Gaure, Simen; Røed, Knut; Zhang, Tao

Working Paper

Time and causility: A Monte Carlo assessment of the timing-of-events approach

Memorandum, No. 2005,19

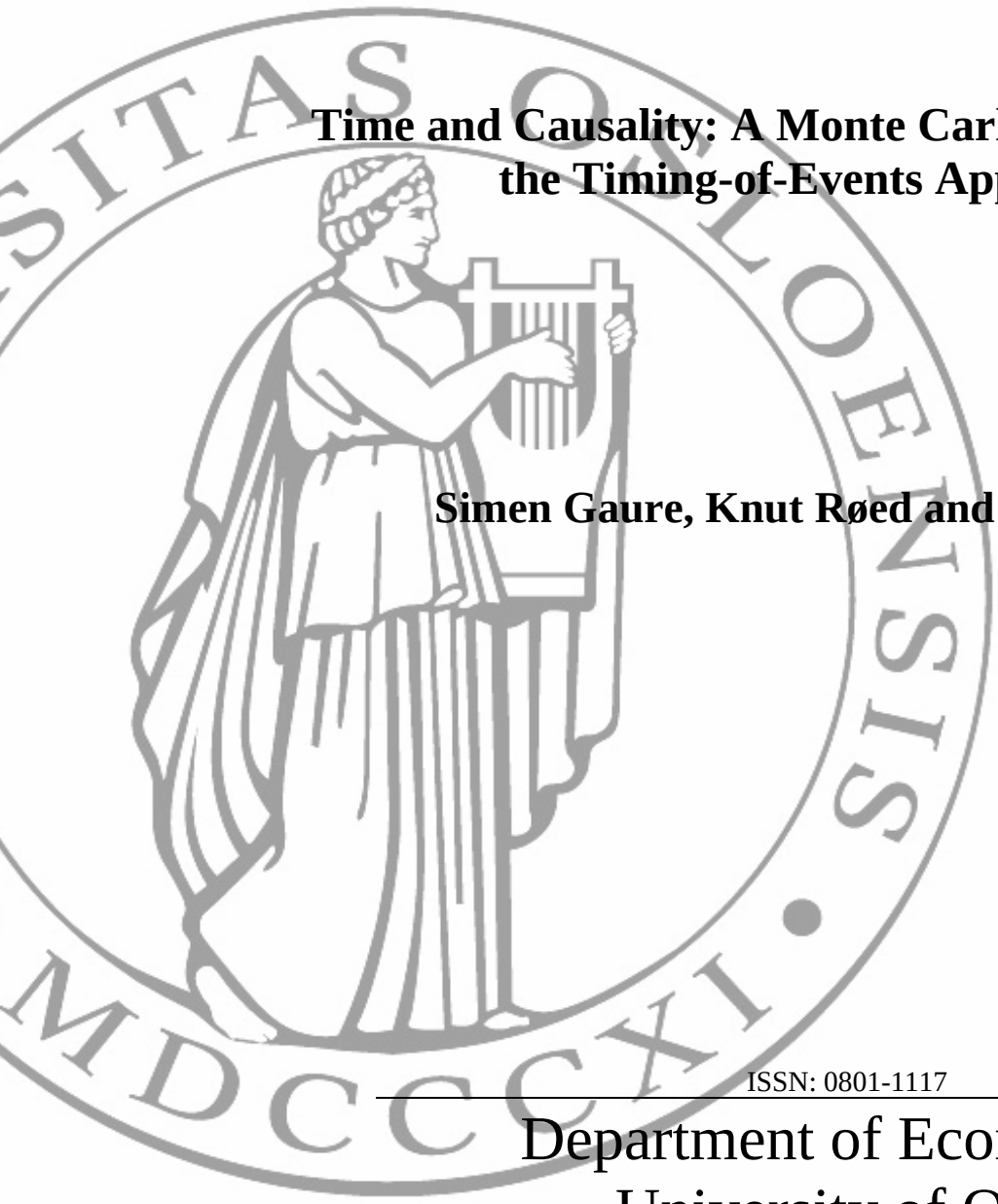
Provided in Cooperation with:
Department of Economics, University of Oslo

Suggested Citation: Gaure, Simen; Røed, Knut; Zhang, Tao (2005) : Time and causility: A Monte Carlo assessment of the timing-of-events approach, Memorandum, No. 2005,19, University of Oslo, Department of Economics, Oslo

This Version is available at:
https://hdl.handle.net/10419/63034

MEMORANDUM

No 19/2005



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

Simen Gaure, Knut Røed and Tao Zhang

ISSN: 0801-1117

Department of Economics
University of Oslo

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.oekonomi.uio.no/
e-mail: econdep@econ.uio.no

In co-operation with
The Frisch Centre for Economic Research

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no/
e-mail: frisch@frisch.uio.no

List of the last 10 Memoranda:

No 18	Tyra Ekhaugen Immigrants on Welfare: Assimilation and Benefit Substitution. 35 pp
No 17	Oddbjørn Raaum, Jon Rogstad, Knut Røed and Lars Westlie Young and Out: An Application of a Prospects-Based Concept of Social Exclusion. 49 pp.
No 16	Michael Hoel Prioritizing public health expenditures when there is a private alternative. 20 pp.
No 15	Hans Jarle Kind, Tore Nilssen and Lars Sørgard Advertising on TV: Under- or Overprovision?. 20 pp.
No 14	Olav Bjerkholt Markets, models and planning: the Norwegian experience. 34 pp.
No 13	Diderik Lund An analytical model of required returns to equity under taxation with imperfect loss offset. 42 pp.
No 12	Hilde C. Bjørnland and Kai Leitemo Identifying the Interdependence between US Monetary Policy and the Stock Market. 34 pp.
No 11	Kari Due-Andresen Tax evasion and labour supply in Norway in 2003: Structural models versus flexible functional form models. 44 pp.
No 10	Steinar Holden and Fredrik Wulfsberg Downward Nominal Wage Rigidity in the OECD. 39 pp.
No 09	Kjell Arne Brekke, Karine Nyborg and Mari Rege The Fear of Exclusion: Individual Effort when Group Formation is Endogenous. 23 pp.

A complete list of this memo-series is available in a PDF® format at:
http://www.oekonomi.uio.no/memo/

2 August 2005

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 Monte Carlo evidence regarding the feasibility of separating causality from selection within non-experimental interval-censored duration data, by means of the nonparametric maximum likelihood estimator (NPMLE). Key findings are: i) the NPMLE is extremely reliable, and it accurately separates the causal effects of treatment and duration dependence from sorting effects, almost regardless of the true unobserved heterogeneity distribution; ii) the NPMLE is normally distributed, and standard errors can be computed directly from the optimally selected model; and iii) unjustified restrictions on the heterogeneity distribution, e.g., in terms of a pre-specified number of support points, may cause substantial bias.

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. The project has also received support from the Research Council's Programme for High Performance Computing, through a grant of computing time. Thanks to Rolf Aaberge, John K. Dagsvik, Christian Göbel, and Angelo Melino for valuable comments and discussions. Correspondence to: Knut Røed, the Ragnar Frisch Centre for Economic Research, Gaustadalléen 21, 0349 Oslo, Norway. E-mail: knut.roed@frisch.uio.no.

1 Introduction

The unit of analysis in this paper is a subject entering into some state (the origin state), and its subsequent stochastic transition to 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 causal effects: The effect of a treatment, and the effect of spell duration. Identification problems arise when relevant subject heterogeneity is not fully controlled for, either because it is unobserved or because it fails to obey parametric restrictions imposed by the researcher. The distribution of uncontrolled heterogeneity obviously changes with the time spent in the origin state, and to the extent that treatment assignment is not fully randomised, it also 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 realistically designed non-experimental data that embody unobserved sorting processes, and also incorporate the ubiquitous problem of interval censoring. In particular, we provide an extensive Monte Carlo assessment of the nonparametric maximum likelihood estimator (NPMLE). The key idea behind this estimator is to approximate the unknown distribution of unobserved heterogeneity by means of a discrete distribution, with the number of support points selected such that the appropriate likelihood function is maximised (Lindsay, 1983; Heckman and Singer, 1984). Although the discrete mixture approach has become quite popular in econometric applications, the method is rarely employed in its truly nonparametric fashion. The standard practice is to pre-specify a (relatively low) number of support points, or to add points until computational problems inhibit further improvement in the likelihood.

Even to the extent that a fully nonparametric approach is pursued, the validity of the resultant parameters has been questioned on the ground that little is known about their sampling distribution. Scientific progress at this point has been held back by computational limitations. Due to the non-concavity of the likelihood function, localisation of the NPMLE is often a demanding task, even for small samples and parsimonious models. Existing Monte Carlo based evidence regarding the performance of NPMLE is therefore typically extracted from small samples and restrictive models (Heckman and Singer, 1984; Huh and Sickles, 1994; Baker and Melino, 2000).

The present paper takes advantage of the parallelisation capabilities of high performance computers, as well as some innovations regarding the computational treatment of large sets of dummy variables, that together facilitate Monte Carlo evaluations of quite a different scale and scope than those already reported in the literature. We find that NPMLE is extremely reliable, almost regardless of the true unobserved heterogeneity distribution, provided that the sample is large and that there is some exogenous variation in the hazard rates. Interval censoring does not cause insurmountable problems for the recovery of structural parameters (although assumptions regarding hazard behaviour within the censored intervals are required in the competing risks case). A particularly useful source of identification, that has received only modest attention in the literature, is the existence of a common calendar time factor in this exogenous variation. We also provide encouraging results regarding the sampling distribution of NPMLE, indicating that estimates of structural parameters are normally distributed, and that the standard errors conditional on the number of support points, also correctly represent the statistical uncertainty of NPMLE. However, we also present some results that may be source for concern: First, we find that the common practice of pre-specifying a low number of support points in the hetero-

geneity distribution (or the usage of restrictive information criteria) may produce unreliable results. Second, the NPMLE is not robust towards deviations from the basic modelling assumptions; e.g., if a mixed proportional hazard (MPH) assumption is imposed on a dataset for which it does not hold, serious bias problems may arise.

The present paper is related to a previous Monte Carlo study by Baker and Melino (2000), who investigated the behaviour of NPMLE for a discrete single risk duration model. They found that NPMLE is likely to fail for small samples when duration dependence is unrestricted, but that the usage of an information criterion with a penalty attached to the number of support points in the heterogeneity distribution may solve the problem. Our results confirm these findings, but they also establish that the problems of estimating both duration dependence and unobserved heterogeneity nonparametrically are indeed confined to small samples. With larger samples, penalties for parameter abundance typically do more harm than good. The present paper extends the results provided by Baker and Melino in several directions. First, although we maintain the assumption that the researcher only observes the outcomes of the statistical processes at discrete points in time, we assume that the data are generated by continuous time hazard rate models. Second, we extend the single risk model to a competing risks framework, with a particular emphasis on treatment effects. Third, while we focus on results obtained for the Mixed Proportional Hazard (MPH) rate family, we also investigate consequences of deviations from the MPH assumption. In particular, we look at the issue of heterogeneous duration dependence and treatment effects. Finally, we assess the sample selection problem that is inherent in most interval censored data, since subjects may enter and leave the origin state between two observation points.

Data-structures similar to the DGP's evaluated in this paper arise in many 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 or a training program. 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, a pay rise, or an early retirement scheme. In our experiments, we focus on situations in which large numbers of observations are available, to facilitate estimation techniques that are as 'nonparametric' 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).

In addition to providing new insights to the scope for nonparametric identification and estimation of duration models, our paper serves as a Monte Carlo evaluation of what has become known as the *timing of events approach*; see Abbring and Van den Berg (2003a). This approach has rapidly gained popularity among micro econometricians, particularly within the field of labour market econometrics. Influential contributions include Card and Sullivan (1988), Gritz (1993), Lillard (1993), Bonnal *et al.* (1997), and Van den Berg *et al*. (2004). Our paper also provides some insights to the usage of NPMLE in general. Finite mixtures are used extensively to

account for subject heterogeneity in models of dynamic discrete choice; see, e.g., Keane and Wolpin (1997), Mroz (1999) and Eckstein and Wolpin (1999).[1]

The remainder of the paper is structured as follows: The next Section describes the data generating process (DGP) that we refer to as our baseline model. Section 3 discusses identification issues, and Section 4 introduces the statistical model and optimisation algorithm used to recover the parameters of the DGP. Section 5 then examines the model's performance, in terms of recovering the true baseline parameters. Section 6 discusses the impact of sample size. Section 7 examines the consequences of modifying the distributions of unobserved heterogeneity, whereas 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 the left-truncation typically encountered in interval censored data. 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

[1] Finite mixtures have also been used extensively within other scientific disciplines, such as biometrics and psychometrics; see Skrondal and Rabe-Hesketh (2004) for an overview.

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 joint 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 day, week, or month. We take this point-in-time sampling into account by generating data that do not record exact transition times, but rather the time interval in which each transition has taken place. In most of our trials, this interval-censoring problem is substantial, and the period-specific transition probabilities typically lie between 5 and 25 per cent. We assume, however, that the underlying continuous time hazard rates are constant within each of these time intervals. We also assume that treatment and final exit cannot occur in the same time interval.

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 statistical approach, we have designed 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 training 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 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).[2]

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

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

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

[2] Note that previous treatment is assumed to be irrelevant when a subject is enrolled again, (i.e., $z_t \neq (1,1)$).

business cycles, seasonal effects, or changes in treatment capacity. The period-specific transition probabilities are equal to

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

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 setting up 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 into treatment 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

Provided that observed durations are accurately recorded, the mixed proportional hazards (MPH) structure of the baseline DGP ensures nonparametric 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 our case, observed durations are interval censored. Identification then clearly hinges on the assumption that the time trajectories of time-varying explanatory covariates are constant within each of the censored time intervals, such that the hazard rates are piecewise constant (An, 2004). In practice, the scope for actually recovering the true parameters from observed data depends on the size of the interval-censoring problem as well as 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 nonparametric 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 individual 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 nonparametrically 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 Data Likelihood Optimisation and Model Evaluation Criteria

The parameters are recovered by means of a nonparametric 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

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

where $\varphi_k(.), k = e, p,$ is defined in (1). Since the distribution of unobserved heterogeneity is assumed unknown to the researcher, we approximate the heterogeneity distribution in a nonparametric 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 (data), 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', in the sense that we are not able to increase the likelihood any further. We have examined alternative methods for verifying that this condition is satisfied. Most of the results presented in this paper are based on the following procedure: At each

stage of the estimation (i.e., after having estimated the model for a *given* number of support points), the appropriateness of an additional point of support is explored by a random search for likelihood improvement obtained by manipulating the location of an extra mass-point with the assigned probability of 0.0001. To start the search, we just copy one of the previously estimated location vectors (selected randomly). We then investigate how the likelihood function changes as we modify one element at a time. For the first element, we pick 50 random numbers in the interval $(-3, 2)$ and compute the resulting likelihood functions. 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 do the same thing for the next element in the location vector. 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 (of all the parameters in the model). If we have not been able to improve the likelihood through direct search, we replace all location vectors and probabilities with random numbers and start the full maximization anyway. We continue adding mass-points in this fashion until there is no improvement in the likelihood. For practical and computational reasons, we consider this to be the case when the log-likelihood increases by less than 0.05.

As an alternative to our direct search for potential improvements in the likelihood function, we have also tried to maximise the Gateaux (directional) derivative, in line with the procedure recommended by Baker and Melino (2000, p. 361), and also used by Heckman and Singer (1984). In our experiments, this procedure frequently produces a new support-point at the boundary of our search area, even though the final likelihood-maximising new point usually is in the interior of the search area. This

problem is closely related to interval and right censoring. To illustrate, consider the case where we have established the existence of *w-1* support points. The Gateaux derivative of the log-likelihood function associated with an additional point *w* is then equal to:

$$G_w = \sum_{i=1}^{N}\left(\frac{L_i(v_w)}{\sum_{l=1}^{w} p_l L_i(v_l)} - 1\right), \quad p_w = 0\,. \tag{5}$$

Now, assume that there are some right-censored spells in the data for which the previous *w-1* support points were poorly chosen, i.e. $\sum_{l=1}^{w} p_l L_i(v_l) \approx 0$. Since there are no transitions to the final destination state in right-censored spells, a highly negative number is obviously a very good choice for the corresponding new location point for these spells, yielding $L_i(v_w) \approx 1$. And since approximately one divided by approximately zero can be a very large number, these spells may easily dominate all other spells in the computation of the Gateaux derivative, even when there are relatively few of them. Due to the interval censored data, there will also be some spells that are terminated in the very first observed time interval. In these cases, a very large *positive* mass-point location yields likelihood functions approximately equal to unity. For this reason, it works better, with our data, to search for new points by maximising the likelihood with the new probability p_w set to a small positive number ρ. Technically, the increase in the log-likelihood approximates the integral of the Gateaux-derivative from $p_w=0$ to $p_w=\rho$.[3]

Irrespective of the usage of the likelihood function or the Gateaux derivative as the target function, it is possible that the way we reduce a two-dimensional search

[3] This can be seen from the lim definition of the Gateaux derivative given by Heckman and Singer (1984, p. 304, prior to Theorem 3.5).

problem to two one-dimensional searches (by checking for improvements associated with partial changes in v_e or v_p separately), leads us to miss potential improvement points associated with joint variations. We have therefore also attempted to perform a two-dimensional search by means of simulated annealing (see, e.g., Goffe *et al.*, 1994). This method clearly improved our ability to identify additional support points in the heterogeneity distribution through direct search, but at a relatively high cost in terms of increased computation time. Since our optimisation algorithm in any case embodies the extra step of a full maximisation (based on random starting values for all heterogeneity parameters), even when the direct search procedure fails to identify increase points, it did not change the final results to any noticeable extent. The main virtue of simulated annealing in this context seems to be that it reduces the need for the 'fallback' procedure of full maximisation. However, it does not eliminate it. This extra step appears to be important regardless of the search algorithm, not only because the search algorithm fails to identify existing increase points, but also because, during most of the estimation process, it is conditioned on wrong parameter values attached to observed explanatory variables (since the correct values have not yet been found). The full maximisation procedure that we use is a very robust combination of Fisher scoring (i.e., Newton-Raphson with the Hessian replaced by the Fisher information matrix) and BFGS.[4]

---

[4] 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. Sometimes, a heterogeneity parameter is esti-

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*.[5] 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 $a_M = \ln(\ln M)$. 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.

Given the absence of parametric restrictions on the period-specific spell duration and calendar time effects in our model, there is a relatively large number of parameters to estimate for each *W*, and almost all of them are attached to mutually exclusive dummy variables. In order to speed up the computations, we have developed

---

mated 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.

[5] Note that we use the total number of period-observations (*M*) in this formula, and not the number of subjects (*N*). This is perhaps not an obvious choice, see Skrondal and Rabe-Hesketh (2004, p. 265) for a discussion, but stands out as the natural thing to do in our context, given the unique information content embedded in each observation.

an optimisation routine that is tailored to nonparametric models, i.e., models in which all (or most) of the explanatory variables are dummy coded. This program builds on the concept of 'implicit dummy variables', which in essence reduces any number of mutually exclusive indicator variables to a single variable. The idea is that we take directly into account that most of the dummy variables are equal to zero most of the time. For example, if, say, S calendar-time dummy variables were to be treated like ordinary dummy variables, we would calculate the sum of these variables multiplied by their associated parameters every time the likelihood is computed. Each time, this would amount to S multiplications and $S-1$ additions, together with $2 \cdot S$ memory look-ups. But since all but one of the dummies are zero, there is only a single non-zero element in the sum. In contrast to human arithmetic, a computer uses equally long time to multiply and add zeros as it uses to multiply and add anything else. It is therefore more efficient to 'tell' the computer that most terms are zero and let it pick the right parameter directly. This is essentially what an 'implicit dummy' does, i.e., it replaces $2 \cdot S-1$ arithmetic operations and $2 \cdot S$ memory lookups with 2 memory lookups. We also utilize the implicit dummy variables when we compute the gradient. The speedup for the gradient computation is comparable to the speedup for the likelihood calculation, since we avoid computing and storing a lot of derivatives in the gradient. The mathematical result is of course exactly the same as if we had used ordinary dummy variables, but the computational cost is, in our case, dramatically reduced.

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 structural 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 (*2x40=80* duration parameters, *2x40=80* calendar time parameters, *2* parameters reflecting the effect of the exogenous covariate *x, 2* treatment effects, plus the parameters of the mixing distribution).[6] Table 3 reports the number of mass-points that were required to satisfy the four alternative model selection 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 shows 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 nonparametrically estimated effects of spell

[6] 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').

duration. The results are presented in the form of mean point estimates, mean estimated standard error, and the fraction of trials that led to rejection of the various true parameter values at a five per cent nominal significance level.[7] 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 reflects that as the spells proceed, *x* becomes correlated with unobserved heterogeneity, even though they are orthogonal to start with (see Cameron and Heckman, 1998, for a discussion of this phenomenon). 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 resulting from 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

[7] Approximate standard errors for the mean estimates can be obtained by dividing the mean standard errors by 10, i.e., the square root of the number of trials.

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 nonparametric 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.[8] A third point to note is that for the AIC and ML criteria, ordinary t-tests tend to reject the true parameters almost in accordance with the nominal significance levels. The standard errors used to perform this exercise (reported in Table 4) are calculated as the square roots of the diagonal elements of the inverted Fisher matrix. These standard errors are conditional on the number of support points in the mixture distribution, and it is far from obvious that they also represent the true uncertainty of the NPMLE. In order to take a closer look at the sampling distribution of NPMLE, we also made 100 data-replications based on exactly 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 repeated the estimation exercise described above. It turned out that i) the estimated parameters were approximately normally distributed, and ii) the estimated standard errors in each trial were virtually identical to the empirical standard deviation of point estimates across the 100 trials. Hence, it appears to

[8] 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).

be the case that standard inference procedures can be applied *as if* the optimally selected mixture model was known a priori. For the four key structural parameters (the effects of the exogenous covariate and of treatment), these points are illustrated in Figure 2 and Table 5, respectively. The normality test (Hendry and Doornik, 1996, pp 209-210) was applied on the 100 replications of all the estimated parameters attached to observed covariates. At the 5 per cent level, 10 of these 160 tests (6.25 per cent) rejected normality.[9]

Figure 2 around here

Table 5 around here

A fourth 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, Table 4 shows that models selected on the basis of HQIC or BIC tend to reject the true parameters much more often than indicated 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 inter-

[9] We did not attempt to test directly for multivariate normality, but it is known that all univariate marginal distributions of a multivariate normal distribution are themselves univariate normal (Johnson and Wichern, 1992).

est to investigate how models with just a few (predetermined) number of support points perform. For the four key 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 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 indeed saturated in terms of a maximum likelihood or a penalized maximum likelihood criterion, in line with the results provided by Lindsay (1983) and Heckman and Singer (1984).

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 is sometimes 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 be continuous (as is the case in our baseline model). There may, however, be other properties of the estimated heterogeneity distribution that have a more substantive interpretation. An important point to note regarding the model's ability to identify the distribution of unobserved heterogeneity is that although our 100 trials of data

generation and estimation returned the same (correct) structural parameters, they returned different heterogeneity parameters in terms of location vectors and probabilities, even in the experiments where all the 100 datasets were based on exactly the same drawing of unobserved heterogeneity. This implies that the discrete mixture distribution is not unique. Yet, this finding is not in conflict with the well-known theoretical result that proves uniqueness (Lindsay, 1983), since there are important features of our experiments, i.e., interval censored data, competing risks, and non-bounded mass-point locations, that do not conform to the assumptions underlying this proof. Figure 5 illustrates how the estimated discrete distribution mimics the true continuous distribution of unobserved heterogeneity. The two upper panels plot the marginal cumulative distribution functions (CDF's) of $\left(1-\exp(-\exp(v_e))\right)$ and $\left(1-\exp(-\exp(v_p))\right)$ for the true (bivariate normal) DGP and for (a randomly selected) estimated discrete heterogeneity distribution from one of the 100 trials.[10] The other 99 trials produced similar, but far from identical estimated discrete distributions. However, as illustrated in the two lower panels of Figure 5, when we merge the 100 estimated discrete distributions into a single one, it replicates the true CDF quite well.

Figure 5 around here

Even though the distribution function clearly cannot be perfectly replicated in a single trial, other distribution parameters, such as the first and second order moments, may be more robust. In a treatment evaluation setting, it is often desirable to characterise the selection process into treatment in terms of, say, a correlation coeffi-

[10] Note that $1-\exp(-\exp(v_k)) \approx \exp(v_k)$ for small $\exp(v_k)$. The reason why we have focused on these functions, rather than the hazard proportionality terms ($\exp(v_k)$) themselves is explained below. Note also that $1-\exp(-\exp(v_k))$ is the period-probability of making a transition to state *k* for a reference individual (with *x=0*, *d=1*, and *t=reference*) given that no competing destination exists, as a function of the unobserved covariate.

cient. Hence, it is clearly of interest to examine the extent to which such parameters can be recovered from observed data. Table 6 reports the first and second order moments of $\left(1-\exp(-\exp(v_e)),1-\exp(-\exp(v_p))\right)$ for the true and the estimated heterogeneity distributions. The moments of the estimated heterogeneity distribution are, on average, very close to the moments of the true heterogeneity distribution. And the correlation coefficient calculated on the basis of the estimated heterogeneity distribution seems to be a consistent estimator for the true correlation coefficient. But, while the first order moments appear to be consistently estimated regardless of the information criterion, consistent estimates of the second order moments depend on the least restrictive criterions (AIC or ML) being chosen.

Table 6 around here

The reason why we have not reported distribution measures, such as first and second order moments, for the heterogeneous proportionality terms themselves, i.e., $\left(\exp(v_e),\exp(v_p)\right)$, is that they typically turn out to be irrecoverable from interval-censored data. The explanation for this is that it is empirically impossible to differentiate between distinct large values of heterogeneity locations $\left(v_e,v_p\right)$, since they yield indistinguishable transition probabilities in any discrete time interval. The problem is illustrated in Figure 6, where we have plotted the functional relationship between period-specific transition probabilities and the log integrated hazard rate (disregarding the issue of competing risks). It is clear that any integrated hazard rate exceeding the number of two implies virtually the same transition probability of unity; hence the particular numbers that enter the estimated unobserved heterogeneity vectors at high levels are selected with extreme statistical uncertainty. Nevertheless, the exact locations of such values have of course enormous impact on the calculation of the mo-

ments of the $\left(\exp(v_e), \exp(v_p)\right)$ distribution. In fact, since we have not restricted $\left(v_e, v_p\right)$ to be finite, the moments of the $\left(\exp(v_e), \exp(v_p)\right)$ distribution, as well as the moments of the $\left(v_e, v_p\right)$ distribution may not even exist.

Figure 6 around here

6 The Role of Sample Size

So far, the analysis has been based on datasets comprising 50,000 subjects. In this section, we consider the impact of sample size, by comparing results based on five alternative sample sizes, containing from 5,000 to 5,000,000 subjects. The main results are summarised in Table 7, 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 seems to increase monotonously with sample size for all information criteria, but at a low and declining rate (the elasticity of the number of support points with respect to the number of subjects is on average around 0.1). The rate of increase is also lower the more restrictive is the information criterion.

Table 7 around here

The mean errors that are presented in Table 7 are all based on the same total number of subjects, irrespective of sample size used in each trial, namely 5,000,000. When we look at sample sizes of only 5,000 subjects, 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 7 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, which approximately corresponds to a 17 per cent over-evaluation of hazard rates at durations ≥ 2, relative to the first period), 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 7, 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; i.e., BIC and HQIC 'under-corrects', ML over-corrects, and AIC (almost) hits the target.

Figure 7 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. We have drawn unobserved heterogeneity and calendar time effects only once for each model type, so that the DGP used to generate spells is the same across the 10 trials. The main purpose 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 nonparametric identification of treatment effects and duration dependence. 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 8. 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 8 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 select 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 9. The bottom line is that NPMLE robustly recovers the true structural parameters, including treatment effects, irrespective of the way unobserved het-

erogeneity is distributed in the data. As illustrated in Figure 8, 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 9 around here

Figure 8 around here

It may be of interest to take a closer look at the results from model v) (see the bottom part of Table 9), 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 model's 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 in the final destination hazard (5 per cent in the DGP) was identified in all the 10 trials, while the presence of defective risks in the treatment hazard (1 per cent in the DGP) 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. This clearly reflects the non-uniqueness

property discussed in Section 5; i.e., different combinations of mass-point locations and probabilities are equally consistent with data.

8 The Role of the True Causal Effects

In this section, we show that the consistence of NPMLE does not hinge on the specific selection of true effects embedded in the baseline model. But first, we examine the importance of true variation in the sources of model identification, i.e., in observed exogenous variables. We have already established that we do not need to observe the subject-specific 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 parameters, 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:

$$WMAE_k = \frac{1}{R}\sum_r \sum_d \frac{\frac{1}{\hat{\psi}_{kdr}}}{\sum_d \frac{1}{\hat{\psi}_{kdr}}} \left| \hat{\lambda}_{kdr} - \lambda_{kd} \right| . \qquad (6)$$

Some illustrative results are provided in Table 10. As expected, the manipulation of the sources of identification primarily affects the estimates of the spell duration baseline parameters 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 nonparametric identification of spell duration effects. Auto-correlated calendar time effects do not reduce the scope for identification.

Table 10 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 11 and Figure 9.

Table 11 around here

Figure 9 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 varies according to 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. Provided that the model is correctly specified – including the appropriate interaction terms – the true parameters are 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 10, 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 10 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 11 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), can produce results that have no convenient interpretation. In particular, the NPMLE of an assumed homogeneous parameter that is really heterogeneous in the DGP, does not necessarily represent an average of the underlying true parameters. The reason for this is, of course, that the parameter heterogeneity in-

duces 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 11 around here

A particularly interesting case to look at is that with heterogeneous treatment effects; see, e.g., Heckman *et al*. (1999) for a survey. 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). We illustrate this point by modifying the DGP, such that treatment effects are, indeed, heterogeneous. To be specific, we assume that on-treatment and post-treatment effects are equal, i.e., $\alpha_1 = \alpha_2 = \alpha$, and that α is independently normally distributed across subjects with mean 0.2 and variance 0.2. This implies that 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 $ATE = E[\exp(\alpha_i)] = \exp(0.2 + 0.1) = 1.35$, i.e., a 35 per cent increase in the hazard rate. To avoid inessential complications, we simplify the DGP in this case,

by assuming that the unobserved intercepts (v_e, v_p) are uncorrelated. Based on these assumptions, we generate 100 new artificial datasets (containing 50,000 subjects each), and then estimate the parameters of the model, as before. The results are displayed in Table 12. They reveal that our estimates of treatment effects are negatively biased compared to the true ATE, although the deviation is not particularly large in this case. As expected, the bias is larger for post-treatment effects than for on-treatment effects, since subjects entering into post-treatment are negatively selected from the start (given the perfect correlation between on-treatment and post-treatment effects in this example).

Table 12 around here

A corollary of the sorting-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 homogeneous (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.

A natural solution to the problem of heterogeneous treatment effects is to model this heterogeneity explicitly; i.e., interpret treatment effects as state-specific contributions to the distribution of unobserved heterogeneity (random coefficients). This approach raises, however, some new identification and estimation problems, particularly regarding the presence of defective transition probabilities. It is beyond to

scope of this paper to evaluate a random coefficient approach to estimation of average treatment effects, but we consider this to be a fruitful avenue for future research.

10 Interval Censoring and Left-Truncation

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 starting and ending between two observation-points – are never recorded. This is sometimes referred to as left-truncation, and it 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 25 per cent of the subjects are lost. Figure 12 presents the results from 10 trials based on each of these models. The two upper panels illustrate what happens with the estimated duration parameters (in the final destination hazard) when the sample selection problem is disregarded, in the sense that the selected samples are treated as if they were un-selected. The NPML estimators then fail to remove the spurious nega-

tive duration dependence. Other parameters are also substantially 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 –0.878 (with a mean standard error of 0.012). When 25 per cent of the spells are unobserved, the estimate is –0.794 (with a mean standard error of 0.016) (recall that the true value is –1).

Figure 12 around here

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)

$$f(v_i \mid d > 1) = \frac{\exp\left(-\sum_{k \in K_{it}} \varphi_k(t, 1, x_{it}, z_{it}, v_{ki}) \right)}{\underset{v_i}{E}\left[\exp\left(-\sum_{k \in K_{it}} \varphi_k(t, 1, x_{it}, z_{it}, v_{ki}) \right) \right]} f(v_i), \qquad (7)$$

and the likelihood function takes the form

$$L \mid d>1 = \prod_{i=1}^{N} \underset{v_i}{E}\left[\frac{\exp\left(-\sum_{k \in K_{it}} \varphi_k(t,1,x_{it},z_{it},v_{ki}) \right)}{\underset{v_i}{E}\left[\exp\left(-\sum_{k \in K_{it}} \varphi_k(t,1,x_{it},z_{it},v_{ki}) \right) \right]} L_i\left(v_i \mid d>1\right) \right], \qquad (8)$$

where $L_i\left(v_i \mid d>1\right)$ can be obtained from Equation (3). Hence, the solution to the left-truncation 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 12 illustrate what happens with the estimated duration dependence parameters when we maximise the likelihood function in (8). The negative bias is now almost removed, even in the model with as much as 25 per cent censoring. And the effects of other parameters are again also correctly recovered (not shown). For example, when 10 per cent of the spells are unobserved, the effect of *x* on the final destination hazard is now estimated to –0.982 (with mean standard error 0.010). When 25 per cent of the spells are unobserved, the estimate is –0.986 (with mean standard error 0.020).

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 knowl-

edge, 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

$$prob(\sum_k y_{kit} = 0 \mid d = 1, x_{it}, v_i) = \int_0^1 \exp(-(1-s)\sum_k \varphi(t,1,x_{it},0,v_i)ds$$
$$= \frac{1-\exp\left(-\sum_k \varphi(t,1,x_{it},0,v_i)\right)}{\sum_k \varphi(t,1,x_{it},0,v_i)} \tag{9}$$

11 Conclusion

Based on comprehensive Monte Carlo experiments, we conclude that, for a correctly specified model, the nonparametric maximum likelihood estimator (NPMLE) robustly recovers the true treatment effects from non-experimental interval-censored 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.

We have also demonstrated that NPMLE is fragile towards unjustified restrictions, and, in particular, that 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 (nonparametric) 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 mixture distribution are not unique. They are estimated with large statistical uncertainty and have no convenient interpretation. But the first and second order moments of the transition probability distribution (conditioned on observed covariates) appear to be both unique and recoverable.
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 left-truncation (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 specifying the likeli-

hood 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. Deviations from the proportionality assumption that are unaccounted for in the model may cause 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 Nonparametric Identification of Treatment Effects in Duration Models. *Econometrica*, Vol. 71, 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, 701-710.

An, M. Y. (2004) Likelihood-Based Estimation of a Proportional-Hazard, Competing-Risk Model with Grouped Duration Data. Economics Working Paper Archive

at WUSTL in its series Urban/Regional with number 0407013.

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.

Bonnal, L., Fougere, D. and Serandon, A. (1997) Evaluating the Impact of French Employment Policies on Individual Labour Market Histories. *Review of Economic Studies*, Vol. 64, 683-713.

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.

Cameron, S. V. and Heckman, J. J. (1998) Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males. *Journal of Political Economy*, Vol. 106, 262-333.

Card, D. and Sullivan, D. (1988) Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment. *Econometrica*, Vol. 56, 497-530.

Eckstein, Z. and Wolpin, K. I. (1999) Why Youths Drop Out of High School: The Impact of Preferences, Opportunities, and Abilities. *Econometrica*, Vol. 67, 1295-1339.

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.

Goffe, W. L., Ferrier, G. D. and Rogers, J. (1994) Global Optimization of Statistical Functions with Simulated Annealing. *Journal of Econometrics*, Vol. 60, 65-99.

Gritz, R. M. (1993) The Impact of Training on the Frequency and Duration of Employment. *Journal of Econometrics*, Vol. 57, 21-51.

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.

Hendry, D. F. and Doornik, J. A. (1996) *Empirical Econometric Modelling Using PcGive 9.0 for Windows*. International Thomson Business Press.

Huh, K. and Sickles, R. C. (1994) Estimation of the Duration Model by Nonparametric Maximum Likelihood, Maximum Penalized Likelihood, and Probability Simulators. *The Review of Economics and Statistics*, Vol. 76, 683-694.

Johnson, R. A. and Wichern, D. W. (1992) *Applied Multivariate Statistical Analysis*. Prentice-Hall.

Kean, M. P. and Wolpin, K. I. (1997) The Career Decision of Young Men. *Journal of Political Economy*, Vol. 105, 473-522.

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.

Lillard, L. A. (1993) Simultaneous Equations for Hazards. *Journal of Econometrics*, Vol. 56, 189-217.

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.

Mroz, T. A. (1999) Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of Dummy Endogenous Variable on a Continuous Outcome. *Journal of Econometrics*, Vol. 92, 233-274.

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 Reser-

voirs of Scientific Knowledge? *Economic Journal*, Vol. 113. (2003), F258-F281.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalised Latent Variable Modelling. Multilevel, Longitudinal, and Structural Equations Models.* Chapman & Hall/CRC, Interdisciplinary Statistics Series.

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 den Berg, G.J., B. Van der Klauw and J. Van Ours (2004) Punitive Sanctions and the Transition from Welfare to Work, *Journal of Labor Economics,* Vol. 22, 211-241.

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 Nonparametric 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
Estimated and observed statistical uncertainty
Results from 100 trials based on a unique drawing of unobserved heterogeneity and calendar time effects from the baseline DGP

	AIC			ML		
	Mean estimated S.E for point estimate	Observed stand. dev. for point estimate	Observed stand. dev. in estimated S.E.	Mean estimated S.E for point estimate	Observed stand. dev. for point estimate	Observed stand. dev. in estimated S.E.
β_e	0.023	0.024	0.0009	0.024	0.024	0.0008
β_p	0.021	0.019	0.0002	0.021	0.019	0.0002
α_1	0.032	0.033	0.0008	0.032	0.032	0.0007
α_2	0.036	0.038	0.0009	0.037	0.036	0.0008

Table 6
The lower order moments of the estimated heterogeneity distribution with respect to the first-period normalised transition probabilities $\bar{q}_{ki} = 1 - \exp(-\exp(v_{ki}))$
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 $\bar{q}_{ei}$	0.139	0.137	0.005	0.138	0.005	0.139	0.005	0.138	0.005
Mean $\bar{q}_{pi}$	0.071	0.071	0.003	0.071	0.003	0.072	0.003	0.072	0.003
Var $\bar{q}_{ei}$	0.016	0.010	0.004	0.013	0.003	0.015	0.003	0.017	0.003
Var $\bar{q}_{pi}$	0.005	0.004	0.001	0.005	0.001	0.005	0.001	0.005	0.001
Corr $(\bar{q}_{ei}, \bar{q}_{pi})$	0.252	0.481	0.216	0.336	0.136	0.278	0.114	0.237	0.079

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

Table 7
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 8
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 9
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 10
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.
α_1=0	0.001	0.042	0.024	0.037	0.037	0.051	0.024	0.051	-0.024	0.037
α_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 11
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 12
Estimated Effects of Treatment.
Results from 100 trials based on a modified baseline model with heterogeneous treatment effects

	DGP	AIC		ML	
Mean # of support points		6.8		9.7	
	log(ATE)	Mean Est.	Mean S.E.	Mean Est.	Mean S.E.
α_1	0.3	0.276	0.032	0.273	0.032
α_2	0.3	0.248	0.046	0.245	0.047

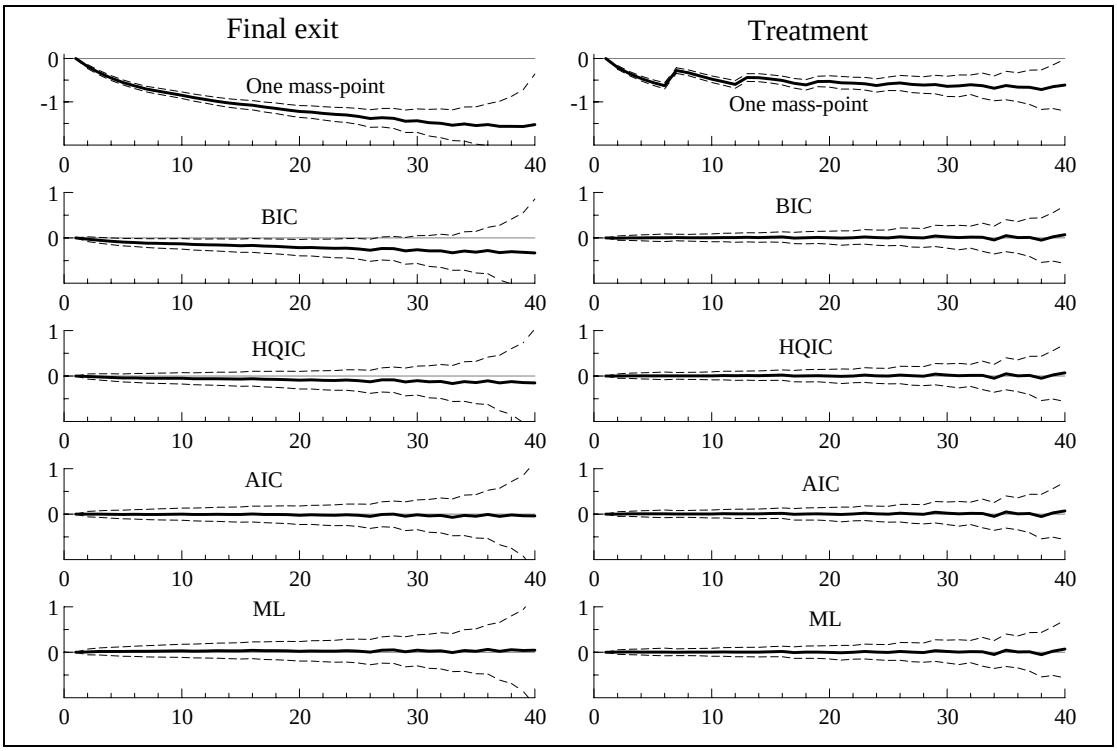


Figure 1. Average estimated effects of spell duration (point estimates with 95% confidence intervals, based on observed standard deviation from the 100 trials). The true effects are equal to zero for all durations.

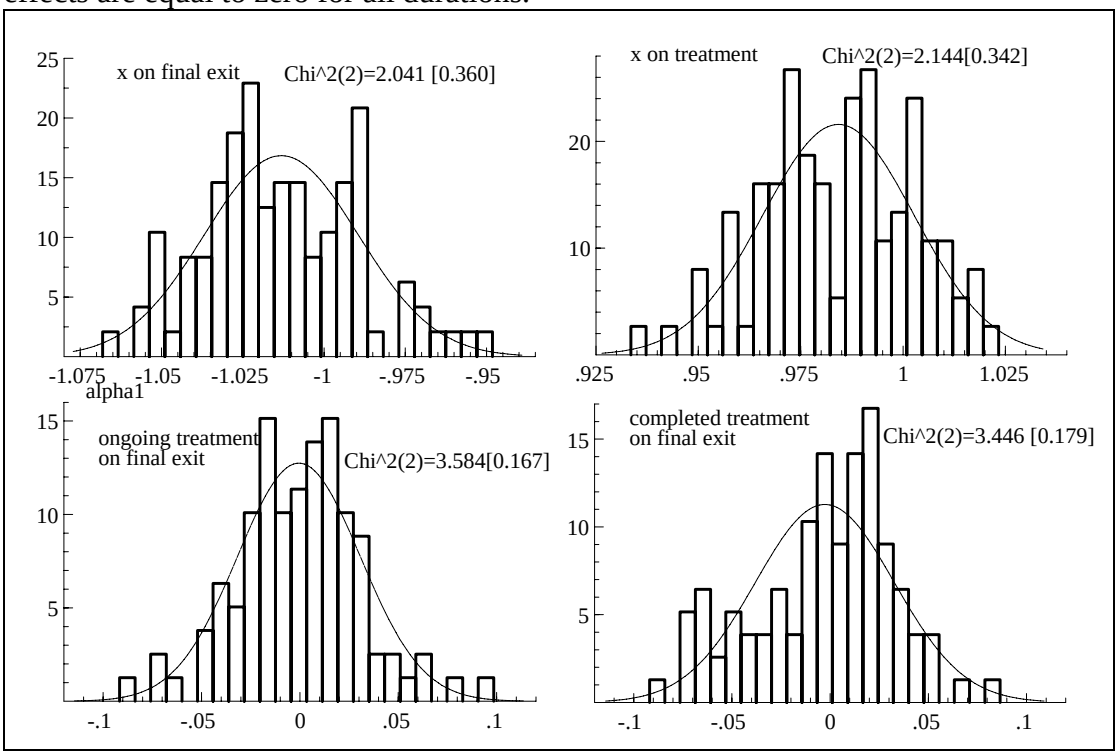


Figure 2. Distribution of the estimates of the four structural parameters, based on the ML criterion, and normal densities (with Chi^2(2) normality tests)

Note: The graphs are based on results from 100 trials, based on a unique drawing of unobserved heterogeneity and calendar time effects from the baseline DGP. Each histogram contains 25 bars. The normality tests are described in Hendry and Doornik (1996, pp 209-210).

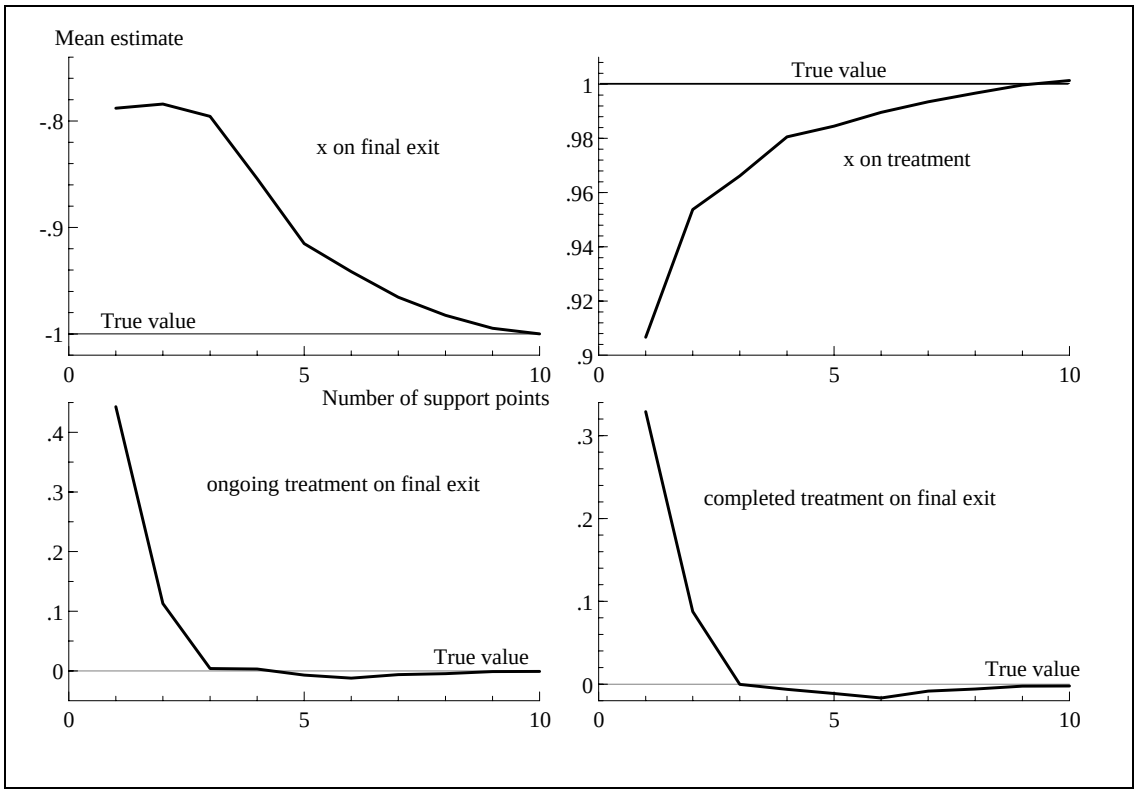


Figure 3. Mean estimates (over 100 trials) of the four structural parameters as functions of the number of support points in the unobserved heterogeneity distribution (1 support points corresponds to a model without unobserved heterogeneity).

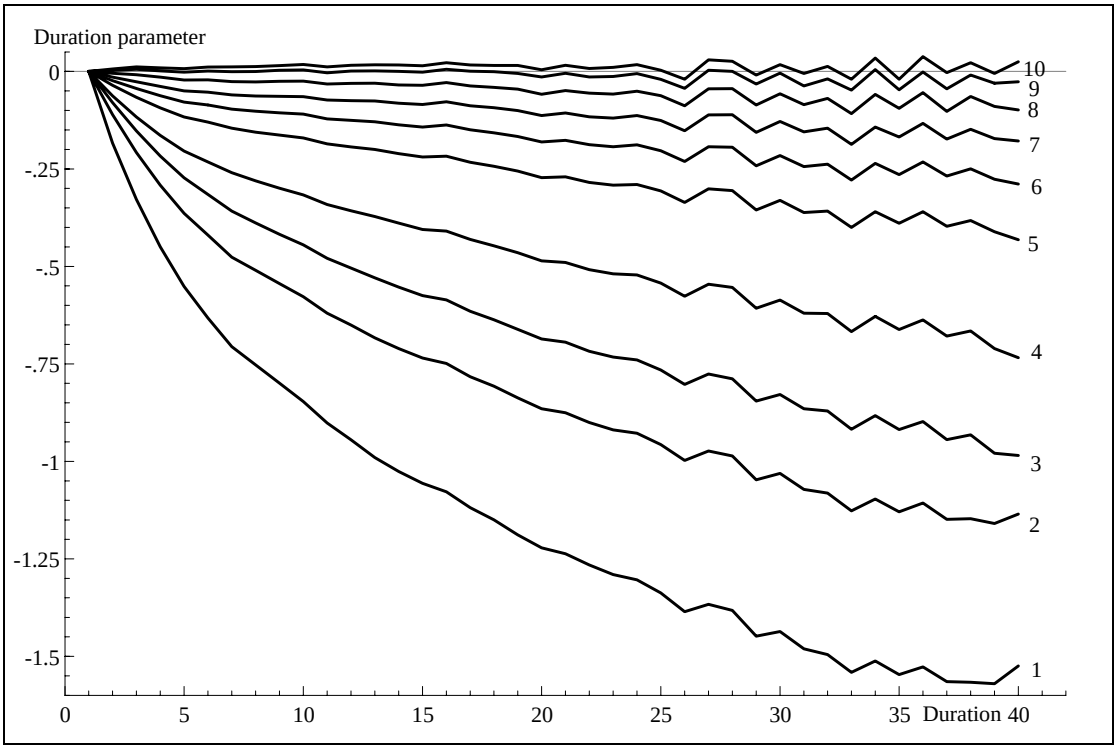


Figure 4. Average estimated duration parameters in the final destination hazard, with from 1 to 10 support points in the unobserved heterogeneity distribution. The true parameters are all equal to zero.

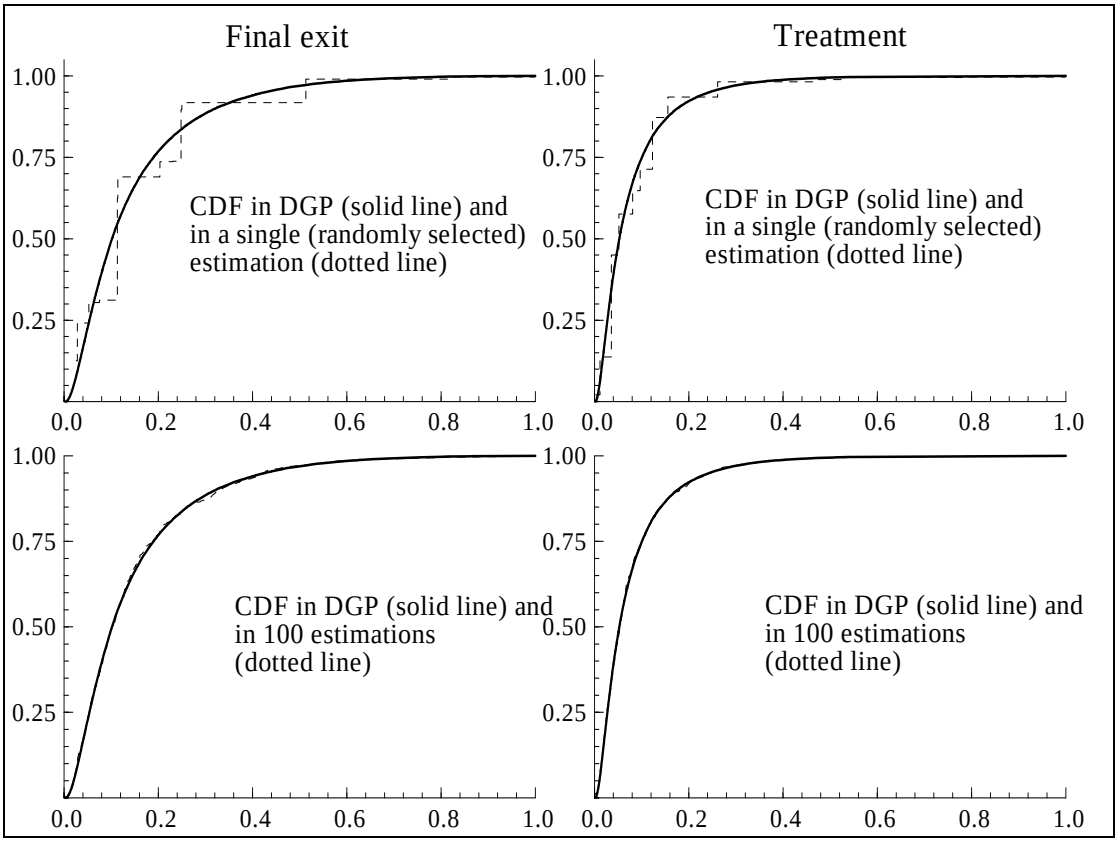


Figure 5. Marginal cumulative distribution functions for unobserved heterogeneity *(1-exp(-exp(v_k))* in true DGP and in estimated models (based on the Maximum Likelihood criterion).

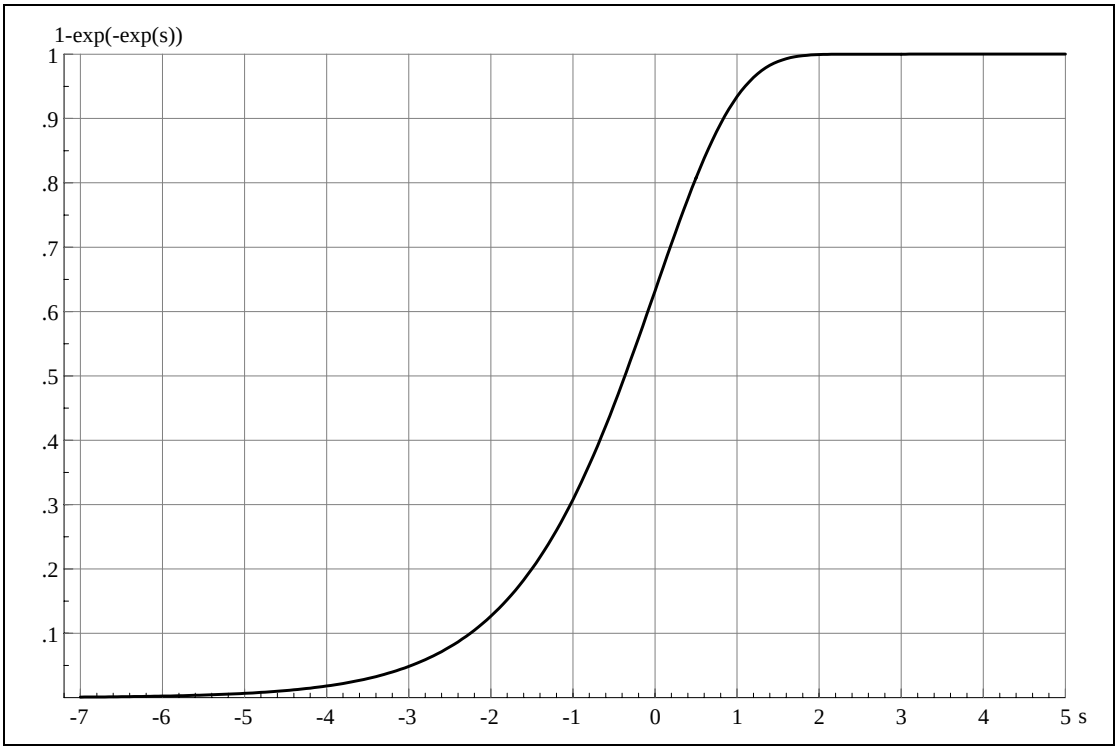


Figure 6. Single risk transition probability during a discrete time interval, as a function of the log integrated hazard rate *s*.

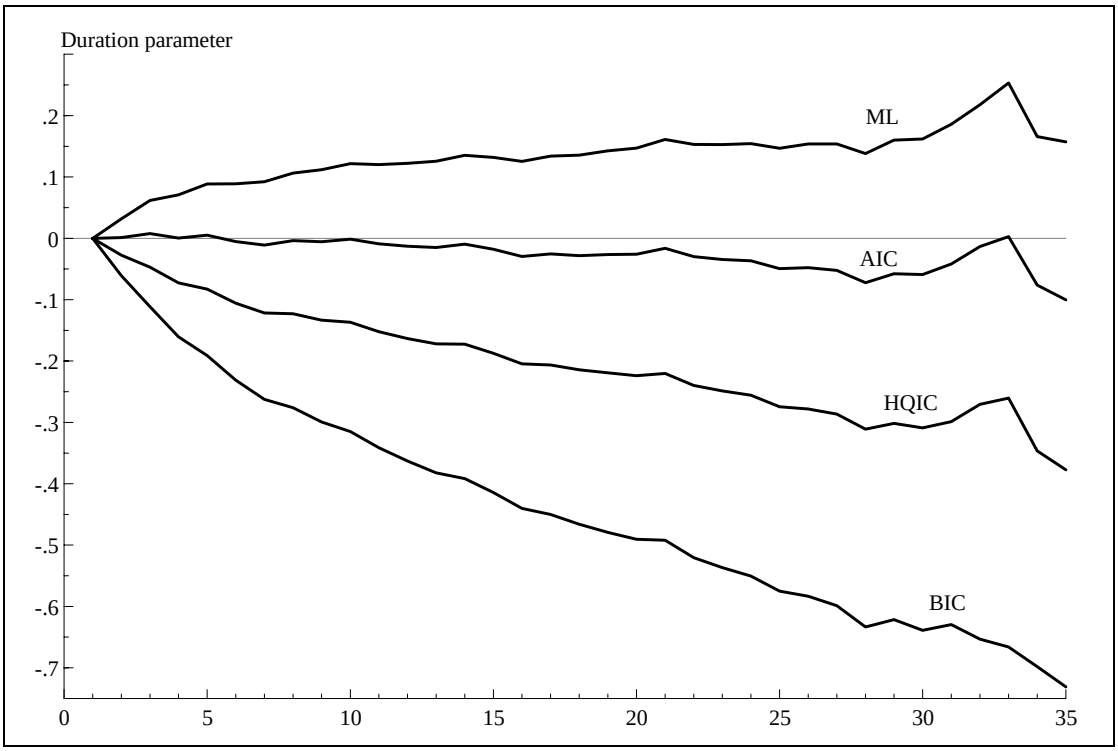


Figure 7. Average estimated duration parameters in the final destination hazard, based on 500 samples with 10,000 subjects in each sample.

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.

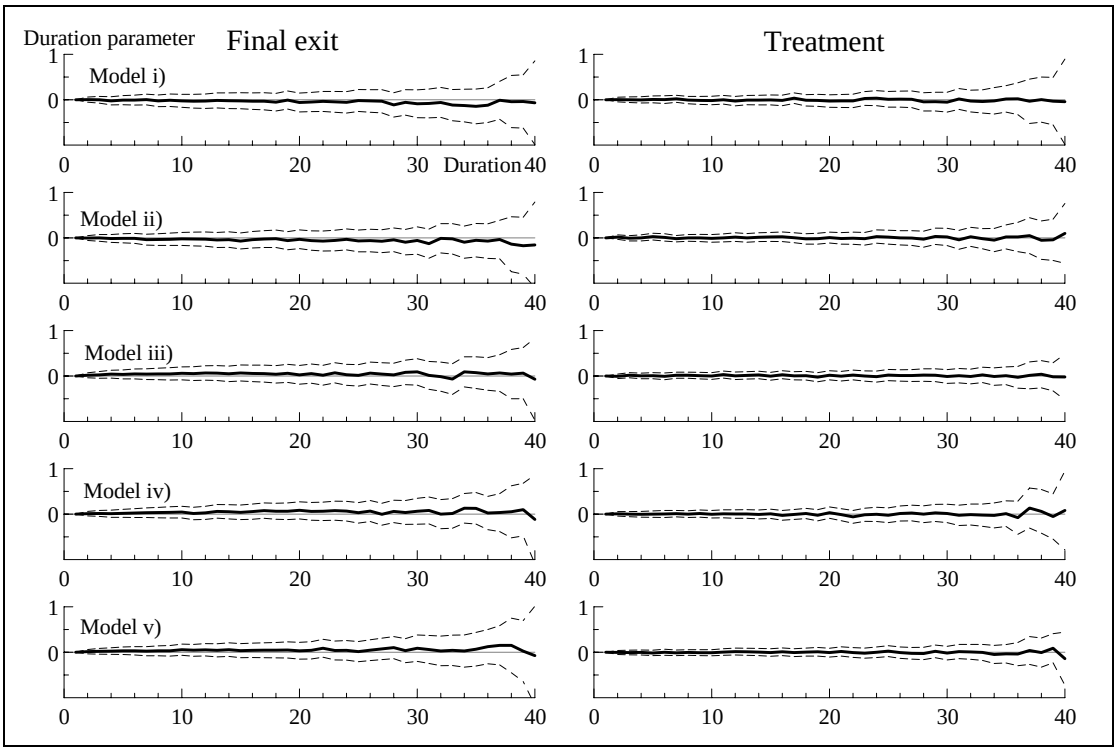


Figure 8. Average estimated effects of spell duration (with 95 per cent confidence intervals), according to the Maximum Likelihood criterion (based on average point es-

timates and standard errors over 10 trials for each model). The true effects are equal to zero for all durations

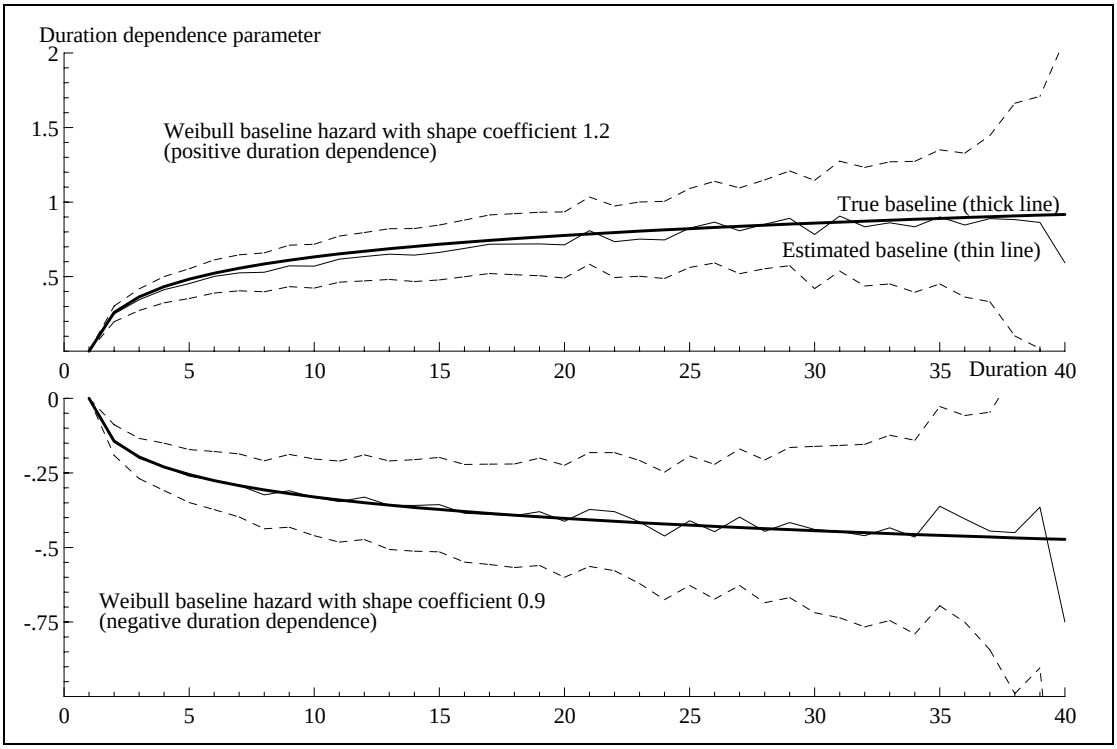


Figure 9. Estimated duration dependence parameters according to the Maximum Likelihood criterion (with 95 per cent confidence intervals) in final destination hazard when the true baseline exhibits positive or negative duration dependence (average based on 10 trials)

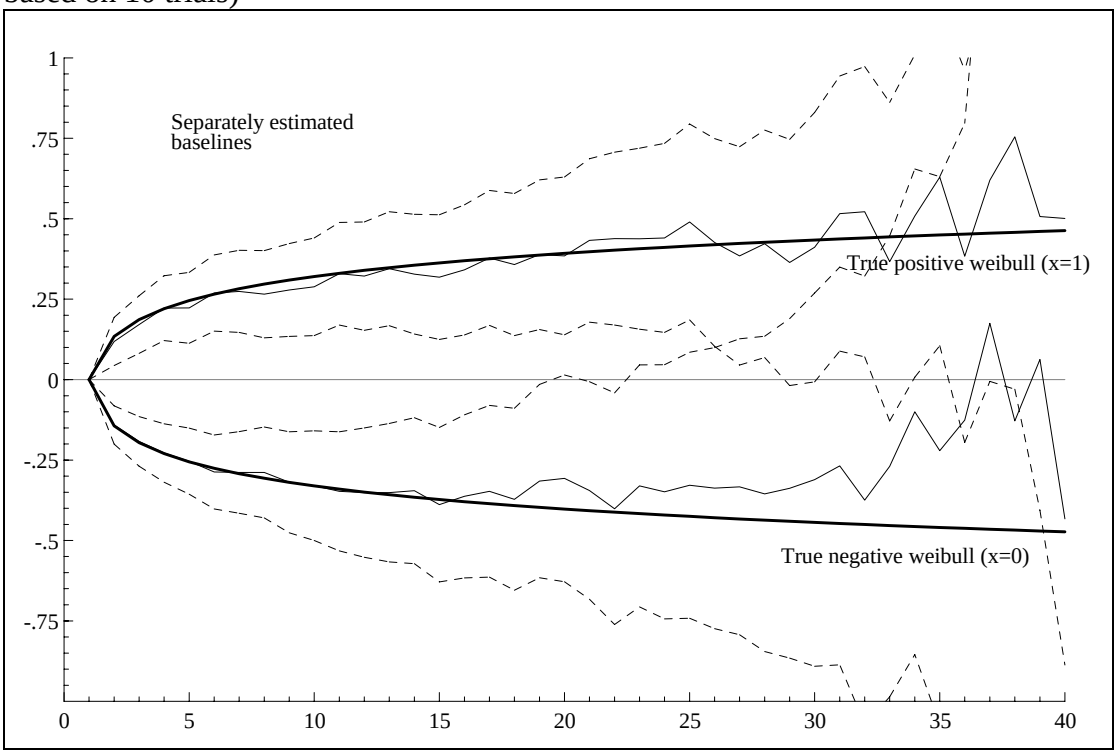


Figure 10. Estimated group-specific duration dependence parameters according to the Maximum Likelihood criterion (with 95 per cent confidence intervals) in final desti-

nation hazard when the baseline exhibits positive duration dependence for *x=1* and negative duration dependence when *x=0* (average based on 10 trials)

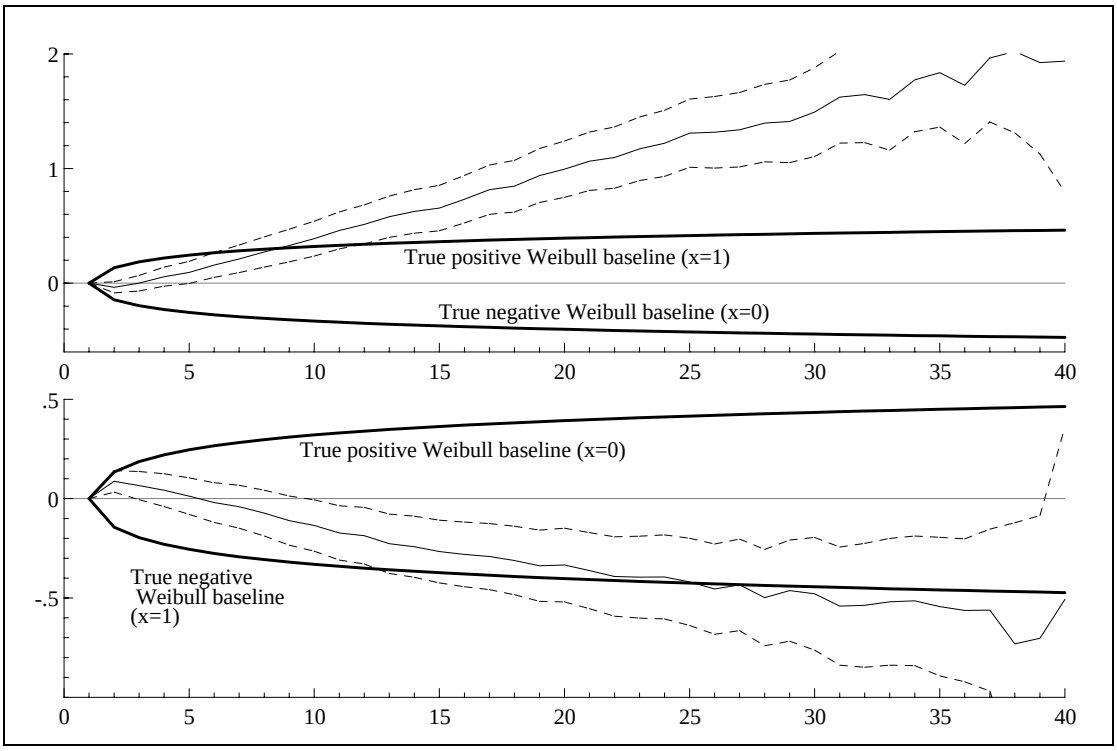


Figure 11. Estimated common duration dependence parameters according to the Maximum Likelihood criterion (with 95 per cent confidence intervals) in final destination hazard when the true baseline exhibits positive duration dependence for *x=1* and negative duration dependence when *x=0* (upper panel) 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).

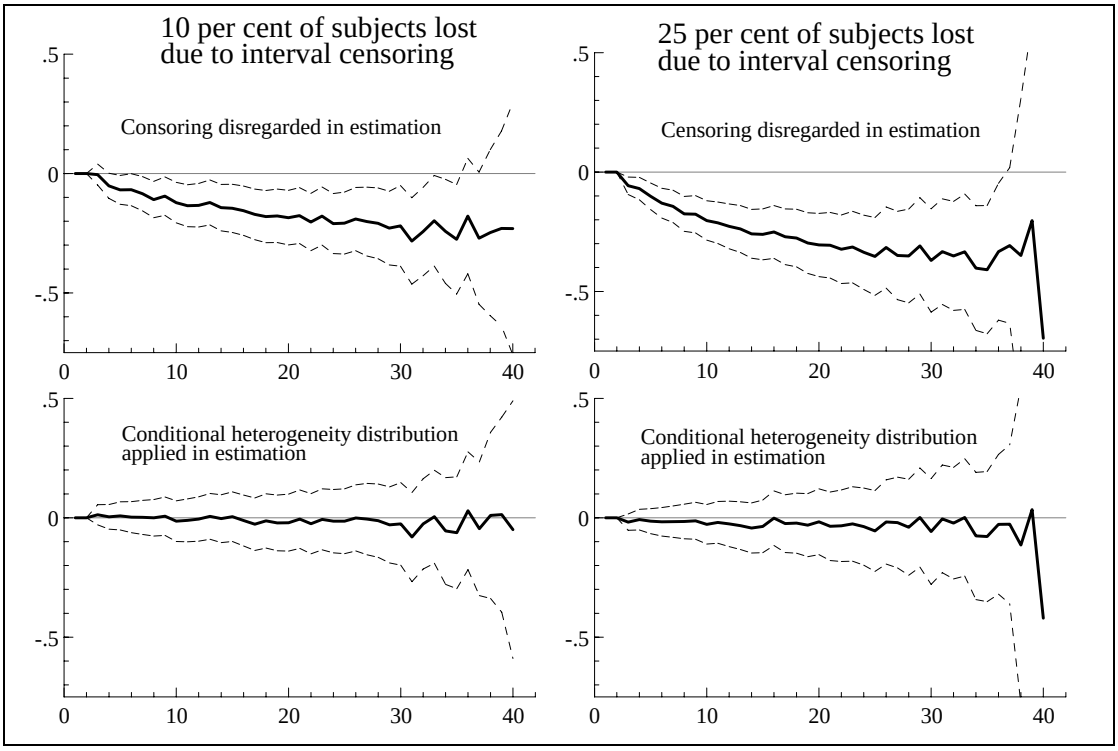


Figure 12. Mean of estimated final destination duration dependence parameters with (lower panels) and without (upper panels) correction for sample selectivity due to interval censoring (Maximum Likelihood criterion, with 95 per cent confidence intervals).

Note: The results are based on 10 trials. The DGP is a baseline model with 100,000 subjects to start with. The presented parameters are selected on the basis of the ML criterion.