Lecture 1: Covariate Adjustment

Bryan S. Graham, UC - Berkeley & NBER

September 4, 2015

The attempt by social scientists to uncover causal relationships with linear regression methods is a century old enterprise.[1] One hundred years later this enterprise remains as controversial as it was at its outset (e.g., Freedman, 1997). Two key challenges arise. The first is foundational. The second is technical, but by no means trivial. The second will be the focus of this lecture, but we will nevertheless begin with a quick overview of the first.

When can association be used to infer causation?

Let Y be an outcome of interest, X a policy variable (which may be vector valued with discrete and/or continuous elements), and W a vector of "control" variables or observed "confounders". Under random sampling of $Z = (W', X', Y)'$ the econometrician can eventually learn the *proxy variable regression* (PVR) function

$$\mathbb{E}\left[Y \middle| W = w, X = x\right] = q\left(w, x\right) \tag{1}$$

at all points in the joint support of W and X.

From (1) the econometrician can then compute the contrast

$$\Delta\left(x', x; w\right) = q\left(x', w\right) - q\left(x, w\right) \tag{2}$$

for any (x, w) and (x', w) pair in the joint support.

This contrast corresponds to the difference in average outcomes between two subpopulations. These two subpopulations are *identical* in terms of W, but *differ* in terms of the "policies" X to which they have been exposed.

[1] An early reference is George Udny Yule's (1897) investigation into the causes of poverty ("pauperism") in England.

Our first, foundational, question is when does $\Delta(x', x; w)$ also measure the average effect (on the subpopulation with $W = w$) of intervening to *change* the policy from x to x'? The proxy variable regression provides us with a measure of association; does this measure also predict the effects of interventions?

Rubin (1977) provided conditions under which (2) has the desired causal interpretation for the case where X is binary-valued (see also Barnow, Cain and Goldberger (1980) and the references therein). The extension of his basic argument to the case where X is vector valued with discrete and/or continuous elements is straightforward and sketched here. There are a number of subtleties involved in the argument, which I ignore. Holland (1986) provides a useful, if somewhat dated, reference.

I begin by positing the existence of an individual-specific potential response function

$$Y(x) = m(x, U). \tag{3}$$

The function $Y_i(x)$ gives individual $i's$ potential outcome to policy x. Individual $i's$ observed outcome, in a slight abuse of notation, is

$$Y_i = Y_i(X_i), \tag{4}$$

where X_i is the actual policy she faces. The right-hand-side of (3) is a structural equation representation of the response function. Here U is a (potentially very large) vector of pre-policy individual-attributes that generate heterogeneity in responses. The structural equation representation is without loss of generality since we are free to make U as rich as needed.

With this notation we can re-write 2 as

$$\Delta(x', x; w) = \mathbb{E}[m(x', U) | W = w, X = x'] - \mathbb{E}[m(x, U) | W = w, X = x].$$

The first expectation to the right of the equality is an average over U within the $W = w$ and $X = x'$ subpopulation, the second is an average over U within the $W = w$ and $X = x$ subpopulation. A causal interpretation of $\Delta(x', x; w)$ requires that the distribution of U is identical in these two subpopulations. If agents exercise some control over the policy they face, then equality of these two distributions may be difficult to justify.

If the policy X varies independently of *all other determinants* of outcome heterogeneity U conditional on the unobserved controls W :

$$U \perp X | W, \tag{5}$$

then

$$\begin{aligned}\mathbb{E}\left[Y\middle|W,X=x\right] &= \mathbb{E}\left[m\left(x,U\right)\middle|W,X=x\right] \\ &= \mathbb{E}\left[m\left(x,U\right)\middle|W\right],\end{aligned}$$

with the second line an implication of (5). Under (5) the difference (2) now equals

$$\Delta\left(x',x;w\right)=\mathbb{E}\left[m\left(x',U\right)-m\left(x,U\right)\middle|W=w\right],$$

which does equal the average effect of an intervention on X from $X=x$ to $X=x'$ (within the subpopulation homogenous in $W=w$).

Condition (5) is rather strong. It implies, for example, that W includes *all* variables which simultaneously determine the outcome and selection into different policies. Certain assignment mechanisms ensure (5) holds; for example if policies are randomly assigned conditional on W.

It can sometimes be helpful to think of (5) as an 'as if' conditional random assignment assumption. This can be useful for evaluating the credibility of a particular empirical exercise. It can also be limiting for economic applications, where it may be natural to assume that agents choose X purposefully.

As an example, inspired by Imbens (2004), assume that agents choose the input X to maximize expected profits:

$$X=\arg\max_{x\in\mathbb{R}^1}\mathbb{E}\left[m\left(x,U\right)-c\left(x,W,V\right)\middle|W,V\right].$$

Here we associate the outcome $Y=m\left(X,U\right)$ with revenue and $c\left(x,W,V\right)$ is a cost function. The agent's information set consists of W – which is observed by the econometrician – and V – which is not. Here the agent knows her cost function perfectly, but she may only imperfectly predict the revenue effects of different input choices.

To keep the discussion simple, assume that the problem is well-defined with unique solution

$$X=k\left(W,V\right).$$

If $V\perp U|W$, then condition (5) will hold. In words: if the unobserved determinants of input costs (V) vary independently of the unobserved determinants of the outcome (U) *conditional* on the observed confounders W, then condition (5) holds. This example suggests that any variables which predict *both* cost and outcome heterogeneity would be useful to include in W.

Average structural function (ASF)

Given the potential outcome structure (3), condition (5) implies that within subpopulations homogenous in W, the observed association between outcomes Y and inputs X coincides with the hypothetical effect that we would have observed if X were, instead, intervened upon directly.

It is useful to summarize these conditional on W effects by averaging them. The *average structural function* (ASF) coincides with the expected outcome associated with assignment to policy $X = x$ for a random draw from the population

$$m^{\text{ASF}}(x) = \mathbb{E}[m(x, U)]. \tag{6}$$

When X is binary, the difference

$$\begin{aligned} \text{ATE} &= m^{\text{ASF}}(1) - m^{\text{ASF}}(0) \\ &= \mathbb{E}[m(1, U) - m(0, U)] \\ &= \mathbb{E}[Y(1) - Y(0)] \end{aligned}$$

coincides with the familiar average treatment effect (ATE) estimand from the program evaluation literature (e.g., Imbens, 2004).

Under condition (5) we can recover the ASF by "covariate adjustment". It is the mechanics of doing so that will occupy us here. A key result is that, under (5), the ASF has a *partial mean* representation. Newey (1994) defines a partial mean to be the average of a conditional expectation function over the marginal distribution of some covariates, holding the others fixed. Consider the partial mean of the proxy variable regression function over the marginal distribution of confounders:

$$\begin{aligned} \mathbb{E}[q(W, x)] &= \mathbb{E}[\mathbb{E}[Y | W, X = x]] \\ &= \mathbb{E}[\mathbb{E}[m(x, U) | W, X = x]] \\ &= \mathbb{E}[\mathbb{E}[m(x, U) | W]] \\ &= \mathbb{E}[m(x, U)]. \end{aligned} \tag{7}$$

The first equality follows by the definition of the PVR. The second by (3), the third by condition (5), and the fourth by the law-of-iterated expectations.

For the equality $\mathbb{E}[q(W, x)] = \mathbb{E}[m(x, U)]$ to be well-defined we require a support condition.

Let $\mathbb{S}_W(x)$ denote the set of w values observed among individuals assigned to policy $X = x$:

$$\mathbb{S}_W(x) = \{w : f(w|x) > 0\}.$$

Let $\mathbb{W}$ denote the marginal support of W. We require that these two supports coincide

$$\mathbb{S}_W(x) = \mathbb{W}, \tag{8}$$

at any x for which we seek to learn $m^{\text{ASF}}(x)$. Condition (8) ensures that for any group of units defined in terms of W, at least some experience policy $X = x$. This condition implies that $\mathbb{E}[q(W, x)]$ is identified by the joint distribution of W, X and Y.

It is helpful to connect these arguments to those you may be familiar with from the program evaluation literature. When $X \in \{0, 1\}$ is binary we have, by Bayes' Law, $f(w|x = 1) = p(w) f(w)/Q$ with $p(w) = \Pr(X = x|W = w)$ and $Q = \Pr(X = 1)$. To learn about the ASF at $X = 1$ we require that the propensity score $p(w)$ is positive for all $w \in \mathbb{W}$. To learn about the ASF at $X = 0$ a similar argument implies that the propensity score must be less that one at all values of w. Therefore, identifying the average treatment effect requires the *overlap* condition

$$0 < p(w) < 1$$

for all $w \in \mathbb{W}$.

Covariate adjustment

Imbens (2004) and Imbens and Wooldridge (2009) survey methods of covariate adjustment appropriate for when X is binary. I will not revisit this material here. Instead I wish to discuss how one might undertake covariate adjustment when X includes non-binary (including continuously-valued) components. Notwithstanding its relevance for empirical research, this is a surprisingly understudied area.

Newey (1994) presents kernel estimators for partial means like $\mathbb{E}[q(W, x)]$. These estimators converge at a slower than $\sqrt{N}$ rate, with the precise rate depending on the number of continuously valued components in X. This is not a limitation of methods, but rather one imposed by the problem. Under (3), (5), and (8) $\sqrt{N}$ consistent estimation of $m^{\text{ASF}}(x)$ is only possible when X is discretely-valued.

Under additional, semiparametric, restrictions on the potential response function (3), $\sqrt{N}$ estimation of the ASF may be possible. This is the approach that will be pursued here.

Specifically, I will developed results for when the potential response function is of the form

$$Y(x) = m(x, U) = x'\beta_0 + U. \tag{9}$$

Equation (9) is restrictive relative to (3) above. First, individuals respond to changes in X linearly. When X is scalar and binary-valued, linearity is not restrictive, but when X is vector-valued and/or includes non-binary components it generally *is* restrictive. Second, responses to changes in X are *homogenous* across units. All heterogeneity in the response function is confined to the intercept. While each individual in the population has their own response function, all such response functions are linear and parallel to one another. When X is binary-valued (9) is called a *constant additive treatment effect* assumption.

Clearly (9) is a strong assumption. It is introduced here primarily to make a difficult estimation problem easier. This, of course, does not mean it is a good assumption for empirical work. Under the null of no causal effect of X on Y, then (9) is not restrictive. An implication of this observation is that an analysis based on (9) can be used to test for causal effects.

To identify β_0 we need two main assumptions. The first is a specialization of (5) above.

Assumption 1. (Mean Independence) *For all $x \in \mathbb{X}$ and $w \in \mathbb{W}$*

$$\mathbb{E}[U|W = w, X = x] = \mathbb{E}[U|W = w] = h_0(w).$$

Note that we do not impose any restrictions of the form of $h_0(W)$. The second assumption specializes (8) above.

Assumption 2. (Conditional Variation of X) *Let $v_0(w) = \mathbb{V}(X|W = w)$; the matrix $\mathbb{E}[v_0(W)]$ is positive definite.*

This assumption implies that the policy X will vary conditional on $W = w$ for a non-trivial fraction of all possible subpopulations defined in terms of W.

The semiparametric model defined by (9) and Assumptions 1 and 2 coincides with the following partially linear model (PLM)

$$Y = X'\beta_0 + h_0(W) + V, \ \mathbb{E}[V|W, X] = 0. \tag{10}$$

Model (10) is among the most-widely studied semiparametric models (see Newey (1990) for an overview and references). A seminal example of its use in economics is provided by Olley and Pakes (1996) (cf., Wooldridge, 2009). Let Y_t be the log of firm output in period t, X_t the log of variable inputs, and W_{1t} the log of capital. Capital, a semi-fixed input, evolves

according to the law-of-motion

$$W_{1t} = (1-\delta) W_{1t-1} + W_{2t-1},$$

where δ is the rate-of-depreciation and W_{2t} is period t investment. The period t stock of capital depends on the prior period's capital stock and choice of investment level.

The production function is Cobb-Douglas, which after taking logs, yields the output equation

$$Y_t = X_t'\beta_0 + W_{1t}'\gamma_0 + A_t + U_t$$

where A_t is log productivity and U_t is an unforecastable production shock that is realized after input decisions are made:

$$\mathbb{E}\left[U_t \middle| W_t, X_t\right] = 0.$$

Olley and Pakes (1996) show that, under certain conditions, the firm's investment rule can be inverted to obtain:

$$A_t = g_{0t}\left(W_t\right).$$

Period t log productivity is some function of period t capital stock and investment levels. If we set $h_{t0}\left(W_t\right) = W_{1t}'\gamma_0 + g_{0t}\left(W_t\right)$ we have

$$Y_t = X_t'\beta_0 + h_{t0}\left(W_t\right) + U_t.$$

The elasticity of output with respect to variable inputs X_t can be recovered by a semiparametric regression of Y_t onto X_t and a nonparametric function of W_t (which includes both capital stock and investment). The Olley and Pakes (1996) example indicates that (10) can be given strong micro-foundations.

Let $\sigma_0^2\left(w,x\right) = \mathbb{V}\left(Y \middle| W = w, X = x\right)$, Chamberlain (1992) calculated a semiparametric efficiency bound for β_0 of

$$\mathcal{I}\left(\beta_0\right) = \mathbb{E}\left[\frac{XX'}{\sigma^2\left(W,X\right)}\right] - \mathbb{E}\left[\frac{\mathbb{E}\left[\frac{1}{\sigma^2(W,X)}X\middle| W\right]\mathbb{E}\left[\frac{1}{\sigma^2(W,X)}X\middle| W\right]'}{\mathbb{E}\left[\frac{1}{\sigma^2(W,X)}\middle| W\right]}\right], \tag{11}$$

with a corresponding *efficient score* of (cf., Ma, Chiou and Wang, 2006)

$$\mathbb{S}_\beta^{\text{eff}}\left(Z,\beta_0,g_0\left(W,X\right)\right) = \left(X - \frac{\mathbb{E}\left[\omega_0\left(W,X\right)X\middle| W\right]}{\mathbb{E}\left[\omega_0\left(W,X\right)\middle| W\right]}\right)\omega_0\left(W,X\right)\rho\left(Z,\beta_0,h_0\left(W\right)\right), \tag{12}$$

where

$$\rho\left(Z,\beta,h\left(W\right)\right)=Y-X'\beta-h\left(W\right)$$

and $\omega_0\left(w,x\right)=1/\sigma_0^2\left(w,x\right)$ and

$$g\left(W,X\right)=\left(h\left(W\right),\omega\left(W,X\right),\mathbb{E}\left[\omega_0\left(W,X\right)|\,W\right],\mathbb{E}\left[\omega_0\left(W,X\right)X|\,W\right]\right)'.$$

An outline of the derivation of (11) and (12) is provided in Appendix A.

For the balance of what follows I will assume that the conditional variance of Y given W and X is constant (i.e., homoscedasticity), but that this fact is not part of the prior restriction used to calculate the efficiency bound. The inference procedures we develop will not be sensitive to the homoscedasticity assumption.

Under homoscedasticity the efficient score simplifies to

$$\mathbb{S}^{\text{eff}}\left(Z,\beta_0,e_0\left(W\right),h_0\left(W\right)\right)=\frac{X-e_0\left(W\right)}{\sigma_0^2}\rho\left(Z,\beta_0,h_0\left(W\right)\right), \tag{13}$$

where $e_0\left(w\right)=\mathbb{E}\left[X|\,W=w\right]$ is the mean of the policy variable given confounders. When X is scalar and binary-valued $e_0\left(w\right)$ is the propensity score. I will call $e_0\left(w\right)$ the generalized propensity score here.

For estimation it will be convenient to impose a parametric restriction on $e_0\left(w\right)$. Because the distribution of X given W is ancillary for β_0, this restriction does not change the information bound (cf., Newey, 1990).

Assumption 3. (Generalized Propensity Score) *$f\left(x|\,w;\phi\right)$ is a parametric family of densities indexed by $\phi\in\Phi\subset\mathbb{R}^{\dim(\phi)}$ with (i) $f_0\left(x|\,w\right)=f(x|\,w;\phi_0)$ at some unique $\phi_0\in\text{int}\left(\Phi\right)$, (ii) a maximum likelihood estimate (MLE) of ϕ_0 equal to*

$$\hat{\phi}=\arg\max_{\phi\in\Phi}\sum_{i=1}^{N}\ln f\left(X_i|\,W_i;\phi\right)$$

with a score vector of $\mathbb{S}_\phi\left(X|\,W;\phi\right)=\nabla_\phi f\left(X|\,W;\phi\right)/f\left(X|\,W;\phi\right)$, (iii) $\hat{\phi}\xrightarrow{p}\phi_0$ with $\mathbb{E}\left[\mathbb{S}_i\mathbb{S}_i'\right]$ non-singular and the asymptotically linear representation

$$\sqrt{N}\left(\hat{\phi}-\phi_0\right)=\mathbb{E}\left[\mathbb{S}_{\phi i}\mathbb{S}_{\phi i}'\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbb{S}_{\phi i}+o_p\left(1\right). \tag{14}$$

where $\mathbb{S}_{\phi i}=\mathbb{S}_\phi\left(X_i|\,W_i;\phi_0\right)$.

Sometimes Assumption 3 can be made to hold by design; as in a randomized experiment

with known assignment rule. In other situations Assumption 3 reflects the fact that our knowledge about the nature of selection is stronger than that regarding the form of the outcome equation. That is, our prior knowledge regarding the form of $e_0(w)$ is sharper than it is for $h_0(w)$ (cf., Robins, Mark and Newey, 1992).

In principle we could proceed without Assumption 3. Chamberlain (1986) and Robinson (1988) provide estimators for this case. However, if W includes many components, as is common in empirical work, these methods are impractical as they involve preliminary high-dimensional nonparametric estimation steps.[2]

E-Estimation

I will work under (9) and Assumption 1, 2 and 3. Let $e\left(w;\hat{\phi}\right)$ denote the maximum likelihood estimate of the generalized propensity score. Newey (1990) proposed the estimator:

$$\hat{\beta}_{\mathrm{E}} = \left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - e\left(W_i;\hat{\phi}\right)\right)X_i'\right]^{-1} \times \left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - e\left(W_i;\hat{\phi}\right)\right)Y_i\right]. \tag{15}$$

Robins, Mark and Newey (1992) explore the properties of (15) in detail, calling $\hat{\beta}_{\mathrm{E}}$ the "E-Estimate". Computation of $\hat{\beta}_{\mathrm{E}}$ is straightforward:

Algorithm 1. E-Estimation

1. Compute the maximum likelihood estimate of ϕ_0. Construct $e\left(W_i,\hat{\phi}\right)$ for $i = 1,\ldots,N$.
2. Compute the linear instrumental variables regression fit of Y_i onto $X_i - e\left(W_i,\hat{\phi}\right)$ using X_i as the instrument (and excluding the constant term). The coefficient on $X_i - e\left(W_i,\hat{\phi}\right)$ equals $\hat{\beta}_{\mathrm{E}}$.

To better understand the advantages and disadvantages of E-Estimation it is helpful to consider an alternative approach. Covariate adjustment in the semiparametric model defined by (9) and Assumption 1, 2 and 3 typically proceeds as follows. The researcher first augments her prior by assuming a *specific* functional form for $h_0(w) = h(w;\eta_0)$. The most common assumption is that $h(w;\eta) = w'\eta$. Under this assumption the ordinary least squares fit of Y onto X and W provides a consistent estimate of β_0. By the results of Chamberlain (1987), the OLS estimate is also semiparametrically efficient (under homoscedasticity).

[2] Belloni, Chernozhukov and Hansen (2014) study estimation under the assumption that $e_0(w)$ and $h_0(w)$ are well-approximated by a low-dimensional vector of basis functions.

Now consider an approach which, like E-Estimation, does not impose any prior restrictions on the form of $h_0(w)$. It is clear that such an approach can do no better, in terms of asymptotic precision, than the supremum of asymptotic variances across the set of all possible parametric submodels for $h_0(W)$. That is the best we can do without a parametric assumption on $h_0(w)$, can be no better than the worst we can do when making such an assumption. Chamberlain (1992) showed that this supremum is given by the inverse of (11), which, when evaluated under the homoscedasticity assumption, equals

$$\mathcal{I}(\beta_0)^{-1} = \sigma^2 / \mathbb{E}[v_0(W)]. \tag{16}$$

To summarize: if we correctly specify $h(w;\eta_0)$ we can do no worse than (16), while if we cannot *a priori* restrict the form of $h_0(w)$ we can do no better than (16).

If we do, indeed, have sharp prior information about the form of $h_0(w)$, then we should incorporate that information into our estimation procedure. Unfortunately when W is very high-dimensional this is rarely the case. Furthermore if we base estimation on an incorrectly specified parametric model for $h_0(w)$, our estimate of β will generally be inconsistent. These considerations argue for an approach which does not require the analyst to make possibly untenable *a priori* assumptions about the form of $h_0(w)$. The E-estimator of Robins, Mark and Newey (1992) provides such an approach, albeit one which presumes sufficient prior knowledge regarding the form of the selection process to make maintaining Assumption 3 credible. This describes at least some empirical settings. Furthermore there are other principled arguments for basing identification of causal effects on the propensity score (cf., Imbens and Rubin, 2015).

Large sample properties of $\hat{\beta}_{\mathrm{E}}$

Let $m(Z,\phi,\beta) = (X - e(W;\phi))(Y - X'\beta)$. Some basic manipulation gives

$$\begin{aligned}\hat{\beta}_{\mathrm{E}} &= \left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - e\left(W_i;\hat{\phi}\right)\right)X_i'\right]^{-1} \\ &\quad\times\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - e\left(W_i;\hat{\phi}\right)\right)(Y - X'\beta_0)\right] \\ &\quad+\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - e\left(W_i;\hat{\phi}\right)\right)X_i'\right]^{-1} \\ &\quad\times\left[\frac{1}{N}\sum_{i=1}^{N}\left(X_i - e\left(W_i;\hat{\phi}\right)\right)X_i'\right]\beta_0 \\ &= \beta_0 + \mathbb{E}\left[v_0\left(W\right)\right]^{-1}\times\left[\frac{1}{N}\sum_{i=1}^{N} m\left(Z_i,\hat{\phi},\beta_0\right)\right] + o_p\left(1\right),\end{aligned}$$

where the last equality follows from Assumption 2, the LLN, and a Slutsky Theorem. A second mean value expansion in $\hat{\phi}$ about ϕ_0 gives

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}_{\mathrm{E}} - \beta_0\right) &= \mathbb{E}\left[v_0\left(W\right)\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m\left(Z_i,\hat{\phi},\beta_0\right) \\ &= \mathbb{E}\left[v_0\left(W\right)\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m\left(Z_i,\phi_0,\beta_0\right) \\ &\quad+\mathbb{E}\left[v_0\left(W\right)\right]^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial m\left(Z_i,\phi_0,\beta_0\right)}{\partial\phi}\right\}\sqrt{N}\left(\hat{\phi} - \phi_0\right) + o_p\left(1\right).\end{aligned}$$

Now observe that

$$\mathbb{E}\left[m\left(Z,\phi_0,\beta_0\right)|\, W = w\right] = \mathbb{E}\left[\left(X - e\left(W;\phi_0\right)\right)\left(Y - X'\beta_0\right)|\, W = w\right] = 0,$$

or in integral form:

$$\int m\left(z,\phi_0,\beta_0\right) f_0\left(y|\, w,x\right) f\left(x|\, w;\phi_0\right)\mathrm{d}x\mathrm{d}y = 0. \tag{17}$$

Differentiating (17) through the integral with respect to ϕ gives:

$$\begin{aligned}\int\frac{\partial}{\partial\phi'}m\left(z,\phi_0,\beta_0\right) f_0\left(y|\, w,x\right) f\left(x|\, w;\phi_0\right)\mathrm{d}x\mathrm{d}y &= -\int m\left(z,\phi_0,\beta_0\right)\mathbb{S}_{\phi}\left(x|\, w;\phi_0\right)' \\ &\quad\times f_0\left(y|\, w,x\right) f\left(x|\, w;\phi_0\right)\mathrm{d}x\mathrm{d}y,\end{aligned}$$

or, equivalently,

$$\mathbb{E}\left[\frac{\partial m\left(Z,\phi_0,\beta_0\right)}{\partial\phi}\middle| W=w\right]=-\mathbb{E}\left[m\left(Z,\phi_0,\beta_0\right)\mathbb{S}_\phi'\middle| W=w\right], \tag{18}$$

which is a Generalized Information Matrix Equality (GIME) result (e.g., Newey, 1990, p. 104).

Using (14) and (18) we have

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}_{\mathrm{E}}-\beta_0\right) &= \mathbb{E}\left[v_0\left(W\right)\right]^{-1}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}m_i\right.\\ &\quad\left.-\mathbb{E}\left[m\mathbb{S}_\phi'\right]\mathbb{E}\left[\mathbb{S}_\phi\mathbb{S}_\phi{}'\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbb{S}_{\phi i}\right\}+o_p\left(1\right)\\ &= \frac{\mathbb{E}\left[v_0\left(W\right)\right]^{-1}}{\sqrt{N}}\sum_{i=1}^{N}\left\{m_i-\mathbb{E}\left[m\mathbb{S}_\phi'\right]\mathbb{E}\left[\mathbb{S}_\phi\mathbb{S}_\phi'\right]^{-1}\mathbb{S}_{\phi i}\right\}+o_p\left(1\right)\end{aligned} \tag{19}$$

for $m_i=m\left(Z_i,\phi_0,\beta_0\right)$.

One implication of the asymptotically linear representation (19) is that practitioners can ignore sampling error in $\hat{\phi}$ and get conservative confidence intervals. Similar results have been shown numerous times for the binary X case (e.g., Robins, Rotnitzky and Zhao, 1994; Wooldridge, 2007). A second implication is that over-parameterizing the conditional distribution of X given W will not decrease asymptotic precision.

It turns out that we can get a more insightful expression of (19). One that also suggests valuable guidelines for applied work. Observe that $m\left(Z,\phi_0,\beta_0\right)=\psi_\beta\left(Z,\beta_0,\phi_0,h_0\left(W\right)\right)+r\left(W,X,\phi_0,h_0\left(W\right)\right)$ with

$$\psi_\beta\left(Z,\beta,\phi,h\left(W\right)\right)=\left(X-e\left(W;\phi\right)\right)\left(Y-X'\beta-h\left(W\right)\right) \tag{20}$$

and

$$r\left(W,X,\phi,h\left(W\right)\right)=\left(X-e\left(W;\phi\right)\right)h\left(W\right). \tag{21}$$

Let $r_0\left(W,X\right)=r\left(W,X,\phi_0,h_0\left(W\right)\right)$ and note that $\mathbb{E}\left[r_0\left(W,X\right)\middle| W\right]=0$. Note further that $\mathbb{S}_\phi$ is also conditionally mean zero given W. Next observe that

$$\mathbb{E}\left[\frac{\partial\psi_\beta}{\partial\phi}\middle| W\right] = -\frac{\partial e\left(W;\phi_0\right)}{\partial\phi}\mathbb{E}\left[\rho\left(Z,\beta_0,h_0\left(W\right)\right)\middle| W\right]=0.$$

With these preliminaries in place, start with the integral

$$\int \psi_\beta f_0(y|x,w) f(x|w;\phi_0) f_0(w) = 0$$

and differentiate with respect to ϕ:

$$\int \frac{\partial \psi_\beta}{\partial \phi'} f_0(y|x,w) f(x|w;\phi_0) f_0(w) = -\int \left\{\psi_\beta \mathbb{S}_\phi'\right\} f_0(y|x,w) f(x|w;\phi_0) f_0(w) = 0.$$

Putting all these pieces together gives

$$\mathbb{E}\left[m\mathbb{S}_\phi'\right] = \mathbb{E}\left[\psi_\beta \mathbb{S}_\phi'\right] + \mathbb{E}\left[r\mathbb{S}_\phi'\right] = \mathbb{E}\left[r\mathbb{S}_\phi'\right].$$

Plugging this into our influence function (19) we get

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}_{\mathrm{E}} - \beta_0\right) &= \frac{\mathbb{E}\left[v_0(W)\right]^{-1}}{\sqrt{N}} \sum_{i=1}^N \left\{m_i - \mathbb{E}\left[m\mathbb{S}_\phi'\right]\mathbb{E}\left[\mathbb{S}_\phi\mathbb{S}_\phi'\right]^{-1}\mathbb{S}_{\phi i}\right\} + o_p(1) \\ &= \frac{\mathbb{E}\left[v_0(W)\right]^{-1}}{\sqrt{N}} \sum_{i=1}^N \left\{\psi_{\beta i} + \left[r_i - \mathbb{E}\left[r\mathbb{S}_\phi'\right]\mathbb{E}\left[\mathbb{S}_\phi\mathbb{S}_\phi'\right]^{-1}\mathbb{S}_{\phi i}\right]\right\} + o_p(1) \quad (22)\end{aligned}$$

An immediate implication of (22) is

$$\sqrt{N}\left(\hat{\beta}_{\mathrm{E}} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \mathcal{I}(\beta_0)^{-1} + \mathbb{E}\left[(r - \Pi_{r\mathbb{S}}\mathbb{S}_\phi)(r - \Pi_{r\mathbb{S}}\mathbb{S}_\phi)'\right]\right) \quad (23)$$

with $\Pi_{r\mathbb{S}} = \mathbb{E}\left[r\mathbb{S}_\phi'\right]\mathbb{E}\left[\mathbb{S}_\phi\mathbb{S}_\phi'\right]^{-1}$. Here $\mathcal{I}(\beta_0)$ corresponds to the information bound (16) evaluated under the auxiliary homoscedasticity assumption.

Local semiparametric efficiency

In general $\hat{\beta}_{\mathrm{E}}$ will not attain the bound for the model defined by (9) and Assumption 1, 2 and 3 (where the bound is evaluated under the homoscedasticity assumption). However the form of the asymptotic variance function provides insight into the structure of inefficiency and, consequently, how to construct approximately efficient estimators.

For concreteness assume that a generalized linear model (GLM) with a canonical response function is used to model the generalized propensity score. Let $k(w)$ be the vector of linearly independent basis functions entering the GLM link function. For example, in the Poisson case $e(w;\phi) = \exp\left(k(w)'\phi\right)$. Given this GLM structure we have (assuming X is scalar to

keep the notation simple)

$$\mathbb{S}_\phi = (X - e(W;\phi_0))\, k(W). \tag{24}$$

Now consider the following assumption.

Assumption 4. (Response Model) $h_0(w) = k(w)'\pi_0$ *for all* $w \in \mathbb{W}$.

Under (24) and Assumption 4 (again assuming X is scalar to keep the notation simple)

$$\begin{aligned}\mathbb{E}\left[r\mathbb{S}_\phi'\right] &= \mathbb{E}\left[(X - e(W;\phi_0))^2 h_0(W)\, k(W)'\right] \\ &= \pi_0'\mathbb{E}\left[(X - e(W;\phi_0))^2 k(W)\, k(W)'\right] \\ &= \pi_0'\mathbb{E}\left[\mathbb{S}_\phi\mathbb{S}_\phi'\right]\end{aligned}$$

and hence

$$\mathbb{E}\left[(r - \Pi_{r\mathbb{S}}\mathbb{S}_\phi)(r - \Pi_{r\mathbb{S}}\mathbb{S}_\phi)'\right] = 0.$$

The E-Estimator is *locally efficient* at (24) and Assumption 4. By locally efficient I mean that $\hat{\beta}_{\mathrm{E}}$ is a consistent estimate of β_0 in the semiparametric model defined by (9) and Assumption 1, 2 and 3. If Assumption 4 also "happens to be true" in the population sampled from (but is not part of the prior restrictions), then $\hat{\beta}_{\mathrm{E}}$ attains the semiparametric efficiency bound of (16). See Newey (1990) or Tsiatis (2006) for more details on the local efficiency concept.

Since, in general, the GLM with canonical link assumption is a natural one for practitioners, the main implication of this results is that one should include good predictors of the outcome variable as well as the policy variable in the generalized propensity score model.

Algorithm 2. Locally Efficient E-Estimation

1. Assume that the generalized propensity score takes the GLM with canonical response form. Let the vector of basis functions in the response, $k(w)$, be the union of all linearly independent predictors of both X and Y.
2. Compute the maximum likelihood estimate of ϕ_0. Construct $e\left(W_i, \hat{\phi}\right)$ for $i = 1, \ldots, N$.
3. Compute the linear instrumental variables regression fit of Y_i onto $X_i - e\left(W_i, \hat{\phi}\right)$ using X_i as the instrument (and excluding the constant term). The coefficient on $X_i - e\left(W_i, \hat{\phi}\right)$ equals $\hat{\beta}_{\mathrm{E}}$.

This E-Estimate is consistent for β_0 under (9) and Assumption 1, 2 and 3. It is locally efficient at Assumption 4.

Double Robustness

In turns out that consideration of Assumption 4 has a benefit which extends beyond efficiency. Suppose that the researcher incorrectly specifies the generalized propensity score (i.e., Assumption 3 does not hold in the population), but that Assumption 4 is true. In such a situation $\hat{\beta}_{\mathrm{E}}$ continues to provide a consistent estimate of β_0. We say that the E-Estimate is *doubly robust*.

To understand this property consider the population mean of the efficient score where $e_0(w)$ is replaced with some other function $e_*(w)$:

$$\begin{aligned}\mathbb{E}\left[\frac{X-e_*(W)}{\sigma^2}\left(Y-X'\beta_0-h_0(W)\right)\right] &= \mathbb{E}\left[\frac{X-e_*(W)}{\sigma^2}\mathbb{E}\left[\left(Y-X'\beta_0-h_0(W)\right)\middle|W,X\right]\right]\\ &= 0.\end{aligned}$$

Hence a method of moments estimator based on the quasi-efficient score, with an misspecified generalized propensity score estimate but the true $h_0(W)$, remains consistent for β_0.

Let $\hat{\phi}$ be the quasi-MLE estimate of $\phi_* \neq \phi_0$. By the properties of the first order conditions of GLM problem we have that $\sum_{i=1}^N \left(X_i - e\left(W_i;\hat{\phi}\right)\right) k(W_i) = 0$. The E-Estimate solves

$$\begin{aligned}0 &= \frac{1}{N}\sum_{i=1}^N \left(X_i - e\left(W_i;\hat{\phi}\right)\right)\left(Y - X_i'\hat{\beta}_{\mathrm{E}}\right)\\ &= \frac{1}{N}\sum_{i=1}^N \left(X_i - e\left(W_i;\hat{\phi}\right)\right)\left(Y - X_i'\hat{\beta}_{\mathrm{E}} - k(W_i)'\pi_0\right),\end{aligned}$$

where the second line follows from the fact that $\sum_{i=1}^N \left(X_i - e\left(W_i;\hat{\phi}\right)\right) k(W_i) = 0$. Note that $\hat{\beta}_{\mathrm{E}}$ is precisely the parameter value which sets the sample mean of the quasi-efficient score to zero.

Covariate adjustment in empirical work

A substantial portion of applied work involves computing the least squares fit of Y onto a vector of policy variables of interest X and additional controls W. When $e_0(w) = w\Pi_{xw}$ is linear in W, the E-Estimate and OLS estimate of β_0 will coincide; however, in generally they will differ. The advantage of E-Estimation is that it provides a principled way for a researcher to (i) incorporate her knowledge about the selection process into estimation, (ii)

make her inferences robust to misspecification of the outcome equation and (iii), through the double robustness property, generate two chances for correct inference. These virtues argue for greater use of E-Estimation in empirical research.

Under (9) and Assumption 1, 2 and 3 heteroscedastic robust confidence intervals computed by a linear instrumental variables program, which implicitly ignore the effects of estimation error in $e\left(W_i;\hat{\phi}\right)$, will be conservative (i.e., have asymptotic coverage greater than $1-\alpha$). If Assumption 4 additionally holds, these confidence intervals will have correct large sample coverage. In the case where Assumption 3 is false, but Assumption 4 is true, failing to correct for sampling error in $e\left(W_i;\hat{\phi}\right)$ may result in undercoverage in large samples. In practice the Bayesian Bootstrap can be used to construct confidence intervals which incorporate all sources of sampling error.

A Derivation of semiparametric efficiency bound

Chamberlain (1992) derived the semiparametric efficiency bound (SEB) for β_0 in the model defined by (9) and Assumption 1 and 2, using a multinomial approximation argument. In this Appendix I sketch the SEB calculation using the general approach developed by Bickel, Klaasen, Ritov and Wellner (1993) as exposited by Newey (1990, Section 3). First, I characterize the nuisance tangent space. Second, I calculate the residual associated with the projection of the score function for β onto the nuisance tangent space. The form of the efficient influence function and variance bound then follows from Theorem 3.2 of Newey (1990).

Step 1: Characterization of the nuisance tangent space

The joint density of $z=(w,z,y)$ is given by

$$f_0(w,x,y) = f_0(y|\,w,x)\,f_0(w,x).$$

Assumption 1 also requires that $f_0(y|\,w,x)$ satisfy the conditional moment restriction

$$\int \rho\left(z,\beta_0,h_0(w)\right) f_0(y|\,w,x)\,\mathrm{d}y = 0,$$

where

$$\rho\left(z,\beta,h(w)\right) = y - x'\beta - h(w).$$

Let $\theta=(\beta',\eta')'$ be the parameters of a parametric submodel with associated score vector

$\mathbb{S}_\theta = \left(\mathbb{S}_\beta', \mathbb{S}_\eta'\right)'$ partitioned conformably. We have $f(w,x,y;\theta) = f_0(w,x,y)$ at $\theta = \theta_0$. The submodel also satisfies the conditional moment restriction

$$\int \rho\left(z, \beta_0, h\left(w;\eta_0\right)\right) f\left(y \middle| w,x;\theta_0\right) \mathrm{d}y = 0. \tag{25}$$

The submodel score vector equals

$$\mathbb{S}_\theta(w,x,y;\theta) = \mathbb{S}_\theta\left(y \middle| w,x\right) + \mathbb{T}_\theta(w,x;\theta)$$

where

$$\begin{aligned}
\mathbb{S}_\theta(w,x,y;\theta) &= \nabla_\theta \log f(w,x,y;\theta), \\
\mathbb{S}_\theta\left(y \middle| w,x;\theta\right) &= \nabla_\theta \log f\left(y \middle| w,x;\theta\right), \\
\mathbb{T}_\theta(w,x;\theta) &= \begin{pmatrix} 0 \\ \nabla_\eta \log f(w,x;\theta) \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbb{T}_\eta(w,x;) \end{pmatrix}.
\end{aligned}$$

The last line above follows because W and X are ancillary to β and hence their marginal density does not depend on β (i.e., $f(w,x;\theta) = f(w,x;\eta)$).

By the usual conditional mean zero property of scores we have

$$\mathbb{E}\left[\mathbb{S}_\theta\left(Y \middle| W,X\right) \middle| W,X\right] = \mathbb{E}\left[\mathbb{T}_\eta(W,X)\right] = 0, \tag{26}$$

where the suppression of (a sub-vector of) θ in a function means it is evaluated at its population value (e.g., $\mathbb{T}_\eta(W,X) = \mathbb{T}_\eta(W,X;\eta_0)$).

Condition (25) imposes additional restrictions on the form of $\mathbb{S}_\theta\left(Y \middle| W,X\right)$ beyond (26). To see the structure of these restrictions differentiate (25) with respect to β and η and evaluate the result at $\theta = \theta_0$. This yields the pair of equalities

$$\begin{aligned}
X &= \mathbb{E}\left[\rho\left(Z,\beta_0,h_0(W)\right) s_\beta\left(Y \middle| W,X\right) \middle| W,X\right] \\
\frac{\partial h\left(W;\eta_0\right)}{\partial \eta} &= \mathbb{E}\left[\rho\left(Z,\beta_0,h_0(W)\right) s_\eta\left(Y \middle| W,X\right) \middle| W,X\right].
\end{aligned} \tag{27}$$

Restrictions (26) and (27) imply that the submodel score vector takes the form

$$\begin{aligned}
\mathbb{S}_\beta(W,X,Y) &= -\frac{X}{\sigma_0^2(W,X)} \rho\left(Z,\beta_{,0} h_0(W)\right) \\
\mathbb{S}_\eta(W,X,Y) &= -\frac{\partial h\left(W;\eta_0\right)}{\partial \eta} \frac{1}{\sigma_0^2(W,X)} \rho\left(Z,\beta_{,0} h_0(W)\right) + \mathbb{T}_\eta(W,X).
\end{aligned} \tag{28}$$

Equations (26), (27) and (28) suggest that the nuisance tangent space is

$$\mathcal{T} = \left\{ \frac{k(W)}{\sigma_0^2(W,X)} \rho(Z, \beta_{,0} h_0(W)) + t(W,X) \right\}, \tag{29}$$

with $k(w)$ an unrestricted vector-valued function of w and $\frac{k(W)}{\sigma_0^2(W,X)} \rho(Z, \beta_{,0} h_0(W))$ and $t(W,X)$ satisfying

$$\begin{aligned} \mathbb{E}\left[\left\{ \frac{\rho(Z, \beta_{,0} h_0(W))}{\sigma_0^2(W,X)} \right\}^2 \|k(W)\|^2 \right] &< \infty \\ \mathbb{E}[t(W,X)] &= 0. \end{aligned}$$

Note also that, by Assumption1, $\frac{k(W)}{\sigma_0^2(W,X)} \rho(Z, \beta_{,0} h_0(W))$ is conditionally mean zero given W and X.

Step 2: Calculation of the efficient influence function

The nuisance tangent set (29) is the sum of two orthogonal components. Projection onto this sum therefore coincides with the sum of the projections onto each component alone. However, as $\mathbb{S}_\beta$ is orthogonal to $t(W,X)$ (since it is conditionally mean zero given W and X), we only need to calculate the projection onto the first component of (29). This projection coincides with the linear regression of $\mathbb{S}_\beta = -X \frac{\rho_0}{\sigma_0^2}$ onto the infinite dimensional vector of functions of the form $\frac{k(W)}{\sigma_0^2} \rho_0$ with $k(W)$ an arbitrary function of W (and where I let $\rho_0 = \rho(Z, \beta_{,0} h_0(W))$ and $\sigma_0^2 = \sigma_0^2(W,X)$ to economize on notation; albeit at the risk of some confusion). This projection coincides with the conditional linear predictor

$$\begin{aligned} \mathbb{E}^*\left[\mathbb{S}_\beta | \frac{\rho_0}{\sigma_0^2}; W \right] &= \mathbb{E}\left[\mathbb{S}_\beta \frac{\rho_0}{\sigma_0^2} \middle| W \right] \times \mathbb{E}\left[\left\{ \frac{\rho_0}{\sigma_0^2} \right\}^2 \middle| W \right]^{-1} \frac{\rho_0}{\sigma_0^2} \\ &= \mathbb{E}\left[\left\{ \frac{\rho_0}{\sigma_0^2} \right\}^2 X \middle| W \right] \times \mathbb{E}\left[\left\{ \frac{\rho_0}{\sigma_0^2} \right\}^2 \middle| W \right]^{-1} \left\{ \frac{\rho_0}{\sigma_0^2} \right\}^2 \\ &= \mathbb{E}\left[\frac{X}{\sigma_0^2} \middle| W \right] \times \mathbb{E}\left[\frac{1}{\sigma_0^2} \middle| W \right]^{-1} \frac{\rho_0}{\sigma_0^2}, \end{aligned}$$

where the last equality follows from the fact that $\mathbb{E}[\rho_0^2 | W, X] = \sigma_0^2$.
The efficient score is the residual vector $\mathbb{S}_\beta^{\text{eff}} = \mathbb{S}_\beta - \mathbb{E}^*\left[\mathbb{S}_\beta | \frac{\rho_0}{\sigma_0^2}; W \right]$, which evaluates to

$$\mathbb{S}_\beta^{\text{eff}}(Z, \beta_0, g_0(W,X)) = \left(X - \frac{\mathbb{E}[\omega_0(W,X) X | W]}{\mathbb{E}[\omega_0(W,X) | W]} \right) \omega_0(W,X) \rho(Z, \beta_0 h_0(W)),$$

as given by (12) in the main text. The information bound for β_0 is given by (cf., Theorem 3.2 of Newey (1990))

$$\begin{aligned}\mathcal{I}(\beta_0) &= \mathbb{E}\left[\mathbb{S}_\beta^{\text{eff}}\left(\mathbb{S}_\beta^{\text{eff}}\right)'\right] \\ &= \mathbb{E}\left[\left(X - \frac{\mathbb{E}[\omega_0 X|W]}{\mathbb{E}[\omega_0|W]}\right)\left(X - \frac{\mathbb{E}[\omega_0 X|W]}{\mathbb{E}[\omega_0|W]}\right)'\omega_0^2\rho_0^2\right] \\ &= \mathbb{E}[\omega_0 XX'] - \mathbb{E}\left[\frac{\mathbb{E}[\omega_0 X|W]\,\mathbb{E}[\omega_0 X|W]'}{\mathbb{E}[\omega_0|W]}\right],\end{aligned}$$

which also coincides with Chamberlain's (1992, p. 569) calculation.

References

[1] Barnow, Burt, Glen Cain and Arthur Goldberger (1980), "Issues in the Analysis of Selectivity Bias,"(1980). "," *Evaluation Studies Review Annual* 5: 42 - 59 (E. W. Stromsdorfer & G. Farkas, Eds.). SAGE Publications, Inc.

[2] Belloni, Alexandre. Victor Chernozhukov and Christian Hansen. (2014). "Inference on treatment effects after selection among high-dimensional controls," *Review of Economic Studies* 81 (2): 608 - 650.

[3] Bickel, Peter J., Chris A. J. Klaassen, Ya'acov Ritov, Jon A. Wellner. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* New York: Springer-Verlag.

[4] Chamberlain, Gary. (1986). "Notes on semiparametric regression," *Mimeo.*

[5] Chamberlain, Gary. (1987). "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics* 34 (3): 305 - 334.

[6] Chamberlain, Gary. (1992). "Efficiency bounds for semiparametric regression," *Econometrica* 60 (3): 567 - 596.

[7] Freedman, David. (1997). "From association to causation via regression," *Advances in Applied Mathematics* 18 (1): 59 – 110.

[8] Graham, Bryan S. (2011). "Efficiency bounds for missing data models with semiparametric restrictions," *Econometrica* 79 (2): 437 - 452.

[9] Hahn, Jinyong. (1998). "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica* 66 (2): 315 - 331.

[10] Holland, Paul W. (1986). "Statistics and causal inference," *Journal of the American statistical Association* 81 (396): 945 - 960.

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

[12] Imbens, Guido W. and Donald B. Rubin. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences An Introduction.* Cambridge: Cambridge University Press.

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

[14] Ma, Yanyuan, Jeng-Min Chiou and Naisyin Wang. (2006). "Efficient semiparametric estimator for heteroscedastic partially linear models," *Biometrika* 93 (1): - 75 - 84.

[15] Newey, Whitney K. (1990). "Semiparametric efficiency bounds," *Journal of Applied Econometrics* 5 (2): 99 - 135.

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

[17] Olley, Steven and Ariel Pakes. (1996). "The dynamics of productivity in the telecommunications equipment industry," *Econometrica* 64 (6): 1263 - 1297.

[18] Robins, James M., Steven D. Mark and Whitney K. Newey. (1992). "Estimating exposure effects by modeling the expectation of exposure conditional on confounders," *Biometrics* 48 (2): 479 - 495.

[19] Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. (1994). "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association* 89 (427): 846 - 866.

[20] Robinson, Peter. M. (1988). "Root-N-consistent semiparametric regression," *Econometrica* 56 (4): 931 - 954.

[21] Tsiatis, Anastasios A. (2006). *Semiparametric Theory and Missing Data.* New York: Springer.

[22] Wooldridge, Jeffrey M. (2003). "Estimating average partial effects under conditional moment independence assumptions, *CEMMAP Working Paper CWP03/04.*

[23] Wooldridge, Jeffrey M. (2007). "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics* 141 (2): 1281 - 1301.

[24] Wooldridge, Jeffrey M. (2009). "On estimating firm-level production functions using proxy variables to control for unobservables," *Economics Letters* 104 (3): 112 – 114.

[25] Yule, G. Udny. (1897). "An Investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades (Part I.)," *Journal of the Royal Statistical Society* 62 (2): 249 - 295.

