Heterogeneous Treatment Effects: Instrumental Variables without Monotonicity?

Tobias J. Klein*
University of Mannheim and IZA

first version: November 28, 2003
this version: April 4, 2007

Abstract

A fundamental identification problem in program evaluation arises when idiosyncratic gains from participation and the treatment decision depend on each other. Imbens and Angrist (1994) were the first to exploit a monotonicity condition in order to identify an average treatment effect parameter using instrumental variables. More recently, Heckman and Vytlacil (1999) suggested estimation of a variety of treatment effect parameters using a local version of their approach. However, identification hinges on the same monotonicity assumption that is fundamentally untestable. We investigate the sensitivity of respective estimates to reasonable departures from monotonicity that are likely to be encountered in practice and relate it to properties of a structural parameter. One of our results is that the bias vanishes under a testable linearity condition. Our findings are illustrated in a Monte Carlo analysis.

JEL Classification: C21.
Keywords: Program evaluation, heterogeneity, dummy endogenous variable, selection on unobservables, instrumental variables, monotonicity, identification.

Address: University of Mannheim, Department of Economics, 68131 Mannheim, Germany. *E-Mail:* klein@econ.uni-mannheim.de.

1 Introduction

1.1 Monotonicity

A fundamental identification problem in program evaluation arises if the treatment decision depends on the idiosyncratic gain from participation even if we condition on observables. This selection into treatments on unobservables precludes the use of the usual econometric tools such as matching type estimators, conventional instrumental variables analysis, and standard simultaneous equations models because their respective estimates of treatment effect parameters are generally biased.

Imbens and Angrist (1994) were the first to exploit monotonicity of the treatment decision in instrumental variables in order to identify a local average treatment effect parameter. These instrumental variables are assumed to be independent of the pair of potential outcomes conditional on covariates in the outcome equation. They have identifying power if, conditional on these covariates, they have an impact on the treatment probability. The monotonicity assumption is that a *hypothetical* change in the instruments either has no impact on a unit's treatment status, or changes its treatment status in the same direction as it does for all other units for which it has an impact.

More recently, Heckman and Vytlacil (1999, 2000, 2005, HV in the remainder) suggested estimation of a variety of treatment effect parameters using a local version of their approach.

Both approaches are in principle able to cope with unobserved dependence between the treatment decision and the outcome. They are intuitive, elegant, and easy to implement. Their generality consists of the fact that neither a parametric specification of the joint distribution of unobservables and observables, nor peculiarities of the data set or the economic question of interest are of need. However, identification in both approaches hinges on the same monotonicity assumption. In general, estimates of treatment effects will be biased if it does not hold.

A violation of the monotonicity assumption is nicely motivated in Example 2 of Imbens and Angrist (1994). When we think of two officials screening applicants for a social program,

we would expect that for every set of characteristics of the applicants (the covariates in the outcome equation) the admission rate differs between the two officials. When it is unlikely that the identity of the official affects the outcome of participation or nonparticipation in the program, then, conditional on the characteristics of the applicant, this identity qualifies as an instrument. Suppose the admission rate for official A was higher than for official B. Then, in this setup monotonicity holds whenever *any* applicant who would have been accepted by official B *is* accepted by official A. Imbens and Angrist (1994) note that "this is unlikely to hold if admission is based on a number of criteria." In this case, monotonicity is violated.[1]

In this paper, we aim at quantifying the degree of violation of the monotonicity assumption in order to investigate the consequences of a violation when we unjustifiably rely on this assumption. In particular, we study the effect of reasonable departures from monotonicity, that are likely to be encountered in practice, on estimates of the marginal, average and local average treatment effect. Importantly, highly sensitive estimates would question the suitability of monotonicity based estimates for applied work as the monotonicity assumption is fundamentally untestable since it is identifying.[2]

The identification strategies for the (local) average treatment effect that are proposed by Imbens and Angrist (1994) and HV differ with respect to the requirements on the support of the treatment probability conditional on the instruments, the so-called propensity score. In particular, conditional on covariates in the outcome equation, HV require derivatives of the expected outcomes with respect to the propensity score at infinitely many values of the propensity score to be identified whereas Imbens and Angrist (1994) base their analysis on a finite set of level estimates.

[1] It has been noted in the literature that there are cases in which monotonicity holds naturally, e.g. if it is known for a subset of units that the treatment probability is either zero or one (Battistin and Rettore, forthcoming, for a discussion). This can occur if there are eligibility rules for participation in a program. In this case, by construction, units can only be induced (not) to take the treatment by changes in eligibility.

[2] HV propose a joint test for monotonicity and the existence of instruments. By itself, monotonicity is fundamentally untestable.

1.2 Local Departures From Monotonicity

The approach in this paper is to study the impact of *local departures* from monotonicity on our estimates. Taylor series approximations to the bias terms are derived. This is in the tradition of local specification error analysis suggested by Kiefer and Skoog (1984).[3] It has also been successfully applied by Chesher (1991), Chesher and Schluter (2002) and Battistin and Chesher (2004) in the context of measurement error. Lately, Chesher and Santos Silva (2002) studied the impact of uncontrolled taste variation in discrete choice models by modelling local departures from a multinomial logit model.

The virtue of this approach is that it allows us to keep in touch with the original structure which implies monotonicity. At the same time, we are able to explore what the sensitivity of monotonicity based estimates depends on when monotonicity is in fact violated. In our case, the original structure consists of selection models of the form

$$D = \mathbb{1}\{\tilde{Q}(\tilde{P}(Z, \sigma U)) \geq \tilde{V}\}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, $\tilde{P} : \mathbb{R}^k \to \mathbb{R}$ is a function (an index) of a k-vector of instrumental variables, Z, $\sigma \geq 0$ is a constant, U and $\tilde{V}$ are continuously distributed scalar random variables, and $\tilde{Q} : \mathbb{R}^2 \to \mathbb{R}$. Z, U and $\tilde{V}$ are assumed to be independently distributed.

For $\sigma = 0$ these selection models imply monotonicity. To see this let w and z be two values of Z with $\tilde{Q}(\tilde{P}(w), 0) < \tilde{Q}(\tilde{P}(z), 0)$.[4] Then, given $\tilde{V}$, D can never change from 1 to 0 if Z changes from w to z. This is the original monotonicity assumption of Imbens and Angrist (1994). Vytlacil (2002) shows that a representation of their set of assumptions in terms of such a selection model does not impose any additional restrictions on the data generating process.

[3] Angrist, Imbens, and Rubin (1996) take a different approach and relate the bias in conventional IV estimates to the proportion of non-compliers, i.e. the units for which monotonicity is violated, and the treatment effect heterogeneity. We feel that our approach is fruitful because it allows us, up to an approximation error, to express the treatment effect of non-compliers in terms of possibly identifiable quantities.

[4] For any random variable A and any vector of random variables B we denote realizations thereof by lowercase letters, the c.d.f. of A evaluated at $A = a$ by $F_A(a)$, the conditional c.d.f. of A given $B = b$ evaluated at $A = a$ by $F_{A|B=b}(a)$, and the respective p.d.f.'s by $f_A(a)$ and $f_{A|B=b}(a)$.

Therefore, without loss of generality, from now on, we represent monotonicity in this form.

Central to our generalization of the conventional selection model is an additional scalar random component σU which gives rise to additional individual heterogeneity by perturbing the nonparametric index $\tilde{P}(Z)$. This perturbation, under appropriate conditions, directly leads to a violation of the monotonicity assumption whenever σU is non-degenerate. For degenerate σU, however, the model is constructed so that it is equivalent to a canonical selection model. A local departure from monotonicity is given by a change from $\sigma = 0$ to $\sigma > 0$.

The following example illustrates this point.

EXAMPLE 1 (Random Coeffcients): Consider the index selection model

$$D = \mathbb{1}\{Z\tilde{\gamma} \geq \tilde{V}\}.$$

The canonical index selection model would postulate that $\tilde{\gamma} = \gamma$, a k-vector of parameters. In a probit model, e.g., the additional assumption is made that $\tilde{V}$ is standard normally distributed. Then, given $\tilde{V}$, if $\tilde{P}(Z) = Z\gamma$ changes from $w\gamma$ to $z\gamma > w\gamma$, D can only change from 0 to 1, remain 0, or remain 1, but can never change from 1 to 0. Now, let $\tilde{\gamma}$ be a vector of random coefficients $\tilde{\gamma} = \gamma(1 + \sigma U)$, $\sigma \geq 0$ with U non-degenerate and independent of Z. This is an example for the generalized selection model motivated above. If $\sigma > 0$, given $\tilde{V}$, D is no longer monotone in Z because now, under fairly general conditions on the distribution of U, there exist realizations u and u' of U such that $w\gamma(1 + \sigma u) > z\gamma(1 + \sigma u)$ while $w\gamma(1 + \sigma u') < z\gamma(1 + \sigma u')$. Consequently, monotonicity is violated. Observe that this model nests the canonical index selection model as the special case in which $\sigma = 0$. A local departure from monotonicity is hence given by an external change from $\sigma = 0$ to a small $\sigma > 0$. □

In this paper, we are interested in the bias of treatment effect parameter estimates that can be attributed to such a violation of the monotonicity assumption. We derive a second order ap-

proximation to respective bias terms in σ about $\sigma = 0$ that can be used to assess the accuracy of monotonicity based estimates without monotonicity. We show that the respective bias depends primarily on the dependence between the individual gains from participation in the program, $Y_1 - Y_0$, and the normalized selection threshold $V = F_{\tilde{V}}(\tilde{V})$ from the selection model, which is normalized to be uniformly distributed. Our results can be expressed in terms of a structural parameter, the so-called marginal treatment effect, $m(v) \equiv \mathbb{E}[Y_1 - Y_0|V = v]$. It was introduced by Björklund and Moffitt (1987) and is the average treatment effect conditional on the selection threshold being equal to a certain value v.

We show that under appropriate assumptions, the bias of monotonicity based estimates is related to the curvature of $m(v)$ and the variance of $\tilde{Q}$ conditional on $\tilde{P}$. A bias correction procedure is available if this variance can be estimated.[5] In case no such prior information is available and one is not willing to make additional assumptions, a sensitivity analysis can still be undertaken by evaluating the obtained expressions—under varying additional assumptions—at different values of this variance.

Example 2 is meant to give the intuition behind the main result.

EXAMPLE 2 (Example 1 continued): For the ease of the exposition let $F_{\tilde{V}}$ be known. An equivalent representation of the index selection model in Example 1 is

$$D = \mathbb{1}\{F_{\tilde{V}}(Z\tilde{\gamma}) \geq V\}$$

with $V = F_{\tilde{V}}(\tilde{V})$. By the independence between Z, $\tilde{V}$ and U and the law of total probability $\Pr(D = 1|Z) = \Pr(V \leq F_{\tilde{V}}(Z\tilde{\gamma})|Z) = \mathbb{E}\left[F_{\tilde{V}}(Z\tilde{\gamma})\middle|Z\right]$. Notice that, since only the left hand side of this equation is identified from observations, only $\mathbb{E}\left[F_{\tilde{V}}(Z\tilde{\gamma})\middle|Z\right]$ is known and in general,

[5] Ichimura and Thompson (1998) consider selection models of the form $D = \mathbb{1}\{Z\tilde{\gamma} \geq \tilde{V}\}$ with Z being independent of $(\tilde{\gamma}, \tilde{V})$ and provide conditions under which the joint distribution of $(\tilde{\gamma}, \tilde{V})$ is identified up to normalizations. These conditions do not involve finite dimensional functional form restrictions. They show that the model can consistently be estimated by maximum likelihood. Following up on Example 1, if $\tilde{\gamma}$ is independent of $\tilde{V}$ and $\tilde{\gamma} = \gamma(1 + \sigma U)$, σ is the standard deviation of $Z\tilde{\gamma}$ conditional on $Z\gamma$ if we normalize the variance of U to be 1.

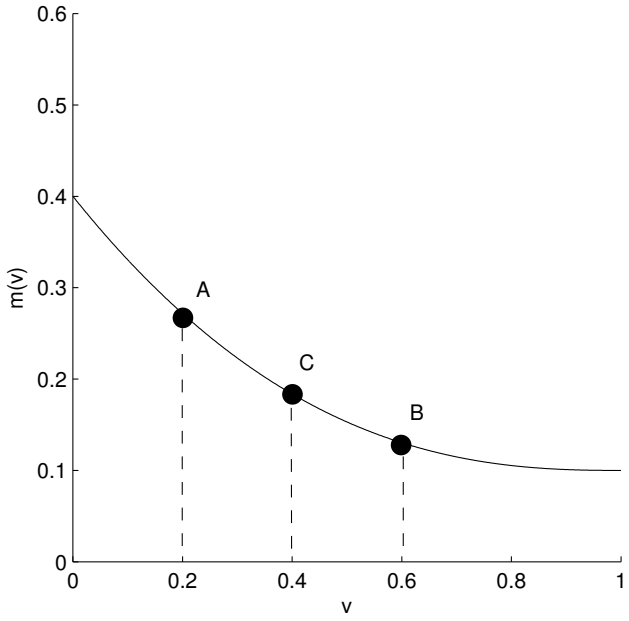

Figure 1: Intuition behind the main result.

$F_{\tilde{V}}(Z\tilde{\gamma})$ is not. Suppose we wanted to obtain an estimate of $m(0.4)$. In Section 2, we will show that $m(0.4)$ can be estimated from observations for which $F_{\tilde{V}}(Z\tilde{\gamma})$ takes on values in a neighborhood around 0.4. Under monotonicity, $\sigma = 0$ and hence $F_{\tilde{V}}(Z\tilde{\gamma})$ is known because it is equal to $F_{\tilde{V}}(Z\gamma)$ which in turn is equal to $\Pr(D = 1|Z)$, recalling that $F_{\tilde{V}}$ is known by assumption. Consequently, $m(0.4)$ is identified because in this case, we can select observations for which $F_{\tilde{V}}(Z\tilde{\gamma}) = F_{\tilde{V}}(Z\gamma) = 0.4$. Now suppose σ and U are such that with respective probability one half either $F_{\tilde{V}}(Z\tilde{\gamma}) = 0.2$ or $F_{\tilde{V}}(Z\tilde{\gamma}) = 0.6$ whenever $\mathbb{E}\left[F_{\tilde{V}}(Z\tilde{\gamma})\middle|Z\right] = 0.4$. Then, monotonicity is violated and from observations with $\mathbb{E}\left[F_{\tilde{V}}(Z\tilde{\gamma})\middle|Z\right]$ taking on values around 0.4 we would estimate $0.5m(0.2) + 0.5m(0.6)$, the convex combination of points A and B in Figure 1. This corresponds to point C only if the marginal treatment effect is linear in V. Only then, our estimate of $m(0.4)$ would still be unbiased even if monotonicity failed to hold. □

1.3 Related Results

Heckman, Urzua, and Vytlacil (2006) discuss in detail that in the framework considered here, if $\sigma = 0$, outcomes of choices are allowed to be heterogeneous in a very general way, but choices itself are not. They therefore advocate the denomination "uniformity" instead of "monotonicity" for this central identifying condition.

Formal representation results have been derived by Vytlacil (2002) who shows that monotonicity can equivalently be expressed in terms of a selection model. Vytlacil (2006) provides a class of nonseparable latent index functions which will have equivalent representations as additively separable or linear index functions. Monotonicity holds for all elements of this class. Central to this representation result is that the impact of instrumental variables on the treatment decision can be separated from the impact of unobservables. Heckman, Urzua, and Vytlacil (2006) discuss this result and relate it to the notion of index sufficiency.

Consequences of a violation of monotonicity have informally been discussed in Angrist, Imbens, and Rubin (1996) who relate the bias in estimates of the local average treatment effect to the proportion of units for which monotonicity does not hold and the difference in local average treatment effects between those units and the ones for which monotonicity holds. In this paper, we try to relate those quantities to structural features of the model, namely σ and properties of the marginal treatment effect.

1.4 Plan of the Paper

Section 2 lays out the formal framework. Section 3 contains the main theoretical results. We illustrate our findings and assess the accuracy of the approximation to the bias term in a Monte Carlo study which is carried out in Section 4. Section 5 concludes.

2 Formal Framework and Identification under Monotonicity

We adopt the usual convention in program evaluation and say that if a unit is not treated, we observe an indicator variable D being equal to zero and a realization of Y_0, and if it is treated, we observe D being equal to one and a realization of Y_1. Usually, Y_0 and Y_1 are referred to as potential outcomes. They are real valued scalar random variables. We write $Y \equiv (1-D)Y_0+DY_1$. Our analysis can be thought of as being conditional on exogenous covariates as, e.g., in Vytlacil (2002).

As we have argued in the introduction we focus on the class of models in which identifying power is derived from exogenous variation in instrumental variables. We denote the k-vector of instrumental variables by Z and express their impact on the treatment decision in terms of a selection model. A conventional selection model is of the form

$$D = \mathbb{1}\{\tilde{P}(Z) \geq \tilde{V}\}$$

with a nonparametric index $\tilde{P}(Z)$, $\tilde{P} : \mathbb{R}^k \to \mathbb{R}$ and scalar $\tilde{V}$. Instead, we consider models where the nonparametric index is perturbed:

$$D = \mathbb{1}\{\tilde{Q}(\tilde{P}(Z), \sigma U) \geq \tilde{V}\}. \tag{1}$$

$\tilde{Q} : \mathbb{R}^2 \to \mathbb{R}$ is a nontrivial function of the nonparametric index $\tilde{P}(Z)$, $\sigma \geq 0$ is a scalar and U is an additional scalar random variable. Under appropriate conditions on $\tilde{Q}$ and the distribution of U monotonicity is violated in these models if $\sigma > 0$.

We make the following assumptions.

ASSUMPTION 1 (Existence of Instruments): *Z is independent of $(Y_0, Y_1, \tilde{V})$.*

ASSUMPTION 2 (Random Noise): *(i) U is independent of Z and $(Y_0, Y_1, \tilde{V})$, (ii) the distribution*

of U is absolutely continuous with respect to Lebesgue measure, and (iii) the support of U is equal to the real line.

ASSUMPTION 3 (Regularity Conditions I): *(i) Y_0 and Y_1 have finite first moments and (ii) the distribution of $\tilde{V}$ is absolutely continuous with respect to Lebesgue measure.*

Assumption 1 presumes (i) that the instruments can be excluded from the outcome equation and (ii) that they are independent of the selection threshold. This is a considerably weaker condition than conditional independence in matching (Rosenbaum and Rubin, 1983), for example, which precludes the dependence of the pair of potential outcomes, (Y_0, Y_1), on unobservables $\tilde{V}$. This dependence is sometimes called selection on unobservables and is allowed for here. Assumption 2(i) formalizes the pure randomness of U. Assumption 3(i) states regularity conditions which ensure that the parameters of interest are well defined. Part (ii) is for convenience and can be relaxed at some notational cost.[6]

(1) along with Assumptions 1 through 3 constitute a simple framework in which violations of monotonicity can occur. This is possible *even* in cases in which U is purely random. Notably, Vytlacil (2006) shows that an equivalent representation of the model, in which monotonicity holds, exists if $\tilde{Q}$ is strictly monotonic in its first argument and the distribution of $\tilde{V}$ is degenerate. This will not be assumed here.

W.l.o.g. we can apply a monotone transformation to the expression in curly brackets of (1). By Assumption 3(ii), $F_{\tilde{V}}(\tilde{v})$ is strictly increasing in $\tilde{v}$ so that an equivalent representation of the model is given by

$$D = \mathbb{1}\{Q(\tilde{P}(Z), \sigma U) \geq V\}$$

with $Q(\tilde{P}(Z), \sigma U) = F_{\tilde{V}}(\tilde{Q}(\tilde{P}(Z), \sigma U))$ and uniformly distributed $V = F_{\tilde{V}}(\tilde{V})$. Observe that this

[6] See HV.

implies that

$$\Pr(D = 1|Z = z, U = u) = \Pr(V \leq Q(\tilde{P}(Z), \sigma U)|Z = z, U = u) = Q(\tilde{P}(z), \sigma u)),$$

where the last equality follows from V being uniformly distributed. This demonstrates that Q is a probability measure. Moreover, under Assumption 2(i) V is independent of U so that by the law of total probability

$$\Pr(D = 1|Z = z) = \Pr(V \leq Q(\tilde{P}(Z), \sigma U)|Z = z) = \Pr(V \leq Q(\tilde{P}(z), \sigma U)) = \mathbb{E}[Q(\tilde{P}(z), \sigma U)].$$

The left hand side of this equation is the so-called propensity score, which we denote by $P(Z)$. Last, we normalize $\tilde{P}(Z)$ and $\tilde{Q}$ so that $P(Z) = \mathbb{E}[Q(P(Z), \sigma U)|Z]$, noting that now $P(Z)$ enters Q instead of $\tilde{P}(Z)$. Moreover, we normalize U to have mean zero and variance 1. Then, σ is the standard deviation of σU. The normalizations are summarized below.

Normalization 1: *Normalize (i) $\tilde{V}$ to be uniformly distributed, (ii) $\tilde{P}$ and $\tilde{Q}$ so that $\mathbb{E}[Q(P(z), \sigma U)] = P(z)$ for any z in the support of Z, and (iii) U so that $\mathbb{E}[U] = 0$, $\mathbb{E}[U^2] = 1$.*

For ease of the exposition, from now on we write P for $P(Z)$. Throughout the paper we will think of P as being a single scalar instrument that can be constructed from the k-vector of instruments as $P = \Pr(D = 1|Z)$. Given the structure of the model and the assumptions, this is innocuous for our purposes.[7] To summarize, under Normalization 1 we can write the selection model as

$$D = \mathbb{1}\{Q(P, \sigma U) \geq V\}. \tag{2}$$

[7]See also the discussion in Heckman, Urzua, and Vytlacil (2006) for this interpretation.

2.1 Properties of the Selection Model

In general, $Q(P, \sigma U)$ is not identified from observations since U and V are not observed and the indicator function that is central to the selection model is not invertible in its argument.

However, if $\sigma = 0$, under Normalization 1(ii) and Assumption 2(i) we have that for any value p of P

$$p = \mathbb{E}[Q(p, 0)] = Q(p, 0) \tag{3}$$

so that in this case we get the trivial result that Q is (locally) identified at $P = p$ and $\sigma U = 0$.

Finally, we shall demonstrate in an example that monotonicity is easily violated in this model.

Example 3: Let

$$Q(P, \sigma U) = P + b(P)\sigma U.$$

Then, if the support of U is equal to the real line and $b(p) \neq b(p')$ monotonicity is violated for $\sigma > 0$ as there exist realizations u and u' in the support of U so that $Q(p, \sigma u) > Q(p', \sigma u)$ while $Q(p, \sigma u') < Q(p', \sigma u')$. Let, w.l.o.g., $p' > p$. Then, the probability that monotonicity is violated is given by

$$\Pr(Q(p, \sigma U) > Q(p'\sigma U)) = \Pr(p + b(p)\sigma U > p' + b(p')\sigma U).$$

If $\sigma = 0$ this is equal to 0 as $p < p'$. If $\sigma > 0$ it is

$$\Pr\left(U > \frac{p' - p}{\sigma(b(p) - b(p'))}\right).$$

If the distribution of U is symmetric about 0 this tends to 0.5 as σ tends to infinity. In the terminology of Angrist, Imbens, and Rubin (1996) this is the fraction of defiars. Finally, observe that the bigger $p' - p$ the "stronger" the instrument and the more likely monotonicity is to hold

in this example. □

2.2 Structural Parameters of Interest

A variety of structural parameters of interest can be expressed in terms of the marginal treatment effect[8]

$$m(v) \equiv \mathbb{E}[Y_1 - Y_0 | V = v]. \tag{4}$$

The marginal treatment effect by itself is of economic interest in many applications.[9] In this paper, we focus on the bias in estimates of the marginal treatment effect, the population average treatment effect,

$$\Delta^{\text{ATE}} \equiv \mathbb{E}[Y_1 - Y_0] = \int_0^1 m(v)\, dv, \tag{5}$$

and the local average treatment effect, for $v_l < v_h$,

$$\Delta^{\text{LATE}}(v_l, v_h) \equiv \mathbb{E}[Y_1 - Y_0 | v_l \leq V \leq v_h] = \frac{1}{v_h - v_l} \int_{v_l}^{v_h} m(v)\, dv. \tag{6}$$

Our results extend easily to other average treatment effect parameters of interest because they can be expressed in terms of the marginal treatment effect, as it is discussed by HV.

Note that by Assumption 3(i) all parameters that are considered here exist.

2.3 Identification of Structural Parameters under Monotonicity

In this subsection we briefly review the identification results by HV and Imbens and Angrist (1994). In both of them structural parameters of interest are related to the expected value of

[8] See HV as well as Heckman, Urzua, and Vytlacil (2006) for a detailed discussion. Angrist, Graddy, and Imbens (2000) derive the marginal treatment effect as the limit form of the local average treatment effect and show that, conversely, the local average treatment effect is an average marginal treatment effect, though not the population average.

[9] For empirical studies of the returns to college education see Björklund and Moffitt (1987), Carneiro, Heckman, and Vytlacil (2005), Carneiro and Lee (2005), and Klein (2006). In this context, V has the interpretation of unobserved ability which both has an impact on the decision of whether to attend college and the return from doing so. The dependence of this return on unobserved ability is of central interest to policy makers.

the outcome conditional on the propensity score. We first turn to the former which is based on derivatives of this conditional expectation with respect to the propensity score. We then present the latter which is based on the conditional expectation itself.

2.3.1 Derivative Based Approach

We show that under monotonicity, i.e. $\sigma = 0$, the marginal treatment effect is identified at values of V which are limit points of the support of P.

Definition 1: *For any random variable A we call $\tilde{a}$ a* limit point of the support of *A if A has a continuous density in a neighborhood around $\tilde{a}$ which is bounded away from zero.*

Note that at $A = \tilde{a}$, if they exist, derivatives of expectations conditional on A are identified from observations.

Under Assumption 1 and 3 and Normalization 1(i) and 1(ii) the marginal treatment effect is identified at $V = p$ if $\sigma = 0$ and p is a limit point of the support of $Q(P, 0)$. This, by (3), is equivalent to the requirement that p is a limit point of the support of P since $Q(p, 0) = p$. To see that under this condition the marginal treatment effect is identified write

$$\mathbb{E}[Y|Q(P,0) = p] = \mathbb{E}[Y_0] + \int_0^p m(v)\, dv, \tag{7}$$

where the integral is equal to

$$p \cdot \mathbb{E}[Y_1 - Y_0|D = 1, Q(P,0) = p] = p \cdot \mathbb{E}[Y_1 - Y_0|V \leq p] = p \cdot \int_0^p m(v)/p\, dv$$

noting that the density of V conditional on $V \leq p$ is $1/p$.

$\mathbb{E}[Y|Q(P,0) = p]$ is differentiable with respect to p since, by Assumption 3 (i), m is inte-

grable with respect to V. Differentiating both sides of (7) with respect to p yields

$$\frac{\partial \mathbb{E}[Y|Q(P,0) = p]}{\partial p} = m(p) \tag{8}$$

by Leibnitz' rule. By (3), the left hand side is equal to $\partial \mathbb{E}[Y|P = p]/\partial p$ and is identified from observations at limit points p of P so that $m(p)$ is identified.

If all p in the open interval $(0, 1)$ are limit points of the support of P the average treatment effect is identified via (5) because it is given by the integral over marginal treatment effects, noting that the probability of V being either 0 or 1 is equal to zero and first moments are finite. This might be the case if a strong continuously distributed instrument is available. Similarly, by (6) the local average treatment effect between p_l and p_h is identified if all p in the open interval (p_l, p_h) are limit points of the support of P.

2.3.2 Level Based Approach

The (local) average treatment effect can also be identified from observations under weaker support conditions. Specifically, let p_l and p_h be two points of support of P with $p_l < p_h$. Imbens and Angrist (1994) show that under Assumption 1 and 3(i), if $\sigma = 0$,

$$\frac{\mathbb{E}[Y|Q(P,0) = p_h] - \mathbb{E}[Y|Q(P,0) = p_l]}{p_h - p_l} = \Delta^{\text{LATE}}(p_l, p_h). \tag{9}$$

Taking limits for $p_l \to p_h$ shows that (9) directly corresponds to (8).

Finally, observe that for the average treatment effect, which is the local average treatment effect for $p_l = 0$ and $p_h = 1$, to be identified from levels we need that 0 and 1 are in the support of P. This might be a reasonable assumption in the presence of eligibility rules (Battistin and Rettore, forthcoming) and mandatory participation.

3 The Impact of Deviations from Monotonicity

In this section we study the impact of local departures from monotonicity on derivative and level based estimates of structural parameters. The generalized selection model that was developed above is central to this analysis. In particular, monotonicity holds if $\sigma = 0$. A local departure from monotonicity is given by a change from $\sigma = 0$ to a small $\sigma > 0$.

We derive approximations to the bias terms by performing second order Taylor series expansions in σ about $\sigma = 0$.

As for notation, partial derivatives of a function $f(a)$ with respect to its argument evaluated at $a = 0$ are denoted by $\partial f(0)/\partial a$. Second and third partial derivatives as well as cross derivatives are denoted accordingly. The approximations will be derived under the following differentiability condition.

Assumption 4 (Differentiability): *(i) $Q(P, \sigma U)$ and $\partial Q(P, \sigma U)/\partial P$ are twice continuously differentiable in σU around $\sigma U = 0$ and (ii) $m(V)$ is three times continuously differentiable.*

The approximation involves a second order approximation of $Q(P, \sigma U)$ at $P = p$ in σ about $\sigma = 0$,

$$Q(p, \sigma U) = Q(p, 0) + (\sigma U) \cdot \frac{\partial Q(p, 0)}{\partial(\sigma U)} + (\sigma U)^2/2 \cdot \frac{\partial^2 Q(p, 0)}{\partial(\sigma U)^2} + o(\sigma^2). \tag{10}$$

Proposition 1 provides conditions on Q under which monotonicity is violated in the absence of an approximation error and relates the fraction of observations for which monotonicity is violated to σ.

Proposition 1: *Let Normalization 1(ii), Assumption 2 and 4(i) hold. Moreover, let*

$$Q(P, \sigma U) = Q(P, 0) + (\sigma U) \cdot \frac{\partial Q(P, 0)}{\partial(\sigma U)} + (\sigma U)^2/2 \cdot \frac{\partial^2 Q(P, 0)}{\partial(\sigma U)^2}.$$

(i) Then, for two values $p < p'$ of P with $\frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} \neq \frac{\partial^2 Q(p',0)}{\partial(\sigma U)^2}$ monotonicity is violated if $\sigma > 0$ and

$$\left(\frac{\partial Q(p',0)}{\partial \sigma U} - \frac{\partial Q(p,0)}{\partial \sigma U}\right)^2 - 2 \cdot \left(p' - p\right) \cdot \left(\frac{\partial^2 Q(p',0)}{\partial(\sigma U)^2} - \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}\right) > 0. \tag{11}$$

(ii) Under this condition, the fraction of observations with $Q(p, \sigma U) > Q(p', \sigma U)$ approaches 1 as σ tends to infinity. (iii) If U is uniformly distributed this fraction is strictly increasing in σ.

Observe from (11) that for a given σ, if $\partial Q(P, \sigma U)/\partial(\sigma U)$ depends on P while $\partial^2 Q(P, \sigma U)/\partial(\sigma U)^2$ is sufficiently small, monotonicity is violated. This is similar to the observation in Example 3, namely that for $\partial^2 Q(P, \sigma U)/\partial(\sigma U)^2 = 0$ monotonicity is always violated. The second and third part of the proposition show that it is reasonable to use σ as a measure for the degree of the violation of monotonicity.

Under Normalization 1(iii)[10],

$$Var(Q(P, \sigma U | P = p)) = \sigma^2 \cdot \left(\frac{\partial Q(p,0)}{\partial(\sigma U)}\right)^2 + o(\sigma^2).$$

For the ease of the exposition we denote the approximation to the left hand side by

$$\sigma_p^2 \equiv \sigma^2 \cdot \left(\frac{\partial Q(p,0)}{\partial(\sigma U)}\right)^2. \tag{12}$$

3.1 Bias of Derivative Based Estimates

The main result is summarized in the following proposition.

[10] Under Assumption 2, $Var(Q(P, \sigma U)|P = p) = Var(Q(p, \sigma U))$. From (10) we have

$$\begin{aligned} Var(Q(p, \sigma U)) &= Var(Q(p,0) + \partial Q(p,0)/\partial(\sigma U) \cdot (\sigma U) + \partial^2 Q(p,0)/\partial(\sigma U)^2 \cdot (\sigma U)^2/2) + o(\sigma^2) \\ &= Var(\partial Q(p,0)/\partial(\sigma U) \cdot (\sigma U)) + o(\sigma^2), \end{aligned}$$

where for the second equality we perform a second order Taylor series expansion in σ about $\sigma = 0$ and let multiples of σ^3 and σ^4 enter the remainder term.

Proposition 2: *Let the selection model be given by* (1) *and let p be a limit point of the support of P. Then, under Assumptions 1-4 and Normalization 1 the bias of the derivative based estimate of* $m(p)$, $\partial\mathbb{E}[Y|P=p]/\partial P$, *is given by*

$$B^{MTE*}(p) = \frac{1}{2}\cdot\sigma_p^2\cdot\frac{\partial^2 m(p)}{\partial p^2} + \frac{1}{2}\cdot\frac{\partial\sigma_p^2}{\partial p}\cdot\frac{\partial m(p)}{\partial p} + o(\sigma^2). \tag{13}$$

Proof. Appendix. □

The approximation to the bias term consists of two parts. The first part is given by the product of the variance of $Q(P,\sigma U)$ conditional on $P = p$ and the second derivative of the marginal treatment effect at $V = p$. This shows that instead of the marginal treatment effect at $V = p$, a weighted average of marginal treatment effects is estimated. The second part of the bias term arises because the conditional variance of Q depends on p. From the formula in Proposition 2 the bias in derivative based estimates of the average and local average treatment effect can be obtained by integrating over values of p, as suggested by (5) and (6).

Corollary 2.1: *Let the selection model be given by* (1) *and let all* $p \in (p_l, p_h)$ *be limit points of the support of P. Then, under Assumptions 1-4 and Normalization 1 the bias of the derivative based estimate of the local average treatment effect between* p_l *and* p_h *is given by*

$$B^{LATE*}(p_l,p_h) = \frac{1}{2}\cdot\frac{1}{p_h-p_l}\cdot\left(\sigma^2_{p_h}/2\cdot\frac{\partial m(p_h)}{\partial V} - \sigma^2_{p_l}/2\cdot\frac{\partial m(p_l)}{\partial V}\right) + o(\sigma^2). \tag{14}$$

If $p_l = 0$ *and* $p_h = 1$ *this is the bias in the derivative based estimate of the average treatment effect,* B_D^{ATE*}.

Proof. Appendix. □

3.2 Bias in Level Based Estimates

We can prove that the bias in level based estimates of treatment effect parameters is equal to the bias in derivative based estimators.

Proposition 3: *Let the selection model be given by* (1) *and let* p_l *and* p_h *be in the support* P*. Then, under Assumptions 1-4 and Normalization 1 the bias of the level based estimate of the local average treatment effect between* p_l *and* p_h*,*

$$\frac{\mathbb{E}[Y|P = p_h] - \mathbb{E}[Y|P = p_l]}{p_h - p_l},$$

is equal to the bias of the derivative based estimate in (14)*. If* $p_l = 0$ *and* $p_h = 1$ *this is again the bias in level based estimates of the average treatment effect.*

Proof. Appendix. □

3.3 Practical Relevance

Under appropriate conditions, the variance of Q conditional on P can be estimated (Ichimura and Thompson, 1998). Then, a bias correction procedure in which we substitute biased estimates for the first and second derivative of the marginal treatment effect is feasible in the sense that the order of the approximation error remains unchanged. This is the case because the approximations to the bias terms are multiples of σ^2 and the order of the approximation error is $o(\sigma^2)$.

If the variance of Q conditional on P is unknown, a sensitivity analysis can be undertaken by calculating the approximation to the bias term for different values of σ_p^2 and $\partial\sigma_p^2/\partial p$.

In general, our analysis has shown that the curvature of the marginal treatment effect determines the magnitude of the bias when monotonicity does not hold. As a rule of thumb, we have that the less curved the marginal treatment effect is in its argument, the less biased estimates are

when monotonicity does not hold. A particularly interesting result is that the bias is of order $o(\sigma^2)$ if the marginal treatment effect is linear in $V = v$ This is a testable condition on the data generating process because it implies that $\mathbb{E}[Y|P = p]$ is quadratic in p.

In fact, this observation yields identifying conditions which do not involve monotonicity but allow for selection on unoberservables. We summarize this finding in a proposition.[11]

Proposition 4: *Let the marginal treatment effect be linear in v so that*

$$\mathbb{E}[Y|Q(P, \sigma U) = q] = \alpha + \beta q + \gamma q^2 \tag{15}$$

for some constants α, β, γ. Moreover, let Assumptions 1-3 hold and let the variance of $Q(P, \sigma U)$ conditional on P be equal to $\tilde{\sigma}$. Then, the marginal, average, and local average treatment effect are identified if the the support of the propensity score contains at least three points.

Proof. By (15),

$$\begin{aligned}
\mathbb{E}[Y|P = p] &= \alpha + \beta \mathbb{E}[Q(P, \sigma U)|P = p| + \gamma \mathbb{E}[Q(P, \sigma U)^2|P = p| \\
&= \alpha + \beta p + \gamma \left(\tilde{\sigma} + p^2 \right) \\
&= \tilde{\alpha} + \beta p + \gamma p^2,
\end{aligned}$$

where $\tilde{\alpha} = \alpha + \gamma \tilde{\sigma}$. Thus, β and γ are identified from observations by the support condition. Consequently, the marginal treatment effect, which is given by the derivative of (15) with respect to q,

$$m(q) = \beta + 2\gamma q,$$

is identified since it is a function of β and γ and can be evaluated at values q. Consequently, the average and local average treatment effect are identified by the relationships (5) and (6). □

[11] As before, we can think of the exposition here as being conditional on exogenous covariates.

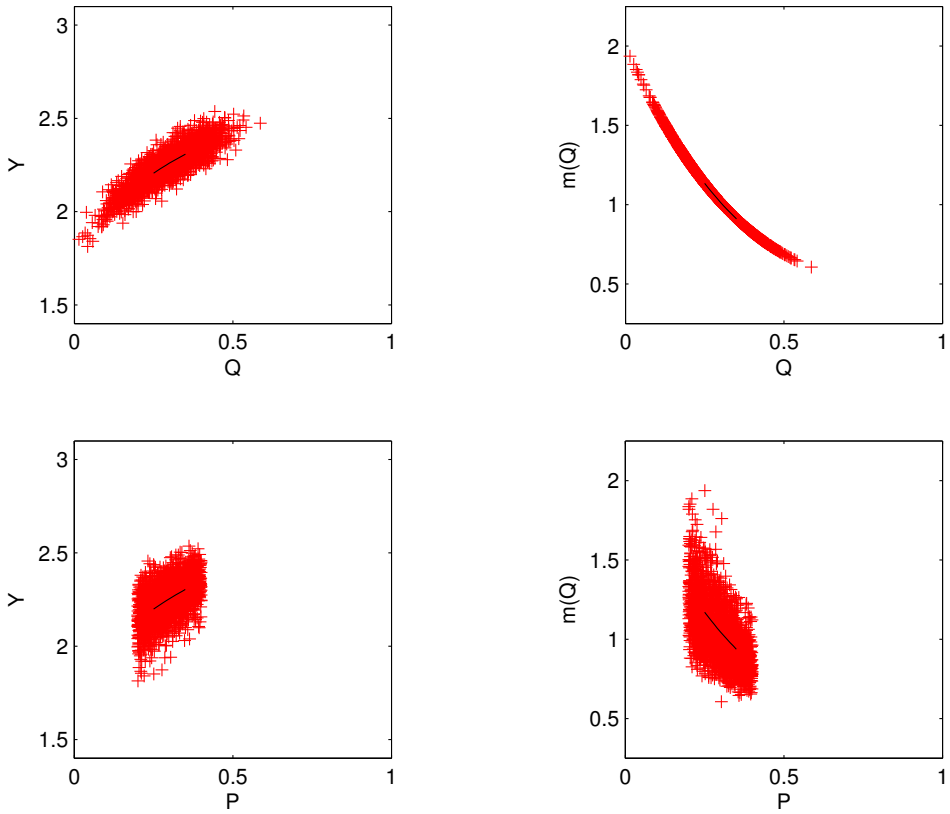

Figure 2: One draw of generated data for $\rho = 3$.

4 Monte Carlo

We simulated data in order to characterize the bias that arises from a violation of the monotonicity assumption as well as the accuracy of our analytical approximation to the bias term. For $R = 1.000$ repetitions and values of a curvature parameter ρ we generated $N = 2.000$ data points. Since biases of estimates of average treatment effect parameters are functions of biases of estimates of the marginal treatment effect we focus on the respective mean bias in estimates of the marginal treatment effect as a function of the curvature ρ.

Specifically, we drew values of P and V from a uniform distribution, with respective support $[0.2, 0.4]$ and $[0, 1]$. Values of U were drawn from a standard normal distribution and $\sigma = 0.1$. We let $Q = P + \sigma U - P \cdot \sigma U$. Next, we calculated $D = \mathbb{1}\{Q \geq V\}$. Notice that by construction, monotonicity of the treatment decision in Q holds whereas monotonicity in P is violated. In the

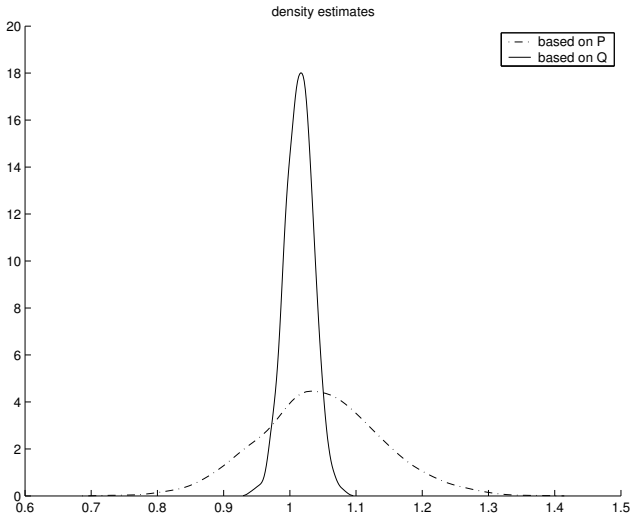

Figure 3: Distribution of estimates of $m(0.3)$ based on P and Q. $\rho = 3$. The true value is 1.0145.

spirit of the empirical results in Carneiro, Heckman, and Vytlacil (2005) we let

$$Y = 2.2 + 0.5Q - \frac{(1-Q)^{\rho+1}}{\rho+1} + \varepsilon,$$

where ε was drawn from a normal distribution with mean 0 and variance 0.05. In their application, the treatment decision is whether to attend college or not. For $\rho = 2$ our simulations yield data similar to theirs. Since monotonicity in Q holds, we get by (8) that the marginal treatment effect at $V = q$ is given by the derivative of $\mathbb{E}[Y|Q]$ with respect to Q, evaluated at $Q = q$:

$$m(q) = 0.5 + 1.5(1-q)^{\rho}.$$

It is decreasing in q. The second and third derivative are $\partial m(q)/\partial V = -1.5\rho(1-q)^{\rho-1}$ and $\partial^2 m(q)/\partial V^2 = 1.5\rho(\rho-1)(1-q)^{\rho-2}$, respectively. Observe that for $\rho = 1$ the marginal treatment effect is linear in q whereas for $\rho > 1$ it is a convex function of q. For low values of q, e.g. $q = 0.3$ this function is the more convex the higher ρ.

For this Monte Carlo, estimates of the marginal treatment effect as well as its first and second

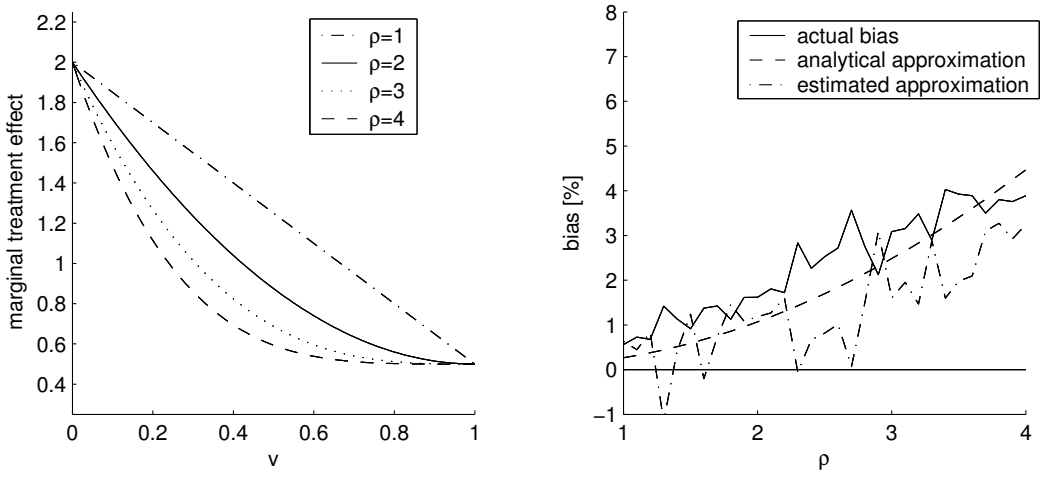

Figure 4: The marginal treatment effect for different values of ρ on the left and the bias as well as the accuracy of the approximation as a function of ρ, for $V = 0.3$, on the right. Respective sample means across $1,000$ repetitions were calculated.

derivative were obtained by fitting a third order polynomial to the data.

Figure 2 shows one draw of generated data for $\rho = 3$. Solid lines are estimated means over all repetitions. On the left, values of Y are plotted against values of P and Q. On the right, the marginal treatment effect evaluated at values of Q, $m(Q)$, is plotted against values of P and Q. Obviously, when we plot $m(Q)$ against Q we get the marginal treatment effect itself. However, plotting $m(Q)$ against P yields a distribution of marginal treatment effects for every P.

Figure 3 shows the distribution of estimates of $m(0.3)$ that are based on P and Q. Monotonicity of D with respect to Q holds by construction, whereas monotonicity of D with respect to P is violated. The figure shows that estimates based on P are in general upward biased and more dispersed.

In Section 3 we have shown that the bias which arises from a violation of the monotonicity assumption is the higher the more convex the marginal treatment effect is in v. Next, in Figure 4, we plot the dependence of the mean bias in estimates of the marginal treatment effect at $V = 0.3$ against ρ. Additionally, we plot the analytical approximation to the bias term, where we treat $m(V)$ as known, and the estimated approximation, where we use an estimate of the

second derivative of the marginal treatment effect at $V = 0.3$ that is based on P. In both cases, the sample variance of Q conditional on P and the derivative thereof with respect to P have been used. Figure 4 demonstrates that the approximation is reasonably accurate.

5 Concluding Remarks

This paper has provided a formal analysis of the consequences of a violation of the monotonicity assumption. Approximations to respective bias terms have been derived. They are functions of features of the underlying structure: a measure for the degree of violation of the monotonicity assumption and the marginal treatment effect. In general, we find that estimates are the more sensitive to violations of monotonicity the more curvature the marginal treatment effect exhibits in V. This analytical result was illustrated in a Monte Carlo study.

The bias can be estimated from the data up to a parameter σ_p and $\partial\sigma_p/\partial p$ without changing the order of the approximation error. Therefore, our results have practical relevance which we summarize in the following three points. First, a bias correction procedure is available if the researcher is willing to make additional assumptions in order to estimate σ_p and $\partial\sigma_p/\partial p$. Second, a sensitivity analysis can be carried out by calculating the bias for different values of σ_p and $\partial\sigma_p/\partial p$. Finally, whenever the marginal treatment effect is linear in V, the bias is of order $o(\sigma^2)$ if the variance of Q conditional on P is constant across values of P. The former condition is testable whereas, as before, the latter condition can be tested if the researcher is willing to make additional assumptions.

Appendix: Proofs

Proposition 1

Proof. We first prove (i). By Normalization 1(ii) and Assumption 2(i) we can, as in (3), replace $Q(p,0)$ and $Q(p',0)$ by p and p', respectively. Monotonicity is violated if the sign of $Q(p',\sigma u) - Q(p,\sigma u)$ depends on u. This is the case if the second order polynomial in u,

$$\begin{aligned} Q(p',\sigma u) - Q(p,\sigma u) = p' + (\sigma u)\cdot\frac{\partial Q(p',0)}{\partial(\sigma U)} + (\sigma u)^2/2\cdot\frac{\partial^2 Q(p',0)}{\partial(\sigma U)^2} \\ -p - (\sigma u)\cdot\frac{\partial Q(p,0)}{\partial(\sigma U)} - (\sigma u)^2/2\cdot\frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}, \end{aligned} \tag{16}$$

has more than 1 root. As $\frac{\partial^2 Q(p',0)}{\partial(\sigma U)^2} \neq \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}$ the roots are given by the quadratic formula as

$$\begin{aligned} u_{1,2} = -\sigma\left(\frac{\partial Q(p',0)}{\partial(\sigma U)} - \frac{\partial Q(p,0)}{\partial(\sigma U)}\right) \\ \pm \frac{\sqrt{\left(\frac{\partial Q(p',0)}{\partial(\sigma U)} - \frac{\partial Q(p,0)}{\partial(\sigma U)}\right)^2 - 2\cdot(p'-p)\cdot\left(\frac{\partial^2 Q(p',0)}{\partial(\sigma U)^2} - \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}\right)}}{\sigma\left(\frac{\partial^2 Q(p',0)}{\partial(\sigma U)^2} - \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}\right)}. \end{aligned} \tag{17}$$

There exists more than 1 root if the discriminant is positive, i.e.

$$\left(\frac{\partial Q(p',0)}{\partial\sigma U} - \frac{\partial Q(p,0)}{\partial\sigma U}\right)^2 - 2\cdot\left(p'-p\right)\cdot\left(\frac{\partial^2 Q(p',0)}{\partial(\sigma U)^2} - \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}\right) > 0,$$

the desired result.

(ii) Relabel the roots so that $u_2 > u_1$. We look for values u for which the difference in (16) is negative. Observe that it is positive for $u = 0$. Therefore, for $u < u_1$ and $u > u_2$ (16) is negative if there exist two roots. Then, the fraction of observations for which monotonicity does not hold is given by

$$1 - (F_U(u_2) - F_U(u_1)). \tag{18}$$

As σ tends to infinity the distance between u_1 and u_2 approaches 1. By Assumption 2(ii) F_U is

continuous so that $\lim_{\sigma\to\infty}(F_U(u_2) - F_U(u_1)) = 0$.

(iii) If U is uniformly distributed we have $F_U(u_1) = u_1$ and $F_U(u_2) = u_2$. Then, (18) becomes $1 - (u_2 - u_1)$ which, by (17), is increasing in σ. □

Proposition 2

We prove Proposition 2 using Lemma 1. Observe that all Taylor series expansions can be performed by the differentiability conditions in Assumption 4.

LEMMA 1: *Under Assumptions 1-4 and Normalization 1*

$$\begin{aligned}
&\frac{\partial}{\partial p}\mathbb{E}[Y|P=p] \\
&\quad = m(Q(p,0)) \\
&\qquad + \sigma^2/2 \cdot \frac{\partial^2 m(Q(p,0))}{\partial V^2} \cdot \left(\frac{\partial Q(p,0)}{\partial(\sigma U)}\right)^2 \\
&\qquad + \sigma^2/2 \cdot \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} \\
&\qquad + \sigma^2 \cdot \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\,\partial P} + o(\sigma^2).
\end{aligned}$$

Proof. The proof is in 5 steps.

First step: Recall from (10) that a second order Taylor series expansion of $Q(p,\sigma u)$ in σ about $\sigma = 0$ yields

$$Q(p,\sigma u) = Q(p,0) + \sigma u \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} + (\sigma u)^2/2 \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} + o(\sigma^2). \tag{19}$$

Second step: By Assumption 2 and Normalization 1(ii) and 1(iii)

$$p = \mathbb{E}[Q(P,\sigma U)|P=p] = \mathbb{E}[Q(p,\sigma U)] = Q(p,0) + \sigma^2/2 \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} + o(\sigma^2). \tag{20}$$

Combining this with (19) yields

$$Q(p,\sigma u) = p + \sigma u \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} + \sigma^2/2 \cdot (u^2-1) \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} + o(\sigma^2) \tag{21}$$

and

$$\frac{\partial}{\partial p} Q(p,\sigma u) = 1 + \sigma u \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\,\partial P} + \sigma^2/2 \cdot (u^2-1) \frac{\partial^3 Q(p,0)}{\partial(\sigma U)^2\,\partial P} + o(\sigma^2). \tag{22}$$

Third step: A second order Taylor series expansion of $\partial \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)]/\partial Q(P,\sigma U)$ in σ about $\sigma = 0$ yields

$$\begin{aligned}
&\frac{\partial}{\partial Q(P,\sigma U)} \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)] \\
&= m(Q(p,\sigma u)) \\
&= m(Q(p,0)) \\
&\quad + \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} \cdot \sigma u \\
&\quad + \frac{\partial^2 m(Q(p,0))}{\partial V^2} \cdot \left(\frac{\partial Q(p,0)}{\partial(\sigma U)} \right)^2 \cdot (\sigma u)^2/2 \\
&\quad + \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} \cdot (\sigma u)^2/2 + o(\sigma^2),
\end{aligned} \tag{23}$$

where the first equality uses (8).

Fourth step: We have

$$\begin{aligned}
&\frac{\partial}{\partial P} \mathbb{E}[Y|P=p] \\
&\quad = \frac{\partial}{\partial P} \mathbb{E}[\mathbb{E}[Y|Q(P,\sigma U)]|P=p] \\
&\quad = \frac{\partial}{\partial P} \mathbb{E}[\mathbb{E}[Y|Q(p,\sigma U)]]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial P} \int \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)]\ f_U(u)\ du \\
&= \int \frac{\partial}{\partial P} \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)]\ f_U(u)\ du \\
&= \int \frac{\partial}{\partial Q(P,\sigma U)} \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)] \cdot \frac{\partial}{\partial P} Q(p,\sigma u)\ f_U(u)\ du \\
&= \int \frac{\partial}{\partial Q(P,\sigma U)} \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)] \cdot \left\{ 1 + \sigma u \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\ \partial P} \right. \\
&\qquad \left. + \sigma^2/2 \cdot (u^2 - 1) \cdot \frac{\partial^3 Q(p,0)}{\partial(\sigma U)^2\ \partial P} \right\}\ f_U(u)\ du + o(\sigma^2),
\end{aligned}
$$

where the first equality is by iterated expectations, the second follows from Assumption 2(i), the third from Assumption 2(ii), the fourth from the integrand being finite (Assumption 3(i)), the fifth applies the chain rule, and the sixth uses (22).

Together with (23), this is

$$
\begin{aligned}
&\frac{\partial}{\partial P} \mathbb{E}[Y|P = p] \\
&= \int \left\{ \frac{\partial}{\partial Q(P,\sigma U)} \mathbb{E}[Y|Q(P,\sigma U) = Q(p,0)] \right. \\
&\qquad + \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} \cdot \sigma u \\
&\qquad + \frac{\partial^2 m(Q(p,0))}{\partial V^2} \cdot \left(\frac{\partial Q(p,0)}{\partial(\sigma U)} \right)^2 \cdot (\sigma u)^2/2 \\
&\qquad \left. + \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} \cdot (\sigma u)^2/2 \right\} \\
&\quad \times \left\{ 1 + \sigma u \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\ \partial P} + \sigma^2/2 \cdot (u^2 - 1) \cdot \frac{\partial^3 Q(p,0)}{\partial(\sigma U)^2\ \partial P} \right\} \\
&\quad \times f_U(u)\ du + o(\sigma^2)
\end{aligned}
$$

and this in turn is

$$
\begin{aligned}
= \int \Bigg\{ & \frac{\partial}{\partial Q(P,\sigma U)} \mathbb{E}[Y|Q(P,\sigma U) = Q(p,0)] \\
& + \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} \cdot \sigma u \\
& + \frac{\partial^2 m(Q(p,0))}{\partial Q(P,\sigma U)^2} \cdot \left(\frac{\partial Q(p,0)}{\partial(\sigma U)} \right)^2 \cdot (\sigma u)^2/2 \\
& + \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} \cdot (\sigma u)^2/2 \\
& + \frac{\partial \mathbb{E}[Y|Q(P,\sigma U) = Q(p,0)]}{\partial Q(P,\sigma U)} \cdot \sigma u \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\, \partial P} \\
& + \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} \cdot \sigma u \cdot \sigma u \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\, \partial P} \\
& + \frac{\partial \mathbb{E}[Y|Q(P,\sigma U) = Q(p,0)]}{\partial Q(P,\sigma U)} \cdot \sigma^2/2 \cdot (u^2 - 1) \cdot \frac{\partial^3 Q(p,0)}{\partial(\sigma U)^2\, \partial P} \Bigg\} \\
& \times f_U(u)\, du + o(\sigma^2),
\end{aligned}
$$

where we already let multiples of σ^2 enter the remainder term. By Normalization 1(iii), $\mathbb{E}[U] = 0$ and $\mathbb{E}[U^2] = 1$, this is

$$
\begin{aligned}
& \frac{\partial}{\partial P} \mathbb{E}[Y|P = p] \\
& \quad = \frac{\partial}{\partial Q(P,\sigma U)} \mathbb{E}[Y|Q(P,\sigma U) = Q(p,0)] \\
& \qquad + \sigma^2/2 \cdot \frac{\partial^2 m(Q(p,0))}{\partial V^2} \cdot \left(\frac{\partial Q(p,0)}{\partial(\sigma U)} \right)^2 \\
& \qquad + \sigma^2/2 \cdot \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} \\
& \qquad + \sigma^2 \cdot \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\, \partial P} + o(\sigma^2).
\end{aligned}
$$

This completes the proof of Lemma 1. □

Proof of Proposition 2. (20) implies that

$$m(p) = m\left(Q(p,0) + \sigma^2/2 \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}\right) + o(\sigma^2).$$

A second order Taylor series expansion thereof in σ about $\sigma = 0$ yields

$$m(p) = m(Q(p,0)) + \sigma^2/2 \cdot \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} + o(\sigma^2). \tag{24}$$

Moreover, (12) implies

$$\frac{\partial \sigma_p^2}{\partial p} = 2\sigma^2 \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\,\partial p}. \tag{25}$$

From (24) and Lemma 1 we have that

$$\begin{aligned}
&\frac{\partial}{\partial p}\mathbb{E}[Y|P=p] - m(p) \\
&\quad = \sigma^2/2 \cdot \frac{\partial^2 m(Q(p,0))}{\partial V^2} \cdot \left(\frac{\partial Q(p,0)}{\partial(\sigma U)}\right)^2 \\
&\qquad + \sigma^2 \cdot \frac{\partial m(Q(p,0))}{\partial V} \cdot \frac{\partial Q(p,0)}{\partial(\sigma U)} \cdot \frac{\partial^2 Q(p,0)}{\partial(\sigma U)\,\partial P} + o(\sigma^2).
\end{aligned}$$

We get the result using (12) and (25). □

Proof of Corollary 2.1. By (6) and (13)

$$\begin{aligned}
B_D^{\text{LATE}*}(p_l, p_h) &= \frac{1}{p_h - p_l} \int_{p_l}^{p_h} \frac{1}{2} \cdot \sigma_p^2 \cdot \frac{\partial^2 m(p)}{\partial p^2} + \frac{1}{2} \cdot \frac{\partial \sigma_p^2}{\partial p} \cdot \frac{\partial m(p)}{\partial p}\, dp + o(\sigma^2) \\
&= \frac{1}{p_h - p_l}\left[\frac{1}{2} \cdot \sigma_p^2 \cdot \frac{\partial m(p)}{\partial V}\right]_{p=p_l}^{p_h}.
\end{aligned}$$

This yields the result. □

Proposition 3

Proof of Proposition 3. The proof consists of 3 steps.

First step: We have

$$
\begin{aligned}
&\mathbb{E}[Y|P=p] \\
&= \mathbb{E}\Big[\mathbb{E}[Y|Q(P,\sigma U)]\Big|P=p\Big] \\
&= \mathbb{E}\Big[\mathbb{E}[Y|Q(P,\sigma U)=Q(p,\sigma U)]\Big] \\
&= \int \mathbb{E}[Y|Q(P,\sigma U)=Q(p,\sigma u)]\; f_U(u)\; du \\
&= \int \mathbb{E}[Y|Q(P,\sigma U)=Q(p,0)+(\sigma u)\cdot\frac{\partial Q(p,0)}{\partial(\sigma U)}+(\sigma u)^2/2\cdot\frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}]\; f_U(u)\; du + o(\sigma^2)
\end{aligned}
$$

where the first equality is by iterated expectations, the second follows from Assumption 2(i), the third from Assumption 2(ii), and the forth uses (10). A second order Taylor series expansion in σ about $\sigma = 0$ yields that this is

$$
\begin{aligned}
\int \Big\{&\mathbb{E}[Y|Q(P,\sigma U)=Q(p,0)] \\
&+(\sigma u)\cdot m(Q(p,0))\cdot\frac{\partial Q(p,0)}{\partial(\sigma U)} \\
&+(\sigma u)^2/2\cdot m(Q(p,0))\cdot\frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2} \\
&+(\sigma u)^2/2\cdot\frac{\partial m(Q(p,0))}{\partial V}\cdot\left(\frac{\partial Q(p,0)}{\partial(\sigma U)}\right)^2\Big\} \\
&\times f_U(u)\; du + o(\sigma^2)
\end{aligned}
$$

and by Normalization 1(iii) this is equal to

$$
\begin{aligned}
&\mathbb{E}[Y|Q(P,\sigma U)=Q(p,0)]+\sigma^2/2\cdot m(Q(p,0))\cdot\left(\frac{\partial Q(p,0)}{\partial(\sigma U)}\right)^2 \\
&+\sigma^2/2\cdot\frac{\partial m(Q(p,0))}{\partial V}\cdot\frac{\partial Q(p,0)}{\partial(\sigma U)}\Big\}+o(\sigma^2).
\end{aligned}
\tag{26}
$$

Second step: By Assumption 2 and Normalization 1(ii) and 1(iii) we get (20) which implies that

$$\mathbb{E}[Y|Q(p,\sigma U)=p]=\mathbb{E}\left[Y\left|Q(P,\sigma U)=Q(p,0)+\sigma^2/2\cdot\frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}\right.\right]+o(\sigma^2).$$

A second order Taylor series expansion thereof in σ about $\sigma=0$ yields

$$\mathbb{E}[Y|Q(P,\sigma U)=p]=\mathbb{E}[Y|Q(P,\sigma U)=Q(p,0)]+\sigma^2/2\cdot m(Q(p,0))\cdot\frac{\partial^2 Q(p,0)}{\partial(\sigma U)^2}+o(\sigma^2), \quad (27)$$

Third step: Combining (3), (26) and (27) we get

$$\mathbb{E}[Y|Q(P,\sigma U)=p]-\mathbb{E}[Y|P=p]=\sigma^2/2\cdot\frac{\partial m(p)}{\partial V}\cdot\left(\frac{\partial Q(p,0)}{\partial(\sigma U)}\right)^2\Big\}. \quad (28)$$

Using this with (9), (12) and (25) yields the result. □

Acknowledgements

I am especially grateful to Andrew Chesher and Enno Mammen for numerous insightful discussions. Moreover, I would like to thank Jaap Abbring, Josh Angrist, Pedro Carneiro, Jim Heckman, Pierre Hoonhout, Jürgen Maurer, Enrico Rettore, Cristina Santos, Ed Vytlacil and seminar participants at University College London as well as conference participants of the 2004 European Meeting of the Econometric Society in Madrid, the 2006 Far Earstern Meeting of the Econometric Society in Beijing and the 2006 European Meeting of the Econometric Society in Vienna for helpful comments. Finally, I would like to thank the Department of Economics at University College London for its hospitality during the academic year 2003/4, the European Commission for financial support through the Marie Curie program and the Deutsche Forschungsgemeinschaft for financial support through SFB/TR 15. The title of previous versions of this paper was "Heterogenous Treatment Effects: Local IV without Monotonicity?"

References

ANGRIST, J. D., K. GRADDY, AND G. W. IMBENS (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67(3), 499–527.

ANGRIST, J. D., G. W. IMBENS, AND D. B. RUBIN (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Society*, 91(434), 444–455.

BATTISTIN, E., AND A. CHESHER (2004): "The Impact of Measurement Error on Evaluation Methods Based on Strong Ignorability," Mimeograph.

BATTISTIN, E., AND E. RETTORE (forthcoming): "Ineligible and Eligible Non-Participants as a Double Comparison Group in Regression-Discontinuity Designs," *Journal of Econometrics*.

BJÖRKLUND, A., AND R. MOFFITT (1987): "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," *Review of Economics and Statistics*, 69(1), 42–49.

CARNEIRO, P., J. J. HECKMAN, AND E. VYTLACIL (2005): "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education," Mimeograph.

CARNEIRO, P., AND S. LEE (2005): "Ability, Sorting and Wage Inequality," c*e*mmap Working Paper CWP 16/05, University College London.

CHESHER, A. (1991): "The effect of measurement error," *Biometrika*, 78(3), 451–462.

CHESHER, A., AND J. M. C. SANTOS SILVA (2002): "Taste Variation in Discrete Choice Models," *Review of Economic Studies*, 69(1), 147–168.

CHESHER, A., AND C. SCHLUTER (2002): "Welfare Measurement and Measurement Error," *Review of Economic Studies*, 69(2), 357–378.

HECKMAN, J. J., S. URZUA, AND E. VYTLACIL (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

HECKMAN, J. J., AND E. J. VYTLACIL (1999): "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

——— (2000): "The Relationship between Treatment Parameters within a Latent Variable Framework," *Economics Letters*, 66(1), 33–39.

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

ICHIMURA, H., AND T. S. THOMPSON (1998): "Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution," *Journal of Econometrics*, 86(2), 269–295.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

KIEFER, N., AND G. R. SKOOG (1984): "Local Asymptotic Specification Error Analysis," *Econometrica*, 52(4), 873–886.

KLEIN, T. J. (2006): "College Education and Wages in the U.K.: Estimating Conditional Average Structural Functions in Nonadditive Models with Binary Endogenous Variables," JEPS Working Paper 06-001.

ROSENBAUM, P. R., AND D. B. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

VYTLACIL, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.

——— (2006): "A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results," *Oxford Bulletin of Economics and Statistics*, 68(4), 515–518.