Essays in Non- and Semiparametric Econometrics

Inauguraldissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenschaften der Universität Mannheim

vorgelegt von

Christoph Rothe

April 2009

.

Abteilungssprecher:	Prof. Dr. Enno Mammen
Referent:	Prof. Dr. Enno Mammen
Koreferent:	Prof. Richard Blundell, Ph.D.
Tag der Verteidigung:	28.05.2009

Acknowledgements

First of all, I would like to thank my adviser Enno Mammen for his guidance and supervision during the process of writing this thesis. He introduced me to the fascinating field of non- and semiparametric econometrics, and my research has benefited a lot from his deep knowledge about statistics. During my time at the Chair of Statistics, I have received all the support I could ever have asked for, and I am very grateful for that.

I also wish to thank Richard Blundell for advice on my thesis and beyond. I benefited a lot from four very encouraging and insightful months in London that did not only broaden my perspective on my own topic but on econometrics in general.

I would also like to thank my colleagues at the Chair of Statistics, in particular Kyusang Yu, Melanie Schienle and Christoph Nagel, for all the lively discussions about econometrics, statistics and essentially anything else. The same goes for my fellow graduate students at the CDSE, who were important contributors to the wonderful research atmosphere I experienced here.

Parts of this thesis have been presented at conferences and seminars over the past three years, and I have received a number of helpful suggestions and comments from the respective audiences that lead to major improvements, which is gratefully acknowledged.

Finally, my family has always been an invaluable source of support. Thank you!

Christoph Rothe

Contents

Chapter 1

Introduction

With important advances made in econometric theory and the rapidly increasing availability of computing power and large datasets, the use of nonparametric and semiparametric techniques has gained considerable importance for applied economic research over the past three decades. Extensive overviews of recent developments in this area are given for example by Pagan and Ullah (1999) or Li and Racine (2007).

The general aim of nonparametric and semiparametric techniques is to weaken the often restrictive assumptions that are imposed in order to be able to use standard econometric methods. In a classical regression framework, for example, it is the aim of the researcher to investigate the functional relationship between the mean of an outcome variable of interest and a number of explanatory quantities. A typical, fully parametric approach to this problem would be to assume that the relationship can be represented through a function known up to a finite number of parameters, which can be estimated from the data by maximum likelihood or the method of least squares. Such a procedure will of course be adequate when the true underlying data generating process can be well approximated by the postulated functional form. It will, however, potentially result in grossly misleading conclusions under misspecification.

In order to avoid this problem, nonparametric methods replace the global parametric restrictions on the functional relationship between the outcome and the regressors with weaker conditions that only require the relationship to be sufficiently smooth in small neighborhoods. A typical requirement could for example be the existence of a second bounded derivative. Various techniques can be used to construct estimators based on

such restrictions, such as kernel smoothing, local polynomial regression or orthogonal series approximations, but throughout this thesis the focus will be on the former class.

Nonparametric estimators are highly flexible, but suffer from the so-called curse of dimensionality. The typical rate at which of such estimators converge to their corresponding population values deteriorates drastically if the dimension of the covariate space becomes larger. Such estimates might therefore not be reliable in practice. One option in this case is to consider an intermediate class of models, that describe certain aspects of the relationship through a parametric structure, while others are left more open. These so-called semiparametric models often represent a reasonable compromise in the sense that they balance the potential risk of misspecification and the potential inaccuracy of the estimator due to the curse of dimensionality.

This thesis contains of three main chapters, which contribute to the literature on non- and semiparametric econometrics. The chapters are self-contained and can be read separately. Chapter 2 and 3 each end with an appendix that contains the more technical arguments.

Chapter 2 proposes a fully nonparametric procedure to evaluate the effect of a counterfactual change in the distribution of some covariates on the unconditional distribution of an outcome variable of interest. Due to the focus on the unconditional distribution, we are able to circumvent the curse of dimensionality even in settings with a large number of covariates. In contrast to other methods, we do not restrict attention to the effect on the mean. In particular, our method can be used to conduct inference on the change of the distribution function as a whole, its moments and quantiles, inequality measures such as the Lorenz curve or Gini coefficient, and to test for stochastic dominance. The practical applicability of our procedure is illustrated via a simulation study and an empirical example.

In Chapter 3, we analyse a semiparametric estimator for the coefficients of a single index binary choice model with endogenous regressors. In order to achieve identification, we employ the control function approach used by Blundell and Powell (2003, 2004). The estimator we propose is a two-step semiparametric maximum likelihood (SML) estimator, that can be seen as a generalization of the popular approach of Klein and Spady (1993). The first step consists of estimating a reduced form equation for the endogenous regressors

and extracting the corresponding residuals. In the second step, the latter are added as control variates to the outcome equation, which is in turn estimated by SML. We establish the estimator's $\sqrt{n}$-consistency and asymptotic normality. In a simulation study, we compare the properties of our estimator with those of existing alternatives, highlighting the advantages of our approach.

In Chapter 4, we study identification of a certain class of parameters, called Unconditional Partial Effects, in nonseparable models with endogenous regressors, using a control variable approach due to Imbens and Newey (2009). We thus extend the work of Firpo, Fortin, and Lemieux (2009), who recently introduced this parameters for models without endogeneity. We also show that these effects can be written in terms of an average derivative of the conditional CDF of the outcome variable Y given the regressors X and the control variable V, where the derivative is taken with respect to X. This representation is useful to give an explicit expression for Unconditional Partial Effects in nonlinear parametric or semiparametric models.

Chapter 2

Nonparametric Estimation of Distributional Policy Effects

2.1 Introduction

In this paper, we propose a fully nonparametric procedure to evaluate the effect of a change in the distribution of some covariates on the unconditional distribution of an outcome variable of interest. We consider a general nonseparable model of the form

$$Y = m(X, \varepsilon) \tag{2.1.1}$$

where Y is the dependent variable of interest, X a vector of regressors and ε an unobserved error term that will usually represent individual heterogeneity. The question we are interested in is: How would the *unconditional* distribution of the dependent variable change if a policy maker could exogenously shift the values of X to some X^*, i.e what is the difference between the distribution of Y and the one of the (counterfactual) random variable

$$Y^* = m(X^*, \varepsilon).$$

We will call any difference between the distribution of Y and Y^* a *distributional policy effect*.

There are numerous examples in applied economics that fit into this rather abstract framework: Ichimura and Taber (2002) study the effect of a change in distribution of

income induced by the introduction of a tuition subsidy on college attendance rates and future earnings; Stock (1991) considers the effects of cleaning up a nearby hazardous waste disposal site on average house prices; DiNardo, Fortin, and Lemieux (1996) analyse how the distribution of wages would have evolved in the United States between 1973 and 1992 if the distribution of workers' characteristics had remained at their 1973 level (see also Donald, Green, and Paarsch (2000), Gosling, Machin, and Meghir (2000) or Machado and Mata (2005)); and Blau and Kahn (1997) consider how much of the gender wage gap would persist if women had the same observable characteristics as men.

The contribution of our paper is to provide a fully nonparametric method to analyse these kind of questions. In contrast to other methods, we neither impose any parametric restrictions on the relation between Y and X, nor do we restrict attention to the policy's effect on the mean. In particular, our results can be used to conduct inference on the the change of the distribution function as a whole, its moments and quantiles, inequality measures such as the Lorenz curve or Gini coefficient, and to test for stochastic dominance. We show that the cumulative distribution function (CDF) of the counterfactual random variable Y^* is identified under some weak restrictions, and propose a two-stage estimator, that does not rely on any parametric specification of the model (2.1.1) and is easy to implement. In the first stage, we estimate the conditional distribution function of Y given X through nonparametric kernel methods. Secondly, we take a simple average of this estimate evaluated at the observed values of X^* to obtain an estimate of the CDF of Y^*. Too see whether the counterfactual distribution differs from the original one, this result can be compared to an estimate of the distribution function of Y, such as the ordinary empirical CDF. Furthermore, any functional of the two CDFs can in turn be estimated by plugging in the respective estimators. For example, in order to obtain an estimate of the quantile function of Y^*, one can simply invert the estimate of the corresponding CDF.

We also provide a complete asymptotic theory for the estimation procedures proposed in this paper. A key result is that although our method is fully nonparametric, it is not affected by the curse of dimensionality: using empirical process theory, we show that our estimates of the functions of interest converge to certain Gaussian processes at the usual parametric rate $\sqrt{n}$ irrespective of the dimension of X. We can therefore expect

the asymptotics to provide a rather accurate approximation to the finite sample distribution even for moderate sample sizes. A further important result is that the ordinary nonparametric bootstrap works in our framework. This allows us to conduct asymptotically valid uniform inference on the entire functions being estimated, and not just some isolated points. We can thus test a number of important hypothesis involving the whole distribution of Y^* and Y, such as stochastic dominance ordering for example. The use of our methodology is illustrated through an empirical example and an extensive simulation study. The latter shows that our estimators and the related inferential procedures have good finite sample properties, even when the sample size is relatively small. Our approach should thus be appealing to applied researchers.

To the best of our knowledge, our paper is the first to consider estimation and inference for general distributional policy effects in a fully nonparametric framework. As such, it complements and extends an extensive literature on the estimation of policy effects in more restrictive settings. Stock (1989) and Imbens and Newey (2009) consider estimation of policy effects on the *mean* of the outcome variable in a nonparametric framework. DiNardo, Fortin, and Lemieux (1996), Donald, Green, and Paarsch (2000), Gosling, Machin, and Meghir (2000) and Machado and Mata (2005) develop policy estimators for more general distributional effects, but rely on various parametric restrictions of the model in (2.1.1). Furthermore, these papers generally focus on estimation rather than inference, and thus do not provide a full asymptotic theory for their procedures. Chernozhukov, Fernandez-Val, and Melly (2008) derive general limit distribution results for estimators of distributional policy effects, but again only for the case that model (2.1.1) is contained in certain parametric classes. In particular, their arguments critically rely on the assumption that the conditional distribution function of Y given X can be be estimated at a parametric rate, which is clearly not possible in our fully nonparametric framework. While using a correctly specified parametric model will obviously result in efficiency gains compared to our fully nonparametric procedure, such estimators will generally be inconsistent when the respective restrictions are violated. This trade-off is discussed in more detail as part of our simulation study.

In another related paper, Firpo, Fortin, and Lemieux (2009) propose a method to estimate the impact of a *marginal* increase in the covariates on the unconditional distribution

of the outcome variable in a framework similar to ours. This parameter is different from the one being estimated in this paper, which corresponds to the effect of a general and *fixed* change in the distribution of the covariates. Furthermore, they use a very different estimation approach based on a linearization of the outcome distribution.

The outline of the paper is as follows. In the next section, we give a more formal description of our problem, define the parameters of interest, show under which conditions they can be identified, and describe the estimation procedure. Section 3 treats the asymptotic properties of our estimate of the distribution function, and Section 4 shows how these results can be used to analyse a wide range of statistics of the CDF. In Section 5 the practical relevance of our procedure is shown through a simulation study and a small-scale empirical application. Finally, Section 6 concludes. All proofs are collected in the Appendix.

2.2 Modelling Framework and Estimation Approach

2.2.1 Model

The setup we consider is as follows: we observe a dependent variable Y and a d-dimensional vector of covariates X, with marginal distribution functions F_Y and F_X, respectively. The dependent variable is assumed to be generated through the nonseparable model

$$Y = m(X, \varepsilon), \tag{2.2.1}$$

where ε is an unobserved error term. We assume that (2.2.1) is either a structural equation that describes the causal relationship between the right-hand and left-hand side variables, or a reduced form equation from a bigger structural system, as in Ichimura and Taber (2002). In a typical microeconometric application, X and ε would correspond to observed and unobserved characteristics of an individual, respectively, and m would describe the decision rule that, given individual characteristics, determines the individual's choice Y. This flexible formulation allows the covariates to exert influence on Y in manifold ways. For example, model (2.2.1) allows for heteroskedasticity or skewness in the conditional distribution of Y given X. It is fully nonparametric in the sense that we do not restrict

the function m or the distribution of the random variables involved to belong to some parametric family.

The values of at least some components of X are assumed to be under control of a (hypothetical) policy maker, and can thus be shifted exogenously to another observed random vector X^* with associated distribution function F_X^*. Substituting X^* for X in (2.2.1), we obtain the counterfactual random variable

$$Y^* = m(X^*, \varepsilon),$$

whose distribution function we denote by F_Y^*. Our interest is in learning (features of) this distribution and comparing it to that of Y.

For applications, the random vector X^* could originate from a number of different sources. First, X^* could be drawn from a different subpopulation corresponding to a different demographic group, geographic region or time period, like workers' characteristics in a different country for example. Second, X^* could be a deterministic transformation of X, i.e. there exists a known function $\pi : \mathbb{R}^d \to \mathbb{R}^d$ such that $X^* = \pi(X)$. Examples of this case include a public program that causes smoking pregnant women to reduce their daily cigarette consumption by, say, half, or a tuition subsidy that is paid out subject to certain eligibility conditions. Third, X^* could be a repeated measurement on the same individual at a different point in time, as it would typically be the case when the data originate from a panel study.

While the specific source of X^* will not matter for our identification argument or the computation of the estimator, it is useful for the asymptotic development to distinguish those cases where X^* and X are stochastically independent, and those where they are not. We will call the former case an *independent policy implementation*, in the sense that the realization of X^* does not depend on the old value of X. The latter case is then accordingly termed as a *dependent policy implementation*.

2.2.2 Objects of Interest

Depending on the application at hand, a researcher might be interested in learning about different features of the distribution of Y^* and Y. Here we list some useful examples for which estimation and inference is discussed in detail below. However, our framework is by

no means limited to these examples, but can be used for any object that can be written as a sufficiently smooth (in the sense described below) functionals of the distribution functions of Y and Y^*.

Our primary objects of interest are the distribution functions F_Y and F_Y^* themselves. Assuming that the latter is identified, we could give a complete description of the effect of the policy on the distribution of the outcome variable by evaluating them directly. Furthermore, we will consider the pointwise difference between the two distribution functions,

$$\Delta_F(y) = F_Y^*(y) - F_Y(y), \tag{2.2.2}$$

which we call the Distribution Function Policy Effect. This quantity can be of interest in at least two respects. First, in some empirical contexts it might be sufficient to consider the effect of the policy on the distribution of the outcome variable at some fixed point only. In development economics for example, if Y is household income and $\bar{y}$ is some fixed income level defined as the poverty line, one might be interested in whether the policy reduces the fraction of households that live under the poverty line, i.e. whether $\Delta_F(\bar{y})$ is negative. Secondly, it might be interesting to test whether $\Delta_F \equiv 0$, and thus the policy has *any* kind of impact on the distribution of the outcome variable at all.

Instead of looking directly at the CDF, it is often more intuitive to consider the unconditional τ-quantiles Q_Y^* and Q_Y of Y^* and Y, respectively, where

$$Q_Y^*(\tau) = \inf\{y \in \mathbb{R} : F_Y^*(y) \geq \tau\} \tag{2.2.3}$$

and Q_Y is defined analogously. Another convenient summary statistic is the corresponding Quantile Policy Effect, which is defined as

$$\Delta_Q(\tau) = Q_Y^*(\tau) - Q_Y(\tau). \tag{2.2.4}$$

This quantity is analogous to the quantile treatment effect in the literature on program evaluation.

In our framework, it is also possible to analyse the effect of the policy on the Lorenz curve and the Gini coefficient. These measures play an important role in the analysis of inequality and poverty. Formally, for a positive random variable with distribution

function F_Y^* the Lorenz curve $L_Y^*(p)$ is defined as the integral over the quantile function up to p divided by μ, the mean of F:

$$L_Y^*(p) = \frac{1}{\mu_Y^*} \int_0^p Q_Y^*(\tau) d\tau = \frac{1}{\int_0^1 F_Y^{*-1}(\tau) d\tau} \int_0^p F_Y^{*-1}(\tau) d\tau, \tag{2.2.5}$$

The corresponding Gini coefficient is defined as twice the area between the Lorenz curve and the uniform distribution line, i.e.

$$G_Y^* = 1 - 2 \int_0^1 L_Y^*(p) dp, \tag{2.2.6}$$

with $G = 0$ implying perfect equality and $G = 1$ implying perfect inequality. The quantities L_Y and G_Y are then defined analogously. Again, we can also consider the Lorenz Curve Policy Effect, given by

$$\Delta_L(p) = L_Y^*(p) - L_Y(p), \tag{2.2.7}$$

and the Gini Policy Effect,

$$\Delta_G = G_Y^* - G_Y. \tag{2.2.8}$$

The final application we consider in this paper is testing for stochastic dominance. This topic is of great practical importance since the results can be used to evaluate the welfare implications of a proposed policy without making strong assumptions about social preferences. In particular, it is well known that if some distributions can be ranked by stochastic dominance, the same ranking is obtained through the corresponding social welfare over a wide range of utility functions (Atkinson 1970). To simplify the notation, define the operator $\mathcal{D}_j(y, \phi)$ that integrates the function ϕ up to order $j-1$ for $j \geq 1$, i.e.

$$\mathcal{D}_1(y, \phi) = \phi(y), \quad \mathcal{D}_2(y, \phi) = \int_0^y \phi(z) dz, \quad \mathcal{D}_3(y, \phi) = \int_0^y \int_0^t \phi(z) dz dt, \quad \text{etc.} \tag{2.2.9}$$

Then F_Y^* is said to dominate F_Y in "j-th order stochastic dominance" sense if $\mathcal{D}_j(y, F_Y^* - F_Y) \leq 0$ for all possible values of y. A Kolmogorov-Smirnov-type test for stochastic dominance can then be carried out by testing whether $\sup_y \mathcal{D}_j(y, F_Y^* - F_Y)$ is negative.

2.2.3 Identification

For the setup considered in this paper, the only issue is whether the distribution functions of Y and Y^* are identified, since identification of the quantities (2.2.2)–(2.2.9) discussed in the previous subsection then follows trivially. While F_Y is obviously identified by the data, the case of F_Y^* is less clear. Identification in nonseparable models is a potentially delicate issue, that has attracted considerable interest in the recent literature (see for example Chesher (2003), Matzkin (2003), Imbens and Newey (2009) or Hoderlein and Mammen (2007)). However, since we are not interested in directly identifying features of m, our problem is much less complicated. In particular, there is no need to impose anymore structure on the function m, such as monotonicity in the unobservables. For identification of our quantity of interest, the following assumption suffices.

Assumption 1 (Identification)**.** *(i) The term ε is independent of both X and X^*. (ii) The support of X^* is a subset of the support of X.*

Proposition 1. *Under Assumption 1, $F_Y^*(y) = \mathbb{E}(F_{Y|X}(y, X^*))$ and is thus identified.*

The first part of Assumption 1 implies that X is exogenous, which is a strong assumption for many applications. It is straightforward to relax this condition by requiring independence to hold only conditional on some other random variable V, which is either observed or estimateable. Such a control variable could be available in a wide range of contexts, as described in Imbens and Newey (2009) for example. These include treatment effect models with selection on observables, triangular simultaneous equation models and certain sample selection models. In all cases, one can extend the result of Proposition 1 to $F_Y^*(y) = \mathbb{E}(F_{Y|X,V}(y, X^*, V))$. Our results on estimation given below then apply immediately when V is observed. The case where V has to be estimated (nonparametrically) from the data is technically much more involved and beyond the scope of this paper.

The second part of Assumption 1 restricts the policy experiments that can be considered to ones for which there is already some experience in the data. This restriction is due to the inability of nonparametric estimators to extrapolate from range of actual observations. While it limits the potential fields of application, without imposing some parametric structure on m this condition seems necessary to obtain point identification

of F_Y^*. However, it is possible to give meaningful bounds on the CDF of Y^* when X^* is allowed to take values outside of the support of X with moderate probability.

2.2.4 Estimation

Our estimation approach is to first construct estimates $\hat{F}_Y^*$ and $\hat{F}_Y$ of the distribution functions F_Y^* and F_Y, respectively, and then estimate any functional of the form $\Gamma = \Gamma(F_Y^*, F_Y)$ through the plug-in method by $\hat{\Gamma} = \Gamma(\hat{F}_Y^*, \hat{F}_Y)$. For example, an estimate of the Quantile Policy Effect Δ_Q can be constructed as

$$\hat{\Delta}_Q(\tau) = \hat{Q}_Y^*(\tau) - \hat{Q}_Y(\tau) = \inf\{y \in \mathbb{R} : \hat{F}_Y^*(y) \geq \tau\} - \inf\{y \in \mathbb{R} : \hat{F}_Y(y) \geq \tau\}.$$

Estimates of all objects of interest defined in eq. (2.2.2)–(2.2.8) can be defined in an analogous manner.

The structure of the data used for the estimation depends on whether we are considering a *dependent* or an *independent policy implementation*. In the former case, the data consist of a sample $\{(Y_i, X_i, X_i^*)\}_{i=1}^n$ of size n from the distribution of (Y, X, X^*). For an independent policy implementation, the data consist of a sample $\{(Y_i, X_i)\}_{i=1}^n$ of size n from the distribution of (Y, X), and a sample $\{X_i^*\}_{i=1}^{n^*}$ of size n^* from the distribution of X^*. While we allow the two samples sizes to be different in this case, we assume for the later asymptotic analysis that they increase proportionally, so that $n^* = n/\lambda$ for some constant λ.

We now turn to the construction of the estimators. Throughout, we will use the notation for an independent policy implementation without loss of generality, since it covers the dependent implementation as the special case with $n^* = n$. Starting with an estimate for the CDF of Y, an obvious candidate is the usual empirical cumulative distribution function (ECDF),

$$\hat{F}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i \leq y\},$$

whose theoretical properties are well-known in the literature. To derive an estimator for F_Y^*, we know from the identification argument in the previous subsection that under Assumption 1

$$F_Y^*(y) = \mathbb{E}(F_{Y|X}(y, X^*)).$$

Following the analogy principle and replacing unknown quantities with suitable sample counterparts, it appears intuitive to use an estimator $\hat{F}_Y^*$ of the form

$$\hat{F}_Y^*(y) = \frac{1}{n^*} \sum_{i=1}^{n^*} \hat{F}_{Y|X}(y, X_i^*),$$

where $\hat{F}_{Y|X}$ is a first-stage nonparametric estimate of the distribution function of Y conditional on X. If all covariates are continuously distributed, we propose to estimate this function by a Nadaraya-Watson-type estimator, i.e.

$$\hat{F}_{Y|X}(y, x) = \frac{\hat{g}_{YX}(y, x)}{\hat{f}_X(x)}$$

where

$$\hat{g}_{YX}(y, x) = \frac{1}{n} \sum_j \mathbb{I}\{Y_j \leq y\} K_{x,h}(X_j - x),$$
$$\hat{f}_X(x) = \frac{1}{n} \sum_j K_{x,h}(X_j - x).$$

Here $\mathbb{I}\{A\}$ is an indicator function that equals one if A is true and zero otherwise, $h = h_n$ is a bandwidth sequence that tends to zero as $n \to \infty$, $K_{x,h}(\cdot) = h^{-d} K_x(\cdot/h)$, and K_x is a higher-order boundary kernel, i.e. a kernel function whose moments up to a certain order are zero, and whose shape adapts when the point of evaluation x is in the vicinity of the boundary of the support of X (see for example Gasser, Müller, and Mammitzsch (1985)). These properties are needed to derive a uniform rate of convergence for our first-step estimator later, uniformly over the *entire* support of Y and X. We will be more precise about the specifics below.

The estimator $\hat{F}_{Y|X}$ can easily be generalized to admit discrete regressors using the conventional frequency method. This entails splitting the sample into subsets, or cells, and then calculating the Nadaraya-Watson estimator within each such subset separately. This procedure is well known to have no effect on the rate of convergence. For notational convenience, we will therefore maintain the assumption that all covariates are continuously distributed.

A computational advantage of our estimator is that it admits a representation as a reweighted version of the usual empirical distribution function. To see this, note that

$\hat{F}^*_Y(y)$ can be written as

$$\begin{aligned}\hat{F}^*_Y(y) &= \frac{1}{n^*}\sum_{i=1}^{n^*}\frac{\sum_{j=1}^n \mathbb{I}\{Y_j \leq y\}K_{X_i^*,h}(X_j - X_i^*)}{\sum_{l=1}^n K_{X_i^*,h}(X_l - X_i^*)} \\ &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\{Y_j \leq y\}w_j\end{aligned}$$

where the weights w_j are given by

$$w_j = \lambda\sum_{i=1}^{n^*}\frac{K_{X_i^*,h}(X_j - X_i^*)}{\sum_{l=1}^n K_{X_i^*,h}(X_l - X_i^*)}.$$

Since the weights do not depend on y, they have to be calculated only once even when $\hat{F}^*_Y$ is evaluated at multiple locations, making the estimator extremely cheap to compute in practice.

A potential caveat when using higher-order kernels is that they are not restricted to be positive, and hence can lead to estimates of F^*_Y which are non-monotone or take values ouside the unit interval in finite samples, which is of course undesirable. This problem can be circumvented by a slight modification of the estimator: if we replace the weights w_j by $\tilde{w}_j = w_j\mathbb{I}\{w_j \geq 0\}/\sum_i(w_i\mathbb{I}\{w_i \geq 0\})$, we obtain a modified estimator

$$\tilde{F}^*_Y(y) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{I}\{Y_j \leq y\}\tilde{w}_j$$

which is constrained to be monotonically increasing and bounded between 0 and 1. We show in the following section that this estimator asymptotically equivalent to $\hat{F}^*_Y$ under standard conditions. For the further analysis, we will therefore assume without loss of generality that there are no issues with non-monotonicity of $\hat{F}^*_Y$.

2.3 Asymptotic Properties

In order to conduct inference on the CDFs and related functionals as a whole, we have to derive the joint asymptotic properties of the two estimators not only at some fixed value, but as a random function. To do so, we first state the assumptions and give some useful preliminary results in the following subsection, before proving the main weak convergence theorem in the next but one. Finally, we discuss inference and the validity of the bootstrap.

2.3.1 Assumptions and Preliminaries

To present our framework, we first have to introduce some notation. For μ a k-vector of nonnegative integers, we define (i) $|\mu| = \sum_{i=1}^{k} \mu_i$, (ii) for any function $\phi(x)$ on $\mathbb{R}^k$, $\partial_x^\mu \phi(x) = \partial^{|\mu|}/(\partial^{\mu_1} x_1, \ldots, \partial^{\mu_k} x_k)\phi(x)$ and (iii) $x^\mu = \prod_{i=1}^{k} x_i^{\mu_i}$. We write "$\xrightarrow{d}$" to denote convergence in distribution of a sequence of random variables, and "$\Rightarrow$"to denote weak convergence of a sequence of random functions.

To prove our main results, we need the following assumptions.

Assumption 2. *The data $\{(Y_i, X_i)\}, i = 1, \ldots, n$ and $\{X_j^*\}, j = 1, \ldots, n^*$ are i.i.d., respectively.*

Assumption 3. *(i) The support of X and X^* are the compact sets $J = \otimes_{i=1}^{d}[\underline{x}_i, \overline{x}_i]$ and $J^* = \otimes_{i=1}^{d}[\underline{x}_i^*, \overline{x}_i^*] \subset J$, respectively. (ii) X has a probability density function $f_X(x)$, which is bounded away from zero on J. (iii) X^* has a probability density function $f_X^*(x)$, which is bounded away from zero on J^*. (iv) The functions $f_X(x)$ and $g(y, x) = F_{Y|X}(y, x)f_X(x)$ are r-times differentiable with respect to x on the interior of J, and the derivatives are uniformly continuous and bounded. (v) The function $f_X^*(x)$ is r-times differentiable with respect to x on the interior of J^*, and the derivatives are uniformly continuous and bounded.*

Assumption 4. *Let $D(c) = \{z : \underline{x} - c \leq z \leq c - \overline{x}\}$. Then the kernel function $K_c : \mathbb{R}^d \to \mathbb{R}$ satisfies (i) $\int_{D(c)} K_c(z)dz = 1$, (ii) $\int_{D(c)} K_c(z)z^\mu dz = 0$ for all $|\mu| = 1, \ldots, r$, (iii) $\int_{D(c)} |K_c(z)z^\mu|dz < \infty$ for $|\mu| = r$, (iv) $K_c(z) = 0$ if $|z| > 1$ (v) $K_c(z)$ is r-times differentiable with respect to both z and c, and the derivatives are uniformly continuous and bounded.*

Assumption 5. *The bandwidth sequence $h = h_n$ satisfies (i) $h \to 0$, (ii) $n^{1/2}h^d/\log(n) \to \infty$ and (iii) $n^{1/2}h^r \to 0$.*

Assumption 2 is standard in microeconometric applications. Assumptions 3 collects conventional smoothness restrictions on the functions being estimated through nonparametric methods at some point in this paper. Note that it implicitly restricts the policies that can be considered to those where both X and X^* are continuously distributed. It is straightforward to show that this condition can be replaced by the assumption that

both random vectors have a probability density function with respect to the same dominating measure. This would allow their components to be discrete or even continuous with some mass points, as long as the policy does not affect the location of the mass point. However, all policies for which X^* has a probability density function with respect to a different dominating measure than X are excluded in this framework. The fourth assumption prescribes the use of a higher-order boundary kernel, which is required to remove asymptotic bias from our first-step estimator. Note that the boundary correction is not necessary when J^* is a compact subset of the *interior* of J. Finally, the last assumption determines the rate at which the bandwidth sequence converges to zero. If h is of the form $h = cn^{-\delta}$ for some constants $c, \delta > 0$, then in order for Assumption 6 to hold we need that $\delta \in (1/2r, 1/2d)$ which in turn requires that the order of the kernel exceeds the dimension of X, so that the interval is not empty.

These assumptions are convenient, because they allow us to prove the following proposition, which is an important ingredient for the further arguments. Similar results have been obtained by Härdle, Janssen, and Serfling (1988) and Newey (1994a), to mention a few.

Proposition 2. *Under Assumption 1-5, we have that*

$$
\begin{aligned}
&i) \quad \sup_{y\in\mathbb{R}} \sup_{x\in J} |\hat{g}_{YX}(y,x) - g_{YX}(y,x)| = O_p\left(\left(\frac{\log n}{nh^d}\right)^{1/2} + h^r\right) \\
&ii) \quad \sup_{x\in J} |\hat{f}_X(x) - f_X(x)| = O_p\left(\left(\frac{\log n}{nh^d}\right)^{1/2} + h^r\right) \\
&iii) \quad \sup_{y\in\mathbb{R}} \sup_{x\in J} |\hat{F}_{Y|X}(y,x) - F_{Y|X}(y,x)| = O_p\left(\left(\frac{\log n}{nh^d}\right)^{1/2} + h^r\right)
\end{aligned}
$$

The proposition provides an explicit uniform rate of convergence for the first-step estimates. In particular, under Assumption 5 the proposition implies that the difference between $\hat{F}_{Y|X}$ and $F_{Y|X}$ vanishes at a rate faster than $n^{-1/4}$, whereas the corresponding bias disappears faster than $n^{-1/2}$, both uniformly over the two arguments. Thus our first-stage nonparametric estimator satisfies the well-known minimal convergence rates given by Newey (1994b).

Another important preliminary result is the asymptotic equivalence of the estimator $\hat{F}_Y^*$ and its modified version $\tilde{F}_Y^*$ introduced in Section 2.4. The following proposition

implies that the limit results derived for $\hat{F}_Y^*$ in the following section will also apply to $\tilde{F}_Y^*$, which has the practical advantage of being a proper distribution function.

Proposition 3. *Under Assumption 1-5, we have that*

$$\sup_{y\in\mathbf{R}} |\hat{F}_Y^*(y) - \tilde{F}_Y^*(y)| = o_p(n^{-1/2}).$$

2.3.2 Main Result

In this section, we derive the limit behaviour of our estimates of the distribution functions of Y^* and Y. In particular, we show that the bivariate random function

$$y \mapsto \sqrt{n}\left(\hat{\mathbf{F}}(y) - \mathbf{F}(y)\right) \tag{2.3.1}$$

converges weakly to some Gaussian process, where we will use the notation that $\hat{\mathbf{F}} = (\hat{F}_Y^*, \hat{F}_Y)^T$, $\mathbf{F} = (F_Y^*, F_Y)^T$ and $y = (y_1, y_2)^T$.

The main complication for deriving this result originates from the process' first component, the normalized estimate of the CDF of Y^*, which is given by

$$\sqrt{n}(\hat{F}_Y^*(y_1) - F_Y^*(y_1)) = \sqrt{n}\left(\frac{1}{n^*}\sum_{i=1}^{n^*} \hat{F}_{Y|X}(y_1, X_i^*) - F_Y^*(y_1)\right).$$

The properties of this expression are not straightforward to derive, since our estimator $\hat{F}_Y^*(y_1)$ is not a sum of independent terms: $\hat{F}_{Y|X}(y_1, X_i^*)$ does not only depend on the ith but on all observations. In the appendix, we therefore construct an asymptotic representation for our estimate, which decomposes $\hat{F}_Y^*$ into the following three parts:

$$\begin{aligned}\sqrt{n}(\hat{F}_Y^*(y_1) - F_Y^*(y_1)) = &\frac{1}{\sqrt{n^*}}\sum_{i=1}^{n^*} \sqrt{\lambda}\left(F_{Y|X}(y_1, X_i^*) - F_Y^*(y_1)\right) \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{f_X^*(X_i)}{f_X(X_i)}(\mathbb{I}\{Y_i \le y_1\} - F_{Y|X}(y_1, X_i)) + o_p(1).\end{aligned}$$

The first term on the right hand side is the one we would obtain if the function $F_{Y|X}$ was known and did not have to be estimated from the data.[1] It accounts for the uncertainty in our estimate that is induced by replacing the expectation with a sample average. The second term is an adjustment term that accounts for the uncertainty in our estimate of

[1]The factor $\sqrt{\lambda}$ is an artefact of scaling the estimator by $\sqrt{n}$ instead of $\sqrt{n^*}$.

$F_{Y|X}$. The last term is $o_p(1)$ uniformly in y and thus asymptotically negligible.[2] Using this decomposition and the definition of the ECDF, we arrive at the following Theorem:

Theorem 1. *If Assumtions 1-5 hold, then*

$$\sqrt{n}\left(\hat{\mathbf{F}}(\cdot) - \mathbf{F}(\cdot)\right) \Rightarrow \mathbb{F}_o(\cdot)$$

where $\mathbb{F}_o$ is a two dimensional Gaussian process with mean zero and covariance function

$$\Psi^F(y, y') = \mathbb{E}(\psi^F(y, Z)\psi^F(y', Z)^T),$$

where $Z = (Y, X, X^*)$ and $\psi^F(y, Z) = \left(\psi_1^F(y_1, Z), \psi_2^F(y_2, Z)\right)^T$ is given by

$$\psi_1^F(y_1, Z) = \sqrt{\lambda}(F_{Y|X}(y_1, X^*) - F_Y^*(y_1)) + \frac{f_X^*(X)}{f_X(X)}(\mathbb{I}\{Y \leq y_1\} - F_{Y|X}(y_1, X)),$$
$$\psi_2^F(y_2, Z) = \mathbb{I}\{Y \leq y_2\} - F_Y(y_2),$$

and the convergence is in $D(-\infty, \infty) \times D(-\infty, \infty)$.

The theorem shows that in large samples the behaviour of our random function (2.3.1) can be approximated by a bivariate correlated Gaussian process, whose second component is easily seen to be an ordinary F_Y-Brownian Bridge. The explicit form of the covariance function Ψ^F depends on whether we consider a dependent or independent policy implementation, but the influence function is the same in both cases. Although our estimator depends in part on high-dimensional nonparametric components, we obtain the $\sqrt{n}$-rate of convergence that one would typically obtain for standard parametric estimators, irrespective of the dimension of X. Our estimates are thus not affected by the curse of dimensionality, and hence we can expect the asymptotics to be a rather accurate approximation even when the sample size is only moderate relative to the dimension of X.

An immediate implication of Theorem 1 is that our estimator of the Distribution Function Policy Effect converges to a Gaussian process as well. That is, we obtain that

$$\sqrt{n}(\hat{\Delta}_F(\cdot) - \Delta_F(\cdot)) \Rightarrow (1, -1)\mathbb{F}_o(\cdot, \cdot).$$

[2]For the case of a fixed y, a similar decomposition for averages of nonparametrically estimated functions is shown by Newey (1994a).

by simply applying the continuous mapping theorem (CMT). In order to analyse the properties of estimates of more general functionals of the form $\Gamma(\hat{\mathbf{F}})$, one can use the following Theorem, which is an application of the Functional Delta Method (van der Vaart 2000, Theorem 20.8):

Theorem 2. *Suppose that the conditions of Theorem 1 hold, and let* Γ *be a Hadamard differentiable functional mapping from* $D(-\infty,\infty) \times D(-\infty,\infty)$ *to some normed space* S*, with derivative* $\Gamma'_{\mathbf{F}}$*. Then*

$$\sqrt{n}\left(\Gamma(\hat{\mathbf{F}})(\cdot) - \Gamma(\mathbf{F})(\cdot)\right) \Rightarrow \Gamma'_{\mathbf{F}}(\mathbb{F}_o)(\cdot) \equiv \mathbb{G}_o(\cdot),$$

where $\mathbb{G}_o$ *is a Gaussian process with mean zero and covariance function*

$$\Psi^{\Gamma}(y,y') = \mathbb{E}(\psi^{\Gamma}(y,Z)\psi^{\Gamma}(y',Z)^T)$$

with $\psi^{\Gamma}(y,Z) = \Gamma'_{\mathbf{F}}(\psi^F)(y,Z)$*, and the convergence is in* $S \times S$*.*

The Hadamard differentiability condition in Theorem 2 requires the functional of interest to be sufficiently smooth around the true value $\mathbf{F}$. Roughly speaking, this means that Γ can locally be well approximated by some continuous linear functional $\Gamma'_{\mathbf{F}}$, in the sense that

$$\frac{\Gamma(\mathbf{F}+ts_t)-\Gamma(\mathbf{F})}{t} \to \Gamma'_{\mathbf{F}}(s) \quad \text{as } t \to 0$$

for all functions $s_t \to s$ (see van der Vaart (2000, p. 296) for a precise definition). As we will see below, this condition can be verified for all our applications of interest under mild additional conditions. Also note that the theorem allows for functionals Γ that map into $\mathbb{R}^k$ instead of some function space. In this case $\mathbb{G}_o$ is simply a k-variate normal distribution, and $\Psi^{\Gamma}(y,y') \equiv \Psi^{\Gamma}$ is its covariance matrix.

2.3.3 Inference

The results in Theorem 1 and 2 immediately provide the basis for conducting pointwise inference on certain features of the counterfactual distribution, by using the standard normal approximation. This might already be sufficient in some empirical contexts. In development economics for example, if Y is household income and $\bar{y}$ is some fixed income

level defined as the poverty line, one could simply be interested in whether the policy reduces the fraction of households that live under the poverty line, i.e. whether $\Delta_F(\bar{y})$ is negative. Then it follows from the above results that the corresponding estimate $\hat{\Delta}_F(\bar{y})$ is asymptotically normal with mean zero and standard error $\sqrt{(1,-1)\Psi^F(\bar{\mathbf{y}},\bar{\mathbf{y}})(1,-1)^T/n}$, where $\bar{\mathbf{y}} = (\bar{y},\bar{y})$. Given a consistent estimate of the covariance function, one could thus test a null hypothesis such as H_0: "The policy does not change the proportion of people that earn below $\bar{y}$" through an ordinary t-statistic using standard normal critical values.

Many important hypotheses, however, cannot adequately be tested by considering only a fixed number of isolated points. This includes the hypothesis that the policy has no effect whatsoever, or that it leads to an improvement in a stochastic dominance sense. A related problem that involves the entire functions being estimated is the construction of uniform confidence bands, that cover the true function with some prespecified probability.

One possibility to address these problems would be to simulate the limiting processes in Theorem 1 and 2 by so-called multiplier methods (see e.g. van der Vaart and Wellner (1996, Section 2.9)). A disadvantage of this approach is that it requires explicit calculation and consistent estimation of the covariance function, which can be a cumbersome task for some applications. A natural alternative to using simulation methods is to conduct inference using a form of the bootstrap, for which one does not necessarily need to be able to give an explicit characterization of the limiting distribution of the process of interest. In particular, by using the bootstrap one can circumvent explicit specification of the covariance function.

In this paper, we propose using a simple nonparametric (or empirical) bootstrap scheme, which is based on resampling the original observations. For the implementation, we again have to take possible dependencies between X and X^* into account. For a dependent policy implementation, the bootstrap data is given by a sample $\{(Y_{b,i}, X_{b,i}, X^*_{b,i})\}_{i=1}^n$ drawn with replacement from $\{(Y_i, X_i, X^*_i)\}_{i=1}^n$, whereas for an independent policy implementation it consists of two samples $\{(Y_{b,i}, X_{b,i})\}_{i=1}^n$ and $\{X^*_{b,i}\}_{i=1}^{n^*}$ drawn with replacement from $\{(Y_i, X_i)\}_{i=1}^n$ and $\{X^*_i\}_{i=1}^{n^*}$, respectively. In both cases, the resampled data is then used to calculate the bootstrap estimates of F^*_Y and F_Y, which are denoted by $\hat{\mathbf{F}}_\mathbf{b} = (\hat{F}^*_{Y,b}, \hat{F}_{Y,b})$, using the estimator described in Section 2.4. The distribution of $\hat{\mathbf{F}}_\mathbf{b}$ can then be determined through the usual repeated resampling of the data, and used as

an approximation of the distribution of $\hat{\mathbf{F}}$. The following Theorem gives a theoretical justification for this approach.

Theorem 3. *Suppose that the conditions of Theorem 1 hold. Then*

$$\sqrt{n}\left(\hat{\mathbf{F}}_{\mathbf{b}}(\cdot) - \hat{\mathbf{F}}(\cdot)\right) \Rightarrow \mathbb{F}_o(\cdot),$$

conditional on the data, in probability, and the convergence is in $D(-\infty,\infty)\times D(-\infty,\infty)$. *Furthermore, under the conditions of Theorem 2,*

$$\sqrt{n}\left(\Gamma(\hat{\mathbf{F}}_{\mathbf{b}})(\cdot) - \Gamma(\hat{\mathbf{F}}(\cdot))\right) \Rightarrow \Gamma'_{\mathbf{F}}(\mathbb{F}_o)(\cdot),$$

conditional on the data, in probability[3]*, and the convergence is in* $S \times S$.

Theorem 3 states that the nonparametric bootstrap is not only valid for the original problem of conducting inference on the estimates of the CDFs directly, but that it can also be used to analyse the properties of general functionals of the CDFs as well. As a simple example, consider the problem of forming a uniform $1-\alpha$ confidence band for the Distribution Function Policy Effect $\Delta_F(\cdot)$. To this end, let $\hat{\Delta}_{F,b}(y) = \hat{F}^*_{Y,b}(y) - \hat{F}_{Y,b}(y)$ the bootstrapped Distribution Function Policy Effect, and define the pointwise variance of $\hat{\Delta}_{F,b}$ with respect to bootstrap sampling as $\sigma^2(y) = \mathrm{Var}_b(\hat{\Delta}_{F,b}(y))$. It then follows directly from Theorem 3 that a uniform $1-\alpha$ confidence band for Δ_F is given by

$$CB_{1-\alpha}(y) = [\hat{\Delta}_F(y) - c\sigma(y), \hat{\Delta}_F(y) + c\sigma(y)],$$

where c is the smallest positive constant that satisfies

$$P_b\left(\sup_y \left|(\hat{\Delta}_{F,b}(y) - \hat{\Delta}_F(y))\right| / \sigma(y) \leq c\right) \geq 1-\alpha \tag{2.3.2}$$

and P_b is the probability with respect to bootstrap sampling. In practice, the unknown quantities c and $\sigma^2(\cdot)$ can be approximated with the usual resampling techniques.

2.4 Application to Objects of Interest

In this section, we use results from Theorem 1 – 3 do discuss the remaining applications of interest from Section 2.2. We show that in each case plug-in type estimators con-

[3]See van der Vaart and Wellner (1996, Section 3.9.3) for a precise definition of conditional weak convergence in probability.

verge to a Gaussian limit, and that the nonparametric bootstrap can be used to conduct asymptotically valid inference, under appropriate additional regularity conditions.

2.4.1 Quantiles

To analyse the properties of the estimators of quantiles and Quantile Policy Effect, we use the fact that inversion operator that transforms a CDF into its corresponding quantile function is a Hadamard differentiable functional, which gives us the following proposition:

Proposition 4. *Assume that (i) F_Y^* and F_Y are both continuously differentiable with strictly positive derivative f_Y^* and f_Y, and (ii) Y^* and Y have compact support, then*

$$\sqrt{n}(\hat{\mathbf{Q}}(\cdot) - \mathbf{Q}(\cdot)) \Rightarrow -\left(\frac{\mathbb{F}_{o1}}{f_Y^*}, \frac{\mathbb{F}_{o2}}{f_Y}\right) \circ \mathbf{Q} \equiv \mathbb{Q}_o$$

and

$$\sqrt{n}(\hat{\mathbf{Q}}_{\mathbf{b}}(\cdot) - \hat{\mathbf{Q}}(\cdot)) \Rightarrow \mathbb{Q}_o(\cdot)$$

conditional on the data, in probability. Here $\mathbb{Q}_o$ is a Gaussian process with mean zero and covariance function

$$\Psi^Q(\tau, \tau') = \mathbb{E}(\psi^Q(\tau, Z)\psi^Q(\tau, Z)^T)$$

where

$$\psi^Q(\tau, Z) = \left(\frac{\psi_1^F(Q_Y^*(\tau_1), Z)}{f_Y^*(Q_Y^*(\tau_1))}, \frac{\psi_2^F(Q_Y(\tau_2), Z)}{f_Y(Q_Y(\tau_2))}\right)^T.$$

and the convergence is in $\ell^\infty(0,1) \times \ell^\infty(0,1)$.

We thus have the familiar result that the influence function of the quantile process is just the influence function of the corresponding distribution function divided by the density of the variable of interest, and evaluated at the respective quantile. The assumption that Y^* and Y have compact support can be relaxed at the cost of restricting convergence to subsets of the unit interval. Furthermore, it follows directly from the proposition and the continuous mapping theorem that our estimator of the Quantile Policy Effect satisfies

$$\sqrt{n}(\hat{\Delta}_Q(\cdot) - \Delta_Q(\cdot)) \Rightarrow (1, -1)\mathbb{Q}_o(\cdot),$$

and that the nonparametric bootstrap is valid in this case as well. Thus, one can construct uniform confidence bands on the Quantile Policy Effect in exactly the same manner as for the Distribution Function Policy Effect.

2.4.2 Inequality measures

In this section, we apply Theorem 1–3 to analyse the effect of a proposed policy on the Lorenz curve and the Gini coefficient. Our approach is similar to that of Barrett and Donald (2000) and Bhattacharya (2007), who in different contexts also obtained weak convergence results using the functional delta method. Many other inequality measures such as the Theil index for example can be treated by analogous arguments. We start by deriving the asymptotic properties of the estimated Lorenz curves.

Proposition 5. *Assume that (i) F_Y^* and F_Y are both continuously differentiable with derivatives f_Y^* and f_Y, respectively, (ii) these derivatives are strictly positive on any compact subset of $(0,\infty)$, (iii) Y^* and Y have finite second moments, and (iv) it holds that*

$$\lim_{y\to\infty}\frac{(1-F_Y(y))^{1+c}}{f_Y(y)}=\lim_{y\to 0}\frac{F_Y(y)^{1+c}}{f_Y(y)}=0$$

for some $0<c<1$, and similarly for Y^*. Furthermore, define the process $\mathbb{H}_o$ as

$$\mathbb{H}_o(p)=\int_0^p \mathbb{Q}_o(\tau)d\tau$$

for $p=(p_1,p_2)^T$, and the integral is understood to be taken componentwise. Then

$$\sqrt{n}\left(\hat{\mathbf{L}}(\cdot)-\mathbf{L}(\cdot)\right)\Rightarrow\left(\frac{\mathbb{H}_{o1}(\cdot)}{\mu_Y^*}-\frac{L_Y^*(\cdot)}{\mu_Y^*}\mathbb{H}_{o1}(1),\frac{\mathbb{H}_{o2}(\cdot)}{\mu_Y}-\frac{L_Y(\cdot)}{\mu_Y}\mathbb{H}_{o2}(1)\right)\equiv\mathbb{L}_o(\cdot),$$

and

$$\sqrt{n}\left(\hat{\mathbf{L}}_{\mathbf{b}}(\cdot)-\hat{\mathbf{L}}(\cdot)\right)\Rightarrow\mathbb{L}_o(\cdot),$$

conditional on the data, in probability. Here $\mathbb{L}_o$ is a Gaussian process with mean zero and covariance function

$$\Psi^L(p,p')=\mathbb{E}(\psi^L(p,Z)\psi^L(p',Z)^T)$$

where $\psi^L(p,Z)=(\psi_1^L(p_1,Z),\psi_2^L(p_2,Z))^T$ is given in the Appendix, and the convergence is in $\ell^\infty(0,1)\times\ell^\infty(0,1)$.

The additional assumptions we make are used by Bhattacharya (2007) in order to establish Hadamard-differentiability of the functional that translates a CDF into its Lorenz

curve. Note that in contrast to the quantile process, one does not have to assume that the support of Y and Y^* is compact, and thus that their density functions are bounded away from zero, to obtain weak convergence in $\ell^\infty(0,1) \times \ell^\infty(0,1)$. Instead, the tail condition (iv), which is fulfilled by many distributions commonly used to describe income distributions such as log-normal or Pareto, suffices.

Proposition 4 could again be used to conduct inference on the Lorenz curve as a whole. Also, as in the case of the quantiles considered above, it follows from the continuous mapping theorem that our estimator of the Lorenz Policy Effect satisfies

$$\sqrt{n}(\hat{\Delta}_L(\cdot) - \Delta_L(\cdot)) \Rightarrow (1,-1)\mathbb{L}_o(\cdot),$$

and that the nonparametric bootstrap is valid in this case as well. Here the bootstrap could be used for example to construct a one-sided confidence band $\hat{\Delta}_L$, which could be used to test the hypothesis of Lorenz dominance, i.e. that $\Delta_L(p) \leq 0$ for all $p \in (0,1)$. Another direct consequence of the proposition is the distribution of the Gini coefficient, which again follows simply from the continuous mapping theorem.

Proposition 6. *Under the same conditions as Proposition 4,*

$$\sqrt{n}(\hat{\mathbf{G}} - \mathbf{G}) \xrightarrow{d} N(0, \Psi^G) \quad \textit{and} \quad \sqrt{n}(\hat{\mathbf{G}}_\mathbf{b} - \hat{\mathbf{G}}) \xrightarrow{d} N(0, \Psi^G),$$

conditional on the sample, in probability, where the limiting distribution is bivariate normal with mean zero and covariance matrix

$$\Psi^G = 4\mathbb{E}\left(\int_0^1 \psi^L(p,Z)dp \int_0^1 \psi^L(p,Z)^T dp\right).$$

2.4.3 Testing for Stochastic Dominance

The limit results from Section 3 can also be used for various testing problems. Here we adapt the methods of Barrett and Donald (2003) and consider Kolmogorov-Smirnov-type statistics to test stochastic dominance of F_Y^* over F_Y for any prespecified order, with critical values obtained via the bootstrap. Other approaches are discussed, for example, by McFadden (1989), Anderson (1996), Davidson and Duclos (2000), Abadie (2002) and Linton, Maasoumi, and Whang (2005).

The test statistics we consider are based on pointwise comparisons of appropriate measures of distance between $\hat{F}_Y^*$ and $\hat{F}_Y$ over their entire common support, which we assume to be the compact interval $[0, \bar{y}]$[4]. Using the operator $\mathcal{D}_j(y, \phi)$ defined in (2.9), the hypothesis of "j-th order stochastic dominance" of F_Y^* over F_Y can then be formulated as

$$\begin{aligned} H_0^j: &\quad \mathcal{D}_j(y, F_Y^* - F_Y) \leq 0 \quad \forall y \in [0, \bar{y}], \\ H_1^j: &\quad \mathcal{D}_j(y, F_Y^* - F_Y) > 0 \quad \exists y \in [0, \bar{y}], \end{aligned}$$

for $j = 1, 2, \ldots$. The corresponding test statistics are given by

$$KS_j = \sqrt{n} \sup_{y \in [0, \bar{y}]} \mathcal{D}_j(y, \hat{F}_Y^* - \hat{F}_Y) = \sqrt{n} \max_{y \in Y_1, \ldots Y_n} \mathcal{D}_j(y, \hat{F}_Y^* - \hat{F}_Y),$$

where the last equality follows from the fact that by construction both $\hat{F}_Y^*$ and $\hat{F}_Y$ are piecewise constant functions that jump at observed values of Y only.

Our aim is to reject H_0^j whenever KS_j exceeds some critical value. Note that there can be many different combinations of distribution functions F_Y^* and F_Y such that H_0^j is true, and we are thus testing a composite null hypothesis. However, it is easy to see that the least favourable case in this context corresponds to $F_Y^* = F_Y$. An asymptotically valid test procedure can thus be based on bootstrapping the test statistic under the least favourable null. In particular, one can calculate the bootstrap p-value as

$$\hat{p}_j = B^{-1} \sum_{b=1}^{B} \mathbb{I}\{\hat{KS}_{j,b} > KS_j\},$$

where

$$\hat{KS}_{j,b} = \sqrt{n} \max_{y \in Y_1, \ldots Y_n} \Big(\mathcal{D}_j(y, \hat{F}_{Y,b}^* - \hat{F}_{Y,b}) - \mathcal{D}_j(y, \hat{F}_Y^* - \hat{F}_Y)) \Big)$$

is the realisation of the test statistic when calculated from the bootstrap sample. Our test decision can then be based on the following rule:

$$\text{Reject } H_0^j \text{ if } \hat{p}_j < \alpha \text{ for some prespecified significance level } \alpha. \tag{2.4.1}$$

The following proposition delivers a theoretical justification for this approach.

[4]Focussing on random variables that take only positive values seems natural, since stochastic dominance tests are usually applied to income or wealth distributions. The upper limit on the support is needed for the proofs and is usually not restrictive for empirical applications.

Proposition 7. *For any $j = 0, 1, 2, \ldots$ and $\alpha < 1/2$, the decision rule* (2.4.1) *is a test of H_0^j vs. H_1^j that has (i) asymptotic size of at most α and (ii) is consistent against any fixed alternative.*

2.5 Numerical Examples

2.5.1 Simulation Study

Setup

In order to demonstrate the usefulness of our proposed estimation procedures, we conduct a number of simulation experiments to assess their finite sample properties. Specifically, we simulate a vector $X = (X_1, X_2, X_3)$ of conditioning variables, where the three components are i.i.d. standard exponentially distributed and truncated at 3, and generate the dependent variable of interest Y through a linear model with conditional heteroskedasticity as

$$Y = 6 - 2X_1 + X_2 + \sigma(X)\varepsilon, \qquad \sigma^2(X) = X_1 + X_2, \tag{2.5.1}$$

and ε follows a standard exponential distribution. Note that Y is restricted to be positive in this setup, and that X_3 is an irrelevant regressor. We consider two dependent policy implementations, where X^* is a determininistic transformation of the orginal X-values:

- *Policy 1:* X_3 is reduced by 50%: $\pi_1(x_1, x_2, x_3) = (x_1, x_2, .5x_3)$.
- *Policy 2:* All regressors are reduced by 50%: $\pi_2(x_1, x_2, x_3) = .5(x_1, x_2, x_3)$.

Since X_3 does not appear in the data generating process of Y, the first policy has no effect on the distribution of Y. In contrast, Policy 2 highly affects the dependent variable, leading to a distribution of Y^* that second-order stochastically dominates that of Y. The corresponding distribution functions F_Y and F_Y^* are plotted in Figure 2.5.1.

For both policies, we consider the applications of interest described in Section 2.2. Here the CDFs of Y^* and Y, and the Distribution Function Policy Effect Δ_F are estimated over the equidistant grid $\{3, 3.05, 3.1, \ldots, 8.95, 9\}$, whereas for the quantiles of Y^* and Y, the Quantile Policy Effect, the Lorenz curve of Y^* and Y and the Lorenz Curve Policy

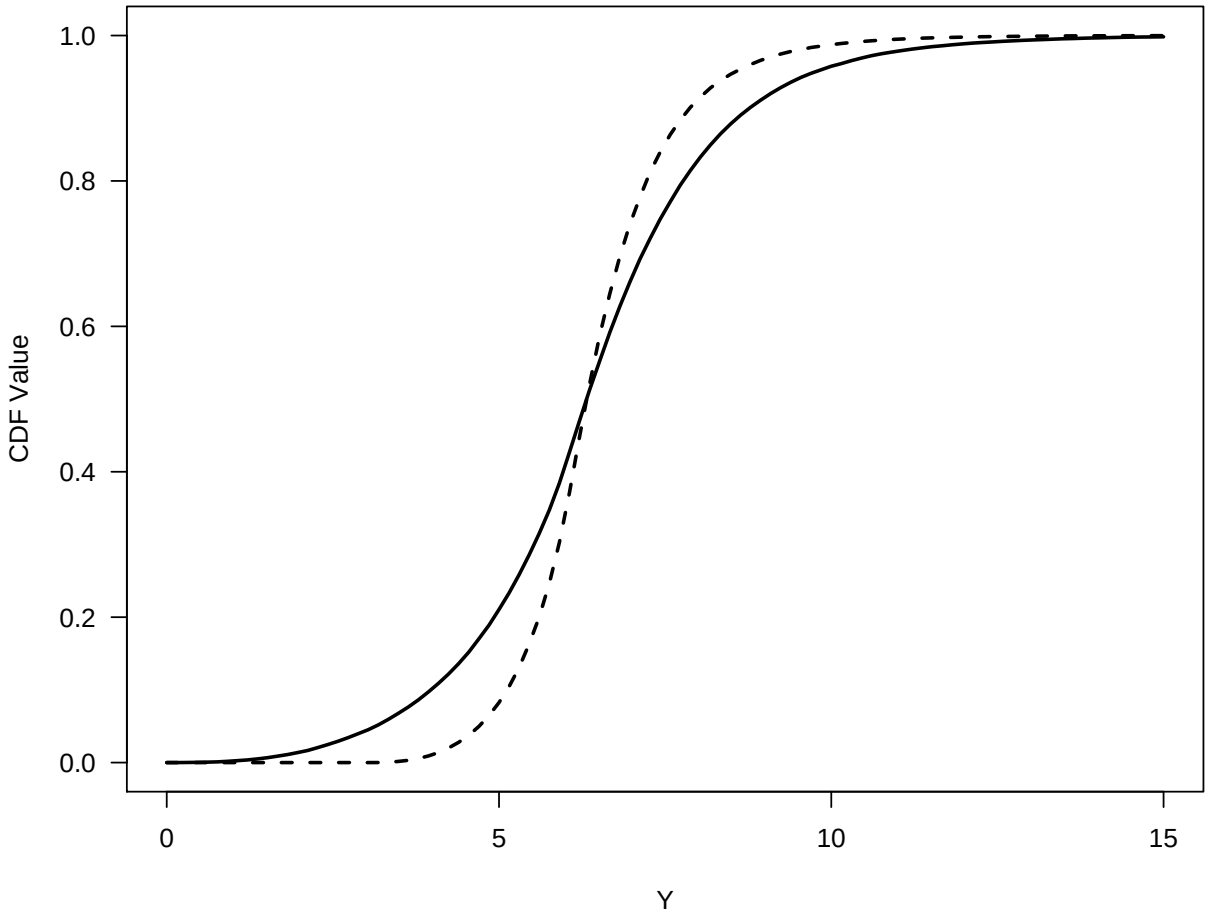

Figure 2.5.1: Plot of F_Y (solid line) and F_Y^* (dashed line) for Policy 2.

Effect the grid $\{.1, .11, .12, \ldots, .89, .9\}$ is used. Moreover, we consider tests for first- and second-order stochastic dominance of F_Y^* over F_Y. That is, in each simulation run we test the hypothesis $H_0^j : \mathcal{D}_j(y, F_Y^* - F_Y) \leq 0$ for $j = 1, 2$. We use the sample sizes $n = 100$ and $n = 400$, and set the number of Monte Carlo replications to 1000. In each replication, we use the nonparametric bootstrap with $B = 1000$ repetitions to obtain uniform 90% confidence bands for the functionals of interest (and approximate p-values in case of the stochastic dominance tests).

In order to implement our estimators, we have to specify a kernel function and a bandwidth sequence h that are compatible with the assumptions made in Section 3. A kernel function that satisfies both the higher-order and the boundary correction property

is given by

$$K_c(z) = \prod_{i=1}^{d} e_1^T S_{c_i}^{-1} (1, z_i, \ldots, z_i^p)^T \kappa(z_i),$$

where $S_c = (\mu_{j+l,c})_{0 \leq j,l \leq p}$ is a matrix of kernel constants $\mu_{j,c} = \int_{D(c)} z^j \kappa(z) dz$, $e_1 = (1, 0, \ldots, 0)^T$ is the unity vector, $p = r - 3$ and $\kappa(z)$ is a standard univariate kernel function that satisfies the remaining regularity conditions of Assumption 4. This is the product of univariate equivalent kernels of a local polynomial regression estimator (see Fan and Gijbels (1996) for more details). For our simulations, we let $\kappa(z)$ be the usual Epanechnikov kernel and choose $p = 1$, which implies that $K_c(z)$ is a fourth order kernel.

Regarding the bandwidth, our asymptotic results only prescribe a rate at which h tends to zero, but are silent about its optimal size in finite samples. Here we use a bandwidth of the form $h = cn^{-\delta}$, which requires that $\delta \in (1/2r, 1/2d) = (1/8, 1/6)$ in order for Assumption 5 to be fulfilled. Absent further guidelines, we choose $h = 1.5\sigma_x n^{-1/7}$ for our simulations, where σ_x is the standard deviation of the respective covariates. Informal robustness checks suggest that the results are not too sensitive with respect to this choice.

Results

In Table 1, we present the result of our simulation study regarding the properties of our estimates of the CDFs of Y^* and Y, the Distribution Function Policy Effect, the quantiles of Y^* and Y, the Quantile Policy Effect, the Lorenz curve of Y^* and Y, the Lorenz curve policy effect, the Gini coefficients of Y^* and Y and the Gini Policy effect. In each case, we report Monte Carlo estimates of the integrated bias (IBias), the root integrated mean squared error (RIMSE), and the coverage rate of a uniform confidence band with nominal coverage level of 90% (Cov. Rate).

Although the sample sizes we consider are relatively small, our estimators exhibit reasonable finite sample properties. Finite sample biases are generally small and decrease rapidly with the sample size. Also note that increasing the sample size from $n = 100$ to $n = 400$, i.e. by a factor of four, roughly halves the magnitude of the RIMSE for all quantities under consideration, which indicates that convergence to the true values indeed takes places at rate $\sqrt{n}$. The empirical coverage rate of the uniform 90% bootstrap

Table 2.1: Simulation Results: Properties of Nonparametric Policy Estimators

	Policy 1					
	$n = 100$			$n = 400$		
	IBias	RIMSE	CR	IBias	RIMSE	CR
F_Y^*	0.879	9.469	0.855	0.414	5.358	0.894
F_Y	0.654	7.926	0.813	0.349	3.975	0.862
Δ_F	0.367	5.059	0.952	0.277	3.553	0.964
Q_Y^*	4.495	30.421	0.900	1.000	16.063	0.898
Q_Y	2.143	22.777	0.843	0.626	11.611	0.861
Δ_Q	3.557	21.288	0.973	0.453	11.185	0.941
L_Y^*	0.150	1.008	0.833	0.066	0.542	0.877
L_Y	0.135	0.842	0.805	0.049	0.435	0.846
Δ_L	0.015	0.566	0.972	0.016	0.340	0.964
G_Y^*	0.353	1.770	0.885	0.154	0.950	0.901
G_Y	0.310	1.493	0.860	0.115	0.769	0.871
Δ_G	0.042	0.954	0.955	0.039	0.596	0.947
	Policy 2					
	$n = 100$			$n = 400$		
	IBias	RIMSE	CR	IBias	RIMSE	CR
F_Y^*	2.169	10.531	0.862	0.899	5.211	0.857
F_Y	0.266	7.698	0.847	0.326	3.978	0.865
Δ_F	2.109	7.513	0.859	0.699	3.717	0.876
Q_Y^*	2.698	20.428	0.868	1.098	9.877	0.861
Q_Y	0.580	22.265	0.875	0.591	11.466	0.853
Δ_Q	2.833	21.669	0.948	0.811	10.731	0.920
L_Y^*	0.063	0.668	0.825	0.016	0.337	0.856
L_Y	0.014	0.843	0.812	0.012	0.429	0.865
Δ_L	0.059	0.780	0.855	0.012	0.379	0.878
G_Y^*	0.098	1.150	0.877	0.007	0.586	0.873
G_Y	0.028	1.483	0.879	0.026	0.762	0.870
Δ_G	0.126	1.361	0.864	0.019	0.664	0.885

Note: Integrated Bias and RIMSE figures have been multiplied by 100 to improve readability.

Table 2.2: Simulation Results: Rejection rates of KS-type tests for stochastic dominance

		Policy 1		*Policy 2*	
		KS_1	KS_2	KS_1	KS_2
$\alpha = .05$	$n = 100$	.036	.040	.345	.011
	$n = 400$	.048	.041	.917	.012
$\alpha = .1$	$n = 100$	.074	.083	.538	.025
	$n = 400$	.098	.089	.983	.011

confidence bands is generally close to the nominal level. This procedure should thus be able to provide reliable inference even in small samples.

Table 2.2 presents the simulation results on the stochastic dominance tests. For each test, we report the empirical rejection rates for the nominal levels $\alpha = .05$ and $\alpha = .1$. Recall that for the first policy $F_Y^* \equiv F_Y$ so that H_0^j is true for $j = 1, 2$, whereas under Policy 2 F_Y^* is dominating F_Y in a second-order stochastic dominance sense, but not in a first-order one, so that only H_0^2 holds in this case. For Policy 1, both tests are conservative, but their empirical size gets closer to the respective nominal level as the sample size increases. Under Policy 2, the KS_1 test has non-trivial power for $n = 100$ and rejects the null in almost all simulation runs for $n = 400$. The rejection rates of the KS_2 test are substantially below their nominal values, which comes as no surprise as the test can only be expected to have correct size under the least favourable null hypothesis.

Comparison with Approach based on Quantile Regression

Without any point of reference, it is admittedly difficult to judge whether the finite sample properties of our estimators are "good". In this section, we therefore briefly compare them with those of an estimator based on a first-step linear quantile regression (LQR), as discussed in Machado and Mata (2005), Melly (2005) and Chernozhukov, Fernandez-Val, and Melly (2008). Instead of using a kernel estimator, this method obtains an estimate of the conditional distribution function $F_{Y|X}$ by inverting an estimate of the conditional quantile function $Q_{Y|X}$, which is assumed to be linear in the regressors at every quantile, i.e. it imposes the restriction that $Q_{Y|X}(\tau|x) = x\beta(\tau)$ for all $\tau \in (0, 1)$. Since this estimator imposes additional parametric restrictions on the relationship between the

Table 2.3: Simulation results: Comparisson of estimators.

	$a = 0$			$a = .5$			$a = 1$		
	IBias	RIMSE	CR	IBias	RIMSE	CR	IBias	RIMSE	CR
NP	1.197	10.623	.920	2.202	13.972	.940	3.363	18.249	.934
LQR	1.883	9.319	.917	8.051	15.082	.702	16.743	26.074	.294

Note: Integrated Bias and RIMSE figures have been multiplied by 100 to improve readability.

dependent variable and the covariates, the result will generally exhibit less finite sample variation than our nonparametric procedure. On the other hand, such an estimator is also more prone to misspecification bias. The purpose of this section is to illustrate this tradeoff.

We compare the properties of the LQR-based estimator to our nonparametric procedure via simulation. For brevity, we restrict attention on the Quantile Policy Effect, and consider only the effect of Policy 2 for $n = 400$. The setup we use is the same as described above, with the exception that the dependent variable is now generated as

$$Y = 6 - 2X_1 + X_2 + aX_2^2 + \sigma(X)\varepsilon, \qquad \sigma^2(X) = X_1 + X_2. \tag{2.5.2}$$

The parameter a governs the complexity of the relationship. For $a = 0$, (2.5.2) is the same as (2.5.1) considered above. In this case, a LQR model would be correctly specified. To illustrate the effect of misspecification, we also consider $a = .5$ and $a = 1$.

The result of the simulations, given in Table 2.3, show that our nonparametric estimator compares favourably with its competitor and performs well uniformly over the different values of a we considered. It has the lowest RIMSE under all designs except for $a = 0$, where it exceeds the RIMSE of the correctly specified LQR-based estimator by about 15%. When the underlying model is not correctly specified, the LQR-based estimator can exhibit a substantial bias, with its magnitude depending on the degree of misspecification. The coverage rates of uniform confidence bands can also deviate significantly from their nominal levels in this case.

2.5.2 Empirical Illustration: The Effect of an Anti-Smoking Campaign on Infant Birthweights

In this section, we illustrate the application of our estimators through an example form public health. We consider a (hypothetical) public policy that successfully induces women who smoke during their pregnancy to cut their average daily cigarette consumption by 75%. Our interest is in the effect of this policy on the distribution of infant birthweight in general, and whether it helps to reduce the incidence of low-birthweight infants, which is usually defined by infants weight at birth falling below 2500 grams (about 5 pounds, 8 ounces). These issues should be of concern to policy makers since low birthweight is known to be associated with a wide range of subsequent health problems, and has even been linked to later educational attainment and labor market outcomes (see for example Almond, Chay, and Lee (2005) or Black, Devereux, and Salvanes (2007)).

The data we use is a subsample of the Detailed Natality Data (June 1997) published by the National Center for Health Statistics. A more extensive analysis of the full data set is given by Abrevaya (2001) and Koenker and Hallock (2001). The subsample we employ comprises 4439 white mothers between age 18 and 45 without any college education, who gave birth to a live, single infant and smoked during their pregnancy. For each woman in this particular subgroup, we record the infant's birthweight (in grams) and the average daily number of cigarettes the mother smoked during the pregnancy, together with other variables that could possibly confound the relationship between birthweight and the level of cigarette consumption. These include the mother's age, mother's weight gain during the pregnancy (in pounds) and whether the mother is married. Table 2.4 presents some descriptive statistics for our full data set. The identifying assumption is that conditional on these variables unobserved factors that influence birthweight are independent of the average amount of cigarettes consumed, given that the mother chooses to smoke during the pregnancy in the first place.

Using the same specifications as for our simulation study in Section 5.1, we estimate the effect of reducing every woman's cigarette consumption by 75% on the distribution of birthweights by our nonparametric procedure. The discrete regressor is accommodated using the conventional frequency method. Figure 2.5.2 presents the estimate of the Quan-

Table 2.4: Descriptive Statistics

	Mean	Std. Dev	Min	Q25	Median	Q75	Max
Birthweight	3176	560.36	457	2889	3204	3515	5245
Cigarettes per day	11.98	7.51	1	6.5	10	20	60
Mother's age	25.15	5.56	18	21	24	29	45
Mother's Weight Gain	30.19	14.07	0	20	30	40	98
Married	0.52	0.49	0	–	–	–	1

Note:$N = 4439$. Married is an indicator variable for the mother being married.

tile Policy Effect $\Delta_Q(\cdot)$, together with 90% confidence intervals for every point, and a 90% uniform confidence band, both obtained via the bootstrap with $B = 1000$ replications. The graph suggest that the policy increases infant birthweights over almost the entire range of quantiles considered. Moreover, the point estimate suggests that particularly the low quantiles would benefit by such a campaign, with an estimated increase in the 10% quantile of about 140 grams compared to only 70 grams at the 90% quantile. However, since the associated confidence band is relatively wide, one cannot reject the hypothesis that $\Delta_Q(\cdot)$ is constant, and thus there is no significant evidence of heterogeneous policy effects.

As mentioned above, another quantity of interest is the proportion of low-weight birth incidents, which amounts to 9.28% in the subpopulation under consideration. The see how a change in smoking habit would affect this share, we simply estimate the Distribution Function Policy Effect $\Delta_F(\cdot)$ and evaluate it 2500 grams. As a result, we obtain that

$$\hat{\Delta}_F(2500) = -0.0290 \quad \text{with} \quad s.e.(\hat{\Delta}_F(2500)) = 0.0092.$$

The corresponding 90% confidence interval is given by $CI_{.9} = (-0.0429, -0.0140)$. This implies that the policy would reduce low-weight birth incidents by roughly one third, which is a substantial amount. While the confidence interval is again relatively wide, it is substantially to the left of zero, which indicates that the policy should be effective for reducing low-birthweight incidents.

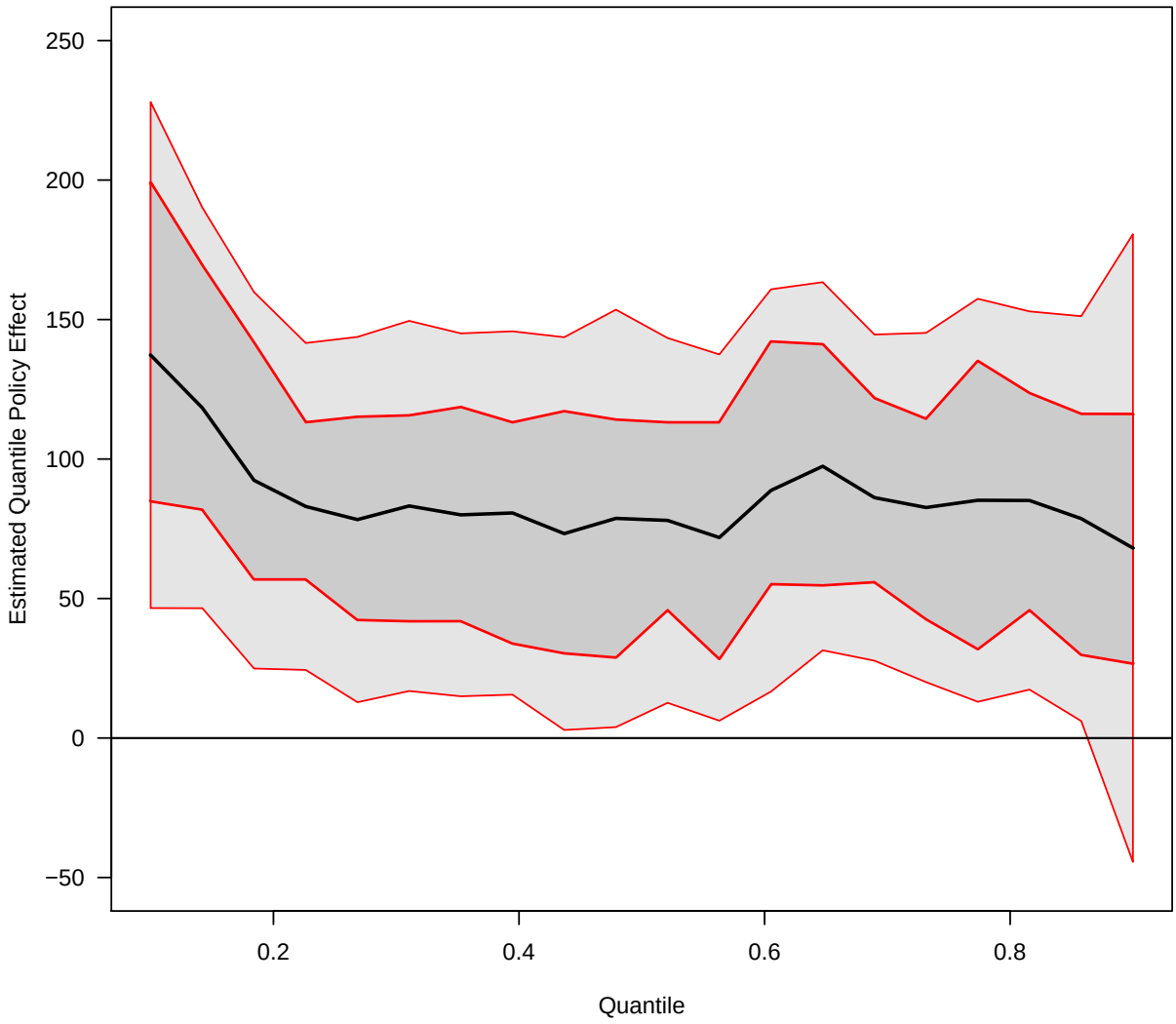

Figure 2.5.2: Estimated Quantile Policy Effect on Infant Birthweight (solid line), with pointwise 90% confidence bands (dark-grey area) and uniform 90% confidence bands (light-grey area) based on 1000 bootstrap replications.

2.6 Conclusions

In this paper, we have proposed a fully nonparametric way to assess the effect of an exogenous change in the distribution of the covariates on the unconditional distribution of a dependent variable of interest. The method can be used to conduct asymptotically valid inference on various kinds of statistics that can be written as a sufficiently smooth functional of the CDF. It is straightforward to implement and performs surprisingly well in simulations when the sample size is relatively small, even compared to correctly specified parametric estimators. This is in sharp contrast to classical nonparametric

methods, whose use is prohibitive due to the curse of dimensionality in situations with many regressors and few observations.

Appendix

Proof of Proposition 1. From the definition of a distribution function, the law of iterated expectations, and that ε is independent of both X and X^*, we obtain that

$$\begin{aligned} F_Y^*(y) &= P(m(X^*, \varepsilon) \leq y) \\ &= \int P(m(X^*, \varepsilon) \leq y | X^* = x) dF_X^*(x) \\ &= \int P(m(x, \varepsilon) \leq y) dF_X^*(x) \\ &= \int P(m(X, \varepsilon) \leq y | X = x) dF_X^*(x) \\ &= \mathbb{E}(F_{Y|X}(y, X^*)), \end{aligned}$$

where $F_{Y|X}$ is the conditional distribution function of Y given X. Since the data (Y, X) certainly identify this function at any point (y, x) in their support, and X^* only takes values in a subset of the support of X with probability 1, this implies that F_Y^* is identified. □

Proof of Proposition 2. The statement can be shown using standard kernel smoothing theory. First, one can show that the rate of the bias of each estimator is uniformly of the order $O(h^r)$ by standard Taylor expansion arguments. The fact that the rate is the same in both the interior and the vicinity of the boundary of the support of X is a consequence of the use of a boundary kernel. Second, the rate of the stochastic part can be shown to be $O_p(\log(n)/(nh^d))$ by using the arguments from, say, Newey (1994a) for (ii), and Härdle, Janssen, and Serfling (1988) or Akritas and Van Keilegom (2001) for (i) and (iii). Taken together, these results imply the rates given in the proposition. □

Proof of Proposition 3. Consider for notational simplicity the case that $\lambda = 1$, and definde the function w through

$$w(x) = \frac{1}{n} \sum_{i=1}^{n^*} \frac{K_{X_i^*, h}(x - X_i^*)}{\hat{f}_X(X_i^*)},$$

so that $w_j = w(X_j)$. Noting that $\hat{F}_Y^* \equiv \tilde{F}_Y^*$ when $w(x)$ is positive for every value of x, we obtain

that

$$\begin{aligned} P(\hat{F}_Y^* = \tilde{F}_Y^*) &\geq P(\inf_{x\in J} w(x) \geq 0) \\ &\geq P(\inf_{x\in J} \hat{f}_X(x) \geq c \cap \inf_{x\in J} \hat{f}_X^*(x) \geq 0) \\ &\geq P(\inf_{x\in J} \hat{f}_X(x) \geq c) + P(\inf_{x\in J} \hat{f}_X^*(x) \geq 0) - 1 \end{aligned} \tag{2.6.1}$$

for some constant $c > 0$, uniformly in y. We now show that both probabilities in (2.6.1) are of the order $1 + o_p(n^{-1/2})$ if c is chosen sufficiently small, which gives the desired result.

To see this for the first term in (2.6.1), note that the density of X is bounded away from zero on J, i.e. there exists some $\delta > 0$ such that $\inf_{x\in J} f_X(x) > \delta$. Now, set $c = \delta/2$. It is a well-known result that $\mathbb{E}(\hat{f}_X(x)) = f_X(x) + O(h^r)$ uniformly in x, and we thus know that $\inf_x \mathbb{E}(\hat{f}_X(x)) > \frac{3}{4}\delta$ for some sufficiently large n. This also means that for sufficiently large n the event $\{\inf_x \hat{f}_X(x) \geq c\}$ is implied by the event $\{\inf_x(\hat{f}_X(x) - \mathbb{E}(\hat{f}_X(x))) \geq -c/2\}$, which is in turn implied by the event $\{\sup_x |\hat{f}_X(x) - \mathbb{E}(\hat{f}_X(x))| < c/2\}$. That is, it holds that

$$P(\inf_{x\in J} \hat{f}_X(x) \geq c) \geq P(\sup_{x\in J} |\hat{f}_X(x) - \mathbb{E}(\hat{f}_X(x))| < c/2).$$

Now, since J is compact, it can be covered by $v_n \leq \gamma_1 n^d$ open balls with radius n^{-1}, for some $\gamma_1 > 0$. The kth of these balls, with midpoint $x_{n,k}$, is denoted by

$$J_{n,k} = \{x \in \mathbb{R}^d : \|x - x_{n,k}\| \leq n^{-1}\}.$$

Then we have that

$$\sup_{x\in J} |\hat{f}_X(x) - \mathbb{E}(\hat{f}_X(x))| \leq \max_{1\leq k\leq v_n} \sup_{x\in J_{n,k}} |\hat{f}_X(x) - \mathbb{E}(\hat{f}_X(x))|.$$

For $x \in J_{n,k}$, it follows from the triangle inequality that

$$|\hat{f}_X(x) - \mathbb{E}(\hat{f}_X(x))| \leq |\hat{f}_X(x) - \hat{f}_X(x_{n,k})| + |\hat{f}_X(x_{n,k}) - \mathbb{E}(\hat{f}_X(x_{n,k}))| + |\mathbb{E}(\hat{f}_X(x_{n,k})) - \mathbb{E}(\hat{f}_X(x))|.$$

Since the kernel function K has bounded partial derivatives, the first and third term on the right-hand side of the last equation can be bounded as follows:

$$\begin{aligned} |\hat{f}_X(x) - \hat{f}_X(x_{n,k})| &\leq C\frac{|x - x_{n,k}|}{h^{d+1}} \leq Cn^{-1}h^{-d-1}, \\ |\mathbb{E}(\hat{f}_X(x_{n,k})) - \mathbb{E}(\hat{f}_X(x))| &\leq C\frac{|x - x_{n,k}|}{h^{d+1}} \leq Cn^{-1}h^{-d-1} \end{aligned}$$

for some $C > 0$. Thus

$$\sup_{x\in J} |\hat{f}_X(x) - \mathbb{E}(\hat{f}_X(x))| \leq \max_{1\leq k\leq v_n} |\hat{f}_X(x_{n,k}) - \mathbb{E}(\hat{f}_X(x_{n,k}))| + 2Cn^{-1}h^{-d-1}. \tag{2.6.2}$$

Using a result from Bosq (1998, p.47–48), we also have that

$$\begin{aligned} P(\max_{1\leq k\leq v_n} |\hat{f}_X(x_{n,k}) - \mathbb{E}(\hat{f}_X(x_{n,k}))| > c/4) &\leq \sum_{k=1}^{v_n} P(|\hat{f}_X(x_{n,k}) - \mathbb{E}(\hat{f}_X(x_{n,k}))| > c/4) \\ &\leq v_n C \exp(-\gamma_2 \sqrt{nh^d}) \\ &\leq C \exp(-\gamma_2 \sqrt{nh^d} + \gamma_1 \log(n)) \end{aligned} \tag{2.6.3}$$

for some $\gamma_2 > 0$. Since $n^{-1}h^{-d-1} \to 0$ by Assumption 5, (2.6.2) and (2.6.3) together imply that for n sufficiently large

$$\begin{aligned} P(\sup_{x\in J} |\hat{f}_X(x) - \mathbb{E}(\hat{f}_X(x))| < c/2) &\geq P(\max_{1\leq k\leq v_n} |\hat{f}_X(x_{n,k}) - \mathbb{E}(\hat{f}_X(x_{n,k}))| < c/4) \\ &\geq 1 - C \exp(-\gamma_2 \sqrt{nh^d} + \gamma_1 \log(n)) \\ &= 1 + o_p(n^{-1/2}) \end{aligned}$$

if $nh^d/\log(n)^2 \to \infty$, which is again ensured through Assumption 5. An analogous argument then applies to the second term in (2.6.1). This completes our proof. □

Proof of Theorem 1. In this section, we will briefly switch to the operator notation typically used in the empirical process literature. In particular, for $A_1, \ldots, A_n$ an i.i.d. sequence of random variables taking values in $(\mathcal{A}, \mathcal{B})$ with distribution P and for some measureable function $\phi : \mathcal{X} \to \mathbb{R}$ we will write

$$P\phi = \int \phi dP, \qquad \mathbb{P}_n\phi = \frac{1}{n}\sum_i \phi(Z_i), \qquad \mathbb{G}_n\phi = \sqrt{n}(\mathbb{P}_n - P)\phi$$

for the expectation, empirical measure and empirical process at ϕ, respectively. Furthermore, we write $\bar{o}_p(a_n)$ as a shorthand notation for "$o_p(a_n)$ uniformly in $y \in \mathbb{R}$".

The difficulties in deriving the limit behaviour of $\mathbb{F}_n$ arise from the fact that $\hat{F}_Y^*$ is itself a random function that has been estimated from the data. Using the notation described above, the first component of $\mathbb{F}_n$ can be rewritten as

$$\sqrt{n}(\hat{F}_Y^* - F_Y^*) = \sqrt{n}(\mathbb{P}_{n^*}^* \hat{F}_{Y|X} - P^* F_{Y|X})$$

where P^* is the distribution of X^* and both $\hat{F}_{Y|X}$ and $F_{Y|X}$ are seen as functions of X^* indexed by y, i.e. $X^* \mapsto \hat{F}_{Y|X}(y, X^*)$ and similarly for $F_{Y|X}$. The above expression can be decomposed into the following three terms,

$$\begin{aligned} \sqrt{n}(\mathbb{P}_{n^*}^* \hat{F}_{Y|X} - P^* F_{Y|X}) &= \sqrt{\lambda}\mathbb{G}_{n^*}^*(\hat{F}_{Y|X} - F_{Y|X}) + \sqrt{\lambda}\mathbb{G}_{n^*}^* F_{Y|X} + \sqrt{n}P^*(\hat{F}_{Y|X} - F_{Y|X}) \\ &= T_1 + T_2 + T_3, \end{aligned}$$

which can now be analyzed separately. Beginning with T_1, it is shown below in Lemma 1 that

$$\sup_{y\in\mathbb{R}}\left|\mathbb{G}^*_{n^*}(\hat{F}_{Y|X}-F_{Y|X})\right| = o_p(1)$$

and thus the first term on the right hand side vanishes as n tends to infinity. The second term is equal to

$$T_2 = \frac{1}{\sqrt{n^*}}\sum_{i=1}^{n^*}\sqrt{\lambda}(F_{Y|X}(y,X_i^*)-F_Y^*(y)),$$

which does not contain any unknown functions and is thus easy to handle. Finally, Lemma 2 establishes that

$$T_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{f_X^*(X_i)}{f_X(X_i)}(\mathbb{I}\{Y_i\le y\}-F_{Y|X}(y,X_i))+\bar{o}_p(1).$$

To derive the asymptotic distribution for the actual process of interest, we have to distinguish between a dependent and an independent policy implementation, and we have to introduce some more notation. First, define the functions

$$\begin{aligned}
\xi_{11}(y) &: X^* \mapsto \sqrt{\lambda}F_{Y|X}(y,X^*)\\
\xi_{12}(y) &: X^* \mapsto 0\\
\xi_{21}(y) &: (Y,X) \mapsto \frac{f_X^*(X)}{f_X(X)}(\mathbb{I}\{Y\le y\}-F_{Y|X}(y,X))\\
\xi_{22}(y) &: (Y,X) \mapsto \mathbb{I}\{Y\le y\}-F_Y(y)
\end{aligned}$$

and let $\xi_1 = (\xi_{11},\xi_{12})^T$ and $\xi_2 = (\xi_{21},\xi_{22})^T$ and $\xi = \xi_1+\xi_2$. Second, define the classes of functions $\mathcal{P}_{ij} = \{\xi_{ij}(y); y\in\mathbb{R}\}$ for $i,j = 1,2$ and $\mathcal{P}_i = \{\xi_i(y); y\in\mathbb{R}^2\}$ for $i = 1,2$ and $\mathcal{P} = \{\xi(y); y\in\mathbb{R}^2\}$. Then it can be shown that each of these classes of functions is Donsker by combining the results in the Examples 19.6, 19.9 and 19.20 in van der Vaart (2000), and by noting that if two classes of functions are Donsker then so is their Cartesian product (van der Vaart 2000, p. 270).

Now consider the case of an independent policy implementation. Letting P be the distribution of (Y,X), the process of interest can be written as

$$\sqrt{n}\left(\hat{\mathbf{F}}-\mathbf{F}\right) = \mathbb{G}^*_{n^*}\xi_1+\mathbb{G}_n\xi_2+o_p(1).$$

Since $\mathcal{P}_1$ and $\mathcal{P}_2$ are Donsker, the first and second term on the right hand side converge to two independent Gaussian process. Thus, by the continuous mapping theorem, the entire right hand side of the last equation converges to the sum of these to Gaussian processes, which is

again a Gaussian process by independence. It is then straightforward to see that this has mean zero and covariance function as stated in the Theorem.

For a dependent policy implementation, let $\bar{P}$ be the joint distribution of (Y, X, X^*) and recall that $\lambda = 1$ in this case. The process can then be written as

$$\sqrt{n}\left(\hat{\mathbf{F}} - \mathbf{F}\right) = \bar{\mathbb{G}}_n \xi + o_p(1).$$

which converges to a Gaussian process since $\mathcal{P}$ is Donsker. It is again straightforward to see that the limiting process has again mean zero and covariance function as stated in the Theorem (with $\lambda = 1$). This completes our proof. □

We now prove the two lemmas used in the above argument.

Lemma 1. *Under the conditions of Theorem 1, it holds that*

$$\sup_{y \in \mathbb{R}} \left| \mathbb{G}^*_{n^*}(\hat{F}_{Y|X} - F_{Y|X}) \right| = o_p(1)$$

Proof. By Lemma 19.24 in van der Vaart (2000), the statement of the lemma follows if (i) the sequence of random functions $\hat{F}_{Y|X}$ takes its values in some Donkser class $\mathcal{F}$, and if (ii)

$$\sup_{y \in \mathbb{R}} P^*(\hat{F}_{Y|X} - F_{Y|X})^2 = o_p(1).$$

This last condition obviously holds in our case because

$$\begin{aligned} P^*(\hat{F}_{Y|X} - F_{Y|X})^2 &= \int (\hat{F}_{Y|X}(y,x) - F_{Y|X}(y,x))^2 dF^*_X(x) \\ &= o_p(n^{-1/2}) \end{aligned}$$

uniformly in y and x since $||\hat{F}_{Y|X}(y,x) - F_{Y|X}(y,x)||_\infty = o_p(n^{-1/4})$ by Proposition 2, and the integration takes place over a compact set.

To see that the first condition holds, define the class $\mathcal{F}$ by

$$\mathcal{F} = \{\hat{F}_{Y|X}(y, \cdot) : y \in \mathbb{R}\},$$

and note that by Assumption 4 each of it's elements is r times continuously differentiable, and the derivatives are uniformly bounded. Hence $\mathcal{F} \subset C^r_M(J^*)$, the space of all functions on J^*, the support of X^*, whose derivatives up to order r are uniformly bounded by some constant M. But this class is Donsker if $r > d/2$, where d is dimension of J^*, as shown in van der Vaart (2000, Example 19.9). However, our assumptions on the bandwidth already require that $r > d$, so that this condition easily holds in our case. □

Lemma 2. *Under the conditions of Theorem 1,*

$$\sqrt{n}P^*(\hat{F}_{Y|X} - F_{Y|X}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbb{I}\{Y_i \leq y\} - F_{Y|X}(y, X_i))\frac{f_X^*(X_i)}{f_X(X_i)} + o_p(1)$$

uniformly in y.

Proof. Switching back to the usual notation, we have that

$$\sqrt{n}P^*(\hat{F}_{Y|X} - F_{Y|X}) = \sqrt{n}\int \hat{F}_{Y|X}(y,x) - F_{Y|X}(y,x)dF_X^*(x).$$

Here and in the following, all integrals are understood to be taken over the entire $\mathbb{R}^d$. In order to derive an expression for the integral in the above equation, it is useful to split it up into two components.

$$\begin{aligned}
&\int \hat{F}_{Y|X}(y,x) - F_{Y|X}(y,x)dF_X^*(x) \\
&= \int \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \leq y\} - F_{Y|X}(y,X_i))\frac{K_{x,h}(x - X_i)}{\hat{f}_X(x)}dF_X^*(x) \\
&+ \int \frac{1}{n}\sum_i (F_{Y|X}(y,X_i) - F_{Y|X}(y,x))\frac{K_{x,h}(x - X_i)}{\hat{f}_X(x)}dF_X^*(x) \\
&= A + B
\end{aligned}$$

We start with analysing the term A. Using the fact that X is a continuous random variable with density function f_X, and applying a second order Taylor expansion of $1/\hat{f}_X(x)$ around $1/f_X(x)$, we obtain

$$\begin{aligned}
A &= \int \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \leq y\} - F_{Y|X}(y,X_i))\frac{K_{x,h}(x - X_i)}{\hat{f}_X(x)}f_X^*(x)dx \\
&= \int \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \leq y\} - F_{Y|X}(y,X_i))\frac{K_{x,h}(x - X_i)}{f_X(x)}f_X^*(x)dx \\
&- \int \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \leq y\} - F_{Y|X}(y,X_i))(\hat{f}_X(x) - f_X(x))\frac{K_{x,h}(x - X_i)}{f_X(x)^2}f_X^*(x)dx \\
&+ o_p(n^{-1/2}) \\
&= A_1 + A_2 + o_p(n^{-1/2})
\end{aligned}$$

where the last term is $o_p(n^{-1/2})$ uniformly in x and y since $||\hat{f}_X(x) - f_X(x)||_\infty = o_p(n^{-1/4})$, $|\mathbb{I}\{Y_i < y\} - F_{Y|X}(y, X_i)| \leq 1$, and integration takes place over a compact set.

To derive an expression for A_1, define $u(x) = f_X^*(x)/f_X(x)$, let $u^{(\mu)}(x) = \partial_x^\mu u(x)$ and $K_c^{(\mu)}(x) = \partial_c^\mu K_c(x)$. Using standard techniques from the kernel smoothing literature, we obtain

that

$$
\begin{aligned}
A_1 &= \int \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i)) K_{x,h}(x - X_i) u(x) dx \\
&= \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i)) \int K_{zh+X_i}(z) u(zh + X_i) dz \\
&= \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i)) \int (K_{X_i}(z) + zhK^{(1)}_{X_i}(z) + \ldots + (zh)^r K^{(r)}_{\xi}(z)) \\
&\qquad \times (u(X_i) + zhu^{(1)}(X_i) + \ldots + (zh)^r u^{(r)}(\xi)) dz \\
&= \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i))(u(X_i) + O_p(h^r)) \\
&= \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i)) \frac{f^*(X_i)}{f(X_i)} + o_p(n^{-1/2})
\end{aligned}
$$

where ξ is some value between X_i and $X_i + zh$. Here the second-to-last equality follows from interchanging the order of differentiation and integration (which in turn follows by dominated convergence) and using the kernel properties, and the last equality holds because $O_p(h^r) = o_p(n^{-1/2})$ by Assumption 5.

Next, we consider the term A_2. Plugging in the definition of $\hat{f}_X$, we obtain

$$
\begin{aligned}
A_2 &= \int \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i))(\hat{f}(x) - f(x)) \frac{K_h(x - X_i)}{f_X(x)^2} f^*_X(x) dx \\
&= \frac{1}{n}\sum_i (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i)) \int \left(\frac{1}{n}\sum_j K_{x,h}(X_j - x) - f(x) \right) K_{x,h}(x - X_i) \frac{f^*_X(x)}{f_X(x)^2} dx \\
&= \frac{1}{n^2}\sum_{i,j} (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i)) \int (K_{x,h}(X_j - x) - f(x)) K_{x,h}(x - X_i) \frac{f^*_X(x)}{f_X(x)^2} dx \\
&= \frac{1}{n^2}\sum_{i,j} (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i)) \\
&\qquad \times \left(\int K_{x,h}(X_j - x) K_{x,h}(x - X_i) \frac{f^*_X(x)}{f_X(x)^2} dx - \int K_{x,h}(x - X_i) \frac{f^*_X(x)}{f_X(x)} dx \right) \\
&= \frac{1}{n^2}\sum_{i,j} (\mathbb{I}\{Y_i \le y\} - F_{Y|X}(y, X_i))(A_{21} - A_{22})
\end{aligned}
$$

Applying the same kind of argument as for the derivation of A_1, we obtain

$$
A_{22} = \frac{f^*_X(X_i)}{f_X(X_i)} + o_p(n^{-1/2})
$$

Turning to the first term, defining $v(x) = f^*_X(x)/f^2_X(x)$, using similar arguments as before, we

obtain that

$$\begin{aligned} A_{21} &= \int v(x) K_{x,h}(X_j - x) K_{x,h}(x - X_i) dx \\ &= \int v(zh + X_i) K_{zh+X_i,h}(X_j - X_i - zh) K_{zh+X_i}(z) dz \\ &= v(X_i) K_{X_i,h}(X_j - X_i) + \bar{o}_p(n^{-1/2}). \end{aligned}$$

Hence we have shown that

$$\begin{aligned} A_2 &= \frac{1}{n^2} \sum_{i,j} v(X_i)(\mathbb{I}\{Y_i \leq y\} - F_{Y|X}(y, X_i)) \left(K_{X_i,h}(X_j - X_i) - f_X(X_i)\right) + \bar{o}_p(n^{-1/2}) \\ &= U_n(y) + \bar{o}_p(n^{-1/2}). \end{aligned}$$

To proceed, we need to introduce some further notation. Let $\bar{f}_X(x) = \mathbb{E}(K_{x,h}(X_j - x))$, and define $H(Z_i, Z_j; y, h) = v(X_i)(\mathbb{I}\{Y_i \leq y\} - F_{Y|X}(y, X_i)) \left(K_{X_i,h}(X_j - X_i) - \bar{f}_X(X_i)\right)$. Then, because by Proposition 2 we have that $\bar{f}_X(x) - f_X(x) = o_p(n^{-1/2})$ uniformly in x, we can write $U_n(y)$ as

$$U_n(y) = \frac{1}{n^2} \sum_i \sum_{j \neq i} H(Z_i, Z_j; y, h) + \frac{1}{n^2} \sum_i H(Z_i, Z_i; y, h) + o_p(n^{-1/2}). \tag{2.6.4}$$

It is straightforward to see that the first term on the right hand side of (2.6.4) is a degenerate second order U-process. It then follows from Corollary 4 in Sherman (1994), a Uniform Law of Large Numbers for U-processes, that

$$\sup_{y \in \mathbb{R}} \left| \frac{1}{n^2} \sum_i \sum_{j \neq i} H(Z_i, Z_j; y, h) \right| = O_p(h^{-d} n^{-1}).$$

To see that the conditions of Sherman (1994, Corollary 4) are satisfied, let

$$\mathcal{H} = \{h^d H(\cdot; y, h), y \in \mathbb{R}, h > 0\}.$$

Since $h^d H(\cdot; y, h)$ is uniformly bounded as a function of both y and h, the class $\mathcal{H}$ has a constant (and hence square integrable) envelope function. The class also satisfies the so-called euclidean property required for Corollary 4, by Assumption 3 and 4, Lemma 22(ii) in Nolan and Pollard (1987) and Lemma 2.14 of Pakes and Pollard (1989). Hence the conditions of the corollary are satisfied.

One can furthermore directly see that the second term in (2.6.4) is also $O_p(h^{-d} n^{-1})$ uniformly in y, and thus $A_{22} = \bar{o}_p(n^{-1/2})$ because $O_p((nh^d)^{-1}) = o_p(n^{-1/2})$ by Assumption 5. This completes our argument regarding the term A. Turning to the term B, it can be shown that $B = o_p(n^{-1/2})$ uniformly in y through similar reasoning. □

Proof of Theorem 2. The statement of the Theorem follows directly from the Functional Delta Method (see Section 3.9 in van der Vaart and Wellner (1996)) □

Proof of Theorem 3. To see the first assertion of the theorem, consider the case of a dependent policy implementation. Recall from the proof of Theorem 1 that

$$\sqrt{n}\left(\hat{\mathbf{F}} - \mathbf{F}\right) = \bar{\mathbb{G}}_n \xi + o_p(1).$$

Since $\mathcal{P} = \{\xi(y); y \in \mathbb{R}^2\}$ is Donsker, the result then follows from Theorem 3.6.1 in van der Vaart and Wellner (1996). An analogous argument can be made for the case of an independent policy implementation. The second assertion of the theorem is then simply a consequence of the Function Delta Method for the bootstrap, see Theorem 3.9.11 in van der Vaart and Wellner (1996). □

Proof of Proposition 3. By Lemma 21.4 in van der Vaart (2000), under the conditions of the proposition the map Γ with

$$\Gamma(\phi) = (\phi_1^{-1}, \phi_2^{-1}) \quad \text{and } \phi_i^{-1}(\tau) = \inf\{y : \phi_i(y) \geq \tau\}, \quad i = 1, 2$$

is Hadamard differentiable at $\mathbf{F}$ tangentially to $C(0,1) \times C(0,1)$, with derivative

$$\phi \mapsto \Gamma'_{\mathbf{F}}(\phi) = -\left(\frac{\phi_1}{f_Y^*}, \frac{\phi_2}{f_Y}\right) \circ \mathbf{F}^{-1}.$$

Thus, by Theorem 2 the joint quantile process $\sqrt{n}(\hat{\mathbf{Q}} - \mathbf{Q})$ converges weakly to $\mathbb{Q}_o$ in $\ell^\infty(0,1) \times \ell^\infty(0,1)$. The validity of the bootstrap then follows directly from Theorem 3. □

Proof of Proposition 4. Our process of interest $\sqrt{n}(\hat{\mathbf{L}} - \mathbf{L})$ can be rewritten as

$$\sqrt{n}(\hat{\mathbf{L}} - \mathbf{L}) = \sqrt{n}(\Gamma_L(\hat{\mathbf{F}}) - \Gamma_L(\mathbf{F}))$$

with the "Lorenz functional" Γ_L defined as

$$\begin{aligned}\Gamma_L(\phi)(p) &= \left(\int_0^{p_1} \phi_1^{-1}(\tau)d\tau, \int_0^{p_2} \phi_2^{-1}(\tau)d\tau\right) \cdot \left(\frac{1}{\int_0^\infty \phi_1^{-1}(\tau)d\tau}, \frac{1}{\int_0^\infty \phi_2^{-1}(\tau)d\tau}\right) \\ &\equiv S(\phi)(p) \cdot \mu^{-1}(\phi)\end{aligned}$$

for $p = (p_1, p_2)^T$. Using this notation, we can rewrite the Lorenz process as

$$\sqrt{n}(\hat{\mathbf{L}} - \mathbf{L}) = \mu^{-1}(\hat{\mathbf{F}}) \cdot \sqrt{n}(S(\hat{\mathbf{F}}) - S(\mathbf{F})) - \mu^{-1}(\hat{\mathbf{F}}) \cdot \mathbf{L} \cdot \sqrt{n}(\mu(\hat{\mathbf{F}}) - \mu(\mathbf{F}))$$

The asymptotic properties of this expression can now be derived by looking at the individual components. First, since $\mu^{-1}(\cdot)$ is a continuous functional, it follows from Theorem 1 and the continuous mapping theorem that

$$\mu^{-1}(\hat{\mathbf{F}}) \to \mu^{-1}(\mathbf{F}) \equiv (1/\mu_Y^*, 1/\mu_Y),$$

where μ_Y^* and μ_Y are the unconditional means of Y^* and Y, respectively. Second, using a result from Bhattacharya (2007, Claim 1), it follows that the map S is Hadamard differentiable at $\mathbf{F}$ tangentially to $C[0,1] \times C[0,1]$ with derivative

$$\phi \mapsto S'_{\mathbf{F}}(\phi)(p) = \left(\int_0^{p_1} \frac{\phi_1(F_Y^{*-1}(\tau))}{f_Y^*(F_Y^{*-1}(\tau))} d\tau, \int_0^{p_2} \frac{\phi_2(F_Y^{-1}(\tau))}{f_Y(F_Y^{-1}(\tau))} d\tau \right).$$

Note that this is the componentwise integral over the Hadamard derivative of the quantile operator from the proof of Proposition 3. Applying Theorem 2, we obtain that

$$\sqrt{n}(S(\hat{\mathbf{F}}) - S(\mathbf{F})) \Rightarrow \mathbb{H}_o.$$

Finally, we have that

$$\sqrt{n}(\mu(\hat{\mathbf{F}}) - \mu(\mathbf{F})) \Rightarrow \mathbb{H}_o(1)$$

because $\mu(\cdot) = S(\cdot)(1)$. Taking the last results together, the first statement of the proposition follows from Slutzky's Theorem. The validity of the bootstrap then follows again directly from Theorem 3. □

Proof of Proposition 5. From the continuous mapping theorem, it follows that

$$\sqrt{n}(\hat{\mathbf{G}} - \mathbf{G}) \Rightarrow \int_0^1 \mathbb{L}_o(p) dp.$$

Since $\mathbb{L}_o$ is a Gaussian process, the term on the right-hand-side is a normally distributed random variable, with mean zero and variance as given in the proposition. The validity of the bootstrap follows again from Theorem 3. □

Proof of Proposition 6. Our proof follows essentially the same lines as the one of Proposition 3 in Barrett and Donald (2003). First, note that the map $\phi \mapsto \mathcal{D}_j(\cdot, \phi)$ is a linear functional of a Hadamard differentiable mapping. This follows by induction, since $\mathcal{D}_1(\cdot, \phi)$ is the identity and thus Hadamard differentiable, and the integral operator that transforms $\mathcal{D}_{j-1}(\cdot, \phi)$ into $\mathcal{D}_j(\cdot, \phi)$ is linear. It thus follows from Theorem 2 that

$$\sqrt{n}\mathcal{D}_j(\cdot, \hat{F}_Y^* - \hat{F}_Y) = \mathcal{D}_j(\cdot, \sqrt{n}\hat{\Delta}_F) \Rightarrow \mathcal{D}_j(\cdot, (1,-1)\mathbb{F}_o).$$

Furthermore, since the map $\phi \mapsto \sup_y |\phi(y)|$ is continuous, it follows from the CMT that

$$KS_j \xrightarrow{d} \sup_y \mathcal{D}_j(y, (1,-1)\mathbb{F}_o) = \mathbb{KS}_j.$$

On the other hand, invoking similar arguments and Theorem 3, we obtain for the bootstrapped test statistic that

$$KS_{j,b} \xrightarrow{d} \mathbb{KS}_j, \tag{2.6.5}$$

conditional on the sample, in probability. Using the same arguments as in Barrett and Donald (2003, p.102), the distribution of $KS_{j,b}$ is absolutely continuous on $(0,\infty)$, and $c_j(\alpha)$ defined by $P(KS_j > c_j(\alpha)) = \alpha$ is finite for $\alpha < 1/2$. Then the event that $\hat{p}_j < \alpha$ is equivalent to the event $KS_j > \hat{c}_j(\alpha)$, where

$$\inf\{t : P_b(KS_{j,b} > t) > \alpha\} = \hat{c}_j(\alpha) \xrightarrow{p} c_j(\alpha)$$

by (2.6.5). Then, under the least favorite null hypothesis,

$$\begin{aligned}
\lim_{n\to\infty} P(\text{reject } H_0^j) &= \lim_{n\to\infty} P(KS_j > \hat{c}_j(\alpha)) \\
&= \lim_{n\to\infty} P(KS_{j,b} > c_j(\alpha)) + \lim_{n\to\infty} (P(KS_j > \hat{c}_j(\alpha)) - P(KS_j > c_j(\alpha))) \\
&= \alpha + \lim_{n\to\infty} P(KS_j \in (\hat{c}_j(\alpha), c_j(\alpha))) \\
&= \alpha
\end{aligned}$$

since $\hat{c}_j(\alpha) \xrightarrow{p} c_j(\alpha)$. This proves assertion (i) of the proposition. Under the alternative, there is an additional drift term such that the distribution of $KS_{j,b}$ diverges, and $\lim_{n\to\infty} P(\text{reject } H_0^j) = 1$ in this case, since $c_j(\alpha)$ is finite. This proves assertion (ii). □

Influence Function of the Lorenz Curve. The influence function ψ^L of the bivariate Lorenz process has a lengthy expression, but can be calculated using standard rules of calculus. First, one has to compute the influence function ψ^H of the Gaussian process $\mathbb{H}_o$ defined in Proposition 5 by integrating over the influence function ψ^Q of the quantile process. It then follows from the product rule that

$$\begin{aligned}
\psi_1^L(p, Z_i) &= \frac{1}{\mu_Y^*}\psi_1^H(p, Z_i) - \frac{L_Y^*(p)}{\mu_Y^*}\psi_1^H(1, Z_i), \\
\psi_2^L(p, Z_i) &= \frac{1}{\mu_Y}\psi_2^H(p, Z_i) - \frac{L_Y(p)}{\mu_Y}\psi_2^H(1, Z_i).
\end{aligned}$$

Using the definiton that $\alpha(p,x) = Q_Y^*(p)F_{Y|X}(Q_Y^*(p),x) - \mathbb{E}(\mathbb{I}\{Y \leq Q_Y^*(p)\}Y|X = x)$, the bivariate function ψ^H can be written as

$$\begin{aligned}
\psi_1^H(p,Z_i) &= (pQ_Y^*(p) - H_Y^*(p)) - \alpha(p,X_i^*) \\
&\quad - \frac{f_X^*(X_i)}{f_X(X_i)}\left(\mathbb{I}\{Y \leq Q_Y^*(p)\}(Q_Y^*(p) - Y_i) - \alpha(p,X_i)\right), \\
\psi_2^H(p,Z_i) &= (pQ_Y(p) - H_Y(p)) - \left(\mathbb{I}\{Y \leq Q_Y(p)\}(Q_Y(p) - Y_i)\right),
\end{aligned}$$

where $H_Y(p) = \int_0^\tau Q_Y(\tau)d\tau$ and $H_Y^*(p)$ is defined analogously. Noting that $\alpha(1,x) = E(Y|X = x)$, we furthermore obtain that

$$\begin{aligned}
\psi_1^H(1,Z_i) &= (\bar{y} - \mu_Y^* - \alpha(1,X_i^*)) - \frac{f_X^*(X_i)}{f_X(X_i)}(\bar{y} - Y_i - \alpha(1,X_i)), \\
\psi_2^H(1,Z_i) &= Y_i - \mu_Y,
\end{aligned}$$

which completes the description of the components of ψ^L. □

Chapter 3

Semiparametric Estimation of Binary Response Models with Endogenous Regressors

3.1 Introduction

This paper is concerned with the semiparametric estimation of the coefficients of a single index binary response model with endogenous regressors when identification is achieved via the control function approach put forward by Blundell and Powell (2004). The type of model we consider is of the form

$$Y = \begin{cases} 1 & \text{if } Y^* = X'\theta_o - U > 0 \\ 0 & \text{else,} \end{cases}$$

where Y is an indicator of the sign of a latent variable Y^* generated through a linear model with regressors X, vector of parameters θ_o and error term U. Our interest is in the estimation of the (normalized) coefficients θ_o, which is a semiparametric problem in the sense that the distribution of the unobservable variables is not assumed to belong to some parametric family. Furthermore, we do not assume that the error is independent of the regressors since we want to allow some components of X to be endogenous and thus correlated with U. To account for endogeneity, a control function approach introduces additional control variables, such as residuals from a reduced form of the endogenous

variables for example, as covariates into the outcome equation. Within this class of models, the only estimator that has been suggested so far is the one proposed by Blundell and Powell (2004), which is an extension of the Ahn, Ichimura, and Powell (1996) "matching" estimator.

This paper contributes to the literature by proposing a new two-step semiparametric maximum likelihood (SML) estimator. The procedure, which is also suggested but not further developed in Blundell and Powell (2004), is an extension of the Klein and Spady (1993) estimator, which achieves the semiparametric efficiency bound in the exogenous case. The first step consists of estimating the control variables through an auxiliary regression, which can either be fully nonparametric, or incorporate some parametric restrictions. In the second step, these are added nonparametrically to the equation of interest, which is in turn estimated by semiparametric maximum likelihood. Compared with the Blundell-Powell estimator, our procedure exploits the restrictions implied by the model more effectively, and does not require high-dimensional smoothing. The estimator possesses the classic asymptotic properties of $\sqrt{n}$-consistency and asymptotic normality, and valid standard errors and test statistics can be obtained via a nonparametric bootstrap procedure. Through a simulation study, we show that using our SML approach yields a considerable gain in terms of finite sample performance over other existing semiparametric estimators for binary choice models with endogenous regressors in many empirically relevant settings. The procedure should thus be appealing to applied researchers.

Binary response models play a prominent role in microeconometrics and are therefore the focus of an extensive literature. Estimation is typically carried out using standard Logit or Probit procedures, assuming that the distribution of the error term follows some parametric law and that X and U are independent. Having an estimator like ours that relies on neither of these two assumptions is of considerable practical importance since both might be inappropriate for many empirical applications.

First, economic theory usually provides no guidance about the functional form of the distribution of the error term, but misspecifications will generally result in inconsistent estimates for likelihood-based approaches. A number of semiparametric estimators have therefore been proposed which do not impose parametric restrictions on the distribution

of U. Such estimators include Semiparametric Least Squares (Ichimura 1993), Semiparametric Maximum Likelihood (Klein and Spady (1993), Ai (1997)), Average Derivative estimators (Stoker (1986), Powell, Stock, and Stoker (1989)), the Maximum Score estimator (Manski 1975) and the semiparametric estimator for discrete regressors of Horowitz and Härdle (1996), to mention a few.

Second, when the binary choice model arises in the context of a system of triangular or fully simultaneous equations, or certain measurement error models, some components of X will typically be endogenous, violating the independence assumption. Although neglecting this problem will again render the usual estimates inconsistent, it has received much less attention in literature. If one has access to an instrumental variable, an *ad-hoc* solution often recommended in econometrics textbooks would be to estimate a linear probability model by two-stage least squares (2SLS), although this procedure is generally inconsistent and might imply choice probabilities that are not between 0 and 1. More adequate estimators that are widely used have been proposed by Smith and Blundell (1986), Rivers and Vuong (1988) and Newey (1987), but they require fairly strong parametric distributional assumptions.

A semiparametric way of recovering the index coefficients that does not assume the unobservables to follow any parametric law is provided by Newey (1985). The approach requires a correctly specified parametric reduced form with homoskedastic error terms, where in particular the latter condition can be restrictive in practice. More recently, Lewbel (2000) proposed a simple to implement semiparametric procedure for estimating θ_o when X contains a continuously distributed, strictly exogenous "special regressor" that satisfies a large support condition. While this approach has the advantage that it allows the endogenous variable to be discrete or even binary, in many applications there might be no exogenous variable which qualifies as a "special regressor".

The control function approach that we use in this paper was proposed by Blundell and Powell (2004). The general idea of using residuals from a reduced form of the regressors to account for endogeneity is well established in parametric econometrics and has recently been used in the identification and estimation of various non- and semiparametric models with endogenous regressors (e.g. Newey, Powell, and Vella (1999), Blundell and Powell (2003), Chesher (2003), Das, Newey, and Vella (2003), Florens, Heckman, Meghir, and

Vytlacil (2008), Imbens and Newey (2009), Blundell and Powell (2007), Lee (2007)). It has the drawback that it requires the endogenous regressor to be continuously distributed, but other variables, including the instruments, can well be discrete.

The plan of this paper is as follows. In the next section, we specify the model being used. In Section 3, we show how identification is achieved and describe our SML approach to estimation. Asymptotic properties of our estimator are analyzed in Section 4. In Section 5, we discuss a number of extensions of our setup, while Section 6 deals with implementation issues and presents the results of our simulation study. The application of our procedure is illustrated via an empirical example in Section 7. Finally, Section 8 concludes.

3.2 The Model

The setup we consider in this paper is a linear single-index binary response model with an arbitrary large number of endogenous regressors, similar to the one of Blundell and Powell (2004). It is given by:

$$Y \quad = \quad \mathbb{I}\{X'\theta_o - U \geq 0\}, \tag{3.2.1}$$

where Y is the binary dependent variable, X is d_x-dimensional vector of regressors, U is an unobserved random error term, and $\mathbb{I}\{A\}$ is the indicator function that equals 1 when A is true and 0 otherwise. Furthermore, there is a d_e-dimensional subvector X^e of X that contains the endogenous variables, in the sense that these are potentially correlated with U. We think of (3.2.1) as a structural equation, describing the causal relationship between the right-hand and left-hand side variables, and refer to it in the following as the outcome equation.

Since it is clear from the exogenous case that we can only hope to identify the index coefficients θ_o up to a multiplicative constant, we normalize the coefficient on the first component of X to unity, i.e. we assume that $\theta_o = (1, \beta_o)$.[1] The object of interest that we want to estimate is the remaining vector of coefficients β_o. Also, for notational convenience, we use $X\beta_o$ as a shorthand for $(1, \beta_o)'X$.

[1] This choice is of course totally arbitrary. In general, we could normalize the coefficient on any of the regressors as long we can be sure that its true value is different from zero.

Without making further assumptions, it is generally not possible to identify β_o in equation (3.2.1). To this end, we assume the existence of a control variable. That is, we assume that U and X are independent conditional on some (unobserved) random d_v-vector V, that can be written as an identified function of X^e and some vector of exogenous instruments Z, which may include some of the exogenous components of X:

$$U \perp X|V \text{ for some } V = v_o(X^e, Z). \tag{3.2.2}$$

Such a control variable can be available under various circumstances, but the specific source is not important for the construction and analysis of our estimator. We only require that the function v_o is identified and can be estimated by some $\hat{v}$ satisfying a "high-level" condition given below, which can be easily verified under very general circumstances.

The leading case in which such a control variable will typically be available is when the endogenous regressors are generated through a second equation as

$$X^e = m_o(Z) + V, \quad \mathbb{E}(V|Z) = 0, \tag{3.2.3}$$

where m_o is a conditional mean function. This function can either be left unspecified, in which case (3.2.3) is the standard nonparametric regression model, or assumed to satisfy some parametric or semiparametric restrictions. For example, it is possible to specify (3.2.3) as a single-index model, with $m_o(Z) = \tilde{m}_o(Z'\alpha_o)$ for some unknown function $\tilde{m}_o$ and an unknown vector of parameters α_o, or as a fully parametric nonlinear regression model, with $m_o(Z) = \tilde{m}(Z, \alpha_o)$ for a function $\tilde{m}$ that is known up to a finite dimensional parameter α_o.

It has been shown by Blundell and Powell (2004) that under the distributional exclusion restriction that

$$\Pr(U < c|X, Z) = \Pr(U < c|V), \tag{3.2.4}$$

for all c, the error term $V = X^e - m_o(Z) \equiv v_o(X^e, Z)$ is a control variable that satisfies condition (3.2.2). This restriction is more flexible than a "full independence" condition like $(U, V) \perp Z$, since it allows for example the variance of V to be a function of the instruments. However, it retains the general drawback of the control function approach

that one has to correctly specify the relevant instrumental variables Z in (3.2.3), and that the endogenous regressor has to be continuous, since otherwise the distribution of V and thus its relation with U will in general depend upon Z, which violates (3.2.4).

A specification like (3.2.3)–(3.2.4) is plausible in a number of contexts. For example, equations (3.2.1) and (3.2.3) could be seen as a triangular system of structural equations, with (3.2.3) describing the causal mechanism that determines the values of the endogenous regressor. Alternatively, such a specification could also arise when the latent variable Y^* and X^e are jointly determined through a system of simultaneous equations. In this case, equation (3.2.3) would be a reduced form equation resulting from an equilibrium condition. Another option would be a classical measurement error framework such as

$$\begin{aligned} Y &= \mathbb{I}\{\tilde{X}^{e\prime}\theta_{o1} + Z_1'\theta_{o2} - \epsilon_1 \geq 0\} \\ X^e &= \tilde{X}^e + \epsilon_2 \\ \tilde{X}^e &= m_o(Z) + \epsilon_3, \end{aligned}$$

where X^e is a noisy version of the unobserved regressor $\tilde{X}^e$ measured with error ϵ_2. This model is equivalent to (3.2.1) and (3.2.3) with $U = \epsilon_1 + \epsilon_2$ and $V = \epsilon_2 + \epsilon_3$.

While in this paper we will focus on control variables emerging from a structure like the one in (3.2.3), they might also appear under different circumstances, as pointed out by Imbens and Newey (2009). For example, as shown in Newey (2007), in a sample selection model where Y is only observed conditional on a selection variable $S = \mathbb{I}\{m(Z) > U^*\}$ being equal to one, and (U, U^*) is independent of Z, the selection probability $P = \Pr(S = 1|Z)$ is a control variable in the sense of condition (3.2.2). Such models can hence be treated in our framework as well.

3.3 Identification and Estimation Approach

3.3.1 Identification

The most important consequence of the restriction (3.2.2) is that the conditional expectation of the dependent variable Y given the observable variables X and V can be written as a function of the linear index $X\beta_o$ and the control variables V. Denoting the

conditional distribution function of U given V by G_o, we can write

$$\mathbb{E}(Y|X,V) = \mathbb{E}(\mathbb{I}\{U \leq X\beta_o\}|X,V) = \mathbb{E}(\mathbb{I}\{U \leq X\beta_o\}|V) = G_o(X\beta_o, V), \qquad (3.3.1)$$

and thus reduce the dimension from $d_x + d_v$ to $1 + d_v$.

This restriction is also useful for identifying β_o. In particular, it is clear that our parameter of interest is identified by the data if the following condition holds:

Identification Condition (IC). *There exists a unique interior point $\beta_o \in \mathcal{B}$ such that the relationship $\mathbb{E}(Y|X,V) = \mathbb{E}(Y|X\beta_o, V)$ holds for $(X,Z) \in \mathcal{A}$, a set with positive probability.*

Thus, what remains to establish identification of β_o is to give conditions on the primitives of the model under which IC is fulfilled. It turns out that for this purpose, in addition to requiring that v_o is identified, only the standard regularity conditions for identification of single-index binary response models are needed. The reason is that we are not dealing with an actual multiple-index model: although the function G_o has $1+d_v$ arguments, only the first one contains index parameters to be identified. We therefore have the following theorem.

Theorem 1 (Identification)**.** *The parameter β_o in the model* (3.2.1)–(3.2.2) *is identified in the sense that the identification condition IC holds, if the following conditions are satisfied:*

i) The function G_o is differentiable and strictly increasing in its first argument on a set $\mathcal{A}$ with positive probability under the distribution of X.

ii) Conditional on the control variable V, the vector X contains at least one continuously distributed component $X^{(1)}$ with nonzero coefficient.

iii) The span of the remaining components $X^{(-1)}$ contains no proper linear subspace which has probability 1 under the distribution of X.

The proof, which is analogous to the argument in Manski (1988), is given in the appendix. Note that when the control variables emerge from a structure like (3.2.3)–(3.2.4), the fact that the endogenous regressors are continuously distributed is not sufficient for

condition (ii) to be fulfilled. Instead, it is required that additionally either one of the exogenous regressors or the "fitted value" $m_o(Z)$ from the reduced form is continuously distributed as well[2]. To see this, assume that all exogenous regressors and instruments are discrete. Then $X = (X^e, X^{(-e)}) = (m_o(Z)+V, X^{(-e)})$ is discretely distributed conditional on V, which violates condition (ii).

3.3.2 The Estimator

To motivate the estimator, assume for the moment that the function G_o was known and that V was observable. If observations are stochastically independent, it would then be straightforward to estimate β_o by maximizing the log-likelihood function

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \log(G_o(X_i\beta, V_i)) + (1 - Y_i)\log(1 - G_o(X_i\beta, V_i)) \tag{3.3.2}$$

with respect to β. When G_o and V are unknown, this approach is clearly not feasible. However, generalizing the idea of Klein and Spady (1993), we can approximate the objective function by replacing all unknown quantities with appropriate estimates.

To make this idea more precise, we have to introduce some notation. For any candidate value of β and some function v, define $W(\beta, v) = (X\beta, v(X^e, Z))$, and set

$$G(w|\beta, v) = \mathbb{E}(Y|W(\beta, v) = w).$$

Furthermore, we use the convention that arguments indexing a function are dropped when they are evaluated at their true value, i.e. $G(w|\beta) = G(w|\beta, v_o)$, $G(w) = G(w|\beta_o)$, $W(\beta) = W(\beta, v_o)$, $W = W(\beta_o)$ etc. Using this notation, we have that $G_o(X\beta_o, V) = G(W(\beta_o, v_o)|\beta_o, v_o) \equiv G(W)$. The idea is now to replace the term $G_o(X_i\beta, V_i)$ in (3.3.2) by a nonparametric kernel estimate $\hat{G}(\hat{W}_i(\beta)|\beta, \hat{v})$, where $\hat{W}(\beta) = W(\beta, \hat{v})$ and $\hat{v}$ is itself a (possibly nonparametric) estimate of v_o from a preliminary estimation stage. Note that the function $G(W(\beta, v)|\beta, v)$ and its estimate depend on β both through its first argument, which determines the point of evaluation, and its second one, which influences the shape of the function.

Since we have made no assumptions about the structural form generating the control variates, we also do not impose a specific estimation procedure. Instead, we simply assume

[2]I would like to thank an anonymous referee for pointing this out.

the existence of an estimator $\hat{v}$ of v_o satisfying some high-level conditions given below. Then for any β and v, a nonparametric kernel estimate of $G(\cdot|\beta, v)$ can be obtained as

$$\hat{G}(w|\beta, v) = \hat{N}(w|\beta, v)/\hat{D}(w|\beta, v)$$

where

$$\begin{aligned}\hat{N}(w|\beta, v) &= \frac{1}{n}\sum_{j=1}^{n} K_h\left(W_i(\beta, v) - w\right) Y_j, \\ \hat{D}(w|\beta, v) &= \frac{1}{n}\sum_{j=1}^{n} K_h\left(W_i(\beta, v) - w\right).\end{aligned}$$

Here $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function on $\mathbb{R}^{1+d_v}$ and h is a bandwidth sequence that goes to zero as n goes to infinity. The exact specifications are given below. Substituting this estimate into equation (3.3.2) we obtain the semiparametric likelihood function

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tau_i(Y_i \log(\hat{G}(\hat{W}_i(\beta)|\beta, \hat{v})) + (1 - Y_i)\log(1 - \hat{G}(\hat{W}_i(\beta)|\beta, \hat{v}))),$$

and define our estimator $\hat{\beta}$ of β_o as the maximizer of this objective function:

$$\hat{\beta} = \underset{\beta \in \mathcal{B}}{\operatorname{argmax}}\, L_n(\beta). \tag{3.3.3}$$

Here $\tau_i = \mathbb{I}\{(X_i, Z_i) \in \mathcal{X}\}$ is a trimming term that equals 1 whenever the values of (X_i, Z_i) lie within an appropriately chosen compact set $\mathcal{X}$ and 0 otherwise. In particular, the set is chosen such that the probability limit of $\hat{G}$ is bounded away from zero and one on $\mathcal{X}$.

While the maximization in (3.3.3) can be carried out using standard numerical optimization procedures, it is certainly computationally expensive, since we have to run n nonparametric regressions for *every* iteration step. A further complication is the possible presence of local maxima in the objective function. We discuss these issues in more detail in the simulation study.

3.4 Asymptotic Properties

In this section, we establish the asymptotic properties of our estimator. We start with stating the assumptions and then give results on consistency, asymptotic normality and variance estimation. Here we only sketch our proofs and delegate rigorous arguments to the Appendix.

3.4.1 Assumptions and Preliminaries

Before we present our framework, we have to introduce some more notation. For μ a k-vector of nonnegative integers, we define (i) $|\mu| = \sum_{i=1}^{n} \mu_i$, (ii) for any function $f(x)$ on $\mathbb{R}^k$, $\partial_x^\mu f(x) = \partial^{|\mu|}/(\partial^{\mu_1} x_1, \ldots, \partial^{\mu_k} x_k) f(x)$ and (iii) $x^\mu = \prod_{i=1}^{n} x_i^{\mu_i}$. Furthermore, we write ∂_k as a shorthand for ∂_{w_k} for $k = 1, 2$. We can now state the assumptions for our asymptotic analysis.

Assumption 1. *The sample observations $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ are a sequence of independent and identically distributed random vectors generated according to the model defined in equation* (3.2.1) – (3.2.2)*. The model is identified in the sense that IC holds.*

Assumption 2. *The parameter space $\mathcal{B}$ is a compact subset of $\mathbb{R}^{d_x - 1}$ and β_o is an element of its interior.*

These are standard regularity conditions in the semiparametrics literature.

Assumption 3. *i) For all $\beta \in \mathcal{B}$, the distribution of the random vector $W(\beta)$ admits a density function $D(w|\beta)$ with respect to the Lebesque measure.*

ii) For all $\beta \in \mathcal{B}$, $D(w|\beta)$ is r times continuously differentiable in w, and the derivatives are uniformly bounded: $\sup_{w,\beta} |\partial_w^\mu D(w|\beta)| < \infty$ $\forall \mu$ with $|\mu| \leq r$.

iii) For all $\beta \in \mathcal{B}$, $G(w|\beta)$ is r times continuously differentiable in w, and the derivatives are uniformly bounded: $\sup_{w,\beta} |\partial_w^\mu G(w|\beta)| < \infty$ $\forall \mu$ with $|\mu| \leq r$.

iv) $D(w|\beta)$ and $G(w|\beta)$ are twice continuously differentiable in β.

Assumption 3 collects some conventional smoothness restrictions on the functions being estimated through kernel methods. The higher-order differentiability conditions are needed to obtain certain uniform convergence rates on the estimates of $G(\cdot|\beta)$ and its derivatives.

Assumption 4. *For $\mathcal{X}$ a compact subset of the support of (X, Z), define $W(\mathcal{X}) = \{w \in \mathbb{R}^{1+d_v} : \exists (x, z) \in \mathcal{X}, \beta \in \mathcal{B}$ s.t. $w = (x\beta, v_o(x^e, z))\}$. Then $\mathcal{X}$ is chosen such that:*

i) $\inf_{w \in W(\mathcal{X}), \beta \in \mathcal{B}} D(w|\beta) > 0$

ii) $\inf_{w\in W(\mathcal{X}),\beta\in\mathcal{B}} G(w|\beta) > 0$ *and* $\sup_{w\in W(\mathcal{X}),\beta\in\mathcal{B}} G(w|\beta) < 1$.

Assumption 4 prescribes a fixed trimming procedure, which significantly simplifies the derivation of the asymptotic properties. Since trimming is generally considered to be of minor practical importance and thus is often disregarded in empirical applications, this seems to be a mild restriction. However, at the cost of a more complicated proof it would be possible to replace the fixed trimming function $\tau_i = \mathbb{I}\{(X_i, Z_i) \in \mathcal{X}\}$ with a random, data dependent one that tends to one as the sample size increases. Using results from e.g. Pakes and Pollard (1989), one could for example implement a trimming procedure on the basis of the upper and lower sample quantiles of the data, as in Lee (1995).

Assumption 5. *The matrix*

$$\Sigma = \mathbb{E}\left[\tau \partial_\beta G(W) \partial_{\beta'} G(W) \left(G(W)(1-G(W))\right)^{-1}\right]$$

is positive definite.

Assumption 5 ensures the non-singularity of the asymptotic covariance matrix of our final estimator. Note that here and in the following the notation $\partial_\beta G(W)$ is understood to denote the derivative of $G(W(\beta)|\beta)$ with respect to both occurrences of β, evaluated at $\beta = \beta_o$.

Assumption 6. *The kernel functions* $K : \mathbb{R}^{d_v+1} \to \mathbb{R}$ *satisfies (i)* $\int K(z)dz = 1$*, (ii)* $\int K(z)z^\mu dz = 0$ *for all* $|\mu| = 1, \ldots, r-1$*, (iii)* $\int |K(z)z^\mu| dz < \infty$ *for* $|\mu| = r$*, (iv)* $K(z) = 0$ *if* $|z| > 1$ *(v)* $K(z)$ *is* r *times continuously differentiable.*

Assumption 7. *The bandwidth vector* $h = (h_1, \ldots, h_{d_v+1})$ *satisfies* $h_i = c_i n^{-\delta}$*,* $i = 1, \ldots, d_v + 1$*, for some constants* $c_i > 0$ *and* δ *such that* $1/2r < \delta_i < 1/(2 + 6d_v)$*.*

The last two assumptions define a standard bias-reducing kernel of order r, which is used for reducing asymptotic bias in the estimates of G and its derivatives, and determine the rate at which the bandwidth sequences go to zero as $n \to \infty$. In order to ensure that the set of possible values for δ is not empty, a sufficient condition is that $r > 1 + 3d_v$.

Assumption 8. *i) The estimate* $\hat{v}$ *of* v_o *satisfies*

$$\hat{v}(X_i^e, Z_i) - v_o(X_i^e, Z_i) \equiv \hat{V}_i - V_i = \frac{1}{n}\sum_{j=1}^{n} \omega_n(Z_i, Z_j)\psi_j + r_{in},$$

with

$$\max_i \tau_i \|r_{in}\| = o_p(n^{-1/2}) \quad \text{and} \quad \max_i \tau_i \|\hat{V}_i - V_i\| = o_p(n^{-1/4}),$$

where $\psi_j = \psi(X_j^e, Z_j)$ *is an influence function with* $\mathbb{E}(\psi_j|Z_j) = 0$ *and* $\mathbb{E}(\psi_j^2|Z_j) < \infty$*, and the weights* $\omega_n(Z_i, Z_j)$ *satisfy* $\mathbb{E}(\|\omega_n(Z_i, Z_j)\|^2) = o(n)$*.*

ii) There exists a space $\mathcal{V}$*, such that* $\Pr(\hat{v} \in \mathcal{V}) \to 1$ *and* $\int_0^\infty \sqrt{\log N(\lambda, \mathcal{V}, \|\cdot\|_\infty)} d\lambda < \infty$*, where* $N(\lambda, \mathcal{V}, \|\cdot\|_\infty)$ *is the covering number with respect to the* L_∞*-norm of the class of functions* $\mathcal{V}$*, i.e. the minimal number of balls with* $\|\cdot\|_\infty$*-radius* λ *needed to cover* $\mathcal{V}$*.*

This assumption is a high-level condition on the estimator of the control variables. The first part states that the estimator admits a certain asymptotic expansion, whereas the second part requires the estimator to take values in some well-behaved function space with probability approaching 1.

These conditions can be shown to be fulfilled for various scenarios discussed in Section 2. For example, assume that $X^e = m_o(Z) + V$ with $E(V|Z) = 0$, $\hat{V}_i = \hat{v}(X_i^e, Z_i) = X_i^e - \hat{m}(Z_i)$, $\hat{m}$ is the usual Nadaraya-Watson estimator, and $\mathcal{V}$ is the class of all functions f taking the form $f(x^e, z) = x^e - g(z)$ for some function g whose partial derivatives up to order p exist and are uniformly bounded. Then, under certain assumptions on the kernel and the bandwidth, the first part of Assumption 8 is fulfilled with

$$\omega_n(Z_i, Z_j) = \kappa_b(Z_i - Z_j) f_Z(Z_i)^{-1} \text{ and } \psi_j = -(X_j^e - m_o(Z_j)) = -V_j,$$

where f_Z is the density function of the vector of instruments Z, κ is a kernel function and b is the bandwidth. Also, $\Pr(\hat{v} \in \mathcal{V}) \to 1$ in this case if the kernel function has uniformly bounded partial derivatives up to order p. The remaining requirement then follows from Corollary 2.7.4 in van der Vaart and Wellner (1996) if $p > d_z/2$. Similar arguments can also be used when m_o is specified in a semiparametric way, for example as a single-index or partially linear model, or when other nonparametric smoothers, such as local polynomials are used (see e.g. Kong, Linton, and Xia (2009)).

On the other hand, when $m_o(z) = m(z, \alpha_o)$ is known up to some vector of parameters α_o, under standard regularity conditions for nonlinear regression models we obtain that

part (i) is fulfilled with

$$\omega_n(Z_i, Z_j) = \partial_\alpha m(Z_i, \alpha_o)\mathbb{E}(\partial_\alpha m(Z, \alpha_o)\partial_\alpha m(Z, \alpha_o)')^{-1}\partial_\alpha m(Z_j, \alpha_o)' \text{ and } \psi_j = -V_j,$$

whereas part (ii) is true when m satisfies a Lipschitz conditions with respect to α, as shown van der Vaart and Wellner (1996, Theorem 2.7.11).

3.4.2 Consistency and Asymptotic Normality

To establish consistency, we take the usual route and first show that the estimated likelihood function $L_n(\beta)$ converges uniformly to a nonrandom limit function $L(\beta)$. Secondly, we show that this function attains a unique maximum at β_o, which implies both that β_o is identified and that $\hat{\beta}$ is consistent. This is formally stated in the following theorem:

Theorem 2 (Consistency)**.** *Under Assumptions 1 – 8, it holds that* $\hat{\beta} = \beta_o + o_p(1)$ *as* $n \to \infty$.

Showing that $\hat{\beta}$ is also asymptotically normal requires a somewhat more involved argument. Our strategy is to use general results on semiparametric estimation procedures given in Chen, Linton, and Van Keilegom (2003). As shown in the Appendix, this requires deriving uniform rates of convergence for the nonparametric estimates of the link function $G(\cdot|\beta)$ and its derivatives. This constitutes the main difficulty for the proof, since the estimates of $G(\cdot|\beta)$ are in turn based on possibly non- or semiparametrically generated regressors $\hat{V}$.

Intuitively, the asymptotic normality result follows from the following argument. From a standard Taylor expansion of the semiparametric score function $S_n(\beta) = \partial_\beta L_n(\beta)$ around the true parameter values β_o we obtain, after rearranging terms,

$$\sqrt{n}(\hat{\beta} - \beta_o) = -(\partial_{\beta,\beta} L_n(\bar{\beta}))^{-1}\sqrt{n}\partial_\beta L_n(\beta_o), \tag{3.4.1}$$

where $\bar{\beta}$ is some intermediate value between $\hat{\beta}$ and β_o. Starting with the first term on the right-hand-side of (3.4.1), it follows from the uniform convergence results on $\hat{G}(\cdot|\beta, \hat{v})$ and its derivatives, and the consistency of $\hat{\beta}$ and $\hat{v}$, that it converges in probability to some matrix, i.e.

$$\partial_{\beta\beta} L_n(\bar{\beta}) \xrightarrow{p} \Sigma,$$

where the limit is positive definite by Assumption 5. Continuing with the second term in (3.4.1), it is shown in the Appendix that

$$\begin{aligned}\sqrt{n}\partial_\beta L_n(\beta_o) = &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{Y_i - G(W_i)}{G(W_i)(1-G(W_i))}(\tau_i\partial_\beta G(W_i) - \mathbb{E}(\tau_i\partial_\beta G(W_i)|W_i)) \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\xi_{1i} - \xi_{2i})\psi_i + o_p(1),\end{aligned}$$

where ψ_i is the influence function from Assumption 8, and

$$\begin{aligned}\xi_{1i} &= \mathbb{E}(\tau\partial_\beta G(W)\partial_2 G(W)(G(W)(1-G(W)))^{-1}\omega_n(Z,Z_i)|Z_i), \\ \xi_{2i} &= \mathbb{E}(\mathbb{E}(\tau\partial_\beta G(W)|W)\partial_2 G(W)(G(W)(1-G(W)))^{-1}\omega_n(Z,Z_i)|Z_i).\end{aligned}$$

Taken together, and applying a Central Limit Theorem, we obtain the following result:

Theorem 3 (Asymptotic Normality)**.** *Under Assumptions 1–8, it holds that*

$$\sqrt{n}(\hat{\beta} - \beta_o) \xrightarrow{d} N(0, \Omega)$$

where

$$\Omega = \Sigma^{-1}(\Psi_1 + \Psi_2)\Sigma^{-1}$$

and

$$\begin{aligned}\Sigma &= \mathbb{E}\left[\frac{\tau\partial_\beta G(W_i)\partial_{\beta'} G(W_i)}{G(W_i)(1-G(W_i))}\right], \\ \Psi_1 &= \mathbb{E}\left[\frac{(\tau\partial_\beta G(W) - \mathbb{E}(\tau\partial_\beta G(W)|W))(\tau\partial_\beta G(W) - \mathbb{E}(\tau\partial_\beta G(W)|W))'}{G(W_i)(1-G(W_i))}\right] \\ \Psi_2 &= \mathbb{E}\left[(\xi_{1i} - \xi_{2i})\psi_i\psi_i'(\xi_{1i} - \xi_{2i})'\right].\end{aligned}$$

It is instructive to compare our asymptotic variance matrix to that of an infeasible maximum likelihood estimator using the true functions $G(\cdot|\beta)$ and v_o. If we define $\tilde{\Sigma}$ be equal to Σ with $\tau \equiv 1$, the variance of such an estimator would be $\tilde{\Sigma}^{-1}$. In general, our matrix Ω will be larger for two reasons. First, due to the fixed trimming procedure our estimator does not use all available observations, which obviously results in a loss of efficiency. Second, there is an additional penalty in terms of asymptotic variance for only using an estimate of the function v_o. However, there is no penalty term for estimating the unknown link function $G(\cdot|\beta)$, which is also the case when all regressors are exogenous.

To see this, let $\tilde{\Omega}$ be equal to Ω with $\tau \equiv 1$, and define $\tilde{\Psi}_1$, $\tilde{\Psi}_2$, $\tilde{\xi}_{1i}$ and $\tilde{\xi}_{2i}$ analogously. Then it follows from the fact that $\mathbb{E}(\partial_\beta G(W)|W) = 0$ (see Klein and Spady (1993, p. 403)) that $\tilde{\Sigma} = \tilde{\Psi}_1$ and $\tilde{\Psi}_2 = \mathbb{E}[\tilde{\xi}_{1i}\psi_i\psi_i'\tilde{\xi}_{1i}']$. Thus, if we neglect the effect of trimming, the asymptotic covariance matrix of our estimator would be $\tilde{\Omega} = \tilde{\Sigma}^{-1} + \tilde{\Sigma}^{-1}\tilde{\Psi}_2\tilde{\Sigma}^{-1}$, where the presence of the second term $\tilde{\Sigma}^{-1}\tilde{\Psi}_2\tilde{\Sigma}^{-1}$ is due to using an estimate of v_o. Since this term is generally nonnegative definite, the variance will be larger than it would be if v_o was known and thus the control variable V was observed.

3.4.3 Variance estimation

In order to be able to conduct inference on $\hat{\beta}$, an estimate of the asymptotic variance matrix is needed, but since Ω depends on a number of unknown functions in a relatively complicated way, a direct sample moment estimator would be hard to implement. However, the results in Chen, Linton, and Van Keilegom (2003) justify the use of an ordinary nonparametric bootstrap procedure to calculate confidence regions for the unknown parameters. To be specific, let $\{(Y_i^*, X_i^*, Z_i^*)\}_{i=1}^n$ be the bootstrap sample drawn randomly with replacement from the original data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, and let $\hat{v}^*$ and $\hat{G}^*(\cdot|\beta, v)$ be the same estimators as $\hat{v}$ and $\hat{G}(\cdot|\beta, v)$ but based on the bootstrap data. Also, define the bootstrap estimator $\hat{\beta}^*$ as

$$\hat{\beta}^* = \underset{\beta\in\mathcal{B}}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \tau_i^*(Y_i^* \log(\hat{G}^*(W_i(\beta, \hat{v}^*)|\beta, \hat{v}^*)) + (1 - Y_i^*)\log(1 - \hat{G}^*(W_i(\beta, \hat{v}^*)|\beta, \hat{v}^*))).$$

Then it can be shown using Theorem B in Chen, Linton, and Van Keilegom (2003) and similar arguments as in the proof of our Theorem 3, that $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ has the same asymptotic limiting distribution as $\sqrt{n}(\hat{\beta} - \beta_o)$.

A general disadvantage of using such resampling techniques for a semiparametric optimization estimator like ours is that they can be extremely costly from a computational point of view. For practical applications, the following approximation might thus be useful. Note that the complicated functional form of Ω is mainly an effect of the fixed trimming procedure. Yet when only a small amount of observations is trimmed, this effect should be small. In particular, when $\tau = 1$ for most observations, then $\mathbb{E}(\tau\partial_\beta G(W)|W)) \approx 0$ and by continuity the matrix Ω can be well approximated by

$\bar{\Omega} = \Sigma^{-1} + \Sigma^{-1}\bar{\Psi}_2\Sigma^{-1}$, where $\bar{\Psi}_2 = \mathbb{E}[\xi_{1i}\psi_i\psi_i'\xi_{1i}']$. Under our assumptions stated above, the matrix Σ can be consistently estimated by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \tau_i \frac{\partial_\beta \hat{G}(W_i(\hat{\beta}|\hat{\beta},\hat{v})\partial_{\beta'}\hat{G}(W_i(\hat{\beta})|\hat{\beta},\hat{v})}{\hat{G}(W_i(\hat{\beta}|\hat{\beta},\hat{v})(1-\hat{G}(W_i(\hat{\beta})|\hat{\beta},\hat{v}))}.$$

Estimating the matrix $\bar{\Psi}_2$ is more difficult when using only the "high-level" condition on the control function from Assumption 8. However, when imposing more structure on the estimates of the control variables, the shape of the terms ξ_i and ψ_i can usually be made more explicit, and thus suggest a potential estimator. Consider for example the case where $\hat{V}_i = X_i^e - \hat{m}(Z_i)$ is the residual from a nonparametric reduced for equation estimated by some kernel method, as in (3.2.3). Then $\psi_i = -V_i$, and it is easy to show that $\xi_{1i} = \mathbb{E}(\tau_i\partial_\beta G(W_i)\partial_2 G(W_i)(G(W_i)(1-G(W_i)))^{-1}|Z_i)$. Accordingly, one could estimate $\bar{\Psi}_2$ by

$$\hat{\bar{\Psi}}_2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_{1i}\hat{\psi}_i\hat{\psi}_i'\hat{\xi}_{1i}',$$

where $\hat{\psi}_i = -\hat{V}_i$, and $\hat{\xi}_{1i}$ is defined as the fitted value of some nonparametric kernel regression of $\tau_i\partial_\beta\hat{G}(W)\partial_2\hat{G}(W)(\hat{G}(W)(1-\hat{G}(W)))^{-1}$ on Z. Under suitable regularity conditions, one can verify that a Law of Large Numbers holds for $\hat{\bar{\Psi}}_2$ in this case.

3.5 Some Extensions of the Structure of the Model

For the ease of exposition, we have chosen a formulation our model in Section 2 that is simple but also restrictive in many ways, yet various aspects can easily be generalized. First, the linear relationship in the outcome equation (3.2.1) could be replaced with a nonlinear one, such as

$$Y = \mathbb{I}\{g(X,\theta_o) - U > 0\}$$

for some known function g, at the cost of a slightly more complicated normalization of the parameters (see Ichimura 1993, Klein and Spady 1993).

Second, we could replace the conditional independence restriction in (3.2.2) with the alternative, slightly weaker version

$$U \perp X|(V, X\beta_o). \tag{3.2.2b}$$

This would allow for a limited degree of dependence between X and U even when conditioning on the control variable V, as long as this dependence is restricted to run through the index values (as would be the case under index heteroskedasticity, for example). In particular, it would still be possible to write $\mathbb{E}(Y|X,V)$ as some function G_o of $X\beta_o$ and V, but now this function would not be confined to be monotone in its first argument. As our estimator (in contrast to the one of Blundell and Powell) does not explicitly use the properties of a distribution, it automatically works under (2.2b) as well. We illustrate this point in more detail in our simulation study.

Finally, in this paper we focus on the estimation of the (normalized) index coefficients β_o. Another object of practical interest one could consider would be the choice probability for some exogenously determined value of the regressors $X = \bar{x}$. Blundell and Powell (2004) call this the *average structural function (ASF)*, and show that it is identified as the partial mean of G_o with respect to the distribution of the control variable V,

$$ASF(\bar{x}) = \int G_o(\bar{x}\beta_o, V)dF_V, \tag{3.5.1}$$

provided that the support of V does not vary with $\bar{x}$. The estimation of this object is discussed in more detail in Imbens and Newey (2009).

3.6 Simulation Study

3.6.1 Setup

In order to demonstrate the usefulness of our proposed estimator for applications to finite samples, we report the results of three simulation experiments in this section. Apart from our SML procedure, we also consider Blundell and Powell's (2004) semiparametric "matching" estimator, the "Two-Stage-Probit" estimator of Smith and Blundell (1986) or Rivers and Vuong (1988), and Two-Stage Least Squares (2SLS) estimation of a linear probability model, which is frequently used in applied work. These are intended to serve as points of reference.

For the three simulations, we always use the same specification for the regressors and instruments, but change the properties of the joint distribution of the error terms (U, V). The dependent variable is generated by a binary response model with two covariates in

the outcome equation, of which one is endogenous, and two additional instruments in a linear reduced form equation:

$$\begin{aligned} Y_1 &= \mathbb{I}\{X^e + Z_1\beta_o > U\}, \\ X^e &= \alpha_{o0} + Z_1\alpha_{o1} + Z_{21}\alpha_{o2} + Z_{22}\alpha_{o3} + V. \end{aligned}$$

The true parameter values $\beta_o = 1$ and $\alpha_o = (1, 2/3, 2/3, 1/3)'$ are held constant across simulations. The exogenous variables are independent, with Z_1 being exponentially distributed, truncated from above at 3, and standardized to have mean zero and variance two, and Z_{21}, Z_{22} are standard normal. In order to ensure a sensible comparison, all estimators are based on the OLS residuals from the reduced form equation. For the error distributions, we simulate V as $N(0,1)$ and $U = U^* + V$, where we use the following specifications for U^*:

- Design I: $U^* \sim N(0, 5)$
- Design II: $U^* \sim 0.8N(-1, .6) + 0.2N(4, 2)$
- Design III: $U^* \sim N(0, \exp(0.1 + 0.5X\beta_o))$

Design I implies a jointly normal distribution of (U, V) and is the one under which a Probit should give the best results. The second design is a mixture of two normal distributions, resulting in a right-skewed and bimodal density of U. It is constructed such that the Probit estimator should be markedly biased, and we thus expect a comparatively better performance of the semiparametric procedures. For the third design, the variance of U conditional on V is a function of the linear index. It is included to show that our estimator also works when the restriction in (3.2.2) is replaced with its weaker version (2.2b) (see section 5).

While these designs correspond to very different distributions, they are chosen such that some features are approximately the same. In particular, it holds that $\text{Var}(U) \approx 6, \text{Var}(Y_2 + Z_1) \approx 4.5$, $\text{Cor}(U, V) \approx 0.4$ and $\text{Cor}(U, Y_2) \approx 0.25$. With the multiple R^2 in the reduced form regression being about 0.6, we are in a situation with relatively strong instruments. In all three cases, we consider the sample sizes $n = 250, 500, 1000$, and set the number of replications to 1000.

3.6.2 Implementation Issues

In order to implement our SML estimator, we have to select a kernel function and the bandwidth parameters. In particular, our assumptions require the use of higher-order kernels to eliminate asymptotic bias. However, when using higher-order kernels to calculate $\hat{G}(\cdot|\beta, \hat{v})$, some observations will be given a negative weight and the result is not confined to be between zero and one, which of course causes problems when taking logarithms. For our simulations, we therefore consider two approaches to circumvent this this problem. The first one employs an idea from Klein and Spady (1993) and consists of minimizing a modified criterion function $\tilde{L}_n$, where

$$\tilde{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tau_i(Y_i \log(\hat{G}(\hat{W}_i(\beta)|\beta)^2) + (1 - Y_i)\log((1 - \hat{G}(\hat{W}_i(\beta, \hat{v})|\beta, \hat{v}))^2)).$$

The corresponding estimator $\tilde{\beta}$ can easily be shown to be consistent and having the same limiting distribution as our SML estimator $\hat{\beta}$. In particular, note that both are solutions of the same first-order condition. We refer to this estimator as SML-1 below.

As a second possibility, we simply compute our estimator as described above, but without the use of higher-order kernel functions. This is motivated by the frequently made observation that while higher-order kernel might be required from a theoretical point of view in many semiparametric applications, the resulting estimators often tend to have inferior finite sample properties compared to those based on standard kernels (see Marron (1994) or Jones and Signorini (1997)). Thus, although strictly speaking not compatible with our asymptotic analysis, we also consider this approach for our simulations. It is referred to as SML-2 below.

Regarding the choice of the bandwidth parameters $h = (h_1, h_2, \ldots, h_{d_v+1})$, for our simulation study we follow Härdle, Hall, and Ichimura (1993) and Delecroix, Hristache, and Patilea (2005), and consider the following pragmatic approach: we treat the bandwidths as additional parameters of the estimated likelihood and perform the maximization with respect to both β and h. That is, we use the first component of

$$\left(\hat{\beta}, \hat{h}\right) = \underset{(\beta,h)\in B\times\mathbb{R}_+^{d_v+1}}{\operatorname{argmax}} L_n(\beta, h)$$

as our estimator. While we do not claim any optimality of this approach for our problem at hand, the method seems to perform well in applications to finite samples, as shown by

our simulation study. A further advantage is that it can also serve as an informal test for endogeneity: when X^e is actually exogenous, typically a large value will be chosen for the bandwidth, because in this case $G_o(X\beta, V)$ does not vary with V. As an alternative, one could also experiment with various multiples of $n^{-\delta}$, but practitioners are generally reluctant to do so because it involves a large degree of subjectivity.

In line with most of the literature in this field, no trimming is used. We investigated various forms of trimming, but found no substantial effect on the performance of the estimator in our simulation scenarios. This result is common when evaluating the finite sample properties of semiparametric estimators of single-index models in general. However, the use of trimming might be beneficial in practice if the data contains some extreme outliers, as they can have a substantial impact on the estimate of the link function and the chosen bandwidth. In this case, a trimming procedure could for example be implemented on the basis of the upper and lower sample quantiles of the data, as mentioned above[3].

The numerical optimization is carried out using a Gauss-Newton type algorithm as implemented in the software package *R* 2.5.1. We use the Probit results as starting values for the index coefficients and .4 for the bandwidths. To guard against the algorithm converging to possible local maxima, we also use half and twice the starting values values to compute the estimator, and retain the result that gives the highest value of the objective function. However, it turns out that in our simple setup the values coincide in most runs.

All other estimators were implemented as described in the respective literature. For the Blundell-Powell estimator, we use Least Squares Cross Validation to determine the bandwidth for the nonparametric regression part, and $1.06\sigma_w n^{-1/5}$ for the "matching" part, which corresponds to the specification in their empirical application.

3.6.3 Results

To facilitate comparison of our SML estimator with the other procedures, we make use of a different normalization than the one described in Section 2: instead of setting the coefficient of the endogenous variable to one, we rescale the estimates of the coefficients

[3]In the presence of extreme outliers, the use of trimming should of course be beneficial even for correctly specified parametric estimators

Table 3.1: Simulation Results Design I

		MEAN	SD	RMSE	MAD	25%	50%	75%	CR
$n = 250$	SML-1	1.073	0.443	0.449	0.366	0.722	1.000	1.336	0.997
	SML-2	1.008	0.248	0.248	0.199	0.854	1.012	1.177	0.895
	Probit	1.011	0.187	0.188	0.149	0.889	1.007	1.122	0.897
	2SLS	1.089	0.186	0.206	0.165	0.956	1.090	1.204	0.844
	BP	0.735	0.389	0.471	0.391	0.459	0.722	0.969	0.787
$n = 500$	SML-1	1.061	0.463	0.467	0.371	0.672	0.999	1.333	0.975
	SML-2	1.001	0.163	0.163	0.131	0.895	1.002	1.116	0.913
	Probit	0.999	0.137	0.137	0.111	0.901	1.000	1.093	0.883
	2SLS	1.079	0.138	0.159	0.128	0.979	1.084	1.172	0.789
	BP	0.812	0.287	0.343	0.279	0.602	0.810	1.027	0.675
$n = 1000$	SML-1	1.010	0.444	0.444	0.356	0.667	0.983	1.332	0.945
	SML-2	1.002	0.120	0.120	0.095	0.926	0.999	1.080	0.901
	Probit	1.003	0.094	0.094	0.077	0.936	1.009	1.070	0.904
	2SLS	1.082	0.094	0.125	0.103	1.018	1.088	1.147	0.745
	BP	0.857	0.195	0.242	0.197	0.733	0.851	0.981	0.582

such that the sum of their absolute values is equal to 2, which corresponds to the sum of the magnitude of the true coefficients. The reason for this change is that using ratios of estimated coefficients results in a number of extreme outliers for the Blundell-Powell estimator that corrupt the analysis. With the new normalization, the estimates are much more well behaved.

The results of the simulation experiments are given in Tables 3.1– 3.3. For each estimator of $\beta_o = 1$, we report the mean value (MEAN), standard deviation (SD), root mean squared error (RMSE), median absolute deviation (MAD), the 25%, 50% and 75% sample quantiles, and the coverage rate (CR) of a bootstrap confidence interval with nominal level of 90%, obtained via the percentile method from 200 bootstrap replications.

Some general conclusions can be drawn from these results. First, although the SML-1 estimator has slightly better bias properties than SML-2, it also has a substantially higher variability in all three designs. Thus, in terms of RMSE or MAD, the SML estimator based on standard kernels uniformly dominates the one using higher-order kernels. Secondly, the SML-2 estimator compares favourably with the other alternatives and performs well uniformly over the different models we consider. It has the lowest

Table 3.2: Simulation Results Design II

		MEAN	SD	RMSE	MAD	25%	50%	75%	CR
$n = 250$	SML-1	1.053	0.459	0.462	0.371	0.667	0.999	1.334	0.995
	SML-2	1.122	0.311	0.334	0.264	0.911	1.107	1.317	0.913
	Probit	1.209	0.267	0.339	0.265	1.038	1.200	1.368	0.784
	2SLS	1.286	0.246	0.377	0.312	1.120	1.282	1.437	0.672
	BP	1.019	0.576	0.576	0.497	0.546	0.990	1.558	0.915
$n = 500$	SML-1	1.084	0.449	0.457	0.366	0.695	1.000	1.344	1.000
	SML-2	1.088	0.229	0.245	0.194	0.924	1.073	1.258	0.890
	Probit	1.204	0.178	0.271	0.220	1.081	1.199	1.313	0.724
	2SLS	1.285	0.170	0.332	0.288	1.169	1.278	1.383	0.523
	BP	1.061	0.509	0.512	0.429	0.699	1.022	1.459	0.928
$n = 1000$	SML-1	1.026	0.412	0.413	0.333	0.668	0.980	1.328	0.989
	SML-2	1.054	0.165	0.173	0.136	0.941	1.045	1.157	0.897
	Probit	1.200	0.135	0.241	0.207	1.104	1.205	1.281	0.506
	2SLS	1.277	0.128	0.305	0.278	1.188	1.283	1.355	0.272
	BP	1.065	0.355	0.360	0.293	0.817	1.067	1.341	0.913

Table 3.3: Simulation Results Design III

		MEAN	SD	RMSE	MAD	25%	50%	75%	CR
$n = 250$	SML-1	1.052	0.428	0.431	0.349	0.672	1.000	1.333	1.000
	SML-2	1.101	0.234	0.255	0.195	0.945	1.080	1.246	0.921
	Probit	1.203	0.216	0.296	0.236	1.050	1.191	1.337	0.767
	2SLS	1.242	0.204	0.317	0.260	1.114	1.235	1.358	0.677
	BP	1.016	0.548	0.548	0.472	0.546	1.008	1.508	0.918
$n = 500$	SML-1	1.006	0.396	0.397	0.316	0.671	0.986	1.329	0.985
	SML-2	1.068	0.170	0.183	0.145	0.941	1.059	1.193	0.896
	Probit	1.200	0.144	0.247	0.210	1.105	1.189	1.291	0.587
	2SLS	1.238	0.136	0.274	0.241	1.152	1.231	1.324	0.431
	BP	1.094	0.464	0.473	0.396	0.731	1.112	1.453	0.922
$n = 1000$	SML-1	0.973	0.372	0.373	0.287	0.672	0.970	1.243	1.000
	SML-2	1.036	0.109	0.114	0.090	0.961	1.032	1.108	0.911
	Probit	1.188	0.103	0.215	0.190	1.117	1.182	1.257	0.371
	2SLS	1.226	0.097	0.246	0.227	1.159	1.223	1.292	0.193
	BP	1.098	0.329	0.343	0.273	0.865	1.074	1.329	0.881

RMSE under all designs but the first, where it exceeds the RMSE of the correctly specified Probit by about 20%. In addition, the confidence intervals' coverage rates are remarkably close to the nominal level for all sample sizes and designs in the study. Third, the Probit estimator performs best when the parametric model is correctly specified, as is the case in Design I, and least well when the deviations from this model are most extreme. In general, its variance tends to be somewhat smaller than that of the SML estimator, but the bias is higher. Thus, when the bias induced by the misspecification is not too large, it tends to give reasonably good estimates. The bootstrap confidence intervals on the other hand have coverage rates far below their nominal level in the misspecified cases, and can thus be misleading in practice. Fourth, the Blundell-Powell estimators' performance is generally inferior to our that of our SML-2 procedure. For the relatively small sample sizes we consider, its RMSE and MAD also exceed the ones of the misspecified parametric estimators. For larger samples however, one would expect this relation to revert, since, at least for the second and third design, the Blundell-Powell estimator has a relatively small bias. Since the bootstrap confidence intervals perform satisfactory as well, this procedure could then be a useful alternative to SML in large samples, since then the latter is hard to compute. Moreover, it should be possible to improve the performance of the Blundell-Powell estimator through more effective rules for selecting the smoothing parameters, which is an important topic for future research. Finally, the 2SLS estimator turns out to have a low variance, but it is markedly biased in the second and third design. Consequently, the confidence intervals' coverage rates are far below their nominal values in this case. Although this estimator is applied frequently in empirical applications, one should thus be very careful when interpreting the results.

3.7 An Empirical Application: Home-ownership and Income in Germany

As an empirical application, we study the role of household income on the decision to rent an apartment or house versus owning it. The data we use are taken from the 2004 wave of the German Socioeconomic Panel (GSOEP), an extensive longitudinal survey of

Table 3.4: Descriptive Statistics

Variable	Mean	Std.Dev.	Min	Max
Homeowner	0.599	0.490	0	1
ln(total income)	7.853	0.324	6.397	9.473
Age	40.613	5.374	30	50
Children in HH	0.848	0.359	0	1
Education of wife				
Low degree	0.482	0.498	0	1
Intermediate degree	0.415	0.493	0	1
High degree	0.103	0.304	0	1
Wife Working	0.699	0.459	0	1

Notes: Sample size is $n = 981$. Education dummys indicate the highest of the three main secondary school tracks in Germany completed by the wife: *Hauptschulabschluss* ("low degree"), *Realschulabschluss* ("intermediate degree") or *Abitur* ("high degree"; university entry qualification)). "Wife Working" is an indicator that takes the value 1 when the wife has done any for-pay work in 2004.

households in Germany similar to the Panel Study of Income Dynamics (PSID) in the United States. The sample we use consists of 981 married men aged 30 to 50 that are working full time and have completed at most the lowest secondary school track of the German education system. Our dependent variable Y is an indicator that takes the value of 1 if a person owns its residence, and 0 if it is rented. The covariates X we are controlling for are the 2004 average total monthly income of the corresponding household (X^e), the person's age in years (Z_{11}) and an indicator for the presence of children younger than 16 in the household (Z_{12}). Generally speaking, home ownership should be determined by the permanent component of the income stream, of which monthly income is only a noisy measure. Therefore, we treat income as a mismeasured and thus potentially endogenous variable and employ dummy variables for the wife's education level (Z_{21}) and employment status (Z_{22}) as instruments. These human capital variables should be strongly related to the household income but have no direct influence on the housing decision. Some descriptive statistics for these variables are given in Table 3.4.

A priori, we would expect that all three regressors are positively related with homeownership for the following reasons: First, buying a house is associated with high financial costs including down payments, mortgage interests and repayments, maintenance costs and transaction costs such as notary fees and transfer taxes. Particularly in the first few

years after buying a home, these costs can exceed the costs of renting an equivalent place considerably. Thus, a higher level of income is needed to acquire a house in the first place. Second, the transition from renting to home-ownership is usually a one-time, non-reversible event associated with the family lifecycle. Thus, the proportion of home-owners should increase, other things equal, with age. Finally, it is well known that parenthood is a trigger for buying a home, and hence families with children should be more likely to own their residence.

For our application, we normalize the coefficient on the indicator for the presence of children to unity. Hence the model is given by

$$\begin{aligned} Y &= \mathbb{I}\{X^e\beta_{o1} + Z_{11}\beta_{o2} + Z_{12} \geq U\}, \\ X^e &= m_o(Z) + V. \end{aligned}$$

We consider specifying the reduced form for the endogenous regressor both parametrically as a linear model and in a fully nonparametric way. Since the resulting residuals are relatively similar, in Table 3.4 only report OLS estimates of α_o when the mean function is specified as $m_o(z) = z'\alpha_o$.

We then estimate the unknown parameter vector β_o by SML. Following the results from our simulations, we consider only the SML-2 estimator. The estimated coefficients $\hat{\beta}$ and their corresponding standard errors are given in the second column of Table 3.5. For the purpose of comparison, we also estimate the outcome equation by SML without taking the potential endogeneity into account, i.e. we use the ordinary Klein-Spady estimator with kernel and bandwidth specification analogous to the ones described in the preceding section. Finally, we also report results from applying the Blundell-Smith estimator and a standard probit model in the fourth and firth column of Table 3.5, respectively.

We can see that under all specifications the general tendencies we described above are confirmed. However, the difference between the estimates of β_o with and without controlling for endogeneity are quite substantial. Consider for example the estimates obtained by SML. After accounting for endogeneity, the coefficient on income is about twice as large as before (relative to the coefficient on the child indicator). Using a fully parametric approach leads to a quantitatively similar conclusion.

To illustrate the impact of such a change in coefficients, we consider the implications

Table 3.5: Estimation Results from Semiparametric and Fully Parametric Procedures

Variable	Reduced Form	SML estimates		Probit estimates	
	X^e	$\Pr(Y\|V)$	$\Pr(Y)$	$\Pr(Y\|V)$	$\Pr(Y)$
log(Total Income)	—	3.8533	1.9118	4.7923	2.1343
		(1.3338)	(.7310)	(1.5135)	(.5571)
Age	.0117	.0982	.1916	0.0863	0.2076
	(.0017)	(.0889)	(.0439)	(0.0209)	(.0257)
Children in HH	.0911	1.0000	1.0000	1.0000	1.0000
	(.0194)	—	—	—	—
$\hat{V}$ (Control variable)	—	—	—	−3.0348	—
				(1.3048)	
Education of wife					
Intermediate degree	.0642	—	—	—	—
	(.0185)				
High degree	.1291	—	—	—	—
	(.0298)				
Wife Working	.0911	—	—	—	—
	(.0194)				
R^2	.1072	—	—	—	—
F-statistic	23.42 (5, 975 df)	—	—	—	—
Bandwidth	—	$h = (0.04, .21)$	$h = .03$	—	—

Notes: Standard errors (based on bootstrap 500 bootstrap replications for SML and the usual formulas otherwise) in parentheses. Baseline category for Education of wife is "low degree".

for the Average Structural Function (ASF), which gives the choice probabilities when the value of the regressors X is fixed at some exogenously determined value $\bar{x}$. As mentioned in Section 5, this object is identified as a partial mean of the link function G_o with respect to V. Following the advice of Imbens and Newey (2009), we estimate the ASF by

$$\hat{ASF}(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} \hat{G}_{LL}(\bar{x}\hat{\beta}, \hat{V}_i|\hat{\beta}, \hat{v}), \tag{3.7.1}$$

where $\hat{G}_{LL}(\bar{x}\hat{\beta}, \hat{V}_i|\hat{\beta}, \hat{v})$ is the local linear estimator of $E(Y|X\beta, V)$ evaluated at $\bar{x}\hat{\beta}$ and $\hat{V}_i$, and the bandwidth is chosen by least squares cross-validation.

In Figure 1, the estimated ASF is plotted from the 5% to the 95% quantile of the income distribution for a man aged 40 with children. We can see that the two models imply vastly different probabilities of home-ownership, particularly in the lower half of the income distribution. For a monthly household net income of 1800 EUR (corresponding to a log income of about 7.5), the probability of owning the residence reduces from 50% to roughly 20% when controlling for endogeneity. This difference diminishes as we move up the income distribution, and for values of income larger than 2500 EUR (which corresponds to a log income of about 7.8), the predictions from the two models are qualitatively similar.

3.8 Concluding Remarks

This paper presents a semiparametric maximum likelihood procedure for the estimation of the coefficients of a single index binary choice model with endogenous regressors. We discuss how identification is achieved via a control function approach, and derive the asymptotic properties of the new estimator. In our Monte Carlo experiments, the new estimator performs well in comparison with other related procedures.

One of the major issues of our estimator is its computational complexity when applied in settings with many regressors and/or observations. In this case, even evaluating the likelihood function at a specific point is very time consuming, and the function might have several local maxima. However, these problems are not specific to our SML estimator but are encountered in general when computing semiparametric optimization estimators such as the ones by Ichimura (1993) or Klein and Spady (1993). For these estimators, a

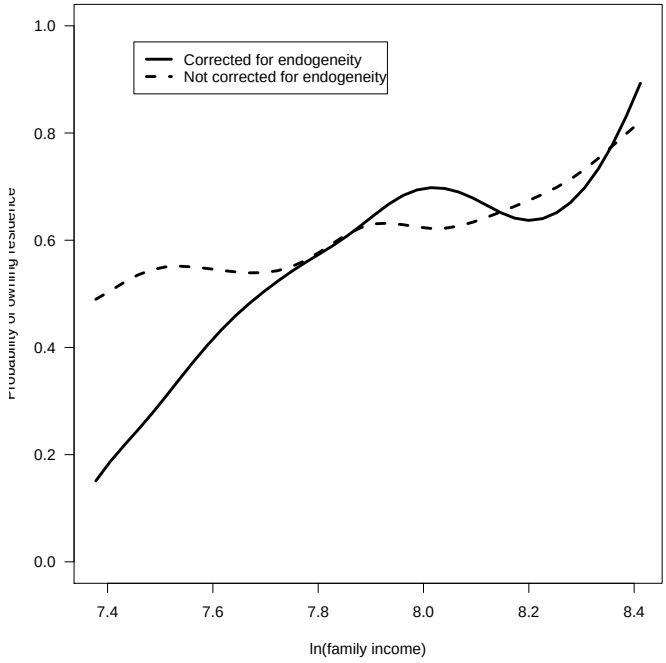

Figure 3.7.1: Estimated probability of owning the residence for a man aged 40 with children.

number of suggestions have been made to improve their numerical properties, such as e.g. the use of Fast Fourier Transforms or binning techniques (see Ichimura and Todd (2007) for a comprehensive overview). All of these approaches could in general be adapted to our estimator as well.

It might also be possible to avoid the use of numerical optimization routines altogether. In a recent paper, Xia (2006) shows that the computationally much simpler rMAVE procedure of Xia, Tong, Li, and Zhu (2002) achieves the same asymptotic variance as the Klein and Spady estimator when applied to a standard binary choice model without endogenous regressors. Again, it should be possible to adapt this technique to our problem and thus reduce the computational complexity.

Appendix

Proof of Theorem 1. The proof of the theorem is analogous to the argument in Manski (1988): First, note that that $V = v_o(X^e, Z)$ is identified by assumption. Now assume that there exists a $\tilde{\beta} \in \mathcal{B}$ such that $\mathbb{E}(Y|X,V) = \mathbb{E}(Y|X\tilde{\beta},V) = \mathbb{E}(Y|X\beta_o,V) \equiv G_o(X\beta_o,V)$. Then there must exist a function $H(\cdot,V)$ that is strictly monotone for all V, such that $X^{(1)} + X^{(-1)'}\tilde{\beta} =$

$H(X^{(1)} + X^{(-1)'}\beta_o, V)$. Differentiating both sides of this equation with respect to $X^{(1)}$ for $X \in \mathcal{A}$, we see that $H(\cdot, V)$ must be the identity function since $X^{(1)}$ is continuously distributed conditional on V, and thus $X^{(-1)'}\tilde{\beta} = X^{(-1)'}\beta_o$. By condition (iii), this relation can hold with probability one only if $\tilde{\beta} = \beta_o$. □

We now turn to the proofs of the consistency and asymptotic normality result. First, we give some useful preliminary results on uniform rates of convergence for nonparametric estimators based on generated regressors. Second, we show consistency via a classical, direct argument. Third, we prove asymptotic normality of our estimator by showing that our problem fits the framework of Chen, Linton, and Van Keilegom (2003).

Lemma 1. *Under Assumption 1-8, we have that uniformly in $w \in \mathcal{W}$ and $\beta \in \mathcal{B}$, respectively, i) $\hat{D}(w|\beta, \hat{v}) - \hat{D}(w|\beta) = o_p(n^{-1/4})$, ii) $\partial_\beta \hat{D}(w|\beta, \hat{v}) - \partial_\beta \hat{D}(w|\beta) = o_p(n^{-1/4})$, iii) $\partial_k \hat{D}(w|\beta, \hat{v}) - \partial_k \hat{D}(w|\beta) = o_p(n^{-1/4})$ for $k = 1, 2$, iv) $\hat{N}(w|\beta, \hat{v}) - \hat{N}(w|\beta) = o_p(n^{-1/4})$, v) $\partial_\beta \hat{N}(w|\beta, \hat{v}) - \partial_\beta \hat{N}(w|\beta) = o_p(n^{-1/4})$, vi) $\partial_k \hat{N}(w|\beta, \hat{v}) - \partial_k \hat{N}(w|\beta) = o_p(n^{-1/4})$ for $k = 1, 2$.*

Proof. We only proof the first result, as the remaining ones can by shown analogously. Using the definition of $\hat{D}$ and Hölder's inequality, it follows that

$$\begin{aligned}
|\hat{D}(w|\beta, \hat{v}) - \hat{D}(w|\beta)| &= \left| \frac{1}{nh^{d_v}} \sum_{i=1}^{n} \partial_2 K_h(X_i\beta - w_1, \tilde{V}_i - w_2)(\hat{V}_i - V_i) \right| \\
&\leq h^{-d_v} \left(\frac{1}{n} \sum_{i=1}^{n} (\partial_2 K_h(X_i\beta - w_1, \tilde{V}_i - w_2))^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{V}_i - V_i)^2 \right)^{1/2} \\
&= h^{-d_v} T_1 \times T_2,
\end{aligned}$$

where $\tilde{V}_i$ is some value between V_i and $\hat{V}_i$. It is then easy to show that $T_1 = O_p(1)$ uniformly in β and w. Now consider T_2. Substituting the "high-level" representation for $\hat{V}_i - V_i$ from Assumption 8 into the expression, and applying Jensen's inequality and the usual projection arguments for U-Statistics, we obtain that

$$\begin{aligned}
T_2^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} \sum_{j=1}^{n} \omega_n(Z_i, Z_j)\psi_j + o_p(n^{-1/2}) \right)^2 \\
&\leq \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_n(Z_i, Z_j)^2 \psi_j^2 + o_p(n^{-1}) \\
&= O_p(n^{-1})
\end{aligned}$$

It then follows together with Assumption 7 that $h^{-d_v} T_1 T_2 = O_p(h^{-d_v} n^{-1/2}) = o_p(n^{-1/4})$, as claimed. □

Lemma 2. *Under Assumption 1–8, (i)*

$$\sup_{w\in\mathcal{W},\beta\in\mathcal{B}} |\hat{G}(w|\beta,\hat{v}) - G(w|\beta)| = o_p(1)$$

and (ii)

$$\sup_{w\in\mathcal{W},\|\beta-\beta_o\|\leq\delta_n} |\hat{G}(w|\beta,\hat{v}) - G(w|\beta)| = o_p(n^{-1/4})$$

$$\sup_{w\in\mathcal{W},\|\beta-\beta_o\|\leq\delta_n} |\partial_\beta\hat{G}(w|\beta,\hat{v}) - \partial_\beta G(w|\beta)| = o_p(n^{-1/4})$$

$$\sup_{w\in\mathcal{W},\|\beta-\beta_o\|\leq\delta_n} |\partial_1\hat{G}(w|\beta,\hat{v}) - \partial_1 G(w|\beta)| = o_p(n^{-1/4})$$

for all $\delta_n = o(1)$.

Proof. This follows from standard kernel smoothing theory together with Lemma 1. □

Proof of Theorem 2. To show that $\hat{\beta}$ is consistent, we first define an infeasible version of the semiparametric likelihood function, with $\hat{G}(\hat{W}_i(\beta)|\beta,\hat{v})$ replaced with its probability limit $G(W_i(\beta)|\beta)$:

$$\tilde{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tau_i(Y_i \log(G(W_i(\beta)|\beta)) + (1-Y_i)\log(1-G_i(W(\beta)|\beta))).$$

The difference between $\tilde{L}_n(\beta)$ and $L_n(\beta)$ goes to zero uniformly in β, for $n\to\infty$, because

$$\begin{aligned}
\sup_{\beta\in B}|L_n(\beta)-\tilde{L}_n(\beta)| &\leq \left(\inf_{\beta\in B}\min_i \left\{\hat{G}(\hat{W}_i(\beta)|\beta,\hat{v}), 1-\hat{G}(\hat{W}_i(\beta)|\beta,\hat{v}), G(W_i(\beta), 1-G(W_i(\beta))\right\}\right) \\
&\quad \times \sup_{\beta\in B}\max_i \tau_i|\hat{G}(\hat{W}_i(\beta)|\beta,\hat{v}) - G(W_i(\beta)|\beta)| \\
&= O_p(1)\sup_{\beta\in B}\max_i \tau_i|\hat{G}(W_i(\beta)|\beta,\hat{v}) - G(W_i(\beta)) + \partial_2\hat{G}(\tilde{W}_i(\beta)|\beta,\hat{v})|\beta)(\hat{V}_i - V_i))| \\
&= o_p(1)
\end{aligned}$$

by Lemma 2 and Assumption 8, where $\tilde{W}_i(\beta)$ is some value between $\hat{W}_i(\beta)$ and $W_i(\beta)$. Furthermore, since $\tilde{L}_n(\beta)$ is an ordinary parametric likelihood function, by a standard uniform law of large numbers, e.g. Lemma 2.4 in Newey and McFadden (1994), it converges uniformly in β to its expectation, i.e. we have

$$\sup_{\beta\in B}|\tilde{L}_n(\beta) - L(\beta)| = o_p(1),$$

where

$$L(\beta) = \mathbb{E}(L_n(\beta)) = \mathbb{E}(\tau_i\left[Y_i\log(G(W(\beta)|\beta)) + (1-Y_i)\log(1-G(W(\beta)|\beta))\right])$$

is a non-random function that is continuous in β. Taken together, it follows from the triangle inequality that

$$\sup_{\beta \in B} |L_n(\beta) - L(\beta)| = o_p(1),$$

which implies that $\hat{\beta}$ is consistent whenever $L(\beta)$ attains a unique maximum at β_o. By the law of iterated expectations,

$$L(\beta) = \mathbb{E}(\tau \left[G_o(X\beta_o, V) \log(G(W(\beta)|\beta)) + (1 - G_o(X\beta_o, V)) \log(1 - G(W(\beta)|\beta)) \right]),$$

and the term in square brackets attains its maximum whenever the relation $G(W(\beta)|\beta) = G_o(X\beta_o, V)$ holds. By Assumption 1, this is the case if and only if $\beta = \beta_o$. The statement of the Theorem then follows from the usual consistency argument, e.g. Theorem 2.1 in Newey and McFadden (1994). □

We now turn to the proof of asymptotic normality of our estimator $\hat{\beta}$. This is done by verifiying the conditions of Theorem 2 in Chen, Linton, and Van Keilegom (2003) in Lemma 3–8. Similar arguments are used by Linton, Sperlich, and Van Keilegom (2008), who consider semiparametric estimation of a transformation model. Their problem is technically related to ours since they also consider a semiparametric maximum likelihood estimator based on non-parametrically generated regressors, but the actual model is very different.

We start with introducing some further notation. First, we have to define a criterion function depending on β and some unknown nuisance function, whose population value is equal to zero at the true parameter values. To this end, write $\gamma = (\gamma_1, \ldots, \gamma_4)$ for a generic collection of nuissance functions, and define $\gamma_\beta = (\partial_1 G(\cdot|\beta), \partial_\beta G(\cdot|\beta), G(\cdot|\beta), v_o)$, $\gamma_o = \gamma_{\beta_o}$, and $\hat{\gamma}_\beta = (\partial_1 \hat{G}(\cdot|\beta, \hat{v}), \partial_\beta \hat{G}(\cdot|\beta, \hat{v}), \hat{G}(\cdot|\beta, \hat{v}), \hat{v}_o)$, and $\hat{\gamma}_o = \hat{\gamma}_{\beta_o}$. Then, for any γ, let

$$S_n(\beta, \gamma) = \frac{1}{n} \sum_{i=1}^{n} s(Y_i, X_i, Z_i, \beta, \gamma)$$

where

$$\begin{aligned} s(Y_i, X_i, Z_i, \beta, \gamma) = &(\gamma_1(X_i\beta, \gamma_4(X_i, Z_i))\tilde{X}_i + \gamma_2(X_i\beta, \gamma_4(X_i, Z_i))) \\ &\times \frac{Y_i - \gamma_3(X_i\beta, \gamma_4(X_i, Z_i))}{\gamma_3(X_i\beta, \gamma_4(X_i, Z_i))(1 - \gamma_3(X_i\beta, \gamma_4(X_i, Z_i)))} \end{aligned}$$

and note that

$$S_n(\beta, \hat{\gamma}_\beta) = \partial_\beta L_n(\beta),$$

i.e. $S_n(\beta, \hat{\gamma}_\beta)$ is the score corresponding to our likelihood function $L_n(\beta)$. Furthermore, define the population version of the criterion function as

$$S(\beta, \gamma) = \mathbb{E}(S_n(\beta, \gamma)).$$

Finally, we have to define an appropriate space for the nuissance functions γ. Denote this space by $\Gamma = \Gamma_1 \times \mathcal{V}$, where $\mathcal{V}$ is defined in Assumption 8 and Γ_1 is the class of all functions $f : \mathbb{R}^{1+d_v} \to \mathbb{R}$ whose partial derivatives up to order $\alpha > (1 + d_v)/2$ exist and are uniformly bounded by some constant M. This class of functions is typically denoted by $C_M^\alpha(\mathbb{R}^2)$ in the literature (see e.g. van der Vaart and Wellner (1996, p. 154)). A norm $\|\cdot\|_\Gamma$ on the space Γ that satisfies the requirements of Chen, Linton, and Van Keilegom (2003) can be defined as

$$\|\gamma\|_\Gamma = \sup_{\beta \in \mathcal{B}} \max\{\|\gamma_1\|_\infty, \ldots, \|\gamma_4\|_\infty\}.$$

Note that our Assumption 3 and 8 are sufficient to ensure that $\gamma_o \in \Gamma$.

We can now prove the Lemmas needed to verify the conditions of Theorem 2 in Chen, Linton, and Van Keilegom (2003).

Lemma 3 (Condition (2.1))**.** $\|S_n(\hat{\beta}, \hat{\gamma}_\beta)\| = \inf_{\beta \in \mathcal{B}} \|S_n(\beta, \hat{\gamma}_\beta)\| + o_p(n^{-1/2})$

Proof. This is trivially satisfied since $\|S_n(\hat{\beta}, \hat{\gamma}_\beta)\| = 0$ by construction. □

Lemma 4 (Condition (2.2))**.** *The ordinary derivative $S_\beta(\beta, \gamma_\beta) = \partial S(\beta, \gamma_\beta)/\partial\beta$ exists in a neighborhood of β_o, is continuous at $\beta = \beta_o$, and the matrix $S_\beta(\beta_o, \gamma_{\beta_o})$ is of full rank.*

Proof. This follows directly from Assumptions 3 and 5. □

Lemma 5 (Condition (2.3))**.** *The pathwise derivative $\dot{S}(\beta, \gamma_\beta)$ of $S(\beta, \gamma_\beta)$ exists in all directions $\gamma - \gamma_\beta$, and satisfies: (i)*

$$\|S(\beta, \gamma) - S(\beta, \gamma_\beta) - \dot{S}(\beta, \gamma_\beta)[\gamma - \gamma_\beta]\| \leq c\|\gamma - \gamma_\beta\|_\Gamma^2$$

for all $\beta \in \mathcal{B}$ with $\|\beta - \beta_o\| \leq \delta_n$, all γ with $\|\gamma - \gamma_o\|_\Gamma \leq \delta_n$, some positive sequence $\delta_n = o(1)$, and some constant $c < \infty$; and (ii)

$$\|\dot{S}(\beta, \gamma_\beta)[\gamma - \gamma_\beta] - \dot{S}(\beta_o, \gamma_o)[\gamma - \gamma_o]\| \leq o(1)\delta_n.$$

Proof. Using standard rules for calculating pathwise derivatives, we obtain after some calculations that

$$\begin{aligned}\dot{S}(\beta,\gamma_\beta)[\gamma] = \mathbb{E}\Bigg[&\tau\frac{Y-G(W(\beta)|\beta)}{G(W(\beta)|\beta)(1-G(W(\beta)|\beta))}(\gamma_1(W(\beta))\tilde{X}+\gamma_2(W(\beta)))\\ &-\tau\partial_\beta G(W(\beta)|\beta)\frac{1}{G(W(\beta)|\beta)(1-G(W(\beta)|\beta))}\gamma_3(W(\beta))\\ &-\tau\partial_\beta G(W(\beta)|\beta)\frac{(Y-G(W(\beta)|\beta))(1-2G(W(\beta)|\beta))}{G(W(\beta)|\beta)(1-G(W(\beta)|\beta))}\gamma_3(W(\beta))\\ &-\tau\partial_\beta G(W(\beta)|\beta)\frac{1}{G(W(\beta)|\beta)(1-G(W(\beta)|\beta))}\partial_2 G(W(\beta)|\beta)\gamma_4(X^e,Z)\\ &-\tau\partial_\beta G(W(\beta)|\beta)\frac{(Y-G(W(\beta)|\beta))(1-2G(W(\beta)|\beta))}{G(W(\beta)|\beta)(1-G(W(\beta)|\beta))}\partial_2 G(W(\beta)|\beta)\gamma_4(X^e,Z)\\ &+\tau\,\frac{Y-G(W(\beta)|\beta)}{G(W(\beta)|\beta)(1-G(W(\beta)|\beta))}\partial_{\beta,2}G(W(\beta)|\beta)\gamma_4(X^e,Z)\Bigg].\end{aligned}$$

Furthermore, since $\mathbb{E}(Y-G(W))=0$, it follows from the Law of Iterated Expectations that

$$\dot{S}(\beta_o,\gamma_o)[\gamma] = \mathbb{E}\left[-\frac{\tau\partial_\beta G(W)}{G(W)(1-G(W))}\gamma_3(W)-\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}\gamma_4(X^e,Z)\right].$$

The two inequalities then follow immediately by using that under our assumptions the functions involved satisfy a Lipschitz property. □

Lemma 6 (Condition (2.4))**.** *$\hat{\gamma}\in\Gamma$ with probability tending to one; and $\|\hat{\gamma}-\gamma_o\|_\Gamma = o_p(n^{-1/4})$.*

Proof. The first part follows directly from the definition of the estimators and the smoothness conditions imposed on the kernel function, whereas the second part is a consequence of Lemma 2 and Assumption 8. □

Lemma 7 (Condition (2.5'))**.** *For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,*

$$\sup_{\|\beta-\beta_o\|\le\delta_n,\|\gamma-\gamma_o\|\le\delta_n}\|S_n(\beta,\gamma)-S(\beta,\gamma)-S_n(\beta_o,\gamma_o)\| = o_p(n^{-1/2})$$

Proof. This statement follows from Theorem 3 in Chen, Linton, and Van Keilegom (2003). To verify the conditions of that theorem one first has to show that

$$\mathbb{E}\left(\sup_{\|\bar{\beta}-\beta\|<\delta,\|\bar{\gamma}-\gamma\|_\Gamma<\delta}\left|s(Y,X,Z,\bar{\beta},\bar{\gamma})-s(Y,X,Z,\beta,\gamma)\right|^2\right)\le K\delta^2$$

for all $(\beta,\gamma)\in\mathcal{B}\times\Gamma$, all $\delta>0$ and some constant $K>0$. This follows from the differentiability of the functions of which s is composed and the mean value theorem.

Secondly, one has to show that

$$\int_0^\infty\sqrt{\log N(\lambda,\Gamma,\|\cdot\|_\Gamma)}d\lambda<\infty,$$

where $N(\lambda, \Gamma, \|\cdot\|_\Gamma)$ is the minimal number of balls with $\|\cdot\|_\Gamma$-radius λ needed to cover Γ. This is a consequence of a result in van der Vaart and Wellner (1996, Corollary 2.7.4) and Assumption 8. □

Lemma 8 (Condition (2.6))**.**

$$\sqrt{n}(S_n(\beta_o, \gamma_o) + \dot{S}(\beta_o, \gamma_o)[\hat{\gamma} - \gamma_o]) \xrightarrow{d} N(0, \Omega)$$

Proof. Note that as shown in the proof of Lemma 5, we have that

$$\dot{S}(\beta_o, \gamma_o)[\gamma] = -\mathbb{E}\left(\frac{\mathbb{E}(\tau\partial_\beta G(W)|W)}{G(W)(1-G(W))}\gamma_3(W)\right) - \mathbb{E}\left(\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}\gamma_4(X^e, Z)\right),$$

and hence

$$\begin{aligned}\dot{S}(\beta_o, \gamma_o)[\hat{\gamma}_o - \gamma_o] &= -\mathbb{E}\left(\frac{\mathbb{E}(\tau\partial_\beta G(W)|W)}{G(W)(1-G(W))}(\hat{G}(W|\hat{v}) - G(W))\right) \\ &\quad - \mathbb{E}\left(\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}(\hat{V} - V)\right) \\ &\equiv -A_1 - A_2.\end{aligned}$$

To simplify the notation, let $t(w) = \mathbb{E}(\tau\partial_\beta G(W)|W = w)/(G(w)(1-G(w)))$. Then we have that

$$\begin{aligned}A_1 &= \int t(w)(\hat{G}(w|\hat{v}) - G(w))D(w)dw \\ &= \int t(w)((\hat{N}(w|\hat{v}) - N(w)) - \frac{N(w)}{D(w)}(\hat{D}(w|\hat{v}) - D(w)))dw + o_p(n^{-1/2}) \\ &= \int t(w)(\hat{N}(w) - N(w))dw \qquad (3.8.1) \\ &\quad - \int t(w)G(w)(\hat{D}(w) - D(w))dw \qquad (3.8.2) \\ &\quad + \int t(w)(\hat{N}(w|\hat{v}) - \hat{N}(w))dw \qquad (3.8.3) \\ &\quad - \int t(w)G(w)(\hat{D}(w|\hat{v}) - \hat{D}(w))dw + o_p(n^{-1/2}) \qquad (3.8.4)\end{aligned}$$

Now consider the term in (3.8.1). Due to the use of higher-order kernels, the difference between $N(w)$ and $\mathbb{E}(\hat{N}(w))$ is of the order $o(n^{-1/2})$ uniformly in w. Hence

$$\begin{aligned}\int t(w)(\hat{N}(w) - N(w))dw &= \int t(w)(\hat{N}(w) - \mathbb{E}(\hat{N}(w)))dw + o(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^n \int t(w)(Y_i K_h(W_i - w) - \mathbb{E}(Y_i K_h(W_i - w))dw + o(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^n t(W_i)Y_i - \mathbb{E}(t(W)G(W)) + o_p(^{-1/2})\end{aligned}$$

where the last equality follows from standard change-of-variables and Taylor-expansion arguments. Similarly, one obtains for the term in (3.8.2) that

$$\int t(w)G(w)(\hat{D}(w) - D(w))dw = \frac{1}{n}\sum_{i=1}^{n} t(W_i)G(W_i) - \mathbb{E}(t(W)G(W)) + o_p(^{-1/2}).$$

Next, inserting the definition of the respective estimators, we obtain for the term in (3.8.3) that

$$\begin{aligned}
\int t(w)(\hat{N}(w|\hat{v}) - \hat{N}(w))dw &= \frac{1}{n}\sum_{i=1}^{n} Y_i \int t(w)(K_h(X_i\beta_o - w_1, \hat{V}_i - w_2) - K_h(X_i\beta_o - w_1, V_i - w_2))dw \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i \int t(X_i\beta_o - rh, V_i - sh)(K(r, s + (\hat{V}_i - V_i)/h) - K(r,s))drds \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(\hat{V}_i - V_i)/h \int t(X_i\beta_o - rh, V_i - sh)\partial_s K(r,s)drds + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(\hat{V}_i - V_i) \int \partial_2 t(X_i\beta_o - rh, V_i - sh)K(r,s)drds + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(\hat{V}_i - V_i)\partial_2 t(W_i) + o_p(n^{-1/2}),
\end{aligned}$$

where the 2nd to 5th line follow by substitution, a Taylor expansion of the kernel, partial integration, and the higher order property of the kernel, respectively. The last expression is then equal to

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_n(Z_i, Z_j)\psi_j Y_i \partial_2 t(W_i) + o_p(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\omega_n(Z, Z_i)\mathbb{E}(Y|X,Z)\partial_2 t(W)|Z_i)\psi_i + o_p(n^{-1/2})$$

using Assumption 8 and common projection arguments for U-statistics. Finally, one can use similar arguments to show that the term in (3.8.4) is equal to

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\omega_n(Z, Z_i)\partial_2 l(W)|Z_i)\psi_i + o_p(n^{-1/2}),$$

where $l(w) = G(w)t(w)$, and thus the terms in (3.8.3)–(3.8.4) are equal to $n^{-1}\sum_{i=1}^{n}\xi_{2i}\psi_i + o_p(n^{-1/2})$ since $\mathbb{E}(Y|X,Z) = G(W)$.

Now consider the term A_2. It follows directly from Assumption 8 that

$$\begin{aligned}
A_2 &= \mathbb{E}\left(\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}(\hat{v}(X^e, Z) - v(X^e, Z)\right) \\
&= \int \frac{\tau\partial_\beta G(x\beta_o, v(x,z))\partial_2 G(x\beta_o, v(x,z))}{G(x\beta_o, v(x,z))(1 - G(x\beta_o, v(x,z)))}\frac{1}{n}\sum_{i=1}^{n}\omega_n(z, Z_i)\psi_i dF_{X,Z}(x,z) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}\omega_n(Z, Z_i)|Z_i\right)\psi_i + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\xi_{1i}\psi_i + o_p(n^{-1/2}),
\end{aligned}$$

where $F_{X,Z}$ is the joint CDF of (X, Z). Taken together, we have shown so far that

$$\begin{aligned}
&\sqrt{n}(S_n(\beta_o, \gamma_o) + \dot{S}(\beta_o, \gamma_o)[\hat{\gamma} - \gamma_o]) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - G(W_i)}{G(W_i)(1 - G(W_i))} (\tau_i \partial_\beta G(W_i) - \mathbb{E}(\tau_i \partial_\beta G(W_i)|W_i)) \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\xi_{1i} - \xi_{2i})\psi_i + o_p(1).
\end{aligned}$$

The statement of the Lemma then follows from applying an ordinary CLT, since ψ_i and $Y_i - G(W_i)$ are orthogonal. □

Proof of Theorem 3. The results in Theorem 2 and Lemma 3 – 8 imply that the conditions of Theorem 2 in Chen, Linton, and Van Keilegom (2003) are fulfilled, which in turn implies the statement of the theorem. □

Chapter 4

Identification of Unconditional Partial Effects in Nonseparable Models

4.1 Introduction

An important feature of many interesting economic models is that they do not imply an econometric specification with additively separable disturbance terms when they are taken to data. The properties of such nonseparable models have therefore received considerable interest in the recent literature, being investigated by Chesher (2003), Matzkin (2003), Chesher (2005), Chernozhukov and Hansen (2005), Hoderlein and Mammen (2007), Chernozhukov, Imbens, and Newey (2007) and Imbens and Newey (2009), amongst others. One of the most important issues in this context is how to accommodate the presence of endogenous regressors, which are frequently encountered in microeconometric applications. A possible approach is the use of control variable techniques, which are discussed in detail by Imbens and Newey (2009). They establish identification of various quantities of interest in triangular simultaneous equation models under relatively general conditions. These quantities include the Average Structural Function, the Quantile Structural Function, Average Derivatives and Policy Effects.

In this paper, we show that a further interesting class of parameters can be identified

under general conditions in their framework: the Unconditional Partial Effects. These parameters have recently been introduced to the literature by Firpo, Fortin, and Lemieux (2009) in the exogenous case, and correspond to the following thought experiment: suppose that every member of the population would experience the same exogenous marginal increase in one of its observable characteristics. How would this affect the unconditional distribution of the outcome variable? To give a concrete example, a researcher might be interested in the effect of a marginal increase in everybody's income on some feature of the distribution of consumption, such as its moments, quantiles, Gini coefficient or other measures of inequality. As pointed out by Firpo, Fortin, and Lemieux (2009), such summary measures are of interest for policy analysis, where the focus is on aggregate as opposed to individual effects of a variable.

Firpo, Fortin, and Lemieux (2009) show that in a setting without endogenous variables, Unconditional Partial Effects are identified, showing that they can be represented by the average derivative of a projection of the recentered influence function of the statistic of interest on the regressors. We demonstrate that this result can be generalized to the triangular nonseparable models discussed in Imbens and Newey (2009) using their control variable approach (other papers that use control variable techniques in non- or semiparametric setting include Blundell and Powell (2003), Blundell and Powell (2004), Blundell and Powell (2007) and Florens, Heckman, Meghir, and Vytlacil (2008)). As a further contribution, this paper also provides a slightly different representation of Unconditional Partial Effects compared to the one given in Firpo, Fortin, and Lemieux (2009). We show that these parameters can be written in terms of the average derivate of the conditional CDF of the outcome variable given the regressors and the control variable (where the derivate is taken with respect to the regressors). This representation is useful to give an explicit expression for Unconditional Partial Effects when further parametric or semiparametric restrictions are imposed on the model. This representation is by no means specific to the setting with endogenous variables but holds under full exogeneity as well, with obvious simplifications. We illustrate this point by considering the linear quantile regression model as an example.

The remainder of this paper is organized as follows. In the next section, we describe the model and give a precise definition of Unconditional Partial Effects. Identification is

discussed in Section 3. The final section concludes.

4.2 Model and Parameters of Interest

The model we consider in this paper is essentially the same as in Imbens and Newey (2009). We observe a scalar outcome variable of interest denoted by Y, which is linked to a random vector $X = (X_1, Z_1)$ of observable determinants and an unobserved disturbance term ε through the structural equation

$$Y = g(X, \varepsilon). \tag{4.2.1}$$

The subvector X_1 of X is potentially endogenous and assumed to be determined through a reduced form equation,

$$X_1 = h(Z, \eta) \tag{4.2.2}$$

where η is another unobserved disturbance and $Z = (Z_1, Z_2)$ is a vector of instruments that excert influence on X_1 in a sense to be made precise below, but are independent of the error terms. As in Imbens and Newey (2009), no restrictions on the dimesionality of ε are imposed, allowing for general forms of unobserved heterogeneity. However, for identification purposes it will be necessary to impose such a restriction on the disturbance in (4.2.2), as discussed below. To simplify the notation, we will focus in the following on the case with $X = X_1$ consisting of a single endogenous regressor only, but all arguments can easily be generalized to allow for the presence of multiple endogenous regressors or additional exogenous ones.

The parameters we are interested in correspond to the effect of a marginal increase in X on some feature $\Gamma(F_Y)$ of the unconditional distribution of Y. That is, for some constant $\delta \neq 0$, define the counterfactual random variable Y_δ as

$$Y_\delta = g(X + \delta, \varepsilon).$$

Denote the CDF of Y and Y_δ by F_Y and $F_{Y,\delta}$, respectively, and let $\Gamma(\cdot)$ be a functional of interest. For example Γ could be the functional that maps a CDF into one of it moments, or into its quantile function. With this notation, we can now formally define an Unconditional Partial Effect.

Definition 1 (Unconditional Partial Effect)**.** *For any functional* $\Gamma : D(-\infty, \infty) \to S$*, where* S *is some normed space, the quantity*

$$\theta_\Gamma = \lim_{\delta \to 0} \frac{\Gamma(F_{Y,\delta}) - \Gamma(F_Y)}{\delta} \tag{4.2.3}$$

is called the Unconditional Partial Effect of X on $\Gamma(F_Y)$*, provided that the limit in* (4.2.3) *exists.*

4.3 Identification

In order to identify the Unconditional Partial Effects in models with endogeneity, we can use control variable techniques developed in Imbens and Newey (2009). Generally speaking, a control variable is an identified random vector that is able to absorb the dependence between the regressors and the unobserved disturbance term in the outcome equation (4.2.1), in the sense that X and ε will be independent conditional on the control variable. Imbens and Newey (2009) show that in the triangular model such a control variable is available under certain restrictions on the second equation. We repeat their result here for completeness.

Lemma 1 (Imbens and Newey, 2009)**.** *Suppose that* $h(z, \cdot)$ *is strictly increasing for all values of* z*, that* η *is continuously distributed with strictly increasing CDF, and that* $Z \perp (\varepsilon, \eta)$*. Then* $\varepsilon \perp X|V$*, where* $V = F_{X|Z}(X, Z)$*.*

The reason $V = F_{X|Z}(X, Z)$ has the properties of a control variable in our model is that the exclusive source of dependence between X and ε is their joint dependence on the disturbance term η from equation (4.2.2). However, under the conditions of Lemma 1, V is simply a one-to-one transformation of η, which in turn implies the result.

The conditional independence property can be used to derive an explicit representation for $F_{Y,\delta}$. Using the structure of the model and the law of iterated expectations, we obtain

that

$$\begin{aligned} F_{Y,\delta}(y) &= \int \Pr(m(X+\delta,\varepsilon) \leq y | X = x, V = v) dF_{X,V}(x,v) \\ &= \int \Pr(m(X,\varepsilon) \leq y | X = x+\delta, V = v) dF_{X,V}(x,v) \\ &= \int F_{Y|X,V}(y, x+\delta, v) dF_{X,V}(x,v) \\ &= \mathbb{E}(F_{Y|X,V}(y, X+\delta, V)). \end{aligned}$$

This implies that the function $F_{Y,\delta}$ is identified if the support of the random vector $(X+\delta, V)$ is contained in the support of (X,V). For identification of the Unconditional Partial Effect, it will be sufficient that this condition holds for small values of δ only. The role of this support condition is to ensure that there is a sufficient amount of variation in the endogenous regressors induced by the instruments. To see this, assume for a moment that Z does not excert any influence on X. Then $V = F_{X|Z}(X,Z) \equiv t(X)$ is simply a transformation of the endogenous regressor. While the conditional independence condition $X \perp \varepsilon | V$ will still hold in this case, the joint support of X and V is now given by $\{(x, t(x)) : x \in supp(X)\}$, which is generally not a subset of $\{(x+\delta, t(x)) : x \in supp(X)\}$ for any $\delta \neq 0$.

In order to derive a general formula for the Unconditional Partial Effect of X on $\Gamma(F_Y)$ for some general functional Γ, we first consider the simplest case where $\Gamma = id$ is the identity mapping, i.e. $\Gamma(F) = F$. Then

$$\begin{aligned} \theta_{id}(y) &= \lim_{\delta \to 0} \frac{F_{Y,\delta}(y) - F_Y(y)}{\delta} \\ &= \lim_{\delta \to 0} \frac{\mathbb{E}(F_{Y|X,V}(y, X+\delta, V)) - \mathbb{E}(F_{Y|X,V}(y, X, V))}{\delta} \\ &= \mathbb{E}(\partial_x F_{Y|X,V}(y, X, V)) \end{aligned}$$

where the last equality follows by dominated convergence. The Unconditional Partial Effect of X on F_Y is thus simply the average derivative of the conditional CDF of Y given X and V, where the derivative is taken with respect to X. We formally state this preliminary finding in the following lemma.

Lemma 2. *Suppose that the conditions of Lemma 1 hold, and that for some $c > 0$ and $\delta \in (-c, c)$ the support of $(X+\delta, V)$ is contained in the support of (X,V). Then*

$$\theta_{id}(\cdot) = \mathbb{E}(\partial_x F_{Y|X,V}(\cdot, X, V))$$

and is thus identified.

Using the last result, one can now easily extend the analysis of Unconditional Partial Effects to more general quantities $\Gamma(F_Y)$, if $\Gamma(\cdot)$ is sufficiently "smooth". In particular, we consider functionals that satisfy a Hadamard differentiability condition, where Γ is called Hadamard differentiable at F if there exists a continuous linear functional Γ'_F such that

$$\lim_{\delta \to 0} \left\| \frac{\Gamma(F + \delta h_\delta) - \Gamma(F)}{\delta} - \Gamma'_F(h_\delta) \right\| = 0 \tag{4.3.1}$$

for all sequences of function $h_\delta \to h$ such that $F + \delta h_\delta$ is contained in the domain of Γ for some sufficiently small value of δ. See van der Vaart (2000, Chapter 20.2) for further details.

To derive a general representation of Unconditional Partial Effects on $\Gamma(F_Y)$, define the function h_δ through $h_\delta = (F_{Y,\delta} - F_Y)/\delta$. We then obtain that

$$\begin{aligned} \theta_\Gamma &= \lim_{\delta \to 0} \frac{\Gamma(F_{Y,\delta}) - \Gamma(F_Y)}{\delta} \\ &= \lim_{\delta \to 0} \frac{\Gamma(F_Y + \delta h_\delta) - \Gamma(F_Y)}{\delta} \\ &= \Gamma'_F(\theta_{id}), \end{aligned}$$

where the last equality follows from the continuous mapping theorem since $h_\delta \to \theta_{id}$ as shown above. That is, we can identify general Unconditional Partial Effects by using the effect of X on the unconditional CDF of Y as a building block. We formalize this finding in the following Theorem.

Theorem 1. *Suppose that the conditions of Lemma 2 hold, and that the functional* Γ *is Hadamard differentiable at* F_Y *with derivative* Γ'_F*. Then the Unconditional Partial Effect of* X *on* $\Gamma(F_Y)$ *is given by* $\theta_\Gamma = \Gamma'_F(\theta_{id})$*.*

This representation of the Unconditional Partial Effect of X on $\Gamma(F_Y)$ given in Theorem 1 is convenient for two reasons. First, results on Hadamard differentiablity are widely available in the literature for many functionals of interest. Under appropriate conditions, this smoothness property is fulfilled for moments and quantiles, but also for inequality measures like the Gini coefficient and the Lorenz curve.

Second, the above representation is particularly useful when further parametric or semiparametric restrictions are imposed on the relationship of the outcome variable and the regressors. In this case, the Unconditional Partial Effect of X on F_Y itself is usually still easy to compute, and results for other statistics of interest follow immediately from Theorem 1. Our representation thus allows us to establish a tight link between the Unconditional Partial Effects and the structural features of the model. This result is not specific for models with endogeneity, but applies analogously to the exogenous case where the control variable V is not present. On the other hand, the general representation in Firpo, Fortin, and Lemieux (2009) for the exogenous case, using a projection of the recentered influence function of $\Gamma(F_Y)$ on the regressors, can be much more difficult to evaluate for specific models.

We now illustrate this last point by considering the case where the model in equation (4.2.1) is a standard linear quantile regression model (see Koenker (2005)). That is, suppose for the moment that ε is now a scalar random variable, normalized to be uniformly distributed on $[0, 1]$, and that

$$g(X, \varepsilon) = \beta_1(\varepsilon) + X\beta_2(\varepsilon),$$

where $\beta_1(\cdot)$ and $\beta_2(\cdot)$ are strictly monotonic functions. The form of (4.2.2) can remain unchanged. Using standard arguments, one obtains that under this specification we have that

$$\partial_x F_{Y|XV}(y, x, v) = -f_{Y|XV}(y, x, v)\beta_2(F_{Y|XV}(y, x, v))$$

and thus the Unconditional Partial Effect of X on F_Y is given by

$$\theta_{id}(\cdot) = -\mathbb{E}(f_{Y|XV}(\cdot, X, V)\beta_2(F_{Y|XV}(\cdot, X, V))).$$

Now consider the Unconditional Partial Effect of X on $\Gamma(F_Y)$, where $\Gamma(F)[\tau] = F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$ is the functional that transfers a CDF into its quantile function. Then under some standard restrictions (ensuring e.g. uniqueness of the quantiles) this map is Hadamard differentiable at F_Y with derivative

$$\phi \mapsto \Gamma'_{F_Y}(\phi) = -\left(\frac{\phi}{\partial_y F_Y}\right) \circ F_Y^{-1},$$

which leads to the following expression for the Unconditional Partial Effect:

$$\theta_\Gamma(\tau) = -\frac{\theta_{id}(F_Y^{-1}(\tau))}{f_Y(F_Y^{-1}(\tau))} = \frac{\mathbb{E}(f_{Y|XV}(F_Y^{-1}(\tau), X, V)\beta_2(F_{Y|XV}(F_Y^{-1}(\tau), X, V)))}{f_Y(F_Y^{-1}(\tau))}.$$

Note that this is a weighted average of the function β_2 evaluated at $F_{Y|XV}(F_Y^{-1}(\tau), X, V)$, which can be interpreted as the "rank" of $F_Y^{-1}(\tau)$ in the distribution of Y conditional on X and V. Firpo, Fortin, and Lemieux (2009) obtain a similar result for the exogenous case through more involved arguments (compare their Proposition 1). However, while their arguments apply to the specific case where $\Gamma(F_Y)$ is the quantile function only, our analysis can easily generalized to other statistics, such as the Lorenz curve or the Gini coefficient, as long as the Hadamard differentiability condition holds.

4.4 Conclusions

In this paper, we established the identification of Unconditional Partial Effects introduced by Firpo, Fortin, and Lemieux (2009) in general nonseparable models with endogenous regressors using a control variable approach due to Imbens and Newey (2009). We also show that these effects can be written in terms of an average derivative of the conditional CDF of the outcome variable Y given the regressors X and the control variable V, where the derivative is taken with respect to X. This representation is useful to give an explicit expression for Unconditional Partial Effects in nonlinear parametric or semiparametric models.

Bibliography

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models.," *Journal of the American Statistical Association*, 97(457), 284–293.

Abrevaya, J. (2001): "The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes," *Empirical Economics*, 26(1), 247–257.

Ahn, H., H. Ichimura, and J. Powell (1996): "Simple Estimators for Monotone Index Models," *manuscript, Department of Economics, UC Berkeley*.

Ai, C. (1997): "A Semiparametric Maximum Likelihood Estimator," *Econometrica*, 65(4), 933–963.

Akritas, M., and I. Van Keilegom (2001): "Non-parametric Estimation of the Residual Distribution," *Scandinavian Journal of Statistics*, 28(3), 549–567.

Almond, D., K. Chay, and D. Lee (2005): "The Costs of Low Birth Weight," *The Quarterly Journal of Economics*, 120(3), 1031–1083.

Anderson, G. (1996): "Nonparametric Tests of Stochastic Dominance in Income Distributions," *Econometrica*, 64(5), 1183–1193.

Atkinson, A. (1970): "On the Measurement of Inequality," *Journal of Economic Theory*, 2(3), 244–263.

Barrett, G., and S. Donald (2000): "Statistical Inference with Generalized Gini Indices of Inequality and Poverty," Working Paper.

——— (2003): "Consistent Tests for Stochastic Dominance," *Econometrica*, 71(1), 71–104.

Bhattacharya, D. (2007): "Inference on inequality from household survey data," *Journal of Econometrics*, 137(2), 674–707.

Black, S., P. Devereux, and K. Salvanes (2007): "From the Cradle to the Labor Market? The Effect of Birth Weight on Adult Outcomes," *The Quarterly Journal of Economics*, 122(1), 409–439.

Blau, F., and L. Kahn (1997): "Swimming Upstream: Trends in the Gender Wage Differential in the 1980s," *Journal of Labor Economics*, 15(1), 1.

Blundell, R., and J. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, 2.

——— (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71(3), 655–679.

Blundell, R., and J. Powell (2007): "Censored Regression Quantiles with Endogenous Regressors," *Journal of Econometrics*, 141(1), 65–83.

Bosq, D. (1998): *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer.

Chen, X., O. Linton, and I. Van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Chernozhukov, V., I. Fernandez-Val, and B. Melly (2008): "Inference on Counterfactual Distributions," Working Paper.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

Chernozhukov, V., G. Imbens, and W. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139(1), 4–14.

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71(5), 1405–1441.

——— (2005): "Nonparametric Identification under Discrete Variation," *Econometrica*, 73(5), 1525–1550.

Das, M., W. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70(1), 33–58.

Davidson, R., and J. Duclos (2000): "Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality," *Econometrica*, 68(6), 1435–1464.

Delecroix, M., M. Hristache, and V. Patilea (2005): "On Semiparametric M-estimation in Single-Index Regression," *Journal of Statistical Planning and Inference*, 136(3), 730–769.

DiNardo, J., N. Fortin, and T. Lemieux (1996): "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, 64(5), 1001–1044.

Donald, S., D. Green, and H. Paarsch (2000): "Differences in Wage Distributions between Canada and the United States: An Application of a Flexible Estimator of Distribution Functions in the Presence of Covariates," *Review of Economic Studies*, 67(4), 609–633.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.

Firpo, S., N. Fortin, and T. Lemieux (2009): "Unconditional Quantile Regressions," *Econometrica*, forthcoming.

Florens, J., J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76(5), 1191–1206.

Gasser, T., H. Müller, and V. Mammitzsch (1985): "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(2), 238–252.

Gosling, A., S. Machin, and C. Meghir (2000): "The Changing Distribution of Male Wages in the UK," *Review of Economic Studies*, 67(4), 635–666.

Hoderlein, S., and E. Mammen (2007): "Identification of Marginal Effects in Nonseparable Models Without Monotonicity," *Econometrica*, 75(5), 1513–1518.

Horowitz, J., and W. Härdle (1996): "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates.," *Journal of the American Statistical Association*, 91(436), 1632–1640.

Härdle, W., P. Hall, and H. Ichimura (1993): "Optimal Smoothing in Single-Index Models," *Annals of Statistics*, 21(1), 157–178.

Härdle, W., P. Janssen, and R. Serfling (1988): "Strong Uniform Consistency Rates for Estimators of Conditional Functionals," *Annals of Statistics*, 16(4), 1428–1449.

Ichimura, H. (1993): "Semiparametric least squares(SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58(1-2), 71–120.

Ichimura, H., and C. Taber (2002): "Semiparametric Reduced-Form Estimation of Tuition Subsidies," *American Economic Review*, 92(2), 286–292.

Ichimura, H., and P. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," *Handbook of Econometrics*, 6.

Imbens, G., and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, fortcoming.

Jones, M., and D. Signorini (1997): "A Comparison of Higher-Order Bias Kernel Density Estimators.," *Journal of the American Statistical Association*, 92(439).

Klein, R., and R. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61(2), 387–421.

Koenker, R. (2005): *Quantile Regression*. Cambridge University Press.

Koenker, R., and K. Hallock (2001): "Quantile Regression," *Journal of Economic Perspectives*, 15(4), 143–156.

Kong, E., O. Linton, and Y. Xia (2009): "Uniform Bahadur Representation for Local Polynomial Estimates of M-regression and its Application to the Additive Model," *Econometric Theory*, in press.

Lee, L. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65(2), 381–428.

Lee, S. (2007): "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 141(2), 1131–1158.

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

Li, Q., and J. Racine (2007): *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Linton, O., E. Maasoumi, and Y. Whang (2005): "Consistent Testing for Stochastic Dominance under General Sampling Schemes," *Review of Economic Studies*, 72(3), 735–765.

Linton, O., S. Sperlich, and I. Van Keilegom (2008): "Estimation of a Semiparametric Transformation Model," *Annals of Statistics*, 36(2), 686–718.

Machado, J., and J. Mata (2005): "Counterfactual Decomposition of Changes in Wage Distributions using Quantile Regression," *Journal of Applied Econometrics*, 20(4), 445–465.

Manski, C. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3(3), 205–228.

Marron, J. (1994): "Visual Understanding of Higher-Order Kernels," *Journal of Computational and Graphical Statistics*, 3(4), 447–458.

Matzkin, R. (2003): "Nonparametric Estimation of Nonadditive Random Functions," *Econometrica*, 71(5), 1339–1375.

McFadden, D. (1989): "Testing for Stochastic Dominance," *Studies in the Economics of Uncertainty (in honor of J. Hadar), Springer-Verlag*.

Melly, B. (2005): "Decomposition of Differences in Distribution using Quantile Regression," *Labour Economics*, 12(4), 577–590.

Newey, W. (1985): "Semiparametric Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Annales de l'INSEE*, 59(60), 219–235.

——— (1987): "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 36(3), 231–250.

——— (1994a): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10(2), 233–253.

——— (1994b): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62(6), 1349–1382.

——— (2007): "Nonparametric Continuous/Discrete Choice Models," *International Economic Review*, 48(4), 1429–1439.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simulataneous Equations Models," *Econometrica*, 67(3), 565–603.

Nolan, D., and D. Pollard (1987): "U-processes: Rates of Convergence," *Annals of Statistics*, 15(2), 780–799.

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

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027–1057.

Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–1430.

Rivers, D., and Q. Vuong (1988): "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39(3), 347–366.

Sherman, R. (1994): "Maximal Inequalities for Degenerate U-Processes with Applications to Optimization Estimators," *Annals of Statistics*, 22(1), 439–459.

Smith, R., and R. Blundell (1986): "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, 54(3), 679–686.

Stock, J. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84(406).

——— (1991): "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits," *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics, Cambridge*.

Stoker, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54(6), 1461–1481.

van der Vaart, A. (2000): *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer.

Xia, Y. (2006): "Asymptotic Distributions For Two Estimators Of The Single-Index Model," *Econometric Theory*, 22(06), 1112–1137.

Xia, Y., H. Tong, W. Li, and L. Zhu (2002): "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, 64(3), 363–410.

Eidesstattliche Erklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig angefertigt und mich keiner anderen als der in ihr angegebenen Hilfsmittel bedient zu haben. Insbesondere sind sämtliche Zitate aus anderen Quellen als solche gekennzeichnet und mit Quellenangaben versehen.

Mannheim, 3. April 2009

Christoph Rothe

Lebenslauf

Personliche Daten

Christoph Rothe

geboren am 6. März 1981 in Dortmund

Ausbildung

10/2005 – heute	Promotion in Volkswirtschaftslehre an der Universität Mannheim Mitglied des CDSE / CDSEM
10/2000 – 05/2005	Studium der Statistik an der Universität Dortmund Abschluß als Diplom-Statistiker
02/2003 – 12/2003	Studium der Statistik an der University of Auckland, Neuseeland
06/1992 – 06/2000	Abitur am Ruhr-Gymnasium Witten

Mannheim, 3. April 2009