

Research Note

Marginal Treatment Effects from a Propensity Score Perspective

Xiang Zhou

Harvard University

Yu Xie

Princeton University

We offer a propensity score perspective to interpret and analyze the marginal treatment effect (MTE). Specifically, we redefine MTE as the expected treatment effect conditional on the propensity score and a latent variable representing unobserved resistance to treatment. As with the original MTE, the redefined MTE can be used as a building block for constructing standard causal estimands. The weights associated with the new MTE, however, are simpler, more intuitive, and easier to compute. Moreover, the redefined MTE immediately reveals treatment effect heterogeneity among individuals at the margin of treatment, enabling us to evaluate a wide range of policy effects.

I. Introduction

An essential feature common to all empirical social research is variability across units of analysis. Individuals differ not only in background

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under grant R01-HD-074603-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors benefited from communications with Daniel Almirall, Matthew Blackwell, Jennie Brand, James Heckman, Jeffrey Smith, Edward Vytlačil, and three anonymous reviewers.

Electronically published October 28, 2019

[*Journal of Political Economy*, 2019, vol. 127, no. 6]

© 2019 by The University of Chicago. All rights reserved. 0022-3808/2019/12706-0011\$10.00

characteristics but also in how they respond to a particular treatment, intervention, or stimulation. Moreover, individuals may self-select into treatment on the basis of their anticipated treatment effects in a way that is not captured by observed covariates. This is likely when individuals (or their agents) possess more knowledge than the researcher about their gains (or losses) from treatment and act on it (Roy 1951; Björklund and Moffitt 1987; Heckman and Vytlačil 2007a). To study heterogeneous treatment effects in the presence of unobserved self-selection, Heckman and Vytlačil (1999, 2001a, 2005, 2007b) have developed a structural approach that builds on the marginal treatment effect (MTE). Under a latent index model of treatment assignment, the MTE is defined as the expected treatment effect given observed covariates and a latent variable representing unobserved, individual-specific resistance to treatment. A wide range of causal parameters, such as the average treatment effect (ATE) and the treatment effect of the treated (TT), can be expressed as weighted averages of MTE. Moreover, MTE can be used to evaluate ATEs for individuals at the margin of indifference to treatment, thus allowing the researcher to assess the efficacy of marginal policy changes (Carneiro, Heckman, and Vytlačil 2010, 2011).

In the MTE framework, the latent index model ensures that all unobserved determinants of treatment status are summarized by a single latent variable and that the variation of treatment effect by this latent variable captures all of the unobserved treatment effect heterogeneity that may cause selection bias. Our basic intuition is that, under this model, treatment effect heterogeneity that is consequential for selection bias occurs only along two dimensions: (1) the observed probability of treatment (i.e., the propensity score) and (2) the latent variable for unobserved resistance to treatment. In other words, after unobserved selection is factored in through the latent variable, the propensity score is the only dimension along which treatment effect may be correlated with treatment status. Therefore, to identify population- and subpopulation-level causal effects, such as ATE and TT, it would be sufficient to model treatment effect as a bivariate function of the propensity score and the latent variable. In this paper, we show that such a bivariate function is not only analytically sufficient but also crucial to the evaluation of policy effects.

Specifically, we redefine MTE as the expected treatment effect conditional on the propensity score (instead of the entire vector of observed covariates) and the latent variable representing unobserved resistance to treatment. This redefinition offers a novel perspective to interpret and analyze MTE that supplements the current approach. First, although projected onto a unidimensional summary of covariates, the redefined MTE is sufficient to capture all of the treatment effect heterogeneity that is consequential for selection bias. Thus, as with the original MTE, it can also be used as a building block for constructing standard causal parameters,

such as ATE and TT. The weights associated with the new MTE, however, are simpler, more intuitive, and easier to compute. Second, by discarding treatment effect variation that is orthogonal to the two-dimensional space spanned by the propensity score and the latent variable, the redefined MTE is a bivariate function, easier to visualize than the original MTE. Finally, the redefined MTE immediately reveals treatment effect heterogeneity among individuals who are at the margin of treatment. As a result, it can be used to evaluate a wide range of policy effects with little analytical twist and to design policy interventions that optimize the marginal benefits of treatment. To facilitate practice, we also provide an R package, *localIV*, for estimating the redefined MTE as well as the original MTE via local instrumental variables (LIVs; Zhou 2019).

Of course, this paper is not the first to characterize the selection problem using the propensity score. Since the seminal work of Rosenbaum and Rubin (1983), propensity-score-based methods, such as matching, weighting, and regression adjustment, have been a mainstay strategy for drawing causal inferences in the social sciences. In a series of papers, Heckman and colleagues have established the key roles of the propensity score in a variety of econometric methods, including matching, control functions, instrumental variables (IVs), and the MTE approach (Heckman and Robb 1986; Heckman and Hotz 1989; Heckman and Navarro-Lozano 2004; Heckman 2010). In the MTE approach, for example, incremental changes in the propensity score serve as LIVs that identify the MTE at various values of the unobserved resistance to treatment. Moreover, the weights with which MTE can be aggregated up to standard causal parameters depend solely on the conditional distribution of the propensity score given covariates. In this paper, we show that the propensity score offers not only a tool for identification but also a perspective from which we can better summarize, interpret, and analyze treatment effect heterogeneity due to both observed and unobserved characteristics.

II. MTEs: A Review

The MTE approach builds on the generalized Roy model for discrete choices (Roy 1951; Heckman and Vytlačil 2007a). Consider two potential outcomes, Y_1 and Y_0 ; a binary indicator for treatment status, D ; and a vector of pretreatment covariates, X . The variable Y_1 denotes the potential outcome if the individual were treated ($D = 1$), and Y_0 denotes the potential outcome if the individual were not treated ($D = 0$). We specify the outcome equations as

$$Y_0 = \mu_0(X) + \epsilon, \quad (1)$$

$$Y_1 = \mu_1(X) + \epsilon + \eta, \quad (2)$$

where $\mu_0(X) = \mathbb{E}[Y_0|X]$, $\mu_1(X) = \mathbb{E}[Y_1|X]$, the error term ϵ captures all unobserved factors that affect the baseline outcome (Y_0), and the error term η captures all unobserved factors that affect the treatment effect ($Y_1 - Y_0$). Treatment assignment is represented by a latent index model. Let I_D be a latent tendency for treatment, which depends on both observed (Z) and unobserved (V) factors:

$$I_D = \mu_D(Z) - V, \tag{3}$$

$$D = \mathbb{1}(I_D > 0), \tag{4}$$

where $\mu_D(Z)$ is an unspecified function and V is a latent random variable representing unobserved, individual-specific resistance to treatment, assumed to be continuous with a strictly increasing distribution function. The Z vector includes all of the components of X , but it also includes IVs that affect only the treatment status D . The key assumptions associated with equations (1)–(4) are as follows:

ASSUMPTION 1 (Independence). (ϵ, η, V) are statistically independent of Z given X .

ASSUMPTION 2 (Rank condition). $\mu_D(Z)$ is a nontrivial function of Z given X .

The latent index model characterized by equations (3) and (4), combined with assumptions 1 and 2, is equivalent to the Imbens and Angrist (1994) assumptions of independence and monotonicity for the interpretation of IV estimands as local ATEs (LATEs; Vytlacil 2002).

To define the MTE, it is best to rewrite the treatment assignment equations (3) and (4) as

$$\begin{aligned} D &= \mathbb{1}(F_{V|X}(\mu_D(Z)) - F_{V|X}(V) > 0) \\ &= \mathbb{1}(P(Z) - U > 0), \end{aligned} \tag{5}$$

where $F_{V|X}(\cdot)$ is the cumulative distribution function of V given X and $P(Z) = \Pr(D = 1|Z) = F_{V|X}(\mu_D(Z))$ denotes the propensity score given Z . The expression $U = F_{V|X}(V)$ is the quantile of V given X , which by definition follows a standard uniform distribution. From equation (5) we can see that Z affects treatment status only through the propensity score $P(Z)$. The MTE is defined as the expected treatment effect conditional on pretreatment covariates $X = x$ and the normalized latent variable $U = u$:

$$\begin{aligned} \text{MTE}(x, u) &= \mathbb{E}[Y_1 - Y_0|X = x, U = u] \\ &= \mathbb{E}[\mu_1(X) - \mu_0(X) + \eta|X = x, U = u] \\ &= \mu_1(x) - \mu_0(x) + \mathbb{E}[\eta|X = x, U = u]. \end{aligned}$$

Since U is the quantile of V , the variation of $\text{MTE}(x, u)$ over values of u reflects how treatment effect varies with different quantiles of the unobserved resistance to treatment.

A wide range of causal parameters, such as ATE and TT, can be expressed as weighted averages of $\text{MTE}(x, u)$. To obtain population-level causal effects, $\text{MTE}(x, u)$ needs to be integrated twice, first over u given $X = x$ and then over x . The weights for integrating over u are detailed in Heckman, Urzua, and Vytlačil (2006a). It bears noting that the estimation of weights can be challenging in practice (except for the ATE case), as it involves estimating the conditional density of $P(Z)$ given X and the latter is usually a high-dimensional vector.

Given assumptions 1 and 2, $\text{MTE}(x, u)$ can be nonparametrically identified using the method of LIVs.¹ Specifically, for any (x, u) within the support of the joint distribution of X and $P(Z)$, $\text{MTE}(x, u)$ can be identified as the partial derivative of $\mathbb{E}[Y|X = x, P(Z) = p]$ with respect to p :

$$\text{MTE}(x, u) = \left. \frac{\partial \mathbb{E}[Y|X = x, P(Z) = p]}{\partial p} \right|_{p=u}.$$

In practice, however, it is difficult to condition on X nonparametrically, especially when X is high-dimensional. Therefore, in most empirical work using LIV, it is assumed that (X, Z) is jointly independent of (ϵ, η, V) (e.g., Carneiro and Lee 2009; Carneiro et al. 2011; Maestas, Mullen, and Strand 2013). Under this assumption, the MTE is additively separable in x and u :

$$\begin{aligned} \text{MTE}(x, u) &= \mu_1(x) - \mu_0(x) + \mathbb{E}[\eta|X = x, U = u] \\ &= \mu_1(x) - \mu_0(x) + \mathbb{E}[\eta|U = u]. \end{aligned} \tag{6}$$

The additive separability not only simplifies estimation but also allows $\text{MTE}(x, u)$ to be identified over $\text{supp}(X) \times \text{supp}(P(Z))$ (instead of $\text{supp}(X, P(Z))$). The above equation also suggests a necessary and sufficient condition for the MTE to be additively separable:

ASSUMPTION 3 (Additive separability). $\mathbb{E}[\eta|X = x, U = u]$ does not depend on x .

This assumption is implied by (but does not imply) the full independence between (X, Z) and (ϵ, η, V) (for a similar discussion, see Brinch et al. [2017]).

In most applied work, the conditional means of Y_0 and Y_1 given X are further specified as linear in parameters: $\mu_0(X) = \beta_0^T X$ and $\mu_1(X) = \beta_1^T X$. In this case, $\mathbb{E}[Y|X = x, P(Z) = p]$ can be written as

¹ An alternative method to nonparametrically identify the MTE is based on separate estimation of $\mathbb{E}[Y|P(Z), X, D = 0]$ and $\mathbb{E}[Y|P(Z), X, D = 1]$ (see Heckman and Vytlačil 2007b; Brinch, Mogstad, and Wiswall 2017).

$$\mathbb{E}[Y|X = x, P(Z) = p] = \beta_0^T x + (\beta_1 - \beta_0)^T xp + \underbrace{\int_0^p \mathbb{E}[\eta|U = u] du}_{K(p)},$$

where $K(p)$ can be estimated either nonparametrically or with some functional form restrictions.² The MTE is then estimated as

$$\widehat{\text{MTE}}(x, u) = (\hat{\beta}_1 - \hat{\beta}_0)^T x + \hat{K}'(u), \tag{7}$$

where $K'(u)$ is the derivative of $K(p)$ evaluated at u . Heckman et al. (2006) provide a detailed discussion of different estimation methods.

III. A Redefinition of MTE

Under the generalized Roy model, a single latent variable U not only summarizes all unobserved determinants of treatment status but also captures all the treatment effect heterogeneity by unobserved characteristics that may cause selection bias. In fact, the latent index structure implies that all the treatment effect heterogeneity that is consequential for selection bias occurs along only two dimensions: (1) the propensity score $P(Z)$ and (2) the latent variable U representing unobserved resistance to treatment. This is directly reflected in equation (5): a person is treated if and only if her propensity score exceeds her (realized) latent resistance u . Therefore, given both $P(Z)$ and U , treatment status D is fixed (either 0 or 1) and thus independent of treatment effect:

$$Y_1 - Y_0 \perp\!\!\!\perp D|P(Z), U.$$

Thus, to characterize selection bias, it is sufficient to model treatment effect as a bivariate function of the propensity score and the latent variable U . We redefine MTE as the expected treatment effect given $P(Z)$ and U :

$$\widetilde{\text{MTE}}(p, u) \triangleq \mathbb{E}[Y_1 - Y_0|P(Z) = p, U = u].$$

Compared with the original MTE, $\widetilde{\text{MTE}}(p, u)$ is a more parsimonious representation of all the treatment effect heterogeneity that is relevant for selection bias. Moreover, by discarding treatment effect variation that is orthogonal to the two-dimensional space spanned by $P(Z)$ and U , $\widetilde{\text{MTE}}(p, u)$ is a bivariate function, easier to visualize than $\text{MTE}(x, u)$.

As with $\text{MTE}(x, u)$, $\widetilde{\text{MTE}}(p, u)$ can also be used as a building block for constructing standard causal parameters, such as ATE and TT. However, compared with the weights on $\text{MTE}(x, u)$, the weights on $\widetilde{\text{MTE}}(p, u)$ are

² When the analysis is conditional on X and the instruments $Z \setminus X$ are discrete, $P(Z)$ can take only a finite number of values. In this case, functional form restrictions have to be imposed on $P(Z)$. See Brinch et al. (2017).

simpler, more intuitive, and easier to compute. The weights for ATE, TT, and treatment effect of the untreated (TUT) are shown in the first three rows of table 1.³ To construct $\text{ATE}(p)$, we simply integrate $\widetilde{\text{MTE}}(p, u)$ against the marginal distribution of U —a standard uniform distribution. To construct $\text{TT}(p)$, we integrate $\widetilde{\text{MTE}}(p, u)$ against the truncated distribution of U given $U < p$. Similarly, to construct $\text{TUT}(p)$, we integrate $\widetilde{\text{MTE}}(p, u)$ against the truncated distribution of U given $U \geq p$. To obtain population-level ATE, TT, and TUT, we further integrate $\text{ATE}(p)$, $\text{TT}(p)$, and $\text{TUT}(p)$ against appropriate marginal distributions of $P(Z)$. For example, to construct TT, we integrate $\text{TT}(p)$ against the marginal distribution of the propensity score among treated units.

IV. Identification and Estimation of $\widetilde{\text{MTE}}(p, u)$

As with $\text{MTE}(x, u)$, the regions over which $\widetilde{\text{MTE}}(p, u)$ is identified depend on whether assumption 3 (additive separability) is invoked.⁴ Let us first look at the case without additive separability. From assumption 1, we know that $U \perp\!\!\!\perp P(Z)|X$. Since U follows a standard uniform distribution for each $X = x$, we also have $U \perp\!\!\!\perp X$. By the rules of conditional independence, we have $U \perp\!\!\!\perp X|P(Z)$. Using this fact and the law of total expectation, we can write $\widetilde{\text{MTE}}(p, u)$ as

$$\begin{aligned} \widetilde{\text{MTE}}(p, u) &= \mathbb{E}_{X|P(Z)=p, U=u}[\mathbb{E}[Y_1 - Y_0|P(Z) = p, U = u, X]] \\ &= \mathbb{E}_{X|P(Z)=p}[\mathbb{E}[\mu_1(X) - \mu_0(X) + \eta|P(Z) = p, U = u, X]] \\ &= \mathbb{E}_{X|P(Z)=p}[\mathbb{E}[\mu_1(X) - \mu_0(X) + \eta|U = u, X]] \quad (8) \\ &\quad (\text{because } (\eta, U) \perp\!\!\!\perp P(Z)|X) \\ &= \mathbb{E}_{X|P(Z)=p}[\text{MTE}(X, u)]. \end{aligned}$$

Thus, $\widetilde{\text{MTE}}(p, u)$ is no more than the conditional expectation of $\text{MTE}(x, u)$ given $P(Z) = p$. As discussed above, with assumptions 1 and 2, $\text{MTE}(x, u)$ is identified over the support of the joint distribution of X and $P(Z)$. Thus, for a given u , $\text{MTE}(x, u)$ is identified if and only if $x \in \text{supp}(X|P(Z) = u)$. Yet to evaluate $\widetilde{\text{MTE}}(p, u)$ from equation (8), we need to know $\text{MTE}(x, u)$ for all $x \in \text{supp}(X|P(Z) = p)$. Therefore, for a given (p, u) pair, we can identify $\widetilde{\text{MTE}}(p, u)$ if $\text{supp}(X|P(Z) = p) \subseteq \text{supp}(X|P(Z) = u)$. For general $p \neq u$, this condition can be quite restrictive. However, for

³ When $P(Z) = p$ is treated as a random variable, weights of the same form as in table 1 can be used to construct population-level causal parameters directly from $\text{MTE}(x, u)$ (see Mogstad, Santos, and Torgovitsky 2017).

⁴ We thank an anonymous reviewer for suggesting the identification conditions for $\widetilde{\text{MTE}}(p, u)$.

TABLE 1
WEIGHTS FOR CONSTRUCTING \widehat{ATE} , TT, TUT, PRTE,
AND MPRTTE FROM $MTE(p, u)$

Quantity of Interest	Weight
$ATE(p)$	$h_{ATE}(p, u) = 1$
$TT(p)$	$h_{TT}(p, u) = 1(u < p)/p$
$TUT(p)$	$h_{TUT}(p, u) = 1(u \geq p)/1 - p$
$PRTE(p, \lambda(p))$	$h_{PRTE}(p, u) = 1(p \leq u < p + \lambda(p))/\lambda(p)$
$MPRTTE(p)$	$h_{MPRTTE}(p, u) = \delta(u - p)$

NOTE.— $\delta(\cdot)$ is the Dirac delta function.

the particular case in which $p = u$, this condition is trivially satisfied. Thus, for any $p \in \text{supp}(P(Z))$, $\widehat{MTE}(p, p)$ can be identified as

$$\widehat{MTE}(p, p) = \mathbb{E}_{X|P(Z)=p} \frac{\partial \mathbb{E}[Y|X, P(Z) = p]}{\partial p},$$

which is a univariate function of p that reflects the effects of treatment among individuals who are at the margin of indifference to treatment. As we will see, it plays a prominent role in the evaluation of policy effects.

When assumption 3 is invoked (as in most empirical work with MTE), $MTE(x, u)$ is identified for any $(x, u) \in \text{supp}(X) \times \text{supp}(P(Z))$. That is, for each $u \in \text{supp}(P(Z))$, $MTE(x, u)$ is identified over the marginal support of X . Thus, for any $(p, u) \in \text{supp}(P(Z)) \times \text{supp}(P(Z))$, we can identify $\widehat{MTE}(p, u)$ through equation (8). Since $MTE(x, u)$ can now be partitioned into a function of x and a function of u , evaluation of equation (8) will be straightforward. For example, when $\mu_0(X)$ and $\mu_1(X)$ are specified as linear in parameters, $MTE(x, u)$ can be estimated as equation (7). To obtain estimates of $\widehat{MTE}(p, u)$, we need only one more step: fit a nonparametric curve of $(\hat{\beta}_1 - \hat{\beta}_0)^T x$ with respect to \hat{p} (e.g., using a local linear regression) and combine it with existing estimates of $K'(u)$.

V. Policy-Relevant Causal Effects

The redefined MTE can be used not only to construct standard causal parameters but also, in the context of program evaluation, to draw implications for the ways in which the program should be revised in the future. To predict the impact of an expansion (or a contraction) in program participation, one needs to examine treatment effects for those individuals who would be affected by such an expansion (or contraction). To formalize this idea, Heckman and Vytlačil (2001b, 2005) define the policy-relevant treatment effect (PRTE) as the mean effect of moving from a baseline policy to an alternative policy per net person shifted into treatment; that is,

$$\text{PRTE} \triangleq \frac{\mathbb{E}(Y|\text{Alternative Policy}) - \mathbb{E}(Y|\text{Baseline Policy})}{\mathbb{E}(D|\text{Alternative Policy}) - \mathbb{E}(D|\text{Baseline Policy})}.$$

They further show that under the generalized Roy model, the PRTE depends on a policy change only through its impacts on the distribution of the propensity score $P(Z)$. Specifically, conditional on $X = x$, the PRTE can be written as a weighted average of $\text{MTE}(x, u)$, where the weights depend only on the distribution of $P(Z)$ before and after the policy change. Within this framework, Carneiro et al. (2010) further define the marginal PRTE (MPRTE) as a directional limit of the PRTE as the alternative policy converges to the baseline policy. Denoting by $F(\cdot)$ the cumulative distribution function of $P(Z)$, they consider a set of alternative policies indexed by a scalar α , $\{F_\alpha : \alpha \in \mathbb{R}\}$ such that F_0 corresponds to the baseline policy. The MPRTE is defined as

$$\text{MPRTE} = \lim_{\alpha \rightarrow 0} \text{PRTE}(F_\alpha).$$

We follow their approach to analyzing policy effects but without conditioning on X . While Carneiro et al. (2010) assume that the effects of all policy changes are through shifts in the conditional distribution of $P(Z)$ given X , we focus on policy changes that shift the marginal distribution of $P(Z)$ directly. In other words, we consider policy interventions that incorporate individual-level treatment effect heterogeneity by values of $P(Z)$, whether their differences in $P(Z)$ are determined by their baseline characteristics X or the IVs $Z \setminus X$. In a companion paper (Zhou and Xie, forthcoming), we provide a more detailed comparison between these two approaches.

Specifically, let us consider a class of policy changes under which the i th individual's propensity of treatment is boosted by $\lambda(p_i)$ (in a way that does not change her treatment effect), where p_i denotes her propensity score $P(z_i)$ and $\lambda(\cdot)$ is a positive, real-valued function such that $p + \lambda(p) \leq 1$ for all $p \in [0, 1)$. Thus, the policy change nudges everyone in the same direction, and two persons with the same baseline probability of treatment share an inducement of the same size. For such a policy change, the PRTE given $P(Z) = p < 1$ and $\lambda(p)$ becomes

$$\text{PRTE}(p, \lambda(p)) = \mathbb{E}[Y_1 - Y_0 | p(Z) = p, p \leq U < p + \lambda(p)].$$

As with standard causal parameters, $\text{PRTE}(p, \lambda(p))$ can be expressed as a weighted average of $\text{MTE}(p, u)$:

$$\text{PRTE}(p, \lambda(p)) = \frac{1}{\lambda(p)} \int_p^{p+\lambda(p)} \widetilde{\text{MTE}}(p, u) du.$$

Here the weight on u is constant (i.e., $1/\lambda(p)$) within the interval of $[p, p + \lambda(p))$ and zero elsewhere.

To examine the effects of marginal policy changes, let us consider a sequence of policy changes indexed by a real-valued scalar α . Given $P(Z) = p$, we define the MP RTE as the limit of PRTE($p, \alpha\lambda(p)$) as α approaches zero:

$$\begin{aligned} \text{MP RTE}(p) &= \lim_{\alpha \rightarrow 0} \text{PRTE}(p, \alpha\lambda(p)) \\ &= \mathbb{E}(Y_1 - Y_0 | P(Z) = p, U = p) \\ &= \widetilde{\text{MTE}}(p, p). \end{aligned}$$

Hence, we have established a direct link between MP RTE(p) and $\widetilde{\text{MTE}}(p, u)$: at each level of the propensity score, the MP RTE is simply the $\widetilde{\text{MTE}}$ at the margin where $u = p$. As shown in the last row of table 1, MP RTE(p) can also be expressed as a weighted average of $\widetilde{\text{MTE}}(p, u)$ using the Dirac delta function. This quantity, as noted in the previous section, can be nonparametrically identified even without the assumption of additive separability.

The relationships between ATE, TT, TUT, and MP RTE are graphically illustrated in figure 1. Figure 1A shows a shaded gray plot of $\widetilde{\text{MTE}}(p, u)$ for heterogeneous treatment effects in a hypothetical setup. In this plot, both the propensity score p and the latent resistance u (both ranging from 0 to 1) are divided into 10 equally spaced strata, yielding 100 grids, and a darker grid indicates a higher treatment effect. The advantage of such a shaded gray plot is that we can use subsets of the 100 grids to represent meaningful subpopulations. For example, we present the grids for treated units in figure 1B, untreated units in figure 1C, and marginal units in figure 1D. Thus, evaluating ATE, TT, TUT, and MP RTE simply means taking weighted averages of $\widetilde{\text{MTE}}(p, u)$ over the corresponding subsets of grids.

VI. Treatment Effect Heterogeneity among Marginal Entrants

For policy makers, a key question of interest would be how MP RTE(p) varies with the propensity score p . To gain some intuition, let us consider the functional structure of MP RTE(p) under the assumption of additive separability. Substituting equation (6) into equation (8), we can see that MP RTE(p) consists of two components:

$$\text{MP RTE}(p) = \mathbb{E}[\mu_1(X) - \mu_0(X) | P(Z) = p] + \mathbb{E}[(\eta | U = p)]. \quad (9)$$

The first component reflects treatment effect heterogeneity by the propensity score, and the second component reflects treatment effect heterogeneity by the latent resistance U . Among marginal entrants, $P(Z)$ is equal to U so that these two components fall on the same dimension.

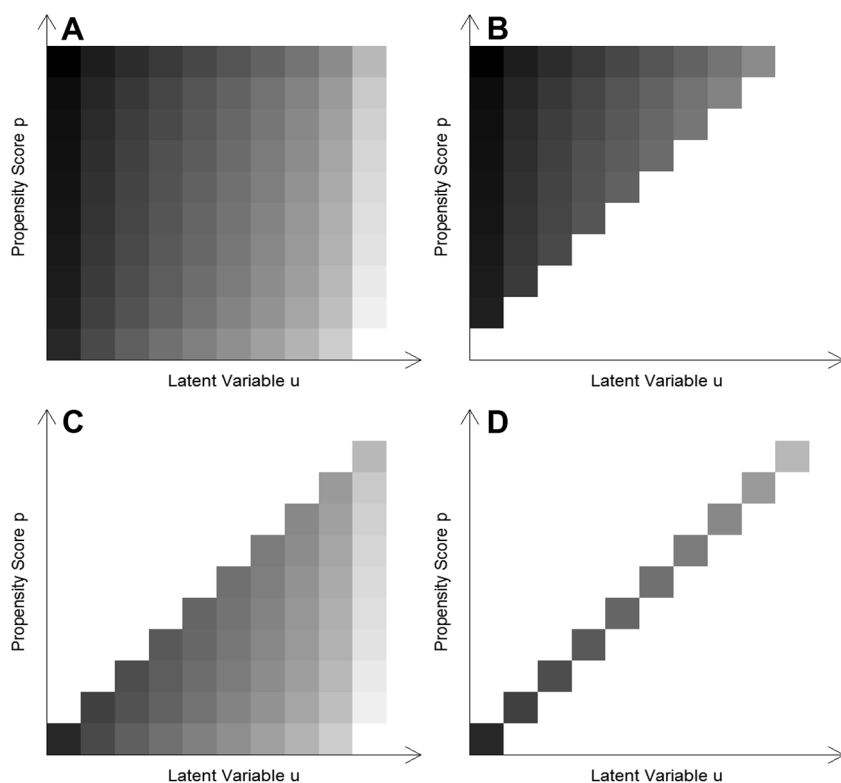


FIG. 1.—Illustration of treatment effect heterogeneity by propensity score $P(Z)$ and latent variable U . A darker color indicates a higher treatment effect.

To see how the two components combine to shape $MPRTE(p)$, let us revisit the classic example on the economic returns to college. In the labor economics literature, a negative association has often been found between η and U , suggesting a pattern of “positive selection”; that is, individuals who benefit more from college are more motivated than their peers to attend college (e.g., Willis and Rosen 1979; Blundell, Dearden, and Sianesi 2005; Moffitt 2008; Carneiro et al. 2011; Heckman, Humphries, and Veramendi 2016). In this case, the second component of equation (9) would be a decreasing function of p . On the other hand, the literature has not paid much attention to the first component, concerning whether individuals who by observed characteristics are more likely to attend college also benefit more from college. A number of observational studies have suggested that nontraditional students, such as racial and ethnic minorities or students from less educated families, experience higher returns to college than traditional students, although interpretation of such findings remains controversial because of potential unobserved selection

biases (e.g., Bowen and Bok 1998; Attewell and Lavin 2007; Maurin and McNally 2008; Brand and Xie 2010; Dale and Krueger 2011). However, if the downward slope in the second component were sufficiently strong, $MPRTE(p)$ would also decline with p . In this case, we would paradoxically observe a pattern of “negative selection”: among students who are at the margin of attending college, those who by observed characteristics are less likely to attend college would actually benefit more from college.

To better understand the paradoxical implication of self-selection, let us revisit figure 1. From figure 1A, we can see that in the hypothetical data, treatment effect declines with u at each level of the propensity score, suggesting an unobserved self-selection. In other words, individuals may have self-selected into treatment on the basis of their anticipated gains. On the other hand, at each level of the latent variable u , treatment effect increases with the propensity score, indicating that individuals who by observed characteristics are more likely to be treated also benefit more from the treatment. This relationship, however, is reversed among the marginal entrants. As shown in figure 1D, among the marginal entrants, those who appear less likely to be treated (bottom left grids) have higher treatment effects. This pattern of negative selection at the margin, interestingly, is exactly due to an unobserved positive selection into treatment.

VII. Policy as a Weighting Problem

In Section V, we used $\lambda(p)$ to denote the increment in treatment probability at each level of the propensity score p . Since $MPRTE(p)$ is defined as the pointwise limit of $PRTE(p, \alpha\lambda(p))$ as α approaches zero, the mathematical form of $\lambda(p)$ does not affect $MPRTE(p)$. However, in deriving the population-level (i.e., unconditional) $MPRTE$, we need to use $\lambda(p)$ as the appropriate weight. To see this, let us consider the overall $PRTE$ for a given α . Since the size of inducement $\alpha\lambda(p)$ reflects the share of individuals who are induced into treatment (“compliers”) given $P(Z) = p$, the overall $PRTE$ is a weighted average of $PRTE(p, \alpha\lambda(p))$ with weights $\alpha\lambda(p)$:

$$\begin{aligned}
 PRTE_\alpha &= \frac{\int_0^1 \alpha\lambda(p)PRTE(p, \alpha\lambda(p)) dF_p(p)}{\int_0^1 \alpha\lambda(p) dF_p(p)} \\
 &= \frac{\int_0^1 \lambda(p)PRTE(p, \alpha\lambda(p)) dF_p(p)}{\int_0^1 \lambda(p) dF_p(p)},
 \end{aligned}$$

where $F_p(\cdot)$ denotes the marginal distribution function of the propensity score. We then define the population-level $MPRTE$ as the limit of $PRTE_\alpha$

as α approaches zero. Under some regularity conditions,⁵ we can take the limit inside the integral:

$$\begin{aligned} \text{MPRTE} &= \lim_{\alpha \rightarrow 0} \text{PRTE}_{\alpha} \\ &= \frac{\int_0^1 \lambda(p) \lim_{\alpha \rightarrow 0} \text{PRTE}(p, \alpha\lambda(p)) dF_p(p)}{\int_0^1 \lambda(p) dF_p(p)} \\ &= \frac{\int_0^1 \lambda(p) \text{MPRTE}(p) dF_p(p)}{\int_0^1 \lambda(p) dF_p(p)}. \end{aligned}$$

Thus, given the estimates of $\text{MPRTE}(p)$, a policy maker may apply the above formula to design an expression for $\lambda(\cdot)$ to boost the population-level MPRTE. For example, if it were found that the marginal return to college declines with the propensity score p , a college expansion targeted at students with lower values of p (say, a means-tested financial aid program) would be more effective overall than a uniform expansion of college attendance in the population (Zhou and Xie, forthcoming).⁶

In practice, for a given policy $\lambda(p)$, we can evaluate the above integral directly from sample data, using

$$\text{MPRTE} \approx \frac{\sum_i \text{MPRTE}(\hat{p}_i) \lambda(\hat{p}_i)}{\sum_i \lambda(\hat{p}_i)},$$

where \hat{p}_i is the estimated propensity score for unit i in the sample. When the sample is not representative of the population by itself, sampling weights need to be incorporated in these summations.

VIII. Conclusion

Through a redefinition of MTE using the propensity score, we presented a new perspective to interpret and analyze heterogeneous treatment effects in the presence of unobserved selection. The redefined MTE treats observed and unobserved selection symmetrically and parsimoniously

⁵ A sufficient (but not necessary) condition is that $\widetilde{\text{MTE}}(p, u)$ is bounded over $[0, 1] \times [0, 1]$. By the mean value theorem, $\text{PRTE}(p, \alpha\lambda(p))$ can be written as $\widetilde{\text{MTE}}(p, p^*)$, where $p^* \in [p, p + \alpha\lambda(p)]$. Thus, $\text{PRTE}(p, \alpha\lambda(p))$ is also bounded. By the dominated convergence theorem, the limit can be taken inside the integral.

⁶ Admittedly, the discussion here provides no more than a theoretical guide to practice. In the real world, designing specific policy instruments to produce a target form of $\lambda(p)$ can be a challenging task.

summarizes all of the treatment effect heterogeneity that is relevant for selection bias. As with the original MTE, the redefined MTE can serve as a building block for evaluating aggregate causal effects. Yet the weights associated with the new MTE are simpler, more intuitive, and easier to compute. Finally, the new MTE immediately reveals treatment effect heterogeneity among individuals who are at the margin of treatment, thus enabling us to design more cost-effective policy interventions.

References

- Attewell, Paul, and David Lavin. 2007. *Passing the Torch: Does Higher Education for the Disadvantaged Pay Off across the Generations?* New York: Russell Sage Found.
- Björklund, Anders, and Robert Moffitt. 1987. "The Estimation of Wage Gains and Welfare Gains in Self-Selection." *Rev. Econ. and Statis.* 69 (1): 42–49.
- Blundell, Richard, Lorraine Dearden, and Barbara Sianesi. 2005. "Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey." *J. Royal Statis. Soc. A* 168 (3): 473–512.
- Bowen, William G., and Derek Bok. 1998. *The Shape of the River: Long-Term Consequences of Considering Race in College and University Admissions.* Princeton, NJ: Princeton Univ. Press.
- Brand, Jennie E., and Yu Xie. 2010. "Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education." *American Sociological Rev.* 75 (2): 273–302.
- Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. 2017. "Beyond LATE with a Discrete Instrument." *J.P.E.* 125 (4): 985–1039.
- Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil. 2010. "Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin." *Econometrica* 78 (1): 377–94.
- . 2011. "Estimating Marginal Returns to Education." *A.E.R.* 101 (773): 2754–81.
- Carneiro, Pedro, and Sokbae Lee. 2009. "Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality." *J. Econometrics* 149 (2): 191–208.
- Dale, Stacy, and Alan B. Krueger. 2011. "Estimating the Return to College Selectivity over the Career Using Administrative Earnings Data." Working Paper no. 17159 (June), NBER, Cambridge, MA.
- Heckman, James J. 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *J. Econ. Literature* 48 (2): 356–98.
- Heckman, James J., and V. Joseph Hotz. 1989. "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *J. American Statis. Assoc.* 84 (408): 862–74.
- Heckman, James J., John Eric Humphries, and Gregory Veramendi. 2016. "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking." *J.P.E.* 126 (S1): S197–S246.
- Heckman, James, and Salvador Navarro-Lozano. 2004. "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models." *Rev. Econ. and Statis.* 86 (1): 30–57.
- Heckman, James J., and Richard Robb. 1986. "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes." In *Drawing Inferences from Self-Selected Samples*, edited by Howard Wainer, 63–107. New York: Springer.

- Heckman, James J., Sergio Urzua, and Edward J. Vytlačil. 2006. "Understanding Instrumental Variables in Models with Essential Heterogeneity." *Rev. Econ. and Statis.* 88 (3): 389–432.
- Heckman, James J., and Edward J. Vytlačil. 1999. "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects." *Proc. Nat. Acad. Sci. USA* 96 (8): 4730–34.
- . 2001a. "Local Instrumental Variables." In *Nonlinear Statistical Modeling—Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, edited by Cheng Hsiao, Kimio Morimune, and James L. Powell, 1–46. Cambridge: Cambridge Univ. Press.
- . 2001b. "Policy-Relevant Treatment Effects." *A.E.R.* 91 (2): 107–11.
- . 2005. "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica* 73 (3): 669–738.
- . 2007a. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics*, vol. 6, edited by J. J. Heckman and E. E. Leamer, 4779–4874. Amsterdam: North-Holland.
- . 2007b. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments." In *Handbook of Econometrics*, vol. 6, edited by J. J. Heckman and E. E. Leamer, 4875–5143. Amsterdam: North-Holland.
- Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.
- Maestas, Nicole, Kathleen J. Mullen, and Alexander Strand. 2013. "Does Disability Insurance Receipt Discourage Work? Using Examiner Assignment to Estimate Causal Effects of SSDI Receipt." *A.E.R.* 103 (5): 1797–1829.
- Maurin, Eric, and Sandra McNally. 2008. "Vive la Révolution! Long-Term Educational Returns of 1968 to the Angry Students." *J. Labor Econ.* 26 (1): 1–33.
- Moffitt, Robert. 2008. "Estimating Marginal Treatment Effects in Heterogeneous Populations." *Ann. Econ. et Statis.* 2008 (91/92): 239–61.
- Mogstad, Magne, Andres Santos, and Alexander Torgovitsky. 2017. "Using Instrumental Variables for Inference about Policy Relevant Treatment Effects." Working Paper no. 23568, NBER, Cambridge, MA.
- Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.
- Roy, Andrew Donald. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Econ. Papers* 3 (2): 135–46.
- Vytlačil, Edward. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70 (1): 331–41.
- Willis, Robert J., and Sherwin Rosen. 1979. "Education and Self-Selection." *J.P.E.* 87, no. 5, pt. 2 (October): S7–S36.
- Zhou, Xiang. 2019. "local IV: Estimation of Marginal Treatment Effects Using Local Instrumental Variables." R package version 0.2.1. <http://cran.r-project.org/package=localIV>.
- Zhou, Xiang, and Yu Xie. Forthcoming. "Heterogeneous Treatment Effects in the Presence of Self-Selection: A Propensity Score Perspective." *Sociological Methodology*.