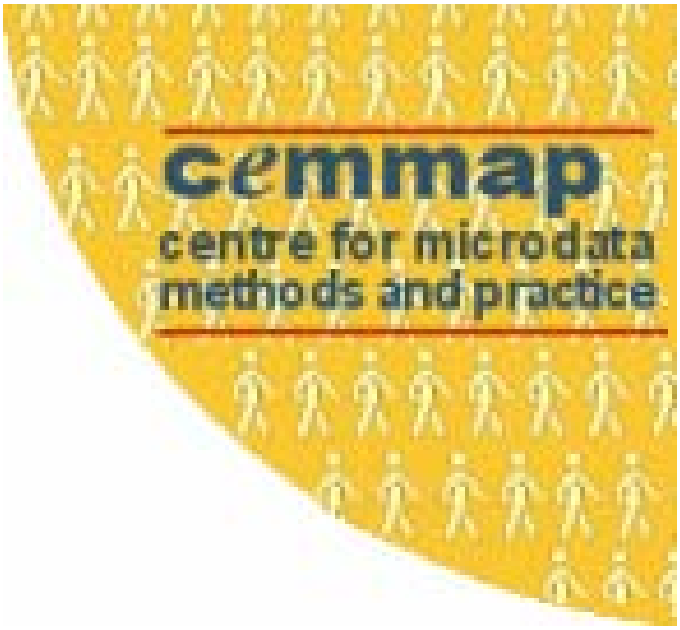

STATISTICAL TREATMENT CHOICE: AN APPLICATION TO ACTIVE LABOUR MARKET PROGRAMMES

Markus Frölich

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP24/06

Statistical Treatment Choice: An Application to Active Labour Market Programmes

Markus Frölich

Department of Economics, University of St.Gallen

June 22, 2006

Abstract: Choosing among a number of available treatments the most suitable for a given subject is an issue of everyday concern. A physician has to choose an appropriate drug treatment or medical treatment for a given patient, based on a number of observed covariates X and prior experience. A case worker in an unemployment office has to choose among a variety of available active labour market programmes for unemployed job seekers. In this paper, two methodological advancements are developed: First, this methodology permits to combine a data set on previously treated individuals with a data set on new clients when the regressors available in these two data sets do not coincide. It thereby incorporates additional regressors on previously treated that are not available for the current clients. Such a situation often arises due to cost considerations, data confidentiality reasons or time delays in data availability. Second, statistical inference on the recommended treatment choice is analyzed and conveyed to the agent, physician or case worker in a comprehensible and transparent way. The implementation of this methodology in a pilot study in Switzerland for choosing among active labour market programmes (ALMP) for unemployed job seekers is described.

Keywords: Statistical treatment rules, active labour market policies
JEL classification: C13, C14

The author is also affiliated with the Swiss Institute for International Economics and Applied Economic Research (SIAW), the Institute for the Study of Labor (IZA), Bonn and the Institute for Labour Market Policy Evaluation (IFAU), Uppsala. I am grateful for discussions and comments to Stefanie Behncke, Michael Lechner and Heidi Steiger. This research is supported by the Swiss State Secretariat for Economic Affairs (seco) and the Marie Curie Individual Fellowship MEIF-CT-2004-006873. Address for correspondence: Markus Frölich, University of St. Gallen, Bodanstrasse 8, SIAW, 9000 St. Gallen, Switzerland; markus.froelich@unisg.ch, www.siaw.unisg.ch/froelich

1 Introduction

Choosing among a number of available treatments the most suitable for a given subject is an issue of everyday concern. A physician has to choose an appropriate medical drug treatment for a given patient, based on a number of observed covariates X and prior experience. As a second example, consider choosing among the different types of rehabilitation therapies available for persons with alcohol related problems. As a third example, which will guide the application in this paper, we examine the choice of an active labour market programme for an unemployed job seeker. In many countries the case workers in charge have a number of different training programmes at their disposal to which they can assign an unemployed person to increase her chances to find a job soon. These treatment options often include job search training, language training, computer training, vocational skills training, further training, re-training as well as employment programmes, interim jobs, etc. In addition, there is the option of not assigning any programme. Participation in such programmes is often mandatory if assigned by the case worker.[1] In all these situations the best treatment choice may depend on the characteristics of the individual and may thus differ from individual to individual. Statistics may help in attaining better choices. Statistical predictions of treatment outcomes on an individual basis may be communicated to the physician, case worker or the jobseeker to produce more informed treatment choices.

Providing such estimates of treatment effects for various demographic groups to physicians has a long history in the medical literature, but these are often based on randomized trials and reported only for very broadly defined demographic groups. In recent years there has been a strong interest in using statistical tools in other fields and in particular for assigning active labour market programmes, where usually no experimental data is available and where covariate information should be accounted for in much more detail than considering only broad demographic groups. This interest in profiling and targeting of active labour market programmes is demonstrated by several recent publications, e.g. the book "Targeting Employment Services" (Eberts, O'Leary, and Wandner 2002) or OECD (1998), DOL (1999), Berger, Black, and Smith (2001), Rudolph and Müntnich (2001), Colpitts (2002), Eberts (2002), Eberts, O'Leary, and DeRango (2002), Wandner (2002), Black, Smith, Berger, and Noel (2003), Manski (2000, 2004), Frölich, Lechner, and Steiger (2003), Plesca and Smith

[1] Noncompliance may result in suspension of unemployment benefits.

(2005) and Lechner and Smith (2006).[2]

In this paper, two methodological advancements are developed and then applied to the choice among Swiss active labour market programmes. First, a methodology is developed to assist treatment choice in a situation where a large and informative data base with information on many characteristics W is available for deriving statistical predictions but only a limited number of covariates X is observed for the individual for whom a choice has to be made. Second, statistical inference on the recommended treatment choice is analyzed.

The situation where a large set of characteristics W for previously treated is available but recommendations are to be based on a smaller set of covariates X occurs in many settings where an expert has to make a choice but has only limited access to the entire knowledge data base e.g. due to limitations in reporting, confidentiality or data privacy reasons, costs of measuring covariates, time delays in data availability or different measurement scales.

For example, consider that recommendations about the best treatment choice for women and for men are to be derived from a large drug trial. Here, X is gender and W refers to additional covariate information collected during the trial. If the drug trial had not been randomized, the additional covariate information will often be very important to obtain unbiased gender-specific estimates of the treatment effects in that they control for *selection bias*. In the other example, X may be a set of information the case worker has about an unemployed person, whereas W may contain additional information on earnings and employment histories obtained from data bases not accessible to the case worker. Again, since allocation to labour market programmes has usually been non-random in the past, incorporating this additional W information is important to account for non-random selection.

In the following, a methodology is developed to include this additional covariate information, which is applicable in linear and in non-linear models. This is important because linear models are often not appropriate if the outcome variable is binary or bounded, e.g. patient's survival status or the employment status of the unemployed person.

[2] For references on targeting of treatments in biometrics and statistics and in other fields see e.g. Wald (1950), Brownell and Wadden (1991), Velicer, Prochaska, Bellis, DiClemente, Rossi, Fava, and Steiger (1993), Kreuter and Strecher (1996), ProjectMatchResearchGroup (1997), Breslin, Sobell, Sobell, Cunningham, Sdao-Jarvie, and Borsoi (1999), Kreuter, Strecher, and Glassman (1999), Velicer and Prochaska (1999), Thall, Sung, and Estey (2002), Murphy (2003) and Rush (2005). In most of this literature it is assumed that *all* confounding variables are observed to identify treatment effects. In this paper we allow for unobserved confounders.

This methodology is then applied to assisting case workers in choosing among active labour market programmes (ALMP) for unemployed jobseekers. ALMP have been introduced in many countries during the early 1990s to combat problems of high and persistent unemployment or low earnings of disadvantaged groups through the public provision of training, job creation schemes, subsidized jobs and wage subsidies. Such programmes exist in the USA on a relatively small scale (e.g. the Job Training Partnership Act, JTPA, Heckman, Ichimura, Smith, and Todd (1998)) and in many European countries on a much larger scale. Recent evaluation studies often found these programmes to be relatively unsuccessful on average, but also concluded that some individuals may benefit more from training than others, see e.g. Heckman, Smith, and Clements (1997), Gerfin and Lechner (2002) and Gerfin, Lechner, and Steiger (2005) for evidence on treatment effect heterogeneity. There has been a recent trend emphasizing the need for a better targeting of these programmes, in other words for choosing more carefully the most adequate programme for each unemployed person on an individual basis. This trend is also supported by studies which found the current allocations made by case workers to be suboptimal, see e.g. Frölich, Lechner, and Steiger (2003) and Lechner and Smith (2006). Several countries have expressed their interest in using statistical systems in supporting the choice of adequate programmes, and various approaches often based on a simple profiling strategy are and have been implemented. Switzerland decided to pilot a statistical targeting system and conducted a randomized field study in 2005 based on the methodology developed in this paper. First evaluation results will be available in 2007.[3]

This paper develops the methodology for statistical treatment choice. Section 2 analyzes the treatment choice setting and the selection problem and develops the econometric methodology for identification and estimation. Section 3 gives more information on Swiss labour market policies, and Section 4 describes the implementation and application in more detail. Section 5 concludes.

2 Optimal treatment choice

Suppose there are R different and mutually exclusive treatments. An individual i at time t needs to receive one of these treatments and requests advice in choosing the best treatment. The

[3]The pilot study covers only a randomly selected subset of jobseekers, thereby permitting an experimental evaluation of the impact of the targeting system itself.

treatments may be different drugs or medical therapies for an individual with a heart disease. They may be different types of training or employment programmes for an unemployed person. Or they may represent different educational tracks to choose from for a young person leaving school. It may be that the individual i chooses the treatment for herself or that an agent, e.g. a physician or a case worker in the employment office, makes the choice. One of the available treatment options will often be *not* to take any drug or training now, but leaving the option for later. Hence, this option of "no treatment" at time t, i.e. of deferring the choice for later, is considered as being one of the R treatment options. Let

$$Y^1_{i,t+\tau}, ..., Y^R_{i,t+\tau}$$

be the *potential* outcomes (Rubin 1974) for individual i at some time $t+\tau$, e.g. the survival status or the employment status. $Y^1_{i,t+\tau}$ is the outcome that individual i would *realize* if taking treatment one. Similarly, $Y^2_{i,t+\tau}$ is the realized outcome if taking treatment two, and so on. These potential outcomes are unknown ex-ante, but even ex-post only one of them can be observed: the potential outcome corresponding to the treatment that has actually been taken. These outcomes are assumed to be scalars, but they can be indices combining several different outcome variables, e.g. a weighted average of survival status at different points in time, perhaps combined with a measure of the *costs* of treatment.

The optimal treatment for individual i would be

$$r^*_i = \arg\max_r \; Y^r_{i,t+\tau},$$

which is unknown since the potential outcomes $Y^r_{i,t+\tau}$ are not known ex ante. Nevertheless, if we observe some covariates $X_{i,t}$ e.g. age and gender,[4] we may be able to predict the expected potential outcomes

$$E\left[Y^r_{i,t+\tau}|X = X_{i,t}\right] \qquad \text{for} \qquad r = 1, ..., R$$

and estimate the expected optimal treatment as[5]

$$r^*(X_{i,t}) = \arg\max_{r\in\{1,..,R\}} \; E\left[Y^r_{t+\tau}|X = X_{i,t}\right].$$

[4] These may also contain information on past values of covariates, e.g. previous receipt of treatment, health and employment history etc.

[5] Manski (2000, 2004) examined optimal treatment choice from a normative perspective by analyzing how a benevolent central planner would allocate individuals to treatments such that social welfare would be maximized. Since the planner can discriminate between individuals only on the basis of observable characteristics X, the treatment allocation will be a mapping from X to the available treatments $\{1, .., R\}$. Manski shows that if

These estimates can then be made available to the agents or the individuals themselves, e.g. through the internet, to assist them in their choices.

For estimating $r^*(X_{i,t})$ two issues are of interest: First, consistent estimation of the expected *potential* outcomes $E\left[Y^r_{t+\tau}|X = X_{i,t}\right]$ and second, information on the statistical precision of the estimated $r^*(X_{i,t})$. In the following, the subscripts $t+\tau$ and t are suppressed and assumed to be implicitly included in Y and X. Hence, X may contain a time trend and a seasonality component etc.

For estimating the conditional expectation functions $E\left[Y^r|X\right]$ we may resort to data on previous treatment recipients, i.e. to the observed outcomes of individuals who received some treatment in the past. For these previously treated we often have more detailed information available than for the current clients, which may help to obtain more reliable estimates. Apart from observing their realized outcomes Y we may not only know their covariates X, e.g. age and gender, but additionally a vector of further characteristics W, e.g. entire health histories, treatment histories, employment and earnings histories, subjective assessments, information on family background etc. These additional covariates W are *not* available for individual i at the time when predictions are to be made, hence the predictions of expected potential outcomes can only be based on the covariates X. Nevertheless, the observed W of the past treatment recipients may be very helpful for identification and/or precision of the estimates, and a methodology to including such W variables in the estimation of *nonlinear* models is developed in the following. Before that, a few examples are discussed why some covariate information W may be available for past treatment recipients but not for current clients i at the time t when having to choose among the treatments; or why it may not be useful to include them in X.

First, data may not be available for data security reasons. It may be that additional administrative data on past participants can be accessed in an anonymized form for estimating the statistical system but that this data base shall not directly be linked to the software producing the predictions for individual i. In the application to ALMP, social security data with information on entire employment and earnings histories of past participants has been

the planner aims to maximize utilitarian welfare, the optimal treatment choice is assigning each individual to that programme that promises the largest expected potential outcome *conditional* on the individual's observed characteristics, i.e. to maximize $E[Y^r|X]$.

made available for estimating parameters but is not available for the day-to-day operations in the employment offices. Similarly, in a large clinical trial detailed data may have been collected, e.g. by a private research company, that shall not be made publicly accessible. Instead, the company may be requested to publish estimates of $E[Y^r|X]$ for X defining different age and gender groups. Think of a (non-randomized) drug trial where the company publishes results for women, men and children with different degrees of sickness.

Second, data on some variables may be available only with a delay. E.g. the social security data in our application is usually reported and compiled only about one or two years later. In addition, there might be amendments or corrections to the reported data some time later, and administrative data, collected for different purposes, often needs to be cleaned and made consistent before use. These regressors are not yet available when the decision needs to be taken.

Third, it may be expensive to collect the additional information. Think again of a large clinical multi-purpose trial where data have been collected on many parameters and many different tests have been conducted, e.g. sponsored by a public research agency with the aim to answer several research questions with the same data. For future day-to-day treatment choices for new clients it would be impractical and unreasonably expensive to collect all these information.

Fourth, even when the data is available it may not be appropriate to include all variables in X because of structural changes in the relationship $E[Y^r|X]$ between the time when the estimation data was collected and the time when the predictions need to be made. If there is a substantial time gap,[6] the coding or measurement of some of the variables might have changed. Unless this variable can be re-coded, it might be doubtful whether $E[Y^r|X]$ with the old coding of X is useful for predicting $E[Y^r|X]$ with the new coding of X. Nevertheless, this variable can still be used as a W variable since the W variables are only used within the population of past treated and not for projections into the future. Another reason might be that the

[6] In our application, predictions on employment chances are made for 12 months ahead. Hence, the time gap has to be at least one year, i.e. the length of the observation window. In addition, for obtaining a reasonable sample size, it would be useful to include all inflows over a longer time period, e.g. at least half a year or one year. The data need to be cleaned and the statistical system estimated, adding another few months to the time gap. In addition, once the system is estimated it may be used for a while (e.g a year or two) before it is updated. Hence, the minimum time gap is at the very least two years.

conditional expectation function $E[Y^r|X]$ itself may have changed over time. It may still be that expected outcomes conditional on e.g. only age and gender remained relatively stable through time,[7] but that this is not the case if we were to condition additionally on several other characteristics. To give an example, consider the binary treatment choice between a training programme versus no programme at all, i.e. $R = 2$. In the past, training was provided in two variants: Those individuals with a contribution to the unemployment funds of less than two years received a very ineffective and cheap job search training whereas those who contributed for more than two years participated in a very expensive and effective coaching and placement programme. In the meantime this differentiation by contribution time was abolished and everyone is assigned randomly to any of these two training programmes, with equal probability in a way to ensure that the total budget did not change. Hence, *conditional* on contribution time the optimal treatment decision might have been very different in the past than it is now, whereas unconditionally there was no change over time. In this situation, including contribution time in X would lead to biased estimates, whereas it should still be included in W.

For any of these reasons, we may thus not be able to base our treatment choice recommendations on estimates of $E[Y^r|W, X]$ but only on $E[Y^r|X]$ since information on W is not available for individual i at the time t when a choice has to be made. Nevertheless, the additional data on W for the previous participants will often be indispensable for consistent estimation of $E[Y^r|X]$ for reasons of remaining selection bias, as explained in the following. Let $\{(Y_j, X_j, W_j, D_j)\}_{j=1}^N$ be the available data on previous treatment recipients, where $D_j \in \{1, ..., R\}$ indicates the received treatment.[8] $Y_j \equiv Y_j^{D_j}$ is the realized outcome, i.e. the potential outcome corresponding to the treatment D_j that was actually received. X_j and W_j are the observed covariates.

If the treatment had *not* been randomly administered in the past,[9] the potential outcome Y^d among those who decided to take treatment d usually is different from those who decided

[7] Or at least the differences $E[Y^{r'}|X] - E[Y^{r''}|X]$ which is what we are interested in since changes in the levels do not affect the optimal treatment choice.

[8] Which, as discussed above, may also include the treatment: "not receiving any drugs" or "not participating in any training".

[9] In an experimental setup with randomized assignment, identification is straightforward. Using the JTPA experimental data, Plesca and Smith (2005) examine targeting in this situation.

not to take treatment d:

$$E[Y^d|D=d] \;\neq\; E[Y^d|D\neq d] \;\neq\; E[Y^d],$$

e.g. it may be that those who decided to participate in training may be more motivated or higher skilled than those who did not. This is the well known selection problem (Heckman and Robb 1985, Manski 1993).

If the set of covariates X contains only a few characteristics such as age and gender for example, it will usually still be the case that

$$E[Y^d|X, D=d] \;\neq\; E[Y^d|X],$$

e.g. among women of a certain age the better skilled received training, leading to selection bias. Since the potential outcome Y^d can only be observed for those who actually took treatment d and are counterfactual for everyone else, the expected potential outcome $E\left[Y^d|X\right]$ is not identified. However, if the sets X and W together contain all confounding variables, i.e. all variables that affected treatment choice D as well as the potential outcomes Y^d, conditioning on X *and* W eliminates selection bias:

$$E[Y^d|X, W, D=d] = E[Y^d|X, W]. \tag{1}$$

This assumption is also known as *selection on observables* (Barnow, Cain, and Goldberger 1981), *ignorable treatment assignment* (Rosenbaum and Rubin 1983) or as *conditional independence assumption* (Lechner 1999). Since $E[Y^d|X, W, D=d] = E[Y|X, W, D=d]$, the expected potential outcomes $E[Y^d|X, W]$ are identified.

Plausibility of this ignorability assumption often requires a very rich and informative database with many variables W. Many of these variables are, however, not available when it comes to making a choice decision for individual i and can thus not be included in X. Therefore we need to identify $E\left[Y^d|X\right]$ instead. This can be achieved by integrating out the W variables as

$$E[Y^d|X] = \int E[Y^d|X, W] \cdot dF_{W|X} = \int E[Y|X, W, D=d] \cdot dF_{W|X}.$$

Hence, the expected potential outcomes are nonparametrically identified, provided that $Supp(W|X, D=d) = Supp(W|X)$ or equivalently that

$$\Pr(D=d|W, X=x) > 0 \qquad \text{a.s.} \qquad \text{for all } d$$

for every x where predictions need to be made. This is a common support condition.

Hence, in principle $E\left[Y^d|X\right]$ is *nonparametrically* identified and could be estimated using nonparametric regression for $E\left[Y|X,W,D=d\right]$ and weighting this regression function by an estimate of $dF_{W|X}$. If X is *discrete* with only a few different mass points, e.g. gender by age groups, $E\left[Y^d|X\right]$ can be estimated separately for each value of X e.g. by conventional matching estimators as in Heckman, Ichimura, and Todd (1998) and Imbens (2004). $E\left[Y|X,W,D=d\right]$ is estimated by nonparametric regression and $dF_{W|X}$ is estimated by the empirical distribution function of W in the $X=x$ subpopulation, which gives:

$$\widehat{E\left[Y^d|X=x\right]} = \frac{1}{N_x}\sum_{j:X_j=x}\hat{m}_{d,x}(W_j)$$

where N_x is the number of observations with $X_j=x$ and $\hat{m}_{d,x}(w)$ is a nonparametric regression estimator of $m_{d,x}(w)=E\left[Y|X=x,W=w,D=d\right]$.

However, if one intends to obtain finer predictions in the sense that there are many more X-partitions in the population, generated by continuous regressors and/or discrete regressors with many mass points, as in our application, this nonparametric approach to integrate out the W characteristics may not work well anymore. The number of observations with $X_j=x$ would be very small or *zero* and estimating $dF_{W|X}$ by the empirical distribution function of W in the $X=x$ subpopulation will not be possible anymore or would be very imprecise. A more involved nonparametric density estimate of $dF_{W|X}$ that also incorporates observations with $X_j\neq x$ but very close to x would be required. However, integrating out the nonparametric density $d\hat{F}_{W|X}$ may then lead to rather imprecise estimates of $E\left[Y^r|X\right]$. In this situation, *parametric* or semiparametric approaches may be more appropriate to obtain less variable estimates.

In addition, there is also a practical concern about nonparametric estimation in that it may be too time consuming. If X contains a large predictor set, it will no longer be feasible to tabulate all estimates of $E\left[Y^r|X\right]$ for $r=1,...,R$, rather they have to be provided e.g. through a database via the internet. Estimating $E\left[Y^r|X\right]$ for all possible values of X will be computationally inefficient, and it would be more appropriate to estimate $E\left[Y^r|X\right]$ on demand, i.e. at that time when for a patient or an unemployed person with certain characteristics X_i a decision has to be taken. With a large database and an inference procedure with stochastic simulators for the critical values, as is discussed later, nonparametric estimation can be slow.

Another reason might be data security concerns. With a parametric specification, the data $\{(Y_j, X_j, W_j, D_j)\}_{j=1}^N$ are needed only once to estimate the coefficients. Thereafter $\hat{E}[Y^r|X]$ can be calculated from the estimated coefficients, the full data set is no longer needed and can be disconnected from the software producing the predictions, as is the case in our application. A nonparametric approach would always require direct access to the full dataset for estimating $E[Y^r|X]$.

Out of these various reasons, employing a parametric specification for $E[Y^r|X]$ might be helpful to obtain faster and more precise predictions. Let the expected potential outcomes be parametrically specified as

$$E[Y^r|X = x] \doteq \varphi(x; \theta^r) \qquad \text{for } r = 1, ..., R$$

where φ is a known function and θ^r an unknown coefficient vector of known finite dimension k.[10] To obtain precise predictions of Y^r in L_2 distance one would like to choose the coefficients θ^r as

$$\theta_*^r = \arg\min_\theta E\left[\left(Y^r - \varphi(X;\theta)\right)^2\right] \tag{2}$$

or equivalently

$$\theta_*^r = \arg\min_\theta E\left[\left(E[Y^r|X] - \varphi(X;\theta)\right)^2\right].$$

However, estimation of θ_*^r is not feasible since the potential outcomes Y^r are not observed: Y^r is only observed for those individuals who received treatment r but not for all the other individuals. Nevertheless, one can show that the minimizer of (2) is identical to the minimizer of an expression that does not contain any potential outcomes:

Theorem 1 *The two minimizers in the following expression are identical:*

$$\arg\min_\theta E\left[\left(Y^r - \varphi(X;\theta)\right)^2\right] = \arg\min_\theta E\left[\left(\frac{Y \cdot 1(D = r)}{p^r(X, W)} - \varphi(X;\theta)\right)^2\right], \tag{3}$$

where

$$p^r(x, w) = \Pr(D = r|X = x, W = w).$$

[10] The function φ could also be permitted to be different for each treatment r. For reasons of comparability the same functional form φ is used for all treatments $r = 1, ..., R$, and the differences arise through different estimated θ^r.

This can be shown by noting that

$$E\left[\frac{Y \cdot 1\left(D=r\right)}{p^r(X,W)}|X\right] = E\left[Y^r|X\right]. \tag{4}$$

Proof in appendix.

Having estimated the coefficients θ^r for all treatments $r = 1, ..., R$, the expected potential outcomes can be predicted for an individual i as

$$\hat{Y}_i^1, ..., \hat{Y}_i^R$$

where

$$\hat{Y}_i^r = \hat{E}\left[Y^r|X = X_i\right] = \varphi(X_i; \hat{\theta}^r)$$

and the optimal treatment for individual i is estimated to be

$$\hat{r}_i^* = \arg\max_r \hat{Y}_i^r.$$

This information can then be provided to the individual or to the agent to assist treatment choice.

In addition to these predictions themselves, it may also often be of interest to have some information about the statistical precision in the estimation of $\hat{r}_i^*$. If $\hat{r}_i^*$ is very imprecisely estimated, the agent or the individual may not want to trust these estimates very much and may use other information to base her decision on.[11] On the other hand, if $\hat{r}_i^*$ is very precisely estimated, the agent will be more likely to follow these statistical predictions.

For practical purposes it is important to convey this information about statistical precision to the individual, agent or case worker in a simple and accessible way. Providing case workers with standard errors or variance-covariance matrices would not be appropriate since case workers are usually not trained in thinking in terms of confidence intervals or statistical tests. As a more transparent alternative, we suggest to group the available treatments into three categories: 'good', 'intermediate' and 'bad' treatments, based on the results of a Multiple

[11] Other aspects of the treatment that have not been included in the outcome variable Y^r might then be considered as well, e.g. the costs of treatment, including opportunity costs, or other variables that are difficult to quantify or to measure. There might also be other considerations to be taken into account such as waiting times, quantity restrictions, supply constraints, etc.

Comparison With the Best (MCB) analysis. These results can easily be shown and explained to case workers and other decision makers.

Let r_i^* denote the (unknown) best treatment for each individual

$$r_i^* = \arg\max_r Y_i^r.$$

We would like to know how likely the best treatment r_i^* and the *estimated* best treatment $\hat{r}_i^*$ coincide. If this probability is not very large, we also would like to know which other treatments might be good as well. In other words, we would like to estimate a set $\hat{S}_i$ of treatments that contains the best treatment with high probability

$$\Pr(r_i^* \in \hat{S}_i) \geq 1 - \alpha,$$

where $1 - \alpha$ is the confidence level. A multiple comparison with the best approach, see Hsu (1996) and Horrace and Schmidt (2000), produces estimates of the set $\hat{S}_i$ as well as confidence intervals for $Y_i^{r_i^*} - Y_i^r$ such that

$$\Pr\left(r_i^* \in \hat{S}_i \quad \text{and} \quad \hat{L}_{i,r} \leq Y_i^{r_i^*} - Y_i^r \leq \hat{U}_{i,r} \quad \text{for all} \quad r = 1, ..., R\right) \geq 1 - \alpha,$$

where $\hat{L}_{i,r}, \hat{U}_{i,r}$ are estimated lower and upper bounds. The estimates of $\hat{S}_i$ and $\hat{L}_{i,r}$ thus distinguish three types of treatments. A treatment r with $r \in \hat{S}_i$ belongs to the set of best treatments. A treatment r with $\hat{L}_{i,r} > 0$, i.e. $Y_i^{r_i^*} > Y_i^r$, is clearly *worse* than the best treatment. Finally, treatments with $r \notin \hat{S}_i$ but $\hat{L}_{i,r} = 0$, which implies $Y_i^{r_i^*} \geq Y_i^r$, are intermediate in that they do not belong to the set of best treatments nor are they clearly worse than the best treatment. For details see Horrace and Schmidt (2000).

We therefore suggest to provide the case worker with the predictions $\hat{Y}_i^1, ..., \hat{Y}_i^R$ together with an estimate of the sets of 'good', 'intermediate' and 'bad' treatments. If $\hat{S}_i$ contains only a single element, the case worker can be rather confident that the estimated best treatment $\hat{r}_i^*$ is likely to be the best choice. If $\hat{S}_i$ contains a few treatments, at least he knows which is the estimated best treatment, which other treatments might be good as well and which treatments are probably worse. On the other hand, if $\hat{S}_i$ contains (almost) all treatments, the case worker knows that the information available in the statistical system is insufficient and too unreliable to be of much assistance for this individual. In this situation the case worker may want to follow other guidelines for treatment choice, e.g. waiting times, supply constraints, personal preferences, programme goals that are not easily quantifiable (and thus

cannot be included in the statistical selection), etc. The cardinality of the set $\hat{S}_i$ may vary from individual to individual according to the characteristics X_i and it is quite likely that for some individuals $\hat{S}_i$ will be a singleton, whereas for other individuals the set $\hat{S}_i$ may contain all available treatments. Hence, we can distinguish between individuals where the statistical system provides precise estimates and individuals where it fails to provide useful information.

The following sections describe how this statistical targeting system was implemented for Swiss active labour market policies and piloted in several employment offices in 2005.

3 Application to active labour market programmes

In many countries active labour market policies have been introduced during the 1990s to combat the problems of high and persistent unemployment or low earnings of disadvantaged groups. Active labour market programmes may comprehend job search training, placement services, counselling, training in computer skills, language training, vocational training, employment programmes (job creation schemes), wages subsidies etc. These courses may be of a few weeks up to several months duration and aim to increase job search intensity and effectiveness, increase human capital or ameliorate its deterioration, increase the number of employer contacts or provide psychological support to increase employability. Such training programmes may be implemented on a limited scale such as the Job Training Partnership Act (JTPA, Heckman, Ichimura, Smith, and Todd (1998)) in the USA or on a large scale as e.g. in Germany or Sweden.

Many countries introduced ALMP on a large scale, expecting them to reduce mass unemployment rapidly. The initial enthusiasm has waned in the recent years since several evaluation studies found rather moderate or even negative effects.[12] These results have prompted several changes in the design of ALMP: programmes have been modified, negative incentive mechanisms reduced[13] and individuals were assigned less frequently to such programmes. There has

[12] There is usually a substantial lock-in effect in the form of reduced job search whilst in the programme that would have to be compensated for by a considerably higher job finding rate after the training, which often does not seem to be the case. See e.g. Bloom, Orr, Bell, Cave, Doolittle, Lin, and Bos (1997), Fay (1996), Gerfin and Lechner (2002), Lechner (2000) or Puhani (1999), among many others.

[13] In several countries, participation in ALMP extended the entitlement period for unemployment benefits in that another entitlement period was granted after participating in ALMP for a sufficiently long time. ALMP were then a route to obtain benefits for extended time periods.

been a general trend towards providing ALMP only or predominantly to those individuals, who are expected to benefit from it, reflecting the belief that ALMP are neither beneficial for everyone nor harmful to everyone. To support such a more deliberate targeting of ALMP, statistical *profiling systems* have been piloted in several countries, often with mixed results.[14]

Profiling attempts to estimate the risk of becoming long-term unemployed when not receiving any assistance and assigns those unemployed who are most at risk to ALMP. Implicit is the assumption that those least likely to become long-term unemployed do not benefit (much) from ALMP, whereas those with the largest risk will benefit most from these programmes. This implicit assumption may often not be true as found e.g. in Berger, Black, and Smith (2001), Black, Smith, Berger, and Noel (2003) or Rudolph and Müntnich (2001).[15] Basically, profiling systems base their decision only on estimates of the potential outcome Y^1, where $r = 1$ represents the treatment "no participation in ALMP". The potential outcomes $Y^2, ..., Y^R$ for participation in different ALMP are not estimated and thereby neglected.

The aim of the *targeting system* developed in this paper is to estimate Y_i^1 along with $Y_i^2, ..., Y_i^R$ for every individual and to determine the ALMP that provides the highest outcome.[16] This targeting system was implemented in a pilot study in Switzerland that took place from May to December 2005.[17] Targeting systems that were based on similar objectives have partly been implemented in Canada and the USA. (A similar system is currently developed in Germany.) The Frontline Decision Support System (FDSS) in the USA, see Eberts, O'Leary, and DeRango (2002), predicts expected earnings for different

[14] See e.g. OECD (1998), de Koning (1999), DOL (1999), Berger, Black, and Smith (2001), Rudolph and Müntnich (2001), Colpitts (2002), Eberts (2002), Eberts, O'Leary, and DeRango (2002), Wandner (2002), Black, Smith, Berger, and Noel (2003) and Plesca and Smith (2005).

[15] For example, Berger, Black, and Smith (2001) and Black, Smith, Berger, and Noel (2003), in their analysis of the worker profiling system in the USA, find a relatively good predictability of long-term unemployment, but do not find evidence for programme effects and profiling scores being correlated or even being strictly monotonously related. It seems that individuals in the middle ranges of the profiling score benefitted most from treatment. Profiling is likely to perform even worse if a variety of different and heterogenous programmes ($R > 2$) is available. In a model project in Germany, no positive effects of case management on the reemployment chances of people identified to be at risk of getting long-term unemployed were found (Rudolph and Müntnich 2001).

[16] For a further discussion on profiling and targeting systems see Frölich, Lechner, and Steiger (2003), Plesca and Smith (2005) and Lechner and Smith (2006).

[17] Early results of the evaluation of this pilot study are expected in the first half of 2007.

training programmes using OLS regressions and was piloted in Georgia in 2002.[18] Canada developed its Service and Outcome Measurement System (SOMS) from 1994 to 1999 (Colpitts 2002), which was designed as a support system for service delivery staff who still had full discretionary power. A huge database had been constructed by merging a number of different datasets. SOMS, however, was never implemented mainly because of data security concerns, and the SOMS database had to be deleted in 2002 by a ruling of the Privacy Commissioner. This indicates that data security may be a sensitive issue and should be taken seriously when developing a targeting system. The system developed in this paper permits to incorporate additional covariate information, which may be available from social security data or other sources, in the estimation process without the need for having them available when predicting outcomes. A huge database may be necessary once for estimation, but can be disconnect afterwards. It also provides information to the case workers about the statistical precision of the estimated best programme.

3.1 ALMP in Switzerland

In Switzerland, the unemployment rate had been very low during most of the past century until it increased with the recession of the early 1990s to levels not seen before. It reached a peak at 5.7% in 1997 and stayed around 3.5 to 4% from 2003 to 2006. This triggered a complete revision of the Swiss unemployment insurance system in 1996, which made the provision of active labour market programmes a first priority: The federal states (cantons) were forced to provide a minimum number of active labour market programme places, and participation was made mandatory for every unemployed person if allocated to a programme by the case worker. (Allocation to a programme is at the case worker's full discretion, and non-compliance leads to a suspension of benefit payments.) A first evaluation of these Swiss active labour market programmes in Gerfin and Lechner (2002) and Gerfin, Lechner, and Steiger (2005) found negative employment effects of some of the programmes and positive effects for others. In an evaluation of the effectiveness of case workers in allocating individuals to programmes, Lechner and Smith (2006) found that case workers did not seem to be very successful in selecting

[18] The pilot study in Georgia was discontinued for "several reasons" and "was not in place long enough to undergo a rigorous evaluation" (Eberts and Randall 2005). Nevertheless, Eberts and Randall (2005) also mention a similar project directed towards welfare recipients where a randomized pilot study found large positive impacts of the statistical system.

the most beneficial programme and indicated a substantial potential for improvement.[19]

Based on these and other evaluation results, the Swiss State Secretariat for Economic Affairs (seco) initiated a pilot study on targeting active labour market services in 21 employment offices: Case workers should be assisted in their treatment choices with statistical information.

3.2 Categories of labour market programmes

A large number of different programmes is available in Switzerland and these programmes might also vary somewhat from region to region. The official classification distinguishes 43 different types, of which most are training or employment programmes. To incorporate regional differences in these programmes and in the composition of unemployed and the local labour market situations, the statistical system was estimated separately for five different regions: Basel, Bern, Geneva, St.Gallen and Zurich. In addition, separate estimates for jobseekers with mother tongue identical to the local language (German or French) and for those with a different mother tongue were derived. In the following only the results for jobseekers with non-German mother tongue in *Basel city* are shown exemplary. (The results for the other regions are available from the author.) In Basel the ALMP are categorized into six (R=6) different groups:

1 No programme
2 Job search and personality courses
3 Language skills training
4 Computer skills training
5 Further training
6 Employment programmes

The first treatment 'No programme' means that the jobseeker is not allocated to any ALMP in this month, but leaving the option for the future, if still unemployed then. This category could therefore also be labelled as 'waiting' or 'no programme now but perhaps later'. This has to be distinguished from a treatment 'no programme at all' or 'no programme for the next 12 months' or 'no programme for the entire unemployment spell'. Such a programme does not exist in the above list out of two reasons: First, forgoing the option to choose a labour market programme at a later time is not really a practical option for a case worker. The case

[19] Bell and Orr (2002) found similarly for the USA that case workers may often not be systematically selecting those into treatment who would benefit most from it.

worker meets the jobseeker about once a month and decides about actions to be taken then. Sequential plans may be developed but at every meeting the latest information and events are incorporated to update such plans. Second, identifying the effect of a treatment 'no programme for the next 12 months' is more difficult than for a treatment 'no programme now but perhaps later' because of the dynamic nature of the job search. When examining previous participants in 'no programme for the next 12 months', many of them had been lucky enough to find a job before a programme had been assigned. Hence, this group may contain a larger proportion of good risks or individuals successful in the job search. For a further discussion see Fredriksson and Johansson (2003) and Sianesi (2004).

The treatment categories two to six contain active programmes.[20] The second treatment consists of a variety of often short-term basic courses, including training in effective job search strategies and resume writing and more intensive personality courses, which provide psychological backing for handling the shock of becoming unemployed and coaching in developing new perspectives to entering the labour market. These courses may be tailored to different groups (manual workers, management) and offered in different languages.

The third treatment contains language and communication skills training for foreigners (including alphabetization courses, basic skills in dealing with Swiss administrations and vocational language courses for low educated foreigners[21]) as well as courses in foreign languages at different levels. Treatment group four, computer training, refers mostly to general courses in office applications such as word processing and spread sheet calculations, but also stock-keeping and order management software. The fifth treatment consists of further training in the jobseeker's occupation and are often of one week to two months duration. (Re-training to a new profession is not offered by Swiss ALMP.)

The sixth treatment consists of subsidized employment programmes or job creation schemes in a sheltered labour market, usually of three to six months duration. This includes activities in cantonal and municipal administrations (including hospitals, kindergartens, schools, nursing homes) and non-regular workplaces in charitable, cultural, recycling, environmental protection or other non-profit organizations. Internships are also included in this category.

[20] Only courses of at least five days duration are included. Shorter courses are included in the no programme category. Such may be short evening courses that provide information on the duties and rights of unemployed or language proficiency tests for assessing the need for a language course or its appropriate level.

[21] Learning occupation specific vocabulary e.g. in the construction or hotel and restaurant industry.

Given the large number of active labour market programmes available in Switzerland the above grouping into only 5 broad categories may appear rather rough. There are several reasons for not choosing very narrow categories, though. One reason is statistical precision in that the number of observations available in the dataset would be very small for some courses. But there are also more substantial issues. First, all of the R available treatments should make sense for every jobseeker. If one of the treatments were defined as a language course for foreigners, it would not be a reasonable option for a Swiss jobseeker and no predictions should be made as such a programme would be dismissed from the outset. The choice set $R \in \{1, ..., 6\}$ would thus depend on the characteristics X_{it} and has to be treated as a function of X_{it}, which would complicate the implementation. By defining a category *language skills training* which includes German, French and foreign language courses, this category becomes feasible for every jobseeker, and the X_{it} characteristics (e.g. mother tongue, profession) define which type of language course or further training is appropriate.

A second reason is that the case worker may actually have much better information for choosing the exact course out of a broader category. The statistical system may be able to estimate how much the labour market values different types of training, but cannot recommend whether an advanced or intermediate English course would be more appropriate. The case worker may also know better about local waiting lists or supply constraints that are to be taken into account when allocating a course.

Third, in the pilot study employment predictions are made for the year 2005/06 based on data on participants of the years 2001 to 2003. During these years, some of these courses have been modified and providers have changed in several details. But the broader structure of these programmes remained largely unchanged. Therefore we do not want to define treatments too narrowly, as specific courses may be rather different today.[22]

In addition to defining the treatments, another fundamental parameter of the system is the definition of the outcome variable. We define the outcome variable $Y_{i,t+\tau}$ for individual i

[22] The above treatments contain only programmes that a case worker can actively assign. The Swiss labour market policy also provides a few other instruments, such as subsidies for temporary jobs (interim jobs), regular jobs (settling-in allowances) and self-employment assistance. These are not included in the statistical system since the former are largely contingent upon that a job has already been found (and thus cannot be assigned directly by the case worker) and since the occurrence of self-employment assistance is relatively rare and the selection problem more difficult to handle.

when a decision is taken at time t as the number of months in *stable* employment within the subsequent 12 months, divided by 12. An employment spell is considered as stable if it is of at least three months duration. Hence, jobs of very short duration e.g. a few days or weeks are not considered as a positive outcome. This is largely in line with the official aims of the Swiss State Secretariat for Economic Affairs (seco), which emphasizes rapid re-employment but avoiding re-registration of unemployment.

3.3 Data and variables

Two types of datasets are required for implementing the statistical system. First, an extensive data base on previously treated is needed containing sufficient information on X and W variables to make the conditional independence assumption (1) valid. This requires individual information on personal characteristics and labour market histories as well as a sufficient number of observations for precise estimation. The second dataset refers to the new clients for whom predictions about their expected outcomes shall be made. For them only information on the X variables is needed, at the time the treatment has to be chosen.

The first dataset for the estimation of the coefficients θ^r consists of the entire population of individuals that registered as job seekers at an employment office anytime during January 2001 to December 2003. For these 460442 persons, information from the unemployment insurance information system (AVAM/ASAL) is available up to December 2004. This data is matched with information from the social security records (AHV) for the period January 1990 to December 2002. These combined data sources contain very detailed information on registration and de-registration of unemployment, benefit payments and sanctions, participation in ALMP, eleven years employment histories with monthly information on earnings and employment status (employed, unemployed, non-employed, self-employed) and a lot of information on socioeconomic characteristics including qualification, education, language skills (mother tongue, proficiency of foreign languages), job position, experience, profession, industry and an employability rating by the case worker.

The data for the new clients during the pilot study in 2005 is based on the unemployment insurance information system for all new jobseekers and is updated every two weeks. It does not contain any social security information, and some information on previous participation in ALMP and interim jobs becomes available only with a delay. These variables are thus only

available for the 460442 past treatment participants and therefore can be included only as W regressors.

Table 3.1 contains descriptive statistics on selected X and W variables for the 460442 past participants. About half of the jobseekers are female and 44% married. Switzerland has a large population of foreigners and they represent almost 40% of the jobseekers. (This is one of the reasons why information on language proficiency is important.) The X variables contain information on education, qualification, language skills, employability, insured earnings and information about the previous occupation. Very detailed information on employment history is available from the social security data, which however can only be used as W variables. It shows that more than half of the jobseekers were never unemployed before (in 1991 to 2000), while 28% had at least two unemployment spells.

— Table 3.1 about here —

— Table 3.2 about here —

Table 3.2 lists all the X and W variables that were used in the estimation. The X variables contain individual characteristics, but also variables characterizing the season (*month*) and the local labour market (*regional unemployment rate, industry unemployment rate, industry vacancy rate*). These variables are important since the employment data from 2002 to 2004 are used for predicting employment outcomes in 2005/06. These variables should reflect the business cycle reasonably well, which anyhow did not fluctuate very much: The Swiss unemployment rate was very stable around 3.5 to 4% during the entire period 2003 to 2006.[23]

Since jobseekers can be assigned to a treatment anytime during their unemployment spell in our observation window 2001 to 2003, we can define $D_{jt}, X_{jt}, W_{jt}, Y_{j,t+\tau}$ on a monthly basis[24] yielding 36 panel observations for every person. This gives a total of 16.6 million panel observations (460442 individuals times 36 months).

In any month a jobseeker might have been assigned to a programme $D_{jt} \in \{2, 3, 4, 5, 6\}$ or it might have been decided to continue job search without an active labour market programme $D_{jt} = 1$, at least until the next counselling interview takes place. In most of these months,

[23] Most of the observations are from the year 2003, whereas only 13% of the observations in decision relevant situations are from 2001.

[24] In principle, even on a daily basis.

however, no decision was taken at all (D_{jt} undefined) e.g. because the individual was already in a programme of longer duration or had found a job. Such months do not represent choice relevant situations since the case worker would not initiate a training or employment programme under these circumstances.[25] Let S_{jt} indicate whether person j in month t was in a choice relevant situation or not. S_{jt} is zero (1) if the person is not entitled to participate in ALMP because of not being registered as unemployed or not having contributed sufficiently to the unemployment funds (minimum contribution duration is 12 months) or having exhausted the unemployment benefits entitlement period or (2) if she already participates in a training or employment programme of longer duration or (3) if she is temporarily employed in an interim job or (4) if she de-registers anytime during this month. Otherwise, $S_{jt} = 1$. After deleting all panel observations with $S_{jt} = 0$,[26] 2.3 million observations remain. In most of these months, treatment 1 (i.e. no programme) was selected.

In a next step, the sample was restricted to those regions where the pilot employment offices were located and focussed on the population with strongest labour force attachment (age between 20 and 60, not disabled, unemployed for less than 2 years and not exhausted entitlement, not being foreigner with less than yearly permit). Since the pilot study took place only during spring to autumn 2005, the winter months December to February were deleted to avoid modelling the winter peak.

4 Implementation

4.1 Estimation

The implementation of the targeting system proceeds in two steps. First, all the coefficients θ^r for the parametric specifications of $E[Y^r|X]$ are estimated on the basis of the previously described data set. Second, once these estimates and their variances have been obtained, expected potential outcomes and best treatment choices can be predicted for new clients.

[25] In principle, a case worker might already start planning the next training programme while the jobseeker is still in training. In practice this is very unlikely, though, since jobseekers should be given ample time for job search after every programme (including temporary jobs which release the financial burden on the unemployment system) and also due to time constraints on the side of the case workers. At worst we lose a few atypical observations.

[26] And also those where a programme labelled "other courses" or a subsidy for a temporary job (interim jobs) or a regular job (settling-in allowances) or a self-employment assistance started. See Footnote 22 above.

The estimation of the coefficients θ^r for the expected potential outcomes $E[Y^r|X]$ itself proceeds in two steps, separately for each of the R programmes. In a first step the propensity score is estimated by maximum likelihood logit by specifying

$$p^r(x, w; \beta^r) = \Lambda\left(x'\beta_x^r + w'\beta_w^r\right),$$

where $\beta^r = (\beta_x^r, \beta_w^r)$, and $\Lambda(u) = \frac{1}{1+e^{-u}}$ is the logistic function, and x includes a constant. Since the propensity score enters inversely in the estimation of the function $\varphi(x; \theta^r)$, very small estimated propensity scores could affect the results unduly. Therefore the estimated propensity scores $\hat{p}_j^r$ are capped at 0.02 of the mean of the propensity scores in the $D = r$ subpopulation.[27]

In the next step, the conditional expectation functions $E[Y^r|X]$ are estimated using the relationship between potential outcomes Y^r and observed outcomes Y in (4):

$$E[Y^r|X] = E\left[\frac{Y \cdot 1(D = r)}{p^r(X, W)}|X\right].$$

The parametric specification of $E[Y^r|X]$ should take the particularities of the outcome variable into account. Since the outcome variable is defined as the number of months in stable employment in the following year, divided by twelve, it is bounded between 0 and 1. A simple logit specification does not fit this outcome variable very well, though, since there is a large mass point at zero months of employment. (About two thirds of the observations.) It appears more appropriate to consider this outcome as a result of two processes: First, finding a job

$$\Pr(Y^r > 0|X)$$

and second keeping this job for a number of months[28]

$$E[Y^r|Y^r > 0, X].$$

We use a logit model for the *binary* variable $1(Y^r > 0)$, i.e. the probability of finding a job is specified as

$$\Pr(Y^r > 0|X = x) = \Lambda\left(x'\theta_1^r\right) \tag{5}$$

[27] I.e. if the estimate $\hat{p}_j^r$ was below 2% of the subpopulation mean, it was set to $0.02 \cdot$ *subpopulation mean*. Variations of this threshold did not affect the results much. At the same time, all other estimated propensity scores are reduced to ensure that $\sum \frac{1(D_j=r)}{\hat{p}_j^r}$ remains unchanged by this capping.

[28] There are also individuals who found a job, lost it, became unemployed and found another job. This is rather rare, though, given the short observation window of twelve months.

where θ_1^r is a vector of unknown coefficients. The length of keeping this job $Y^r|Y^r > 0$ is bounded between zero and at most one year during our observation window. To implement this restriction, we use a logistic function for

$$E\left[Y^r|Y^r > 0, X = x\right] = \Lambda\left(\alpha^r + \gamma^r x'\theta_1^r + x_2'\theta_2^r\right), \tag{6}$$

where α^r, γ^r and θ_2^r are unknown coefficients, and x_2 is a subset of x. Without the term $x_2'\theta_2^r$ in the expression (6), the expected job duration would be assumed to depend on the characteristics X only through the same *single index* $x'\theta_1^r$ as the probability of finding a job (5). Including some variables X_2 in specification (6) does permit that these variables have a different impact on job duration than on the job finding probability.[29] Let θ^r denote all coefficients together $\theta^r = (\theta_1^{r\prime}, \theta_2^{r\prime}, \alpha^r, \gamma^r)$. With these two specifications, the conditional expectation is given by $E\left[Y^r|X\right] = E\left[Y^r|Y^r > 0, X\right] \cdot \Pr\left(Y^r > 0|X\right)$. Notice that whereas each of the two logistic functions is symmetric, the implied specification for $E\left[Y^r|X\right]$ is usually asymmetric. This gives the two moment conditions

$$E\left[\frac{1\left(Y > 0\right) \cdot 1\left(D = r\right)}{p^r(X, W)} - \Lambda\left(X'\theta_1^r\right) \;|X\right] = 0 \tag{7}$$

and

$$E\left[\frac{Y \cdot 1\left(D = r\right)}{p^r(X, W)} - \Lambda\left(X'\theta_1^r\right)\Lambda\left(\alpha^r + \gamma^r X'\theta_1^r + X_2'\theta_2^r\right) \;|X\right] = 0. \tag{8}$$

These moment conditions identify θ^r, given estimates of β^r for the propensity scores.

Since observations are independent across individuals but not over time, for estimating standard errors of $\hat{\theta}^r$ it is convenient to stack all the observations for the different months for the same individual in a vector of moment conditions. This will also easily permit us to take the effect of the first step estimation of β^r on the variance of $\hat{\theta}^r$ into account. As the winter months December, January and February have been left out to avoid modelling the winter seasonal effects for the different professions,[30] there are 27 months for each individual. Yet, only in some of these months an individual may have been in a choice relevant situation, as indicated by the variable S_{jt}, which was defined in Section 3. The moment condition for individual j in month t is then

$$m_{jt}^r = m^r(Z_{jt}, \theta^r; \beta^r) = \begin{pmatrix} S_{jt} \cdot \left(\frac{1(Y_{jt}>0)\cdot 1(D_{jt}=r)}{p_{jt}^r} - \Lambda(X_{j1}'\theta_1^r)\right) \\ S_{jt} \cdot \left(\frac{Y_{jt}1(D_{jt}=r)}{p_{jt}^r} - \Lambda(X_{jt}'\theta_1^r)\Lambda(\alpha^r + \gamma^r X_{jt}'\theta_1^r + X_{2jt}'\theta_2^r)\right) \end{pmatrix}$$

[29] Including too many variables in X_2 may result in convergence problems of the GMM estimator, though.

[30] The pilot sudy took place during spring, summer and autumn only.

where $Z_{jt} = (Y_{jt}, D_{jt}, X_{jt}, W_{jt}, S_{jt})$ contains all the data for individual i in month j, with

$$E[m^r(Z_{jt}, \theta_0^r; \beta_0^r) \mid X_{jt}] = 0.$$

Stacking the moment conditions for the 27 months for the same individual

$$m_j^r = m^r(Z_j, \theta^r; \beta^r) = \left(m_{j1}^{r\prime}, ..., m_{j27}^{r\prime}\right)' \tag{9}$$

where $Z_j = (Z_{j1}, ..., Z_{j27})$ and assuming strict exogeneity of X_{jt} gives

$$E[m^r(Z_j, \theta_0^r; \beta_0^r) \mid X_j] = 0,$$

where $X_j = (X_{j1}, ..., X_{j27})$. The vectors m_j^r are independent across individuals and conventional results for GMM estimators with iid data apply. Taking β_0^r as given (since β_0^r is just-identified from the logit specification of the propensity score), the optimal *unconditional* moment function would be

$$g_j^r = g^r(Z_j, \theta^r; \beta_0^r) = E\left[\frac{\partial m^r(Z_j, \theta_0^r; \beta_0^r)}{\partial \theta^{r\prime}} | X_j\right]' \cdot \left(E\left[m^r\left(Z_j, \theta_0^r; \beta_0^r\right) m^r\left(Z_j, \theta_0^r; \beta_0^r\right)' \mid X_j\right]\right)^{-1} \cdot m^r(Z_j, \theta^r; \beta_0^r) \tag{10}$$

with

$$E[g^r(Z_j, \theta_0^r; \beta_0^r)] = 0,$$

see Newey and McFadden (1994). The corresponding GMM estimator would estimate θ^r by setting the mean of the moment function $\frac{1}{N}\sum g_j^r$ to zero.

Implementing this estimator, however, requires estimates of the optimal instrument matrix as given in (10). The first term, i.e. the expected conditional gradient of the moment function is simple to obtain, because for (9) it follows that

$$E\left[\frac{\partial m^r(Z_j, \theta_0^r; \beta_0^r)}{\partial \theta^{r\prime}} | X_j\right] = \frac{\partial m^r(Z_j, \theta_0^r; \beta_0^r)}{\partial \theta^{r\prime}},$$

where then estimators of θ_0^r and β_0^r can be plugged in. The second term in (10), i.e. the conditional variance matrix of the moment function is more difficult to obtain and nonparametric estimation of the conditional variance would be computationally very expensive. Since consistent estimation of this term is only needed for efficiency but not for consistency, one can expect to obtain consistent and relatively efficient estimates if the main features of the variance matrix in (10) are incorporated in the estimator: moment functions that have a large variance

should receive less weight and moment functions with a small variance should receive a larger weight. Since the variance of m_{jt}^r largely depends on the term $p^r(X,W)$ that appears in the denominator, the variance will be large if $p^r(X,W)$ is small. Therefore, the contributions of m_{jt}^r should be downweighted if $p^r(X,W)$ is small. Since the formulae (10) contains the conditional variance given X and not given X and W, we estimate by logit the probability

$$\rho^r(x) = \Pr\left(D = r | X = x\right).$$

Observations with small estimated $\rho^r(X_{jt})$ are downweighted by multiplying the moment function with $e^{-\frac{0.02}{\rho^r(X_{jt})}}$. Let $\mathcal{P}^r(X_j)$ be the 54×54 diagonal matrix containing the weighting factors $e^{-\frac{0.02}{\rho^r(X_{jt})}}$ for every month, i.e.

$$\mathcal{P}^r(X_j) = diag(e^{-\frac{0.02}{\rho^r(X_{j1})}}, e^{-\frac{0.02}{\rho^r(X_{j1})}}, e^{-\frac{0.02}{\rho^r(X_{j2})}}, e^{-\frac{0.02}{\rho^r(X_{j2})}}, ..., e^{-\frac{0.02}{\rho^r(X_{j27})}}, e^{-\frac{0.02}{\rho^r(X_{j27})}}).$$

Thus, given first step estimates of β^r and $\mathcal{P}^r(X_j)$, the coefficients $\hat{\theta}^r$ are chosen to satisfy:

$$\sum_j \frac{\partial m^r(Z_j, \hat{\theta}^r; \hat{\beta}^r)}{\partial \theta^r}' \cdot \mathcal{P}^r(X_j) \cdot m^r(Z_j, \hat{\theta}^r; \hat{\beta}^r) = 0$$

or numerically equivalently as

$$\hat{\theta}^r = \arg\min_\theta \sum_j m^r(Z_j, \theta; \hat{\beta}^r)' \cdot \mathcal{P}^r(X_j) \cdot m^r(Z_j, \theta; \hat{\beta}^r).$$

For inference on $\hat{\theta}^r$ and thus for deriving sets of optimal treatment choices via MCB, the variance matrix of $\hat{\theta}^r$ needs to be estimated. Here, the first step estimation of β^r needs to be accounted for. This can be done conveniently using the framework of Newey and McFadden (1994, Section 6.2). The first step maximum likelihood estimator $\hat{\beta}^r$ is equivalent to the just-identified GMM estimator using the scores of the log-likelihood function as moment conditions

$$E\left[S_{jt} \cdot \left(1\left(D_{jt} = r\right) - \Lambda(X_{jt}'\hat{\beta}_x^r + W_{jt}'\hat{\beta}_w^r)\right) \cdot (X_{jt}', W_{jt}')'\right] = 0.$$

Stacking these moment conditions on top of the moment conditions m_j^r, the joint variance matrix of all moment conditions can be estimated as the outer product of the gradients. From this the asymptotic variance of $\hat{\theta}^r$ can be obtained according to Newey and McFadden (1994, Section 6.2). Similarly we can derive the *joint* variance matrix for all $\hat{\theta}^1, ..., \hat{\theta}^R$, which is needed for inference on the estimated best programme.

The following tables give some details about the estimations for Basel city. The estimation sample contained 8796 unemployed persons with non-German mother tongue registered in Basel, corresponding to 46406 panel observations in choice relevant situations, of which more than 40000 received treatment 1 (no programme), see Table 4.1. Treatment 5 (further training) on the other hand was received only by 183 observations, and treatment 4 (PC courses) by 454 observations. These small sample sizes are the result of having restricted the estimation sample to only those observations living in Basel and thus sharing the same local labour market and the same unemployment insurance administration. This is the price to pay when taking local differences in labour market programmes and their administration and in local labour market conditions serious. It will lead to less biased estimates for jobseekers living in Basel at the expense of a larger variance. Most of these treatments start within the first year of unemployment, with the exception of treatment 6 (employment programmes), which is rarely administered early in the unemployment spell (i.e. for a job search duration of less than 90 days) and rather frequently to long-term unemployed.

Regarding the outcome variable Y, the observed number of months in stable employment is zero for about two thirds of the observations and about half a year for the others. The employment outcomes are worst for treatments 3, 5 and 6.[31]

—— Table 4.1 about here ——

Table 4.2 gives some descriptive statistics of the estimated selection probabilities $p^r(X, W)$. These probabilities represent the channel through which the W variables enter in the estimation to control for selection bias. Since p^r enters inversely in the estimation equation, there is a concern that very small values of $\hat{p}^r$ might lead to a large variance of the estimates of θ^r. Table 4.2 shows that this turned out to be little concern here since only 10 of the 46406 observations needed to be capped. The first column shows the mean of $\hat{p}^r$ among all observations for the different treatments r. Since the estimates of $\hat{p}_j^r$ for observations with $D_j \neq r$ do not matter because $\frac{1}{\hat{p}_j^r}$ is multiplied with $1\left(D_j = r\right)$, the following columns concentrate on the distribution of $\hat{p}^r$ among the $D_j = r$ observations only. The mean of $\hat{p}^r$ and the quantiles of $\hat{p}^r$ relative to the mean are given. Values of the ratio above 50 are capped at 50.[32] This occurred with only 10 observations and the 99% percentile was below 20 for almost all treatments.

[31] A table with descriptive statistics on all X and W variables is available from the author.

[32] Or in other words, values of the ratio of $\hat{p}_j^r$ to the mean below 0.02 are increased to 0.02.

—— Table 4.2 about here ——

Table 4.3 shows an excerpt of the estimated θ^r coefficients, including the estimates of α^r and γ^r.[33] The coefficients α^r and γ^r determine the shape of the conditional expectation function and were introduced because a simple logistic function did not appear appropriate to model the large number of observations with zero employment outcome. Since the estimated functional forms vary by treatment, the magnitudes of the other θ^r coefficients cannot be compared across treatments but only within each treatment. The duration of job search reduces employment chances when in treatment 1 (no programme). It is insignificant for the other treatments, perhaps because job search duration is also interacted with several other variables. Age is associated with reduced employment chances, as is the status of being foreigner. German proficiency as well as the number of foreign languages increases employment prospects, as does the level of qualification. The local labour market conditions clearly impact on employment prospects, but are not always significant. Several variables could not be included in the estimations for some treatment groups due to lacking degrees of freedom. This is particularly the case for treatment 5 with the smallest number of observations. At the bottom of the table, job search duration and gender are included additionally as x_2 variables in expression (6), which models the expected job stability given employment has been found. The negative coefficient for female in $\hat{\theta}_1^r$ and the positive estimate in $\hat{\theta}_2^r$, for treatment r=1, indicates that women are less likely to obtain employment but if they do so their jobs seem to be of longer duration.

—— Table 4.3 about here ——

Near to the bottom of Table 4.3, also the estimated α^r and γ^r coefficients are shown, which determine the shape of the conditional expectation function $E[Y^r|X]$. The resulting shape as a function of the index[34] is shown exemplary for treatment 1 in the following graph, which clearly depicts the non-symmetry of the relationship. For a very large number of observations Y^r is zero. If the index surpasses a certain threshold, $E[Y^r|X]$ increases steeply to 0.6, where the curve begins to flatten and becomes rather flat at values of $E[Y^r|X]$ above 0.9. This corresponds to our modeling strategy since there are only very few observations that keep employment for 12 months with almost certainty.

[33] Tables with all estimated β^r and θ^r coefficients are available from the author.

[34] Ignoring the θ_2 coefficients.

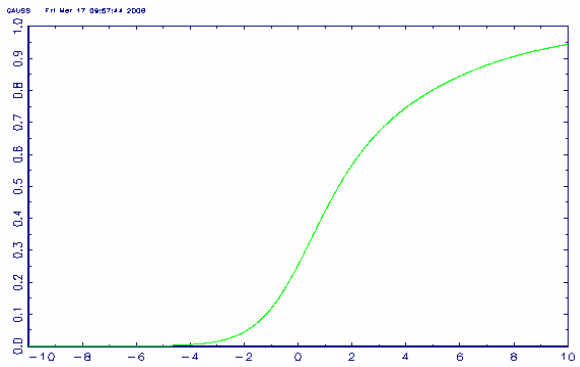

Figure 1: Estimated shape of conditional expectation function for treatment 1

These coefficient estimates will be used in the next section to predict employment chances for new clients in the pilot study. Before embarking upon predictions for new clients, the following tables provide a casual inspection of the predictions for the previous 46406 observations.[35] Table 4.4 shows the correlations among the predicted $E\left[Y^r|X\right]$ for the 46406 observations. All outcomes are positively correlated with the "no programme" outcome $E\left[Y^1|X\right]$, which is what we expected since individuals with generally good labour market chances (without training) are also very likely to enjoy good prospects with training, and vice versa. The correlations are far from one, however, which is an indication of treatment effect heterogeneity and may imply that the optimal treatment is different for different people.

—— Table 4.4 about here ——

Table 4.5 gives average prediction errors obtained by comparing $E\left[Y^r|X\right]$ with the observed Y for those observations with $D = r$. The prediction errors are smallest for treatment 3 (language courses), perhaps because there is relatively little variation in the outcomes among the language course participants since Y is zero for very many of them. The median absolute prediction error is about 0.04 to 0.17, the mean absolute error is about 0.14 to 0.23. These prediction errors are relatively large and indicate the statistical uncertainty in predicting the future employment outcomes.

[35] One may want to keep in mind that these are "in-sample" predictions, in that estimation and validation samples are partly identical. The predictions for treatments 2 to 6, however, are almost out-of-sample predictions, since only about 183 to 2556 observations are used for estimating θ^r which are then used for prediction for the other about 45000 observations.

—— Table 4.5 about here ——

4.2 Prediction of best treatment choices

Given these estimates $\hat{\theta}^1, ..., \hat{\theta}^R$ and their estimated variances, expected potential employment outcomes can be predicted for new unemployed persons. This system was piloted from May/June to December 2005 in 21 employment offices in Switzerland. Predictions were made biweekly for all registered jobseekers in the pilot offices and conveyed to the case workers via the Internet. (In fact, predictions were made only for half of the jobseekers, with the other half functioning as a control group for whom no predictions were made accessible. These two groups were randomly selected via a randomization of their case workers.) When predicting the employment chances for jobseeker i the information up to time t was taken into account in X_{it}. This covariate information changes over time not only in that the length of the current unemployment spell increases over time but also due to participation in ALMP, changes in the personal situation etc.

For individual i the predicted employment chances

$$\hat{Y}_i^1, ..., \hat{Y}_i^R$$

are computed from $\hat{\theta}^1, ..., \hat{\theta}^R$, and the set of best treatments $\hat{S}_i$ and of worst treatments is derived by multiple comparison with the best. The information provided to the case worker to assist the treatment choice for jobseeker i is in the following form

	$E[Y^r\|X]$
No programme	**0.56**
Job search and personality courses	0.23
Language skills training	0.34
Computer skills training	**0.50**
Further training	0.48
Employment programmes	0.25

where the programmes in the set $\hat{S}_i$ are marked in **bold** and the programmes with strictly positive lower bound ($\hat{L}_{i,r} > 0$) are stated in small font.

Predictions were made only for those jobseekers who belonged to the population on which the estimations were based. In particular, no predictions were made for jobseekers below the

age of 20 and above 60 since the pilot version of the statistical system was aimed at prime age individuals with attachment to the labour market. In addition, no predictions were displayed for those treatments r where the characteristics X_{it} were very different from those characteristics of the previous participants. In such a situation, the predictions $\hat{E}[Y^r|X = X_{it}]$ would be out of the support of the data on which the estimates were based and might therefore be highly biased. This support condition was implemented by noting that the propensity score

$$\rho^r(x) = \Pr(D = r|X = x)$$

provides a convenient *one-dimensional* representation of the distribution of X in the $D_j = r$ subpopulation, as it has been often used in the evaluation literature, see Rosenbaum and Rubin (1983). As mentioned in the previous section, $\rho^r(x)$ was estimated from the data sample $\{(Y_j, X_j, W_j, D_j)\}_{j=1}^N$. Let $f_{\rho^r|D=r}$ be the density function of $\rho^r(x)$ in the $D_j = r$ subpopulation, i.e. among those who actually received treatment r, and let $\hat{f}_{\rho^r|D=r}$ be the empirical density function. The support of ρ^r in the $D_j = r$ subpopulation is estimated by trimming 0.5% of the data on either side (at least 5 observations). In other words, a jobseeker i with characteristics X_{it} is considered as 'in-support' if $\hat{\rho}^r(X_{it})$ is within the 0.5 and 99.5 percentile of the empirical distribution of $f_{\rho^r|D=r}$. Otherwise the jobseeker is considered as being too different from the previous participants to predict $\hat{E}[Y^r|X = X_{it}]$ reliably. Note that predictions might still be made for the other treatments: Jobseeker i might be very different from previous participants in r but not from those in s. In particular if some treatments had very selected intakes of previous participants, e.g. only foreigners in language courses, there might be many jobseekers (i.e. all non-foreigners) who may be considered as out of support. On the other hand, the previous participants in the treatment "no programme" were so heterogenous that hardly any jobseeker would be considered as out of support.

The following Table 4.6 shows the results for the 2303 jobseekers registered in Basel city on August 23, 2006.[36] The first five rows give descriptive statistics of the predictions of $\hat{Y}^r$ for the 2303 jobseekers. With an average outcome $\hat{Y}^1$ of 2.7 months, treatment 1 (no programme) is the best of all treatments on average, whereas treatment 3 (language courses) is worst on average with only 2 months of expected employment. In fact, treatment 1 also seems to be best at different quantiles of the distributions of $\hat{Y}^r$: the median of $\hat{Y}^1$ for the 2303 jobseekers

[36] Only those jobseekers for whom predictions were made. About the same number are in the control group.

is larger than the median of $\hat{Y}^r$ of any other treatment. This is also the case for the lower and upper quartile. This does not imply, however, that treatment 1 is best for everyone since predictions are not perfectly rank correlated. Although treatment 1 may be a reasonable choice for everyone, there might for each individual still be a better choice.

This is visible in the next row in the table where each jobseeker is hypothetically allocated to the treatment with highest prediction. Here it can be seen that treatment 5 (further training) is predicted to be best for 25% of all clients, and treatment 3 (language courses) is still predicted to be best for 10% of all clients. This allocation was based entirely on the predicted outcomes $\hat{Y}_i^r$ and ignored any estimation uncertainty. The row below shows the treatment allocation that would arise if everyone were allocated randomly to a treatment within the estimated set $\hat{S}_i$.

—— Table 4.6 about here ——

Finally, the last rows show the cardinality of the sets of 'best' treatments $\hat{S}_i$, 'worst' treatments (i.e. those with $\hat{L}_{i,r} > 0$) and 'intermediate' treatments (i.e. those that belong to neither of the other two sets). For 781 of the 2303 jobseekers (= 40%) the cardinality of $\hat{S}_i$ is one, i.e. there is a single treatment that is uniquely predicted to be the best. For 683 persons $\hat{S}_i$ contains two treatments. On the other hand, for 11 persons $\hat{S}_i$ contains all six treatments and thus provides no information for treatment choice. For 65 persons $\hat{S}_i$ contains five treatments and is thus almost without information. Overall, this indicates that the statistical system rather often provides statistically useful predictions, but not for all clients, though.[37]

5 Conclusions

In this paper a methodology for statistical treatment choice has been developed, and its implementation to choosing active labour market programmes has been described. The developed methodology has two advantages over available targeting systems: First, it permits to combine a data set on previously treated individuals with a data set on new clients when the regressors available in these two data sets do not coincide. It thereby incorporates additional information

[37] Eventually, the degree of statistical precision that the case worker perceives through the cardinality of $\hat{S}_i$ depends on the choice of the confidence level. Since there is little guidance about choosing the confidence level, it was randomized among the case workers in the pilot study to enable an ex post estimation of the optimal degree of pretended statistical precision that facilitated highest employment outcomes.

on previously treated that are not available for the current clients. Such a situation often arises e.g. due to cost considerations, data confidentiality reasons or time delays in data availability. Second, statistical inference on the recommended treatment choice is analyzed and conveyed to the agent, physician or case worker in an intelligible and transparent way. The implementation of this methodology in a pilot study in Switzerland for choosing among active labour market programmes (ALMP) for unemployed job seekers has been described, where evaluation results will be available from 2007.

A Proof of Theorem 1

It is to show that

$$\arg\min_{\theta} E\left[\left(Y^r - \varphi(X;\theta)\right)^2\right] = \arg\min_{\theta} E\left[\left(\frac{Y \cdot 1\left(D=r\right)}{p^r(X,W)} - \varphi(X;\theta)\right)^2\right]. \tag{11}$$

First it is shown that

$$\begin{aligned} E\left[Y \cdot 1\left(D=r\right)|X,W\right] &= E\left[Y|X,W,D=r\right] \cdot \Pr\left(D=r|X,W\right) \\ &= E\left[Y^r|X,W\right] \cdot \Pr\left(D=r|X,W\right), \end{aligned}$$

where the last equality follows from the conditional independence assumption (1). Hence,

$$E\left[\frac{Y \cdot 1\left(D=r\right)}{p^r(X,W)}|X,W\right] = E\left[Y^r|X,W\right]$$

and it follows that

$$E\left[\frac{Y \cdot 1\left(D=r\right)}{p^r(X,W)}|X\right] = E\left[Y^r|X\right]. \tag{12}$$

Now consider the second term in (11), which can be written as

$$\begin{aligned} & E\left[\left(\frac{Y \cdot 1\left(D=r\right)}{p^r(X,W)} - \varphi(X;\theta)\right)^2\right] \\ = \; & E\left[\left(\frac{Y \cdot 1\left(D=r\right)}{p^r(X,W)} - E\left[Y^r|X\right] + E\left[Y^r|X\right] - \varphi(X;\theta)\right)^2\right] \end{aligned}$$

$$\begin{aligned} = E\left[\left(\frac{Y \cdot 1\left(D=r\right)}{p^r(X,W)} - E\left[Y^r|X\right]\right)^2\right] &+ 2E\left[\left(\frac{Y \cdot 1\left(D=r\right)}{p^r(X,W)} - E\left[Y^r|X\right]\right)\left(E\left[Y^r|X\right] - \varphi(X;\theta)\right)\right] \\ &+ E\left[\left(E\left[Y^r|X\right] - \varphi(X;\theta)\right)^2\right]. \end{aligned}$$

The first term does not depend on the coefficients θ and thus does not affect the minimizer of the entire expression. The second term is zero by iterated expectations and (12). Only the third term remains, which proves the equality of Theorem 1.

References

Barnow, B., G. Cain, and A. Goldberger (1981): "Selection on Observables," *Evaluation Studies Review Annual*, 5, 43–59.

Bell, S., and L. Orr (2002): "Screening (and creaming?) applicants to job training programs: the AFDC homemaker-home health aide demonstrations," *Labour Economics*, 9, 279–301.

Berger, M., D. Black, and J. Smith (2001): "Evaluating Profiling as a Means of Allocating Government Services," in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 59–84. Physica/Springer, Heidelberg.

Black, D., J. Smith, M. Berger, and B. Noel (2003): "Is the Threat of Reemployment Services More Effective Than the Services Themselves? - Evidence from Random Assignment in the UI System," *American Economic Review*, 93, 1313–1327.

Bloom, H., L. Orr, S. Bell, G. Cave, F. Doolittle, W. Lin, and J. Bos (1997): "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study," *Journal of Human Resources*, 32, 549–576.

Breslin, F., M. Sobell, L. Sobell, J. Cunningham, J. Sdao-Jarvie, and D. Borsoi (1999): "Problem drinkers: Evaluation of a stepped-care approach," *Journal of Substance Abuse*, 10, 217–232.

Brownell, K., and T. Wadden (1991): "The heterogeneity of obesity: Fitting treatments to individuals," *Behavior Therapy*, 22, 153–177.

Colpitts, T. (2002): "Targeting Reemployment Services in Canada: The Service and Outcome Measurement System (SOMS) Experience," in *Targeting Employment Services*, ed. by R. Eberts, C. O'Leary, and S. Wandner, pp. 283–301. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

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

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. (2002): "Using Statistical Assessment Tools to Target Services to Work First Participants," in *Targeting Employment Services*, ed. by R. Eberts, C. O'Leary, and S. Wandner, pp. 221–244. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Eberts, R., C. O'Leary, and K. DeRango (2002): "A Frontline Decision Support System for One-Stop Career Centers," in *Targeting Employment Services*, ed. by R. Eberts, C. O'Leary, and S. Wandner, pp. 221–244. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Eberts, R., C. O'Leary, and S. Wandner (2002): *Targeting Employment Services*. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Eberts, R., and W. Randall (2005): "After the doors close: assisting laid-off workers to find jobs," *Economic Perspectives*, 6/22/2005.

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.

Fredriksson, P., and P. Johansson (2003): "Program Evaluation and Random Program Starts," *IFAU Discussion Paper 2003:1*.

Frölich, M., M. Lechner, and H. Steiger (2003): "Statistically Assisted Programme Selection - International Experiences and Potential Benefits for Switzerland," *Swiss Journal of Economics and Statistics*, 139, 311–331.

Gerfin, M., and M. Lechner (2002): "Microeconometric Evaluation of the Active Labour Market Policy in Switzerland," *Economic Journal*, 112, 854–893.

Gerfin, M., M. Lechner, and H. Steiger (2005): "Does subsidised temporary employment get the unemployed back to work? An econometric analysis of two different schemes," *Labour Economics*, 12, 807–835.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J., H. Ichimura, and P. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Heckman, J., and R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labour Market Data*, ed. by J. Heckman, and B. Singer. Cambridge University Press, Cambridge.

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.

Horrace, W., and P. Schmidt (2000): "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. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Kreuter, M., and V. Strecher (1996): "Do tailored behavior change messages enhance the effectiveness of health risk appraisals?: results from a randomized trial," *Health Education Research*, 11, 97–105.

Kreuter, M., V. Strecher, and B. Glassman (1999): "One size does not fit all: the case for tailoring print materials," *Annals of Behavioral Medicine*, 21, 276–283.

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

——— (2000): "An Evaluation of Public Sector Sponsored Continuous Vocational Training Programs in East Germany," *Journal of Human Resources*, 35, 347–375.

Lechner, M., and J. Smith (2006): "What is the value added by caseworkers?," *Labour Economics*, forthcoming.

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.

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

——— (2004): "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 72, 1221–1246.

Murphy, S. (2003): "Optimal dynamic treatment regimes," *Journal of Royal Statistical Society Series B*, 65, 331–366.

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): *Early Identification of Jobseekers at Risk of Long-term Unemployment: The Role of Profiling*. OECD Proceedings, Paris.

Plesca, M., and J. Smith (2005): "Rules versus discretion in social programs: empirical evidence on profiling in employment and training programs," unpublished, University of Maryland.

ProjectMatchResearchGroup (1997): "Matching alcoholism treatments to client heterogeneity: Project MATCH posttreatment drinking outcomes," *Journal of Studies on Alcohol*, 58, 7–29.

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.

Rudolph, H., and M. Müntnich (2001): "Profiling zur Vermeidung von Langzeitarbeitslosigkeit: Erste Ergebnisse aus einem Modellprojekt," *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, 4/2001, 530–553.

Rush, A. (2005): "Algorithm-guided treatment in depression: TMAP and STAR*D," in *Therapieresistente Depressionen - Aktueller Wissensstand und Leitlinien für die Behandlung in Klinik und Praxis*, ed. by M. Bauer, A. Berghofer, and M. Adli. Springer, Heidelberg.

Sianesi, B. (2004): "An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s," *The Review of Economics and Statistics*, 86, 133–155.

Thall, P., H. Sung, and E. Estey (2002): "Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials," *Journal of the American Statistical Association*, 97, 29–39.

Velicer, W., and J. Prochaska (1999): "An expert system intervention for smoking cessation," *Patient Education and Counseling*, 36, 119–129.

Velicer, W., J. Prochaska, J. Bellis, C. DiClemente, J. Rossi, J. Fava, and J. Steiger (1993): "An expert system intervention for smoking cessation," *Addictive Behaviors*, 18, 269–290.

Wald, A. (1950): *Statistical Decision Functions*. Wiley, New York.

Wandner, S. (2002): "Targeting employment services under the workforce investment act," in *Targeting Employment Services*, ed. by R. Eberts, C. O'Leary, and S. Wandner, pp. 1–25. W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.

Table 3.1: Descriptive statistics of jobseekers in 2001 to 2003, 460442 persons

			Means or shares (%)
Female			45
Age in years			34.9
Married			44
No. of dependents (incl. him/herself)			2.04
Not disabled			98.3
Swiss nationality			62
Foreigner with residence permit			24
Foreigner with yearly permit			11
Qualification	unskilled		25
	semiskilled		15
	skilled without degree		4
	skilled with degree		56
Education	less than 7 years		4
	8 to 11 years		20
	secondary vocational		31
	secondary academic		2
	tertiary vocational		5
	tertiary academic		4
	no information		35
First foreign language:			
German, French, Italian			59
English, Spanish, Portuguese			22
other			2
Employability rating:		(very) good	13
		medium	71
		(very) difficult	15
Insured earnings (CHF)			3940

		Means or shares (%)
Profession (only selected professions)		
Metals		7
Health care		3
Construction		4
Transportation		3
Restaurants		13
Entrepreneurs, senior officials, justice		5
Painting, technical drawing		4
Office and computer		19
Retail trade		8
Public services [a)]		4
Teaching, education		2
Unemployment history 1991-2000		
Number of unemployment spells		1.19
Average duration (months)		2.88
No unemployment in 1991-2000		55
More than once unemployed in 91-2000		28
Unemployment benefits in CHF	1991	314
	1992	960
	1999	1860
	2000	1632
Employment history 1991-2000		
Number of employment spells		2.60
Average duration (months)		39.6
No employment in 1991-2000		8
More than one employment in 91-2000		59
Earnings (CHF)	1991	24402
	1992	25025
	1999	34476
	2000	37930

[a)] Public services: security, cleaning, clerical, social work.

Table 3.2: Variables X for predictions and W for controlling selection bias

X Variables for predictions	W Variables for controlling selection bias
Month	Unemployment insurance contribution duration (months)
Duration of job search since registration	Hours worked in last employment (% of full-time equivalent)
Gender, age	Available for work, hours (% of full-time equivalent)
Marital status: Indicators for single and married	Number of the current registered unemployment spell
Number of dependent persons	Number of months until benefit exhaustion
Permit: foreigner with yearly or permanent permit	
Nationality: Indicators for Southern Europe, EU-Countries, former Yugoslavia, Eastern Europe	Earnings in last year
	Unemployment benefits in last year
Mother tongue: Indicators for German, French, Italian, Spanish, Portuguese, Albanian	Number of months with positive earnings in last year
	Number of months with unemployment benefits in last year
Number of foreign languages	Number of months without social security entry in last year
Oral and written proficiency in German, French and English: very good, good, basic, none	- these 5 variables also for the year before last year
Education: ≤7 years, 8-11 y, secondary vocational, second. academic, tertiary vocational, tertiary acad., no information	Number of employment spells in last 5 years
	Average duration of these employment spells
Qualification: Indicators for unskilled, semiskilled and skilled (with and without accepted degree)	Number of unemployment spells in last 5 years
	Average duration of these unemployment spells
Profession: Several indicators for learned, previous and preferred profession	Nonemployment ≥ 6 months in last 5 years
	Indicator of uninterrupted social security entries last 5 years
Job position: Indicators for self employed, management, craftsman, labourer	- these 6 variables also for the past 6 to 10 years
Insured earnings	
Experience in current and preferred occupation	Self-employment in last 10 years
Employability rating by case worker: very good, good, intermediate, difficult, very difficult	Indicator for continually increasing income
	Month of first entry in social security (since 1990)
Indicator for: further qualification needed	interacted with age > 35 and with mother tongue
Concordance of previous and preferred occupation	
Preferred hours of work: part/full time	Fraction of months employed since first entry in social sec.
	Mean wage in months when employed
<u>*Recent unemployment history (from AVAM)*</u>	Fraction of months unemployed since first entry in social s.
Duration of job search in last 2 years	Mean unemployment benefits in months when unemployed
Number of unemployment spells in last 2 years	
Number of sanction days in months *t*-3 to *t* -24	Year of naturalization
Number of interim jobs in months *t*-3 to *t*-24	Number of social security numbers (e.g. due to marriage)
Number of employment programmes in months *t*-3 to *t*-24	
Days in employment programmes in months *t*-3 to *t*-24	
Number of short/long courses in months *t*-3 to *t*-24	
<u>*Local labour market*</u>	
Cantonal unemployment rate	
Industry unemployment rate	
Index of vacancies in industry	

Table 4.1: Number of observations and employment outcomes in Basel city

Treatment	1	2	3	4	5	6
Number of observations	40655	2556	1319	454	183	1239
Duration job search ≤ 90 days	14371	1039	366	127	45	137
Duration job search 90-365 days	19801	1388	852	276	111	720
Duration job search > 365 days	6483	129	101	51	27	382
Outcome variables						
no employment Y=0	67 %	66 %	74 %	66 %	71 %	73 %
months of employment E[Y|Y>0]	0.55	0.52	0.43	0.51	0.49	0.53

Table 4.2: Descriptive statistics of the estimated $p^r(X,W) = P(D = r \mid X,W)$

	All obs	Observations in *D*=*r* subsample only								
	$\hat{p}^r_{jt}$	$\hat{p}^r_{jt}$	Quantiles of $\frac{Mean(\hat{p}_{jt}{}^r \mid D_{jt} = r)}{\hat{p}_{jt}{}^r}$							Number of observations
r	Mean	Mean	0.01	0.025	0.05	0.95	0.975	0.99	max	capped
1	0.88	0.88	0.90	0.91	0.92	1.13	1.17	1.22	1.51	0
2	0.055	0.096	0.36	0.41	0.48	5.26	8.67	18.44	185.9	6
3	0.028	0.084	0.32	0.36	0.41	9.52	15.11	23.77	276.8	4
4	0.01	0.02	0.23	0.29	0.37	5.64	7.51	12.40	21.30	0
5	0.004	0.007	0.20	0.27	0.35	4.40	5.26	6.34	10.53	0
6	0.027	0.049	0.30	0.37	0.43	4.42	5.93	8.75	45.96	0

Table 4.3: Estimated θ^r coefficients (selected variables only, t-values in parenthesis)

Treatment	1	2	3	4	5	6
Constant	0.60 (2.27)	-0.23 (0.27)	-0.36 (0.50)	-1.28 (0.6)	-0.21 (0.12)	-1.29 (0.88)
Job search duration	-0.56 (4.98)	0.43 (0.71)	0.28 (0.95)	0.29 (0.28)	0.55 (1.09)	0.69 (1.70)
Female	-0.17 (6.12)	-0.05 (0.67)	-0.75 (4.82)	0.41 (2.08)	0.77 (2.30)	-0.52 (4.23)
Age	-0.71 (0.59)	-0.80 (0.19)	-0.31 (0.37)	-0.55 (0.05)	-0.27 (0.14)	-0.98 (0.16)
Age squared	-1.90 (1.21)	-0.52 (0.10)	.	-0.19 (0.01)	.	-0.69 (0.09)
Married	0.14 (2.88)	-0.09 (0.67)	1.86 (8.95)	-0.89 (3.49)	.	0.31 (1.22)
Single	0.31 (6.00)	-0.08 (0.60)	1.55 (5.62)	0.10 (0.32)	-0.77 (1.15)	0.24 (0.89)
No. dependent persons	-0.04 (3.15)	0.01 (0.36)	-0.25 (4.28)	0.59 (5.11)	-0.40 (1.53)	0.03 (0.55)
Foreigner with yearly permit	-0.31 (6.45)	0.11 (0.73)	-0.10 (0.35)	-1.02 (2.85)	-0.07 (0.15)	0.41 (1.67)
Foreigner with residence permit	-0.33 (7.77)	-0.17 (1.37)	-0.97 (3.43)	.	.	-0.11 (0.63)
German proficiency: good	0.15 (4.65)	0.06 (0.60)	0.51 (3.21)	.	.	0.03 (0.24)
German proficie: good or better	0.03 (0.82)	0.17 (2.14)	.	0.75 (3.31)	.	.
No. of foreign languages	0.06 (3.69)	0.09 (2.27)	0.15 (1.70)	.	.	0.22 (3.08)
Employability: difficult	-0.58 (8.01)	-0.50 (2.38)	0.27 (1.41)	0.51 (0.98)	.	-0.35 (0.98)
Qualification: unskilled	-0.02 (0.51)	-0.15 (1.61)	-0.08 (0.54)	-0.83 (2.99)	.	0.07 (0.51)
Skilled with degree	0.14 (3.66)	-0.12 (1.16)	0.30 (1.75)	0.37 (1.87)	0.40 (1.09)	0.81 (4.92)
Insured earnings	0.02 (1.87)	0.04 (1.10)	-0.18 (3.38)	-0.09 (1.58)	0.03 (0.35)	-0.10 (2.31)
Cantonal unemployment rate	-1.98 (8.66)	-0.57 (0.92)	-1.96 (1.62)	-0.22 (0.16)	-0.38 (0.11)	-0.53 (0.46)
Index of vacancies in industry	0.19 (0.73)	0.27 (0.25)	1.16 (0.94)	.	.	-0.24 (0.13)
Industry unemployment rate	-0.24 (2.76)	-0.60 (2.35)	-0.14 (0.29)	-0.26 (0.41)	0.15 (0.16)	0.65 (1.44)
γ	0.28 (14.5)	0.07 (1.16)	1.21 (6.35)	0.99 (6.46)	0.52 (1.17)	0.20 (2.75)
α	0.03 (0.99)	0.04 (0.54)	-0.77 (2.38)	0.00 (0.01)	0.13 (0.21)	-0.07 (0.50)
X_2 : Job search duration	0.00 (0.61)	-0.04 (1.66)	-0.23 (3.88)	0.21 (3.08)	-0.21 (2.92)	-0.05 (2.12)
X_2 : Gender	0.22 (10.6)	0.07 (1.33)	0.59 (3.04)	-0.08 (0.36)	0.3 (0.90)	0.32 (2.67)

Note: Only selected variables shown. Full table available from the author. A dot · marks variables not included in the estimations.

Table 4.4: Correlations between the predicted outcomes for the 46406 observations

Treatment	1	2	3	4	5	6
1	1.000	0.594	0.401	0.168	0.313	0.534
2	0.594	1.000	0.313	0.047	0.131	0.410
3	0.401	0.313	1.000	0.041	-0.057	0.239
4	0.168	0.047	0.041	1.000	-0.065	0.171
5	0.313	0.131	-0.057	-0.065	1.000	0.075
6	0.534	0.410	0.239	0.171	0.075	1.000

Table 4.5: Prediction error in the estimation sample

	1	2	3	4	5	6
Prediction error						
Mean squared error	0.085	0.083	0.057	0.084	0.069	0.070
Median squared error	0.030	0.030	0.002	0.009	0.010	0.017
Mean absolute error	0.230	0.227	0.136	0.185	0.176	0.191
Median absolute error	0.173	0.173	0.041	0.094	0.099	0.131

Table 4.6: Descriptive statistics of the predictions for the 2303 jobseekers

Treatment	1	2	3	4	5	6
Predictions (months)						
Mean	2.69	2.26	2.02	2.33	2.45	2.46
Stddeviation	0.92	0.96	1.10	1.52	1.46	1.05
Q_{25}	1.97	1.51	1.01	1.12	1.28	1.64
Median	2.52	2.10	1.69	1.71	1.99	2.23
Q_{75}	3.27	2.86	2.93	3.13	3.27	3.07
Treatment allocation according to predictions, when choosing treatment ...						
with highest prediction	20.2 %	11.5 %	10.7 %	16.3 %	25.4 %	19.6 %
within set S_i of best treatments	18.3 %	12.1 %	9.7 %	17.2 %	24.7 %	18.0 %

Cardinality of sets of best, intermediate and worst treatments

Number of obs with cardinality	0	1	2	3	4	5	6
Best treatments (S_i)		781	683	483	280	65	11
Intermediate treatments	1208	643	315	120	17	0	0
Worst treatments	188	307	435	347	338	688	0