Gaure, Simen; Røed, Knut; van den Berg, Gerard J.; Zhang, Tao

Working Paper

Estimation of heterogeneous treatment effects on hazard rates

IZA Discussion Papers, No. 4794

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Gaure, Simen; Røed, Knut; van den Berg, Gerard J.; Zhang, Tao (2010) : Estimation of heterogeneous treatment effects on hazard rates, IZA Discussion Papers, No. 4794, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/35857

DISCUSSION PAPER SERIES

IZA DP No. 4794

Estimation of Heterogeneous Treatment Effects on Hazard Rates

Simen Gaure
Knut Røed
Gerard J. van den Berg
Tao Zhang

February 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Estimation of Heterogeneous Treatment Effects on Hazard Rates

Simen Gaure
Ragnar Frisch Centre for Economic Research

Knut Røed
Ragnar Frisch Centre for Economic Research and IZA

Gerard J. van den Berg
University of Mannheim, IFAU-Uppsala and IZA

Tao Zhang
Ragnar Frisch Centre for Economic Research

Discussion Paper No. 4794
February 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4794
February 2010

ABSTRACT

Estimation of Heterogeneous Treatment Effects on Hazard Rates*

Consider a setting where a treatment that starts at some point during a spell (e.g. in unemployment) may impact on the hazard rate of the spell duration, and where the impact may be heterogeneous across subjects. We provide Monte Carlo evidence on the feasibility of estimating the distribution of treatment effects from duration data with selectivity, by means of a nonparametric maximum likelihood estimator with unrestricted numbers of mass points for the heterogeneity distribution. We find that specifying the treatment effect as homogenous may yield misleading average results if the true effects are heterogeneous, even when the sorting into treatment is appropriately accounted for. Specifying the treatment effect as a random coefficient allows for precise estimation of informative average treatment effects including the program's overall impact on the mean duration.

JEL Classification: C31, C41, J64, C63

Keywords: duration analysis, unobserved heterogeneity, program evaluation, nonparametric estimation, Monte Carlo simulation, timing of events, random effects

Corresponding author:

Knut Røed
The Ragnar Frisch Centre for Economic Research
Gaustadalléen 21
0349 Oslo
Norway
E-mail: knut.roed@frisch.uio.no

* This research has received financial support from the Norwegian Research Council through the research project "Labor market exclusion".

1. Introduction

With the increasing availability of administrative data sources in many countries, it may be foreseen that in the future, non-experimental evaluations of social programs can be performed at a large scale to a very low cost. However, the fruitfulness of this approach depends on the extent to which methodological difficulties associated with non-experimental analyses can be satisfactorily solved. In particular, non-experimental evaluations of social programs must cope with self-selection and unobserved heterogeneity. Valid instruments are often not available, since variables that affect the outcome of interest and can be observed by a researcher, may also have affected the selection process into the program. Direct comparisons of pre and post program outcomes, e.g., by means of fixed effects estimators, are typically not appropriate, since participation in a social program often results directly from lack of success during the pre program period. Two empirical strategies dominate the literature. The first is to use observed characteristics to establish the best possible control groups (through, e.g., propensity score matching), and hope that any remaining uncontrolled heterogeneity in the participation process is orthogonal to unobserved heterogeneity in the outcome of interest. The second strategy is to model unobserved characteristic as random effects (mixture distributions) and use the exact timing of events (and not only their occurrence) to disentangle causality from sorting. It has been shown that within a mixed proportional hazard (MPH) framework, the latter approach does not require access to instrumental variables.

Modeling unobserved heterogeneity is of course only meaningful if it is nonparametrically identified (unless prior knowledge about the functional form of its distribution is available). Substantial progress has been made regarding our understanding of the identification problem. An important contribution to this literature is the so-called “timing-of-events approach”, which provides identification results for hazard rate models with endogenous treatments. A substantial part of this literature assumes *homogenous*

treatment effects, i.e., a situation where unobserved "intercepts" in the treatment and final-destination hazards induce a spurious correlation between treatment propensity and outcome measures, but where the effect of interest is the same for all subjects, conditioned on observed covariates. However, for most social programs, the assumption of a homogenous effect is not convincing. Even among subjects that are equal according to observed characteristics, we typically expect treatment effects to vary. Abbring and Van den Berg (2003) and Richardson and Van den Berg (2008) show that mixed proportional hazard rate models with selective durations until treatment and *heterogeneous* treatment effects are also nonparametrically identified under sets of regularity assumptions. Richardson and Van den Berg (2008) estimate such models, where the unobserved heterogeneity distribution is a multivariate discrete distribution with a fixed number of points of support.[1] Discrete-time duration analyses often adopt joint normality and/or factor loading assumptions concerning the heterogeneity distribution; see e.g. Carneiro *et al*. (2003) and Aakvik *et al.* (2005).

If the researcher is primarily interested in the average treatment effect (or the average treatment effect among the treated), a homogeneity assumption might be justified on the ground that it traces out the average effect of interest. However, in non-linear settings, the mean effect is not necessarily equal to the effect on the average individual. Little is known regarding the appropriate interpretation of estimated homogenous effects in cases where the true effects are heterogeneous. In this paper we show that we cannot expect it to capture an average treatment effect for the population of potential participants as a whole (ATE) or an average treatment effect among the treated (ATET). Moreover, to the extent that the unobserved participation propensity is correlated to the treatment effect, the imposition of a homogenous effect may yield a significant bias in simulation-based program effect statistics.

[1] For matching estimation with duration outcomes and effect heterogeneity, see Fredriksson and Johansson (2008) and Crépon et al. (2009).

The purpose of the present paper is to evaluate the scope for inference on heterogeneous treatment effects within the "timing-of-events" framework by means of the nonparametric maximum-likelihood estimator (NPMLE). We set up Monte Carlo experiments aimed at shedding light on the extent to which key summary statistics and distributional parameters can be uncovered from observed data.[2] There are three latent variables in all our data-generating processes (DGPs): the intercept in the treatment hazard, the intercept in the final-destination hazard, and the proportional treatment effect (the shift in the final-destination hazard resulting from treatment). These three variables follow a joint distribution that is assumed unknown to the researcher. A key finding of our paper is that a number of relevant treatment effect statistics, including ATE, ATET, and a host of simulation-based program effects, can be reliably uncovered from the data by means of the *full-dimensional* NPMLE, i.e., without restrictions on the joint distribution of the three unobserved determinants. We also find that a two-dimensional factor loading model performs well. In terms of robustness, the two-dimensional factor loading model even appears to be superior to the full-dimensional model. However, a one-dimensional factor loading model performs poorly in our experiments. Imposing perfect correlation between latent variables is therefore not advisable, unless this restriction is justified by prior knowledge.

Unfortunately, it turns out to be difficult to evaluate the sampling distribution of the treatment effect statistics that we estimate in this paper. We have not been able to compute reliable standard errors for either ATE, ATET, or for the simulation-based program effects, except by means of nonparametric or semiparametric bootstrap. We also find that the sampling distributions of interest display significant deviations from normality. This problem is related to the non-concavity of the likelihood function, which implies that we in some of

[2] As such, the paper builds on the literature in which inference of duration models with unobserved heterogeneity is assessed using Monte Carlo simulations; see Baker and Melino (2000), Gaure, Røed and Zhang (2007) and Van den Berg, Caliendo and Uhlendorff (2009).

the trials either fail to identify the global optimum or that sampling error causes the "wrong" optimum to represent the global maximum. A typical finding from our nonparametric bootstrap exercises is that more than 95 percent of the trials end up yielding a set of normally distributed treatment effect statistics, while the rest of the trials produce completely different results. We conclude from this experience that nonparametric (or semiparametric) bootstrap should be part of a standard estimation procedure, not only to evaluate statistical uncertainty, but also to ensure that the original result (based on the full sample) does not belong to the group of outliers. Statistical inference cannot be made without discretionary judgment regarding identification and handling of outlier results.

The next section describes the modeling framework and provides definitions for the key treatment and program effect statistics. Section 3 describes the data generating process and Section 4 outlines our estimation strategy. Section 5 presents the results and Section 6 concludes.

2. The Modeling Framework and Treatment Effect Measures

The models we examine in this paper portray a subject entering into an origin state, and describe its subsequent transition intensity into a destination state. During occupation of the origin state, a treatment may occur that affects the final destination transition intensity. For simplicity, it is assumed that treatment only occurs once, i.e., realization of a treatment removes the subject from the risk of subsequent treatment. There are unobserved jointly distributed covariates that describe the subjects' two transition propensities and their treatment effects.

Our starting point is a simple continuous-time multivariate mixed proportional hazard-rate model (MMPH). The two events that can occur are transitions to the final-destination state (e) and the treatment state (p). While the former transition terminates the spell, the latter does not. The event of a treatment may, however, cause a change in the final-destination

hazard. Let i be the index for subjects ($i=1,2,\ldots,N$) and let d index spell duration ($d \in R_+$). In its simplest form, the model is described in terms of two hazard rates:

$$\begin{aligned} \theta_{eid} &= \exp\left(\beta_e x_i + z_{id}\alpha_i + v_{ei}\right), \\ \theta_{pid} &= (1-z_i)\exp\left(\beta_p x_i + v_{pi}\right), \end{aligned} \tag{1}$$

where z_{id} is the treatment indicator, i.e., $z_{id} = 1$ if treatment has been implemented (and zero otherwise). The vector x_i contains observed covariates. To avoid inessential complications, we abstract from duration dependence and time-varying covariates. Gaure *et al.* (2007) show that duration dependencies can be robustly uncovered from observed duration data by means of the MMPH model, and that time-varying exogenous covariates significantly improve the foundation for nonparametric identification. The triple $(v_{ei}, v_{pi}, \alpha_i)$ constitutes the three unobserved subject-specific characteristics in terms of the final-destination hazard propensity, the treatment hazard propensity, and the treatment effect, respectively. We assume that the unobserved covariates and treatment effects are time-invariant and independent of observed characteristics; hence, (1) may be interpreted as a random coefficients model.

An important distinction made in this paper is that between a "treatment effect" and a "program effect". A treatment effect (TE) is the actual or hypothetical effect of being subject to treatment. A program effect (PE) is the expected impact of a given program structure before the actual timing of treatment is revealed. In the literature, TE and PE are often referred to as the ex post and ex ante effects. (Here, we use "ex ante" to refer to the situation before treatment in general.) While each subject's treatment effect is conditional on participation (although it can be estimated for non-participants as well) and hence independent of the statistical process determining participation, the subjects' program effects clearly depend on the selection process and the overall intensity of treatment.

A natural measure of a time-invariant subject-specific treatment effect in this model is the proportional change in the final-destination hazard caused by treatment; i.e.

$$TE_i = \exp(\alpha_i). \tag{2}$$

The average treatment effect (ATE) is equal to

$$ATE = \underset{i \in N}{E}[TE_i]. \tag{3}$$

ATE as defined in (3) is well-defined provided that there are no defective risks, either in the participation or in the final outcome transition processes.[3] Clearly, a defective participation risk $(\Pr(v_{pi} = -\infty) > 0)$ makes it impossible to identify the treatment effect in the corresponding location vector, since any treatment effect α_i in the location vector $(v_{el}, -\infty, \alpha_l)$ fits the data likelihood equally well. A defective ex ante (before treatment) outcome hazard $(\Pr(v_{ei} = -\infty) > 0)$ leads to similar problems, since any finite treatment effect ($\alpha_i < \infty$) in the location vector $(-\infty, v_{pl}, \alpha_l)$ fits the data likelihood equally well. A particular problem may arise if the ex ante hazard in a location vector is zero while the ex post (after treatment) hazard is positive, in which case $\alpha_i = \infty$ in that location vector. If positive probability is attributed to $\alpha_i = \infty$, we obtain that $ATE = \infty$. In principle, these problems can be circumvented by restricting hazards to be non-defective. However, it is not obvious how this should be done in practice. Moreover, with finite datasets, the issue of "defective risks" is more a matter of degree than of kind. As the hazards in question approach zero, it becomes more and more difficult to identify the associated treatment effect with any precision. A related problem is that ATE attributes the same weight to all proportional treatment effects, regardless of the baseline hazard to which they are multiplied. It may be argued that a big proportional impact on an almost defective hazard rate is of little interest from a policy point of view, particularly if there are competing risks or censoring processes implying that the event in question is almost certain not to take place anyway.

---

[3] Abbring and Van den Berg (2005) and Van den Berg, Bozio and Costa Dias (2010) develop a range of measures for average treatment effects on duration outcomes in a non-parametric potential-outcome framework.

ATE also ignores variation in treatment propensity; hence, the effect for a subject with high treatment propensity is attributed the same weight as the effect for a subject with low (or even defective) treatment propensity. But a similar effect measure can in principle be provided for the population of actually treated subjects, thereby providing the average treatment effect on the treated (*ATET*). Let N_Δ be the set of actually treated subjects. We then have that

$$ATET = \underset{i \in N_\Delta}{E}\left[TE_i\right]. \tag{4}$$

Estimation of ATET requires, however, that the members of the treatment group are identified and equipped with the appropriate conditional joint distribution of α_i. This can be achieved by means of simulation. We return to this that later on.

While the subject-specific treatment effects are naturally evaluated in terms of the proportional shift in the hazard rate (or in remaining expected duration) caused by the treatment, the program effects are most naturally evaluated in terms of the program's overall effects on the ex ante expected durations. The latter depend on the distribution of final destination hazards, treatment hazards, and treatment effects, and hence on the joint distribution of $(x_{id}, v_{ei}, v_{pi}, \alpha_i)$. Let $D_i^0 = E\left[D_i \mid v_{pi} = -\infty\right]$ be the expected length of subject *i's* spell if treatment never occurs and let $D_i^\Delta = E\left[D_i\right]$ be the expected length of such a spell given the true enrolment process and the true effect of treatment. The program effect on the expected duration for subject *i* is then equal to

$$PE_i = D_i^0 - D_i^\Delta. \tag{5}$$

In order to evaluate average program effects, we take the mean of individual effects and divide by the scale of the program in terms of the overall frequency of treatments. Hence, we define the *average program effects on absolute duration* as the mean duration effect per treatment:

$$APE^{AD} = \frac{\underset{i \in N}{E}\left[PE_i\right]}{P}, \tag{6}$$

where P is the fraction of treated subjects. In some settings, it seems natural to evaluate the average program effect relative to mean duration without the program effect; i.e.

$$APE^{RD} = \frac{APE^{AD}}{\underset{i \in N}{E}\left[D_i^0\right]}. \tag{7}$$

The *selection process* into treatment needs to be identified both for the purpose of disentangling the causal treatment effects from unobserved sorting and for the purpose of estimating program effects. But characterization of the selection process may also be of interest in its own right, e.g., in order to assess the extent to which program slots are allocated to those who need them most and/or to those with the largest treatment effects. The selection process is thus most naturally evaluated in terms of its relationship to the final-destination hazard and in relation to the treatment effects. From a policy perspective, it is typically the actual features of the selection process that matter, and not the extent to which it can be decomposed into factors that are observed or unobserved by the researcher. Hence, we focus on selection measures that incorporate both observed and unobserved determinants. We examine s*election on the final-destination hazard* (*SFH*) and *selection on the treatment effect (STE)*. SFH is examined in terms of the statistical association between $\left(\exp(\beta_e x_i + v_{ei}\right)$ and $\left(\exp(\beta_p x_i + v_{pi}\right)$, while STE is examined in terms of the association between $\left(\exp(\alpha_i\right)$ and $\left(\exp(\beta_p x_i + v_{pi}\right)$; conf. Equation (1).[4] As summary statistics for SFH and STE, we compute correlation and concordance coefficients. While correlation coefficients may have the most convenient interpretation, they are highly sensitive towards extreme values in the estimated

[4] Note that in the more general case with duration dependence and/or time-varying covariates, these associations depend on time/duration, even when the subject-specific effects are assumed time-invariant. This can be handled by standardizing on a particular spell duration or by integrating the hazards over time/duration.

heterogeneity distribution, which (as discussed above) may be determined on a weak empirical basis.

3. Data Generation

The strategy we pursue in this paper is that we create a large number of artificial subjects, in terms of $(x_i, v_{ie}, v_{ip}, \alpha_i)$ on the basis of various subject generating processes (SGP's). Each SGP is characterized by a particular joint distribution of the unobserved characteristics $(v_{ei}, v_{pi}, \alpha_i)$. After the subjects have been constructed, they participate in an event history lottery, where treatment times and durations are drawn randomly on the basis of specified hazard-rates. The lottery is based on repeated calculations of the two pseudo survival functions. From (1), we have that

$$\begin{aligned} S_{ei}(d) &= \exp\left(-\int_0^d \exp\left(\beta_e x_i + v_{ei}\right) du \right) = \exp\left(-d \exp\left(\beta_e x_i + v_{ei}\right)\right), \\ S_{pi}(d) &= \exp\left(-\int_0^d \exp\left(\beta_p x_i + v_{pi}\right) du \right) = \exp\left(-d \exp\left(\beta_p x_i + v_{pi}\right)\right). \end{aligned} \tag{8}$$

To generate durations and treatment times, we draw the survival probabilities and invert the two pseudo survival functions; i.e., we replace the left hand side of (8) with [*1-u*$_k$], where u_k (*k=e,p*) are random drawings from a uniform [*0,1*] distribution, and solve for the resultant latent durations; see Crépon *et al.* (2005, p. 19). If the duration until treatment is shorter than the duration until exit, a treatment occurs, and a new duration (for the remaining spell) is generated with $z_{id} = 1$. This lottery is what we refer to as the observation generating process (OGP), and it creates the datasets used for estimation purposes. Together SGP and OGP constitute the data generating process (DGP).

In order to construct datasets for analysis, we specify a *baseline DGP* which will be used to examine the key properties of the NPMLE. The baseline DGP consists of 50,000 subjects. The researcher observes three exogenous time-invariant covariates (x_1, x_2, x_3). All

three are subject to independent normal distributions with means and variances equal to (0,0.25), (0,0.25), (1,1), respectively, and with causal coefficients $\beta_{1e}=1,\ \beta_{1p}=1,\ \beta_{2e}=1,$ $\beta_{2p}=-1, \beta_{3e}=0.5,\ \beta_{3p}=1$. This satisfies the condition for model identification spelled out in Abbring and Van den Berg (2003, p. 1505). The two intercepts (v_e, v_p) are also subject to normal distributions with variances equal to 1. We assume that they are negatively correlated with $\left(corr(v_{ei}, v_{pi})=-0.5\right)$ such that there is a negative selection to treatment on unobserved covariates. For the treatment effect, we deliberately construct a non-standard and intricate distribution. The treatment parameter α is drawn from three alternative normal distributions with mean 0.0, 0.2, and 0.6, respectively. Which of the three distributions a subject draws from is determined by its position in the v_p-distribution; the higher the v_p, the higher is the expected α.[5] Hence, we have introduced a positive selection on the treatment effect. The resultant average treatment effects are $ATE=1.36<E[ATET]=1.55$ (recall that while *ATE* is a parameter, *ATET* is a stochastic variable). The distribution of treatment effects in the baseline DGP is illustrated in Figure 1 (the distribution of treatment effects among the treated is obtained by repeating the OGP 120 times). The means of (v_e, v_p) are scaled such that the mean expected duration until final exit (in the absence of treatment) is approximately equal to 13, and such that the mean treatment probability is equal to 0.37. Note that since both the treatment effect and the final destination propensity are correlated to the treatment propensity, there is also a (negative) correlation between the treatment effect and the final destination hazard in the DGP.

- Figure 1 around here -

[5] Subjects belonging to percentiles [0,33] in the v_p-distribution draw α from the N(0.0,0.0025) distribution (mean, standard deviation), subjects belonging to percentiles (33,66] draw from the N(0.2,0.0025) distribution and the rest draw from the N(0.6,0.0225) distribution.

4. Estimation

We put ourselves in the position of a researcher who has access to data with accurately measured spell durations, treatment times and observed covariates, but no information about the distribution of unobserved covariates and treatment effects. The researcher's data window is also assumed limited, such that spells lasting longer than 100 periods are right-censored. Based on this restricted information set, our researcher's aim is to uncover reliable information and make statistical inferences regarding the true treatment and program effects and the sorting into treatment.

Since the distribution of unobserved heterogeneity is assumed completely unknown, it is modeled nonparametrically with the aid of a discrete distribution (Lindsay, 1983; Heckman and Singer, 1984); i.e., by means of nonparametric maximum likelihood estimators (NPMLE). In this section, we first briefly explain how we employ this method in the most general case of a completely unrestricted three-dimensional vector of unobserved heterogeneity. We then discuss an alternative and more restrictive – but potentially also more robust – modeling strategy based on reduced heterogeneity dimensionality.

4.1 The full-dimensional nonparametric maximum likelihood estimator

Let d_i be individual *i's* observed spell-duration, and s_i the realized duration of treatment (if it occurred). For a non-treated subject, the contribution to the likelihood function (conditional on unobserved characteristics) is

$$L_i(v_e, v_p, \alpha) \mid d_i < s_i = \exp\Big(-d_i\left(\exp\left(\beta_e x_i + v_{ei}\right) + \exp\left(\beta_p x_i + v_{pi}\right)\right)\Big)\exp\left(\beta_e x_i + v_{ei}\right), \tag{9}$$

and for a treated subject, the contribution is

$$\begin{aligned} L_i(v_e, v_p, \alpha) \mid d_i > s_i = \exp\Big(-s_i\left(\exp\left(\beta_e x_i + v_{ei}\right) + \exp\left(\beta_p x_i + v_{pi}\right)\right)\Big)\exp\left(\beta_p x_i + v_{ei}\right) \\ \times\exp\Big(-(d_i - s_i)\left(\exp\left(\beta_e x_i + \alpha_i + v_{ei}\right)\right)\Big)\exp\left(\beta_e x_i + \alpha_i + v_{ei}\right) \end{aligned} \tag{10}$$

Let W be the (a priori unknown) number of support points in this distribution and let $\{(v_{el}, v_{pl}, \alpha_l), p_l\}$, $l = 1, 2, ...W$, be the associated location vectors and probabilities. In terms of observed variables (data), the likelihood function is then given as

$$L = \prod_{i=1}^{N} E[L_i(v_{el}, v_{pl}, \alpha_l)] = \prod_{i=1}^{N} \sum_{l=1}^{W} p_l L_i(v_{el}, v_{pl}, \alpha_l), \quad \sum_{l=1}^{W} p_l = 1. \tag{11}$$

Our estimation procedure is to maximize this function with respect to all the model and heterogeneity parameters repeatedly for alternative values of *W*; see Gaure *et al.* (2007) for details. The maximization is unconstrained, in the sense that we do not restrict the parameter space for the unobserved covariates to be consistent with non-defective risks. When we interpret the estimation results, however, we do take into account that some parameters of interest cannot be identified on the basis of heterogeneity vectors containing defective (or close to defective) risks; see Section 2. To determine the "optimal" number of support points, we start out with *W=1*, and then expand the model with new support points until the model is "saturated", in the sense that we are not able to increase the likelihood any further. We then chose a preferred model (the number of support points) on the basis of the Akaike information criterion (AIC). This choice is motivated both by theoretical considerations and empirical evidence. In particular, we have seen from Gaure *et al.* (2007) that AIC performs well in our finite mixture models. Following Burnham and Anderson (2002, Section 6.9.6) we compute AIC with the actual number of parameters estimated. E.g., with W points of support, only W-1 parameters are estimated for the probabilities in the mixture.

Standard errors attached to non-random coefficients (β_e, β_p) can be calculated directly from the inverted Fisher matrix. Gaure *et al*. (2007) show that for parameters attached to observed covariates, the inverted Fisher matrix from the optimally selected model provides standard errors that can be used for standard statistical inference. By means of the delta-method, a similar procedure can be devised for treatment effect statistics, insofar as they can

be expressed directly as a function of estimated parameters. Hence, it is possible to compute standard errors for ATE, but not for the other – simulation based – treatment and program effect statistics. However, since little is known about the distribution of the parameters characterizing the treatment effect distribution, we have little a priori knowledge about the performance of these standard errors and their applicability for statistical inference. In particular, we have no reason to expect standard inference procedures to be valid.

4.2 Reduced heterogeneity dimensionality

Estimation of a full-dimensional model typically requires large computational resources. In some applications it is also questionable whether the data-based foundation for nonparametric identification is sufficiently strong for a truly nonparametric model to yield robust results. These considerations may motivate the researcher to reduce the dimensionality of the accounted for unobserved heterogeneity. There are two ways of implementing this idea. The first is to assume that at least one of the heterogeneous parameters is really homogenous. In the treatment evaluation literature, for example, it is common practice to specify the treatment effect as homogenous (fixed), conditioned on observed covariates; see, e.g. Abbring and Van den Berg (2004), Van den Berg *et al.* (2004), Røed and Raaum (2006), and Rosholm and Svarer (2008). The second is to use factor loading, i.e., to assume that the full vector of unobservables depend (linearly) on a lower number of generic unobservables; see, e.g., Carneiro *et al*. (2003) and Aakvik *et al.* (2005). To illustrate, assume that the researcher specifies unobserved heterogeneity in terms of two generic unobserved covariates, and that these two are the final destination and treatment propensities (v_e, v_p), respectively. The two-dimensional linear factor loading model is then specified by assuming that the treatment effect (α) is a linear function of (v_e, v_p), i.e.,

$$\alpha_i = \alpha_0 + \alpha_e v_{ei} + \alpha_p v_{pi}, \tag{12}$$

where $(\alpha_0, \alpha_e, \alpha_p)$ are parameters to be estimated. Alternatively, the researcher could express the two generic unobservables as being outside the domain of particular hazard rates and specify all three intercepts (v_e, v_p, α) as distinct linear functions of them. However, given required normalizations, it turns out that this model would be equivalent to the one described here; see Appendix for proof. It follows that it is also immaterial which of the three unobserved variables that is selected for factor loading.

With some modifications, the NPMLE estimation procedure is the same for the factor loading model as for the full-dimensional model.

5. Uncovering the Baseline Model from observed data

The purpose of this section is to assess the ability of NPMLE to uncover key properties of the true model. We use the baseline DGP described in Section 3 to generate 120 distinct datasets, $S_1, ..., S_{120}$, from a common SGP with 50,000 subjects. Each of the 120 datasets is subject to the estimation procedures set out in Section 4. And each estimated model is then used in a simulation exercise to compute summary statistics like *ATET* and *APE*, and selection statistics like *SFH* and *STE*. We present the results in three steps. First, Section 5.1 describes the alternative models' ability to uncover correct summary statistics in terms of the various treatment effects (ATE, ATET, APE) and sorting parameters (SFH, STE). Section 5.2 then discusses the scope for valid statistical inference. Finally, Section 5.3 assesses the models' performance in terms of uncovering the underlying distribution functions of the treatment effects, i.e., $F(TE_i)$.

5.1 Point estimates for summary statistics

Table 1 summarizes our key findings. Note first that if the researcher simply disregards unobserved heterogeneity (Column I), all the summary statistics are estimated with huge

biases. The researcher is led to erroneously conclude that the treatment in question significantly reduces the final destination hazard (reflecting the negative selection on unobservables in the DGP). Introducing one-dimensional heterogeneity does not alleviate the problem very much, regardless of whether factor loading is used to account for heterogeneous treatment effects (Column III) or not (Column II). The researcher still gets the signs of both treatment and program effects completely wrong. A one-dimensional model simply lacks the flexibility required to account for the two sorting processes (on the final destination hazard and on the treatment effect) taking place simultaneously in the baseline DGP.

- Table 1 around here -

Introducing two-dimensional heterogeneity, however, may improve the model's performance substantially. A popular way of doing this (see references in the previous section) is to specify the treatment effect as homogenous, while estimating the joint distribution of v_e, v_p nonparametrically (or with a fixed number of support points). This procedure is designed to eliminate bias arising from sorting on the final destination hazard, but – by construction – it cannot eliminate any bias arising from sorting on the treatment effect. The results presented in Table 1, Column IV, indicate that estimates of the assumed homogenous treatment effect tend to resemble the true ATE. Average program effects, however, are substantially underrated (along with ATET), since the model disregards the selection on the treatment effect (which happens to be positive in the baseline DGP). Hence, it is tempting to conclude that the common practice within the timing-of-events literature of specifying the treatment effect as homogeneous (conditioned on observed covariates) is defensible insofar as the researcher is only interested in the average treatment effect (ATE). However, it turns out that the ATE-interpretation of the estimated treatment effect is not robust. This is illustrated in Table 2, where we compare the estimated homogeneous treatment effects with the true ATEs and ATETs under different assumptions regarding the sorting

processes. It is clear that the resemblance between the estimated homogenous effect and the true ATE in our baseline model occurred by coincidence. Insofar as there is systematic sorting into the program on the treatment effect ($corr(v_p, \alpha) \neq 0$), the estimated homogenous effect generally deviates from both ATE and ATET. The size of the deviations depends on the relative importance of the various unaccounted for sorting processes. To the extent that treatment effects are correlated with the treatment and final destination propensities in the DGP, this is partly picked up by the nonparametrically estimated (v_e, v_p) distribution, effectively confounding the effects of the treatment. With positive selection on the treatment effect, the homogenous estimator tends to lie below both ATE and ATET. With negative selection, it tends to lie above ATET.

- Table 2 around here -

Introducing linear factor loading in the two-dimensional model improves the performance significantly; see Table 1, Column V. All the summary statistics are then estimated without noticeable bias, including ATE, ATET and the program effects. Even the two sorting statistics are relatively close to their true values. The full-dimensional NPMLE also produce very reliable results; see Table 1, Column VI. But there are no evident gains in precision compared to the more parsimonious two-dimensional factor loading model. Moreover, the standard deviations for the various estimates (taken across the 120 trials) tend to be somewhat larger for the full-dimensional than for the two-dimensional model, suggesting that the reduced dimensionality yields some gains in robustness.

Figures 2 and 3 show the distribution of estimation errors for the six key statistics ATE, ATET, APE^{AD}, APE^{RD}, SFH, and STE (the latter two based on Kendall's τ), for the two-dimensional factor loading model and the full-dimensional model, respectively. It is evident that most of the estimated statistics are heavily concentrated around their true values, and apart from a few "outlier" results, most of the statistics seem to be close to normally

distributed. This suggests that the models may be applied for statistical inference. The existence of outliers is worrying, however, and indicates that robustness needs to be assessed.

- Figure 2 around here -

- Figure 3 around here –

5.2 Robustness and statistical inference

In order for a researcher to make statistical inference on the basis of estimation on a single dataset, standard errors and/or confidence intervals are required. As discussed in Section 4, it is in principle possible to compute standard errors for ATE (since ATE is function of the estimated parameters and no simulation is required) by means of the Delta-method. As it turns out, however, these standard errors are not sufficiently reliable for either the full-dimensional or the two-dimensional models. In most of the trials (around 60-80 percent), the estimated standard errors do not deviate more than ± 0.2 from the observed standard deviation across all trials. But in many of the remaining trials, the deviation is extremely large (and in some very few cases, the estimated standard error approaches infinity). These difficulties are actually not very surprising, given ATE's sensitivity towards extreme values. And they suggest that statistical inference must be based on bootstrap techniques. This is obviously also the case for the other, simulation based, summary statistics.

We examine the impact of the sampling distribution by means of nonparametric bootstrap; i.e., we draw artificial samples with replacement from the observed data and re-estimate the model several times, each time followed by a new simulation. Given the computational costs involved, we have chosen to do the nonparametric bootstrap on ten randomly selected datasets only (implying 120×10=1200 estimation and simulation trials). To illustrate our findings, we present in Table 3 the key results generated from one of these 10 trials (still randomly selected), focusing on the two most promising estimation strategies; i.e., the two-dimensional linear factor loading model and the full-dimensional model. As it turned

out, the randomly selected dataset had generated estimated treatment and program effect statistics somewhat below their true values. These errors were maintained through the bootstrap trials, indicating that they resulted from sampling error in the original full sample. For the two-dimensional factor loading model, the bootstrap standard deviations tended to resemble the standard deviations across the original 120 datasets, suggesting that the bootstrap procedure does produce valid standard errors. But again, outlier results in some cases drive the standard errors completely off target. For the full-dimensional model, we simply had to "remove" some extreme outliers before sensible summary statistics could be calculated for the treatment effects. Moreover, even for the two-dimensional factor loading model, some bootstrap samples ended up yielding results that were completely off target. To illustrate, Figure 4 shows the complete results for the estimated ATE from the 10×120 bootstrap samples. While we had no apparent outliers in the bootstrap based on Dataset 1 (which forms the basis for Table 3), it is clear that such outliers do appear in some of the other bootstrap trials (in particular, datasets 2, 5, and 8). This obviously gives rise to a non-normal distribution of estimates, complicating statistical inference.

- Table 3 around here -

- Figure 4 around here -

We assume that various outlier detection techniques may be applied to eliminate atypical results, such that statistical inference can be based on the remaining estimates. However, we cannot always expect outliers to be as easily detected as they apparently are in Figure 4. Hence, it is probably difficult to eliminate the need for subjective judgment as a basis for statistical inference. The existence of outliers probably arises from the fact that the likelihood function to be maximized is not globally concave, and that no algorithm known to the authors can ensure, in reasonable time, that a global optimum has really been found. It may therefore be important (in actual applications) to ensure that the vector of estimates based

on the full sample does not by accident belong to a group of outliers. This can be done by means of the nonparametric bootstrap. Hence, the nonparametric bootstrap serves a dual purpose here; first, to ensure that the original estimates are not atypical from bootstrap-based estimates; and second, to facilitate statistical inference.

Given the large computational costs often associated with just a single estimation of NMPLE, the nonparametric bootstrap will for some applications be prohibitively expensive and time-consuming. As an alternative, we also explore the properties of a semiparametric bootstrap technique based on the following procedure: The model is estimated only once, namely on the complete (original) dataset. On the basis of this estimation, repeated drawings are made from the vector of parameters attached to observed characteristics (β_e, β_p), for which the joint normality assumption is likely to hold; see Gaure *et al.*, 2007. For each drawing of these parameters, the heterogeneity distribution is then re-estimated by means of conditional nonparametric likelihood maximization (AIC). To save computational resources, this latter step may be conditioned on the number of support points in the heterogeneity distribution. Finally, the resultant parameter sets are used for repeated simulations to compute the statistics of interest.

It turns out that the semiparametric bootstrap performs well, even when the re-estimation of the heterogeneity distribution is performed conditional on the number of support points. The empirical standard deviations across summary statistics computed from the 120 draws/re-estimations are close to the "true" standard deviations reported in Table 1. The risk of obtaining outliers also seems to be reduced compared to the nonparametric bootstrap. Hence, the semiparametric bootstrap may provide a foundation for assessment of sampling variability.

5.3 Distribution functions for treatment and program effects

We now turn to the issue of uncovering the *distribution* of individual treatment effects TE_i. The NPMLE can obviously not provide a correct distribution of treatment effects since it gives a discrete representation (typically with around 8-12 support points) of the underlying continuous distribution. However, in order to check whether the underlying distributions are really identified in practice, we can again use the bootstrap; i.e., we can collect all the 120 estimated bootstrap distributions generated from a single dataset into a single distribution, and compare it to the true one (DGP).

- Figure 5 around here -

Figure 5 provides some illustrative results based on the two-dimensional factor loading model (the full-dimensional model produces similar results) and the nonparametric bootstrap. The upper panels plot the distribution functions of TE_i for all subjects (panel a) and for treated subjects (panel b) based on a single estimation on the original Dataset 1. The lower panels plot the distribution functions that arise when the results from all the 120 bootstrap trials based on Dataset 1 are merged into one single distribution. For comparison, all panels also plot the true distributions. Since we already know that average treatment effects were somewhat underestimated on Dataset 1, it is no surprise that the estimated distributions do not match the true distributions perfectly. A more interesting finding is that the estimated distributions seem to have a larger probability mass in the central area of the distribution than the true distribution. The estimated distributions are also smoother than the true distributions, i.e., they fail to pick up the small “humps” deliberately imposed in the DGP.

6. Conclusion

We have shown that if the true effects are heterogeneous, then an estimated homogeneous treatment effect may not be informative on outcome measures of interest. Unless the

distribution of treatment effects is independent of the unobserved treatment and final destination propensities, the estimated treatment effect will generally be a biased estimator for both the average treatment effect and for the average treatment effect on the treated.

We have also shown that it is possible to obtain a number of informative treatment effect statistics from observed data by estimating the distribution of treatment effects jointly with the distribution of other unobserved covariates. This can be done by means of a full-dimensional nonparametric maximum likelihood estimator or by a factor loading model of reduced dimensionality. In our trials, the latter approach turns out to yield more robust results than the former. It is essential, however, that the factor loading dimensionality is sufficient to represent the relevant sorting processes with some flexibility. A factor loading model with only a single latent variable performs poorly.

The nonparametric specification of the distribution of treatment effects and the other unobserved covariates also makes it possible to examine the sorting into treatment, both with respect to the final outcome of interest and with respect to the size of the treatment effect. We argue that it is important to uncover these sorting processes in order to assess and understand the overall impacts of a program. Characterizing the sorting process may also be of interest in its own right, and we show that correlation measures are reasonably well estimated.

An important limitation to nonparametric modeling of unobserved treatment effects is that the resultant treatment effect statistics are subject to an unknown sampling distribution, making it difficult to perform statistical inference. This problem is amplified by the fact that the likelihood function subject to maximization is not globally concave, implying that there is a likelihood of ending up at a non-global local maximum. We argue that nonparametric or semiparametric bootstrap techniques can be successfully applied both to ensure that a given estimation result is not based on a non-global maximum and to provide some basis for statistical inference. It appears that most treatment effect statistics are approximately normally

distributed apart from outlier results due to selection of a non-global optimum. However, we have not been able to provide a recipe for statistical inference that is completely free from subjective judgment.

Appendix

Invariance of factor loading model with respect to selection of loading factors

We have written the W-vector α as a linear combination of v_e and v_p:

$$\alpha = V \begin{bmatrix} 1 & \dots & 1 \\ v_e^1 & \dots & v_e^W \\ v_p^1 & \dots & v_p^W \end{bmatrix}. \tag{13}$$

An alternative would be to write α , v_e and v_p as linear combinations of freely estimated parameter vectors ϕ_1 and ϕ_2.

$$\begin{bmatrix} \alpha' \\ v'_e \\ v'_p \end{bmatrix} = \begin{bmatrix} V_\alpha \\ V_e \\ V_p \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 \\ \phi_1^1 & \dots & \phi_1^W \\ \phi_2^1 & \dots & \phi_2^W \end{bmatrix} \tag{14}$$

(where apostrophes do not indicate transposes, just names).

Given *V*, v_e and v_p as in (13) we may take

$$\begin{aligned} V_\alpha &= V \\ V_e &= [0 \quad 1 \quad 0] \\ V_p &= [0 \quad 0 \quad 1] \\ \phi_1 &= v_e \\ \phi_2 &= v_p \end{aligned}$$

in (14); thus (13) is a special case of (14).

On the other hand, given (14) we may take $v_e = v'_e, v_p = v'_p$, let

$$A = \begin{bmatrix} 1 & \dots & 1 \\ & v'_e & \\ & v'_p & \end{bmatrix}$$

and solve $\alpha' = VA$ for *V* in (13) to see that (14) is a special case of (13), thus the formulations are equivalent. To see that we may actually find such a *V*, we note that AA^t is invertible if *A*

has rank 3, thus $V = \alpha' A^t \left(AA^t \right)^{-1}$. That A has rank 3 is merely that v_e and v_p are linearly independent (together with a constant term). In any case we do not get more from (14) than from (13). This argument easily generalizes to higher dimensions.

References

Aakvik, A., Heckman, J., and Vytlacil, E.J. (2005) Estimating Treatment Effects for Discrete Outcomes when Responses to Treatment Vary: An Application to Norwegian Vocational Rehabilitation Programs. *Journal of Econometrics*, Vol. 125, 15-51.

Abbring, J.H. and Van den Berg, G.J. (2003) The Nonparametric Identification of Treatment Effects in Duration Models. *Econometrica*, Vol. 71, 1491-1517.

Abbring, J.H. and Van den Berg, G.J. (2004) Analyzing the effect of dynamically assigned treatments using duration models, binary treatment models, and panel data models. *Empirical Economics*, Vol. 9, No. 1, 5-20.

Abbring, J.H. and Van den Berg, G.J. (2005) Social Experiments and Instrumental Variables with Duration Outcomes, Working paper, IFAU, Uppsala.

Baker, M. and Melino, A. (2000) Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study. *Journal of Econometrics*, Vol. 96, 357-393.

Burnham, K.P. and Anderson D.R. (2002) *Model Selection and Mulitmodel Inference*, Springer.

Carneiro, P., Hansen, K., and Heckman, J. (2003) Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice. *International Economic Review*, Vol. 44, 361-422.

Crépon, B., Dejemeppe, M., and Gurgand, M. (2005) Counselling the Unemployed: Does it Lower Unemployment Duration and Recurrence? IZA Discussion Paper No. 1796.

Crépon, B., Ferracci, M., Jolivet, G., and Van den Berg, G.J. (2009) Active Labor Market Policy Effects in a Dynamic Setting, *Journal of the European Economic Association*, Vol. 7, 595-605.

Fredriksson, P. and Johansson, P. (2008) Dynamic Treatment Assignment – The Consequences for Evaluations Using Observational Data, *Journal of Business and Economic Statistics*, Vol. 26, 435-445.

Gaure, S., Røed, K., and Zhang, T. (2007) Time and Causality – A Monte Carlo Evaluation of the Timing-of-Events Approach. *Journal of Econometrics*, Vol. 141, 1159–1195.

Heckman, J. and Singer, B. (1984) A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, Vol. 52, 271-320.

Lindsay, B. G. (1983) The Geometry of Mixture Likelihoods: A General Theory. *The Annals of Statistics*, Vol. 11, 86-94.

Richardson, K. and Van den Berg, G.J. (2008) Duration Dependence Versus Unobserved Heterogeneity in Treatment Effects: Swedish Labor Market Training and the Transition Rate to Employment, Working paper 2008-07, IFAU Uppsala.

Rosholm, M. and Svarer, M. (2008) The Threat Effect of Active Labour Market Programmes. *Scandinavian Journal of Economics*, Vol. 110, No. 2, 385-401.

Røed, K. and Raaum, O., (2006) Do Labour Market Programmes Speed up the Return to Work?, Oxford Bulletin of Economics & Statistics , Vol. 68, No. 5, 541-68

Van den Berg, G.J., Van der Klaauw, B., and Van Ours, J.C. (2004) Punitive Sanctions and the Transition Rate from Welfare to Work. *Journal of Labor Economics*, Vol. 22, No. 1, 211-241.

Van den Berg, G.J., Caliendo, M. and Uhlendorff, A. (2009) Matching or Duration Models? A Monte Carlo Study. Mimeo.

Van den Berg, G.J., Bozio, A., and Costa Dias, M. (2010) Policy Discontinuity and Duration Outcomes, Mimeo.

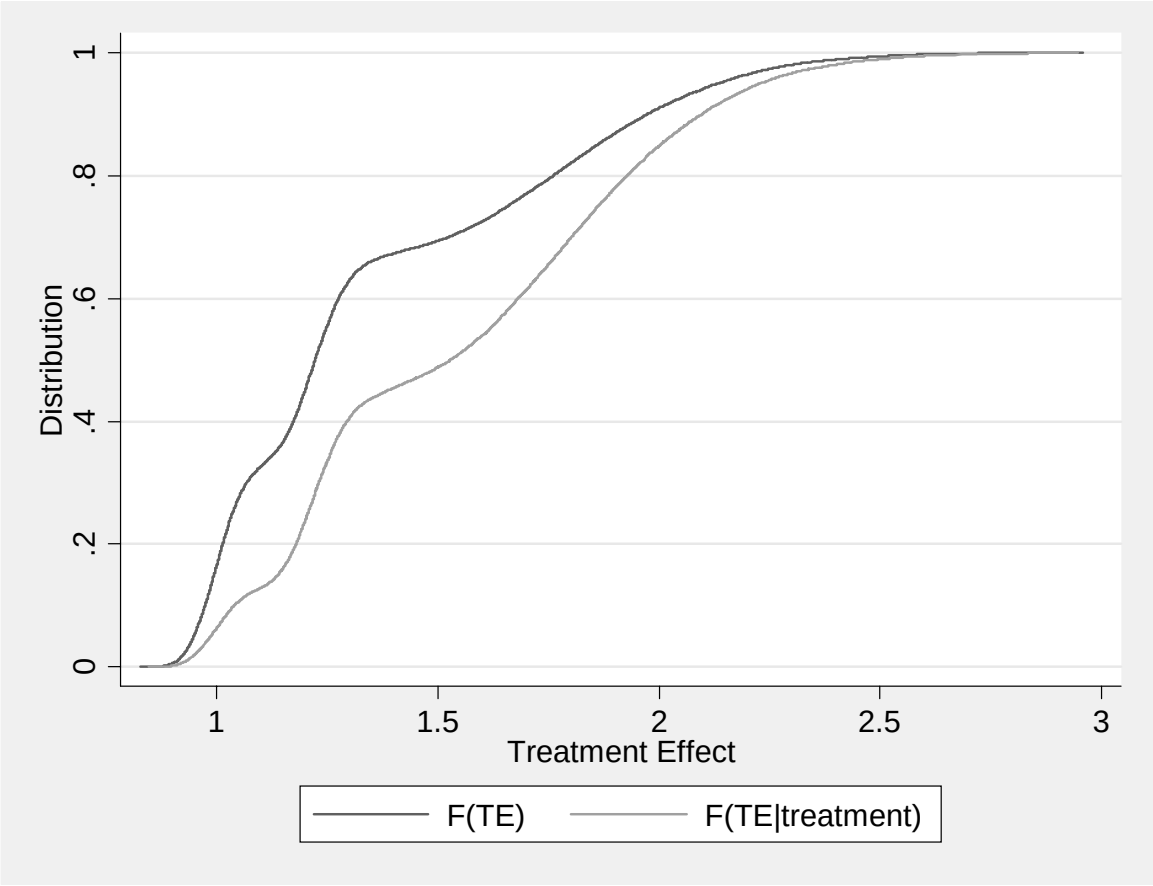

Figure 1. Cumulative distribution functions (CDF) for treatment effects and treatment effects among the treated in the baseline DGP

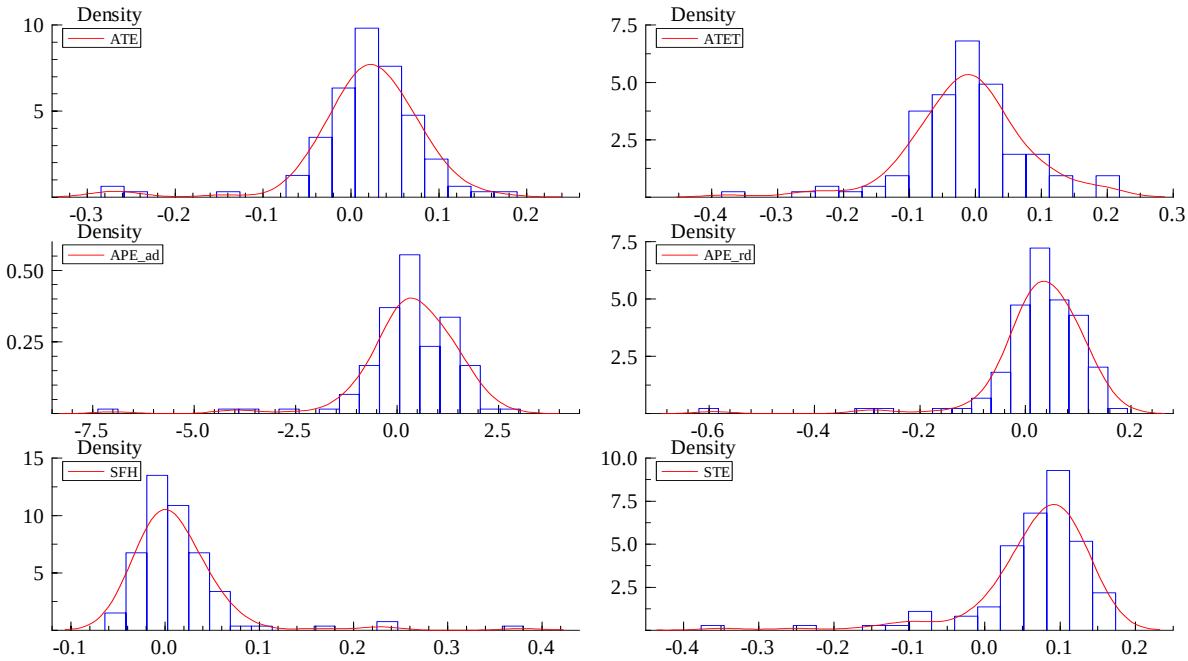

Figure 2. The distribution of estimation errors over the 120 trials. Two-dimensional linear factor loading model.

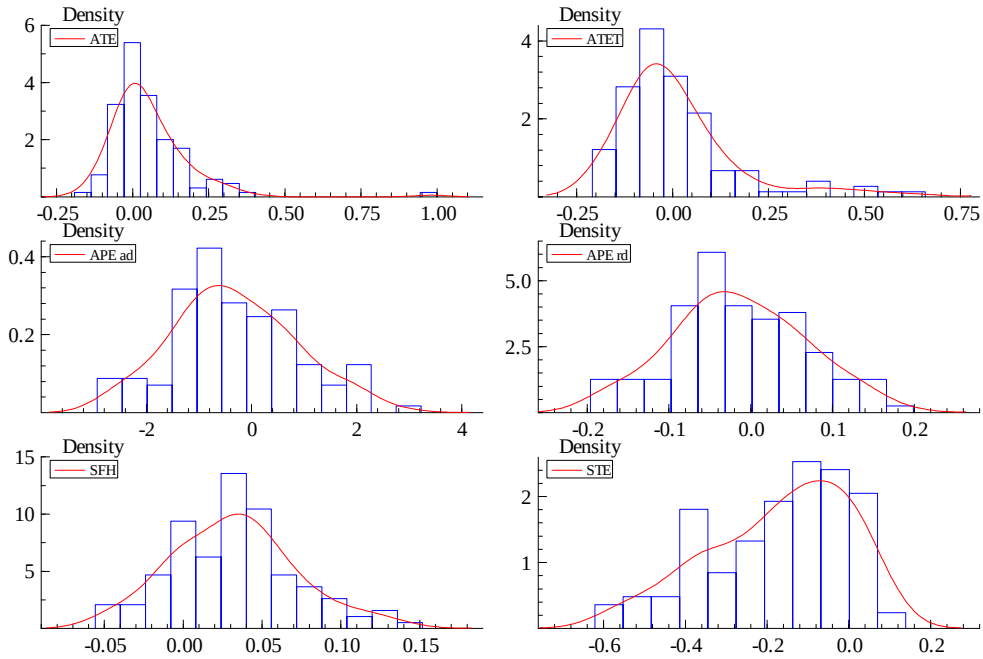

Figure 3. The distribution of estimation errors over the 120 trials. Full-dimensional NPMLE model.

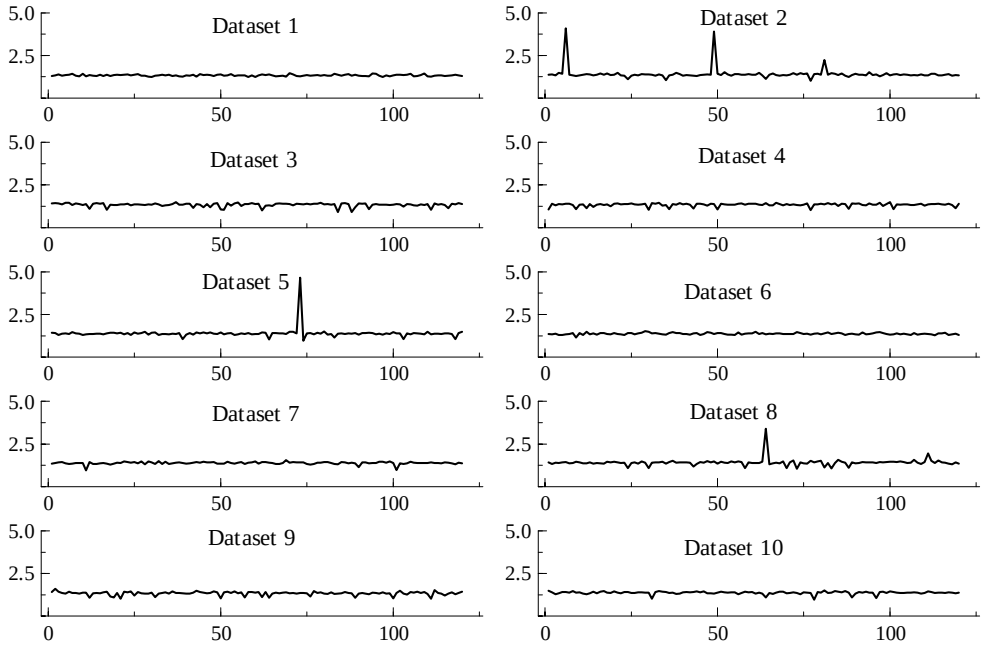

Figure 4. ATE-estimates based on 10×120 bootstrap samples.

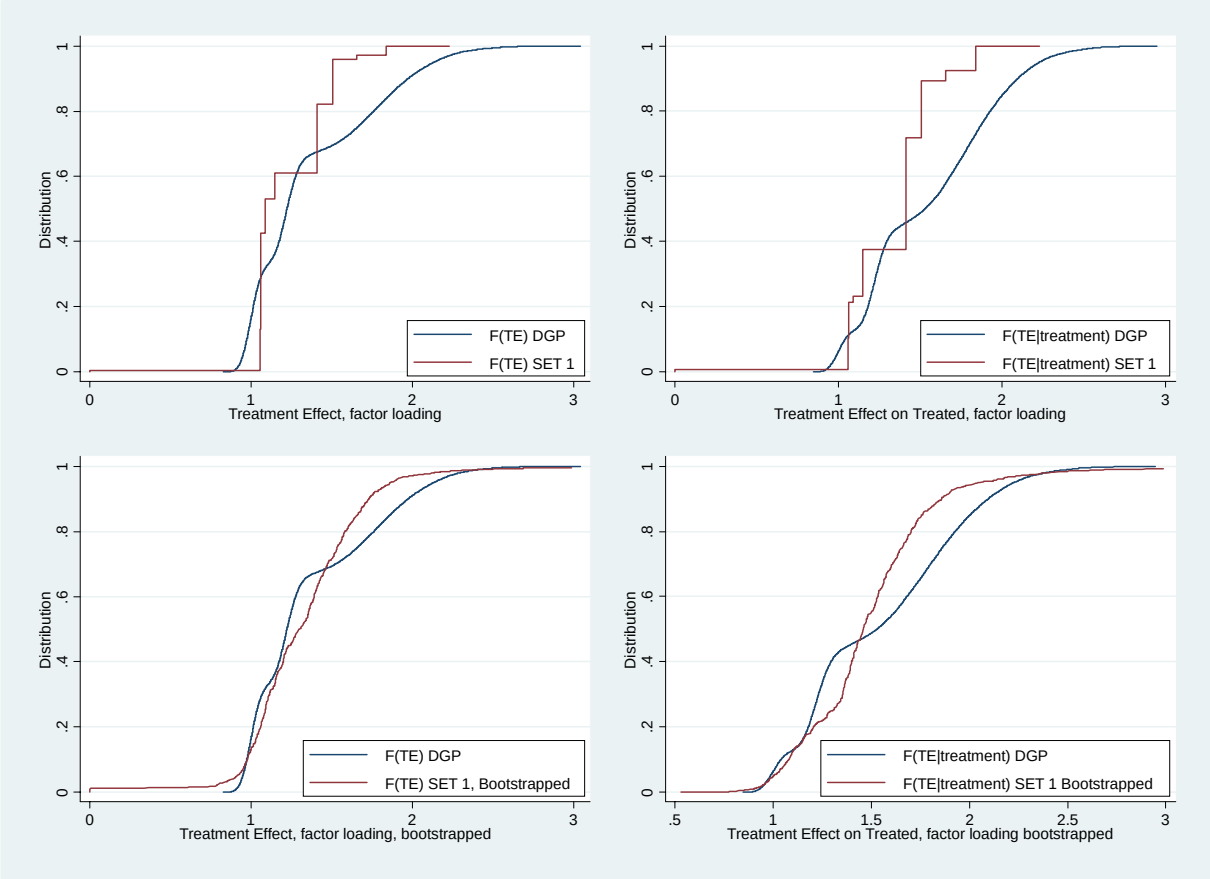

Figure 5. Distribution of estimated treatment effects based on a single full-sample estimation (upper panels) and on 120 bootstrap estimations (lower panels). Dataset 1. Two-dimensional linear factor loading model.

Table 1
Summary Statistics for Treatment and Program Effects and Sorting. Results from 120 trials.

	DGP	No unobserved heterogeneity		One-dimensional NPMLE				Two-dimensional NPMLE				Full-dimensional NPMLE	
		I		II Fixed α		III Linear factor loading		IV Fixed α		V Linear factor loading		VI	
Mean number of support points	-	1		4.7		4.9		10.0		10.2		9.9	
		Mean est.	**St. d.**	**Mean est.**	**St. d.**	**Mean est.**	**St. d.**	**Mean est.**	**St. d.**	**Mean est.**	**St. d.**	**Mean est.**	**St. d.**
Treatment effects													
ATE	**1.360**	0.604	0.008	0.816	0.019	0.830	0.020	1.367	0.031	1.378	0.065	1.412	0.134
StDevTE	**0.394**	-	-	-	-	0.060	0.017	-	-	0.341	0.104	0.809	0.919
ATET	**1.547**	0.604	0.008	0.816	0.019	0.817	0.018	1.367	0.031	1.535	0.088	1.555	0.153
StDevTET	**0.405**	-	-	-	-	0.056	0.015	-	-	0.368	0.125	0.867	1.159
Program effects													
APE^{AD}	**5.081**	-5.521	0.268	-2.198	0.378	-2.495	0.317	3.818	0.391	5.396	1.152	4.753	1.122
APE^{RD}	**0.399**	-0.644	0.035	-0.230	0.041	-0.263	0.036	0.322	0.031	0.426	0.087	0.384	0.078
Sorting													
Selection on final hazard (SFH)													
correlation	**0.003**	0.573	0.008	0.550	0.047	0.566	0.034	0.054	0.045	0.015	0.063	0.057	0.093
concordance	**0.003**	0.489	0.007	0.544	0.010	0.547	0.010	0.078	0.041	0.017	0.055	0.035	0.039
Selection on treatment (STE)													
correlation	**0.339**	-	-	-	-	0.351	0.021	-	-	0.466	0.104	0.162	0.174
concordance	**0.449**	-	-	-	-	0.303	0.018	-	-	0.515	0.075	0.277	0.171
Expected durations													
$E[D_i^0]$	**12.736**	8.573	0.085	9.590	0.138	9.500	0.127	11.863	0.437	12.583	0.520	12.298	0.459
$E[D_i^\Delta]$	**10.869**	10.605	0.106	10.402	0.123	10.421	0.115	10.431	0.365	10.589	0.227	10.548	0.166
Participation rate (P)	**0.368**	0.368	0.003	0.370	0.003	0.369	0.003	0.375	0.015	0.370	0.008	0.368	0.005

Note: In order to calculate ATE and ATET for the two-dimensional and the full-dimensional models, we have truncated the estimated heterogeneity distributions to avoid defective risks and, hence, irrelevant treatment effects. This has been implemented by disregarding location vectors with $v_e < -5$ or $v_p < -5$, and rescale the probability distributions accordingly.; see Section 2 for a discussion.

Table 2
Estimated homogenous treatment effects when the true effects are heterogeneous. Results from 20 trials on 7 different DGPs

	Corr(v_p,α)	Corr(v_p,v_e)	Var(α)	True ATE (DGP)	True ATET (DGP)	Mean estimate of exp(α)	St. dev. estimate of exp(α)
I	0.86	-0.5		1.43	1.74	1.41	0.09
II	0.86	0	0.18	1.43	1.65	1.33	0.03
III	0.82	0.5	0.07	1.36	1.43	1.26	0.04
IV	-0.86	-0.5		1.43	1.13	1.24	0.03
V	-0.81	0	0.07	1.36	1.23	1.33	0.04
VI	-0.82	0.5	0.07	1.36	1.28	1.40	0.04
VII	0	-0.5	0.23	1.34	1.34	1.33	0.04

Note: Estimates of exp(α) are based on a model in which the distribution of (v_p,v_e) is estimated nonparametrically (NPMLE).

Table 3
Results for nonparametric bootstrap on a single randomly selected dataset (dataset No. 1)

		Two-dimensional linear factor loading model				Full dimensional model			
	DGP	Estimate from selected dataset	St. Dev. 120 original datasets	Mean bootstrap estimate selected dataset	St. Dev. 120 bootstrap trials on selected dataset	Estimate from selected dataset	St. Dev. 120 original datasets	Mean bootstrap estimate selected dataset	St. Dev. 120 bootstrap trials on selected dataset
Treatment effects									
ATE	**1.360**	1.320	0.065	1.338	0.047	1.354	0.134	1.395*	0.197*
StDevTE	**0.394**	0.275	0.104	0.298	0.107	0.906	0.919	1.179*	2.882*
ATET	**1.548**	1.474	0.088	1.492	0.071	1.401	0.153	1.455*	0.106*
StDevTET	**0.405**	0.293	0.125	0.319	0.126	0.483	1.159	1.051*	3.041*
Program effects									
APE^{AD}	**5.062**	4.903	1.152	5.227	0.829	3.778	1.122	4.061	0.852
APE^{RD}	**0.399**	0.420	0.087	0.415	0.056	0.333	0.078	0.336	0.063
Sorting									
Selection on final hazard (SFH)									
correlation	**0.003**	0.006	0.063	0.024	0.035	0.041	0.093	0.045	0.045
concordance	**0.003**	0.046	0.055	0.031	0.034	0.068	0.039	0.055	0.038
Selection on treatment (STE)									
correlation	**0.339**	0.510	0.104	0.496	0.057	0.032	0.174	0.064	0.125
concordance	**0.449**	0.516	0.075	0.528	0.042	0.306	0.171	0.227	0.138
Expected durations									
$E[D_i^0]$	**12.682**	11.664	0.520	12.549	0.440	11.331	0.459	12.014	0.336
$E[D_i^\Delta]$	**10.823**	9.897	0.227	10.609	0.260	9.965	0.166	10.518	0.220
Participation rate (P)	**0.367**	0.360	0.008	0.371	0.009	0.362	0.005	0.369	0.007

* Four extreme outlier results were removed prior to the computation of these statistics.