.

Semiparametric Estimation of Selectivity Models

Wissenschaftliche Arbeit
zur Erlangung des Grades eines Diplom-Volkswirtes
an der Fakultät für Wirtschaftswissenschaften und Statistik
der Universität Konstanz.

ü

Bearbeitungszeit: 11. Juni bis 11. August 1998
1. Gutachter: Prof. Dr. Winfried Pohlmeier
2. Gutachter: PD Dr. Werner Smolny

Konstanz, den 6. August 1998

Contents

Acknowledgments

For technical support and ready advice on technical questions I am deeply grateful to Dr. Marcia Schafgans, London School of Economics, to Frank Hettich and to Martin Bodenstein, both University of Konstanz, Germany. Furthermore I thank Ms. Liliana Smau for her exceptional patience.

Chapter 1

Introduction

Commonly the purpose of econometrics is to investigate a, perhaps economical, relationship between some explanatory variables and some dependent variables for a certain population. Were both the explanatory and the dependent variables known for all population members, this analysis would be trivial with all conceivable information at hand. Usually only a (small) subgroup of this population, the sample, is observed or interviewed and estimation of the relationship for the whole population must restrict itself to this available information. Two distinct types of samples can be distinguished: random (or representative) samples and non-random samples. Random samples are drawn completely randomly from the population and thus are representative for the whole population, as every population member has the same probability to be within the sample. All other samples, where the probability to be included in the sample varies among the population members, are accordingly non-random samples. Commonly this variation or unevenness is of unknown form. At a first glance only random samples appear to be meaningful for extrapolation to the *entire population*, since non-random samples do not mirror the true population members' composition, though truly random samples may frequently be difficult or not at all to obtain.

For instance, when comparing male and female wage structures for certain age groups, almost random samples could be drawn from the male population, if almost all men work. This is often inconceivable for females, since not all women work and wage data on non-working women are unknown or unreliable. Or, when evaluating the quality of education, care must be taken to account for the (at least partly) self-selection of the students and it must be recognized that the students of the competing educational institutions are as groups inherently, perhaps invisibly, different. Sample selection of any non-random form always leads to these inherently different groups, both for self-selection, e.g. union/non-union membership and for samples who are selected by any other authority, like the program administrator of manpower training programs.

One tool to cope with non-random samples, where the in-sample probability is unknown, are selectivity models, also known as sample selection models. Selectivity models always consist of two parts. One is the structural part, embodying the desired population relationships. The second, the selection part, takes account of the non-representative nature of the present non-random sample, to model explicitly the process of observations entering into the sample and to determine the in-sample probabilities for all population members. This compound allows to correct the non-randomness of the present sample and to yield *representative* estimates for the population relationships.

Consider, for instance, an analysis of the structural wage determinants for the population of workers that are selected either into the agriculture, manufacturing or service sector. Accordingly the whole population is partitioned into three self-selected subpopulations, from which random samples can be drawn. These subpopulations are self-selected since the workers themselves decide (or at least influence) the sector in which they work, with other words into which sector they select into. Hence a random sample drawn from the subpopulation of agricultural workers would not be representative neither for the entire population nor for the subpopulations manufacturing workers, service workers. The tricky part of the estimation of the structural relationships arises through the composition of the present dataset, which is a conjunction of some, in our case three, random part-samples, each representative for a certain subpopulation, but as a compound neither representative for any subpopulation nor for the whole population at all.

These structural wage relationships are not to be confused with a mere tableau of separate and unconnected wage equations, with one wage equation for agricultural workers, that is valid only for population members who actually work in the agricultural sector, one for manufacturing and one for service workers. Those could easily be estimated from the random part-samples for each subpopulation. Those wage equations, each conditioned on the corresponding selection, would approve only for the corresponding subpopulations and that approach will be briefly broached as the two-part model in the chapter Parametric Estimation. The structural wage approach of the selectivity model strives much further, since its' relationships are valid and representative for the *entire population*. For any individual drawn randomly from the population it's sectoral job choice is illuminated and it's earning prospects in each sector. These wage relationships inherently provide potential (or shadow) outcomes for all population members for all three sectors, although even in an "infinite" sample, containing all information from all population members, only one third of all potential wages can be observed, since each member is selected into exactly one of the three sectors. Hence the sector chosen should be attributed as an additional individual's characteristic and even with all other characteristics equal an individual working in manufacturing is to distinguish from someone working in

the service sector.

Assuming for a moment that the structural relationships have been estimated already, then potential wages could be predicted. These are wages that an individual, selected into a certain sector, would expect to earn in the both other sectors, if this individual would decide to switch the sector. By this, it could be evaluated how much earning loss[1] an individual would suffer, would it be forced to work in an other sector, which, for instance, could occur if the agricultural sector becomes eliminated by market forces, provided that no other general-equilibrium effects, scarcity effects etc. have changed the structural relationships. With these potential wages at hand structural differences or treatment effects can be evaluated. The evaluation of individual causal treatment effects, that is the outcome difference between the two hypothetical stages: the outcome achieved, if the individual would have participated and the outcome achieved, if the *same* individual would not have participated, is of high relevance in clinical trials, medicine and all social programs, like manpower training. It is often modelled by a selectivity model with two sectors (in sector one the participants, in the other sector the non-participants). This difference is called the, by definition unobservable, causal effect of a treatment, since the participation in treatment shifts the individual into the other category and causes by this a different outcome, with all other characteristics, parameters etc. unchanged.

This motivation exposed some of the opportunities of selectivity models, though the difficulties with their estimation should not be overseen. Considering the two part structure, the selection part models the sectoral choice of any individual by assigning the individual to the chosen sector on the basis of this individual's characteristics. Once the individual's selection is accomplished, the structural part explains the outcomes for this individual. By including all individual characteristics in both parts the selection mechanism and the corresponding wages were completely determined, provided true relationships exist at all. But often, not all individuals' characteristics are observed and also random, unobservable influences may disturb the longed for strict deterministic relationships. This is usually taken into consideration by announcing stochastical population relationships and the introduction of stochastical error terms capturing these both effects. Through these error terms, appearing in both the selection and the structural part, arises the protagonizing estimation problem, since signals between the selection and the structural parts may be channelled unnoticed via a link between the error terms, introducing an unobserved *selectivity bias*. To demonstrating a striking example, assume that, in a two-sector-setting (employed/unemployed) where the wages for the employed and the employment participation are analyzed, the both error terms exclusively represent the impact of an unobserved individuals' characteristic labelled motivation on the employment participation and the wage, respectively. If higher motivation

[1] A gain seems often unlikely since the individuals have self-selected their opportune sector.

leads to higher wages and also to an increased willingness to work, both error terms would be positively correlated and the observed wages of the working subpopulation would be higher than expected by their observed characteristics through the double effect of the unobserved motivation and would be too high to represent the entire population, all other characteristics the same. In like manner, for a negative correlation between the selection error and the structural error the observed sample would understate the true population relationship. This selectivity bias cannot be attributed to any observed characteristics and demands a particular treatment of the error terms.

The early approaches in the 1970s, regarding this error term channel, aimed to model this link by parametrically specifying the joint error distribution up to a finite amount of coefficients. Regularly the bivariate normal distribution has been imposed with the unknown covariance reflecting the selectivity link (Heckman, 1974,[17] and 1976,[18]). In the 1980s the conjecture emerged, also foreseen by Heckman, that this distributional restriction with the only flexibility being the error term correlation $\in [-1, 1]$, might be too strict to fully comprehend the error terms' link and eliminate the selectivity bias, which has then been underlined by several studies, e.g. Arabmazar and Schmidt (1982,[4], Goldberger (1983,[15]) and Schafgans (1997,[46])). This awareness fueled the development of semiparametric techniques to estimating selectivity models, that leave substantially more freedom to the form of this error term link, hence achieving the elimination of the selectivity bias under more general assumptions. Parametric modelling proceeds by fully specifying every unknown function and error term distribution to a parametric family with a finite amount of unknown parameters, e.g. intercept and slope coefficients for a linear functional specification or mean and standard deviation for normal distribution family. Fully nonparametric models depreciate such arbitrary assumptions and are satisfied with smoothness and regularity conditions on the functions and densities. Semiparametric models are a half-breed of both since they specify parts of the model parametrically but leave others unspecified. In semiparametric selectivity modelling the regressor functions are commonly specified and the error term densities left nonparametric, though this pattern is not cogent.

In this spirit it is an objective of this work to provide a thorough theoretical illumination of the properties of dichotomous (= 2 sector) selectivity models, but it's aims are striking further, as it intends to deliver a practical guide of how to apply semiparametric estimation in these dichotomous selectivity models. To this end the main results of the most substantial techniques are presented in conjunction with all the relevant formulae required to put these estimators to use.

Attention will be given to dichotomous models, as these are the most widely applied, have received by far the most of scientific reflection and allow to be covered without ramification into different specifications of selectivity models such as ordered

and unordered selection models etc. In addition, as the size of this thesis is restricted, broaching also more general selectivity models would involve a reduction of the space devoted to the presentation of the dichotomous models, then running danger to miss the above outlined objectives of a thorough and comprehensive exposition of the dichotomous model. The inclined reader with further ambitions be referred to the overview article of Vella (1998,[50]).

This trade off was in like manner settled in favour to practical ends, when leaving out theoretical discussions, properties and proofs of questionable essentiality. Nevertheless, a more general model will be sketched in the succeeding chapter Model Framework, though afterwards the scope of contemplation will be restricted to selectivity models with selection into either one of two categories.

To this end the thesis is organized as follows: Chapter two introduces the formal framework used and restricts the discussion to selectivity models with only two sectors. Chapter three covers then the theoretical needs to identify the structural relationships in infinite samples. Chapter four sketches briefly the parametric estimation techniques and the two-part model, that may suit as an alternative if the selectivity model disapproves the identification requirements. Chapter five describes in detail the more promising available semiparametric estimators for the selectivity model. In Chapter six the asymptotic and finite sample properties of these semiparametric estimators are outlined, while Chapter seven illustrates how these estimators are applied in practical work. Chapter eight presents some final conclusions.

Chapter 2

Model Framework

The difficulties arising from selectivity when investigating structural population relationships and methods to overcome these should now be analyzed within a formal setting for selectivity models with cross-sectional data.

Assume the existence of J distinct and excluding categories into which each individual will be selected into exactly one. This could be, for instance, union-membership, non-membership ($J = 2$) or different job sectors like agriculture, manufacturing and services ($J = 3$), where always it is to ensure that all individuals fit into exactly one. Furthermore assuming that for each category a unique population relationship between some explanatory variables and a continuous outcome variable exists, the system of population relationships can be sketched as:

$$\begin{aligned} Y_{i1}^* &= f_1(X_i) + \varepsilon_{i1} \qquad \text{with } E[\varepsilon_{i1}|X_i] = 0 \\ Y_{i2}^* &= f_2(X_i) + \varepsilon_{i2} \qquad \text{with } E[\varepsilon_{i2}|X_i] = 0 \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ Y_{iJ}^* &= f_J(X_i) + \varepsilon_{iJ} \qquad \text{with } E[\varepsilon_{iJ}|X_i] = 0. \end{aligned} \tag{2.1}$$

The functions $f_1, .., f_J$ represent the unknown population relationship for each category, determining the corresponding continuous outcome variables $Y_{i1}^*, .., Y_{iJ}^*$ through the observed explanatory regressors X_i of dimension $\Re^p$, which for convenience are defined as the conjunction of all relevant explanatory variables for all categories.

Within each population relationship j, henceforth called structural equation, a mean-zero one-dimensional random error term ε_{ij} is added to catch for unobserved and unobservable individual characteristics and to account for measurement errors or other influences which are not part of the population relationship. The continuous dependent variables $Y_{i1}^*, .., Y_{iJ}^*$ are earmarked by a * as potential or latent outcomes, inasmuch as only exactly one of these outcomes is observed as Y_i, depending on which category the individual i has been selected into. They represent latent outcomes as the outcome Y_{ij}^* would be observed if the individual would be in category j, but otherwise denote unobserved hypothetical expectations. In contrast to the

following chapter Identification, here all characteristics, potential outcomes Y_{ij}^* and other observed features are non-stochastical realizations of the underlying process and only the error terms are random variables.

Defining for each individual i a discrete variable $d_i \in \{1, .., J\}$ as selection indicator, then for an individual selected into category j, hence $d_i = j$, only the outcome $Y_i = Y_{ij}^*$ would be observed. More generally, all the actually observed outcomes Y_i are given by:

$$Y_i = \sum_{j=1}^{J} Y_{ij}^* \cdot 1(d_i = j), \qquad i = 1..N, \tag{2.2}$$

with $1(\cdot)$ being the binary indicator function and N the sample size. Y_i, $i = 1..N$ is the sample of observed characteristics.

Here arises the selectivity problem as a missing data problem, since the true population relationships $f_j(\cdot)$ are of interest, but only limited observations on their outcomes are available. To restricting the scope to these available observations Y_i the model should be transformed by taking expectations conditional on observability.

$$\begin{aligned} E[Y_{i1}^*|X_i, d_i &= 1] = f_1(X_i) + E[\varepsilon_{i1}|X_i, d_i = 1] \\ E[Y_{i2}^*|X_i, d_i &= 2] = f_2(X_i) + E[\varepsilon_{i2}|X_i, d_i = 2] \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\ E[Y_{iJ}^*|X_i, d_i &= J] = f_J(X_i) + E[\varepsilon_{iJ}|X_i, d_i = J]. \end{aligned} \tag{2.3}$$

Now, easily can be seen that estimating the structural equations by OLS[1] is likely to be inconsistent by omitting the regressors $E[\varepsilon_{ij}|X_i, d_i = j]$ [2], which are generally not zero, even though $E[\varepsilon_{ij}|X_i] = 0$. Defining a continuous selection correction term $\lambda_{ij} \equiv E[\varepsilon_{ij}|X_i, d_i = j]$ for each structural equation $j \in \{1, .., J\}$, this term takes account of the non-zero average error term for the subsample of all individuals selected into the same category j. With this unknown (non-linear) selection correction term the original model can be augmented by this beforehand omitted additional regressor to:

$$\begin{aligned} Y_{i1}^* &= f_1(X_i) + \lambda_{i1} + \xi_{i1} \qquad \text{with } E[\xi_{i1}|X_i, d_i = 1] = 0 \\ Y_{i2}^* &= f_2(X_i) + \lambda_{i2} + \xi_{i2} \qquad \text{with } E[\xi_{i2}|X_i, d_i = 2] = 0 \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\ Y_{iJ}^* &= f_J(X_i) + \lambda_{iJ} + \xi_{iJ} \qquad \text{with } E[\xi_{iJ}|X_i, d_i = J] = 0, \end{aligned} \tag{2.4}$$

with $\xi_{ij} = \varepsilon_{ij} - \lambda_{ij}$ and $E[\xi_{ij}|X_i, d_i = j] = E[\varepsilon_{ij} - \lambda_{ij}|X_i, d_i = j] = E[\varepsilon_{ij} - E[\varepsilon_{ij}|X_i, d_i = j]|X_i, d_i = j] = 0$.

[1] Ordinary Least Squares, e.g. by specifying $f_j(X_i) = X_i'\beta$.

[2] Inconsistency occurs if the omitted regressor is non-orthogonal to f_j.

To complete this model a selection mechanism determining the selection status of each individual has to be supplemented. Let the selection equation

$$d_i = \varphi(Z_i, u_i) \qquad \text{with } d_i \textit{ discrete } \in \{1, .., J\} \tag{2.5}$$

being a flexible representation of the selection process, with Z_i selection regressors of dimension $\Re^q$, perhaps identical with X_i, u_i an univariate unobserved error term and $\varphi(\cdot)$ a selection function mapping in behalf of both arguments into the discrete space $1, .., J$. It is assumed that the discrete indicator d_i and the variables Z_i are the only information available about the selection process. Consequently no more details about the intensity of choice, individual uncertainty or doubts between two categories etc. can be extracted from the dataset. With this selection mechanism complement the selectivity model looks promising as the error terms ξ_{ij} are mean-zero conditional on observability and the unknown selection correction terms λ_{ij} have been somehow connected with the selection procedure.

$$\begin{aligned} Y_{i1}^* &= f_1(X_i) + \lambda_1(Z_i) + \xi_{i1} \\ Y_{i2}^* &= f_2(X_i) + \lambda_2(Z_i) + \xi_{i2} \\ &\vdots \qquad\qquad \vdots \\ Y_{iJ}^* &= f_J(X_i) + \lambda_J(Z_i) + \xi_{iJ} \\ d_i &= \varphi(Z_i, u_i) \quad \in \{1, .., J\} \\ Y_i &= \sum_{j=1}^{J} Y_{ij}^* \cdot 1(d_i = j) \end{aligned} \tag{2.6}$$

with (Y_i, d_i, X_i, Z_i) observed for each individual $i = 1..N$.

A main drawback with specifying and estimating this general model with multiple categories is not with the number of structural equations, that could all be estimated separately, but with the specification of the selection equation, becoming more obvious soon. Henceforth in the remainder of this work attention will be restricted to the handling of dichotomous selectivity models, that are models with just two different categories to be selected into, like participants and non-participants, union-members and non-members. Besides it's high popularity in practical work, the selection procedure can easily and quite generally be implemented by a threshold-crossing mechanism:

$$d_i = 1(Z_i'\gamma - u_i > \tau) \quad \in \{0, 1\}. \tag{2.7}$$

A continuous, one-dimensional index value $Z_i'\gamma - u_i$ is computed from the selection variables Z_i and compared with an unknown threshold. Being the index value above this threshold the individual is selected into category one, otherwise into the other category. Although the threshold τ [3] being unknown, the probability being in group one increases monotonously with this index value and vice versa.

[3] The threshold τ is non-identified, since it is undistinguishable from an intercept γ_o of $Z_i'\gamma$.

In the case of three or more different categories such a simple threshold-crossing mechanism cannot be implemented on a general basis. Now the selection procedure has to be scrutinized with care, which could be for instance a multinomial choice on basis of individual utility maximization by comparing all potential outcomes Y_{ij}^* and choosing the most suited. In this setting the different categories would be unordered and the selection equation could be $d_i = \arg\max_j [Y_{ij}^*]$, if outcome maximization is implied by the individual preferences. On the other hand, the system of different categories may entail a fixed order of preference or assignment, which could then be modelled by multiple thresholds. Hence different selection procedures would require different modelling of the selection equation, which would burst the frame of this work.

Concentrating on dichotomous selectivity models, sometimes called switching regression models, in selection correction form:[4]

$$\begin{aligned} Y_{i1}^* &= f_1(X_i) + \lambda_1(Z_i) + \xi_{i1} \\ Y_{i0}^* &= f_0(X_i) + \lambda_0(Z_i) + \xi_{i0} \\ d_i &= \varphi(Z_i, u_i) \qquad \in \{0,1\} \\ Y_i &= d_i \cdot Y_{i1}^* + (1 - d_i) \cdot Y_{i0}^* \end{aligned} \tag{2.8}$$

with (Y_i, d_i, X_i, Z_i) observed for all individuals.

For means being explained in the chapter Identification, an index function $\theta : \Re^q \longrightarrow \Re$ is often introduced to replace Z_i by $\theta(Z_i)$ in both the structural equations and the selection equation:

$$\begin{aligned} Y_{i1}^* &= f_1(X_i) + \lambda_1(\theta(Z_i)) + \xi_{i1} \\ Y_{i0}^* &= f_0(X_i) + \lambda_0(\theta(Z_i)) + \xi_{i0} \\ d_i &= \varphi(\theta(Z_i), u_i). \end{aligned} \tag{2.9}$$

This single index function can be any continuous and differentiable function mapping from $\Re^q$ to $\Re$-space. For convenience $\theta(Z_i) = Z_i'\gamma$ is the usual choice with γ an unknown coefficient vector of dimension q. This modification, together with the assumption that this single index form of the selectivity model represents the true model, yields a considerable reduction of the dimensionality of the identification issue and improved estimator properties.

Two peculiar variants of this dichotomous selectivity model are to be mentioned. The first, the *censored* selectivity model, which will be treated as the base-line

Secondly the discrete choice coefficients are identified only up to a scale parameter σ, the unknown standard deviation. Accordingly, the threshold is commonly set to zero, $\tau \equiv 0$, except when an intercept γ_0 is precluded.

[4] To conform with the literature both categories will henceforth be denoted by the binary numbers 0 and 1.

model henceforward, occurs when the potential outcomes Y_i^* of one category are not at all observed within the sample. This occurs if the outcomes of individuals who are selected into this category are unknown, unobserved or not truly recorded, perhaps due to unobservability of say negative values or as a consequence of the way the dataset has been established. For instance, when investigating structural wage equations commonly wages are only observed for the category of officially employed workers, while wages for non-workers are not recorded, though they may be working in non-registered jobs and hide information on earnings. Obviously the structural relationship of the censored category cannot be identified, but the selection equation and the other category's structural equation are unaffected by censoring as still all the relevant characteristics (d_i, X_i, Z_i) are recorded for all individuals.

The second variant are *truncated* selectivity models, where one category is completely missing in the dataset in such a way that neither Y_i^* nor X_i or Z_i are available for individuals selected into this category. Often it may be that even the number of those mysterious individuals selected into this category is unknown. In contrast to the censored model the complete absence of information on a certain category does not only prohibit the identification of this category's structural relationship, but worsens furthermore the estimation of the selection equation and the other structural equation considerably, since no data on the number of non-participants nor on their characteristics Z_i is in reach. As truncated models have not received much attention in the research literature and as their estimates may be pretty imprecise, they will not be investigated further. The inclined reader be referred to the estimator of Ichimura and Lee (1991,[22]), which is suited for truncated binary selectivity models.

An adjacent variant modelling selectivity is a 2-category model with Tobit type selection mechanism where more information on the selection process is available from the dataset through a continuously observed participation indicator. Although the selection will finally distinguish only between participants and non-participants, the degree or intensity with which this corresponding group was chosen is indicated. Exemplifying this by specifying the selection equation of the standard dichotomous selectivity model as

$$d_i = 1(Z_i'\gamma - u_i > 0), \tag{2.10}$$

with the selection mapping function $\varphi(\cdot) = 1(\cdot > \tau)$ and the unidentified threshold τ set to zero. In the non-Tobit type model only (Y_i, d_i, X_i, Z_i) are observed. On the other hand, in the Tobit type model

$$\begin{aligned} h_i &= Z_i'\gamma - u_i \\ d_i &= 1(h_i > 0), \end{aligned} \tag{2.11}$$

additionally the value h_i can be observed, interpretable as an indicator of the individual intensity or sureness of the selection decision. An h_i close to zero suggests

indifference, unsureness or doubts, whereas h_i large exhibiting clear-cut, doubtless preferences without ambiguity. With continuous information on h_i available more precise estimation of the selection process is conceivable. Another valuable advantage of this Tobit-type model is that it's identification does not hinge on exclusion restrictions in contrast to the standard selectivity model, see Lee (1994,[27]).

A practical application of a censored Tobit type selection model is the analysis of wage determinants where for the employed individuals (category 1) not only their corresponding wage is observed but additionally the number of hours worked h_i, which is zero by definition for the non-employed. Nevertheless, the occurrence of these Tobit type selectivity models appears to be rather an exception in the literature and will not be examined henceforward. Consult Lee for an estimator and further details.

Chapter 3

Identification

Before setting forth with estimating the selectivity model it should be clarified at first what is of interest and second what at all can be identified. The term identification is generally referred to in a theoretical concept of estimation in infinite (asymptotic) samples which is the best case scenario where all conceivable information through observations is available. The expression identification if often also applied in a much looser sense as the exactitude or precision with which a coefficient is estimated in finite samples, hence is related to the efficiency of an estimator.

Before announcing the objects of interest, the base-line dichotomous model in selection correction form should be reconsidered. With respect to the interpretation of the model this chapter Identification, in contrast to the previous and the following chapters, departs from quite a different point of view. Not an available dataset from which the true process has to be investigated is the starting point; but rather by originating from the true underlying relationships and by analyzing the way of how observations come into existence the identifiable parts of the model are determined, which could then be estimated from a proper sample. To stress this different approach the applied notation will be switched from the sample data representation with (Y_i, d_i, X_i, Z_i) actually observed non-stochastical realizations to the representation of the true process with (Y, d, X, Z) random variables. The former notation will be adopted again in the following chapters, where estimation from a given sample is the focus. Consequently the model looks:

$$\begin{aligned} Y_1^* &= f_1(X) + \varepsilon_1 \\ Y_0^* &= f_0(X) + \varepsilon_0 \\ d &= \varphi(Z, u) \quad \in \{0, 1\} \end{aligned} \tag{3.1}$$

with the (homoskedasticity) assumption that the distribution of $(\varepsilon_1, \varepsilon_0, u)$ is independent of (X, Z), henceforward labelled: **[IA]** *Independence of the error terms and the regressors*, Powell (1994,[42], p.2476).

Considering now the objects of interest. Most researchers devote their attention

entirely to the estimation of the structural relationships, the functions $f_1(X)$, $f_0(X)$, though other features of this model may be of interest as well, like the selection mechanism, individual participation probabilities, error term distributions etc.

To deal now with the difficulties with identifying the selectivity model, Manski (1993,[32]) elucidates the selection problem in detail. What is at the heart of interest is the relationship between some exogenous individual characteristics (X, Z) and the potential outcome variables (Y_1^*, Y_0^*), which can be represented by the conditional probability densities $pr(Y_1^*|X, Z)$ and $pr(Y_0^*|X, Z)$. These probabilities can be rewritten to

$$\begin{aligned} pr(Y_1^*|X,Z) &= pr(Y_1^*|X,Z,d=1)pr(d=1|X,Z) + pr(Y_1^*|X,Z,d=0)pr(d=0|X,Z) \\ pr(Y_0^*|X,Z) &= pr(Y_0^*|X,Z,d=1)pr(d=1|X,Z) + pr(Y_0^*|X,Z,d=0)pr(d=0|X,Z). \end{aligned}$$

Now the selection problem becomes apparent, as only $pr(d = 1|X, Z)$, $pr(d = 0|X, Z)$, $pr(Y_1^*|X, Z, d = 1)$ and $pr(Y_0^*|X, Z, d = 0)$ can be identified since data on these values are available in infinite samples, while $pr(Y_1^*|X, Z, d = 0)$ and $pr(Y_0^*|X, Z, d = 1)$ can by definition never be observed. As a consequence, without further assumptions standard methods are unable to identify $pr(Y_1^*|X, Z)$ or $pr(Y_0^*|X, Z)$ because necessary observations are inherently missing.

To cope with this selection problem, early suggestions were to assume independence of (Y_1^*, Y_0^*) and d conditional on (X, Z), called ignorable nonresponse by statisticians. This means that the selection indicator does not entail any unobserved information about the structural outcome or more exemplary that a worker who has chosen to select into the manufacturing sector would earn the same wage in the service sector as an other worker who chose voluntarily the service sector, all observable characteristics equal. This disregards all unobservable influences like motivation and opportune selection and implies the unimportance of selection on the structural relationships, since $pr(Y_1^*|X, Z, d = 1) = pr(Y_1^*|X, Z, d = 0) = pr(Y_1^*|X, Z)$ and analog for Y_0^*. This assumption is equivalent to independence of $(\varepsilon_1, \varepsilon_0)$ and u conditional on (X, Z)[1], see Manski (1993,[32], p.79). This naive assumption can only be justified if the selection process is characterized by an assignment depending either entirely on observed characteristics or by a random selection, such that both subsamples are representative of the whole population in all respects including the unobservable characteristics. In the more common case that some unobservable characteristics guide the selection, e.g. when individuals self-select themselves or are assigned by someone else to these categories, the two subgroups will be inherently different. Though these differences must not necessarily be revealed by the observable characteristics but rather could be entailed invisibly in the unobservable error terms.

[1] This assumption should not be confused with the [IA] independence assumption of $(\varepsilon_1, \varepsilon_0, u)$ and (X, Z).

Heckman (1990,[19]) proofed that this naive assumption is not essential for identification. In his nonparametric Identification Theorem he shows that the distributions $F_{\varepsilon_1,u}(\cdot)$, $F_{\varepsilon_0,u}(\cdot)$ and the functions $f_1(X)$, $f_0(X)$ and γ are *nonparametrically* identified from the observable distributions $F(Y_1^*|X,Z,d=1)$, $F(Y_0^*|X,Z,d=0)$ and the distribution of d through $pr(d=1|Z)$. Yet the full joint distribution $F_{\varepsilon_1,\varepsilon_0,u}(\cdot)$ is not. His theorem rests on the independence assumption of errors and regressors [IA] and a further condition precluding that all the information entailed in Z is completely covered by X, which would be the case if Z were a subset of X. Hence at least one variable must be contained in Z that influences the selection and $\lambda(\cdot)$, but is not included in X and does not enter into the outcome equation via the structural relationship f. This exclusion restriction of at least one variable will also be seen as crucial for semiparametric identification, as it allows to isolate $f_1(X)$, $f_0(X)$ from selection influences due to Z. Heckman further notes, that the continuity of the distribution of Z and the continuity of the index $\theta(Z) = Z'\gamma$ over the whole $\Re$-line is crucial.

Since the identification of the whole conditional probability distribution of Y_1^*, Y_0^* requires usually too much information to be estimated from samples of reasonable size, the focus has been limited to the consideration of their conditional first moments $E[Y_1^*|X,Z]$, $E[Y_0^*|X,Z]$, that can also be decomposed into

$$\begin{aligned} E[Y_1^*|X,Z] &= E[Y_1^*|X,Z,d=1]\cdot pr(d=1|X,Z) + E[Y_1^*|X,Z,d=0]\cdot pr(d=0|X,Z) \\ E[Y_0^*|X,Z] &= E[Y_0^*|X,Z,d=1]\cdot pr(d=1|X,Z) + E[Y_0^*|X,Z,d=0]\cdot pr(d=0|X,Z), \end{aligned}$$

With only $E[Y_1^*|X,Z,d=1]$, $E[Y_0^*|X,Z,d=0]$ and the selection probabilities observable, but not $E[Y_1^*|X,Z,d=0]$ and $E[Y_0^*|X,Z,d=1]$, these moments are not identified. Inserting the selection correction term $\lambda_j(Z) = E[\varepsilon_j|Z,d=j]$, for $j=0,1$[2], introduced in the previous chapter and defining the expected structural outcome for the random variable as $\bar{f}_j(X) = E[f_j(X)]$, for $j=0,1$, further $\bar{X} = E[X]$, then some of these moments can be expressed as:

$$\begin{aligned} E[Y_1^*|X,Z] &= \bar{f}_1(X) \\ E[Y_1^*|X,Z,d=1] &= \bar{f}_1(X) + \lambda_1(Z) \\ E[Y_0^*|X,Z] &= \bar{f}_0(X) \\ E[Y_0^*|X,Z,d=0] &= \bar{f}_0(X) + \lambda_0(Z). \end{aligned} \tag{3.2}$$

Once more it becomes obvious, that standard OLS techniques are incapable to identify $\bar{f}_1(X)$, since $\bar{f}_1(X) + \lambda_1(Z)$ is observed only in conjunction, except when the omitted regressor $\lambda_1(Z)$ is orthogonal to $\bar{f}_1(X)$; $E[\,\bar{f}_1(X)\cdot\lambda_1(Z)] = 0$, which in practical matters can hardly ever be justified.[3]

[2] Notice that $\lambda(Z)$ is not a random variable.

[3] Henceforward the exposition will center entirely on the structural equation for category one ($d=1$) to avoid tedious repetitions for the second category, that should be treated analogously.

The central features of the identification issue can now easily be seen from the representation

$$E[Y_1^*|X, Z, d = 1] = \bar{f}_1(X) + \lambda_1(Z). \tag{3.3}$$

1. To identify $\bar{f}_1(X)$ up to an intercept, $\bar{f}_1(X)$ must be distinguishable from $\lambda_1(Z)$. This requires the absence of multicollinearity between $\bar{f}_1(\cdot)$ and $\lambda_1(\cdot)$.

2. Identification of an intercept of $\bar{f}_1(X)$, which is crucial for comparison with other categories, requires to tell apart the intercept of $\bar{f}_1(X)$ from an intercept of $\lambda_1(Z)$.

Two variants have been employed to solve the first issue by ensuring absence of collinearity. The parametric variant specifies the function $\bar{f}_1(X)$ and the bivariate error densities up to a finite number of unknown parameters, at once implying a specification for the function $\lambda_1(Z) = E[\varepsilon_1|Z, d = 1]$. For the common choice of $f_1(X) = X'\beta_1$ and $\lambda_1(Z)$ the inverse Mill's ratio (implied by a bivariate normality of the error terms), $\bar{f}_1(X)$ is linear and $\lambda_1(Z)$ non-linear, dismissing collinearity regardless of the overlap of X and Z, e.g. $X = Z$ is allowed. Here the non-linearity of $\lambda_1(Z)$ is completely sufficient to identify $\bar{f}_1(X)$.

The second variant is the semiparametric model with $\bar{f}_1(X)$ parametrically specified, but the error term distribution and $\lambda_1(Z)$ unspecified. Here identity of X and Z, that is $X = Z$, which would be the natural choice for many economical applications, cannot be permitted, since collinearity between $\bar{f}_1(X)$ and $\lambda_1(Z)$ may occur as the function $\lambda_1(\cdot)$ is completely unspecified. For instance in a semiparametric model with $f_1(X) = X'\beta$ linear, nothing prohibits the selection correction term $\lambda_1(\cdot)$ to be linear as well. In this case in equation (3.3) $\bar{X}'\beta$ and $\lambda_1(X)$ are impossible to tell apart. To be more precise Identification fails if the variables in Z are a subset of X, see Powell (1989,[41], p.6f). Then for a linear structural relationship $f_1(X) = X'\beta$ and a true selection correction term that follows the form $\lambda_1(Z) = \lambda_1(Z'\gamma)$, it could occur that a linear combination $X'\alpha$ exists with it's conditional expectation only depending on the argument $Z'\gamma$ of the correction term: $E[X'\alpha|Z] = a(Z'\gamma)$. If Z is a subset of X then conditioning on Z is equivalent to conditioning on the subset of X

$$E[X'\alpha|Z] = E[X'\alpha|subset(X)] = E[X'\alpha] = \bar{X}'\alpha.$$

Equation (3.3) could be written as

$$\begin{aligned} E[Y_1^*|X, Z, d = 1] &= \bar{X}'\beta + \lambda_1(Z'\gamma) = \bar{X}'\beta + E[X'\alpha|Z] - E[X'\alpha|Z] + \lambda_1(Z'\gamma) \\ &= \bar{X}'\beta + \bar{X}'\alpha + [-a(Z'\gamma) + \lambda_1(Z'\gamma)] \qquad \{\text{for } Z \text{ a subset of } X\} \\ &= \bar{X}'(\beta + \alpha) + \lambda^*(Z'\gamma). \end{aligned}$$

Since λ_1 is not specified it cannot be distinguished whether $\lambda_1(Z'\gamma)$ or $\lambda^*(Z'\gamma)$ is the true selection correction term and β is not identified.

Identification of a semiparametric model is then only conceivable if an exclusion restriction can be invoked. That is, at least one (continuous) variable exists, that influences the selection process significantly but has no effect on the potential outcomes Y_1^*, Y_0^*, with other words Z contains at least one variable that is not included in X. Now, the influence of a variation in the excluded variable(s) on $\lambda_1(Z)$ and the unaffectedness of $\bar{f}_1(X)$ nail down the selection correction term $\lambda_1(\cdot)$ and allow it to be told apart from the unchanged $\bar{f}_1(X)$. However, this requires sufficient oscillation in the excluded variable(s) to track the different reaction.

The frequent choice in economical applications is X as a subset of Z. Since it often is conjectured that the selection process $\varphi(\cdot)$ depends, besides some other variables Z', also on the potential outcomes Y_1^*, Y_0^*, hence $\varphi = \varphi(Y_1^*, Y_0^*, Z', u)$. Inserting $Y_1^* = f_1(X)$, $Y_0^* = f_0(X)$ reveals that all variables in X must also be included in Z to justify the reduced form $\varphi = \varphi(Z, u)$. Whether X now is a proper subset of Z or identical $X = Z$ depends on prior information about the selection mechanism and the structural relationships or behavioural theory. From economical theory it is seldom to justify, especially when individuals select themselves, that any variable in Z influencing the selection assignment should have no effect on the (potential) structural outcomes and can be dismissed from X. Identification in semiparametric models, however, requires an exclusion restriction to avoid collinearity, thus X should be a proper subset of Z for identification's sake. An example for an exclusion restriction that is often employed in the analysis of wage structure and labour market participation of women is the husband's income. It is conjectured that the husband's income influences the woman's decision whether to look for a payed job or stay at home, commit herself to a charity etc., though it is assumed that the woman's pay if she decided to work and found a job is completely independent of the husband's income. Quite often objections to particular exclusion assumptions can be raised, for instance repercussions if the woman's pay affects also the husband's income.

The second identification issue has been the intercept of $f_1(X)$. For the parametric model the intercepts of $f_1(X)$ and $\lambda_1(Z)$ are easily distinguished due to the implied specification of $\lambda_1(\cdot)$.

In the semiparametric case identification is less straightforward as $\lambda_1(Z)$ may contain an intercept as well. Here the selection procedure should be scrutinized to find certain values z in the support of Z for which the propensity score, that is synonymous to the selection probability to category 1, is one: $pr(d = 1|Z = z) = E[d|Z = z] = 1$. For observations with characteristic-values z the selection correction term is zero, $\lambda_1(Z = z) = 0$, since conditioning on (X, z) or on $(X, z, d = 1)$ makes no difference, as $Z = z$ implies $d = 1$. All individuals with characteristics z will select cogent and unambiguously into category 1 and the expected value of Y_1^* reduces to $E[Y_1^*|X, Z = z, d = 1] = \bar{f}_1(X) + \lambda_1(z) = \bar{f}_1(X)$, which entails the

isolated intercept of $f_1(X)$. For a single-index specification of the selection process, the index $\theta(Z)$ must go to infinity, if the support of the error term u is unbounded, for some $Z = z$ to attain selection probabilities of one, implying $\lambda_1(\theta(z)) = 0$. Hence a from above unbounded support of the index is required to identify the intercept and sufficient probability mass in the upper tail of the index distribution is needed for estimation. The dataset must provide sufficient variation in Z_i and a decent number of observations with very large index values tending to infinity. This strategy has been labelled "identification at infinity" by Chamberlain (1986,[6]). Other procedures have been suggested to identify the intercept like Gallant/Nychka (1987,[13]) or Chen (1996,[8]), though those require stricter assumptions.

Single Index Restriction [**SIR**]

As has been mentioned, when modelling selectivity the three approaches parametric, semiparametric and nonparametric are to distinguish. Parametric models specify all functions and error densities up to a finite amount of coefficients and density nuisance parameters, that can then be estimated by Maximum Likelihood or related methods, whereas nonparametric models leave both $f_1(X)$ and $\lambda_1(Z)$ completely undetermined. Semiparametric models in contrast specify only parts, usually the function $f_1(X)$ and the selection mechanism while leaving the error term distributions and hence $\lambda_1(Z)$ unrestricted. An important class within the semiparametric selectivity models, relaxing the strong independence of errors and explanatory variables assumption [IA], are the so-called index models, see Manski (1989,[31], p.356 and 1993,[32], p.81).

The single index assumption on the error term signifies that the error term density f_{ε_1}depends on the regressors X, Z only via an index function $\theta(Z)$, of lower dimension than Z, but is otherwise independent of the regressors X and Z: $f_{\varepsilon_1|X,Z,d=1} = f_{\varepsilon_1|\theta(Z),d=1}$. This also implies

$$\lambda_1(Z) = E[\varepsilon_1|X, Z, d = 1] = E[\varepsilon_1|\theta(Z), d = 1] = \lambda_1(\theta(Z)) \qquad \text{[SIR]}. \tag{3.4}$$

It is assumed that the impact of the selection variables Z of $\Re^q$ on the error term ε_1 can be truly represented by a function $\theta(Z) : \Re^q \to \Re^{q'}$ with $q' < q$. The prevailing choice is $q' = 1$. This requires that all the influence of Z on the error terms in the structural equations enters only through this index, hence the complete information about a proper subspace of $\Re^q$ carries all the relevant information entailed in Z of $\Re^q$. θ works like an aggregator, transferring all distinct selection variables to the same unit of measurement and permits comparison of their impacts. The index restriction on the errors [SIR] replaces the stronger assumption of independence of the errors and the regressors [IA] and with the *same* index restriction θ imposed on the structural error terms $\varepsilon_1, \varepsilon_0$ and on the selection error u the model in single

index form looks:

$$\begin{aligned} Y_1^* &= f_1(X) + \lambda_1(\theta(Z)) + \xi_1 \\ Y_0^* &= f_0(X) + \lambda_0(\theta(Z)) + \xi_0 \\ d &= \varphi(\theta(Z), u). \end{aligned} \tag{3.5}$$

The motivation for the single index assumption becomes apparent from Manski (1993,[32], p.81). Consider the pairs of realizations $(X = x_1, Z = z_1)$ and $(X = x_2, Z = z_2)$ with x_1, z_1, x_2, z_2 so far arbitrary numbers. Then the expected outcomes of Y_1^* are

$$\begin{aligned} E[Y_1^*|x_1, z_1] &= f_1(x_1) \\ E[Y_1^*|x_1, z_1, d = 1] &= f_1(x_1) + \lambda_1(\theta(z_1)) \\ E[Y_1^*|x_2, z_2] &= f_1(x_2) \\ E[Y_1^*|x_2, z_2, d = 1] &= f_1(x_2) + \lambda_1(\theta(z_2)) \end{aligned} \tag{3.6}$$

Now assume that $(x_1, z_1), (x_2, z_2)$ are chosen such that for both pairs $d = 1$ and $\theta(z_1) = \theta(z_2)$, but $x_1 \neq x_2$, implying the identity of both selection correction terms $\lambda_1(\theta(z_1)) = \lambda_1(\theta(z_2))$. Subtracting both observable expectations from each other yields:

$$\begin{aligned} E[Y_1^*|x_1, z_1, d = 1] - E[Y_1^*|x_2, z_2, d = 1] &= f_1(x_1) - f_1(x_2) + \{\lambda_1(\theta(z_1)) - \lambda_1(\theta(z_2))\} \\ &= f_1(x_1) - f_1(x_2) = E[Y_1^*|x_1, z_1] - E[Y_1^*|x_2, z_2]. \end{aligned} \tag{3.7}$$

Thus, by taking differences of distinct observations but with the same index value $\theta(z_1) = \theta(z_2)$, the selection bias vanishes and the structural function remains, which is identified by $E[Y_1^*|X, Z, d = 1]$ up to the intercept, that gets lost through differencing. The value of this finding depends on the size of the set $\{(x_1, z_1), (x_2, z_2) | \theta(z_1) = \theta(z_2), x_1 \neq x_2\}$, corresponding to the likelihood to obtain alike index values for distinct x_1, x_2, which itself depends on the dimension of the index function $\theta(\cdot)$. For the (usually) continuous domain of $\theta(\cdot)$ the probability of finding distinct realizations z_1, z_2 of Z with $\theta(z_1) = \theta(z_2)$ is zero. However, given the selection correction function $\lambda_1(\theta(Z))$ is well-behaved and smooth, close indices $\theta(z_1) \approx \theta(z_2)$ correspond to close selection correction terms $\lambda_1(\theta(z_1)) \approx \lambda_1(\theta(z_2))$, introducing only an arbitrarily small neglected error $[\lambda_1(\theta(z_1)) - \lambda_1(\theta(z_2))]$ into above conception, controlled by the closeness definition of the indices. The smaller the space into which $\theta(Z)$ maps the vector Z from $\Re^q$, the denser the observations get in comparison to the vector X of $\Re^p$-space and the more possibilities arise to find pairs of realizations of (X, Z) with distinct x but the same (or very close) index $\theta(z)$. This holds also for the common specification with X being a subvector of Z, as long as the domain of θ has a lower dimension than X; $\dim(\theta) < p$. The prevailing index choice is $\theta : \Re^q \to \Re^1$, mapping onto the real line to make the observations sufficiently dense.

This one-dimensional index is in semiparametric settings habitually parametrically specified as $\theta(Z) = Z'\gamma$. Though this is not essential, since a nonparametric index like the propensity score $E[d|Z]$, e.g. in Ahn/Powell, Robinson suffices. At the one extreme, $\theta : \Re^q \to \Re^0$ could be chosen, assigning every observation the same $\theta(z)$. This assumption would generate abundant observations with identical $\theta(z)$ and $\lambda_1(\theta(z))$ for identification, but seems implausible since the implication of a constant λ regardless of (X, Z) is inadequate for a selectivity model. In the other extreme, the no-index assumption $\theta : \Re^q \to \Re^q$, $\theta(Z) = Z$, together with X being a subvector of Z helps nothing to identification of $f_1(\cdot)$, as equality of the indices $\theta(z_1) = \theta(z_2)$ implies at once $z_1 = z_2 \Leftrightarrow x_1 = x_2$.

Although a single index assumption [SIR] is not essential for identification, it improves the accuracy of estimation considerably. These index models further allow to account for non iid error terms, as long as heteroskedasticity of unknown form depends only on the single index. This relaxes the strong homoskedasticity assumption [IA] of independence of the error term distribution $f_{\varepsilon_1,\varepsilon_0,u}(\cdot)$ of the exogenous variables $X, Z$by permitting *conditional* heteroskedasticity. [4]

Besides these prerequisites for identification, all estimators impose additional estimator-specific identifying conditions to attain certain estimators' properties. Common to all semiparametric and nonparametric estimators is their need for ample smoothness of the functions $f_1(\cdot)$, $\lambda_1(\cdot)$, the index $\theta(Z)$, the propensity score $E[d|Z]$ and of the densities of the regressors and the error terms. Smoothness, including continuous differentiability up to a certain order, is an essential ingredient of nonparametric algorithms since these proceed by extrapolating from observed data points to their local neighbourhood. Nonparametric regression methods are used to estimate the unspecified parts of the semiparametric model, like $\lambda_1(Z)$. They compute estimates by averaging over those observed dependent variables whose explanatory variables lie within a neighbourhood of the point of interest, hereby implicitly assuming that an observed datapoint entails information about it's adjacent neighbourhood and that the dependent variables of observations, whose explanatory variables are close, have similar values. For a discontinuous function, however,

[4] For X being a subset of Z, the imposition of an exclusion restriction can also be interpreted as a particular restriction on $f_1(\cdot)$, see Powell (1994,[42], p.2482). Powell decomposes the variables of Z into the subvectors X_1 and X_2, with X_1 containing the common variables with X and X_2 capturing the excluded variables and he defines $\theta_2 : \Re^q \to \Re^p$, $\theta_2(Z) = X_1$ as an index representation of the exclusion assumption:

$$Y_1^* = f_1(\theta_2(Z)) + \lambda_1(Z) + \xi_1.$$

Nevertheless, exclusion restriction and the single-index restriction [SIR] should not be confused, since the exclusion restriction is an assumption on the structural relationship $f_1(\cdot)$, essential for semiparametric identification, while the index restriction is an assumption on the error term $E[\varepsilon_1|X, Z, d = 1] = E[\varepsilon_1|\theta(Z), d = 1]$, which is a relaxation of the [IA] independence assumption $E[\varepsilon_1|X, Z, d = 1] = E[\varepsilon_1|d = 1]$.

this assumption does no longer hold, as observations that are very close in their explanatory variables may have very distant dependent outcomes. Besides these smoothness conditions, the existence of a sufficient number of finite moments of the random variables X, Z and the error distributions of ε, u is often imposed.

Concluding this chapter with a brief summary of the conditions for semiparametric identification, two basic ingredients stand out. One is the exclusion restriction on the structural relationship f_1 to prevent multicollinearity, the other is an assumption on the error terms preventing heteroskedasticity either by the [IA] independence assumption of errors and regressors $E[\varepsilon_1|X, Z, d = 1] = E[\varepsilon_1|d = 1]$ or by the weaker [SIR] single-index assumption $E[\varepsilon_1|X, Z, d = 1] = E[\varepsilon_1|\theta(Z), d = 1]$, allowing heteroskedasticity conditional on the index, see Powell (1994,[42], p.2507). Then the typical selectivity model looks:

$$\begin{aligned} Y_1^* &= f_1(\theta_2(Z)) + \lambda_1(\theta(Z)) + \xi_1 \\ Y_0^* &= f_0(\theta_2(Z)) + \lambda_0(\theta(Z)) + \xi_0 \\ d &= \varphi(\theta(Z), u) \end{aligned} \tag{3.8}$$

with $\theta_2(Z) = X$ the exclusion condition and $\theta(Z) : \Re^q \to \Re$ the single index.

Chapter 4

Parametric Estimation

To fully understand the motivation, advantages and drawbacks of semiparametric estimation the parametric approach to investigating selectivity must be presented in advance as a benchmark case, that has dominated the modelling of selectivity until the early 1980s. Returning back to the sample data notation for the base-line dichotomous censored selectivity model, the structural relationship $f(X_i)$ shall be estimated from a given sample (Y_i, d_i, X_i, Z_i) of size N, where the data has been sorted such that the first $1, .., n$ individuals are participants (category $d_i = 1$, Y_i observed), while the remaining $(n+1), .., N$ observations are censored ($d_i = 0$, Y_i unobserved, but X_i, Z_i still observed). The parametric way proceeds by specifying all functions and error distributions to finite dimensional families, such that only a finite amount of coefficients and density nuisance parameters remains to be estimated. A convenient choice is $f(\cdot)$ and $\theta(\cdot)$ linear, $\varphi(\cdot)$ the binary threshold crossing indicator and a bivariate normal distribution for the error terms, implying independence of the errors and regressors [IA] :[1]

$$\begin{aligned} Y_i &= X_i'\beta + \varepsilon_i \\ d_i &= 1(Z_i'\gamma + u_i > 0) \\ (\varepsilon_i, u_i) &\sim N\left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & 1 \end{pmatrix}\right), \end{aligned} \tag{4.1}$$

with Y_i only observed for $d_i = 1$. Here σ_u^2 is set to one as it is not identified in the binary response model. The model can easily be rewritten to

$$\begin{aligned} Y_i &= X_i'\beta + \sigma_{\varepsilon u}\lambda(Z_i'\gamma) + \xi_i \\ d_i &= 1(Z_i'\gamma + u_i > 0), \end{aligned} \tag{4.2}$$

where $\lambda(\cdot)$ is the inverse Mill's ratio, implied by the bivariate normality of (ε_i, u_i):

$$\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}. \tag{4.3}$$

[1] For ease of exposition Y_{1i}^* is replaced by Y_i henceforth.

$\phi(\cdot)$, $\Phi(\cdot)$ are the univariate density and distribution function, respectively, of the standard normal distribution $N(0,1)$. The model is fully identified through the *non-linearity* of $\lambda(\cdot)$ and does not require any exclusion restriction to prevent multicollinearity.

Heckman (1974,[17] and 1976,[18]) suggested two popular estimators, an efficient Maximum Likelihood based estimator and the inefficient but simple Heckman two-step, which have been widely applied in empirical work. The ML estimator can readily be applied by maximizing the log likelihood function: $\ln L(\beta, \gamma, \sigma_{\varepsilon u}, \sigma_\varepsilon^2) =$

$$\frac{1}{N}\sum_{i=1}^{N}\left[d_i \cdot \ln \int_{-Z_i'\gamma}^{\infty} \phi_{\varepsilon u}(Y_i - X_i'\beta, u)du \quad + \quad (1-d_i)\cdot \ln \int_{-\infty}^{\infty}\int_{-Z_i'\gamma}^{\infty} \phi_{\varepsilon u}(\varepsilon, u)dud\varepsilon\right]$$

with $\phi_{\varepsilon u}$ the density function of the bivariate normal, see Vella (1998,[50], p.130). The ML approach reaches the Cramer Rao bound, producing efficient estimates $\hat{\beta}, \hat{\gamma}, \hat{\sigma}_{\varepsilon u}, \hat{\sigma}_\varepsilon^2$.

Besides ML Heckman proposed a simple two step method where in the first step the binary selection equation is estimated by Probit over the full sample $1..N$ to obtain estimates of $\hat{\gamma}$. Then $\hat{\lambda}_i = \frac{\phi(Z_i'\hat{\gamma})}{\Phi(Z_i'\hat{\gamma})}$ can be calculated for each observation with $d_i = 1$ (subsample $1..n$) and inserted into the structural equation for λ as the additional regressor:

$$Y_i = X_i'\beta + \sigma_{\varepsilon u}\hat{\lambda}_i + \xi_i'. \tag{4.4}$$

Now $\hat{\beta}$ and $\hat{\sigma}_{\varepsilon u}$ can be estimated by OLS, with a significant coefficient estimate $\hat{\sigma}_{\varepsilon u}$ indicating that selectivity was inherent.

The parametric estimators have the appealing property, besides being ready and quick to implement, that they are more efficient than their semiparametric counterparts, producing more accurate estimates as long as the model is correctly specified. On the other hand parametric estimates have been shown to be possibly seriously inconsistent if the parametric form assumptions either on the functions $f(\cdot)$ and $\theta(\cdot)$ or on the joint error distribution are incorrect. Although specification tests give some hint about the credibility of the chosen specification, their power may be limited and the search for a proper specification for the functions and the errors until the model is correctly specified may be tedious. Several studies testified the sensitivity of the parametric techniques to heteroskedasticity and non-normality of the joint error distribution, e.g. Arabmazar and Schmidt (1982,[4], Goldberger (1983,[15]) and Schafgans (1997,[46]).

Two further studies (Leung and Yu, 1996,[28] and Nawata 1993,[36]) caution against the use of parametric estimation, especially the Heckman two-step, too precipitated from quite another point of view. Both studies reveal substantial weaknesses of the Heckman two-step estimator that has led to poor performance in some Monte Carlo studies. They attribute these weak results to a disadvantageous peculiarity of the inverse Mill's ratio. Although the inverse Mill's ratio is in theory a

non-linear function, it can be approximated extremely well by a linear function over a wide range of it's support, which is exposed by two plots (Leung and Yu, p.16 and Nawata, p.212). These both graphs exhibit the almost linearity of $\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ over the main body of it's support. A closer observation of the correlation between $\lambda(Z_i'\gamma)$ and $Z_i'\gamma$ revealed high collinearity with a correlation coefficient very close to one for many different index ranges. For the usual parametric specification $X_i = Z_i$ without exclusion restrictions, the required non-linearity of $\lambda(Z_i'\gamma)$ constitutes to the only identifying element for the structural equation. With these findings it is unsurprising that identification for the Heckman two-step may be very weak due to near collinearity, if the index values $Z_i'\gamma$ vary only over a small range within the sample, leading to very imprecise coefficient estimates with large variances and low significance levels. Furthermore Leung and Yu (p.213) conjecture on basis of their results that also the ML estimator seems to be "impaired by collinearity" through a near-linear selection correction term λ, though less severe than for the Heckman two-step.

If insufficient variation of the index $Z_i'\gamma$ within the dataset makes the Heckman two-step estimates imprecise and unreliable, various solutions are at hand. First, the Maximum Likelihood approach seems to be much less affected by multicollinearity. An other way, justifiably excluding one (or more) variables from X_i helps to separate $X_i'\beta$ from $Z_i'\gamma$. A final alternative would be to reject the selectivity model and proceed, for instance, with the "two-part" model described in Leung and Yu, where instead of the unconditional structural relationship $Y_i = X_i'\beta + \varepsilon_i$ with $E[\varepsilon_i|X_i] = 0$ the conditional equation $Y_{i|d_i=1} = X_i'\beta_c + \eta_i$ with $E[\eta_i|X_i, d_i = 1] = 0$ is estimated. The two part model can readily be estimated by standard techniques (Probit, OLS), since the selection equation and the conditional outcome equation are disconnected by definition and can be estimated completely separately, in contrast to the selectivity model where selectivity links both parts. The conditional outcome function, embodied in the coefficients β_c, however has a different interpretation than β, since it explains a relationship for only a subpopulation, the group of participants and does not represent potential outcomes for the whole population. Hence the two-part model is less informative and incapable to predict the full range of potential outcomes, but it is estimable without difficulties.

These both studies reveal that one of the advantages of the parametric approach, the non-reliance on exclusion restrictions, that are by economic theory often hard to justify, may be of fragile value.

Chapter 5

Semiparametric Estimation

Recalling the censored dichotomous selectivity model in the selection correction representation with single index

$$\begin{aligned} d_i &= \varphi(\theta(Z_i), u_i) \longrightarrow \text{mapping into } \{0,1\} \\ Y_i &= f(X_i) + \lambda(\theta(Z_i)) + \xi_i, \end{aligned} \tag{5.1}$$

with the potential outcome Y_i only observed for individuals with $d = 1$. The functions of interest f, λ, φ are unknown, with $f(\cdot)$ representing the structural relationship, $\lambda(\cdot)$ the selection correction term, $\varphi(\cdot)$ mapping the selection parameters into the discrete space $\{0,1\}$ and $\theta(\cdot)$ symbolizing a specified index function to reduce the dimensionality of the selection parameters.

The fully parametric approach, yet presented in the preceding chapter, allows to identify the relationship between the explanatory variables X_i and the dependent variable Y_i through fully specifying both the functions f, λ, φ and the distributions of ξ_i and u_i to finite dimensional families, such that the beforehand unknown functions and distributions are determined up to a finite amount of coefficients and density nuisance parameters, which are the object of estimation. While these parametric models have fine properties when mirroring exactly the true model such as $\sqrt{n}$-consistency[1] and efficiency, they incorporate serious drawbacks when misspecified. Their disadvantages range from inconsistency of the coefficient estimates, when misspecified either in the functional forms or in the unobservable error distributions, to the weak identifying power of the inverse Mill's ratio due to it's near linearity.

This induced the emergence of nonparametric methods which pursue the estimation of the statistical relationship between the explanatory variables and the dependent variable(s) under only very general smoothness and moment assumptions, without specifying any density function or any function of the model. Frankly speaking, they permit to attain consistency in more comprehensive models, while

[1] Meaning a convergence of the coefficient estimates towards the true values at a rate of $N^{-\frac{1}{2}}$, with N being the relevant sample size.

sacrificing efficiency in comparison to the parametric estimation. Although developed to counter the parametric procedure's flaws these strategies have their weaknesses in their own. Nonparametric estimates are commonly less precise, embody a convergence rate slower than $N^{-\frac{1}{2}}$, represent merely a statistical relationship that may be hard to interpret when it's shape does not conform to economic theory and preclude prediction outside the support of the explanatory variables in the underlying dataset. Consult Powell (1994,[42]) for an extensive introduction in semiparametric estimation, headed by a brief comparison of parametric, semiparametric and nonparametric approaches.

As a mixture between these both flanks the semiparametric approach seeks to combine some advantages of both the fully parametric and the completely nonparametric ones, by parametrically specifying only parts of the model, while applying nonparametric estimation issues to the remainder parts. In the selection correction form of the model (equations 5.1) the semiparametric approach regularly proceeds by specifying the functions $f(\cdot), \varphi(\cdot)$ and $\theta(\cdot)$, while leaving the selection correction function $\lambda(\cdot)$ and the errors' densities generic. It has inherited some rewards as well as flaws of both flanks. It allows consistent estimates under weaker restrictions, reducing the room for misspecification. Although it's parametric parts are still sensitive to erroneous specification, it's pre-fixed shape is usually easier to interpret. In addition it's estimators usually achieve a convergence rate of $N^{-\frac{1}{2}}$, though they are not as efficient as the estimates of a truly specified parametric model.

In contrast to the parametric approach, the two step estimation strategies are clearly prevailing within the semiparametric context.

For the censored selection model these techniques proceed by initially specifying the parametric parts of the model and by ordering the whole sample of N observations such that the first $1, .., n$ observations represent participants with $d_i = 1$ and Y_i observed, whereas the remaining $(n+1), .., N$ observations are the non-participants with $d_i = 0$ and Y_i unobserved.

- Then, by using a step one estimator over the full sample of $1..N$ observations, the coefficients $\hat{\gamma}$ of the selection equation are computed.
- Once estimates of γ have been extracted, execution of a step two estimator over the participation subsample of the $1..n$ observations supplies the $\hat{\beta}$-estimates of the structural equation.
- As will be seen below many step two estimators are able to identify only the slope coefficients of the structural equation. Therefore an additional intercept estimator has to be applied to the participation subsample $(1..n)$ if an estimate $\hat{\beta}_0$ of the structural equation's intercept is requested (step 2b).

As the step one and step two estimators are in most cases separate building-blocks, they may be combined in many ways. However, as the properties of the desired $\hat{\beta}$-estimates of the structural equation depend on the attributes of both estimators and their conjunction, a sensitive combination should be chosen.

Now some of the more promising estimators of both the selection and the structural equations are illustrated in detail. Although a plenty of the research literature will be surfaced, the emphasize has been laid on a thorough exhibition of just the principal estimators of selectivity models, rather than on an exposure of every estimation method presented in the recent years. For an overview article in that sense the survey of Vella (1998,[50]) be once more recommended.

For the ease of exposition all estimation techniques will be elucidated within the same model, the selection correction form with an index function parameterized by a linear index $\theta(\cdot)$, the indicator function as the selection mapping function $\varphi(\cdot) = 1(\cdot > 0)$ and a parametric linear structural function $f(\cdot)$. With $\lambda(\cdot)$ and the distributions of ξ_i and u_i left unspecified, the model looks:

$$\begin{array}{ll} d_i = 1(Z_i'\gamma > u_i) & \text{for } i = 1..N \\ Y_i = X_i'\beta + \lambda(Z_i'\gamma) + \xi_i & \text{for } i = 1..n \end{array} \tag{5.2}$$

The (uncensored) selection model with two observed categories would be treated similarly, with the sample ordered into the corresponding subsamples of individuals with $d_i = 1$ and $d_i = 0$:

$$\begin{array}{ll} d_i = 1(Z_i'\gamma > u_i) & \text{generating the two subsamples} \\ Y_{i1} = X_i'\beta_1 + \lambda_1(Z_i'\gamma) + \xi_{i1} & i = 1, .., n \\ Y_{i0} = X_i'\beta_0 + \lambda_0(Z_i'\gamma) + \xi_{i0} & i = (n+1), .., N. \end{array} \tag{5.3}$$

The step one estimator computes then in like manner the $\hat{\gamma}$-coefficients of the choice equation over the whole sample $1..N$, whereas the succeeding step two estimator estimates separately both structural equations over the disjoint subsamples $1, .., n$ and $(n+1), .., N$, respectively, to obtain $\hat{\beta}_1$ and $\hat{\beta}_0$. Efficiency gains through simultaneous estimation of both outcome equations Y_{i1} and Y_{i0} might be attainable, but are unlikely to be substantial because of the already segregated estimation procedure.

As mentioned in the chapter Identification, most estimators rely on the single index assumption, that makes the observations, loosely speaking, more dense, by mapping the relevant choice characteristics Z_i from $\Re^q$-space onto the real line, $\theta(\cdot) : \Re^q \longrightarrow \Re$. This contraction makes the selection regressors much more comparable and reduces the dimensionality considerably when proceeding further from the selection equation to the structural equation. Within this setting the linear single index function $\theta(Z_i, \gamma) = Z_i'\gamma$ has been specified to conform with most of

the literature. Be aware that most semiparametric estimators do not generally require linearity of the index $\theta(\cdot)$ and $f(\cdot)$, though some of them have to be modified somewhat when employing in nonlinear specifications.

Before reading ahead the reader unfamiliar with nonparametric Kernel regression techniques is advised to work through the appendix Introduction in Kernel Regression at first.

5.1 Semi-nonparametric approach: Gallant/Nychka

Before coming to the prevailing two step estimators, the interesting though rarely applied one step approach of Gallant and Nychka (1987,[13]), by many authors referred to as semi-nonparametric, should be briefly contrasted. Building on the ideas of Lee (1982,[26]) Gallant and Nychka approximate the joint density of (ε_i, u_i) by a polynomial of growing order. On basis of the [IA] independence of errors and regressors assumption[2] the bivariate error term density $f_{\varepsilon,u}$ is approximated by

$$f^*_{\varepsilon,u} = \left(\sum_{i,j=0}^{K} \alpha_{ij} \cdot \varepsilon^i u^j \right) \phi_\varepsilon \phi_u \tag{5.4}$$

with ϕ_ε, ϕ_u univariate normal densities and the unknown coefficients α_{ij}. The coefficients α_{ij} must ensure that ${f^*_{\varepsilon,u}}^3$ is a proper density function, integrating up to one and nonnegative. This flexible approximation comes arbitrarily close to many non-normal error term distributions for a large enough number of approximating terms $(K+1)^2$, as long as it's tails are not thicker than those of the t-distribution. Distributions with fatter tails cannot be handled by this series, see Gerfin (1996,[14], p.323). With this specification and for a selected K the selectivity model can be estimated by Maximum Likelihood like a parametric model. What distinguishes Gallant/Nychka as semiparametric is that K grows to infinity as the sample size increases, approaching an infinite number of unknown density parameters. To achieve consistency K must grow at a fast enough rate. Through this specification the estimator is able to estimate $(\hat{\gamma}, \hat{\beta}, \hat{\beta}_0)$ at once, though it has the major drawback that it's asymptotic distribution has not been derived so far. For a finite K beforehand pre-selected by the econometrician this approach boils down to a fully parametric model, $\sqrt{n}$-consistent and asymptotically normal if mirroring the true error distribution, though generally asymptotically biased. Yet for K unrestricted, growing

[2] The [IA] assumption can be relaxed to include heteroskedasticity as Melenberg and van Soest (1993,[33], p.15f) demonstrate.

[3] Although appearing so, this series does not nest the parametric approach with bivariate normality for $\sigma_{\varepsilon u} \neq 0$ since the product of two univariate normal densities $\phi_\varepsilon \phi_u$ is the root of the approximation, which is easier to handle for the computation of the likelihood than the bivariate normal density $\phi_{\varepsilon u}$, see Melenberg and van Soest (1993,[33], p.13).

with the sample size and to be determined by the estimator itself the asymptotic distribution is unknown. In practical work Gallant/Nychka's idea is usually applied in form of several nested parametric specifications with different K out of which the most appealing is finally selected, see Melenberg and van Soest (1993,[33]) or Gabler, Laisney and Lechner (1993,[12]).

5.2 Step 1-Estimation of Selection Equation

The first step to estimating the selectivity model is to analyze the selection assignment mechanism, through which non-random samples emerge and introduce the selectivity bias into the structural equations. This connection between some explanatory variables and the discrete selection assignment to one of a finite number of categories, here two, is embraced by the binary selection equation over the full sample $1..N$

$$d_i = 1\left(Z_i'\gamma > u_i\right) \rightarrow \{0,1\}, \tag{5.5}$$

and it is well explored in the discrete choice literature with a variety of distinct techniques suggested to estimate the choice coefficients γ [4] and the propensity score $pr(d_i = 1|Z_i) = E[d_i|Z_i]$. They differ by the assumptions on the error term u_i and it's conditional distribution function $F_{u|Z}$.

In its most general form the propensity score could be estimated nonparametrically by a multivariate Kernel as

$$\hat{E}[d_i|Z_i] = \frac{\frac{1}{N}\sum_{j=1}^{N} d_j \cdot \frac{1}{h} \cdot K\left(\frac{Z_i - Z_j}{h}\right)}{\frac{1}{N}\sum_{j=1}^{N} \frac{1}{h} \cdot K\left(\frac{Z_i - Z_j}{h}\right)} \tag{5.6}$$

under only very weak smoothness assumptions on $F_{u|Z}$. Yet as expected this estimator's generality does not come along without shortcomings. The nonparametric estimates are susceptible to the curse of dimensionality since the selection equation is likely to contain many regressors Z_i, leading to high inaccuracy that is particularly serious inasmuch as the succeeding estimation steps ground on these outcomes. Besides the low precision, the nonparametric approach is only able to identify $E[d_i|Z_i]$ but not γ and $F_{u|Z}$, which contain useful information about the selection process, see Gerfin (1996,[14], p.322). The latter disadvantage, however, may be acceptable if not the selection process by itself is of interest but only it's impact on the structural equation. Then the propensity score is sufficient for the identification of the structural relationships, as the estimators Ahn/Powell and Robinson assert on basis of nonparametric estimates of the selection equation.

[4] Recall that γ is identified only up to scale since only the sign of $Z_i'\gamma - u_i$ is observed, such that an intercept in γ cannot be distinguished from an unknown threshold τ, which consequently is set to zero.

In the other extreme, $F_{u|Z}$ can be parametrically specified up to a finite amount of parameters. Then the selection process boils down, for instance, to the Probit or Logit model depending on the actual specification. This parametric assumptions on $F_{u|Z}$ are to a large extent arbitrary and only limitedly verifiable and make the estimates prone to inconsistency if inadequately imposed.

Still specifying $F_{u|Z}$ but by a flexible series approximation with the number of density parameters approaching infinity is the subject of Gallant/Nychka, here exemplary applied on the binary choice model. For a fixed number of approximating terms this corresponds to a fully parametrically specified selection equation.

A many of alternative estimators have been developed that leave the conditional distribution $F_{u|Z}$ unspecified and identify the selection equation via some additional conditions on u_i. With respect to these conditions Horowitz (1993,[21], p.50) distinguishes two classes. One class of estimators has as common feature the single index assumption on u_i [SIR], signifying that the distribution $F_{u|Z}$ depends only on an (one-dimensional) index, usually $Z_i'\gamma$, thus $F_{u|Z} = F_{u|Z'\gamma}$. This dimension reduction avoids the curse of dimensionality and the propensity score can be rewritten as $E[d_i|Z_i] = E[d_i|Z_i'\gamma] = pr(d_i = 1|Z_i'\gamma) = F_{u|Z'\gamma}(Z_i'\gamma)$. Herein belong the Klein/Spady (1993,[24]), Ichimura (1993,[23]), Powell/Stock/Stoker (1989,[43]) and some more. Many achieve a convergence rate of $N^{-\frac{1}{2}}$, but their single-index restriction forbids heteroskedasticity of unknown form except if it depends entirely on the index $Z_i'\gamma$.

The second class, allowing any form of heteroskedasticity, hence labelled "arbitrarily heteroskedastic models" by Horowitz, contains estimators that impose merely the weak zero-median assumption on the error distribution $median(u|Z) = 0$, including the M-Score of Manski (1975,[29] and 1985,[30]) and the smoothed M-Score of Horowitz (1992,[20]). Although this class appears interesting due to its tolerance against heteroskedasticity, it will be disregarded henceforth since it's estimators fall short of $\sqrt{n}$-convergence. This convergence rate is requested to avoid an efficiency loss in the subsequent estimation steps, because the properties of any step two estimator deteriorate with the weaknesses of the first step. Manski's M-Score converges at a rate of $N^{-\frac{1}{3}}$ while the estimator of Horowitz attains a rate somewhere between $N^{-\frac{2}{5}}$ and $N^{-\frac{1}{2}}$.

The article of Gerfin (1996,[14]) provides an exemplifying comparison of some corresponding estimators. Due to scarcity of space only the Klein/Spady estimator is presented here, which is semiparametrically efficient.

5.2.1 Klein/Spady 1993

Klein/Spady (1993,[24]) proposed an estimator of the selection equation $d_i = \varphi(\theta(Z_i), u_i)$ that is semiparametric by specifying the index function $\theta(\cdot)$ but leaving the distri-

bution of the error term u_i unrestricted. With $\varphi(\cdot)$ the usual threshold crossing indicator function and the unidentified threshold set to zero the selection equation looks

$$d_i = 1\left(\theta(Z_i, \gamma) > u_i\right). \tag{5.7}$$

Due to the single index restriction heteroskedasticity of unknown form can only be accounted for if it depends entirely on the assumed index $\theta(Z_i)$, for instance $\theta(Z_i) = Z_i'\gamma$. Klein and Spady's estimation procedure is motivated on the parametric ML approach where the coefficients γ are estimated by maximizing the likelihood

$$\begin{aligned} \max_{\gamma} L(\gamma) &= \sum_{i=1}^{N} d_i \cdot \ln pr(d_i = 1|\theta_i) + (1 - d_i) \cdot \ln(1 - pr(d_i = 1|\theta_i)) \qquad (5.8) \\ &= \sum_{i=1}^{N} d_i \cdot \ln pr(Z_i'\gamma > u_i|Z_i'\gamma) + (1 - d_i) \cdot \ln(1 - pr(Z_i'\gamma > u_i|Z_i'\gamma)). \end{aligned}$$

As in contrast to the fully parametric approach the density of the error term u_i and thus the probability function is unspecified, a quasi likelihood function with an estimated error term density is constructed, that is a smooth function of γ and approximates the parametric likelihood. Maximizing this quasi likelihood yields $\hat{\gamma}_{Klein/Spady}$ via[5]

$$\max_{\gamma} quasiL(\gamma) = \sum_{i=1}^{N} d_i \cdot \ln\left[\hat{pr}(d_i = 1|Z_i)^2\right] + (1 - d_i) \cdot \ln\left[(1 - \hat{pr}(d_i = 1|Z_i))^2\right]. \tag{5.9}$$

Notice that the probabilities in equation (5.9) are squared, since the probability estimates might be negative if a bias-reducing, higher order Kernel is employed for their estimation. The probability function $pr(d_i = 1|Z_i) = pr(d_i = 1|Z_i'\gamma) = pr(Z_i'\gamma > u_i|Z_i'\gamma)$ can be rewritten by Bayes theorem as

$$pr(Z_i'\gamma > u_i|Z_i'\gamma) = \frac{pr(Z_i'\gamma > u_i) \cdot g_{\theta|d=1}(Z_i'\gamma|Z_i'\gamma > u_i)}{g_\theta(Z_i'\gamma)}, \tag{5.10}$$

where g_θ is the density function of the index $Z'\gamma$ and $g_{\theta|d=1}$ the density of the index conditional on the selection status $d_i = 1$. This is equivalent to

$$pr(d_i = 1|Z_i) = \frac{pr(d_i = 1) \cdot g_{\theta|d=1}(Z_i'\gamma|d_i = 1)}{g_\theta(Z_i'\gamma)}. \tag{5.11}$$

With the densities g_θ and $g_{\theta|d=1}$ nonparametrically estimable by univariate Kernels and $pr(d_i = 1)$ replaced by it's sample average the quasi likelihood function can be computed for any coefficient vector γ. The vector γ that maximizes equation (5.9) is the Klein/Spady estimate $\hat{\gamma}_{Klein/Spady}$. To computing the probability density functions Klein and Spady insert standard nonparametric density estimation algorithms

[5] In the original article a trimming function had been included, though Monte Carlo results of Klein and Spady indicate that neglecting trimming has little effect.

to acquire

$$\hat{pr}(d_i = 1|Z_i) = \frac{\frac{1}{N-1}\sum_{j\neq i}^{N} d_j \cdot \frac{1}{h} \cdot K\left(\frac{Z_i'\gamma - Z_j'\gamma}{h}\right)}{\frac{1}{N-1}\sum_{j\neq i}^{N} \frac{1}{h} \cdot K\left(\frac{Z_i'\gamma - Z_j'\gamma}{h}\right)} \tag{5.12}$$

and recommend either a higher order Kernel or an adaptive local smoothing Kernel. (Consult Silverman (1986,[49]) for an intuitive introduction in density estimation). To achieve desired asymptotical properties the true probability distribution $pr(Z_i'\gamma > u_i|Z_i'\gamma)$ must be continuously differentiable with bounded derivatives. Furthermore small estimated densities that may perturb the outcomes should be adjusted or trimmed, though as admitted in their Monte Carlo study (p.406) the estimator performs pretty well without it. Then the estimator is consistent, asymptotically normal and semiparametrically efficient,

$$\sqrt{n}(\hat{\gamma}_{Klein/Spady} - \gamma) \xrightarrow{d} N(0, V_{Klein/Spady}). \tag{5.13}$$

If the estimated probability function is inserted into equation (5.9) the quasi likelihood behaves like a likelihood function. Hence the variance matrix can be estimated in like manner as in the parametric case, for instance, as the outer product of the gradient or by the Hessian matrix. Also likelihood ratio tests can be performed.

5.3 Step 2-Estimation of Structural Equation

The objective of this section will be to present a range of well-behaved semiparametric estimators of the coefficients β for a single structural equation:

$$Y = X'\beta + \lambda(Z'\gamma) + \xi \qquad \text{with } E[\xi|X, Z, d = 1] = 0, \tag{5.14}$$

for the triple (Y, X, Z) of random variables with support $(\Re \times \Re^p \times \Re^q)$. This corresponds to the estimation of a censored selection model with the sample data representation $Y_i = X_i'\beta + \lambda(Z_i'\gamma) + \xi_i$, semi-parameterized by a linear single index $(\theta(Z_i'\gamma) = Z_i'\gamma)$ and a linear functional relationship between X_i and Y_i, while $\lambda(\cdot)$ and ξ_i have been retained unspecified.

In the sections 5.3 and 5.4 some of these methods for estimating the structural equation with regard to selectivity are introduced, which part into two distinct divisions:

- The techniques in section 5.3, suggested by Powell, Newey, Ahn/Powell and Robinson, center on the estimation of the slope parameters of β, while they are unable to identify the intercept of β, denoted β_0 hereafter. This is due to the virtual indistinguishability of an intercept β_0 from an intercept of the unspecified function $\lambda(\cdot)$, easily be seen by the following transformation:

$$Y_i = \beta_0 + X_i'\beta + \lambda(Z_i'\gamma) + \xi_i = X_i'\beta + \tilde{\lambda}(Z_i'\gamma) + \xi_i, \tag{5.15}$$

with $\tilde{\lambda}(\cdot) = \lambda(\cdot) + \beta_0$. As either $\lambda(\cdot)$ and $\tilde{\lambda}(\cdot)$ are valid within this context, special techniques are required to estimate β_0. Henceforth, this section is devoted to the estimation of the slope parameters and β and $\hat{\beta}$ will be treated as not containing an intercept. *An application of these methods should be exercised excluding an intercept within* β. Only the estimator of Chen is due to an additional symmetry condition able to identify β and β_0 and should be performed including an intercept.

- The succeeding section 5.4 will then be devoted to the estimation of β_0 exclusively, with methods designed by Andrews and Schafgans and Heckman, which require preliminary estimators of $\hat{\beta}$ and $\hat{\gamma}$.

Due to this distinction the whole estimation process breaks into three steps if identification of both the slope coefficients β and the intercept β_0 is requested.

The bunch of estimators of the slope coefficients β can furthermore be broken down into two main branches with respect to their identification strategies. While one camp of estimators (Newey, Lee, Cosslett) seeks to identify β through replacing the selection bias $\lambda(\cdot)$ by an infinite series approximation, the second group (Powell, Robinson, Chen) relies on the idea of differencing out the selection bias $\lambda(\cdot)$ by various Kernel methods. In any case it should be kept in mind that all the ongoing calculations are performed on the subsample of participants ($d_i = 1$) with n observations.

5.3.1 Powell 1989

Powell (1989,[41]) motivated his estimator of the slope coefficients β on the idea, that under the single index assumption individuals with identical indices should have identical correction terms λ, as their λ are assumed to depend entirely on their indices $Z'\gamma$.

Ideally, given two individuals i and j with $Z_i'\gamma = Z_j'\gamma$, their expected outcomes Y_i and Y_j could be compared disregarding their selectivity bias, since their conditional outcome difference $Y_i - Y_j$

$$Y_i - Y_j = (X_i'\beta + \lambda(Z_i'\gamma) + \xi_i) - (X_j'\beta + \lambda(Z_j'\gamma) + \xi_j) \tag{5.16}$$

would simplify with $Z_i'\gamma = Z_j'\gamma \Rightarrow \lambda(Z_i'\gamma) = \lambda(Z_j'\gamma)$ to:[6]

$$\begin{array}{lll} Y_i - Y_j & = (X_i - X_j)'\beta + (\xi_i - \xi_j) & \Rightarrow \\ E[Y_i - Y_j | Z_i, Z_j] & = (X_i - X_j)'\beta. & \end{array} \tag{5.17}$$

[6] $E[\xi_i|Z_i, Z_j]$ and in analogy $E[\xi_j|Z_i, Z_j]$ are zero, since random sampling has been assumed signifying that the error term ξ_i is independent of the other individuals in the sample, hence $E[\xi_i|Z_i, Z_j] = E[\xi_i|Z_i]$ which is zero as only participants ($d_i = 1$) are contemplated.

Consequently, by differencing out the selectivity bias terms β is identified and can be estimated by instrumental variables[7].

Pursuing this inspiration, in an initial step the entire participation subsample would be subdivided into groups of individuals with identical indices. This is followed by the construction of an instrumental variable vector $W_i \in \Re^p$ for each individual i out of their $Z_i \in \Re^q$ variables, such that W_i is of same dimension as X_i. Finally for each group containing more than one entry instrumental variable regression of $(Y_i - Y_j)$ on $(X_i - X_j)$ with $(W_i - W_j)$ as the instrument would be performed for all pairs i,j within the same group, yielding consistent estimates of β as for all included pairs their respective selectivity biases $\lambda(Z_i'\gamma) = \lambda(Z_j'\gamma)$ cancel out. However, the depicted algorithm cannot be implemented directly since pairs of individuals with matching indices cannot be detected from the data. On the one hand, the index has been required as continuously distributed (see Identification assumptions) as a supposed demand for $\sqrt{n}$-consistent step one estimation of the selection equation. Accordingly the occurrence of pairs with equal indices is of probability zero. In addition, as γ is unknown and has to be approximated by $\hat{\gamma}$ from the sample even truly existing pairs i,j with $Z_i'\gamma = Z_j'\gamma$ would probably not be discovered, as well as in reverse manner the equality of any $Z_i'\hat{\gamma}$ and $Z_j'\hat{\gamma}$ does not surely imply the equality of $Z_i'\gamma = Z_j'\gamma$.

Nevertheless, under certain smoothness conditions on the function $\lambda(\cdot)$, one can conjecture that the term $(\lambda(Z_i'\gamma) - \lambda(Z_j'\gamma))$ should vanish as $Z_i'\hat{\gamma}$ and $Z_j'\hat{\gamma}$ come close, provided $\hat{\gamma}$ being a consistent estimate of γ. Hence observations j with estimated indices within a near neighbourhood of $Z_i'\hat{\gamma}$ should allow a guess of β as their bias differences $\lambda(Z_i'\gamma) - \lambda(Z_j'\gamma)$, being likely close to zero, disturb little.

As the concept of closeness of two estimated indices $Z_i'\hat{\gamma}$ and $Z_j'\hat{\gamma}$ relies on the supplied data, it seems appropriate employing the Kernel weighting method covering the whole sample of participants of size n. The computation proceeds by including all $\binom{n}{2}$ pairs of individuals and averaging β over those pairs whose indices are reasonably close, while ascribing zero or little weight to observations that depart considerably. The Kernel function, whose statistical properties and behaviour as a weighting device are well explored, assigns an averaging weight $\hat{\varpi}_{ij}$ to every pair i,j according to their indices' distance measured as $Z_i'\hat{\gamma} - Z_j'\hat{\gamma}$, that decreases symmetrically with higher deviation.

[7] IV estimation instead of OLS is neccessary since the regressor $(\xi_i - \xi_j)$ is unavailable.

These weights $\hat{\varpi}_{ij}$ calculated, $\hat{\beta}$ can be estimated by a weighted instrumental variable estimator making use of the earlier defined instruments W_i:

$$\begin{aligned}\hat{\beta}_{Powell} &= \left[\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\hat{\varpi}_{ij}\cdot(W_i-W_j)(X_i-X_j)'\right]^{-1} \\ &\cdot\left[\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\hat{\varpi}_{ij}\cdot(W_i-W_j)(Y_i-Y_j)\right].\end{aligned} \tag{5.18}$$

This expression, strikingly similar to the usual IV estimator $(W'X)^{-1}(W'Y)$, takes via the computed weights account of only those pairs that are not too distant apart and upweights closer observations with supposedly more similar selectivity bias terms, $\lambda(Z_i'\gamma)-\lambda(Z_j'\gamma)\approx 0$.

For the computation of the weights $\hat{\varpi}_{ij}$ a symmetric, twice differentiable Kernel of $4th$ order[8] and bounded support has been proposed, with a suitable choice of the bandwidth parameter h which converges to zero but not too rapid as the sample size n grows to infinity, that $nh^8\to 0$, whereas $nh^6\to\infty$. For practical purposes the bandwidth choice may be guided by generalized cross-validation, see Craven and Wahba (1979,[11]), with the selection of the Kernel function conforming with the mentioned properties. The weights $\hat{\varpi}_{ij}$ are calculated as

$$\hat{\varpi}_{ij}=\frac{1}{h}K\left(\frac{Z_i'\hat{\gamma}-Z_j'\hat{\gamma}}{h}\right). \tag{5.19}$$

Powell proves that the $\hat{\beta}_{Powell}$ estimator is $\sqrt{n}$-consistent and asymptotically normal under an appropriately chosen Kernel and some other subsequent conditions. These results mainly rest on the single index assumption, continuous distribution of the index $Z_i'\gamma$ with continuous but bounded density function, as well as an exclusion restriction and smoothness conditions on the function $\lambda(\cdot)$. Then, provided a $\sqrt{n}$-consistent step one $\hat{\gamma}$-estimate has been generated the estimate of β converges as:

$$\sqrt{n}(\hat{\beta}_{Powell}-\beta)\xrightarrow{d}N(0,V_{Powell}). \tag{5.20}$$

In addition a path to estimating the asymptotic covariance matrix is accommodated, which allows for asymptotic inference on basis of the normal distribution. However, as the formulae is a bit lengthy, the derivation is transferred to the appendix.

5.3.2 Newey Series 1988

Newey (1988,[37]) proposes in his article two single-index estimators of the slope coefficients β of the structural equation on the motivation of series approximation

[8] Recall that this bias-reducing Kernel implies negative Kernel weights for some arguments and that accordingly this Kernel function cannot be a density. (Consult the appendix Kernel Regression for further readings.)

to the selection correction term λ_i. The idea of replacing the unknown function $\lambda(\cdot)$ by a sum of known basic functions p_k with unknown coefficients η_k,

$$\lambda(\cdot) \approx \sum_{k=1}^{K} \eta_k \cdot p_k(\cdot), \tag{5.21}$$

with the number of terms K growing to infinity as the sample size increases, has beforehand been employed by Lee (1982,[26]) and Cosslett (1991,[10]). Lee furthermore influenced with his article many subsequent developments in the field of semiparametric estimation of selectivity models, like the approximation of the bivariate error term density later on pursued by Gallant and Nychka (1987,[13]). Cosslett approximated $\lambda(\cdot)$ with dummy variables (step function) by classifying all individuals into certain groups depending on into which interval their index falls, with the intervals and number of groups being a by-product of his step one estimator. Building on their work, Newey developed a similar series estimator $\hat{\beta}_{Series}$ and also an estimator $\tilde{\beta}_{GMM}$ based on the generalized method of moments, which is efficient in the semiparametric sense and will be presented later on. For the series estimator Newey worked out conditions on the basic functions to attain desired estimator's properties.

The series estimator $\hat{\beta}_{Series}$ aims to replace $\lambda(Z_i'\gamma)$ by a consistent approximation with suitable functions $\sum \eta_k \cdot p_k(\cdot)$, where η_k are the unknown coefficients and $p_k(\cdot)$ the known smooth basic functions depending only on the index. Given any consistent step one estimate of γ the index values $Z_i'\hat{\gamma}$ can be calculated and also the value of the series approximation $\sum \eta_k \cdot \hat{p}_{ki}$. Assuming that the infinite series $\sum_{k=1}^{\infty} \eta_k \cdot p_k(Z_i'\gamma)$ renders $\lambda(Z_i'\gamma)$ exactly the structural equation $Y_i = X_i'\beta + \lambda(Z_i'\gamma) + \xi_i$ can be rewritten as:

$$Y_i = X_i'\beta + \sum_{k=1}^{K} \eta_k \cdot p_k(Z_i'\gamma) + \xi_i' \qquad \text{with } \xi_i' = \left\{ \sum_{k=K+1}^{\infty} \eta_k \cdot p_k(Z_i'\gamma) + \xi_i \right\}. \tag{5.22}$$

Inserting the estimated values $\hat{p}_{ki}$ for $p_k(Z_i'\gamma)$, the unknown coefficients β, η can then be estimated by OLS.

It remains the choice of suitable basic functions $p_k(\cdot)$, like

$$p_k(Z_i'\gamma) = \left[\frac{\phi(Z_i'\gamma)}{\Phi(Z_i'\gamma)} \right] \cdot (Z_i'\gamma)^{k-1}, \tag{5.23}$$

which has been suggested by Lee (1982,[26]), or Newey's advice:

$$p_k(Z_i'\gamma) = [\tau(Z_i'\gamma)]^k \tag{5.24}$$

with $\tau(\cdot)$ a monotonic and bounded function within $[-1, 1]$, for instance $\tau(Z_i'\gamma) = 2\Phi(Z_i'\gamma) - 1$. Newey presumes that uniformly bounded functions, like the latter $2\Phi(\cdot) - 1$, are probably more robust to outliers than functions that grow unbounded,

as the one suggested by Lee, and he alerts of mixed series. Hence, he suggests the use of monotonic bounded functions, though the series of Lee has the nice property of nesting the parametric Heckman two-step. For consistency K must grow to infinity with increasing sample size. More important may be the right choice of the parameter K, the number of terms included, which may be guided by cross-validation to choose K to minimize $CV(K)$ (formulae see appendix). Newey proves his series estimator being $\sqrt{n}$-consistent and asymptotically normal,

$$\sqrt{n}(\hat{\beta}_{Series} - \beta) \xrightarrow{d} N(0, V_{Newey\_Series}) \tag{5.25}$$

under the following conditions. The computation of the estimated variance matrix $\hat{V}_{Series}$ will be provided in the appendix.

As usual this estimator, constructed on the single index assumption, requires exclusion restrictions, together with continuous distribution of the index and a many times continuously differentiable bivariate error term density. Furthermore some prerequisites on the chosen basic functions should be noticed. Setting forth with the monotonic functions $p_k(Z_i'\gamma) = [\tau(Z_i'\gamma)]^k$ bounded within $[-1, 1]$, the function $\tau(\cdot)$ must be twice continuously differentiable with bounded derivatives on the support of the index. In addition $\tau(\cdot)$ should be bounded as well. Besides some other conditions, these ensure a vanishing asymptotic bias and asymptotic normality.

5.3.3 Newey GMM 1988

A second semiparametric estimator has been proposed by Newey (1988,[37]) which bases on the generalized method of moments and makes a more thorough use of the available information implied by the single-index assumption [SIR]. As the estimator becomes quite technical in it's derivation, it's operationalization is banned to the appendix and only the underlying motivation is sketched here.

As Newey remarks, his single-index series estimator $\hat{\beta}_{Series}$ makes only use of the implication that the conditional expectation of ε_i is a function of the index $Z_i'\gamma$ only:

$$E[\varepsilon_i | X_i, Z_i, d_i = 1] = \lambda(Z_i'\gamma). \tag{5.26}$$

However the single index assumption entails more information, as it declares that the whole distribution of ε_i, not only it's first moment, depends only on the index, $f_{\varepsilon|Z} = f_{\varepsilon|Z'\gamma}$. This means that the conditional expectation of *any* (known) function m of ε_i is determined by the index $Z_i'\gamma$ alone:

$$E[m(\varepsilon_i) | X_i, Z_i, d_i = 1] = m'(Z_i'\gamma), \tag{5.27}$$

with $m'(Z_i'\gamma)$ the unknown conditional expectation function corresponding to $m(\varepsilon_i)$. Or, put differently, that the residual $m(\varepsilon_i) - m'(Z_i'\gamma)$ has conditional expectation

zero,

$$E[m(\varepsilon_i) - m'(Z_i'\gamma) \quad |X_i, Z_i, d_i = 1] = 0. \tag{5.28}$$

This equation (5.28) further implies that the residual $(m(\varepsilon_i) - m'(Z_i'\gamma))$ is orthogonal to Z_i and X_i and that *any* function a of Z_i is uncorrelated with this residual in the participation subsample[9]:

$$E[a(Z_i) \cdot \{m(\varepsilon_i) - m'(Z_i'\gamma)\}] = 0. \tag{5.29}$$

Inserting $\varepsilon_i = Y_i - X_i'\beta$, the equation $E[a(Z_i) \cdot \{m(Y_i - X_i'\beta) - m'(Z_i'\gamma)\}] = 0$ could be used to construct a method of moment estimator, were γ and the function $m'(\cdot)$ known. Replacing then γ and $m'(Z_i'\gamma)$ with consistent preliminary estimates $\hat{\gamma}$ and a series approximation $m'(Z_i'\hat{\gamma}) = \sum_{h=1}^{K} \eta_h \cdot p_h(Z_i'\hat{\gamma})$ with unknown coefficients $\eta_1, .., \eta_h$, respectively, to obtain:

$$E[a(Z_i) \cdot \{m(Y_i - X_i'\beta) - \sum_{h=1}^{K} \eta_h p_h(Z_i'\hat{\gamma})\}] = 0. \tag{5.30}$$

Now a general method of moment estimator can be established for chosen $\beta, \eta_1, .., \eta_h$ by it's sample moment

$$\hat{g}_1(\beta, \eta_1, .., \eta_h) = \frac{1}{n} \sum_{i=1}^{n} \left[a(Z_i) \cdot \{m(Y_i - X_i'\beta) - \sum_{h=1}^{K} \eta_h p_h(Z_i'\hat{\gamma})\} \right], \tag{5.31}$$

which should be zero if $\beta, \eta_1, .., \eta_h$ were the true coefficient values. $\hat{g}_1$ indicates departures from orthogonality for a particular choice of $\beta, \eta_1, .., \eta_h$. Hence the squared departures are an indicator of how good the chosen coefficients fit the single-index selectivity model. The β estimates are obtained by minimizing these squared departures over all admissible values of $\beta, \eta_1, .., \eta_h$:

$$\min_{\beta, \eta_1, .., \eta_h} \hat{g}_1^2. \tag{5.32}$$

This estimator becomes more efficient the more "supposedly uncorrelated" functions $a(Z_i)$ and the more moment functions $m(\varepsilon_i)$ are included, for which the coefficients β have to be adjusted such that all these functions are pairwise orthogonal. The single-value function $a(Z_i)$ is replaced by a column vector $A(Z_i)$ of functions

[9] In this respect the presentation of Newey (1988,[37]) and it's reproduction in Melenberg and van Soest (1993,[33], p.23ff) are inconsistent referring to the employed sample. Newey performs the estimator over the whole sample of size N, while Melenberg and van Soest apply it to the participation subsample. I follow Melenberg and van Soest's exposition by basing the ongoing calculations on the participation subsample of n observations.

of Z_i and dimension r and a series of moment functions $m_1(\varepsilon_i), .., m_J(\varepsilon_i)$ is introduced with their corresponding conditional expectations $m'_1(Z'_i\gamma), .., m'_J(Z'_i\gamma)$. Rowing the moment function residuals $m_j(\varepsilon_i) - m'_j(Z'_i\gamma), \quad j = 1..J$ into a column vector $R(\varepsilon_i, Z'_i\gamma)$ of size J, equation (5.29) can be extended to

$$E[R(\varepsilon_i, Z'_i\gamma) \otimes A(Z_i)] = \mathbf{0}_{(J \cdot r \times 1)}. \tag{5.33}$$

This equation signifies that for the right choice of the coefficients β all J moment residuals should be orthogonal to all elements of the vector $A(Z_i)$. Once more, the unknown coefficients γ and the functions $m'_j(\cdot)$ are to be replaced by $\hat{\gamma}$ and series approximations $m'_j(Z'_i\hat{\gamma}) = \sum_{h=1}^{K} \eta_{jh} \cdot p_h(Z'_i\hat{\gamma})$. Defining $\hat{p}_i$ as the column vector of the approximating functions $(p_1(Z'_i\hat{\gamma}), .., p_K(Z'_i\hat{\gamma}))'$ and η as the unknown $(J \times K)$ coefficient matrix $\eta = ((\eta_{11}, \eta_{12}, .., \eta_{1K})', (\eta_{21}, \eta_{22}, .., \eta_{2K})', .., (\eta_{J1}, \eta_{J2}, .., \eta_{JK})')'$ the moment residual vector can be approximated by

$$\hat{R}(\varepsilon_i, Z'_i\hat{\gamma}) = \begin{bmatrix} m_1(Y_i - X'_i\beta) \\ \vdots \\ m_J(Y_i - X'_i\beta) \end{bmatrix} - \eta \cdot \hat{p}_i. \tag{5.34}$$

Then the GMM estimator is founded on the equation

$$g(\beta, \eta) = E[\hat{R}(\varepsilon_i, Z'_i\hat{\gamma}) \otimes \hat{A}(Z_i)] = \mathbf{0}, \tag{5.35}$$

with it's sample moment

$$\hat{g}(\beta, \eta) = \frac{1}{n} \sum_{i=1}^{n} \hat{R}(\varepsilon_i, Z'_i\hat{\gamma}) \otimes \hat{A}(Z_i) \tag{5.36}$$

indicating for particular values of β, η the deviation from zero correlation that would be attained by the true values of β and η. The GMM estimator proceeds by choosing values β, η that minimize the squared deviations

$$\min_{\beta,\eta} \hat{g}'\hat{W}\hat{g}, \tag{5.37}$$

with $\hat{W}$ a random norming matrix.

To ease the computational burden Newey modifies this procedure somewhat by extracting the parameters η out of the moment residual vector $\hat{R}$ such that only the coefficients β have to be estimated. Furthermore he derives a form for the GMM estimator $\tilde{\beta}_{GMM}$ that is based on his series estimator $\hat{\beta}_{Series}$ augmented linearly by an additional term. It just remains the choice of $m_j(\varepsilon_i)$, $\hat{A}(Z_i)$ and $\hat{W}$ on which the estimates depend. Newey aims to attain the semiparametric efficiency bound through a smart choice of $m_j, \hat{A}$ and $\hat{W}$. $\hat{W}$ is chosen as an estimate of the inverse asymptotic variance matrix of $\hat{g}(\beta)$. The vector of "supposedly uncorrelated" functions $\hat{A}(Z_i)$, for which β will be adjusted such that these functions are uncorrelated with the moment residuals $\hat{R}(\varepsilon_i, Z'_i\hat{\gamma})$, is constructed as

$$\hat{A}(Z_i) = \mathbf{a}(Z'_i\hat{\gamma}) \otimes Z_i, \tag{5.38}$$

with $\mathbf{a}(\cdot)$ a column vector of particular functions of the index: $\mathbf{a}(Z_i'\gamma) = (a_1(Z_i'\gamma), .., a_L(Z_i'\gamma))'$. To achieve the efficiency bound the functions $m_j(\varepsilon_i)$ and $a_l(Z_i'\gamma)$ must be bounded and three times differentiable with bounded derivatives. Newey suggests $m_j(\varepsilon_i) = m_0(\varepsilon_i)^j$ and $a_l(Z_i'\gamma) = a_0(Z_i'\gamma)^{l-1}$ with the root function, for instance, $m_0(\cdot) = a_0(\cdot) = 2\Phi(\cdot) - 1$, see Newey (p.25).

Under all these conditions the GMM estimator $\tilde{\beta}_{GMM}$ is *efficient*, provided that the preliminary step one estimator $\hat{\gamma}$ has been efficient in semiparametric sense, like Klein/Spady.

$$\sqrt{n}(\tilde{\beta}_{GMM} - \beta) \xrightarrow{d} N(0, V_{Newey\_GMM}) \tag{5.39}$$

The formulae to implement the $\tilde{\beta}_{GMM}$ estimator on basis of the preliminary Newey series estimates $\hat{\beta}_{Series}$ and it's variance matrix are given in the appendix.

5.3.4 Ahn/Powell 1993

The Ahn/Powell (1993,[1]) estimator distinguishes itself from the beforehand presented pure single-index estimators by retaining the single-index assumption only halfway. The typical single-index estimator specifies the same parametric form index function $\theta(Z) : \Re^q \to \Re$ both for the selection equation and for the structural equation, such that the selection mapping function $\varphi(\cdot)$ and the selection correction term $\lambda(\cdot)$ have support $\Re$. On the other hand the Ahn/Powell estimator imposes only a single index condition for the structural equation, that $\lambda(\cdot)$ has support $\Re$ with the propensity score $pr[d_i = 1|Z_i] = E[d_i|Z_i]$ being the index. And does not specify an index restriction $\theta(\cdot)$ for the selection equation, but rather estimates the participation probability nonparametrically, which is used as the index in the second step. By letting the selection function $\varphi(Z_i) : \Re^q \to \{0, 1\}$ being almost any smooth function mapping directly from $\Re^q$ to the binary space instead of prespecifying an index's functional shape ahead, the Ahn/Powell estimator is more general than the beforehand presented.

While the selection equation is estimated by Kernel methods to derive the participation probability $pr[d_i = 1|Z_i]$, the second step is very similar to the Powell estimator. Here individuals with similar participation probabilities are presumed to have similar selection correction terms λ since $\lambda = \lambda(E[d|Z])$ is assumed as a function depending only on the propensity score index, replacing the similarity of indices $Z'\gamma$ in Powell 1989 arisen through the parametric single-index assumption $pr[d = 1|Z] = F_u(Z'\gamma)$. The new element brought in by Ahn and Powell is the flexibility in the selection mechanism that permits accounting for unknown heteroskedasticity in the selection equation of any form, while the estimation of the structural equation retains the typical single index implication that all heteroskedasticity must enter through the index, with other words general heteroskedasticity depending on the regressors Z_i of the selection equation is prohibited in the structural relationship.

Their base-line model consists of an unspecified binary choice selection equation to be estimated nonparametrically and a structural equation of single index form:

$$\begin{aligned} d_i &= \varphi(Z_i, u_i) \rightarrow \{0,1\} \quad \text{for } i = 1..N \\ Y_i &= X_i'\beta + \lambda(\theta_i) + \xi_i, \quad \text{for } i = 1..n, \end{aligned} \tag{5.40}$$

with $\theta_i \in \Re$ representing the single index. With regard to the binary choice setting the participation probability $\theta(Z_i) = E[d_i|Z_i] \in [0,1]$ is selected as index and inserted in the structural equation:

$$Y_i = X_i'\beta + \lambda(E[d_i|Z_i]) + \xi_i \quad . \tag{5.41}$$

The step one estimation, producing the desired propensity score index $\hat{\theta}_i$, proceeds by nonparametric Kernel weighting within a neighbourhood of the choice characteristics Z_i (see appendix Kernel Regression). *Take notice that this step covers the full sample* $1..N$.

$$\hat{\theta}_i = \hat{E}[d_i|Z_i] = \left[\frac{1}{N}\sum_{j=1}^{N} d_j \cdot K_{ij}\right]\left[\frac{1}{N}\sum_{j=1}^{N} K_{ij}\right]^{-1} \tag{5.42}$$

with the weights K_{ij} obtained by a *multivariate* Kernel $K(\cdot)$:

$$K_{ij} = \frac{1}{h}K\left(\frac{Z_i - Z_j}{h}\right). \tag{5.43}$$

Given these estimates $\hat{\theta}_i = \hat{E}[d_i|Z_i]$ the structural equation $Y_i = X_i'\beta + \lambda(\hat{\theta}_i) + \xi_i$ is estimated in the second step, *now over the participation subsample* $1..n$, following exactly the Powell estimator with $\hat{\theta}_i$ as index instead of $Z_i'\hat{\gamma}$.

Although this estimator is obviously less efficient than the Powell estimator through it's use of nonparametric techniques, it still can achieve under stronger conditions $\sqrt{n}$-consistency and asymptotic normality, avoiding a worsening inefficiency with growing sample size:

$$\sqrt{n}(\hat{\beta}_{Ahn/Powell} - \beta) \xrightarrow{d} N(0, V_{Ahn/Powell}) \tag{5.44}$$

(An estimator for the asymptotic covariance matrix enclosed in the appendix). This result requires at first all the conditions imposed on the Powell estimator plus several more to ensure a uniform convergence of the step one Kernel estimator and an asymptotic bias of order below $\sqrt{n}$. The additional assumptions of Ahn and Powell ((3.8)..(3.11)) demand very high smoothness properties of the densities of the index θ, the selection correction form $\lambda(\theta_i)$ and other constructions and a Kernel of very high order, with all these demands increasing rapidly with the amount of continuous variables in Z_i. Let m denote the number of continuous selection variables in Z_i, than the Kernel function must be of order higher than $6 \cdot m$ and guarantee uniform convergence to reduce the asymptotic bias introduced by Kernel estimation.

Furthermore some products of $\theta_i, \lambda_i, etc.$ with the conditional densities of these continuous regressors must be more than $12 \cdot m$ times differentiable. These remarks alert that when employing this estimator the amount and the degree of measurability (dummy,integer,continuous) of the regressors should be handled with care to avoid the curse of dimensionality, as the accuracy and consistency of the estimates is likely to be sensitive to this choice.

For the step one estimation Ahn and Powell propose the multivariate Kernel:

$$K(z) = \sum_{j=1}^{6m+1} c_j \cdot \phi_q(z) \qquad \phi_q(\cdot) \sim N(0, \alpha_j \cdot A), \tag{5.45}$$

$\phi_q(z)$ being the q-variate normal density, A an arbitrary $q \times q$ positive definite matrix, α_j arbitrary, positive scalars and c_j be chosen such that

$$\begin{aligned} \sum_{j=1}^{6m+1} c_j \cdot \alpha_j^l \qquad & = 1 \text{ for } l = 0 \\ & = 0 \text{ for } l = 1, .., 6m. \end{aligned} \tag{5.46}$$

5.3.5 Robinson 1988

In contrast to all beforehand elucidated estimators Robinson's (1998,[45]) strategy aims to estimate the selectivity model without invoking any index restriction [SIR] on the error terms:

$$Y_i = X_i'\beta + \lambda(Z_i) + \xi_i. \tag{5.47}$$

Although appearing more general than the single-index estimators his estimator however relies on the stronger [IA] independence of the errors and regressors assumption (see his Theorem: (iii), p.939). One possible relaxation of the independence assumption by introducing an index restriction into Robinson's estimator has been suggested by Powell (1989,[41]) and would generate an estimator equivalent to Powell's. Though other approaches to relaxing the independence assumption could be imaginable as well. Robinson's idea is to subtract the statistical expectations of the structural equation (5.47) from the observed values for each individual to eliminate the selection correction bias $\lambda(Z_i)$. Taking expectations from equation (5.47):

$$E[Y_i|Z_i] = E[X_i|Z_i]'\beta + E[\lambda(Z_i)|Z_i] + E[\xi_i|Z_i] \tag{5.48}$$

$$= E[X_i|Z_i]'\beta + \lambda(Z_i) + 0 \tag{5.49}$$

$$\Rightarrow (Y_i - E[Y_i|Z_i]) = (X_i - E[X_i|Z_i])'\beta + \xi_i^{'}.$$

Inserting estimates of the conditional expectations $\hat{E}[Y_i|Z_i]$ and $\hat{E}[X_i|Z_i]$, that can be estimated nonparametrically similarly to the propensity score $\hat{E}[d_i|Z_i]$, allows OLS estimation of the equation

$$(Y_i - \hat{E}[Y_i|Z_i]) = (X_i - \hat{E}[X_i|Z_i])'\beta + \xi_i^{'} \tag{5.50}$$

over the participation sample $1..n$ to yield $\hat{\beta}$

$$\begin{aligned}\hat{\beta}_{Robinson} &= \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{E}[X_i|Z_i])\cdot(X_i - \hat{E}[X_i|Z_i])'\right]^{-1} \\ &\quad \cdot\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{E}[X_i|Z_i])\cdot(Y_i - \hat{E}[Y_i|Z_i])'\right].\end{aligned} \tag{5.51}$$

In an initial step a nonparametric density estimate of the conditioning selection variables $\hat{f}_z(Z_i)$ is prepared, then the conditional means are estimated. Take notice that in contrast to the Kernel estimation of the propensity score, which is estimated making use of the full sample $1..N$, see Ahn/Powell, here both $E[Y_i|Z_i]$ and $E[X_i|Z_i]$ are extracted from the participation subsample $1..n$, because Y_i is undefined for non-participants and both means have to be offsprings of the same (sub)sample.

$$\hat{f}_z(Z_i) = \frac{1}{nh^q}\sum_{j=1}^{n} K\left(\frac{Z_i - Z_j}{h}\right), \qquad \text{with } q = \dim(Z_i). \tag{5.52}$$

$$\hat{E}[X_i|Z_i] = \frac{\frac{1}{nh^q}\sum_{j=1}^{n} X_i \cdot K\left(\frac{Z_i - Z_j}{h}\right)}{\hat{f}_z(Z_i)}, \qquad \text{with } \hat{E}[Y_i|Z_i] \text{ analog,} \tag{5.53}$$

making use of a multivariate Kernel $K\left(\frac{Z_i - Z_j}{h}\right)$ with bandwidth h, that weighs the distance of the two vectors Z_i, Z_j. Note that observations with small estimated densities below a certain threshold $\left|\hat{f}_z(Z_i)\right| \leq b$ [10] should be deleted from the sample and the above estimates adjusted. To attain desired properties he suggests a particular higher order multivariate Kernel to reduce the bias at a fast enough rate, a multiplicative Nadaraya-Watson Kernel $K(z) : \Re^q \to \Re$ that is constructed as the product of the univariate Kernel weights of all elements of an q-dimensional vector:

$$K(z) = \prod_{d=1}^{q} k(z_d), \tag{5.54}$$

with z_d the d'th element of the vector z and an bounded univariate Kernel $k(z_d) : \Re \to \Re$. For the nonparametric estimates the Kernel $K\left(\frac{Z_i - Z_j}{h}\right)$ weighs the distance of the two vectors $Z_i, Z_j \in \Re^q$ by multiplying all the univariate Kernel weights of the pairwise differences $(Z_{id} - Z_{jd})$ for all elements $d = 1, .., q$ of these both vectors.

Robinson proves his estimator being $\sqrt{n}$-consistent and asymptotically normal, with an estimator of $\hat{V}_{Robinson}$ in the appendix,

$$\sqrt{n}(\hat{\beta}_{Robinson} - \beta) \xrightarrow{d} N(0, V_{Robinson}), \tag{5.55}$$

under the independence of errors and regressors [IA] assumption, smoothness and finite moment requirements and a bias-reducing suitable choice of the univariate

[10] The density estimate $\hat{f}_z(Z_i)$ could be negative for some Z_i due to the use of a higher order Kernel.

Kernel function $k(\cdot)$ which should be of order $q-1$, where q is the number of selection variables.

In his outlook about heteroskedasticity Robinson acknowledges the rigour of the independence assumption [IA] and suggests relaxation to allow for heteroskedasticity of ε_i conditional on X_i, Z_i, though he does not set forth this thought much further. If his estimation strategy could be extended to embrace conditional heteroskedasticity on X_i, Z_i and still retaining reasonable estimator properties, this would generate a competitive and more general estimator than the single-index estimators, which permit only heteroskedasticity conditional on the index. Otherwise Robinson's idea can be augmented by an index assumption [SIR] to relax the independence assumption [IA], which yields an estimator similar to Powell's.

5.3.6 Chen 1996

As the original paper Chen (1996,[8]) has not been accessible the following vague descriptions are taken from Moretti (1997,[35]). In contrast to the other step two estimators Chen's is able to estimate both the slope coefficients and the intercept of the structural equation at the same time by imposing additional to the usual mean-zero assumption on the errors the condition that the error terms are jointly distributed *symmetrically* around zero. The estimator is motivated on Powell's (1989,[41]) differencing scheme to eliminate the selection correction term. There, pairs of individuals i, j are tracked that have identical (or at least close) indices $Z_i'\gamma = Z_j'\gamma$, implying identical participation probabilities $pr[d_i = 1|Z_i] = pr[Z_i'\gamma > u_i|Z_i'\gamma] = F_u(Z_i'\gamma) = pr[d_j = 1|Z_j]$ *and* identical selection bias terms $\lambda_i = E[\varepsilon_i|Z_i'\gamma, d_i = 1] = E[\varepsilon_i|Z_i'\gamma, Z_i'\gamma > u_i] = \lambda_j$. With the additional symmetry condition pairs i, j with symmetric indices $Z_j'\gamma = -Z_i'\gamma$ will have identical bias terms $\lambda_i = \lambda_j$ but different propensity scores $pr[d_i = 1|Z_i] \neq pr[d_j = 1|Z_j]$. Through this differencing scheme Chen is now able to delete the selection bias without simultaneously eliminating all information about the intercept. In all other regards similar to Powell Chen estimates β and β_0 also with $\sqrt{n}$-consistency and asymptotic normality.

5.4 Step 2b-Estimation of the Intercept

The intercept β_0 of the structural equation is commonly of little interest in censored selectivity models, as it's absolute value being without any companion to compare with. Here the estimation of the slope coefficients $\hat{\beta}$ is in most cases all that is desired. In selectivity models where the outcomes of more than one category are observed, however, the identification of the intercept is crucial for predicting individual counterfactual outcomes by simulating the hypothetical (potential) outcome expected were this individual selected into an other category. These counterfactual

outcomes are indispensable for measuring structural differentials between categories for certain individuals or groups of individuals, that are requested for quantifying causal effects[11] in evaluation, for instance of manpower training programs, or for measuring the effect of unionism on wages etc.

The estimation of the intercepts has long been neglected. Besides Gallant and Nychka (1987,[13]) who estimate both the slope coefficients and the intercept simultaneous nonparametrically and refraining from Chen (1996,[8]), the only further estimators able to cope with the intercept β_0 have been Heckman (1990,[19]) and Andrews/Schafgans (1996,[2]).

5.4.1 Heckman 1990

Heckman motivated his strategy on the idea that the participation probability increases monotonously with the index value $\theta(Z_i) = Z_i'\gamma$. Hence for a subgroup of individuals with very high indices $Z_i'\gamma$ their participation probability comes close to one. Would some individuals achieve a probability of exactly one the nature of selectivity would vanish for this subgroup, as their choice or selection assignment is deterministically fixed by their observed characteristics and there is no need to model selectivity stemming form unobservable characteristics. For these observations ε_i would be mean-zero regardless of influences through the selection equation error u_i and their selection correction term λ_i be zero; $\lambda_i = E[\varepsilon_i|X_i, Z_i, d_i = 1] = E[\varepsilon_i|X_i, Z_i] = 0$, since Z_i implies $d_i = 1$. For this fraction of the dataset the intercept β_0 is identified as it can be distinguished from an intercept in $\lambda(\cdot)$, since $\lambda(\cdot) \equiv 0$ for these observations. Heckman suggested to estimate initially the selection equation coefficients $\hat{\gamma}$ (step one) over the full sample of size N to compute the indices $Z_i'\hat{\gamma}$. Then any step two estimator with appealing properties is performed on the participation subsample of size n to yield slope coefficient estimates $\hat{\beta}$ of the structural equation. By further trimming the participation subsample such that only observations with their index surpassing a certain threshold, $Z_i'\hat{\gamma} > b$, enter into the estimation of the intercept (step 2b), it is conjectured that for a threshold b sufficiently high all these remaining observations have a participation probability close to one, implying a selection correction term λ roughly zero. For b growing to infinity with increasing sample size, only a decreasingly small fraction of the dataset with indices $Z_i'\hat{\gamma}$ approaching infinity is employed to compute the intercept $\hat{\beta}_0$.

This exposes the main disadvantage of Heckman's and also Andrews/Schafgans's approach, in contrast to Gallant/Nychka who approximate the bivariate error term density, as identification hinges on sufficient distribution mass in the upper tail of the index $Z_i'\gamma$ which must be unbounded from above, and the estimation often rests

[11] The causal effect is the outcome difference attributed to a switch of the individual's category, if the treatment or program is the cause of this switch.

on just a handful of observations surpassing the growing threshold b which may be hard to distinguish from unreasonable outliers. Theoretically identification is only achieved for observations with infinite indices whose participation probability is then exactly one. This idea has been subsumed as "identification at infinity" by Chamberlain (1986,[6]).

Over the remaining fraction of the participation subsample surpassing the threshold b the intercept estimate $\hat{\beta}_0$ is calculated as:

$$\hat{\beta}_0^{Heckman} = \frac{\sum_{i=1}^{n} (Y_i - X_i'\hat{\beta}) \cdot 1(Z_i'\hat{\gamma} > b)}{\sum_{i=1}^{n} 1(Z_i'\hat{\gamma} > b)}. \tag{5.56}$$

The semiparametric Heckman intercept estimator $\hat{\beta}_0^{Heckman}$ under the [IA] independence assumption of errors and regressors has been proven to be consistent and asymptotically normal by Schafgans and Zinde-Walsh (1998,[48]), cited after Schafgans (1998,[47]).

5.4.2 Andrews/Schafgans 1996

Andrews and Schafgans (1996,[2] and 1997,[3]) extended Heckman's idea by replacing the trimming threshold through a weighting function that gives observations with higher indices increasing weight to achieve desired estimator properties. Their estimator with $s(\cdot)$ a weighting function and b' a smoothing parameter is

$$\hat{\beta}_0^{Andrews/Schafgans} = \frac{\sum_{i=1}^{n} (Y_i - X_i'\hat{\beta}) \cdot s(Z_i'\hat{\gamma} - b')}{\sum_{i=1}^{n} s(Z_i'\hat{\gamma} - b')}. \tag{5.57}$$

Appearing as a mere modification of Heckman's, it is more flexible and better explored and is consistent and asymptotically normal

$$\frac{\sqrt{n}E[s(Z_i'\gamma - b')]}{\{Var[\varepsilon_i] \cdot E[s^2(Z_i'\gamma - b')]\}^{\frac{1}{2}}}(\hat{\beta}_0^{Andrews/Schafgans} - \beta_0) \xrightarrow{d} N(0,1) \tag{5.58}$$

under the following conditions. Andrews and Schafgans initially impose the [IA] independence of errors and regressors assumption, though a relaxation is introduced later on. They demand $\sqrt{n}$-consistent and asymptotically normal preliminary estimators $(\hat{\gamma}, \hat{\beta})$ and require that the index $Z_i'\gamma$ is unbounded from above and has sufficient distribution mass in the upper tail. The bandwidth parameter b' must tend to infinity with growing sample size to guarantee that a decreasing fraction of observations with their indices approaching infinity is employed to form the intercept estimate $\hat{\beta}_0$. Lower and upper bounds for the convergence rate of b' are imposed. Furthermore the weighting function $s(\cdot)$ must fulfil certain properties, which are:

$s(\cdot) : \Re \to [0,1]$ is nondecreasing, three times differentiable and $s(x) = 0$ for $x \leq 0$ and $s(x) = 1$ for $x \geq c$ for some $c > 0$. A suitable choice would be

$$s(x) = \left\{ \begin{array}{ll} 1 - \exp\left(-\frac{x}{b'-x}\right) & 0 < x < b' \\ 0 & x \leq 0 \\ 1 & x \geq b' \end{array} \right| . \tag{5.59}$$

In contrast to other semiparametric step one and step two estimators $\sqrt{n}$ convergence is usually not achieved, which is unsurprising since with increasing sample size the bandwidth parameter b' must also increase, trimming of more and more observations that the number of remaining observations grows slower than the sample size n. However $n^{-\frac{1}{3}}$ convergence, yet even $\sqrt{n}$-convergence, can be attained, provided that the distribution of the index $Z_i'\gamma$ has a sufficiently thick upper tail compared to the upper tail of the distributions of ε_i and u_i. If the index $Z_i'\gamma$, ε_i and u_i have Weibull distributed upper tails of the same thickness the estimator converges with $n^{-\frac{1}{3}}$. For the index's upper tail thicker than the errors' even $n^{-\frac{1}{2}}$ is achieved, see Andrews and Schafgans (1997,[3], p.2,8). Worth noting is also that asymptotic normality is not attained when the index's upper tail is thinner than the errors'. These remarks emphasize once more the importance of a continuous and widespread distribution of the index $Z_i'\gamma$ with abundant tail observations for the both presented intercept estimators, which may be exceptional in practical work. The estimator of Gallant and Nychka, that does not found itself on the philosophy of identification at infinity, would be a highly competitive rival if it's asymptotic distribution could be established.

A relaxation of the [IA] independence of errors and regressors assumption to the [SIR] index assumption broadens the scope of this estimator and requires only slightly stronger conditions. For this extension all beforehand presented results hold and an estimator of the variance of the intercept estimate $\hat{V}ar[\hat{\beta}_0]$ is enclosed in the appendix. Andrews and Schafgans' finding that the joint asymptotic variance-covariance matrix of $(\hat{\gamma}, \hat{\beta}, \hat{\beta}_0)$ is blockdiagonal simplifies substantially the implementation of hypothesis testing etc., since the intercept's variance is unconnected with the variances of both other preliminary estimators $(\hat{\gamma}, \hat{\beta})$.

A weakness of both presented estimators is the absence of a formal rule of how to select a proper value for the trimming or bandwidth parameter b, b', respectively. Since the estimators' results may be sensitive to this choice the semiparametric estimates could then appear arbitrary in this light. Schafgans (1997,[46] and 1998,[47]) illuminates this connection in detail in her Monte Carlo studies and some of her results are brought up below in the chapter Properties.

Chapter 6

Properties of these Estimators

This chapter will provide a brief but comprehensive assessment of the estimators' properties. At first it should be considered what are desired properties of an estimator and what are serious deficiencies. As Powell (1994,[42], p.2460) remarks, that finite sample properties of semiparametric estimators are difficult to derive due to the generality of the imposed error term restrictions, thus almost all results rest on asymptotic theory for infinite samples, which may or may not be a good guide to finite sample behaviour. Only extensive Monte Carlo studies can shed a light in this respect.

Consistency, efficiency and a known asymptotic distribution of the coefficient estimates are commonly regarded as the most important assets of an estimator, while other features like small sample behaviour, ease of implementation or computational effort to estimating the coefficients are in addition highly appreciated. Furthermore a consistent way to compute the estimates' variance matrix is expected.

Consistency: Consistency has been usually demanded as the minimum condition of an estimator worthwhile to be examined. Nevertheless it may be questionable whether an efficient but slightly inconsistent parametric estimator should all the time be dismissed in favour of an inefficient but consistent nonparametric one. unbiasedness even in small samples would be an asset, but is usually not achieved by semiparametric estimators. Due to their reliance on nonparametric algorithms small samples issues are hard to deal with. To evade the asymptotic bias, habitually introduced by these nonparametric tools, these estimators often make use of bias-reducing higher order Kernels.

Inference: Besides estimating coefficient values the precision of these estimates, the credibility or with other words the range in which the true values are likely to be, is another important characteristic of an estimator. For inference on the estimates their distribution has to be known. In contrast to the linear regression model where small sample distributions on the basis of the t-distribution are established for normal error terms, the theory of semiparametric estimation concentrates in

this respect on asymptotic results solely, that may approximate the small sample behaviour. Though, extensive Monte Carlo studies are needed to produce reliable guidelines for small samples. The initial step for inference is to find out the asymptotic distribution of a stabilized transformation of the estimates. This results in an establishment of the estimator's convergence rate and it's asymptotic distribution. The rate of convergence exposes the speed at which the estimator's variance shrinks to zero as the sample size increases and gives some clue about the estimator's accuracy in finite samples. A rate of $N^{-\frac{1}{2}}$ (often referred to as $\sqrt{n}$-consistency) is the benchmark achieved by parametric estimators and the more promising semiparametric ones, signifying that a quadrupling of the sample size halfs the estimates' variance. Many semiparametric estimators and the nonparametric techniques do not achieve this rate and converge slower towards their true values, resulting in less precise estimates given the same sample size. Besides this convergence rate the estimator's asymptotic distribution is of interest. Some semiparametric estimators have been proven as asymptotically normally distributed around the true coefficient values with an asymptotic variance-covariance matrix. In contrast, the distribution of the parametric estimators depends on their specification of the bivariate error term density, habitually bivariate normal. For executing inference in practical work a consistent estimator of the estimates' asymptotic variance matrix is requested, on which inference must be based since proper small sample theory has not been established for the semiparametric case and finite-sample variance estimators are lacking. To determine efficiency between these estimators and relate it to a theoretical efficiency bound the asymptotic variance matrices of estimators with the same convergence rate are compared.

Trade-off: Consistency against Efficiency: Quite often a trade-off between consistency and efficiency, stronger assumptions yield more efficient estimates if the assumptions are true for the cost of inconsistency if they are not, prohibits an unambiguous assessment of the superiority of one estimator above an other. Yet when comparing estimators with the same set of assumptions the most efficient is to prefer.

Efficiency bounds: Theoretical efficiency bounds represent the lowest variance conceivable within a certain class of estimators, though not always achieved by any known estimator. For the parametric models the Cramer-Rao is the corresponding efficiency bound that is attained by Maximum Likelihood. Chamberlain (1986,[6]) established the efficiency bound for the class of semiparametric models where the single index restriction [SIR] is implemented both in the structural equation and in the selection equation by a linear index function $\theta(Z_i)$. The GMM estimator of Newey achieves this bound and is consequently the most efficient within the semiparametric single-index class. Estimators that impose less structure on the selection process by not assuming an index restriction on the selection equation and utilizing a nonparametric index for the structural equation cannot attain this

bound. Such a nonparametric index could be the propensity score $\theta(Z_i) = E[d_i|Z_i]$, employed in Ahn/Powell. Newey and Powell (1993,[40]) derived efficiency bounds for the selectivity model with nonparametric index for different identifying conditions on the error terms, once for independence of the errors (ε_i, u_i) from the regressors and once for the weaker mean-independence of errors and regressors $E[(\varepsilon_i, u_i)|Z_i] = 0$. They also testify that so far neither Ahn/Powell nor any other known estimator achieves this bound.

Assumptions and other features: Linked with an estimator's properties are obviously it's inherent assumptions, regarding to model specification, conditions on the regressors (like exclusion restriction) and so forth. Loosely speaking, an estimator can easily acquire more precise and efficient results the more assumptions it invokes. In this respect all the presented semiparametric estimators stand rather on the same footing. They all demand exclusion restriction, similar smoothness conditions and rely (or can be modified to) on the single-index assumption. Only the estimator of Chen stands out, which requires additionally a symmetry condition on the joint error term distribution. The Ahn/Powell estimator, for being very generous to the selection mechanism, demands in return rather hefty smoothness conditions to counter the curse of dimensionality and still achieve $\sqrt{n}$-convergence rate. Other qualities of an estimator are the computer time consumed for it's execution and the simplicity of it's use. Estimators with a closed form representation are advantageous since they are straightforward to program and calculate their estimates directly instead of searching for a global maximum like the Maximum Likelihood variants, avoiding convergence problems and often the need for starting values. A shortcoming common to all semiparametric estimators are their need for bandwidth or smoothing parameters, which have to be chosen lastly by the econometrician itself as no infallible automatic rule is available. This makes the estimates dependent on the econometrician's experience, less comparable and perhaps even somewhat arbitrary.

Now in order of their presentation some semiparametric estimators' properties are outlined:

The semi-nonparametric ML estimator Gallant/Nychka, approximating the bivariate density, is the only one-step estimator brought up and is capable to identify besides γ and the slope parameters β also the intercept β_0 without relying on sparse and perhaps unreliable tail observations of the index, which are the basement of the "identification at infinity" philosophy. It is implemented via ML and can handle most non-normal distributions through sufficient approximating terms. Yet it is unable to cope with distributions with very fat tails (thicker than t-distribution). The estimator imposes the independence of errors and regressors assumption [IA], though can be augmented to allow for heteroskedasticity as sketched in Melenberg

and van Soest (1993,[33], p.15). It's major drawback is that it's asymptotic distribution is unknown, prohibiting proper inference. Yet for a beforehand arbitrarily fixed number of approximating terms K this approach corresponds to the fully parametric case with standard errors readily computed and all it's other advantages and flaws. Melenberg and van Soest further mention that the computational burden grows with K^4, that since the likelihood function is non-concave potential convergence problems to the global maximum may occur and that the estimator is not semiparametrically efficient.

For the separate estimation of the selection equation only the Klein/Spady has been illustrated exemplary. This single-index estimator is $\sqrt{n}$-consistent, asymptotically normal and attains the semiparametric efficiency bound. As a (quasi) ML estimator, in the same way as for ML, no closed form representation is available. The Klein/Spady has been quite popular with many researchers and appears to comport well in applied work, see Melenberg and van Soest (1993,[33]) or Lanot and Walker (1998,[25]).

A variety of step two estimators for the structural equation's slope coefficients β have been presented. Almost all of them are $\sqrt{n}$-consistent and asymptotically normal if not indicated otherwise. For all a closed form representation exists.

The single-index estimator of Powell sidesteps the estimation of the selection correction term λ by eliminating it through pairwise differencing. The construction of the estimator stimulates the conjecture that for a proper behaviour in small samples the indices of the individuals in the participation subsample must be dense or reasonably close to each other. This seems important since the idea of cancelling out the selection correction term works exactly only for identical indices and becomes increasingly fragile for distant individuals. If the observations are isolated and widely spread the Kernel algorithm has to perform over a wider neighbourhood to reach a sufficient amount of individuals and the assumption of identical correction terms λ for all these individuals becomes less credible. Thus I suppose that this estimator works better in settings with the true distribution of the index concentrated around it's mean and little mass in it's tails.[1]

The Newey series estimator is also a single-index estimator approximating the selection correction term by a polynomial on basis of preliminary step one estimates $\hat{\gamma}$ and performing the second step with OLS. The estimator is straightforward and easily implemented since the coefficients β are estimated by OLS and quickly computed estimates of the variance matrix are available, see Melenberg and van Soest (1993, [33]). When choosing powers of the inverse Mill's ratio as basis functions this series estimator nests the parametric bivariate normal case.

[1] Note that such a distribution is the opposite to the requirements of the pure intercept estimators Andrews/Schafgans and Heckman, that solicitate abundant distribution mass in the upper tail of the index distribution.

The Newey GMM imposes also the single-index assumption though utilizes the entailed information more thoroughly to attain the semiparametric efficiency bound. This GMM is the only semiparametrically efficient step two estimator presented. However it is more complicated to compute and requires Newey's series estimator for preliminary estimates, yet through it's closed form it is straightforward to implement.

Ahn/Powell is a modification of the Powell estimator with loosened single index restriction that is imposed only halfway. The nonparametric index alias the propensity score $E[d_i|Z_i]$ is invoked only in the structural equation, while it is estimated nonparametrically in the step one estimation of the selection equation. This proceeding leaves substantially more flexibility and robustness to the selection mechanism, but may deteriorate the estimator's properties. $\sqrt{n}$-consistency can be attained, though the corresponding requirements are fairly demanding and become increasingly severe with the number of variables in Z_i. Additionally a Kernel of very high order has to be applied. Hence the nonparametric estimation of the propensity score is less efficient than Klein/Spady and becomes increasingly inaccurate for high-dimensional Z_i. Besides this, the use of the higher order, bias-reducing Kernel may produce puzzling participation probability estimates for some individuals that are below zero or larger than one. Furthermore Melenberg and van Soest (1993,[33], p.25) remark that the calculation of this estimator's variance matrix may take up a lot of computer time.

Robinson bypasses any index restriction and estimates the conditional first moments of d_i, Y_i, X_i nonparametrically by multivariate Kernel methods. Instead of the single index assumption, however, he imposes the stronger independence assumption. Due to it's ample use of nonparametric techniques this estimator is especially susceptible to the curse of dimensionality.

The single-index estimator of Chen is able to identify both the slope coefficients β and the intercept β_0 jointly with $\sqrt{n}$-convergence, rivaling the pure intercept estimators of Andrews/Schafgans and Heckman that often fall short of $\sqrt{n}$-consistency. To achieving this the estimator has tightened up the mean-zero condition on the errors to the assumption that the error terms are distributed symmetrically around the origin. Except for the reduced generality through this stricter assumption this estimator is in other respects similar to Powell.

Two pure intercept estimators have been illuminated, both relying on the "identification at infinity"-philosophy. Charlier, Melenberg and van Soest (1997,[7], p.8 Footnote) remark that a fruitful application of these estimators requires a good many observations in the sample with participation probabilities close to one. Otherwise the approaches of Chen or Gallant/Nychka might be more promising.

Both intercept estimators have been shown to be consistent, asymptotically normal and asymptotically independent of the previous estimation stages, though both

fall usually short of $\sqrt{n}$-consistency. Their convergence rate improves with thicker upper tails of the index $Z_i'\gamma$ and with thinner upper tails of the errors ε_i, u_i. Under certain conditions Andrews/Schafgans can then achieve $\sqrt{n}$-rate. Two Monte Carlo studies by Schafgans (1997,[46], 1998,[47]) have provided useful insights into the less explored finite sample properties. The both semiparametric estimators $\hat{\beta}_0^{Heckman}$ and $\hat{\beta}_0^{Andrews/Schafgans}$ are compared to the parametric Heckman two-step and reveal as a major drawback of the semiparametric approach the absence of a rule-of-thumb to determine a proper value for the bandwidth parameter that is essential to both estimators. In contrast to both previous steps ($\hat{\gamma}$ and $\hat{\beta}$ estimation) where generalized cross-validation or the proposals of Powell and Stoker (1996,[44]) are a valuable guide, this task remains still unaddressed for the semiparametric estimation of the intercept β_0. Hence Schafgans compares these estimators over a large range of possible bandwidth parameter values and consequently comes up with rather vague results. No explicit comparison between the both semiparametric estimators has been pursued, though I would cautiously conjecture from her graphs and indications that in terms of Root Mean Squared Error measure the Andrews/Schafgans estimator often performs better than the semiparametric Heckman estimator than vice versa, see Schafgans' graphs 1,2,3,4,6 (1998,[47]).

Chapter 7

Application of these Estimators

After having learned much about semiparametric estimators and their properties it remains to be discussed how to proceed when employing them in practical matters. At first some general remarks about model specification and the choice of covariates will be given. Hereafter some advice to an appropriate choice of the estimation chain $(\hat{\gamma}, \hat{\beta}, \hat{\beta}_0)$ is added. A brief demonstration of a recent research application on wage structures to illustrate these estimation steps will complete this chapter. It should be noted that although many semiparametric selectivity estimators have already been developed in the late 1980s applications are still rare.

Model Specification:

Initially, before commencing with concrete specifications, it should be questioned if the selectivity model is the most appropriate model or wether an other, e.g. the two-part model, should be preferred or additionally explored. The specification as a selectivity model requires for identification plausible exclusion restrictions when applying semiparametric methods, which frequently may be hard to justify as Melenberg and van Soest (1996,[34]) document when modelling vacation participation [selection part] and vacation expenditures [structural part], since in economical applications all economical, demographical, regional and other variables are often likely to affect both the choice decision and the outcomes of the structural equations. Therefore they reject the selectivity model and proceed with the two-part model broached in chapter Parametric Estimation. This two-part model is unaffected by selectivity and can be estimated easily by OLS, though it's coefficients have a different meaning than in the selectivity model. Since those β-coefficients mirror just the conditional outcome relationship for the subpopulation of vacationers while in the selectivity model the outcome equation reflects the true potential outcome structure for the entire population, usable for structural analysis and for if-then simulations. As they remark (p.67) the appropriate model depends rather on the object of interest. If modelling potential outcomes is meaningful as in the case of latent wages, that reflect the wage an unemployed person would earn given she finds a job, or if

selectivity must be modelled for structural analysis or evaluating treatment effects the selectivity model is adequate, provided exclusion restrictions can be justifiably imposed. On the other hand, in the vacation example potential vacation expenditure for a family that does not participate in vacation does not make sense, since the family has decided to spend nothing by their own decision not to participate in vacation. Then the two-part model is more appealing.

The concrete semiparametric model specification depends largely on the available dataset, whether outcomes are censored or even truncated, on the number of categories and so forth. The most promising approach would be to specify a starting model and verify its appropriateness by tests on the estimation results. The usual way is the specification of a homoskedastic and linear, single-index selectivity model, where both the structural relationship and the index function are a linear combination of not only the important individual characteristics but can also include squares or powers of some variables and cross-terms. This permits ready estimation as standard linear methods are applied and leaves also sufficient flexibility for the shape of the relationship between explanatory and the dependent variables through the inclusion of interaction-terms. Having obtained estimates for this model specification tests like the score tests of Chesher and Irish (1987,[9]) should be carried out to verify homoskedasticity or the single index assumption, which allows for heteroskedasticity depending only on the index $Z_i'\gamma$. If these tests do not support the standard specification a ready remedy is often to modify the composition of the covariates. If this does not bring relief the two weaknesses of the semiparametric approach: the restrictiveness against heteroskedasticity and the parametric specification of f and θ should be tackled, either by embracing heteroskedasticity (Ahn/Powell, Gallant/Nychka, nonparametric models) or trying other than linear forms for f and θ. For example Ahn/Powell helps to distinguish whether the difficulties lie with the selection or the structural part.

Exclusion Restriction:

For identifying β and $\lambda(\cdot)$ in the structural equation absence of multicollinearity between X_i and λ_i must be ensured. Multicollinearity does not occur when the correction function $\lambda(\cdot)$ is sufficiently nonlinear over the support of the explanatory variables in the available dataset. This however can generally not be guaranteed for the semiparametric estimators, since they leave the function $\lambda(\cdot)$ completely unspecified and nothing prohibits $\lambda(\cdot)$ from being collinear with X_i. Even for the fully parametric approach identification may be weak due to the almost linearity of the inverse Mill's ratio over a considerable parameter space resulting in imprecise estimates, see chapter Parametric Estimation.

A second way to preclude multicollinearity between $\lambda(\cdot)$ and X_i is the assumption of an exclusion restriction, guaranteeing that the regressor sets X_i and Z_i do not completely overlap. As the selection equation is customarily modelled in a reduced

form representation of a deeper behavioural conception leaning on the structural outcome relationships all the variables in X_i have to be included in the selection equation regressors Z_i. On the contrary, variables that are contained in Z_i may be excluded from the structural equation regressors X_i if they are unimportant for the structural relationship, while essential for the selection decision. These exclusions restrictions theoretically secure identification, though practical experience advises that some care should be taken when selecting dismissable variables.

At first, it should be definitely sure that the excluded variables have no influence on the outcome equation, with other words that these omitted variables would have a zero coefficient in the true model. If no such covariates can be found that plausibly, by economical or behavioural theory, do not enter in the structural equation but do concern in the selection equation the semiparametric specification of the selectivity model seems inadequate and either parametric ML, which does not require an exclusion restriction or an other model should be chosen. Furthermore, the identifying effect of exclusion restrictions increases considerably with the number of covariates excluded from the structural equation, with their degree of measurability (binary,integer,continuous) and with their size of support within the underlying dataset. In applied work seldomly more than one exclusion restriction can be justified, yet, the removal of just one dummy variable may help little as it's support of at most two different values leaves much room for collinearity within each of the two resulting subsamples. At least one *continuous* variable excluded with ample variation within the sample has been proposed by many practitioners, unlike for instance a continuous variable unemployment that may take only a very limited number of different values within one cross-section. Some examples now:

Lanot and Walker (1998,[25]) exclude in their 2-category analysis of the effect of union membership [selection] on wages [structure] the regressors "unearned-income" and number and age of children from the structural wage equation. They claim that these characteristics do not influence the market wage, but conjecture that the decision to join a union may depend on these as union supplied services, like insurance or security may matter different.

Moretti (1997,[35]) investigates the wage differential [structure] between seasonal and permanent jobs [selection] in agriculture and assumes that the unemployment rate, which varies due to pooled cross-sections and the "unemployment insurance dummy" have no effect on the wage but on the job-preferences of the individuals.

Melenberg and van Soest (1993,[33]) examine wages [structure] and labour participation [selection] of women and exclude "other family income" as unconnected with the wage but crucial for the decision whether to work at all.

Newey, Powell and Walker (1990,[39]) analyze the labour supply of women in a censored selectivity model with labour participation as the selection part and the annual working hours as the structural equation. Excluded from the structural part

were the variables "years of labour force experience" and some other.

Charlier, Melenberg and van Soest (1997,[7]) exclude the non-continuous variable "household's head's education level" when inspecting the determinants of housing expenditure [structure] for the two groups owners and renters [selection], since they conjecture that this characteristic inspires the choice between owning or renting but should not concern the amount of spending.

In contrast, Melenberg and van Soest (1996,[34]) reject the imposition of any exclusion restriction and therewith the selectivity model when modelling both the choice whether to go on vacation and the expenditure on vacation [structure], since no regressor has been found that plausibly by economic theory would have an import on only one of these decisions and not on the other.

Single Index Assumption:

The single index assumption is crucial for many semiparametric estimators and is slightly weaker than the [IA] condition of independence between the errors (ε_i, u_i) and the explanatory variables (X_i, Z_i), as it allows for conditional heteroskedasticity. The SIR's merit is the dimension reduction that it brings to the estimation process.

The estimation of the selection equation simplifies considerably with the index assumption. The functional form of the selection mechanism becomes completely pre-specified up to merely q unknown coefficients and only the distribution of the selection error u_i stays vague. With a linear index the different selection variables become "proportional" and comparable since their impact on the index and on the selection mechanism is now measured in the same unit of measurement, the index unit, easing the estimation and the estimates' interpretation. This means for example that for coefficients $\binom{0.1}{0.2}$ for the selection variables vector $\binom{A}{B}$, B has twice as much influence on the selection assignment than A and that this relationship stays the same with proportional variation of A and B.

The second step estimation of the structural equation gains also enormously from the index restriction since the unspecified selection correction term $\lambda(Z_i)$, that could be almost any function of Z_i, no longer depends on Z_i of dimension q but entirely on the one-dimensional index $Z_i'\gamma$, escaping the curse of dimensionality. On theoretical grounds a continuous support of the index distribution is required. Nevertheless, the single index condition may be too strict, for instance, to embrace heteroskedasticity that cannot be explained completely as a function of the index. If heteroskedasticity tests, see Chesher and Irish (1987,[9]) reject the single index model more general estimators can be employed, though they lose quickly efficiency (Ahn/Powell, Gallant/Nychka with heteroskedasticity, multivariate Kernel regression or other nonparametric techniques).

Choice of Regressors:

The selection of the variables to include in the both equations goes hand in hand with the model assumptions made and the choice of the semiparametric estimator. All relevant characteristics should be included, squares and cross-terms may be added to enrich the linear specification. For the single-index assumption at least one of the significant regressors in the selection equation should be observed continuously to ensure continuous distribution of the index and at least one continuous regressor should be found that is significant in the selection equation and can justifiably be excluded from the structural equation. For estimators that do not impose the index restriction on the selection part, like Ahn/Powell or Robinson, however, the number of regressors and their degree of measurability (binary,integer,continuous) should be kept small, since their nonparametric methods' convergence rates deteriorate substantially with higher dimensional regressor vectors and therewith the accuracy of their estimates.

Dual Strategy:

When it comes to the decision whether to proceed with a parametric or semiparametric specification it should be recollected that the parametric procedures, like the Heckman two-step or Maximum Likelihood, are easily implemented, quickly computed and are more efficient than semiparametric techniques. Accordingly these estimates are regularly more precise, the intercept is identified at once with the slope coefficients and prediction outside the support of the explanatory variables of the dataset is possible. Furthermore even the more susceptible Heckman two-step appears in it's both steps to be considerably robust to misspecification of the distribution in many studies, e.g. Newey, Powell and Walker (1990,[39]). Though other investigations, like Klein/Spady (1993,[24]) or Schafgans (1998,[47]), testified a substantial inconsistency when the distribution of the error terms departed seriously from the bivariate normal.

On this finding, many researchers prefer a "dual strategy" to check this trade-off by analyzing both the parametric and semiparametric estimation results of the same model specification[1] and select for each step the more appealing estimates. Habitually the dual strategy opens with the parametric Probit (Heckman's two-step first step) and a semiparametric step one estimator, like Klein/Spady, of the selection equation. The results of the semiparametric estimator are used as a benchmark for the reliability of the parametric estimates since both estimates should be identical in the case that the parametric specification mirrors the true model when consequently both approaches are consistent.

In this spirit casual observation and distance tests between $\hat{\gamma}^{semi}$and $\hat{\gamma}^{para}$e.g. by

[1] Differing only with regard to the error terms' distributional assumptions.

χ^2-tests under the null-hypothesis of equality of both coefficient estimates[2] measure the closeness between both results and give some hint about a serious misspecification of the parametric model. Given that the equality-hypothesis is not rejected it is recommended to go ahead with the parametric coefficient estimates as parametric models can be estimated more efficiently than semiparametric ones. On the other hand, if the null is rejected, which seems to occur rather rarely, it is to proceed with the semiparametric results $\hat{\gamma}^{semi}$since the parametric estimates may be inconsistent. The same procedure should be applied for the second step, where the parametric Heckman's second step OLS is judged by a semiparametric step two estimator of the structural equation. If equality-tests of the slopes do not reject the parametric estimates $\hat{\beta}^{para}$and $\hat{\beta}_0^{para}$should be maintained as final results. Otherwise remain with $\hat{\beta}^{semi}$and estimate further the intercept $\hat{\beta}_0^{semi}$in a final step.

As a result, the dual strategy can be summarized as measuring in each separate estimation step the credibility of the more efficient parametric results and continuing with these parametric estimates if no evidence for inconsistency has been detected, otherwise with the semiparametric ones.

Choice of the semiparametric estimator:

Although proper Monte Carlo studies about the estimators' finite sample behaviour are rarely available some guidelines can be extracted from their asymptotic properties and practicing researchers' comments brought up in the chapter Properties.

For the estimation of the selection equation many discrete choice estimators have been suggested over the time. The Klein/Spady has been pretty popular with practical men and appeals as well in finite samples as in asymptotic theory: $\sqrt{n}$-consistent, semiparametric efficient and normally distributed.

The favourite candidate for the estimation of the slope coefficients β are the both algorithms of Newey. At first his series estimator which is straightforward to implement and computationally little demanding should be computed to acquire a first glimpse at the estimates. If then more precise results are demanded the semiparametrically efficient GMM can be performed on top of these initial outcomes. This is a great advantage over the other estimators that are less efficient and do not have this upgrading device at their disposal. Gallant/Nychka might be interesting

[2] A χ^2-test of equality of two coefficient vectors $\hat{\beta}_1 = \hat{\beta}_2$ with the test statistic calculated as $(\hat{\beta}_1 - \hat{\beta}_2)'[\hat{V}(\hat{\beta}_1 - \hat{\beta}_2)]^{-1}(\hat{\beta}_1 - \hat{\beta}_2)$ has been executed by Newey, Powell and Walker (1990,[39]) in their analysis of female labor supply in both estimation steps on the corresponding parametric and semiparametric estimates $(\hat{\gamma}, \hat{\beta})$. Equality of $\hat{\gamma}^{para}$with $\hat{\gamma}^{semi}$ and $\hat{\beta}^{para}$with $\hat{\beta}^{semi}$ was not rejected and the parametric estimates appeared consistent. Be aware that β did not include an intercept since β_0 was not identified by the applied semiparametric methods. For the χ^2-test, obviously, both vectors should have been transferred to the same norming, e.g. the first element's coefficient set to one, if the parametric and semiparametric estimators come up with different norming defaults, like $\hat{\sigma}_u^2 = 1$ in the Probit and $\hat{\gamma}_1 = 1$ for Klein/Spady.

either, though it's asymptotic distribution is unknown.

The Ahn/Powell is an alternative if more flexibility for the selection mechanism is indicated. If tests on heteroskedasticity reject the single index assumption the Ahn/Powell can be used as a benchmark for the Newey and the parametric estimators to detect whether heteroskedasticity is severe either only in the selection or the structural part or contaminating both. However the curse of dimensionality makes Ahn/Powell infeasible in applications with many regressors.

The Chen estimator is an appealing rival to Andrews/Schafgans and Gallant/Nychka when it comes to the estimation of the intercept β_0, provided that it's additional symmetry assumption seems trustworthy by it's slope coefficient estimates $\hat{\beta}$ in line with a more general semiparametric slope coefficient estimator, e.g. Newey.

Estimating the intercept:

If an estimate of the intercept is required the more general methods are those proposed by Heckman and Andrews/Schafgans. However, as these estimators rely on sufficient and trustable observations in the tails of the index distribution and make use of only a small portion of the data the estimator of Chen may be considered to overcome these weaknesses. Chen's additional condition that the joint distribution of the error terms is symmetric around the origin allows the estimation of both the slope parameters and the intercept at the same time. This extra requirement makes this estimator less general compared with the other estimators, yet it bridges the estimation of the slopes and of the intercept. My proposal for selecting the adequate intercept estimator exploits this link by initially employing the Chen estimator accompanied by a more general semiparametric step two slope-estimator and examining whether both sets of slope coefficient estimates are alike by testing for equality of the slopes of $\hat{\beta}^{Chen}$ and $\hat{\beta}^{other}_{semi}$ using a χ^2-test. If all $\hat{\beta}^{Chen}$ slope estimates are close to the slope estimates of the more general estimator and the equality hypothesis has not been rejected the extra symmetry assumption of the Chen estimator seems to be undisturbing and consequently also the estimated intercept $\hat{\beta}^{Chen}_0$ convincing. Otherwise for equality $\hat{\beta}^{Chen} = \hat{\beta}^{other}_{semi}$ rejected, the estimation of β_0 should proceed with the intercept estimator of Andrews/Schafgans.

Choice of smoothing parameters:

The need to select manually appropriate values for the smoothness parameters is a serious deficiency of all semiparametric and nonparametric techniques inasmuch as estimation results may be sensitive to the choice of these nuisance parameters, and the more of these have to be chosen the worse it gets in practical matters.

These smoothing parameters may be the bandwidth (window width) choice for Kernel algorithms, the number of terms included in a series approximation to the bivariate density (Gallant/Nychka) or to the selection correction term $\lambda(\cdot)$ (Newey and others) or the number of moments included in Newey's GMM. For the estimation of

the selection equation, commonly by Kernel regression, *generalized cross-validation* introduced by Craven and Wahba (1979,[11]) is straightforward to accomplish and produces good suggestions for the bandwidth h_n. If the estimation of the structural equation is performed on the basis of Kernel methods cross-validation however does not always advise the optimal Kernel bandwidth (Moretti, 1997,[35], p.16) and Powell and Stoker (1996,[44]) suggest alternative ways for determining the smoothness parameters. The appropriate number of terms in series approximation is habitually determined by estimating nested models differing only by the amount of approximating terms and selecting the most appealing model on the basis of significance levels of the approximating terms, likelihood ratios or other goodness-of-fit criterions.

However these guides do not always lead automatically to the right choice and consequently a low number of smoothness parameters (at best none) is an appreciated quality of a semiparametric estimator in practical terms. Particularly difficult is the choice of the bandwidth parameter for the estimation of the intercept $\hat{\beta}_0$ in Heckman (1990,[19]) and Andrews/Schafgans (1996,[2]). No formal method has been developed to select the trimming bound. To have a rough guide to chose this parameter for a present dataset Schafgans (1997,[46], p.30f or 1998,[47], p.12f) sketches a track how model pretesting may help, though no rule-of-thumb has been derived and further research is necessary to clear up this incertitumbre. An examination of her graphs reveals a crude lower bound for this bandwidth parameter corresponding to approximately 50% censoring, meaning that the trimming parameter b' should be chosen sufficiently high that at least 50% of the n observations of the participation subsample are trimmed of for the intercept estimation.

An exemplifying illustration of one recent research study where some of these estimators have been put to use allows to acquire an insight into the practical value of these techniques with regard to empirical work. The inclined reader may further consult the studies mentioned with the exclusion restriction examples.

In their study Lanot and Walker (1998,[25]) seek to quantify the effect of union membership on wages on the basis of around 20,000 observations of manual workers. They apply both parametric and semiparametric estimators within a selectivity model and additionally OLS in a linear regression model simply regressing the wages on the explanatory variables in both sectors $w = X'\beta + \varepsilon$, disregarding potential endogeneity of the union membership status and neglecting selectivity correction. Their selectivity model is characterized by a choice equation determining union membership status and two wage equations, one for union members and one for non-members, with the excluded variables "Unearned Income" and the number and age of children for the semiparametric estimators. The parametric Heckman two-step and the semiparametric estimation chain Klein/Spady, Newey Series and

Andrews/Schafgans have been performed and compared with each other and with the selectivity disregarding OLS estimator. On these results they concluded that the parametric selectivity specification is robust since recorded estimates have been very similar across all methods based on the selectivity model. The both OLS single equation regressions differ from the results of the selectivity model as was already expected due to the significance of the selection correction terms λ_{union}, $\lambda_{non-union}$. Interestingly, sizeable deviations between the selectivity model and the OLS linear regression model occurred only for the estimates of the both intercepts while the slope estimates have been remarkably stable among both models. This difference in the estimated intercepts is the origin of much higher wage differentials between union members and non-members (around 25%) in the selectivity model than in the selectivity-neglecting OLS wage equation regressions (differential about 10%). This exemplifies the value of identifying the intercept when comparing different categories and the importance of allowing for endogenous union membership where workers select into their categories on the basis of expected wages. Lanot and Walker show also the cruciality of a continuous excluded variable by altering the exclusion restrictions, and they furthermore repeat the same analysis on 10% subsamples of the impressing origin sample of 20,000 observations. Their similar estimation results point in favour of the small sample properties of the semiparametric approach.

Chapter 8

Conclusions

After having introduced the selectivity model as a means to capture structural population relationships in an environment where no representative samples are feasible, the semiparametric approach has been contrasted against the parametric and nonparametric philosophy. The dichotomous selectivity model together with it's censored variant was vested with a formal model in chapter two and delimited against models with more than two categories, truncated and Tobit-type selectivity models. Chapter three laid the groundwork for semiparametric estimation by tackling the identification issues of what can be identified in asymptotic samples under which assumptions. The exclusion restriction and the single index restriction were introduced. Chapter four sketched briefly the parametric alternative and exposed also a lesser known shortcoming of the inverse Mill's ratio. Semiparametric estimation from finite samples was the focus of chapter five. A variety of the most promising semiparametric estimators for the different estimation steps has been illuminated. Chapter six summarized the available properties of these estimators and chapter seven explicated some relevant issues to be considered when putting these estimators to use, together with a recommendation which estimator to apply.

It remains a final assessment of the value of the semiparametric approach within the selectivity model context in comparison to the parametric procedures. Although on theoretical grounds the parametric approach has certain deficiencies, in practical matters it seems to be considerably robust and to work very well in most of the cases. In a recent article Newey (1997,[38]) shows that even in spite of misspecified, inconsistent step one estimates the step two estimates can still be consistent, cited after Vella (1998,[50], p.141): "This is an important result as it allows the estimation of the first step to be conducted under some maintained distributional assumption and the second step estimates remain consistent even if this assumption is violated." Nevertheless, serious inconsistency can occur if the error terms are grossly misspecified. Specification tests may verify the validity of the assumed specification but do not provide any remedy if the particular specification is rejected.

Here the semiparametric approach breaches in and delivers alternative, more general though less efficient, estimates as a reference to unveil a severe inconsistency of the parametric estimates. Thus the dual strategy is recommended, where the semiparametric estimators is attributed the protagonizing role in selecting the more adequate estimates.

It is to remark that so far no sophisticated Monte Carlo studies have been executed, properly comparing parametric and semiparametric estimators' behaviour in finite samples. More research in this respect is appreciated and awaited to permit a careful assessment of both rivaling approaches and to develop guidelines which estimators to apply under which circumstances.

Despite this caution to applying semiparametric methods for verifying the parametric results still and all it should be kept in mind that the validity of semiparametric techniques still depends on the correct specification of their parametric parts, as they are still semi - *parametric* models in contrast to the nonparametric environment. Along with the mentioned marked robustness of the parametric approaches this remark alerts that a sound specification of the functional forms is of higher weight than the handling of the error terms, as Newey, Powell and Walker (1990,[39]) close their article with the words:

> "The results reported here can be viewed as further evidence that specification of the regression function and set of instrumental variables appears to be more important than specification of the error distribution for these data."

Chapter 9

Appendix - Introduction in Kernel Regression

Since nonparametric regression by Kernel and related methods is a vast field only some excerpts can be drafted here and cannot replace their thorough study. For this consult the article of Bierens (1987,[5]) or the textbooks from Härdle (1991,[16]). For a very illustrative and intuitive introduction in nonparametric density estimation the book of Silverman (1986,[49]) be highly recommended. The main distinction between the nonparametric and the parametric philosophy referring to the investigation of a relationship between some observed variables is that the nonparametric approach estimates this relationship, whereas the parametric approach assumes it. Parametric form specification forces some external structure on the data that may blur the true relationship between dependent and explanatory variables and leads to hypocritically innocent and good-looking estimation results that merely reproduce the arbitrary external structure plugged in by the econometrician. To it's advocation it must be said that smartly chosen parametric specifications, however, have been shown to be fairly robust and only negligibly inconsistent within a considerable range of misspecification. Hence for applied work nonparametric regression rather functions as a tool to verify the appropriateness of a particular parametric specification and helps otherwise to select a suitable parametric specification. In this respect nonparametric regression is more helpful and informative than specification tests, since specification tests merely reject or do not a certain specification while nonparametric regression further indicates how the true relationship probably looks like.

Nonparametric regression is the estimation of the true relationship between a random dependent variable Y and a random explanatory vector X on the basis of a sample $\{(Y_i, X_i)\}_{i=1..n}$ and the regression equation $E[Y_i|X_i] = g(X_i) \Rightarrow$

$$Y_i = g(X_i) + \varepsilon_i, \tag{9.1}$$

$E[\varepsilon_i|X_i] = 0$, where neither the function $g(\cdot)$ nor the distribution of the error term

ε_i are anyhow specified.

Nonparametric regression is also called nonparametric smoothing since it originates from the pure data points (Y_i, X_i) and instead of connecting all points by a zigzag curve it constructs a smooth curve "between" these data points to emphasize the underlying pattern. Here already appears the trade-off between *variability* and *bias* that will accompany all aspects of nonparametric methods. A smoothed curve that is very fidel to the available dataset by staying close to the zigzag form has a small bias as it does not depart much from the observed points, that somehow embody the true relationship. Also it does not flatten peaks and troughs too much. However it has a high variability due to it's zigzag shape, which may be hard to interpret and perhaps blurs a simple true relationship. On the other side oversmoothed curves have a low variability but may have lost too much fine structure to reproduce the true relationship, resulting in a higher asymptotic bias. To smoothing such a curve into the chaos of a scatter plot several *local averaging* techniques: Kernel smoothing, adaptive Kernel, nearest neighbours and so forth, have been developed. Here only Kernel smoothing will be covered. How the smoothed curve $\hat{g}(X)$ finally looks like depends then on the Kernel function and the choice of a smoothing parameter.

Beginning with univariate X_i, the smoothing motivation rests on the assumption that the observed data points (Y_i, X_i) represent the true relationship's most likely outcome Y_i for a certain X_i and that the true relationship itself is smooth, that is the dependent variable Y is constant in any infinitely small neighbourhood of X. Then it should be valid to select a pair (Y_i, X_i) and to project Y_i from this X_i into a close neighbourhood around X_i, with this projection becoming less credible the farther the projection is apart from X_i. Such a projection into their corresponding neighbourhoods can be done for all the observations in the sample. Now, for any "point" x in the support of the explanatory variable X all these local projections of Y should be added if they reach thus far to attain the point x. To compute an average for the dependent variable $\hat{y}|x$, corresponding to x, this sum has to be divided by the number of projections that have been summed up. By repeating this procedure for all values x in the support of X the estimated conditional average $\hat{Y}|X$ is established, representing the nonparametric regression estimate $\hat{E}[Y|X] = \hat{g}(X)$.

The outlined proceeding would correspond to a Kernel regression with a bounded uniform Kernel, simply trimming projections that are too far apart. However the deterioration of the projections' credibility with the distance to it's source has not been taken into proper consideration. This can be solved by adjusting the projections' weights in the computation of a *weighted* average of the nearby projections, that are downweighted the farther their origin is apart. Contributing these weights to every projection for the calculation of the weighted average is the task of the Kernel function.

The Kernel function $K(u)$ is a symmetric, continuous and bounded function that integrates up to one, often a density function. It's shape reflects the assignment of the averaging weights according to the projections' distance. With this Kernel weighting function the common Nadaraya-Watson regression estimator looks:

$$\hat{y} = \hat{g}(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \cdot \frac{1}{h} \cdot K\left(\frac{x-X_i}{h}\right)}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} \cdot K\left(\frac{x-X_i}{h}\right)}, \tag{9.2}$$

where for any value x of the explanatory variable the numerator sums up the projections Y_i weighted with regard to their distance $|x - X_i|$ by the Kernel function $K(\cdot)$ and the denominator divides the weighted sum of projections by the sum of the weights to obtain the weighted average. Note that the denominator by itself represents the density estimate of the explanatory variable X. The bandwidth (window width) parameter h plays a crucial role in nonparametric regression as it defines the size of the close neighbourhood for the Kernel function. For h small this neighbourhood is small, only a few projections enter into the reach of x and the regression curve retains it's zigzag shape, preserving the fine-structure but also exhibiting much random noise. For h very large the Kernel downweights or trims only a few distant projections such that the fine structure will be overshadowed by a great many projections of quite distant observations.[1]

Finding the smooth curve that fits best the present data has been the objective of extensive statistical research and depends largely on the bandwidth parameter h and to a lesser extent on the choice of the Kernel function $K(\cdot)$.

Kernel function:

Beginning with the Kernel function for a univariate explanatory variable X. For any symmetric function K to suit as a Kernel it should be bounded and integrate to one:

$$\begin{aligned} &\int_{-\infty}^{\infty} |K(u)|\, du < \infty \\ &\int_{-\infty}^{\infty} K(u) du = 1 \\ &K(-u) = K(u). \end{aligned} \tag{9.3}$$

The usual choice for K is a density function which satisfies above conditions and is non-negative all over it's support. This non-negativity is particularly appealing since it delivers always non-negative Kernel weights and is especially useful for nonparametric density estimation, as a weighted average of densities remains a proper

[1] Adaptive Kernels are a smart tool where the bandwidth h is not hold constant over the whole support of X but adjusted to the density of observations in it's neighbourhood, similar to the k-nearest neighbours technique. In areas where observations are sparse the averaging window is enlargened to smooth the random noise while it is shortened in dense areas to not oversmooth all the fine structure.

density function. To improve asymptotic properties of this nonparametric estimator extensive use of higher-order Kernels has been made. Kernels of order higher than 2 reduce the asymptotic bias and improve the convergence rate, though therewith comes an increase of the asymptotic variance. A Kernel function is of order m if it's first $m-1$ moments are zero. More specific, if $K(\cdot)$ satisfies:

$$\begin{array}{ll} \int_{-\infty}^{\infty} |K(u)|\, du < \infty & \int_{-\infty}^{\infty} |u|^m \cdot |K(u)|\, du < \infty \\ \int_{-\infty}^{\infty} K(u)du = 1 & \int_{-\infty}^{\infty} u^m \cdot K(u)du \neq 0 \\ \int_{-\infty}^{\infty} u^j \cdot K(u)du = 0 \quad \text{for } j = 1, .., m-1 & \end{array} \tag{9.4}$$

the Kernel function is said to be of order m.

The frequent density function Kernel is of order 2, since it integrates to one, has mean zero (first moment) but a positive variance (second moment). As for nonparametric regression commonly symmetric Kernels are employed, that give projections from the left the same weight as projections from the right, only Kernel functions of order $m = 2, 4, 6, 8, ..$ fulfil this quality. Also it should be kept in mind that Kernel functions of order higher than 2 always generate negative Kernel weights for some areas, usually the flanks, of it's support to meet above criteria. This is particularly disturbing for the estimation of densities or probabilities, like the participation probability of the selection equation, that may lead to estimated probabilities below zero or larger than one.

Optimal, in an asymptotic sense, Kernels have been established for different m, like the Epanechnikov for $m = 2$, though the use of an other Kernel, which may be easier to handle or compute, involves with it usually only a minor efficiency loss. Some even Kernel functions are presented exemplary. Odd Kernels of order $m = 3, 5, 7, ..$ are asymmetric and useful for the estimation of odd derivatives, though not contemplated here. Let d denote the dimension of the explanatory vector X.

Univariate ($d = 1$) Kernels of order $m = 2$. (Density Kernels, with non-negative domain):

$$\begin{array}{ll} \text{Epanechnikov} & K(u) = \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}u^2) \cdot 1(|u| \leq \sqrt{5}) \\ \text{Gaussian} & K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \\ \text{Quartic} & K(u) = \frac{15}{16}(1 - u^2)^2 \cdot 1(|u| \leq 1). \end{array} \tag{9.5}$$

Univariate ($d = 1$) Kernels of order $m = 4$. (Non-density Kernels, with their domain including negative Kernel weights):

$$\begin{aligned} K(u) &= \frac{15}{32}(3 - 10u^2 + 7u^4) \cdot 1(|u| \leq 1) \\ K(u) &= \frac{21}{64}(1 - 5\left(\frac{u}{5}\right)^2 + 7\left(\frac{u}{5}\right)^4 - 3\left(\frac{u}{5}\right)^6) \cdot 1(|u| \leq 5) \end{aligned} \tag{9.6}$$

Also the jackknife estimator $\check{g}(X)$ that is the weighted average of two order-2 Kernels with different bandwidths, generates Kernels of order of at least 4, see Powell (1989,[41]):

$$K(u) = \frac{a^2}{a^2-1}\cdot\left[k(u) - a\cdot k(\frac{u}{a})\right] \qquad \text{for any positive } a \neq 1 \text{ and univariate Kernel } k(\cdot) \tag{9.7}$$

Multivariate ($d > 1$) Kernels can be constructed as the product of univariate Kernels. For a vector $U = (u_1, .., u_d)'$ the multivariate product Kernel K for any univariate Kernel $k(\cdot)$ becomes:

$$K(U) = \prod_{j=1}^{d} k(u_j). \tag{9.8}$$

Another multivariate Kernel is the multivariate standard normal density (order $m = 2$):

$$K(U) = \frac{1}{\sqrt{2\pi}^d}\exp\left(-\frac{1}{2}U'U\right). \tag{9.9}$$

Bierens (p.112,125) presented a method to construct a Kernel of arbitrary dimension d and arbitrary order $m \geq 4$ as:

$$K(U) = \sum_{j=1}^{m/2} \frac{\theta_j \exp\left[-\frac{1}{2}U'\hat{V}^{-1}U/\sigma_j^2\right]}{\sqrt{2\pi}^d \cdot |\sigma_j|^d \sqrt{\det \hat{V}}} \tag{9.10}$$

with $\hat{V}$ estimated as the sample variance and θ_j, σ_j chosen to fulfil the moment conditions:

$$\begin{aligned} &\hat{V} = \tfrac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})' \\ &\sum_{j=1}^{m/2}\theta_j = 1 \\ &\sum_{j=1}^{m/2}\theta_j\sigma_j^{2l} = 0 \qquad \text{for } l = 1, 2, .., \tfrac{m-2}{2} \end{aligned} \tag{9.11}$$

Bandwidth parameter:

To attain asymptotically unbiased (=consistent) regression results the introduced smoothing bias must converge to zero, though the asymptotic bias (of the stabilized transformation generally does not). Since this bias depends on the degree of smoothing and is lower for short-sighted local averaging that retains a more zigzag form of the smoothed regression curve than for far-sighted averaging that misses all the fine structure through oversmoothing, the window width must shorten once sufficient observations' projections are within reach. Hence h must tend to zero the denser the observations get, though only that fast that always a sufficient amount of observations are in it's neighbourhood to avoid spurious regression peaks through single

outliers that gain too much weight. As a consequence the bandwidth h should be a function of the sample size n and for consistency of the regression must satisfy:

$$\lim_{n\to\infty} h = 0 \qquad\qquad \lim_{n\to\infty} nh^d = \infty, \tag{9.12}$$

with d the dimension of the explanatory vector X, for univariate X here $r = 1$.

To achieve asymptotic normality and then furthermore the maximum rate of convergence the bandwidth has to suit certain proportions to the sample size n. With this optimal bandwidth $h_{opt} = h_{opt}(n)$ as a function of the sample size the stabilized transformation of the nonparametric estimate $\hat{g}(X)$ attains: (Bierens, p.109)

$$n^{\frac{2}{d+4}} \left(\hat{g}(x) - g(x)\right) \xrightarrow{d} N(B, V) \tag{9.13}$$

with B the asymptotic bias and V the asymptotic variance. Hence the nonparametric estimate is biased and it's bias converges to zero at a rate $\frac{2}{d+4}$. Also the "curse of dimensionality" becomes apparent since the convergence rate deteriorates with the dimension d of the explanatory vector X.

The convergence rate can be improved and the bias reduced by picking a Kernel of higher order, $m = 2, 4, 6, ...$ With Bierens (p.112) the convergence rate will be

$$\frac{m}{2m + d}. \tag{9.14}$$

Hence the convergence comes closer to $\sqrt{n}$-rate with the order m of the Kernel applied, at the same time the asymptotic variance increases with a higher order Kernel. Bierens further covers the case that all variables in the explanatory vector X are discrete or integer, which happens quite often in microeconometrical applications. Provided the optimal bandwidth has been chosen the appropriate Kernel estimator will be asymptotically unbiased, asymptotically normal and converges even with $\sqrt{n}$-rate, see Bierens (p.117).

All these results hinge on the optimal choice of the bandwidth parameter h, that is of much higher importance than the Kernel choice. In practical work however this asymptotic theory helps little to determine the optimal bandwidth h_{opt}. Lower and upper bounds for the path of the optimal bandwidth with sample size n are useless if only one sample is present. Hence automatic bandwidth-selection algorithms for a given finite sample have been longed for, with the generalized cross-validation the most well-known candidate, discussed in Härdle (1991,[16]) or Craven and Wahba (1979,[11]). Cross-validation, nicknamed as leave-one-out method, founds itself on the motivation that picking a data point (Y_j, X_j) out of the sample and deleting this data point temporarily from the sample, a regression estimate can be calculated from the remaining sample $\{(Y_i, X_i)\}_{i\neq j, i=1..n}$ of sample size $n-1$, that is the whole sample *without* this observation (Y_j, X_j), for any arbitrarily chosen bandwidth h. This regression estimate $\hat{g}_j(X)$ can now be used to predict $\hat{Y}_j = \hat{g}_j(X_j)$ and compared

with Y_j and the goodness of prediction depends on the chosen bandwidth. This prediction can iteratively be computed for all observations Y_j, $j = 1..n$ on basis of their corresponding leave-one-out $n-1$ subsamples. Thereby for every observation (Y_j, X_j) Y_j itself is predicted by making use of all other observations except this observation itself. The motivation is that for a properly chosen bandwidth either part of the sample should be able to predict the other since both are realizations of the same underlying process. Once all predictions $\hat{Y}_j$ have been supplied the overall goodness of prediction of Y is measured for the whole sample and the whole procedure repeated over a grid of different bandwidths h. As data-driven optimal bandwidth will be picked that h that delivered the best prediction.

Following Härdle, the bandwidth is chosen to minimize a mean squared error criterion, that is the sum $MSE = bias^2 + variance$, approximated by the average squared prediction error. Be $\{(Y_i, X_i)\}_{i=1..n}$ the sample, the regression function can be estimated as the leave-one-out regression estimate $\hat{g}_j(X;h)$ for any $j \in \{1,..,n\}$ and any positive bandwidth h from the rest sample $\{(Y_i, X_i)\}_{i\neq j, i=1..n}$ as:

$$\hat{g}_j(x;h) = \frac{\frac{1}{n-1}\sum\limits_{i\neq j}^{n} Y_i \cdot \frac{1}{h} \cdot K\left(\frac{x-X_i}{h}\right)}{\frac{1}{n-1}\sum\limits_{i\neq j}^{n} \frac{1}{h} \cdot K\left(\frac{x-X_i}{h}\right)}. \tag{9.15}$$

Y_j can be predicted from this leave-one-out regression estimate as $\hat{Y}_j = \hat{g}_j(X_j;h)$. This prediction will be performed iteratively for all $j = 1..n$ leave-one-out regression estimates $\hat{g}_j(X;h)$ and the average squared prediction error be calculated, adjusted by a non-negative weighting function $w(x)$, for instance $w(x) = 1\left(\left|x - \frac{1}{2}\right| < 0.4\right)$, that trims or downweights observations which are close to the margin of the support of X within the sample.[2] The result is the cross-validation statistic for a particular bandwidth h. Minimizing this adjusted average squared prediction error over all admissible bandwidth parameter values, commonly executed by a grid search with unidistant grid points over a suitable range for h, yields the cross-validated optimal bandwidth h_{CV} that converges, though slowly, to the asymptotically optimal bandwidth h_{opt}, (Härdle, p.157ff):

$$\min_h CV(h) = \frac{1}{n}\sum_{j=1}^{n} \left(Y_j - \hat{g}_j(X_j;h)\right)^2 \cdot w(X_j). \tag{9.16}$$

Although this automatic bandwidth-selection algorithm delivers in most cases good results a subsequent visual inspection of the resulting regression curve with h_{CV} is strongly advised. Other bandwidth-selection procedures are covered in Härdle.

[2] This is recommended since the Kernel averaging procedure becomes asymmetric at the boundary of the support of X within the sample where observations are missing either to the left or to the right. Although adaptations exist that take this vacuum into consideration the regression estimate is less reliable in these frontier areas.

Multivariate nonparametric regression:

For the multivariate case, $\dim(X) > 1$, the basic ideas stay the same, just making use of a multivariate Kernel. The local averaging procedure proceeds now over a multidimensional neighbourhood and the neighbourhood size is defined by different bandwidth parameters for each dimension. (Though many algorithms ignore this). But swiftly the curse of dimensionality affects the estimate since with higher dimension a given number of observations becomes quickly sparse and isolated within the enormously enlargened space. How fast observations become rare and how the sample size must grow with higher dimension to recover the same accuracy illustrates Silverman (p.92ff) for the related field of nonparametric density estimation. For the univariate case and X standard normally distributed nearly 90% of the distribution mass lies within a range of $[-1.6, 1.6]$ around the origin whereas this figure is only 1% for X of dimension 10 with all the rest scattered in the tails. To achieve the same accuracy as for the univariate nonparametric density estimation the sample must be about 55 times larger for X of dimension 4 and about 210,000 times larger for $\dim(X) = 10$. These figures unambiguously unveil that proper nonparametric applications are condemned to low dimensional settings.

Chapter 10

Appendix - Variance Matrices

Powell 1989

The asymptotic covariance matrix V_{Powell} of Powell's step two estimator can be computed as follows:

$$\hat{V}_{Powell} = \hat{A}_{WX}^{-1}\left[\hat{C}_{\varsigma\varsigma} - \hat{B}_{WZ}\hat{C}_{\varsigma\psi} - \hat{C}_{\varsigma\psi}\hat{B}'_{WZ} + \hat{B}_{WZ}\hat{C}_{\psi\psi}\hat{B}'_{WZ}\right]\hat{A}_{WX}^{-1}, \tag{10.1}$$

where the $\hat{C}_{\varsigma\varsigma}$, $\hat{C}_{\psi\psi}$ terms denote the sample variances of the series $\hat{\varsigma}_i$ and $\hat{\psi}_i$ respectively, and $\hat{C}_{\varsigma\psi}$ the sample covariance between these both series:

$$\hat{C}_{\varsigma\psi} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varsigma}_i\hat{\psi}_i, \tag{10.2}$$

with $\hat{C}_{\varsigma}$, $\hat{C}_{\psi\psi}$ analog. Whereas the series $\hat{\varsigma}_i$ can be constructed from this estimator, the series $\hat{\psi}_i$ relies on the preliminary estimator $\hat{\gamma}$ which is supposed to conform to an asymptotically linear representation

$$\hat{\gamma} = \gamma + \frac{1}{n}\sum_{i=1}^{n}\psi(d_i, Z_i, \gamma) + o_p(n^{-\frac{1}{2}}) \tag{10.3}$$

(see Powell (1989,[41]) Corollary 5.1, p.24), as do most of the $\sqrt{n}$-consistent discrete choice estimators, for instance the estimator of Powell, Stock and Stoker (1989,[43]). Given such a series $\hat{\psi}_i$ is available, therewith the remaining parts are computed straight ahead as:

$$\hat{\varsigma}_i = \frac{2}{n-1}\sum_{j=1,j\neq i}^{n}\frac{1}{h}K\left(\frac{(Z_i - Z_j)'\hat{\gamma}}{h}\right)(W_i - W_j)(\hat{v}_i - \hat{v}_j), \tag{10.4}$$

$$\hat{A}_{WX} = \binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\hat{\varpi}_{ijn}\cdot(W_i - W_j)(X_i - X_j), \tag{10.5}$$

$$\hat{B}_{WZ} = \binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\hat{v}_{ijn}\cdot(\hat{v}_i - \hat{v}_j)(W_i - W_j)(Z_i - Z_j), \tag{10.6}$$

$$\hat{v}_{ijn} = \frac{1}{h^2} K^{(1)} \left(\frac{(Z_i - Z_j)'\hat{\gamma}}{h} \right), \qquad \hat{v}_i = Y_i - X_i'\hat{\beta} \tag{10.7}$$

with $K^{(i)}$ denoting the Kernel functions i^{th} derivative.

Newey Series 1988

Define

$$\begin{aligned} \hat{w}_i &= (X_i', \hat{p}_{1i}, \hat{p}_{2i}, .., \hat{p}_{Ki})' \\ \hat{W} &= (\hat{w}_1, \hat{w}_2, .., \hat{w}_n)' \\ \hat{\Theta} &= (\hat{\beta}', \hat{\eta}_1, \hat{\eta}_2, .., \hat{\eta}_K)'. \end{aligned} \tag{10.8}$$

With this definitions the choice of K is guided by cross-validation and K should be chosen to minimize $CV(K)$:

$$\begin{aligned} &\min_K CV(K) = \sum_{i=1}^{n} \left[\frac{(1-2\hat{\delta}_i)\cdot \hat{e}_i}{1-\hat{\delta}_i} \right]^2 \\ &\hat{\delta}_i = \hat{w}_i'(\hat{W}'\hat{W})^{-1}\hat{w}_i \\ &\hat{e}_i = Y_i - \hat{w}_i'\hat{\Theta}. \end{aligned} \tag{10.9}$$

Also the asymptotic variance matrix of Newey's series estimator is estimated as:

$$\begin{aligned} \hat{V}_{Newey\_Series} &= [I_p, 0] \left(\frac{\hat{W}'\hat{W}}{n} \right)^{-1} A \left(\frac{\hat{W}'\hat{W}}{n} \right)^{-1} [I_p, 0]' \\ A &= \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i \hat{w}_i' \cdot (Y_i - \hat{w}_i'\hat{\Theta})^2 + \hat{H}\hat{V}[\hat{\gamma}]\hat{H}' \\ \hat{H} &= \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i \cdot \left[\frac{d\left(\sum_{k=1}^{K} \hat{\eta}_k \cdot p_k(Z_i'\hat{\gamma}) \right)}{d(Z_i'\gamma)} \right] \cdot \left[\frac{d(Z_i'\hat{\gamma})}{d\gamma'} \right], \end{aligned} \tag{10.10}$$

where I_p is the identity matrix of dimension p and $\hat{V}[\hat{\gamma}]$ is the consistent variance estimate of the applied step one estimator $\hat{\gamma}$. Newey remarks that this variance $\hat{V}[\hat{\beta}]$ is a sum of two components, the "White-heteroskedasticity corrected covariance matrix" and a term inherited form the estimation of γ. Here can be seen that the accuracy of the series estimator relies on the precision of the selection equation estimation.

Newey GMM 1988

The $\tilde{\beta}_{GMM}$ of Newey is computed as follows. Recall that $\hat{p}_i(Z_i'\hat{\gamma})$ has been defined as the series approximation in column vector form

$$\hat{p}_i = (p_1(Z_i'\hat{\gamma}), .., p_K(Z_i'\hat{\gamma}))'. \tag{10.11}$$

Let $\hat{p} = (\hat{p}_1, .., \hat{p}_n)'$ be the $n \times k$ matrix including all the participation subsample observations, $m(\varepsilon_i)$ be the column vector of all the J moment functions

$$m(\varepsilon_i) = (m_1(\varepsilon_i), .., m_J(\varepsilon_i))' \tag{10.12}$$

and $\hat{a}_i$ the column vector of the "supposedly uncorrelated" functions

$$\hat{a}_i = (a_1(Z_i'\hat{\gamma}), .., a_L(Z_i'\hat{\gamma}))'. \tag{10.13}$$

It should be remembered that efficiency hinges on a suitable choice of the functions $m_j(\varepsilon_i)$ and $a_l(Z_i'\hat{\gamma})$, which have been recommended to follow the form $m_j(\varepsilon_i) = m_0(\varepsilon_i)^j$ and $a_l(Z_i'\hat{\gamma}) = a_0(Z_i'\hat{\gamma})^{l-1}$ with the root functions, for instance, $m_0(\cdot) = a_0(\cdot) = 2\Phi(\cdot) - 1$.

The estimator $\tilde{\beta}_{GMM}$ bases on the Newey series estimator $\hat{\beta}_{Series}$, that has to be performed beforehand to obtain initial starting values. Then the following expressions have to be computed:

$$\hat{Z}_{iK} = \left[\hat{p}_i'(\hat{p}'\hat{p})^{-1}\sum_{i=1}^{n}\hat{p}_i Z_i'\right]' \tag{10.14}$$

$$\hat{m}_i = m(Y_i - X_i'\,\hat{\beta}) \tag{10.15}$$

$$\hat{m}_{iK} = \left[\hat{p}_i'(\hat{p}'\hat{p})^{-1}\sum_{i=1}^{n}\hat{p}_i\hat{m}_i'\right]' \tag{10.16}$$

$$\hat{m}_{i\beta} = -\left[\frac{dm(Y_i - X_i'\,\hat{\beta})}{d(Y_i - X_i'\,\hat{\beta})}\right]\cdot X_i' \tag{10.17}$$

$$\hat{m}_{i\beta K} = \left[\hat{p}_i'(\hat{p}'\hat{p})^{-1}\sum_{i=1}^{n}\hat{p}_i\left(\frac{dm(Y_i - X_i'\,\hat{\beta})}{d(Y_i - X_i'\,\hat{\beta})}\right)'\right]'\cdot X_i' \tag{10.18}$$

$$\hat{g}_i = (\hat{m}_i - \hat{m}_{iK})\otimes\hat{a}_i\otimes(Z_i - \hat{Z}_{iK}) \tag{10.19}$$

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{d\hat{p}_i(Z_i'\hat{\gamma})}{d(Z_i'\gamma)}\right)'(\hat{p}'\hat{p})^{-1}\sum_{i=1}^{n}\hat{p}_i\hat{m}_i'\right\}'\otimes\hat{a}_i\otimes(Z_i - \hat{Z}_{iK}) \tag{10.20}$$

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_i\hat{g}_i' + \hat{H}\hat{V}(\hat{\gamma})\hat{H}' \tag{10.21}$$

$$\hat{G} = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}_{i\beta} - \hat{m}_{i\beta K})\otimes\hat{a}_i\otimes(Z_i - \hat{Z}_{iK}) \tag{10.22}$$

The estimator $\tilde{\beta}_{GMM}$ is defined as

$$\tilde{\beta}_{GMM} = \hat{\beta}_{Series} - (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}\cdot\hat{G}'\hat{\Omega}^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{g}_i \tag{10.23}$$

with it's variance V_{GMM} estimated as

$$\hat{V}_{GMM} = (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1} \tag{10.24}$$

Ahn/Powell 1993

To not confusing the first and the second step the estimation process is briefly recalled. Since the variance depends on the merger of both estimation steps the asymptotic variance estimate is computed over the full sample $1..N$. To conform with the Powell estimator, that performs over the participation subsample $1..n$, the weights $\hat{\varpi}_{ij}$ in the structural equation estimation will be set to zero for all non-participants ($d_i = 0$).

In the first step the selection equation is estimated over the full sample $1..N$ to obtain the propensity score $\hat{\theta}_i = \hat{E}[d_i|Z_i]$ by a multivariate Kernel $K_1(\cdot)$ implying Kernel weights K_{ij}:

$$K_{ij} = \frac{1}{h_1} K_1\left(\frac{Z_i - Z_j}{h_1}\right) \tag{10.25}$$

$$\hat{\theta}(Z_i) = \frac{\frac{1}{N}\sum_{j=1}^{N} d_j \cdot K_{ij}}{\frac{1}{N}\sum_{j=1}^{N} K_{ij}}. \tag{10.26}$$

These estimated propensity scores can be outside their logical range $[0, 1]$ due to the use of a higher-order Kernel.

In the second step, basing on Powell 1989, the structural equations are pairwise compared:

$$Y_i - Y_j = (X_i - X_j)'\beta + \left[\lambda(\hat{\theta}_i) - \lambda(\hat{\theta}_j)\right] + (\xi_i - \xi_j) \tag{10.27}$$

For observations with close indices $\hat{\theta}_i, \hat{\theta}_j$ their selection correction terms cancel out and β is estimated by instrumental variable regression with $W_i = W(Z_i)$ as instruments. Pairwise Kernel weights $\hat{\varpi}_{ij}$, which are only non-zero if both i and j are participants ($d = 1$) and depend then on the indices' distance from each other, are obtained by an univariate Kernel $k_2(\cdot)$ as:

$$\hat{\varpi}_{ij} = \frac{1}{h_2} k_2\left(\frac{\hat{\theta}_i - \hat{\theta}_j}{h_2}\right) \cdot d_i \cdot d_j. \tag{10.28}$$

Now β is estimated by the estimator of Powell. Though formally over the full sample $1..N$, effectively only the participation subsample $1..n$ is used since for all non-participants their multiplicative Kernel weights $\hat{\varpi}_{ij}$ are zero.

$$\hat{\beta} = \hat{S}_{WX}^{-1} \cdot \hat{S}_{WY} \tag{10.29}$$

$$\begin{aligned}\hat{S}_{WX} &= \binom{N}{2}^{-1} \sum_{i=1}^{N}\sum_{j=i+1}^{N} \hat{\varpi}_{ij} \cdot (W_i - W_j)(X_i - X_j)' \\ \hat{S}_{WY} &= \binom{N}{2}^{-1} \sum_{i=1}^{N}\sum_{j=i+1}^{N} \hat{\varpi}_{ij} \cdot (W_i - W_j)(Y_i - Y_j).\end{aligned} \tag{10.30}$$

With all this groundwork prepared the asymptotic variance of Ahn/Powell can be calculated over the full sample $1..N$:

$$\hat{V}_{Ahn/Powell} = \frac{1}{4}(\hat{S}_{WX})^{-1} \cdot \hat{W} \cdot ((\hat{S}_{WX})^{-1})', \tag{10.31}$$

where the matrix $\hat{S}_{WX}$ matrix stems from the step two Powell estimator β, equation (10.30). The matrix $\hat{W}$ is computed:

$$\hat{W} = \frac{1}{2N}\sum_{i=1}^{N}\left[\hat{\psi}_i + \hat{\zeta}_i\hat{e}_i\right]\left[\hat{\psi}_i + \hat{\zeta}_i\hat{e}_i\right]' \tag{10.32}$$

$$\hat{\psi}_i = \frac{1}{N-1}\sum_{j=1}^{N}\hat{\varpi}_{ij} \cdot (\hat{v}_i - \hat{v}_j) \cdot (W_i - W_j), \tag{10.33}$$

where $\hat{e}_i$ and $\hat{v}_i$ are the residuals of the step one and step two estimation, respectively:

$$\begin{aligned}\hat{e}_i &= d_i - \hat{\theta}_i \\ \hat{v}_i &= Y_i - X_i'\beta.\end{aligned} \tag{10.34}$$

Furthermore

$$\begin{aligned}\hat{\zeta}_i &= \frac{1}{N-1}\sum_{j=1}^{N}\sum_{l=1}^{N}\alpha_{jl} \\ \alpha_{jl} &= \left(\frac{1}{h_2}\right)^2 k_2'\left(\frac{\hat{\theta}_i - \hat{\theta}_j}{h_2}\right) \cdot d_j \cdot d_l \\ &\quad \times K_{jl} \cdot (\hat{v}_j - \hat{v}_l) \cdot (W_j - W_l) \cdot \left[\sum_{d=1}^{N} K_{jd}\right]^{-1}.\end{aligned} \tag{10.35}$$

k_2' is the first derivative of the univariate Kernel k_2 that has been employed in the step two estimation and K_{jl} are the Kernel weights of the step one.

Robinson 1988

$$\hat{V}_{Robinson} = \hat{\sigma}^2 \cdot S_{X-\hat{X}}^{-1} \tag{10.36}$$

$$S_{X-\hat{X}}^{-1} = \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{E}[X_i|Z_i]) \cdot (X_i - \hat{E}[X_i|Z_i])'\right]^{-1} \tag{10.37}$$

$$\begin{aligned}\hat{\sigma}^2 &= S_{Y-\hat{Y}-(X-\hat{X})'\hat{\beta}} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{E}[Y_i|Z_i] - (X_i - \hat{E}[X_i|Z_i])'\hat{\beta}\right)^2.\end{aligned} \tag{10.38}$$

Andrews/Schafgans 1996 (intercept estimator)

The variance of the intercept estimate for the [SIR] single index assumption is computed as

$$\hat{V}_{Andrews/Schafgans} = \frac{\hat{V}_\varepsilon \cdot \frac{1}{n} \sum_{i=1}^{n} s^2(Z_i'\hat{\gamma} - b')}{\left[\frac{1}{n} \sum_{i=1}^{n} s(Z_i'\hat{\gamma} - b')\right]^2}, \tag{10.39}$$

with

$$\hat{V}_\varepsilon = \frac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - X_i'\hat{\beta})^2 \cdot s^2(Z_i'\hat{\gamma} - b')}{\sum_{i=1}^{n} s^2(Z_i'\hat{\gamma} - b')}. \tag{10.40}$$

Bibliography

[1] Ahn, H., and James Powell (1993), "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism", *Journal of Econometrics*, 58, 3-29.

[2] Andrews, Donald, and Marcia Schafgans (1996), "Semiparametric Estimation of a Sample Selection Model", Cowles Foundation Discussion Paper, No. 1119, Yale University.

[3] ——— (1997), "Semiparametric Estimation of the Intercept of a Sample Selection Model", Forthcoming, *Review of Economic Studies* (1998).

[4] Arabmazar, A., and P. Schmidt (1982), "An Investigation of the Robustness of the Tobit Estimator to Non-Normality", *Econometrica*, 50, 1055-1063.

[5] Bierens, Herman J. (1987), "Kernel Estimators of Regression Functions", in *Advances in Econometrics: Fifth World Congress, Vol. 1*, 99-144, ed. Truman F. Bewley, Cambridge: Cambridge University Press.

[6] Chamberlain, G. (1986), "Asymptotic Efficiency in Semiparametric Models with Censoring", *Journal of Econometrics*, 32, 189-218.

[7] Charlier, E., B. Melenberg, and A. van Soest (1997), "An Analysis of Housing Expenditure Using Semiparametric Cross-Section Models", unpublished paper, Department of Econometrics and CentER, Tilburg University.

[8] Chen, S. (1996), "Distribution-free Estimation of the Random Coefficient Dummy Endogenous Variable Model", unpublished paper, Hong Kong University of Science and Technology.

[9] Chesher, A., and M. Irish (1987), "Residual Analysis in the Grouped and Censored Normal Linear Model", *Journal of Econometrics*, 34, 33-61.

[10] Cosslett, Stephen R. (1991), "Semiparametric Estimation of a Regression Model with Sample Selectivity", in *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, eds. William A. Barnett, James Powell, and George Tauchen, Cambridge: Cambridge University Press.

[11] Craven, P., G. and Wahba (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation", *Numerical Mathematics*, 31, 377-403.

[12] Gabler, S., F. Laisney, and M. Lechner (1993), "Seminonparametric Estimation of Binary-Choice Models with an Application to Labor-Force Participation", *Journal of Business and Economic Statistics*, 11, 61-80.

[13] Gallant, Ronald, and Douglas Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation", *Econometrica*, 55, 363-390.

[14] Gerfin, Michael (1996), "Parametric and Semi-Parametric Estimation of the Binary Response Model of Labour Market Participation", *Journal of Applied Econometrics*, 11, 321-339.

[15] Goldberger, A.S. (1983), "Abnormal Selection Bias", in *Studies in Econometrics, Time Series and Multivariate Statistics*, eds. S. Karlin, T. Amemiya, and L.A. Goodman, New York: Wiley.

[16] Härdle, W. (1991), *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

[17] Heckman, James (1974), "Shadow Prices, Market Wages and Labor Supply", *Econometrica*, 42, 679-694.

[18] ——— (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models", *Annals of Economic and Social Measurement*, 5, 475-492.

[19] ——— (1990), "Varieties of Selection Bias", *American Economic Review, Papers and Proceedings*, 80, 313-318.

[20] Horowitz, Joel L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, 505-531.

[21] ——— (1993), "Semiparametric Estimation of a Work-Trip Mode Choice Model", *Journal of Econometrics*, 58, 49-70.

[22] Ichimura, Hidehiko, and Lung-Fei Lee (1991), "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation", in *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, eds. William A. Barnett, James Powell, and George Tauchen, Cambridge: Cambridge University Press.

[23] Ichimura, Hidehiko (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models", *Journal of Econometrics*, 58, 71-120.

[24] Klein, Roger, and Richard Spady (1993), "An Efficient Semiparametric Estimator of the Binary Response Model", *Econometrica*, 61, 387-423.

[25] Lanot, Gauthier, and Ian Walker (1998), "The Union/Non-Union Wage Differential: An Application of Semi-parametric Methods", *Journal of Econometrics*, 84, 327-349.

[26] Lee, Lung-Fei (1982), "Some Approaches to the Correction of Selectivity Bias", *Review of Economic Studies*, 49, 355-372.

[27] ——— (1994), "Semiparametric Two-Stage Estimation of Sample Selection Models Subject to Tobit-Type Selection Rules", *Journal of Econometrics*, 61(2), 305-344.

[28] Leung, Siu-Fai, and Shihti Yu (1996), "On the Choice between Sample Selection and Two-Part Models", *Journal of Econometrics*, 72, 107-128.

[29] Manski, Charles (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205-228.

[30] ——— (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator", *Journal of Econometrics*, 27, 313-334.

[31] ——— (1989), "Anatomy of the Selection Problem", *Journal of Human Resources*, 24(3), 343-360.

[32] ——— (1993), "The Selection Problem in Econometrics and Statistics", in *Handbook of Statistics, Vol. 11*, eds. G.S. Maddala, C.R. Rao, and H.D. Vinod, Elsevier Science Publishers.

[33] Melenberg, Bertrand, and Arthur van Soest (1993), "Semi-Parametric Estimation of the Sample Selection Model", Tilburg University, CentER for Economic Research, Discussion Paper No. 9334.

[34] ——— (1996), "Parametric and Semi-parametric Modelling of Vacation Expenditures", *Journal of Applied Econometrics*, 11(1), 59-76.

[35] Moretti, Enrico (1997), "Do Wages Compensate for Risk of Unemployment? Parametric and Semiparametric Evidence from Seasonal Jobs", unpublished paper, Department of Economics, University of California at Berkely.

[36] Nawata, K. (1993), "A Note on the Estimation of Models with Sample Selection-Biases", *Economics Letters*, 42, 15-24.

[37] Newey, Whitney (1988), "Two Step Series Estimation of Sample Selection Models", unpublished paper, Princeton University.

[38] ——— (1997), "Consistency of Two-Step Sample Selection Estimator Despite Misspecification of Distribution", unpublished paper.

[39] Newey, Whitney, and James Powell, James Walker (1990), "Semiparametric Estimation of Selection Models: Some Empirical Results", *American Economic Review, Papers and Proceedings*, 80(2), 324-328.

[40] Newey, Whitney, and James Powell (1993), "Efficiency Bounds for some Semiparametric Selection Models", *Journal of Econometrics*, 58(1/2), 169-184.

[41] Powell, James (1989), "Semiparametric Estimation of Censored Selection Models", Department of Economics, unpublished paper, University of Wisconsin-Madison.

[42] ——— (1994), "Estimation of Semiparametric Models", in *Handbook of Econometrics, Vol. 4*, 2444-2523, eds. R.F. Engle, and D.L. McFadden, Amsterdam: North-Holland.

[43] Powell, James, J.H. Stock, and T.M. Stoker (1989), "Semiparametric Estimation of Index Coefficients", *Econometrica*, 57, 1403-1430.

[44] Powell, James, and T.M. Stoker (1996), "Optimal Bandwidth Choice for Density-Weighted Averages", *Journal of Econometrics*, 75, 291-316.

[45] Robinson, Peter (1988), "Root-N-Consistent Semiparametric Regression", *Econometrica*, 56, 931-954.

[46] Schafgans, Marcia (1997), "Semiparametric Estimation of a Sample Selection Model: A Simulation Study", Sticerd Discussion Paper, No. EM/97/326, London School of Economics.

[47] ——— (1998), "A Monte Carlo Study of the Semiparametric Estimation of the Intercept of a Sample Selection Model", unpublished paper, Department of Economics, London School of Economics.

[48] Schafgans, Marcia, and V. Zinde-Walsh (1998), "On Intercept Estimation in the Sample Selection Model", unpublished paper, Department of Economics, London School of Economics.

[49] Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[50] Vella, Francis (1998), "Estimating Models with Sample Selection Bias: A Survey", *Journal of Human Resources*, 33(1), 127-169.