24 September 2004
Work in progress, comments welcome

Time and Causality: A Monte Carlo Assessment of the Timing-of-Events Approach*

Simen Gaure
Centre for Information Technology Services, University of Oslo and
The Ragnar Frisch Centre for Economic Research, Oslo,

Knut Røed
The Ragnar Frisch Centre for Economic Research, Oslo,

Tao Zhang
The Ragnar Frisch Centre for Economic Research, Oslo

Abstract

We present new and encouraging Monte Carlo evidence regarding the feasibility of separating causality from selection within non-experimental event history data, based on the Non-Parametric Maximum Likelihood Estimator (NPMLE). Provided that the model is correctly specified and that no unjustified restrictions are imposed on the distribution of unobserved heterogeneity, the effect of non-random treatment on subsequent transition propensities can be accurately recovered from observed data. However, the approach is vulnerable towards misspecification, and sources of non-modelled unobserved heterogeneity can cause substantial biases in the estimated parameters.

Keywords: NPMLE, treatment effect

JEL Classification: C14, C15, C41,

---

* This paper is part of the project 'Mobilizing labour force participation', financed by the Norwegian Research Council. Thanks to John K. Dagsvik for valuable comments. Correspondence to: Knut Røed, the Ragnar Frisch Centre for Economic Research, Gaustadalleen 21, 0349 Oslo, Norway. E-mail: knut.roed@frisch.uio.no.

1 Introduction

In this paper, we provide Monte Carlo evidence regarding the extent to which the timing of stochastic processes and their outcomes convey information that enable us to identify and estimate causal parameters. The unit of analysis in this paper is a subject entering into some state (the origin state), and its subsequent stochastic transition into another state (the destination state). We are interested in how non-random events during the occupation of the origin state affect the probability of making a transition to the destination state. The paper focuses on two types of events. The first is a *treatment*, which we may think of as some kind of (induced) change in the economic environment relevant for the transition propensity. The second is the outcome of the stochastic transition process itself, as reflected in the *duration* of still progressing origin-state-spells. The problems associated with identification of treatment effects and duration dependence arise from imperfect control for subject heterogeneity, e.g. because parts of this heterogeneity is unobserved by the researcher. The distribution of uncontrolled heterogeneity obviously changes as the time spent in the origin state elapses. And to the extent that the treatment in question is not fully randomised through a controlled experiment, it is also likely that the distribution of uncontrolled heterogeneity varies between the treatment and the non-treatment observations. These problems are well known and described in the literature (see e.g. Heckman et al, 1999), and they will not be further elaborated here. The purpose of the present paper is to evaluate identification and estimation strategies for non-experimental data. Our results suggest that causal effects can be accurately and reliably recovered from non-experimental data, provided that the model is correctly specified and that there exist some sources of exogenous variation in transition propensities. A particularly useful source of iden-

tification, that has received only modest attention in the literature, is the existence of a common calendar time source of variation in hazard rates (provided, of course, that calendar time is not perfectly correlated to process time). Our encouraging results hold for a large variety of unobserved heterogeneity distributions, including distributions containing defective risks. However, we also find that unjustified restrictions on the model structure or on the heterogeneity distribution may lead to seriously biased results.

The evaluation presented in this paper builds on artificial observations for which the true Data Generating Process (DGP) is known. Data-structures that are similar to the DGP's evaluated in this paper arise in many types of real-world applications. The most obvious situation to think of is perhaps that of an individual entering into an origin state of e.g. unemployment, welfare participation, or sickness absence. In these cases, the destination state is typically that of ordinary employment, while the treatment may be a benefit sanction, a labour market program, or some kind of medication. Another example is an individual entering into the origin state of a job, and thereafter consider whether to quit this job for another, or to pull out of the labour force (retire). In this case, the treatment could be a promotion or a pay rise. In our experiments, we focus on situations in which large numbers of observations are available, to facilitate estimation techniques that are as 'non-parametric' as possible. With respect to the examples referred to above, that kind of data are now, in many countries, accessible from administrative registers, and such registers are likely to play an important role in future micro-econometric research; see e.g. Røed and Raaum (2003b). But even with the best conceivable data at hand, no statistical model can be completely non-parametric, hence an important concern in our evaluation is the extent to which formal non-parametric identification results carry over to realistically de-

signed data and models. Our paper serves as a Monte Carlo evaluation of what has become known as the *timing of events approach* (Abbring and Van den Berg, 2003a). The strategy of the paper is to use non-parametric estimation techniques to recover true causal parameters, under alternative (realistic) assumptions about the data. Our paper is closely related to a previous Monte Carlo study by Baker and Melino (2000), who investigated the behaviour of the nonparametric maximum likelihood estimator (NPMLE) for a discrete single risk duration model with unobserved heterogeneity and unknown duration dependence. Baker and Melino concluded that NPMLE in many cases resulted in a substantial bias in estimated duration dependence as well as in the effects of observed heterogeneity, but that the usage of an information criterion with a penalty attached to the number of support points in the heterogeneity distribution could lead to a dramatic improvement. Our findings suggest that NPMLE is extremely reliable as long as the model is correctly specified (apart from the unknown distribution of unobserved heterogeneity), but that misspecification of the model (for example in the form of a pre-specified number of support points in the heterogeneity distribution) can lead to serious bias problems. We also find that the 'over-parameterisation' problem identified by Baker and Melino (2000) is a strictly small sample problem. As the dataset becomes large, the difference between the Maximum Likelihood and the Maximum Penalised Likelihood estimators disappears (as long as the penalty is moderate).

The paper is structured as follows. The next Section describes the Data Generating Process (DGP) that we refer to as the baseline model. Section 3 discusses identification issues, and Section 4 outlines the statistical model used to recover the parameters of the DGP. Section 5 then discusses the models' ability recover the true baseline model parameters. Section 6 looks at the influence of sample size. Section 7

studies the impact of complicating the statistical distribution of unobserved heterogeneity, while Section 8 looks at the consequences of changing the causal parameters in the DGP in ways that potentially can affect the scope for identification. Section 9 discusses more fundamental deviations in the DGP from the basic assumptions underlying the estimated model (such as the proportional hazards assumption). Section 10 explores the consequences of, and suggests a remedy for, sample-selection due to interval censoring. Finally, Section 11 concludes.

2 The Data Generating Processes

The setting of our analysis is the following: There is an observation window of Q calendar time periods for which the researcher has access to records of entries into an origin state and subsequent transitions into a treatment state p and/or a final destination state e. The treatment may (or may not) have a causal effect on the hazard rate into the final destination state, both during (on-treatment effect) and after (post-treatment effect) the treatment. The length of the treatment (if no exit occurs to the final destination state) is assumed predetermined and observed. The first cohort of entrants is monitored up to Q periods, the second Q-1 periods and so on, until the last cohort, which is monitored only 1 period. Still active spells are censored at the end of the observation window. The transition rate probabilities for each subject are governed by underlying continuous time hazard rates, which again are determined by five factors: calendar time (t), spell duration (d), an observed time-invariant covariate (x), treatment status (z), and a two-dimensional vector of time-invariant unobserved covariates (v). It is the two unobserved variables that embody the selection problem. They are drawn from a simultaneous probability distribution.

An important aspect of real data is that they rarely conform to the idea of continuous time measurement. Real data records are typically updated at particular points

in time, such as by the end of each week or month. We take this point-in-time sampling into account by generating data that do not record exact transition times, but instead record in which discrete time interval each transition has taken place. We assume, however, that the underlying continuous time hazard rates are constant (or develop smoothly) within each of these time intervals.

We generate a number of different datasets characterised by different types of (and degrees of) calendar time effects, different degrees of duration dependence, different treatment effects and different distributions of unobserved heterogeneity. Although we stress the generality of the timing of events approach, we have chosen to design the artificial data such that they resemble genuine administrative register data (that we are familiar with), in which the origin state is *open unemployment*, the treatment state is a *labour market programme*, and the destination state is *regular employment*. The size of the observation window, the level of the period-specific transition rates, and the magnitudes of the various causal effects, are chosen roughly to match that kind of data.

Since the processes under study are assumed to be observed only at a finite number of discrete points in time, we set up the DGP in terms of grouped (discrete) hazard rates. Let $\varphi_k(t,d,x,z_t,v_k)$ denote the period-specific integrated hazard rate, integrated over the time interval (t-1,t] governing the transition to state $k=e,p$, given that the spell duration by the end of this interval is d periods and given the observed explanatory variable x and the unobserved scalar v_k, and given the treatment status z_t. The treatment status has two dimensions as captured by the indicator variables $z_t=(z_{1t},z_{2t})$. The variable z_{1t} is equal to 1 during treatment (and 0 otherwise), while z_{2t} is equal to 1 after a treatment is completed (and 0 otherwise). Note that previous treatment is assumed to be irrelevant while a subject is enrolled again, (i.e. $z_t \neq (1,1)$).

In most of the datasets that we generate, the underlying hazard rates are proportional in the effects of calendar time, spell duration, observed heterogeneity, unobserved heterogeneity and treatment. The integrated period-specific hazard rates φ_k can then be written as

$$\begin{aligned}\varphi_e(t,d,x_{it},z_{it},v_{ei}) &= \exp\left(\beta_e x_{it} + \sigma_{et} + \lambda_{ed} + \alpha z_{it} + v_{ei}\right), \\ \varphi_p(t,d,x_{it},z_{it},v_{pi}) &= \exp\left(\beta_p x_{it} + \sigma_{pt} + \lambda_{pd} + v_{pi}\right)\end{aligned}, \qquad (1)$$

where σ_{kt} and λ_{kd} are the period-specific calendar time and duration dependence parameters, respectively, and α is the vector of treatment effects. Note that there are two dimensions of time in this model, process time (*d*) and calendar time (*t*). Calendar time should not be thought of a causal factor itself, but rather as a proxy for all external influences that jointly affect the hazard rates of the population at risk, such as business cycles, seasonal effects, or changes in treatment capacity. The period-specific transition probabilities are equal to

$$p_k(t,d,x_{it},z_{it},v_{ki}) = \left(1-\exp\left(-\sum_{k\in K_{it}}\varphi_k(t,d,x_{it},z_{it},v_{ki})\right)\right)\frac{\varphi_k(t,d,x_{it},z_{it},v_{ki})}{\sum_{k\in K_{it}}\varphi_k(t,d,x_{it},z_{it},v_{ki})}, \qquad (2)$$

where $K_{it} = \{p,e\}$ for $z_{it} = (0,0)$ (no treatment so far) or $z_{it} = (0,1)$ (completed treatment) and $K_{it} = \{e\}$ for $z_{it} = (1,0)$ (ongoing treatment).

We start out by generating a baseline model, which is described in Table 1. There is neither duration dependence nor treatment effects in the baseline model (i.e. constant hazard rates and irrelevant treatment), but there is negative selection on the observed covariate and positive selection on the unobserved covariates. The positively correlated unobservables will – if unaccounted for - produce a spurious pattern of negative duration dependence and favourable treatment effects.

Table 1 around here

3 Identification

The Mixed Proportional Hazards (MPH) structure of the baseline DGP ensures non-parametric identification of both treatment effects (Abbring and Van den Berg, 2003a) and duration dependence (Elbers and Ridder, 1982; Heckman and Honoré, 1989; Abbring and Van den Berg, 2003b). In practice, the scope for actually recovering the true parameters from observed data depends on the degree of exogenous variation in the hazard rates stemming from observed covariates. In our model, there are two observed sources of exogenous variation in hazard rates; the time invariant (and subject-specific) covariate x and the calendar time period t.

Even though the identification results referred to above are all derived from the requirement of time-invariant covariates only, an important aim of the present paper is to explore the potential for non-parametric identification embedded in calendar time variation in hazard rates as well. Intuitively, time-varying covariates can recover the influences of unobserved heterogeneity because, for a population of subjects with common spell duration above zero, it will be the case that the present distribution of unobserved heterogeneity depends on hazard rates experienced earlier in the spells, while current transition rates do not (Van den Berg and Van Ours, 1994; 1996). Hence, as pointed out in a similar context by Eberwein et al (1997, p. 663), time-varying variables naturally provide an exclusion restriction in the sense that past values of these variables affect the current transition probabilities only through the selection process. As a result, mixed hazard rate models may be non-parametrically identified even in the absence of the proportionality assumption (McCall, 1994; Brinch, 2000). Time-varying covariates may therefore provide a more robust source of identification than time-invariant covariates.

4 The Statistical Method and Model Evaluation Criteria

The parameters are recovered by means of a Non-Parametric Maximum Likelihood (NPML) technique. Each subject contributes to the analysis with a number of observations equal to the number of periods at risk of making a transition of some sort. Each observation is described in terms of calendar time, spell duration, the value of explanatory variables and an *outcome* (generated by the drawings described in the previous section). Let y_{kit} be an outcome indicator variable which is equal to 1 if the corresponding observation period ended in a transition to state *k*, and zero otherwise, and let N_i be the set of potential transition periods observed for subject *i*. The contribution to the likelihood function formed by a particular subject, conditional on the vector of unobserved variables $v_i = (v_{ei}, v_{pi})$ can then be formulated as

$$L_i(v_i) = \prod_{i \in N_i} \left[\begin{array}{l} \prod_{k \in K_{it}} \left[\left[\left(1 - \exp\left(- \sum_{k \in K_{it}} \varphi_k(t, d, x_{it}, z_{it}, v_{ki}) \right) \right) \frac{\varphi_k(t, d, x_{it}, z_{it}, v_{ki})}{\sum_{k \in K_{it}} \varphi_k(t, d, x_{it}, z_{it}, v_{ki})} \right]^{y_{kit}} \right] \\ \times \left[\exp\left(- \sum_{k \in K_{it}} \varphi_k(t, d, x_{it}, z_{it}, v_{ki}) \right) \right]^{\left(1 - \sum_{k \in K_{it}} y_{kit} \right)} \end{array} \right] \quad (3)$$

Since the distribution of unobserved heterogeneity is assumed unknown to the researcher, we approximate the heterogeneity distribution in a non-parametric fashion with the aid of a discrete distribution (Lindsay, 1983; Heckman and Singer, 1984). Let *W* be the (a priori unknown) number of support points in this distribution and let $\{v_l, p_l\}$, $l = 1, 2, ...W$, be the associated location vectors and probabilities. In terms of observed variables, the likelihood function is then given as

$$L = \prod_{i=1}^{N} \underset{v_i}{E}[L_i(v_i)] = \prod_{i=1}^{N} \sum_{l=1}^{W} p_l L_i(v_l), \quad \sum_{l=1}^{W} p_l = 1. \quad (4)$$

Our estimation procedure is to maximise this function with respect to all the model and heterogeneity parameters repeatedly for alternative values of *W*. We start out with *W=1*, and then expand the model with new support points until the model is 'saturated'. Our maximization method is a combination of Fisher scoring (i.e. Newton-Raphson with the Hessian replaced by the Fisher information matrix) and BFGS.[1] Our approach for identifying the 'correct' number of support points differs from the one used by Baker and Melino (2000). Rather than performing a separate maximization with respect to the heterogeneity distribution before maximizing with respect to all parameters, we make a limited random search in the following simple-minded fashion:

1. We copy an old non-defective (randomly selected) mass-point and assign it a probability of 0.0001.
2. For the first element in the location vector, we pick 50 random numbers between minus 3 and 2 and compute the resulting likelihood functions.
3. If one of the random numbers yields an improvement of the likelihood, we use it (if more than one yields an improvement, we use the best); otherwise, we keep the old one. We then go back to (2) for the next element in the location vector.

---

[1] For the Fisher scoring we have modified Xie and Schlick's TNPACK (from http://www.netlib.org), and for BFGS we have used Zhu, Byrd, Lu and Nocedal's LBFGS-B. Both of these methods have their strengths and weaknesses. Fisher scoring usually converges fast, and the Fisher matrix is easy to compute since we anyway do analytic gradients. BFGS converges slower, but is much more robust. Typically we start out with 100 iterations of BFGS, then we switch to Fisher scoring. Normally, a maximum is found with 3-10 iterations of Fisher. However, some models are harder, in particular when the number of mass-points increases. If we haven't converged after 50 Fisher iterations, we switch back to BFGS, this time with more iterations. We switch back and forth a couple of more times, and this is usually sufficient for convergence. Experience has shown that both BFGS and Fisher may at times get stuck, apparently due to an ill-conditioned Hessian in certain regions, or because the Fisher matrix is too different from the Hessian. By switching between them we often manage to move out of the problematic region.

4. If we have improved the likelihood through this exercise, we use the newly found heterogeneity distribution, together with the previously estimated parameters, as an initial vector, and start the full maximization.
5. If we have not been able to improve the likelihood through direct search, we replace the location vectors and probabilities with random numbers and start the full maximization anyway.[2]
6. We continue adding mass-points until there is no improvement in the likelihood. For practical and computational reasons, we consider this to be the case when the likelihood increases by less than 0.05.

According to the Maximum Likelihood (ML) criterion the model is saturated when the likelihood cannot be made any larger by adding additional support points to the heterogeneity distribution. However there is some discussion in the literature about the need for information criteria that 'punish' parameter abundance (Leroux, 1992; Baker and Melino, 2000). Let $\hat{\mu}_W$ be the vector of parameter estimates derived from a model with W support points in the heterogeneity distribution and let $l(\hat{\mu}_W)$ be the corresponding log-likelihood function. A general form of a Maximum Penalised Likelihood Criterion is $l(\hat{\mu}_W) - a_M(\#\,parameters)$, where a_M is a penalty function derived from the total number of observations M. Baker and Melino (2000) propose to use either the Bayesian Information Criterion (BIC) or the Hannan-Quinn Information Criterion (HQIC) in order to avoid "over-parameterisation" of the heterogeneity distribution. The BIC uses the penalty function $a_M = 0.5 \ln M$, while the HQIC uses

[2] Sometimes, a heterogeneity parameter is estimated as a large negative number (< -20). This is numerically problematic. When we encounter this, we mark the offending parameter as 'negative infinity' and keep it out of further estimation. The 'negative infinity' mark is kept when we add new mass-points. This also implies that we allow defective risks to be present in the data.

$a_M = \ln(\ln M)$. Zhang (2003) found, however, that the much "milder" penalty provided by the Akaike Information Criterion (AIC), with $a_M = 1$ perform better than BIC and HQIC in a setting similar to the one used here.

5 Recovering the Baseline Model from Observed Data

The aim of this section is to assess the statistical model's ability to uncover the true causal parameters in repeated trials of data generation and estimation. For this purpose, we generate 100 distinct datasets from the assumptions of the baseline model, each with 50,000 subjects. Some key characteristics of these datasets are described in Table 2. The average size is 492,000 observations, implying that the average duration of origin state spells is 9.8 periods (including right-censored spells, which made up 28.9 per cent of all spells). Despite their common DGP, the random drawings of unobserved heterogeneity and calendar time effects ensure that the data sets differ a lot. While the smallest dataset has an average spell duration of only 3.2 periods, and, hence, contains as little as 160,000 observations, the largest has an average spell duration of 17.3 periods and contains 864,000 observations. There is also a substantial variation between the datasets in the fraction subject to treatment (from 4 to 87 per cent) and in the degree of censoring (from 4 to 72 per cent).

Table 2 around here

We let each of the 100 datasets be subject to the estimation procedure described in section 3. For each trial, around 165-200 parameters are estimated, depending on the number of support points in the mixing distribution.[3] Table 3 reports the number of mass-points that were required to satisfy the four alternative model selec-

[3] In some of the datasets, there are also some coefficients, particularly attached to some of the last calendar time and spell duration parameters, that cannot be estimated due to lack of variation in outcomes (or "empty cells").

tion criteria, BIC, HQIC, AIC and ML, in the 100 trials. While the most restrictive information criterion, BIC, typically requires 4-7 support points, the least restrictive criterion, ML, typically requires 10-14.

Table 3 around here

Table 4 presents the main results regarding the four structural parameters of interest, and some summary statistics regarding the two duration baselines, while Figure 1 presents a more detailed picture of the non-parametrically estimated effects of spell duration. A first point to note is that the biases induced by failing to control for unobserved heterogeneity are large, not only in the estimated effects of treatment and spell duration, but also in the estimated effects of the exogenous covariate x. The latter results from the fact that subjects with $x=1$ require a high unobserved exit propensity in order to exit fast, while subjects with $x=0$ tend to make this exit quickly anyway; hence x becomes correlated with unobserved heterogeneity as the spells proceed (even though they are orthogonal to start with). It is sometimes claimed that the resulting bias is likely to be small insofar as the duration baseline is sufficiently flexible (see e.g. Narendranathan and Stewart, 1993; Arulampalam and Stewart, 1995); but the results above, which are based on a completely flexible duration baseline, show that this should not be taken for granted. Treatment effects are of course also biased by the selection related to the dependence between the unobserved employment and treatment propensities. Without controls for unobserved heterogeneity, we would typically draw the false conclusions that treatment increases the hazard rate to the final destination state by $100(\exp(0.443)-1)=55.7\%$ during the treatment, and by $100(\exp(0.324)-1)=39.0\%$ afterwards. We would also draw the false conclusion that there is strong negative duration dependence in both hazard rates, particularly in the final destination hazard, which declines with as much as

$100(\exp(-1.53)-1) = -78.3\%$ during the first 36 time periods. The estimated treatment baseline declines less, reflecting that subjects transiting to the treatment state return to the risk set when the treatment is completed. This also explains the peculiar step-wise rises in the estimated treatment hazard that occur as treatment participants (who, on average, are positively selected with respect to the unobserved treatment propensity) return to the origin state (after five periods of participation), and are again exposed to the risk of treatment.

Figure 1 around here

A second point to note is that the biases are eliminated by means of non-parametric control for unobserved heterogeneity, but that only the models with little or no penalty for parameter abundance (AIC and ML) eliminate the biases *completely*. Both the AIC and the ML criteria perform remarkable well, in the sense that they reliably return unbiased estimates close to the true parameter values.[4] The reported standard errors are also close to reflecting the true statistical uncertainty.[5] As a result, the standard t-tests tend to reject the true parameters almost in accordance with the nominal significance levels. There is, however, a slight tendency to over-reject the truth for some of the parameters in the final destination hazard. The reason for this is that the degree of statistical uncertainty tends to be underestimated in some of the replications. This seems particularly to be the case in the replications based on very small datasets (with very few period-observations for each subject). Figure 2 describes the

[4] Although not shown here, it may be noted that the model also recovered the true calendar time parameters with great precision. These parameters may in some cases have an interesting interpretation, e.g. in the form of business and/or seasonal cycles; see Gaure and Røed (2003).

[5] To verify this statement, we also made 100 data-replications based on the same population and economic environment (i.e. we drew the heterogeneity terms and calendar time effects only once, implying that only the transition 'lottery' was replicated) and compared the resultant estimated standard errors with the observed standard deviation. In this case, the estimated standard errors were almost exactly the same in each replication, and equal to the empirical standard deviation of the 100 estimates.

distribution of the 100 estimates for the four key structural parameters derived from the ML criterion, by means of histograms, and compares them to normal densities. The typical picture is that the distributions are "almost" normal, but that a few outliers disturb the picture. These "outliers" are typically generated from the models with few observations and very short spells.

Figure 2 around here

A third point to note is that there does not seem to be a great risk of 'over-correcting' for unobserved heterogeneity, in the sense that e.g. the negative duration bias imposed by neglected heterogeneity is replaced by a positive bias. On the contrary, there is a substantial risk of 'under-correcting' for unobserved heterogeneity when information criteria with large penalties for additional parameters are used. In particular, models selected on the basis of HQIC or BIC tend to reject the true parameters much more often than suggest by nominal significance levels. For example, at the five per cent nominal level, HQIC rejects the true value of β_e in 30 per cent of our replications, while BIC rejects the true value in as much as 75 per cent of the cases.

Given that the search for the optimal number of support points requires substantial computational resources – and hence that the number of support points in actual applications is often specified a priori as at most two or three - it may be of interest to investigate how models with just a few (predetermined) number of support points perform. For the four structural parameters, this is illustrated in Figure 3. It turns out that two support points is clearly insufficient to identify any of the parameters, while three points seem to do a good job in revealing the two treatment effects. However, a low number of support points seems utterly inadequate in order to identify the true spell duration effects. This is illustrated in Figure 4, where we have plotted

the average estimated duration parameters associated with the final destination hazard for models incorporating from 1 to 10 support points in the heterogeneity distribution. It is clear that the negative duration bias diminishes as more support points are included, but only the most flexible models (with up to 10 points) are able to remove it completely. Hence, in order to correctly disentangle duration dependence and selection, it seems to be essential that the heterogeneity distribution is saturated in terms of a maximum likelihood or a penalized maximum likelihood criterion.

Figure 3 around here

Figure 4 around here

Although the main purpose of applied research typically is to recover structural parameters of the type discussed above, it may sometimes also be of interest to recover properties of the heterogeneity distribution itself. It is of course not meaningful to interpret the mass-point distribution literally in terms of representing a corresponding number of distinct subject types, since the underlying true heterogeneity distribution may very well be continuous (as is the case in our baseline model). It is also interesting to note that although our 100 trials of data generation and estimation returned the same (correct) structural parameters, they returned very different heterogeneity parameters in terms of mass-point locations and probabilities. This is related to a fundamental symmetry in the likelihood function, implying that different combinations of mass-point locations and probabilities are observationally equivalent (a trivial example of this symmetry amounts to swap the locations, as well as the probabilities, of any two mass-points). But, even though the mass-point locations and probabilities themselves are not directly interpretable, there may be other properties of the estimated heterogeneity distribution that have a more substantive interpretation. Obvious candidates are the lower order moments of the distribution of the unobserved propor-

tionality terms (i.e. of the $\left(\exp(v_e),\exp(v_p)\right)$ distribution). In particular, in a treatment effect setting, it may be of interest to characterise the selection process into treatment in terms of, say, a correlation coefficient. Table 5 report our estimates of the first and second order moments of the heterogeneity distributions.[6] While the means are correctly, and also robustly, recovered, the second order moments are not always well represented. For the Maximum Likelihood criterion, the estimated correlation coefficients between the two latent variables are, on average, fairly close to the true value; hence the estimate seems to be consistent. However, there is a large statistical uncertainty associated with this parameter, and in our 100 trials, the estimated correlation coefficient ranged from a minimum of 0.05 to a maximum of 0.94 (the true value being around 0.40).

Table 5 around here

6 The Role of Sample Size

So far, the analysis has been based on datasets containing 50,000 subjects. In this section, we look at the impact of sample size, by comparing results based on five different sample sizes, containing from 5,000 to 5,000,000 subjects. The main results are summarised in Table 6, where we present the average number of support points in the estimated heterogeneity distribution for each model, as well as mean errors for some key parameters. The estimated number of support points seem to increase monoto-

[6] Note that it is empirically impossible to distinguish between different 'large' positive locations for v_e or v_p, since the exponential functional form by which they affect the hazard rate in any case imply that such numbers are associated with transition probabilities equal to unity (irrespective of other characteristics). At the same time, large positive locations for v_e or v_p may have a very strong impact on the calculations of first and second order moments (even when the associated probability is close to zero). For this reason, we have chosen to set an upper limit on these numbers before moments are calculated by replacing locations larger than 2 by the number 2. The exact selection of cut-off point does not matter much in practice, since these locations are typically attributed extremely low probability.

nously with sample size for all information criteria, suggesting that the sample size may have a substantive influence on the way our statistical model interprets the data.

Table 6 around here

The mean errors that are presented in Table 6 are all based on the same total number of subjects, irrespective of sample size, namely 5,000,000. When we look at sample sizes of only 5,000, we thus generate and estimate the model 1,000 times, and the reported mean errors are averages taken over all these trials. At the other extreme, when we look at sample sizes of 5,000,000, we only make a single trial. This means that if the parameter estimates are unbiased irrespective of sample size, the mean errors should be the same, and close to zero, for all sample sizes. However, Table 6 reveals that the mean errors do depend on sample size. The larger is the sample, the smaller are the mean errors, irrespective of the model selection criterion. Moreover, the larger is the sample, the less important is the selection of information criterion (for sufficiently large samples, all information criteria perform remarkably well). For small samples (5,000 or 10,000 subjects), there is a substantial risk of obtaining biased results, and the selection of information criterion seems to be of paramount importance. Like Baker and Melino (2000), we find that the ML criterion tends to 'over-correct' for unobserved heterogeneity in small-sample situations, and that a substantial improvement can be achieved by relying on an information criterion that penalises the number of parameters in the heterogeneity distribution. This is most clearly seen by looking at the mean errors associated with the final destination spell duration baseline (λ_{ed}). For example, for sample sizes of 10,000, we see that the ML criterion produces a positive bias in the spell duration parameters (on average equal to 0.173), while the AIC criterion delivers correct results. However, more restrictive information criteria (BIC and HQIC) tend to 'under-correct' for unobserved heterogeneity, and,

hence ,fail to remove the negative duration bias. This is more clearly illustrated in Figure 5, where we have plotted the mean duration parameter estimates from the 500 trials with sample sizes of 10,000 subjects. The pattern is the same for sample sizes of 5,000; BIC and HQIC 'under-corrects', ML over-corrects, and AIC (almost) hits the target.

Figure 5 around here

Our results suggest that AIC is the safest information criterion to rely on, particularly when samples are small. However, it is difficult to assess the generality of this result. The 'optimal' information criterion may be DGP-specific.

7 The Role of the Heterogeneity Distribution

In this section, we present some estimation results obtained from models with unobserved heterogeneity distributions that deviate from the baseline case. For each model, we repeat data generation and estimation 10 times only, in order to limit our usage of computational resources. To avoid too much variation in the number of observations from trial to trial (which would make comparisons between different models awkward), we have drawn unobserved heterogeneity and calendar time effects only once for each model type. The main aim of this section is to assess the extent to which the relatively optimistic identification results from the previous section holds for more challenging classes of heterogeneity distributions. In the presentation of our results, we restrict attention to parameter estimates based on AIC and ML (it is still the case that these criteria perform best). We first complicate the heterogeneity problem without changing the DGP, by assuming that the researcher does not observe the exogenous explanatory variable x; hence x is transformed into an unobserved (dichotomous) covariate, which, together with the bivariate normal covariate, now constitutes the unobserved heterogeneity distribution. Note that the researcher in this case is assumed

not to have access to any subject-specific exogenous covariates at all; hence it is only the calendar time dummy variables that ensure non-parametric identification of treatment effects and duration dependence effects. Even though the unobserved heterogeneity distribution is more complicated in this case, it is not unambiguously the case that the number of support points required to satisfy the two model selection criteria increases. The maximum likelihood criterion ended up requiring from 10 to 14 points, while the AIC required from 7 to 11 points, very much in line with the requirements when x was observed. The results regarding the treatment effects are presented in Table 7. These effects are still robustly identified, although standard errors are larger than what was the case when x was observed. The same conclusion applies to the spell duration baselines (not shown). Hence, with some exogenous variation in hazard rates over calendar time, no subject-specific covariates are required in order to identify treatment and spell duration effects.

Table 7 around here

Before we modify the DGP in order to include more complicated heterogeneity distributions, we take a look at the case in which the DGP does not contain any unobserved heterogeneity at all. When this is the case, a model without heterogeneity is obviously appropriate, but it could nevertheless be the case that we erroneously found some unobserved heterogeneity to be present. Indeed, when we used the maximum likelihood criterion for model selection, only one out of 10 trials ended up rejecting the presence of unobserved heterogeneity completely. In six of the trials, three support points were identified. However, the identified support points were either located closely together (almost indistinguishable), or the attached probability to the "deviating" mass-points was close to zero; hence the structural parameters of interest were not biased at all. When, we used the penalized likelihood criterion (AIC) to se-

lect model, all 10 trials ended up correctly rejecting the presence of unobserved heterogeneity.

We now briefly assess the consequences of complicating the unobserved heterogeneity distribution. We do this by presenting five illustrative example distributions. The first four examples are based on various combinations of continuous (Normal or Gamma) and discrete heterogeneity distributions. The last example is a pure discrete simultaneous distribution, in which some of the support points involve defective risks. A more detailed description of the various models and the main results are provided in Table 8. The bottom line is that the true structural parameters, including treatment effects, are robustly recovered from the data irrespective of the way unobserved heterogeneity is distributed. As illustrated in Figure 6, this also applies to the duration dependence parameters. These results also hold for a number of other heterogeneity distributions that we have tried; hence we conclude that the precise nature of the heterogeneity distribution is unimportant with respect to identification of our baseline model.

Table 8 around here

Figure 6 around here

It may be of interest to take a closer look at the results from model v), since this is the only model in which the DGP is actually based on a discrete heterogeneity distribution of the type used in the estimation procedure. Hence, this model could potentially be fully recovered from the data, in the sense that the correct mass-point locations and probabilities were identified. A particularly interesting issue is the models' ability to recover the true fraction of defective risks, since this fraction sometimes may be of substantive importance. As it turned out, the presence of defective risks (5 per cent in the DGP) in the final destination hazard was identified in all the 10 trials,

while the presence of defective risks (1 per cent in the DGP) in the treatment hazard was identified in 9 out of the 10 trials. In most cases, the corresponding estimated probability was also close to the true fraction of defective risks, particularly in the hazard with the largest defective risks fraction. However, it is not generally the case that the true mass-point locations are recovered. And none of the model selection criteria were particularly good at identifying the true number of support points (both criteria found the correct number of points in 2 out of the 10 trials only); the maximum likelihood criterion tended to return too many points, while AIC tended to return too few points. Given that other parts of the model are reliably recovered, this must reflect the fundamental symmetry property discussed in the previous section; i.e. that different combinations of mass-point locations and probabilities are equally consistent with data.

8 The Role of the True Causal Effects

As pointed out in the introduction to this paper, there are two substantive sources of non-parametric identification of treatment and spell duration effects in our artificial data: The exogenous subject specific covariate *x*, and the variation in hazard rates over calendar time σ_t. In this section, we start out by investigating how our ability to identify the true causal effects changes as we manipulate these two identification sources. We have already established that we do not need to observe the exogenous covariate *x*. We now proceed by also reducing the degree of variation in the calendar time component (while keeping the degree of variation in unobserved heterogeneity, which now also incorporates the variable *x*, constant) and by looking at possible consequences of calendar time effects being auto-correlated. Given the number of estimated models, we do not present complete graphical results for the spell duration pa-

rameters, but focus instead on the Weighted Mean Absolute Error (WMAE) of these parameters, using the inverse of the estimated standard errors as weights. Let $(\hat{\lambda}_{kdr}, \hat{\psi}_{kdr})$ be the estimated spell duration parameter and standard error corresponding to transition k and spell duration d in trial r. For R trials, $WMAE_k$ is defined as follows:

$$WMAE_k = \frac{1}{R}\sum_r \sum_d \frac{\frac{1}{\hat{\psi}_{kdr}}}{\sum_d \frac{1}{\hat{\psi}_{kdr}}} \left| \hat{\lambda}_{kdr} - \lambda_{kd} \right|. \tag{5}$$

Some illustrative results are provided in Table 9. As expected, the manipulation of the sources of identification primarily affects the estimates of spell duration baseline for the final destination state. The smaller is the variance of the calendar time parameters, the less precise are the estimates, and the larger is the expected mean absolute error in the estimated duration effects. This reflects that a reduction in the impact of calendar time variation reduces the data-based foundation for non-parametric identification of spell duration effects. Auto-correlated calendar time effects do not reduce the scope for identification.

Table 9 around here

The results presented so far are based on models in which treatment and duration effects are all equal to zero in the data generating process. But the conclusions do not depend on this assumption. We have also estimated models on DGP's containing positive and negative duration dependence and positive and negative treatment effects. Some illustrative results are provided in Table 10 and Figure 7.

Table 10 around here

Figure 7 around here

9 Non-Proportional Models and Parameter Heterogeneity

In this section, we look at the consequences of introducing into the DGP deviations from two of the basic assumptions underlying our statistical model, namely the assumptions of proportional hazards and of homogeneous causal parameters. These two assumptions are of course closely related, since heterogeneity in causal effects – e.g. such that the effect of spell duration depends on the value of the exogenous covariate x – represents a violation of the proportionality assumption. But, as long as parameter heterogeneity (and non-proportionality) is related to observed explanatory variables only, no new fundamental difficulties arise. As long as the correct model is specified – including the appropriate interaction terms – the true parameters will be recovered. We illustrate this point by modifying the DGP, such that subjects with low final exit propensity ($x=1$) are attributed positive duration dependence in the final destination hazard (Weibull baseline with shape parameter equal to 1.1), while subjects with high exit propensity ($x=0$) are attributed negative duration dependence (Weibull baseline with shape parameter equal to 0.9). As illustrated in Figure 8, when separate baselines are estimated for the two groups, we are still able to recover the true parameters (although the degree of statistical uncertainty obviously increases). This result holds true for other types of non-proportionalities as well.

Figure 8 around here

More serious problems arise if we take into account that the statistical model we use may represent a simplification of the true DGP, in the sense that there exist sources of non-proportionality that are not modelled. To illustrate, let us return to the issue of heterogeneity in duration dependence effects (according to the value of x), but this time assume that the researcher erroneously restricts the model to be fully proportional. Figure 9 illustrates the rather dismaying results obtained in this case. The upper

panel presents the estimated common duration parameters for the case discussed above, i.e. with positive duration dependence attributed to subjects with low unobserved exit propensity and negative duration dependence attributed to subjects with high unobserved exit propensity. The estimates are far off any conceivable 'compromise' between the two true baselines. The lower panel presents the estimation results for the case in which negative duration dependence is attributed to subjects with high exit propensity (and vice versa). The results are more promising in this case. But unfortunately, the general conclusion that we draw from this and other similar exercises, is that parameter heterogeneity in the DGP that is unaccounted for in the estimated model (either because it is unobserved or because the appropriate interaction term is not included in the model), produces results that have no convenient interpretation. The NPMLE of an assumed constant parameter that is really a random coefficient in the DGP does not necessarily represent any reasonable average of the underlying true parameters. The reason for this is of course that the parameter heterogeneity induces a source of unobserved heterogeneity that is not controlled for; and this heterogeneity entails a sorting effect of exactly the same kind as the sorting effect following from disregarding unobserved heterogeneity in the first place. Subjects with high parameter values leave the risk set first, leaving behind subjects with lower parameter values.

Figure 9 around here

A particularly interesting case to look at is that with heterogeneous treatment effects. Assume, for example, that the true treatment effects (α_1, α_2), rather than being the same for all subjects, are subject to some kind of probability distribution. It follows directly from the sorting argument referred to above that our estimators $(\hat{\alpha}_1, \hat{\alpha}_2)$ cannot be expected to represent average treatment effects in this case. Once subjects have entered into the treatment state, those with the highest effects exit first,

and a negatively selected group – in terms of treatment effects – is left behind. Hence, if the treatment effects are distributed independently of other variables in the model (including the two unobserved scalar variables v_e and v_p), the estimated effect will typically be negatively biased (compared to the true mean). A corollary of this argument is that, if the researcher allows the treatment effect to depend on time since entry or completion (which is in fact common practice in the treatment evaluation literature, see e.g. Van Ours, 2001; Richardson and Van den Berg, 2001; Lalive et al, 2002), the existence of effect heterogeneity will induce a negative duration bias in the estimated treatment effect. This reflects that it is difficult to distinguish empirically between heterogeneous (but constant) and common (but declining) treatment effects. Further complications arise if the distribution of treatment effects is not independently distributed from other sources of unobserved heterogeneity in the model.

Since we have already seen, in previous sections, that the first-order moment of an unobserved heterogeneity distribution can be reliably recovered from data, a natural solution to the problem of heterogeneous treatment effects is to model this heterogeneity explicitly. The treatment effects are then interpreted as state-specific contributions to the distribution of unobserved heterogeneity. We illustrate this point within a slightly simplified version of the baseline model, where we assume that the effects of ongoing and completed treatment are the same, i.e., it is only one treatment effect for each subject. Let $\alpha_i = \alpha_{1i} = \alpha_{2i}$ be the treatment effect for subject *i*. Furthermore, let α_i be independently distributed according to a normal distribution with mean 0.2 and variance 0.2. This means that, in this example, roughly two thirds of the subjects have positive treatment effects. The average treatment effect (ATE), as measured by the proportionality factor in the final destination hazard rate is $E[\exp(\alpha_i)] = \exp(0.2 + 0.1) = 1.35$, i.e., a 35 per cent increase in the hazard rate. Based

on these assumptions, we generate 100 new artificial datasets (containing 50,000 subjects each), and then estimate the parameters of the model. But this time, the treatment effect is not modelled as a parameter, but as a random coefficient, subject to some unknown joint probability distribution. Hence, the vector of unobserved covariates consists of three elements in this case, $v_i = (v_{ei}, v_{pi}, \alpha_i)$, and the parameters of this distribution is estimated in exactly the same way as described in Section 4. The only difference is that there are now three elements, rather than two, in each location vector.

The main results from this exercise, regarding the treatment effects, are described in Table 11. The table shows that the mean treatment effect (ATE) is consistently evaluated by the estimated parameters of the discrete mass-point distribution. A problem with this estimator, however, is that little is known about its sampling distribution. From our trials, we note that the standard deviation associated with the ATE-estimate of 1.361 (according to the ML criterion) is 0.088. The individual estimates ranged from a minimum of 1.135 to a maximum of 1.581; hence, based on our data, it seems that ATE is recovered in a fairly reliable fashion. However, in order facilitate statistical inference, more knowledge about the sampling distribution of heterogeneity parameters is required. It may also be noted that we are not able to recover the true variance of the treatment distribution. This is unsurprising, given our failure to recover second order moments of the heterogeneity distribution in previous sections.

Table 11 around here

10 Interval Censoring and Lost Subjects

So far, we have assumed that all spells belonging to the DGP under consideration are observed by the researcher, and that their starting times can be accurately measured. In practice, interval censoring usually means that some very short spells - those start-

ing and ending between two observation-points – are never recorded. This implies that the sample available to the researcher is selected. In particular, unobserved heterogeneity can no longer be assumed independent of either observed covariates or calendar time, since the impact of unobserved heterogeneity during the censored period – in terms of actual transitions - depends on the values of all other explanatory variables.

The problem can be assessed within the framework of our Monte Carlo experiments by assuming that all first-period records are unobserved. Hence, subjects are observed conditional on the spell lasting more than one period. We illustrate the consequences of such a sampling scheme by estimating a version of the baseline model (with 100,000 subjects to start with), under two alternative assumptions about the size of the sample selection problem. In the first example, the final destination hazard rates are scaled such that approximately 10 per cent of the subjects are lost due to exits in the first (unobserved) period of their spell. In the second example, as much as 20 per cent of the subjects are lost. We only make a single experiment for each of these DGP's, since this suffices for making our points. The two upper panels of Figure 10 illustrate what happens with the estimated duration parameters when the sample selection problem is disregarded, in the sense that the selected sample is treated as if it was un-selected. The NPML estimators then fail to remove the spurious negative duration dependence completely. Other parameters are also biased. For example, when 10 per cent of the spells are unobserved, the effect of the exogenous covariate x on the final destination hazard is estimated (according to the ML criterion) to –0.93 (0.02). When 20 per cent of the spells are unobserved, the estimate is –0.86 (0.02) (recall that the true value is –1).

The solution to this sample selection problem is to set up the likelihood function directly in terms of the true conditional probabilities. Let $L_i(v_i \mid d>1)$ be the likelihood contribution formed by subject *i*, conditional on survival during the first (censored) period and conditional on the vector of unobservables. In order to integrate out unobserved heterogeneity in this case, we need to take into account that it can no longer be assumed independent of other variables in the model (due to the sorting process that has already occurred). The conditional distribution of unobserved heterogeneity can be derived from Bayes' theorem. Let $f(v_i)$ be the joint density of v_i to start with (i.e. for the entire uncensored population). We can then write the conditional density as (we assume, for simplicity, that subjects exiting to the treatment state between two observation points are also lost)

$$f(v_i \mid d>1) = \frac{\exp\left(-\sum_{k\in K_{it}} \varphi_k(t,1,x_{it},z_{it},v_{ki})\right)}{\underset{v_i}{E}\left[\exp\left(-\sum_{k\in K_{it}} \varphi_k(t,1,x_{it},z_{it},v_{ki})\right)\right]} f(v_i)\,, \tag{6}$$

and the likelihood function takes the form

$$L \mid d>1 = \prod_{i=1}^{N} \underset{v_i}{E}\left[\frac{\exp\left(-\sum_{k\in K_{it}} \varphi_k(t,1,x_{it},z_{it},v_{ki})\right)}{\underset{v_i}{E}\left[\exp\left(-\sum_{k\in K_{it}} \varphi_k(t,1,x_{it},z_{it},v_{ki})\right)\right]} L_i\left(v_i \mid d>1\right)\right], \tag{7}$$

where $L_i\left(v_i \mid d>1\right)$ can be obtained from Equation (3). Hence, the solution to the interval left-censoring problem is to multiply the conditional likelihood contribution for each subject with the probability of being observed (conditional on *v*), and divide by the expected probability of being observed (with *v* integrated out). It is clear, however, that an additional assumption regarding the spell duration baseline is called for, since there is no foundation in the data for inferences about the first-period exit rate.

A natural assumption to make (in the absence of a parametrically specified baseline) is that the spell duration effect for the first period is equal to that of the second period (a similar assumption is required regarding the calendar time effects associated with the very first calendar period in the dataset). The two lower panels in Figure 10 illustrate what happens with the estimated duration dependence parameters when we maximise the likelihood function in (7). The negative bias is now completely removed. And the effects of other parameters are again also correctly recovered. For example, when 10 per cent of the spells are unobserved, the effect of x on the final destination hazard is now estimated to –1.02 (0.03). When 20 per cent of the spells are unobserved, the estimate is –0.98 (0.02).

In practice, the researcher may not have exact information about the duration a sampled subject has been at risk at the time of sampling, since it may have entered into the origin state at any time between the two observation points. In this case, additional assumptions are required regarding the distribution of the flow into and out of the origin state during the censored time interval. In the absence of additional knowledge, the most natural assumption to make is that entrances to the origin state are uniformly distributed over the censored interval, and that the hazard rates are constant within the same interval. We can then write the probability of survival to the first observation point after entry as

$$\begin{aligned} prob(\sum_k y_{kit} = 0 \mid d = 1, x_{it}, v_i) &= \int_0^1 \exp(-(1-s)\sum_k \varphi(t,1,x_{it},0,v_i)ds \\ &= \frac{1-\exp\left(-\sum_k \varphi(t,1,x_{it},0,v_i)\right)}{\sum_k \varphi(t,1,x_{it},0,v_i)} \end{aligned} \tag{8}$$

11 Conclusion

Based on comprehensive Monte Carlo experiments, we conclude that, for a correctly specified model, the Non-Parametric Maximum Likelihood Estimator (NPMLE) robustly recovers the true treatment effects from non-experimental event history data, even when there are large unobserved sorting problems involved. We also find that the degree of duration dependence can be recovered, without parametric restrictions on either duration dependence or unobserved heterogeneity. Our results are encouraging compared to previous studies, and suggest that event history analysis may represent a powerful tool for solving the difficult problem of disentangling causality from sorting, based on non-experimental data. It is a well-known fact that the Mixed Proportional Hazard (MPH) model is non-parametrically identified, provided that a relevant exogenous variable exits. We have shown in this paper that a subject-specific exogenous covariate is not required if there exists a calendar-time source of variation in hazard rates (as long as calendar time is not perfectly correlated to process time).

We have also demonstrated that the NPML estimator is fragile towards unjustified restrictions, and, in particular, that any non-modelled sources of unobserved heterogeneity (e.g. in the form of random slope parameters) may produce substantial bias in causal parameters. We emphasise in particular, the following:

1. It is essential that the number of support points in the unobserved heterogeneity distribution is selected according to an appropriate information criterion. A pre-specified (low) number of support points may result in substantial bias, particularly with respect to the estimated duration dependence parameters.
2. The most reliable information criterion is the likelihood itself, or the likelihood supplemented by a weak penalty for parameter abundance (such as the

Akaike Information Criterion). With small samples, a stronger penalty may be required (e.g. the Hannan-Quinn Information Criterion).

3. It is not the case that a flexible (non-parametric) baseline hazard is sufficient for ensuring that uncontrolled heterogeneity does not bias parameter estimates attached to exogenous covariates. The typical situation is that an error in the estimation of one parameter (or one set of parameters) contaminates other parameters as well.
4. The individual parameters of the estimated discrete unobserved heterogeneity distribution, are estimated with enormous statistical uncertainty and have no convenient interpretation. The only robustly estimated property of this distribution is its mean.
5. For parameters reflecting the influence of observed covariates, it is the case that the standard errors calculated conditional on the given number of (optimally chosen) support points in the heterogeneity distribution, also reflect the unconditional statistical uncertainty.
6. Sample selection caused by interval censoring (the failure to sample spells that start and stop between two observation points) may cause substantial bias in all parameter estimates. This problem can be solved by setting up the likelihood function in terms of the appropriate conditional distribution of unobserved heterogeneity.
7. Deviations from the proportional hazards assumption are not problematic, as long as these deviations are accounted for in the formulation of the model.
8. It is possible to consistently recover the mean of a heterogeneous treatment effect distribution, by means of modelling the treatment effects as subject-

specific unobserved covariates. However, little is known about the sampling distribution of the estimator.

9. Deviations from the proportionality assumption that are unaccounted for in the model may cause substantial bias in all parameter estimates.

The latter of these points constitutes a rather serious challenge for event history analysis in social (non-experimental) sciences, and suggests, unfortunately, that results gathered by means of this statistical technique can rarely be considered definitive. In practice, it is typically impossible for the researcher to take all potential interaction effects and all potential sources of parameter heterogeneity into account. Most statistical models represent simplifications of the true DGP rather than an exact representation. Hence, the risk of estimating a wrongly specified model is acute. This also implies that robustness should always be considered a key concern in the assessment of results based on NPMLE.

References

Abbring, J. H. and Van den Berg, G. J. (2003a) The Non-Parametric Identification of Treatment Effects in Duration Models. *Econometrica*, Vol. 71, No. 5 (September), 1491-1517.

Abbring, J. H. and Van den Berg, G. J. (2003b) The Identifiability of the Mixed Proportional Hazards Competing Risks Model. *Journal of Royal Statistical Society, Series B*, Vol. 65, No. 3, 701-710.

Arulampalam, W. and Stewart, M. B. (1995). The determinants of individual unemployment durations in an era of high unemployment. *Economic Journal*, Vol. 105, 321-332.

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

Brinch, C. (2000) Identification of structural duration dependence and unobserved heterogeneity with time-varying covariates'. Memorandum No. 20/2000, Department of Economics, University of Oslo.

Elbers, C. and Ridder, G. (1982) True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model. *Review of Economic Studies*, Vol. 64, 403-409.

Gaure, S. and Røed, K. (2003) How Tight is the Labour Market? A Micro-Based Macro Indicator. Memorandum No. 9/2003, Department of Economics, University of Oslo.

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.

Heckman, J. J. and Honoré, B. E. (1989) The Identifiability of the Competing Risks Model. *Biometrika*, Vol. 76, 325-330.

Heckman, J. J., Lalonde, R. J. and Smith, J. A. (1999) The economics and econometrics of active labor market programs. In O. Ashenfelter and D. Card (Eds.) Handbook of Labor Economics, Vol. 3a, North-Holland.

Lalive, R., Van Ours, J. C. and Zweimüller, J. (2002) The Impact of Active Labor Market Programs on the Duration of Unemployment. Institute for Empirical Research in Economics, University of Zurich, Working paper No. 41.

Leroux, B. G. (1992) Consistent Estimation of a Mixing Distribution. *The Annals of Statistics*, Vol. 20, 1350-1360.

Lindsay, B. G. (1983) The Geometry of Mixture Likelihoods: A General Theory. *The*

Annals of Statistics, Vol. 11, 86-94.

McCall, B. P. (1994) Identifying State Dependence in Duration Models. American Statistical Association 1994, Proceedings of the Business and Economics Section, 14-17.

Narendranathan, W. and Stewart, M. B. (1993) Modelling the Probability of Leaving Unemployment: Competing Risks Models with Flexible Base-line Hazards. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, Vol. 42, 63-83.

Richardson, K. and Van den Berg, G. J. (2001) The Effect of Vocational Employment Training on the Individual Transition Rate from Unemployment to Work. *Swedish Economic Policy Review*, Vol. 8, 175-213.

Røed, K. and Raaum, O. (2003a) The Effect of Programme Participation on the Transition Rate from Unemployment to Employment. Memorandum No. 13/2003, Department of Economics, University of Oslo.

Røed, K. and Raaum, O. (2003b) 1. Administrative Registers – Unexplored Reservoirs of Scientific Knowledge? *Economic Journal*, Vol. 113. (2003), F258-F281.

Van den Berg, G. J. and Van Ours, J. C. (1994). Unemployment Dynamics and Duration Dependence in France, The Netherlands and the United Kingdom. *Economic Journal*, Vol. 104, 432-443.

Van den Berg, G. J. and Van Ours, J. C. (1996). Unemployment Dynamics and Duration Dependence. *Journal of Labor Economics*, Vol. 14, 100-125.

Van Ours, J. (2001) Do Active Labor Market Policies Help Unemployed Workers to Find and Keep Regular Jobs? In: Michael Lechner and Friedhelm Pfeiffer (Eds.), Econometric Evaluation of Labour Market Policies, Physica-Verlag,

125-152.

Zhang, T. (2003) A Monte Carlo Study on Non-Parametric Estimation of Duration Models with Unobserved Heterogeneity. Memorandum No. 25/2003, Department of Economics, University of Oslo.

Table 1
Properties of the baseline DGP

Sample size	50,000 subjects
Data window size (number of periods observed)	40 periods
Entrance into origin state	Randomly distributed over the 40 periods (with probability 1/40 for each period)
Observed covariate	Subjects are randomly attributed x=1 with a probability of 0.5, otherwise x=0. The covariate has a negative effect on the final destination hazard, and a positive effect on the treatment hazard, such that $\beta_e = -1, \beta_p = 1$
Calendar time effects	For each of the 40 periods, the parameters σ_{et} and σ_{pt} are independently distributed drawings from the standard normal distribution.
Spell duration effects	There are no spell duration effects, i.e. $\lambda_{ed} = \lambda_{pd} = 0 \forall d$
Duration of treatment	The treatment lasts for five periods (unless a transition to the final destination occurs). Thereafter, the subjects return to the origin state.
Treatment effects	There are no treatment effects, i.e. $\alpha = (0, 0)$.
Unobserved heterogeneity	The vector of unobserved covariates (v_e, v_p) is distributed according to a bivariate normal distribution with means (c_e, c_p), variances (1,1) and correlation coefficient 0.5. The means (c_e, c_p) are normalised such that, when *x* is zero and the calendar time effect is zero, the transition probabilities are equal to 0.1 (to final destination) and 0.05 (to treatment).
Transitions	Transition probabilities are calculated from Equation (2). Actual transitions are generated by comparing the transition probabilities with random drawings from a uniform distribution on [0,1].

Table 2
Descriptive Summary Statistics for the 100 Data Sets Generated by the Baseline DGP

	Mean	Minimum	Maximum
Average spell duration	9.84	3.20	17.28
Fraction subject to treatment	0.47	0.04	0.87
Average duration until treatment (conditional on treatment)	9.49	3.71	15.13
Fraction censored	0.29	0.04	0.72

Table 3
The distribution of the required number of support points according to Maximised Penalised Likelihood and Maximum likelihood model selection criteria (100 trials)

Required # support points	BIC	HQIC	AIC	ML
3	2	0	0	0
4	16	3	1	0
5	36	16	0	0
6	32	16	10	0
7	14	25	15	1
8	0	31	27	6
9	0	9	23	5
10	0	0	13	17
11	0	0	7	17
12	0	0	2	21
13	0	0	1	13
14	0	0	1	11
15	0	0	0	6
16	0	0	0	1
17	0	0	0	0
18	0	0	0	1
19	0	0	0	1
Average # support points	5.4	6.9	8.5	11.7

Table 4
Estimated Effects of Exogenous Covariate and Endogenous Treatment
Results from 100 trials based on the baseline DGP

		Without control for unobserved heterogeneity			BIC			HQIC			AIC			ML		
	True value	Mean Est.	Mean S.E.	Reject at 5%	Mean Est.	Mean S.E.	Reject at 5%	Mean Est.	Mean S.E.	Reject at 5%	Mean Est.	Mean S.E.	Reject at 5%	Mean Est.	Mean S.E.	Reject at 5%
β_e	**-1**	-0.788	0.012	100	-0.932	0.020	75	-0.972	0.023	30	-0.992	0.025	13	-1.007	0.027	8
β_p	**1**	0.907	0.012	100	0.987	0.020	16	0.993	0.020	9	0.996	0.020	6	0.999	0.021	5
α_1	**0**	0.443	0.017	100	-0.009	0.032	18	-0.006	0.034	8	-0.006	0.035	8	-0.003	0.037	4
α_2	**0**	0.329	0.025	100	-0.014	0.037	19	-0.010	0.039	15	-0.008	0.041	6	-0.004	0.042	6
$\lambda_{ed}, \forall d$				100			45			20			11			8
$\lambda_{pd}, \forall d$				100			5			5			5			5

Note: The "Reject at 5% column contains the per cent of the replications that led to models for which the null hypothesis corresponding to the true parameter value was rejected at the five per cent nominal significance level.

Table 5
The lower order moments of the estimated heterogeneity distribution
Results from 100 trials based on the baseline DGP

	DGP	BIC		HQIC		AIC		ML	
		Mean Est.	St. Dev.	Mean Est.	St. Dev.	Mean Est.	St. Dev.	Mean Est.	St. Dev.
Mean $\exp(v_e)$	0.177	0.167	0.012	0.174	0.012	0.181	0.015	0.188	0.014
Mean $\exp(v_p)$	0.089	0.088	0.011	0.089	0.009	0.093	0.010	0.096	0.011
Var $\exp(v_e)$	0.054	0.030	0.061	0.049	0.065	0.090	0.103	0.127	0.095
Var $\exp(v_p)$	0.014	0.021	0.079	0.028	0.067	0.046	0.080	0.061	0.084
Corr($\exp(v_e)$, $\exp(v_p)$)	0.399	0.0596	0.0202	0.482	0.195	0.476	0.220	0.423	0.226

Note: The constant terms c_e and c_p are included in the heterogeneity distributions.

Table 6
Mean errors (estimated minus true) of estimated parameters under alternative sample sizes

# subjects	5,000				10,000				50,000				500,000				5,000,000			
# samples	1,000				500				100				10				1			
	BIC	HQIC	AIC	ML	BIC	HQIC	AIC	ML	BIC	HQIC	AIC	ML	BIC	HQIC	AIC	ML	BIC	HQIC	AIC	ML
Mean W	3.1	4.3	6.0	10.1	3.6	5.0	6.7	10.7	5.4	6.9	8.5	11.7	8.4	9.6	11.3	13.4	12	12	14	16
β_e	0.167	0.090	0.002	-0.099	0.151	0.075	0.013	-0.040	0.068	0.029	0.008	-0.006	0.006	0.000	-0.002	-0.003	0.003	0.003	0.002	0.002
β_p	-0.046	-0.029	-0.015	0.001	-0.029	-0.016	-0.006	0.004	0.013	0.007	-0.004	-0.001	0.000	0.002	0.003	0.003	0.000	0.000	0.000	0.000
α_1	-0.036	-0.044	-0.044	-0.032	-0.020	-0.025	-0.023	-0.016	-0.009	-0.007	-0.006	-0.003	0.004	0.004	0.006	0.006	-0.002	-0.002	-0.002	-0.001
α_2	-0.045	-0.050	-0.047	-0.030	-0.031	-0.036	-0.029	-0.019	-0.014	-0.010	-0.008	-0.004	0.003	0.004	0.006	0.006	-0.002	-0.002	-0.002	-0.002
$\lambda_{ed}, \forall d$	-0.467	-0.207	0.070	0.383	-0.452	-0.188	0.001	0.173	-0.206	-0.086	-0.017	0.029	-0.049	-0.031	-0.024	-0.019	-0.007	-0.007	-0.003	-0.003
$\lambda_{pd}, \forall d$	0.087	0.088	0.085	0.080	0.052	0.046	0.042	0.039	0.009	0.010	0.010	0.008	0.000	0.000	0.000	0.000	0.005	0.005	0.005	0.005
$E[expv_e]$	-0.024	-0.009	0.014	0.041	-0.019	-0.005	0.012	0.027	-0.010	-0.003	0.004	0.011	-0.002	0.000	0.004	0.006	0.000	0.000	0.000	0.000
$E[expv_p]$	-0.003	0.004	0.013	0.021	-0.002	0.004	0.010	0.014	-0.001	0.001	0.004	0.007	-0.001	-0.001	0.001	0.001	-0.001	-0.001	0.000	0.000
$V[expv_e]$	0.000	0.057	0.161	0.305	-0.016	0.029	0.100	0.174	-0.024	-0.006	0.036	0.073	-0.048	-0.042	0.025	0.035	-0.010	-0.010	-0.007	-0.007
$V[expv_p]$	0.018	0.061	0.120	0.165	0.012	0.048	0.082	0.106	0.007	0.014	0.032	0.047	-0.005	0.000	0.011	0.009	-0.002	-0.002	-0.002	-0.002
Corr.	0.475	0.285	0.109	-0.025	0.412	0.218	0.102	-0.017	0.196	0.083	0.077	0.024	0.055	0.041	0.063	-0.037	-0.028	-0.028	-0.021	-0.033

Table 7
Estimated Effects of Endogenous Treatment
Results from 10 trials based on the baseline DGP, with all subject specific exogenous characteristics unobserved

		Without control for unobserved heterogeneity		AIC		ML	
	True value	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.
α_1	0	0.193	0.014	-0.000	0.041	0.001	0.042
α_2	0	0.123	0.021	-0.012	0.047	-0.009	0.048

Table 8
Estimated Effects of Exogenous Covariate and Endogenous Treatment
Results from 10 trials based on the baseline model with different modified heterogeneity distributions

	True value	Without control for unobserved heterogeneity		AIC		ML	
		Mean Est.	Mean S.E.	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.
Model i) Perfectly correlated discrete with five equally likely support points at $(-1, -\frac{1}{2}, 0, \frac{1}{2}, 1)$ plus bivariate normal drawing (as in baseline model)							
Average number of support points				8.6		14	
β_e	-1	-0.737	0.010	-0.980	0.024	-0.993	0.025
β_p	1	0.968	0.018	1.007	0.024	1.010	0.023
α_1	0	0.660	0.015	0.031	0.034	0.032	0.035
α_2	0	0.509	0.023	0.023	0.039	0.028	0.041
Model ii) Independent discrete with five equally likely support points (as in model i), but with independent drawings for the two unobservables) and bivariate normal							
Average number of support points				10.3		14.0	
β_e	-1	-0.672	0.011	-0.989	0.025	-0.999	0.026
β_p	1	0.860	0.011	1.018	0.022	1.019	0.022
α_1	0	0.353	0.014	0.015	0.034	0.014	0.035
α_2	0	0.261	0.022	0.008	0.041	0.008	0.040
Model iii) Independent Gamma and perfectly negatively correlated discrete							
Average number of support points				9.7		12.2	
β_e	-1	-0.588	0.011	-0.999	0.026	-1.008	0.027
β_p	1	0.748	0.010	1.008	0.020	1.008	0.020
α_1	0	-0.182	0.016	-0.009	0.038	-0.014	0.038
α_2	0	-0.120	0.022	0.002	0.042	-0.002	0.043
Model iv) Truncated bivariate normal. Based on the baseline model, but the five upper percentiles in the v_e-distribution are deleted from the dataset.							
Average number of support points				7.9		11.2	
β_e	-1	-0.794	0.011	-1.021	0.022	-1.031	0.023
β_p	1	0.941	0.013	0.990	0.020	0.991	0.021
α_1	0	0.374	0.014	-0.014	0.031	-0.013	0.032
α_2	0	0.312	0.020	-0.015	0.035	-0.014	0.037
Model v) Discrete with 7 (v_e, v_p) support points at (-100, 0.5), (-1, 0.5), (-0.5, 1), (0, 0), (0.5, -1), (1, -0.5), and (0.5, -100); the first point with a probability of 0.05, the last point with 0.01 and the others with a probability of 0.188.							
Average number of support points				5.6		7.9	
β_e	-1	-0.618	0.011	-1.003	0.021	-1.013	0.022
β_p	1	0.817	0.012	1.003	0.015	1.002	0.015
α_1	0	-0.404	0.015	-0.000	0.027	-0.011	0.029
α_2	0	-0.368	0.021	0.003	0.032	-0.008	0.033

Table 9
The Role of Exogenous Calendar Time Variation
Estimates based ML criterion

	$Var\ \sigma_{et}=1$ $Var\ \sigma_{pt}=1$		$Var\ \sigma_{et}=0.25$ $Var\ \sigma_{pt}=0.25$		$Var\ \sigma_{et}=0.01$ $Var\ \sigma_{pt}=0.01$		$Var\ \sigma_{et}=0$ $Var\ \sigma_{pt}=0$		Random walk*	
	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.
$\alpha_1=0$	0.001	0.042	0.024	0.037	0.037	0.051	0.024	0.051	-0.024	0.037
$\alpha_2=0$	-0.009	0.048	0.009	0.035	0.035	0.053	0.009	0.053	-0.036	0.044
	WMAE		WMAE		WMAE		WMAE		WMAE	
Dur. eff. fin. dest.	0.104		0.190		0.555		0.603		0.155	
Dur. eff. treatment	0.046		0.040		0.041		0.035		0.048	

- In the random walk model calendar time effects are generated as $\sigma_{kt}=\sigma_{kt-1}+\varepsilon_{kt}$, where ε_{kt} is standard normal with variance 0.25
-

Table 10
Estimated Effects of Treatment
Results from 10 trials with baseline model modified to contain positive or negative treatment effects

	True value	Without control for unobserved heterogeneity		AIC		ML	
		Mean Est.	Mean S.E.	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.
Positive effects							
Average number of support points				9.0		12.9	
α_1	0.2	0.559	0.014	0.200	0.031	0.199	0.032
α_2	0.2	0.473	0.021	0.212	0.037	0.213	0.038
Negative effects							
Average number of support points				8.8		12.0	
α_1	-0.2	0.247	0.015	-0.204	0.032	-0.204	0.033
α_2	-0.2	0.179	0.021	-0.199	0.037	-0.200	0.038
Negative on-treatment effects, positive post-treatment effects							
Average number of support points				8.5		11.3	
α_1	-0.2	0.235	0.015	-0.210	0.032	-0.210	0.032
α_2	0.2	0.519	0.020	0.169	0.035	0.170	0.036

Table 11
Estimated mean and variance of the unobserved treatment effect distribution.
Results from 100 trials based on a modified baseline model with heterogeneous treatment effects

	DGP	AIC		ML	
Mean # of support points		7.1		11.1	
		Mean Est.	St. Dev.	Mean Est.	St. Dev.
$E[\exp(\alpha_i)]$	1.351	1.388	0.099	1.361	0.088
$Var[\exp(\alpha_i)]$	0.406	0.538	0.197	0.553	0.189

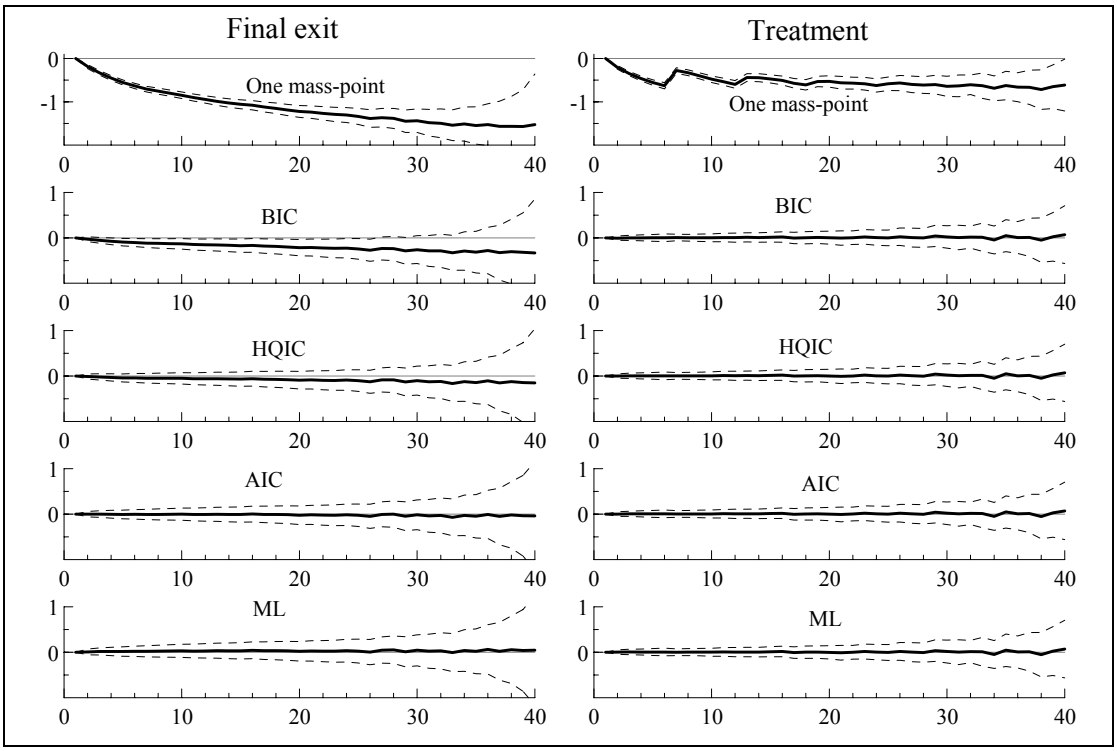

Figure 1. Average estimated effects of spell duration (point estimates with 95% confidence intervals, based on observed standard deviation from the 100 trials). The true effects are equal to zero for all durations.

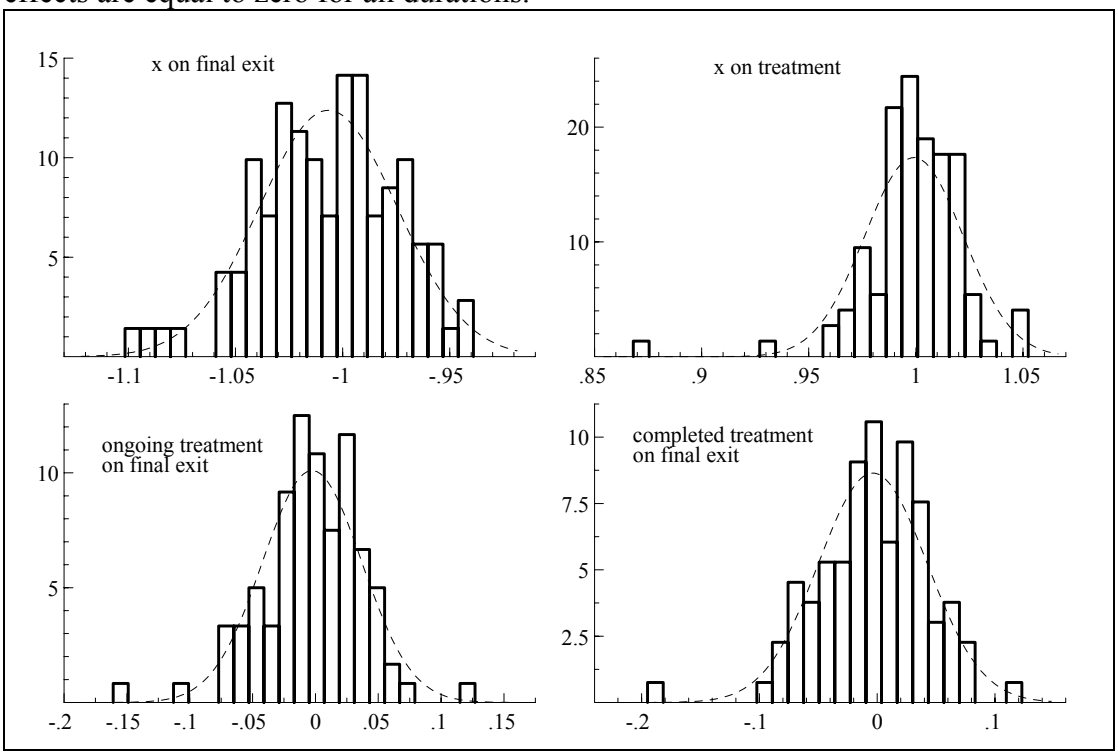

Figure 2. Distribution of the estimates of the four structural parameters, based on the ML criterion, and normal densities (with the same mean and standard deviation)
Note: Each histogram contains 25 bars.

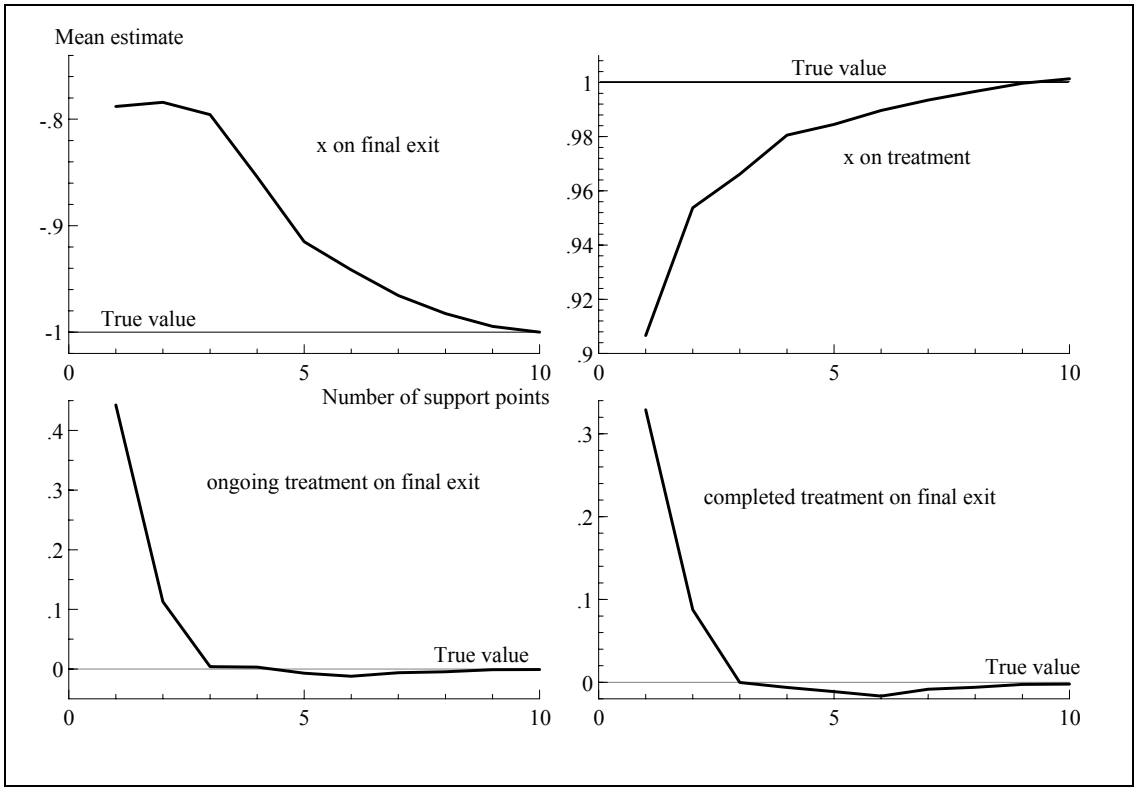

Figure 3. Mean estimates (over 100 trials) of the four structural parameters as functions of the number of support points in the unobserved heterogeneity distribution (1 support points corresponds to a model without unobserved heterogeneity).

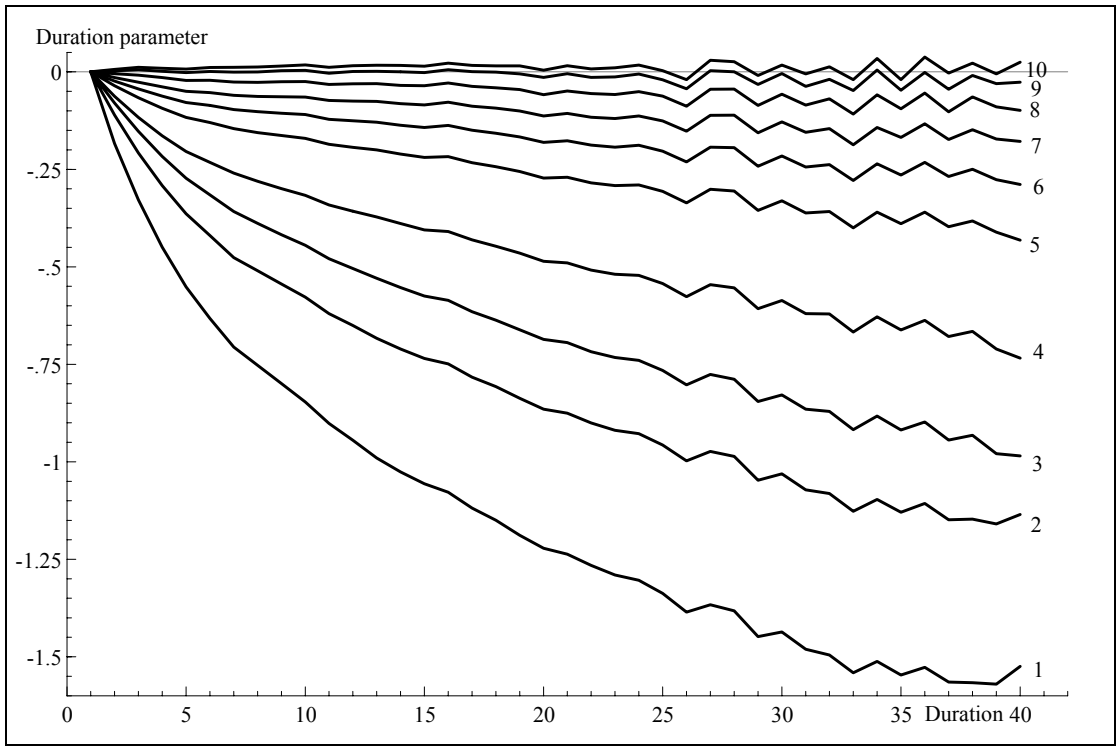

Figure 4. Average estimated duration parameters in the final destination hazard, with from 1 to 10 support points in the unobserved heterogeneity distribution. The true parameters are all equal to zero.

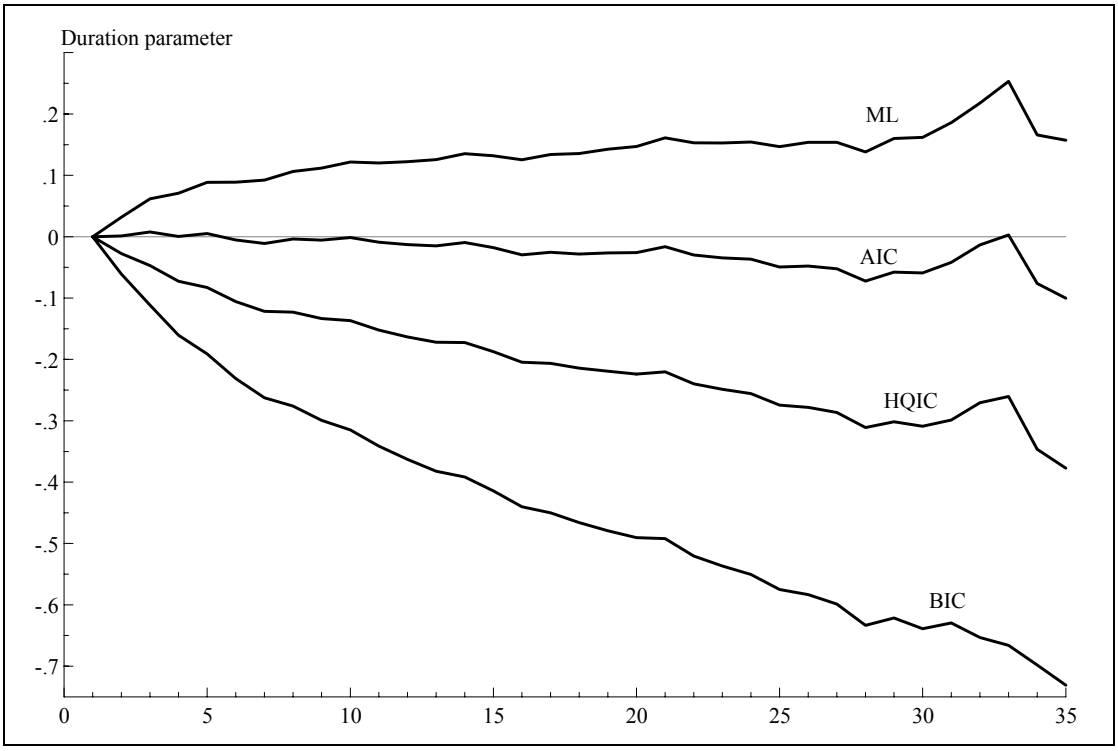

Figure 5. Average estimated duration parameters in the final destination hazard, based on 500 samples with 10,000 subjects in each sample.

Note: We only report estimates associated with the first 35 periods, since the number of observations of durations above 35 periods in each sample is too small to obtain sensible estimates.

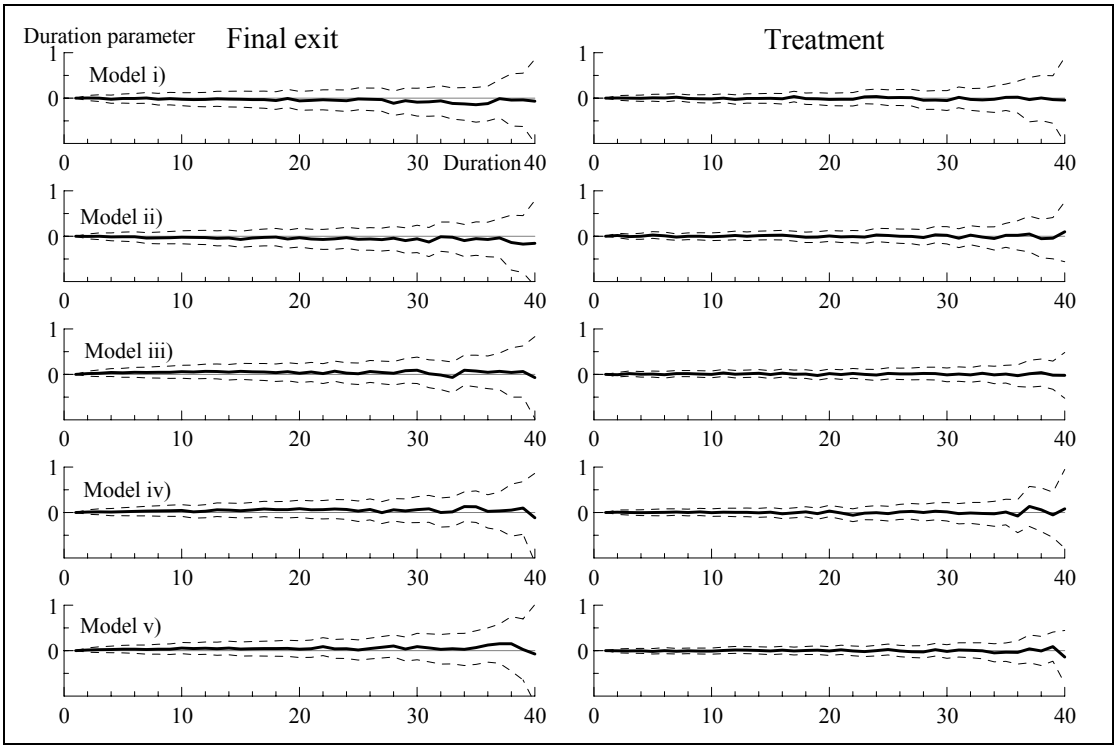

Figure 6. Average estimated effects of spell duration (with 95 per cent confidence intervals), according to the Maximum Likelihood criterion (based on average point estimates and standard errors over 10 trials for each model). The true effects are equal to zero for all durations.

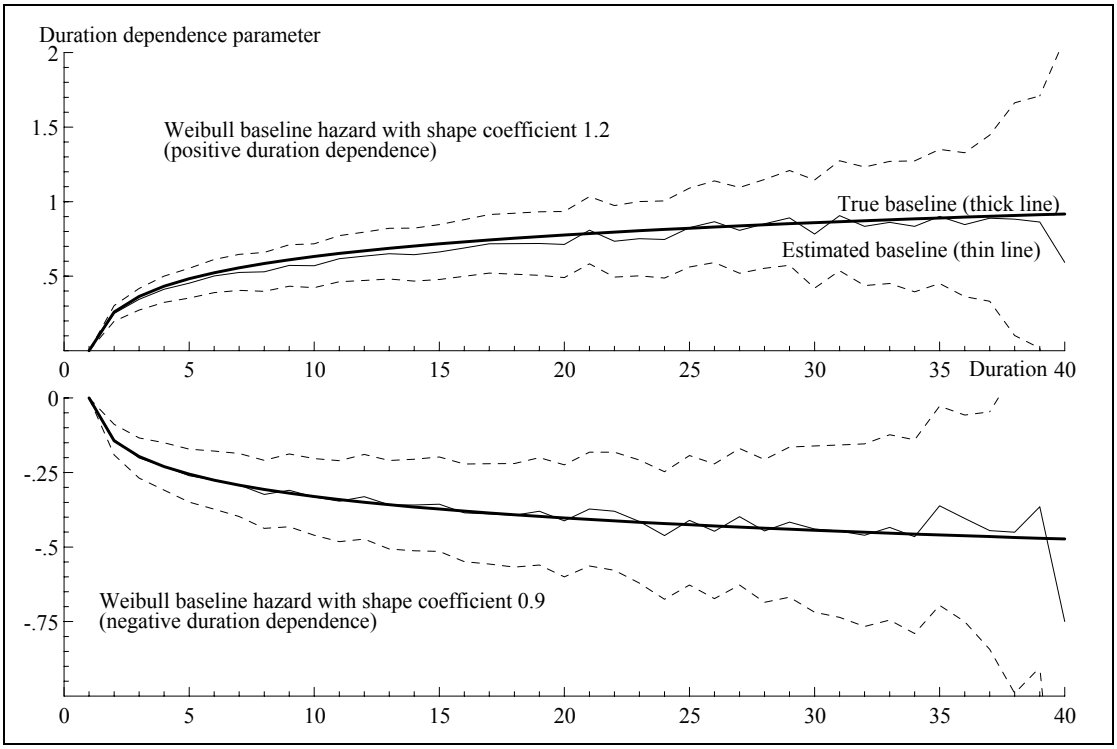

Figure 7. Estimated duration dependence parameters according to the Maximum Likelihood criterion (with 95 per cent confidence intervals) in final destination hazard when the true baseline exhibits positive or negative duration dependence (average based on 10 trials)

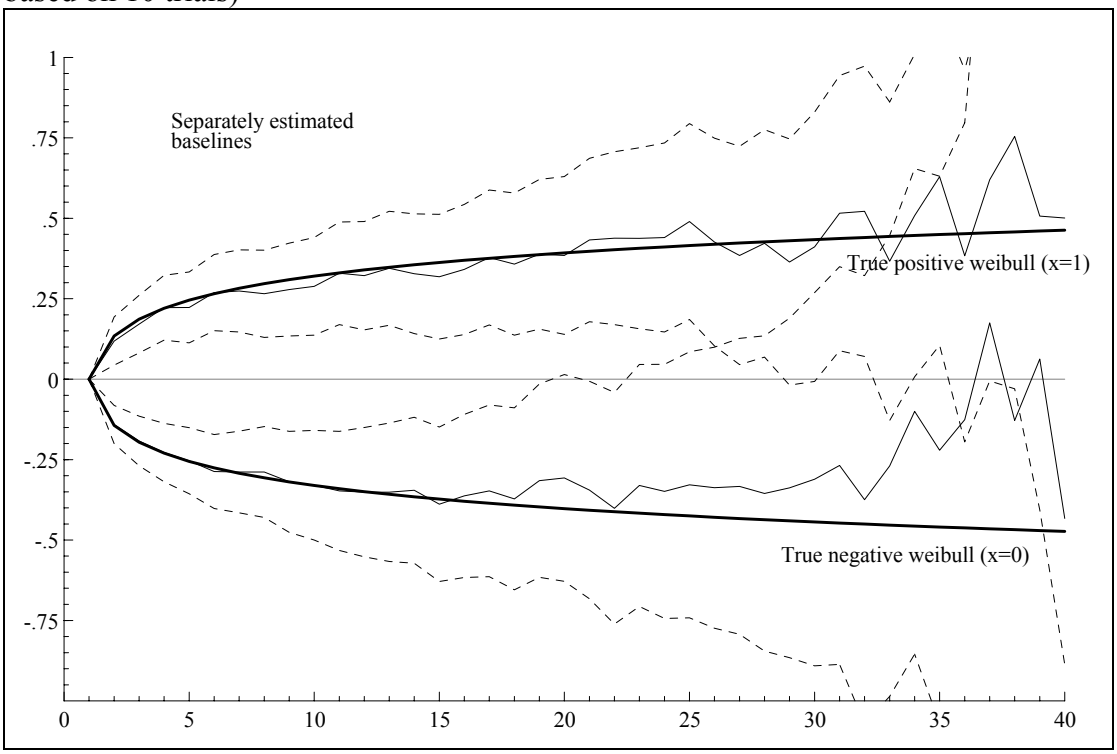

Figure 8. Estimated group-specific duration dependence parameters according to the Maximum Likelihood criterion (with 95 per cent confidence intervals) in final destination hazard when the baseline exhibits positive duration dependence for *x=1* and negative duration dependence when *x=0* (average based on 10 trials)

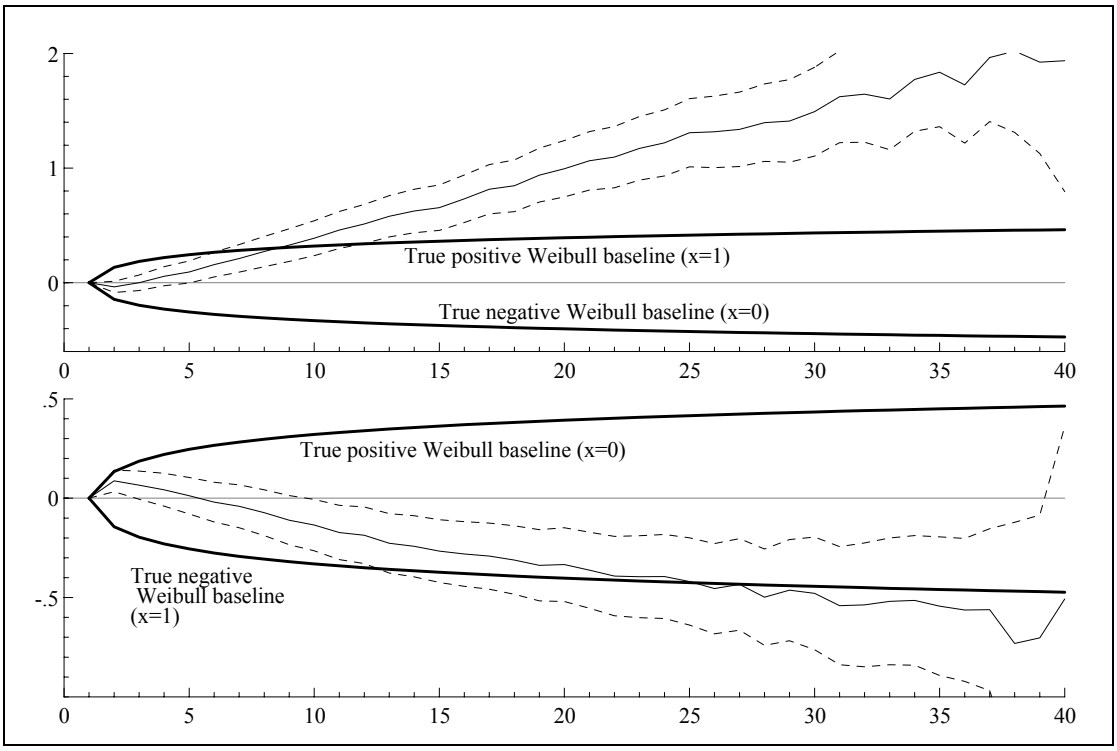

Figure 9. Estimated common duration dependence parameters according to the Maximum Likelihood criterion (with 95 per cent confidence intervals) in final destination hazard when the true baseline exhibits positive duration dependence for *x=1* and negative duration dependence when *x=0* (upper panel) and when the true baseline exhibits positive duration dependence for *x=0* and negative duration dependence when *x=1* (lower panel) (average based on 10 trials).

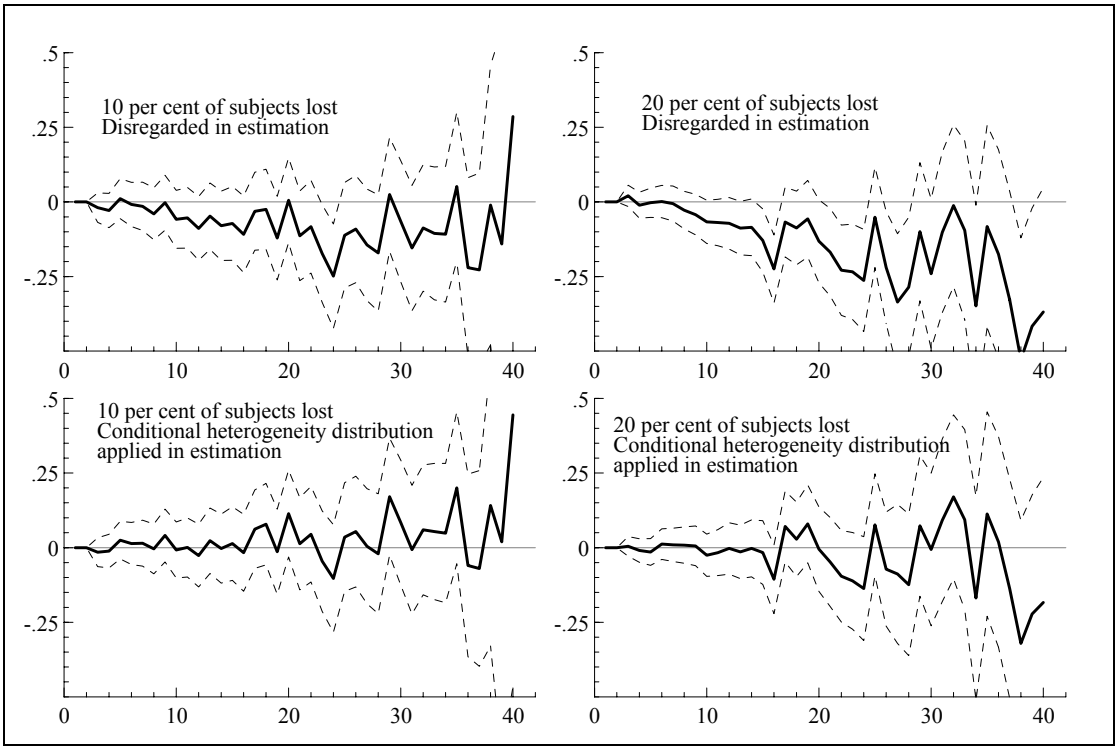

Figure 10. Estimated duration dependence parameters with (lower panels) and without (upper panels) correction for sample selectivity (Maximum Likelihood criterion, with 95 per cent confidence intervals).

Note: The DGP is a baseline model with 100,000 subjects to start with.