Heterogeneous treatment effects: instrumental variables without monotonicity?

Klein, Tobias J.

Heterogeneous treatment effects: Instrumental variables without monotonicity?

Tobias J. Klein

PII: S0304-4076(09)00219-X
DOI: 10.1016/j.jeconom.2009.08.006
Reference: ECONOM 3249

To appear in: *Journal of Econometrics*

Received date: 22 September 2008
Revised date: 30 April 2009
Accepted date: 10 August 2009

Heterogeneous Treatment Effects: Instrumental Variables without Monotonicity?

Tobias J. Klein[*]
Netspar, CentER, Tilburg University

August 5, 2009

Abstract

Imbens and Angrist (1994) were the first to exploit a monotonicity condition in order to identify a local 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. We investigate the sensitivity of respective estimates to random departures from monotonicity. Approximations to respective bias terms are derived. In an empirical application the bias is calculated and bias corrected estimates are obtained. The accuracy of the approximation is investigated in a Monte Carlo study.

JEL Classification: C21.

Keywords: Program evaluation, heterogeneity, identification, dummy endogenous variable, selection on unobservables, instrumental variables, monotonicity, nonseparable index selection model.

[*] *Address:* Tilburg University, Department of Econometrics and OR, PO Box 90153, 5000 LE Tilburg, The Netherlands. *E-Mail:* T.J.Klein@uvt.nl.

1 Introduction

1.1 Monotonicity

A fundamental identification problem in program evaluation arises if, condition on observables, the treatment decision depends on the idiosyncratic gain from participation. 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) suggested to exploit monotonicity of the treatment decision in instrumental variables in order to identify a local average treatment effect parameter. The instrumental variables are assumed to be independent of the pair of potential outcomes conditional on covariates. 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 an individual's treatment status, or changes it in the same direction as it does for all other individuals for which it has an impact. Under monotonicity observed exogenous variation in instrumental variables identifies the average treatment effect for the subset of individuals who would have been affected by such a change.

More recently, Heckman and Vytlacil (1999) proposed estimation of a variety of treatment effect parameters using a local version of this approach. Both approaches are in principle able to cope with selection on unobservables. They are intuitive, elegant, and easy to implement. Their generality consists of the fact that a parametric specification of the joint distribution of unobservables and observables is not needed. However, identification hinges on monotonicity and in general, estimates of treatment effect parameters will be biased if it does not hold. Therefore, assessing the sensitivity of monotonicity based estimates lends credibility to an empirical study if monotonicity cannot safely be assumed. In this paper, we assume that violations of monotonicity occur at random and derive results that can be used for such an assessment.

There are situations in which monotonicity holds naturally. Battistin and Rettore (2008) discuss the case in which eligibility for participation in a program is not related to the pair of potential outcomes conditional on covariates. Then, if we make the thought experiment of letting an individual become eligible for a treatment, this can either have no effect or it can change the treatment status from not being treated to being treated. Therefore, monotonicity holds by construction. This is a special case of the regression discontinuity design where discontinuous changes in the treatment probability are exploited (for a discussion see Hahn, Todd, and van der Klaauw, 2001).

While monotonicity holds naturally in some situations it can easily be violated in others. Such a violation is discussed in Example 2 of Imbens and Angrist (1994). There are two officials, A and B, who screen applicants for a social program. Suppose that applicants are randomly assigned to the officials. Then, it is unlikely that the identity of the official affects the outcome of participation or nonparticipation in the program. Suppose that at the same time the officials have some discretion when making the treatment decision. Then, it could be that the admission rate for B is higher than for A. Under these conditions the identity qualifies as an instrument because it is related to the treatment decision but unrelated to the potential outcomes.

In this example, monotonicity holds whenever *any* applicant who would have been accepted by official A *is* accepted by official B. It is violated if official A sometimes accepts individuals who would not have been accepted by official B. This could be because A, unlike B, values an unobserved characteristic of the applicant that is in fact unrelated to the outcome. An example of such a characteristic could be companionableness for a blue collar worker. Another reason why monotonicity could fail is because A makes a mistake and erroneously accepts the applicant. In both cases, from the point of view of the econometrician, *additional randomness* causes monotonicity to fail (both the characteristic and the errors are unobserved). In this paper we will assume that this additional randomness is not related to the potential outcomes.[1] However, we still allow for a dependence between the treatment decision and the potential outcomes.

Typically, only the treatment decision that has actually been made is observed, and therefore violations of monotonicity are not directly observable. Korn and Baumrind (1998) study recommendations of orthodontists to extract a tooth. Interestingly, they asked two orthodontists about their assessments and are thus able to provide direct evidence for a failure of monotonicity.[2] As before we can think of two orthodontists, say A and B. B has a higher extraction rate. Monotonicity holds if the set of patients for which A recommends extraction is a subset of the set of patients for which B recommends extraction, or equivalently if the set of observed *and unobserved* characteristics of the applicants for which A recommends extraction is a subset of the set of characteristics of the applicants for which B recommends extraction. Again, monotonicity could fail if an unobserved patient characteristic affects the decision of one orthodontist but not of the other. E.g., B might not recommend to extract a tooth because a patient is very sensitive to pain, while A does not base his decision on this, and therefore recommends extraction.[3] Monotonicity fails in that case because on average B recommends extraction more often than A. Another reason for a violation of monotonicity could be that clinicians are not always consistent when making recommendations. Korn, Teeterb, and Baumrind (2001, Table 3) provide evidence for this. In particular, they find that the stated treatment preference of a clinician is not stable, meaning that the same clinician's treatment recommendation at a later point in time does not agree with his earlier one. Monotonicity is violated if for that reason on one day only clinician A recommends the treatment for a given patient and on another day only clinician B recommends the treatment for a patient with the exact same observed and unobserved characteristics. Again, the assumption in this paper will be that the additional randomness that causes monotonicity to fail, which arises because B takes sensitivity to pain into account or because of self-disagreement, is unrelated to the success of the treatment.

In this paper, the sensitivity of monotonicity based estimates of treatment effect parameters to departures from monotonicity is investigated. The main assumption is that violations of monotonicity occur at random. Approximations to bias terms that depend on estimable quantities are derived. The practical relevance of the results is illustrated in an empirical application and

[1] There is a parallel to the literature on measurement error. The model here corresponds to a model with classical measurement error. It would be interesting to relax this assumption, as it is to study non-classical measurement error, but this is beyond the scope of the paper.

[2] This is the leading example in Small and Tan (2007). They show that under a weaker stochastic monotonicity assumption the sign of the instrumental variables estimator is equal to the sign of the treatment effect if this sign is the same for all individuals.

[3] I am grateful to a referee for coming up with this example.

the accuracy of the approximations is investigated in a Monte Carlo study.

1.2 Selection Models

Let D be a binary treatment indicator, Z be a vector of instruments, and V be a an unobservable random variable.[4] Then, selection models of the form

$$D = 1\{P(Z) - V \geq 0\},$$

where $1\{\cdot\}$ is the indicator function and $P(Z)$ is a nonparametric function, imply monotonicity of D in Z. This is because changing Z from w to z affects D only through $P(Z)$ and hence changes D in the same direction for all individuals for whom it has an effect. Here, the index $P(Z) - V$ is *additively separable* in a component that depends *only* on observables Z and a component that is given by the unobservable V. Monotonicity is implied by all models which are of this form. Conversely, Vytlacil (2002, 2006) shows that we can always represent monotonicity in terms of such a selection model and that this does not impose any additional restrictions on the data generating process.

In this paper we study the impact of *local departures* from monotonicity on monotonicity based estimates of the marginal, average and local average treatment effect. These local departures are expressed in terms of the generalized selection model

$$D = 1\{Q(P(Z), \sigma U) - V \geq 0\}.$$

Z, U and V are assumed to be mutually independently distributed. $\sigma \geq 0$ is a parameter. The index in this model is *nonseparable* in $P(Z)$ and the unobservable U if $\sigma > 0$. The model is still additively separable in V and also the assumption that the instruments affect the treatment decision only though $P(Z)$ is maintained. Therefore, setting $\sigma = 0$ implies monotonicity in $P(Z)$.

It is meaningful to think of a change from $\sigma = 0$ to a small $\sigma > 0$ as a local departure from monotonicity because there is a model with $\sigma = 0$ if, and only if, monotonicity holds. In Appendix B this is discussed in more detail. Studying the effect of local departures from a structure is in the tradition of local specification error analysis that was proposed by Kiefer and Skoog (1984).[5] The virtue of this approach is that it allows us to keep in touch with the original structure.

We can think of the generalized selection model as a reduced form. The underlying structure could be, e.g., a random coefficient model that satisfies certain assumptions.[6] In such a model monotonicity can fail because on average a change in Z might increase the probability to observe $D = 1$ given Z while at the same time the additional randomness in the random coefficients may

[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 $f_{A|B=b}(a)$.

[5] It has also been applied by Chesher (1991), Chesher and Schluter (2002) and Battistin and Chesher (2004) in the context of measurement error. Chesher and Santos Silva (2002) study the impact of uncontrolled taste variation in discrete choice models by modeling local departures from a multinomial Logit model.

[6] See Appendix B and C for details.

induce some individuals to change D from 1 to 0 (Heckman and Vytlacil, 2005; Heckman, Urzua, and Vytlacil, 2006). In Section 4 we postulate that the underlying structure is a random coefficient Logit model and use this model to estimate the variance of $Q(P(Z), \sigma U)$ given $P(Z)$. This allows us to estimate the bias of monotonicity based estimates (up to an approximation error) using the results that are developed in this paper because then the bias only depends on quantities that can be estimated.

1.3 Related Results

The relationship between additive separability of the index and monotonicity has been discussed in several papers. Vytlacil (2002) shows that the original set of assumptions by Imbens and Angrist (1994) can equivalently be expressed in terms of a selection model with an additively separable index. Vytlacil (2006) characterizes a class of nonseparable latent index functions which have equivalent representations as additively separable 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.

Consequences of violations of monotonicity have been discussed before but no results that are directly useful for applied work have been derived. Angrist, Imbens, and Rubin (1996) relate the bias of estimates of the local average treatment effect to the proportion of individuals for whom monotonicity does not hold and the difference in local average treatment effects between those individuals and the ones for which monotonicity holds.[7] Both quantities are unknown. Heckman and Vytlacil (2001, 2005) notice that without monotonicity instrumental variables still identify a weighted average of marginal treatment effects. Their argument is briefly reviewed in Section 2.3.

1.4 Plan of the Paper

Section 2 lays out the formal framework, reviews the identification results under monotonicity, and discusses why monotonicity based estimates could be biased if monotonicity does not hold. Section 3 contains the main theoretical results. Their practical relevance is illustrated in an application in Section 4. Finally, in Section 5 the accuracy of the approximation to the bias term is assessed in a Monte Carlo study. Section 6 concludes. Appendix A contains the proofs. The generalized selection model is studied in more detail in Appendix B. Appendix C contains further technical details that are related to the application and the random coefficient Logit model that is used there.

2 Formal Framework

We adopt the usual convention in program evaluation and say that if an individual 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

[7] The local average treatment effect is defined in Section 2.1.

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).

We focus on the class of models in which identifying power is derived from exogenous variation in instrumental variables Z. The generalized reduced form of the selection model is given by

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

Q is a function of the nonparametric index $P(Z)$, $\sigma \geq 0$ is a scalar, and U and V are both scalar random variables. Here we impose that the instruments affect the treatment decision only through their impact on $P(Z)$, and not individually.[8] Appendix B contains a discussion of the properties of the model and a representation result.

We make the following assumptions.

A 1 (Existence of Instruments): *Z is independent of* (Y_0, Y_1, V).

A 2 (Regularity Conditions): *(i) Y_0 and Y_1 have finite first moments and (ii) the distribution of V is absolutely continuous with respect to Lebesgue measure.*

A 3 (Random Noise): *U is independent of Z and* (Y_0, Y_1, V).

Assumption 1 and 2 are standard, see Heckman and Vytlacil (1999, 2000, 2001, 2005) for details and a discussion. Assumption 3 imposes that failures of monotonicity occur at random. In the two examples that were given in the Introduction this is plausible, but this need not be the case in general.[9] Assumption 3 would be violated if, e.g., in the example with the two officials companionableness was related to ability, and only one of the official would take it into account.[10] Importantly, Assumption 3 still allows for a dependence between the potential outcomes and V.[11]

[8] This is innocuous under monotonicity but restrictive if monotonicity is violated. However, we could think of $P(Z)$ as being one of the instruments, and the analysis as being conditional on all other instruments. Local instrumental variables estimation would then exploit variation in P, holding the other instruments fixed. This would resemble the exposition in Appendix D in Heckman and Vytlacil (2005). This appendix discusses local instrumental variables estimation for the random coefficient model.

[9] It would be interesting to relax this assumption, but this is beyond the scope of this paper because it would complicate the analysis considerably. Besides, in the general case in which U and V are allowed to be related there are conceptual difficulties to define the parameters of interest. See also footnote 16.

[10] If both would take it into account in the same way then we could think of it as entering V.

[11] (1) and Assumption 1 imply that those individuals with low values of V are *a priori* more likely to be treated. V is allowed to be related to the potential outcomes. Among all individuals with treatment probability $P = p$ and $V = v$ some will be treated and some won't. E.g., if there are two individuals and only the first one is treated, then we have $Q(p, \sigma u_1) > v > Q(p, \sigma u_2)$, where u_1 is the value of U for the first individual and u_2 is the value of U for the second individual. The assumption is that U is independent of V. That is, what determines who of the two is

It will be convenient to impose the following normalizations.

N 1: *Normalize V to be uniformly distributed with support* $[0,1]$*,* U *so that* $\mathbb{E}[U] = 0$, $\mathbb{E}[U^2] = 1$*, and* $P(Z)$ *and* $Q(P(Z), \sigma U)$ *such that* $\mathbb{E}[Q(P(Z), \sigma U)] = P(Z)$*.*

They resemble the ones in Heckman and Vytlacil (1999, 2000, 2001, 2005) and Vytlacil (2002). Under these assumptions and normalizations $P(Z)$ is equal to $\Pr(D = 1|Z)$, the so-called propensity score.[12] For the ease of the exposition, from now on we usually write P for $P(Z)$. We can think of P as being a single scalar instrument that can be constructed from Z. Given the structure of the model and the assumptions, this is innocuous for our purposes.[13]

$Q(P, \sigma U)$ is the probability of choosing $D = 1$ given Z and σU. This is because by Assumption 1 V is independent of P, by Assumption 3 V is independent of U, and by Normalization 1 the marginal distribution of V is the uniform distribution. Moreover, if $\sigma = 0$, we have that for any value p of P

$$\mathbb{E}[Q(p, 0)] = Q(p, 0) = p \tag{2}$$

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

Proposition 5 in Appendix B is a slight extension of Vytlacil (2006). It says that there is always a function $Q(P, \sigma U)$ with $\sigma = 0$, if and only if, monotonicity holds. Therefore, it is meaningful to think of a change from $\sigma = 0$ to a small $\sigma > 0$ as a local departure from monotonicity.

2.1 Parameters of Interest

A variety of structural parameters of interest can be expressed in terms of the marginal treatment effect[14]

$$m(v) \equiv \mathbb{E}[Y_1 - Y_0|V = v]. \tag{3}$$

treated, if only one of the two is treated, is unrelated to V.

[12] We have

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

where the first equality follows from (1), the second equality follows from Assumption 1 and 3, and the third equality holds because by Normalization 1 V is uniformly distributed and because by Assumption 3 V is independent of U. Normalization 1 implies that the right hand side of this equation is equal to $P(Z)$.

[13] See also footnote 8 and the discussion in Heckman, Urzua, and Vytlacil (2006).

[14] See Heckman and Vytlacil (1999, 2000, 2001, 2005) 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 of marginal treatment effects, though not the population average.

In many applications, the marginal treatment effect itself is of interest.[15] In this paper, we focus on the bias of estimates of the marginal treatment effect, the population average treatment effect,

$$\mathbb{E}[Y_1 - Y_0] = \int_0^1 m(v)\,dv, \tag{4}$$

and the local average treatment effect,

$$\mathbb{E}[Y_1 - Y_0 | v_l \leq V \leq v_h] = \frac{1}{v_h - v_l}\int_{v_l}^{v_h} m(v)\,dv, \tag{5}$$

where $v_l < v_h$. Our results extend to all other treatment effect parameters if they can be expressed as functions of the marginal treatment effect, see Heckman and Vytlacil (1999, 2000) for details.

By Assumption 2(i) all parameters that are considered here exist. Moreover, they are well defined irrespective of any violations of monotonicity since the index in the generalized reduced form (1) still depends additively on V.[16]

2.2 Identification of Structural Parameters under Monotonicity

Next, we briefly review the respective identification result by Heckman and Vytlacil (1999) and Imbens and Angrist (1994). We first turn to the former which is based on derivatives of the expected value of the outcome conditional on $P = p$ with respect to p. We then present the latter which is based on the difference in this conditional expectation between two different values of P.

2.2.1 Derivative Based Approach

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

D 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.

Let p be a limit point of the support of P. Under Assumption 1 and Normalization 1 (2) implies that p is also a limit point of the support of $Q(P, 0)$. To see that under this condition the

[15] For empirical studies of the returns to college education see Björklund and Moffitt (1987), Carneiro and Lee (2009), and Klein (2009). In this context, V has the interpretation of being a measure of unobserved ability which is related to both the decision to attend college and the wage return to college education. The dependence of this return on unobserved ability is of central interest to policy makers.

[16] Heckman and Vytlacil (2001, 2005) discuss the conceptual difficulties that arise if the index is also nonseparable in V. This case is not considered here.

marginal treatment effect is identified write

$$\mathbb{E}[Y|Q(P,0) = p] = \mathbb{E}[Y_0] + \int_0^p m(v)\, dv, \tag{6}$$

where the integral is equal to

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

The first equality follows from (1) and Assumption 1. For the second equality we use 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 2(i), m is integrable with respect to V. Differentiating both sides of (6) with respect to p yields

$$\frac{\partial \mathbb{E}[Y|Q(P,0) = p]}{\partial p} = m(p) \tag{7}$$

by Leibnitz' rule. By (2), the left hand side is equal to $\partial \mathbb{E}[Y|P = p]/\partial p$ and is identified at the limit point p of P so that $m(p)$ is identified. $\partial \mathbb{E}[Y|P = p]/\partial p$ is called the *local instrumental variables (LIV) estimator* of $m(p)$. It is local because we only consider values of P in a neighborhood around p.

If all p in the open interval $(0, 1)$ are limit points of the support of P the average treatment effect is identified via (4) 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 result is useful if at least one continuously distributed instrument that shifts the treatment probability from 0 to 1 is available. Similarly, by (5) 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.

2.2.2 Level Based Approach

The local average treatment effect is also identified 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 2(i), if $\sigma = 0$,

$$\frac{\mathbb{E}[Y|Q(P,0) = p_h] - \mathbb{E}[Y|Q(P,0) = p_l]}{p_h - p_l} = \mathbb{E}[Y_1 - Y_0|p_l \leq V \leq p_h]. \tag{8}$$

Taking limits for $p_l \rightarrow p_h$ yields (7). Conversely, (8) can be obtained from (5) and (7), with limits of integration being given by p_l and p_h. By (2) the left hand side is identified so that the local average treatment effect is identified as well.

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 and mandatory participation (Battistin and Rettore, 2008) but does in general not hold.

2.3 Failure of the LIV estimator under Non-Monotonicity

Instrumental variables don't identify the marginal treatment effect if the index of the selection model is nonseparable in unobservables.[17] To see this we can derive an expression similar to (6) involving $Q(P, \sigma U)$ instead of $Q(P, 0)$. We have

$$\mathbb{E}[Y|Q(P, \sigma U)] = \mathbb{E}[Y_0] + \int_0^{Q(P,\sigma U)} m(v)\ dv. \tag{9}$$

The LIV estimator of $m(p)$ is given by

$$\frac{\partial \mathbb{E}[Y|P = p]}{\partial p} = \frac{\partial \mathbb{E}\Big[\mathbb{E}[Y|Q(P, \sigma U)]\Big|P = p\Big]}{\partial p} = \frac{\partial \mathbb{E}\left[\int_0^{Q(P,\sigma U)} m(v)\ dv\Big|P = p\right]}{\partial p}.$$

This can be written as

$$\frac{\partial \mathbb{E}[Y|P = p]}{\partial p} = \int m(Q(p, \sigma u))\frac{\partial Q(p, \sigma u)}{\partial p} f_U(u)du \tag{10}$$

and is in general not equal to the marginal treatment effect evaluated at $V = p$, $m(p)$.

Angrist, Imbens, and Rubin (1996) provide a similar expression for the level based estimator. They distinguish between the group of individuals whose treatment decision changes, in reaction to a positive change in P, from 0 to 1 and term them “compliers”, and the group of individuals whose treatment decision changes from 1 to 0, and call them “defiers.” In (10) defiers have $\partial Q(p, \sigma u)/\partial p < 0$ and compliers have $\partial Q(p, \sigma u)/\partial p > 0$.

The LIV estimand is the right hand side of (10). The marginal treatment effect evaluated at $Q(p, \sigma u)$, $m(Q(p, \sigma u))$, is the same for compliers and defiers. The LIV estimator puts a negative weight on the marginal treatment effect for defiers and a positive weight on the marginal treatment effect for compliers (Heckman and Vytlacil, 2001, 2005). However, this by itself does not yield to a bias.[18] The bias arises because we average over different marginal treatment effects.

The result in (10) is not directly useful because recovering $m(p)$ is a non-linear inverse problem and even if a unique solution exists estimating $m(p)$ is a nontrivial task because the solution may not be a continuous function in $\partial \mathbb{E}[Y|P = p]/\partial p$. So far, no results have been developed for this particular case.

In the next section, instead of solving this inverse problem, an approximation to the difference between the quantity that is estimated by the LIV estimator and $m(p)$ is derived. This difference depends on the variance of $Q(P, \sigma U)$ conditional on $P = p$ and is zero if $\sigma = 0$. The

[17] In Appendix B we discuss that under monotonicity we can always find a selection model that is separable in unobservables (Proposition 5). We also discuss that irrespective of whether monotonicity holds it depends on the choice of Q whether estimates are biased.

[18] E.g., the marginal treatment effect could be the same for everybody, i.e. $m(v) = \bar{m}$. Then,

$$\int m(Q(p, \sigma u))\, \frac{\partial Q(p, \sigma u)}{\partial p}\, f_U(u)\, du = \bar{m} \int \frac{\partial Q(p, \sigma u)}{\partial p}\, f_U(u)\, du = \bar{m},$$

where we use that $p = \int Q(p, \sigma u)\, f_U(u)\, du$ gives $\int \frac{\partial Q(p,\sigma u)}{\partial p}\, f_U(u)\, du = 1$ when we differentiate both sides with respect to p.

obtained result can be used to calculate a bias corrected (up to the approximation error) estimate of the marginal treatment effect if this variance can be estimated. We also provide results for the level based estimates.

3 The Impact of Deviations from $\sigma = 0$

In this section we study the impact of local departures from $\sigma = 0$ on derivative and level based estimates of structural parameters. For this, second order approximations in σ about $\sigma = 0$ are performed. The approximations will be derived under the following differentiability conditions.

A 4 (Differentiability): *(i) $Q(p, \sigma u)$ is continuously differentiable in p around $\sigma u = 0$, (ii) $Q(p, \sigma u)$ and $\partial Q(p, \sigma u)/\partial p$ are twice continuously differentiable in σu around $\sigma u = 0$, and (iii) $m(v)$ is three times continuously differentiable.*

In the first part of Appendix A we establish that

$$var(Q(P, \sigma U)|P = p) = \sigma^2 \cdot \left(\frac{\partial Q(p, \sigma u)}{\partial(\sigma u)}\bigg|_{\sigma=0} \right)^2 + o(\sigma^2). \tag{11}$$

Here and later the remainder terms are denoted by $o(\sigma^2)$ and have the property that $o(\sigma^2)/\sigma^2$ goes to zero as σ^2 goes to zero. For the ease of the exposition we denote the approximation to the left hand side by

$$\sigma_p^2 \equiv \sigma^2 \cdot \left(\frac{\partial Q(p, \sigma u)}{\partial(\sigma u)}\bigg|_{\sigma=0} \right)^2 . \tag{12}$$

3.1 Bias of Derivative Based Estimates

The main analytic result of this paper is an approximation to the bias of LIV estimates of the marginal treatment effect, i.e. an approximation to

$$\frac{\partial \mathbb{E}[Y|P = p]}{\partial p} - m(p).$$

P 1: *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 A. □

The approximation to the bias consists of two parts. The first part is 1/2 times the product of the variance of $Q(P, \sigma U)$ conditional on $P = p$ and the second derivative of the marginal treatment effect evaluated at $V = p$. This shows that instead of $m(p)$ a weighted average of marginal treatment effects is estimated. If $m(p)$ is (locally) convex then estimates tend to be upward biased whereas if $m(p)$ is (locally) concave then they tend to be downward biased. The second part of the bias term is $1/2$ times the product of the derivative of the variance of $Q(P, \sigma U)$ conditional on $P = p$ with respect to p and the derivative of the marginal treatment effect evaluated at $V = p$. This shows that the bias is related to the dependence of the variance of $Q(P, \sigma U)$ conditional on $P = p$ on p. E.g., if this variance decreases in p, which is usually the case if p is high, and if the marginal treatment effect is negatively sloped, then estimates tend to be upward biased.

From the formula in Proposition 1 an approximation to the bias of derivative based estimates of the average and local average treatment effect can be obtained by integrating over values of p, as suggested by (4) and (5).

C 1.1: *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*

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

If $p_l = 0$ and $p_h = 1$ this is the bias of the derivative based estimate of the average treatment effect.

Proof. Appendix A. □

3.2 Bias of Level Based Estimates

As discussed before, derivative and level based estimates of the average and local average treatment effect are closely related. We can prove that approximations to respective biases of level based estimates of the average and local average treatment effect are equal to approximations to biases of the corresponding derivative based estimates.

P 2: *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 as given in (14)*. If $p_l = 0$ and $p_h = 1$ this is the bias of level based estimates of the average treatment effect.*

Proof. Appendix A. □

3.3 Practical Relevance

The bias terms that were derived above depend on unknown quantities. Nevertheless, the results have practical relevance because in (13) and (14) we can replace the first and second derivative of the marginal treatment effect by their biased estimates without changing the order of the approximation error. It follows from (7) that these biased estimates are given by the second and third derivative of $\mathbb{E}[Y|P = p]$ with respect to p, respectively. These derivatives can be estimated nonparametrically but estimates of derivatives are generally less precise than estimates of the function itself (see, e.g., Pagan and Ullah, 1999). The order of the approximation error in (13) and (14) remains unchanged because the second order approximations (in σ about $\sigma = 0$) to the respective biases of those estimates are multiples of σ^2 and therefore enter the remainder term in (13) and (14) because the approximation to the bias that uses unknown quantities is already a multiple of σ^2.[19] Hence, we have

$$B^{\text{MTE}}(p) = \frac{1}{2} \cdot \sigma_p^2 \cdot \frac{\partial^3 \mathbb{E}[Y|P = p]}{\partial p^3} + \frac{1}{2} \cdot \frac{\partial \sigma_p^2}{\partial p} \cdot \frac{\partial^2 \mathbb{E}[Y|P = p]}{\partial p^2} + o(\sigma^2) \tag{15}$$

and

$$B^{\text{LATE}}(p_l, p_h) = \frac{1}{2} \cdot \frac{1}{p_h - p_l} \cdot \left(\sigma_{p_h}^2 \cdot \frac{\partial^2 \mathbb{E}[Y|P = p_h]}{\partial p_h^2} - \sigma_{p_l}^2 \cdot \frac{\partial^2 \mathbb{E}[Y|P = p_l]}{\partial p_l^2} \right) + o(\sigma^2). \tag{16}$$

A sensitivity analysis can be undertaken by calculating the respective approximation to the bias term for different values of σ_p^2 and $\partial \sigma_p^2 / \partial p$. Furthermore, under appropriate additional conditions, the variance of $Q(P, \sigma U)$ conditional on P can be estimated.[20] Then, a bias correction procedure (up to the approximation error) is feasible. In the empirical application that is presented in Section 4 a random coefficient Logit model is used to estimate σ_p^2.

3.4 Identification without Monotonicity

The obtained results show that the curvature of the marginal treatment effect is a key determinant of the magnitude of the bias of monotonicity based estimates if σ differs from 0. In the following proposition it is established that under the condition that the marginal treatment effect is linear in v all treatment effect effect parameters considered above are identified if we have prior knowledge on σ_p^2. Evaluating (9) for $Q(P, \sigma U) = q$ and differentiating with respect to q

[19] See also the references in footnote 5. Notice that by (12) σ_p^2 is a multiple of σ^2.

[20] Mixing models can be used to estimate σ_p^2. Matzkin (2007) surveys this literature. Fox and Gandhi (2008) provide a framework that is useful to assess whether a model is identified and provide identification results for the binary choice model. Briesch, Chintagunta, and Matzkin (2007) show identification when there is a Lewbel (2000) type special regressor and propose a nonparametric estimator. Ichimura and Thompson (1998) contains an early identification result for the random coefficient model which heavily relies on linearity of the index. Gautier and Kitamura (2008) propose a new estimator for the density of the random coefficients in such a model. Those papers also contain numerous further references. See also Appendix B for further details.

shows that this is equivalent to requiring that $\mathbb{E}[Y|Q(P,\sigma U) = q]$ is quadratic in q. Importantly, monotonicity or the absence of selection on unoberservables needs *not* to be assumed here.

P 3: *Let the marginal treatment effect be linear in v so that*

$$\mathbb{E}[Y|Q(P,\sigma U) = q] = \alpha + \beta q + \gamma q^2 \tag{17}$$

for some constants α, β, γ. Moreover, let Assumptions 1-3 and Normalization 1 hold and let the variance of $Q(P,\sigma U)$ conditional on $P = p$ be equal to σ_p^2. Let there be a subset $\mathscr{P}$ of the support of P that contains at least three points and assume that either (i) σ_p^2 is constant for all $p \in \mathscr{P}$ or (ii) σ_p^2 is known for all $p \in \mathscr{P}$ and $\sigma_p^2 + p^2$ varies with p but is not linear in it. Then, the marginal, average, and local average treatment effect are identified.

Proof. By (17), Assumption 1, Assumption 3, and Normalization 1

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

Hence, if σ_p^2 is known and $\sigma_p^2 + p^2$ varies with p but is not linear in it the parameters, α, β and γ are identified by the support condition. They can be estimated using a linear regression of the outcome on p and $(\sigma_p^2 + p^2)$. If σ_p^2 is unknown but constant for all $p \in \mathscr{P}$ we can write

$$\mathbb{E}[Y|P = p] = \tilde{\alpha} + \beta p + \gamma p^2,$$

where $\tilde{\alpha} = \alpha + \gamma\sigma_p^2$, which shows that also under this condition β and γ are identified. They can be estimated using a linear regression of the observed outcome on a constant term, p and p^2. The marginal treatment effect is given by the derivative of (17) with respect to q,

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

and is identified since it is a function of β and γ that can be evaluated at arbitrary values for q. Finally, the average and local average treatment effect are identified by the relationships (4) and (5). □

This result can be regarded as an extension of the results by Wooldridge (1997, 2003) and Heckman and Vytlacil (1998) who propose to estimate the average treatment effect in a linear model using standard instrumental variables techniques. However, for binary treatments their assumptions imply that the marginal treatment effect is constant across different values of v.

The homoskedasticity condition (i) that is used here is unlikely to hold globally. To see this let the support be given by the set $\{0, 0.5, 1\}$. By construction, the variance of $Q(P,\sigma U)$ is zero for $P = 0$ and $P = 1$, but could be positive for $P = 0.5$. However, this condition could hold locally, e.g. if the estimation procedure is based on $\mathscr{P} = \{0.45, 0.5, 0.55\}$.

A specification test can be constructed under condition (i) provided that there are two subsets $\mathscr{P}_1$ and $\mathscr{P}_2$ for which this condition holds. Then, two sets of estimates of β and γ can be obtained. Both are consistent under the null hypothesis of no misspecification and a specification test is given by the test for equality of the respective estimates of β and γ. For a specification test under condition (ii) σ_p^2 could be included as an additional regressor. The test is then a test for the equality of the coefficients on p^2 and σ_p^2.

4 An Empirical Application

In this section, parts of the Angrist and Krueger (1991) data are used to illustrate the practical relevance of the results that were obtained above.[21] The data are from the 1980 census and contain information for individuals born between 1930 and 1939. The original paper contains a detailed description of the data and detailed summary statistics.

Angrist and Krueger (1991) aim at estimating the effect of compulsory schooling on wages. They argue that an individual's quarter of birth can be excluded from the wage equation (exclusion restriction) and show empirically that actual schooling is linked to the quarter of birth (rank condition). This is due to compulsory schooling requirements, in particular entry regulations. On these grounds the quarter of birth is used as an instrument for schooling in a regression of log weekly earnings on completed years of schooling. The exclusion restriction corresponds to Assumption 1 in this paper and has been questioned by Bound and Jaeger (2000). It will, however, be maintained here.

In our application, we investigate the effect of completing more than 9 years of education on wages. This is reasonable because the most common age at school entry is about 6 years (Table B.1. in Angrist and Krueger, 1992) and the most common compulsory schooling attendance age is 16 years (Appendix 2 in Angrist and Krueger, 1991). If students are born early in the year and enter school at the age of 6, then they will be in 10th grade when they turn 16 because the school year typically starts in September and ends in July. They thus have 9 years of completed education if they drop out as soon as possible. Conversely, if they are born in the fourth quarter and enter school in the year they turn 6 they will have 10 years of completed education when they turn 16. A sizable fraction of those who are born in the fourth quarter are held back for one year (Barua and Lang, 2008). They enter school at the age of almost 7 years and will have completed 9 years of education when they turn 16.

Figure 1 shows that the effect of being born in the fourth instead of the first quarter on the state specific probability to obtain more than 9 year of schooling is positive in most states. This effect is significant primarily in the states in which this probability is low. Arguably, these are states in which compulsory schooling requirements are binding for a sizable part of the population.[22]

[21] The data set can be downloaded at http://econ-www.mit.edu/files/397.

[22] The compulsory schooling age does not vary within states, but the exact school entry regulations an individual faces do. Regulations differ across states. However, Barua and Lang (2008) show that state level regulations have not been strictly enforced. E.g., in third quarter cutoff states children may enter school in the year in which they reach the entry age provided that they do so in the first three quarters. In those states many children that were born in the fourth quarter still entered school in the year in which they turned 6. Also regression results show that the coefficient estimates on interaction terms between the quarter of birth and the type of state regulation are generally not significant. Therefore, we do not focus on differences in regulations across states in this application.

Variation in the quarter of birth provides a natural experiment (provided that it is unrelated to ability) because it generates variation in the age at school entry, which then yields to variation in the years of schooling. This is because individuals can drop out of school on the day they reach the legal dropout age. Individuals who were born earlier in the year are older on average when they enter school (Table 2 in Angrist and Krueger, 1992). Therefore, they reach the legal dropout age after less schooling. From this it follows that monotonicity holds if for each individual the decision to have more than 9 years of completed schooling is monotonically increasing in the quarter of birth. This assumption is implausible for at least two reasons.

First, the relationship between age at school entry and quarter of birth could be affected by unobserved differences in start age policies across schools (footnote 4 in Angrist and Krueger, 1991). In particular, schools are not obliged to admit individuals as soon as they turn 6 before the state-wide cutoff date.[23] In fourth quarter cutoff states, e.g., some school districts can choose not to accept children that turn 6 in the fourth quarter (because they are considered too young to enter school or because the number of applicants exceeds school capacity) and thus delay their admission until the following year.[24] Children who cannot enter in the year they turn 6 will be in the oldest who enter school in the subsequent year, and not the youngest. They will just have started 10th grade when they turn 16 and might therefore decide to drop out immediately, with 9 years of completed schooling. At the same time children who are born in the first quarter will be in the middle of 10th grade when they turn 16 and might therefore decide to complete this grade and thus have more than 9 years of completed education. This is just an example that shows that while on average those born in the last quarter are more likely to complete more than 9 years of education it could be that a hypothetical change from the first to the fourth quarter *induces* some individuals *not* to complete more than 9 years of education. In this example, this is due to unobserved differences in school policies. From the point of view of the econometrician these unobserved differences give rise to additional randomness which causes monotonicity to fail.[25] In the generalized selection model this additional randomness is part of U. Assumption 3 in this context is that the child's ability is not related to this additional randomness. The distribution of ability conditional on entering school is still allowed to differ across quarters of birth, see footnote 11.

Second, children may be held back by their parents (Angrist and Krueger, 1992). From the point of view of the econometrician this has the same effect as an unobserved entry regulation that does not allow a child to enter school in the year in which it turns 6.[26] For an example suppose that there are some parents, say of type A, who prefer that their child enters school at an early age because this is the simplest way of having the child taken care off during the day,

[23] I am grateful to a referee for pointing this out.

[24] In these states the law says that children may in general enter school if they turn 6 before the *end of* the fourth quarter. First, second and third quarter cutoff states are defined accordingly.

[25] In principle, one could deal with this by exploiting monotonicity within schools if the identity of the school was known. However, the identity of the school is unobserved in our data.

[26] See also Barua and Lang (2008) for further discussion of this point. Table 1 in that paper shows the distribution of entry age into kindergarten for the 1952 and 1953 birth cohort. 80% of the children who are born in the first quarter enter kindergarten in the year in which they turn 5. In fourth quarter cutoff states about 45% of the children born in the fourth quarter still enter kindergarten at the age of 5 and about 50% enter at the age of 6. Children normally stay in kindergarten for one year and enter school thereafter. These numbers are consistent with the entry age differences reported in Table 2 in Angrist and Krueger (1992).

while other parents, say of type B, think that entering school too early may be detrimental to their children.[27] Whether parents are of type A or B matters for children that are born later in the year and may cause monotonicity to fail. Assumption 3 in this context is that the parent type is independent of the child's ability. It is not that entry age is independent of ability. Assumption 3 it is violated if children of parents who would like to send their child to school as early as possible are systematically of higher (or lower) ability.

We proceed as follows. First, the probability to stay in school for more than 9 years, P, is estimated using a random coefficient Logit model. In a second step potentially biased derivative based estimates of the marginal treatment effect as well as level based estimates of the local average treatment effect are obtained. Then, the results that were derived in this paper, in particular (15) and (16), are used to calculate an approximation to the respective bias. Finally, bias corrected estimates of the marginal and local average treatment effect are calculated.

The model for observed earnings, Y, is given by

$$Y = D \cdot (Y_1 - Y_0) + X\beta, \tag{18}$$

where X is a vector of exogenous covariates that contains a constant term, the year of birth, and state of birth indicators.[28] This specification allows earnings levels to depend on the year and state of birth but imposes that the return to obtaining more than 9 years of schooling is the same across states and does not depend on the exact year of birth.[29]

Let Z be a vector consisting of quarter of birth indicators and X. The selection model is the random coefficient Logit model

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

where $\tilde{V}$ follows the logistic distribution and $\tilde{\gamma}$ is a vector of random coefficients for the quarter of birth indicators and the year of birth. For reasons of tractability (there are 50 states and the District of Columbia) the coefficients on the state indicators are assumed to be constants. The random elements are assumed to be distributed according to a multivariate normal distribution denoted by $\phi(\cdot|\gamma, \Sigma)$ with mean γ and variance-covariance matrix Σ. We assume that Z is independent of $(Y_0, Y_1, \tilde{V})$. This is stronger than the restrictions in Assumption 1 since now Z includes the vector X of covariates. The interpretation of this assumption is that the distribution of $\tilde{V}$ is the same across states of birth, years of birth, and quarters of birth.[30] Finally, it is assumed that the random coefficients are independent of Z and $(Y_0, Y_1, \tilde{V})$. This corresponds to Assumption 3 from above. In this model, monotonicity can be violated even if all coeffi-

[27] I am grateful to a referee for coming up with this example.

[28] Linearity in the year of birth is imposed. Figure I and V in Angrist and Krueger (1991) suggest that this is a good approximation. To check this a Logit model with quarter of birth dummies, the year of birth, year of birth dummies, and state of birth dummies was estimated to explain whether individuals complete more than 9 years of education. The year of birth dummies were not jointly significant ($p = 0.31$). Then, the wage was regressed on the same variables and again the year of birth dummies were not jointly significant ($p = 0.64$).

[29] This follows common practice in first applications of LIV estimation such as Carneiro and Lee (2009). See Klein (2009) for a generalization in which the return is allowed to depend on the interaction between covariates and V.

[30] Assumption 1 is made conditional on X, which is considerably weaker. Without strengthening it we have to perform the analysis separately for each state. This is beyond the limits of the data set because the support of P in a given state is always smaller than the unconditional support where we consider all states at the same time.

cients on the quarter of birth indicators are positive. This is because the effect of being born in the fourth instead of the first quarter, e.g., is given by the difference in the respective random coefficients. The model is estimated using simulated maximum likelihood. In Appendix B and C we describe how $Q(P, \sigma U)$ can be derived from the random coefficient model and how $var(Q(p, \sigma U))$ is estimated.

Table 1 contains parameter estimates for a specification with 4 quarter of birth indicators and independently distributed random coefficients.[31] Specification (1) and (2) do not include state of birth indicators while specification (3) and (4) do. Except for the coefficient on being born in the 1st quarter in specification (2) results are very similar across all four specifications.[32] Throughout, the impact of the year of birth is positive and being born later in the year has a positive impact on staying in school. Moreover, all estimates of the standard deviation of the random coefficients are significantly different from zero. In the subsequent analysis specification (4) will serve as the basis for the second stage estimates.

Table 2 shows the implied mean and variance of the impact of the quarter of birth on the index $Z\tilde{\gamma}$. Column (4) indicates that the effect of being born early in the year is negative on average but is positive for a sizable part of the population. Taken at face value the results show that being born in the first instead of the fourth quarter has a negative impact on the likelihood to stay in school for 56 per cent of the individuals and a positive impact for 44 per cent of the individuals. Monotonicity is violated for the latter 44 per cent.

Figure 2 shows the distribution of fitted values of the probability to have more than 9 years of completed schooling, P. A sizable part of the population will continue to attend school with a probability larger than 85 per cent. We have seen in Figure 1 that those individuals mainly live in states in which the quarter of birth has no significant impact on P. In those states the instrument has no identifying power. Therefore, we will focus on individuals with P below 85 per cent in the subsequent analysis.

Figure 3 summarizes individual differences in the variance of the individual probability to attend more than 9 years of schooling in the population. To explore how this variance depends on the level of P we regressed it on a constant and a set of fourth order splines in the estimated P. The knots are equidistant and located at 0.73, 0.78 and 0.83. Thereby, we average over the sample distribution of values of Z that share the same value of P. This yields estimates of the population average σ_p^2. The first derivative of the estimate is obtained analytically. Results in Figure 4 show that the higher P the lower the variance of the individual probability.

These estimates of σ_p^2 and $\partial\sigma_p^2/\partial p$ will be used to calculate the bias of monotonicity based estimates of the marginal and local average treatment effect. Before doing so we obtain the ordinary least squares and the linear instrumental variables estimate of the return to more than 9 years of schooling. Results are reported in Table 3. The instrumental variables estimate is

[31] Notice that the correlation across random coefficients is not identified because only one quarter of birth indicator takes on the value 1 at a time. Results that were obtained using a specification in which the index depends linearly on the year and the quarter of birth, allowing for a correlation across random coefficients, are very similar. Importantly, we run the risk of obtaining biased estimates as soon as the variance of two random coefficients differs from zero. This is the case throughout.

[32] Marginal effects are very close to one another across specifications. Taking the fourth quarter as the baseline, being born in the first quarter decreases the probability to have more than 9 years of schooling by 1.7 to 1.9 per cent, being born in the second quarter decreases it by 1.4 to 1.5 per cent, and being born in the third quarter decreases it by 0.5 to 0.6 per cent.

higher than the ordinary least squares estimate. Under monotonicity this would reflect that average returns for those for whom the compulsory schooling requirement is binding are exceptionally high (Imbens and Angrist, 1994; Harmon and Walker, 1999). We have illustrated that monotonicity cannot safely be assumed here. Hence, it is useful to obtain bias corrected estimates of the marginal and local average treatment effect. Towards obtaining those second stage estimates notice that (18) implies that

$$\mathbb{E}[Y|X,P] = \kappa(P) + X\beta.$$

$\kappa(P)$ and β are estimated using an ordinary least squares regression of the log weekly wage on the year of birth, state indicators, as well as fourth order splines in the estimated P. The knots were chosen as before. Throughout, confidence intervals do not account for the first stage estimation error. In light of the precision of the first stage estimates reported in Table 1 we do not expect this to have severe consequences. In Section 2, we have established that the LIV estimate of the marginal treatment effect is given by the first derivative of $\kappa(P)$ with respect to its argument. The second and third derivative are therefore LIV estimates of the first and second derivative of the marginal treatment effect with respect to its argument. The advantage of using splines is that these derivatives as well as their standard errors can be obtained analytically. Figure 5 shows those estimates.

Figure 6 and 7 shows potentially biased monotonicity based estimates of the marginal and local average treatment effect, respectively, and bias corrected estimates.[33] The confidence intervals for the bias corrected estimates are wider than for the original estimate and not shown here. The dependence of biased and bias corrected estimates on V is positive. The interpretation of this is that in terms of wages bad types, who are characterized by high values of V and for whom compulsory schooling requirements are binding, profit more from additional education than good types do. Bias corrected estimates of the marginal and local average treatment effect are lower for low values of V and higher for high values of V.

In this application the interpretation of both monotonicity based and bias corrected estimates is qualitatively the same. However, as has previously been argued, monotonicity cannot safely be assumed here so that it is important to assess the sensitivity of estimates to departures from monotonicity. This application has shown that this is feasible provided that the researcher is willing to make additional assumptions that allow him to estimate $var(Q(p, \sigma U))$.

5 Monte Carlo

The approximation to the bias of the marginal treatment effect in Proposition 1 depends on the unknown first and second derivative of $m(v)$. It has been noted that these unknown quantities can be replaced by biased estimates without altering the order of the approximation error. The goal of this Monte Carlo study is to assess the effect of doing so on the accuracy of the approximation to the bias term. For this reason realizations of P and $var(Q(p, \sigma U))$ will be treated as known and $m(v)$ is specified to be a quadratic function in v. This implies that the approximation error

[33] It has been argued before that the instrument might not be valid for values of P above 0.85. Therefore, we will only report estimates of treatment parameters for values of P that are below 0.85.

comes from replacing the aforementioned derivatives by their biased estimates. The analysis will be carried out for the approximation to the bias of LIV estimates of $m(0.5)$.

The design is the following. There are 10,000 simulated data sets with 2,000 observations, respectively. To generate these data we start by drawing values p of P from a uniform distribution with support $[0.4, 0.6]$. Then, values u of U are drawn from a standard normal distribution and values q of Q are calculated according to

$$q = p + 0.75 \cdot \sigma u - p \cdot \sigma u. \tag{19}$$

We vary σ between 0 and 0.5. (19) implies that the approximation in (11) is exact so that

$$\sigma_p^2 = var(Q|P = p) = (0.75 - p)^2 \cdot \sigma^2 \tag{20}$$

and

$$\frac{\partial \sigma_p^2}{\partial p} = \frac{\partial var(Q|P = p)}{\partial p} = -2 \cdot (0.75 - p) \cdot \sigma^2.$$

$m(v)$ is specified to be a second order polynomial in v and we impose $m(0.5) = 0$ and $\partial m(v)/\partial v = -1$. This implies

$$m(v) = 3\rho v^2 - (3\rho + 1) \cdot v + \frac{3\rho + 2}{4}, \tag{21}$$

where ρ is a parameter which is related to the curvature of $m(v)$. To assess the dependence of the accuracy of the approximation on this curvature ρ will be varied between 0, for which $m(v)$ is linear in v, and 10. The first and second derivative of $m(v)$ are given by

$$\frac{\partial m(v)}{\partial v} = 6\rho v - 3\rho - 1$$

and

$$\frac{\partial^2 m(v)}{\partial v^2} = 6\rho,$$

respectively. This shows that $m(v)$ is the more convex the higher ρ. Figure 8 shows the dependence of $m(v)$ on ρ.

This specification for $m(v)$ implies a particular functional form for $\mathbb{E}[Y|Q]$. We impose that $\mathbb{E}[Y|Q = 0] = 0$. Then, (9) implies

$$\mathbb{E}[Y|Q = q] = \int_0^q m(v)\, dv = \int_0^q 3\rho v^2 - (3\rho + 1) \cdot v + \frac{3\rho + 2}{4}\, dv = \rho q^3 - \frac{3\rho + 1}{2} \cdot q^2 + \frac{3\rho + 2}{4} \cdot q.$$

Realizations y of Y are calculated as $y = \mathbb{E}[Y|Q = q] + \varepsilon$, where ε is drawn from a normal distribution with mean 0 and variance 0.02. Figure 9 shows one draw of generated data for $\rho = 4$ and $\sigma_{0.5}^2 = 0.01$.[34] On the left hand side, values of Y are plotted against values of P and Q. On the right hand side, values of the marginal treatment effect evaluated at values of Q are plotted against values of P and Q, respectively. Obviously, when we plot values $m(q)$ against q

[34] This is the value that was obtained in the empirical application for $p \approx 0.85$, see Figure 4.

we get the marginal treatment effect itself. However, plotting these values against p shows that there is a distribution of marginal treatment effects evaluated at Q for every value of P. This relates to the intuition that is developed in Section 2.3. There we have argued that (10) shows that the LIV estimator may be biased because it averages over marginal treatment effects.

LIV estimates were obtained by regressing realizations of Y on a third order polynomial in realizations of P. This is the correct specification when $\sigma = 0$. From the coefficient estimates an estimate of the first, second and third derivative of the fitted value, evaluated at $P = 0.5$, is obtained. These are biased (except for $\sigma = 0$) estimates of $m(0.5)$, its first derivative, and its second derivative, respectively. The bias of estimates of $m(0.5)$ is given by the LIV estimate because, by construction, the true value is 0 for all ρ. We calculate the approximation to the bias using (15).

Since the approximations in (11) and (13) are exact we can directly assess the approximation error that stems from using (potentially) biased LIV estimates of the first and second derivative of $m(v)$ instead of the unknown quantities when calculating the approximation to the bias. In Figure 10, the bias and the approximation that uses biased estimates are plotted against ρ. This is done for $\sigma^2_{0.5} = 0.0056$ and $\sigma^2_{0.5} = 0.01$. Table 4 shows the dependence between the mean bias and the mean of the estimated approximation for $\rho = 4$. The mean difference between the two is reported in the last column. Standard deviations for respective means are reported in parentheses. Both the Figure and the table show that the approximation error is approximately linear in ρ and is increasing in ρ and $\sigma^2_{0.5}$.

Finally, Table 5 reports, for different combinations of ρ and $\sigma^2_{0.5}$, the percentage of the bias that is corrected using (15). It shows that up to about $\sigma^2_p = 0.01$ the cure is better than the disease in the sense that the bias corrected estimate is closer to the true value than the biased estimate. Finally, the sign of the bias is the same as the sign of the approximation if the number reported in this table is positive. It is worth noting that for all combinations of ρ and $\sigma^2_{0.5}$ the approximation to the bias is of the same sign as the bias.

6 Concluding Remarks

Recently developed instrumental variables estimators of treatment effect parameters hinge on the assumption that conditional on exogenous covariates a hypothetical changes in instrumental variables either has no effect on an individual's treatment decision or changes it in the same direction as it does for all other individuals for which is has an effect. Under this assumption exogenous variation in instrumental variables identifies the average treatment effect for the subpopulation of individuals who would have been affected by this change. The monotonicity condition is economically interpretable but not directly testable. If it cannot safely be assumed it is important to assess how monotonicity based estimates are affected by a failure of this assumption.

It is now well understood that under monotonicity treatment decisions can be represented using a binary choice model with an index that is additively separable in observables and unobservables. Conversely, a binary choice model with an index that is additively separable in observables and unobservables implies monotonicity. In this paper, we have investigated the effect of local departures from such an additively separable structure on monotonicity based estimates of the marginal, average and local average treatment effect. Second order approximations to respective bias terms have been derived under the assumption that violations of monotonicity occur at random. In the application we have shown that under additional assumptions that allow the researcher to estimate a selection model with a nonseparable index a bias correction procedure is feasible. The Monte Carlo study has demonstrated that whereas the approximations might be inaccurate when the departure from a structure that implies monotonicity is big, they are still accurately predicting the sign of the bias. Consequently, even if deviations from monotonicity are substantial the results are useful to assess whether estimates that were obtained under the monotonicity assumption constitute a lower or an upper bound for the structural parameter of interest.

Such sensitivity analyses can be used to show that the major findings of an empirical study are robust to violations of monotonicity. This also applies to studies in which identification is based on natural experiments. While these experiments often generate credible exogenous variation the impact of this variation on individual decision making may depend on many factors that are not recorded. Hence, violations of monotonicity are possible. Carrying out a sensitivity analysis and calculating the bias can therefore greatly enhance the credibility of the results.

Appendix A: Proofs

We first derive approximations to $Q(p, \sigma u)$, $\partial Q(p, \sigma u)/\partial p$, and $var(Q(P, \sigma U)|P = p)$. All approximations are of the second order in σ and about $\sigma = 0$. The remainder terms are denoted by $o(\sigma^2)$ and have the property that $o(\sigma^2)/\sigma^2$ goes to zero as σ^2 goes to zero. The necessary differentiability conditions are stated in Assumption 4.

The second order approximation to $Q(p, \sigma u)$ is given by

$$Q(p, \sigma u) = Q(p, 0) + \sigma u \cdot \left.\frac{\partial Q(p, \sigma u)}{\partial(\sigma u)}\right|_{\sigma=0} + (\sigma u)^2/2 \cdot \left.\frac{\partial^2 Q(p, \sigma u)}{\partial(\sigma u)^2}\right|_{\sigma=0} + o(\sigma^2). \tag{22}$$

Next, we derive an approximation to p. Under Normalization 1 and Assumption 3 we have, using (22),

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

Under Normalization 1 $\int u \; f_U(u)\, du = 0$ and $\int u^2 \; f_U(u)\, du = 1$ so that

$$p = Q(p, 0) + \sigma^2/2 \cdot \left.\frac{\partial^2 Q(p, \sigma u)}{\partial(\sigma u)^2}\right|_{\sigma=0} + o(\sigma^2). \tag{23}$$

Combining (22) with (23) yields

$$Q(p, \sigma u) = p + \sigma u \cdot \left.\frac{\partial Q(p, \sigma u)}{\partial(\sigma u)}\right|_{\sigma=0} + \sigma^2/2 \cdot (u^2 - 1) \cdot \left.\frac{\partial^2 Q(p, \sigma u)}{\partial(\sigma u)^2}\right|_{\sigma=0} + o(\sigma^2) \tag{24}$$

and from this it follows that

$$\frac{\partial Q(p,\sigma u)}{\partial p} = 1 + \sigma u \cdot \left.\frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)\,\partial p}\right|_{\sigma=0} + \sigma^2/2 \cdot (u^2 - 1)\left.\frac{\partial^3 Q(p,\sigma u)}{\partial(\sigma u)^2\,\partial p}\right|_{\sigma=0} + o(\sigma^2). \tag{25}$$

Finally, we derive an approximation to the variance of $Q(P,\sigma U)$ conditional on $P = p$. Under Assumption 3

$$var(Q(P,\sigma U)|P = p) = var(Q(p,\sigma U)) = \int (Q(p,\sigma u) - p)^2\ f_U(u)\,du.$$

Using (23) and (24) we get that this is equal to

$$\int \left(\sigma u \cdot \left.\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\right|_{\sigma=0} + \sigma^2/2 \cdot (u^2 - 1) \cdot \left.\frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)^2}\right|_{\sigma=0}\right)^2\ f_U(u)\,du + o(\sigma^2).$$

If we let multiples of σ^3 and σ^4 enter the remainder term this simplifies to

$$\int \left(\sigma u \cdot \left.\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\right|_{\sigma=0}\right)^2\ f_U(u)\,du + o(\sigma^2).$$

Under Normalization 1 $\int u^2\ f_U(u)\,du = 1$ so that

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

The following lemma will be used in the proof of Proposition 1.

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

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

Proof. Under Assumptions 1-4 and Normalization 1 evaluating (9) at $Q(P,\sigma U) = Q(p,\sigma u)$ and differentiating with respect to $Q(p,\sigma u)$ yields, by Leibnitz' rule,

$$\frac{\partial \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)]}{\partial Q(p,\sigma u)} = m(Q(p,\sigma u)). \tag{26}$$

A second order approximation in σ about $\sigma = 0$ gives

$$\begin{aligned}
&\frac{\partial \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)]}{\partial Q(p,\sigma u)} \\
&\quad = m(Q(p,0)) + \sigma u \cdot \frac{\partial m(Q(p,0))}{\partial Q(p,0)} \cdot \left.\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\right|_{\sigma=0} \\
&\qquad + (\sigma u)^2/2 \cdot \left\{ \frac{\partial^2 m(Q(p,0))}{\partial Q(p,0)^2} \cdot \left(\left.\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\right|_{\sigma=0} \right)^2 + \frac{\partial m(Q(p,0))}{\partial Q(p,0)} \cdot \left.\frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)^2}\right|_{\sigma=0} \right\} + o(\sigma^2).
\end{aligned} \tag{27}$$

This will be used below.

Towards establishing the result we have

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

where the first equality is by iterated expectations, the second follows from Assumption 3, the fourth from the integrand being finite, which is implied by Assumption 2, and the fifth applies the chain rule.

Together with (25) and (27) this yields

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

By expanding and including multiples of σ^3 and σ^4 in the remainder term we get that this is equal to

$$\begin{aligned}
\int \Bigg\{ & m(Q(p,0)) + \sigma u \cdot \frac{\partial m(Q(p,0))}{\partial Q(p,0)} \cdot \left.\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\right|_{\sigma=0} \\
& + (\sigma u)^2/2 \cdot \left\{ \frac{\partial^2 m(Q(p,0))}{\partial Q(p,0)^2} \cdot \left(\left.\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\right|_{\sigma=0} \right)^2 + \frac{\partial m(Q(p,0))}{\partial Q(p,0)} \cdot \left.\frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)^2}\right|_{\sigma=0} \right\} \\
& + \sigma u \cdot m(Q(p,0)) \cdot \left.\frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)\ \partial p}\right|_{\sigma=0} + (\sigma u)^2 \cdot \frac{\partial m(Q(p,0))}{\partial Q(p,0)} \cdot \left.\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\right|_{\sigma=0} \cdot \left.\frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)\ \partial p}\right|_{\sigma=0} \\
& + \sigma^2/2 \cdot (u^2 - 1) \cdot m(Q(p,0)) \cdot \left.\frac{\partial^3 Q(p,\sigma u)}{\partial(\sigma u)^2\ \partial p}\right|_{\sigma=0} \Bigg\} f_U(u)\, du + o(\sigma^2).
\end{aligned}$$

By Normalization 1 $\int u\ f_U(u)\ du = 0$ and $\int u^2\ f_U(u)\ du = 1$ so that the result follows. □

Proof of Proposition 1

Proof. (23) implies that

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

Note that on the right hand side the first derivative of the argument of m evaluated at $\sigma = 0$ is zero. Hence, a second order Taylor series expansion in σ about $\sigma = 0$ yields

$$m(p) = m(Q(p,0)) + \sigma^2/2 \cdot \frac{\partial m(Q(p,0))}{\partial Q(p,0)} \cdot \frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)^2}\bigg|_{\sigma=0} + o(\sigma^2). \tag{28}$$

Moreover, (12) implies

$$\frac{\partial\sigma_p^2}{\partial p} = 2\sigma^2 \cdot \frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\bigg|_{\sigma=0} \cdot \frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)\,\partial p}\bigg|_{\sigma=0}. \tag{29}$$

From Lemma 1 and (28) we get

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

We get the result using (2), (12) and (29). □

Proof of Corollary 1.1

Proof. By (5) and (13)

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

This yields the result. □

Proof of Proposition 2

Proof. We have

$$\begin{aligned}\mathbb{E}[Y|P=p] &= \mathbb{E}\Big[\mathbb{E}[Y|Q(P,\sigma U)]\Big|P=p\Big] \\ &= \mathbb{E}\Big[\mathbb{E}[Y|Q(p,\sigma U)]\Big] \\ &= \int \mathbb{E}[Y|Q(P,\sigma U) = Q(p,\sigma u)]\ f_U(u)\ du,\end{aligned}$$

where the first equality is by iterated expectations and the second follows from Assumption 3. A second order Taylor series expansion in σ about $\sigma = 0$, using (26), yields that this is equal to

$$\int \Bigg\{ \mathbb{E}[Y|Q(P,\sigma U) = Q(p,0)] + \sigma u \cdot m(Q(p,0)) \cdot \frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\bigg|_{\sigma=0} \\ + (\sigma u)^2/2 \cdot \left\{ m(Q(p,0)) \cdot \frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)^2}\bigg|_{\sigma=0} + \frac{\partial m(Q(p,0))}{\partial Q(p,0)} \cdot \left(\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\bigg|_{\sigma=0}\right)^2 \right\}\Bigg\} \; f_U(u)\; du + o(\sigma^2).$$

By Normalization 1 $\int u \;\; f_U(u)\; du = 0$ and $\int u^2 \;\; f_U(u)\; du = 1$ so that

$$\mathbb{E}[Y|P = p] = \mathbb{E}[Y|Q(P,\sigma U) = Q(p,0)] \\ + \sigma^2/2 \cdot \left\{ m(Q(p,0)) \cdot \frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)^2}\bigg|_{\sigma=0} + \frac{\partial m(Q(p,0))}{\partial Q(p,0)} \cdot \left(\frac{\partial Q(p,\sigma u)}{\partial(\sigma u)}\bigg|_{\sigma=0}\right)^2 \right\} + o(\sigma^2). \tag{30}$$

This will be used below.

Under Assumption 3 and Normalization 1 we get (23) and hence

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

A second order Taylor series expansion thereof in σ about $\sigma = 0$, noting that the first derivative of the conditional expectation with respect to σ evaluated at $\sigma = 0$ is equal to zero, yields

$$\mathbb{E}[Y|Q(P,\sigma U) = p] = \mathbb{E}[Y|Q(P,\sigma U) = Q(p,0)] + \sigma^2/2 \cdot m(Q(p,0)) \cdot \frac{\partial^2 Q(p,\sigma u)}{\partial(\sigma u)^2}\bigg|_{\sigma=0} + o(\sigma^2). \tag{31}$$

Using (30) and (31) we get

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

Combining this with (2), (8), and (12) yields the result. □

Appendix B: Properties of the Generalized Selection Model

In this appendix we discuss the properties of the selection model

$$Q = 1\{Q(P,\sigma U) \geq V\}$$

in more detail. Importantly, under Assumption 1 and 3 it imposes a particular form of *index sufficiency*:

$$\Pr(Y \in A|Z, D) = \Pr(Y \in A|P(Z), D)$$

for any A. Index sufficiency means that the instruments affect the treatment decision and the outcome only though $P(Z)$. It holds trivially if Z is a scalar and $P(Z)$ is one to one, like in Imbens and Angrist (1994) where the instrument is binary. Index sufficiency is in principle testable as soon as $var(Q(p,\sigma U))$

can be estimated because then it is possible to test whether

$$var(Q(P(z^a), \sigma U)) = var(Q(P(z^b), \sigma U))$$

for any two z^a and z^b such that $P(z^a) = P(z^b)$.

We now show that many selection models that satisfy index sufficiency, i.e. are of the form $D = 1\{\mu_D(P, U_D) \geq \tilde{V}\}$ with possibly vector valued U_D have a representation $D = 1\{Q(P, \sigma U) \geq V\}$ with a scalar U.

Proposition 4 (Representation): *Any selection model that is of the form $D = 1\{\mu_D(P, U_D) \geq \tilde{V}\}$ with a possibly vector valued U_D has an equivalent representation $D = 1\{Q(P(Z), \sigma U) \geq V\}$ with a scalar U provided that (i) $\tilde{V}$ is continuously distributed, (ii) P is independent of $\tilde{V}$, and (iii) U_D is continuously distributed independent of P and $\tilde{V}$. Moreover, Assumption 1, 2(ii), 3, and Normalization 1 hold for this equivalent representation.*

Proof. By (i) $F_{\tilde{V}}$ is strictly increasing and hence $\mu_D(p, U_D) \geq \tilde{V}$ is equivalent to $F_{\tilde{V}}(\mu_D(p, U_D)) \geq V$, where V is uniformly distributed on $[0, 1]$. Write $\tilde{\mu}_D(p, U_D) \equiv F_{\tilde{V}}(\mu_D(p, U_D))$.

Define

$$\tilde{U} \equiv F_{\tilde{\mu}_D(p,U_D)}(\tilde{\mu}_D(p, U_D)).$$

By (i) and (iii) the distribution of $\tilde{\mu}_D(p, U_D)$ is continuous so that $\tilde{U}$ is uniformly distributed independent of P. Moreover, $F_{\tilde{\mu}_D(p,U_D)}$ is strictly increasing and hence invertible. Denoting the inverse function of $F_{\tilde{\mu}_D(p,U_D)}$ by $F^{-1}_{\tilde{\mu}_D(p,U_D)}$ we have

$$\tilde{\mu}_D(p, U_D) = F^{-1}_{\tilde{\mu}_D(p,U_D)}(\tilde{U}).$$

Denote the right hand side of this equation by $\tilde{Q}(p, \tilde{U})$. This yields

$$\tilde{\mu}_D(p, U_D) = \tilde{Q}(p, \tilde{U}) \tag{32}$$

which will be used below. Pick a strictly increasing distribution function F_U. Denoting the inverse function of F_U by F_U^{-1} we can define

$$U \equiv F_U^{-1}(\tilde{U}) = F_U^{-1}\left(F_{\tilde{\mu}_D(p,U_D)}(\tilde{\mu}_D(p, U_D))\right).$$

Notice that, by construction, $\tilde{U} = F_U(U)$ and hence, by (32),

$$\tilde{\mu}_D(p, U_D) = \tilde{Q}(p, \tilde{U}) = \tilde{Q}(p, F_U(U)).$$

From this we can see that there exists a function Q, a random variable U and a scalar $\sigma \geq 0$ such that Normalization 1 holds,

$$\tilde{Q}(p, F_U(U)) = Q(p, \sigma U),$$

and hence

$$\tilde{\mu}_D(p, U_D) = Q(p, \sigma U).$$

It follows that the models $D = 1\{\mu_D(P, U_D) \geq \tilde{V}\}$ and $D = 1\{Q(P(Z), \sigma U) \geq V\}$ are observationally equivalent. □

A direct implication of Proposition 4 is that once we have estimated a selection model that is of the form $D = 1\{\mu_D(P, U_D) \geq \tilde{V}\}$ and satisfies index sufficiency we can construct

$$var(Q(p, \sigma U)) = var(F_{\tilde{V}}(\tilde{\mu}(p, U_D)). \tag{33}$$

The following proposition shows in which sense it is meaningful to think of a deviation from $var(Q(p, \sigma U)) = 0$ as a deviation from monotonicity.

Proposition 5 (Monotonicity): *Under the assumptions of Proposition 4 any selection model that is of the form* $D = 1\{\mu_D(P, U_D) \geq \tilde{V}\}$ *has an equivalent representation* $D = 1\{Q(P(Z), \sigma U) \geq V\}$ *with* $\sigma = 0$ *and* $var(Q(p, \sigma U)) = 0$ *if, and only if, monotonicity holds.*

Proof. The "if" part follows from Vytlacil (2006). The "only if" part holds because monotonicity can only be violated if $var(Q(p, \sigma U)) > 0$. A necessary condition for this is $\sigma > 0$. □

The proposition does not say that $\sigma > 0$ or $var(Q(p, \sigma U)) > 0$ is sufficient for a violation of monotonicity. It could be that $Q(P, \sigma U)$ is specified in a way such that monotonicity holds but $var(Q(p, \sigma U)) > 0$. Then, the results that are presented in Section 4 continue to hold. That is, if the selection model is specified to be

$$D = 1\{Q^a(P, \sigma U) \geq V^a\},$$

and $var(Q^a(p, \sigma U)) > 0$ then estimates of $\mathbb{E}[Y_1 - Y_0|V^a = v^a]$ and the other treatment parameters could be biased. This is because we still have

$$\frac{\partial \mathbb{E}[Y|P = p]}{\partial p} \neq \mathbb{E}[Y_1 - Y_0|V^a = p].$$

However, Proposition 5 implies that under monotonicity another representation could be found such that[35]

$$D = 1\{Q^b(P, 0) \geq V^b\}$$

and as shown in Section 2.2.1 we have

$$\frac{\partial \mathbb{E}[Y|P = p]}{\partial p} = \mathbb{E}[Y_1 - Y_0|V^b = p].$$

$\mathbb{E}[Y_1 - Y_0|V^b = p]$ is a meaningfully re-defined parameter of interest. Such a re-definition is only possible under monotonicity (Heckman and Vytlacil, 2005, p. 718).

Appendix C: Further Details on the Application

In Section 4 we use the random coefficient Logit model

$$D = 1\{Z\tilde{\gamma} \geq \tilde{V}\} \tag{34}$$

[35] Under monotonicity we have that $\partial Q^a(p, \sigma u)/\partial p$ is positive for all u in the support of U. But then, Q^a is invertible in its first argument and we can rewrite the selection model as $D = 1\{P(Z) \geq Q^{a,-1}(V, \sigma U)\}$, where $Q^{a,-1}$ is the inverse of Q^a with respect to its first argument.

to estimate $var(Q(p, \sigma U))$. $\tilde{V}$ follows the logistic distribution so that

$$\Pr(D = 1|Z, \tilde{\gamma}) = \frac{\exp(Z\tilde{\gamma})}{1 + \exp(Z\tilde{\gamma})}.$$

This implies

$$\Pr(D = 1|Z; \gamma, \Sigma) = \int \left(\frac{\exp(Z\tilde{\gamma})}{1 + \exp(Z\tilde{\gamma})} \right) \phi(\tilde{\gamma}|\gamma, \Sigma)\, d\tilde{\gamma}, \tag{35}$$

where $\phi(\tilde{\gamma}|\gamma, \Sigma)$ is the joint normal density with mean γ and variance-covariance matrix Σ evaluated at $\tilde{\gamma}$. In the estimation step (35) is simulated for a given set of parameters (γ, Σ) and given the observed values z_i of Z, where i indexes individuals. Then the simulated log likelihood for the sample is maximized over the choice of those parameters.[36]

The results in this paper are derived under the assumption of index sufficiency. Index sufficiency means that the instruments affect the outcome only through their impact on the treatment probability. It holds within a state because the probability to complete more than 9 years of education is increasing in the quarter of birth. Proposition 4 and (33) then show how we can simulate $var(Q(p, \sigma U))$. For this we use

$$\mu_D(p, \tilde{\gamma}) = \int_{z \in \mathscr{Z}_p} \frac{\exp(z\tilde{\gamma})}{1 + \exp(z\tilde{\gamma})} f_z(z)\, dz,$$

where $U_D = \tilde{\gamma}$ and $\mathscr{Z}_p$ is the set of values of Z such that $P(Z) = p$.[37]

Finally, we provides a testable sufficient condition for index sufficiency for the more general case when the instrument is not a scalar and $\mathscr{Z}_p$ has more than one element.

Proposition 6: *Suppose that the selection model is given by* (34)*, that Z, $\tilde{\gamma}$ and $\tilde{V}$ are mutually independent, and that Z and $\tilde{\gamma}$ are independent of* (Y_0, Y_1)*. Then, index sufficiency holds if for any two values z^a and z^b of Z such that $P(z^a) = P(z^b) = p$ we have that the distribution of $z^a\tilde{\gamma}$ is equal to the distribution of $z^b\tilde{\gamma}$.*

Proof. Define $W^a \equiv z^a\tilde{\gamma}$ and $W^b \equiv z^b\tilde{\gamma}$. By the assumption that the random coefficients are independent of Y_1 and $\tilde{V}$ we have that W^a and W^b are independent of Y_1 and $\tilde{V}$, respectively. Therefore, index sufficiency,

$$Pr(Y_0 \in A|P = p, D = 0) = Pr(Y_0 \in A|W^a < \tilde{V}) = Pr(Y_0 \in A|W^b < \tilde{V})$$

and

$$Pr(Y_1 \in A|P = p, D = 1) = Pr(Y_1 \in A|W^a \geq \tilde{V}) = Pr(Y_1 \in A|W^b \geq \tilde{V}),$$

holds W^a and W^b are equal in distribution. □

[36] See Train (2003) for details. I used the STATA program mixlogit by Arne Risa Hole to estimate γ and Σ.

[37] This notation is for the general case in which $P(Z)$ is not one to one. If $P(Z)$ is one to one then $\mathscr{Z}_p$ contains only one value of Z (provided that p is in the support of P).

Acknowledgements

This paper is based on the first chapter of my Ph.D. thesis that I defended in July 2006 at the University of Mannheim. I am especially grateful to Andrew Chesher and Enno Mammen for numerous insightful discussions. Moreover, I would like to thank Jaap Abbring, Josh Angrist, Otilia Boldea, Pedro Carneiro, Pavel Čížek, Jim Heckman, Pierre Hoonhout, Jürgen Maurer, Enrico Rettore, Cristina Santos, Ed Vytlacil, seminar participants, and conference participants of various conferences for helpful comments. 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. Finally, I would like to thank the associate editor and the referees for their very helpful comments.

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.

Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings," *Quarterly Journal of Economics*, 106(4), 979–1014.

——— (1992): "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 87(418), 328–336.

Barua, R., and K. Lang (2008): "School Entry, Educational Attainment and Quarter of Birth: A Cautionary Tale of LATE," Working Paper, Boston University, Boston, MA, USA.

Battistin, E., and A. Chesher (2004): "The Impact of Measurement Error on Evaluation Methods Based on Strong Ignorability," Working Paper, University College London, London, UK.

Battistin, E., and E. Rettore (2008): "Ineligible and Eligible Non-Participants as a Double Comparison Group in Regression-Discontinuity Designs," *Journal of Econometrics*, 142(2), 715–730.

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.

Bound, J., and D. A. Jaeger (2000): "Do Compulsory School Attendance Laws Alone Explain the Association between Quarter of Births and Earnings?," *Research in Labor Economics*, 19, 83–108.

Briesch, R. A., P. K. Chintagunta, and R. L. Matzkin (2007): "Nonparametric Discrete Choice Models with Unobserved Heterogeneity," Working Paper, UCLA.

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(2), 191–208.

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.

Fox, J. T., and A. Gandhi (2008): “Identifying Heterogeneity in Economic Choice and Selection Models Using Mixtures,” Working Paper, University of Chicago, Chicago, IL, USA.

Gautier, E., and Y. Kitamura (2008): “Nonparametric Estimation in Random Coefficients Binary Choice Models,” CREST Working Paper 2008-15, CREST, Paris, France.

Hahn, J., P. Todd, and W. van der Klaauw (2001): “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design,” *Econometrica*, 69(1), 201–209.

Harmon, C., and I. Walker (1999): “The Marginal and Average Returns to Schooling in the UK,” *European Economic Review*, 43(4-6), 879–887.

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 (1998): “Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling,” *Journal of Human Resources*, 33(4), 974–987.

——— (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.

——— (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. by C. Hsiao, K. Morimune, and J. Powell, pp. 1–46. Cambridge University Press, Cambridge.

——— (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. (2009): “College Education and Wages in the U.K.: Estimating Conditional Average Structural Functions in Nonadditive Models with Binary Endogenous Variables,” Working Paper, Tilburg University.

Korn, E. L., and S. Baumrind (1998): “Clinician Preferences and the Estimation of Causal Treatment Differences,” *Statistical Science*, 13, 209–227.

Korn, E. L., D. M. Teeterb, and S. Baumrind (2001): “Using Explicit Clinician Preferences in Nonrandomized Study Designs,” *Journal of Statistical Planning and Inference*, 96(1), 67–82.

Lewbel, A. (2000): “Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables,” *Journal of Econometrics*, 97(1), 145–177.

Matzkin, R. L. (2007): “Heterogeneous Choice,” in *Advances in Economics and Econometrics, Theory and Applications, Ninth World Congress of the Econometric Society*, ed. by R. Blundell, W. K. Newey, and T. Persson. Cambridge University Press.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge, United Kingdom.

Small, D. S., and Z. Tan (2007): “A Stochastic Monotonicity Assumption for the Instrumental Variables Method,” Working Paper, University of Pennsylvania, Philadelphia, PA, USA.

Train, K. E. (2003): *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, UK.

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

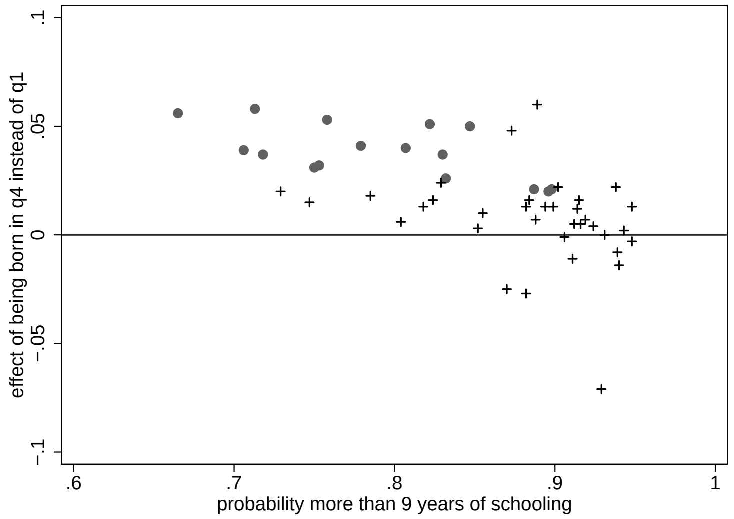
——— (2006): "A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results," *Oxford Bulletin of Economics and Statistics*, 68(4), 515–518.

Wooldridge, J. M. (1997): "On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model," *Economics Letters*, 56(2), 129–133.

——— (2003): "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," *Economics Letters*, 79(2), 185–191.

Tables and Figures

Empirical Application



Effect of being born in the 4th instead of the 1st quarter on the probability to attend more than 9 years of schooling, plotted against probability for those who are born in the first quarter. Each data point in this figure represents one state. The dots represent significant effects at the 5 per cent level. The pluses represent insignificant effects. Results were obtained using a regression of an indicator variable for more than 9 years of schooling on a full set of interaction terms between state and quarter of birth indicators. 329,509 observations.

Figure 1: Effect of being born in 4th instead of 1st quarter.

	(1)	(2)	(3)	(4)
year of birth (mean)	0.062	0.077	0.067	0.076
	(0.002)**	(0.003)**	(0.002)**	(0.002)**
year of birth (std.)		0.016		0.014
		(0.001)**		(0.002)**
born in 1st quarter (mean)	-0.369	1.669	-1.182	-1.293
	(0.061)**	(0.300)**	(0.068)**	(0.085)**
born in 1st quarter (std.)		3.697		0.736
		(0.302)**		(0.084)**
born in 2nd quarter (mean)	-0.343	-0.708	-1.163	-1.425
	(0.061)**	(0.095)**	(0.068)**	(0.081)**
born in 2nd quarter (std.)		0.411		0.069
		(0.285)		(0.257)
born in 3rd quarter (mean)	-0.270	-0.657	-1.085	-1.252
	(0.061)**	(0.161)	(0.068)**	(0.083)**
born in 3rd quarter (std.)		0.305		0.552
		(0.084)**		(0.106)**
born in 4th quarter (mean)	-0.224	-0.469	-1.026	-1.144
	(0.061)**	(0.089)**	(0.068)**	(0.086)**
born in 4th quarter (std.)		0.713		0.667
		(0.104)**		(0.100)**
state of birth indicators	no	no	yes	yes

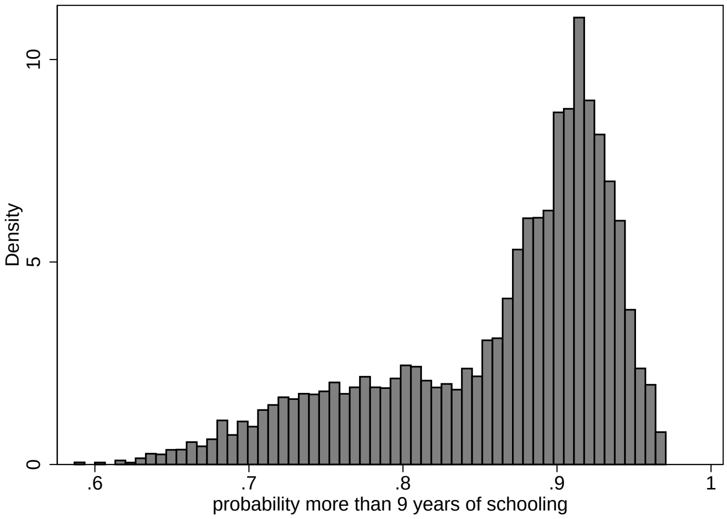
329,509 observations. Standard errors in parentheses. * significant at 5%; ** significant at 1%. Column (1) and (3) are Logit estimates. Column (2) and (4) are Logit estimates with independently normally distributed random coefficients. Estimates of means and standard deviations of random coefficients are reported.

Table 1: First stage estimates.

	(2)	(4)
1st instead of 4th quarter		
mean	2.138	-0.149
standard deviation	3.765	0.993
percentage positive	71%	44%
2nd instead of 4th quarter		
mean	-0.239	-0.281
standard deviation	0.823	0.671
percentage positive	39%	34%
3rd instead of 4th quarter		
mean	-0.188	-0.108
standard deviation	0.775	0.866
percentage positive	40%	45%

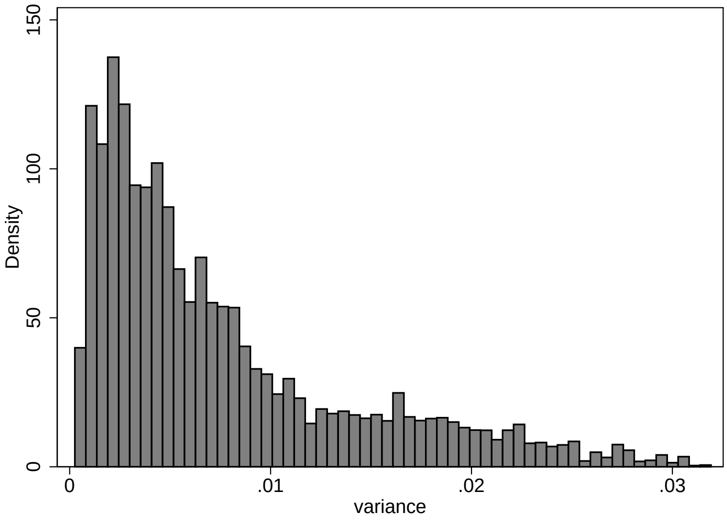
Mean difference in the coefficient on quarter of birth indicators and sum of estimated standard deviations of the respective random coefficients are reported. Obtained from the estimation results reported in column (2) and (4) of Table 1. The labeling of the columns has been adopted. The percentage positive is given by one minus the cumulative normal distribution with the respective mean and standard deviation evaluated at zero.

Table 2: Impact of hypothetical change in quarter of birth.



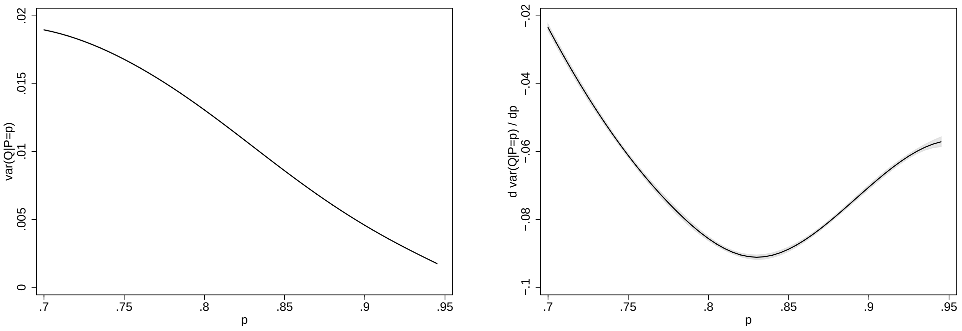
Obtained from the estimation results reported in column (4) of Table 1.

Figure 2: Fitted probability of attending more than 9 years of schooling.



Obtained from the estimation results reported in column (4) of Table 1.

Figure 3: Individual variance of fitted probability.



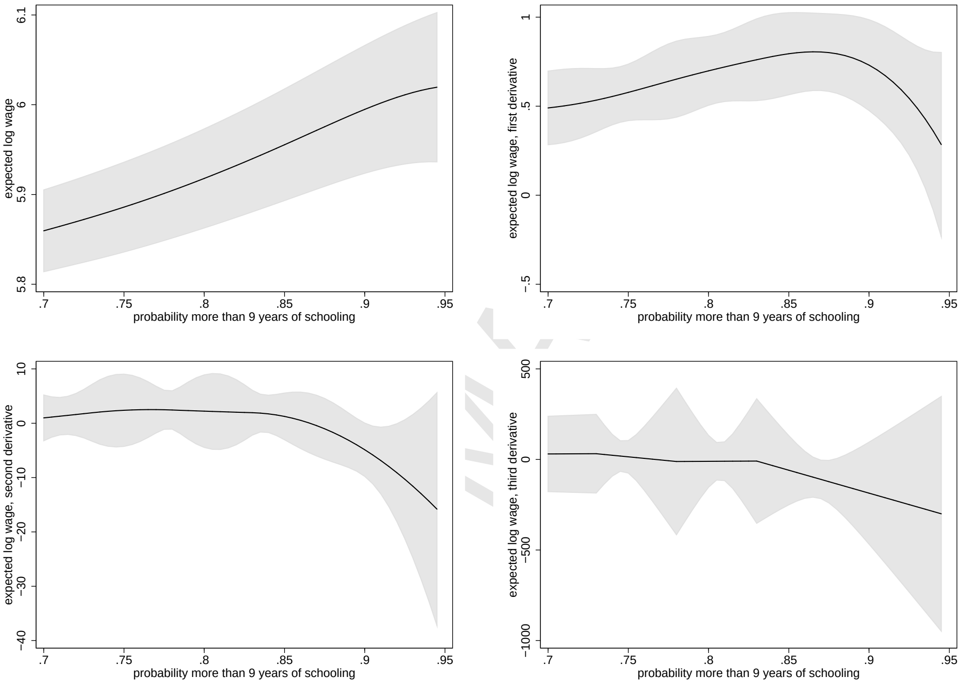
For these plots the simulated value of $var(Q(p, \sigma U))$ was regressed on a set of fourth order splines in the simulated value p of P. The simulations are based on the parameter estimates reported in column (4) of Table 1. 95 per cent confidence intervals are reported (they are very narrow). They do not account for the first stage estimation and simulation error.

Figure 4: Level of individual variance and its derivative plotted against fitted probability.

	effect of more than 9 years of schooling
ordinary least squares	0.395
	(0.002)**
linear instrumental variables estimator	0.884
	(0.119)**

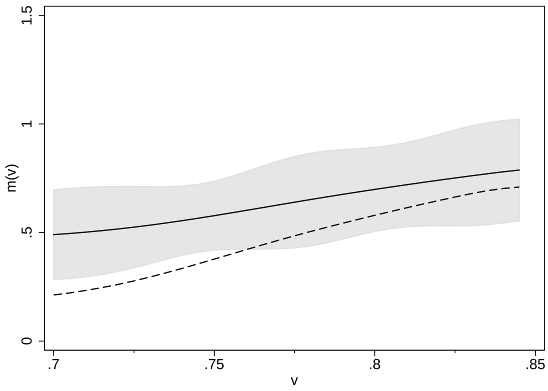
329,509 observations. Standard errors in parentheses. * significant at 5%; ** significant at 1%. A full set of state of birth indicators as well as the year of birth was included as covariates. The linear instrumental variables estimate was obtained using the efficient 2 step GMM estimator with quarter of birth indicators as instruments. Hanson's J statistic (test of overidentifying restrictions): 6.112, p-value: 0.047.

Table 3: OLS and standard IV estimates.



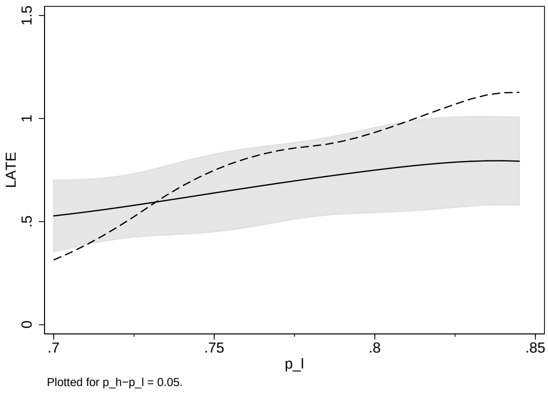
Results were obtained using an ordinary least squares regression of the log weekly wage on the year of birth, state of birth indicators, as well as fourth order splines in the simulated value of the probability to attend more than 9 years of schooling, as specified in (35). 95 per cent confidence intervals are reported. These do not account for the first stage estimation and simulation error.

Figure 5: Second stage results.



The solid line is the biased estimate and the dashed line is the bias corrected estimate. Results are presented for v between 0.7 and 0.85. 95 per cent confidence intervals for the biased estimate are reported. These do not account for the first stage estimation and simulation error. Results were obtained using the first and second stage estimation results as well as (15).

Figure 6: Biased and bias corrected estimate of the marginal treatment effect.



The local average treatment effect is calculated for a 5 per cent change in the probability of attending more than 9 years of schooling. The solid line is the biased estimate and the dashed line is the bias corrected estimate. Results are presented for p_l between 0.7 and 0.85. 95 per cent confidence intervals for the biased estimate are reported. These do not account for the first stage estimation and simulation error. Results were obtained using the first and second stage estimation results as well as (16).

Figure 7: Biased and bias corrected estimate of the local average treatment effect.

Monte Carlo Study

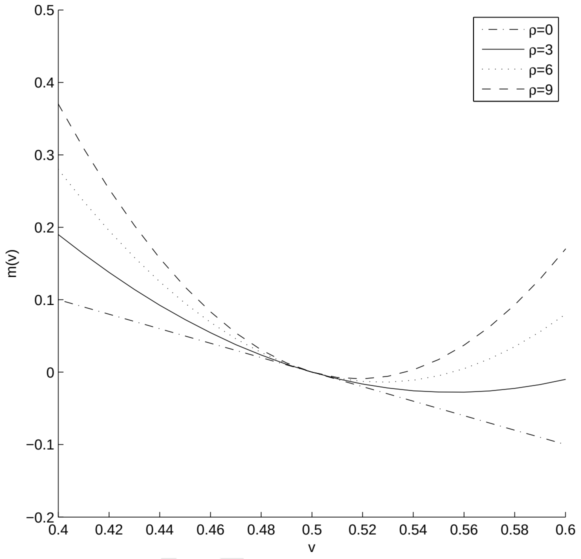


Figure 8: Marginal treatment effect for different values of ρ.

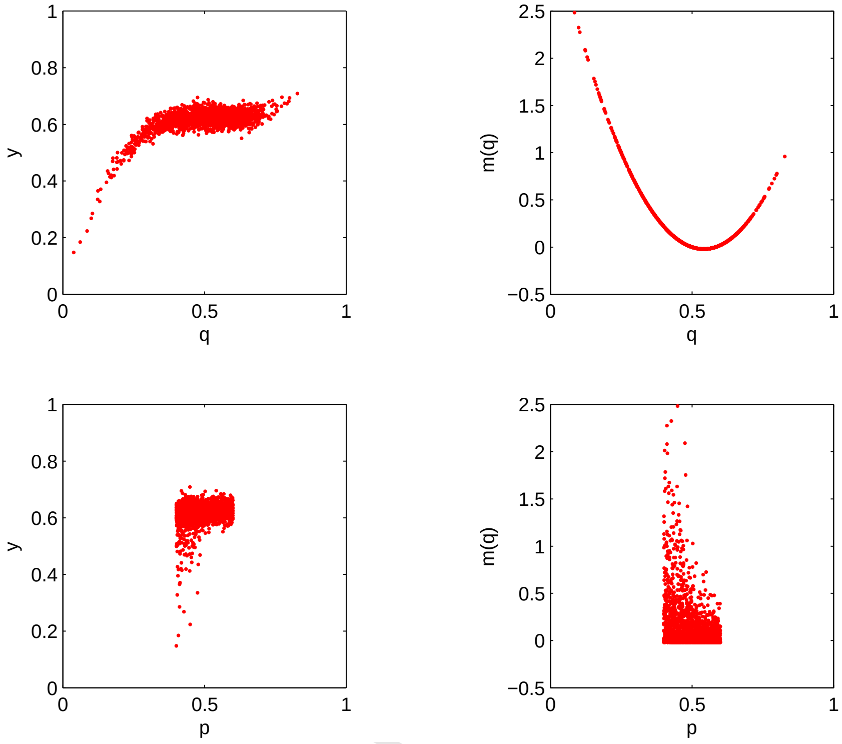


Figure 9: One draw of generated data for $\rho = 4$ and $\sigma^2_{0.5} = 0.01$.

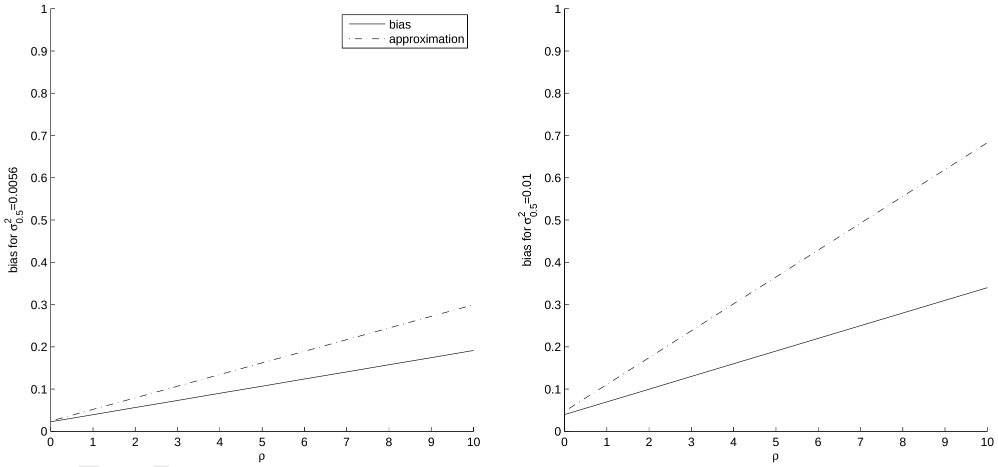


Figure 10: Mean bias and approximations for different values of ρ and $\sigma^2_{0.5}$.

σ	$\sigma^2_{0.5}$	bias	approximation	difference
0.000	0.000	0.0000	0.0000	0.0000
		(0.0002)	(0.0000)	(0.0002)
0.050	0.000	0.0027	0.0025	-0.0002
		(0.0002)	(0.0000)	(0.0002)
0.100	0.001	0.0104	0.0104	0.0001
		(0.0002)	(0.0001)	(0.0001)
0.150	0.001	0.0227	0.0252	0.0025
		(0.0002)	(0.0001)	(0.0001)
0.200	0.003	0.0399	0.0489	0.0090
		(0.0002)	(0.0003)	(0.0001)
0.250	0.004	0.0628	0.0832	0.0205
		(0.0002)	(0.0005)	(0.0002)
0.300	0.006	0.0901	0.1344	0.0443
		(0.0003)	(0.0008)	(0.0006)
0.350	0.008	0.1220	0.2057	0.0837
		(0.0003)	(0.0014)	(0.0011)
0.400	0.010	0.1599	0.3017	0.1417
		(0.0004)	(0.0023)	(0.0019)
0.450	0.013	0.2024	0.4279	0.2255
		(0.0005)	(0.0037)	(0.0032)
0.500	0.016	0.2502	0.5930	0.3428
		(0.0006)	(0.0057)	(0.0051)

Bias, approximation (15) using biased estimates of first and second derivative of $m(V)$ evaluated at $V = 0.5$, and the difference between the two. Reported for $\rho = 4$. $\sigma^2_{0.5}$ was calculated according to (20). Respective means were calculated from 10,000 simulations. Standard deviations of means are reported in parentheses.

Table 4: Dependence of bias and approximation on σ.

						ρ					
σ	$\sigma^2_{0.5}$	1	2	3	4	5	6	7	8	9	10
0.050	0.000	85	89	92	93	94	95	96	97	97	97
0.100	0.001	93	97	99	101	102	102	103	103	104	104
0.150	0.001	106	109	110	111	112	112	113	113	113	114
0.200	0.003	116	119	121	123	123	124	124	125	125	125
0.250	0.004	119	126	130	133	134	136	137	137	138	138
0.300	0.006	131	141	146	149	151	153	154	155	156	157
0.350	0.008	143	158	164	169	171	174	175	176	177	178
0.400	0.010	159	175	183	189	192	195	197	199	200	201
0.450	0.013	173	193	205	211	216	220	222	224	226	227
0.500	0.016	189	215	228	237	243	247	250	253	255	257

This table shows, for a given ρ and σ, the percentage of the bias that is corrected using the approximation (15) with biased estimates of the first and second derivative of $m(V)$ evaluated at $V = 0.5$. $\sigma^2_{0.5}$ was calculated according to (20). Respective means were calculated from $10,000$ simulations.

Table 5: Accuracy of the approximation.