynamic evaluation problems unfolds sequentially. At each time period some individuals continue on with the particular protocol of interest, while others deviate. To the extent that these two groups differ, covariate imbalance is introduced. Each of the averages in the lines above corrects for a particular component of this imbalance. This correction occurs backwards: we undo the imbalance introduced in the final period first, then the next to last period and so on.

Our goal is to recover the value of the average structural function (ASF) at treatment sequence x_0^T. We begin, at "line $T+1$", by averaging the final outcome conditional on V, W_0^T and $X_0^T = x_0^T$. This yields $\tilde{m}_{T+1}\left(V, W_0^T, x_0^T\right)$, which coincides with the proxy variable regression (PVR) function introduced above.

Next we compute $\tilde{m}_T\left(V, W_0^{T-1}, x_0^T\right)$; this average corrects for covariate imbalance induced by treatment selection in the final period. More precisely, suppose we wish to identify the average potential outcome associated with final period treatment $X_T = x_T$ *conditional* on having previously followed treatment sequence $X_0^{T-1} = x_0^{T-1}$ *and* on having period $T-1$ characteristics W_0^{T-1} and V. Within this specific subpopulation of units, the only possible

covariate imbalance is in W_T, the distribution of which may differ across the those who continue on to final period treatment $X_T = x_T$ and those that do not. We can correct for any such imbalance by averaging over W_T. Since we are only concerned with the subpopulation that previously followed protocol $X_0^{T-1} = x_0^{T-1}$ and has period $T-1$ characteristics W_0^{T-1} and V, the appropriate average over W_T conditions on these covariates. Hence $\tilde{m}_T\left(V, W_0^{T-1}, x_0^T\right)$ gives the average counterfactual outcome that would have been observed if *all* units that followed protocol $X_0^{T-1} = x_0^{T-1}$ and have period $T-1$ characteristics W_0^{T-1} and V had, possibly contrary to fact, continued on with treatment $X_T = x_T$ in the final period.

Next we compute $\tilde{m}_{T-1}\left(V, W_0^{T-2}, x_0^T\right)$; this corresponds to the average potential outcome associated with assignment to $X_T = x_T$ and $X_{T-1} = x_{T-1}$ within the subpopulation that previously followed the treatment protocol $X_0^{T-2} = x_0^{T-2}$ *and* with period $T-2$ characteristics W_0^{T-2} and V. This subpopulation is larger than the one considered in the previous calculation. Specifically, it includes all those units represented in the prior average as well as those units who began to deviate from the x_0^T target treatment assignment in period $T-1$. To recover the mean potential outcome in this subpopulation we average $\tilde{m}_T\left(V, W_0^{T-1}, x_0^T\right)$ with respect to the conditional distribution of W_{T-1} given V, W_0^{T-2} and $X_0^{T-2} = x_0^{T-2}$. Recall that $\tilde{m}_T\left(V, W_0^{T-1}, x_0^T\right)$ gives the average outcome associated with final period treatment $X_T = x_T$ conditional on having previously followed protocol $X_0^{T-1} = x_0^{T-1}$ and having period $T-1$ characteristics W_0^{T-1} and V. This subpopulation is a selected one relative to our new target one because not all individuals in the subpopulation with $X_0^{T-2} = x_0^{T-2}$ and homogenous in W_0^{T-2} and V go onto to treatment $X_{T-1} = x_{T-1}$. We can correct for this by appropriately undoing our conditioning on W_{T-1}.

We continue recursively in this way until we recover the population-wide average outcome associated with treatment sequence $X_0^T = x_0^T$. At each stage our average of potential outcomes becomes more and more representative of the population of interest.

A parametric example

Let $T = 1$ and, for simplicity, set $V = \oslash$. We can construct a very simple example of G-Computation when both the proxy variable regression function $q\left(w_0^1, x_0^1\right)$ and the auxiliary mean regression $\mathbb{E}\left[W_1 \middle| W_0 = w_0, X_0 = x_0\right]$ are linear in their arguments:

$$\begin{aligned} q\left(W_0^1, x_0^1\right) &= \gamma_0 + W_1'\gamma_{w1} + W_0'\gamma_{w0} + x_1'\gamma_{x1} + x_0'\gamma_{x0} \\ \mathbb{E}\left[W_1 \middle| W_0, X_0 = x_0\right] &= \Delta_0 + \Delta_{w0}W_0 + \Delta_{x0}x_0. \end{aligned}$$

Under these parametric specifications we have

$$\begin{aligned}
\tilde{m}_2\left(W_0^1, x_0^1\right) &= \gamma_0 + W_1'\gamma_{w1} + W_0'\gamma_{w0} + x_1'\gamma_{x1} + x_0'\gamma_{x0} \\
\tilde{m}_1\left(W_0, x_0^1\right) &= \gamma_0 + \mathbb{E}\left[W_1 \middle| W_0, X_0 = x_0\right]'\gamma_{w1} + W_0'\gamma_{w0} + x_1'\gamma_{x1} + x_0'\gamma_{x0} \\
&= \gamma_0 + \left\{\Delta_0 + \Delta_{w0}W_0 + \Delta_{x0}x_0\right\}'\gamma_{w1} + W_0'\gamma_{w0} + x_1'\gamma_{x1} + x_0'\gamma_{x0} \\
&= \left(\gamma_0 + \Delta_0'\gamma_{w1}\right) + W_0'\left(\gamma_{w0} + \Delta_{w0}'\gamma_{w1}\right) + x_1'\gamma_{x1} + x_0'\left(\gamma_{x0} + \Delta_{x0}'\gamma_{w1}\right) \\
\tilde{m}_0\left(x_0^1\right) &= \left\{\left(\gamma_0 + \Delta_0'\gamma_{w1}\right) + \mathbb{E}\left[W_0\right]'\left(\gamma_{w0} + \Delta_{w0}'\gamma_{w1}\right)\right\} + x_1'\gamma_{x1} + x_0'\left(\gamma_{x0} + \Delta_{x0}'\gamma_{w1}\right) \\
&= m^{\text{ASF}}\left(x_0^1\right).
\end{aligned}$$

Note that in this example the coefficient on X_1 in the proxy variable regression correctly identifies the causal effect of X_1 on the outcome, however the coefficient on X_0 does not identify the correct causal effect. The correct effect is given by $(\gamma_{x0} + \Delta_{x0}'\gamma_{w1})$, which includes the direct effect γ_{x0} as well as the indirect effect "mediated" by W_1.

Estimation based on the G-Computation Formula

Basing estimation of Lemma 1 is not-straightforward. The most common method is fully parametric and uses simulation to compute the integral in (8) (e.g., Taubman, Robins, Mittleman and Hernán, 2009).

Algorithm 1. *Parametric G-Computation*

1. Specify and fit by maximum likelihood (ML) the $t = 1, \ldots, T$ parametric models

$$f\left(w_1 \middle| v, w_0, x_0; \eta_1\right), \ldots, f\left(w_T \middle| v, w_0^{T-1}, x_0^{T-1}; \eta_T\right).$$

2. Specify and fit by maximum likelihood (ML) the parametric model $f\left(y \middle| v, w_0^T, x_0^T; \lambda\right)$.

3. Let x_0^T be the treatment sequence of interest. Compute the following sequence of simulated random variables

   (a) $\left(V_0^{(s)}, W_0^{(s)}\right)$, a random draw from the empirical distribution.

   (b) $W_1^{(s)}$, a random draw from the distribution defined by the density $f\left(w_1 \middle| V^{(s)}, W_0^{(s)}, x_0; \hat{\eta}_1\right)$.

   (c) For $t = 2, \ldots, T$ draw $W_t^{(s)}$ at random from the distribution defined by the density $f\left(w_t \middle| V^{(s)}, W_0^{(s)}, \ldots, W_{t-1}^{(s)}, x_0^{t-1}; \hat{\eta}_t\right)$.

(d) Draw $Y^{(s)}\left(x_0^T\right)$ from the distribution defined by the density

$$f\left(y|\, V^{(s)}, W_0^{(s)}, \ldots, W_T^{(s)}, x_0^T; \hat{\lambda}\right).$$

(e) Repeat steps (a) to (d) S times.

4. Estimate the average structural function at treatment sequence x_0^T from the S simulation draws as

$$\hat{m}^{\text{ASF}}\left(x_0^T\right) = \frac{1}{S}\sum_{s=1}^{S} Y^{(s)}\left(x_0^T\right).$$

Other functionals of the distribution of potential outcomes can be computed in step 4 in the obvious way. Assuming correct specification of the distributions fitted in steps 1 and 2, there remain two sources of error in $\hat{m}^{\text{ASF}}\left(x_0^T\right)$: sampling error and simulation error. The latter can be made arbitrarily small by choosing S to be very large (albeit at a computational cost). To account for the sampling error it is probably easiest to use the bootstrap (i.e., repeat steps 1 to 4 using B different bootstrap samples). As Algorithm 1 is likelihood-based, posterior uncertainty also could be assessed using Bayesian methods. In principle, a fully nonparametric implementation of G-Computation is also feasible.

I am aware no applications of G-Computation methods within economics and only a handful in biostatistics. The exposition here provides some indication of why: the method is subtle to understand and not simple to implement (particularly when W_t is high dimensional). Nevertheless G-Computation provides a coherent and attractive generalization of the familiar idea of covariate adjustment via marginal integration to dynamic settings. Given the proliferation of data with longitudinal structures, and the inherently dynamic nature of many policies of interest to economists, it seems fair to say that Lemma 1 should be more widely known among economists.

Inverse probability weighting

Robins (2000), maintaining Assumptions 1 and 2 proposes an inverse probability of treatment type estimator for the ASF. Just as the G-Computation approach generalizes static identification arguments for the ASF based on marginal integration, Robins (2000) generalizes familiar re-weighting methods for covariate adjustment (e.g., Rosenbaum, 1987) in static models to dynamic ones.

We begin by specifying a parametric model for the ASF directly

$$m^{\text{ASF}}\left(x_0^T\right) = g\left(x_0^T; \beta_0\right). \tag{9}$$

Here $g(\cdot;\beta)$ is a known family of link functions indexed by the finite dimensional parameter β. Since we are assuming that the number of logically feasible treatment sequences is finite, we can make our parametric model nonparametric through saturation. In practice working with a restrictive model for the ASF aides in both identification and effect interpretation.

Recall that the support of possible treatment assignments $\mathbb{X}_0^T$ is discrete and finite. Let

$$e_0\left(v, w_0^T, x_0^T\right) = \prod_{t=0}^{N} \frac{f\left(x_t \middle| v, w_0^t, x_0^{t-1}\right)}{f\left(x_t \middle| x_0^{t-1}\right)} \tag{10}$$

be a stabilized weight. The numerator in (10) is a particular probability of treatment measure. Specifically $f(x_0|\, w_0, v)$ gives the probability of assignment to $X_0 = x_0$ given the period 0 information set of the econometrician; $f\left(x_t|\, w_0^t, x_0^{t-1}, v\right)$ gives the probability of assignment to $X_t = x_t$ given the period t information set of the econometrician. Hence the denominator equals a recursively updated probability for treatment sequence $X_0^T = x_0^T$. The denominator equals the marginal probability of treatment sequence $X_0^T = x_0^T$. Under pure random assignment $e_0\left(v, w_0^T, x_0^T\right) \equiv 1$ for all units. Under sequential conditional exogeneity $e_0\left(v, w_0^T, x_0^T\right) < 1$ for treatment sequences which occur rarely conditional on controls and $e_0\left(v, w_0^T, x_0^T\right) > 1$ for those which occur frequently.

Robins (2000) proves the following Lemma.

Lemma 2. (Identification by Reweighting) *Under (2), Assumptions 1, and Assumption 2:*

$$\mathbb{E}\left[\frac{h\left(X_0^T\right)}{e_0\left(V, W_0^T, X_0^T\right)}\left(Y - m^{\text{ASF}}\left(X_0^T\right)\right)\right] = 0. \tag{11}$$

Proof. Using the joint density factorization implied by Assumptions 1 we can re-write (11)

as

$$
\begin{aligned}
& \mathbb{E}\left[\frac{h\left(X_0^T\right)}{e_0\left(V, W_0^T, X_0^T\right)}\left(Y - m^{\text{ASF}}\left(X_0^T\right)\right)\right] \\
= & \sum_{x_0 \in \mathbb{X}_0} \cdots \sum_{x_T \in \mathbb{X}_T} \int_u \int_v \int_{w_0} \cdots \int_{w_T} \frac{h\left(x_0^T\right)\left(m\left(x_0^T, u\right) - m^{\text{ASF}}\left(x_0^T\right)\right)}{e_0\left(v, w_0^T, x_0^T\right)} \\
& \times \left\{\prod_{t=1}^{T} f\left(x_t \middle| v, w_0^t, x_0^{t-1}\right) f\left(w_t \middle| u, v, w_0^{t-1}, x_0^{t-1}\right)\right\} \\
& \times f\left(x_0 \middle| v, w_0\right) f\left(u, v, w_0\right) \mathrm{d}w_T \cdots \mathrm{d}w_0 \mathrm{d}v \mathrm{d}u \\
= & \sum_{x_0 \in \mathbb{X}_0} \cdots \sum_{x_T \in \mathbb{X}_T} \int_u \int_v \int_{w_0} \cdots \int_{w_T} h\left(x_0^T\right)\left(m\left(x_0^T, u\right) - m^{\text{ASF}}\left(x_0^T\right)\right) \\
& \times \left\{f\left(x_0\right) \prod_{t=1}^{T} f\left(x_t \middle| x_0^{t-1}\right)\right\} \\
& \times f\left(w_t \middle| u, v, w_0^{t-1}, x_0^{t-1}\right) f\left(u, v, w_0\right) \mathrm{d}w_T \cdots \mathrm{d}w_0 \mathrm{d}v \mathrm{d}u \\
= & \sum_{x_0 \in \mathbb{X}_0} \cdots \sum_{x_T \in \mathbb{X}_T} \left[\int_u h\left(x_0^T\right)\left(m\left(x_0^T, u\right) - m^{\text{ASF}}\left(x_0^T\right)\right) f\left(u\right) \mathrm{d}u\right] \\
& \times f\left(x_0\right) \prod_{t=1}^{T} f\left(x_t \middle| x_0^{t-1}\right).
\end{aligned}
$$

Under Assumption 2 and other regularity conditions, the integrand in the expression to the right of the first equality will be finite. This fact, the form of the weights, and Assumption 1 gives the second equality. The third equality follows after integrating over $v, w_0, w_1, \ldots, w_T$. The claim then follows since, by (3),

$$
\int_u m\left(x_0^T, u\right) f\left(u\right) \mathrm{d}u = m^{\text{ASF}}\left(x_0^T\right),
$$

which ensures that the integrand to the right of the last equality is zero for all $x_0^T \in \mathbb{X}_0^T$. □

Estimation based on IPW

For simplicity consider the case where X_t is binary-valued. In this case the $e_0\left(V, W_0^T, X_0^T\right)$ weights may be estimated by assembling a sequence of logistic regression fits. The numerator corresponds to the marginal probability the sequence X_0^T, which may be estimated as a cell mean. Assume that the outcome is continuously-valued. We specify a parametric form for

the ASF of, for example,

$$\begin{aligned} m^{\text{ASF}}\left(x_0^T\right) &= g\left(x_0^T;\beta\right) \\ &= \alpha+\gamma\left(\sum_{t=0}^{T} x_t\right). \end{aligned}$$

This model is restrictive. If $T=1$, it asserts that the average potential outcome associated with $X_0^1=(0,1)'$ is the same as that associated with $X_0^1=(1,0)'$. Only the number of times treated "matters". A non-parametric parameterization of the ASF would involve 2^{T+1} parameters, the specification above involves just two. Assuming the stabilized weights are correctly specified, Lemma 2 implies that even if our ASF is incorrectly-specified we will still recover a projection of the true ASF onto our family of approximating functions.

With our estimated weights in hand, we recover estimates of α_0 and γ_0 by computing the weighted least squares fit of Y onto a constant and $\left(\sum_{t=0}^{T} X_t\right)$ using weight $1/\hat{e}\left(V,W_0^T,X_0^T\right)$. The standard errors reported by our program will be incorrect as they do not incorporate the effects of sampling error in $\hat{e}\left(V,W_0^T,X_0^T\right)$. A simple way to account for this is to use a bootstrap.

Additional reading

Daniel et al. (2010) provide a user-oriented introduction to methods for dealing with time varying confounders. The online supplement of this paper includes STATA scripts. Two empirical examples of parametric G-Computation are Moore et al. (2008) and Taubman, Robins, Mittleman and Hernán (2009). Sharkey and Elwert (2011), Wodtke, Harding and Elwert (2011) and Wodtke (2013) apply IPW methods to learn about neighborhood effects. Introductions to inverse probability weighting for sequential treatments are provided by Robins, Hernán and Brumback (2000) and Hernán, Brumback and Robins (2001). Robins (1999) is an accessible, albeit somewhat dated, introduction to his theoretical work in this area. Some of the theoretical arguments introduced here, especially the G-Computation formula, also arise in mediation analysis. Baron and Kenny (1986) is a widely-cited early reference. Tchetgen and Shpitser (2012) a recent theoretical reference.

References

[1] Bang, Heejung and James M. Robins. (2005). "Doubly robust estimation in missing data and causal inference models," *Biometrics* 61 (4): 962 - 972.

[2] Baron, Reuben M., and David A. Kenny (1986). "The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations," *Journal of Personality and Social Psychology* 51 (6): 1173 - 1182.

[3] Blundell, Richard W. and James L. Powell. (2003). "Endogeneity in nonparametric and semi- parametric regression models," *Advances in Economics and Econometrics: Theory and Applications* II: 312 - 357. (M. Dewatripont, L.P. Hansen, S. J. Turnovsky, Eds.). Cambridge: Cambridge University Press.

[4] Daniel, R. M., S. N. Cousens, B. L. De Stavola, M. G. Kenward and J. A. C. Sterne. (2012). "Methods for dealing with time-dependent confounding," *Statistics in Medicine* 32 (9): 1584 - 1618.

[5] Hernán, Miguel A., Babette Brumback and James M. Robins. (2001). "Marginal structural models to estimate joint causal effect of nonrandomized treatments," *Journal of the American Statistical Association* 96 (454): 440 - 448.

[6] Imbens, Guido W. (2000). "The role of the propensity score in estimating dose-response functions," *Biometrika* 87 (3): 706 - 710.

[7] Imbens, Guido W. (2004). "Nonparametric estimation of average treatment effects under exogeneity: a review," *Review of Economics and Statistics* 86 (1): 4 - 29.

[8] Imbens, Guido W., and Jeffrey M. Wooldridge. (2009). "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature* 47 (1): 5 - 86.

[9] Lechner, Michael (2009). "Sequential causal models for the evaluation of labor market programs," *Journal of Business and Economic Statistics* 27 (1): 71 - 83.

[10] Moore, Kelly, Romain Neugebauer, Fred Lurmann, Jane Hall, Vic Brajer, Sianna Alcorn and Ira Tager. (2008). "Ambient ozone concentrations cause increased hospitalization for asthma in children: an 18-year study in southern California," *Environmental Health Perspectives* 116 (8): 1063 - 1070.

[11] Newey, Whitney K. (1994). "Kernel estimation of partial means and a general variance estimator," *Econometric Theory* 10 (2): 233 - 253.

[12] Robins, James. (1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect," *Mathematical Modelling* 7 (9–12): 1393 - 1512.

[13] Robins, James M. (1997). "Causal inference from complex longitudinal data," *Latent Variable Modeling and its Applications to Causality*: 69 -117 (M. Berkane, Ed.). New York: Springer-Verlag.

[14] Robins, James M. (1999). "Association, causation, and marginal structural models," *Synthese* 121 (1-2): 151 - 179.

[15] Robins, James M. (2000). "Marginal structural models versus structural nested models as tools for causal inference," *Statistical Models in Epidemiology, the Environment, and Clinical Trials: The IMA Volumes in Mathematics and its Applications* 116: 95 - 133 (M. E. Halloran & D. Berry, Eds.). New York: Springer.

[16] Robins, James M. Miguel Angel Hernán and Babette Brumback. (2000). "Marginal structural models and causal inference in epidemiology," *Epidemiology* 11 (5): 550 - 560.

[17] Rosenbaum, Paul R. (1984). "The consequences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of the Royal Statistical Society A* 147 (5): 656 - 666.

[18] Rosenbaum, Paul R. (1987). "Model-based direct adjustment," *Journal of the American Statistical Association* 82 (398): 387 - 394.

[19] Rosenbaum, Paul R. and Donald B. Rubin. (1983). "The central role of the propensity score in observational studies for causal effects," *Biometrika* 70 (1): 41 - 55.

[20] Rubin, Donald B. (1977). "Assignment to treatment group on the basis of a covariate," *Journal of Educational Statistics* 2 (1): 1 - 26.

[21] Sharkey, Patrick and Felix Elwert. (2011). "The legacy of disadvantage: multigenerational neighborhood effects on cognitive ability," *American Journal of Sociology* 116 (6): 1934 - 81.

[22] Taubman, Sarah L., James M. Robins, Murray A. Mittleman and Miguel A. Hernán. (2009). "Intervening on risk factors for coronary heart disease: an application of the parametric g-formula," *International Journal of Epidemiology* 38 (6): 1599 - 1611.

[23] Tchetgen, Eric J. Tchetgen and Ilya Shpitser. (2012). "Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness and sensitivity analysis," *Annals of Statistics* 40 (3): 1816 - 1845.

[24] Wodtke, Geoffrey T., David J. Harding and Felix Elwert. (2011). "Neighborhood effects in temporal perspective: the impact of long-term exposure to concentrated disadvantage on high school graduation," *American Sociological Review* 76 (5): 713 - 736.

[25] Wodtke, Geoffrey T. (2013). "Duration and timing of exposure to neighborhood poverty and the risk of adolescent parenthood," *Demography* 50 (5): 1765 - 1788.

[26] Wooldridge, Jeffrey M. (2005). "Unobserved heterogeneity and estimation of average partial effects," *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*: 27 - 55 (D.W.K. Andrews & J.H. Stock, Eds.). Cambridge: Cambridge University Press.

