Treatment Choice
based on semiparametric evaluation methods

Markus Frölich, SIAW, Universität St. Gallen,
markus.froelich@unisg.ch
www.siaw.unisg.ch/lechner/froelich
Dufourstr. 48, CH-9000 St. Gallen, Switzerland

Preliminary and incomplete !
Comments are highly appreciated.

Last changes: July 25th, 2000

Abstract:

This paper proposes a new semiparametric estimator to determine the optimal programme for an individual, who faces the decision problem to choose exactly one out of a variety of available programmes. In a first step hypothetical outcomes are predicted for this individual on the basis of realised outcomes of past programme participants. While nonparametric estimation of average potential outcomes for various subpopulations is standard in the evaluation literature, estimating individual potential outcomes conditional on a high-dimensional explanatory vector usually requires a parametric specification. The proposed estimator combines a parametric specification of the conditional outcomes with nonparametrically estimated average outcomes within the GMM framework. This allows to test whether the model is correctly specified and shall improve estimation when misspecified. In a second step the programme attaining highest utility is determined. Finally the estimator is applied to Swedish rehabilitation programmes.

Keywords: Targeting programmes, profiling, programme evaluation, matching, treatment effect, conditional independence assumption

JEL classification: C44, C14, J60, J24

I am grateful to Bo Honoré, Francois Laisney, Michael Lechner, Christof Luchsinger, Ruth Miquel, and Jeff Smith for valuable discussions, comments, and suggestions. Furthermore, I thank Joachim Inkmann and Vassilis Hajivassiliou for providing me with Gauss code to verify of my own routines of the GHK simulator.

1 Introduction

This paper deals with programme selection or treatment choice[1], where individual *optimal-programme* recommendations are derived on basis of observed outcomes of past participants, which are contaminated with selection bias due to non-random programme selection. Consider the situation where an individual has to participate in exactly one of R mutually exclusive and exhaustive programmes, which usually include a 'no-programme' alternative. Either the individual herself chooses the programme or is assigned to a certain programme by another individual or institution. In any case the decision maker wants to select that programme which maximizes after-treatment utility.[2] To choose the most adequate programme he needs to predict ex ante for all R programmes the hypothetical outcomes, which she would acquire would she participate in this respective programme. A natural way to predict these hypothetical outcomes would be based on the realised outcomes of former programme participants. However, in the absence of experimental data, it must be taken into account that individuals are usually not randomly selected into the programmes.

Targeted programme assignment has received considerable attention in recent years for instance with respect to active labour market policies. In some countries, e.g. Australia, Canada, Netherlands, USA, statistical models are used to assign unemployed persons to programmes like job search assistance, training, and employment programmes. In most cases these models intend to predict the probability of becoming long-term unemployed on basis of individual and regional characteristics and allocate the individuals with highest long-term unemployment risk to the most intensive programmes[3]. Implicit is the assumption that the long-term unemployed or those likely going to become are those who gain most from participation in more intensive programmes.[4] However, these models are often rather ad-hoc and usually rely on parametric specifications and few explanatory variables. Berger, Black, and Smith (2000) criticise the weak explanatory power of the operating unemployment profiling systems in the USA in predicting unemployment duration. They show that relevant information for predicting unemployment duration, e.g. employment histories, is neglected in most systems.[5]

The necessity of predicting treatment effects on an individual basis has for instance been indicated in Heckman, Smith, and Clements (1997), which found considerable treatment effect heterogeneity among individuals in the Job Training Partnership Act programme (JTPA, USA) and rejected the assumption of a constant treatment effect. While clearly beneficial to some participants the JTPA programme appeared to be harmful to other participants, in the sense that non-participation would have been more advantageously to them. Also Black, Smith, Berger, and Noel (1999) reject the constant treatment effect assumption with respect to worker profiling services in Kentucky.

[1] The words progamme and treatment are used synonymously throughout the text.

[2] The focus of this paper is entirely on outcomes. Utility during programme participation is neglected.

[3] Exceptions are the Service and Outcome Measurement System (SOMS) in Canada and the Frontline Decision Support System (FDSS) in the USA, which directly predict hyptothetical outcomes on an individual basis to propose a suited programme. For details consult: Australia (OECD 1998), Canada (Colpitts 1999), Netherlands (de Koning 1999), USA: Worker Profiling and Reemployment Services (DOL 1999, Black, Smith, Berger, and Noel 1999), Welfare-to-Work (Eberts 1998), FDSS (Eberts and O'Leary 1999).

[4] This does not hold, for instance, for the worker profiling system in the USA (Black, Smith, Berger, and Noel 1999, Berger, Black, and Smith 2000), where those in the middle ranges of the profiling score gain most and the treatment effect becomes even negative for individuals with high long-term unemployment risk. Obviously, it is also assumed that the programmes are beneficial to the participants, which often is at least doubtful as many evaluation studies show, that sometimes even find negative treatment effects (see e.g. Fay (1996), Gerfin and Lechner (2000), Lechner (2000), Puhani (1999)).

[5] For instance, the profiling system of Pennsylvania relies on only 8 explanatory variables to predict the probability of UI benefit exhaustion and does not even contain race, age, and gender, as prohibited by law (O'Leary, Decker, and Wandner 1998).

A more thorough theoretical foundation of treatment choice can be found in the work of Manski Manski (1997, 1999, 2000) and Dehejia (1999), and from a different viewpoint in Berger, Black, and Smith (2000). Berger, Black, and Smith (2000) analyse the use of profiling as an indirect method for the selection of the optimal treatment. Instead of being directly guided by potential outcomes, profiling proceeds by allocating individuals to the programmes on basis of their profiling scores, which is a 'need' or 'risk' indicator, e.g. the probability of becoming long-term unemployed, indicating how urgent active measures are. A close and positive relationship between the profiling variable and the treatment effects is supposed, in the sense that individuals with high profiling scores are those who gain most from treatment. Profiling might be a convenient way to select programme participants if such a close relation between the profiling variable and the outcome variables of interest exists and estimation of this profiling variable is more precise than direct estimation of the outcome variables of interest. However, single-score profiling is unlikely to work well and transparent, if a variety of different and heterogenous programmes is available to choose from or if multiple outcome variables refer to conflicting programme goals.[6]

Particularly if some programmes are harmful to some individuals, the equity argument often brought forward to argue that (potential) long-term unemployed are worst off and therefore most in need for intensive programmes might become contradictory. This equity argument is also questioned in Berger, Black, and Smith (2000) who show within the framework of a simple search model that long-term unemployment and welfare must not necessarily be negatively correlated.

The work of Manski and Dehejia is directly concerned with potential outcomes. Dehejia (1999) analyzed the GAIN experiment (Greater Avenues for Independence, USA) by explicitly considering the treatment decision problem of an individual. Still based on a parametric model, he used a Bayesian approach to take careful account of uncertainty in the individual decision making situation and looked for first-order stochastic dominance relationships between participation and non-participation. Manski, focusing on identification under weakest assumptions but neglecting estimation issues, analyzes the individual treatment selection problem and statistical selection rules. In Manski (1997) he derives bounds on the individual treatment effects under weak assumptions like monotonicity or concavity which may in some cases be sufficiently informative to establish dominance relationships between certain treatments for a particular person, thus enriching the information set of the decision maker. Statistical treatment selection rules, which are allocation rules based on statistical models that directly assign clients to treatments, are at the centre of Manski (1999, 2000). In Manski (2000) he analyses treatment choice under ambiguity and dominance and optimality of selection rules. Assigning each individual to the treatment with the highest expected potential outcome conditional on the observed covariates $E[Y^r|X = x]$ is optimal in an utilitarian sense. Decentralised self-selection is at least as good as a statistical selection rule if the individuals know more about their covariates than the statistician, have rational expectations about their potential outcomes, maximize expected utility and are risk neutral and if their objectives and preferences do not differ from the central planner's. This makes it, at least in some cases, questionable whether unguided self-selection is a wise selection process. Further, dominance relationships between feasible treatment rules are derived for a simple setting with a binary covariate, a binary treatment and a binary outcome variable. In Manski (1999) he compares two selection rules, one taking account of covariate information and the other one neglecting it.

This paper is motivated by a more practical approach to develop a robust statistical decision support system for choosing among multiple programmes, more elaborated and based on weaker

[6] To illustrate the difference between profiling and targeting based on estimated potential outcomes consider unemployment duration. Profiling would estimate for an unemployed individual the expected unemployment duration if the individual does not participate in treatment and assigns those individuals with the highest expected duration to treatment. Targeting would estimate unemployment duration in non-participation state as well as expected duraction if the individual would participate in treatment and selects individuals on basis of both estimated durations or their difference.

assumptions than the currently existing systems, but requiring more structure than Manski's work. A new semiparametric estimator for selecting the optimal programme is developed and exemplary applied to Swedish rehabilitation programmes.

To be able to discuss individually optimal programmes first the potential outcome framework of Rubin (1974) shall be introduced. Let R be the number of different programmes of which each individual has to choose exactly one. Suppose that for every individual potential (after-treatment) outcomes exist for all R programmes, denoted by random vectors $Y_i^1, Y_i^2, .., Y_i^R \in \Re^V$, of which eventually only that outcome vector will be observed that corresponds to the programme in which the individual participates. These *potential outcome vectors* Y^r contain a variety of different outcome variables and may include besides economic and monetary indicators also health, social, and psychological variables corresponding to the multiplicity of goals to which the programmes are directed. Let $X \in \Re^k$ denote a fine, i.e. detailed, set of exogenous individual characteristics. Then the expected potential outcome $E[Y^r|X = X_i]$ conditional on the characteristics of person i is a good approximation for the expected potential outcome $E[Y_i^r]$ for person i. Let $u(\cdot) : \Re^V \mapsto \Re$ denote a utility function mapping the V-dimensional outcome vector Y^r into an one-dimensional utility index and assuming that utility depends only on the expected outcomes, then the optimal programme can closely be approximated by

$$r_i^* = \underset{r \in \{1,..,R\}}{\arg\max} u_i \left(E[Y^r|X = X_i] \right).$$

Accordingly, the proposed estimator consists of two steps. First the expected potential outcomes $E[Y^r|X]$ are estimated. Since observations on Y^r are only available for former participants in programme r, only $E[Y^r|X, D = r]$ is identified, where $D \in \{1, ..., R\}$ is the participation indicator that indicates the programme in which a former participant has participated. Assuming that conditioning on a fine vector of explanatory characteristics X removes all selection bias, such that $E[Y^r|X] = E[Y^r|X, D = r]$, the potential outcomes are nonparametrically identified from the observed outcomes. This conditional independence assumption is central to many evaluation studies (e.g. Angrist (1998), Dehejia and Wahba (1999), Lechner (1999a)). However, as is well known from nonparametric statistics, nonparametric estimation of the expected value of Y^r conditional on a high-dimensional predictor X requires very large sample sizes. Due to this difficulty the profiling and targeting literature has by and large employed fully parametric specifications, whereas the microeconometric evaluation literature has concentrated on the nonparametric estimation of unconditional expectations EY^r (e.g. average treatment effects) and of conditional expectations on low-dimensional stratifying attributes, like men vs. women (see Angrist and Krueger (1999), Heckman, LaLonde, and Smith (1999), Manski (1995)). Estimating these average treatment outcomes is further eased by the balancing score property of the participation probabilities. As shown by Rosenbaum and Rubin (1983) for $R = 2$ and by Imbens (1999) and Lechner (1999b) for the multiple treatment case, conditional independence also implies that $E[Y^r|p^r(X)] = E[Y^r|p^r(X), D = r]$, where $p^r(X) = pr(D = r|X)$ is the probability that an individual with characteristics X participates in programme r. Thus, conditioning on the one-dimensional participation probability p^r is sufficient to avoid selection bias. By the law of iterated expectations the average treatment outcome is identified by $EY^r = E_{p^r(X)} E[Y^r|p^r(X), D = r]$ and can be estimated by first regressing Y^r on the observed participation probabilities p^r and weighting this estimate with the density of p^r in the population. Heckman, Ichimura, and Todd (1998) proposed local polynomial regression for the nonparametric first step estimation of $E[Y^r|p^r(X), D = r]$ instead of inefficient, though popular, pair-matching and developed asymptotic distribution theory, which can be applied to the estimate of the average treatment effect $\hat{E}Y^r$.

Since expected potential outcomes can be estimated for broad subpopulations without relying on any parametric assumptions, but nonparametric estimation conditional on a fine vector of characteristics X is practically infeasible, this paper suggests to combine parametric specifications of the conditional outcomes $E[Y^r|X]$ with nonparametrically estimated average outcomes to test whether the parametric model is correctly specified and to improve the coefficient estimates in case of misspecification. Both models are combined within the GMM framework and the statistical properties of the GMM estimator are derived in sections 2 and 3.[7]

After the expected potential outcomes have been estimated, determining the optimal treatment as the second step is considered in section 4. In case that the utility function u_i is known or the potential outcomes Y^r contain only one variable, it is tested whether for a particular individual one programme is jointly significantly better than all other programmes. This can easily be extended to derive a semi-ordering of best, intermediate, and worst programmes by multiple-comparisons-with-the-best (MCB, see Horrace and Schmidt (1996) or Hsu (1996)). If the utility function u_i is unknown, it is analysed whether a certain programme dominates other programmes in all outcome variables. Finally, in section 5 this semiparametric estimator is applied to Swedish rehabilitation programmes and section 6 concludes.

2 Statistical modelling

2.1 Nonparametric identification by the conditional independence assumption

Since for any individual which participated in programme D only the potential outcome Y^D can be observed but never any of the counterfactual outcomes Y^s, where $s \neq D$, the expected outcomes $E[Y^r|X]$ are generally not identified, but only $E[Y^r|X, D = r]$. Identifying $E[Y^r|X]$ from the observations Y_i^r of individuals who have participated in this programme ($D_i = r$), requires either a local instrumental variable (Heckman and Vytlacil 1999), which influences programme choice but does not affect the potential outcomes, or a sufficiently fine conditioning vector X which controls for all selection effects, such that no selection on unobservables remains (see e.g. Manski 1993). Following the latter approach requires that *all* exogenous variables which simultaneously influence the selection process and the potential outcomes are included in X, usually resulting in a conditioning vector of high dimension. This is expressed in the *conditional independence assumption* for multiple programmes (Imbens 1999, Rubin 1974):

$$Y^r \perp\!\!\!\perp \mathbf{1}(D = r)\,|X \qquad\qquad \forall r \in \{1..R\}, \tag{1}$$

where $\perp\!\!\!\perp$ denotes statistical independence and $1(\cdot)$ is the indicator function. It states that given the characteristics X knowledge whether an individual selected into programme r or into any other programme contains no further information about her potential outcome, i.e. conditional on X treatment selection is random. With other words, treatment selection depends on the potential outcomes only to the extent to which they can be anticipated on basis of the exogenous characteristics X, but not on an anticipation based on unobserved characteristics.

The plausibility of this assumption hinges on the extent to which individuals deliberately or unconsciously select into programmes on basis of characteristics related to their potential outcomes and to which extent these characteristics are observed. If self-selection is limited or if detailed information about participants' characteristics and behaviour is available this assumption may be reasonable.

Conditional independence implies $E[Y^r|X = x] = E[Y^r|X = x, D = r]$, such that the expected potential outcomes are identified from outcomes of former participants for all x values, for which there is positive probability to being selected into programme r. Thus $E[Y^r|X = x]$ is identified only for

[7] This differs from combining Micro and Macro data as in Imbens and Lancaster (1994), since they combine moments from two different datasets, whereas here moments from two different models based on the same data are combined. The nonparametric model is the just-identified model, while the parametric model introduces additional structure to the nonparametrically identified model.

all $x \in S^r$, where $S^1, S^2, .., S^R$ with $S^r \equiv supp\ f_{X|D=r} = \{x : f_{X|D=r}(x) > 0\}$ denotes the support of X in the subpopulation of participants in treatment r and $f_{X|D=r}(x)$ is the density of X in the subpopulation participating in treatment r. $S \equiv supp\ f_X = \{x : f_X(x) > 0\}$ is the support in the population.

2.2 Nonparametric estimation of average treatment outcomes

Although the expected potential outcomes $E[Y^r|X]$ are nonparametrically identified by the conditional independence assumption, nonparametric regression on a multivariate predictor becomes quickly imprecise with increasing dimension of X. However, at least in economic applications with observational data, controlling for a large number of characteristics is mandatory to eliminate selection on unobservables. In consequence most evaluation studies have abstained from estimating conditional outcomes and concentrated on average treatment outcomes[8] EY^r for broad subpopulations, which can be estimated more precisely (even with $\sqrt{n}$-convergence Heckman, Ichimura, and Todd (1998)).

The dimension-reducing balancing score property of the participation probabilities facilitates the estimation of average treatment outcomes as the conditional independence assumption (1) also implies that

$$E[Y^r|p^r(X = x)] = E[Y^r|p^r(X = x), D = r] \qquad \forall x \in S^r, \tag{2}$$

where $p^r(x) \equiv pr(D = r|X = x)$ is the probability to be selected into treatment r (Rosenbaum and Rubin (1983) for the case $R = 2$ and Imbens (1999) and Lechner (1999b) for multiple programmes). I.e. if treatment assignment is ignorable conditional on X, then it is also ignorable conditional on the participation probability $p^r(X)$. With known or consistently estimated participation probabilities the expected outcomes conditional on the (one-dimensional) participation probability $p^r(X)$ can easily be estimated by nonparametric regression[9]. By the law of iterated expectations the average treatment outcome EY^r is identified for the population with characteristics in the support S^r by

$$E_{S_r}Y^r = \underset{X|X\in S^r}{E} E[Y^r|p^r(X)] = \underset{X|X\in S^r}{E} E[Y^r|p^r(X), D = r] = \frac{\underset{X|X\in S}{E}\left(E[Y^r|p^r(X), D = r] \cdot 1(X \in S^r)\right)}{pr(X \in S^r|X \in S)}. \tag{3}$$

and can be estimated as

$$\hat{E}_{S_r}Y^r = \frac{n^{-1}\sum_i \hat{m}^r(\hat{p}^r(X_i)) \cdot 1(X_i \in \hat{S}^r)}{n^{-1}\sum_i 1(X_i \in \hat{S}^r)}, \tag{4}$$

where $\left(X_i, D_i, Y_i^{D_i}\right)$ are a random sample of n independent observations and $\hat{S}^r$, $\hat{p}^r$ and $\hat{m}^r(.)$ are preliminary estimates of the support S^r, the participation probability $p^r(x)$, and the regression curve $m^r(p)$ $=E[Y^r|p^r(X) = p]$, respectively. All operations with respect to the vector Y^r are defined as *element-wise* in Y^r. Notice that outside the support S^r the conditional outcome $E[Y^r|X]$ is not identified, such that the population expectation of Y^r is not a useful concept.

[8] The focus of most evaluation studies is on treatment effects, which are the difference between two potential outcomes.

[9] It still remains the problem to estimate $pr(D = r|X)$ without strong parametric assumptions. However, at least in the simulations of Todd (1999), comparing parametric and semiparametric estimators of the participation probability, the parametric estimation was quite robust and misspecification of the participation probability seemed to be less serious.

2.3 Parametric modelling of potential outcomes

In contrast to the microeconometric evaluation literature, the operating profiling and programme-targeting systems aim to predict potential outcomes specifically for a particular individual and are usually based on fully parametric models. Since the joint distribution of the potential outcome vectors is not identified, the potential outcomes can be modelled separately for each programme r by specifying a vector-valued function $h^r(x, \theta^r)$, with coefficient vector θ^r of dimension k, and finding values for θ^r, such that $h^r(x, \theta^r)$ approximates $E[Y^r|X = x]$ as well as possible. The approximation is based on the participants in programme r, since only for them the outcome Y^r can be observed. Consequently the approximation to the expected potential outcome may be weak for x values which are distant from the region where the density mass of the participants in programme r lies.

In the case that the parametric functions h^r are correctly specified, i.e. there exist coefficient vectors $\{\theta_0^r\}_{r=1..R}$ such that under the conditional independence assumption (1)

$$E[Y^r|X = x, D = r] = E[Y^r|X = x] = h^r(x, \theta_0^r) \qquad \forall x \in S, \quad \forall r \in \{1..R\}, \tag{5}$$

then standard asymptotic theory for parametric estimators can be employed to show convergence of $h^r(x, \hat{\theta}^r)$ to the true expected outcomes $h^r(x, \theta_0^r)$, where the R outcome relationships can be estimated independently from the subsamples of respective participants.

If no coefficient vectors $\{\theta_0^r\}_{r=1..R}$ exist that satisfy (5) then the parametric model is misspecified and at least one of the functions h^r is not sufficiently flexible to embrace the true mean function for all x. Necessarily any estimate $h^r(x, \hat{\theta}^r)$ must be biased for at least some x values and the quality of the approximation to the mean function could be deficient with respect to the whole population, if the distribution of X in the population and in the subpopulation of participants in programme r differ (cf. Section 3.3 below).

2.4 The semiparametric model

Extracting the essence of both the nonparametric and the fully parametric model, it follows that in observational studies of reasonable sample size the nonparametric estimation of $E[Y^r|X = x]$ is often extremely difficult due to the high dimension of X, while the nonparametric estimation of $E[Y^r|p^r = p]$ and consequently of $E[Y^r]$ is fairly straightforward. On the other hand, by parametrically specifying the outcome relationships $E[Y^r|X = x]$ can easily be estimated, though at the cost that the results may be biased due to misspecification of the functional forms. The semiparametric model proposed below combines the parametric equations with nonparametric estimates of $E[Y^r]$.

If both models were correct, their predicted mean outcomes should converge to the same limit. I.e. the average treatment outcome with respect to the population (within S^r) once estimated nonparametrically as

$$\hat{E}_{S_r} Y^r = \frac{\sum_i \hat{m}^r(\hat{p}^r(X_i)) \cdot 1(X_i \in \hat{S}^r)}{\sum_i 1(X_i \in \hat{S}^r)}$$

and once implied by the parametric model

$$\hat{E}_{S_r} Y^r = \frac{\sum_i h^r(X_i, \hat{\theta}^r) \cdot 1(X_i \in \hat{S}^r)}{\sum_i 1(X_i \in \hat{S}^r)}$$

should be identical. This implies the equality condition

$$\sum_i \left(h^r(X_i, \hat{\theta}^r) - \hat{m}^r(\hat{p}^r(X_i)) \right) \cdot 1(X_i \in \hat{S}^r) = 0. \tag{6}$$

Supposing consistent estimates of $p^r(\cdot)$ and $m^r(\cdot)$, a violation of the equality condition (6) would indicate a misspecification of the parametric model (5).

Such an equality condition must also hold for any subpopulation defined on the X characteristics. To state the equality conditions (6) concisely for L different subpopulations define a vector-valued indicator function $\Lambda(x)$ of dimension $L \times 1$, which is one in the components $\Lambda_l(x)$ for which x is part of the corresponding subpopulation and zero otherwise. An example of this multidimensional indicator function defining the 3 populations: whole population, subpopulation of males, and subpopulation aged 40 to 50 would be

$$\Lambda(x) = \begin{pmatrix} 1 \\ 1(\text{gender} = \text{male}) \\ 1(\text{age} \in [40, 50]) \end{pmatrix}. \tag{7}$$

Then the semiparametric model can be summarized as:

$$\begin{array}{l} \text{Estimate } h^r(x, \theta^r) \text{ from the observations } (X_i, Y_i^r) \text{ with } D_i = r \\ \text{sucht that} \\ \sum_i \left(\Lambda(X_i) \otimes h^r(X_i, \hat{\theta}^r) - \hat{\mathbf{m}}_{VL}^r(\hat{p}^r(X_i)) \right) \cdot 1(X_i \in \hat{S}^r) = 0 \qquad \forall r \in \{1..R\}, \end{array} \tag{8}$$

where $\hat{\mathbf{m}}_{VL}^r$ is a column vector of length VL of all stacked nonparametric estimates of $E[Y^r|p^r]$, , such that the first V elements correspond to the outcome vector for population one, the second V elements to the outcome vector for population two, and so forth. I.e. $\hat{\mathbf{m}}_{VL}^r(\hat{p}^r(X_i)) = (\hat{m}_1^{r\prime}(\hat{p}^r(X_i) \cdot \Lambda_1(X_i), .., \hat{m}_l^{r\prime}(\hat{p}^r(X_i) \cdot \Lambda_l(X_i), .., \hat{m}_L^{r\prime}(\hat{p}^r(X_i) \cdot \Lambda_L(X_i))'$ is the column vector of all estimated outcome variables $\hat{m}_l^{r\prime}(\hat{p}^r(X_i) = (\hat{m}_{1l}^{r\prime}(\cdot), .., \hat{m}_{vl}^{r\prime}(\cdot), .., \hat{m}_{Vl}^{r\prime}(\cdot))'$ in all populations $l = 1, .., L$, where $\hat{m}_{vl}^r(\hat{p}^r)$ is an estimator of the expectation $E[Y_v^r|p^r(X) = \hat{p}^r, \Lambda_l(X) = 1]$ of the v-th variable of the potential outcome vector Y^r in the l-th subpopulation defined by $\Lambda_l(X)$ conditional on the participation probability. $\otimes$ is the Kronecker product operator.

Obviously, the larger the number of subpopulations L the more detailed information about subgroup heterogeneity will enter into the model. On the other hand, the smaller the subpopulations get the less precise the nonparametric estimates will be and their additional value as equality restrictions for the parametric model will decline.

3 Estimation of the semiparametric model

3.1 GMM Estimator

The semiparametric model of (8) can neatly be expressed in the framework of the general method of moments (GMM). The GMM estimator of θ^r is constructed from two sets of moments. The first set of moment conditions emerges from the parametric equations as $E[Y^r - h^r(X, \theta_0^r)|X = x, D = r] = 0$, supposed the parametric functions are correctly specified. Conditioning on $D = r$ is essential, since otherwise the outcome Y^r is not observed. This implies that

$$E[A^r(X) \cdot (Y^r - h^r(X, \theta_0^r)|D = r] = 0, \tag{9}$$

with $A^r(x)$ any vector-valued function of x. Suppose that $A^r(x)$ is a $k \times 1$ vector-valued function, such that a GMM estimator based on only these moments would be just identified, since θ_0^r contains k coefficients.

Combining these moment conditions with the equality conditions discussed in the previous section leads to the moment vector g_n^r

$$g_n^r(\theta^r, \hat{\mathbf{m}}_{VL}^r, \hat{S}^r) = n^{-1} \sum_i \begin{pmatrix} A^r(X_i) \cdot (Y_i - h^r(X_i, \theta^r)) \cdot 1(D_i = r) \\ (\Lambda(X_i) \otimes h^r(X_i, \theta^r) - \hat{\mathbf{m}}_{VL}^r(\hat{p}^r(X_i)) \cdot 1(X_i \in \hat{S}^r) \end{pmatrix}, \qquad (10)$$

of length $k + VL$. The upper part of the moment vector, i.e. the first k elements, corresponds to the moment conditions (9) multiplied by n_r/n. The lower part of the moment vector correspond to the VL equality conditions as introduced in the previous section. Remember that V is the number of outcome variables in the potential outcome vectors and that L is the number of (sub)populations considered. $\hat{\mathbf{m}}_{VL}^r(\hat{p}^r(\cdot))$ is the column vector of the stacked nonparametric estimates of $E[Y^r|p^r(\cdot)]$ for all considered subpopulations. While the moments emanating from the parametric equations are evaluated over the subsample with $D_i = r$, the equality condition between the nonparametric and parametric estimates provides a set of moments for the whole population. Adding these equality conditions as additional moments to the moments of the parametric model shall ensure that the coefficients θ^r are estimated such that parametric equations and the nonparametric estimates predict similar outcomes. Then θ_0^r can be estimated by GMM as the solution to

$$\hat{\theta}_n^r = \arg\min_{\theta^r} g_n^r(\theta^r, \hat{\mathbf{m}}_{VL}^r, \hat{S})' \hat{W}^r g_n^r(\theta^r, \hat{\mathbf{m}}_{VL}^r, \hat{S}),$$

where $\hat{W}^r$ is a weighting matrix with *plim* $\hat{W}^r = W^r$ positive semidefinite.

Since the joint distribution of the potential outcomes is not identified the joint distribution of the moment vectors $g_n^1(\theta^1), .., g_n^R(\theta^R)$ is also not identified. If the potential outcomes $Y^1, .., Y^R$ are assumed to be independent then the moment vectors $g_n^1(\theta^1), .., g_n^R(\theta^R)$ are uncorrelated (see Lemma 3 in the appendix). With g_n^r and $\hat{\theta}_n^r$ asymptotically normal, as shown below, θ^r can be estimated separately for each r without loss in efficiency. Thus, in the following the estimation of $\{\theta^r\}_{r=1..R}$ proceeds independently for each r (except for the participation probability model, which is estimated only once).

3.2 Properties of the GMM estimator

First the statistical properties of the GMM estimator are investigated when the parametric outcome relationship h^r is correctly specified, i.e. there exists a true parameter vector θ_0^r in a compact parameter space such that

$$H_0^r: \quad h^r(x, \theta_0^r) = E[Y^r|X = x] = E[Y^r|X = x, D = r] \qquad \forall x \in S^r \quad \text{with } \theta_0^r \in \Theta^r.$$

Under this hypothesis the moment function (10) has expectation zero at the true values $\theta_0^r, \mathbf{m}_{VL}^r$, i.e. $Eg_n^r(\theta_0^r, \mathbf{m}_{VL}^r, S^r) = 0$, and $\sqrt{n}$-consistency and asymptotic normality of the coefficient estimates $\hat{\theta}^r$ will be shown. Proofs are found in Appendix A. Furthermore with a suited weighting matrix the GMM statistic is asymptotically χ^2 distributed and the J-test of overidentifying restrictions (Hansen 1982) can be used to test the null-hypothesis that the model is correctly specified, i.e. whether the parametric model prognosticates potential outcomes which are in line with the nonparametric estimates.

Theorem 1 (Consistency) *If*
(i) the parametric function $h^r(x,\theta^r)$ is continuous in θ^r over a compact parameter space Θ^r,
(ii) has a unique solution $\theta_0^r \in \Theta^r$ such that $h^r(x,\theta^r) = E[Y^r|X=x]$ $\forall x \in S^r$ if and only if $\theta^r = \theta_0^r$,
(iii) for each subpopulation defined by $\Lambda(x)$ the moments $E \sup\limits_{\theta^r \in \Theta^r} \|A^r(X) \cdot (Y^r - h^r(X,\theta^r))\|$ and also $E \sup\limits_{\theta^r \in \Theta^r} \|h^r(X,\theta^r) \cdot 1(X \in S^r) - E[Y^r 1(X \in S^r)]\|$ exist, and the number of subpopulations is finite,
(iv) a consistent estimator of $\mathbf{m}^r_{VL}(p^r)$ and S^r is available, and
(v) the weighting matrix $\hat{W}^r$ converges in probability to a positive definite matrix.
then the GMM estimator with moment vector (10) is consistent.

Remark 1 *Assumption (iii) of Theorem 1 could be relaxed to the form given in Corollary 5. However, in its current form it is more intuitive, where the first moment existence condition is the condition that would apply if only the parametric model would be estimated by GMM and the second condition requires that the parametric function re-centred by its mean has finite expectation for all admissible coefficient values.*

To establish the limit distribution of the GMM estimator I draw on the results of Heckman, Ichimura, and Todd (1998). To apply their results the preliminary estimators $\hat{\mathbf{m}}^r_{VL}$ and $\hat{p}^r$ are restricted to the class of asymptotically linear estimators with trimming (see Definition 1 in the appendix). Heckman, Ichimura, and Todd (1998) have shown that local polynomial estimators, e.g. kernel or local linear regression, belong to this class and that $\hat{\mathbf{m}}^r_{VL}(\hat{p}^r(x))$ with $\hat{p}^r(x)$ estimated either parametrically or nonparametrically by local polynomial regression is also asymptotically linear with trimming.

For the following theorem define $n_{l,r}$ as the number of participants in treatment r who belong to the l-th subpopulation, let $\Psi_{l,p}$ denote the influence function stemming from the estimation of the participation probabilities $p^r(\cdot)$ and $\Psi_{l,m}$ denote the influence function from estimating $m^r(\cdot)$ (see Corollaries 6 and 7 in the appendix).

Theorem 2 (Asymptotic Normality) *Suppose the estimator $\hat{m}^r(\hat{p}^r(x))$ of $E[Y^r|p^r(X=x)]$ is asymptotically linear with trimming for all outcome variables in all considered subpopulations of the form*
$[\hat{m}^r_l(\hat{p}^r(x)) - m^r_l(p^r(x))] \cdot \Lambda_l(x) 1(x \in \hat{S}^r) =$

$$= n_{l,r}^{-1} \sum_j \Psi^r_{l,m}(Y^r_j, D_j, X_j; x) + n^{-1} \sum_j \Psi^r_{l,p}(Y^r_j, D_j, X_j; x) + \hat{b}^r_l(x) + \hat{R}^r_l(x), \tag{11}$$

with $E[\Psi^r_{l,p}(Y^r_j, D_j, X_j; X)|X=x] = 0$, $E[\Psi^r_{l,m}(Y^r_j, D_j, X_j; X)|X=x] = 0$, plim $n^{-\frac{1}{2}} \sum \hat{b}^r_l(X_j) = b^r_l < \infty$, and $n^{-\frac{1}{2}} \sum \hat{R}^r_l(X_j) = o_p(1)$. Furthermore suppose that
(i) $VL \cdot Var\left(\Psi^r_{l,m}(Y^r_j, D_j, X_j; X_i)\right) = o(n) = VL \cdot Var\left(\Psi^r_{l,p}(Y^r_j, D_j, X_j; X_i)\right)$ for each outcome variable v and in each subpopulation l,
(ii) some regularity conditions on J (see appendix),
(iii) the parametric model is correctly specified $H_0^r : h^r(x,\theta_0^r) = E[Y^r|X=x]$ $\forall x \in S$, where $\theta_0^r \in$ interior of Θ^r, with Θ^r compact in $\Re^k$, h^r continuously differentiable with bounded derivative in a neighbourhood of θ_0^r, $E\left[\|A^r(X)(Y - h^r(X,\theta^r)\|^2\right] < \infty$, and $G^{r\prime} W^r G^r$ nonsingular where G^r is the expected gradient of the moment vector
(iv) $\lim\limits_{n\to\infty} \frac{n_{l,r}}{n} = \lambda_{l,r}$ with $0 < \lambda_{l,r} < \infty$ for each subpopulation $l = 1,..,L$,
(v) the estimated support $\hat{S}^r = \{x : \hat{f}_X(x) \geq q_0\}$ is estimated such, that $\sup_{x\in S}\left|\hat{f}_X(x) - f_X(x)\right|$ converges a.s. to zero where $S = \{x : f_X(x) \geq q_0\}$, $\hat{f}_X$ is a kernel density estimate with kernel with

moments 1 through k equal to zero, and f_X is $k+1$ times continuously differentiable with $(k+1)$-th derivative Hölder continuous,
(vi) $\hat{W}^r$ converges to W^r p.d.
then the GMM estimator $\hat{\theta}^r = \arg\min_{\theta^r} g_n^r(\theta^r, \hat{\mathbf{m}}_{VL}^r, \hat{S}^r)' \hat{W}^r g_n^r(\theta^r, \hat{\mathbf{m}}_{VL}^r, \hat{S}^r)$ with moment vector (10) is asymptotically normally distributed with

$$n^{\frac{1}{2}}(\hat{\theta}^r - \theta_0^r) \xrightarrow{d} N\left(\left[\begin{array}{c} \mathbf{0}_k \\ -(G^{r\prime}W^rG^r)^{-1}G^{r\prime}W^r\mathbf{b}_{VL}^r \end{array}\right], (G^{r\prime}W^rG^r)^{-1}G^{r\prime}W^rE[J^rJ^{r\prime}]W^rG^r(G^{r\prime}W^rG^r)^{-1}\right) \tag{12}$$

where

$$\begin{aligned} J^r &= g^r(Z^r, \theta_0^r, \mathbf{m}_{VL}^r) \qquad (13) \\ &- \left(\begin{array}{c} \mathbf{0}_k \\ \lambda_{1,r}^{-1} \cdot E[\Psi_{1,m}^r(Y_1^r, D_1, X_1; X_2)|Y_1^r, D_1, X_1] + E[\Psi_{1,p}^r(Y_1^r, D_1, X_1; X_2)|Y_1^r, D_1, X_1] \\ \vdots \\ \lambda_{L,r}^{-1} \cdot E[\Psi_{L,m}^r(Y_1^r, D_1, X_1; X_2)|Y_1^r, D_1, X_1] + E[\Psi_{L,p}^r(Y_1^r, D_1, X_1; X_2)|Y_1^r, D_1, X_1] \end{array}\right). \end{aligned}$$

Remark 2 *If the number of outcome variables and the number of subpopulations does not grow with sample size, then assumption (i) reduces to*
$\lim_{n\to\infty} n^{-1}Var\left(\Psi_{l,m}^r(Y_j^r, D_j, X_j; X_i)\right) = 0 = \lim_{n\to\infty} n^{-1}Var\left(\Psi_{l,p}^r(Y_j^r, D_j, X_j; X_i)\right).$

Remark 3 *Also the moment vector g_n^r is asymptotically normal*

$$n^{\frac{1}{2}} g_n^r(\theta_0^r, \hat{\mathbf{m}}_{VL}^r, \hat{S}^r) \xrightarrow{d} N\left(\left[\begin{array}{c} \mathbf{0}_k \\ \mathbf{b}_{VL}^r \end{array}\right], EJ^rJ^{r\prime}\right). \tag{14}$$

Remark 4 *The bias term could be made to vanish, by choosing a bandwidth sequence h_n that converges faster to zero. Hence by smoothing less the nonparametric regression curve m the bias can be abolished. Less smoothing however will lead to a more wiggly curve and the absolute value of the first derivative will on average increase. This could lead to a drastic increase in the variance of the estimation since the sampling imprecision of the discrete choice step enters multiplicatively with the slope of the regression curve m. If the propensity score were known, one would like to estimate the regression line m with little bias to track the true line of the regression curve. However, with $p(x)$ imprecisely estimated one wants to estimate the regression curve m with less variance such that $\hat{m}(\hat{p}(x)+\varepsilon) \approx \hat{m}(\hat{p}(x))$, since p is not estimated precisely. This will lead generally to higher bias and lower variance and the less precise the first step estimation of the discrete choice model is the less variance one wants. To allow for more smoothing, the bandwidth sequence must be allowed to decrease at a lower rate to zero, resulting in the bias term in the limit distribution. However, the bandwidth sequence must converge sufficiently fast to zero to ensure that the average bias term multiplied by $\sqrt{n}$ is nonstochastic.*

By standard GMM theory the asymptotically efficient GMM estimator chooses $W = [EJ^rJ^{r\prime}]^{-1}$, which simplifies the asymptotic variance to $\left(G'[EJJ']^{-1}G\right)^{-1}$. With this weighting matrix the GMM statistic multiplied by the sample size is χ^2 distributed with number of freedoms equal to the number of overidentifying restrictions VL

$$n \cdot g_n^{r\prime}\hat{\Omega}^r g_n^r \xrightarrow{d} \chi^2_{(VL)},$$

with $\hat{\Omega}^r$ a consistent estimate of the efficient weighting matrix $[EJ^r J^{r\prime}]^{-1}$. This statistic gives a criterion of how well the parametric fit conforms with the nonparametric averages and rejection of the J-test is an indication that the parametric model might be misspecified. The efficient weighting matrix $[EJ^r J^{r\prime}]^{-1}$ can be estimated by its sample average $n\,[\sum J_i^r J_i^{r\prime}]^{-1}$. Evaluation of J_i^r however requires expected values of the influence functions $\Psi_{l,p}$ and $\Psi_{l,m}$, which themselves can also be estimated by sample averages. The influence functions depend on the employed estimators for the participation probabilities p^r and the regression curves m^r. These are given in the appendix for maximum likelihood estimation of p^r and kernel regression estimation of $m^r(\hat{p}^r)$.

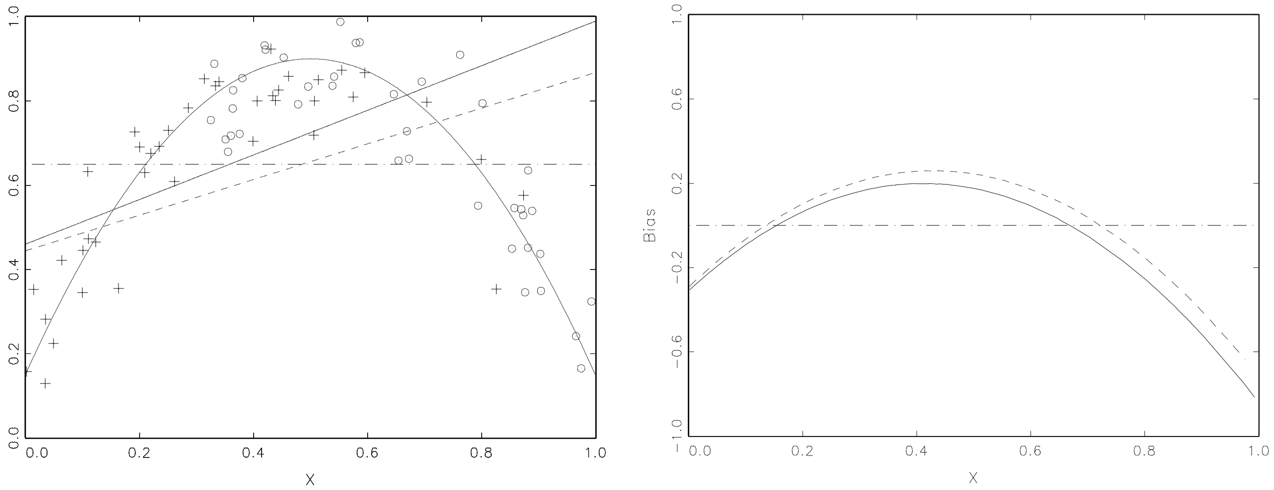
3.3 Properties under misspecification

In the previous section the properties of the GMM estimator have been derived under correct specification of the outcome relationship h^r. In this case, however, substantial gains in precision by adding the equality moments to the moments of the parametric model are unlikely, since the average programme outcomes are unknown and $\hat{m}^r(\hat{p}^r)$ needs to be estimated. But if the conditional mean function is misspecified, i.e.

$$H_1^r: \quad \forall \theta^r \in \Theta^r \quad \exists x \in S \quad \text{such that} \quad h^r(x, \theta^r) \neq E[Y^r|X=x],$$

then considerable reductions in Mean Squared Bias should be possible. In this case the class of functions spanned by $h^r(\cdot)$ does not contain the true mean function and since there is no "true" θ_0^r for which h^r and the true mean function would coincide, any estimate $h^r(x, \hat{\theta}^r)$ must be biased for at least some x values.

Figure 3.1: Bias of purely parametric estimator and of the GMM estimator



Estimated regression lines of parametric and GMM estimator.

Bias as a function of x.

Left picture: Parabolic line is true regression curve with true mean 0.65 (horizontal dashed line). Observations of participants depicted by + signs, unobserved observations of non-participants by circles. Estimated regression lines: OLS (solid) and GMM (dashed). Right picture: Bias as a function of x of OLS (solid) and GMM (dashed)

This situation is illustrated by an example in Figure 3.1 (left picture). There Y^r and X are both one-dimensional and the true mean function $E[Y^r|X=x]$ is parabolic. The parametric function is specified as linear. Data on 80 observations are available, of which 40 participated in programme r and the remaining 40 participated in any other programme. The observed (Y_i^r, X_i) pairs of the participants are marked by + signs, the pairs (Y_i^r, X_i) of the non-participants are marked by circles. Notice that the counterfactual outcome Y_i^r of the non-participants is unobservable and only shown for illustration. The density mass of X of the participants lies to the left of that of the non-participants.

The true mean EY^r is 0.65, which could be estimated nonparametrically. Estimating the purely parametric model, e.g. by OLS, on basis of the observed outcomes of the participants yields the solid regression line. Estimating the semiparametric model, which includes the information on average potential outcomes, gives the dashed regression line. In the right picture of Figure 3.1 the conditional bias $b^r(x) = E[Y^r - h^r(x, \hat{\theta}^r)|X = x]$ as a function of x is displayed. As expected, the OLS regression line approximates the true mean function better in regions where the participants are concentrated, but departs seriously from the true mean function for x large. The GMM regression line stays on average closer to the true mean EY^r and is less biased for large values of x. Mean squared bias $E\left[b^r(x)^2\right]$ as a criterion of approximation quality is significantly lower for the GMM regression line than for the OLS regression line.

Assessing the quality of the approximation by expected squared bias with respect to the distribution of X in the whole population appears sensible, as the potential outcomes shall be predicted for any individual drawn randomly from the population to select a suited programme. However, although intuitive, a general proof of the superiority of the semiparametric model to the purely parametric model seems difficult and shall be developed in future research.[10] Obviously substantial gains in efficiency or approximation quality can only be expected, if the density distributions $f_X(x)$ and $f_{X|D=r}(x)$ differ, since an estimator based on the purely parametric model seeks to minimize mean squared error with respect to $f_{X|D=r}(x)$, while $f_X(x)$ would be the relevant weighting function. Then the average outcome predicted by the parametric model usually will differ from the true mean EY^r. Hence, although the parametric approximation will in most cases be more precise for the participants in programme r, it will on average be biased and might be unreliable for x values where the density $f_{X|D=r}(x)$ of the participants is small. In this sense it is suspected, that including nonparametric estimates of average expected outcomes EY^r for various subpopulations as additional information will improve the approximation to the true conditional mean function. As such the semiparametric GMM estimator should be more robust to misspecification than an estimator based only on the parametric model.

4 Choosing the optimal programme

After satisfactory specifications of the functions $h^r(\cdot)$ have been found and the coefficients θ^r have been estimated, the expected potential outcomes can be estimated for a particular person with characteristics x as

$$\hat{Y}^r(x) = \hat{E}[Y^r|X = x] = h^r(x, \hat{\theta}^r) \qquad \forall r \in \{1..R\}. \tag{15}$$

Since the $\hat{\theta}^r$ are asymptotically normal with known covariance matrices and uncorrelated among each other the approximate distribution of $\hat{Y}^r(x)$ for given x can easily be simulated.

For determining the optimal programme $r^* = \arg\max u\left(Y^r(x)\right)$ consider first the case where either the utility function is known, e.g. a weighting function assigning different weights to the V components of the potential outcome vector, or where the potential outcome vector is a scalar random variable, i.e. $V = 1$, and it is known that the utility function is monotone. Then the probability that a certain programme r generates higher utility than any other programme is

$$pr\left\{u\left(\hat{Y}^r(x)\right) \geq u\left(\hat{Y}^1(x)\right), u\left(\hat{Y}^r(x)\right) \geq u\left(\hat{Y}^2(x)\right), ..., u\left(\hat{Y}^r(x)\right) \geq u\left(\hat{Y}^R(x)\right)\right\}. \tag{16}$$

[10] Notice that it is fairly easy to show that absolute average bias $|Eb^r(x)|$ will usually be smaller with the semiparametric model than with the parametric model. However, average unbiasedness of the estimated potential outcomes is of limited practical use, if an adequate programme needs to be choosen for a particular person, since average unbiasedness could also be achieved with large positive biases for some x values and large negative biases for other x values.

If for instance $\hat{Y}^r(x)$ can be approximated as multivariate normal and if the utility function is a linear weighting function then (16) corresponds to the probabilities of the likelihood function of a multinomial probit model. Thus the dominance probability that programme r is the best programme given characteristics x can readily be computed for all programmes. A test whether programme r is optimal at the significance level α corresponds to the dominance probability of programme r being larger or equal to $1 - \alpha$.

For larger numbers of available programmes R it will often be the case that no programme dominates all other with high probability, but that a semi-ordering into best, intermediate, and worst programmes is possible. The subset of best programmes jointly dominates all other programmes with high probability, but among these best programmes no statistically significant ordering is possible. Such a clustering of programmes is accomplished by multiple-comparisons-with-the-best (MCB) techniques (Horrace and Schmidt (1996), Hsu (1996)), which are not further considered in this paper.

In the case where the utility function u is unknown it might be best to directly provide the decision maker with the estimation results, e.g. in the form of Table 4.1. This table displays 95% confidence intervals of the estimated potential outcomes for a given person for $V = 2$ outcome variables and $R = 3$ available programmes. The 95% confidence intervals provide the decision maker with some indication of the estimation precision. Although determining the optimal programme will in most cases be impossible without engaging in valuing the different outcome variables[11], still various forms of dominance relationships can be derived, if the utility function is monotone in all arguments and it is known for all outcome variables whether utility is nondecreasing or nonincreasing in this variable. For each of the V outcome variables separately, e.g. each row in Table 4.1, the best and the worst programmes can be determined as described above. Furthermore, probabilities can be derived that a certain programme r dominates another programme s jointly in all outcome variables, which is simply $pr\left(\hat{Y}^r(x) \geq \hat{Y}^s(x)\right)$ if the utility function is nondecreasing in all arguments. Again, if this probability reaches $1 - \alpha$ implies that programme r statistically significantly dominates programme s. Analogously, this dominance concept can be extended to joint dominance of programme r over a set of programmes s and if programme r jointly dominates all other programmes significantly then programme r can be declared as the optimal programme without knowing the exact form of the utility function.

Table 4.1: Expected potential outcomes for a particular individual

	Programme 1	Programme 2	Programme 3
Re-employment chances	0.12-0.18	0.10-0.19	0.18-0.22
Programme costs	910-990	90-95	650-670

Note: Entries refer to 95% confidence intervals.

[11] For instance, how much weight is placed on programme costs certainly depends on who is bearing the cost. For a discussion why the statistician should provide disaggregate information and should not aggregate all these information into a single number, as in a cost-benefit analysis, see for instance Mohr (1999). Furthermore the decision maker might have additional information about the characteristics of the individual facing treatment which did not enter into the statistical system. Incorporating this supplementary information into the decision making will be easier with disaggregate information than with aggregated results. Especially in the case of (partial) self-selection, where the individual knows best about her unobserved characteristics, disaggregate information is constructive to complement personal knowledge with statistical recommendations on basis of observed covariates. This might also provide incentives for individuals to reveal their characteristics truthfully to be able to obtain better statistical information for their personal decision making.

5 Application to Swedish Rehabilitation Programmes

In this section the Swedish rehabilitation policy is briefly examined, optimal choices of rehabilitation programmes are elaborated, and it is assessed, to which degree the current participants-to-programme selection was efficient. However, it should be mentioned in advance that due to limited sample size only a very broad categorization of rehabilitation programmes is feasible, such that policy recommendations will stay rather general.

The Swedish rehabilitation policy distinguishes between vocational and non-vocational rehabilitation and is directed towards individuals with reduced work capacity due to long-term sickness of at least one month. Non-vocational rehabilitation contains medical rehabilitation as well as social programmes for individuals with alcohol, drug or psychiatric problems and intends to re-establish independency of the sick individual from medical or therapeutical assistance. Vocational rehabilitation consists of workplace training and occupation-related educational measures and aims at restoring lost working capacity and fostering independence of financial social assistance. The Swedish National Social Insurance Board is responsible for the co-ordination of vocational rehabilitative measures and grants them only if recovery within less than a year is expected and the recovered work-capacity seems economically valuable to the Swedish labour market.[12] In 1994 a retrospective survey of about 75'000 long-term sickness cases between 1991 to 1994 has been conducted by the Swedish National Social Insurance Board of which $N = 6287$ observations in Western Sweden have been selected. A full description of the relevant institutions, the dataset, and the justification of the conditional independence assumption is found in the evaluation study Frölich, Heshmati, and Lechner (2000a). In that study the rehabilitative measures were grouped into the six categories passive, workplace, educational, medical, social, and no rehabilitation, and the effects of rehabilitation on the two outcome variables re-employment and re-integration into the labour force were estimated. In general no positive average treatment effects of rehabilitation were found, but sometimes even negative effects, particularly of educational rehabilitation.

Since the categories passive, educational, and social contained only rather few participants, the rehabilitation programmes are summarised here into only three categories: *No rehabilitation*, *vocational rehabilitation (VR)*, and *non-vocational rehabilitation (NVR)*.[13] Notice that within these categories programme heterogeneity is still substantial, e.g. workplace training appeared to be much more successful than educational rehabilitation in Frölich, Heshmati, and Lechner (2000a, b). Concentrating on a single outcome variable *re-employment at the end of the sickness spell*, in a first step the expected potential outcomes conditional on observed characteristics shall be estimated. Table 5.1 gives the number of participants in the three rehabilitation groups and the share of participants who engaged in regular employment at the end of their sickness spell. These unadjusted re-employment shares provide a first impression of the magnitude of re-employment after long-term sickness, and represent the *gross* success rate of the rehabilitation programmes not taking into account that the characteristics of the participants differ substantially among the programmes. On this account about 46% of all long-term sick achieve re-employment after sickness. The employment rate is lower for the participants in vocational rehabilitation (46.7%) than for the non-participants (48.3%), and of the participants in non-vocational rehabilitation only 40.5% encountered re-employment. However, such figures might be expected if the individuals participating in rehabilitation generally face poorer labour market prospects. Thus, Table 5.1 might display merely

[12] For instance, if the sick individual was occupied in a declining sector and successful re-training to a different sector seems improbable, then vocational measures are denied.

[13] These groups correspond to Frölich, Heshmati, and Lechner (2000a) as follows: No rehabilitation (contains also the passive rehabilitation programmes), vocational rehabilitation (contains workplace and educational rehabilitation), and non-vocational rehabilitation (contains medical and social programmes).

selection effects and in particular does not provide any information about which of these programmes are recommendable for a particular person.[14]

Table 5.1: Number of observations and unadjusted re-employment shares

	All	None	Vocational	Non-vocational
# Observations	6287	3502	1478	1307
Re-employment	46.3	48.3	46.7	40.5

Note: Share of transitions to employment of all sickness cases (in %).

To estimate the binary outcomes Y^r the regression curve is parametrically specified as a probit with different coefficients for each of the three potential outcomes:

$$\begin{array}{lll} E[Y^{None}|X=x] & = h^{None}(x,\theta^{No}) & = \Phi(x'\theta^{No}) \\ E[Y^{VR}|X=x] & = h^{VR}(x,\theta^{VR}) & = \Phi(x'\theta^{VR}) \\ E[Y^{NVR}|X=x] & = h^{NVR}(x,\theta^{NVR}) & = \Phi(x'\theta^{NVR}) \end{array} \tag{17}$$

For the instrument matrix $A^r(X_i)$ the instruments of the scores of the log likelihood $\frac{\partial \ln l(x'\theta^r)}{\partial \theta^r} = \frac{\phi(x'\theta^r)}{\Phi(x'\theta^r)(1-\Phi(x'\theta^r))} x \cdot (y - \Phi(x'\theta^r))$ are taken, which yields the moment function (10):

$$g_n^r = n^{-1} \sum_i \begin{pmatrix} \frac{\left(Y_i - \Phi(X_i'\theta^r)\right) \cdot \phi(X_i'\theta^r)}{\Phi(X_i'\theta^r)(1-\Phi(X_i'\theta^r))} X_i \cdot 1(D_i = r) \\ \left(\Lambda(X_i) \cdot \Phi(X_i'\theta^r) - \hat{\mathbf{m}}_{VL}^r(\hat{p}^r(X_i))\right) \cdot 1(X_i \in \hat{S}^r) \end{pmatrix}, \tag{18}$$

with $r = \{None, VR, NVR\}$. The upper part of (18) corresponds to the parametric specification (17) while the lower part ensures that the estimated coefficients of the parametric model are in line with nonparametrically estimated average outcomes for various subpopulations. Since by Lemma 3 and Theorem 2 the moment vectors g_n^r are uncorrelated and asymptotically normal, the three potential outcome relationships can be estimated by separate GMM estimators.

Preliminary estimates of $\hat{p}^r(x)$, $\hat{\mathbf{m}}_{VL}^r(p)$ and $\hat{S}^r(x)$ are required for (18), which are contained in Appendix B. The participation probabilities $p^r(\cdot)$ are estimated by maximum likelihood using a flexible multinomial probit model, with no rehabilitation as the reference group.[15] Table B.1 shows the estimated coefficients from which the estimated participation probabilities $\hat{p}^r$ are computed for each observation. Table B.2 provides the correlation coefficients between these estimated participation probabilities, and Figure B.1 displays the distribution of these probabilities. Non-participation and participation in rehabilitation programmes are negatively correlated, while the propensity to vocational or non-vocational rehabilitation are almost uncorrelated. The support regions $S^r(x)$ are approximated by the supports of the estimated participation probabilities $\hat{p}^r$ and are delimited by the minimum and the maximum of the estimated participation probabilities in the respective treatment groups (Table B.3 in the appendix). Table B.3 also shows that almost all observations lie within the estimated supports $\hat{S}^r$.

With these estimated participation probabilities the regression curves $m^r(p^r) = E[Y^r|p^r] = E[Y^r|p^r, D = r]$ are nonparametrically estimated. Using only the observations which participated in

[14] It should be mentioned that among the vocational rehabilitation programmes workplace training is substantially superior to educational programmes (Frölich, Heshmati, and Lechner 2000a).

[15] The multinomial probit model is highly flexible among discrete choice models and particularly does not hinge on the Independence of Irrelevant Alternative assumptions, as e.g. the multinomial logit model does. It is estimated by weighted simulated maximum likelihood using the GHK simulator for simulating the multivariate normal distribution function, with 400 random numbers for each observation and each choice equation (see Börsch-Supan and Hajivassiliou 1993). Using simulators for the multivariate c.d.f. is not strictly necessary with only three choice alternatives, but would become essential were the rehabilitation programmes grouped into more than 4 categories.

programme r, their observed outcomes Y^r are regressed on their estimated participation probabilities p^r by Nadaraya-Watson kernel estimator. These regression curves are also estimated for various subpopulations (Table 5.2), where $m_l^r(p) = E[Y^r|p^r(X) = p, \Lambda_l(X) = 1]$ is estimated by

$$\hat{m}_l^r(p) = \frac{\sum Y_j^r K\left(\frac{\hat{p}^r(X_j)-p}{h}\right) \cdot \Lambda_l(X_j) \cdot 1(D_j = r)}{\sum K\left(\frac{\hat{p}^r(X_j)-p}{h}\right) \cdot \Lambda_l(X_j) \cdot 1(D_j = r)}$$

where K is a symmetric weight function[16] and h is a bandwidth parameter. The bandwidth h is chosen by penalised cross-validation, which performed best in a simulation study of Frölich (2000).[17] In Figure B.2 the estimated $\hat{m}_l^r(\hat{p}^r)$ are plotted within their respective support regions.

For illustration the nonparametrically estimated average potential outcomes are displayed in Table 5.2, which are computed as

$$\hat{E}_{S_r}\left[Y^r|\Lambda_l(X) = 1\right] = \frac{\sum \hat{m}_l^r(\hat{p}^r(X_i)) \cdot \Lambda_l(X_i) \cdot 1(X_i \in \hat{S}^r)}{\sum \Lambda_l(X_i) \cdot 1(X_i \in \hat{S}^r)}, \tag{19}$$

using the results of (??). These average potential outcomes indicate that for most subpopulations no-rehabilitation is the superior treatment and, not surprisingly, non-vocational rehabilitation the worst in re-integrating long-term sick into the labour market. This ranking, however, might not be true for those in need of vocational rehabilitation, for those who have been sick for more than 60 days in the previous six months, and for the age group 36-45 years.

Table 5.2: Nonparametrically estimated average outcomes for selected subpopulations

Population	Observations	$\hat{E}Y^{None}$	$\hat{E}Y^{Vocational}$	$\hat{E}Y^{Non-VR}$
All	6227	47.1	45.1	40.6
Employed	5006	57.0	52.2	50.5
Men	2819	47.6	43.4	40.0
Age 46-55	2260	47.7	45.9	40.3
Manufacturing	1974	50.5	43.6	42.4
Rehabilitation needed	1996	36.3	46.1	37.6
Älvsborgslän	2998	48.9	45.5	42.6
Värmlandslän	1912	52.6	48.2	41.4
Previously long-term sick	1760	44.8	50.6	32.4
Rural community	1350	45.6	40.1	41.4
Age 36-45	1341	36.0	36.5	26.8

Note: Estimated mean outcomes for various subpopulations within the support $\hat{S}^r$. Number of observations corresponds to the smallest number of subpopulation members after trimming at the supports $\hat{S}_l^{None}$, $\hat{S}_l^{VR}$, or $\hat{S}_l^{NVR}$, respectively. Rehabilitation needed refers to the non-medical VR recommendation: VR needed and defined (in Table B.1).

With $\hat{\mathbf{m}}_{VL}^r(p)$ and $\hat{S}^r$ estimated the GMM estimator can be employed to estimate the coefficients θ^{No}, θ^{VR}, θ^{NVR} of the three potential outcome relationships. *The following results are preliminary*

[16] Throughout this study the Epanechnikov kernel $K(u) = \frac{3}{4}\left(1-u^2\right) 1_{[-1,1]}(u)$ is always employed because of its optimality properties (Fan, Gasser, Gijbels, Brockmann, and Engel 1997).

[17] The Akaike penalised cross validation selector chooses the bandwidth which minimises $CV(h) = \sum(Y_i - \hat{m}(\hat{p}(X_i)) \cdot \exp(2/nh)$ (Pagan and Ullah 1999, p. 119). In the simulations of Frölich (2000) a local linear variant of Seifert and Gasser (1996, 2000) actually turned out to be more precise in small samples than the kernel or local linear estimator, while local linear regression was particularly disappointing and performed often even worse than one-to-one pair matching. However, since it would complicate the estimation of the GMM covariance matrix, it is not implemented here.

and exemplary and will be replaced soon. All three outcome equations h^{None}, h^{VR}, h^{NVR} contain the same 5 explanatory variables (plus constant) and 6 equality moments, corresponding to the first six populations in Table 5.2, are included in the GMM estimation. Accordingly the number of moments in (10) is 12 and the number of overidentifying restrictions is 6. The explanatory variables are gender, previously employed, previously long-term sick, sickness degree 100%, and medical diagnosis is injury.

First, neglecting the equality conditions and regressing the observed outcomes on the explanatory variables by standard Probit provides the coefficients θ^r_{Probit} of the purely parametric model as starting values for the 2-step GMM (Tables 5.3a,b,c). The GMM estimates θ^r_1 are computed with the identity matrix I_{12} as weighting matrix. Then the efficient weighting matrix is estimated as $\hat{\Omega}^r_1(\theta^r_1) = [\hat{E}J^r(\theta^r_1)J^{r\prime}(\theta^r_1)]^{-1}$. J^r is computed by formula (13) with the influence functions $\Psi^r_{l,p}$ stemming from the maximum likelihood estimation of the participation probabilities p^r given by (38) in the appendix and the influence functions $\Psi^r_{l,m}$ according to the Nadaraya-Watson kernel regression of $m^r(p^r) = E[Y^r|p^r]$ given by (39). These influence functions take account of the additional variance due to the estimation of the expected outcome conditional on the estimated participation probability and need to be computed for all subpopulations $l \in \{1..L\}$. With this estimated efficient weighting matrix $\hat{\Omega}^r_1$ GMM is re-estimated to obtain $\hat{\theta}^r_2(\hat{\Omega}^r_1)$. To verify the integrity of the GMM estimates the weighting matrix is also re-estimated as $\hat{\Omega}^r_2(\theta^r_2) = [\hat{E}J^r(\theta^r_2)J^{r\prime}(\theta^r_2)]^{-1}$ and standard errors of $\hat{\theta}^r_2$ are computed as well with respect to $\hat{\Omega}^r_1$ as with respect to $\hat{\Omega}^r_2$.

Tables 5.3a,b,c report estimated coefficients and standard errors of the parametric probit and the semiparametric GMM for the three potential outcomes Y^{None}, Y^{VR}, and Y^{NVR}. Except for gender and sickness degree most coefficients are significant with t-values above 2. The first step GMM results θ^r_1 exhibit a similar pattern as the probit estimates θ^r_{Probit}, with (almost) no sign changes occurring between θ^r_{Probit} and θ^r_1. The standard errors of θ^r_1 are generally somewhat larger than those of θ^r_{Probit}. In contrast, the second step GMM results θ^r_2 are surprisingly almost identical to the probit estimates. A closer examination of the estimated weighting matrix $\hat{\Omega}^r_1$ (not reported here) clarified the reasons for this behaviour. The influence functions $\Psi^r_{l,m}$ stemming from the nonparametric estimation of $m^r(p^r)$ increased those elements of $EJ^rJ^{r\prime}$ corresponding to the equality moments of the moment vector (10) drastically, whereas the moments according to the parametric model are unaffected of this, see formula (13). As a consequence, very small weights are placed on the equality moments in the estimate of the weighting matrix $[EJ^rJ^{r\prime}]^{-1}$, leading to estimates governed by the moments of the parametric model. This effect is even more prevalent the smaller the number of observed participants in programme r. Interestingly, the influence $\Psi^r_{l,p}$ of the estimation of the participation probabilities had only a very small impact on the weighting matrix $\hat{\Omega}^r_1$ and on the GMM results $\hat{\theta}^r_1$ and $\hat{\theta}^r_2$, which might have been expected from the Figures B.2 in the appendix where the slopes of $m^r(p^r)$ are small in most cases. However, it is not obvious here whether the efficient second-step GMM estimates are preferable to the first-step GMM estimates. Apart from known numerical problems with the estimation of Ω^r_1 when the number of overidentifying restrictions is large, the weights assigned by Ω^r_1 to the moment conditions may not correspond to the credibility weights, which the econometrician attributes to the parametric and the nonparametric moments, since the weights Ω^r_1 reflect only variance but not bias. Obviously, the variance of the parametric model is lower than that of the nonparametric estimates and if the parametric model is correctly specified Ω^r_1 is right to pick the parametric model. However, the greater robustness of the nonparametric estimates is not incorporated in Ω^r_1 and in case of misspecification of the parametric model placing almost zero weights on the equality conditions does not appear sensible.

Table 5.3a: Coefficients for the potential outcome Y^{None} with 6 equality moments

Variables	θ^{None}_{Probit}	$(Std.)$	θ^{None}_1	$(Std.)$	θ^{None}_2	(Std_1, Std_2)
Constant	-1.87	(0.10)	-1.75	(0.15)	-1.86	(0.10 , 0.10)
Male	0.10	(0.05)	0.07	(0.06)	0.12	(0.05 , 0.05)
Employed	1.78	(0.08)	1.70	(0.12)	1.78	(0.08 , 0.08)
Previously long-term sick	-0.24	(0.06)	-0.38	(0.07)	-0.25	(0.06 , 0.06)
Sickness degree 100%	0.29	(0.06)	0.21	(0.07)	0.27	(0.06 , 0.06)
Diagnosis: Injury	0.54	(0.07)	0.75	(0.08)	0.54	(0.07 , 0.07)

Note: QML robust standard errors of probit estimates θ^{None}_{Probit}. θ^{None}_1 are the GMM estimates with identity matrix as weighting matrix. θ^{None}_2 are the GMM estimates with asymptotically efficient weighting matrix. Standard errors of the GMM estimates are based on the asymptotic covariance matrix. The covariance matrix of θ^{None}_1 is computed as $(G'IG)^{-1}G'I\hat{\Omega}_1^{-1}IG(G'IG)^{-1}$. For θ^{None}_2 two standard errors are reported. The first stems from the covariance matrix $(G'\hat{\Omega}_1 G)^{-1}$ and the second from $(G'\hat{\Omega}_1 G)^{-1}G'\hat{\Omega}_1\hat{\Omega}_2^{-1}\hat{\Omega}_1 G(G'\hat{\Omega}_1 G)^{-1}$.

Table 5.3b: Coefficients for the potential outcome Y^{VR} with 6 equality moments

Variables	θ^{VR}_{Probit}	$(Std.)$	θ^{VR}_1	$(Std.)$	θ^{VR}_2	(Std_1, Std_2)
Constant	-1.10	(0.16)	-1.14	(0.27)	-1.09	(0.16 , 0.16)
Male	-0.06	(0.07)	-0.06	(0.08)	-0.06	(0.07 , 0.07)
Employed	1.13	(0.11)	1.09	(0.15)	1.13	(0.11 , 0.11)
Previously long-term sick	-0.20	(0.08)	-0.15	(0.09)	-0.21	(0.08 , 0.08)
Sickness degree 100%	0.07	(0.12)	0.17	(0.19)	0.08	(0.11 , 0.11)
Diagnosis: Injury	0.25	(0.10)	0.15	(0.13)	0.24	(0.10 , 0.10)

Note: See Table 5.3a

Table 5.3c: Coefficients for the potential outcome Y^{NVR} with 6 equality moments

Variables	θ^{NVR}_{Probit}	$(Std.)$	θ^{NVR}_1	$(Std.)$	θ^{NVR}_2	(Std_1, Std_2)
Constant	-1.67	(0.18)	-2.41	(0.77)	-1.68	(0.20 , 0.18)
Male	0.09	(0.08)	0.02	(0.10)	0.09	(0.08 , 0.08)
Employed	1.76	(0.15)	2.27	(0.62)	1.77	(0.17 , 0.15)
Previously long-term sick	-0.46	(0.10)	-0.48	(0.14)	-0.48	(0.10 , 0.10)
Sickness degree 100%	-0.07	(0.11)	0.19	(0.26)	-0.07	(0.11 , 0.11)
Diagnosis: Injury	0.22	(0.11)	0.43	(0.21)	0.21	(0.11 , 0.11)

Note: See Table 5.3a

To verify the parametric specifications Table 5.4 presents the results of the J-tests for overidentifying restrictions for the first and the second step GMM estimates. The J-tests with respect to the second-step estimates $\hat{\theta}^r_2$ are reported as well with respect to $\hat{\Omega}^r_1$ as with $\hat{\Omega}^r_2$, which should lead to similar conclusions. For the potential outcome in absence of any rehabilitation Y^{None} the parametric specification is clearly rejected at the 1% level for $\hat{\theta}^{None}_1$ and $\hat{\theta}^{None}_2$. For the potential outcome with vocational rehabilitation the parametric specification cannot be rejected at the 10% level as well for $\hat{\theta}^{VR}_1$ as for $\hat{\theta}^{VR}_2$. For non-vocational rehabilitation the J-test strongly rejects with respect to $\hat{\theta}^{NVR}_1$ but does not reject with respect to $\hat{\theta}^{NVR}_2$. This might indicate that the first step GMM estimates do not satisfy the equality conditions but the second step estimates do. However, it might also be an indication that the first step estimates with the identity matrix as weighting matrix have stayed more attached to the nonparametric averages, but with $\hat{\Omega}^r_1$ assigning almost all weight to the parametric model the incompatibility of the first step estimates with the observations on the participants in non-vocational rehabilitation (on which the parametric model is based) becomes evident. For the second step estimates $\hat{\theta}^r_2$ this does not occur, since as well $\hat{\theta}^r_2$ as $\hat{\Omega}^r_1$ nearly ignore the nonparametric estimates and therefore do not reject the J-test. Overall the J-test based on $\hat{\theta}_1^r$ is probably the more reliable indicator of the plausibility of the parametric specification.

Table 5.4: Tests for overidentifying restrictions

	None	Vocational	Non-VR
$n \cdot g_n^r(\hat{\theta}_1^r)'\hat{\Omega}_1^r g_n^r(\hat{\theta}_1^r)$	50.89	7.24	22.39
$n \cdot g_n^r(\hat{\theta}_2^r)'\hat{\Omega}_1^r g_n^r(\hat{\theta}_2^r)$	33.85	5.05	6.25
$n \cdot g_n^r(\hat{\theta}_2^r)'\hat{\Omega}_2^r g_n^r(\hat{\theta}_2^r)$	33.62	5.02	6.64

6 degrees of freedom. $\chi^2_{(6)}$ critical values are: $10.64_{(\alpha=0.10)}$, $12.59_{(\alpha=0.05)}$, $14.45_{(\alpha=0.25)}$, $16.81_{(\alpha=0.01)}$.

On this account it would appear advisable to search for more adequate specifications of the potential outcomes Y^{None} and Y^{NVR}. *This is currently in progress, though for explanatory purposes I continue with the specification of Tables 5.3a,b,c.* Having found a proper specification of the potential outcomes the expected individual treatment outcomes can be computed and the optimal programme (or a set of optimal programmes) can be determined for any individual. Since the probit specification is a strictly increasing function of its argument $x'\theta^r$ the probability that programme r dominates all other programmes for an individual with characteristics x is identical to $\Pr(x'\hat{\theta}^r \geq x'\hat{\theta}^{No}, x'\hat{\theta}^r \geq x'\hat{\theta}^{VR}, x'\hat{\theta}^r \geq x'\hat{\theta}^{NVR})$. These dominance probabilities are computed for all 6287 individuals of the data set and Table 5.5 gives the number of individuals for which no-rehabilitation, vocational rehabilitation, and non-vocational rehabilitation, respectively, are optimal at the significance level specified in the first column. For instance, for 2187 individuals the hypothesis that participation in either vocational or non-vocational rehabilitation would improve their re-employment chances can be rejected at the 5% level and for 1864 individuals even at the 1% level. (This corresponds to their dominance probabilities of no-rehabilitation being larger or equal than 95% and 99%, respectively.) For slightly more than 1000 individuals vocational rehabilitation is their optimal programme and non-vocational rehabilitation is for no-one optimal to foster re-employment prospects at the end of sickness spell.

Table 5.5: Number of individuals for whom programme is optimal

α-level	None	Vocational	Non-VR
20 %	3893	1245	176
10 %	2196	1082	0
5 %	2187	1044	0
1 %	1864	1020	0

The cell entries give the number of the 6287 individuals for which the programme in the respective column is the optimal one, with significance level given in the first column.

When comparing these predicted optimal programmes with the treatments actually received it is found that of the 2187 individuals for whom no-rehabilitation is optimal at the 5% level 1174 received no rehabilitation, while 574 participated in vocational and 439 in non-vocational rehabilitation. Of those 1044 individuals for whom vocational rehabilitation is optimal 599 did not participate in rehabilitation, whereas 193 received vocational and 252 non-vocational rehabilitation. Taken together, 1367 of the 3231 individuals for whom a unique optimal programme could have been determined with high probability received their optimal treatment, while 1864 participated in a sub-optimal programme, implying an efficiency ratio of *42.3%*. This efficiency ratio is virtually identical when derived analogously for the 1, 5, and 20% significance level, respectively.

6 Conclusions and further research

This paper has developed a new semiparametric approach to determining the optimal programme among a number of available programmes for a particular individual, based on observations from past programme participants. Such a statistical system is useful in at least two respects. It can be employed ex ante as a decision support system to provide recommendations when a particular individual faces the choice between various treatments. Such a system can also be used ex post to assess the efficiency of the selection process, which had assigned the past programme participants to the programmes. Measuring the degree to which individuals have received their optimal treatment could be utilized as a performance indicator for monitoring and assessing agencies, which either assign individuals to available programmes or advise them in choosing a programme, e.g. local public employment services. Inefficient participant-to-treatment allocation might suggest a re-organization of the selection process, e.g. from individual self-selection to assignment or vice versa.[18]

The proposed estimator has been applied to rehabilitation programmes in Sweden, which had been grouped into just three categories (no rehabilitation, vocational rehabilitation, non-vocational rehabilitation) on the grounds of small sample size. Only one outcome variable, returning to regular employment at the end of sickness, has been considered. It turned out that for nearly 2200 of the 6287 individuals of the data set no-rehabilitation would have been the optimal treatment, for just above 1000 individuals vocational rehabilitation would have been optimal, and for no-one non-vocational rehabilitation would have been most recommendable. For the remaining over-3000 individuals none of the three treatments was jointly significantly superior to the others at the 10% significance level, but for another approximately 2000 persons an optimal programme could be determined at the 20% significance level. For the remaining persons a semi-ordering could be established by multiple-comparisons-with-the-best. Contrasting these results with the treatments that the individuals have actually received it is found that only about 42% of all individuals for whom an unrivalled optimal programme was found received their optimal treatment. If re-employment at the end of sickness spell would be the only relevant outcome variable of interest (which surely is not the case e.g. for medical rehabilitation) and if the model specification were correct (which in the current version of the paper is not yet the case), one would be tempted to conclude that the selection process in place during 1991 to 1994 in Sweden, assigning about 58% of the easy-to-classify cases to sub-optimal rehabilitation programmes, was highly inefficient.

In future work multiple comparisons with the best shall be incorporated and it shall be addressed which and how many of the overidentifying equality moments should be included in the GMM estimator and how the proposed estimator behaves under misspecification.

A Appendix: Theorems and corollaries

A.1 Proofs of Theorems

Lemma 3 *If the potential outcomes $Y^1, .., Y^R$ are independent, then the moment vectors $g_n^1(\theta^1), ..,$ $g_n^R(\theta^R)$ are uncorrelated.*

Proof. $E\left[(g_n^r(\theta^r) - Eg_n^r(\theta^r)) \cdot (g_n^s(\theta^s) - Eg_n^s(\theta^s))'\right] = EE[(g_n^r(\theta^r) - Eg_n^r(\theta^r)) \cdot (g_n^s(\theta^s) - Eg_n^s(\theta^s))'$ $||X_1, .., X_n, D_1, .., D_n] = n^{-2}\sum_i\sum_j EE\left[\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} |X_1, .., X_n, D_1, .., D_n\right],$

[18] Notice that efficiency is with respect to the relevant utility function. Thus, although self-selection might appear ideal when characteristics such as motivation or work commitment are unobservable to the case worker, it might also be that programme costs and personal monetary and non-monetary costs are valued differently in the utility functions of the individual and that of the programme operator.

with $A_{11} = (A^r(X_i) \cdot (Y_i^r - h^r(X_i,\theta^r)) \cdot 1(D_i = r) - E\left[A^r(X)(Y^r - h^r(X,\theta^r))1(D = r)\right]) \cdot$ $\Big(A^s(X_j) \cdot (Y_j^s - h^s(X_j,\theta^s)) \cdot 1(D_j = s) - E\left[A^s(X)(Y^s - h^s(X,\theta^s))1(D = s)\right]\Big)'$, $A_{22} =$ $\Big((\Lambda(X_i) \otimes h^r(X_i,\theta^r) - \hat{\mathbf{m}}^r_{VL}(\hat{p}^r(X_i))1(X_i \in \hat{S}^r) - E\left[(\Lambda(X) \otimes h^r(X,\theta^r) - \hat{\mathbf{m}}^r_{VL}(\hat{p}^r(X))1(X \in \hat{S}^r)\right]\Big) \cdot$ $\Big((\Lambda(X_j) \otimes h^s(X_j,\theta^s) - \hat{\mathbf{m}}^s_{VL}(\hat{p}^s(X_j))1(X_j \in \hat{S}^s) - E\left[(\Lambda(X) \otimes h^s(X,\theta^s) - \hat{\mathbf{m}}^s_{VL}(\hat{p}^s(X))1(X \in \hat{S}^s)\right]\Big)'$,
$A_{12} = (A^r(X_i) \cdot (Y_i^r - h^r(X_i,\theta^r)) \cdot 1(D_i = r) - E\left[A^r(X)(Y^r - h^r(X,\theta^r))1(D = r)\right]) \cdot$ $\Big((\Lambda(X_j) \otimes h^s(X_j,\theta^s) - \hat{\mathbf{m}}^s_{VL}(\hat{p}^s(X_j))1(X_j \in \hat{S}^s) - E\left[(\Lambda(X) \otimes h^s(X,\theta^s) - \hat{\mathbf{m}}^s_{VL}(\hat{p}^s(X))1(X \in \hat{S}^s)\right]\Big)'$, and A_{21} analogously.

Multiplying out and taking the expectation conditional on $X_1,..,X_n,D_1,..,D_n$ all terms cancel to zero, by noticing that from the independence of the potential outcomes $Y^1,..,Y^R$ it follows with expectations conditional on $X_1,..,X_n$ that $E[(Y^r - h^r(X_i,\theta^r)) \cdot (Y^s - h^s(X_j,\theta^s))'] = E[(Y^r - h^r(X_i,\theta^r))]$ $\cdot\, E[(Y^s - h^s(X_j,\theta^s))']$, $E[(Y^r - h^r(X_i,\theta^r)) \cdot ((\Lambda(X_j) \otimes h^s(X_j,\theta^s) - \hat{\mathbf{m}}^s_{VL}(\hat{p}^s(X_j)))'] = E[(Y^r - h^r(X_i,\theta^r))] \cdot E[((\Lambda(X_j) \otimes h^s(X_j,\theta^s) - \hat{\mathbf{m}}^s_{VL}(\hat{p}^s(X_j)))']$, and $E[((\Lambda(X_i) \otimes h^r(X_i,\theta^r) - \hat{\mathbf{m}}^r_{VL}(\hat{p}^r(X_i)))$ $\cdot\, ((\Lambda(X_j) \otimes h^s(X_j,\theta^s) - \hat{\mathbf{m}}^s_{VL}(\hat{p}^s(X_j)))'] = E[(\Lambda(X_i) \otimes h^r(X_i,\theta^r) - \hat{\mathbf{m}}^r_{VL}(\hat{p}^r(X_i))] \cdot E[(\Lambda(X_j) \otimes h^s(X_j,\theta^s) - \hat{\mathbf{m}}^s_{VL}(\hat{p}^s(X_j))']$, since $\hat{\mathbf{m}}^r_{VL}$ is estimated from the observed outcomes Y^r and $\hat{\mathbf{m}}^s_{VL}$ is estimated from the observed outcomes Y^s. ■

Consistency of the GMM estimator

Recall the moment vector

$$g_n^r(\theta^r, \hat{S}^r, \hat{\mathbf{m}}^r_{VL}) = n^{-1}\sum_i \begin{pmatrix} A^r(X_i) \cdot (Y_i - h^r(X_i,\theta^r)) \cdot 1(D_i = r) \\ (\Lambda(X_i) \otimes h^r(X_i,\theta^r) - \hat{\mathbf{m}}^r_{VL}(\hat{p}^r(X_i)) \cdot 1(X_i \in \hat{S}^r) \end{pmatrix},$$

where $\hat{\mathbf{m}}^r_{VL}(\hat{p}^r(X_i)) = (\hat{m}_1^{r\prime}(\hat{p}^r(X_i) \cdot \Lambda_1(X_i),..,\hat{m}_l^{r\prime}(\hat{p}^r(X_i) \cdot \Lambda_l(X_i),..,\hat{m}_L^{r\prime}(\hat{p}^r(X_i) \cdot \Lambda_L(X_i))'$ is the column vector of length VL of all estimated outcome variables $\hat{m}_l^{r\prime}(\hat{p}^r(X_i) = (\hat{m}^{r\prime}_{1l}(\cdot),..,\hat{m}^{r\prime}_{vl}(\cdot),..,\hat{m}^{r\prime}_{Vl}(\cdot))'$ in all populations $l = 1,..,L$.

Define

$$\hat{\boldsymbol{\alpha}}^r(\hat{S}^r, \Lambda) = n^{-1}\sum_i \hat{\mathbf{m}}^r_{VL}(\hat{p}^r(X_i)) \cdot 1(X_i \in \hat{S}^r),$$

as the column vector of length VL which converges in probability to

$$\underset{n\to\infty}{plim}\hat{\boldsymbol{\alpha}}^r(S^r, \Lambda) = E\left(\Lambda(X) \otimes E[Y^r|X] \cdot 1(X \in S_r)\right) \equiv \boldsymbol{\alpha}_0^r(S^r, \Lambda).$$

Then the moment function can be written as

$$g_n^r(\theta^r, \hat{\boldsymbol{\alpha}}^r) = n^{-1}\sum_i \begin{pmatrix} A^r(X_i) \cdot (Y_i - h^r(X_i,\theta^r)) \cdot 1(D_i = r) \\ \Lambda(X_i) \otimes h^r(X_i,\theta^r) \cdot 1(X_i \in \hat{S}^r) \end{pmatrix} - \begin{pmatrix} \mathbf{0}_k \\ \hat{\boldsymbol{\alpha}}^r(\hat{S}^r, \Lambda) \end{pmatrix}. \tag{20}$$

In analogy to Newey and McFadden (1994) the proof proceeds in three steps. In Corollary 4 sufficient conditions for the consistency of an extremum estimator $\hat{\theta}^r(\hat{\boldsymbol{\alpha}}^r) = \arg\min_{\theta^r \in \Theta^r} \hat{Q}_n^r(\theta^r, \hat{\boldsymbol{\alpha}}^r)$ are laid down. In Corollary 5 these sufficient conditions are specified more precisely for a GMM estimator, which is a member of the class of extremum estimators. Finally, in Theorem 1 it is shown that the specific GMM estimator based on the above moment vector does under mild conditions satisfy the conditions of Corollaries 4 and 5. Since the different coefficient vectors θ^r are estimated separately the superscripts r will henceforth frequently be suppressed to ease notation.

First, sufficient conditions are given for a generic extremum estimator of the form $\hat{\theta}(\hat{\boldsymbol{\alpha}}) = \arg\min \hat{Q}_n(\theta, \hat{\boldsymbol{\alpha}})$ to be consistent, where $\hat{Q}_n$ is a stochastic objective function and $\hat{\boldsymbol{\alpha}}$ a nonparametric preliminary estimate of $\boldsymbol{\alpha}_0$. Be Q_0 the nonstochastic limit function of $\hat{Q}_n$ and θ_0 and $\boldsymbol{\alpha}_0$ the true values. Suppose that the estimator $\hat{\boldsymbol{\alpha}}$ converges in probability to $\boldsymbol{\alpha}_0$. Define $B(\boldsymbol{\alpha}_0)$ as an arbitrarily small ball around $\boldsymbol{\alpha}_0$. Consistency of $\hat{\boldsymbol{\alpha}}$ implies that with probability approaching one $\hat{\boldsymbol{\alpha}}$ lies in the ball $B(\boldsymbol{\alpha}_0)$, hence

$$\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\alpha}} \in B(\boldsymbol{\alpha}_0)) = 1. \tag{21}$$

Corollary 4 *If*
(i) $Q_0(\theta, \boldsymbol{\alpha})$ is uniquely minimized at $(\theta_0, \boldsymbol{\alpha}_0)$,
(ii) $\theta_0 \in \Theta$, with Θ the compact parameter space,
(iii) $Q_0(\theta, \boldsymbol{\alpha})$ is continuous,
(iv) $\hat{Q}_n(\theta, \boldsymbol{\alpha})$ converges uniformly in Θ to $Q_0(\theta, \boldsymbol{\alpha})$ for all $\boldsymbol{\alpha} \in B(\boldsymbol{\alpha}_0)$, hence

$$\lim_{n\to\infty} \Pr\left(\sup_{\theta\in\Theta}\left|\hat{Q}_n(\theta, \boldsymbol{\alpha}) - Q_0(\theta, \boldsymbol{\alpha})\right| < \varepsilon_1\right) = 1 \qquad \forall \boldsymbol{\alpha} \in B(\boldsymbol{\alpha}_0) \qquad \text{with } \varepsilon_1 > 0, \tag{22}$$

(v) plim $\hat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_0$,
then the estimator $\hat{\theta}(\hat{\boldsymbol{\alpha}}) = \arg\min_{\theta\in\Theta} \hat{Q}_n(\theta, \hat{\boldsymbol{\alpha}})$ converges in probability to θ_0.

Proof. With $\hat{\boldsymbol{\alpha}}$ consistent it follows by the Slutzky theorem that also the nonstochastic function $Q_0(\theta_0, \hat{\boldsymbol{\alpha}})$ is convergent:

$$\lim_{n\to\infty} \Pr\left(|Q_0(\theta_0, \hat{\boldsymbol{\alpha}}) - Q_0(\theta_0, \boldsymbol{\alpha}_0)| < \varepsilon_2\right) = 1 \qquad \text{with } \varepsilon_2 > 0 \tag{23}$$

First it is shown, that $Q_0(\hat{\theta}, \hat{\boldsymbol{\alpha}})$ converges to $Q_0(\theta_0, \boldsymbol{\alpha}_0)$ from above. Write $Q_0(\hat{\theta}, \hat{\boldsymbol{\alpha}}) - Q_0(\theta_0, \boldsymbol{\alpha}_0)$ as $\left(Q_0(\hat{\theta}, \hat{\boldsymbol{\alpha}}) - \hat{Q}_n(\hat{\theta}, \hat{\boldsymbol{\alpha}})\right) + \left(\hat{Q}_n(\hat{\theta}, \hat{\boldsymbol{\alpha}}) - \hat{Q}_n(\theta_0, \hat{\boldsymbol{\alpha}})\right) + \left(\hat{Q}_n(\theta_0, \hat{\boldsymbol{\alpha}}) - Q_0(\theta_0, \hat{\boldsymbol{\alpha}})\right) + (Q_0(\theta_0, \hat{\boldsymbol{\alpha}}) - Q_0(\theta_0, \boldsymbol{\alpha}_0))$. From the uniform convergence assumption (22) together with (21) it follows that with probability approaching 1 (w.p.a.1) $\left|\hat{Q}_n(\hat{\theta}, \hat{\boldsymbol{\alpha}}) - Q_0(\hat{\theta}, \hat{\boldsymbol{\alpha}})\right| < \varepsilon_1$ and w.p.a.1 $\left|\hat{Q}_n(\theta_0, \hat{\boldsymbol{\alpha}}) - Q_0(\theta_0, \hat{\boldsymbol{\alpha}})\right| < \varepsilon_1$. From (23) it follows that w.p.a.1 $|Q_0(\theta_0, \hat{\boldsymbol{\alpha}}) - Q_0(\theta_0, \boldsymbol{\alpha}_0)| < \varepsilon_2$. The term $\hat{Q}_n(\hat{\theta}, \hat{\boldsymbol{\alpha}}) - \hat{Q}_n(\theta_0, \hat{\boldsymbol{\alpha}})$ is negative by the definition of the estimator with $\hat{Q}_n(\hat{\theta}, \hat{\boldsymbol{\alpha}}) = \min_{\theta\in\Theta} \hat{Q}_n(\theta, \hat{\boldsymbol{\alpha}})$. Thus the first, third and fourth terms are w.p.a.1 smaller than an arbitrarily small number and the second term is smaller than zero. Accordingly it follows with $\varepsilon \equiv \max(\varepsilon_1, \varepsilon_2)$

$$Q_0(\hat{\theta}, \hat{\boldsymbol{\alpha}}) < Q_0(\theta_0, \boldsymbol{\alpha}_0) + 3\varepsilon \qquad \text{w.p.a.1.} \tag{24}$$

The following reasoning is similar in spirit to that of Theorem 2.1 in Newey and McFadden (1994). Let $\mathcal{N}$ be any open subset of Θ with $\theta_0 \in \mathcal{N}$ and let $\mathcal{N}^c = \Theta \backslash \mathcal{N}$ be its complement. From $\mathcal{N}^c$ compact and $Q_0(\theta, \boldsymbol{\alpha})$ continuous it follows that $\inf_{\theta\in\mathcal{N}^c} Q_0(\theta, \boldsymbol{\alpha}_0) > Q_0(\theta_0, \boldsymbol{\alpha}_0)$, since θ_0 uniquely minimizes Q_0. Choosing $3\varepsilon = \inf_{\theta\in\mathcal{N}^c} Q_0(\theta, \boldsymbol{\alpha}_0) - Q_0(\theta_0, \boldsymbol{\alpha}_0)$ it follows w.p.a.1 that $Q_0(\hat{\theta}, \hat{\boldsymbol{\alpha}}) < \inf_{\theta\in\mathcal{N}^c} Q_0(\theta, \boldsymbol{\alpha}_0)$. This means that w.p.a.1 $\hat{\theta}$ cannot be element of $\mathcal{N}^c$ and thus $\hat{\theta} \in \mathcal{N}$ must hold. Hence for sufficiently small ε all open subsets of Θ which contain θ_0 also w.p.a.1 contain $\hat{\theta}$ and all subsets of Θ which do not contain θ_0 also w.p.a.1 do not contain $\hat{\theta}$. Thus $\hat{\theta}$ converges in probability to θ_0. ■

Now the sufficient conditions of Corollary 4 are specified for a generic GMM estimator.

Corollary 5 *Suppose*
(i) $\hat{\boldsymbol{\alpha}}$ *is a consistent estimator of* $\boldsymbol{\alpha}_0$ *and* $B(\boldsymbol{\alpha}_0)$ *a ball around* $\boldsymbol{\alpha}_0$*, such that* $\lim_{n\to\infty} \Pr(\hat{\boldsymbol{\alpha}} \in B(\boldsymbol{\alpha}_0)) = 1$*,*
(ii) the data $Z_i = (X_i, Y_i)$ *are iid,* $\hat{W} \xrightarrow{p} W$*, where* W *a positive semidefinite matrix,*
(iii) $WEg(Z, \theta, \boldsymbol{\alpha}) = 0$ *if and only if* $\theta = \theta_0$ *and* $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$*,*
(iv) $\theta_0 \in \Theta$*, with* Θ *compact,*
(v) $g(Z, \theta, \boldsymbol{\alpha})$ *is continuous in* θ *and* $\boldsymbol{\alpha}$
(vi) $E\left(\sup_{\alpha\in B(\alpha_0)}\sup_{\theta\in\Theta}\|g(Z,\theta,\boldsymbol{\alpha})\|\right) < \infty$*,*
then the GMM estimator of $\hat{Q}_n(\theta, \hat{\boldsymbol{\alpha}}) = \left(\frac{1}{n}\sum g(Z_i, \theta, \hat{\boldsymbol{\alpha}})\right)' \hat{W} \left(\frac{1}{n}\sum g(Z_i, \theta, \hat{\boldsymbol{\alpha}})\right)$ *with limit function* $Q_0(\theta, \boldsymbol{\alpha}) = (Eg(Z, \theta, \boldsymbol{\alpha}))' W (Eg(Z, \theta, \boldsymbol{\alpha}))$ *satisfies the conditions of Corollary 4 and the GMM estimator is consistent.*

Proof. Showing that the conditions of Corollary 4 are satisfied follows with only minor modifications Lemma 2.4 and Theorem 2.6 of Newey and McFadden (1994) and is here omitted. ■

Proof. [of Theorem 1] It must be shown, that all conditions of Corollary 5 are satisfied. Conditions (i), (ii), (iv) and (v) of Corollary 5 are satisfied by assumption. If $\hat{m}$ is consistent then also $\hat{\boldsymbol{\alpha}}$. The identification condition (iii) and the moment existence condition (vi) need to be analysed. Condition (iii) requires that W is strictly positive definite. Then the upper part of the moment vector (20), which is independent of $\boldsymbol{\alpha}$, can only have expectation zero if it represents the true mean function. By assumption (ii) on the parametric equation $h(x, \theta)$ this can only be the case if $\theta = \theta_0$. But then the equality conditions, the lower part of (20), can only be zero if $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$.

Turning now to the moment existence condition (vi):

$$\begin{aligned}
& E \sup_{\alpha\in B(\alpha_0)}\sup_{\theta\in\Theta}\|g(Z,\theta,\boldsymbol{\alpha})\| \\
= \;& E \sup_{\alpha\in B(\alpha_0)}\sup_{\theta\in\Theta}\left\| n^{-1}\sum_i \begin{pmatrix} A(X_i)\cdot(Y_i - h(X_i,\theta))\cdot 1(D_i = r) \\ \Lambda(X_i)\otimes h(X_i,\theta)\cdot 1(X_i\in\hat{S}^r) - \hat{\boldsymbol{\alpha}} \end{pmatrix}\right\| \\
\approx \;& E \sup_{\alpha\in B(\alpha_0)}\sup_{\theta\in\Theta}\left\| n^{-1}\sum_i \begin{pmatrix} A(X_i)\cdot(Y_i - h(X_i,\theta))\cdot 1(D_i = r) \\ \Lambda(X_i)\otimes h(X_i,\theta)\cdot 1(X_i\in S^r) - \hat{\boldsymbol{\alpha}} \end{pmatrix}\right\| \\
\leq \;& E \sup_{\alpha\in B(\alpha_0)}\sup_{\theta\in\Theta}\begin{pmatrix} \|A(X)\cdot(Y - h(X,\theta))\cdot 1(D = r)\| \\ +\sum_v^V\sum_l^L |\Lambda_l(X)\cdot h_v(X,\theta)\cdot 1(X\in S^r) - \hat{\boldsymbol{\alpha}}_{vl}| \end{pmatrix},
\end{aligned}$$

where $h_v(X, \theta)$ is the v-th outcome variable, $\Lambda_l(X)$ the l-th element of the multivariate indicator function (7) and $\hat{\boldsymbol{\alpha}}_{vl}$ the element of the estimated $\hat{\boldsymbol{\alpha}}$-vector corresponding to the v-th outcome variable and the l-th subpopulation.

$$\leq E \sup_{\alpha\in B(\alpha_0)} \sup_{\theta\in\Theta} \left(\begin{array}{c} \|A(X)\cdot(Y-h(X,\theta))\| \\ +\sum_v^V\sum_l^L |\Lambda_l(X)\cdot h_v(X,\theta)\cdot 1(X\in S^r) - \boldsymbol{\alpha}_{vl,0}| + |\boldsymbol{\alpha}_{vl,0} - \hat{\boldsymbol{\alpha}}_{vl}| \end{array} \right)$$

$$\leq E\sup_{\theta\in\Theta} \|A(X)\cdot(Y-h(X,\theta))\| + \sum_v^V\sum_l^L E\sup_{\theta\in\Theta} |\Lambda_l(X)\cdot h_v(X,\theta)\cdot 1(X\in S^r) - \boldsymbol{\alpha}_{vl,0}|$$
$$+E\sup_{\alpha\in B(\alpha_0)} |\boldsymbol{\alpha}_{vl,0} - \hat{\boldsymbol{\alpha}}_{vl}|,$$

if all these terms have finite expectations. The last term is finite, since the size of the ball $B(\boldsymbol{\alpha}_0)$ around $\boldsymbol{\alpha}_0$ can be chosen arbitrarily small. Hence, if for each outcome variable and each subpopulation $E\sup\|A(X)\cdot(Y-h(X,\theta))\| < \infty$ and $E\sup|h(X,\theta)\cdot 1(X\in S_r) - \boldsymbol{\alpha}_0| < \infty$ and the number of subpopulations is finite, then the moment existence condition (vi) of Corollary 5 is satisfied. ■

Definition 1 ***Asymptotic linearity with trimming*** *Heckman, Ichimura, and Todd (1998)*
An estimator $\hat{\theta}(x)$ of the conditional expectation function $\theta_0(x) = E[Y|X=x]$ is asymptotically linear with trimming $1(x\in\hat{S})$ iff there is a function ψ_n and stochastic terms $\hat{b}(x)$ and $\hat{R}(x)$ that satisfy the following conditions:
(i) $[\hat{\theta}(x) - \theta_0(x)]\cdot 1(x\in\hat{S}) = n^{-1}\sum_{j=1}^{n}\psi_n(Y_j,X_j;x) + \hat{b}(x) + \hat{R}(x)$
(ii) $E[\psi_n(Y_j,X_j;X)|X=x] = 0$
(iii) $\underset{n\to\infty}{plim} n^{-\frac{1}{2}}\sum_{j=1}^{n}\hat{b}(X_j) = b < \infty$
(iv) $n^{-\frac{1}{2}}\sum_{j=1}^{n}\hat{R}(X_j) = o_p(1)$

Condition (i) demands that the deviation of the estimator from its true mean can be decomposed into three components: the sum of a local influence function, a local bias term and a local residual term. The local influence function ψ_n represents the influence of a particular observation on the deviation of the estimator from its true value and is well known from the class of asymptotically linear estimators as covered in Newey and McFadden (1994). Observe that the local influence function ψ_n is allowed to vary with n for instance through bandwidth parameters that converge with sample size. The second term $b(x)$ is a local bias component, whose population average must be zero and the limit distribution of its average multiplied with $\sqrt{n}$ must be degenerate and converging to a finite nonstochastic number b. Parametric estimators are often locally unbiased with $b(x)$ zero. For many nonparametric estimators the $\sqrt{n}$ average bias could be made to be zero by choosing a bandwidth sequence that converges faster to zero, though in terms of mean squared error even more smoothing would be desirable, but this would imply a non-degenerate asymptotic distribution of the $\sqrt{n}$ average bias. Heckman, Ichimura, and Todd (1998) have shown that under some smoothness conditions the local polynomial regression estimator is asymptotically linear with trimming (see Corollary 6), and that the local polynomial regression estimator $\hat{m}^r(\hat{p}^r(x))$ of $E[Y^r|p^r(X=x)]$ on the estimated participation probability is also asymptotically linear with trimming if the estimator of the participation probability model is asymptotically linear with trimming (Corollary 7).

Proof. [of Theorem 2] To improve readability the superscripts r are frequently suppressed. The GMM estimator $\hat{\theta}_n = \arg\min_{\theta} g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{S})' \hat{W}_n g_n(\theta, \hat{\mathbf{m}}_{VL}, \hat{S})$ can be expressed by its first order condition

$$G_n(\hat{\theta}, \hat{S})' \hat{W} \cdot g_n(\hat{\theta}, \hat{\mathbf{m}}_{VL}, \hat{S}) = 0, \tag{25}$$

where $G_n = \frac{\partial g_n(\hat{\theta}, \hat{\mathbf{m}}_{VL}, \hat{S})}{\partial \theta'}$ is the gradient of g_n with respect to θ,which does not depend on $\hat{\mathbf{m}}_{VL}$.

Applying the mean value theorem to $g_n(\hat{\theta}, \hat{\mathbf{m}}^r_{VL}, \hat{S})$ around the true coefficient vector θ_0 and inserting this into the first order condition (25) yields, with $\bar{\theta}$ on the line between $\hat{\theta}$ and θ_0,

$$G_n(\hat{\theta}, \hat{S})' \hat{W} \cdot \Big[g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{S}) + G_n(\bar{\theta}, \hat{S}) \cdot (\hat{\theta} - \theta_0) \Big] = 0. \tag{26}$$

Solving for $\hat{\theta} - \theta_0$ gives

$$n^{\frac{1}{2}}(\hat{\theta} - \theta_0) = -\Big(G_n(\hat{\theta}, \hat{S})' \hat{W} G_n(\bar{\theta}, \hat{S}) \Big)^{-1} G_n(\hat{\theta}, \hat{S})' \hat{W} \cdot n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{S}). \tag{27}$$

By inserting the moment vector (10) the last term can be written as

$$\begin{aligned} n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{S}) = {} & n^{-\frac{1}{2}} \sum g(Z_i, \theta_0, \mathbf{m}_{VL}, S) \qquad (28) \\ & + n^{-\frac{1}{2}} \sum \begin{pmatrix} \mathbf{0}_k \\ (\Lambda(X_i) \otimes h(X_i, \theta_0) - \mathbf{m}_{VL}(p(X_i))) \cdot \Big[1(X_i \in \hat{S}) - 1(X_i \in S) \Big] \end{pmatrix} \\ & - n^{-\frac{1}{2}} \sum_i \begin{pmatrix} \mathbf{0}_k \\ (\hat{\mathbf{m}}_{VL}(\hat{p}(X_i)) - \mathbf{m}_{VL}(p(X_i))) \cdot 1(X_i \in \hat{S}^r) \end{pmatrix}, \end{aligned}$$

where $Z_i = (Y_i, D_i, X_i)$. The first term corresponds to the case where $\mathbf{m}_{VL}$ and S where known, the second term corrects for the estimation of the support and the third term takes account for the estimation of m and p. The second term goes to zero by assumption (v), since $n^{-\frac{1}{2}} \sum(\Lambda(X_i) \otimes h(X_i, \theta_0) - \mathbf{m}_{VL}(p(X_i))) \cdot 1(X_i \in \hat{S})$ and $n^{-\frac{1}{2}} \sum(\Lambda(X_i) \otimes h(X_i, \theta_0) - \mathbf{m}_{VL}(p(X_i)) \cdot 1(X_i \in S)$ are asymptotically equivalent as shown by Heckman, Ichimura, and Todd (1998, p. 291). Since the nonparametric estimator is asymptotically linear with trimming for each outcome variable and in all relevant subpopulations it follows that

$$\begin{aligned} n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{S}) = n^{-\frac{1}{2}} \sum g(Z_i, \theta_0, \mathbf{m}_{VL}, S) + o_p(1) \qquad (29) \\ -n^{-\frac{1}{2}} \sum_i \begin{pmatrix} \mathbf{0}_k \\ n_{1,r}^{-1} \sum_j \Psi_{1,m}(Y_j, D_j, X_j; X_i) + n^{-1} \sum_j \Psi_{1,p}(Y_j, D_j, X_j; X_i) + \hat{b}_1(X_i) + \hat{R}_1(X_i) \\ \vdots \\ n_{L,r}^{-1} \sum_j \Psi_{L,m}(Y_j, D_j, X_j; X_i) + n^{-1} \sum_j \Psi_{L,p}(Y_j, D_j, X_j; X_i) + \hat{b}_L(X_i) + \hat{R}_L(X_i) \end{pmatrix}. \end{aligned}$$

Consider first the elements corresponding to population l of the latter term in (29)

$$\frac{n^{-\frac{1}{2}}}{n_{l,r}} \sum_i \sum_j \Psi_{l,m}(Y_j, D_j, X_j; X_i) + n^{-\frac{3}{2}} \sum_i \sum_j \Psi_{l,p}(Y_j, D_j, X_j; X_i) + n^{-\frac{1}{2}} \sum_i \hat{b}_l(X_i) + n^{-\frac{1}{2}} \sum_i \hat{R}_l(X_i), \tag{30}$$

which can be reformulated as

$$\begin{aligned} = \;& \frac{n^{\frac{3}{2}}}{2n_{l,r}} \left\{ n^{-2} \sum_i \sum_j \left(\Psi^{j,i}_{l,m} + \Psi^{i,j}_{l,m} \right) \right\} + \frac{n^{\frac{1}{2}}}{2} \left\{ n^{-2} \sum_i \sum_j \left(\Psi^{j,i}_{l,p} + \Psi^{i,j}_{l,p} \right) \right\} \\ & + n^{-\frac{1}{2}} \sum_i \hat{b}_l(X_i) + n^{-\frac{1}{2}} \sum_i \hat{R}_l(X_i) \end{aligned}$$

with $\Psi_{l,m}(Y_j, D_j, X_j; X_i)$ abbreviated as $\Psi^{j,i}_{l,m}$ and $\Psi_{l,p}(Y_j, D_j, X_j; X_i)$ as $\Psi^{j,i}_{l,p}$. Notice that the last term converges to zero and the third term to a nonstochastic bias term. The first two terms in curly brackets are von Mises statistics (e.g. see Serfling 1980) which are asymptotically equivalent to their corresponding U-statistics and if $E \left\| \Psi^{j,i}_{l,m} + \Psi^{i,j}_{l,m} + \Psi^{j,i}_{l,p} + \Psi^{i,j}_{l,p} \right\|^2 = o(n)$ holds, then these terms are asymptotically equivalent to the projection of the corresponding U-statistic (Corollary 8). Assumption (i) ensures that this condition is satisfied. By assumption (i) $E \left\| \Psi^{j,i}_{l,m} + \Psi^{i,j}_{l,m} \right\|^2 = o(n)$ since with Euclidean norm this term equals $E \left\| \Psi^{j,i}_{l,m} + \Psi^{i,j}_{l,m} \right\|^2 = E \sum_{v=1}^{V} \left(\Psi^{j,i}_{vl,m} + \Psi^{i,j}_{vl,m} \right)^2$ where $\Psi^{j,i}_{vl,m}$ denotes the influence function on outcome variable v in subpopulation l. With $E\Psi^{j,i}_{l,m} = 0$ this term corresponds to the sum of the variances $\sum_v Var\left(\Psi^{j,i}_{vl,m} + \Psi^{i,j}_{vl,m} \right) \leq \sum_v 4Var\left(\Psi^{j,i}_{vl,m} \right) \leq 4V \cdot \max_v Var\left(\Psi^{j,i}_{vl,m} \right)$. As assumption (i) is to hold for each outcome variable and each subpopulation this term is $o(n)$. The same reasoning does hold analogously for $E \left\| \Psi^{j,i}_{l,p} + \Psi^{i,j}_{l,p} \right\|^2$ and since covariances between the influence functions for p and m are bounded by their variances the projection theorem is applicable.

Replacing the von Mises statistics by the projection of the corresponding U-statistics expression (30) can be stated as

$$\begin{aligned} = \;& \frac{n^{\frac{1}{2}}}{n_{l,r}} \sum_{i=1}^{n} E[\Psi_{l,m}(Y_i, D_i, X_i; X_j) | Y_i, D_i, X_i] + n^{-\frac{1}{2}} \sum_{i=1}^{n} E[\Psi_{l,p}(Y_i, D_i, X_i; X_j) | Y_i, D_i, X_i] \qquad (31) \\ & + n^{-\frac{1}{2}} \sum_i \hat{b}_l(X_i) + n^{-\frac{1}{2}} \sum_i \hat{R}_l(X_i) + o_p\left(\frac{n}{n_{l,r}} \right) + o_p(1), \end{aligned}$$

to which a central limit theorem is directly applicable, where the asymptotic distribution is determined by the first two terms since all other terms converge in probability

To be able to apply the U-statistic projection theorem simultaneously to all subpopulations l in expression (29) requires that the full influence function vector containing all outcome variables and all subpopulations has expected squared norm of order $o(n)$, which is satisfied by assumption (i) since the squared norm is smaller or equal than the sum of the squared norms of all vector elements, which by assumption are $o(n)$.

Defining J_i as

$$J_i = g(Z_i, \theta_0, \mathbf{m}_{VL}) - \begin{pmatrix} \mathbf{0}_k \\ \lambda_{1,r}^{-1} \cdot E[\Psi_{1,m}(Y_i, D_i, X_i; X_j) | Y_i, D_i, X_i] + E[\Psi_{1,p}(Y_i, D_i, X_i; X_j) | Y_i, D_i, X_i] \\ \vdots \\ \lambda_{L,r}^{-1} \cdot E[\Psi_{L,m}(Y_i, D_i, X_i; X_j) | Y_i, D_i, X_i] + E[\Psi_{L,p}(Y_i, D_i, X_i; X_j) | Y_i, D_i, X_i] \end{pmatrix},$$

it follows by the multivariate Lindeberg-Feller central limit theorem (Greene 1997, Theorem 4.14) under the regularity conditions that $EJ_iJ_i' < \infty \; \forall i$, that all mixed third moments of the multivariate distribution are finite, that $\lim\limits_{n\to\infty} n^{-1}\sum\limits_{i=1}^{n} EJ_iJ_i' = EJJ'$ a finite and positive definite matrix, and that $\lim\limits_{n\to\infty}\left(\sum\limits_{i=1}^{n} EJ_iJ_i'\right)^{-1} EJ_iJ_i' = 0 \; \forall i$, that the moment function (10) $g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{S})$ is asymptotically normal:

$$n^{\frac{1}{2}} g_n(\theta_0, \hat{\mathbf{m}}_{VL}, \hat{S}) \xrightarrow{d} N\left(\begin{bmatrix} \mathbf{0}_k \\ \mathbf{b}_{VL} \end{bmatrix}, EJJ'\right). \tag{32}$$

It remains to show that G_n and $\hat{W}$ converge in probability to G and W, respectively. The gradient of the moment vector (10) is

$$G_n(\theta, \hat{S}) = n^{-1} \sum \begin{pmatrix} A(X_i)\cdot(Y_i - \frac{\partial h(X_i,\theta)}{\partial\theta'})\cdot 1(D_i = r) \\ \Lambda(X_i) \otimes \frac{\partial h(X_i,\theta)}{\partial\theta'} \cdot 1(X_i \in \hat{S}) \end{pmatrix} \tag{33}$$

$$= n^{-1} \sum \begin{pmatrix} A(X_i)\cdot(Y_i - \frac{\partial h(X_i,\theta)}{\partial\theta'})\cdot 1(D_i = r) \\ \Lambda(X_i) \otimes \frac{\partial h(X_i,\theta)}{\partial\theta'} \cdot 1(X_i \in S) \end{pmatrix} + \begin{pmatrix} \mathbf{0}_k \\ \Lambda(X_i) \otimes \frac{\partial h(X_i,\theta)}{\partial\theta'} \cdot \Big[1(X_i \in \hat{S}) - 1(X_i \in S)\Big] \end{pmatrix}.$$

The latter term converges to zero since the first derivative of $h(.)$ is bounded by assumption and $1(X_i \in \hat{S})$ converges to $1(X_i \in S)$. The first term converges to G and the weighting matrix converges to W by assumption. Hence the GMM estimator is asymptotically normal

$$n^{\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(\begin{bmatrix} \mathbf{0}_k \\ -(G'WG)^{-1}G'W\mathbf{b}_{VL} \end{bmatrix}, (G'WG)^{-1}G'WE[JJ']WG(G'WG)^{-1}\right).$$

■

Corollary 6 (Asymptotic linearity of $\hat{m}^r(p^r)$, Heckman, Ichimura, and Todd (1998))
Assuming that
(i) sampling of (Y_j^r, X_j, D_j) is iid with finite variance of Y_j^r, and $X_j \in \Re^k$
(ii) the regression function $m^r(p^r)$ is twice continuously differentiable with second derivative Hölder continuous,
(iii) the stochastic bandwidth sequence a_{n_r} satisfies $\operatorname{plim}\limits_{n_r\to\infty} \frac{a_{n_r}}{h_{n_r}} = a_0 > 0$ for some deterministic sequence $\{h_{n_r}\}$ that satisfies $\frac{n_r h_{n_r}}{\ln n_r} \to \infty$ and $\lim n_r h^4 < \infty$,
(iv) the kernel function K is compact and symmetric, $\int K(u)du = 1$, $\int uK(u)du = 0$.
(v) the estimated support $\hat{S}^r = \{x : \hat{f}_X(x) \geq q_0\}$ is estimated such, that $\sup_{x\in S}\left|\hat{f}_X(x) - f_X(x)\right|$ converges a.s. to zero where $S = \{x : f_X(x) \geq q_0\}$, $\hat{f}_X$ is a kernel density estimate with kernel with moments 1 through k equal to zero, and f_X is $k+1$ times continuously differentiable with $(k+1)$-th derivative Hölder continuous,
(vi) $m^r(\cdot)$ is estimated at interior points,
then the local polynomial regression estimator $\hat{m}^r(p^r(x))$ of polynomial order ≤ 1[19] is asymptotically

[19] The local polynomial regression estimator of order 0 is the Nadaraya-Watson Kernel estimator and the local polynomial regression estimator of order 1 is the local linear estimator.

linear with trimming:

$$(\hat{m}^r(p^r(x)) - m^r(p^r(x)))\,1(x \in \hat{S}^r) = n_r^{-1}\sum_j \psi_m(Y_j^r, p^r(X_j); p^r)\cdot 1(D_j = r) + \hat{b}_m(p^r) + \hat{R}_m(p^r).$$

Corollary 7 (Asymptotic linearity of $\hat{m}^r(\hat{p}^r)$, Heckman, Ichimura, and Todd (1998)) *If (i) an estimator $\hat{p}^r(x)$ of the participation probability is asymptotically linear with trimming*

$$(\hat{p}^r(x) - p^r(x))\,1(x \in \hat{S}^r) = n^{-1}\sum_j \psi_p(D_j, X_j; x) + \hat{b}_p(x) + \hat{R}_p(x),$$

(ii) $\frac{\partial \hat{m}^r(p^r)}{\partial p^r}$ and $\hat{p}^r(x)$ are uniformly consistent and converge to $\frac{\partial m^r(p^r)}{\partial p^r}$ and $p^r(x)$, with $\frac{\partial m^r(p^r)}{\partial p^r}$ continuous,
(iii) $\operatorname*{plim}_{n_r\to\infty} n_r^{-\frac{1}{2}}\sum_j \hat{b}_m(p^r(X_j))1(D_j = r) = b_m$, (iv) $\operatorname*{plim}_{n\to\infty} n^{-\frac{1}{2}}\sum_j \frac{\partial m^r(p^r(X_j))}{\partial p^r}\cdot \hat{b}_p(p^r(X_j)) = b_{m_p}$,
(v) $\operatorname*{plim}_{n\to\infty} n^{-\frac{1}{2}}\sum_j \left[\frac{\partial \hat{m}^r(\bar{p}^r(X_j))}{\partial p^r} - \frac{\partial m^r(p^r(X_j))}{\partial p^r}\right]\cdot \hat{R}_p(X_j) = 0$,
(vi) $\operatorname*{plim}_{n\to\infty} n^{-\frac{3}{2}} \sum_l\sum_j \left[\frac{\partial \hat{m}^r(\bar{p}^r(X_j))}{\partial p^r} - \frac{\partial m^r(p^r(X_j))}{\partial p^r}\right]\cdot \psi_p(D_l, X_l; X_j) = 0$, where $\bar{p}^r(x)$ is a function defined by a Taylor's expansion of $\hat{m}^r(\hat{p}^r(x))$ around $p^r(x)$,
then also the estimator $\hat{m}^r(\hat{p}^r(x))$ of $m^r(p^r(x)) = E[Y^r|p^r(X = x)]$ is asymptotically linear with trimming: $[\hat{m}^r(\hat{p}^r(x)) - m^r(p^r(x))]\cdot 1(x \in \hat{S}_r)$

$$= n_r^{-1}\sum_j \psi_m(Y_j^r, p^r(X_j); p^r)1(D_j = r) + \frac{\partial m^r(p^r(x))}{\partial p^r}\cdot n^{-1}\sum_j \psi_p(D_j, X_j; x) + \hat{b}(x) + \hat{R}(x) \qquad (34)$$

and
$\operatorname*{plim}_{n\to\infty} n^{-\frac{1}{2}}\sum_j \hat{b}(X_j) = b_m + b_{mp}$

Remark 5 *Corollary 7: If the participation probability is estimated either nonparametrically by local polynomial regression or parametrically, e.g. by maximum likelihood, then the conditions (i) to (vi) are satisfied (Heckman, Ichimura, and Todd 1998). In the latter case the local bias $\hat{b}_p(x)$ is zero.*

Corollary 8 (Asymptotic equivalence of V-statistic, U statistic and its projection) *Let $H_n(x_1, x_2)$ be a* symmetric *function, $X_1, .., X_n$ iid random vectors. A natural estimator of EH_n is the one-sample U-statistic*

$$U_n = \binom{n}{2}^{-1}\sum_{1\le i<j\le n} H_n(X_i, X_j)$$

The associated von Mises statistic is

$$V_n = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} H_n(X_i, X_j)$$

and the projection of the U-statistic is defined as

$$\hat{U}_n = \frac{n-2}{n}EH_n + \frac{2}{n}\sum_{i=1}^{n} EH_n(X_1, X_2|X_1 = X_i).$$

If $E\,\|H_n\|^2 = o(n)$ then $n^{\frac{1}{2}}(U_n - \hat{U}_n) = o_p(1)$ and $n^{\frac{1}{2}}(V_n - \hat{U}_n) = o_p(1)$. See Hoeffding (1948) and Serfling (1980). Extended by Powell, Stock, and Stoker (1989) to allow H_n to depend on sample size.

A.2 Influence functions of particular estimators

First the influence function for the parametric estimation of the participation probabilities p^r is derived. Thereafter the influence function of kernel regression estimator of m^r is given. As shown in Newey and McFadden (1994, p.2141 ff) Maximum Likelihood estimators of parametric regression models are asymptotically linear in the sense

$$n^{\frac{1}{2}}(\hat{\beta}-\beta_0) = n^{-\frac{1}{2}}\sum_{j=1}^{n}\psi(Y_j,X_j)+o_p(1),$$

with 'global' influence function

$$\psi(Y_j,X_j) = -\left[E\frac{\partial^2 \ln f(Y_j,X_j|\beta_0)}{\partial\beta\partial\beta'}\right]^{-1}\frac{\partial \ln f(Y_j,X_j|\beta_0)}{\partial\beta}. \tag{35}$$

This influence function is global in the sense that it affects the estimate of the coefficients β and not only the estimate of the conditional mean at a certain point x. Let $\mu(x,\beta)$ denote a parametric mean function, then under some regularity conditions the 'local' asymptotically linear representation can immediately be obtained by an expansion of $\mu(x,\beta)$. Under some regularity conditions it follows that

$$\begin{aligned} n^{-\frac{1}{2}}\left(\mu(x,\hat{\beta})-\mu(x,\beta_0)\right) &= n^{-\frac{1}{2}}\frac{\partial\mu(x,\beta_0)}{\partial\beta'}\sum_{j=1}^{n}\psi(Y_j,X_j)+o_p(1) \qquad (36)\\ &= -n^{-\frac{1}{2}}\frac{\partial\mu(x,\beta_0)}{\partial\beta'}\left[EH\right]^{-1}\sum_{j=1}^{n}\frac{\partial \ln f(Y_j,X_j;\beta_0)}{\partial\beta}+o_p(1), \end{aligned}$$

where EH is the expected Hessian at β_0 in (35). Thus a Maximum Likelihood estimator is asymptotically linear with trimming with zero local bias $\hat{b}_p(x)=0$.

The influence function of the kernel regression estimator in the one-dimensional regression setting is (Heckman, Ichimura, and Todd 1998):

$$\begin{aligned} \psi_m(Y_j^r,p^r(X_j);p^r(x)) &= \left(Y_j^r - E[Y_j^r|p^r(X_j),D_j=r]\right)\frac{K\left(\frac{p^r(X_j)-p^r}{h_{n_r}}\right)}{\underset{X_j|D=r}{E}K\left(\frac{p^r(X_j)-p^r}{h_{n_r}}\right)}1(D_j=r)1(x\in S^r)\\ &= \left(Y_j^r - m^r(p^r(X_j))\right)\frac{1}{h_{n_r}}\frac{K\left(\frac{p^r(X_j)-p^r}{h_{n_r}}\right)}{E\hat{f}_{p^r|D=r}(p^r(x))}1(D_j=r)1(x\in S^r), \end{aligned}$$

since $h^{-1}EK\left(\frac{p^r(X_j)-p^r(x)}{h}\right)=\int h^{-1}K\left(\frac{p^r(u)-p^r(x)}{h}\right)\cdot f_{p|D=r}(p^r(u))du$ where $f_{p|D=r}$ is the density of p^r conditional on $D=r$ and $\hat{f}_{p^r|D=r}(\cdot)$ denotes a kernel density estimate using the same bandwidth h_{n_r}. Noting that by continuity of the density $\int h^{-1}K\left(\frac{p^r(u)-p^r(x)}{h}\right)\cdot f_{p|D=r}(p^r(u))du$ converges to $f_{p|D=r}(p^r(x))\cdot\int K(u)\,du$ (Pagan and Ullah (1999), p. 362, 364 or Parzen (1962)) and since the kernel function is supposed to integrate to one, the influence function converges to:

$$\psi_m(Y_j^r,p^r(X_j);p^r(x)) \longrightarrow \left(Y_j^r - m^r(p^r(X_j))\right)\frac{1}{h_{n_r}}\frac{K\left(\frac{p^r(X_j)-p^r}{h_{n_r}}\right)}{f_{p^r|D=r}(p^r(x))}1(D_j=r)1(x\in S^r).$$

The corresponding bias function $\hat{b}_m(p^r)$ is

$$b_m(p^r) = h_{n_r}^2 \left(m^{r(1)}(p^r)\frac{f^{(1)}_{p^r|D=r}(p^r)}{f_{p^r|D=r}(p^r)} + \frac{1}{2}m^{r(2)} \right),$$

where $f^{(1)}$ denotes the first derivative of the density and $m^{r(1)}$ and $m^{r(2)}$ are the first and second derivative, respectively, of m^r.

A.3 Combination of both estimators:

From Appendix A.2 it follows that if $m^r(p^r(x))$ is estimated for a subpopulation defined as $\{x|\Lambda_l(x) = 1\}$ by a combination of Maximum Likelihood estimation for the participation probability and Kernel regression for the conditional regression curve and the assumptions of corollaries 2.1 and 2.2 are valid for this subpopulation, then the resulting estimate $\hat{m}_l^r(\hat{p}^r(x))$ of $E[Y^r|p^r(X = x), \Lambda_l(x) = 1]$ is asymptotically linear with trimming:

$$\begin{aligned}&[\hat{m}_l^r(\hat{p}^r(x)) - m_l^r(p^r(x))] \cdot \Lambda_l(x)1(x \in \hat{S}^r)\\ &\quad= n_{l,r}^{-1}\sum_j \Psi^r_{l,m}(Y_j^r, D_j, X_j; x) + n^{-1}\sum_j \Psi^r_{l,p}(Y_j^r, D_j, X_j; x) + \hat{b}_l^r(x) + \hat{R}_l^r(x), \qquad (37)\end{aligned}$$

with

$$\Psi^r_{l,p}(Y_j^r, D_j, X_j; x) = -\frac{\partial m_l^r(p^r(x))}{\partial p^r}\frac{\partial p^r(x,\beta_0)}{\partial \beta'}[EH]^{-1}\frac{\partial \ln f(D_j, X_j; \beta_0)}{\partial \beta} \cdot \Lambda_l(x)1(x \in S^r), \qquad (38)$$

where $Ep^r(x)$ is parametrically specified as $p^r(x, \beta_0)$ and EH is the expected Hessian at β_0. Further

$$\Psi^r_{l,m}(Y_j^r, D_j, X_j; x) = \frac{\Lambda_l(X_j)1(D_j = r)}{h_{n_{l,r}}}\left(Y_j^r - m_l^r(p^r(X_j))\right)K\left(\frac{p^r(X_j) - p^r(x)}{h_{n_{l,r}}}\right) \cdot \frac{\Lambda_l(x)1(x \in S^r)}{E\hat{f}_{p^r|D=r,\Lambda_l=1}(p^r(x))}, \qquad (39)$$

where $\hat{f}_{p^r|D=r,\Lambda_l=1}(p^r(x))$ is the kernel density estimate in the common subpopulation of the participants in treatment r and the subpopulation defined by $\Lambda_l(x)$ obtained with the same bandwidth $h_{n_{l,r}}$. It holds that $E[\Psi^r_{l,p}(Y_j^r, D_j, X_j; X)|X = x] = 0$, $E[\Psi^r_{l,m}(Y_j^r, D_j, X_j; X)|X = x] = 0$ and $\underset{n_{l,r}\to\infty}{plim}\ n_{l,r}^{-\frac{1}{2}}\sum_j \hat{b}_m^r(p^r(X_j))\Lambda_l(X_j)1(D_j = r) = b < \infty$ and $n^{-\frac{1}{2}}\sum_{j=1}^{n}\hat{R}^r(X_j) = o_p(1)$. To ease notation, multiple outcome variables, i.e. Y^r being a vector, are treated sequentially one after the other and are stacked thereafter to the vector $(\hat{\mathbf{m}}_{VL}^r - \mathbf{m}_{VL}^r)1(x \in \hat{S}^r)$.

B Appendix: Swedish Rehabilitation Programmes

Table B.1: Estimation of a multinomial probit model, No Rehabilitation as reference group

Variable		Vocational	Non-vocational
Constant		**-2.53**	**-3.19**
Age:	18-35 years	*0.17*	0
	46-55 years	**-0.20**	*-0.14*
Citizenship:	Swedish born	*0.15*	0
Marital status:	widowed	-0.31	0
Occupation in:	Manufacturing	-0.10	0
Employment status:	employed	**0.34**	0
Qualifying income:	(in SEK/1000)	**0.14**	0
Previous sickness days	31-60 days	0	*0.29*
(in last 6 months):	> 60 days	*0.18*	(0.07)
Previous participation	in vocational rehabilitation	**0.31**	*0.28*
Unemployment rate	(in %)	0	**0.10**
County:	Hallandslän	0	**2.36**
	Bohuslän	-0.11	**0.84**
	Älvsborgslän	0	**-0.38**
	Göteborgskommun	**-0.41**	**1.18**
Community type:	urban / suburban region	**-0.39**	**-1.08**
	major / middle large city	**-0.24**	**-0.49**
	industrial city	0	**0.68**
Sickness registrated	in 1991/92	0	**0.27**
Sickness registration	by psych./social med. centre	0	**0.37**
Sickness degree:	100% sick leave	**0.48**	0
Indications of	alcohol or drug abuse	-0.18	*0.28*
Diagnosis:	musculoskeletal	0	**0.38**
	injuries	0.14	*0.21*
Case assessed by:	employer	**0.59**	**0.31**
	insurance office	**0.40**	0
	IO on behalf of the employer	**0.32**	*0.20*
	not needed	**-0.55**	**-0.41**
Medical VR	wait and see	(-0.10)	**0.59**
recommendation:	VR needed and defined	**1.38**	**1.12**
	eligible to disability pension	0	0.38
Non-medical VR	VR needed and defined	1**.68**	**0.55**
recommendation:	eligible to disability pension	0	-0.34
VR prevented by:	medical reasons	0	**0.52**
	other factors	0	*0.32*
Medical **and** non-medical recommendation:	wait and see	**0.54**	0
	VR needed and defined	**-1.38**	*-0.43*

Note: VR stands for vocational rehabilitation. **Bold** coefficients are significant at the 1% level (2 sided-test), numbers in *italics* are significant at the 5% level, coefficients in brackets () are insignificant at the 10% level. Weighted simulated maximum likelihood estimates based on the GHK simulator with 400 replications (Börsch-Supan and Hajivassiliou 1993). 2 Cholesky factors have been estimated (i.e. the maximum number of identified elements. Value of log-likelihood: -4610.9. Coefficients of the group no rehabilitation and the coefficients marked with 0 in the table are fixed to zero. Inference is based on the QML covariance matrix (Manski and Lerman 1977).

In this appendix additional estimation results are provided. In Table B.1 the estimated coefficients of the multinomial programme-choice probit model are given, where no rehabilitation is the reference group. Different sets of explanatory variables have been tried and insignificant variables have been deleted subsequently. From the estimation results of the multinomial probit model the estimated participation probabilities $\hat{p}^{None}(X_i)$, $\hat{p}^{VR}(X_i)$, $\hat{p}^{NVR}(X_i)$ are computed for all observations. Table B.2 shows the correlation coefficients between these estimated probabilities. Non-participation is strongly negatively correlated with vocational and non-vocational rehabilitation, while vocational and non-vocational rehabilitation are nearly uncorrelated. This indicates that selection to the programmes is clearly influenced by observed characteristics, separating the individuals in need for rehabilitation from those unlikely to receive rehabilitation.

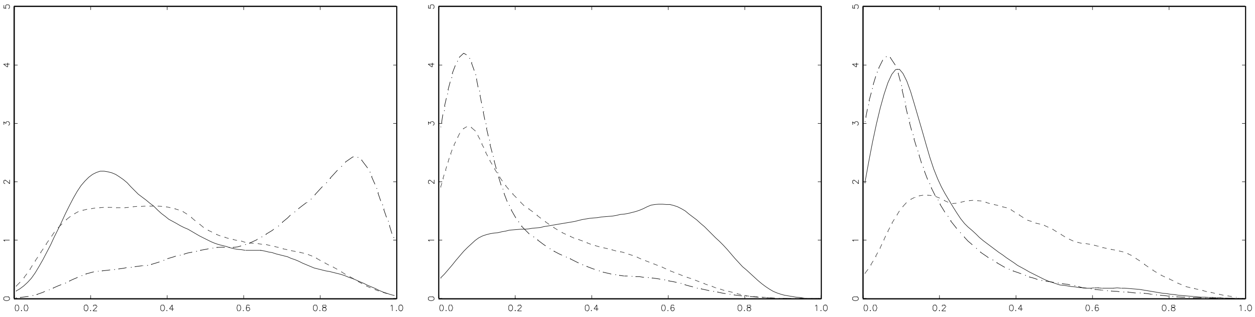
Table B.2: Implied correlation matrix of the participation probabilities

	None	Vocational	Non-vocational
None	1	-0.72	-0.61
Vocational		1	-0.11

Note: Sample correlation coefficients between the estimated participation probabilities.

Nonparametric estimation of the supports of $f_{X|D=r}(x)$ in all subpopulations appears difficult with X containing this many variables. As in many evaluation studies the supports of $f_{X|D=r}$ are approximated by the supports of the participation probabilities $\hat{p}^r$. Figure B.1 shows kernel density estimates of the estimated participation probabilities $\hat{p}^r$ for all three treatment groups, i.e. the long-dashed line in the left picture displays the distribution of $\hat{p}^{None}$ in the group of participants in no rehabilitation, the solid line represents $\hat{p}^{None}$ for the participant group in vocational rehabilitation, and the short-dashed line stands for the participants in non-vocational rehabilitation. Expectedly, the density masses for the respective participants lie most to the right, but cover most of the region where the participation probabilities of the respective non-participants are located. The supports $S^r(p^r)$ are approximated by the interval delimited by the smallest and largest estimated participation probability $\hat{p}^r$ among the participants in treatment r. These cut-off points $p^r_{\text{min}}, \hat{p}^r_{\text{max}}$ are given in Table B.3 and it is seen that for all three categories more than 99% of all observations lie within the estimated support $\hat{S}^r$. The support intervals are estimated analogously for all considered subpopulations and apart from few exceptions the estimated supports cover always about 99% of all observations. The estimated supports for each subpopulation are discernible from Figures B.2, where the estimated regression curves are plotted only within the respective support region.

Figure B.1: Distribution of the estimated participation probabilities



Density of $\hat{p}^{None}$ *Density of $\hat{p}^{Vocational}$* *Density of $\hat{p}^{Non-vocational}$*

Note: Kernel density estimates of the estimated participation probabilities in all treatment groups: Non-participants (long-dashed), Vocational group (solid), Non-vocational group (short-dashed). Bandwidth=0.10.

Table B.3: Estimated supports of the participation probabilities

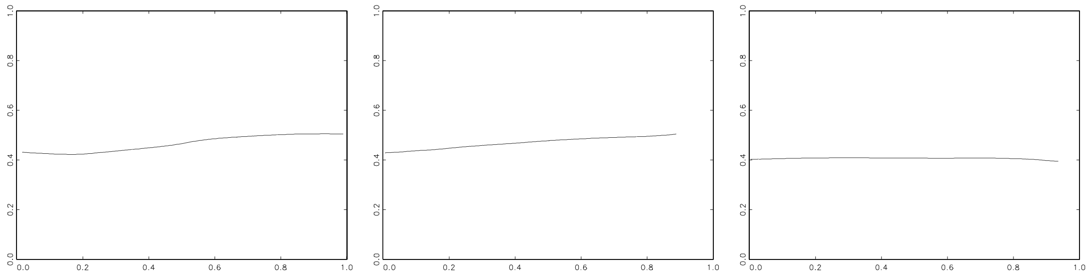
	$\hat{S}^{None}$	$\hat{S}^{Vocational}$	$\hat{S}^{Non-vocational}$
Minimum $\hat{p}^r$ (in %)	1.7	0.6	0.5
Maximum $\hat{p}^r$ (in %)	99.4	89.3	94.0
Observations in support	6283	6231	6227

Note: Minimum corresponds to the smallest estimated participation probability $\hat{p}^r$ in the subsample of participants in programme r. Maximum is defined analogously. Number in support gives the number of observations of the full sample whose estimated participation probability p^r lies within the estimated support S^r.

After the participation probabilities have been computed the conditional expectations $E[Y^r|p^r]$ are estimated by Nadaraya-Watson kernel regression for all considered subpopulations. The bandwidth value is selected by penalised cross-validation according to Frölich (2000). The estimated regression curves $\hat{m}_l^r$ are graphed within the respective support regions in Figure B.2. In most cases the cross-validation selector has chosen quite large bandwidth values, leading to smooth regression curves. For non-vocational rehabilitation the shapes suggest (at least for some subpopulations) that re-employment chances decrease with higher propensity to non-vocational rehabilitation, which could indicate more serious health problems. On the other hand the expected outcomes for non-participation and participation in vocational rehabilitation slope vaguely upwards with higher propensity to participate in these programmes, indicating better labour market prospects of these individuals. Table 5.2 in Section 5 gives the average potential outcomes which are obtained by integrating the regression curves in Figure B.2

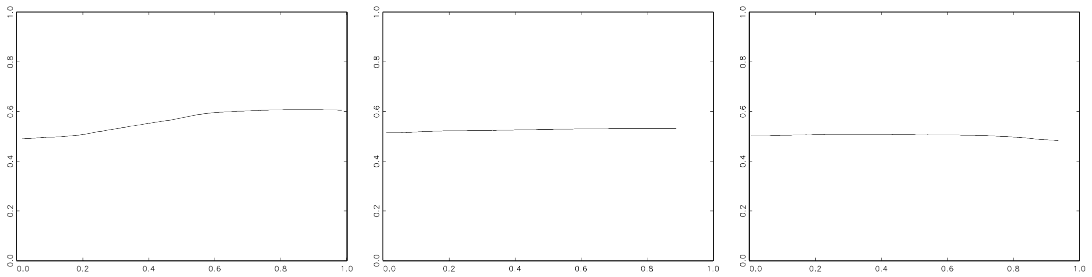
Figure B.2: Estimated regression curves with $\hat{m}_l^r(p^r)$ on the abscissa and p^r on the ordinate

$\hat{m}^r$ *for the whole population*



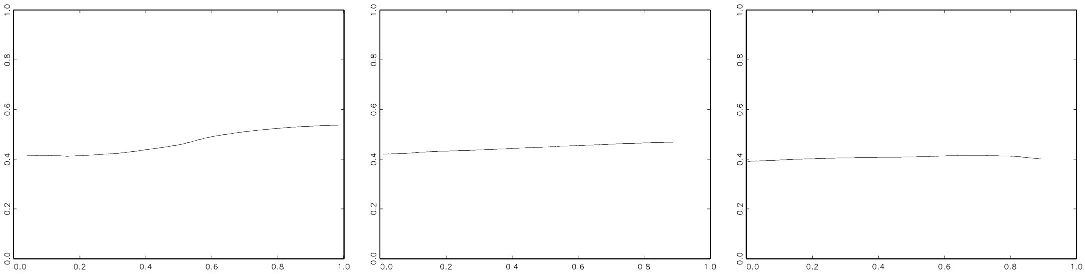
$\hat{m}^{None}$; h=0.39 $\hat{m}^{Vocational}$; h=0.45 $\hat{m}^{Non-vocational}$; h=0.54

$\hat{m}^r$ *for the subpopulation: Employed*



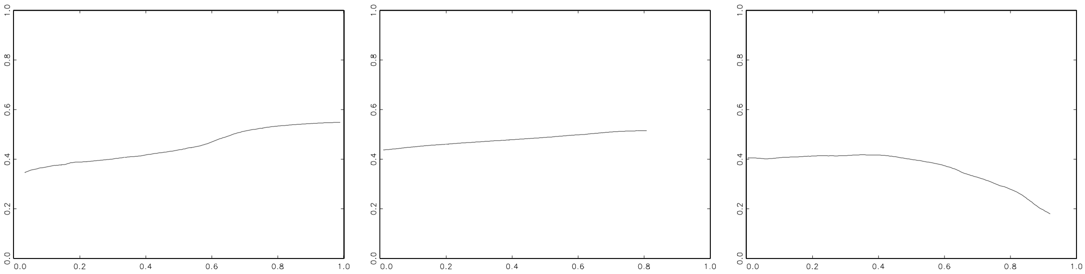
$\hat{m}^{None}$; h=0.39 $\hat{m}^{Vocational}$; h=0.54 $\hat{m}^{Non-vocational}$; h=0.51

$\hat{m}^r$ for the subpopulation: Men



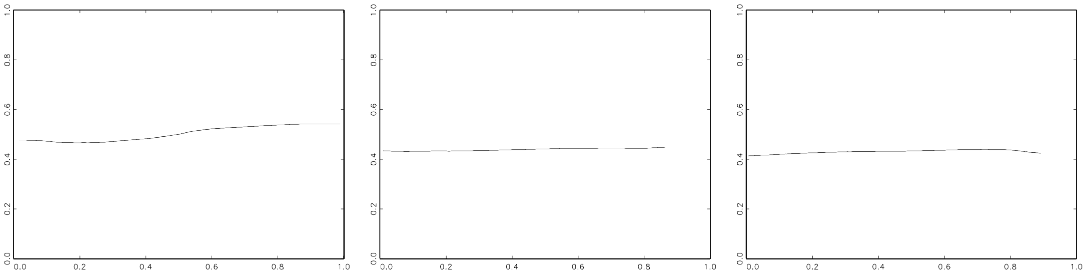
$\hat{m}^{None}$; h=0.36 $\hat{m}^{Vocational}$; h=0.54 $\hat{m}^{Non-vocational}$; h=0.54

$\hat{m}^r$ for the subpopulation: Age 46-55 years



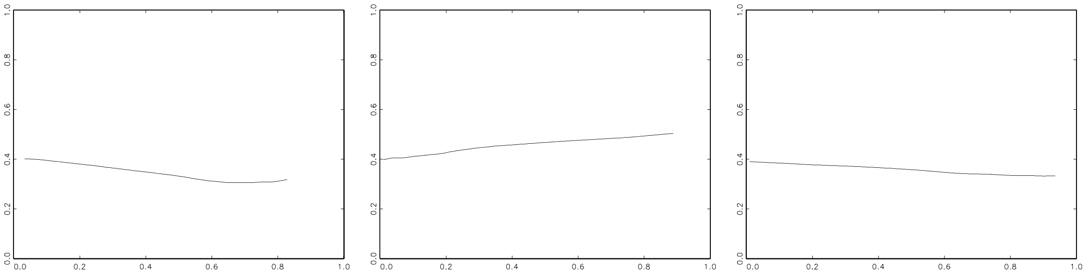
$\hat{m}^{None}$; h=0.30 $\hat{m}^{Vocational}$; h=0.54 $\hat{m}^{Non-vocational}$; h=0.30

$\hat{m}^r$ for the subpopulation: Occupation in manufacturing sector



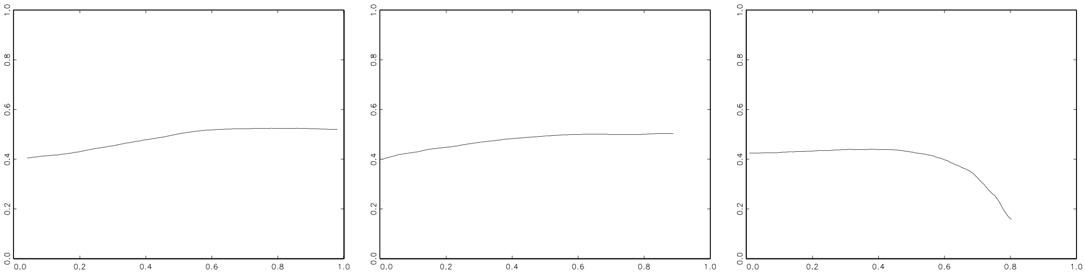
$\hat{m}^{None}$; h=0.39 $\hat{m}^{Vocational}$; h=0.54 $\hat{m}^{Non-vocational}$; h=0.54

$\hat{m}^r$ for the subpopulation: Vocational rehabilitation recommended



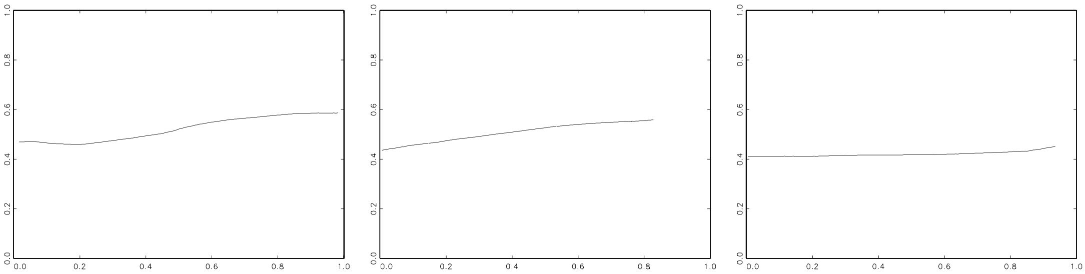
$\hat{m}^{None}$; h=0.33 $\hat{m}^{Vocational}$; h=0.42 $\hat{m}^{Non-vocational}$; h=0.54

$\hat{m}^r$ for the subpopulation: Living in rural community



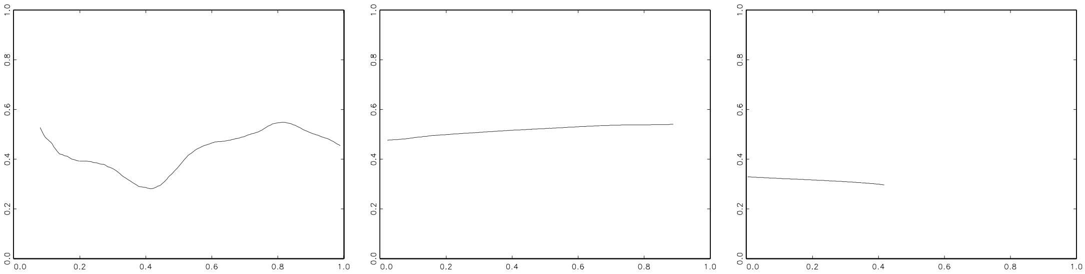
$\hat{m}^{None}$; h=0.39 $\hat{m}^{Vocational}$; h=0.39 $\hat{m}^{Non-vocational}$; h=0.27

$\hat{m}^r$ for the subpopulation: Age 36-45 years



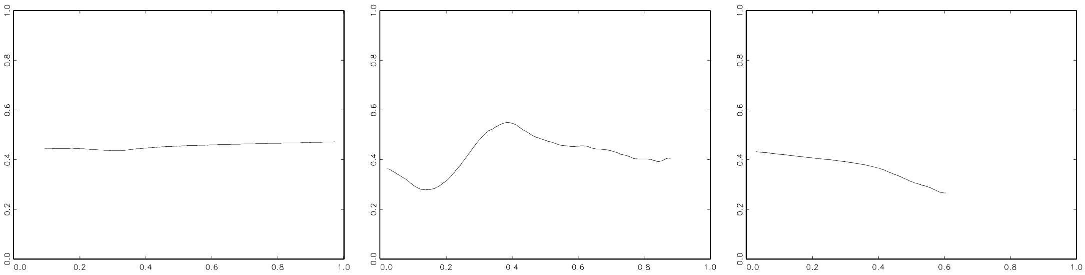
$\hat{m}^{None}$; h=0.39 $\hat{m}^{Vocational}$; h=0.45 $\hat{m}^{Non-vocational}$; h=0.54

$\hat{m}^r$ for the subpopulation: Living in county Älvsborgslän



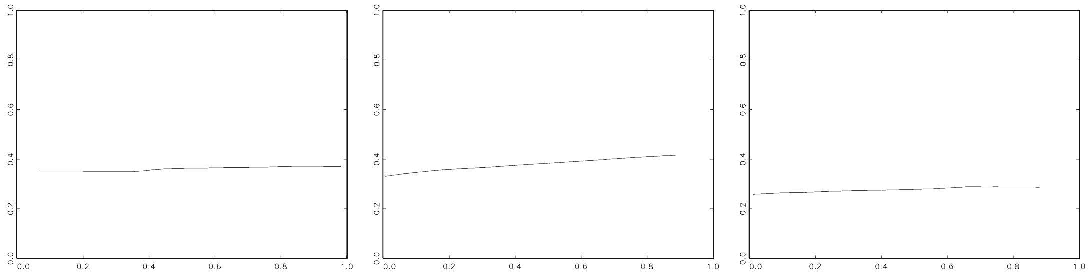
$\hat{m}^{None}$; h=0.12 $\hat{m}^{Vocational}$; h=0.54 $\hat{m}^{Non-vocational}$; h=0.54

$\hat{m}^r$ for the subpopulation: Living in county Värmlandslän



$\hat{m}^{None}$; h=0.54 $\hat{m}^{Vocational}$; h=0.15 $\hat{m}^{Non-vocational}$; h=0.36

$\hat{m}^r$ for the subpopulation: More than 60 sickness days in previous six months



$\hat{m}^{None}$; h=0.48 $\hat{m}^{Vocational}$; h=0.54 $\hat{m}^{Non-vocational}$; h=0.54

References

Angrist, J. (1998): "Estimating Labour Market Impact of Voluntary Military Service using Social Security Data," *Econometrica*, 66, 249–288.

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

Berger, M., D. Black, and J. Smith (2000): "Evaluating Profiling as a Means of Allocating Government Services," mimeo, University of Western Ontario.

Black, D., J. Smith, M. Berger, and B. Noel (1999): "Is the Threat of Training more Effective than Training itself?," mimeo, University of Western Ontario.

Börsch-Supan, A., and V. Hajivassiliou (1993): "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models," *Journal of Econometrics*, 58, 347–368.

Colpitts, T. (1999): "Targeting Reemployment Services in Canada: The Service and Outcome Measurement System (SOMS) Experience," mimeo, Department of Human Resources Development, Ottawa, Canada.

de Koning, J. (1999): "The chance-meter: Measuring the Individual Chance of Long-term Unemployment," Netherlands Economic Institute, Rotterdam.

Dehejia, R. (1999): "Program Evaluation as a Decision Problem," *NBER working paper*, 6954.

Dehejia, R., and S. Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes," *Journal of American Statistical Association*, 94, 1053–1062.

DOL (1999): *Evaluation of Worker Profiling and Reemployment Services Policy Workgroup: Final Report and Recommendations.* U.S. Department of Labor, Employment and Training Administration, Washington D.C.

Eberts, R. (1998): "The Use of Profiling to Target Services in State Welfare-to-Work Programs: An Example of Process and Implementation," *W.E. Upjohn Institute for Employment Reserach Working Paper*, 98-52.

Eberts, R., and C. O'Leary (1999): "A Frontline Decision Support System for One-Stop Career Centers," mimeo, W.E. Upjohn Institute for Employment Reserach.

FAN, J., T. GASSER, I. GIJBELS, M. BROCKMANN, AND J. ENGEL (1997): "Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency," *Annals of the Institute of Mathematical Statistics*, 49, 79–99.

FAY, R. (1996): "Enhancing the Effectiveness of Active Labour Market Policies: Evidence from Programme Evaluations in OECD Countries," *Labour Market and Social Policy Occasional Papers, OECD*, 18.

FRÖLICH, M. (2000): "Nonparametric Treatment Evaluation: Matching versus Local Polynomial Regression," mimeo, Universität St. Gallen.

FRÖLICH, M., A. HESHMATI, AND M. LECHNER (2000a): "A Microeconometric Evaluation of Rehabilitation of Long-term Sickness in Sweden," *Discussion Paper, Volkswirtschaftliche Abteilung, Universität St. Gallen*, 2000-04.

——— (2000b): "Mikroökonometrische Evaluierung berufsbezogener Rehabilitation in Schweden," forthcoming in Schweizerische Zeitschrift für Volkswirtschaft und Statistik 3/2000.

GERFIN, M., AND M. LECHNER (2000): "Microeconometric Evaluation of the Active Labour Market Policy in Switzerland," *Discussion Paper, Volkswirtschaftliche Abteilung, Universität St. Gallen*, 2000-10.

GREENE, W. (1997): *Econometric Analysis*. Prentice Hall, 3 edn.

HANSEN, L. (1982): "Large Sample Properties of Generalized Method of Moment Estimators," *Econometrica*, 50, 1029–1054.

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

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

HECKMAN, J., J. SMITH, AND N. CLEMENTS (1997): "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535.

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

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

HORRACE, W., AND P. SCHMIDT (1996): "Multiple Comparisons with the Best, with Economic Applications," *Journal of Applied Econometrics*, 15, 1–26.

HSU, J. (1996): *Multiple Comparisons: Theory and Methods*, vol. 1. Chapman and Hall, London.

IMBENS, G. (1999): "The Role of the Propensity Score in Estimating Dose-Response Functions," *NBER, Technical Working Paper*, 237, forthcoming in Biometrika.

IMBENS, G., AND T. LANCASTER (1994): “Combining Micro and Macro Data in Microeconometric Models,” *Review of Economic Studies*, 61, 655–680.

LECHNER, M. (1999a): “Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification,” *Journal of Business and Economic Statistics*, 17, 74–90.

——— (1999b): “Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption,” *Discussion Paper, Volkswirtschaftliche Abteilung, Universität St. Gallen*, 9908, forthcoming in M. Lechner and F. Pfeiffer (eds.), Econometric Evaluations of Active Labour Market Policies in Europe, Physica, Heidelberg.

——— (2000): “An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany,” *forthcoming in Journal of Human Resources.*

MANSKI, C. (1993): “The Selection Problem in Econometrics and Statistics,” in *Handbook of Statistics*, ed. by G. Maddala, C. Rao, and H. Vinod. Elsevier Science Publishers.

——— (1995): *Identification Problems in the Social Sciences.* Harvard University Press, Cambridge, Mass.

——— (1997): “Monotone Treatment Response,” *Econometrica*, 65, 1311–1334.

——— (1999): “Statistical Treatment Rules for Heterogenous Populations: With Application to Randomized Experiments,” mimeo, Department of Economics, Northwestern University.

——— (2000): “Identification Problems and Decisions under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice,” *Journal of Econometrics*, 95, 415–442.

MANSKI, C., AND S. LERMAN (1977): “The Estimation of Choice Probabilities from Choice-Based Samples,” *Econometrica*, 45, 1977–1988.

MOHR, L. (1999): “The Impact Profile Approach to Policy Merit,” *Evaluation Review*, 23, 212–249.

NEWEY, W., AND D. MCFADDEN (1994): “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, ed. by R. Engle, and D. McFadden. Elsevier, Amsterdam.

OECD (1998): “The Early Identification of Jobseekers who are at Greatest Risk of Long-term Unemployment in Australia,” in *Early Identification of Jobseekers at Risk of Long-term Unemployment: The Role of Profiling*, pp. 31–61. OECD Proceedings, Paris.

O’LEARY, C., P. DECKER, AND S. WANDNER (1998): “Reemployment Bonuses and Profiling,” *W.E. Upjohn Institute for Employment Reserach Working Paper*, 98-51.

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

PARZEN, E. (1962): “On Estimation of a Probability Density and Mode,” *Annals of Mathematical Statistics*, 33, 1065–1076.

POWELL, J., J. STOCK, AND T. STOKER (1989): “Semiparametric Estimation of Index Coefficients,” *Econometrica*, 57, 1403–1430.

PUHANI, P. (1999): *Evaluating Active Labour Market Policies: Empirical Evidence for Poland during Transition.* Physica, Heidelberg.

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

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

SEIFERT, B., AND T. GASSER (1996): "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of American Statistical Association*, 91, 267–275.

——— (2000): "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, forthcoming.

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

TODD, P. (1999): "...semiparametric propensity score," mimeo, University of Chicago.