Universität St.Gallen

Nonparametric IV estimation of local average treatment effects with covariates

Markus Frölich

September 2002 Discussion paper no. 2019

Department of Economics University of St. Gallen

Editor: Prof. Jörg Baumberger
University of St. Gallen
Department of Economics
Bodanstr. 1
CH-9000 St. Gallen
Phone ++41 71 224 22 41
Fax ++41 71 224 28 85
Email joerg.baumberger@unisg.ch

Publisher: Forschungsgemeinschaft für Nationalökonomie
an der Universität St. Gallen
Dufourstrasse 48
CH-9000 St. Gallen
Phone ++41 71 224 23 00
Fax ++41 71 224 26 46

Electronic Publication: www.fgn.unisg.ch/public/public.htm

Nonparametric IV estimation of local average treatment effects with covariates

Markus Frölich[1]

Author's address:

Markus Frölich
Swiss Institute for International Economics and Applied
Economic Research (SIAW)
Dufourstrasse 48
CH-9000 St. Gallen
Tel. ++41 71 2242342
Fax ++41 71 2242298
Email markus.froelich@unisg.ch
Website www.siaw.unisg.ch/froelich, www.markusfroelich.de

[1] The author is also affiliated with the Institute for the Study of Labor (IZA), Bonn. Financial support from the Swiss National Science Foundation (Project NSF 4043-058311) is gratefully acknowledged. I would like to thank Michael Lechner, Patrick Puhani and seminar participants at the University of Konstanz for discussions, comments and suggestions.

Abstract

In this paper nonparametric instrumental variable estimation of local average treatment effects (LATE) is extended to incorporate confounding covariates. Estimation of local average treatment effects is appealing since their identification relies on much weaker assumptions than the identification of average treatment effects in other nonparametric instrumental variable models. Including covariates in the estimation of LATE is necessary when the instrumental variable itself is endogenous (e.g. when the instrument is self-selected). However, all previous approaches to handle covariates in the estimation of LATE rely on parametric or semiparametric methods. In this paper, a nonparametric estimator for the estimation of LATE with covariates is suggested that is root-n asymptotically normal and efficient.

Keywords

Instrumental variables, LATE, evaluation, treatment effect, matching, unobserved heterogeneity

JEL Classification

C13, C14

1 Introduction

Instrumental variables regression is a fundamental approach to causal reasoning in econometrics. In many applications one wants to uncover the *causal* relationship between a variable D and an outcome variable Y, where the variable D is itself endogenous. For example, if D is years of schooling and Y is wages, it is of interest to learn by how much wages increase due to an additional year of schooling. In another example, where D is union membership and Y is wages, one would like to know which wages would be observed if the union members were non-members or vice versa. Or, if D represents participation in a training programme and Y is subsequent employment status, it is of interest how the employment probability is affected by participation in the training programme. In these examples, one would like to know how D causally affects Y, i.e. how an *exogenous* variation in D would change the variable Y. Since the variable D is endogenous, a regression of Y on D does not uncover a causal (structural) relationship. Nevertheless, if a variable Z exists that affects only D but not Y, then an exogenous variation in Z induces an exogenous variation in D and thus overcomes the endogeneity of D. Such a variable Z is called an instrumental variable and has been exploited in numerous studies to identify the effects of D on Y.

If the values of the instrumental variable Z are assigned completely at random, any variation in Z is exogenous and thus generates an unconfounded variation in D, which identifies the relationship between D and Y. For example, Hearst, Newman, and Hulley (1986) and Angrist (1990) use the Vietnam era conscription lottery as an instrument to identify the effects of mandatory military conscription on subsequent civilian mortality and earnings. Imbens, Rubin, and Sacerdote (2001) use 'winning a prize in the lottery' as an instrument to identify the effects of unearned income on subsequent labour supply, earnings and consumption behaviour. In both examples the instrument is randomly assigned (by a lottery).

However, in many applications the instrument Z itself is endogenous and confounded with D or Y. For example, college proximity may be used as an instrument to identify the returns to schooling, noting that living close to a college during childhood may induce some children to go to college but is unlikely to affect the wages earned in their adulthood directly (Card 1995). Nevertheless, the instrument college proximity is not randomly assigned but chosen by the parents. Their choice, however, might itself be related to characteristics that affect their children's subsequent wages directly. Parental education is another example of an instrumental variable that is often used to identify the returns to schooling. It appears reasonable that parental schooling itself has no direct impact on their children's wages. Nevertheless, it is likely to be correlated with parents' profession, family income and wealth, which may directly affect the wage prospects of their offspring. In these cases it is necessary to control for these confounding covariates X to handle the endogeneity of the instrumental variable Z.

Conventional approaches to accommodate covariates X in instrumental variables estimation (for example two-stage least squares) proceed by specifying functional form restrictions on the conditional expectation functions of Y and D. Recently, *nonparametric* identification and estimation in instrumental variable models, avoiding such delicate functional form assumptions, has received a lot of interest, see Newey and Powell (first draft 1988, revised 2002), Newey, Powell, and Vella (1999), Das (2000), Blundell and Powell (2001), Darolles, Florens, and Renault (2001), Imbens and Newey (2001), Florens (2002) and Florens, Heckman, Meghir, and Vytlacil (2002). However, their approaches still impose identifying assumptions which may not be satisfied in many applications. Most models rely on additive separability in the error term, which amounts to assuming that, conditional on X, the relationship between D and Y is identical for each individual up to an intercept. In other words, a *constant treatment effect* for individuals with the same value of X is assumed. Additively-separable models, thus, rule out unobserved heterogeneity and therefore may not be appropriate in many applications. In non-separable models, however, identification essentially requires that the instrument is sufficiently powerful to move the value of D_i (for any individual i) over the entire support of the variable D, see Blundell and Powell (2001), Florens, Heckman, Meghir, and Vytlacil (2002) and Imbens and Newey (2001).[1] Yet, such powerful instruments are often not available. In this case, the relationship between D and Y can only be uncovered for the subpopulation that reacts on changes of the instrument Z.

This is the concept of the *local average treatment effect* (LATE) of Imbens and Angrist (1994). The local average treatment effect is the mean effect on Y of a change in D for the subpopulation of compliers, where the compliers are all individuals whose value of D would change if the instrument Z were modified exogenously. In spite of its appealing properties, however, the LATE-concept has not been fully extended to accommodate covariates X. Although identification of local average treatment effects with covariates is straightforward, nonparametric

[1] This is similar to the identification-at-infinity argument in selection models.

estimation with covariates has not been attempted so far. All previous approaches to incorporate covariates for estimating local average treatment effects (e.g. Abadie (2001), Angrist, Graddy, and Imbens (2000), Hirano, Imbens, Rubin, and Zhou (2000), Yau and Little (2001) among many others) always resorted to parametric or semiparametric approaches. They refrained from nonparametric estimation methods because of their low precision and the curse of dimensionality.

In this paper it is shown that the average treatment effect for the compliers can be estimated fully nonparametrically at $\sqrt{n}$-convergence rate even with covariates. The proposed estimator is asymptotically normal and efficient. Thus using nonparametric regression to accommodate covariates X in the estimation of local average treatment effects does not give rise to the curse of dimensionality. This result is similar to the $\sqrt{n}$-convergence of nonparametric matching estimators in the treatment evaluation literature, see Heckman, Ichimura, and Todd (1998). (Indeed, the proposed conditional LATE estimator corresponds to a ratio of two matching estimators.)

Section 2 gives an overview of nonparametric instrumental variables methods. Section 3 introduces the nonparametric conditional LATE estimator and derives its properties. In addition, several extensions are discussed. Section 4 concludes.

2 Instrumental variable regression

Nonparametric instrumental variable estimation of the relationship between an endogenous variable D and an outcome variable Y is often analyzed in an *additively separable* model

$$Y_i = \varphi\left(D_i, X_i\right) + u_i,$$

where i denotes an element (unit/individual) of the population, D_i is the observed value of the endogenous regressor for unit i, X_i is the observed value of a (possibly empty) set of exogenous variables and u_i is an error term. φ is a structural (causal) function based on the following *potential outcomes* concept: For unit i the variables D, X and u take the values D_i, X_i and u_i and the outcome $Y_i = \varphi\left(D_i, X_i\right) + u_i$ is observed. If the variable D_i were manipulated by an external intervention to take the value d (without changing the values of X_i and u_i), the outcome $\varphi\left(d, X_i\right) + u_i$ would be observed. For instance, if $D \in \{0,1\}$ is union status and Y is wages, the *potential wages* for unit i are the wage that unit i would receive if union membership were set by some external intervention to 0 and the wage that unit i would receive if union membership were set to 1, ceteris paribus. Analogously, if D is years of schooling, the potential wages for unit i are the wages that unit i would receive if years of schooling of unit i were set externally to different levels. Accordingly, the difference $\varphi(d=1,x) - \varphi(d=0,x)$ is the *causal effect* of union-membership or of one versus zero years of schooling, respectively, on wages, for units with characteristics $X = x$. The defining feature of the additively separable model is that it is conceived that D_i could be manipulated without affecting X_i or u_i.

If instrumental variables Z are available that are related to D but unrelated to u (given X)

$$E[u|Z,X] = 0,$$

the structural function φ is identified, see Newey and Powell (first draft 1988, revised 2002), Newey, Powell, and Vella (1999), Das (2000), Darolles, Florens, and Renault (2001), Florens (2002) and the survey article of Blundell and Powell (2001).[2]

[2] However, consistent estimation of φ is rather difficult, because the reduced form relationship $E[Y|Z=z,X=x] = \int \varphi(d,x)\, dF_{d|z,x}$ needs to be inverted with respect to the integral operator. Since the inverse of the integral operator is discontinuous, small changes in $E[Y|Z,X]$ may give rise to large changes in φ. To overcome this ill-posed inverse problem, Newey and Powell (first draft 1988, revised 2002) restrict the set of possible $\varphi(d)$ functions and introduce a series approximation to $\varphi(d)$. Darolles, Florens, and Renault (2001), Florens (2002) propose a regularization of the integral equation for estimating $\varphi(d)$. Das (2000) avoids the discontinuity problem by restricting the endogenous regressor D_i to be discrete. Newey, Powell, and Vella (1999) consider an alternative additively-separable model, which is more restrictive but easier to estimate than the model based on the assumption $E[u|Z,X] = 0$ alone. They specify additionally an additively-separable structure for the endogenous regressor: $D_i = \zeta(Z_i, X_i) + v_i$ and suppose that $E[v|Z,X] = 0$ and $E[u|v,Z,X] = E[u|v]$. The basic assumption of this model is that the endogeneity of D (i.e. the difference $E[\varphi(d)] \neq E[\varphi(d)\,|D=d]$) is entirely driven by v and not by any interaction between v and Z or X. This model implies that the error term v_i is identified as $v_i = D_i - E[D|Z=Z_i, X=X_i]$ and hence

$$\begin{aligned} E[Y|D=d,Z=z,X=x] &= \varphi(d,x) + E[u|D=d,Z=z,X=x] \\ &= \varphi(d,x) + E[u|Z=z,X=x,V=d-E[D|Z=z,X=x]\ ] \\ &= \varphi(d,x) + E[u|V=d-E[D|Z=z,X=x]\ ] \\ &= \varphi(d,x) + \xi(v) \end{aligned}$$

where $\xi(v)$ is a function of v and $v(d,z,x) = d - E[D|Z=z,X=x]$. The function $\xi(v)$ is called the *control function* and captures all the endogeneity that comes through the relation between u_i and v_i. Independent variation in v and x, induced by variation in the instruments z, separates the structural function $\varphi(d,x)$ from the endogeneity correction term $\xi(v)$.

However, the assumption of an additively-separable structure is often inadequate, since it heavily restricts the permitted heterogeneity among units: Although units may differ in their unit-specific error terms u_i, the causal effect of setting D_i to d_1 versus setting D_i to d_2 is supposed to be identical for all units with the same x value: $\varphi(d_1, x) - \varphi(d_2, x)$. This amounts to assuming that the effect of union membership or of an additional year of schooling is identical for all units i with the same characteristics X. This *constant treatment effect assumption* (conditional on x) is in many situations, depending on which X variables are observed by the econometrician, rather implausible (Heckman 1997). For example, the return to schooling may be interacted with unobserved ability. In such cases, different units react differently on an external intervention on D, and the structural function φ itself varies among units:

$$Y_i = \varphi_i(D_i).$$

The function φ_i is still conceived in a counterfactual sense: $\varphi_i(D_i)$ is the observed outcome and $\varphi_i(d)$ is the outcome that would be observed if D_i had been set to d. However, φ_i may now differ among units in an arbitrary way. Identification of the average structural function $E[\varphi_i(d)]$ or of average treatment effects $E[\varphi_i(d_1) - \varphi_i(d_2)]$ in such non additively-separable models is analyzed by Blundell and Powell (2001), Florens, Heckman, Meghir, and Vytlacil (2002)[3] and Imbens and Newey (2001), using a control function approach. Suppose D_i is generated by

$$D_i = \zeta(Z_i, v_i), \tag{1}$$

where v_i is an error term. If $\zeta(Z_i, v_i)$ is assumed to be additive in Z_i and v_i (Florens, Heckman, Meghir, and Vytlacil 2002) or to be strictly monotone in v_i (Imbens and Newey 2001), v_i can be identified from D_i and Z_i. If one assumes that the endogeneity of D_i is generated entirely through the error term v_i, similarly to Newey, Powell, and Vella (1999), then the endogeneity (i.e. $E[\varphi_i(d)] \neq E[\varphi_i(d) \,|D_i = d]$) can be controlled for by conditioning on v:

$$E[\varphi_i(d) \,|D_i = d, V_i = v] = E[\varphi_i(d) \,|V_i = v]. \tag{2}$$

[3] Florens, Heckman, Meghir, and Vytlacil (2002) analyze identification of $\partial E[\varphi_i(d)]/\partial d$ when D is continuous. They also consider a local instrumental variable (LIV) condition, which essentially assumes that the average treatment effect and a local average treatment effect (discussed below) are identical.

With this assumption, the average structural function $E[\varphi_i(d)]$ can be identified by

$$\begin{aligned}\int E[Y|D=d,V=v]\,dF_v &= \int E[\varphi_i(d)\,|D_i=d,V_i=v]\,dF_v \\ &= \int E[\varphi_i(d)\,|V_i=v]\,dF_v = E[\varphi_i(d)],\end{aligned}$$

since v_i is identified by the additivity or monotonicity assumption in (1).[4]

Yet, a central condition for identification is that the conditional expectation $E[Y|D=d, V=v]$ is defined at every v in the support of V. This requires that the support of the distribution of V given $D=d$ is the same as the support of the marginal distribution of V.

This condition, thus, requires either that the distribution of V is somehow restricted or, otherwise, that every unit i could be induced to take the value $D_i=d$ through a change in Z_i. In other words, the instrument Z_i must be sufficiently powerful to move the regressor D_i, for every unit i, to any value d where $E[\varphi_i(d)]$ shall be estimated (Imbens and Newey 2001).

2.1 Local average treatment effects

However, in many applications the instruments available are not so powerful. It is often highly unreasonable to assume that *all* units i could be induced, through a modification of the instrument Z_i, to change D_i to a particular value. Consider the situation where D and Z are binary. The above assumption would require that all units switch D from 0 to 1 or vice versa if Z is changed from 0 to 1. If this assumption does not hold, the relationship between D and Y can be analyzed only for the subpopulation which is affected by the instrumental variable. Hence an average causal effect can no longer be identified for the full population, but only a *local average treatment effect* (LATE) for the subpopulation of units that could be induced to change D through a variation in the instrumental variable.

The local average treatment effect has been introduced by Imbens and Angrist (1994) and further analyzed by Angrist and Imbens (1995), Angrist, Imbens, and Rubin (1996), Imbens and Rubin (1997), Heckman and Vytlacil (1999), Abadie (2001) and Imbens (2001), among others. Most of the discussion on LATE focuses on the case where the instrumental variable Z itself is exogenous, i.e. not confounded with D or Y. Identification of local average treatment

[4] If $\zeta(Z_i,v_i)$ is monotone in v_i, then $v_i=F_{D|Z}(D_i,Z_i)$, see Imbens and Newey (2001). If $\zeta(Z_i,v_i)=\xi(Z_i)+v_i$ is additive, then $v_i=D_i-E[D|Z=Z_i]$, see Blundell and Powell (2001) and Florens, Heckman, Meghir, and Vytlacil (2002).

effects with confounding covariates X has been discussed in Angrist and Imbens (1995), Heckman and Vytlacil (1999), Abadie (2001) and Imbens (2001). However, *nonparametric estimation* of local average treatment effects with confounding covariates X has not been attempted so far. Extensions to embed covariates X in the estimation of LATE have usually resorted to parametric or to semiparametric approaches. Angrist, Graddy, and Imbens (2000) and Yau and Little (2001) incorporate covariates by assuming that they enter linearly in the conditional expectation functions. Hirano, Imbens, Rubin, and Zhou (2000) suggest to model the probability of being an always-taker, never-taker or a complier given covariates X by a trinomial logistic distribution and to model the outcome distributions separately for these types. Abadie (2001) initially introduces covariates in a nonparametric way but proposes parametric and semiparametric methods to avoid the curse of dimensionality of nonparametric regression.

In the next section a fully nonparametric estimator of the local average treatment effect in the presence of covariates X is proposed, which is $\sqrt{n}$-consistent, asymptotically normal and efficient. These results advocate the use of nonparametric regression in LATE estimation.

Before presenting the estimator, the identifying conditions of local average treatment effects are discussed and motivated. Consider first the case where the endogenous regressor $D \in \{0,1\}$ and the instrument $Z \in \{0,1\}$ are both binary. (Extensions are discussed in Section 3.2). D could be attending/not attending college and Z could be living close to or far from a college. The value of D might be influenced by the instrumental variable: $D_i = \zeta_i(Z_i)$, where ζ_i is unknown and might differ among individuals. To allow for a more compact notation, let $D_{i,Z_i} \equiv \zeta_i(Z_i)$ denote the observed value of D for unit i, and let $D_{i,z} \equiv \zeta_i(z)$ denote the potential value the endogenous regressor would take if Z_i were set exogenously to z.[5] According to the potential values of D the population can be partitioned into 4 subpopulations: Children with $D_{i,0} = D_{i,1} = 1$ will attend college irrespective of the distance to it. On the other hand, children with $D_{i,0} = D_{i,1} = 0$ will not attend college. Children with $D_{i,0} = 0$ and $D_{i,1} = 1$ go to college only if living close to it, whereas children with $D_{i,0} = 1$ and $D_{i,1} = 0$ attend college only if living far away from it. Thus each unit can be classified either as an always-taker, a

[5] This notation is very similar to Imbens and Angrist (1994) and Imbens (2001), with the exception that the arguments D and Z are indicated by super- and subscripts to avoid confusing the order of these arguments.

never-taker, a complier or a defier. Let τ_i denote the type of unit i:

Definition of types

$\tau_i = n$	if $D_{i,0} = 0$ and $D_{i,1} = 0$	Never-taker
$\tau_i = c$	if $D_{i,0} = 0$ and $D_{i,1} = 1$	Complier
$\tau_i = d$	if $D_{i,0} = 1$ and $D_{i,1} = 0$	Defier
$\tau_i = a$	if $D_{i,0} = 1$ and $D_{i,1} = 1$	Always-taker.

Since the units of type always-taker and of type never-taker cannot be induced do change D through a variation in the instrumental variable, the impact of D on Y can at most be ascertained for the subpopulations of compliers and defiers. Denote the observed outcome for unit i as $Y_{i,Z_i}^{D_i} \equiv Y_i = \varphi_i(D_i, Z_i)$ and let $Y_{i,z}^d \equiv \varphi_i(d, z)$ denote the potential outcomes. Y_{i,Z_i}^d is the outcome that would be observed if D_i were set exogenously to d, and $Y_{i,z}^d$ is the outcome observed if both D_i and Z_i were fixed externally. On the other hand, $Y_{i,z}^{D_i}$ is the outcome if only the instrument were set exogenously.

The conceptual difference between $Y_{i,z}^d$ and $Y_{i,z}^{D_i}$ is that in the former case both Z and D are fixed by external intervention, whereas in the latter case only Z is set exogenously and D_i is determined by the behaviour of unit i. In other words, the former potential outcomes isolate the direct effect of the instrument Z on Y, while the latter combine the direct effect and the indirect effect of Z on Y via the endogenous regressor D.

With a variety of assumptions, the average treatment effect on the subpopulation of compliers can be identified. The following exposition proceeds conditional on a (possibly empty) set of covariates X, since the instrumental variable assumptions may often be satisfied only conditional on confounding covariates. (When the set of covariates X is empty, the derivation corresponds to the unconditional identification.) Under the following assumptions, the average treatment effect on the subpopulation of compliers with characteristics X is identified, see Imbens (2001).

[Assumption 1: Exogenous covariates] The covariates X are exogenous in the sense that

$$X_{i,D_i,Z_i} = X_{i,d,z} \qquad \forall d, z,$$

where $X_{i,d,z}$ is the potential value of X that would be observed for unit i if D_i and Z_i were set by external intervention.

Assumption 1 precludes that X_i itself is caused by the instrument Z_i or the endogenous regressor D_i. In other words, the value of X_i would remain the same even if Z_i or D_i were manipulated externally.[6] Furthermore assume the following:

[Assumption 2: Monotonicity] The subpopulation of defiers has probability measure zero:

$$P(\tau = d) = 0.$$

[Assumption 3: Existence of compliers] The subpopulation of compliers has positive probability measure:

$$P(\tau = c) > 0.$$

[Assumption 4: Unconfounded type] The relative size of the subpopulations always-takers, never-takers and compliers is independent of the instrument:

$$P(\tau_i = t | X_i = x, Z_i = 0) = P(\tau_i = t | X_i = x, Z_i = 1) \qquad \text{for } t \in \{a, n, c\}.$$

[Assumption 5: Mean exclusion restriction] The potential outcomes are mean independent of the instrumental variable Z in each subpopulation:

$$\begin{aligned} E\left[Y^0_{i,Z_i} | X_i = x, Z_i = 0, \tau_i = t\right] &= E\left[Y^0_{i,Z_i} | X_i = x, Z_i = 1, \tau_i = t\right] \qquad \text{for } t \in \{n, c\} \\ E\left[Y^1_{i,Z_i} | X_i = x, Z_i = 0, \tau_i = t\right] &= E\left[Y^1_{i,Z_i} | X_i = x, Z_i = 1, \tau_i = t\right] \qquad \text{for } t \in \{a, c\}. \end{aligned}$$

[Assumption 6: Common support] The support of X is identical in both subpopulations:

$$\mathit{Supp}(X | Z = 1) = \mathit{Supp}(X | Z = 0).$$

An equivalent representation of the common support condition is that the conditional probability $\pi(x) = P(Z = 1 | X = x)$ is bounded away from 0 and 1 for all x with positive density: $0 < \pi(x) < 1 \; \forall x$ with $f_x(x) > 0$.

Assumptions 2 and 3 rule out the existence of subpopulations that are affected by the instrument in an opposite direction. Since changes in the instrument Z would trigger changes in D as well for the compliers as for the defiers, but with opposite sign, any causal effect on the compliers could be offset by opposite flows of defiers. Monotonicity ensures that the effect of Z on D has the same direction for all units. The monotonicity and the existence assumption together ensure that $D_{i,1} \geq D_{i,0}$ for all i and that the instrument has an effect on D, such

[6] Implicit in the discussion on estimating the effect of D on Y is the assumption that D_i is not caused by Y_i and that Z_i is neither caused by D_i nor by Y_i.

that $D_{i,1} > D_{i,0}$ for at least some units. When college proximity is used as an instrument to identify the returns to attending college, monotonicity requires that any child which would not have attended college if living close to a college, would also not have done so if living far from a college. The existence assumption requires that the college attendance decision depends, for at least some children, on the proximity to the nearest college.

The mean exclusion restriction (Assumption 5) rules out a direct effect of Z on Y. It combines two conceptually distinct assumptions: An exclusion restriction on the unit level and an unconfoundedness assumption on the population level. Rewrite Assumption 5 for the potential outcome Y_i^1 as

$$\begin{aligned} & E\left[Y_{i,0}^1 | X_i = x, Z_i = 0, \tau_i = t\right] \\ = \; & E\left[Y_{i,1}^1 | X_i = x, Z_i = 0, \tau_i = t\right] && \text{(Assumption 5a)} \\ = \; & E\left[Y_{i,1}^1 | X_i = x, Z_i = 1, \tau_i = t\right] && \text{(Assumption 5b)} \end{aligned}$$

for $t \in \{a, c\}$. The first equality sign (Assumption 5a) corresponds to an exclusion assumption on the *unit level.* It is assumed that the potential outcome Y_{i,Z_i}^1 for unit i is unaffected by an exogenous change in Z_i. For example, if $Y_{i,0}^1 = Y_{i,1}^1$, college proximity itself has no direct effect on the child's wages in its later career. Assumption 5a thus rules out any systematic impact of Z on the potential outcomes on a unit level. Assumption 5b, on the other hand, represents an unconfoundedness assumption on the *population level.* It assumes that the potential outcome $Y_{i,1}^1$ is identically distributed in the subpopulation of units for whom the instrument Z_i actually takes the value 0 and in the subpopulation of units with $Z_i = 1$. This assumption rules out selection effects that are related to the potential outcomes: The families who decided to reside close to a college should be identical in all characteristics (that affect their children's subsequent wages) to the families who decided to live far from a college. Thus, whereas Assumption 5b refers to the composition of units for whom $Z = 1$ or $Z = 0$ is observed, Assumption 5a refers to how the instrument affects the outcome Y of a particular unit.

Assumption 5b is trivially satisfied if the instrument Z is randomly assigned. Nevertheless randomization of Z does not guarantee that the exclusion assumption holds on the unit level (Assumption 5a). On the other hand, if Z is chosen by the unit itself, selection effects may often invalidate Assumption 5b: The families who decide to reside nearer to or farther from a college might be rather different, for example, if the districts were colleges are located also offer other job opportunities. In this case it is necessary to include in X all variables that affect the choice of residence Z as well as the potential outcomes Y^0_{i,Z_i} and Y^1_{i,Z_i}.

Assumption 4 allows to identify the effect of Z on D and to estimate the fraction of compliers. It requires that the fraction of always-takers, never-takers and compliers is independent of the instrument. Without conditioning on covariates X this assumption may often be invalidated because of selection effects, unless the instrument Z is randomly assigned. For example, parents who would like their children to attend college but could not avoid that their children might not want to go if living too far away, might decide to reside closer to a college. In this case, the subpopulation living close to a college would contain a higher fraction of compliers than those living far away. Validity of Assumption 4 requires that the vector X contains all variables that affect the (ex ante) choice of residence Z as well as the type τ (which is determined by $D_{i,0}$, $D_{i,1}$).

Thus in applications where the instrumental variables Z itself is endogenous, conditioning on a vector of confounding covariates X will be necessary to identify a local average treatment effect. Assumption 6 ensures that it is well defined for all x.

Under these assumptions, the size of the complier-subpopulation with characteristics x is identified as

$$P\left(\tau = c|X = x\right) = E\left[D|X = x, Z = 1\right] - E\left[D|X = x, Z = 0\right], \tag{3}$$

and the local average treatment effect $\gamma(x)$ on the compliers with characteristics x is

$$\gamma(x) = E\left[Y^1_{i,Z_i} - Y^0_{i,Z_i}\ |X_i = x, \tau_i = c\right] = \frac{E\left[Y|X = x, Z = 1\right] - E\left[Y|X = x, Z = 0\right]}{E\left[D|X = x, Z = 1\right] - E\left[D|X = x, Z = 0\right]}, \tag{4}$$

see Imbens and Angrist (1994) or Appendix A.1.[7]

If the instrument Z is not confounded with D or Y, the Assumptions 2 to 5 are valid without conditioning on any covariates X, and $\frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]}$ identifies the average treatment effect for the subpopulation of units who are induced to switch from $D_i = 0$ to 1 when the instrument changes from $Z_i = 0$ to 1. For example, if Z represents college proximity, the returns to attending college are estimated only for those children who go to college when living near a college and who would not attend college otherwise.

[7] For identification of $\gamma(x)$, Assumption 3 needs to hold conditional on X. For the identification of the local average treatment effect γ on *all* compliers (discussed below), $P\left(\tau = c\right) > 0$ suffices, because any x values with $P(\tau = c|X = x) = 0$ receive zero weight in the weighting function.

3 Nonparametric conditional LATE estimation

In many applications, however, the instrumental variable is endogenous and the Assumptions 2 to 5 are invalid without conditioning on confounding covariates X. In this case, (4) identifies only the treatment effect $\gamma(x)$ for the compliers with characteristics $X = x$. More interesting, however, would often be an estimate of the average treatment effect for the subpopulation of *all compliers*, which is the largest subpopulation for which a treatment effect is identified. To obtain the average treatment effect γ for all compliers, the conditional effects $\gamma(x)$ need to be weighted by the distribution of x in the all-compliers subpopulation:

$$\gamma = E\left[Y_{i,Z_i}^1 - Y_{i,Z_i}^0 \ |\tau_i = c\right] = \int \gamma(x) \cdot dF_{x|\tau=c},$$

where $F_{x|\tau=c}$ denotes the distribution function of X in the subpopulation of all compliers. By Bayes' theorem γ can be written as

$$\gamma = \int \gamma(x) \cdot \frac{P(\tau = c|X = x)}{P(\tau = c)} dF_x.$$

Inserting (4) and noting that the fraction of compliers corresponds to (3) gives

$$\gamma = \frac{\int \left(E\left[Y|X = x, Z = 1\right] - E\left[Y|X = x, Z = 0\right]\right) f_x(x)dx}{P(\tau = c)}.$$

With $P(\tau = c) = \int P(\tau = c|X = x)\, dF_x$ and using (3), the local average treatment effect on all compliers is

$$\gamma = \frac{\int \left(E\left[Y|X = x, Z = 1\right] - E\left[Y|X = x, Z = 0\right]\right) f_x(x)dx}{\int \left(E\left[D|X = x, Z = 1\right] - E\left[D|X = x, Z = 0\right]\right) f_x(x)dx}. \tag{5}$$

Hence by Assumption 6 the average treatment effect on all compliers is nonparametrically identified.[8]

Define the conditional mean functions $m_z(x) = E[Y|X = x, Z = z]$ and $\mu_z(x) = E[D|X = x, Z = z]$ and let $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ be corresponding nonparametric regression estimators thereof. A nonparametric imputation estimator of γ is

$$\hat{\gamma} = \frac{\sum_i \hat{m}_1(X_i) - \hat{m}_0(X_i)}{\sum_i \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)},$$

[8] Estimating population average parameters in IV models has also been suggested for instance by Das (2000) and others. Those estimators, adapted to the binary setting, are of the form $\int \gamma(x)dF_x = \int \frac{E[Y|X=x,Z=1]-E[Y|X=x,Z=0]}{E[D|X=x,Z=1]-E[D|X=x,Z=0]} f_x(x)dx$, which has no properly defined causal meaning.

where the expected values $E[Y|X,Z]$ and $E[D|X,Z]$ are imputed for each observation X_i.

Using the observed values Y_i and D_i as estimates of $E[Y_i|X_i, Z=z]$ and $E[D_i|X_i, Z=z]$ whenever $z = Z_i$, gives the conditional LATE estimator $\hat{\gamma}$

$$\hat{\gamma} = \frac{\sum\limits_{i:Z_i=1} (Y_i - \hat{m}_0(X_i)) - \sum\limits_{i:Z_i=0} (Y_i - \hat{m}_1(X_i))}{\sum\limits_{i:Z_i=1} (D_i - \hat{\mu}_0(X_i)) - \sum\limits_{i:Z_i=0} (D_i - \hat{\mu}_1(X_i))}. \tag{6}$$

The estimator $\hat{\gamma}$ corresponds to a ratio of two *matching estimators*, which are frequently used in treatment evaluation to estimate average treatment effects when the endogeneity of the regressor D can be completely controlled for by observed covariates (Angrist and Krueger 1999, Heckman, LaLonde, and Smith 1999).[9]

3.1 Properties of the conditional LATE estimator

In the following two theorems $\sqrt{n}$-asymptotic normality and efficiency of the estimator $\hat{\gamma}$ is shown. In Theorem 1 the asymptotic distribution of $\hat{\gamma}$ is derived and conditions on the preliminary estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ are given. In Theorem 2 the semiparametric efficiency bound for the estimation of the local average treatment effect γ is derived, which is identical to the asymptotic variance of the estimator $\hat{\gamma}$. Finally, it is shown that nonparametric kernel regression and local linear regression estimators satisfy the conditions of Theorem 1. Hence the local average treatment effect γ can be estimated without any functional form assumptions at the parametric rate.

Theorem 1 (Asymptotic normality of $\hat{\gamma}$) *Suppose that*
i) the local average treatment effect γ is identified,
ii) $\{(Y_i, D_i, Z_i, X_i)\}_{i=1}^n$ are iid with $Z_i \in \{0,1\}$ and $\lim\limits_{n\to\infty} \frac{n_1}{n_0} = \frac{P(Z=1)}{P(Z=0)}$, where $n_z = \sum\limits_{i=1}^n 1(Z_i = z)$
iii) the moments $E[Y|X,Z]$, $D[Y|X,Z]$, $Var[Y|X,Z]$, $Var[D|X,Z]$ and $Cov[Y,D|X,Z]$ exist for all $x \in Supp(X)$ and $z \in \{0,1\}$,
iv) the nonparametric regression estimator $\hat{m}_1$ of $m_1(x) = E[Y|X=x, Z=1]$ can be written

[9] An alternative estimator based on weighting by the probability $\pi(x) = P(Z=1|X=x)$ is

$$\hat{\gamma} = \frac{\sum \frac{Y_i Z_i}{\pi(X_i)} - \frac{Y_i(1-Z_i)}{1-\pi(X_i)}}{\sum \frac{D_i Z_i}{\pi(X_i)} - \frac{D_i(1-Z_i)}{1-\pi(X_i)}}.$$

This estimator is not further studied, because weighting estimators appear to have worse finite-sample properties than imputation estimators, see Frölich (2001).

in the asymptotically linear form

$$\hat{m}_1(x) - m_1(x) = \frac{1}{n_1} \sum_{j:Z_j=1} \xi_1^m(Y_j, X_j, x) + b_1^m(x) + R_1^m(x),$$

with the properties

iv A) $E\left[\xi_1^m(Y_j, X_j, X)\ |X = x, Z_j = 1\right] = 0$

iv B) $E\left[\xi_1^m(Y_j, X_j, X_i)^2 | Z_j = 1, Z_i = 0\right] = o(n)$

iv C) $\frac{1}{\sqrt{n_0}} \sum\limits_{i:Z_i=0} b_1^m(X_i) = o_p(1)$

iv D) $\frac{1}{\sqrt{n_0}} \sum\limits_{i:Z_i=0} R_1^m(X_i) = o_p(1)$

iv E)

$$E\left[\xi_1^m(Y_j, X_j, X_i) | Y_j, X_j, Z_j = 1, Z_i = 0\right] = (Y_j - m_1(X_j)) \frac{f_{x|z=0}(X_j)}{f_{x|z=1}(X_j)} + o_p(1),$$

and analogously for $\hat{m}_0$, $\hat{\mu}_1$, $\hat{\mu}_0$. *Then the estimator (6) of the local average treatment effect* γ *is asymptotically normal distributed*

$$\sqrt{n}\,(\hat{\gamma} - \gamma) \to N\left(0, \mathcal{V}\right) \tag{7}$$

with asymptotic variance

$$\mathcal{V} = \frac{1}{\Gamma^2} E\left[\frac{\sigma_{Y_1}^2(X) - 2\gamma\sigma_{Y_1D_1}^2(X) + \gamma^2\sigma_{D_1}^2(X)}{\pi(X)} + \frac{\sigma_{Y_0}^2(X) - 2\gamma\sigma_{Y_0D_0}^2(X) + \gamma^2\sigma_{D_0}^2(X)}{1 - \pi(X)}\right] + \frac{1}{\Gamma^2} E\left[\left(m_1(X) - m_0(X) - \gamma\mu_1(X) + \gamma\mu_0(X)\right)^2\right]$$

where $\Gamma = \int \left(\mu_1(x) - \mu_0(x)\right) f_x(x)dx$ *is the denominator in (5),* $\pi(x) = P\left(Z = 1|X = x\right)$ *is the probability that* Z *takes the value one given* X, *and* $\sigma_{Y_z}^2(x)$, $\sigma_{D_z}^2(x)$ *and* $\sigma_{Y_zD_z}^2(x)$ *are the conditional variances and covariances in the* $Z = z$ *subpopulation:* $\sigma_{Y_1}^2(X) = Var\left[Y|X = x, Z = 1\right]$, *and* $\sigma_{Y_1D_1}^2(x) = Cov\left[Y, D|X = x, Z = 1\right]$ *etc. (Proof in Appendix A.2.)*

Condition (i) requires that the local average treatment effect γ is identified by the Assumptions 1 to 6 discussed previously. Condition (ii) supposes random sampling. This condition readily could be relaxed to allow for stratified sampling on X and/or Z,[10] as long as the population density function $f_x(x)$ can be recovered, for example through known sampling weights. Condition (iii) requires the existence of the first two conditional moments of Y and D.

[10] For example, iid sampling within $(Y, D, X)|Z = 1$ and $(Y, D, X)|Z = 0$.

Condition (iv) gives conditions on the nonparametric estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$. These conditions are satisfied by kernel and local linear regression as discussed below. $\xi_z^m(Y_j, X_j, x)$ is the mean-zero local influence function of $\hat{m}_z(x)$, which captures its variance. $b_z^m(x)$ is the local bias of $\hat{m}_z(x)$, and $R_z^m(x)$ is a residual term. Condition (iv B) requires that the variance of ξ_z^m does not grow too fast. Condition (iv C) constrains the local bias term to be of order $o_p(n^{-\frac{1}{2}})$. This assumption could be relaxed to permit a local bias term of order $O_p(n^{-\frac{1}{2}})$, but this would introduce an asymptotic bias term in (7). Analogously, the residual term could be permitted to be of order $O_p(n^{-\frac{1}{2}})$ in Condition (iv D). The condition (iv E) is not necessary for $\sqrt{n}$-asymptotic normality, but is imposed to characterize the variance expression $\mathcal{V}$.

In the following theorem the semiparametric variance bound for the estimation of the local average treatment effect γ is derived. As this variance bound is equal to the asymptotic variance $\mathcal{V}$ of $\hat{\gamma}$ under the conditions of Theorem 1, the conditional LATE estimator $\hat{\gamma}$ is efficient.

Theorem 2 (Efficiency of $\hat{\gamma}$) *The semiparametric variance bound of γ is $\mathcal{V}$. (Proof in Appendix A.3).*

It remains to find nonparametric estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ that satisfy condition (iv) of Theorem 1. Heckman, Ichimura, and Todd (1998, Theorem 3) analyzed the local polynomial kernel regression estimator and showed its asymptotic linearity. Under the conditions that
A1) $m_1(x)$ is $\bar{p}$-times continuously differentiable and its $\bar{p}$-th derivative is Hölder continuous,[11] where $\bar{p} > k$ and k is the number of continuous regressors in X,
A2) the bandwidth sequence h_{n_1} satisfies $n_1 h_{n_1}^k / \ln n_1 \to \infty$ and $n_1 h_{n_1}^{2\bar{p}} \to 0$,
A3) the Kernel function K is symmetric, compact and Lipschitz continuous,[12]
the local polynomial regression estimator $\hat{m}_1$ of polynomial order $\bar{p}$ is asymptotically linear

$$\hat{m}_1(x) - m_1(x) = \frac{1}{n_1} \sum_{j:Z_j=1} \xi_1^m(Y_j, X_j, x) + b_1^m(x) + R_1^m(x), \tag{8}$$

with $b_1^m(x) = o\left(h_{n_1}^{\bar{p}}\right)$ and $\frac{1}{\sqrt{n_1}} \sum\limits_{j:Z_j=1} R_1^m(x) = o_p(1)$.

Asymptotic linearity of local polynomial estimators with polynomial order $p < \bar{p}$ requires additionally that

[11] Hölder continuity of a function $\zeta(x)$ at x_0 means that there exist $\alpha \in (0,1]$ and $C > 0$ such that $|\zeta(x) - \zeta(x_0)| \leq C \cdot \|x - x_0\|^\alpha$ for all x.

[12] Lipschitz continuous means Hölder continuous of order $\alpha = 1$.

A4) the Kernel function K has moments of order 1 through $\bar{p}-1$ equal to zero,[13]

A5) the density f_x of X is $\bar{p}$-times continuously differentiable with its $\bar{p}$-th derivative Hölder continuous,

A6) $m_1(x)$ is estimated at an interior point of the support $Supp(X|Z=1)$.[14]

Under conditions A1 to A6 also the local polynomial estimator with $p<\bar{p}$ is asymptotically linear (8) with $b_1^m(x)=O\left(h_{n_1}^{\bar{p}}\right)$ and $\frac{1}{\sqrt{n_1}}\sum\limits_{j:Z_j=1} R_1^m(x)=o_p(1)$, and satisfies the conditions (iv A), (iv C) and (iv D) of Theorem 1.

It remains to verify conditions (iv B) and (iv E). The influence function for the *Nadaraya-Watson kernel* estimator and for the *local linear estimator* with product kernel is

$$\xi_1^m(Y_j,X_j,x)=(Y_j-m_1(X_j))\cdot\frac{K\left(\frac{X_j-x}{h_{n_1}}\right)}{h_{n_1}^k f_{x|z=1}(x)\lambda}+o_p(1),$$

where $\lambda=\int K(u)du$ (Heckman, Ichimura, and Todd 1998). Verify first condition (iv E):

$$\begin{aligned}
&E\left[\xi_1^m(Y_j,X_j,X_i)|Y_j,X_j,Z_j=1,Z_i=0\right]\\
&=(Y_j-m_1(X_j))\cdot E\left[\frac{K\left(\frac{X_j-X_i}{h_{n_1}}\right)}{h_{n_1}^k f_{x|z=1}(X_i)\lambda}\,|X_j,Z_j=1,Z_i=0\right]+o(1)\\
&=\frac{(Y_j-m_1(X_j))}{\lambda}\cdot\int\frac{K\left(\frac{X_j-X_i}{h_{n_1}}\right)}{h_{n_1}^k f_{x|z=1}(X_i)}f_{x|z=0}(X_i)\cdot dX_i+o(1)\\
&=\frac{(Y_j-m_1(X_j))}{\lambda}\cdot\int\frac{K(u)}{h_{n_1}^k}\frac{f_{x|z=0}(X_j-uh_{n_1})}{f_{x|z=1}(X_j-uh_{n_1})}\cdot h_{n_1}^k du+o(1)
\end{aligned}$$

with the change in variables $(X_j-X_i)/h_{n_1}=u$. Since the densities $f_{x|z=0}$, $f_{x|z=1}$ are continuously differentiable, $f_{x|z=0}(X_j-uh_{n_1})=f_{x|z=0}(X_j)+O(h_{n_1})$ by a Taylor series expansion and thus

$$\begin{aligned}
&= \frac{(Y_j-m_1(X_j))}{\lambda}\cdot\int\frac{f_{x|z=0}(X_j)+O(h)}{f_{x|z=1}(X_j)+O(h)}K(u)\,du+o(1)\\
&= \frac{(Y_j-m_1(X_j))}{\lambda}\cdot\int\frac{f_{x|z=0}(X_j)}{f_{x|z=1}(X_j)}K(u)\,du+o(1)\\
&= (Y_j-m_1(X_j))\frac{f_{x|z=0}(X_j)}{f_{x|z=1}(X_j)}+o(1).
\end{aligned}$$

[13] Conditions A3 and A4 are satisfied by most compact higher-order kernels. For a discussion on higher-order kernels see, for instance, Pagan and Ullah (1999).

[14] This requires a trimming function to trim observations in the boundary.

Condition (iv B) is a rather weak condition and can be verified using a similar argument:

$$
\begin{aligned}
& E\left[(Y_j - m_1(X_j))^2 \cdot \frac{\frac{1}{h_{n_1}^{2k}} K\left(\frac{X_j - X_i}{h_{n_1}}\right)^2}{f_{x|z=1}^2(X_i)\lambda^2} \, |Z_j = 1, Z_i = 0\right] + o(1) \\
= & \frac{1}{\lambda^2} E\left[\frac{\sigma_{Y_1}^2(X_j)}{f_{x|z=1}^2(X_i)} \cdot \frac{1}{h_{n_1}^{2k}} K\left(\frac{X_j - X_i}{h_{n_1}}\right)^2 \, |Z_j = 1, Z_i = 0\right] + o(1) \\
= & \frac{1}{\lambda^2} \int\int \frac{\sigma_{Y_1}^2(X_j)}{f_{x|z=1}^2(X_i)} \cdot \frac{1}{h_{n_1}^{2k}} K\left(\frac{X_j - X_i}{h_{n_1}}\right)^2 f_{x|z=1}(X_j) f_{x|z=0}(X_i) \, dX_j dX_i + o(1) \\
= & \frac{1}{h_{n_1}^k} \frac{1}{\lambda^2} \int\int \sigma_{Y_1}^2(X_j) K(u)^2 \frac{f_{x|z=1}(X_j) f_{x|z=0}(X_j - uh_{n_1})}{f_{x|z=1}^2(X_j - uh_{n_1})} dX_j du + o(1) \\
= & o(1),
\end{aligned}
$$

with the change in variables $(X_j - X_i)/h_{n_1} = u$. Hence the conditional LATE estimator $\hat{\gamma}$ with a Nadaraya-Watson or local linear regression estimator (with product kernel) of the conditional expectation functions $m_z(x)$ and $\mu_z(x)$ is $\sqrt{n}$-asymptotically normal and efficient.

3.2 Extensions

The conditional LATE estimator $\hat{\gamma}$ corresponds to a ratio of two *matching* estimators, which have been thoroughly studied in the treatment evaluation literature, see e.g. Rubin (1974), Heckman, Ichimura, and Todd (1998) or Frölich (2001). As is well known for matching estimators, adjusting for the distribution of the X characteristics is equivalent to adjusting for the distribution of the propensity score $\pi(x) = P(Z = 1|X = x)$. This is the balancing score property of the propensity score (Rosenbaum and Rubin 1983, Imbens 2000, Lechner 2001), which implies that

$$
\begin{aligned}
\gamma &= \frac{\int (E[Y|X = x, Z = 1] - E[Y|X = x, Z = 0]) f_x(x)dx}{\int (E[D|X = x, Z = 1] - E[D|X = x, Z = 0]) f_x(x)dx} \\
&= \frac{\int (E[Y|\pi(X) = \rho, Z = 1] - E[Y|\pi(X) = \rho, Z = 0]) \cdot f_{\pi(x)}(\rho)d\rho}{\int (E[D|\pi(X) = \rho, Z = 1] - E[D|\pi(X) = \rho, Z = 0]) \cdot f_{\pi(x)}(\rho)d\rho},
\end{aligned}
$$

where $f_{\pi(x)}$ is the density function of $\pi(x)$ in the population. An advantage of the latter representation is that the conditional expectation functions $E[Y|\pi(X) = \rho, Z]$ and $E[D|\pi(X) = \rho, Z]$ depend only on the *one-dimensional* propensity score $\pi(x)$ and no longer on the full set of covariates. Hence these conditional expectation functions can be estimated more precisely than $E[Y|X, Z]$ and $E[D|X, Z]$, particularly if X is high-dimensional.

Replacing $\hat{m}_z(X_i)$ and $\hat{\mu}_z(X_i)$ in (6) by corresponding estimates of $E[Y|\pi_i, Z = z]$ and $E[D|\pi_i, Z = z]$, where $\pi_i = \pi\left(X_i\right)$, gives the *propensity-score-matching* LATE estimator $\hat{\gamma}_\pi$.

However, in most applications the propensity score $\pi\left(x\right)$ is unknown and needs to be estimated. The propensity score LATE estimator $\hat{\gamma}_\pi$ is then computed with the estimated propensity scores $\hat{\pi}_i$. In this case the estimator $\hat{\gamma}_\pi$ is still $\sqrt{n}$-asymptotically normal, provided the propensity scores were consistently estimated by parametric or by local polynomial regression. This follows from the results in Heckman, Ichimura, and Todd (1998), which extend to estimated propensity scores and show that local polynomial regression on the estimated propensity score is still asymptotically linear.

Another result in Heckman, Ichimura, and Todd (1998) covers also the situation where the common support condition (Assumption 6) is not satisfied and the supports $Supp\left(X|Z = 1\right)$ and $Supp\left(X|Z = 0\right)$ are unknown. Then the local average treatment effect is only identified with respect to the complier subpopulation with characteristics x belonging to the common support. Since the common support is unknown, Heckman, Ichimura, and Todd (1998) introduce an estimator of it and give conditions under which the local polynomial regression estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ are asymptotically linear (with trimming). If these conditions are satisfied, the nonparametric LATE estimators remain $\sqrt{n}$-asymptotically normal.

Finally, consider situations where the instrumental variable Z and/or the endogenous regressor D are non-binary. Below it will be seen that the conditional LATE estimator $\hat{\gamma}$ is also applicable for estimating complier average treatment effects when D is discrete (with bounded support) and Z is non-binary or vector-valued. First of all, pair-wise comparisons for any two different values z, z' of Z can always be conducted, to estimate the average treatment effect on the subpopulation which is induced to change D by a change in the instrument from z to z'. In the case where Z is continuous, estimates of $\lim_{z \to z'} E[Y^1 - Y^0\ |D_{i,z} = 0, D_{i,z'} = 1]$ give the marginal treatment effect on the subpopulation which is just about to change D (Heckman and Vytlacil 2001). Estimating a variety of pair-wise local average treatment effect provides some indications about treatment effect heterogeneity in the population. However, instead of a multitude of pair-wise effects, one would often prefer to estimate the (aggregate) average treatment effect for the *largest* subpopulation for which an effect can be identified.

Local average treatment effect with non-binary instrument Z

Consider first the situation where D is binary and Z a scalar, non-binary variable. If Z is discrete (and the Assumptions 2 to 5 hold without conditioning on X), Imbens and Angrist (1994) have shown that the conventional linear instrumental variables estimator can be written as a weighted average of pair-wise local average treatment effects. However, as pointed out in Heckman and Vytlacil (2001), the weighting implicit in the linear instrumental variables estimator does not correspond to a well-defined causal parameter. Particularly, the linear IV estimator does not estimate the average treatment effect on the compliers. Instead the appropriate estimator that estimates the average treatment effect in the subpopulation of *all* compliers (with characteristics X) is

$$\frac{E\left[Y|X, Z=z_{\max}\right]-E\left[Y|X, Z=z_{\min}\right]}{E\left[D|X, Z=z_{\max}\right]-E\left[D|X, Z=z_{\min}\right]}$$

when the support of Z is $Supp(Z)=(z_{\min}, z_{\max})$. This corresponds to (4), where the endpoints of the support of Z are used as the binary instrument. To illustrate the argument, consider the case were Z is discrete and can take the values $Z\in\{0,1,2\}$. (A full elaboration for continuous Z is given in Heckman and Vytlacil (2001)). The monotonicity assumption now requires that $D_{i,2}\geq D_{i,1}\geq D_{i,0}$, and according to their reaction on Z the population can be partitioned into 4 types (excluding the defiers, which have probability measure zero):

Definition of types

$\tau_i = n$	if $D_{i,0}=0$ and $D_{i,1}=0$ and $D_{i,2}=0$	Never-taker
$\tau_i = c_{12}$	if $D_{i,0}=0$ and $D_{i,1}=0$ and $D_{i,2}=1$	Complier at 1-2
$\tau_i = c_{01}$	if $D_{i,0}=0$ and $D_{i,1}=1$ and $D_{i,2}=1$	Complier at 0-1
$\tau_i = a$	if $D_{i,0}=1$ and $D_{i,1}=1$ and $D_{i,2}=1$	Always-taker.

The subpopulations of never-takers and always-takers do not react on changes in the instrumental variable and hence the effect cannot be identified for these subpopulations. The subpopulation of compliers consists now of two groups: Those units who change D_i from 0 to 1 when the instrument Z_i is changed from 0 to 1 and those units who react when the instrument Z_i is changed from 1 to 2. The average treatment effect on the first group of compliers (conditional on X) is identified as

$$E\left[Y^1-Y^0|X,\tau=c_{01}\right]=\frac{E\left[Y|X,Z=1\right]-E\left[Y|X,Z=0\right]}{E\left[D|X,Z=1\right]-E\left[D|X,Z=0\right]} \tag{9}$$

and the treatment effect on the second group of compliers is

$$E\left[Y^1 - Y^0|X, \tau = c_{12}\right] = \frac{E\left[Y|X, Z=2\right] - E\left[Y|X, Z=1\right]}{E\left[D|X, Z=2\right] - E\left[D|X, Z=1\right]}. \tag{10}$$

To obtain the average treatment effect on both complier groups, the average effect on the first group and on the second group need to be weighted by their relative sizes:

$$\begin{aligned} E\left[Y^1 - Y^0|X, \tau = c_{01} \text{ or } c_{12}\right] &= E\left[Y^1 - Y^0|X, \tau = c_{01}\right] \cdot P\left(\tau = c_{01} \; |X, \tau = c_{01} \text{ or } c_{12}\right) \\ &\quad + E\left[Y^1 - Y^0|X, \tau = c_{12}\right] \cdot P\left(\tau = c_{12} \; |X, \tau = c_{01} \text{ or } c_{12}\right). \end{aligned}$$

Noting that $P(\tau = c_{01} \; |X, \tau = c_{01} \text{ or } c_{12}) = P(\tau = c_{01} \; |X) / \left(P(\tau = c_{01}|X) + P(\tau = c_{12}|X)\right)$ and that $E[D|X, Z=1] - E[D|X, Z=0] = P(\tau = c_{01}|X)$ and $E[D|X, Z=2] - E[D|X, Z=1] = P(\tau = c_{12}|X)$ and $E[D|X, Z=2] - E[D|X, Z=0] = P(\tau = c_{01}|X) + P(\tau = c_{12}|X)$ it follows that

$$\begin{aligned} E\left[Y^1 - Y^0|X, \tau = c_{01} \text{ or } c_{12}\right] &= E\left[Y^1 - Y^0|X, \tau = c_{01}\right] \cdot \frac{E\left[D|X, Z=1\right] - E\left[D|X, Z=0\right]}{E\left[D|X, Z=2\right] - E\left[D|X, Z=0\right]} \\ &\quad + E\left[Y^1 - Y^0|X, \tau = c_{12}\right] \cdot \frac{E\left[D|X, Z=2\right] - E\left[D|X, Z=1\right]}{E\left[D|X, Z=2\right] - E\left[D|X, Z=0\right]} \\ &= \frac{E\left[Y|X, Z=2\right] - E\left[Y|X, Z=0\right]}{E\left[D|X, Z=2\right] - E\left[D|X, Z=0\right]}, \end{aligned}$$

after inserting (9) and (10). Hence the average treatment effect on both complier groups is identified by a binary instrumental variable estimator of the type (4), with the binary instrument corresponding to the endpoints of the support of Z. An analogous reasoning as in (5) leads to

$$\gamma = \frac{\int \left(E\left[Y|X=x, Z=z_{\max}\right] - E\left[Y|X=x, Z=z_{\min}\right]\right) f_x(x) dx}{\int \left(E\left[D|X=x, Z=z_{\max}\right] - E\left[D|X=x, Z=z_{\min}\right]\right) f_x(x) dx}, \tag{11}$$

which is the average treatment effect on the *largest* subpopulation for which an effect can be identified.[15]

[15] A bias-variance trade-off in the estimation of the local average treatment effect with non-binary Z becomes visible from (11). Although (11) incorporates the proper weighting of the different complier subgroups and leads to an unbiased estimator of γ, only observations with Z_i equal (or close) to $z_{\min}$ or $z_{\max}$ are used for estimation. Observations with Z_i between the endpoints $z_{\min}$ and $z_{\max}$ are neglected, which might lead to a large variance. Variance could be reduced, at the expense of a larger bias, by weighting the complier subgroups differently or by choosing larger bandwidth values for the estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$. This is beyond the scope of this paper and will be analyzed in future research.

A similar reasoning applies when the instrumental variable Z is vector valued and the monotonicity condition holds with respect to all components of Z. Angrist and Imbens (1995) have shown that the linear instrumental variables estimator with multiple instruments can be written as a weighted average of pair-wise binary local average treatment effect estimators. However, again, the weights are not derived from the definition of a meaningful causal parameter, as argued in Heckman and Vytlacil (2001). Since the different instrumental variables act trough their effect on D, a convenient way to summarize the different components of Z is to consider the probability $p(z,x) = P(D=1|X=x, Z=z)$ that D takes the value 1 given X and Z. The average treatment effect for the largest subpopulation (with characteristics x) for which a treatment effect can be identified is the subpopulation of all units for whom $D_{i,z}$ is a non-trivial function of z (denoted by type $\tau = c$). The average treatment effect for this all-compliers subpopulation with characteristics x is

$$E\left[Y^1 - Y^0|X, \tau = c\right] = \frac{E\left[Y|X, Z = z_x^u\right] - E\left[Y|X, Z = z_x^l\right]}{E\left[D|X, Z = z_x^u\right] - E\left[D|X, Z = z_x^l\right]},$$

where $z_x^l = \min\limits_z p(z,x)$ corresponds to the value of Z where the probability that $D=1$ is lowest (i.e. where only the always-takers have $D=1$) and $z_x^u = \max\limits_z p(z,x)$ corresponds to the value of Z where only the never-takers have $D=0$, see Heckman and Vytlacil (2001). An analogous reasoning as in (5) and (11) gives then the average treatment effect on the subpopulation of all compliers as

$$\gamma = \frac{\int \left(E\left[Y|X=x, Z = z_x^u\right] - E\left[Y|X=x, Z = z_x^l\right]\right) f_x(x)dx}{\int \left(E\left[D|X=x, Z = z_x^u\right] - E\left[D|X=x, Z = z_x^l\right]\right) f_x(x)dx}.$$

Local average treatment effect with non-binary regressor D

For the situation where D is discrete and takes on more than two distinct values (and Z is binary), Angrist and Imbens (1995) have shown that the estimator corresponding to the right-hand side of (4) identifies a weighted average causal effect for the subpopulation of compliers. With D taking many different values, the *compliance intensity* can differ among units. Some units might be induced to change from $D_i = d$ to $D_i = d+1$ as a reaction on changing Z_i from 0 to 1. Other units might change, for example, from $D_i = d'$ to $D_i = d'+2$. Suppose D is years of schooling and Z a valid instrument that influences the schooling decision (for example the quarter-of-birth instrument in Angrist and Krueger (1991)). If Z were changed exogenously, some units might respond by increasing school attendance by an additional year. However, some units might increase school attendance even by two or three years. Furthermore, even if Z were set to zero for all units, they would attend different years of schooling. Hence a change in Z induces a variety of different reactions in D, which cannot be disentangled. Consequently only a weighted average of these effects can be identified. Suppose D is discrete and takes values in $D \in \{0, .., K\}$. According to their reaction on a change in Z from 0 to 1, the population can be partitioned into the types:

$$\tau_i = c_{k,l} \qquad \text{if } D_{i,0} = k \text{ and } D_{i,1} = l. \tag{12}$$

Assuming monotonicity, the defier-types $c_{k,l}$ for $k > l$ do not exist. The types $c_{k,k}$ represent those units that do not react on a change in Z (these are the always-takers and the never-takers in the setup where D is binary). The types $c_{k,l}$ for $k < l$ are the compliers, which comply by increasing D_i from k to l. These compliers comply at different base levels k and with different intensities $l - k$. Define the weighted average treatment effect $\gamma_w(x)$ for the compliers with characteristics x as[16]

$$\gamma_w(X) = \frac{\sum_{k}^{K}\sum_{l>k}^{K} E\left[Y^l - Y^k | X, \tau = c_{k,l}\right] \cdot P\left(\tau = c_{k,l}|X\right)}{\sum_{k}^{K}\sum_{l>k}^{K} (l-k) \cdot P\left(\tau = c_{k,l}|X\right)}. \tag{13}$$

$\gamma_w(x)$ is the effect of the induced treatment change divided by the intensity of compliance, averaged over the different complier groups $c_{k,l}$. In the returns to schooling example, $\gamma_w(x)$ is the expected return to one additional year of schooling. It is defined as the average of the return to one additional year of schooling (divided by one), for those who continue schooling by one year, and the return to two additional years of schooling (divided by two), for those who extend schooling by two years, and the return to three additional years of schooling (divided by three), for those who prolong schooling by three years etc. Thus, the effect of the increase in schooling is divided by the *intensity of compliance* (i.e. how many additional years of schooling) for each complying unit, to obtain the average effect of *one* additional year of schooling, for units with characteristics x. The denominator of $\gamma_w(x)$ thus represents the number of *intensity-weighted compliers.*

[16] The presentation in Angrist and Imbens (1995) looks different from the definition of γ_w used here, as they present the effect in terms of overlapping subpopulations. Nevertheless, both definitions are equivalent.

When the unconfoundedness and the exclusion restrictions (Assumptions 4 and 5) are extended to hold for all types defined in (12), $\gamma_w(x)$ is identified as[17]

$$\gamma_w(X) = \frac{E\left[Y|X, Z=1\right] - E\left[Y|X, Z=0\right]}{E\left[D|X, Z=1\right] - E\left[D|X, Z=0\right]}. \tag{14}$$

To obtain the weighted average effect for the subpopulation of all compliers (i.e. all subpopulations $c_{k,l}$ for $k < l$), one would need to weight $\gamma_w(x)$ by the distribution of X in the complier subpopulation:

$$\int \gamma_w(x) \cdot dF_{x|complier},$$

where $F_{x|complier}$ is the distribution of X in the all-compliers subpopulation.

Unfortunately, the distribution of X in the all-compliers subpopulation is not identified if D takes more than 2 different values. In particular, the size of the all-compliers subpopulation is no longer identified by the distribution of D and Z. Consider the following example: For D taking values in $\{0, 1, 2\}$, the population can be partitioned in the subpopulations: $\{c_{00}$, c_{01}, c_{02}, c_{11}, c_{12}, $c_{22}\}$ with the all-compliers subpopulation consisting of $\{c_{01}$, c_{02}, $c_{12}\}$. The two partitions $\{c_{00}$, c_{01}, c_{02}, c_{11}, c_{12}, $c_{22}\}$ = $\{0.1$, 0.1, 0.3, 0.3, 0.1, $0.1\}$ and $\{0.1$, 0.2, 0.2, 0.2, 0.2, $0.1\}$ generate the same distribution of D given Z: $P(D=0|Z=0) = 0.5$, $P(D=1|Z=0) = 0.4$, $P(D=2|Z=0) = 0.1$, $P(D=0|Z=1) = 0.1$, $P(D=1|Z=1) = 0.4$, $P(D=2|Z=1) = 0.5$. However, the size of the all-compliers subpopulation is different for the two partitions (0.5 and 0.6, respectively). Hence the size of the all-compliers subpopulation is not identified from the observable variables, see also Imbens and Rubin (1997).

Nevertheless, if one defines the all-compliers subpopulation in terms of *compliance intensity units*, the distribution of X in this complier subpopulation is identified. In the intensity-weighted complier subpopulation, each complier is weighted by its compliance intensity. In the case where $D \in \{0, 1, 2\}$, the subpopulation $c_{0,2}$ receives twice the weight of the subpopulation $c_{0,1}$. In the years-of-schooling example, the subpopulation $c_{0,2}$ complies with 2 additional years of schooling. If the returns to a year of schooling are the same for each year of schooling, a unit which complies with 2 additional years can be thought of as an observation that measures

[17] The proof is immediate, noting that the population is partitioned by $\tau = c_{k,l}$ for $k \leq l$. Hence $\sum\limits_{k\leq l} P\left(\tau = c_{k,l}|X\right) = 1$ and $E[Y|X, Z=1] = \sum\limits_{k\leq l} E[Y|X, Z=1, \tau = c_{k,l}] P\left(\tau = c_{k,l}|X, Z=1\right) = \sum\limits_{k\leq l} E[Y^l|X, \tau = c_{k,l}] P\left(\tau = c_{k,l}|X\right)$ by the exclusion and the unconfoundedness assumption. Analogously, $E[Y|X, Z=0] = \sum\limits_{k\leq l} E[Y^k|X, \tau = c_{k,l}] P\left(\tau = c_{k,l}|X\right)$ and $E[D|X, Z=1] = \sum\limits_{k\leq l} l \cdot P\left(\tau = c_{k,l}|X\right)$ and $E[D|X, Z=0] = \sum\limits_{k\leq l} k \cdot P\left(\tau = c_{k,l}|X\right)$.

twice the effect of one additional year of schooling. Or, as two (correlated) measurements of the return to a year of schooling. Unless these two measurements are perfectly correlated, the unit which complies with 2 additional years contributes more to the estimation of the return to schooling than a unit which complies with one additional year. Consequently, the units that comply with more than one year should receive a higher weight when averaging the return to schooling over the distribution of X. If each unit is weighted by its number of additional years, the weighted distribution function of X in the all-compliers subpopulation, in the case where $D \in \{0,1,2\}$, is

$$f^w_{x|complier} = \frac{f_{x|\tau=c_{0,1}} P(\tau = c_{0,1}) + f_{x|\tau=c_{1,2}} P(\tau = c_{1,2}) + 2 f_{x|\tau=c_{0,2}} P(\tau = c_{0,2})}{P(\tau = c_{0,1}) + P(\tau = c_{1,2}) + 2P(\tau = c_{0,2})}$$

or in the general case

$$f^w_{x|complier} = \frac{\sum\limits_k^K \sum\limits_{l>k}^K f_{x|\tau=c_{k,l}} P(\tau = c_{k,l}) \cdot (l-k)}{\sum\limits_k^K \sum\limits_{l>k}^K P(\tau = c_{k,l}) \cdot (l-k)}. \tag{15}$$

Using Bayes' theorem $f_{x|\tau=c_{k,l}} = P(\tau = c_{k,l}|X)\, f_x / P(\tau = c_{k,l})$, the weighted distribution function of X in the all-compliers subpopulation is

$$f^w_{x|complier} = \frac{\sum\limits_k^K \sum\limits_{l>k}^K P(\tau = c_{k,l}|X) \cdot (l-k)}{\sum\limits_k^K \sum\limits_{l>k}^K P(\tau = c_{k,l}) \cdot (l-k)} \cdot f_x, \tag{16}$$

where $f_x(x)$ is the density function of X in the full population. With this weighted distribution function the treatment effect in the subpopulation of all compliers is identified by (13) and (14):

$$\begin{aligned}
\gamma_w &= \int \gamma_w(x) \cdot f^w_{x|complier}(x) dx \\
&= \int \frac{E[Y|X=x, Z=1] - E[Y|X=x, Z=0]}{\sum\limits_k \sum\limits_{l>k} P(\tau = c_{k,l}) \cdot (l-k)} \cdot f_x(x) dx \\
&= \frac{\int E[Y|X=x, Z=1] - E[Y|X=x, Z=0] \cdot f_x(x) dx}{\int \sum\limits_k \sum\limits_{l>k} P(\tau = c_{k,l}|X=x) \cdot (l-k) \cdot f_x(x) dx} \\
&= \frac{\int (E[Y|X=x, Z=1] - E[Y|X=x, Z=0]) \cdot f_x(x) dx}{\int (E[D|X=x, Z=1] - E[D|X=x, Z=0]) \cdot f_x(x) dx},
\end{aligned}$$

which is identical to expression (5). Hence the conditional LATE estimator (6) is also applicable when D is a discrete random variable taking more than 2 different values.

4 Conclusions

In this paper nonparametric instrumental variables estimation of local average treatment effects has been extended to accommodate confounding covariates X, which is necessary whenever the instrumental variable Z itself is endogenous. A nonparametric conditional LATE estimator has been proposed and its asymptotic properties have been derived.

Identification of properties of the relationship between an endogenous regressor D and an outcome variable Y based on instrumental variables is appealing, since it allows for (almost) arbitrary unobserved heterogeneity in the population. Although the average treatment effect on the subpopulation of compliers is usually not the primary causal parameter of interest, it is often the only causal effect that is identified. Alternative nonparametric instrumental variable regression models impose assumptions (conditional constant treatment effect, instrument moves regressor over entire support) that are often not credible in many applications. Then only treatment effects for subpopulations that react on changes of the instrument can be identified.

Identification and estimation of local average treatment effects is often discussed without covariates. However, when the instrument is not randomly assigned but itself endogenous, the instrumental variable assumptions are only valid conditional on a vector of confounding covariates X. Usually covariates X have been included via parametric modelling. The proposed conditional LATE estimator, in contrast, incorporates covariates X in a fully *nonparametric* way. This estimator corresponds to a ratio of two matching estimators and it is $\sqrt{n}$-consistent, asymptotically normal and efficient. In addition, a propensity score matching LATE estimator has been presented, which might perform better in small samples.

Extensions to cases where the endogenous regressor or the instrumental variable are non-binary have also been considered. The proposed conditional LATE estimator remains applicable even in these situations (although its interpretation may become more involved). These results support the use of nonparametric regression methods in the estimation of local average treatment effects.

A Appendix

A.1 Derivation of the local average treatment effect $\gamma(x)$

It is shown that the Assumptions 1 to 5 identify the local average treatment effect $\gamma(x)$ according to (4). The derivation is similar to Imbens and Angrist (1994), conditional on X. Using the exogeneity of X (Assumption 1) and the partitioning of the population into the subpopulations always-takers, never-takers, compliers and defiers, the expected value of Y given X and Z can be written as

$$\begin{aligned}
E\left[Y_i|X_i=x,Z_i=z\right] &= E\left[Y_{i,Z_i}^{D_i}|X_i=x,Z_i=z,\tau_i=n\right]\cdot P\left(\tau_i=n|X_i=x,Z_i=z\right)\\
&\quad +E\left[Y_{i,Z_i}^{D_i}|X_i=x,Z_i=z,\tau_i=c\right]\cdot P\left(\tau_i=c|X_i=x,Z_i=z\right)\\
&\quad +E\left[Y_{i,Z_i}^{D_i}|X_i=x,Z_i=z,\tau_i=d\right]\cdot P\left(\tau_i=d|X_i=x,Z_i=z\right)\\
&\quad +E\left[Y_{i,Z_i}^{D_i}|X_i=x,Z_i=z,\tau_i=a\right]\cdot P\left(\tau_i=a|X_i=x,Z_i=z\right)\\
&= E\left[Y_{i,Z_i}^{0}|X_i=x,Z_i=z,\tau_i=n\right]\cdot P\left(\tau_i=n|X_i=x\right)\\
&\quad +E\left[Y_{i,Z_i}^{D_i}|X_i=x,Z_i=z,\tau_i=c\right]\cdot P\left(\tau_i=c|X_i=x\right)\\
&\quad +E\left[Y_{i,Z_i}^{D_i}|X_i=x,Z_i=z,\tau_i=d\right]\cdot P\left(\tau_i=d|X_i=x\right)\\
&\quad +E\left[Y_{i,Z_i}^{1}|X_i=x,Z_i=z,\tau_i=a\right]\cdot P\left(\tau_i=a|X_i=x\right)
\end{aligned}$$

where the second equality makes use of the assumption of unconfounded type (Assumption 4). By the mean exclusion restriction (Assumption 5) the potential outcomes are independent of Z in the always- and in the never-taker subpopulation. Hence when taking the difference $E[Y|X,Z=1]-E[Y|X,Z=0]$ the respective terms for the always- and for the never-takers cancel:

$$\begin{aligned}
&E\left[Y_i|X_i=x,Z_i=1\right]-E\left[Y_i|X_i=x,Z_i=0\right]\\
&=\left(E\left[Y_{i,Z_i}^{1}|X_i=x,Z_i=1,\tau_i=c\right]-E\left[Y_{i,Z_i}^{0}|X_i=x,Z_i=0,\tau_i=c\right]\right)\cdot P\left(\tau_i=c|X_i=x\right)\\
&+\left(E\left[Y_{i,Z_i}^{0}|X_i=x,Z_i=1,\tau_i=d\right]-E\left[Y_{i,Z_i}^{1}|X_i=x,Z_i=0,\tau_i=d\right]\right)\cdot P\left(\tau_i=d|X_i=x\right)
\end{aligned}$$

and exploiting the mean exclusion restriction for the compliers and the defiers gives

$$\begin{aligned}
&=E\left[Y_{i,Z_i}^{1}-Y_{i,Z_i}^{0}\;|X_i=x,\tau_i=c\right]\cdot P\left(\tau_i=c|X_i=x\right)\\
&\qquad\qquad -E\left[Y_{i,Z_i}^{1}-Y_{i,Z_i}^{0}\;|X_i=x,\tau_i=d\right]\cdot P\left(\tau_i=d|X_i=x\right). \quad (17)
\end{aligned}$$

Hence the difference $E\left[Y|X,Z=1\right]-E\left[Y|X,Z=0\right]$ represents the difference between the average treatment effect on the compliers (who switch from $D_i=0$ to 1 as a reaction on a change in the instrument from 0 to 1) and the average treatment effect on the defiers (who switch from $D_i=1$ to 0). This is the net average treatment effect on all units that are induced to switch D due to a change in the instrumental variable. An estimate of (17) is not very informative since, for example, an estimate of zero could be the result of D having no effect on Y as well as the result of D having a large impact which is offset by opposite flows of compliers and defiers. Hence the exclusion restriction (Assumption 5) is not sufficient to isolate a meaningful treatment effect. However, as (17) indicates, a treatment effect could be identified if either no compliers $P\left(\tau_i=c\right)=0$ or no defiers $P\left(\tau_i=d\right)=0$ existed. If an instrumental variable is found that affects *all* units in the 'same direction', e.g. that either induces units to switch to $D_i=1$ or leaves D_i unchanged, but does not induce any unit to switch to $D_i=0$, then the average treatment effect on the responsive subpopulation is identified.

The monotonicity assumption (Assumption 2) rules out the existence of defiers. It follows from (17) that

$$\gamma(x)=E\left[Y^1_{i,Z_i}-Y^0_{i,Z_i}\,|X_i=x,\tau_i=c\right]=\frac{E\left[Y_i|X_i=x,Z_i=1\right]-E\left[Y_i|X_i=x,Z_i=0\right]}{P\left(\tau_i=c|X_i=x\right)}.$$

Noticing that $E\left[D|X,Z=0\right]=P(D=1|X,Z=0)=P\left(\tau=a|X\right)+P\left(\tau=d|X\right)$ and $E[D|X,Z=1]=P(D=1|X,Z=1)=P\left(\tau=a|X\right)+P\left(\tau=c|X\right)$, the relative size of the subpopulation of compliers is identified as

$$P\left(\tau_i=c|X_i=x\right)=E\left[D_i|X_i=x,Z_i=1\right]-E\left[D_i|X_i=x,Z_i=0\right],$$

and it follows that the average treatment effect in the subpopulation of compliers is

$$\gamma(x)=\frac{E\left[Y_i|X_i=x,Z_i=1\right]-E\left[Y_i|X_i=x,Z_i=0\right]}{E\left[D_i|X_i=x,Z_i=1\right]-E\left[D_i|X_i=x,Z_i=0\right]}.$$

A.2 Proof of Theorem 1

Let $\hat\gamma=\frac{\hat\Delta}{\hat\Gamma}$ denote the estimator (6) of the average treatment effect on the compliers (5):

$$\gamma=\frac{\Delta}{\Gamma}=\frac{\int\left(m_1(x)-m_0(x)\right)f_x(x)dx}{\int\left(\mu_1(x)-\mu_0(x)\right)f_x(x)dx}.$$

To derive the asymptotic distribution of $\hat\gamma$, note that $\hat\gamma-\gamma$ can be written as

$$(\hat\gamma-\gamma)=\frac{\hat\Delta}{\hat\Gamma}-\frac{\Delta}{\Gamma}=\left(\frac{\hat\Delta-\Delta}{\Gamma}-\gamma\frac{\hat\Gamma-\Gamma}{\Gamma}\right)\cdot\left(1-\frac{\hat\Gamma-\Gamma}{\hat\Gamma}\right). \tag{18}$$

The derivation proceeds in two steps. First it is shown that the last term $\left(1 - \frac{\hat{\Gamma}-\Gamma}{\hat{\Gamma}}\right)$ in (18) is $1 + o_p(1)$. Hence the first-order behaviour of $\hat{\gamma} - \gamma$ is determined by the term $\frac{\hat{\Delta}-\Delta}{\Gamma} - \gamma\frac{\hat{\Gamma}-\Gamma}{\Gamma}$ in

$$\hat{\gamma} - \gamma = \left(\frac{\hat{\Delta} - \Delta}{\Gamma} - \gamma\frac{\hat{\Gamma} - \Gamma}{\Gamma}\right) \cdot (1 + o_p(1)) . \tag{19}$$

In the second step the asymptotic distribution of this first-order term is derived.

In a preliminary step the term $\hat{\Delta} - \Delta$ is analyzed. (The derivations for $\hat{\Gamma} - \Gamma$ are analogous.) Write $\hat{\Delta} - \Delta$ as

$$\begin{aligned}\hat{\Delta} - \Delta &= \frac{1}{n}\left(\sum_{i:Z_i=1} Y_i + \sum_{i:Z_i=0} \hat{m}_1(X_i) - \sum_{i:Z_i=0} Y_i - \sum_{i:Z_i=1} \hat{m}_0(X_i)\right) - E\left[m_1(X) - m_0(X)\right] \\ &= \frac{1}{n}\left(\sum_{i:Z_i=1} (Y_i - m_1(X_i)) + \sum_{i:Z_i=0} (\hat{m}_1(X_i) - m_1(X_i))\right) \\ &\quad - \frac{1}{n}\left(\sum_{i:Z_i=0} (Y_i - m_0(X_i)) + \sum_{i:Z_i=1} (\hat{m}_0(X_i) - m_0(X_i))\right) \\ &\qquad + \frac{1}{n}\sum_i (m_1(X_i) - m_0(X_i)) - E\left[m_1(X) - m_0(X)\right],\end{aligned}$$

and introducing the asymptotic linear representation of the nonparametric estimators $\hat{m}_z$

$$\begin{aligned}&= \frac{1}{n}\sum_{i:Z_i=1} (Y_i - m_1(X_i)) - \frac{1}{n}\sum_{i:Z_i=0} (Y_i - m_0(X_i)) \\ &\quad + \frac{n_0}{n}\frac{1}{n_0 n_1}\sum_{i:Z_i=0}\sum_{j:Z_j=1} \xi_1^m(Y_j, X_j, X_i) + \frac{1}{n}\sum_{i:Z_i=0} \left[b_1^m(X_i) + R_1^m(X_i)\right] \\ &\quad - \frac{n_1}{n}\frac{1}{n_0 n_1}\sum_{i:Z_i=1}\sum_{j:Z_j=0} \xi_0^m(Y_j, X_j, X_i) - \frac{1}{n}\sum_{i:Z_i=1} \left[b_0^m(X_i) + R_0^m(X_i)\right] \\ &\qquad + \frac{1}{n}\sum_i (m_1(X_i) - m_0(X_i)) - E\left[m_1(X) - m_0(X)\right].\end{aligned}$$

The terms $\frac{1}{n_0 n_1}\sum_{i:Z_i=0}\sum_{j:Z_j=1} \xi^m$ represent mean-zero two-sample U-statistics, to which a projection theorem can be applied:

$$\begin{aligned}\frac{1}{n_0 n_1}\sum_{i:Z_i=0}\sum_{j:Z_j=1} \xi_1^m(Y_j, X_j, X_i) &= \frac{1}{n_0}\sum_{i:Z_i=0} E\left[\xi_1^m\left(Y_1, X_1, X_2\right)|X_2 = X_i\right] \\ &\quad + \frac{1}{n_1}\sum_{j:Z_j=1} E\left[\xi_1^m\left(Y_1, X_1, X_2\right)|Y_1 = Y_j, X_1 = X_j\right] + o_p(n^{-\frac{1}{2}}) \\ &= \frac{1}{n_1}\sum_{j:Z_j=1} E\left[\xi_1^m\left(Y_j, X_j, X\right)|Y_j, X_j\right] + o_p(n^{-\frac{1}{2}}),\end{aligned}$$

where the latter equality follows from condition (iv A) of Theorem 1. Application of the projection theorem requires that $E\left[\|\xi_1^m(Y_j,X_j,X_i)\|^2\right]=o(n)$, see Hoeffding (1948), Serfling (1980) or Heckman, Ichimura, and Todd (1998), which is satisfied from condition (iv B) of Theorem 1.

With this projection theorem it follows that

$$\begin{aligned}\hat{\Delta}-\Delta=\frac{1}{n}\sum_{i:Z_i=1}(Y_i-m_1(X_i))-\frac{1}{n}\sum_{i:Z_i=0}(Y_i-m_0(X_i))\\+\frac{n_0}{nn_1}\sum_{j:Z_j=1}E\left[\xi_1^m\left(Y_j,X_j,X\right)|Y_j,X_j\right]-\frac{n_1}{nn_0}\sum_{j:Z_j=0}E\left[\xi_0^m\left(Y_j,X_j,X\right)|Y_j,X_j\right]\\+\frac{1}{n}\sum_i(m_1(X_i)-m_0(X_i))-E\left[m_1(X)-m_0(X)\right]+o_p(n^{-\frac{1}{2}}),\quad(20)\end{aligned}$$

where it has also been taken into account that the average bias and residual terms $\frac{1}{n}\sum b^m(X_i)$ and $\frac{1}{n}\sum R^m(X_i)$ are $o_p(n^{-\frac{1}{2}})$ by the conditions (iv C) and (iv D) of Theorem 1.

By a weak law of large numbers, sample means converge to their expectations and thus $\hat{\Delta}-\Delta=o_p(1)$ and analogously $\hat{\Gamma}-\Gamma=o_p(1)$. Hence

$$\frac{\hat{\Gamma}-\Gamma}{\hat{\Gamma}}=\left(1+\frac{\hat{\Gamma}-\Gamma}{\Gamma}\right)^{-1}\cdot\frac{\left(\hat{\Gamma}-\Gamma\right)}{\Gamma}=O_p(1)\cdot o_p(1)=o_p(1),$$

because $(1+o_p(1))^{-1}=O_p(1)$ (van der Vaart 1998, p. 13). This implies (19).

Hence the leading term in (19) is $\left(\frac{\hat{\Delta}-\Delta}{\Gamma}-\gamma\frac{\hat{\Gamma}-\Gamma}{\Gamma}\right)$. The approximation to $\sqrt{n}\left(\hat{\gamma}-\gamma\right)$ up to first order is thus

$$\sqrt{n}\left(\hat{\gamma}-\gamma\right)=\sqrt{n}\left(\frac{\hat{\Delta}-\Delta}{\Gamma}-\gamma\frac{\hat{\Gamma}-\Gamma}{\Gamma}\right).$$

Inserting the expression (20) and the analogous expression for $\hat{\Gamma}-\Gamma$ gives

$$\sqrt{n}\left(\hat{\gamma}-\gamma\right)=\frac{\sqrt{n}}{\Gamma}\frac{1}{n}\sum_i\Xi_i$$

up to first order, where

$$\begin{aligned}\Xi_i&=&Z_i\left(Y_i-m_1(X_i)\right)-(1-Z_i)\left(Y_i-m_0(X_i)\right)+m_1(X_i)-m_0(X_i)\\&&+\frac{n_0}{n_1}Z_iE\left[\xi_1^m\left(Y_j,X_j,X\right)|Y_j,X_j\right]-\frac{n_1}{n_0}\left(1-Z_i\right)E\left[\xi_0^m\left(Y_j,X_j,X\right)|Y_j,X_j\right]\\&&-\gamma Z_i\left(D_i-\mu_1(X_i)\right)+\gamma\left(1-Z_i\right)\left(D_i-\mu_0(X_i)\right)-\gamma\left(\mu_1(X_i)-\mu_0(X_i)\right)\\&&-\frac{\gamma n_0}{n_1}Z_iE\left[\xi_1^d\left(Y_j,X_j,X\right)|Y_j,X_j\right]+\frac{\gamma n_1}{n_0}\left(1-Z_i\right)E\left[\xi_0^d\left(Y_j,X_j,X\right)|Y_j,X_j\right]\\&&-E\left[m_1(X)-m_0(X)\right]+\gamma E\left[\mu_1(X)-\mu_0(X)\right].\end{aligned}$$

Note that the last row is zero, because $-E\left[m_1(X)-m_0(X)\right]+\gamma E\left[\mu_1(X)-\mu_0(X)\right]=-\Delta+\gamma\Gamma=-\Delta+\frac{\Delta}{\Gamma}\Gamma=0$. By condition (iv E) the influence functions ξ^m and ξ^d have the form

$$\begin{aligned} E\left[\xi_1^m(Y_j,X_j,X_i)|Y_j,X_j,Z_j=1,Z_i=0\right] &= (Y_j-m_1(X_j))\frac{f_{x|z=0}\left(X_j\right)}{f_{x|z=1}\left(X_j\right)}+o_p(1) \\ &= (Y_j-m_1(X_j))\frac{P\left(Z=1\right)}{P\left(Z=0\right)}\frac{1-\pi(X_j)}{\pi(X_j)}+o_p(1) \end{aligned}$$

where $\pi(x)=P\left(Z=1|X=x\right)$ is the probability that Z takes the value one given characteristics X. With $\frac{n_0}{n_1}=\frac{P(Z=0)}{P(Z=1)}+o_p(1)$ it follows that

$$\begin{aligned} \Xi_i &= Z_i\left(Y_i-m_1(X_i)+(Y_i-m_1(X_i))\frac{1-\pi(X_i)}{\pi(X_i)}\right) \\ &\quad -(1-Z_i)\left(Y_i-m_0(X_i)+(Y_i-m_0(X_i))\frac{\pi(X_i)}{1-\pi(X_i)}\right) \\ &\quad -\gamma Z_i\left(D_i-\mu_1(X_i)+(D_i-\mu_1(X_i))\frac{1-\pi(X_i)}{\pi(X_i)}\right) \\ &\quad +\gamma\left(1-Z_i\right)\left(D_i-\mu_0(X_i)+(D_i-\mu_0(X_i))\frac{\pi(X_i)}{1-\pi(X_i)}\right) \\ &\quad +m_1(X_i)-m_0(X_i)-\gamma\left(\mu_1(X_i)-\mu_0(X_i)\right) \end{aligned}$$

plus an $o_p(1)$ term. Collecting terms gives

$$\begin{aligned} \Xi_i = Z_i&\left(\frac{(Y_i-m_1(X_i))-\gamma\left(D_i-\mu_1(X_i)\right)}{\pi(X_i)}\right) \\ &+(1-Z_i)\left(\frac{\gamma\left(D_i-\mu_0(X_i)\right)-(Y_i-m_0(X_i))}{1-\pi(X_i)}\right) \\ &\qquad +m_1(X_i)-m_0(X_i)-\gamma\left(\mu_1(X_i)-\mu_0(X_i)\right). \end{aligned}$$

Computing the variance of Ξ_i gives:

$$\begin{aligned} &Var\left(\Xi_i\right) \\ &=E\left[Z_i^2\left(\frac{(Y_i-m_1(X_i))-\gamma\left(D_i-\mu_1(X_i)\right)}{\pi(X_i)}\right)^2\right] \\ &+E\left[(1-Z_i)^2\left(\frac{\gamma\left(D_i-\mu_0(X_i)\right)-(Y_i-m_0(X_i))}{1-\pi(X_i)}\right)^2\right] \\ &+E\left[\left(m_1(X_i)-m_0(X_i)-\gamma\left(\mu_1(X_i)-\mu_0(X_i)\right)\right)^2\right] \\ &+E\left[Z_i\left(\frac{(Y_i-m_1(X_i))-\gamma\left(D_i-\mu_1(X_i)\right)}{\pi(X_i)}\right)(1-Z_i)\left(\frac{\gamma\left(D_i-\mu_0(X_i)\right)-(Y_i-m_0(X_i))}{1-\pi(X_i)}\right)\right] \\ &+E\left[Z_i\left(\frac{(Y_i-m_1(X_i))-\gamma\left(D_i-\mu_1(X_i)\right)}{\pi(X_i)}\right)\left(m_1(X_i)-m_0(X_i)-\gamma\left(\mu_1(X_i)-\mu_0(X_i)\right)\right)\right] \\ &+E\left[(1-Z_i)\left(\frac{\gamma\left(D_i-\mu_0(X_i)\right)-(Y_i-m_0(X_i))}{1-\pi(X_i)}\right)\left(m_1(X_i)-m_0(X_i)-\gamma\left(\mu_1(X_i)-\mu_0(X_i)\right)\right)\right], \end{aligned}$$

where the fourth term is zero because $Z(1-Z)=0$ and the fifth and sixth terms are zero conditional on X and Z. Since $Z^2=Z$, it follows

$$
\begin{aligned}
Var\,(\Xi) = E&\left[Z\frac{\left[(Y-m_1(X))-\gamma\,(D-\mu_1(X))\right]^2}{\pi(X)^2}\right] \\
&+E\left[(1-Z)\,\frac{\left[\gamma\,(D-\mu_0(X))-(Y-m_0(X))\right]^2}{(1-\pi(X))^2}\right] \\
&+E\left[\left(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X)\right)^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&=E\left[E\left[E\left[\left[(Y-m_1(X))-\gamma\,(D-\mu_1(X))\right]^2\ |X,Z=1\right]\cdot\frac{P\,(Z=1|X)}{\pi(X)^2}\ |X\right]\right] \\
&\quad+E\left[E\left[E\left[\left[\gamma\,(D-\mu_0(X))-(Y-m_0(X))\right]^2\ |X,Z=0\right]\cdot\frac{P\,(Z=0|X)}{(1-\pi(X))^2}\ |X\right]\right] \\
&\quad+E\left[\left(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X)\right)^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&=E\left[\frac{\sigma^2_{Y_1}(X)-2\gamma\sigma^2_{Y_1D_1}(X)+\gamma^2\sigma^2_{D_1}(X)}{\pi(X)}+\frac{\sigma^2_{Y_0}(X)-2\gamma\sigma^2_{Y_0D_0}(X)+\gamma^2\sigma^2_{D_0}(X)}{1-\pi(X)}\right] \\
&\quad+E\left[\left(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X)\right)^2\right]
\end{aligned}
$$

by iterated expectations, where $\sigma^2_{Y_1}(x)=Var[Y|X,Z=1]=E[(Y-m_1(X))^2\,|X=x,Z=1]$, and $\sigma^2_{Y_1D_1}(x)=Cov[Y,D|X,Z=1]$, and $\sigma^2_{D_1}(x)$, $\sigma^2_{Y_0}(x)$, $\sigma^2_{D_0}(x)$, $\sigma^2_{Y_0D_0}(x)$ defined analogously.

Applying the Lindberg-Levy central limit theorem to

$$\sqrt{n}\,(\hat{\gamma}-\gamma)=\frac{\sqrt{n}}{\Gamma}\frac{1}{n}\sum_i\Xi_i$$

gives that the estimator $\hat{\gamma}$ is root-n asymptotically normal

$$\sqrt{n}\,(\hat{\gamma}-\gamma)\to N\,(0,\mathcal{V})$$

with asymptotic variance

$$
\begin{aligned}
\mathcal{V}=\frac{1}{\Gamma^2}E&\left[\frac{\sigma^2_{Y_1}(X)-2\gamma\sigma^2_{Y_1D_1}(X)+\gamma^2\sigma^2_{D_1}(X)}{\pi(X)}+\frac{\sigma^2_{Y_0}(X)-2\gamma\sigma^2_{Y_0D_0}(X)+\gamma^2\sigma^2_{D_0}(X)}{1-\pi(X)}\right] \\
&+\frac{1}{\Gamma^2}E\left[\left(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X)\right)^2\right].
\end{aligned}
$$

A.3 Proof of Theorem 2

Semiparametric efficiency bounds were introduced by Stein (1956) and developed by Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), Begun, Hall, Huang, and Wellner (1983) and Bickel, Klaassen, Ritov, and Wellner (1993). See also the survey of Newey (1990) or Newey (1994). The approach followed here is similar to Hahn (1998).

The joint density of (Y, D, Z, X) with Z binary can be written as

$$f(y,d,z,x) = f(y,d|z,x)\, f(z|x)\, f(x) = \{f_1(y,d|x)\ \ \pi(x)\}^z \{f_0(y,d|x)\ \ (1-\pi(x))\}^{1-z} f(x)$$

where $f_1(y,d|x) \equiv f(y,d|z=1,x)$ and $\pi(x) = P(Z=1|X=x)$.

Consider a regular parametric submodel indexed by θ with θ_0 corresponding to the true model: $f(y,d,z,x|\theta_0) = f(y,d,z,x)$. The density $f(y,d,z,x|\theta)$ can be written as

$$f(y,d,z,x|\theta) = \{f_1(y,d|x,\theta)\ \ \pi(x,\theta)\}^z \{f_0(y,d|x,\theta)\ \ (1-\pi(x,\theta))\}^{1-z} f(x,\theta),$$

and the corresponding score of $f(y,d,z,x|\theta)$ is

$$\begin{aligned} S(y,d,z,x|\theta) &= \frac{\partial \ln f(y,d,z,x|\theta)}{\partial\theta} \\ &= z \cdot \breve{f}_1(y,d|x,\theta) + (1-z)\cdot \breve{f}_0(y,d|x,\theta) + \frac{z-\pi(x,\theta)}{1-\pi(x,\theta)}\breve{\pi}(x,\theta) + \breve{f}(x,\theta) \end{aligned}$$

where $\breve{f}_1(y,d|x,\theta) = \partial \ln f_1(y,d|x,\theta)/\partial\theta$, and $\breve{f}_0$ analogously, and $\breve{\pi}(x,\theta) = \partial \ln \pi(x,\theta)/\partial\theta$ and $\breve{f}(x,\theta) = \partial \ln f(x,\theta)/\partial\theta$.

At the true value θ_0 the expectation of the score is zero. The tangent space of the model is the set of functions that are mean zero and satisfy the additive structure of the score:

$$\Im = \{z\cdot s_1(y,d|x) + (1-z)\cdot s_0(y,d|x) + (z-\pi(x))\cdot s_\pi(x) + s_x(x)\} \tag{21}$$

for any functions s_1, s_0, s_π, s_x satisfying the mean-zero property:

$\int s_1(y,\mathrm{d}|x)\, f_1(y,\mathrm{d}|x)\, dyd\mathrm{d} = 0\ \forall x$

$\int s_0(y,\mathrm{d}|x)\, f_0(y,\mathrm{d}|x)\, dyd\mathrm{d} = 0\ \forall x$

$\int s_x(x)\, f(x)\, dx = 0$

and $s_\pi(x)$ being a square-integrable measurable function of x.

The *semiparametric variance bound* of γ is the variance of the projection on $\Im$ of a function $\psi(Y,D,Z,X)$ (with $E[\psi] = 0$ and $E[\|\psi(\cdot)\|^2] < \infty$) that satisfies for all regular parametric submodels

$$\frac{\partial \gamma(F_\theta)}{\partial \theta}_{|\theta=\theta_0} = E\left[\psi(Y,D,Z,X)\cdot S(Y,D,Z,X)\right]_{|\theta=\theta_0} \tag{22}$$

Write γ as

$$\begin{aligned}
\gamma &= \frac{\Delta}{\Gamma} = \frac{\int \left(E_\theta\left[Y|X,Z=1\right] - E_\theta\left[Y|X,Z=0\right]\right)\cdot f(x,\theta)dx}{\int \left(E_\theta\left[D|X,Z=1\right] - E_\theta\left[D|X,Z=0\right]\right)\cdot f(x,\theta)dx} \\
&= \frac{\int \left(\int\int y f_1\left(y,d|x,\theta\right)dyd\mathrm{d} - \int\int y f_0\left(y,d|x,\theta\right)dyd\mathrm{d}\right)\cdot f(x,\theta)dx}{\int \left(\int\int d f_1\left(y,d|x,\theta\right)dyd\mathrm{d} - \int\int d f_0\left(y,d|x,\theta\right)dyd\mathrm{d}\right)\cdot f(x,\theta)dx} \\
&= \frac{\int\int\int y f_1\left(y,d|x,\theta\right)f(x,\theta)dyd\mathrm{d}dx - \int\int\int y f_0\left(y,d|x,\theta\right)f(x,\theta)dyd\mathrm{d}dx}{\int\int\int d f_1\left(y,d|x,\theta\right)f(x,\theta)dyd\mathrm{d}dx - \int\int\int d f_0\left(y,d|x,\theta\right)f(x,\theta)dyd\mathrm{d}dx}
\end{aligned}$$

since $E[Y|X,Z=1] = \int\int y f_1\left(y,\mathrm{d}|x\right)dyd\mathrm{d}$.[18]

Computing the pathwise derivative and evaluating it at θ_0 gives:

$$\begin{aligned}
\frac{\partial \gamma(F_\theta)}{\partial \theta}_{|\theta=\theta_0} &= \frac{\frac{\partial\Delta}{\partial\theta}\Gamma - \Delta\frac{\partial\Gamma}{\partial\theta}}{\Gamma^2}_{|\theta=\theta_0} = \frac{\partial\Delta/\partial\theta}{\Gamma} - \gamma\frac{\partial\Gamma/\partial\theta}{\Gamma}_{|\theta=\theta_0} \\
&= \frac{\int\int\int y\left(\dot{f}_1 f + f_1\dot{f}\right)dyd\mathrm{d}dx - \int\int\int y\left(\dot{f}_0 f + f_0\dot{f}\right)dyd\mathrm{d}dx}{\Gamma} \\
&\quad -\gamma\frac{\int\int\int \mathrm{d}\left(\dot{f}_1 f + f_1\dot{f}\right)dyd\mathrm{d}dx - \int\int\int \mathrm{d}\left(\dot{f}_0 f + f_0\dot{f}\right)dyd\mathrm{d}dx}{\Gamma} \\
&= \frac{\int\int\int y\left\{\dot{f}_1 - \dot{f}_0\right\}fdyd\mathrm{d}dx}{\Gamma} - \gamma\frac{\int\int\int \mathrm{d}\left\{\dot{f}_1 - \dot{f}_0\right\}fdyd\mathrm{d}dx}{\Gamma} \\
&\quad +\frac{\int\left(m_1(x) - m_0(x) - \gamma\mu_1(x) + \gamma\mu_0(x)\right)\dot{f}(x)dx}{\Gamma}
\end{aligned}$$

where $\dot{f}_1 = \frac{\partial}{\partial\theta}f_1\left(y,d|x,\theta\right)_{|\theta=\theta_0}$, $\dot{f}_0 = \frac{\partial}{\partial\theta}f_0\left(y,d|x,\theta\right)_{|\theta=\theta_0}$ and $\dot{f} = \frac{\partial}{\partial\theta}f\left(x,\theta\right)_{|\theta=\theta_0}$.

Choose $\psi(Y,D,Z,X)$ as

$$\begin{aligned}
\psi(y,d,z,x) = \frac{z}{\Gamma}\frac{y - m_1(x) - \gamma d + \gamma\mu_1(x)}{\pi(x)} + \frac{1-z}{\Gamma}\frac{\gamma d - \gamma\mu_0(x) - y + m_0(x)}{1-\pi(x)} \\
+ \frac{m_1(x) - m_0(x) - \gamma\mu_1(x) + \gamma\mu_0(x)}{\Gamma}.
\end{aligned} \tag{23}$$

Notice that ψ satisfies (22)

$$\frac{\partial \gamma(F_\theta)}{\partial \theta}_{|\theta=\theta_0} = E\left[\psi(Y,D,Z,X)\cdot S(Y,D,Z,X)\right]_{|\theta=\theta_0}$$

and that ψ lies in the tangent space (21)[19]

$$\psi \in \Im.$$

[18] And analogously for $E[D|X,Z=1]$, $E[Y|X,Z=0]$ and $E[D|X,Z=0]$.

[19] The calculations are available from the author.

Since ψ lies in the tangent space, the variance bound is the expected square of ψ:

$$
\begin{aligned}
E\left[\psi(Y,D,Z,X)^2\right] = \frac{1}{\Gamma^2}E\left[Z\left(\frac{Y-m_1(X)-\gamma D+\gamma\mu_1(X)}{\pi(X)}\right)^2\right] \\
+\frac{1}{\Gamma^2}E\left[(1-Z)\left(\frac{\gamma D-\gamma\mu_0(X)-Y+m_0(X)}{1-\pi(X)}\right)^2\right] \\
+\frac{1}{\Gamma^2}E\left[Z\frac{Y-m_1(X)-\gamma D+\gamma\mu_1(X)}{\pi(X)}\cdot(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))\right] \\
+\frac{1}{\Gamma^2}E\left[(1-Z)\frac{\gamma D-\gamma\mu_0(X)-Y+m_0(X)}{1-\pi(X)}\cdot(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))\right] \\
+\frac{1}{\Gamma^2}E\left[(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2\right]
\end{aligned}
$$

$$
\begin{aligned}
=\frac{1}{\Gamma^2}E\left[E\left[(Y-m_1(X)-\gamma D+\gamma\mu_1(X))^2\ |X,Z=1\right]\frac{P(Z=1|X)}{\pi(X)^2}\right] \\
+\frac{1}{\Gamma^2}E\left[E\left[(\gamma D-\gamma\mu_0(X)-Y+m_0(X))^2|X,Z=0\right]\frac{P(Z=0|X)}{(1-\pi(X))^2}\right] \\
+\frac{1}{\Gamma^2}E\left[(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2\right]
\end{aligned}
$$

by iterated expectations. Defining $\sigma^2_{Y_1}(x)=Var[Y|X,Z=1]$, $\sigma^2_{D_1}(x)=Var[D|X,Z=1]$ and $\sigma^2_{Y_1D_1}(x)=Cov[Y,D|X,Z=1]$ and analogously $\sigma^2_{Y_0}(x)$, $\sigma^2_{D_0}(x)$, $\sigma^2_{Y_0D_0}(x)$ for $Z=0$, gives the asymptotic variance bound:

$$
\begin{aligned}
E\left[\frac{\sigma^2_{Y_1}(X)-2\gamma\sigma^2_{Y_1D_1}(X)+\gamma^2\sigma^2_{D_1}(X)}{\Gamma^2\pi(X)}+\frac{\sigma^2_{Y_0}(X)-2\gamma\sigma^2_{Y_0D_0}(X)+\gamma^2\sigma^2_{D_0}(X)}{\Gamma^2(1-\pi(X))}\right] \\
+E\left[\frac{(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2}{\Gamma^2}\right].
\end{aligned}
$$

References

Abadie, A. (2001): "Semiparametric Instrumental Variable Estimation of Treatment Response Models," mimeo, Harvard University, December 2001.

Angrist, J. (1990): "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence From Social Security Administrative Records," *American Economic Review*, 80, 313–336.

Angrist, J., K. Graddy, and G. 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, 499–527.

Angrist, J., and G. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of American Statistical Association*, 90, 431–442.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472 (with discussion).

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

——— (1999): "Empirical Strategies in Labor Economics," in *The Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1277–1366. North-Holland, New York.

Begun, J., W. Hall, W. Huang, and J. Wellner (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432–452.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. John Hopkins University Press, Baltimore.

Blundell, R., and J. Powell (2001): "Endogeneity in Nonparametric and Semiparametric Regression Models," cemmap working paper 09/01.

Card, D. (1995): "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, ed. by L. Christofides, E. Grant, and R. Swidinsky, pp. 201–222. University of Toronto Press, Toronto.

Darolles, S., J. Florens, and E. Renault (2001): "Nonparametric Instrumental Regression," mimeo, Toulouse.

Das, M. (2000): "Instrumental Variables Estimation of Nonparametric Models with Discrete Endogenous Regressors," mimeo, Columbia University.

Florens, J. (2002): "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables," mimeo, Toulouse.

Florens, J., J. Heckman, C. Meghir, and E. Vytlacil (2002): "Instrumental Variables, Local Instrumental Variables and Control Functions," cemmap working paper 15/02.

Frölich, M. (2001): "Nonparametric Covariate Adjustment: Pair-matching versus Local Polynomial Matching," *University of St. Gallen Economics Discussion Paper Series*, 2000-17.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

HEARST, N., T. NEWMAN, AND S. HULLEY (1986): "Delayed Effects of the Military Draft on Mortality: A Randomized Natural Experiment," *New England Journal of Medicine*, 314, 620–624.

HECKMAN, J. (1997): "Instrumental Variables - A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32, 441–462.

HECKMAN, J., H. ICHIMURA, AND P. TODD (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

HECKMAN, J., R. LALONDE, AND J. SMITH (1999): "The Economics and Econometrics of Active Labour Market Programs," in *The Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1865–2097. North-Holland, New York.

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

——— (2001): "Local Instrumental Variables," in *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. Powell. Cambridge University Press, Cambridge.

HIRANO, K., G. IMBENS, D. RUBIN, AND X. ZHOU (2000): "Assessing the effect of an influenza vaccine in an encouragement design," *Biostatistics*, 1, 69–88.

HOEFFDING, W. (1948): "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics*, 19, 293–325.

IMBENS, G. (2000): "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710.

——— (2001): "Some remarks on instrumental variables," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 17–42. Physica/Springer, Heidelberg.

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

IMBENS, G., AND W. NEWEY (2001): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," mimeo, UCLA and MIT.

IMBENS, G., AND D. RUBIN (1997): "Estimating outcome distributions for compliers in instrumental variables models," *Review of Economic Studies*, 64, 555–574.

IMBENS, G., D. RUBIN, AND B. SACERDOTE (2001): "Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players," *American Economic Review*, 91, 778–794.

KOSHEVNIK, Y., AND B. LEVIT (1976): "On a Non-parametric Analogue of the Information Matrix," *Theory of Probability and Applications*, 21, 738–753.

LECHNER, M. (2001): "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 43–58. Physica/Springer, Heidelberg.

NEWEY, W. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

NEWEY, W., AND J. POWELL (first draft 1988, revised 2002): "Instrumental Variable Estimation of Nonparametric Models," mimeo, MIT and Berkeley.

NEWEY, W., J. POWELL, AND F. VELLA (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

PAGAN, A., AND A. ULLAH (1999): *Nonparametric Econometrics*. Cambridge University Press, Cambridge.

PFANZAGL, J., AND W. WEFELMEYER (1982): *Contributions to a General Asymptotic Statistical Theory*. Springer Verlag, Berlin.

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

RUBIN, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

SERFLING, R. (1980): *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

STEIN, C. (1956): "Efficient Nonparametric Testing and Estimation," in *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. University of California Press, Berkeley.

VAN DER VAART, A. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

YAU, L., AND R. LITTLE (2001): "Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed," *Journal of American Statistical Association*, 96, 1232–1244.