NBER WORKING PAPER SERIES

SEMIPARAMETRICALLY EFFICIENT ESTIMATION OF THE AVERAGE LINEAR REGRESSION FUNCTION

Bryan S. Graham
Cristine Campos de Xavier Pinto

Working Paper 25234
http://www.nber.org/papers/w25234

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2018

We thank Guido Imbens, Pat Kline, Tony Strittmatter and seminar participants at University College London, UC Berkeley and University of St. Gallen for helpful discussion. Financial support from NSF grant SES #1357499 is gratefully acknowledged. The initial draft of this paper was prepared in October of 2016. All the usual disclaimers apply. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Semiparametrically Efficient Estimation of the Average Linear Regression Function
Bryan S. Graham and Cristine Campos de Xavier Pinto
NBER Working Paper No. 25234
November 2018
JEL No. C14,C21,C31

ABSTRACT

Let Y be an outcome of interest, X a vector of treatment measures, and W a vector of pre-treatment control variables. Here X may include (combinations of) continuous, discrete, and/or non-mutually exclusive "treatments". Consider the linear regression of Y onto X in a subpopulation homogenous in $W = w$ (formally a conditional linear predictor). Let $b_0(w)$ be the coefficient vector on X in this regression. We introduce a semiparametrically efficient estimate of the average $\beta_0 = E[b_0(W)]$. When X is binary-valued (multi-valued) our procedure recovers the (a vector of) average treatment effect(s). When X is continuously-valued, or consists of multiple non-exclusive treatments, our estimand coincides with the average partial effect (APE) of X on Y when the underlying potential response function is linear in X, but otherwise heterogenous across agents. When the potential response function takes a general nonlinear/heterogenous form, and X is continuously-valued, our procedure recovers a weighted average of the gradient of this response across individuals and values of X. We provide a simple, and semiparametrically efficient, method of covariate adjustment for settings with complicated treatment regimes. Our method generalizes familiar methods of covariate adjustment used for program evaluation as well as methods of semiparametric regression (e.g., the partially linear regression model).

Bryan S. Graham
University of California - Berkeley
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
bgraham@econ.berkeley.edu

Cristine Campos de Xavier Pinto
Escola de Economia de São Paulo, FGV/SP
Rua Itapeva 474, sala 1200
São Paulo– SP, Brasil, 01332-000
cristine.pinto@fgv.br

A data appendix is available at http://www.nber.org/data-appendix/w25234
A Computer Code is available at https://github.com/bryangraham/ipt

Let Y be a scalar-valued outcome of interest, X a $K \times 1$ vector of policy variables, and W a $J \times 1$ vector of additional controls. For example Y might equal hours worked, X include the real wage rate *and* total unearned income ($K = 2$), and W be a vector of demographic measures capturing heterogeneity in preferences for work (e.g., Pencavel, 1986, Section 4). The goal is to summarize how Y – labor supply – covaries with X – the wage rate and unearned income – "holding the controls W fixed". In a second example, Y might be an end-of-year student mathematics achievement measure, X a vector containing (i) number of days absent from school, (ii) class size and (iii) an indicator for whether the student received supplemental tutoring. Here the vector W might include beginning of school year joint predictors of Y and X (e.g., prior mathematics achievement, socioeconomic background, health indicators, and known determinants of class size and tutoring assignment used by the school). The goal is to summarize how math achievement covaries with attendance, class size and supplemental tutoring conditional on W (cf., Gottfried & Kirksey, 2017).

Following the prototype established by Yule (1899) over one hundred years ago, social scientists typically report the coefficient on X in the (long) least squares fit of Y onto a constant, X, *and* W for this purpose.

When X is a scalar binary variable, the econometrician can choose from – in addition to least squares – an ever more elaborate menu of covariate adjustment methods (see Imbens & Rubin (2015) for a recent textbook introduction). Many of these methods extend naturally to settings where X is multi-valued (e.g., Cattaneo, 2010).

When X is continuously-valued, and/or consists of multiple distinct policy variables ($K \geq 2$), options are fewer (cf., Wooldridge, 2010, Chapter 21.6.3). The partially linear regression (PLM) model

$$Y = X'\beta_0 + h_0(W) + U, \ \mathbb{E}[U|W, X] = 0, \tag{1}$$

represents one semiparametric generalization of (long) linear regression. Chamberlain (1986), in an influential but never published paper, introduced an estimator for β_0 in (1) (cf., Robinson, 1988). In later work he characterized its semiparametric efficiency bound (SEB) (Chamberlain, 1992).

Partially linear regression is widely, albeit heuristically, used in empirical work. Typically researchers proceed by (i) choosing W to be a rich vector of basis functions in the underlying controls (e.g., a vector of polynomial or piecewise polynomial terms) and then (ii) estimate β_0 by least squares. With discretely-valued control variables a saturated specification for $h_0(W)$ is possible, at least when utilizing a very large dataset (e.g., Angrist & Krueger, 1999, Section 2.3.1). A principled variant of this general approach is embodied in the E-Estimation algorithm of Newey (1990) and Robins et al. (1992).

In this paper we propose a different approach to covariate adjustment. Consider a subpopulation homogenous in $W = w$. Within this subpopulation we compute the linear regression of Y onto a constant and X (formally a conditional linear predictor as in Wooldridge (1999)). Let $b_0(w)$ be the coefficient on X in the conditional linear regression for the subpopulation homogenous in $W = w$. We propose a method for identifying and efficiently estimating the average regression coefficient

$$\beta_0 = \mathbb{E}[b_0(W)]. \tag{2}$$

The average is over the marginal distribution of controls, W.

In the absence of controls, the relationship between the linear predictor slope coefficient and the gradient of the (possibly nonlinear) conditional expectation function (CEF) of Y given $X = x$ is well-understood (e.g., Goldberger, 1991; Yitzhaki, 1996). In the presence of controls, this relationship is rather more complicated (cf., Angrist, 1998; Sloczynski, 2017). Our focus on averages of conditional linear predictor coefficients allows for conditioning on W, while also preserving the interpretative transparency of unconditional linear analyses. That is, β_0, as we demonstrate below, is easy to interpret.

When X is binary-valued (multi-valued) β_0 coincides with the (a vector of) average treatment effect(s); estimands familiar from the program evaluation literature (e.g., Hahn, 1998; Imbens, 2000). These estimands have causal interpretations under certain conditions. Modestly extending the analysis of Wooldridge (2004), we show that this causal interpretation generalizes under a (i) heterogenous random coefficients potential outcome structure and (ii) an unconfoundedness-type assumption. These assumptions coincide with their program evaluation counterparts when X is binary- or multi-valued. Our semiparametric model includes both the program evaluation model and the partially linear regression model as special cases.

Our work is also connected to the varying coefficient model of Hastie & Tibshirani (1993). Hastie & Tibshirani (1993) focus on pointwise estimation of $b_0(w)$, while we focus on (efficient) estimation of the average $\beta_0 = \mathbb{E}[b_0(W)]$.

The relationship of our work with that of Wooldridge (2004) is as follows.[2] We both study the same functional of the joint distribution of W, X and Y (see Equation (7) below). Relative to Wooldridge (2004) we provide an average partial effect interpretation of this estimand under (i) weaker assumptions when maintaining a correlated random coefficient potential outcome structure and (ii) a new weighted average partial effect interpretation under a general potential response function structure. These are useful, but relatively modest generalizations. More significantly we (i) provide distribution theory for the estimator

[2]The Wooldridge (2004) paper remains unpublished, but a textbook treatment of the material in it can be found in Chapter 21.6.3 of Wooldridge (2010).

proposed by Wooldridge (2004), (ii) characterize the semiparametric efficiency bound (SEB) for β_0, and (iii) introduce a new locally efficient estimator. The procedure proposed by Wooldridge (2004) is inefficient.

Another feature of our estimator is computational simplicity. Let $\hat{\mu}_W = \frac{1}{N}\sum_{i=1}^{N} W_i$ be the sample mean of W. A common approach to modeling heterogeneous effects in applied work is to compute the least squares fit of Y onto a constant, $W - \hat{\mu}_W$, $(W - \hat{\mu}_W) \otimes X$, and X. As is well-known from textbook treatments on interaction terms in linear regression analysis, centering the control variable vector, W, about is mean in this way ensures that the coefficient on X captures an average effect. This approach essentially coincides with Oaxaca-Blinder type methods of covariate adjustment popular in labor economics (e.g., Kline, 2014). One variant of our procedure involves computing the exact same regression, but where X is instead instrumented with a particular function of its conditional distribution given W (i.e., of the "generalized" propensity score). Theorems 2 and 3 below show that this small modification to a familiar estimation procedure delivers considerable gains.

The next section introduces our average linear regression model. We provide a statistical definition of β_0 as well as sets of assumptions under which it has a causal – average partial effect (APE) – interpretation. Section 2 presents the semiparametric efficiency bound for β_0. Section 3 studies the large sample properties of the Wooldridge (2004) estimator. We also introduce our new estimator and present its large sample properties. Finally, in Section 4, we connect our results with prior work on efficient estimation of average treatments effects as well as the partially linear semiparametric regression model. We end our paper with a small simulation study in Section 5. All proofs are collected in the Appendix or the supplemental materials.

1 Average linear regression model

We begin with a conventional sampling assumption.

Assumption 1. (Random Sampling) *Let* $\left\{(W_i', X_i', Y_i)'\right\}_{i=1}^{\infty}$ *be a sequence of independent and identically distributed random draws from some population* $F_{W,X,Y}$ *with* $\mathbb{E}\left[Y^2 \middle| W = w\right] < \infty$ *and* $\mathbb{E}\left[\|X\|^2 \middle| W = w\right] < \infty$ *for all* $w \in \mathbb{W}$.

The finite moment restrictions included in Assumption 1 ensure that a conditional linear predictor (CLP) is well-defined for all $w \in \mathbb{W}$.

Let

$$e_0(w) = \mathbb{E}\left[X \middle| W = w\right] \tag{3}$$

be the conditional mean of X given $W = w$ and

$$v_0(w) = \mathbb{V}(X|W = w) \tag{4}$$

the corresponding conditional variance. We also require that X vary conditional on $W = w$.

Assumption 2. (Overlap) *For all $w \in \mathbb{W}$ and any non-zero column vector t, $t'v_0(w)t \geq \kappa > 0$.*

Assumption 2 ensures that the CLP is uniquely defined. In the absence of conditioning it is equivalent to linear independence of the elements of X. When X is binary $v_0(W) = e_0(W)(1 - e_0(W))$ with $e_0(W) = \Pr(X = 1|W)$ equal to the propensity score; in this case Assumption 2 coincides with the familiar strong overlap assumption from the program evaluation literature. More generally Assumption 2 implies that X varies conditional on $W = w$ for *all* $w \in \mathbb{W}$.

Under Assumptions 1 and 2 the conditional linear predictor is well-defined for all $w \in \mathbb{W}$. Wooldridge (1999, Section 4) provides a self-contained introduction to conditional linear predictors. The following definition and lemma is taken from Wooldridge (1999).

Definition 1. (Conditional Linear Predictor) The mean squared error minimizing linear predictor of Y given X *conditional* on $W = w$, henceforth the conditional linear predictor (CLP), equals

$$\mathbb{E}^*[Y|X; W = w] \stackrel{def}{\equiv} a_0(w) + X'b_0(w) \tag{5}$$

with

$$\begin{aligned} a_0(w) &\stackrel{def}{\equiv} \mathbb{E}[Y|W = w] - e_0(w)'b_0(w) \\ b_0(w) &\stackrel{def}{\equiv} v_0(w)^{-1}\mathbb{C}(X, Y|W = w). \end{aligned} \tag{6}$$

It is straightforward to show that the prediction error $U = Y - \mathbb{E}^*[Y|X; W]$ is conditionally mean zero and conditionally uncorrelated with X. This property of $\mathbb{E}^*[Y|X; W]$ will prove useful for what follows.

Lemma 1. *Wooldridge (1999, Lemma 4.1). Let $U \stackrel{def}{\equiv} Y - a_0(W) - X'b_0(W)$, then $\mathbb{E}[U|W = w] = 0$ and $\mathbb{E}[XU|W = w] = 0$ for all $w \in \mathbb{W}$.*

Identification of the average regression slope

We begin by presenting a convenient representation of the average slope coefficient $\beta_0 = \mathbb{E}[b_0(W)]$ in terms of the joint distribution of $(W', X', Y)'$. The most direct representation follows directly from (6):

$$\beta_0 = \mathbb{E}\left[v_0(W)^{-1}\mathbb{C}(X, Y \mid W)\right].$$

For our purposes, however, an alternative representation of β_0 is more convenient; both for our semiparametric efficiency bound (SEB) analysis and for the approach to estimation developed below. Using the law of iterated expectations and the definition of conditional covariance we get, under Assumptions 1 and 2,

$$\begin{aligned}\mathbb{E}\left[v_0(W)^{-1}(X - e_0(W))Y\right] &= \mathbb{E}\left[v_0(W)^{-1}\mathbb{E}[(X - e_0(W))Y \mid W]\right] \\ &= \mathbb{E}\left[v_0(W)^{-1}\mathbb{C}(X, Y \mid W)\right].\end{aligned}$$

Applying definition (6) then gives our preferred estimand representation:

$$\beta_0 = \mathbb{E}\left[v_0(W)^{-1}(X - e_0(W))Y\right]. \tag{7}$$

Wooldridge (2004) emphasizes the coincidence between (7) and the average partial effect of X on Y associated with a particular correlated random coefficients (CRC) potential outcomes structure. This endows β_0 with causal meaning. While we also develop this connection below, we wish to initially emphasize that (7) is also just one way of representing a population average of conditional linear predictor coefficients. Under Assumptions 1 and 2 the expectation in (7) is well-defined and β_0 is simply a "statistical" estimand. We are interested in estimating it as precisely as possible.

Causal interpretation

In this subsection we show that (7) admits a causal interpretation under a particular treatment response model and selection on observables type assumption. As noted earlier, this interpretation was previously emphasized by Wooldridge (2004), but under stronger conditions than we maintain here.

Associated with each agent in the target population is an individual-specific potential response function, $Y(x)$, which maps counterfactual values of the input vector X into their corresponding (potential) outcomes. The observed outcome coincides with the value of the potential response function at the observed input level X: $Y = Y(X)$. We assume that

$Y(x)$ is linear in x, but otherwise heterogeneous across individuals:

$$Y(x) = A + x'B, \tag{8}$$

where A and B are an individual-specific intercept and slope vector respectively.

Equation (8) allows for each individual to have their own potential response function, but restricts them to be linear in X. When X is binary, or multi-valued, linearity is unrestrictive. For example, in the binary case, we have the potential outcome under control ($X = 0$) and active ($X = 1$) treatment equal to $Y(0) = A$ and $Y(1) = A + B$. In the multi-valued treatment setting of Imbens (2000) and Cattaneo (2010), with X a vector of treatment indicators for K mutually exclusive treatments, we have $Y(0) = A$ and $Y(k) = A + B_k$ for $k = 1, \ldots, K$. In contrast, when X is ordered, continuously valued, or includes multiple treatments/policies, linearity is restrictive.

Consider the following thought experiment: draw a unit at random and (exogenously) increase the value of the k^{th} component of X by one unit. The expected effect of this intervention is $\mathbb{E}[B_k]$. In the binary- and multi-valued treatment setting $\mathbb{E}[B_k]$ corresponds to an average treatment effect (ATE)

$$\mathbb{E}[B_k] = \mathbb{E}[Y(k) - Y(0)].$$

More generally $\mathbb{E}[B_k]$ equals the *average partial effect* (APE) of a unit increase in X_k. This estimand was introduced in a panel data setting by Chamberlain (1984); general expositions, with additional results, are available in Blundell & Powell (2003) and Wooldridge (2005).

Under the following assumption, in addition to those introduced above, we can show that β_0 coincides with the APE vector, $\mathbb{E}[B]$.

Assumption 3. (Conditional Exogeneity) For all $w \in \mathbb{W}$ and $k, l = 1, \ldots, K$, *and under potential responses of the form given in (8)*

$$\mathbb{C}(A, X_k | W = w) = \mathbb{C}(B, X_k | W = w) = \mathbb{C}(B, X_k X_l | W = w) = 0. \tag{9}$$

Assumption 3 restricts the form of any dependence between the potential response function, $Y(x) = A + x'B$, and the treatment vector actually chosen by the respondent, X. It is a conditional exogeneity or selection on observables type assumption. To see this observe that when X is binary Assumption 3 coincides with the standard mean independence assumption familiar from the program evaluation literature, implying that

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

In the multi-valued treatment setting Assumption 3 also coincides with standard generalizations of the mean independence assumption (cf., Imbens, 2000). See also Section 4 below.

When the linearity of (8) is restrictive, as occurs when X includes continuously-valued components, or non-mutually exclusive binary inputs, Assumption 3 is less restrictive than other possible formulations of conditional exogeneity. For example, Wooldridge (2004, 2010) works with the identifying restrictions

$$\mathbb{E}\left[X\middle|W,A,B\right]=\mathbb{E}\left[X\middle|W\right]=e_0(W),\ \mathbb{V}\left(X\middle|W,A,B\right)=\mathbb{V}\left(X\middle|W\right)=v_0(W) \tag{10}$$

which imply (9), but are generally stronger. An even stronger notion of conditional exogeneity is

$$\mathbb{E}\left[A\middle|X,W\right]=\mathbb{E}\left[A\middle|W\right],\ \mathbb{E}\left[B\middle|X,W\right]=\mathbb{E}\left[B\middle|W\right]. \tag{11}$$

Assumption 3 is (apparently) the weakest assumption necessary to equate β_0 with the average partial effect of X on Y when the potential response function takes form (8). The estimator we introduce below will remain consistent under the stronger restrictions, (10) and (11), but will generally not be semiparametrically efficient in those cases. We elaborate further on this observation below.

Proposition 1. (Average Partial Effect Identification) *Under Assumptions 1, 2 and 3 the average of the CLP coefficients, $\beta_0=\mathbb{E}\left[b_0\left(W\right)\right]$, and the average partial effect (APE), $\mathbb{E}\left[B\right]$, coincide:*

$$\beta_0=\mathbb{E}\left[B\right].$$

Proof. Wooldridge (2004) demonstrates the equality under the stronger restriction (10). Under Assumption 3, however, the proof proceeds differently. Given the linear potential response (8) and by lemma (1), we have the $1+K$ conditional moment restrictions

$$\begin{aligned}\mathbb{E}\left[U\middle|W=w\right] &= \mathbb{E}\left[A-a_0\left(W\right)\middle|w\right]+\mathbb{E}\left[X'\left(B-b_0\left(W\right)\right)\middle|w\right]=0\\ \mathbb{E}\left[XU\middle|W=w\right] &= \mathbb{E}\left[X\left(A-a_0\left(W\right)\right)\middle|w\right]+\mathbb{E}\left[XX'\left(B-b_0\left(W\right)\right)\middle|w\right]=0.\end{aligned} \tag{12}$$

Under Assumption 3 conditions (12) simplify to

$$\begin{aligned}\left\{\mathbb{E}\left[A\middle|w\right]-a_0\left(w\right)\right\}+e_0\left(w\right)'\left\{\mathbb{E}\left[B\middle|w\right]-b_0\left(w\right)\right\}&=0\\ e_0\left(w\right)\left\{\mathbb{E}\left[A\middle|w\right]-a_0\left(w\right)\right\}+\mathbb{E}\left[XX'\middle|w\right]\left\{\mathbb{E}\left[B\middle|w\right]-b_0\left(w\right)\right\}&=0\end{aligned}$$

or, in matrix form,

$$\begin{bmatrix} 1 & e_0(w)' \\ e_0(w) & \mathbb{E}[XX'|w] \end{bmatrix} \begin{pmatrix} \mathbb{E}[A|w] - a_0(w) \\ \mathbb{E}[B|w] - b_0(w) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Under the Assumption 2 the first matrix to the left of the equality is invertible for all $w \in \mathbb{W}$. This implies that $\mathbb{E}[A|W=w] = a_0(w)$ and $\mathbb{E}[B|W=w] = b_0(w)$ for all $w \in \mathbb{W}$. The result follows by iterated expectations. □

Causal interpretation under misspecification

Angrist & Krueger (1999, Section 2.3.1) and Angrist & Pischke (2009, Chapter 3.3) emphasize that when X is a continuously-valued random variable its slope coefficient in the linear predictor of Y onto a constant, X and the vector of "saturated" controls admits a weighted average derivative interpretation when the potential response function takes a general nonlinear form (cf., Angrist et al., 2000, Lemma 5). Angrist and Krueger's (1999) expression is also isomorphic to the probability limit of the E-Estimator of Newey (1990) and Robins et al. (1992)

$$\beta_{\mathrm{E}} = \frac{\mathbb{E}[Y(X - e_0(W))]}{\mathbb{E}[X(X - e_0(W))]} \tag{13}$$

when the partially linear regression structure, equation (1) above, is incorrect.

In this section, using similar arguments to those appearing in Angrist et al. (2000, Lemma 5) and Graham et al. (2010, Lemma A.1), we provide a representation result for β_0 under a general potential response function.

Assume that the potential response function is nonlinear and heterogeneous such that $Y(x) = h(x, U)$. Further assume, stronger than Assumption 3 above, that U is conditionally independent of X given $W = w$ for all $w \in \mathbb{W}$. Blundell & Powell (2003) show that the *partial mean* $\mathbb{E}_W[\mathbb{E}[Y|W, X=x]]$ identifies the average structural function (ASF) $m(x) = \mathbb{E}_U[h(x, U)]$ when the support of W given $X = x$ coincides with its marginal support. Newey (1994) provides an explicit partial mean estimator and derives in asymptotic properties.

Here we show that our average regression slope estimand, β_0, can be expressed as a weighted average of the gradient of $h(X, U)$. This provides a causal interpretation of β_0 under a general potential response function. To present this result we replace Assumption 3 with:

Assumption 4. (Nonlinear Potential Response Function) *(i) X is a continuous scalar random variable with bounded support $\mathbb{X} = [\underline{x}, \overline{x}]$, (ii) the conditional density function*

of X given $W = w$ is bounded and bounded away from zero for all $(w, x) \in \mathbb{W} \times \mathbb{X}$, (iii) $Y = h(X, U)$ with $h(x, u)$ a continuously differentiable function of x for all $(x, u) \in \mathbb{X} \times \mathbb{U}$ and $\underline{h}(u) = h(\underline{x}, u)$ finite for all $u \in \mathbb{U}$, and (iv) U is conditionally independent of X given $W = w$ for all $w \in \mathbb{W}$.

Proposition 2. *(*Weighted Average Derivative Representation*) Under Assumptions 1, 2 and 4*

$$\beta_0 = \mathbb{E}\left[\omega(W, X)\frac{\partial h(X, U)}{\partial x}\right]$$

where

$$\omega(w, x) = \frac{1}{f_{X|W}(x|w)}\frac{\mathbb{E}[X - e_0(W)|W = w, X \geq x]\left(1 - F_{X|W}(x|w)\right)}{\int_{\underline{x}}^{\overline{x}} \mathbb{E}[X - e_0(W)|W = w, X \geq v]\left(1 - F_{X|W}(v|w)\right)\mathrm{d}v}.$$

Proof. See the Supplemental Web Appendix. □

A key feature of the weighting function $\omega(w, x)$ is that its *conditional* mean, $\mathbb{E}[\omega(W, X)|W = w]$, equals 1 for *every* value of $w \in \mathbb{W}$. Furthermore, Lemma A.1 of Graham et al. (2010) implies that, conditional on $W = w$, the weight given to $\frac{\partial h(X,U)}{\partial x}$ is highest for those values of X near its conditional mean, $\mathbb{E}[X|W = w]$, and lowest for those at the boundary of its support, $\underline{x}$ and $\overline{x}$.

These features of the weights appearing in Proposition 2 imply the following intuitive interpretation: (i) for each value of $w \in \mathbb{W}$ compute a weighted average of $\frac{\partial h(X,U)}{\partial x}$, where the average emphasizes values of X near its conditional mean given $W = w$, (ii) average these (weighted average) gradients over the marginal distribution of W. This indicates that β_0 only differs from the unweighted average $\mathbb{E}\left[\frac{\partial h(X,U)}{\partial x}\right]$ due to variation in $\omega(W, X)$ *within* $W = w$ cells. The contribution of each subpopulation, defined in terms of the control, W, mirrors its density in the sampled population. Since W proxies for U in this set-up we *are* averaging over the correct heterogeneity distribution.

More precisely, since $\mathbb{E}[\omega(W, X)|W = w] = 1$, we have that, using the definition of conditional covariance,

$$\beta_0 - \mathbb{E}\left[\frac{\partial h(X, U)}{\partial x}\right] = \mathbb{E}\left[\mathbb{C}\left(\omega(W, X), \frac{\partial h(X, U)}{\partial x}\middle| W\right)\right]. \tag{14}$$

The bias of β_0 for $\mathbb{E}\left[\frac{\partial h(x,U)}{\partial x}\right]$ is therefore solely due to *conditional* covariance between the weight function and the gradient of interest within subpopulations homogenous in W.

In contrast to the one for β_0, the weight function appearing in the weighted average derivative representation result of Angrist & Krueger (1999) or Angrist & Pischke (2009) for β_{E} is only

unconditionally mean zero. This implies that β_{E} averages over the incorrect heterogeneity distribution *as well as* the incorrect policy variable distribution.

$$\begin{aligned}\beta_{\mathrm{E}} - \mathbb{E}\left[\frac{\partial h\left(X,U\right)}{\partial x}\right] =& \mathbb{E}\left[\mathbb{C}\left(\omega\left(W,X\right), \frac{\partial h\left(X,U\right)}{\partial x}\middle| W\right)\right] \\ &+ \mathbb{C}\left(\mathbb{E}\left[\omega\left(W,X\right)|\, W\right], \mathbb{E}\left[\frac{\partial h\left(X,U\right)}{\partial x}\middle| W\right]\right).\end{aligned} \tag{15}$$

If the ultimate object of interest is the average derivative $\mathbb{E}\left[\frac{\partial h(X,U)}{\partial x}\right]$, then, relative to (15), a focus on β_0 eliminates one source of potential bias. Namely that the weight function may over- or under-emphasize various subpopulations defined in terms of their value of the control variable vector W. In this case $\mathbb{E}\left[\omega\left(W,X\right)|\, W\right]$ may not equal one and the second term to the right of the equality in (15) may be non-zero.[3]

Motivating β_0

Our focus on averages of conditional linear predictor slope coefficients is motivated by a combination of principled and pragmatic reasons.

First, the kitchen sink long regression remains a workhorse of everyday empirical social science research. Our model extends kitchen sink regression in an easy to understand way. Relative to the partially linear regression model, our model allows for heterogenous responses of Y to variation in X; a feature likely to be both empirically relevant and *a priori* attractive to researchers.

Second, β_0 has a causal interpretation under additional assumptions. When the potential response function is linear, but heterogeneous across agents, it coincides with an average partial effect (APE) under a selection on observables type assumption. When X is binary- or multi-valued, as in the program evaluation literature, it coincides with the well-known average treatment effect (ATE). Our causal model nests the usual one as a special case, but accommodates continuous and/or multiple treatments as well (albeit under restrictions).

Third, in the presence of misspecification β_0 coincides with a weighted average of the derivative of a general non-linear potential response function. This weighted average derivative is more interpretable than existing representation results; for example those of Angrist & Krueger (1999) for β_{E}.

Fourth, as we show next, β_0 is $\sqrt{N}$ estimable (or regularly identified). This is not the case for, say, a partial mean with a continuous policy variable (e.g., Newey, 1994). Regular

[3]To be clear $\omega\left(W,X\right)$ are different functions in expressions (14) and (15); for its form in the latter case see Angrist & Krueger (1999) or Angrist & Pischke (2009).

identification suggests that estimation is practically feasible and we present one such feasible estimator below.

Ultimately the balance between ease of interpretation under various population assumptions and, as we show below, ease of estimation, provides the strongest case for focusing on β_0.

2 Semiparametric efficiency bound

Using the method of calculation outlined by Bickel et al. (1993) and Newey (1990), we derive the semiparametric variance bound for β_0 of,

$$\mathcal{I}(\beta_0)^{-1} = \mathbb{E}\left[\Omega_0(W)\right] + \mathbb{V}(b_0(W)), \tag{16}$$

where

$$\Omega_0(w) = \mathbb{E}\left[v_0(W)^{-1}\left(X - e_0\left(W\right)\right) UU' \left\{ v_0(W)^{-1}\left(X - e_0\left(W\right)\right)\right\}' \middle| W = w \right].$$

The corresponding efficient influence function equals

$$\begin{aligned} \psi_\beta^{\text{eff}}\left(Z, \beta_0, g_0\left(W\right), h_0\left(W\right)\right) =& v_0\left(W\right)^{-1}\left(X - e_0\left(W\right)\right)\left(Y - a_0\left(W\right) - X' b_0\left(W\right)\right) \\ &+ \left(b_0\left(W\right) - \beta_0\right) \end{aligned} \tag{17}$$

with $Z = (W', X', Y)'$, $g\left(W\right) = (e(W), v(W))$ and $h\left(W\right) = (a\left(W\right), b\left(W\right))$.

Theorem 1. (Semiparametric Efficiency Bound) *The efficient influence function for $\beta_0 = \mathbb{E}\left[b_0\left(W\right)\right]$ in the semiparametric problem established by Definition 1 and Assumptions 1 and 2 equals (17).*

Proof. See Appendix A.

We also have the following corollary, which is similar to a result for the binary case due to Robins et al. (1994), Hahn (1998) and Chen et al. (2008). This corollary will be useful when we discuss locally efficient estimation in Section (3). □

Corollary 1. (Redundancy) *Let $f\left(x \middle| w; \phi\right)$ be a parametric family of conditional distributions for X given W with $f_0\left(x \middle| w\right) = f\left(x \middle| w; \phi\right)$ at some unique $\phi = \phi_0$. The knowledge that $f_0\left(x \middle| w\right)$ is a member of the family $f\left(x \middle| w; \phi\right)$ does not change the semiparametric efficiency bound for β_0.*

Proof. See the Supplemental Web Appendix. □

See Frölich (2004) and Graham et al. (2016) for additional intuition about results like Corollary 1.

Double robustness property of the efficient influence function

Before introducing our estimator in the next section we highlight an important property of the efficient influence function for β_0.

Consider replacing $h_0(W) = (a_0(W), b_0(W))$ in (17) with the incorrect conditional linear predictor coefficients $h_*(W) = (a_*(W), b_*(W))$. Use the notation $U_* = (Y - a_*(W) - X'b_*(W))$ to emphasize that U_* is the prediction error associated with an arbitrary conditional linear predictor (which need not be the mean squared error minimizing one). Note that U_* will not be conditionally mean zero or conditionally uncorrelated with X (i.e., $\mathbb{E}[U_*|W] \neq 0$ and $\mathbb{E}[XU_*|W] \neq 0$). Nevertheless, as long as $e_0(X)$ and $v_0(W)$ equal the true conditional mean and variance of X given W, we have the pair of equalities, using iterated expectations,

$$\begin{aligned}\mathbb{E}\left[v_0(W)^{-1}(X - e_0(W))a_*(W)\right] &= 0\\ \mathbb{E}\left[v_0(W)^{-1}(X - e_0(W))X'b_*(W)\right] &= \mathbb{E}[b_*(W)]\end{aligned}$$

(the second equality follows from the fact that $\mathbb{E}[(X - e_0(W))X'|W] = v_0(W)$).

Therefore (17) remains mean zero even if the nuisance functions $h_0(W) = (a_0(W), b_0(W))$ are replaced by arbitrary functions of W:

$$\mathbb{E}\left[\psi_\beta^{\text{eff}}(Z, \beta_0, g_0(W), h_*(W))\right] = 0. \tag{18}$$

One special choice of $h_*(W)$ is the zero vector. This choice directly recovers the representation of β_0 derived earlier (Equation (7) above). In moment condition form

$$\mathbb{E}\left[v_0(W)^{-1}(X - e_0(W))Y - \beta_0\right] = 0.$$

Next consider replacing $g_0(W) = (e_0(W), v_0(W))$ in (17) with the incorrect conditional mean and variance functions $g_*(W) = (e_*(W), v_*(W))$. Use the notation $U_0 = (Y - a_0(W) - X'b_0(W))$ to emphasize that U_0 is the prediction error associated with the mean squared error minimizing conditional linear prediction of Y given X conditional on W.

By Lemma 1 $\mathbb{E}\left[U_0 \middle| W\right] = 0$ and $\mathbb{E}\left[XU_0 \middle| W\right] = 0$. Therefore

$$\begin{aligned}\mathbb{E}\left[\psi_\beta^{\text{eff}}\left(Z, \beta_0, g_*\left(W\right), h_0\left(W\right)\right)\right] =& \mathbb{E}\left[v_*\left(W\right)^{-1}\left(X - e_*\left(W\right)\right)\left(Y - a_0\left(W\right) - X'b_0\left(W\right)\right)\right] \\ &+ \mathbb{E}\left[\left(b_0\left(W\right) - \beta_0\right)\right] \\ =& \mathbb{E}\left[v_*\left(W\right)^{-1}\mathbb{E}\left[XU_0 \middle| W\right]\right] - \mathbb{E}\left[v_*\left(W\right)^{-1}e_*\left(W\right)\mathbb{E}\left[U_0 \middle| W\right]\right] \\ =& 0.\end{aligned}$$

Hence (17) also remains mean zero even if the nuisance functions $g_0\left(W\right) = \left(e_0\left(W\right), v_0\left(W\right)\right)$ are replaced by arbitrary functions of W.

Moment (17) has the so-called doubly robust property of Scharfstein et al. (1999) (cf., Ruud, 1986). It is mean zero as long as one of the two nuisance functions, $g\left(W\right)$ or $h\left(W\right)$, coincides with its population one. We exploit this property when constructing our estimator in the next section.

3 Estimation

In this section we present a locally semiparametrically efficient estimate of β_0. To motivate the precise form of our estimator we also discuss the estimator proposed by Wooldridge (2004). A textbook presentation of this estimator is available in Chapter 21.6.3 of Wooldridge (2010).

For the purposes of estimation we impose a parametric restriction on the conditional distribution of X given W. Since the distribution of X given W is ancillary for β_0, this parametric restriction does not change the semiparametric efficiency bound (cf., Corollary 1). We call, borrowing nomenclature from related settings (e.g., Hirano & Imbens, 2004), the resulting model for X the *generalized propensity score*.

Assumption 5. (Generalized Propensity Score) *$f\left(x \middle| w; \phi\right)$ is a parametric family of densities indexed by $\phi \in \Phi \subset \mathbb{R}^L$ with (i) $f_0\left(x \middle| w\right) = f\left(x \middle| w; \phi_0\right)$ at some unique $\phi_0 \in \text{int}\left(\Phi\right)$, (ii) a maximum likelihood estimate (MLE) of ϕ_0 equal to*

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

with a score vectors of $\mathbb{S}_\phi\left(X \middle| W; \phi\right) = \nabla_\phi f\left(X \middle| W; \phi\right) / f\left(X \middle| W; \phi\right)$, (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

$$\sqrt{N}\left(\hat{\phi}-\phi_0\right)=\mathbb{E}\left[\mathbb{S}_i\mathbb{S}_i'\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbb{S}_i+o_p\left(1\right) \tag{19}$$

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

Assumption 5 corresponds to a parametric model for the propensity score when X is a binary treatment indicator. More generally Assumption 5 requires the researcher to model the distribution of the policy given controls. Consider a researcher interested in the relationship between regular school attendance and student achievement. In this case Y could be a measure of end-of-school-year achievement, X number of school days absent, and W a vector of joint determinants of achievement and attendance (e.g., family background measures, prior achievement, pre-existing health conditions etc.). In this case the researcher might assume that the distribution of X given W is Poisson with

$$\mathbb{E}\left[X|\,W\right]=\exp\left(k\left(W\right)'\phi_0\right),\ \mathbb{V}\left(X|\,W\right)=\exp\left(k\left(W\right)'\phi_0\right),$$

where $k\left(W\right)$ is a known $L\times1$ vector of functions of W. Estimation of ϕ_0 , and hence $e\left(W;\phi_0\right)$ and $v\left(W;\phi_0\right)$, is by maximum likelihood. In most cases the conditional distribution of X given W can be conveniently modeled by, depending on the nature of X, the appropriate generalized linear model (GLM). When X is multivariate, the outcome of censoring, or has mixed discrete/continuous components, then specifying $f\left(x|\,w;\phi\right)$ may involve considerable work. For complicated likelihoods $e\left(W;\hat{\phi}\right)$ and $v\left(W;\hat{\phi}\right)$ may need to be approximated numerically or by simulation.

The Wooldridge (2004) estimator

Wooldridge (2004) introduced a two-step estimator for β_0. A textbook exposition appears in Wooldridge (2010, Chapter 21.6.3). His procedure is summarized in Algorithm 1.

Wooldridge (2004) does not characterize the asymptotic sampling properties of $\hat{\beta}_W$. In this section, we show that Wooldridge's estimator is not efficient under Assumptions 1, 2 and 5. Furthermore it requires the generalized propensity score to be correctly specified. The structure of this inefficiency and lack of robustness, as well as the form of the efficient influence function derived earlier, guides the construction of our new, locally efficient and doubly robust estimator.

The second step of Algorithm 1 corresponds to finding the $\hat{\beta}_{\mathrm{W}}$ which solves the sample

Algorithm 1 THE WOOLDRIDGE (2004) ESTIMATE OF β_0

1. Compute the maximum likelihood estimate of ϕ_0 and construct $e\left(W_i, \hat{\phi}\right)$ and $v\left(W_i, \hat{\phi}\right)$ for $i = 1, \ldots, N$;
2. Compute linear instrumental variables fit of Y onto X (with no constant) using $v\left(W; \hat{\phi}\right)^{-1}\left(X - e\left(W; \hat{\phi}\right)\right)$ as the instrument for X. The coefficient on X equals $\hat{\beta}$.

moment

$$\frac{1}{N}\sum_{i=1}^{N} \rho\left(Z_i, \hat{\phi}, \hat{\beta}_{\mathrm{W}}\right) = 0, \tag{20}$$

for $\rho(Z, \phi, \beta) = v(W; \phi)^{-1}(X - e(W; \phi))(Y - X'\beta)$. Here $\hat{\phi}$ corresponds to the MLE of ϕ_0 computed in the first step of the procedure. A mean value expansion of (20) in $\hat{\beta}_W$ about β_0 yields

$$\hat{\beta}_{\mathrm{W}} = \beta_0 + \frac{1}{N}\sum_{i=1}^{N} \rho\left(Z, \hat{\phi}, \beta_0\right) + o_p(N^{-1/2}).$$

Rearrangement of terms and a second mean value expansion in $\hat{\phi}$ about ϕ_0 gives

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}_{\mathrm{W}} - \beta_0\right) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \rho(Z, \phi_0, \beta_0) \\ &\quad + \left\{\frac{1}{N}\sum_{i=1}^{N} \frac{\partial \rho\left(Z, \bar{\phi}, \beta_0\right)}{\partial \phi}\right\} \sqrt{N}\left(\hat{\phi} - \phi_0\right) + o_p(1).\end{aligned}$$

Observe that under Assumptions 1 and 2

$$\begin{aligned}\mathbb{E}\left[\rho(Z, \phi_0, \beta_0) \middle| W = w\right] &= \mathbb{E}\left[v(W; \phi_0)^{-1}(X - e(W; \phi_0))(Y - X'\beta_0) \middle| W = w\right] \\ &= b_0(w) - \beta_0\end{aligned}$$

since $\mathbb{E}\left[v(W; \phi_0)^{-1}(X - e(W; \phi_0))X' \middle| W = w\right] = I_K$. In integral form:

$$\int \rho(z, \phi_0, \beta_0) f_0(y| w, x) f(x| w; \phi_0) \mathrm{d}x\mathrm{d}y = b_0(w) - \beta_0. \tag{21}$$

Differentiating (21) through the integral with respect to ϕ gives:

$$\mathbb{E}\left[\left.\frac{\partial\rho\left(Z,\phi_0,\beta_0\right)}{\partial\phi}\right| W=w\right]=-\mathbb{E}\left[\left.\rho\left(Z,\phi_0,\beta_0\right)\mathbb{S}'\right| W=w\right], \tag{22}$$

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

Using (19) and (22) we have

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}_{\mathrm{W}}-\beta_0\right) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\rho_i \\ &\quad -\mathbb{E}\left[\rho\mathbb{S}'\right]\mathbb{E}\left[\mathbb{S}\mathbb{S}'\right]^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbb{S}_i+o_p\left(1\right) \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\rho_i-\mathbb{E}\left[\rho\mathbb{S}'\right]\mathbb{E}\left[\mathbb{S}\mathbb{S}'\right]^{-1}\mathbb{S}_i\right\}+o_p\left(1\right)\end{aligned} \tag{23}$$

for $\rho_i=\rho\left(Z_i,\phi_0,\beta_0\right)$.

Similar to the result of Wooldridge (2007) for the binary X case, this asymptotically linear representation of $\hat{\beta}_{\mathrm{W}}$ implies that if practitioners ignore sampling error in $\hat{\phi}$, they can get conservative confidence intervals. In addition, this expression shows that over-parameterizing the conditional distribution of X given W will not decrease the asymptotic precision $\hat{\beta}_{\mathrm{W}}$.

We show next that $\hat{\beta}_{\mathrm{W}}$ is inefficient for β_0 in the semiparametric model defined by Assumptions 1, 2 and 5. This demonstration of inefficiency usefully provides insight into how to construct a more efficient estimator. We begin by decomposing Wooldridge's (2004) identifying moment into the efficient influence function and a remainder: $\rho\left(Z,\phi_0,\beta_0\right)=\psi_\beta^{\mathrm{eff}}\left(Z,\beta_0,\phi_0,h_0\left(W\right)\right)+r\left(W,X,\beta_0,\phi_0,h_0\left(W\right)\right)$ with

$$\begin{aligned}r\left(W,X,\beta_0,\phi_0,h_0\left(W\right)\right)=&v\left(W;\phi_0\right)^{-1}\left(X-e\left(W;\phi_0\right)\right)\left(a_0\left(W\right)+X'\left(b_0\left(W\right)-\beta_0\right)\right) \\ &-\left(b_0\left(W\right)-\beta_0\right)\end{aligned} \tag{24}$$

Let $r_0\left(W,X\right)=r\left(W,X,\beta_0,\phi_0,h_0\left(W\right)\right)$. Note that $\mathbb{E}\left[\left.r_0\left(W,X\right)\right| W\right]=0$. Note further that $\mathbb{S}$ is also conditionally mean zero given W.

Now observe that for $l=1,\ldots,\dim\left(\phi\right)$

$$\begin{aligned}\frac{\partial\psi_\beta^{\mathrm{eff}}}{\partial\phi_l} &= -v\left(W;\phi_0\right)^{-1}\frac{\partial v\left(W;\phi_0\right)}{\partial\phi_l}v\left(W;\phi_0\right)^{-1}\left(X-e\left(W;\phi_0\right)\right)U \\ &\quad -v\left(W;\phi_0\right)^{-1}\frac{\partial e\left(W;\phi_0\right)}{\partial\phi_l}U,\end{aligned}$$

and hence that

$$\begin{aligned}\mathbb{E}\left[\left.\frac{\partial\psi_{\beta}^{\text{eff}}}{\partial\phi_{l}}\right|W\right] &= -v\left(W;\phi_{0}\right)^{-1}\frac{\partial v\left(W;\phi_{0}\right)}{\partial\phi_{l}}v\left(W;\phi_{0}\right)^{-1}\mathbb{E}\left[\left(X-e\left(W;\phi_{0}\right)\right)U\right|W\right] \quad (25)\\ &\quad -v\left(W;\phi_{0}\right)^{-1}\frac{\partial e\left(W;\phi_{0}\right)}{\partial\phi_{l}}\mathbb{E}\left[U\right|W\right]\\ &= 0\end{aligned}$$

by Lemma 1 above.

Next start with the fact that

$$\int\psi_{\beta}^{\text{eff}}f_{0}\left(y\right|x,w)f\left(x\right|w;\phi_{0})f_{0}\left(w\right)=0.$$

Differentiating through the integral gives the equality

$$\int\frac{\partial\psi_{\beta}^{\text{eff}}}{\partial\phi'}f_{0}\left(y\right|x,w)f\left(x\right|w;\phi_{0})f_{0}\left(w\right)=-\int\left\{\psi_{\beta}^{\text{eff}}\mathbb{S}'\right\}f_{0}\left(y\right|x,w)f\left(x\right|w;\phi_{0})f_{0}\left(w\right)$$

and hence that, using the decomposition of $\rho\left(Z,\phi_{0},\beta_{0}\right)$ introduced above and equation (25),

$$\mathbb{E}\left[\rho\mathbb{S}'\right]=\mathbb{E}\left[\psi_{\beta}^{\text{eff}}\mathbb{S}'\right]+\mathbb{E}\left[r\mathbb{S}'\right]=\mathbb{E}\left[r\mathbb{S}'\right].$$

Plugging this into our influence function we get

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}_{\text{W}}-\beta_{0}\right) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\rho_{i}-\mathbb{E}\left[\rho\mathbb{S}'\right]\mathbb{E}\left[\mathbb{S}\mathbb{S}'\right]^{-1}\mathbb{S}_{i}\right\}+o_{p}\left(1\right)\\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\psi_{\beta,i}^{\text{eff}}+\left[r_{i}-\mathbb{E}\left[r\mathbb{S}'\right]\mathbb{E}\left[\mathbb{S}\mathbb{S}'\right]^{-1}\mathbb{S}_{i}\right]\right\}+o_{p}\left(1\right),\end{aligned}$$

and hence an asymptotic distribution of

$$\sqrt{N}\left(\hat{\beta}_{\text{W}}-\beta_{0}\right)\xrightarrow{D}\mathcal{N}\left(0,\mathcal{I}\left(\beta_{0}\right)^{-1}+\mathbb{E}\left[\left(r-\Pi_{r\mathbb{S}}\mathbb{S}\right)\left(r-\Pi_{r\mathbb{S}}\mathbb{S}\right)'\right]\right) \quad (26)$$

with $\Pi_{r\mathbb{S}}=\mathbb{E}\left[r\mathbb{S}'\right]\times\mathbb{E}\left[\mathbb{S}\mathbb{S}'\right]^{-1}$.

The form of the the limit distribution (26) is similar to that of the familiar inverse probability weighting (IPW) estimator for binary treatments (e.g., Graham et al., 2012, Proposition 3.1). In that context it is well-known that replacing a known propensity score with an estimated one increases precision (Hirano et al., 2003; Hitomi et al., 2008; Graham, 2011). In principle the degree of precision increase is increasing in the complexity/richness of the

fitted propensity score model. Expression (26) indicates that a similar phenomena operates in our setting. If the portion of the efficient influence function that is omitted by the Wooldridge (2004) procedure is well-approximated by a linear combination of the scores used to estimate the propensity score, then the $\hat{\beta}_{\mathrm{W}}$ will be precisely determined. In practice, instead of relying on a possibly overfitted propensity score to yield efficient estimates, it is better to redesign the estimation procedure with efficiency in mind at the outset.

A locally efficient, doubly robust estimator

Our estimator for β_0 requires a working parametric model for the CLP coefficients $a_0(W)$ and $b_0(W)$. Consistency and asymptotic normality of our estimate, $\hat{\beta}$, will not depend on the correctness of this working model, but its limiting variance will. A convenient working model is provided by Assumption 6.

Assumption 6. (CLP Coefficients): $a_0(W) = \alpha_0 + (W - \mu_W)' \gamma_0$ *and* $b_0(W) = \beta_0 + \Delta_0(W - \mu_W)$.

In practice these models for $a_0(W)$ and $b_0(W)$ can be made arbitrarily flexible since W can include a rich set of basis functions (e.g., squares, cross-products etc.) in the underlying controls.

Under Assumption 6 we have that

$$\begin{aligned} \mathbb{E}^*[Y|X;W] &= \alpha_0 + (W - \mu_W)' \gamma_0 + X'(\beta_0 + \Delta_0(W - \mu_W)) \\ &= \alpha_0 + (W - \mu_W)' \gamma_0 + ((W - \mu_W) \otimes X)' \delta_0 + X'\beta_0, \end{aligned} \tag{27}$$

where $\delta_0 = \mathrm{vec}(\Delta_0)$.

Equation (27) implies that, maintaining Assumption 6, one approach to estimating β_0 would be to compute the least squares fit of Y_i onto a constant, $W_i - \mu_W$, all interactions of $W_i - \mu_i$ and X_i, and X_i itself. For the special case where X is a binary treatment indicator, this estimator is familiar to labor economists as a Oaxaca-Blinder average treatment effect (ATE) estimator (e.g., Sloczynski, 2015).[4] Consistency of of this estimator hinges upon Assumption 6 accurately characterizing the sampled population.

In our setting Assumption 6 plays a different role. Unlike in the Oaxaca-Blinder procedure, its validity is not required for consistency, but if it does accurately described the sampled

[4] In this literature researchers typically center W around $\mathbb{E}[W|X=1]$, not the unconditional mean $\mu_W = \mathbb{E}[W]$ as is done here. With this alternative centering the coefficient on X_i will coincide with the average treatment effect on the treated (ATT).

Algorithm 2 Locally Efficient and Doubly Robust Estimation Of β_0

1. Compute the maximum likelihood estimate of ϕ_0 and construct $e\left(W_i, \hat{\phi}\right)$ and $v\left(W_i, \hat{\phi}\right)$ for $i = 1, \ldots, N$;
2. Compute the sample mean $\hat{\mu}_W = \frac{1}{N}\sum_{i=1}^{N} W_i$ and construct $R_i\left(\hat{\mu}_W\right)$ for $i = 1, \ldots, N$;
3. Compute the linear instrumental variables fit of Y_i onto $R_i\left(\hat{\mu}_W\right)$ and X_i using $v\left(W_i; \hat{\phi}\right)^{-1}\left(X_i - e\left(W_i; \hat{\phi}\right)\right)$ as the excluded instrument for X_i. The coefficient on X_i in this fit coincides with $\hat{\beta}$.

population our estimator will be highly efficient. These benefits come at the cost of assuming that prior knowledge regarding the form of the generalized propensity score is available (i.e., maintaining Assumption 5).

To describe our procedure we require some additional notation. Let $\lambda = (\alpha, \gamma', \delta')'$, $R\left(\mu_W\right) = \left(1, \left(W - \mu_W\right)', \left(\left(W_i - \mu_W\right) \otimes X_i\right)'\right)'$ and

$$U_i\left(\mu_W, \lambda, \beta\right) = \left(Y_i - R\left(\mu_W\right)'\lambda - X_i'\beta\right).$$

When $R\left(\mu_W\right)$ is evaluated at the correct population mean of W, we often simply write R.

Our estimator is based upon the $(L + J + 1 + J + JK + K)\times 1$ vector of moment conditions, $m\left(Z_i, \theta\right)$, partitioned into the three ordered sub-vectors:

$$\underset{L\times 1}{m_1(X_i, W_i, \phi)} = \mathbb{S}_\phi\left(X_i \middle| W_i; \phi\right) \tag{28}$$

$$\underset{J\times 1}{m_2(W_i, \mu_W)} = W_i - \mu_W \tag{29}$$

$$\underset{1+J+JK+K\times 1}{m_3(Z_i, \phi, \mu_W, \lambda, \beta)} = \begin{pmatrix} R_i\left(\mu_W\right) \\ v\left(W; \phi\right)^{-1}\left(X - e\left(W; \phi\right)\right) \end{pmatrix} U_i\left(\mu_W, \lambda, \beta\right) \tag{30}$$

where $\theta = \left(\phi, \mu_W, \lambda', \beta\right)'$ with $\dim\left(\theta\right) = L + J + 1 + J + JK + K$.

Equations (28), (29) and (30) constitute a just-identified system. The corresponding method-of-moments estimate of β_0 can be computed in the three simple steps listed in Algorithm 2.

In many cases of interest Algorithm 2 is easily implemented using standard software. Standard errors may be constructed in the usual way for GMM estimators (e.g., Newey & McFadden, 1994; Wooldridge, 2010) or using a bootstrap.

In step 3, if instead we let X_i serve as its own instrument, we get an "Oaxaca-Blinder" type

estimator.

The next theorem summarizes the large sample properties of $\hat{\beta}$. In the statement of the Theorem, Δ_* denotes the limiting pseudo-true value of $\hat{\Delta}$. If Assumption 6 additionally holds then $\Delta_* = \Delta_0$. We also define $\tilde{\epsilon} = v(W)_0^{-1}(X - e_0(W))\epsilon$ where $\epsilon = \{a_0(W) + X'(b_0(W) - \beta_0) - R'\lambda_*\}$ (with λ_* denoting a pseudo-true parameter value). Finally we let $\Pi_{\tilde{\epsilon}\mathbb{S}} = \mathbb{E}[\tilde{\epsilon}_i\mathbb{S}']\mathbb{E}[\mathbb{S}\mathbb{S}']^{-1}$ denote the coefficient matrix associated with the multivariate regression of $\tilde{\epsilon}$ onto the score vector associated with ϕ_0 (the parameter indexing the generalized propensity score).

Theorem 2. (Large Sample Distribution) *Consider the semiparametric problem established by Definition 1 and Assumptions 1, 2, and 5. Let $\hat{\beta}$ be the method of moments estimate of β_0 based upon restrictions (28) to (30). Under regularity conditions (cf., Newey & McFadden, 1994, Theorem 3.4) $\hat{\beta}$ is (i) asymptotically normal with a limiting distribution of*

$$\sqrt{N}\left(\hat{\beta} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \mathbb{E}[\Omega_0(W)] + \Delta_*\mathbb{V}(W)\Delta_*' + \mathbb{E}\left[(\tilde{\epsilon} - \Pi_{\tilde{\epsilon}\mathbb{S}}\mathbb{S})(\tilde{\epsilon} - \Pi_{\tilde{\epsilon}\mathbb{S}}\mathbb{S})'\right]\right), \tag{31}$$

and (ii) locally efficient for β_0 at Assumption 6 with

$$\sqrt{N}\left(\hat{\beta} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \mathcal{I}(\beta_0)^{-1}\right). \tag{32}$$

Proof. See Appendix A. □

Part (ii) of Theorem (2) follows easily from part (i). In the proof we show that ϵ equals the prediction error associated with the mean squared error minimizing linear prediction of $a_0(W) + X'(b_0(W) - \beta_0)$ given $R(\mu_W)$. When Assumption 6 additional holds this prediction error will be identically equal to zero and the third term in the variance expressing appearing in part (i) drops out. Similarly when Assumption 6 holds we have $\Delta_*\mathbb{V}(W)\Delta_*' = \mathbb{V}(b_0(W))$. Together these two observations give part (ii).

Our efficiency bound calculation, Theorem 1, gives the information bound for β_0 without imposing the additional auxiliary Assumption 6. This assumption imposes restrictions on the joint distribution of the data not implied by the baseline model. If these restrictions are added to the prior used to calculate the efficiency bound, then it will generally be possible to estimate β_0 more precisely. Our estimator is not efficient with respect to this augmented model. Rather it attains the bound provided by Theorem 1 if Assumption 6 "happens to be true" in the sampled population, but is not part of the prior restriction used to calculate the bound. Newey (1990, p. 114) discusses the concept of local efficiency in detail. In what follows we will, for brevity, say $\hat{\beta}$ is locally efficient at Assumption 6.

Even if Assumption 6 does not hold precisely, our procedure will be "nearly" efficient when it is approximately true (in which case variability in ϵ about zero is small). A caveat to this claim is that the third variance term in (31) may still be large in this case if $v_0(w)$ is nearly zero for enough values of w. This occurs when overlap is poor, or there exists a lack of variation in the policy variable for some subpopulations defined in terms of $W = w$. Graham et al. (2016) develop this observation more extensively for the special case where X is binary, but similar issues apply in the more general setting considered here.

Our next result formalizes the above observation. It extends our local efficiency result to "near" global efficiency. The basic argument mirrors that given by Chamberlain (1987, Proposition 2) for approximately efficient estimation of conditional moment problems. Presenting this result requires defining a sequence of estimators based upon Algorithm 2.

Let $\mathcal{L}_2$ be the space of functions $f : \mathbb{W} \to \mathbb{R}$ with finite second moment $\mathbb{E}\left[f(W)^2\right] < \infty$. Under Assumptions 1 and 2 the set of feasible conditional linear predictor coefficients lies within this space such that $a : W \to \mathbb{R}^1$ and $b : W \to \mathbb{R}^K$ with $\mathbb{E}\left[a(W)^2\right] < \infty$ and $\mathbb{E}\left[\|b(W)\|^2\right] < \infty$. Let $\{k_j(W)\}_{j=1}^{\infty}$ be a sequence of linearly independent functions of the control variables, each with finite variance. Similar to Chamberlain (1987) we call this sequence complete if, (i) for any $\zeta > 0$ and (ii) any feasible conditional linear predictor coefficients $a(W)$ and $b(W)$ in $\mathcal{L}_2$, there are the real numbers $\alpha, \gamma_1, \ldots, \gamma_J$ and $\delta_{k1}, \ldots, \delta_{kJ}$ for $k = 1, \ldots, K$ such that

$$\mathbb{E}\left[\left\|\delta^{(J)}(W)\right\|^2\right] < \zeta^2, \tag{33}$$

with $\delta^{(J)}(W)$ defined as

$$\delta^{(J)}(W) = \begin{pmatrix} a(W) - \alpha - \sum_{j=1}^{J}\left(k_j(W) - \mu_j\right)\gamma_j \\ b_1(W) - \beta_{01} - \sum_{j=1}^{J}\left(k_j(W) - \mu_j\right)\delta_{1j} \\ \vdots \\ b_K(W) - \beta_{0K} - \sum_{j=1}^{J}\left(k_j(W) - \mu_j\right)\delta_{Kj} \end{pmatrix}. \tag{34}$$

Let $k^{(J)}(W)$ be the $J \times 1$ vector of functions of W with j^{th} element $k_j(W)$. We can construct a sequence of estimators, indexed by J, based upon Algorithm 2 with $k^{(J)}(W)$ replacing W. To do this let $\mu^{(J)} = \mathbb{E}\left[k^{(J)}(W)\right]$ and additionally define

$$R^{(J)} = \left(1, \left(k^{(J)}(W) - \mu^{(J)}\right)', \left(\left(k^{(J)}(W) - \mu^{(J)}\right) \otimes X\right)'\right)'.$$

We can then estimate β_0 by Algorithm 2 with $k^{(J)}(W)$, $\mu^{(J)}$ and $R^{(J)}$ respectively replacing W, μ_W, and $R(\mu_W)$.

Consider the asymptotic precision matrix of this method of moments estimator; from The-

orem 2 we get

$$\begin{aligned}\mathcal{I}^{(J)}\left(\beta_0\right)^{-1} =& \mathbb{E}\left[\Omega_0\left(W\right)\right] + \Delta_*^{(J)}\mathbb{V}\left(k^{(J)}\left(W\right)\right)\left(\Delta_*^{(J)}\right)' \\ &+ \mathbb{E}\left[\left(\tilde{\epsilon}^{(J)} - \Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right)\left(\tilde{\epsilon}^{(J)} - \Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right)'\right].\end{aligned}$$

with $\mathcal{I}^{(J)}\left(\beta_0\right)^{-1} \geq \mathcal{I}\left(\beta_0\right)^{-1}$ (here "$A \geq B$" denotes "$A - B$ is positive semi-definite"). Recall that $\mathcal{I}\left(\beta_0\right)$ is the semiparametric efficiency bound given in Theorem 1. Let $\hat{\beta}^{(J)}$ denote the estimate of β_0 based upon $k^{(J)}\left(W\right)$.

Proposition 3. (Near Efficiency) *If, maintaining the Assumptions of part (i) of Theorem 2,* $\left\{\hat{\beta}^{(J)}\right\}$ *is based upon a linearly independent, complete sequence* $\left\{k_j\left(W\right)\right\}_{j=1}^{\infty}$*, then, for* $\mathbb{X} \times \mathbb{W}$ *a compact subset of* $\mathbb{R}^{K+\dim(W)}$*,*

$$\lim_{J\to\infty} \mathcal{I}^{(J)}\left(\beta_0\right)^{-1} = \mathcal{I}\left(\beta_0\right)^{-1}.$$

Proof. See Appendix A. □

The compact support assumption invoked in the statement of the theorem is used in the proof, but does not appear to be essential.

Proposition 3 leaves unanswered important practical questions, such as how quickly J should increase with N. More generally the question of exactly how to choose the elements of $k^{(J)}\left(W\right)$ in order to achieve good precision in practice remains unanswered. However we expect that many insights from related settings could be applied here (e.g., Belloni et al., 2014).

We conclude this section by demonstrating double robustness in the sense of Scharfstein et al. (1999). If the specification of the generalized propensity score is not correct (i.e., Assumption 5 does not hold), but Assumption 6 is true, then our estimator remains consistent for β_0. Recall that Assumption 6 was initially invoked to ensure local efficiency of our procedure. It turns out that modeling the form of the conditional linear predictor coefficients has the added benefit of ensuring that our estimator remains consistent even if our generalized propensity score model is incorrect. Double robustness results are familiar from the literature on missing data and program evaluation (e.g., Scharfstein et al., 1999; Cattaneo, 2010; Graham, 2011). In these settings X is binary or a vector of mutually-exclusive treatment indicators. Double robustness in our more general setting is perhaps unsurprising, but nevertheless a new result.

To understand this result observe that step 3 of Algorithm 2 corresponds to solving the

sample analog of

$$\mathbb{E}\left[\begin{pmatrix} R(\mu_W) \\ v(W;\phi_*)^{-1}(X - e(W;\phi_*)) \end{pmatrix} U(\mu_W, \lambda_0, \beta_0)\right] = 0$$

for λ_0 and β_0. Here we use the notation ϕ_* to denote that our generalized propensity score model may be miss-specified.

If Assumption 6 holds in the population, then $U_0 = U_i(\mu_W, \lambda_0, \beta_0)$ is a conditional linear predictor (CLP) error and Lemma 1 above applies. Recall that $R(\mu_W) = \left(1, (W - \mu_W)', ((W_i - \mu_W) \otimes X_i)'\right)'$; by Lemma 1 U_0 is uncorrelated with all components of this vector. Likewise, because U_0 is mean independent of W and conditionally uncorrelated with X, we also have that $\mathbb{E}\left[v(W;\phi_*)^{-1}(X - e(W;\phi_*))U\right]$ is mean zero as well. Hence step 3 of Algorithm 2 involves the computation of a correctly specified method-of-moments estimator under Assumption 6; irrespective of whether Assumption 5 additionally holds. Double robustness follows, more or less, directly.

The above discussion also clarifies why, as is sometimes true in practice, sampling variability in our estimator can theoretically be lower than the semiparametric variance bound in Theorem 1 when the generalized propensity score is misspecified, but the form of the CLP coefficients are not. First, recall that the variance bound is computed without making any a priori assumptions about the form of the CLP coefficients. It turns out that in our setting such assumptions generally increase the precision with which β_0 may be estimated. When we invoke the double robustness property of our procedure to ensure consistency we are in a situation where the veracity of Assumption 6 is central. Whereas is the setting covered by Theorem 2, Assumption 6 "may happen to be true", but need not be.

It is instructive to compare our estimator with the "Oaxaca-Blinder-type" one described earlier. The Oaxaca-Blinder procedure necessarily maintains Assumption 6. Since this restriction is part of the prior, it would not be surprising to find that, under correct specification, that the Oaxaca-Blinder estimator is more efficient than ours. For the purposes of developing this point, additionally assume that U_0 is homoscedastic in X and W (but that this is not part of the prior), then – maintaining Assumption 6 – replacing $v(W;\phi_*)^{-1}(X - e(W;\phi_*))$ with X in the above moment would be natural. This replacement leads the researcher to the Oaxaca-Blinder estimator (which will also be efficient in this case). Hence, when Assumption 6 does hold in the sampled population, our procedure will be less efficient that the Oaxaca-Blinder one (at least under homoscedasticity of U_0). Of course, when Assumption 6 does not characterize the sampled population, our procedure remains consistent, while the Oaxaca-Blinder one does not.

Theorem 3. *(*Double Robustness*) Under Assumptions 1 and 2 ,* $\hat{\beta} \xrightarrow{p} \beta_0$ *if either Assumption 5 or 6 holds.*

The proof is straightforward and omitted (see Graham et al. (2012) and Graham et al. (2016) for proofs of related results). As a practical matter using the standard method-of-moments sandwich variance-covariance matrix estimator associated with the moment problem defined by (28), (29) and (30) above will support asymptotically valid inference under the conditions of both Theorems 2 and 3.

4 Examples and special cases

In this section we demonstrate that our semiparametric regression model encompasses several other well-known models.

Example 1: Binary Treatment Effect

Following the program evaluation literature let Y_0 denote the potential outcome under control and Y_1 the potential outcome under active treatment treatment. For each sampled unit we observe either Y_0 or Y_1 but not both. The observed outcome, Y, therefore equals

$$Y = XY_1 + (1 - X)Y_0$$

where X equals 1 if the unit is treated and zero otherwise. Rewriting yields a random coefficients model of

$$Y = A + BX$$

with $A = Y_0$ and $B = Y_1 - Y_0$. The average treatment effect (ATE) equals

$$\beta_0 = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[B].$$

Rosenbaum & Rubin (1983) show that the ATE is identified when $(Y_0, Y_1) \perp X|W$ (unconfoundedness) and $0 < \Pr(X = 1|W = w) < 1$ for all $w \in \mathbb{W}$ (overlap).

When X is binary our Assumption 3 implies unconfoundedness. Assumption 3 implies that X is conditionally uncorrelated with the two potential outcomes. When X is binary this also corresponds to mean and full conditional independence. Next observe that $e_0(W) = \Pr(X = 1|W = w)$ and $v_0(W) = e_0(W)[1 - e_0(W)]$. Hence our Assumption 2 implies that $0 < \kappa \leq e_0(W) \leq 1 - \kappa < 1$ or so called strong overlap.

Now consider Algorithm 2. When X is binary we have that

$$v\left(W,\hat{\phi}\right)^{-1}\left(X-e\left(W,\hat{\phi}\right)\right)=\frac{X}{e\left(W,\hat{\phi}\right)}-\frac{1-X}{1-e\left(W,\hat{\phi}\right)}.$$

Our ATE estimate is the coefficient on X associated with the linear instrumental variables fit of Y onto a constant, $(W-\hat{\mu}_W)$, $(W-\hat{\mu}_W)\cdot X$, and X where $\frac{X}{e\left(W,\hat{\phi}\right)}-\frac{1-X}{1-e\left(W,\hat{\phi}\right)}$ serves as an instrument for X. This estimator is similar to, but distinct from, the weighted least squares (WLS) one introduced by Hirano & Imbens (2001).

Wooldridge (2004) shows, for X binary, that equation (7) coincides with

$$\mathbb{E}\left[v_0\left(W\right)^{-1}\left(X-e_0\left(W\right)\right)Y\right]=\mathbb{E}\left[\frac{XY}{e_0\left(W\right)}-\frac{\left(1-X\right)Y}{1-e_0\left(W\right)}\right],$$

which is the familiar inverse probability weighting (IPW) representation of the average treatment effect (ATE) in, for example, Hirano et al. (2003).

The general form of the efficient influence function given in Theorem 1 above corresponds to the specialized one for the ATE when X is binary derived by, for example, Hahn (1998) and Hirano et al. (2003). Hence our general procedure, as summarized by Algorithm 2, provides a locally efficient and double robust estimator of the ATE. To the best of our knowledge, our proposed estimator is a new one, even in the special case where it identifies the ATE of a binary treatment. Bang & Robins (2005) and Tsiatis (2006) provide introductions to double robust causal inference.

Example 2: Multiple Treatment Effects

Following Imbens (2000), Wooldridge (2004), and Cattaneo (2010) consider finite collection of mutually exclusive treatments indexed by $k\in\{0,1,2,...,K\}$ with $K\in\mathbb{N}$. Associated with these treatments are the $K+1$ potential outcomes, $Y(0),Y(1),\ldots,Y(K)$. The observed outcome is

$$Y=Y(0)+\sum_{k=1}^{K}X_k\left\{Y(k)-Y(0)\right\}$$

where X_k is a binary random variable that equals 1 if treatment $k=0,\ldots,K$ is assigned to the unit and zero otherwise. In this case, we work with the following random coefficient model:

$$Y=A+X^{'}B$$

where $X = (X_1, \ldots, X_K)'$ is a $K \times 1$ vector of treatment indicators and B a corresponding vector of individual treatment effects.

In this setup X is multinomial with a conditional mean of

$$e_0(W) = \begin{pmatrix} \Pr(X_1 = 1|W) \\ \vdots \\ \Pr(X_K = 1|W) \end{pmatrix}$$

and an inverse conditional variance of (cf., Henderson & Searle, 1981)

$$\begin{aligned} v_0(W)^{-1} =& diag\left\{\frac{1}{\Pr(X_1 = 1|W)}, \ldots, \frac{1}{\Pr(X_K = 1|W)}\right\} \\ &+ \frac{1}{1 - \sum_{k=1}^{K} \Pr(X_K = 1|W)} \iota_K \iota_K'. \end{aligned}$$

A little bit of tedious algebra then gives

$$\beta_0 = \mathbb{E}\left[v_0(W)^{-1}(X - e_0(W))Y\right] = \mathbb{E}\begin{bmatrix} \frac{X_1 Y}{\Pr(X_1=1|W)} - \frac{X_0 Y}{\Pr(X_0=1|W)} \\ \vdots \\ \frac{X_K Y}{\Pr(X_K=1|W)} - \frac{X_0 Y}{\Pr(X_0=1|W)} \end{bmatrix},$$

which corresponds to the IPW representation of the ATEs

$$\beta_0 = \begin{pmatrix} \mathbb{E}[Y(1) - Y(0)] \\ \vdots \\ \mathbb{E}[Y(K) - Y(0)] \end{pmatrix},$$

in the multiple treatment setting.

As in the case where X is binary, the general form of the efficient influence function given in Theorem 1 above corresponds to the specialized one derived by Cattaneo (2010). Hence our general procedure also provides a locally efficient and doubly robust estimate of ATEs in the multiple, mutually exclusive, treatments setting.

Example 3: Partially linear model

Chamberlain (1986, 1992) and Robinson (1988) studied the semiparametric partially linear regression model (PLM)

$$Y = X'\beta_0 + h_0(W) + U$$

with $\mathbb{E}[U|W,X]=0$. This model can be represented by the random coefficient model

$$Y = A + X'B$$

where $\mathbb{E}[A|W,X] = a_0(W) = h_0(W)$ and $\mathbb{E}[B|X,W] = b_0(W) = \beta_0$. These assumptions are stronger than those implied by Assumption 3.

To fit this into our framework we replace Assumption 6 with a working CLP model of

$$a_0(W) = h_0(W) = \alpha_0 + (W - \mu_W)'\lambda_0, \;\; b_0(W) = \beta_0.$$

This implies a constant additive treatment effect structure.

Estimation follows Algorithm 2. First compute the MLE of ϕ_0. Second compute the sample means $\hat{\mu}_W = \frac{1}{N}\sum_{i=1}^N W_i$. Finally compute the linear instrumental variables fit of Y onto a constant, $(W - \hat{\mu}_W)$ and X, using $v\left(W,\hat{\phi}\right)^{-1}\left(X - e\left(W,\hat{\phi}\right)\right)$ as an instrument for X. Because of the constant additive treatment effect structure of the PLM we exclude the interactions $(W - \hat{\mu}_W) \otimes X$ from the IV fit computed in the third step.

It is important to recognize that although our procedure invokes the working assumption that the treatment effect is constant in W (i.e., $b_0(w) = \beta_0$ for all $w \in \mathbb{W}$), this assumption is not required for consistency as long as our generalized propensity score model is correct. Put differently although our procedure incorporates the PLM structure, this structure is not part of the maintained prior (albeit the form of the generalized propensity score is part of the prior).

If $b_0(w) = \beta_0$ for all $w \in \mathbb{W}$ and U is conditionally mean zero given *both* W and X (and also has a constant variance), but these are not part of the prior restriction used to calculate the bound, then (16) evaluates to

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

The modified PLM estimator described above, and based on our Algorithm 2, will attain this bound when the true model is a partially linear one.

Chamberlain (1992) gives a bound for β_0 – where the partially linear regression structure *is* part of the prior restriction (but the homoscedasticity assumption is not) – of

$$\mathcal{I}_{\text{plm}}(\beta_0)^{-1} = \sigma^2 \mathbb{E}\left[v_0(W)\right]^{-1}.$$

The difference $\mathcal{I}(\beta_0)^{-1} - \mathcal{I}_{\text{plm}}(\beta_0)^{-1}$ is positive semi-definite. This follows directly from, for example, the Theorem in Groves & Rothenberg (1969) on the expectations of inverse

Table 1: Monte Carlo Designs

Designs	1	2	3	4
α_0	1	1	1.5	1.5
γ_1	1	1	1	1
γ_2	0	0	0.5	0.5
β_0	2	2	2.5	2.5
δ_1	1.22	1.26	1	1.05
δ_2	0	0	0.5	0.5
ϕ_0	0.1	0.1	0.1	0.1
ϕ_1	0.5	0.5	0.5	0.5
ϕ_2	0	0.1	0	0.1

Notes: We specified $a_0(w) = \alpha_0 + \gamma_1(W - \mathbb{E}[W]) + \gamma_2(W^2 - \mathbb{E}[W^2])$ and $b_0(w) = \beta_0 + \delta_1(W - \mathbb{E}[W]) + \delta_2(W^2 - \mathbb{E}[W^2])$ analogous to the formulation given in Assumption 6. Each of the four designs are calibrated such that $\sqrt{\mathcal{I}(\beta_0)^{-1}/N} = 0.05$ when $N = 1,000$.

matrices. Hence although our approach to estimation remains consistent for β_0 when the true regression function takes a partially linear form, it will generally be less efficient than methods which exploit this structure at the outset (e.g., Robinson, 1988; Robins et al., 1992).

5 Finite sample properties

In order to assess the approximation accuracy of Theorems 2 and 3 in finite samples we conducted a small simulation experiment, the results of which we report here. We considered four designs. The outcome was generated according to

$$Y = a_0(W) + b_0(W)X + U$$

with W and U independent standard normal random variables and $a_0(W)$ and $b_0(W)$ either linear (designs 1 and 2) or quadratic (designs 3 and 4) in W. The conditional distribution of X given W was specified as Poisson with parameter $\exp\left(k(W)'\phi\right)$ and $k(W) = (1, W)'$ in designs 1 and 3 and $k(W) = (1, W, W^2)'$ in designs 2 and 4. Complete details on the data generating process are given in Table 1.

We evaluate the performance of three estimators. First we consider a simple "Oaxaca-Blinder" type estimator. Specifically we estimate β_0 by the coefficient on X in the least squares fit of Y onto a constant, $W - \hat{\mu}_W$, $(W - \hat{\mu}_W) \times X$ and X. As in Kline (2014) we appropriately account for the effect of estimating μ_W when constructing standard errors and confidence intervals. This estimator is consistent for the true average partial effect in

both designs 1 and 2. It is also, since U is Gaussian and homoscedastic, efficient in these two designs. Efficiency is in the semiparametric model which, *in addition to* Assumptions 1 and 2, maintains Assumption 6. The variance of the Oaxaca-Blinder estimate therefore lies (weakly) below the bound given by Theorem 1 in these two designs. In designs 3 and 4, where $a_0(W)$ and $b_0(W)$ are quadratic in W, the "Oaxaca-Blinder" estimator is inconsistent.

The second estimator is the generalized inverse probability weighting (GIPW) one due to Wooldridge (2004, 2010). Our implementation tracks our analysis in Section 3. For estimation we correctly assume that the conditional distribution of X given W is Poisson, but set the parameter to $\exp\left(k(W)'\phi\right)$ with $k(W) = (1, W)'$. This is correct in designs 1 and 3, but not 2 and 4. Hence the GIPW estimate of β_0 is consistent in designs 1 and 3, but not 2 and 4. The GIPW is never efficient. Standard errors are constructed using the sample analog of the influence function given in (23) above.

Finally we consider the properties of our locally efficient, doubly robust estimator. Implementing this procedure requires assumptions on both the CLP and the generalized propensity score. We make the same assumptions used to implement the Oaxaca-Blinder and GIPW procedures. Consequently this last estimator is efficient – in the sense of Theorem 1 – in design 1 and consistent in designs 1, 2 and 3. Like all the estimators it is inconsistent in design 4. We construct standard errors using the (sample analog) of the influence function given in Theorem 2; consequently our intervals are conservative in design 2 (where our propensity score model is misspecified).[5]

Each of the four designs are calibrated such that $\sqrt{\mathcal{I}(\beta_0)^{-1}/N} = 0.05$ (0.025) when $N = 1,000$ $(4,000)$. In an asymptotic sense inference on β_0 is equally hard across all the designs considered. We focus on the $N = 1,000$ experiments in our discussion (the quality of the asymptotic approximations predictably improve in the larger sample).

Results from the four designs are reported in Table 2. As expected our DR estimator is median unbiased across Designs 1, 2 and 3. In contrast the Oaxaca-Blinder estimator only performs acceptably in designs 1 and 2, and the GIPW estimator in design 1 and 3. In designs 1 and 2 the variability of the DR estimator is nearly as small as that of the Oaxaca-Blinder one. Neither the DR, nor the GIPW, estimators are expected to be efficient in design 3 but, interestingly, GIPW is more efficient than DR in this case. In design 1, where the DR estimator is locally efficient, its standard deviation is substantially smaller than that of the GIPW estimator (as expected).

[5]We use Python 3.6 to conduct our experiments. Replication code is available in the supplemental materials.

Table 2: Simulation Results

	Panel A, $N = 1,000$				**Panel B**, $N = 4,000$			
	Bias	Std. Dev.	Std. Err.	Coverage	Bias	Std. Dev.	Std. Err.	Coverage
Design 1								
Oaxaca-Blinder	-0.0003	0.0500	0.0496	0.9480	0.0006	0.0252	0.0248	0.9438
GIPW	-0.0008	0.0853	0.0809	0.9438	0.0016	0.0429	0.0419	0.9494
DR	0.0001	0.0507	0.0499	0.9450	0.0009	0.0254	0.0250	0.9448
Design 2								
Oaxaca-Blinder	0.0009	0.0504	0.0497	0.9442	0.0000	0.0251	0.0248	0.9456
GIPW	-0.2597	0.1331	0.1198	0.4634	-0.2613	0.0710	0.0666	0.0258
DR	0.0014	0.0518	0.0561	0.9620	-0.0001	0.0258	0.0283	0.9694
Design 3								
Oaxaca-Blinder	-0.3268	0.0993	0.0899	0.0772	-0.3319	0.0531	0.0481	0.0002
GIPW	-0.0018	0.0830	0.0794	0.9436	-0.0005	0.0428	0.0413	0.9452
DR	-0.0113	0.1099	0.0981	0.9276	-0.0036	0.0570	0.0529	0.9368
Design 4								
Oaxaca-Blinder	-0.2010	0.1571	0.1087	0.5148	-0.2391	0.1255	0.0674	0.0168
GIPW	0.2717	0.1897	0.1216	0.4006	0.3052	0.1335	0.0748	0.0034
DR	0.4123	0.2035	0.1687	0.3076	0.4362	0.1058	0.1013	0.0986
$\sqrt{\mathcal{I}(\beta_0)^{-1}/N}$	0.05				0.0250			

Notes: The bias column reports median bias across all $B = 5,000$ simulations. The Std. Dev. column reports the standard deviation of the point estimates across these simulations, Std. Err. the median estimated standard error, and Coverage the actual coverage of a nominal 95 percent confidence interval (constructed using the estimated point estimate and standard error in the normal way). The standard error associated with a Monte Carlo coverage estimate is $\sqrt{\alpha(1-\alpha)/B}$. With $B = 5,000$ simulations and $\alpha = 0.05$, this results in a standard error of approximately 0.003 or a 95 percent confidence interval of $[0.944, 0.956]$.

Overall the simulation results track our theoretical predictions remarkably closely. Of course exploring the performance of these estimators in the context of real world empirical applications and other, more realistic, simulation designs would be of interest.

6 Conclusion

We have introduced a locally efficient, doubly robust, semiparametric method of estimating averages of conditional linear predictor coefficients. Our estimand, and semiparametric efficiency bound, specialize to familiar counterparts found in the program evaluation literature (e.g. Hahn, 1998; Cattaneo, 2010). While encompassing well-known program evaluation settings, our framework allows for (semiparametric) covariate adjustment in many other settings as well (including ones with few extant alternative methods of such adjustment).

Researchers interested in estimating the average treatment effect (ATE) associated with a binary treatment can apply our methods. While we believe the precise form of our procedure is new even to this familiar setting, it is a variant of the class of augmented inverse probability weighting (AIPW) estimators introduced by Robins et al. (1994) in the missing data context over 20 years ago. The real attraction of Algorithm 2, and the corresponding Theorems 2 and 3 (as well as Proposition 3), is that they apply to models beyond the "classic" program evaluation one of Rosenbaum & Rubin (1983). Multiple, mutually exclusive treatments, as in Imbens (2000) and Cattaneo (2010) are easily handled as a special case. Similarly, maintaining a linear, but heterogeneous, potential response function structure, Algorithm 2 recovers average partial effects (APE) for continuous treatments, multiple non-exclusive treatments, mixtures of binary, discrete and continuous treatments and so on. A weighted average derivative interpretation of our estimand is also available for settings where the linear potential response function structure may not hold (Proposition 2).

We also wish to emphasize that averages of conditional linear predictor coefficients represent a natural, but substantial, generalization of linear predictor coefficients as estimated by the method of least squares. Hence Algorithm 2 also provides a method of flexible covariate adjustment that may be of independent interest even in settings where formal causal inference is not warranted; similar to how least squares is sometimes used for descriptive purposes.

Our work leaves several questions unanswered. First, although the flexible parametric modeling embodied in Assumptions 5 and 6 closely mirrors empirical practice, it would be useful to development methods that leave the generalized propensity score and CLP coefficients non-parametric. It seems likely that results from the binary and multiple treatments case could be extended to apply here (e.g., Hirano et al., 2003; Cattaneo, 2010; Belloni et al.,

2014).

In other work we have shown that first order equivalent estimators may have appreciably different higher order properties in program evaluation settings (Graham et al., 2012). We expect that other locally efficient, doubly robust approaches to estimation for the class of problems considered in this paper are feasible. These approaches may exhibit superior or inferior higher order bias.

Third, maintaining the correlated random coefficient structure, different notions of conditional exogeneity will imply different semiparametric efficiency bounds (when linearity is restrictive). Our decision to work with a weak notion of exogeneity maintains a connection with conditional linear predictors. If a researcher was comfortable with the correlated random coefficient structure, then it would generally be possible to construct more efficient estimates of $\beta_0 = \mathbb{E}[B]$ if she was willing to assume, for example, that $(A, B')' \perp X \big| W = w$ for all $w \in \mathbb{W}$. Such estimators would likely be quite complicated and may have poor finite sample properties.

A Proofs

This appendix contains proofs of the results contained in the main paper. All notation is as defined in the main text unless explicitly noted otherwise. Equation numbering continues in sequence with that established in the main text.

Proof of Theorem 1 (Semiparametric efficiency bound)

In calculating the efficiency bound for β_0 in the semiparametric regression model defined by Definition 1 and Assumptions 1 and 2 of the main text, we follow the approach outlined by Newey (1990, Section 3). First, we characterize the model's tangent space. Second, we demonstrate pathwise differentiability of β_0. The efficient influence function for β_0 equals the projection of this derivative onto the tangent space. In the present case the pathwise derivative lies in the tangent space and hence coincides with the required projection. The result then follows from an application of Theorem 3.1 in Newey (1990).

Step 1: Characterization of the Model Tangent Space:

The joint density function for $z = (w, x, y)$ is given by

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

where $f_0(x, y|\, w)$ denotes the conditional density/mass of $(X = x, Y = y)$ given $W = w$ and $f_0(w)$ is the marginal density/mass of $W = w$.

Consider a regular parametric submodel indexed by η with $f(w, x, y; \eta) = f_0(w, x, y)$ at $\eta = \eta_0$. The submodel joint density equals

$$f(w, x, y; \eta) = f(x, y|\, w; \eta)\, f(w; \eta),$$

with a corresponding score vector of

$$s_\eta(w, x, y; \eta) = s_\eta(x, y|\, w; \eta) + t_\eta(w; \eta) \tag{35}$$

where

$$s_\eta(w, x, y; \eta) = \nabla_\eta f(w, x, y; \eta),\ s_\eta(x, y|\, w; \eta) = \nabla_\eta f(x, y|\, w; \eta),\ t_\eta(w; \eta) = \nabla_\eta f(w; \eta).$$

By the usual (conditional) mean zero property of scores we have that

$$\mathbb{E}\left[s_\eta\left(X,Y|\,W\right)|\,W\right]=\mathbb{E}\left[t_\eta\left(W\right)\right]=0, \tag{36}$$

where the suppression of η in a function indicates that it is evaluated at its population value (e.g., $t_\eta\left(w\right)=t_\eta\left(w;\eta_0\right)$).

The model tangent set is the closed linear span of the set of all such scores. From (35) and (36) this set evidently equals

$$\mathcal{T}=\left\{s\left(x,y|\,w\right)+t\left(w\right)\right\}$$

where $s\left(x,y|\,w\right)$ and $t\left(w\right)$ satisfy the (conditional) moment restrictions

$$\mathbb{E}\left[s\left(X,Y|\,W\right)|\,W\right]=\mathbb{E}\left[t\left(W\right)\right]=0,$$

and also have finite variance.

Step 2: Demonstration of pathwise differentiability:

Under the parametric submodel, $\beta\left(\eta\right)$ is identified by

$$\beta\left(\eta\right)=\int b\left(w;\eta\right)f\left(w;\eta\right)\mathrm{d}w, \tag{37}$$

where $b\left(w;\eta\right)$ satisfies the conditional moment restriction

$$\int\int\begin{pmatrix}1\\x\end{pmatrix}\left(y-a\left(w;\eta\right)-x'b\left(w;\eta\right)\right)f\left(x,y|\,w;\eta\right)\mathrm{d}x\mathrm{d}y=0. \tag{38}$$

Differentiating (37) under the integral and evaluating at $\eta=\eta_0$ gives

$$\frac{\partial\beta\left(\eta_0\right)}{\partial\eta'}=\mathbb{E}\left[\frac{\partial b\left(W;\eta_0\right)}{\partial\eta'}\right]+\mathbb{E}\left[b\left(W;\eta_0\right)t_\eta\left(W;\eta_0\right)\right]. \tag{39}$$

We can derive a close-form expression for $\frac{\partial b(w;\eta_0)}{\partial\eta'}$ in (39) by differentiating (38) with respect to η (and evaluating at $\eta=\eta_0$):

$$\begin{aligned}-\int\int\begin{pmatrix}1\\x\end{pmatrix}\frac{\partial a\left(w;\eta_0\right)}{\partial\eta'}f\left(x,y|\,w;\eta_0\right)\mathrm{d}x\mathrm{d}y-\int\int\begin{pmatrix}x'\\xx'\end{pmatrix}\frac{\partial b\left(w;\eta_0\right)}{\partial\eta'}f\left(x,y|\,w;\eta_0\right)\mathrm{d}x\mathrm{d}y&\\+\int\int\begin{pmatrix}1\\x\end{pmatrix}\left(y-a\left(w;\eta\right)-x'b\left(w;\eta_0\right)\right)s_\eta\left(x,y|\,w;\eta_0\right)f\left(x,y|\,w;\eta_0\right)\mathrm{d}x\mathrm{d}y&=0\end{aligned}$$

Using the matrix inverse

$$\mathbb{E}\left[\begin{array}{cc} 1 & X' \\ X & XX' \end{array}\middle| W=w\right]^{-1} = \left(\begin{array}{cc} 1+e_0(w)' v_0(w)^{-1} & e_0(w)-e_0(w)' v_0(w)^{-1} \\ -v_0(w)^{-1} e_0(w) & v_0(w)^{-1} \end{array}\right)$$

we solve to get

$$\begin{aligned}\left(\begin{array}{c} \frac{\partial a(w;\eta_0)}{\partial \eta'} \\ \frac{\partial b(w;\eta_0)}{\partial \eta'} \end{array}\right) &= \left(\begin{array}{cc} 1+e_0(w)' v_0(w)^{-1} & e_0(w)-e_0(w)' v_0(w)^{-1} \\ -v_0(w)^{-1} e_0(w) & v_0(w)^{-1} \end{array}\right) \\ &\quad \times \mathbb{E}\left[\left(\begin{array}{c} Y-a(W;\eta_0)-X'b(W;\eta_0) \\ X\left(Y-a(W;\eta_0)-X'b(W;\eta_0)\right) \end{array}\right) s_\eta(X,Y|W;\eta_0)\middle| W=w\right].\end{aligned}$$

Evaluating the second row of this expression gives

$$\frac{\partial b(w;\eta_0)}{\partial \eta'} = \mathbb{E}\left[v_0(W)^{-1}(X-e_0(W))(Y-a_0(W)-X'b_0(W))\, s_\eta(X,Y|W)\middle| W=w\right], \tag{40}$$

which, after substituting into (39), yields

$$\begin{aligned}\frac{\partial \beta(\eta_0)}{\partial \eta'} &= \mathbb{E}\left[v_0(W)^{-1}(X-e_0(W))(Y-a_0(W)-X'b_0(W))\, s_\eta(X,Y|W)\right] \\ &\quad +\mathbb{E}\left[b_0(W)\, t_\eta(W)\right].\end{aligned} \tag{41}$$

To demonstrate pathwise differentiability of β, we require $F(W,X,Y)$ such that

$$\frac{\partial \beta(\eta_0)}{\partial \eta'} = \mathbb{E}\left[F(W,X,Y)\, s_\eta(W,X,Y)'\right]. \tag{42}$$

Setting $F(W,X,Y)$ equal to $\psi_\beta^{\text{eff}}(Z,\beta_0,g_0(W),h_0(W))$, as defined in (17) of the main text, we get $\mathbb{E}\left[F(W,X,Y)\, s_\eta(W,X,Y)'\right]$ equal to (41) since, by Lemma 4.1 of Wooldridge (1999),

$$\mathbb{E}\left[(X-e_0(W))(Y-a_0(W)-X'b_0(W))|W\right] = 0$$

and iterated expectations (and the conditional mean zero property of the score $s_\eta(X,Y|W)$) further implies that $\mathbb{E}\left[(b_0(W)-\beta_0)\, s_\eta(X,Y|W)\right]=0$.

Step 3: Verification that the conjectured influence function equals the required projection:

Observe that $\psi_{\beta}^{\text{eff}}\left(Z,\beta_0,g\left(W\right),h\left(W\right)\right)$ lies in the model tangent space. Its first term is conditionally mean zero given W and hence plays the role of $s\left(X,Y|\,W\right)$. Its second term is a mean zero function of W alone and hence plays the role of $t\left(W\right)$. Since $\psi_{\beta}^{\text{eff}}\left(Z,\beta_0,g_0\left(W\right),h_0\left(W\right)\right)\in\mathcal{T}$, its projection onto $\mathcal{T}$ equals itself. Since equation (9) of Newey (1990, p. 106) is satisfied the result follows from his Theorem 3.1.

Proof of Theorem 2 (Large sample properties of $\hat{\beta}$)

Recall that $\lambda=\left(\alpha,\gamma',\delta'\right)'$ and

$$\underset{(1+J+JK)\times 1}{R}=\left(1,\left(W-\mu_W\right)',\left(\left(W-\mu_W\right)\otimes X\right)'\right)'.$$

In what follows we let λ_* denote value of λ which solves the just-identified population moments (28), (29) and (30). If Assumption 6 additionally holds in the sampled population, then we use λ_0 to denote the population value of λ. In this case λ_0 correctly specifies the form of the CLP of Y given X conditional on W.

In the Supplemental Web Appendix we show, without maintaining Assumption 6, that

$$\begin{aligned}\lambda_* =&\mathbb{E}\left[RR'\right]^{-1}\mathbb{E}\left[R\left(Y-X'\beta_0\right)\right]\\ =&\mathbb{E}\left[RR'\right]^{-1}\mathbb{E}\left[R\left(a_0\left(W\right)+X'\left(b_0\left(W\right)-\beta_0\right)\right)\right].\end{aligned}\tag{43}$$

Equation (43) implies that $R'\lambda_*$ is the mean squared error minimizing linear predictor of $a_0\left(W\right)+X'\left(b_0\left(W\right)-\beta_0\right)$ given R. This interpretation of λ_* is all that is required for the first part of Theorem 2.

We will also use the notation

$$U_0=\left(Y-R'\lambda_0-X'\beta_0\right)$$

and

$$U_*=\left(Y-R'\lambda_*-X'\beta_0\right).$$

Note that under Assumption 6 U_0 equals a *conditional* linear prediction error. However when Assumption 6 does not hold an implication of (43) is that U_* is still an *unconditional* linear predictor error.

We also use the shorthand $e_0\left(W\right)=e\left(W;\phi_0\right)$ and $v_0\left(W\right)=v\left(W;\phi_0\right)$ in order to simplify

some of the expressions presented below. Finally we use θ_0 to denote *both* $(\phi_0', \mu_W', \lambda_*', \beta_0')'$ and $(\phi_0', \mu_W', \lambda_0', \beta_0')'$, with the relevant case made clear by the context.

Next define the $(1 + J + JK) \times J$ and $K \times J$ matrices

$$\underset{(1+J+JK)\times J}{B_1} = \mathbb{E}\left[R\left\{\gamma_* + (I_J \otimes X)' \delta_*\right\}'\right] - \mathbb{E}\begin{bmatrix} 0 \\ 0 \\ (I_J \otimes X) U_* \end{bmatrix} \tag{44}$$

$$\underset{K\times J}{B_2} = \mathbb{E}\left[\left(v_0(W)^{-1}(X - e_0(W))\right)\left\{\gamma_* + (I_J \otimes X)' \delta_*\right\}'\right]. \tag{45}$$

Using this notation we can write the $(L + J + 1 + J + JK + K) \times (L + J + 1 + J + JK + K)$ Jacobian matrix of the moment vector as

$$M = -\begin{pmatrix} -\mathbb{H}(\phi_0) & 0 & 0 & 0 \\ 0 & I_J & 0 & 0 \\ 0 & -B_1 & \mathbb{E}[RR'] & \mathbb{E}[RX'] \\ \mathbb{E}\left[v_0(W)^{-1}(X - e_0(W)) U_* \mathbb{S}'\right] & -B_2 & 0 & I_K \end{pmatrix},$$

with $\mathbb{H}(\phi_0)$ equal to the $L \times L$ expected Hessian matrix associated with the generalized propensity score log-likelihood. The inverse Jacobian is therefore

$$M^{-1} = -\begin{pmatrix} -\mathbb{H}(\phi_0)^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ -\mathbb{E}[RR']^{-1}\mathbb{E}[RX']\mathbb{E}\left[v_0(W)^{-1}(X - e_0(W)) U_* \mathbb{S}'\right]\mathbb{H}(\phi_0)^{-1} & 0 & 0 \\ \mathbb{E}\left[v_0(W)^{-1}(X - e_0(W)) U_* \mathbb{S}'\right]\mathbb{H}(\phi_0)^{-1} & 0 & 0 \\ I_J & 0 & 0 \\ \mathbb{E}[RR']^{-1}(B_1 - B_2\mathbb{E}[RX']) & \mathbb{E}[RR']^{-1} & -\mathbb{E}[RR']^{-1}\mathbb{E}[RX'] \\ B_2 & 0 & I_K \end{pmatrix}. \tag{46}$$

Under Assumption 5 a key observation is that the expected value of (45) equals

$$\begin{aligned}\mathbb{E}\left[\left(v_0\left(W\right)^{-1}\left(X-e_0\left(W\right)\right)\right)\left\{\gamma_*+\left(I_J\otimes X\right)'\delta_*\right\}'\right] =& \mathbb{E}\left[\left(v_0\left(W\right)^{-1}\left(X-e_0\left(W\right)\right)\right)\gamma_*'\right] \\ &+\mathbb{E}\left[\left(v_0\left(W\right)^{-1}\left(X-e_0\left(W\right)\right)\right)\delta_*'\left(I_J\otimes X\right)\right] \\ =&0+\mathbb{E}\left[\left(v_0\left(W\right)^{-1}\left(X-e_0\left(W\right)\right)\right)\right. \\ &\left.\times\left(\begin{array}{ccc} X'\delta_{1*} & \cdots & X'\delta_{J*}\end{array}\right)\right] \\ =&\left(\begin{array}{ccc}\delta_{1*} & \cdots & \delta_{J*}\end{array}\right)=\Delta_*.\end{aligned}$$

Using this last equality, as well as the fact that under Assumption 5 we have $\mathbb{H}\left(\phi_0\right)=-\mathbb{E}\left[\mathbb{S}\mathbb{S}'\right]$, implies that the last K rows of $-M^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^N m\left(Z_i,\theta_0\right)+o_p\left(1\right)$ equal, after some manipulation,

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}-\beta_0\right)=&\frac{1}{\sqrt{N}}\sum_{i=1}^N\left\{v_0\left(W_i\right)^{-1}\left(X_i-e_0\left(W_i\right)\right)U_{*i}\right. \\ &-\mathbb{E}\left[v_0\left(W\right)^{-1}\left(X-e_0\left(W\right)\right)U_*\mathbb{S}'\right]\mathbb{E}\left[\mathbb{S}\mathbb{S}'\right]^{-1}\mathbb{S}_i \\ &\left.+\Delta_*\left(W_i-\mu_W\right)\right\}+o_p\left(1\right). \qquad (47)\end{aligned}$$

Next observe that we may decompose U_* as

$$\begin{aligned}U_*=&Y-R'\lambda_*-X'\beta_0 \\ =&Y-a_0\left(W\right)-X'b_0\left(W\right) \\ &+\left\{a_0\left(W\right)+X'\left(b_0\left(W\right)-\beta_0\right)-R'\lambda_*\right\} \\ =&U_0+\epsilon.\end{aligned}$$

Since $\mathbb{E}\left[U_*W\right]=0$ by the properties of linear predictors, $\mathbb{E}\left[U_0|\,W\right]=0$ by the properties of *conditional* linear predictors, and $U_*=U_0+\epsilon$, we have that $\mathbb{E}\left[\epsilon W\right]=0$. Defining $\tilde{\epsilon}=v_0\left(W\right)^{-1}\left(X-e_0\left(W\right)\right)\epsilon$ we can re-write 47 as

$$\begin{aligned}\sqrt{N}\left(\hat{\beta}-\beta_0\right)=&\frac{1}{\sqrt{N}}\sum_{i=1}^N\left\{v_0\left(W_i\right)^{-1}\left(X_i-e_0\left(W_i\right)\right)U_{0i}\right. \\ &\left.+\left(\tilde{\epsilon}_i-\Pi_{\tilde{\epsilon}\mathbb{S}}\mathbb{S}_i\right)+\Delta_*\left(W_i-\mu_W\right)\right\}+o_p\left(1\right) \qquad (48)\end{aligned}$$

where $\Pi_{\tilde{\epsilon}\mathbb{S}}=\mathbb{E}\left[\tilde{\epsilon}_i\mathbb{S}'\right]\mathbb{E}\left[\mathbb{S}\mathbb{S}'\right]^{-1}$. This gives the first implication of the Theorem. The second implication follows from the fact that $\epsilon=0$ and $\Delta_*\mathbb{V}\left(W\right)\Delta_*'=\mathbb{V}\left(b_0\left(W\right)\right)$ under Assumption 6.

Proof of Proposition 3 (Near global semiparametric efficiency)

Let $\mathbf{A}$ be an $m \times n$ matrix with $\|\mathbf{A}\|_F = \mathrm{Tr}\left(\mathbf{A}'\mathbf{A}\right)^{1/2}$ denoting the Frobenius matrix norm, $\|\mathbf{A}\|_2$ the spectral norm and recall that $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F$. Let $\mathbf{a}$ be an $n \times 1$ vector with Euclidean norm $\|\mathbf{a}\| = (\mathbf{a}'\mathbf{a})^{1/2}$. We make use of several matrix and probability inequalities in what follows. These are drawn from Hansen (2018, Appendices A & B) unless stated otherwise.

Let t be a non-zero column vector. The difference in the asymptotic variance of the estimate of the linear combination $t'\beta_0$ based upon $R^{(J)}$ and a corresponding semiparametrically efficient estimate is

$$\begin{aligned} t'\mathcal{I}^{(J)}\left(\beta_0\right)^{-1} t - t'\mathcal{I}\left(\beta_0\right)^{-1} t =& t'\Delta_*^{(J)}\mathbb{V}\left(k^{(J)}\left(W\right)\right)\left(\Delta_*^{(J)}\right)' t - t'\mathbb{V}\left(b_0\left(W\right)\right) t \\ &+ t'\mathbb{E}\left[\left(\tilde{\epsilon}^{(J)} - \Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right)\left(\tilde{\epsilon}^{(J)} - \Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right)'\right] t \\ \geq& 0. \end{aligned} \tag{49}$$

We seek to show that this variance difference is also bounded above by something that can be made arbitrarily close to zero.

To start observe that, after some manipulation (see the Supplemental Web Appendix) we can show that

$$\begin{aligned} \mathbb{V}\left(b_0\left(W\right) + \Delta_*^{(J)}\left(k^{(J)}\left(W\right) - \mu^{(J)}\right) - \beta_0\right) =& \Delta_*^{(J)}\mathbb{V}\left(k^{(J)}\left(W\right)\right)\left(\Delta_*^{(J)}\right)' - \mathbb{V}\left(b_0\left(W\right)\right) \\ &- 2\mathbb{E}\left[\left(b_0\left(W\right) - \beta_0\right)\right. \\ &\left.\times\left\{b_0\left(W\right) + \Delta_*^{(J)}\left(k^{(J)}\left(W\right) - \mu^{(J)}\right) - \beta_0\right\}'\right] \end{aligned} \tag{50}$$

Using (50) we can rewrite $t'\mathcal{I}^{(J)}\left(\beta_0\right)^{-1} t - t'\mathcal{I}\left(\beta_0\right)^{-1} t$ as

$$\begin{aligned} &t'\mathbb{V}\left(b_0\left(W\right) + \Delta_*^{(J)}\left(k^{(J)}\left(W\right) - \mu^{(J)}\right) - \beta_0\right) t \\ &+ 2t'\mathbb{E}\left[\left(b_0\left(W\right) - \beta_0\right)\left\{b_0\left(W\right) + \Delta_*^{(J)}\left(k^{(J)}\left(W\right) - \mu^{(J)}\right) - \beta_0\right\}'\right] t \\ &+ t'\mathbb{E}\left[\left(\tilde{\epsilon}^{(J)} - \Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right)\left(\tilde{\epsilon}^{(J)} - \Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right)'\right] t. \end{aligned} \tag{51}$$

Consider the first term in (51). The Quadratic Inequality (QI), Expectation Inequality (EI), and completeness of the sequence $\{k_j\left(W\right)\}_{j=1}^{\infty}$ (see equation (33)) give

$$t'\mathbb{V}\left(b_0\left(W\right)+\Delta_*^{(J)}\left(k^{(J)}\left(W\right)-\mu^{(J)}\right)-\beta_0\right)t\leq C_1\zeta^2, \tag{52}$$

with C_1 a constant.

Next consider the second term in (51). Applying the Cauchy-Schwarz inequality to this term yields

$$\begin{aligned}\left|t'\mathbb{E}\left[\left(b_0\left(W\right)-\beta_0\right)\left\{b_0\left(W\right)+\Delta_*^{(J)}\left(k^{(J)}\left(W\right)-\mu^{(J)}\right)-\beta_0\right\}'\right]t\right| \leq & \mathbb{V}\left(t'b_0\left(W\right)\right)^{1/2} \\ & \times\mathbb{V}\left(t'\left\{b_0\left(W\right)\right.\right. \\ & \left.\left.+\Delta_*^{(J)}\left(k^{(J)}\left(W\right)-\mu^{(J)}\right)-\beta_0\right\}\right)^{1/2}\end{aligned}$$

Again invoking completeness of the sequence $\{k_j\left(W\right)\}_{j=1}^{\infty}$, and also boundedness of the variance of $b_0\left(W\right)$, we then get

$$\left|t'\mathbb{E}\left[\left(b_0\left(W\right)-\beta_0\right)\left\{b_0\left(W\right)+\Delta_*^{(J)}\left(k^{(J)}\left(W\right)-\mu^{(J)}\right)-\beta_0\right\}'\right]t\right|\leq C_2\zeta, \tag{53}$$

with C_2 a constant (which depends on $\mathbb{V}\left(b_0\left(W\right)\right)$).

Finally consider the third term in (49). To analyze this term we start by writing the linear predictor approximation error of $\left(R^{(J)}\right)'\lambda_*^{(J)}$ for $a_0\left(W\right)+X'\left(b_0\left(W\right)-\beta_0\right)$ as

$$\begin{aligned}\epsilon^{(J)}=&\left\{a_0\left(W\right)+X'\left(b_0\left(W\right)-\beta_0\right)-\left(R^{(J)}\right)'\lambda_*^{(J)}\right\} \\ =&a\left(W\right)-\alpha_*^{(J)}-\left(k^{(J)}\left(W\right)-\mu^{(J)}\right)'\gamma_*^{(J)} \\ &+X'\left(b_0\left(W\right)-\beta_0-\Delta_*^{(J)}\left(k^{(J)}\left(W\right)-\mu^{(J)}\right)\right) \\ =&\left(1,X'\right)\delta^{(J)}\left(W\right),\end{aligned}$$

with the final equality following from definition (34). The EI and the fact that, for $\mathbf{a}$ and $\mathbf{b}$ $m\times 1$ vectors $\left\|\mathbf{a}\mathbf{b}'\right\|_F=\left\|\mathbf{a}\right\|\left\|\mathbf{b}\right\|$, then gives

$$\left\|\mathbb{E}\left[\left(\tilde{\epsilon}^{(J)}-\Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right)\left(\tilde{\epsilon}^{(J)}-\Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right)'\right]\right\|\leq\mathbb{E}\left[\left\|\tilde{\epsilon}^{(J)}-\Pi_{\tilde{\epsilon}\mathbb{S}}^{(J)}\mathbb{S}\right\|^2\right]$$

with

$$\tilde{\epsilon}^{(J)}=v_0\left(W\right)^{-1}\left(X-e_0\left(W\right)\right)\left\{\left(1,X'\right)\delta^{(J)}\left(W\right)\right\}.$$

By the norm-reducing property of projection and Schwarz Matrix Inequality (SMI) we further

get that

$$\begin{aligned}\left\|\tilde{\epsilon}^{(J)} - \Pi_{\tilde{\mathbb{S}}}^{(J)}\mathbb{S}\right\| &\leq \left\|\tilde{\epsilon}^{(J)}\right\| \\ &= \left\|v_0\left(W\right)^{-1}\left(X - e_0\left(W\right)\right)\left\{\left(1, X'\right)\delta^{(J)}\left(W\right)\right\}\right\| \\ &\leq \left\|v_0\left(W\right)^{-1}\left(X - e_0\left(W\right)\right)\left\{\left(1, X'\right)\right\}\right\|\left\|\delta^{(J)}\left(W\right)\right\|.\end{aligned}$$

Applying the expectation operator, invoking Assumption 2, and using the compact support assumption for (W, X), finally gives

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\epsilon}^{(J)} - \Pi_{\tilde{\mathbb{S}}}^{(J)}\mathbb{S}\right\|^2\right] \leq& \mathbb{E}\left[\left\|v_0\left(W\right)^{-1}\left(X - e_0\left(W\right)\right)\left\{\left(1, X'\right)\right\}\right\|^2\left\|\delta^{(J)}\left(W\right)\right\|^2\right] \\ \leq& C_3\mathbb{E}\left[\left\|\delta^{(J)}\left(W\right)\right\|^2\right] \\ \leq& C_3\zeta^2 \end{aligned} \tag{54}$$

with $C_3 = \sup_{w,x\in\mathbb{W},\mathbb{X}}\left\|v_0\left(W\right)^{-1}\left(X - e_0\left(W\right)\right)\left\{\left(1, X'\right)\right\}\right\|^2$.
Applying the TI to (51) and using terms (52), (53) and (54) then gives the bound

$$0 \leq t'\mathcal{I}^{(J)}\left(\beta_0\right)^{-1}t - t'\mathcal{I}\left(\beta_0\right)^{-1}t \leq \left(C_1 + C_3\right)\zeta^2 + C_2\zeta. \tag{55}$$

Since ζ is arbitrary the limit of the difference in (55) is zero.

References

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2), 249 – 288.

Angrist, J. D., Graddy, K., & Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies*, 67(3), 499 – 527.

Angrist, J. D. & Krueger, A. B. (1999). *Handbook of Labor Economics*, volume 3, chapter Empirical strategies in labor economics, (pp. 1277 – 1366.). North-Holland: Amsterdam.

Angrist, J. D. & Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962 – 973.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2), 608 – 650.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. New York: Springer-Verlag.

Blundell, R. & Powell, J. L. (2003). *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, volume 2, chapter Endogeneity in nonparametric and semiparametric regression models, (pp. 312 – 357). Cambridge University Press.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2), 138 – 154.

Chamberlain, G. (1984). *Handbook of Econometrics*, volume 2, chapter Panel Data, (pp. 1247 – 1318). North-Holland: Amsterdam.

Chamberlain, G. (1986). *Notes on Semiparametric Regression*. Working paper, University of Wisconsin - Madison.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3), 305 – 334.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, 60(3), 567 – 596.

Chen, X., Hong, H., & Tarozzi, A. (2008). Semiparametric efficiency in gmm models with auxiliary data. *Annals of Statistics*, 36(2), 808 – 843.

Frölich, M. (2004). A note on the role of the propensity score for estimating average treatment effects. *Econometric Reviews*, 23(2), 167 – 174.

Goldberger, A. S. (1991). *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Gottfried, M. A. & Kirksey, J. J. (2017). "when" students miss school: the role of timing of absenteeism on students' test performance. *Educational Researcher*, 46(3), 119 – 130.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79(2), 437 – 452.

Graham, B. S., Imbens, G. W., & Ridder, G. (2010). *Measuring the effects of segregation in the presence of social spillovers: a nonparametric approach*. Working Paper 16499, NBER.

Graham, B. S., Pinto, C., & Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *Review of Economic Studies*, 79(3), 1053 – 1079.

Graham, B. S., Pinto, C., & Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business and Economic Statistics*, 31(2), 288 – 301.

Groves, T. & Rothenberg, T. (1969). A note on the expected value of an inverse matrix. *Biometrika*, 56(3), 690 – 691.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2), 315 – 331.

Hansen, B. (2018). Econometrics.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society B*, 55(4), 757 – 796.

Henderson, H. V. & Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1), 53 – 60.

Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4), 259 – 278.

Hirano, K. & Imbens, G. W. (2004). *Applied Bayesian Modelling and Causal Inference from Missing Data Perspectives*, chapter The propensity score with continuous treatments, (pp. 73 – 84). John Wiley & Sons, Inc.: New York.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4), 1161 – 1189.

Hitomi, K., Nishiyama, Y., & Okui, R. (2008). A puzzling phenomenon in semiparametric estimation problems with infinite-dimensional nuisance parameters. *Econometric Theory*, 24(6), 1717 – 1728.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functio. *Biometrika*, 87(3), 706 – 710.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press.

Kline, P. (2014). A note on variance estimation for the oaxaca estimator of average treatment effects. *Economics Letters*, 122(3), 428 – 431.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2), 99 – 135.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(2), 233 – 253.

Newey, W. K. & McFadden, D. (1994). *Handbook of Econometrics*, volume 4, chapter Large sample estimation and hypothesis testing, (pp. 2111 – 2245). North-Holland: Amsterdam.

Pencavel, J. (1986). *Handbook of Labor Economics*, volume 1, chapter Labor supply of men: a survey, (pp. 3 – 102). North-Holland: Amsterdam.

Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2), 479 – 495.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of American Statistical Association*, 89(427), 846 – 866.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4), 931 – 954.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41 – 55.

Ruud, P. A. (1986). Consistent estimation of limited dependent variable models despite misspecification of distribution. *Journal of Econometrics*, 32(1), 157–187.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models: rejoinder. *Journal of American Statistical Association*, 94(448), 1135 – 1146.

Sloczynski, T. (2015). The oaxaca–blinder unexplained component as a treatment effects estimator. *Oxford Bulletin of Economics and Statistics*, 77(4), 588 – 604.

Sloczynski, T. (2017). *A general weighted average representation of the ordinary and two-stage least squares estimands*. Working paper, Brandies University.

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

Wooldridge, J. M. (1999). Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics*, 90(1), 77 – 97.

Wooldridge, J. M. (2004). *Estimating average partial effects under conditional moment independence assumptions*. Working Paper CWP03/04, CeMMAP.

Wooldridge, J. M. (2005). *Identification and inference for econometric models*, chapter Unobserved heterogeneity and the estimation of average partial effects, (pp. 27 – 55). Number 3. Cambridge University Press: Cambridge.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2), 1281 – 1301.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2nd edition.

Yitzhaki, S. (1996). On using linear regressions in welfare economics. *Journal of Business and Economic Statistics*, 14(4), 478 – 486.

Yule, G. U. (1899). 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(6), 249 – 295.