Labour Economics 41 (2016) 47–60

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

From LATE to MTE: Alternative methods for the evaluation of policy interventions

Thomas Cornelissen [a,*], Christian Dustmann [b,1], Anna Raute [c], Uta Schönberg [d]

[a] Department of Economics and Related Studies, University of York, Heslington, York YO10 5DD, United Kingdom
[b] Department of Economics, University College London and CReAM, 30 Gordon Street, London WC1H 0AX, United Kingdom
[c] Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany
[d] Department of Economics, University College London, CReAM and IAB, 30 Gordon Street, London WC1H 0AX, United Kingdom

ARTICLE INFO

Article history:
Received 13 May 2016
Received in revised form 13 June 2016
Accepted 13 June 2016
Available online 25 June 2016

JEL classification:
C26
I26

Keywords:
Marginal treatment effects
Instrumental variables
Heterogeneous effects

ABSTRACT

This paper provides an introduction into the estimation of marginal treatment effects (MTE). Compared to the existing surveys on the subject, our paper is less technical and speaks to the applied economist with a solid basic understanding of econometric techniques who would like to use MTE estimation. Our framework of analysis is a generalized Roy model based on the potential outcomes framework, within which we define different treatment effects of interest, and review the well-known case of IV estimation with a discrete instrument resulting in a local average treatment effect (LATE). Turning to IV estimation with a continuous instrument, we demonstrate that the 2SLS estimator may be viewed as a weighted average of LATEs and discuss MTE estimation as an alternative and more informative way of exploiting a continuous instrument. We clarify the assumptions underlying the MTE framework, its relation to the correlated random coefficients model, and illustrate how the MTE estimation is implemented in practice.

1. Introduction

Evaluating the causal effects of programs or policy interventions is a central task in empirical microeconomics. A common case is when the program under evaluation takes the form of a binary treatment, such as attending college or attending preschool. Responses to such treatments (and thus the treatment effect) will most likely differ across individuals. For example, more able individuals are likely to have lower costs of learning than low ability individuals and may therefore enjoy larger returns from college attendance. Children from disadvantaged backgrounds may benefit more from the exposure to a high quality child care program than children from advantaged backgrounds.

Even though treatment effects are likely to be heterogeneous, early standard econometric textbooks aimed at applied researchers did not pay much attention to heterogeneous treatment effects (see, e.g., the textbooks by Johnston, 1963; Maddala, 1992). Switching regression models, in which the effects of observed and unobserved characteristics are allowed to differ across states (where the state could be a treatment, and thus the treatment effect would depend on observed and unobserved characteristics), present early approaches of modeling treatment effect heterogeneity and date back to the 1970s (see Quandt, 1972; Heckman, 1976; Lee, 1979). Rubin (1974) defines heterogeneous causal effects at the individual level in terms of potential outcomes and discusses the average treatment effect (ATE) (or "average causal effect") as an interesting parameter in order learn about the "typical" causal effect in a population. The study of Heckman and Robb (1985) is an important early contribution in pointing out that the average treatment effect (ATE) and the average treatment effect on the treated (ATT) are two conceptually distinct parameters that ask different economic policy questions. They analyze a random coefficients treatment effects regression with observed and unobserved heterogeneity in rewards (which they show to be equivalent to the switching regression model with two states) and emphasize that different estimation methods will in general identify different parameters. However, despite these seminal early contributions, much of the applied work continued to assume homogeneous treatment effects, focusing mainly on addressing the problem of endogeneity caused by self-selection into treatment based on unobserved characteristics.

* Corresponding author.
E-mail addresses: thomas.cornelissen@york.ac.uk (T. Cornelissen), c.dustmann@ucl.ac.uk (C. Dustmann), raute@uni-mannheim.de (A. Raute), u.schoenberg@ucl.ac.uk (U. Schönberg).

[1] Christian Dustmann acknowledges funding from the European Research Council (ERC) under advanced grant no. 323992.

http://dx.doi.org/10.1016/j.labeco.2016.06.004
0927-5371/

The "LATE revolution" in the 1990s changed the focus to identification of models when treatment effects are heterogeneous.[2] The early papers in this literature by Imbens and Angrist (1994) and Angrist et al. (1996) raised awareness about potential heterogeneity in returns and clarified the interpretation of IV estimates when treatment effects are heterogeneous. Heckman and Vytlacil (1998), Card (2001), and others proposed a control function approach based on the correlated random coefficient model as an alternative to conventional linear IV estimation which, under stronger assumptions than IV estimation, allows estimation of the ATE and yields some insight into the pattern of selection in the unobservables. The concept of the marginal treatment effect MTE was first introduced by Björklund and Moffitt (1987) in the context of a multivariate-normal switching regression model, in which they defined the "marginal gain" as the gain from treatment for individuals who are shifted into (or out of) treatment by a marginal change in the cost of treatment (i.e., the instrument). It was extended in a series of papers by Heckman and Vytlacil (1999, 2001b, 2005, 2007) who define the MTE as the gain from treatment for individuals shifted into (or out of) treatment by a marginal change in the propensity score (i.e., the predicted probability of treatment, which is a function of the instrument), develop nonparametric estimation methods, and clarify the connection of the switching regime self-selection model and of MTE with IV and LATE.

Since then, the applied literature estimating MTEs has been growing and now includes, in addition to many applications in the economics of education, applications as varied as the effect of foster care on child outcomes (Doyle, 2007), the effect of Disability Insurance receipt on labor supply (Maestas et al., 2013; French and Song, 2014), and the interaction of quantity and quality of children (Brinch et al., forthcoming).[3] Some recent surveys provide insightful discussions about MTE, see for example Blundell and Costa Dias (2009), who discuss MTE among a range of alternative policy evaluation approaches, French and Taber (2011) who discuss treatment effects and MTE and its relation to the Roy model, and the excellent, comprehensive, but technical treatments of MTE in Heckman and Vytlacil (2007) and Heckman et al. (2006), based on the earlier work by Heckman and Vytlacil (1999, 2001a,b, 2005).

Drawing on these earlier papers, we provide here an introduction to the MTE framework, clarifying the discussion based on examples and developing it in a way that we believe is accessible to the applied economist. We commence by proposing a simple framework that allows for treatment effect heterogeneity and define within this framework different treatment effects of interest such as the average treatment effect (ATE) the average treatment effect on the treated (ATT), and the average treatment effect on the untreated (ATU). We next discuss the well-known local average treatment effect (LATE) identified by IV with a binary instrument, before reviewing IV estimation with continuous instruments. We carefully describe how conventional ways of exploiting continuous instruments identify one overall IV effect that can be difficult to interpret and can hide interesting patterns of treatment effect heterogeneity. Based on the example of the correlated random coefficients model, we then discuss the control function approach as an alternative to conventional linear IV estimation. We explain that, under considerably stronger assumptions than conventional IV estimation, the control function estimator of that model identifies a more general effect than IV and reveals some information on the pattern of selection based on unobserved gains. After that, we turn to MTE estimation as a more informative way of exploiting a continuous instrument, which aims at identifying a continuum of treatment effects along the distribution of the individual unobserved characteristic that drives treatment decisions and allows the identification of a variety of treatment parameters such as ATE, ATT, and ATU under potentially no stronger assumptions than IV estimation. We finally illustrate MTE estimation using two examples from the literature.

Our paper is less technical (and therefore also less rigorous) than the previous methodological contributions on MTE. It is written for the applied economist and introduces the method in a simple way, with a strong focus of relating MTE to more conventional IV estimation. The two applications we discuss illustrate to the applied researcher how MTE estimation can be implemented, and which additional insights hidden by IV estimation can be gained from MTE estimation. We also provide a set of lecture slides to accompany this article (available from the authors' personal websites).

2. Instrumental variable estimation with heterogeneous treatment effects

2.1. Framework of analysis and definition of treatment effects

Our general framework is a generalized Roy model based on the potential outcomes model and a latent variable discrete choice model for selection into treatment, as in Heckman and Vytlacil (1999) and most of the subsequent MTE literature.[4] We assume that treatment is a binary variable denoted by D_i. Let Y_{1i} be an individual's outcome under the hypothetical scenario that the individual is treated ($D_i = 1$) and Y_{0i} the outcome under the hypothetical scenario that the individual is not treated ($D_i = 0$). For example, Y_{1i} and Y_{0i} could be an individual's wage in the two hypothetical scenarios that the individual attends college and does not attend college, respectively. We model these potential outcomes as

$$Y_{0i} = \mu_0(X_i) + U_{0i} \tag{1}$$

$$Y_{1i} = \mu_1(X_i) + U_{1i} \tag{2}$$

where $\mu_j(X_i)$ is the conditional mean of Y_{ji} given X_i in treatment state j and U_{ji} captures deviations from that mean implying that $E[U_{ji}|X_i] = 0$.[5]

Consider the following latent variable discrete choice model for selection into treatment, which forms the basis for the MTE approach:

$$D_i^* = \mu_D(X_i, Z_i) - V_i \tag{3}$$

$$D_i = 1 \text{ if } D_i^* \geq 0, \qquad D_i = 0 \text{ otherwise}, \tag{4}$$

where D_i^* is the latent propensity to take the treatment. D_i^* is interpretable as the net gain from treatment (because individuals take the treatment if $D_i^* \geq 0$). The observed variables that affect the treatment decision include the same covariates X_i as the outcome Eqs. (1) and (2), and one or more variables Z_i excluded from the outcome equation. V_i is an i.i.d. error term indicating unobserved heterogeneity in the propensity for treatment. Because the error term V_i enters the selection equation with a negative sign, it embodies unobserved characteristics that make individuals *less* likely to receive treatment. One could thus label V_i unobserved "resistance" or "distaste" for treatment. The condition $D_i^* \geq 0$ of taking the treatment can be rewritten as $\mu_D(X_i, Z_i) \geq V_i$. If we

[2] In their 1994 Econometrica paper, Imbens and Angrist (1994) define the local average treatment effect (LATE) and spell out the assumptions under which IV identifies LATE. Angrist et al. (1996) coined the terms *compliers*, *always-takers*, *never-takers*, and *defiers*. However, the notion that in a world of heterogeneous treatment effects a binary IV identifies the average treatment effect for individuals who switch treatment status in response to changes in the instrument predates these papers. For example, it was already discussed in Angrist's (1990) paper using the Vietnam draft lottery as an IV for veteran status.

[3] Applications in economics of education range from estimating the effects of child care attendance on child performance (Felfe and Lalive, 2015; Noboa-Hidalgo and Urzúa, 2012, and Cornelissen et al., 2016, the effects of secondary schooling attendance on earnings (Carneiro et al., forthcoming), the effects of advanced high school mathematics education on earnings (Joensen and Nielsen, 2016), the effects of mixed-ability schools on long-term health (Basu et al., 2014), the effects of alternative breast cancer treatments on medical costs (Basu et al., 2007), and the returns to attending college (see e.g. Carneiro et al., 2011 for the U.S., Balfe, 2015 for the U.K., Kamhöfer et al., 2015, for Germany, and Nybom, 2014, for Sweden as well as Kaufmann, 2014, on the role of credit constraints in Mexico).

[4] The potential outcome model, often also referred to as the "Rubin causal model," is a building block for the literature on causal inference and goes back to Rubin (1974) and Holland (1986).

[5] The assumption of linear separability of Y_{ji} in $\mu_j(X_i)$ and U_{ji} is common in the applied MTE literature. It provides a simplification of the more general case $Y_{ji} = \mu_j(X_i, U_{ji})$ and makes computation of the aggregate treatment parameters (Eqs. (7)–(10)) and of the MTE weights (Section 4.3) more tractable.

apply the *c.d.f.* of V to this inequality, we get $F_V(\mu_D(X_i,Z_i)) \geq F_V(V_i)$. Both sides of this inequality are now bounded within the 0/1-interval. The left-hand side represents the propensity score, the probability of being treated based on the observed characteristics, and we refer to this term as $P(X_i,Z_i) \equiv F_V(\mu_D(X_i,Z_i))$. The right-hand side, $F_V(V_i)$, represents the quantiles of the distribution of the unobserved distaste for treatment V_i, which we denote by $U_{Di} \equiv F_V(V_i)$. The treatment decision can thus be rewritten as

$$D_i = 1 \text{ if } P(X_i,Z_i) \geq U_{Di}, \qquad D_i = 0 \text{ otherwise.} \tag{5}$$

Individuals take the treatment if the propensity score exceeds the quantile of the distribution of V_i at which the individual is located—that is, if the "encouragement" for treatment based on the observables X_i and Z_i exceeds the unobserved distaste for treatment.

It should be noted that the two potential outcomes Y_{0i} and Y_{1i} are never jointly observed for the same individual. Instead, we observe the realized outcome Y_i,which is equal to either Y_{0i} or Y_{1i} depending on treatment status:

$$Y_i = (1-D_i)Y_{0i} + D_iY_{1i} = Y_{0i} + D_i(Y_{1i}-Y_{0i})$$

This is in essence the switching regression model of Quandt (1972) and Lee (1979). Substituting in for Y_{0i} and Y_{1i} shows that the potential outcome framework can be represented as the regression model

$$Y_i = \mu_0(X_i) + D_i \underbrace{[\mu_1(X_i)-\mu_0(X_i) + U_{1i}-U_{0i}]}_{Y_{1i}-Y_{0i} \equiv \Delta_i} + U_{0i}, \tag{6}$$

in which the coefficient on the treatment dummy varies across individuals and is equal to

$$\Delta_i = Y_{1i}-Y_{0i} = \mu_1(X_i)-\mu_0(X_i) + U_{1i}-U_{0i}.$$

This treatment effect has two components: The average gain of someone with given observed characteristics, $\mu_1(X_i)-\mu_0(X_i)$, and an idiosyncratic individual-specific gain, $(U_{1i}-U_{0i})$.

There are good reasons to expect treatment effect heterogeneity. Consider the example of college education. First, individuals can be heterogeneous in their untreated outcome (Y_{0i}) reflecting differences in their experiences before entering college, such as the quality of their high school education, family background, etc. If the main effect of college attendance is to equalize preexisting differences and to bring everyone to the same level, then Y_{1i} would be more homogeneous than Y_{0i}, and individuals with lower outcomes in the untreated state would have higher treatment effects. Alternatively, it could be that some individuals are more able to benefit from college attendance (maybe because their ability to learn is higher) so that they would have a higher Y_{1i} even if Y_{0i} was similar to that of other individuals. A higher Y_{1i} for a given Y_{0i} could also result from variation in the quality of the treatment, for example, because colleges differ in the quality of their teaching and resources.

A main implication of heterogeneous effects is that summary treatment effects that aggregate over different parts of the population will in general be different from one another. Consider for example the average treatment effect (ATE), the average treatment effect on the treated (ATT), and the average treatment effect on the untreated (ATU).[6] Conditional on $X_i = x$, they are defined as

$$\text{ATE}(x) = E[\Delta_i|X_i = x] = \mu_1(x)-\mu_0(x)$$

$$\text{ATT}(x) = E[\Delta_i|X_i = x, D_i = 1] = \mu_1(x)-\mu_0(x) + E[U_{1i}-U_{0i}|X_i = x, D_i = 1]$$

$$\text{ATU}(x) = E[\Delta_i|X_i = x, D_i = 0] = \mu_1(x)-\mu_0(x) + E[U_{1i}-U_{0i}|X_i = x, D_i = 0]$$

Conditional on $X_i = x$, the ATE is the average treatment effect for an individual with given observed characteristics $X_i = x$, while the ATT is the average treatment effect in the subgroup of the population that participates in the treatment conditional on $X_i = x$. Similarly, the ATU is the average treatment effect in the subgroup of the population that does not participate in the treatment conditional on $X_i = x$. ATE(x) measures how individuals with observed characteristics $X_i = x$ would benefit on average from the treatment if everybody with these observed characteristics were participating in the treatment, or the expected effect if some individuals from the group of individuals with observed characteristics $X_i = x$ were randomly assigned to treatment. ATT(x) measures how those individuals with observed characteristics $X_i = x$ that are currently enrolled in the treatment benefit from it on average. ATU(x) on the other hand answers the question how those individuals with observed characteristics $X_i = x$ who are currently not enrolled would benefit on average from treatment if they participated.

By averaging these parameters over the appropriate distribution of X_i, they can also be defined unconditionally:

$$\text{ATE} = E[\Delta_i] = E[\mu_1(X_i)-\mu_0(X_i)] \tag{7}$$

$$\text{ATT} = E[\Delta_i|D_i = 1] = E[\mu_1(X_i)-\mu_0(X_i)|D_i = 1] + E[U_{1i}-U_{0i}|D_i = 1] \tag{8}$$

$$\text{ATU} = E[\Delta_i|D_i = 0] = E[\mu_1(X_i)-\mu_0(X_i)|D_i = 0] + E[U_{1i}-U_{0i}|D_i = 0] \tag{9}$$

In a linear specification for the conditional mean, that is, $\mu_j(X_i) = X_i\beta_j$, the terms $E[\mu_1(X_i)-\mu_0(X_i)]$, $E[\mu_1(X_i)-\mu_0(X_i)|D_i = 1]$, and $E[\mu_1(X_i)-\mu_0(X_i)|D_i = 0]$ would simplify to $E[X_i](\beta_1-\beta_0)$, $E[X_i|D_i = 1](\beta_1-\beta_0)$, and $E[X_i|D_i = 0](\beta_1-\beta_0)$, respectively.

Sometimes we would like to know the aggregate effect of a *specific policy change*. This is given by the policy-relevant treatment effect (PRTE), see Heckman and Vytlacil (2001a, 2005) and Carneiro et al. (2011). Consider a policy change that affects the propensity score $P(X_i,Z_i)$, but not potential outcomes (Y_{1i},Y_{0i}) or the unobservables of the selection process (V_i). Such a policy will not change the underlying distribution of treatment effects, or preferences for treatment, but by changing the propensity score, the policy will change who selects into treatment based on the selection Eq. (5). Suppose D_i is the treatment choice under the baseline policy, and $\tilde{D}_i$ is the treatment choice under the alternative policy. The PRTE conditional on $X_i = x$ is defined as (see Appendix A for details):

$$\begin{aligned}\text{PRTE}(x) &= \frac{E[Y_i|X_i = x,\ \text{alternative policy}]-E[Y_i|X_i = x,\ \text{baseline policy}]}{E[D_i|X_i = x,\ \text{alternative policy}]-E[D_i|X_i = x,\ \text{baseline policy}]} \\ &= \mu_1(x)-\mu_0(x) \\ &+ \frac{E[U_{1i}-U_{0i}|X_i = x, D \sim_i = 1]E[D \sim_i|X_i = x]-E[U_{1i}-U_{0i}|X_i = x, D_i = 1]E[D_i|X_i = x]}{E[D \sim_i|X_i = x]-E[D_i|X_i = x]}\end{aligned}$$

and the corresponding unconditional effect is

$$\begin{aligned}\text{PRTE} = &\frac{E[\mu_1(X_i)-\mu_0X_i|D \sim_i = 1]E[D \sim_i]-E[\mu_1(X_i)-\mu_0(X_i)|D_i = 1]E[D_i]}{E[D \sim_i]-E[D_i]} \\ + &\frac{E[U_{1i}-U_{0i}|D \sim_i = 1]E[D \sim_i]-E[U_{1i}-U_{0i}|D_i = 1]E[D_i]}{E[D \sim_i]-E[D_i]}\end{aligned} \tag{10}$$

The PRTE is the mean effect of going from a baseline policy to an alternative policy per net person shifted. It also corresponds to a weighted difference between the ATT under the alternative policy and the ATT under the baseline policy.[7]

[6] For an extension of the framework including additional parameters on the cost and the surplus of the treatment, see Eisenhauer et al. (2015).

[7] If a policy only shifts additional people into treatment without shifting anyone out of the treatment, the PRTE is the average effect on the subgroup of individuals shifted into treatment by the policy. In general, a policy may shift some individuals into treatment and some individuals out of treatment. In this case, the PRTE is a net effect in which those shifted out of treatment receive a negative weight. Nevertheless, it is still informative on the aggregate effect of the policy (Heckman and Vytlacil, 2005).

It is important to note that ATE, ATT, ATU, and PRTE would be the same if there was no selection into treatment based on gains—one might imagine that individuals simply do not know their idiosyncratic returns to treatment or simply do not act on them. In reality it however seems likely that, depending on the context, individuals do select into treatment either directly based on gains, or based on characteristics that are related to gains. In consequence, the treatment parameters would in general differ. In the case of college attendance, for example, we would expect individuals who expect higher gains (e.g., higher future wages) from college attendance to be more likely to attend college. Such positive selection on gains is likely to occur based on both observed and unobserved characteristics. Positive selection on "unobserved gains" implies that $U_{1i}-U_{0i}$ is positively related to D_i conditional on X_i, such that $E[U_{1i}-U_{0i}|X_i=x,D_i=1]>0$ and $E[U_{1i}-U_{0i}|X_i=x,D_i=0]<0$, and thus $\text{ATT}(x)>\text{ATE}(x)>\text{ATU}(x)$. Positive selection on "observed gains" implies that $\mu_1(X_i)-\mu_0(X_i)$ is positively related to D_i, and thus ATT > ATE > ATU (provided that $\text{ATT}(x)\geq\text{ATE}(x)\geq\text{ATU}(x)$).

When treatment effects are heterogeneous, it is of primary relevance to spell out which effect a given econometric method identifies. Next, we discuss which parameters linear instrumental variable estimation with a binary instrument and with a continuous instrument identify (Sections 2.2 and 2.3) and contrast these approaches with the control function estimator of the correlated random coefficient model (Section 2.4).

2.2. IV with a binary instrument and LATE

We first apply the IV estimator within subsamples stratified by $X_i=x$, leading to covariate-specific IV estimates, similar to the covariate-specific treatment effects defined in Section 2.1. We then derive one aggregate IV estimator representing an average across values of X_i.

2.2.1. Covariate-specific IV

Let Z_i be a binary instrumental variable. The IV estimator with binary instrument conditional on $X_i=x$ is equal to the Wald estimator

$$\text{Wald}(x)=\frac{E[Y_i|Z_i=1,X_i=x]-E[Y_i|Z_i=0,X_i=x]}{E[D_i|Z_i=1,X_i=x]-E[D_i|Z_i=0,X_i=x]}. \tag{11}$$

In the sample of individuals with $X_i=x$, this estimator divides the average difference in the outcome between individuals with the instrument switched on ($Z_i=1$) and individuals with the instrument switched off ($Z_i=0$) by the same difference in average treatment status. The numerator is also commonly referred to as the "reduced form" and the denominator as the "first stage."

The assumptions under which this ratio estimates a causal effect are well understood, and we state them only briefly here (see, e.g., Angrist and Pischke, 2009 for a detailed discussion). Let D_{0i} denote the potential treatment state of individual i if $Z_i=0$ and D_{1i} the potential treatment state of individual i if $Z_i=1$, so that observed treatment D_i is equal to [8]

$$D_i=Z_iD_{1i}+(1-Z_i)D_{0i}.$$

The following assumptions are required for a causal interpretation of (11):

(i) Independence: $\{Y_{1i},Y_{0i},D_{1i},D_{0i}\}\perp\!\!\!\perp Z_i|X_i$. This assumption first states that the instrument Z_i must be as good as randomly assigned conditional on X_i. Random assignment ensures that the reduced-form effect of Z_i on Y_i has a causal interpretation (conditional on X_i). The independence assumption further states that conditional on X_i, the instrument must affect potential outcomes only through its effect on the treatment probability D_i—which is commonly referred to as the exclusion restriction.[9] The exclusion restriction is necessary for the Wald estimator to identify the causal effect of treatment D_i on Y_i. It should be noted that the exclusion restriction would be violated if treatment effects $Y_{1i}-Y_{0i}$ depended on the instrument.[10] Within the generalized Roy model of Eqs. (1)–(4), the independence assumption may be alternatively written as $(U_0,U_1,V)\perp\!\!\!\perp Z|X$.

(ii) Existence of a first stage: $E[D_{1i}-D_{0i}|X_i]\neq 0$

(iii) Monotonicity (or uniformity): $D_{1i}\geq D_{0i}\,\forall\, i$ or $D_{1i}\leq D_{0i}\,\forall\, i$. This assumption means that all individuals who change their treatment status as a result of a change in the instrument either get all shifted into treatment or get all shifted out of treatment.[11] Here we assume that Z_i is coded in a way that $Z_i=1$ provides an extra encouragement for treatment compared to $Z_i=0$, implying that monotonicity holds in the form of $D_{1i}\geq D_{0i}\,\forall\, i$.

Under these assumptions, the IV estimator in Eq. (11) with a binary instrument applied in a subsample in which the covariates are fixed at $X_i=x$ identifies the covariate-specific local average treatment effect (LATE) defined by

$$\begin{aligned}\text{LATE}(x)&=E[Y_{1i}-Y_{0i}|D_{1i}>D_{0i},X_i=x]\\&=\mu_1(X_i)-\mu_0(X_i)+E[U_{1i}-U_{0i}|D_{1i}>D_{0i},X_i=x]\end{aligned} \tag{12}$$

The subpopulation for which $D_{1i}>D_{0i}$ holds true is called the group of compliers (Angrist et al., 1996). These are individuals whose potential treatment status changes in response to the extra encouragement for treatment as the instrument changes from 0 to 1. They are treated if the instrument is switched on ($D_{1i}=1$) and untreated if the instrument is switched off ($D_{0i}=0$). For example, if the instrument is a dummy variable for a college being located nearby an individual's place of residence, then the LATE is the treatment effect averaged over the group of individuals who attend college if living nearby a college, but who do not attend college if the college is far away. These might be people who are constrained in their resources to take up college far away from their place of residence, as argued by Card (2001), or who feel that their return from college would not warrant the cost of attending college in a faraway location. IV is not informative on the effect for the subgroup of always-takers (defined by $D_{1i}=D_{0i}=1$) and never-takers (defined by $D_{1i}=D_{0i}=0$), who decide in favour (or against) college attendance independently of the value of the instrument. In this example, always-takers could be individuals who estimate their returns as high enough in order to warrant college attendance even in a faraway location, and never-takers would not attend college even in a nearby location. The existence of defiers, defined by $D_{1i}<D_{0i}$, who attend college

[8] Note that potential outcomes are indexed against the treatment state, whereas the potential treatment decision is indexed against the value of the instrument.

[9] To make the distinction between random assignment and exclusion more explicit, Angrist and Pischke (2009) introduce the following notation. Let $Y_i(d,z,x)$ denote the potential outcome of an individual with treatment status $D_i=d$, instrument value $Z_i=z$, and covariate $X_i=x$. The random assignment assumption may then be written as $\{Y_i(D_{1i},1,x),Y_i(D_{0i},0,x),D_{1i},D_{0i}\}\perp\!\!\!\perp Z_i|X_i$, while the exclusion restriction may be written as $Y_i(d,0,x)=Y_i(d,1,x)$.

[10] To see this, consider the following simple example. Suppose Y_{0i} does not depend on the instrument, but treatment effects vary with the instrument such that $Y_{1i}-Y_{0i}=\Delta_1$ if $Z_i=1$ and $Y_{1i}-Y_{0i}=\Delta_0$ if $Z_i=0$. This violates the exclusion restriction. It follows that $E[Y_i|Z_i=1]=E[Y_{0i}]+\Delta_1E[D_i|Z_i=1]$ and $E[Y_i|Z_i=0]=E[Y_{0i}]+\Delta_0E[D_i|Z_i=0]$. Substituting this into the Wald estimator yields $\frac{\Delta_1E[D_i|Z_i=1]-\Delta_0E[D_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$. Because the treatment effect differs for the two values of the instrument, it cannot be factored out of the difference in the numerator and the result is a nonsensically weighted average of Δ_0 and Δ_1, giving positive weight $\frac{E[D_i|Z_i=1]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$ to Δ_1 and negative weight $\frac{-E[D_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$ to Δ_0. Similarly, when using group indicator dummies (say, regions, cohorts, region-year cells, etc.) as instruments, the exclusion restriction requires the treatment effects to be similar across groups (conditional on the control variables). Whether or not this is credible depends on any given application.

[11] The IV monotonicity assumption is an assumption of a unidirectional effect of Z_i on $E[D_i|Z_i]$ across individuals. It is therefore sometimes referred to as uniformity rather than monotonicity assumption (e.g., Heckman and Vytlacil, 2007).

in a faraway location but not in a nearby location is ruled out by the monotonicity assumption.

2.2.2. Aggregating covariate-specific LATEs into one IV effect

The covariate-specific LATEs can be aggregated into one IV effect by applying 2SLS with a fully saturated model in covariates in both the first and second stage and interactions between the instrument and the covariates in the first stage (the "saturate and weight" theorem by Angrist and Imbens, 1995). This produces a variance-weighted average of the covariate-specific LATEs and equals:

$$IV = \sum_{x \in \mathcal{X}} \omega(x) \text{LATE}(x)$$

where $\mathcal{X}$ is the set of all unique values of X_i, and $\omega(x)$ are weights that sum to one and are equal to the contribution of the observations with $X_i = x$ to the variance of the first-stage fitted values.[12] In practice, less saturated models seem to provide a good approximation to the underlying causal relation (see the discussion related to Theorem 4.5.1 in Angrist and Pischke, 2009).

There is an important difference between LATE defined in Eq. (12) and the other treatment parameters defined in the previous section. ATE, ATT, ATU, and PRTE are parameters that answer economic policy questions and are defined independently of any instrument. LATE, on the other hand, is defined by the instrumental variable used (because compliers are defined in relation to the instrument) and therefore does not necessarily answer an economic policy question and does not necessarily represent a treatment parameter for an economically interesting group of the population, criticisms made for example in Heckman (1997), Deaton (2009), and Heckman and Urzúa (2010).

There are, however, special cases in which LATE coincides with economically interesting parameters. The first case is when the instrument is a policy change in which case LATE is equivalent to PRTE defined in Eq. (10) and thus a policy-relevant parameter (Heckman et al., 1999). An example is the paper by Oreopoulos (2006) who uses an increase of the compulsory school-leaving age as a binary instrument. LATE thus captures the effect for individuals induced to stay in school longer by the policy reform and is a PRTE. Interestingly, the case analyzed by Oreopoulos (2006) is at the same time an example for a second special case. Because the increase in the school-leaving age was fully enforced, there were no never-takers. Consequently all untreated are compliers (with the instrument switched off) and in such a case LATE is equal to ATU. An example for the opposite case is a recent paper by Chetty et al. (2016), who evaluate the long–run effects of the Moving To Opportunity (MTO) experiment, which offered randomly selected families housing vouchers to move from high-poverty housing projects to lower-poverty neighborhoods. The random assignment to the treatment group (offer of a voucher) was used as an instrument for the actual treatment decision (in this case the decision to relocate to a lower-poverty neighborhood). Because nobody in the control group had access to the treatment, there were no always-takers, implying that all treated are compliers (with the instrument switched on) and LATE identifies ATT. [13]

[12] The weights are equal to $\omega(X_i) = \frac{p_x \text{Var}(\hat{D}_i | X_i = x)}{\text{Var}(\hat{D}_i)}$, where $\hat{D}_i = E[D_i | X_i, Z_i]$ denotes the first stage fitted value and p_x the population share of individuals with $X_i = x$. These are the same weights as equation 4.5.4 in Angrist and Pischke (2009) in somewhat different notation. It should be noted that conditional on X, all variation in $\hat{D}_i$ comes from the instrument(s) and that $\text{Var}(\hat{D}_i) = \text{Cov}(\hat{D}_i, D_i)$. Therefore, the weight $\omega(x)$ can also be interpreted as the contribution of observations with $X_i = x$ to the first-stage covariance and in that sense the weights are proportionate to how strongly individuals with $X_i = x$ are shifted by the instrument. This is, however, not in general equal to the share of compliers at $X_i = x$ relative to all compliers.

[13] Using treatment assignment as an instrument for actual treatment is common in randomized trials when there is not full compliance with the treatment assignment. Just as in the examples above, LATE identifies ATT (when some members of the treatment group do not take the treatment, but nobody in the control group has access to treatment) or ATU (when all members of the treatment group take the treatment, and some members of the control group gain access to the treatment). These two cases are called "one-sided non-compliance."

2.3. IV with a continuous instrument

2.3.1. Pairwise covariate-specific LATEs

If Z_i is a continuous instrument, then one can exploit any pair of values z and z' of Z_i as a binary instrument calculating the covariate-specific IV estimator

$$\text{Wald}(z, z', x) = \frac{E[Y_i | Z_i = z, X_i = x] - E[Y_i | Z_i = z\prime, X_i = x]}{E[D_i | Z_i = z, X_i = x] - E[D_i | Z_i = z\prime, X_i = x]}. \quad (13)$$

In order for each of these IV estimators to capture the average treatment effect for compliers with a change in the instrument from z to z', Z_i needs to fulfil the IV assumptions discussed in Section 2.2. In particular, the monotonicity (or uniformity) assumption needs to hold between all pairs of values z and z' of Z_i. Denoting by D_{zi} a binary indicator for the potential treatment status of individual i for instrument value $Z_i = z$, the monotonicity assumption requires that for any given pair of values z and z', either $D_{zi} \geq D_{z'i}$, $\forall i$, or $D_{zi} \leq D_{z'i}$, $\forall i$ (Imbens and Angrist, 1994). That is, all individuals whose treatment status is affected by a change of the instrument from z to z' have to either all be shifted into treatment, or all be shifted out of treatment. A treatment choice model that ensures monotonicity to hold between all pairs of values of Z_i is the simple latent index choice model with a linearly separable error term defined in Eqs. (3) and (4). Assuming that a move from z to z' shifts individuals *into* treatment ($E[D_i | Z_i = z, X_i = x] > E[D_i | Z_i = z', X_i = x]$), the associated LATE is[14]

$$\text{LATE}(z, z', x) = E[Y_{1i} - Y_{0i} | D_{zi} > D_{z'i}, X_i = x]. \quad (14)$$

In terms of the latent index choice model, the condition $D_{zi} > D_{z'i}$ (which characterizes compliers in the case where a move from z to z' increases the average treatment probability) is equivalent to $P(z') < U_D < P(z)$. That is, compliers are individuals with intermediate values of the "distaste" for treatment, such that they do not choose treatment when faced with a propensity score value of $P(z')$, but they choose treatment when faced with the higher value $P(z)$. The LATE exploiting pairs of values z and z' (for the case in which a change from z to z' increases average treatment probability) can thus also be written as

$$E[Y_{1i} - Y_{0i} | P(z\prime) < U_D < P(z), X_i = x] \quad (15)$$

Fig. 1, which is based on hypothetical data, helps to illustrate the group of compliers. Assuming a subsample with covariates fixed at $X_i = x$, the figure depicts a continuous instrument Z_i on the horizontal axis varying between 0 and 200. The vertical axis measures the treatment probability, and the solid line displays $E[D_i | Z_i, X_i = x]$, the treatment probability as a function of Z_i. For example, Z_i could be distance to college and D_i college attendance. A reduction of the instrument from $Z_i = 120$ to $Z_i = 90$ raises the probability of treatment from $P(120) = .5$ to $P(90) = .75$. This shifts individuals with $.5 < U_D < .75$ into treatment, which are individuals who are between the 50th and the 75th percentile of the distribution of V. The associated LATE would thus be the treatment effect for this subgroup.

In practice, the possibility of computing all pairwise LATEs with a continuous instrument is obviously limited, as the number of observations in a given sample for every z and z' pair is likely to be small. A useful way of exploiting a continuous instrument is therefore to partition it into discrete groups.[15] Consider partitioning the range of Z_i in Fig. 1 into

[14] Conversely, if a move from z to z' shifts compliers *out of* treatment ($E[D_i | Z_i = z, X_i = x] < E[D_i | Z_i = z', X_i = x]$), then the associated LATE is $E[Y_{1i} - Y_{0i} | D_{zi} < D_{z'i}, X_i = x]$. The only difference is that compliers are now defined by $D_{zi} < D_{z'i}$ instead of $D_{zi} > D_{z'i}$.

[15] It should be noted that simply using Z_i as a continuous instrument in a linear IV estimator $\frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}$ requires an additional type of monotonicity assumption (see condition 3 in Imbens and Angrist, 1994). This only produces a non-negatively weighted combination of LATEs if Z_i has a monotonic association with the treatment probability. One way to ensure this condition holds is to use the propensity score $P(Z)$ as an instrument.

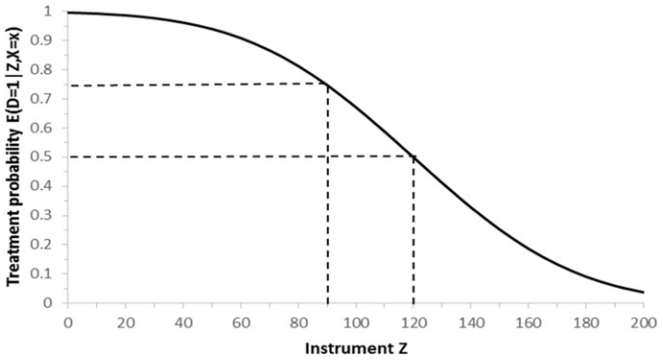

Fig. 1. Treatment probability as a function of a continuous instrument. *Notes:* Based on hypothetical data, the figure shows the effect of a continuous instrument Z on the probability of treatment in a sample with fixed covariates ($E[D = 1,Z,X = x]$). For example, the horizontal axis could represent distance to college and the vertical axis could represent the probability to attend college. *Data source:* Simulated hypothetical data.

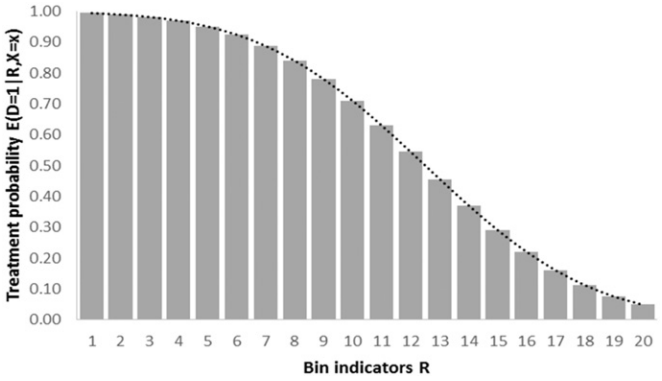

Fig. 2. Treatment probability in discrete bins of a continuous instrument. *Notes:* Based on hypothetical data, the bins in this figure show the probability of treatment in a sample with fixed covariates ($E[D = 1,R,X = x]$) as a function of a discrete variable R, which has been generated by grouping the values of the continuous instrument depicted in Fig. 1 into 20 equally spaced bins. The dotted line reproduces the function depicted in Fig. 1. *Data source:* Simulated hypothetical data.

equally sized bins identified by a bin identifier or grouping variable R_i, which is a function of Z_i and assumes the integer values of 1 to 20 to indicate in which bin a given value of Z_i is situated. This is illustrated in Fig. 2, where the horizontal axis is partitioned into 20 bins, and the bin height indicates the average treatment probability in each bin, $E[D_i|R_i,X_i=x]$. From any pair of two points $R_i=r$ and $R_i=r'$, and with corresponding data on the average outcome by bin, conditional on X_i, a Wald estimator of the form $\frac{E[Y_i|R_i=r,X_i=x]-E[Y_i|R_i=r\prime,X_i=x]}{E[D_i|R_i=r,X_i=x]-E[D_i|R_i=r\prime,X_i=x]}$ can be constructed, each of which identifies LATE(r,r',x), a covariate-specific LATE for compliers with a move of the discretized instrument from r to r'.

2.3.2. Aggregating pairwise (covariate-specific) LATEs into one effect

An efficient way of obtaining an overall IV estimate that aggregates the covariate-specific Wald estimates LATE(r,r',x) across $r-r'$ pairs and across x into one overall effect is provided by 2SLS, using group indicator dummies for the values of R_i as instruments, fully saturating the first and second stage in the covariates, and interacting the instruments in the first stage with the covariates. As discussed in Section 2.2.2, this provides a variance-weighted average of covariate-specific LATEs. To further see how 2SLS using group indicator dummies aggregates the pairwise LATEs across r-–r' pairs, it is useful to abstract from covariates by assuming again a subsample with covariates fixed at $X_i=x$. Fig. 3 based on simulated data, which plots $E[Y_i|R_i,X_i=x]$ against $E[D_i|R_i,X_i=x]$, helps to illustrate how the various Wald estimators are aggregated. The 2SLS estimator can be thought of as fitting a straight line through the points in Fig. 3 using generalized least squares (GLS) estimation because grouped data have a known heteroscedasticity structure (Angrist, 1991). The resulting weights that each covariate-specific LATE receives are positive and sum to one. The weights are positively related to the strength of the first-stage $E[D_i|R_i=r,X_i=x]-E[D_i|R_i=r',X_i=x]$ and to group size (i.e., number of observation in each bin).[16]

Whereas it is fairly straightforward to describe for whom LATE with a single binary instrument is representative (the group of compliers with that instrument), this is no longer the case with a continuous instrument—since the overall IV effect is now representative for compliers with changes between *all* values of the instrument, with different weights attached to groups of compliers at different pairs of values. An aggregate IV estimate may also hide interesting information, such as which pairs of values of the instrument shift a particularly large group of individuals, or a group of individuals with particularly large treatment effects, into treatment.

2.4. Control function approach: the correlated random coefficients model

An alternative to conventional linear IV estimation is to use the instrument to construct a control function, and to include this into the regression alongside the endogenous variable (see Wooldridge (2015) for an overview of control function methods). A well-known model for which a control function estimator has been proposed is the correlated random coefficients model (Card, 2001; Heckman and Vytlacil, 1998; Heckman and Robb, 1985). As we explain below, the control function estimator for this model allows estimation of the ATE and yields some insight into the pattern of selection in the unobservables, albeit under stronger assumption than IV estimation. Consider the outcome of Eq. (6) in which we assume linearity in the regressors, $\mu_0(X_i)=X_i\beta_0$ and $\mu_1(X_i)=X_i\beta_1$, and for a more compact notation rewrite the equation as

$$Y_i = X_i\alpha + D_i\tilde{X}_i\theta + D_i\delta_i + \varepsilon_i, \tag{16}$$

with $\alpha=\beta_0$, $\theta=\beta_1-\beta_0$, $\delta_i=U_{1i}-U_{0i}$, $\varepsilon_i=U_{0i}$, and where $\tilde{X}_i = X_i-\overline{X}$ denotes the covariates centered around their sample means. This is a

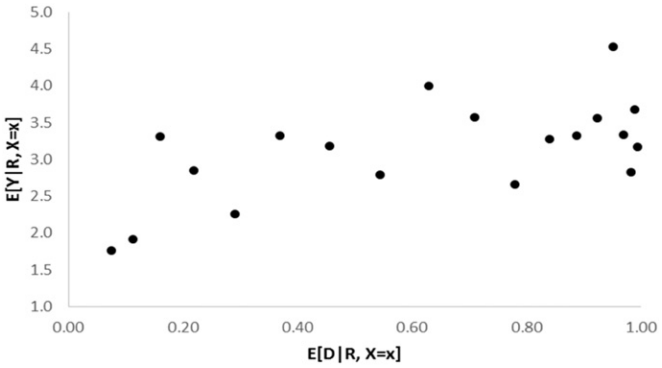

Fig. 3. Grouped data IV. *Notes:* Based on hypothetical data, the figure plots the average outcome against the average treatment probability in a sample with fixed covariates for 20 groups, which are equal to the bins depicted in Fig. 2 and correspond to 20 equally sized bins of an underlying continuous instrument. Grouped data IV can be visualized as fitting a line through these points. *Data source:* Simulated hypothetical data.

[16] A slope estimated by ordinary least squares is equal to a weighted average of all possible combinations of pairwise slopes between any two points, with a larger weight on slopes between points that are further apart on the horizontal axis. This is because $\hat{\beta}^{OLS} = \frac{cov(x,y)}{var(x)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i-y_j)(x_i-x_j)}{\sum_{j=1}^{n}(x_i-x_j)^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(y_i-y_j)}{(x_i-x_j)}(x_i-x_j)^2}{\sum_{j=1}^{n}(x_i-x_j)^2}$. In Fig. 3, the distance between two points on the horizontal axis is exactly equal to the first stage $E[D_i|R_i=r]-E[D_i|R_i=r']$ of the associated LATE, therefore LATEs with a stronger first stage get a higher weight. If in addition the slope is estimated by GLS, then LATEs associated with larger groups receive a higher weight, because GLS weights observations inversely to their variance, and the variance of groups means decreases in group size.

random coefficient model, in which the coefficient δ_i varies across individuals. Decomposing $\delta_i = \overline{\delta} + \tilde{\delta}_i$ into its mean $\overline{\delta} = E[\delta_i]$ and the deviation from the mean $\tilde{\delta}_i = \delta_i - E[\delta_i]$, Eq. (16) can be transformed into a constant coefficient model

$$Y_i = X_i\alpha + D_i\tilde{X}_i\theta + D_i\overline{\delta} + e_i. \tag{17}$$

Here, the coefficient on D_i is defined as the ATE. Because the covariates interacted with D_i are centered around their mean, $\overline{\delta}$ captures the ATE at means of X_i, which in this linear specification is also equal to the unconditional ATE. Deviations from the ATE enter the error term $e_i = D_i\tilde{\delta}_i + \varepsilon_i$. If there is selection based on gains, then $\tilde{\delta}_i$ and D_i are positively correlated, resulting in $E[D_i\tilde{\delta}_i|D_i = 1] > E[D_i\tilde{\delta}_i|D_i = 0]$, and hence Eq. (17) is referred to as the correlated random coefficients model. Any instrument Z_i that affects D_i will in this case also be correlated with the augmented error term e_i. IV estimation of Eq. (17) will therefore yield a biased estimate of $\overline{\delta}$ (the ATE). This is not surprising because, as explained above, when treatment effects are heterogeneous IV estimation does not in general identify the ATE.

In addition to the standard assumptions of independence and existence of a first stage, assume that D_i can be explained by the reduced-form equation

$$D_i = X_i\pi_1 + Z_i\pi_2 + \nu_i, \quad \text{with } E[\nu_i|X_i, Z_i] = 0, \tag{18}$$

and that both of the unobservables in e_i that cause selection bias in Eq. (17) are linearly related to the reduced-form error ν_i:

$$E[\varepsilon_i|\nu_i] = \eta\nu_i \tag{19}$$

$$E\left[\tilde{\delta}_i|\nu_i\right] = \psi\nu_i \tag{20}$$

Eq. (19) describes conventional selection bias. Because $\varepsilon_i = U_{0i}$, the relation between ε_i and ν_i states that individuals who are more likely due to unobserved characteristics to take the treatment differ in their pre-treatment characteristics from individuals who are less likely to take the treatment. Eq. (20) describes the process of selection based on gains and embodies the (rather strong) assumption that the unobserved part of the treatment effect depends linearly on the unobservables that affect the treatment.

As shown in Card (2001), under these assumptions, Eq. (17) can be estimated by OLS including $\hat{\nu}_i$ and $\hat{\nu}_i D_i$ as two additional regressors (control functions), where $\hat{\nu}_i$ is obtained as the predicted residual from Eq. (18) estimated by OLS.[17] The estimate of $\overline{\delta}$ is consistent for the ATE, and the sign of the coefficient on the control function $\hat{\nu}_i D_i$ is informative on the selection pattern (a positive sign implying selection based on gains). This control function approach, which can be implemented with either a binary or a continuous IV, thus yields parameters that are usually not identified by conventional IV. However, it relies on stronger assumptions than those needed for IV estimation, which does not require the assumptions in Eqs. (18)–(20). Moreover, while it estimates ATE, it does not recover other treatment parameters, such as the ATT, ATU, or PRTE. Next, we introduce the concept of marginal treatment effects (MTE) as a more informative way of exploiting a continuous instrument, which uncovers treatment effect heterogeneity more widely than the control function estimator and allows the identification of a variety of treatment parameters under potentially weaker assumptions.

3. Definition of the marginal treatment effect (MTE) and its relation to LATE

3.1. Definition

While LATE aggregates treatment effects over a certain range of the U_D distribution – see Eq. (15) – MTE is defined as the treatment effect at a *particular* value of U_D:

$$\text{MTE}(X_i = x, U_{Di} = u_D) \;=\; E(Y_{1i} - Y_{0i}|X_i = x, U_{Di} = u_D) \tag{21}$$

It is thus the treatment effect for an individual with observed characteristics $X = x$ who are at the u_D-th quantile of the V distribution, implying these individuals are indifferent to receiving treatment when having a propensity score $P(X_i, Z_i)$ equal to u_D.

To better understand what MTEs are, abstract from covariates by assuming that we exploit a subsample with covariates fixed at $X_i = x$. The MTE for $U_{Di} = P(z)$ is the limit of LATE in Eq. (15) for $P(z') \to P(z)$. The MTE at $U_{Di} = P(z)$ is thus, roughly, the LATE identified from a small departure of the propensity score from value $P(z)$ induced by the instrument.[18]

In formal notation, and as shown for example in Heckman et al. (2006) and Carneiro et al. (2011), the MTE is identified by the derivative of the outcome with respect to the propensity score:

$$\text{MTE}(X_i = x, U_{Di} = p) = \frac{\partial \text{E}(Y_i|X_i = x, P(Z_i) = p)}{\partial p} \tag{22}$$

Given that the Wald estimator in Eq. (13) is also a type of derivative of the outcome with respect to the treatment probability (it divides the instrument induced change in the outcome by the instrument induced change in the treatment), it may not be surprising that the MTE is identified by the derivative of the outcome with respect to the propensity score. In the following we provide some additional intuition why the derivative of the outcome with respect to the "*observed* inducement into treatment" (the propensity score) yields the treatment effect for individuals at a given point in the distribution of the *unobserved* resistance to treatment (U_D). At a given propensity score $p = p_0$, individuals with $U_D < p_0$ are treated, while individuals with $U_D = p_0$ are indifferent. Increasing p from p_0 by a small amount dp shifts previously indifferent individuals into treatment, who thus have a marginal treatment effect of $\text{MTE}(U_D = p_0)$. The associated increase in Y equals the share of shifted individuals times their treatment effect: $dY = dp^* \text{ MTE}(U_D = p_0)$. Dividing the change in Y by the change in p (which is, roughly speaking, what a derivative does) thus gives the MTE: $dY/dp = \text{MTE}(U_D = p_0)$. Therefore, the derivative of the outcome with respect to the propensity score yields the MTE at $U_D = p$.

Fig. 3 helps to interpret MTEs in an alternative way. Whereas 2SLS based on the discretized instrument fits a straight line through the grouped values in Fig. 3 (the slope of which is the aggregate IV effect), MTE can be thought of as using very fine "bins" (all available values of the propensity score) and allowing the slope of the curve to differ across values of $P(Z)$. The local slope in a point $P(Z) = P(z)$ then gives the MTE at $U_D = P(z)$.

3.2. Relation to LATE and the importance of a continuous instrument

Identifying the MTE across the full range of U_D between 0 and 1 requires a continuous instrument (at least if one wants to identify the MTE under minimal assumptions, as we discuss in Section 4.2 below). The following example illustrates this. Suppose that treatment is college

[17] Because of the two-step approach, standard errors need to be adjusted or bootstrapped (Wooldridge, 2015). The approach can be modified by explicitly accounting for the binary nature of the endogenous variable and replacing $\hat{\nu}_i$ by a generalized residual based on the inverse Mills ratio from a probit first stage regression (Wooldridge, 2015).

[18] The effect of a marginal change of the instrument as an interesting policy parameter was first introduced as the "marginal gain" in Björklund and Moffitt (1987). It was first defined as a limit form of LATE by Heckman (1997), and its relevance for policy evaluation is emphasized in Heckman and Smith (1998).

attendance, and that individuals continuously differ with respect to their unobserved resistance to college enrolment, U_D. The instrument is distance to college and assume that it varies from living directly next to a college to living very far from a college. Suppose that, as depicted in Fig. 1, when living right next to a college (distance of zero), all individuals attend college, even those with the highest resistance (conditional on X). By contrast, when living far away from a college, only individuals with the lowest resistance attend college (conditional on X_i). Gradually decreasing the distance from living maximally away until living right next to a college will then gradually shift all types into college, starting from the low-U_D types, gradually up to the high-U_D types. Thus, everybody is a complier at some value of the continuous instrument. The wage gains associated with increases in the propensity score that result from the gradual shift in the instrument are informative on the treatment effects of each of the shifted types, and thus the marginal wage increase at a given point (the derivative with respect to p) identifies the MTE for each type.

Compare this continuous instrument with a binary instrument, say an indicator DIST for whether a college is more than 50 miles away (DIST = 1) versus being less than 50 miles away (DIST = 0). Suppose that conditional on $X = x$ the probability of attending college is $P(\text{DIST} = 0) = 0.95$ if it is less than 50 miles away and $P(\text{DIST} = 1) = 0.5$ if it is more than 50 miles away. This instrument shifts types with U_D between 0.5 and 0.95 into treatment (individuals between the 50th and 95th quantile of the distribution of the unobserved resistance to treatment). The associated LATE identifies thus the *average* over the MTE curve between $U_D = 0.5$ and $U_D = 0.95$.

MTE is therefore defined as a continuum of treatment effects along the full distribution of U_D (the individual unobserved characteristic that drives treatment decisions). This has several advantages. First, rather than identifying one aggregate parameter that can mask important heterogeneity in treatment effects, the researcher is able to identify the whole (or at least a substantial part of the) range of individual treatment effects and thus characterize the extent of effect heterogeneity. Second, the MTE can be aggregated into economically interesting treatment effects such as the ATE, ATT, and PRTE, as we show in Section 4.3. Third, by relating the treatment effects to the decision of taking up the treatment measured by the participation probability, the researcher can infer the pattern of selection into treatment in a general manner along the entire unobserved resistance distribution. Estimation of the MTE is therefore more informative than both the conventional IV estimator and the control function estimator of the correlated random coefficients model discussed in Sections 2.2 to 2.4. In the ideal case, in which the instrument varies strongly conditional on X (see Section 4.2), it requires assumptions that are no stronger than the assumptions for conventional IV estimation.

To represent the heterogeneity in gains from treatment based on unobserved characteristics, and how it relates to the unobserved propensity to take up the treatment, one usually plots the MTE on the vertical axis of a graph against U_D on the horizontal axis, with X fixed at given values (say, at means). One important aspect in interpreting an MTE curve is its slope, as this reveals the selection pattern in unobserved characteristics. Recall that U_D are the quantiles of the unobserved resistance for treatment. An MTE curve that falls in U_D would suggest that low-resistance types (who are more likely due to unobserved reasons to participate in the treatment) have a higher treatment effect, and high-resistance types have a lower treatment effect. A falling MTE curve would thus indicate positive selection in unobserved characteristics based on gains—the pattern we typically expect. A rising MTE curve, by contrast, indicates reverse selection on gains in unobserved characteristics, while a flat MTE indicates no selection based on unobserved gains. In general, a non-monotonic shape of the MTE curve is also possible, which would imply a changing pattern of selection across the distribution of U_D. We provide examples of both a falling and a rising MTE curve in Section 5.

U_0, U_1, and V being residuals, their interpretation depends on the observables that are included in the regression. Changes in the variables included in (X,Z) redefine the residuals and thus potentially change the MTE curve. Note however that if Z contains several instruments, then using them one at a time (conditioning on the respective other ones) identifies the same MTE curve (although it could identify different stretches of the MTE curve depending on the range of variation that the different instruments cause in the propensity score).

The analysis of the selection pattern in unobserved characteristics can be complemented by checking for selection on gains (or otherwise) in observed characteristics, simply by checking whether those characteristics that lead to a high $\mu_1(X_i) - \mu_0(X_i)$ in the outcome equations lead to a high $\mu_D(X_i,Z_i)$ in the selection equation (or otherwise).

Next, we discuss the estimation of MTEs, starting with the fully parametric normal model, which is the framework in which MTE was first introduced by Björklund and Moffitt (1987) and which relies on strong distributional assumptions.

4. Estimation of MTE

4.1. The fully parametric normal model

The parametric normal model assumes a joint normal distribution of the error terms U_0, U_1 and V of the outcome and selection equations, $(U_0, U_1, \text{V}) \sim N(0, \Sigma)$, with variance–covariance matrix Σ in which the variance of V is normalized to 1. Moreover, suppose that potential outcomes and the selection equation are based on linear indices, that is $Y_{ji} = X_i\beta_j + U_{ji}$ for j = (0,1), and $D_i^* = (X_i, Z_i)\beta_d - V_i$ (and X_i includes a constant). These assumptions lead to a switching regime normal selection model or Heckman selection model (Heckman, 1976). Eqs. (1)–(4) can be estimated either jointly by maximum likelihood or following a two-step control function procedure. The two-step procedure exploits the fact that the confounding endogenous variation in the error terms of the outcome equations is given by

$$E[U_{0i}|D_i = 0, X_i, Z_i] = E[U_{0i}|V_i \geq (X_i, Z_i)\beta_d, X_i, Z_i] = \rho_0\left(\frac{\phi((X_i, Z_i)\beta_d)}{1-\Phi((X_i, Z_i)\beta_d)}\right), \tag{23}$$

$$E[U_{1i}|D_i = 1, X_i, Z_i] = E[U_{1i}|V_i < (X_i, Z_i)\beta_d, X_i, Z_i] = \rho_1\left(\frac{-\phi((X_i, Z_i)\beta_d)}{\Phi((X_i, Z_i)\beta_d)}\right), \tag{24}$$

where ϕ and Φ denote the p.d.f and c.d.f. of the standard normal distribution, and ρ_0 and ρ_1 are the correlation coefficients between U_{0i} and V_i and U_{1i} and V_i, respectively. Based on an estimate for β_d from a first-stage probit estimation of the selection equation, one can construct estimates of the ratios in parentheses in Eqs. (23) and (24). With these terms added as control functions, the outcome Eqs. (1) and (2) can be estimated by OLS. The ATE conditional on X is then given by $X_i(\hat{\beta}_1 - \hat{\beta}_0)$. The coefficients on the correction terms provide estimates for the correlations ρ_0 and ρ_1. In the normal selection model, the MTE has a parametric representation that follows directly from the joint normal distribution:[19]

$$\text{MTE}(x, u_D) = E(Y_{1i} - Y_{0i}|X_i = x, \text{U}_{\text{D}i} = u_D) = x(\beta_1 - \beta_0) + (\rho_1 - \rho_0)\Phi^{-1}(u_D)$$

Not only is joint normality of $(U_{0i}, U_{1i}, \text{V}_i)$ a strong assumption, it also puts strong restrictions on the shape of the MTE curve, which is simply equal to Φ^{-1}, the inverse of the standard normal c.d.f.,

[19] The joint normal distribution has the property that $EU_{1i}|\text{V}_i = \text{v} = \mu_{U_{1i}} + \frac{\rho_1}{\sigma_V^2}\text{v} - \mu_V$. Given that in this model $\mu_{U_{1i}} = \mu_V = 0$, $\sigma_V^2 = 1$, and $\text{v} = \Phi^{-1}(u_D)$, it follows that $E(U_{1i}|\text{U}_{\text{D}i} = u_D) = \rho_1\Phi^{-1}(u_D)$.

multiplied by a constant $(\rho_1-\rho_0)$, ruling out non-monotonic shapes of the MTE curve. If $\rho_1=\rho_0$, there is no selection based on unobserved gains. If $\rho_1-\rho_0<0$, there is positive selection based on gains, and if $\rho_1-\rho_0>0$, there is reverse selection on gains.

While Björklund and Moffitt (1987) first pointed out that the "marginal gain" is a relevant parameter which can be derived from the switching regime Heckman normal selection model, the subsequent literature has further clarified the definition and interpretation of the MTE and, crucially, has shown how it can be derived under much weaker assumptions (essentially under the same assumptions as conventional IV estimation). We now first describe the ideal case under which the MTE can be estimated nonparametrically under minimal assumptions (which puts high demands on the data), and then the more realistic case of semiparametric or parametric assumptions typically followed in practice (which are usually still weaker than those of the normal selection model).

4.2. Minimal assumptions and nonparametric estimation (the ideal case)

In addition to the assumptions required for a causal interpretation of the IV estimator discussed in Section 2.2, the estimation of MTE requires in the ideal case a continuous instrument Z that has enough variation to generate a propensity score $P(Z)$ with full common support (i.e., that has support in the full unit interval for both treated and untreated individuals) conditional on $X_i=x$. It should be noted that the "conditional on $X_i=x$" means within all unique combinations of the values of the X's—a much stronger requirement than the mere existence of a first stage. Suppose that X contains two dummy variables (say, gender and race), then Z should have strong variation within each of the four cells defined by all possible combinations of the values for gender and race. Obviously, the more regressors are included in X and the more values each regressor assumes, the stronger is this requirement.

The conventional estimation method to identify the MTE is the method of local instrumental variables (LIV; see Heckman and Vytlacil, 1999, 2001b, 2005), which estimates the MTE as the derivative of the outcome equation with respect to the propensity score, where the outcome has been modeled as a flexible function of the propensity score, thus exploiting the representation of the MTE given in Eq. (22).[20]

If a continuous instrument with a large range of variation within cells of $X_i=x$ is available, then the analysis can proceed in subsamples defined by the values of $X_i=x$, thus conditioning perfectly and nonparametrically on X, and identifying a separate MTE curve for each value of $X_i=x$. It should be noted that this allows identifying the MTE in a model with outcome equations of the form $Y_j=\mu_j(X_i,U_{ji})$. This "ideal" estimation approach thus does not rely on the linear separability assumptions embodied in Eqs. (1) and (2). Below we provide a sketch of this estimation method:

a. Split up the sample into the cells defined by $X_i=x$ and repeat the following steps separately within each of the subsamples.
b. Within each sample, estimate the probability of being treated (the propensity score) $P(Z)$ as a function of the excluded instrument(s) Z. Ideally, this might be done nonparametrically. Denote the predicted propensity score by $\hat{p}$.
c. Within each sample, model the outcome Y nonparametrically as a flexible function of $\hat{p}$ (for example by local polynomial regression). Denote the predicted outcome from this flexible function as $\hat{Y}$.
d. Within each sample, obtain MTE $(X_i=x,U_{Di}=p_0)$ as the derivative of $\hat{Y}$ with respect to $\hat{p}$, evaluated at point p_0. Doing this for a grid of values for p_0 from 0 to 1 allows tracing out the MTE curve for the full unit interval.[21]

4.3. Strengthening assumptions for estimation in less ideal cases

The approach outlined in the previous section assumes the availability of an ideal continuous instrument with sufficient variation conditional on $X_i=x$ to generate a propensity score $P(Z)$ with full common support conditional on $X_i=x$. This is rarely available, and additional assumptions need to be made. A first assumption is to not condition on X fully nonparametrically, but in a parametric linear way and model potential outcomes as $Y_{0i}=X_i\beta_0+U_{0i}$ and $Y_{1i}=X_i\beta_1+U_{1i}$ and the selection equation as $D_i^*=(X_i,Z_i)\beta_d-V_i$.

A second assumption restricts the shape of the MTE curve to be independent of X (common across all values of X), except for the intercept of the MTE curve, which is allowed to vary with X. Independence of the shape of the MTE curve across X is implied by the full independence assumption $(X_i,Z_i)\perp\!\!\!\perp(U_{0i},U_{1i},V_i)$, which is stronger than the conditional independence assumption $Z_i\perp\!\!\!\perp(U_{0i},U_{1i},V_i)\mid X_i$ necessary for a causal interpretation of IV and the estimation of MTE in the ideal case. Full independence implies not only that X is exogenous but also that the way in which U_1 and U_0 depend on V, and therefore the shape of the MTE curve, does not depend on X.[22] Alternatively, rather than invoking full independence, one can, in addition to the conditional independence assumption, assume additive separability between an observed and an unobserved component in the expected potential outcomes conditional on $U_D=u_D$ (Brinch et al., forthcoming):

$$E(Y_j|X_i=x,\mathrm{U_{Di}}=u_D)=X_i\beta_j+E(U_{ji}|\mathrm{U_{Di}}),\qquad j=0,1$$

Both the full independence and the linear separability assumption imply that the marginal treatment effect defined in Eq. (21) is additively separable into an observed and an unobserved component:[23]

$$\begin{aligned}\mathrm{MTE}(x,u_D)&=E(Y_{1i}-Y_{0i}|X_i=x,\mathrm{U_{Di}}=u_D)\\&=\underbrace{x(\beta_1-\beta_0)}_{\text{observed component}}+\underbrace{E(U_{1i}-U_{0i}|\mathrm{U_{Di}}=u_D)}_{\text{unobserved component}}.\end{aligned}\tag{25}$$

Exploiting linearity of the outcome in X and a constant shape of the MTE across X (except for a varying intercept) leads to the following outcome equation:

$$E[Y_i|X_i=x,P(Z)=p]=X_i\beta_0+X_i(\beta_1-\beta_0)p+K(p),\tag{26}$$

where $K(p)$ is a nonlinear function of the propensity score. The coefficients on the interaction terms of X_i and p identify $\beta_1-\beta_0$ and show how observed characteristics shift the treatment effect (and thus the intercept of the MTE curve). The fact that $K(p)$ does not depend on X reflects the assumption that the slope of the MTE curve in u_D does not depend on X. Crucially, this allows identifying $K(p)$ across all values of $X_i=x$, instead of within all values of $X=x$, and it therefore only requires *unconditional* full common support of the propensity score (across all values of $X_i=x$), an assumption which is in many applications more

[20] The two-step estimation of the normal selection model described above is an example in which the MTE is estimated by a control function estimator, instead of the local IV estimator. For a more general comparison between local IV and the control function approach to estimate MTE, see Heckman and Vytlacil (2007, section 4.8).

[21] Steps c and d of the estimation algorithm make clear why a continuous instrument that causes variation between 0 and 1 in the propensity score within each cell of unique values of X is required. If $P(Z)$ does not vary between 0 and 1 in each of the cells, then non-parametric estimation of Y as a function of $\hat{p}$ is not possible across the full unit interval, and thus the MTE curve cannot be identified across the full unit interval (which in turn means that aggregate treatment parameters such as the ATE cannot be calculated).

[22] Full independence between (X,Z) and (U_0,U_1,U_D) is for example invoked in Aakvik et al., (2005), Carneiro et al. (2011) and Carneiro et al. (forthcoming).

[23] The choice of the assumption affects the interpretation of the coefficients and error terms of the outcome equations. Under full independence, β_1, β_0, U_{1i}, and U_{0i} are interpreted as structural or causal, whereas under linear separability they are interpreted in terms of a linear projection.

realistically obtainable than full common support *conditional on* $X_i = x$. From Eq. (22), the MTE is then given by

$$\mathrm{MTE}(X_i = x, U_{Di} = p) = \frac{\partial E[Y_i|X_i = x, P(Z) = p]}{\partial p} = x(\beta_1 - \beta_0) + \frac{\partial K(p)}{\partial p}$$

As before, estimation of the outcome equation requires a pre-estimated propensity score from a first-stage estimation in order to estimate the second stage outcome equation given in (26). Estimation of MTE then proceeds by making varying degrees of functional form assumptions on $K(p)$. Heckman et al. (2006) propose a semiparametric estimation method for Eq. (26). A more parametric approach is to model $K(p)$ as a polynomial in p, which nevertheless allows for considerably more flexibility than the parametric normal model described in Section 4.1.

We provide a brief sketch of the semiparametric and parametric polynomial approaches in Appendix B. The semiparametric, parametric polynomial, and the normal model are all implemented in Stata by the user-written *margte* command, and an accompanying Stata Journal article is available (see Brave and Walstrum, 2014). Further documentation on estimation techniques is also available in the supplementary online material of Heckman et al. (2006).[24]

4.4. Aggregating the MTE into treatment parameters

An important advantage of MTE estimation is that the MTE Eq. (21) can be aggregated into weighted averages over X and U_D to generate aggregate treatment parameters, such as ATE, TT, TUT, and PRTE, or the IV effect associated with a given instrument. Heckman and Vytlacil (2005, 2007) present weights that aggregate the MTE curve along the U_D dimension, conditional on $X_i = x$, which then recover aggregate treatment parameters conditional on $X_i = x$. One may want to further aggregate these conditional parameters over the appropriate distribution of X in order to obtain unconditional aggregate treatment parameters. While in theory U_D is continuous (and the MTE weights are therefore often presented in continuous form), an applied researcher will usually calculate the MTE along a grid of values of U_D and will therefore in practice face a discrete distribution of U_D. Here we present unconditional treatment effects computed from a discrete distribution of U_D. We present the IV weights under the assumptions that potential outcomes are linear in X_i (i.e., $\mu_0(X_i) = X_i\beta_0$ and $\mu_1(X_i) = X_i\beta_1$) and that the MTE is linearly separable into its observed and unobserved part, as in Eq. (25), where the unobserved part is normalized to a mean of zero. These assumptions are in line with the applied MTE literature and the strengthened set of assumptions discussed in Section 4.3. We denote the sample size by N, index individual observations by i, denote the propensity score by p_i, and define $\bar{p}$ as the propensity score averaged over all individuals.

An equally weighted average of the MTE over the full distribution of X and U_D yields the unconditional average treatment effect (ATE) defined in Eq. (7):

$$\mathrm{ATE} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} X_i(\beta_1 - \beta_0)}_{\substack{\text{observed component}\\ \text{of MTE at sample means}}} + \underbrace{\frac{1}{100}\sum_{u=1}^{100}(U_{1i} - U_{0i}|\mathrm{U_D} = {}^{u}/_{100})}_{\substack{\text{equally weighted average over unobserved}\\ \text{component of MTE}}}, \tag{27}$$

which designates the expected treatment effect for an individual with average Xs picked at random from the distribution of U_D.

On the other hand, the treatment effect on the treated (TT) defined in Eq. (8) is an average of the MTE over individuals, whose U_D is such that at their given values of $X = x$ and $Z = z$ (and thus a given propensity score, p_i), they choose to take the treatment. It can be represented by

$$TT = \underbrace{\frac{1}{N}\sum_{i=1}^{N}\frac{p_i}{\bar{p}}X_i(\beta_1 - \beta_0)}_{\substack{\text{observed component of MTE}\\ \text{at means of treated}}} + \underbrace{\sum_{u=1}^{100}\frac{P(p>{}^{u}/_{100})}{100\bar{p}}E(U_1 - U_0|\mathrm{U_D} = {}^{u}/_{100})}_{\substack{\text{weighted average over unobserved component of MTE giving}\\ \text{more weight to low-U}_\mathrm{D}\text{ individuals}}} \tag{28}$$

Note that the observed characteristics X_i are weighted such that observations with a higher propensity score (and thus higher treatment probability) get a higher weight—which corresponds to using observed means of X_i of the treated subpopulation, as implied by Eq. (8). In the unobserved component, the weight of a given value of u_D is related to the share of observations that have a propensity score higher than u_D. Thus, low-u_D individuals (with unobserved characteristics that make them more likely to be treated) get a higher weight, and the weight depends on the distribution of the propensity score (note that while U_D is by construction uniformly distributed, the distribution of p is an empirical question).

Replacing $\frac{p_i}{\bar{p}}$ in the observed component by $\frac{1-p_i}{1-\bar{p}}$ and $\frac{P(p>{}^{u}/_{100})}{100\bar{p}}$ in the unobserved component by $\frac{P(p\leq{}^{u}/_{100})}{100(1-\bar{p})}$ yields the equivalent expression for the TUT defined by Eq. (9). The TUT weights the observed part of the treatment effect more strongly for individuals with a low propensity score (and thus low probability of treatment)—which corresponds to using observed means of X_i of the untreated subpopulation, as implied by Eq. (9). The TUT additionally weights the unobserved part more strongly for individuals at the higher end of the U_D distribution who have a stronger unobserved resistance to treatment.

Denoting the average propensity score under two policies by $\bar{p}'$ and $\bar{p}$, the following expression recovers the PRTE defined by Eq. (10) as a weighted difference between the ATTs under the two policies:

$$\begin{aligned}\mathrm{PRTE} = {} & \frac{1}{N}\sum_{i=1}^{N}\frac{(p_i' - p_i)}{\bar{p}' - \bar{p}}X_i(\beta_1 - \beta_0)\\ & + \sum_{u=1}^{100}E(U_1 - U_0|\mathrm{U_D} = {}^{u}/_{100})\left(\frac{P(p'>{}^{u}/_{100}) - P(p>{}^{u}/_{100})}{(\bar{p}' - \bar{p})100}\right)\end{aligned} \tag{29}$$

Both, observed and unobserved characteristics are weighted proportionately to the policy-induced change in the probability of being treated for individuals with given characteristics. Individual observed characteristics X_i are weighted proportionately to the change in the individual propensity score ($p_i' - p_i$), and each value u_D of the unobserved characteristic is weighted proportionately to the change in the probability of being treated at that value, $P(p' > u_d) - P(p > u_d)$.

Finally, it is possible to calculate IV weights, which recover the IV effect when using a specific instrumental variable. Denoting the IV weights using J as an instrument conditional on X and U_D by $\omega^J_{\mathrm{IV}}(x, u_d)$, the IV effect can be expressed as

$$IV = \underbrace{\sum_{i=1}^{N}\omega^J_{\mathrm{IV}}(x_i)X_i(\beta_1 - \beta_0)}_{\substack{\text{observed component of MTE at means}\\ \text{of individuals shifted by the instrument}}} + \sum_{u=1}^{100}\omega^J_{\mathrm{IV}}({}^{u}/_{100})E(U_{1i} - U_{0i}|\mathrm{U_D} = {}^{u}/_{100}) \tag{30}$$

The weights on the observed characteristics are similar to the weights discussed in Section 2.2.2 and are proportionate to the contribution of individuals with $X_i = x$ to the IV first-stage covariance (see footnote 12). The weights on the unobserved part depend on the effect of Z_i on $P(Z_i)$ at different levels of $P(Z_i)$, weighted by the distribution of $P(Z_i)$. More detail on the estimation of these weights is provided in Appendix C.

[24] This is available at http://jenni.uchicago.edu/underiv/.

For the purpose of illustrating the application of MTE we describe two examples from the education literature in more detail, a paper analyzing marginal returns to college education by Carneiro et al. (2011), as well as our own work on the marginal returns to preschool education (Cornelissen et al., 2016). The papers find fundamentally different selection patterns.

5. Two examples from the applied literature

5.1. Example of MTE applied to returns to college education

Carneiro et al. (2011) analyze the marginal returns to college attendance for the United States, based on a sample of white males from the NLSY aged 28–34 years in 1991. The binary treatment, D_i, is defined as having ever been enrolled in college by 1991. Hence, $D_i = 0$ for high school dropouts and high school graduates and $D_i = 1$ for individuals with some college, college graduates as well as postgraduates. The outcome, Y_i, is the log wage in 1991. As instrumental variables (Z_i in our above notation) that enter the selection equation but not the outcome equation, the authors draw on four instruments, some binary and some continuous, that have been used in previous studies on the returns to college attendance. These are on the one hand cost-shifters (i.e., the presence of a four-year college and average tuition fees in public 4-year colleges in the county of residence during adolescents), and on the other hand variables capturing local labor market opportunities at the time the education decision is taken (i.e., the local average earnings and the local unemployment rate).[25] The instrumental variables, which each identify a different part of the MTE curve, are included simultaneously in order to get larger support in the propensity score. Carneiro et al. (2011) further control for individual's socio-economic background and measures of permanent local labor market characteristics (X_i in our notation).

In their main specification, the authors invoke the assumption of full independence $(X,Z) \perp\!\!\!\perp (U_0, U_1, V)$, implying that the shape of the MTE curve does not vary with X, and the MTE can thus be identified over the unconditional (marginal) support of the propensity score (see Section 4.3).[26] They then estimate the MTE using the semiparametric estimation method outlined in Appendix B.1, which allows for a completely flexible shape of the MTE curve.

Fig. 4A depicts the MTE curve $x(\beta_1 - \beta_0) + E(U_{1i} - U_{0i} | U_{Di} = u_D)$ – see Eq. (25) – evaluating x at mean values in the sample. The figure reveals substantial heterogeneity in the returns to college: Whereas individuals with "low resistance" to college (i.e., very low U_D) enjoy returns of 40%, individuals with "high resistance" to college (i.e., very high U_D) lose from college by 20%. This large range of heterogeneity in the treatment effect due to unobserved characteristics would not be visible if looking only at aggregate treatment effects such as ATE. Since these returns refer to individuals with average X, heterogeneity in returns will be even greater when variation in X is taken into account. The downward sloping shape of the MTE curve highlights high gains for individuals likely to enroll in college (low U_D) and lower gains, or even losses, for individuals less likely to enrol in college (high U_D). Thus, individuals positively select into college based on gains, and individuals seem to possess information about their idiosyncratic returns and are able to make informed choices about college attendance.

In a second step, Carneiro et al. (2011) weight and aggregate the MTEs to compute various treatment effect parameters, as described in Section 4.4. Their preferred estimates are based on the normal selection model outlined in Section 4.1, which is less flexible but results in more precise estimates similar to the ones from the semiparametric estimation method. Column (1) in Table 1 summarizes these estimates. The average treatment effect on the treated (TT), which puts most weight on low U_D individuals, shows substantial returns to college of 14% for the average student selecting into college. By contrast, the returns to college for the average individual (i.e., the ATE) are only 6.7% and the returns for the average person who does not attend college (i.e., the TUT) are close to zero and statistically insignificant. Thus, expansion of college to individuals who currently do not attend would not be effective. Carneiro et al. (2011) also recover the IV effect from MTE. In their case, the IV estimate is between the ATT and the ATE, but clearly masks important heterogeneity in returns to college.

5.2. Example of MTE applied to returns to preschool education

Cornelissen et al. (2016) analyzes heterogeneous treatment effects of a universal child care (preschool) program aimed at 3- to 6-year-olds on children's school readiness. They draw on administrative data on children's outcomes from school readiness examinations for the full population of school entry-aged children in one large region in Germany for the years 1994–2002. The authors exploit a reform during the 1990s that entitled *every* child in Germany to a heavily subsidized half-day child care placement from the third birthday to school entry. This reform was enacted in response to a severe shortage of child care slots, which rationed in particular children who wanted to enroll at the earliest possible age (at age 3 years).[27] As a result, the reform greatly increased the share of children enrolling at the earliest possible age and thus attending child care for at least 3 years from 41% to 67% on average over the program rollout period. Correspondingly, the treatment, D_i, is defined as attending child care for at least 3 years and referred to as "early attendance." Their main outcome variable, Y_i, is a measure of overall school readiness (which determines whether the child is held back from school entry for another year). As an instrument (denoted by Z_i in our notation above), the authors use the supply of available child care slots at the municipality-year level measured by the child care coverage rate, a continuous variable.[28] The control variables (X_i in our above notation) include municipality and examination cohort dummies in addition to individual characteristics such as ethnic minority status, and average socio-demographic characteristics and child care quality indicators at municipality-year level.

Similar to the previous example, the authors also exploit the marginal support of the propensity score (rather than the support conditional on X_i in the ideal case), but based on the linear separability assumption described in Section 4.3, rather than the full independence assumption invoked in Carneiro et al. (2011). Their preferred estimation method is the parametric polynomial approach with a second order polynomial in the propensity score (see Appendix B.2). This model restricts the MTE curve to a straight line and thus appears equally restrictive as the normal section model (in that it rules out a non-monotonic shape). To rule out concerns that this restrictive choice drives their results, the authors show that their main pattern of results is robust to estimating more flexible MTE curves by using higher-order polynomials or implementing the semiparametric estimation method.

Fig. 4B depicts the resulting linear MTE curve, evaluated at mean values of X in the sample. In contrast to the previous example, the MTE curve now exhibits an upward sloping shape, indicating a pattern of reverse selection on gains. Whereas children with low resistance to prolonged child care attendance (low U_D) do not gain, improvements in school readiness are substantial for children with a high resistance

[25] The number of IVs is further expanded by interacting these variables with an ability measure (Armed Forces Qualification Test—AFQT), mother's years of schooling, and number of siblings.

[26] The conditional density of the propensity score conditional on values of a linear index in X reveals an extremely narrow support of the propensity score at each value of the index (Figure 2 in Carneiro et al., 2011), preventing estimation of MTE in the ideal case (see Section 4.1). The unconditional (marginal) support of the propensity score, in contrast, encompasses almost the full unit interval (Figure 3 in Carneiro et al., 2011).

[27] Children who wanted to enter at an older age (who may already have waited on the waiting list for one year) were generally given priority.

[28] Linear and squared terms of the instrument are included, and in the main specification both of these terms are interacted with a quadratic in age, gender, and ethnic minority status.

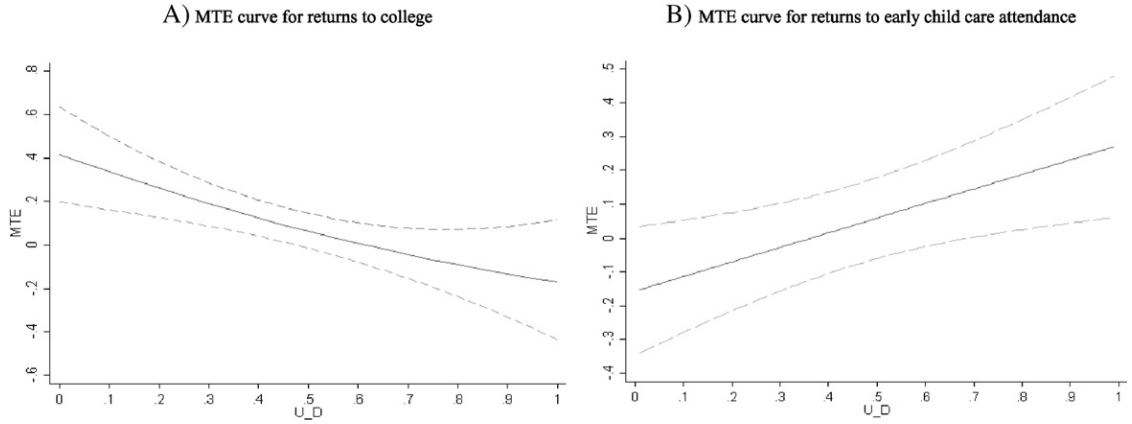

Fig. 4. MTE curves. *Notes:* Part A depicts the MTE curve of Carneiro et al. (2011, Figure 4) for the wage returns to college estimated by the semiparametric method (see Appendix B.1). Part B shows the MTE curve of Cornelissen et al. (2016, Figure 4, Part A) for the returns to early child care attendance on school readiness estimated by the parametric polynomial method (see Appendix B.2). In both figures, the 90% confidence interval is based on bootstrapped standard errors.

to prolonged child care attendance (high U_D). In consequence, the TUT – which indicates that early child care attendance would boost school readiness of children currently not enrolled in child care by 17.3 percentage points – exceeds the ATE and TT, neither of which is statistically significant (see column (2) of Table 1).[29]

As in Carneiro et al. (2011), the linear IV estimate turns out to be similar in magnitude to ATE and masks important heterogeneity in returns. Moreover, the linear IV effect estimated by 2SLS is very similar to the effect obtained when applying the IV weights to the MTE curve, which can be considered a specification check for the functional form of the MTE curve.[30]

The authors confirm a pattern of reverse selection on gains also based on observed characteristics. For example, minority children are 12 percentage points less likely to attend preschool, but their treatment effect is about 12 percentage points higher than for majority children.[31] The authors provide additional evidence that children with a high resistance to attend early child care come from disadvantaged backgrounds and have larger treatment effects because of their worse outcome when not enrolled in child care.

These findings have important policy implications. They first highlight that early child care attendance acts as an equalizer. They also imply that policies that successfully attract children with high resistance not currently enrolled in early child care may yield large returns. Further, programs targeted at minority and disadvantaged children are likely to be more cost effective and beneficial than universal child care programs.

6. Conclusions

Some recent surveys provide insightful discussions about MTE (e.g., Heckman and Vytlacil, 2007; French and Taber, 2011; Blundell and Costa Dias, 2009), and excellent technical treatments of MTE can be found in the papers by Heckman and Vytlacil (1999, 2001b, 2005) and in the application of Carneiro et al. (2011). Drawing on these earlier papers, we provide an introduction to the MTE framework, developing it in a way that we believe is accessible to the applied economist. We also provide a set of lecture slides to accompany this article (available from the authors' personal websites).

Our framework of analysis is a generalized Roy model based on the potential outcomes framework and a latent variable discrete choice model for selection into treatment, assuming – as typically done in empirical applications – linear separability in observables and unobservables. Within this framework, we first define different treatment effects of interest, such as the average treatment effect (ATE), the average treatment effect on the treated (ATT), the average treatment effect on the untreated (ATU), and the policy-relevant treatment effect (PRTE). Next, we review the well-known case of IV estimation with a discrete instrument and highlight that the resulting local average treatment effect (LATE) identified by a binary instrumental variable does not necessarily represent a treatment parameter for an economically interesting group of the population, except in some important specific cases which we discuss. In contrast to ATE, ATT, ATU, and PRTE, which are well defined parameters that answer economic policy questions, LATE is defined by the instrumental variable used.

In a next step, we turn to IV estimation with a continuous instrument and demonstrate that the 2SLS estimator may be viewed as a weighted average of LATEs obtained from *all* possible pairs of values of the continuous instrument. Not only does this estimator lack a straightforward interpretation, but it may also hide interesting information about the pattern of treatment effect heterogeneity. We also contrast IV estimation with a control function estimator for the correlated random coefficients model, which identifies a more general effect than IV (the ATE) and reveals some information on the pattern of selection based on unobserved gains, albeit under stronger assumptions.

Table 1
Treatment effects parameters

	(1) Returns to college	(2) Returns to early child care attendance
ATE	0.067* (0.038)	0.059 (0.072)
TT	0.143*** (0.035)	−0.051 (0.080)
TUT	−0.007 (0.071)	0.173** (0.085)
IV	0.095** (0.039)	0.065 (0.133)

Notes: The table reports the average treatment effect (ATE), the treatment effect on the treated (TT), treatment effect on the untreated (TUT), and the IV estimate from a linear IV specification for the papers presented in Sections 5.1 and 5.2. Column (1) refers to the results reported in Table 5 in Carneiro et al. (2011). Column (2) refers to the results shown in Table 5, column (1) in Cornelissen et al. (2016). Bootstrapped standard errors are reported in parentheses.

* Statistically significant at 0.10 level.
** Statistically significant at 0.05 level.
*** Statistically significant at 0.01 level.

[29] Kline and Walters (forthcoming) uncover a pattern of reverse selection on gains for Head Start attendance when the nontreated state is home care. Aakvik et al. (2005) find reverse selection on gains in the context of a Vocational Rehabilitation training program.

[30] MTE curves derived under different functional form assumptions may yield different weighted IV effects, while neither the IV weights nor the 2SLS estimator depend on the functional form of the MTE curve. A large discrepancy between the weighted IV effect and the 2SLS IV effect may therefore indicate a specification error in the functional form of the MTE curve.

[31] Note, however, that because they do not assume full independence of (X,Z) and (U_0,U_1,V), the partitioning of the treatment effect into the observed and unobserved components has no causal interpretation, meaning that the higher treatment effect for minority children confounds higher treatment effects that are causally due to minority status with those that are due to unobserved characteristics correlated with minority status.

We then discuss MTE estimation as an alternative and more informative way of exploiting a continuous instrument which, unlike IV and control function estimation, allows the identification of a variety of treatment parameters such as ATE, TT, TUT, and PRTE. Instead of aggregating the underlying LATEs into one overall effect, MTE estimation aims at identifying a continuum of treatment effects along the full distribution of the individual unobserved characteristic that drives treatment decisions. We clarify the assumptions underlying the MTE framework, distinguishing between an ideal case, in which the data are rich enough for nonparametric estimation under a set of assumptions no stronger than the general IV assumptions, and a more realistic case in which less ideal data can be exploited using semiparametric and parametric methods (of which we provide a brief sketch) under strengthened assumptions. We finally illustrate how MTE estimation is implemented in practice, and which additional insights can be gained from MTE estimation compared to conventional 2SLS estimation, based on two examples from the applied MTE literature: the wage returns to college attendance and on the effects of preschool attendance on school readiness.

Appendix A. Policy-relevant treatment effects

The policy-relevant treatment effect conditional on X_i, $PRTE(x)$, is the mean effect of going from a baseline policy to an alternative policy per net person shifted:

$$\begin{aligned}
\text{PRTE}x &= \frac{E[Y_i|X_i=x,\quad \text{alternative policy}]-E[Y_i|X_i=x,\quad \text{baseline policy}]}{E[D_i|X_i=x,\quad \text{alternative policy}]-E[D_i|X_i=x,\quad \text{baseline policy}]}\\
&= \frac{E[Y_{0i}+Y_{1i}-Y_{0i}D\sim_i|X_i=x]-E[Y_{0i}+Y_{1i}-Y_{0i}D_i|X_1=x]}{E[D\sim_i|X_i=x]-E[D_i|X_i=x]}\\
&= \frac{E[Y_{1i}-Y_{0i}|X_i=x,D\sim_i=1]E[D\sim_i|X_i=x]-E[Y_{1i}-Y_{0i}|X_i=x,D_i=1]E[D_i|X_i=x]}{E[D\sim_i|X_i=x]-E[D_i|X_i=x]}\\
&= \mu_1(x)-\mu_0(x)\\
&\quad + \frac{E[U_{1i}-U_{0i}|X_i=x,D\sim_i=1]E[D\sim_i|X_i=x]-E[U_{1i}-U_{0i}|X_i=x,D_i=1]E[D_i|X_i=x]}{E[D\sim_i|X_i=x]-E[D_i|X_i=x]}
\end{aligned}$$

The corresponding unconditional effect is

$$\begin{aligned}
\text{PRTE} &= \frac{E[Y_{1i}-Y_{0i}|D\sim_1=1]E[D\sim_i]-E[Y_{1i}-Y_{0i}|D_i=1]E[D_i]}{E[D\sim_i]-E[D_i]}\\
&= \frac{E[\mu_1(X_i)-\mu_0(X_i)|D\sim_i=1]E[D\sim_i]-E[\mu_1(X_i)-\mu_0(X_i)|D_i=1]E[D_i]}{E[D\sim_i]-E[D_i]}\\
&\quad + \frac{E[U_{1i}-U_{0i}|D\sim_i=1]E[D\sim_i]-E[U_{1i}-U_{0i}|D_i=1]E[D_i]}{E[D\sim_i]-E[D_i]}
\end{aligned}$$

Appendix B. Sketch of common estimation methods for the MTE

B.1. Semiparametric estimation

A semiparametric version of estimating (26) consists in the following steps (see for example Appendix B of Heckman et al., 2006, for a more detailed description):

a. Purging X and Xp from the effect of $K(p)$ by regressing each of them in turn on p using local polynomial regression (or a parametric polynomial in p), and predicting the residuals.
b. Regressing Y on the residualized version of X and Xp obtained under a. using a linear regression, and predicting the residual.
c. Regressing the residualized version of Y obtained under b. on p by local polynomial regression to identify $K(p)$.
d. Obtaining the MTE curve as the derivative of $K(p)$.

Note that in order to identify $K(p)$ over the full unit interval by this semiparametric method, one still needs full common unconditional support of the propensity score. If the support of the propensity score is limited (maybe because of limited variation in the instrument), then one possibility is to continue to use the semiparametric method, but identify the MTE only over some sub-range of the unit interval. While this approach reveals useful information on the treatment effects and the selection pattern for the range in which the MTE can be identified, it precludes calculation of aggregate treatment effects such as the ATE, TT, and TUT, as they require aggregating over the full unit interval. Alternatively, one can take more parametric approaches described below, based on which the MTE curve can be extrapolated out of the support of the propensity score. More parametric approaches can also be useful when there is full support and nonparametric and semiparametric approaches are too time-consuming or too demanding on the data (e.g., if results are very sensitive to small changes in the data or specification).

B.2. Parametric polynomial estimation

The parametric polynomial MTE model replaces $K(p)$ in Eq. (26) by a kth-order polynomial in p, so that the outcome equation becomes:

$$Y = X\beta_0 + X(\beta_1-\beta_0)p + \sum_{k=2}^{K}\alpha_k p^k + \upsilon,$$

and as before the MTE curve is the derivative of this equation with respect to p. The higher the degree of the polynomial, the more flexible the MTE curve is estimated. For example, choosing a second order polynomial (K = 2) restricts the MTE curve to be linear, which may hide more flexible patterns, such as a U-shape in the MTE curve. However, strong parametric assumptions are powerful. As shown by Brinch et al. (forthcoming), a linear MTE curve can be identified with a dummy variable instrument (albeit with an alternative estimation method to the conventional LIV method).

Appendix C. Computation of IV weights in the linear separable model

$\omega^J_{\text{IV}}(x_i)$ in Eq. (30) represent the IV weights conditional on X after integrating out U_D, and $\omega^J_{\text{IV}}(^u/_{100})$ represent the IV weights conditional on U_D after integrating out X. We propose the following estimation approach:

1. Running the 2SLS first-stage regression of the treatment D_i on the covariates X_i and the vector of excluded instruments J_i and predicting D_i from this regression.
2. Regressing D_i on the covariates X_i and predicting the residual ϑ from this regression. ϑ aggregates the excluded instruments into one scalar instrument that is orthogonal to the covariates X_i. The bivariate IV estimator $a^{\text{IV}} = \frac{\text{Cov}(Y,\vartheta)}{\text{Cov}(D,\vartheta)}$, using ϑ as a single instrument, reproduces the exact same IV estimate as a 2SLS second stage regression of Y on D_i and X_i.
3. Computing $\omega^J_{\text{IV}}(x_i) = \frac{\frac{1}{N}(D_i-\overline{D})(\vartheta_i-\overline{\vartheta})}{\text{Cov}(D,\vartheta)}$, the weight given by each individual's contribution to the first-stage covariance $\text{Cov}(D,\vartheta)$ divided by this covariance [rationalized by assuming that $Y_i = a + b_iD_i$ and noting that in this case, because $E(\vartheta_i)=0$, $\text{Cov}(Y_i,\vartheta_i) = \text{Cov}(b_iD_i,\vartheta_i) = E(b_iD_i\vartheta_i)$, so $\text{Cov}(Y_i,\vartheta_i) = \frac{1}{N}\sum_i b_iD_i\vartheta_i$. Similarly, $\text{Cov}(D_i,\vartheta_i) = \frac{1}{N}\sum_i D_i\vartheta_i$, meaning that $a^{\text{IV}} = \frac{\text{Cov}(Y_i,\vartheta_i)}{\text{Cov}(D_i,\vartheta_i)} = \frac{1}{N}\sum_i b_i\frac{D_i\vartheta_i}{\frac{1}{N}\sum_i D_i\vartheta_i}$ and b_i is weighted by each individual's contribution to the first-stage covariance.]
4. Computing $\omega^J_{\text{IV}}(^u/_{100})$ as the sample analog of $\frac{1}{100}\frac{[E(\vartheta|p_i>{}^u/_{100})-E(\vartheta_i)]P(p_i>{}^u/_{100})}{\text{Cov}(D,\vartheta)}$ (see, e.g., Eq. (19) and Appendix B.3 in Heckman et al., 2006).

References

Aakvik, A., Heckman, J.J., Vytlacil, E.J., 2005. Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. J. Econ. 125 (1–2), 15–51.

Angrist, J.D., 1990. Lifetime earnings and the Vietnam Era draft lottery: evidence from social security administrative records. Am. Econ. Rev. 80 (3), 313–336.
Angrist, J.D., 1991. Grouped-data estimation and testing in simple labor-supply models. J. Econ. 47 (2–3), 243–266.
Angrist, J.D., Imbens, G.W., 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. J. Am. Stat. Assoc. 90 (430), 431–442.
Angrist, J.D., Pischke, J.-S., 2009. Mostly harmless econometrics: an empiricist's companion. Princeton University Press, Princeton.
Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. J. Am. Stat. Assoc. 91 (434), 444–455.
Balfe, Cathy. 2015. "Heterogeneity, Selection and Advantage in the Graduate and Non-Graduate Labour Market." Unpublished Manuscript, University College London.
Basu, A., Heckman, J.J., Navarro-Lozano, S., Urzúa, S., 2007. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Econ. 16 (11), 1133–1157.
Basu, A., Jones, A.M., Rosa Dias, P., 2014. The roles of cognitive and non-cognitive skills in moderating the effects of mixed-ability schools on long-term health. NBER Working Paper 20811. National Bureau of Economic Research.
Björklund, A., Moffitt, R., 1987. The estimation of wage gains and welfare gains in self-selection models. Rev. Econ. Stat. 69 (1), 42–49.
Blundell, R., Costa Dias, M., 2009. Alternative approaches to evaluation in empirical microeconomics. J. Hum. Resour. 44 (3), 565–640.
Brave, S., Walstrum, T., 2014. Estimating marginal treatment effects using parametric and semiparametric methods. Stata J. 14 (1), 191–217.
Brinch, C.N., Mogstad, M., Wiswall, M., 2016. Beyond LATE with a discrete instrument. J. Polit. Econ. (forthcoming).
Card, D., 2001. Estimating the return to schooling: progress on some persistent econometric problems. Econometrica 69 (5), 1127–1160.
Carneiro, P., Heckman, J.J., Vytlacil, E.J., 2011. Estimating marginal returns to education. Am. Econ. Rev. 101 (6), 2754–2781.
Carneiro, P., Lokshin, M., Ridao-Cano, C., Umapathi, N., 2016. Average and marginal returns to upper secondary schooling in Indonesia. J. Appl. Econ. (http://onlinelibrary.wiley.com/doi/10.1002/jae.2523/abstract, forthcoming).
Chetty, R., Hendren, N., Katz, L.F., 2016. The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment. Am. Econ. Rev. 106 (4), 855–902.
Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2016. "Who Benefits from Universal Childcare? Estimating Marginal Returns to Early Childcare Attendance." Unpublished Manuscript, University College London.
Deaton, A.S., 2009. Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development. NBER Working Paper 14690. National Bureau of Economic Research.
Doyle, J.J., 2007. Child protection and child outcomes: measuring the effects of foster care. Am. Econ. Rev. 97 (5), 1583–1610.
Eisenhauer, P., Heckman, J.J., Vytlacil, E., 2015. The generalized Roy model and the cost-benefit analysis of social programs. J. Polit. Econ. 123 (2), 413–443.
Felfe, Christina, and Rafael Lalive. 2015. "Does Early Child Care Affect Children's Development" Unpublished Manuscript.
French, E., Song, J., 2014. The effect of disability insurance receipt on labor supply. Am. Econ. J. Econ. Pol. 6 (2), 291–337.
French, E., Taber, C., 2011. Identification of models of the labor market. In: Ashenfelter, O., Card, D. (Eds.), Chap. 6 in *Handbook of Labor Economics* 4, Part A. Elsevier, Amsterdam, pp. 537–617.
Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann. Econ. Soc. Meas. 5 (4), 475–492.
Heckman, J.J., 1997. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. J. Hum. Resour. 32 (3), 441–462.
Heckman, J.J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. J. Econ. 30 (1), 239–267.
Heckman, J.J., Smith, J., 1998. Evaluating the Welfare State. Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium, no. 31. Cambridge University Press, Cambridge, p. 241.
Heckman, J.J., Urzúa, S., 2010. Comparing IV with structural models: what simple IV can and cannot identify. J. Econ. 156 (1), 27–37.
Heckman, J.J., Vytlacil, E., 1998. Instrumental variables methods for the correlated random coefficient model: estimating the average rate of return to schooling when the return is correlated with schooling. J. Hum. Resour. 33 (4), 974–987.
Heckman, J.J., Vytlacil, E., 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proc. Natl. Acad. Sci. 96 (8), 4730–4734.
Heckman, J.J., Vytlacil, E., 2001a. Policy-relevant treatment effects. Am. Econ. Rev. 91 (2), 107–111.
Heckman, J.J., Vytlacil, E., 2001b. Local instrumental variables. In: Hsiao, C., Morimune, K., Powell, J. (Eds.), Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya. Cambridge University Press, Cambridge.
Heckman, J.J., Vytlacil, E., 2005. Structural equations, treatment effects, and econometric policy evaluation. Econometrica 73 (3), 669–738.
Heckman, J.J., Vytlacil, E., 2007. 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: Heckman, J.J., Leamer, E.E. (Eds.), Chap. 71 in Handbook of Econometrics. Elsevier, Amsterdam.
Heckman, J.J., LaLonde, R.J., Smith, J.A., 1999. The economics and econometrics of active labor market programs. In: Ashenfelter, O., Card, D. (Eds.), Chap. 31 in *Handbook of Labor Economics* vol. 3, Part A. Elsevier, Amsterdam, pp. 1865–2097.
Heckman, J.J., Urzúa, S., Vytlacil, E., 2006. Understanding instrumental variables in models with essential heterogeneity. Rev. Econ. Stat. 88 (3), 389–432.
Holland, P.W., 1986. Statistics and causal inference. J. Am. Stat. Assoc. 81 (396), 945–960.
Imbens, G.W., Angrist, J.D., 1994. Identification and estimation of local average treatment effects. Econometrica 62 (2), 467–475.
Joensen, J.S., Nielsen, H.S., 2016. Mathematics and gender: heterogeneity in causes and consequences. Econ. J. 126 (593), 1129–1163.
Johnston, J., 1963. Econometric Methods. McGraw-Hill, New York.
Kamhöfer, D.A., Schmitz, H., Westphal, M., 2015. Heterogeneity in marginal non-monetary returns to higher education. Ruhr Economic Papers 591. RWI.
Kaufmann, K.M., 2014. Understanding the income gradient in college attendance in Mexico: the role of heterogeneity in expected returns. Quant. Econ. 5 (3), 583–630.
Kline, P., Walters, C., 2016. Evaluating public programs with close substitutes: the case of Head Start. Q. J. Econ. (forthcoming).
Lee, L.-F., 1979. Identification and estimation in binary choice models with limited (censored) dependent variables. Econometrica 47 (4), 977–996.
Maddala, G.S., 1992. Introduction to Econometrics. second ed Macmillan Publishing Company, New York.
Maestas, N., Mullen, K.J., Strand, A., 2013. Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. Am. Econ. Rev. 103 (5), 1797–1829.
Noboa-Hidalgo, G.E., Urzúa, S.S., 2012. The effects of participation in public child care centers: evidence from Chile. J. Hum. Cap. 6 (1), 1–34.
Nybom, M., 2014. The Distribution of Lifetime Earnings Returns to College. Working Paper 2/2014. Swedish Institute for Social Research, Stockholm University.
Oreopoulos, P., 2006. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. Am. Econ. Rev. 96 (1), 152–175.
Quandt, R.E., 1972. A new approach to estimating switching regressions. J. Am. Stat. Assoc. 67 (338), 306–310.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66 (5), 688–701.
Wooldridge, J.M., 2015. Control function methods in applied econometrics. J. Hum. Resour. 50 (2), 420–445.