The Stata Journal (2018)
18, Number 1, pp. 118–158

Exploring marginal treatment effects: Flexible estimation using Stata

Martin Eckhoff Andresen
Statistics Norway
Oslo, Norway
martin.eckhoff.andresen@gmail.com

Abstract. In settings that exhibit selection on both levels and gains, marginal treatment effects (MTE) allow us to go beyond local average treatment effects and estimate the whole distribution of effects. In this article, I survey the theory behind MTE and introduce the package `mtefe`, which uses several estimation methods to fit MTE models. This package provides important improvements and flexibility over existing packages such as `margte` (Brave and Walstrum, 2014, *Stata Journal* 14: 191–217) and calculates various treatment-effect parameters based on the results. I illustrate the use of the package with examples.

Keywords: st0516, mtefe, margte, heterogeneity, marginal treatment effects, instrumental variables

1 Introduction

Well-known instrumental variables (IVs) methods solve problems of selection on levels, estimating local average treatment effects (LATEs) for instrument compliers even with nonrandom selection into treatment. However, in the more reasonable case with selection into treatment based both on levels and gains, LATEs may represent average treatment effects (ATEs) for very particular subpopulations, and we learn little about the distribution of treatment effects in the population at large.

Marginal treatment effects (MTEs) allow us to go beyond LATEs in settings that exhibit this sort of selection. MTEs are the ATEs for people with either a particular resistance to treatment or at a particular margin of indifference. Thus, MTEs capture heterogeneity in the treatment effect along the unobserved dimension we call resistance to treatment. This is precisely what generates selection on unobserved gains: people who choose treatment because they have a particularly low resistance might have different gains than those with high resistance. Usually at the cost of stronger assumptions than required under standard IVs, we can estimate the full distribution of (marginal) treatment effects and back out the parameters of interest. These include ATEs, ATEs on the treated (ATT) and untreated (ATUT), and policy-relevant treatment effects (PRTEs) from a hypothetical policy that shifts the propensity to choose treatment.

The most common way of estimating MTEs is via the method of local IVs (Heckman and Vytlacil 2007). Alternatively, MTEs can be estimated using the separate approach (Heckman and Vytlacil 2007; Brinch, Mogstad, and Wiswall 2017) or, in the baseline case with joint normal errors, maximum likelihood. When we estimate MTEs using the

separate approach or maximum likelihood, similarities to selection models are apparent; we are basically fitting a traditional selection model. MTEs thus represent old ideas in new wrapping, but they still provide substantial contributions to the interpretation and estimation of models in settings with essential heterogeneity. Unfortunately, commands for flexibly estimating MTEs and using their output are not easily available in popular software such as Stata. Brave and Walstrum (2014) document the package `margte`, which estimates MTEs, but this package has some important limitations.

In this article, I introduce the package `mtefe`, which fits parametric and semiparametric MTE models. These include the joint normal, polynomial, and semiparametric models like Brave and Walstrum (2014); `mtefe` additionally adds the option of spline functions in the MTE for increased flexibility. It can fit all models using either local IVs or the separate approach as well as maximum likelihood for the joint normal model, while the estimation method in `margte` depends on the model. Furthermore, `mtefe` allows for fixed effects using Stata's categorical variables, which is important to isolate exogenous variation in many applications and provides gains in computational speed over generating dummies manually. `mtefe` supports frequency and probability weights, which is important to obtain population estimates in many datasets.

In addition, `mtefe` exploits the full potential of MTEs by calculating treatment-effect parameters as weighted averages of the MTE curve, shedding light on why, for example, the LATE differs from the ATE. Other improvements include calculation of analytic standard errors (that admittedly ignore the uncertainty in the propensity-score estimation), large improvements in computational speed when fitting semiparametric models, and reestimation of the propensity score when bootstrapping for more appropriate inference.

Although quickly becoming a part of the toolbox of the applied econometrician, applied work using MTEs often imposes restrictive parametric assumptions. The Monte Carlo simulations in appendix B illustrates how conclusions might be sensitive to these choices, at least under the particular data-generating processes used. This illustrates the importance of correct functional form and highlights that researchers should probe their results to functional form assumptions when using MTEs as well as base the choice of functional form on detailed knowledge about the case at hand—what might constitute the unobservables and sound economic arguments for the nature of the relationship between these unobservables and the outcome.

This article proceeds as follows: Section 2 reviews the theory behind MTEs, identifying assumptions, estimation methods, and derivation of treatment parameter weights. Section 3 presents the `mtefe` command, its most important options, and examples of its use. Section 4 concludes. Appendix A details the estimation algorithm of `mtefe`, and appendix B contains the Monte Carlo simulations.

2 Estimation of MTEs

MTEs are based on the generalized Roy model:

$$Y_j = \mu_j(X) + U_j \qquad \text{for } j = 0, 1 \tag{1}$$

$$Y = DY_1 + (1 - D)Y_0 \tag{2}$$

$$D = \mathbb{1}\{\mu_D(Z) > V\} \qquad \text{where } Z = (X, Z_-) \tag{3}$$

Y_1 and Y_0 are the potential outcomes in the treated and untreated state, such as log wages with and without a college degree. They are both modeled as functions of observables X, which may contain fixed effects.

Equation (3) is the selection equation and can be interpreted as a latent index because $\mathbb{1}$ is the indicator function. It is a reduced-form way of modeling selection into treatment as a function of observables X and instruments Z_- that affect the probability of treatment but not the potential outcomes.

The unobservable V in the choice equation is a negative shock to the latent index determining treatment. It is often interpreted as unobserved resistance to or negative preference for treatment. As long as the unobservable V has a continuous distribution, we can rewrite the selection equation as $P(Z) > U_D$, where U_D represents the quantiles of V and $P(Z)$ represents the propensity score. U_D, by construction, has a uniform distribution in the population.[1]

Identification of this model requires the following assumption:

Assumption 1: Conditional independence $(U_0, U_1, V) \perp Z_-|X$

The model described by (1)–(3), together with Assumption 1, implies and is implied by the standard assumptions in Imbens and Angrist (1994) necessary to interpret an IV as a LATE. Vytlacil (2002) shows how the standard IV assumptions of relevance, exclusion, and monotonicity[2] are equivalent to some representation of a choice equation as in (3). Thus, the model described above is no more restrictive than the LATE model used in standard IV analysis.

In principle, it is possible to estimate MTEs with no further assumptions. However, this requires full support of the propensity scores in both treated and untreated samples for all values of X. In practice, this is rarely feasible, as shown, for example, in Carneiro, Heckman, and Vytlacil (2011). Instead, most applied articles (Carneiro, Heckman, and Vytlacil 2011; Maestas, Mullen, and Strand 2013; Carneiro and Lee 2009; Cornelissen et al. Forthcoming; Felfe and Lalive 2014) proceed by assuming a stronger assumption:

1. This normalization also ensures that U_D is uniform within cells of X. For details, see Mogstad and Torgovitsky (2017) and Matzkin (2007).
2. Or rather, uniformity, because it is a condition across people, not across realizations of Z; this assumption requires that $\Pr(D = 1|Z = z) \geq \Pr(D = 1|Z = z')$ or the other way around for all people, but it does not require full monotonicity in the classical sense.

Assumption 2: Separability $\mathbb{E}(U_j \mid V, X) = \mathbb{E}(U_j \mid V)$

This can be a strong assumption and must be carefully evaluated by the researcher for each application. Nonetheless, it is clearly less restrictive than, for example, a joint normal distribution of (U_0, U_1, V) as assumed by traditional selection models. This article, along with the applied MTE literature[3] and the `mtefe` command, proceeds by imposing this stronger assumption, as well as working with linear versions of $\mu_j(X) = X\beta_j$ and $\mu_D(Z) = \gamma Z$.[4]

This assumption has two important consequences: First, the MTE is identified over the common support, unconditional on X. Second, the MTEs are additively separable in U_D and X. This implies that the shape of the MTE is independent of X; the intercept, but not the slope, is a function of X.

Using this model, we find that the returns to treatment are simply the difference between the outcomes in the treated and untreated states. MTEs were introduced by Björklund and Moffitt (1987), later generalized by Heckman and Vytlacil (1999, 2001, 2005, 2007), and with the above assumption and index structure, can be defined as

$$\begin{aligned} \text{MTE}(x,u) &\equiv \mathbb{E}(Y_1 - Y_0 | X = x, U_D = u) \\ &= \underbrace{x(\beta_1 - \beta_0)}_{\text{heterogeneity in observables}} + \underbrace{\mathbb{E}(U_1 - U_0 | U_D = u)}_{k(u)\text{: heterogeneity in unobservables}} \end{aligned}$$

They measure average gains in outcomes for people with particular values of X and the unobserved resistance to treatment U_D. Alternatively, the MTE can be interpreted as the mean return to treatment for individuals at a particular margin of indifference. The above expression shows how the separability assumption allows us to separate the treatment effect into one part varying with observables and another part varying across the unobserved resistance to treatment.

MTEs are closely related to LATE. A LATE is the average effect of treatment for people who are shifted into (or out of) treatment when the instrument is exogenously shifted from z to z'. In the above choice model, these people have U_D in the interval $(P(z), P(z'))$. Note how, when $z - z'$ is infinitesimally small, so that $P(z') = P(z)$, the LATE converges to the MTE. A MTE is thus a limit form of LATE (Heckman and Vytlacil 2007).

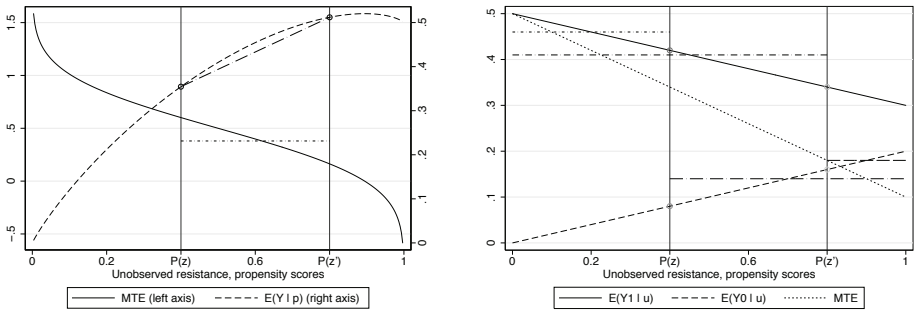
3. Some articles apply a stronger full independence assumption: $(U_0, U_1, V) \perp X, Z_-$. This implies the separability in Assumption 2 but is stricter than necessary. In particular, the separability assumption places no restrictions on the dependence between V and X. See Mogstad and Torgovitsky (2017, in particular sec. 6) for a review and a detailed discussion of the differences between these assumptions.
4. Note, however, that you are free to specify fixed effects, so that a fully saturated model may be specified.

2.1 Estimation using local IV

One way of estimating MTEs is using the method of local IV developed by Heckman and Vytlacil (1999, 2001, 2005). This method identifies the MTE as the derivative of the conditional expectation of Y with respect to the propensity score.

First, consider the intuition for why the derivative identifies the MTE using figure 1a. Abstracting from covariates, the dotted line represents the expected Y for each value of the propensity score, which can be estimated from data given a propensity-score model. When p increases, the probability of getting the treatment increases. If the treatment effect is constant, $\mathbb{E}(Y|p)$ is linear in p. In contrast, under essential heterogeneity, the increase in p has the additional effect that the expectations of the error terms change, resulting in nonlinearities in $\mathbb{E}(Y|p)$. Local IVs therefore identify the MTEs from nonlinearities in the expectation of Y given p.

Now, consider two particular instrument values z and z', generating propensity scores $P(z) = 0.4$ and $P(z') = 0.8$, respectively. This shift of the instrument induces people with unobserved resistance $P(z) < U_D \leq P(z')$ to switch intro treatment. The change in $\mathbb{E}(Y|p)$ relative to the change in propensity scores, $\{\mathbb{E}(Y|p') - \mathbb{E}(Y|p)\}/(p' - p)$, identifies the average of the MTE for U_D in the interval p, p'. The slope of the long dash dot line depicted in the figure is the average derivative over this interval, which is equal to the average of the MTE in the same interval—the short dash dot line. Take this reasoning to the limit by looking at closer and closer values of p and p', and the above expression collapses to the derivative of $\mathbb{E}(Y|p)$. At the point where p' is just an infinitesimal increase over p, the derivative identifies the MTE at the point where $U_D = p$.



(a) Local IV

(b) Separate approach

Figure 1. Identification of MTEs

Formally, using $p = P(Z)$, the index structure, and the separability assumption, we can show that

$$\begin{aligned}\mathbb{E}\{Y|X = x, P(Z) = p\} &= \mathbb{E}\{Y_0 + D(Y_1 - Y_0)|X = x, P(Z) = p\} \\ &= x\beta_0 + x(\beta_0 - \beta_1)p + \underbrace{p\mathbb{E}(U_1 - U_0|U_D \leq p)}_{K(p)} \end{aligned} \tag{4}$$

where we can normalize $\mathbb{E}(U_j) = 0$ as long as X includes an intercept. $K(p)$ is a nonlinear function of p that captures heterogeneity along the unobservable resistance to treatment U_D; if people with different resistances to treatment have different expectations of the error terms, the MTE will be nonlinear. The $K(p)$ notation follows Brinch, Mogstad, and Wiswall (2017).

Taking the derivative of this expression with respect to p and evaluating it at u, we get the MTE

$$\begin{aligned}\frac{\partial \mathbb{E}\{Y|X = x, P(Z) = p\}}{\partial p}|_{p=u} &= x(\beta_1 - \beta_0) + \frac{\partial\{p\mathbb{E}(U_1 - U_0|U_D \leq p)\}}{\partial p}|_{p=u} \\ \text{MTE}(x, u) &= (\beta_1 - \beta_0)x + \underbrace{\mathbb{E}(U_1 - U_0|U_D = u)}_{k(u)}\end{aligned}$$

where $k(u) = \mathbb{E}(U_1 - U_0|U_D = u)$.

The above suggests the following estimation procedure: We start by identifying the selection into treatment based on (3), using a probability model, such as probit or logit; the linear probability model; or even a semiparametric binary choice model. With this estimate of $P(Z)$ in hand, what remains is to make an assumption about the unknown function $K(p) = p\mathbb{E}(U_1 - U_0|U_D \leq p)$,[5] estimate the conditional expectation of Y from (4), and form its derivative to get the MTE. The different functional form assumptions commonly made on $K(p)$ are summarized in table 1. Alternatively, if we are unwilling to make parametric assumptions, the MTE can be estimated using semiparametric methods as summarized in table 2 by estimating (4) as a partially linear model.

2.2 Estimation using the separate approach

Alternatively, as suggested by Heckman and Vytlacil (2007), MTEs can be estimated using the separate approach. This has the benefit of estimating all the parameters of both the potential outcomes so that we can plot these over the distribution of U_D.

First, consider the intuition using figure 1b. Abstracting from covariates, the solid line depicts the true $\mathbb{E}(Y_1|U_D = u)$, the expected value of the outcome in the treated state for a given value of u. Likewise, the dashed line depicts the expected value of the outcome in the untreated state. Next, imagine comparing two particular values of the instrument, z versus z'. Individuals with $Z = z$ have a predicted propensity score of $P(z) = 0.4$, while individuals with z' have $P(z') = 0.8$. We will have both treated and

5. Or usually directly on $k(u)$, and then we exploit that $K(p) = \int_0^p k(u)du$.

untreated individuals in both these groups. For untreated individuals with $Z = z$, we know that their resistance $U_D > P(z) = 0.4$. Therefore, the average outcome among these people informs us of the average of the dashed line between 0.4 and 1, depicted by the long dash dot line. In contrast, the treated individuals with $Z = z$ have resistance $U_D \leq P(z)$ and therefore inform us about the average of Y_1 for people with U_D between 0 and 0.4, depicted by the short dash dot line. Likewise, the people with $Z = z'$ inform us about the average of Y_1 from 0 to 0.8 (dash dot line) and the average of Y_0 from 0.8 to 1 (long dash line). With a continuous instrument, we will have propensity scores all over the $(0, 1)$ interval, and we can identify nonlinear functions in u, but this illustration shows that a binary instrument identifies a linear MTE model. We need only four averages from the groups composed of the treated and untreated individuals with the instrument switched on and off (Brinch, Mogstad, and Wiswall 2017).

In practice, this means specifying some function for the conditional expectations of the error terms and then estimating the conditional expectations of Y_1 and Y_0 in the sample of treated and untreated separately:

$$\mathbb{E}(Y_1|X = x, D = 1) = x\beta_1 + \mathbb{E}(U_1|U_D \leq p) = x\beta_1 + K_1(p) \tag{5}$$
$$\mathbb{E}(Y_0|X = x, D = 0) = x\beta_0 + \mathbb{E}(U_0|U_D > p) = x\beta_0 + K_0(p) \tag{6}$$

I follow the notation in Brinch, Mogstad, and Wiswall (2017) and control selection via the control functions $K_j(p)$. Note that this is the specification we are using when fitting selection models; depending on the specification of $K_j(p)$, this amounts to, for example, Heckman selection or a semiparametric selection model.

To estimate the MTE, we estimate the conditional expectation of Y in the sample of treated and untreated separately using the regression

$$Y_j = X\beta_j + K_j(p) + \epsilon$$

Based on the assumptions on the unknown functions $k_j(u)$, we can infer the functional form of the control function $K_j(p)$ as summarized in table 1. Alternatively, when using semiparametric methods, estimate (5)–(6) using partially linear models as summarized in table 2. In the implementation, this is estimated in a stacked regression to allow for some coefficients to be restricted to be the same in the treated and untreated state. After estimating $K_j(p)$, we can construct the $k_j(u)$ functions and find the MTE estimate from

$$\begin{aligned} \text{MTE}(x, u) &= \mathbb{E}(Y_1|X = x, U_D = u) - \mathbb{E}(Y_0|X = x, U_D = u) \\ &= x(\beta_1 - \beta_0) + k_1(u) - k_0(u) \\ \text{where} \quad k_j(u) &= \mathbb{E}(U_j|U_D = u) \end{aligned}$$

Table 1. Parametric MTE models

Function	Definition	Normal	Polynomial	Polynomial with splines
		$U_0, U_1, V \sim \mathcal{N}(0, \mathbf{\Sigma})$, $\Sigma = \begin{Bmatrix} \sigma_1^2 & & \\ \rho_{01} & \sigma_0^2 & \\ \rho_0 & \rho_1 & 1 \end{Bmatrix}$	$k(u)$ or $k_j(u)$ as Lth-order polynomials with mean 0	$k(u)$ or $k_j(u)$ as Lth-order polynomials ($L \geq 2$) with Q knots for quadratic and higher-order terms at $(h_1, \ldots, h_Q)$ and mean 0
$k(u)$	$\mathbb{E}(U_1 - U_0 \vert U_D = u)$	$(\rho_1 - \rho_0)\Phi^{-1}(u)$	$\sum_{l=1}^{L} \pi_l \left(u^l - \frac{1}{l+1}\right)$	$\sum_{l=1}^{L} \pi_l \left(u^l - \frac{1}{l+1}\right) + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_k^q \left\{(u - h_q)^l \, \mathbb{1}(u \geq h_q) - \frac{(1-h_q)^{l+1}}{(l+1)}\right\}$
$K(p)$	$p\mathbb{E}(U_1 - U_0 \vert U_D \leq p)$	$-(\rho_1 - \rho_0)\varphi\left\{\Phi^{-1}(p)\right\}$	$\sum_{l=1}^{L} \pi_l \frac{p(p^l - 1)}{l+1}$	$\sum_{l=1}^{L} \pi_l \frac{p(p^l+1)}{l+1} + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_k^q \left\{\frac{\mathbb{1}(p \geq h_q)(p-h_q)^{l+1} - (1-h_q)^{l+1} p}{l+1}\right\}$
$k_1(u)$	$\mathbb{E}(U_1 \vert U_D = u)$	$\rho_1 \Phi^{-1}(u)$	$\sum_{l=1}^{L} \pi_{1l} \left(u^l - \frac{1}{l+1}\right)$	$\sum_{l=1}^{L} \pi_{1l} \left(u^l - \frac{1}{l+1}\right) + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_{1l}^q \left\{\mathbb{1}(u \geq h_q)(u - h_q)^l - \frac{(1-h_q)^{l+1}}{(l+1)}\right\}$
$k_0(u)$	$\mathbb{E}(U_0 \vert U_D = u)$	$\rho_0 \Phi^{-1}(u)$	$\sum_{l=1}^{L} \pi_{0l} (u^l - \frac{1}{l+1})$	$\sum_{l=1}^{L} \pi_{0l} (u^l - \frac{1}{l+1}) + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_{0l}^q \left\{\mathbb{1}(u \geq h_q)(u - h_q)^l - \frac{(1-h_q)^{l+1}}{(l+1)}\right\}$
$K_1(p)$	$\mathbb{E}(U_1 \vert U_D \leq p)$	$-\rho_1 \frac{\varphi\left\{\Phi^{-1}(p)\right\}}{p}$	$\sum_{l=1}^{L} \pi_{1l} \frac{p^l - 1}{l+1}$	$\sum_{l=1}^{L} \pi_{1l} \frac{(p^l+1)}{l+1} + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_{1l}^q \left\{\frac{\mathbb{1}(p \geq h_q)(p-h_q)^{l+1} - (1-h_q)^{l+1}}{l+1}\right\}$
$K_0(p)$	$\mathbb{E}(U_0 \vert U_D > p)$	$\rho_0 \frac{\varphi\left\{\Phi^{-1}(p)\right\}}{(1-p)}$	$\sum_{l=1}^{L} \pi_{0l} \frac{p(1-p^l)}{(1-p)(l+1)}$	$\sum_{l=1}^{L} \pi_{0l} \frac{p(1-p^l)}{(1-p)(l+1)} + \sum_{l=2}^{L} \sum_{q=1}^{Q} \pi_{0l}^q \left\{\frac{(1-h_q)^{l+1} p - \mathbb{1}(p \geq h_q)(p-h_q)^{l+1}}{(1-p)(l+1)}\right\}$
MTE(x, u)	$\mathbb{E}(Y_1 - Y_0 \vert U_D = u, X = x)$	$x(\beta_1 - \beta_0) + k(u) = x(\beta_1 - \beta_0) + k_1(u) - k_0(u)$		
$Y_1(x, u)$	$\mathbb{E}(Y_1 \vert U_D = u, X = x)$	$x\beta_1 + k_1(u)$		
$Y_0(x, u)$	$\mathbb{E}(Y_0 \vert U_D = u, X = x)$	$x\beta_0 + k_0(u)$		

Note: Expressions for conditional expectations with different assumptions for the joint distribution of the error terms. $\mathbb{1}(A)$ is the indicator function for the event A. Note that $\pi_l = \pi_{1l} - \pi_{0l}$ and equivalently for the spline coefficients. Calculated as $K(p) = \int_0^p k(u)du$, $K_1(p) = 1/p \int_0^p k_1(u)du$ and $K_0(p) = 1/(1-p) \int_p^1 k_0(u)du$.

Table 2. Semiparametric MTE models

Step	Local IV	Separate approach
Estimating equation	$Y = X\beta_0 + X(\beta_1 - \beta_0) + K(p) + \epsilon$	$Y_j = X\beta_j + K_j(p) + \epsilon$
Double residual regression (Robinson 1988)	local polynomial regressions of Y, X, and $X \times p$ give residuals e_Y, e_X, and $e_{X\times P}$	separate local polynomial regressions of Y and X on p in treated and untreated samples give residuals e_{Y_j} and e_{X_j}, construct $e_Y = De_{Y_1} + (1-D)e_{Y_0}$ and similar for e_X
Estimate β_0, $\beta_1 - \beta_0$ using regression	$e_Y = e_X\beta_0 + e_{X\times p}(\beta_1 - \beta_0) + \epsilon$	$e_Y = e_X\beta_0 + D(\beta_1 - \beta_0)e_X + \epsilon$
Construct residual	$\widetilde{Y} = Y - X\widehat{\beta_0} - X\left(\widehat{\beta_1 - \beta_0}\right)p$	$\widetilde{Y} = Y - X\widehat{\beta_0} - X\left(\widehat{\beta_1 - \beta_0}\right)D$
Estimate K	local polynomial regression of $\widetilde{Y}$ on p, saving level $\widehat{K(p)}$ and slope $\widehat{K'(p)}$	separate local polynomial regressions of $\widetilde{Y}$ on p in treated and untreated samples, saving level $\widehat{K_j(p)}$ and slope $\widehat{K_j'(p)}$
Construct k	$\widehat{k(u)} = \widehat{K'(p)}$	$\widehat{k_1(u)} = \widehat{K_1(p)} + p\widehat{K_1'(p)}$ $\widehat{k_0(u)} = \widehat{K_0(p)} - (1-p)\widehat{K_0'(p)}$
Construct MTE	$\widehat{\text{MTE}}(x,u) = x\left(\widehat{\beta_1 - \beta_0}\right) + \widehat{k(u)}$	$\widehat{\text{MTE}}(x,u) = x\left(\widehat{\beta_1 - \beta_0}\right) + \widehat{k_1(u)} - \widehat{k_0(u)}$

Note: Steps in the estimation of semiparametric MTE models using local IVs or the separate approach. To see the relation between $k_j(u)$ and $K_j(p)$, note that $K_1(p) = \mathbb{E}(U_1|U_D \le p) = 1/p\int_0^p \mathbb{E}(U_1|U_D = u)du \Rightarrow K_1'(p) = -(1/p)K_1(p) + (1/p)k_1(u)$, which leads to $k_1(u) = K_1(p) + pK_1^{'}(p)$. We can find similar expressions for $k_0(u)$.

In principle, it is possible to combine the semiparametric and the polynomial approach by first estimating the β coefficients from the polynomial model and then using semiparametric methods to find K. This is the semiparametric polynomial MTE model, implemented if `polynomial()` and `semiparametric` are specified together in `mtefe`. Although computationally far less complex, there is little theory to think that the semiparametric estimate of this model should be any better than the MTE constructed from the parametric estimates.

2.3 Estimation using maximum likelihood

In the case where we assume joint normality of (U_0, U_1, V), the MTE can be estimated using maximum likelihood like a Heckman selection model. It is straightforward to infer the log-likelihood function. For details, see appendix A.4 or Lokshin and Sajaia (2004).

2.4 Treatment-effect parameters and weights

MTEs unify most common treatment-effect parameters and allow us to recover the parameters from the MTEs. Given that we have estimated $\text{MTE}(x, u)$, we need only estimates of the distribution of U_D given x in a particular population to obtain estimates of the ATE for that population.

In practice, common treatment-effect parameters can be expressed as some weighted average of a particular MTE curve (Heckman and Vytlacil 2007). For the ATE conditional on the event $A = a$ that defines the population relevant for the parameter of interest and $X = x$, we have

$$T_a(x) = \mathbb{E}(Y_1 - Y_0 | A = a, X = x) = \int_0^1 \text{MTE}(x, u_D)\omega_a(x, u)du_D$$
$$\text{where } \omega_a(x, u) = f_{U_D|A=a,X=x}(u)$$

where f is the conditional density of U_D. Because of the additive separability given by Assumption 2 and the linear forms of $\mu_j(X)$, this simplifies to

$$T_a(x) = x(\beta_1 - \beta_0) + \int_0^1 k(u_D)\omega_a(u)du_D$$
$$\text{where } \omega_a(u) = f_{U_D|A=a}(u)$$

In principle, we can calculate the treatment-effect parameters $T_a(x)$ for any value of x, but we are usually interested in the unconditional parameter in the population. In general, we need to integrate over all values of X,[6] but because of the additive separability implied by Assumption 2, we can estimate the average x in the population of interest separately and calculate the unconditional treatment-effect parameters as

$$T_a = \mathbb{E}(Y_1 - Y_0 | A = a) = x_a(\beta_1 - \beta_0) + \int_0^1 k(u_D)\omega_a(u)du_D$$
$$\text{where } \omega_a(u) = f_{U_D|A=a}(u)$$
$$\text{and } x_a = \mathbb{E}(X|A = a)$$

We can estimate x_a using the weighted average $1/N \sum \kappa_i^a x_i$, where κ_i^a is an estimate of the relative probability that event a happens to person i, $\{P(A = a|Z = z_i)\}/\{P(A = a)\}$. In practice, we can estimate both the weighted average x_a and $\omega_a(u)$ by using sample analogs from data on p, Z, and D as summarized in table 3.

6. Note, however, that in cases where the propensity score model is misspecified, the weighted average of the conditional weights may differ from the unconditional weights. See, for example, Carneiro, Heckman, and Vytlacil (2011) and Carneiro, Lokshin, and Umapathi (2017).

Table 3. Unconditional treatment-effect parameters and weights

	Parameter	Event A	$\widehat{\kappa_i}$	$\widehat{\omega}(u)$
ATE	ATE	1	1	$\frac{1}{s}$
ATT	ATT	$U_D \leq p$	$\frac{p_i}{\mathbb{E}(p)}$	$\frac{P(p>u)}{s\mathbb{E}(p)}$
ATUT	ATUT	$U_D > p$	$\frac{1-p_i}{1-\mathbb{E}(p)}$	$\frac{1-P(p>u)}{s\{1-\mathbb{E}(p)\}}$
LATE	Local ATE	$P(z) < U_D \leq P(z')$	$\frac{p(z_i)-p(z_i')}{\mathbb{E}(p')-\mathbb{E}(p)}$	$\frac{P\{p(z')>u\}-P\{p(z)>u\}}{s(\overline{p'}-\overline{p})}$
2SLS	Weighted average LATE		$\frac{\{\widehat{v}_i-\mathbb{E}(\widehat{v})\}(D_i-\overline{D})}{\mathrm{cov}(D,\widehat{v})}$	$\frac{\{\mathbb{E}(\widehat{v}\|p>u)-\mathbb{E}(\widehat{v})\}P(p>u)}{s\times\mathrm{cov}(D,\widehat{v})}$
PRTE	Policy-relevant ATE	$p < U_D \leq p'$	$\frac{p_i'-p_i}{\mathbb{E}(p')-\mathbb{E}(p)}$	$\frac{P(p'>u)-P(p>u)}{s\{\mathbb{E}(p')-\mathbb{E}(p)\}}$
MPRTE1	Marginal PRTE 1	$\lvert\gamma Z - V\rvert < \epsilon$	$\frac{f_V(\gamma Z)}{\mathbb{E}\{f_V(\gamma Z)\}}$	$\frac{f_p(u)f_V\{F_V^{-1}(u)\}}{\mathbb{E}\{f_V(\gamma Z)\}}$
MPRTE2	Marginal PRTE 2	$\lvert p - U\rvert < \epsilon$	1	$f_p(u)$
MPRTE3	Marginal PRTE 3	$\lvert\frac{p}{U} - 1\rvert < \epsilon$	1	$\frac{uf_p(u)}{\mathbb{E}(p)}$

Note: Weights for common treatment effects. Discrete distribution of U_D with s points.

The ATT is the average effect of treatment for the subpopulation that chooses treatment. When one calculates the weighted average of the X in this population, this parameter will weight people with high propensity scores more precisely because they have a higher probability of choosing treatment. Likewise, the weights ω_{ATT} will weight points at the lower end of the U_D distribution higher because a larger share of the population at these values of U_D will choose treatment—they have lower resistance.

In contrast, the ATUT weights individuals with low propensity scores higher when calculating the weighted average of the X. This is precisely because these people, all else the same, have higher probability to be untreated. ω_{ATUT} weights high values of U_D more because these people have high resistance.

A LATE is the average effect of treatment for people who are shifted into (or out of) treatment when the instrument is shifted from z to z'. These are people with U_D in the interval $(P(z), P(z'))$. To be in the complier group, an individual must have resistance between $P(z)$ and $P(z')$. Note how, when $z - z'$ is infinitesimally small, the LATE converges to the MTE, and an MTE is thus a limit form of LATE (Heckman and Vytlacil 2007).

Similarly to the way we can estimate the probability of being a complier in a traditional IV analysis, we can estimate the weights of the linear IV or two-stage least squares (2SLS) parameter. With a continuous instrument, IV is a weighted average over all possible LATEs composed of all $z-z'$ pairs (Angrist and Imbens 1995). Heckman and Vytlacil (2007) show the derivation of these weights; see also section A.6 in the appendix and Cornelissen et al. (2016). In sum, the IV parameter uses a weighted average of X, in which people with large positive or negative values of $\widehat{v}$, a measure of how much the instrument affects propensity scores for each individual, are given more weight. These are precisely the people who have higher probabilities of having their treatment status determined by the instrument. Likewise, values of U_D where people have $\widehat{v}$ above the average, and thus are more likely to be compliers, get a higher weight $\omega_{\text{IV}}(u)$.

Assuming policy invariance (see Heckman and Vytlacil [2007] for a formal definition), we can calculate the PRTE for a counterfactual policy that manipulates propensity scores. This is the expected treatment effect for the people that are shifted into treatment by the new policy relative to the baseline. If the policy is a shift in the instruments themselves, it is natural to use the estimated first stage to evaluate the shift in propensity scores, but propensity scores could be manipulated directly as well. Note that if the policy is a particular set of instrument values z', and the baseline is another set z, the PRTE and LATE are the same. In practice, the PRTE parameter weights the treatment effect of people who are affected more strongly by the alternative policy relative to the baseline.

Lastly, we can use the estimated MTEs to calculate marginal policy relevant treatment effects (MPRTEs), which can be interpreted as average effects of making marginal shifts to the propensity scores. MPRTEs are fundamentally easier to identify than PRTEs (Carneiro, Heckman, and Vytlacil 2010), particularly because they do not require full support; marginal changes to propensity scores will not drive the scores outside the common support.

Carneiro, Heckman, and Vytlacil (2011) suggest three ways to define distance to the margin. The first MPRTE, labeled MPRTE1 in table 3, defines the distance in terms of the differences between the index γZ and the resistance V and corresponds to a marginal change in a variable entering the first stage such as an instrument. MPRTE2 defines the margin as having propensity scores close to the normalized resistance U_D and corresponds to a policy that would increase all propensity scores with a small amount. MPRTE3 defines marginal as the relative distance between the propensity score and U_D and corresponds to a policy that increases all propensity scores by a small fraction.

What we can see from the expressions in table 3 is that an absolute shift in propensity scores uses the observed density of the propensity scores as the weight distribution, while a relative shift will place more weight on the upper part of the U_D distribution precisely because a relative shift increases the propensity scores of people with high initial propensity scores more.

3 The mtefe package

3.1 Syntax

As outlined in the introduction, `mtefe` contains many improvements over earlier MTE packages such as `margte`. The basic syntax of the `mtefe` command mimics that of Stata's `ivregress` and other IV estimators, while the syntax for many of its options resemble the same options in `margte`. All independent variable lists also accept Stata's categorical variables syntax `i.`*varname*. `fweight`s and `pweight`s are supported.

`mtefe` *depvar* [*indepvars*] (*depvar*$_t$ `=` *varlist*$_{iv}$) [*if*] [*in*] [*weight*] [`,`
`polynomial(`#`)` `splines(`*numlist*`)` `semiparametric` `restricted(`*varlist*`)`
`link(`*string*`)` `separate` `mlikelihood` `trimsupport(`#`)` `fullsupport`
`prte(`*varname*`)` `bootreps(`#`)` `norepeat1` `vce(`*vcetype*`)` `level(`#`)` `degree(`#`)`
`ybwidth(`#`)` `xbwidth(`#`)` `gridpoints(`#`)` `kernel(`*string*`)` `first` `second`
`noplot` `omit(`*varlist*`)` `savefirst(`*string*`)` `savepropensity(`*newvar*`)` `savekp`
`saveweights(`*string*`)` `mtexs(`*matrixlist*`)`]

3.2 Options

The most important options that determine which model is fit and how are the following:

`polynomial(`#`)` specifies polynomial MTE models of degree #. In contrast to `margte`, to ensure consistency between estimation procedures, this is the degree of $k(u)$ and in turn the MTE, not the degree of $K(p)$. Although most restrictive, to ensure consistency with `margte`, the default if `polynomial()` is not specified is the joint normal model.

`splines(`*numlist*`)` adds second-order and higher splines to $k(u)$ at the points specified by *numlist*. Use this only for polynomial models of degree ≥ 2. All points in *numlist* must be in the interval $(0, 1)$.

`semiparametric` fits $K(p)$ or $K_j(p)$ using semiparametric methods as described in table 2. This amounts to the semiparametric model if `polynomial` is not specified or the semiparametric polynomial otherwise.

`restricted(`*varlist*`)` specifies that variables in *varlist* be included in both the first and second stages but are restricted to have the same effect in the two states.

`link(`*string*`)` specifies the first-stage model. *string* may be `probit`, `logit`, or `lpm`. The default is `link(probit)`.

`separate` fits the model using the separate approach rather than local IVs.

`mlikelihood` fits the model using the maximum likelihood rather than local IVs. This is appropriate only for the joint normal model.

The command allows many other options, for which I refer the reader to the help file. The `mtefe` package additionally contains the command `mtefeplot`, which plots one or more MTE plots—optionally including treatment parameter weights—based on stored or saved MTE estimates from `mtefe`. The command `mtefe_gendata` generates the data in the following examples and Monte Carlo simulations.

By default, `mtefe` reports analytical standard errors for the coefficients, treatment-effect parameters, and MTEs. These ignore the uncertainty in the estimation of the propensity scores by treating these as fixed in the second stage of the estimation as well as the means of X and the treatment-effect parameter weights that are used for estimating treatment effects. For matching, Abadie and Imbens (2016) show that ignoring the uncertainty in the propensity score increases the standard errors for the ATE, while the impact on other treatment-effect parameters is ambiguous. To the best of my knowledge, we do not know how this omission affects the standard errors in MTE applications, so careful researchers should therefore bootstrap the standard errors using the `bootreps()` option, which reestimates the propensity scores, the mean of X, and the treatment-effect parameter weights for each bootstrap repetition.

3.3 Example output from mtefe

To illustrate the use of `mtefe` and for Monte Carlo simulations in appendix B, let's imagine the following problem: We are interested in the monetary returns to a college education. Unfortunately, college education is endogenous, for example, because higher-ability people do better in the labor market and have more education. Furthermore, people choose education based partly on knowledge about their own gains from college. Thus, the problem exhibits selection on both levels and gains.

Consider distance to college, thought to be a cost shifter for college education. Although traditional in the returns to education literature, there are reasons to doubt the exclusion restriction for this instrument (Carneiro and Heckman 2002). To fix ideas,

suppose that the average distance to college in a district, a measure of rurality, is correlated with average outcomes in the labor market. For example, more rural labor markets could provide worse average employment opportunities, particularly for college jobs. However, if these differences work at the district level, the instrument is valid, conditional on district fixed effects that control for the average variation in distance to college. Thus, the remaining within-district variation in distance to college is a valid instrument for college attendance.

To implement this thought experiment, I draw the average labor market quality for college and noncollege jobs and the average distance to college for each district from a joint normal distribution. The observed distance to college is equal to this district-level average plus some random normal variation, so the within-district variation in distance is a valid instrument. In addition, I generate the error terms U_0, U_1, and V from either a joint normal or a polynomial error structure, where the three errors are correlated and thus generate selection on both levels and gains. Controls X include experience uniformly distributed on $(0, 30)$ and its square. These affect both the selection equation and the outcomes in the two states. The full data-generating process is described in appendix B.

```
. set seed 1234567
. mtefe_gendata, obs(10000) districts(10)
. mtefe lwage exp exp2 i.district (col=distCol)
Parametric normal MTE model                              Observations : 10000
Treatment model: Probit
Estimation method: Local IV
```

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
beta0						
exp	.0278374	.006707	4.15	0.000	.0146903	.0409844
exp2	-.0004643	.0002097	-2.21	0.027	-.0008753	-.0000534
district						
2	-.3713221	.061335	-6.05	0.000	-.4915511	-.2510932
3	-.0308218	.06261	-0.49	0.623	-.15355	.0919065
4	.6884756	.0866584	7.94	0.000	.5186077	.8583435
5	.0262556	.0642906	0.41	0.683	-.0997671	.1522782
6	.4946678	.0633903	7.80	0.000	.37041	.6189255
7	-.2666547	.0613036	-4.35	0.000	-.386822	-.1464873
8	.2437004	.0580991	4.19	0.000	.1298145	.3575864
9	.115046	.0598625	1.92	0.055	-.0022965	.2323885
10	.027915	.0602922	0.46	0.643	-.0902699	.1460999
_cons	3.105918	.0659863	47.07	0.000	2.976572	3.235265

beta1-beta0						
exp	-.0231842	.0100974	-2.30	0.022	-.0429771	-.0033914
exp2	.0006458	.0003238	1.99	0.046	.0000111	.0012804
district						
2	.0529143	.1010226	0.52	0.600	-.1451104	.250939
3	.0531692	.1038045	0.51	0.609	-.1503086	.256647
4	-.0521345	.1184164	-0.44	0.660	-.2842546	.1799855
5	.1319906	.1028638	1.28	0.199	-.0696432	.3336243
6	.0141403	.1050543	0.13	0.893	-.1917872	.2200679
7	.0021696	.1019711	0.02	0.983	-.1977144	.2020536
8	-.4022728	.1008029	-3.99	0.000	-.5998667	-.2046788
9	-.2616856	.1026346	-2.55	0.011	-.4628702	-.0605009
10	-.2654738	.1029734	-2.58	0.010	-.4673225	-.063625
_cons	.5889556	.0967044	6.09	0.000	.3993954	.7785157
k						
mills	-.5492868	.0595307	-9.23	0.000	-.665979	-.4325946
effects						
ate	.3627266	.0239355	15.15	0.000	.3158082	.409645
att	.6166828	.0388462	15.87	0.000	.5405364	.6928291
atut	.0735471	.036977	1.99	0.047	.0010647	.1460295
late	.340083	.0238033	14.29	0.000	.2934237	.3867424
mprte1	.3567228	.0249318	14.31	0.000	.3078514	.4055941
mprte2	.3133149	.0239501	13.08	0.000	.2663677	.360262
mprte3	-.0568992	.0483759	-1.18	0.240	-.1517257	.0379273

Test of observable heterogeneity, p-value	0.0000
Test of essential heterogeneity, p-value	0.0000

Note: Analytical standard errors ignore the facts that the propensity score, the mean of X and the treatment effect parameter weights are estimated objects when calculating standard errors. Consider using bootreps() to bootstrap the standard errors.

Using the data-generating process outlined above, the code uses the commands `mtefe` and `mtefe_gendata` to generate data with a normal error structure and estimate them using the joint normal MTE model and local IVs. `mtefe` first reports the estimated coefficients β_0, $\beta_1 - \beta_0$, and $\rho_1 - \rho_0$; then it reports the treatment-effect parameters as shown in the output. Lastly, `mtefe` reports the p-values for two statistical tests: a joint test of the $\beta_1 - \beta_0$, which can be interpreted as a test of whether the treatment effect differs across X, and a test of essential heterogeneity. The latter is a joint test of all coefficients in $k(u)$.[7]

7. Or, in the case of semiparametric models, a test of whether all MTEs are the same.

Based on the output from `mtefe`, we can straightforwardly evaluate the impact of covariates. Average differences in outcomes across covariates can be interpreted directly from the β_0, just like a regular control variable. For instance, the coefficient for experience in the first panel of the output table indicates that one more year of experience translates into approximately 2.8% higher wages, although the effect is nonlinear, and we cannot say that it is the extra experience that causes the higher wages without strong exogeneity assumptions on X.

Similarly, the $\beta_1 - \beta_0$ can be interpreted as differences in treatment effects across covariate values, just like an interaction between treatment status and a covariate in an ordinary least-squares regression. The coefficient on experience in the second panel of the output table thus indicates that a person with one more year of experience has 2.3% lower gains from college, but again, we cannot give this a causal interpretation, and we must account for the nonlinear effect.

In addition, `mtefe` plots the MTE curve with associated confidence intervals as well as the density of the estimated propensity scores separately for the treated and untreated individuals so that the researcher can evaluate the common support. Examples of these plots are found in figures 2a and 2b. We see that the estimated MTE in this example is downward sloping, with relatively high treatment effects above 1 at the beginning of the U_D distribution, eventually declining to negative effects below -0.5 at the right end of the distribution. This implies an ATE of around 0.36, and the downward sloping pattern is consistent with positive selection on unobservable gains as predicted by a simple Roy model.

Alternatively, we can use the polynomial MTE model and the separate estimation approach or the semiparametric model to relax the joint normal assumption. `mtefe` fits these models if you specify the `polynomial(2)` or the `semiparametric` option,[8] respectively. When one fits MTE models, it is useful to plot several MTE curves in the same figure. This can be done using `mtefeplot`, specifying the names of the saved or stored MTE estimates. An example of this plot is provided in figure 2c for the normal, polynomial, and semiparametric MTE models. MTEs are downward sloping and relatively similar in all three specifications.

As discussed in section 2.4, `mtefe` also estimates the treatment parameter weights and the parameters themselves. One way of illustrating why, for example, the LATE[9] differs from the ATE is by depicting the weights that LATEs put on different parts of the X and U_D distribution. This can be done using `mtefeplot` with the `late` option. The resulting plot is found in figure 2e. Because the MTE curve at the average of X and the MTE curve for compliers practically overlap, it does not seem like the people induced to enroll because of the instrument have different values of X—this is hardly surprising given the data-generating process. Instead, the weight distribution reveals

8. For semiparametric models, the option `gridpoints()` will greatly improve computational speed by performing the first local linear regression at # points equally spaced over the support of p rather than at each and every observed value of p in the population.
9. The term is used somewhat ambiguously here to refer to the linear IV estimate as a weighted average over all possible LATEs from all combinations of two values of the instrument, in line with much of the literature using continuous instruments.

that compliers have a much higher probability to have unobserved resistance in the middle part of the distribution. These people have MTEs slightly below the average, so when taking a weighted average over the MTE curve for compliers, we find that the LATE is lower than the ATE. The compliers to the instrument are people with slightly below-average gains from college.

Furthermore, we can exploit the fact that when using the separate estimation procedure, we identify both $k_1(u)$ and $k_0(u)$. After first fitting the polynomial model using the `separate` option to use the separate approach rather than local IVs, we can plot the resulting potential outcomes using the `separate` option of `mtefeplot`. Because the MTE is simply the difference between Y_1 and Y_0, we can investigate whether the downward sloping trend in the MTE is generated by upward sloping Y_0, downward sloping Y_1, or a combination. From figure 2d, we see that Y_0 is relatively flat, while Y_1 is clearly downward sloping. This indicates that people who have low resistance to treatment do much better than their high-resistance counterparts with a college degree, but relatively similar without. Therefore, these people have higher effects of treatment.

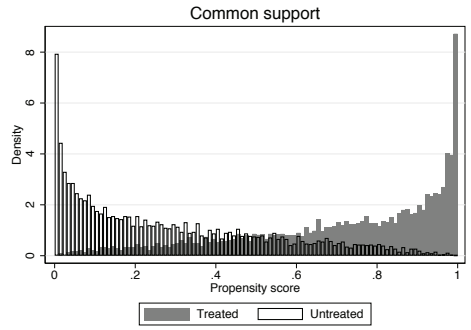
As a last example, consider a hypothetical policy that mandates a maximum distance of 40 miles to the closest college. To estimate the effect of such a policy, we predict the propensity scores from the probit model using the adjusted distance to college:

```
. qui probit col distCol exp exp2 i.district
. generate temp=distCol
. replace distCol=40 if distCol>40
. predict double p_policy
. replace distCol=temp
. mtefe lwage exp exp2 i.district (col=distCol), pol(2) prte(p_policy)
```

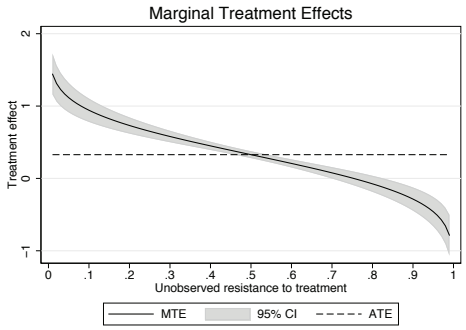
(*output omitted*)

The results of this exercise are depicted in figure 2f. First, note how the MTE curve for the policy compliers practically overlaps the MTE curve at the mean; policy compliers do not seem to have X-values that give them different gains from college than the average. Next, notice the distribution of weights; policy compliers come exclusively from the lower part of the U_D distribution. This is not surprising, given that the people most affected by the reform are people with low propensity scores (driven by high distance to college) before the reform. To be affected by the reform, these people must be untreated under the baseline and thus have U_D's above those low propensity scores. Compared with the people with higher propensity scores, the average U_D for these people are relatively low, generating high weights on the lower part of the U_D distribution. Thus, the potential policy compliers have low U_D's and, subsequently, high treatment effects. Therefore, the expected gain from the reform is larger than the ATE.

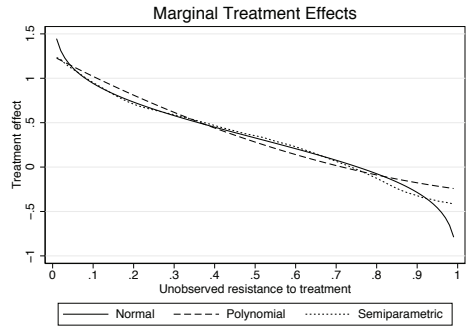
```
. mtefe lwage exp exp2 i.district (col=distCol)
. estimates store normal
. mtefe lwage exp exp2 i.district (col=distCol), polynomial(2) separate
. estimates store polynomial
. mtefe lwage exp exp2 i.district (col=distCol), semiparametric gridpoints(100)
. estimates store semipar
. mtefeplot normal polynomial semipar, memory
> names("Normal" "Polynomial" "Semiparametric")
. mtefeplot polynomial, separate memory
. mtefeplot polynomial, late memory
```



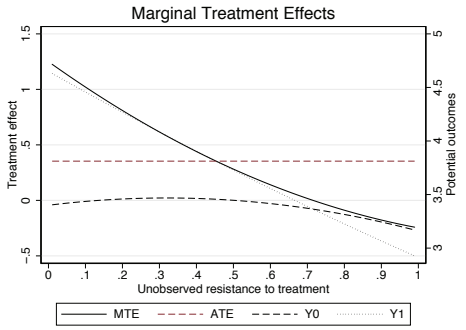
(a) Common support plot, probit



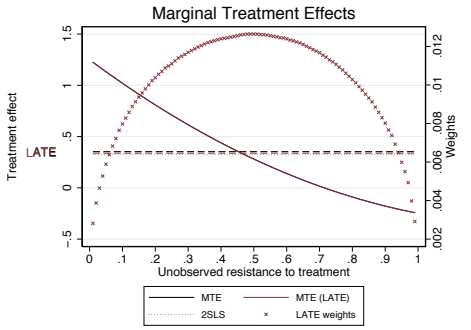
(b) Estimated MTE curve, normal



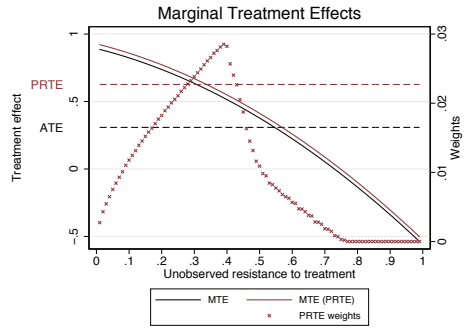
(c) MTE estimates, different models



(d) MTEs and potential outcomes, polynomial



(e) MTE for compliers with LATE weights



(f) Estimated effects of a hypothetical policy

Figure 2. Example figures from `mtefe` and `mtefeplot`

Lastly, we might be interested in the pattern of selection on observable gains. Is it the case that covariates that positively impact treatment choices also positively impact the gains from treatment? It is if $\gamma \times (\beta_1 - \beta_0) > 0$, which happens either if both coefficients are positive or if both are negative. As an example, there is positive selection on the covariate experience in the example reported in figure 2; more experience is associated both with less college (not shown) and with lower treatment effects.

3.4 Postestimation using mtefe

Because `mtefe` saves results in `e()`, estimates are readily available for postestimation. As an example, expected treatment effects can be predicted for every individual with characteristics X, D, and p, using that

$$\begin{aligned}\mathbb{E}(Y_1 - Y_0|X, D, p) \\ &= x(\beta_1 - \beta_0) + D\mathbb{E}(U_1 - U_0|U_D \leq p) + (1 - D)\mathbb{E}(U_1 - U_0|U_D > p) \\ &= x(\beta_1 - \beta_0) + \frac{D - p}{p(1 - p)}K(p)\end{aligned}$$

where I use the fact that both U_1 and U_0 are normalized to have mean zero. Because all these objects are estimated by `mtefe`, we can predict treatment effects for each individual. Use options `savekp` and `savepropensity()` to save the propensity scores and the relevant variables of the $K(p)$ function, then estimate expected treatment effects from the expression above. The resulting variable contains the expected treatment effect given treatment status, propensity scores, and X for each individual. Summarizing this predicted treatment effect among the treated and untreated separately closely matches the ATT and ATUT, respectively.

This procedure highlights the relationship between selection models and MTE—but note how both the ATT and ATUT could be recovered without specifying $K_j(p)$, only the expected difference $K(p)$.

When we use the separate approach, potential outcomes can be predicted directly because both $K(p)$ and $K_0(p)$ are estimated:

$$\begin{aligned}\mathbb{E}(Y_1|X, D, p) &= x\beta_1 + \frac{D - p}{1 - p}K_1(p) \\ \mathbb{E}(Y_0|X, D, p) &= x\beta_0 + \frac{p - D}{p}K_0(p) \\ \text{where } K_1(p) &= \frac{K(p) - (1 - p)K_0(p)}{p}\end{aligned}$$

These expressions can be calculated by predicting $K_j(p)$ and constructing the expected value of the outcome for each individual from the expressions above. The difference between these predicted outcomes closely matches the predicted treatment effect from above and treatment-effect parameters calculated by `mtefe` as weighted averages over the appropriate MTE curves.

4 Conclusion

MTEs are increasingly becoming a part of the toolbox of the applied econometrician when the problem at hand exhibits selection on both levels and unobserved gains. In contrast to traditional linear IV analysis, MTEs uncover the distribution of treatment effects rather than a LATE, which is often of little interest. In practice, this typically comes at the cost of stricter assumptions.

In this article, I outlined the MTEs framework, the relationship to traditional selection models, and three different methods for fitting these models. The documented package, `mtefe`, contains many improvements over existing packages. Among these are support for fixed effects, estimation using both local IVs, the separate approach and maximum likelihood, a larger number of parametric and semiparametric MTE models, support for weights, speed improvements when running semiparametric models, more appropriate bootstrap inference, and improved graphical output. In addition, `mtefe` calculates common treatment-effect parameters such as LATEs, ATT and ATUT, and PRTEs for a user-specified shift in propensity scores to allow the user to exploit the potential of the MTEs framework in understanding treatment-effect heterogeneity.

The Monte Carlo simulations detailed in appendix B show that MTE estimation can be sensitive to function form specifications and that wrong functional form assumptions may result in too high rates of rejection. Note that the two data-generating processes used and the functional form choices made in these simulations are very restrictive, and it is perhaps not surprising that a second-order polynomial cannot approximate a normal very well or the other way around. Nonetheless, they illustrate two things: First, applied researchers should base the choice of functional form on detailed knowledge about the case at hand—what might plausibly constitute the unobservable dimension and economic arguments for how these factors might affect the outcome. Second, researchers should strive to probe the robustness of their results to the functional form choices, using less restrictive models to guide the choice of specification.

5 Acknowledgments

The author wishes to thank Edwin Leuven, Christian Brinch, Thomas Cornelissen, Martin Huber, Katrine Vetlesen Løken, Simen Markussen, Scott Brave, Thomas Walstrum, Yiannis Spanos, Jonathan Norris, and Magne Mogstad for advice, comments, help, and testing at various stages of this project. Brave and Walstrum also deserve many thanks for the package `margte`, which clearly inspired `mtefe`. All errors remain my own. While carrying out this research, I have been associated with the Centre for the Study of Equality, Social Organization, and Performance at the Department of Economics at the University of Oslo. The Centre for the Study of Equality, Social Organization, and Performance is supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 179552.

6 References

Abadie, A., and G. W. Imbens. 2016. Matching on the estimated propensity score. *Econometrica* 84: 781–807.

Angrist, J. D., and G. W. Imbens. 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association* 90: 431–442.

Björklund, A., and R. Moffitt. 1987. The estimation of wage gains and welfare gains in self-selection models. *Review of Economics and Statistics* 69: 42–49.

Brave, S., and T. Walstrum. 2014. Estimating marginal treatment effects using parametric and semiparametric methods. *Stata Journal* 14: 191–217.

Brinch, C. N., M. Mogstad, and M. Wiswall. 2017. Beyond LATE with a discrete instrument. *Journal of Political Economy* 125: 985–1039.

Carneiro, P., and J. J. Heckman. 2002. The evidence on credit constraints in post-secondary schooling. *Economic Journal* 112: 705–734.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil. 2010. Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica* 78: 377–394.

———. 2011. Estimating marginal returns to education. *American Economic Review* 101: 2754–2781.

Carneiro, P., and S. Lee. 2009. Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics* 149: 191–208.

Carneiro, P., M. Lokshin, and N. Umapathi. 2017. Average and marginal returns to upper secondary schooling in Indonesia. *Journal of Applied Econometrics* 32: 16–36.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg. 2016. From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics* 41: 47–60.

———. Forthcoming. Who benefits from universal child care? Estimating marginal returns to early child care attendance. *Journal of Political Economy*.

Felfe, C., and R. Lalive. 2014. Does early child care help or hurt children's development? Discussion Paper No. 8484, Institute for the Study of Labor (IZA). http://ftp.iza.org/dp8484.pdf.

Heckman, J. J., S. Urzua, and E. J. Vytlacil. 2006a. Web supplement to understanding instrumental variables in models with essential heterogeneity: Estimation of treatment effects under essential heterogeneity. http://jenni.uchicago.edu/underiv/documentation_2006_03_20.pdf.

———. 2006b. Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics* 88: 389–432.

Heckman, J. J., and E. J. Vytlacil. 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences* 96: 4730–4734.

———. 2001. Local instrumental variables. In *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. C. Hsiao, K. Morimune, and J. L. Powell, 1–46. New York: Cambridge University Press.

———. 2005. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73: 669–738.

———. 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 *Handbook of Econometrics, vol. 6B*, ed. J. J. Heckman and E. E. Leamer, 4875–5143. Amsterdam: Elsevier.

Imbens, G. W., and J. D. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62: 467–475.

Klein, R. W., and R. H. Spady. 1993. An efficient semiparametric estimator for binary response models. *Econometrica* 61: 387–421.

Lokshin, M., and Z. Sajaia. 2004. Maximum likelihood estimation of endogenous switching regression models. *Stata Journal* 4: 282–289.

Maestas, N., K. J. Mullen, and A. Strand. 2013. Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt. *American Economic Review* 103: 1797–1829.

Matzkin, R. L. 2007. Nonparametric identification. In *Handbook of Econometrics, vol. 6B*, ed. J. J. Heckman and E. E. Leamer, 5307–5368. Amsterdam: Elsevier.

Mogstad, M., and A. Torgovitsky. 2017. Identification and extrapolation with instrumental variables. Unpublished manuscript.

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

Vytlacil, E. 2002. Independence, monotonicity, and latent index models: An equivalence result. *Econometrica* 70: 331–341.

About the author

Martin Eckhoff Andresen is a researcher at Statistics Norway and received his PhD from the University of Oslo in 2017. His research interests are applied econometrics, economics of education, and labor economics.

A Estimation algorithms

The `mtefe` package is inspired by Heckman, Urzua, and Vytlacil (2006a,b) and Brave and Walstrum (2014). The following sections describe the main steps of the estimation using this program:

A.1 First stage and common support

1. Estimate the first stage using a linear probability model, a probit or a logit model of D on Z. In principle, this step could be made more flexible by using, for example, a semiparametric single index model such as Klein and Spady (1993), which is not currently implemented in `mtefe`.
2. Predict the propensity score, $\widehat{p}$.
    a. If using the linear probability, manually adjust propensity scores below 0 to 0 and above 1 to 1.
3. Construct the common support matrix:
    a. If trimming the sample using `trimsupport()`:
        i. Estimate the density of the propensity scores separately in the two samples.
        ii. Remove points of support with the lowest densities until the specified share of each sample has been removed.
        iii. Construct the common support as the points of overlapping support between the two samples.
        iv. Remove observations with estimated propensity scores outside the common support from the estimation sample.
        v. Fit the baseline propensity score model on the trimmed sample again.
    b. If not trimming:
        i. If using a parametric model, use the full support in 0.01 intervals from 0.01 to 0.99.
        ii. If using a semiparametric model, use all points of overlapping support in the treated and untreated samples, from 0.01 to 0.99 in intervals of 0.01.
4. Plot the distribution of propensity scores in the treated and untreated samples, including the trimming limits if used, to visualize the common support.
5. Compute weights for the ATT, ATUT, LATE, MPRTEs, and, if specified, PRTE parameter weights as described in section 2.4.

A.2 Local IV estimation

1. Construct $K(p)$, which depends on your choice of parametric or semiparametric model—see tables 1 and 2.

2. Estimate the conditional expectation of Y given X and p as

$$Y = X\beta_0 + X(\beta_1 - \beta_0)p + K(p) + \epsilon$$

3. From these estimates, recover $k(u) = K'(u)$.

4. Construct the MTE as the derivative of the conditional expectation:

$$\widehat{\text{MTE}(x,u)} = x(\widehat{\beta_1 - \beta_0}) + \widehat{k(u)}$$

A.3 Separate approach

1. Construct $K_1(p)$ and $K_0(p)$, depending on your specification; see tables 1 and 2.

2. Estimate the conditional mean of Y from the stacked regression:

$$Y = X\beta_0 + K_D(p) + D\{X(\beta_1 - \beta_0) + K_D(p)\} + \epsilon$$
$$\text{where } K_D(p) = \begin{cases} K_1(p) & \text{if } D = 1 \\ K_0(p) & \text{if } D = 0 \end{cases}$$

3. From these estimates, recover the $k_j(u)$ functions.

4. Construct the estimates of the potential outcomes and the MTE as

$$\widehat{Y_j}(x,u) = x\widehat{\beta_j} + \widehat{k_j(u)}$$
$$\widehat{\text{MTE}(x,u)} = \widehat{Y_1}(x,u) - \widehat{Y_0}(x,u)$$

A.4 Maximum likelihood estimation

Relevant only for the joint normal model, the `mlikelihood` option implements the maximum likelihood estimator described in Lokshin and Sajaia (2004). The individual log-likelihood contribution is

$$\ell_i = D_i \left[\ln\{F(\eta_{1i})\} + \ln\left\{ \frac{1}{\sigma_1} f\left(\frac{U_{1i}}{\sigma_1} \right) \right\} \right] + (1 - D_i) \left[\ln\{F(-\eta_{0i})\} + \ln\left\{ \frac{1}{\sigma_0} f\left(\frac{U_{0i}}{\sigma_{10}} \right) \right\} \right]$$
$$\text{where } \eta_{ji} = \left(\gamma Z_i + \rho_j \frac{U_{ji}}{\sigma_j} \right) \frac{1}{\sqrt{1 - \rho_j^2}}$$

where f is the standard normal density. This log likelihood can be maximized to give the coefficients $\gamma, \beta_0, \beta_1, \sigma_0, \sigma_1, \rho_0$, and ρ_1. These parameter estimates can be used to construct the MTE and treatment-effect parameters as detailed in table 1.

A.5 Standard errors

When using parametric models, `mtefe` by default calculates standard errors for the MTEs and the treatment-effect parameters from the estimated variance of $\beta_1 - \beta_0$ and the parameters of $k(u)$. This ignores the fact that the propensity score and the means of X themselves are estimated objects. Little is known of the effect of this omission, although Abadie and Imbens (2016) show that ignoring the fact that the propensity scores are estimated when matching increases the standard errors of the ATE, but the effect on other parameters is ambiguous. Alternatively, and preferably, standard errors should be estimated using the bootstrap. Implement this using the `bootreps()` option, or alternatively, use `vce(cluster` *clustvar*`)` if cluster bootstrap is appropriate. This procedure reestimates the propensity score, the mean of X, and the treatment-effect parameter weights for each bootstrap replication and so accounts for the uncertainty from the first stage unless the option `norepeat1` is specified.

A.6 IV weights

To estimate the IV weights, we need measures of the impact of the instrument on each individual conditional on X. We are looking at partitioning the linear first-stage regression

$$D = X\beta_D + \gamma Z_- + \epsilon$$

1. Remove the impact of covariates on D and Z_- by regressing them separately on X and saving the residuals in ϵ_D and ϵ_{Z_-}.

2. Regress the residualized treatment ϵ_D on ϵ_{Z_-} (without a constant) to recover the first-stage estimate of the impact of the instrument on treatment conditional on controls by the Frisch–Waugh–Lowell theorem.

3. The predicted values from this regression, $\widehat{v}$, contains the individual impact of the instrument on treatment conditional on controls. Note how we can recover the first-stage estimate from $\text{cov}(D, \widehat{v})$, the reduced-form estimate from $\text{cov}(Y, \widehat{v})$, and the traditional 2SLS estimate from $\{\text{cov}(Y, \widehat{v})\}/\{\text{cov}(D, \widehat{v})\}$.

4. Compute the weights

$$\widehat{\kappa}_i^{\text{LATE}} = \frac{\left(\widehat{v}_i - \overline{\widehat{v}}\right)\left(D_i - \overline{D}\right)}{\text{cov}(D, \widehat{v})}$$

   Note how κ^{LATE} weights treated individuals with positive $\widehat{v}$ and untreated individuals with negative $\widehat{v}$ more—these are individuals who are more strongly affected by the instruments and so are more likely to be compliers.

5. Compute

$$\widehat{\omega}_{\text{LATE}}(u) = \frac{\{\mathbb{E}(\widehat{v}|p > u) - \mathbb{E}(\widehat{v})\}\, P(p > u)}{s \times \text{cov}(D, \widehat{v})}$$

by replacing expectations and probabilities with sample analogs. This puts more weight on the values of U_D above which people have higher $\widehat{v}$—precisely the people who are more affected by the instruments and so are more likely to be compliers.

This procedure recovers the weights from Cornelissen et al. (2016).

B Monte Carlo simulations

To investigate the properties of the different MTE estimators and models implemented by the `mtefe` package, I simulate data from a data-generating process based on the example described above and detailed in section B.1. For each of the data-generating processes (joint normal or polynomial error structure), I draw 10,000 observations for each of 500 repetitions, randomly allocating each observation to one of 10 districts. The first-stage model that determines selection into treatment is either a linear probability or probit model.

For each repetition, I calculate the difference between the estimated and the true parameter, the estimated analytic and bootstrapped standard errors, and the rejection rates for the true parameter. I fit all models using the separate approach and local IVs to compare any differences.

B.1 Data-generating process

This algorithm determines the data-generating process implemented by `mtefe_gendata` that is part of the `mtefe` package and used in the Monte Carlo simulations and examples in this article. This data-generating process generates selection on both levels and gains as well as a continuous instrument that is valid only conditional on fixed effects.

1. In a sample of N individuals, randomly allocate each to one of G districts with equal probability.

2. Draw average labor market quality Π_j in the college and noncollege labor markets and average distance to college in each district AvgDist from a joint normal distribution:

$$\Pi_0, \Pi_1, \text{AvgDist} \sim \mathcal{N}(0, \mathbf{\Sigma}_f)$$

$$\Sigma_f = \left\{ \begin{array}{ccc} 0.1 & & \\ 0.05 & 0.1 & \\ -0.05 & -0.1 & 10 \end{array} \right\}$$

In the simulations, I draw these once rather than repeat for each simulation to have a true value to compare the estimated coefficients with, but by default, `mtefe_gendata` draws these for each replication unless you specify the parameters.

3. Generate individual distance to college `distCol` as AvgDist plus $N(40, 10)$.
4. Draw `exp` from $U(0, 30)$ and construct `exp2` as its square.
5. Construct the error terms U_0, U_1, and V from one of two error structures:

   Normal Draw randomly from a joint normal model:

$$U_0, U_1, V \sim \mathcal{N}(0, \mathbf{\Sigma})$$
$$\Sigma_f = \left\{ \begin{array}{ccc} .5 & & \\ .3 & .5 & \\ -.1 & -.5 & 1 \end{array} \right\}$$

   Polynomial Generate U_j as second-order polynomials of U_D with mean zero:

   a. Draw $U_D = U(0, 1)$ and generate $V = F^{-1}(U_D)$.
   b. Generate $U_j = \sum_{l=1}^{2} \pi_{jl}[U_D^l - \{1/(l+1)\}] + \epsilon$, where $\epsilon \sim N(0, 0.2)$:
      i. using $\pi_{11} = 0.5$, $\pi_{12} = -0.1$
      ii. and $\pi_{01} = 2$, $\pi_{12} = -1$.

6. Construct the potential outcomes as

$$Y_j = X\beta_j + \Pi_j + U_j$$

$$\begin{array}{cccccc} X & = & \texttt{exp} & \texttt{exp2} & 1 \\ \beta_0 & = & 0.025 & -0.0004 & 3.2 \\ \beta_1 & = & 0.01 & 0 & 3.6 \end{array}$$

7. Determine treatment using one of two binary choice models:

   Probit

$$D = \mathbb{1}\left[\gamma Z > V\right]$$

$$\begin{array}{cccccc} Z & = & \texttt{distCol} & \texttt{exp} & \texttt{exp2} & 1 \\ \gamma & = & -0.125 & -0.08 & 0.002 & 5.59 \end{array}$$

   LPM

$$D = \mathbb{1}\left[\gamma Z > U_D\right]$$

$$\begin{array}{cccccc} Z & = & \texttt{distCol} & \texttt{exp} & \texttt{exp2} & 1 \\ \gamma & = & -0.015 & -0.01 & 0.00025 & 1.17375 \end{array}$$

8. Determine the observed outcome `lwage` as $DY_1 + (1 - D)Y_0$.

B.2 A well-specified baseline

To establish a baseline, I first investigate the case where both the first stage and the polynomial model for $k(u)$ are correctly specified.

The results from these exercises are displayed in table 4. Unsurprisingly, and in line with theory and Monte Carlo evidence from Brave and Walstrum (2014), `mtefe` does a good job at estimating both the coefficients of the outcome equations and the MTEs when both the first stage and the parametric model are correctly specified. The difference between the estimated and true coefficients center at 0 and have low standard deviations. The estimated analytical standard errors come close to the standard deviations of the coefficient. Furthermore, even though the analytical standard errors ignore the fact that the propensity score itself is an estimated object, they fall very close to the bootstrapped standard errors that account for this by reestimating the propensity scores for each bootstrap replication. Note that bootstrapped standard errors are not always higher than analytic standard errors. Rejection rates vary somewhat from parameter to parameter but lie around 0.05 as they should.

Table 4. Monte Carlo estimates: A well-specified baseline

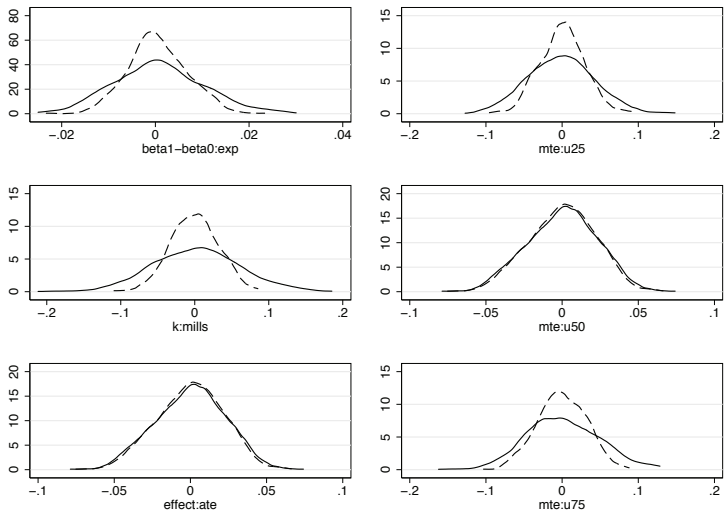
		True model	Normal			Normal			Polynomial			Polynomial		
		Model	Normal			Normal			Polynomial			Polynomial		
		Procedure	Local IV			Separate			Local IV			Separate		
	Coefficient	Truth	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r
β_0	exp	0.025	−0.000	0.006	0.042	−0.000	0.004	0.088	0.000	0.003	0.078	0.000	0.002	0.084
			(0.006)	0.006	0.050	(0.005)	0.004	0.066	(0.003)	0.003	0.044	(0.002)	0.002	0.050
	exp^2	−0.004	0.000	0.000	0.040	0.000	0.000	0.066	−0.000	0.000	0.070	−0.000	0.000	0.096
			(0.000)	0.000	0.050	(0.000)	0.000	0.056	(0.000)	0.000	0.032	(0.000)	0.000	0.044
	3.district	0.34	−0.003	0.062	0.036	−0.001	0.042	0.074	−0.001	0.025	0.062	−0.002	0.016	0.104
			(0.058)	0.058	0.054	(0.045)	0.044	0.060	(0.026)	0.026	0.058	(0.020)	0.019	0.054
	6.district	0.34	−0.004	0.059	0.060	−0.003	0.040	0.072	0.000	0.023	0.078	−0.001	0.015	0.102
			(0.061)	0.058	0.058	(0.044)	0.042	0.060	(0.026)	0.026	0.056	(0.019)	0.019	0.054
	10.district	−0.51	−0.004	0.060	0.052	−0.002	0.041	0.062	0.001	0.024	0.064	0.000	0.016	0.096
			(0.059)	0.057	0.064	(0.043)	0.043	0.058	(0.025)	0.025	0.062	(0.019)	0.019	0.060
	constant	3.2	0.006	0.065	0.042	0.002	0.043	0.078	−0.001	0.026	0.084	0.000	0.017	0.098
			(0.063)	0.062	0.046	(0.047)	0.045	0.064	(0.028)	0.028	0.072	(0.021)	0.020	0.062
$\beta_1 - \beta_0$	exp	−0.015	0.001	0.010	0.042	0.001	0.006	0.076	−0.000	0.004	0.038	0.000	0.002	0.058
			(0.010)	0.010	0.044	(0.006)	0.006	0.070	(0.004)	0.004	0.048	(0.002)	0.002	0.062
	exp^2	0.0004	−0.000	0.000	0.038	−0.000	0.000	0.070	0.000	0.000	0.040	−0.000	0.000	0.050
			(0.000)	0.000	0.046	(0.000)	0.000	0.056	(0.000)	0.000	0.052	(0.000)	0.000	0.056
	3.district	−0.35	0.003	0.106	0.028	0.001	0.060	0.062	−0.000	0.042	0.034	0.001	0.023	0.052
			(0.098)	0.098	0.052	(0.062)	0.061	0.060	(0.039)	0.039	0.048	(0.024)	0.023	0.054
	6.district	1.00	0.007	0.109	0.078	0.005	0.061	0.038	−0.002	0.043	0.068	0.001	0.023	0.062
			(0.114)	0.112	0.066	(0.058)	0.062	0.036	(0.047)	0.047	0.046	(0.024)	0.023	0.066
	10.district	−0.07	0.006	0.107	0.044	0.001	0.060	0.050	−0.001	0.043	0.024	0.001	0.023	0.032
			(0.104)	0.100	0.066	(0.061)	0.061	0.050	(0.037)	0.038	0.042	(0.022)	0.023	0.034
	constant	0.4	−0.010	0.098	0.030	−0.005	0.058	0.062	0.002	0.041	0.042	−0.000	0.023	0.070
			(0.091)	0.094	0.036	(0.060)	0.059	0.050	(0.039)	0.039	0.052	(0.025)	0.024	0.062

Continued on next page

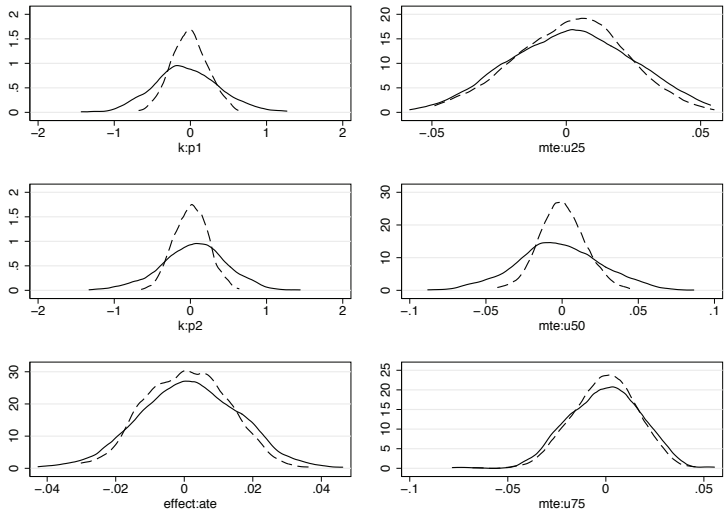
		True model Model Procedure	Normal Normal Local IV			Normal Normal Separate			Polynomial Polynomial Local IV			Polynomial Polynomial Separate		
	Coefficient	Truth	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r
$k(u)$	$\rho_1 - \rho_0$	−0.4	0.002	0.062	0.048	−0.000	0.032	0.048						
			(0.059)	0.059	0.062	(0.032)	0.032	0.050						
	$\pi_{11} - \pi_{01}$	−1.5							−0.051	0.434	0.048	−0.014	0.253	0.028
									(0.431)	0.451	0.038	(0.236)	0.249	0.038
	$\pi_{12} - \pi_{02}$	0.9							0.046	0.433	0.048	0.010	0.248	0.020
									(0.424)	0.439	0.050	(0.218)	0.231	0.046
MTE	MTE($\overline{x}$, 0.05)	1.02	−0.003	0.103	0.042	0.001	0.056	0.052	0.009	0.070	0.048	0.004	0.045	0.078
		0.74	(0.098)	0.099	0.052	(0.056)	0.057	0.050	(0.071)	0.075	0.034	(0.048)	0.050	0.046
	MTE($\overline{x}$, 0.1)	0.88	−0.003	0.081	0.038	0.001	0.045	0.042	0.007	0.052	0.054	0.003	0.035	0.100
		0.67	(0.077)	0.078	0.046	(0.046)	0.047	0.044	(0.054)	0.057	0.030	(0.039)	0.040	0.044
	MTE($\overline{x}$, 0.25)	0.63	−0.001	0.047	0.028	0.000	0.030	0.044	0.002	0.021	0.068	0.002	0.017	0.104
		0.49	(0.044)	0.045	0.044	(0.030)	0.031	0.038	(0.023)	0.023	0.050	(0.020)	0.020	0.056
	MTE($\overline{x}$, 0.5)	0.36	0.000	0.025	0.026	0.000	0.021	0.050	−0.002	0.027	0.060	−0.000	0.016	0.032
		0.29	(0.023)	0.023	0.036	(0.022)	0.022	0.034	(0.027)	0.028	0.048	(0.015)	0.016	0.042
	MTE($\overline{x}$, 0.75)	0.09	0.002	0.050	0.040	−0.000	0.031	0.058	−0.000	0.022	0.012	−0.000	0.018	0.024
		0.19	(0.048)	0.047	0.060	(0.032)	0.031	0.054	(0.018)	0.020	0.018	(0.016)	0.018	0.032
	MTE($\overline{x}$, 0.9)	−0.15	0.003	0.085	0.044	−0.001	0.047	0.054	0.003	0.056	0.032	−0.000	0.037	0.014
		0.19	(0.082)	0.080	0.058	(0.048)	0.046	0.058	(0.050)	0.051	0.048	(0.031)	0.033	0.032
	MTE($\overline{x}$, 0.95)	−0.29	0.004	0.106	0.044	−0.001	0.057	0.046	0.005	0.074	0.030	0.000	0.047	0.016
		0.20	(0.103)	0.101	0.060	(0.058)	0.057	0.052	(0.068)	0.068	0.054	(0.039)	0.041	0.028
	ATE($\overline{x}$)	0.37	0.000	0.025	0.026	0.000	0.021	0.050	0.002	0.014	0.054	0.001	0.010	0.086
		0.36	(0.023)	0.023	0.036	(0.022)	0.022	0.034	(0.014)	0.015	0.048	(0.012)	0.012	0.042

Note: Monte Carlo experiments described in section B. Standard deviations in parentheses. Results based on bootstrapped standard errors are underlined, otherwise analytic. True values shows normal (top) and polynomial (bottom). First stage is probit. 500 replications, each with a sample size of 10,000 in 10 districts.

Comparing the results from local IV estimates with the results from the separate approach reveals one important difference; across the model specifications, estimates seem to be more precise using the separate approach. Both the standard deviations of the coefficients and the estimated standard errors are lower than estimates from local IV, as illustrated for a range of parameters in figure 3. This is surprising, given that more parameters are estimated when using the separate approach than local IV. I have no good explanation for this result, but it warrants more research.



(a) Normal model



(b) Polynomial model

Note: Kernel density plots of estimated coefficients using local IV (solid), the separate approach (dashed) for selected parameters of the MTE models. Well-specified baseline model. 500 Monte Carlo simulations from the normal (a) and polynomial (b) models.

Figure 3. Efficiency of the two estimation methods

B.3 Misspecification of k(u)

The above experiment focused on the case where both the first-stage model and the parametric model for $k(u)$ was correctly specified. To illustrate how the estimators do when one of the two are misspecified, I keep the assumption that the first-stage model is correctly specified but use a misspecified MTE model. In table 5, I generate data from the normal model and estimate the MTE using the polynomial model of second order and vice versa.

Both models do a relatively good job at estimating the coefficients of the outcome equations, and rejection rates for the true β_0 and $\beta_1 - \beta_0$ coefficients are close to 0.05, as they should be. Note, however, that the bootstrap now outperforms the analytic standard errors, yielding rejections rates closer to 5% than the analytic standard errors in most cases.

The MTE estimates, and in turn the ATE, are severely overrejected. Bootstrapping the standard errors and accounting for the uncertainty in the estimates of p produce lower rejection rates, but these are still above 5%. The problem seems to be driven by inconsistent estimates rather than too-low standard errors. The result of this is large rates of overrejection when the functional form is misspecified; average rejection rates are far above 0.05 for a 5% test. The polynomial model seems to do a better job at approximating the normal model than the other way around, as is apparent by the lower rejection rates.

This illustrates the importance of using the correct parametric model and working with flexible specifications. In this case, a higher-order polynomial or a polynomial with splines could have generated a better fit to the underlying model.

Table 5. Monte Carlo estimates: Misspecification of $k(u)$

		True model Model Procedure	Polynomial Normal Local IV			Polynomial Normal Separate			Normal Polynomial Local IV			Normal Polynomial Separate		
	Coefficient	Truth	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r
β_0	exp	0.025	−0.000	0.003	0.086	−0.000	0.002	0.126	−0.000	0.006	0.042	−0.000	0.004	0.068
			(0.003)	0.003	0.054	(0.002)	0.002	0.068	(0.006)	0.006	0.050	(0.005)	0.004	0.050
	exp^2	−0.004	0.000	0.000	0.084	0.000	0.000	0.128	0.000	0.000	0.038	0.000	0.000	0.078
			(0.000)	0.000	0.052	(0.000)	0.000	0.068	(0.000)	0.000	0.046	(0.000)	0.000	0.058
	3.district	0.34	0.002	0.025	0.052	0.001	0.016	0.092	−0.001	0.062	0.040	0.000	0.042	0.088
			(0.026)	0.025	0.054	(0.018)	0.019	0.044	(0.059)	0.058	0.054	(0.046)	0.044	0.066
	6.district	0.34	−0.001	0.023	0.086	−0.004	0.015	0.110	−0.005	0.059	0.070	−0.003	0.040	0.084
			(0.026)	0.026	0.060	(0.018)	0.018	0.062	(0.062)	0.058	0.074	(0.045)	0.043	0.076
	10.district	−0.51	−0.001	0.024	0.050	−0.002	0.016	0.094	−0.002	0.060	0.050	−0.000	0.041	0.054
			(0.025)	0.025	0.040	(0.018)	0.019	0.034	(0.059)	0.057	0.060	(0.044)	0.043	0.046
	constant	3.2	0.007	0.026	0.076	0.031	0.016	0.480	0.011	0.066	0.034	−0.005	0.045	0.060
			(0.028)	0.027	0.070	(0.019)	0.019	0.364	(0.064)	0.063	0.044	(0.047)	0.047	0.046
$\beta_1 - \beta_0$	exp	−0.015	−0.000	0.004	0.062	0.000	0.002	0.054	0.000	0.010	0.058	0.000	0.006	0.060
			(0.004)	0.004	0.064	(0.002)	0.002	0.056	(0.010)	0.010	0.064	(0.006)	0.006	0.052
	exp^2	0.0004	0.000	0.000	0.056	−0.000	0.000	0.058	−0.000	0.000	0.036	−0.000	0.000	0.060
			(0.000)	0.000	0.052	(0.000)	0.000	0.058	(0.000)	0.000	0.052	(0.000)	0.000	0.052
	3.district	−0.35	−0.002	0.042	0.036	0.000	0.023	0.034	0.000	0.106	0.030	−0.001	0.060	0.044
			(0.039)	0.039	0.048	(0.022)	0.023	0.040	(0.097)	0.098	0.046	(0.061)	0.061	0.042
	6.district	1.00	0.001	0.043	0.066	0.003	0.023	0.050	0.005	0.109	0.046	0.002	0.061	0.064
			(0.046)	0.047	0.040	(0.023)	0.023	0.054	(0.109)	0.112	0.040	(0.064)	0.062	0.064
	10.district	−0.07	0.001	0.043	0.028	0.002	0.023	0.052	0.001	0.107	0.040	−0.001	0.060	0.058
			(0.040)	0.038	0.056	(0.023)	0.023	0.054	(0.099)	0.100	0.056	(0.060)	0.061	0.052
	constant	0.4	−0.023	0.039	0.078	−0.030	0.022	0.248	0.003	0.102	0.038	0.035	0.060	0.088
			(0.038)	0.037	0.102	(0.022)	0.023	0.242	(0.101)	0.097	0.056	(0.060)	0.062	0.082

Continued on next page

		True model Model Procedure	Polynomial Normal Local IV			Polynomial Normal Separate			Normal Polynomial Local IV			Normal Polynomial Separate		
	Coefficient	Truth	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r
$k(u)$	$\rho_1 - \rho_0$	N/A*	−0.199	0.025		−0.115	0.012							
			(0.024)	0.024		(0.012)	0.012							
	$\pi_{11} - \pi_{01}$	N/A*							−1.333	1.091		−2.513	0.664	
									(1.074)	1.079		(0.628)	0.672	
	$\pi_{12} - \pi_{02}$	N/A*							0.104	1.087		1.296	0.651	
									(1.058)	1.066		(0.620)	0.647	
MTE	MTE$(\overline{x}, 0.05)$	1.02	−0.073	0.041	0.460	−0.215	0.021	1.000	−0.087	0.175	0.086	0.079	0.117	0.102
		0.74	(0.041)	0.042	0.438	(0.023)	0.023	1.000	(0.173)	0.174	0.088	(0.112)	0.122	0.082
	MTE$(\overline{x}, 0.1)$	0.88	−0.077	0.032	0.672	−0.189	0.017	1.000	−0.008	0.132	0.050	0.109	0.092	0.220
		0.67	(0.033)	0.034	0.644	(0.019)	0.019	1.000	(0.131)	0.131	0.050	(0.089)	0.096	0.194
	MTE$(\overline{x}, 0.25)$	0.63	−0.020	0.019	0.196	−0.080	0.011	1.000	0.040	0.052	0.110	0.043	0.044	0.166
		0.49	(0.020)	0.020	0.148	(0.013)	0.014	1.000	(0.051)	0.050	0.142	(0.046)	0.046	0.140
	MTE$(\overline{x}, 0.5)$	0.36	0.052	0.010	1.000	0.048	0.008	1.000	−0.004	0.068	0.056	−0.073	0.043	0.398
		0.29	(0.011)	0.011	1.000	(0.010)	0.010	1.000	(0.068)	0.067	0.054	(0.043)	0.043	0.398
	MTE$(\overline{x}, 0.75)$	0.09	0.012	0.020	0.084	0.065	0.012	1.000	−0.035	0.056	0.092	−0.026	0.047	0.088
		0.19	(0.019)	0.018	0.106	(0.012)	0.012	1.000	(0.053)	0.052	0.114	(0.047)	0.048	0.082
	MTE$(\overline{x}, 0.9)$	−0.15	−0.106	0.034	0.894	−0.002	0.018	0.030	0.034	0.142	0.046	0.160	0.098	0.372
		0.19	(0.032)	0.031	0.922	(0.017)	0.016	0.054	(0.132)	0.133	0.060	(0.096)	0.096	0.384
	MTE$(\overline{x}, 0.95)$	−0.29	−0.187	0.042	0.990	−0.052	0.022	0.646	0.122	0.186	0.090	0.300	0.124	0.682
		0.20	(0.040)	0.040	0.994	(0.021)	0.020	0.716	(0.174)	0.176	0.104	(0.122)	0.121	0.692
	ATE$(\overline{x})$	0.37	−0.021	0.010	0.588	−0.025	0.008	0.822	0.005	0.036	0.822	0.033	0.026	0.822
		0.36	(0.011)	0.011	0.482	(0.010)	0.010	0.720	(0.031)	0.034	0.042	(0.025)	0.028	0.186

Note: Monte Carlo experiments described in section B. Standard deviations in parentheses. Results based on bootstrapped standard errors are underlined, otherwise analytic. 500 replications, each with a sample size of 10,000 in 10 districts.
* Coefficients reported for $k(u)$ rather than differences.

B.4 Misspecification of P(Z)

Next, I move on to investigate the specification of the first stage. To focus on the problem of misspecification of the first stage, assume that the parametric model for $k(u)$ is correct and second-order polynomial. Instead, I let the functional form of the selection equation be either linear or probit when generating data, and I use the other functional form when fitting the model.

The results from this exercise are provided in table 6. The β parameters are somewhat consistently estimated with rejection rates around 0.05, but there is larger variation across parameters than for the previous experiments. The MTEs, ATE, and parameters of the $k(u)$ function are severely overrejected. Again, this does not seem to be driven by underestimation of standard errors if we compare them with the standard deviation of the estimated coefficients. Rather, it seems to be the coefficients themselves that do not center at 0.

As an alternative misspecification, consider the case where the functional form is correct but where the experience controls are not included in either the first- or the second-stage estimation. This should affect the precision of the coefficients in the propensity score model, but because the omitted regressors are orthogonal to everything else, it should not bias the estimates of other parameters except the constant. The results from this exercise are found in table 7, and fortunately, we see that the results are not sensitive to this sort of omission of variables as long as the exclusion restriction holds.

In conclusion, specification of the propensity score generally receives relatively little attention from articles, but it turns out to be important. Careful researchers should evaluate the robustness of their MTE models using various and flexible first-stage models.

Table 6. Monte Carlo estimates: Misspecification of $P(Z)$

	True first-stage model		Probit			Probit			LPM			LPM		
	First-stage model		LPM			LPM			Probit			Probit		
	Procedure		Local IV			Separate			Local IV			Separate		
	Coefficient	Truth	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r
β_0	exp	0.025	0.001	0.003	0.102	0.000	0.002	0.092	−0.000	0.005	0.058	0.000	0.001	0.100
			(0.003)	0.003	0.070	(0.002)	0.002	0.050	(0.005)	0.005	0.042	(0.002)	0.002	0.060
	$\exp^2$	−0.004	−0.000	0.000	0.102	−0.000	0.000	0.088	0.000	0.000	0.056	−0.000	0.000	0.104
			(0.000)	0.000	0.070	(0.000)	0.000	0.050	(0.000)	0.000	0.050	(0.000)	0.000	0.058
	3.district	0.34	0.024	0.028	0.132	−0.001	0.016	0.110	−0.002	0.047	0.028	−0.000	0.014	0.050
			(0.029)	0.029	0.128	(0.019)	0.019	0.068	(0.041)	0.041	0.052	(0.015)	0.016	0.028
	6.district	0.34	−0.071	0.026	0.736	−0.000	0.015	0.126	0.004	0.045	0.080	0.000	0.014	0.106
			(0.030)	0.030	0.664	(0.019)	0.019	0.060	(0.050)	0.052	0.040	(0.016)	0.016	0.044
	10.district	−0.51	0.002	0.027	0.048	−0.000	0.016	0.084	0.001	0.046	0.016	−0.000	0.014	0.100
			(0.028)	0.028	0.044	(0.019)	0.019	0.046	(0.037)	0.038	0.050	(0.016)	0.016	0.050
	constant	3.2	0.006	0.029	0.056	−0.010	0.017	0.116	−0.067	0.070	0.142	−0.021	0.027	0.172
			(0.030)	0.031	0.050	(0.020)	0.021	0.060	(0.068)	0.071	0.130	(0.032)	0.033	0.084
$\beta_1 - \beta_0$	exp	−0.015	−0.003	0.005	0.092	−0.000	0.002	0.038	0.001	0.009	0.056	0.000	0.002	0.044
			(0.005)	0.005	0.094	(0.002)	0.002	0.040	(0.008)	0.009	0.058	(0.002)	0.002	0.044
	$\exp^2$	0.0004	0.000	0.000	0.086	0.000	0.000	0.042	−0.000	0.000	0.052	−0.000	0.000	0.048
			(0.000)	0.000	0.076	(0.000)	0.000	0.040	(0.000)	0.000	0.048	(0.000)	0.000	0.048
	3.district	−0.35	−0.049	0.050	0.148	0.001	0.023	0.054	0.002	0.091	0.024	−0.001	0.020	0.030
			(0.045)	0.046	0.176	(0.023)	0.023	0.056	(0.078)	0.077	0.056	(0.018)	0.020	0.024
	6.district	1.00	0.143	0.051	0.776	−0.000	0.023	0.048	−0.007	0.092	0.094	0.000	0.020	0.048
			(0.056)	0.054	0.742	(0.024)	0.023	0.052	(0.105)	0.110	0.050	(0.020)	0.020	0.048
	10.district	−0.07	−0.007	0.050	0.036	−0.001	0.023	0.062	−0.003	0.091	0.008	−0.000	0.020	0.044
			(0.045)	0.046	0.050	(0.024)	0.023	0.052	(0.071)	0.071	0.052	(0.019)	0.020	0.046
	constant	0.4	−0.006	0.047	0.038	0.013	0.024	0.092	0.099	0.123	0.102	0.025	0.038	0.120
			(0.044)	0.045	0.048	(0.024)	0.025	0.074	(0.114)	0.120	0.100	(0.040)	0.043	0.070

Continued on next page

		True first-stage model		Probit			Probit			LPM			LPM		
		First-stage model		LPM			LPM			Probit			Probit		
		Procedure		Local IV			Separate			Local IV			Separate		
	Coefficient	Truth		diff	s.e.	r	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r
$k(u)$	$\pi_{11} - \pi_{01}$	−1.5		2.612	0.411	1.000	1.466	0.291	1.000	−1.808	1.459	0.236	−0.481	0.496	0.176
				(0.411)	0.428	1.000	(0.282)	0.297	1.000	(1.474)	1.514	0.206	(0.541)	0.554	0.122
	$\pi_{12} - \pi_{02}$	0.9		−2.488	0.410	1.000	−1.435	0.288	1.000	1.654	1.463	0.202	0.412	0.480	0.136
				(0.398)	0.415	1.000	(0.261)	0.278	1.000	(1.446)	1.501	0.180	(0.494)	0.506	0.120
MTE	MTE($\overline{x}$, 0.05)	0.74		−0.361	0.067	1.000	−0.173	0.049	0.926	0.368	0.275	0.274	0.105	0.107	0.184
				(0.070)	0.073	1.000	(0.056)	0.057	0.852	(0.286)	0.290	0.232	(0.126)	0.130	0.100
	MTE($\overline{x}$, 0.1)	0.67		−0.249	0.051	1.000	−0.110	0.038	0.774	0.290	0.215	0.276	0.084	0.087	0.180
				(0.054)	0.056	0.998	(0.045)	0.045	0.680	(0.224)	0.227	0.234	(0.104)	0.108	0.094
	MTE($\overline{x}$, 0.25)	0.49		0.012	0.022	0.116	0.034	0.017	0.498	0.106	0.080	0.254	0.033	0.042	0.180
				(0.023)	0.024	0.084	(0.022)	0.022	0.366	(0.086)	0.087	0.204	(0.052)	0.055	0.072
	MTE($\overline{x}$, 0.5)	0.29		0.198	0.026	1.000	0.132	0.020	1.000	−0.036	0.037	0.174	−0.010	0.018	0.126
				(0.026)	0.027	1.000	(0.017)	0.019	1.000	(0.038)	0.039	0.136	(0.020)	0.020	0.084
	MTE($\overline{x}$, 0.75)	0.19		0.074	0.024	0.900	0.050	0.019	0.758	0.029	0.090	0.026	−0.001	0.046	0.050
				(0.021)	0.021	0.934	(0.018)	0.018	0.768	(0.079)	0.086	0.042	(0.044)	0.046	0.046
	MTE($\overline{x}$, 0.9)	0.19		−0.150	0.056	0.776	−0.086	0.041	0.580	0.167	0.232	0.090	0.029	0.092	0.046
				(0.049)	0.049	0.866	(0.036)	0.036	0.668	(0.213)	0.229	0.086	(0.087)	0.091	0.050
	MTE($\overline{x}$, 0.95)	0.20		−0.250	0.073	0.944	−0.145	0.052	0.820	0.230	0.295	0.098	0.043	0.113	0.046
				(0.065)	0.065	0.972	(0.045)	0.046	0.880	(0.273)	0.292	0.102	(0.107)	0.111	0.050
	ATE($\overline{x}$)	0.37		−0.005	0.015	0.056	0.014	0.011	0.274	0.099	0.092	0.188	0.024	0.032	0.134
		0.36		(0.015)	0.016	0.046	(0.013)	0.014	0.168	(0.090)	0.094	0.158	(0.036)	0.038	0.072

Note: Monte Carlo experiments described in section B. Standard deviations in parentheses. Results based on bootstrapped standard errors are underlined, otherwise analytic. 500 replications, each with a sample size of 10,000 in 10 districts. Error structure is polynomial.

Table 7. Monte Carlo estimates: Omissions in the specification of $P(Z)$

	True first-stage model		LPM			LPM			Probit			Probit		
	First-stage model		LPM			LPM			Probit			Probit		
	Procedure		Local IV			Separate			Local IV			Separate		
	Coefficient	Truth	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r
β_0	3.district	0.34	−0.003	0.051	0.018	0.000	0.015	0.088	−0.000	0.026	0.072	−0.000	0.017	0.096
			(0.044)	0.045	0.036	(0.018)	0.018	0.056	(0.028)	0.027	0.058	(0.020)	0.020	0.052
	6.district	0.34	−0.005	0.049	0.096	−0.000	0.015	0.088	−0.001	0.025	0.076	−0.001	0.016	0.108
			(0.058)	0.055	0.066	(0.018)	0.018	0.056	(0.028)	0.028	0.056	(0.019)	0.020	0.054
	10.district	−0.51	−0.001	0.050	0.020	0.000	0.015	0.096	−0.002	0.025	0.064	−0.002	0.016	0.124
			(0.041)	0.042	0.046	(0.018)	0.018	0.056	(0.027)	0.026	0.056	(0.020)	0.020	0.066
	constant	3.2	0.258	0.055	0.994	0.259	0.025	1.000	0.254	0.021	1.000	0.255	0.014	1.000
			(0.053)	0.055	0.992	(0.031)	0.032	1.000	(0.023)	0.022	1.000	(0.017)	0.017	1.000
$\beta_1 - \beta_0$	3.district	−0.35	0.005	0.098	0.016	−0.001	0.022	0.060	0.002	0.045	0.042	0.001	0.024	0.042
			(0.083)	0.085	0.038	(0.022)	0.022	0.066	(0.043)	0.042	0.058	(0.024)	0.024	0.048
	6.district	1.00	0.009	0.099	0.110	0.000	0.022	0.052	0.002	0.046	0.058	0.001	0.025	0.034
			(0.121)	0.115	0.076	(0.023)	0.022	0.050	(0.049)	0.050	0.042	(0.023)	0.024	0.042
	10.district	−0.07	0.002	0.098	0.008	−0.001	0.022	0.036	0.002	0.045	0.030	0.002	0.024	0.058
			(0.075)	0.079	0.042	(0.021)	0.022	0.038	(0.043)	0.041	0.050	(0.025)	0.024	0.064
	constant	0.4	−0.111	0.100	0.166	−0.113	0.037	0.842	−0.104	0.034	0.872	−0.107	0.020	1.000
			(0.089)	0.094	0.206	(0.039)	0.042	0.784	(0.033)	0.032	0.896	(0.020)	0.020	1.000
$k(u)$	$\pi_{11} - \pi_{01}$	−1.5	0.106	1.162	0.058	0.197	0.501	0.086	0.028	0.458	0.054	0.136	0.269	0.074
			(1.191)	1.211	0.050	(0.532)	0.567	0.058	(0.471)	0.476	0.050	(0.260)	0.271	0.076
	$\pi_{12} - \pi_{02}$	0.9	−0.109	1.167	0.064	−0.202	0.488	0.064	−0.038	0.457	0.048	−0.147	0.264	0.058
			(1.183)	1.196	0.060	(0.490)	0.519	0.050	(0.455)	0.465	0.044	(0.238)	0.251	0.074

Continued on next page

True first-stage model														
First-stage model														
Procedure			LPM			LPM			Probit			Probit		
			LPM			LPM			Probit			Probit		
			Local IV			Separate			Local IV			Separate		
Coefficient	Truth	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r	diff	s.e.	r
MTE($\overline{x}$, 0.05)	0.74	−0.016	0.220	0.062	−0.033	0.107	0.100	0.000	0.074	0.082	−0.015	0.048	0.092
		(0.227)	0.234	0.038	(0.123)	0.132	0.040	(0.079)	0.079	0.054	(0.054)	0.054	0.054
MTE($\overline{x}$, 0.1)	0.67	−0.012	0.172	0.068	−0.025	0.086	0.100	0.001	0.055	0.080	−0.010	0.037	0.106
		(0.178)	0.184	0.036	(0.102)	0.108	0.038	(0.061)	0.060	0.054	(0.043)	0.044	0.058
MTE($\overline{x}$, 0.25)	0.49	−0.001	0.068	0.068	−0.006	0.042	0.124	0.003	0.022	0.070	0.003	0.018	0.112
		(0.072)	0.075	0.048	(0.052)	0.055	0.048	(0.024)	0.024	0.054	(0.022)	0.022	0.050
MTE($\overline{x}$, 0.5)	0.29	0.005	0.035	0.086	0.006	0.020	0.098	0.003	0.028	0.070	0.009	0.017	0.076
		(0.039)	0.037	0.066	(0.022)	0.022	0.074	(0.029)	0.030	0.062	(0.016)	0.017	0.092
MTE($\overline{x}$, 0.75)	0.19	−0.003	0.079	0.030	−0.008	0.046	0.050	−0.002	0.023	0.064	−0.002	0.019	0.072
		(0.073)	0.074	0.044	(0.046)	0.046	0.052	(0.023)	0.021	0.082	(0.021)	0.019	0.076
MTE($\overline{x}$, 0.9)	0.19	−0.014	0.191	0.050	−0.028	0.093	0.054	−0.007	0.060	0.026	−0.018	0.040	0.064
		(0.184)	0.186	0.056	(0.090)	0.092	0.056	(0.055)	0.055	0.056	(0.037)	0.036	0.100
MTE($\overline{x}$, 0.95)	0.20	−0.019	0.241	0.054	−0.037	0.114	0.054	−0.009	0.079	0.022	−0.025	0.051	0.060
		(0.234)	0.236	0.058	(0.110)	0.112	0.058	(0.073)	0.073	0.042	(0.046)	0.045	0.096
MTE													
ATE($\overline{x}$)	0.37	−0.004	0.071	0.050	−0.011	0.033	0.074	0.000	0.015	0.060	−0.003	0.011	0.108
	0.36	(0.069)	0.073	0.044	(0.035)	0.038	0.040	(0.016)	0.016	0.056	(0.013)	0.013	0.054

Note: Monte Carlo experiments described in section B, where exp and $\exp^2$ are omitted from the model. Standard deviations in parentheses. Results based on bootstrapped standard errors are underlined, otherwise analytic. 500 replications, each with a sample size of 10,000 in 10 districts. Error structure is polynomial.