DEPARTMENT OF ECONOMICS

Working Paper

Left-Censoring in Duration Data: Theory and Applications

Anna Christina D'Addio and Michael Rosholm

Working Paper No. 2002-5

ISSN 1396-2426

UNIVERSITY OF AARHUS • DENMARK

INSTITUT FOR ØKONOMI

AFDELING FOR NATIONALØKONOMI - AARHUS UNIVERSITET - BYGNING 350
8000 AARHUS C - ☎ 89 42 11 33 - TELEFAX 86 13 63 34

WORKING PAPER

Left-Censoring in Duration Data: Theory and Applications

Anna Christina D'Addio and Michael Rosholm

Working Paper No. 2002-5

DEPARTMENT OF ECONOMICS

SCHOOL OF ECONOMICS AND MANAGEMENT - UNIVERSITY OF AARHUS - BUILDING 350
8000 AARHUS C - DENMARK ☎ +45 89 42 11 33 - TELEFAX +45 86 13 63 34

Left-Censoring in Duration Data: Theory and Applications

Anna Cristina D'Addio
CIM - Aarhus School of Business
Department of Economics,
University of Aarhus

Michael Rosholm
Department of Economics,
University of Aarhus

March, 2002*

Abstract

In this paper, we discuss how to best exploit the information contained in spells that are in progress when an observation period begins, that is, left-censored and left-truncated duration data. We provide a survey of censoring and truncation mechanisms in event history models. We describe some approaches that have been suggested in the literature to deal with left-censoring. Our contribution is the description of ways to use additional information to obtain more efficient parameter estimates using the left-censored informations, and particularly, the derivation of the associated likelihood expressions. In order to use the information efficiently, we often resort to the stationarity assumption. Hence, we provide a Hausman test for this assumption. The second part of the paper briefly presents some empirical examples which demonstrates the efficiency gains associated with the use of the information contained in the left-censored observations. In particular, we show how the use of some additional pieces of information allows us to obtain more efficient estimates of the parameters of interest. In doing this, we use the information reported in the waves 1990-1992 of the French Labour Force Surveys on young French individuals.

JEL: C13, C41, C51, C81

Keywords: left-censoring, left-truncation, conditional likelihood, hazard rate, stationarity

*Mailing address: Department of Economics, University of Aarhus, building 326, 8000 Aarhus C, Denmark. E-mail: MRosholm@econ.au.dk (Rosholm) and adaddio@econ.au.dk (D'Addio). The authors acknowledge financial support from the Marie Curie Fellowship Scheme of the European Commission and the Danish Research Agency (the SUE grant). We are especially grateful to Bart Cockx, Bo Honoré, Thierry Magnac, Michel Mouchart as well as to the participants at the Econometric Society European Meeting 2001 and the Meeting of the Nordic Econometric Society 2001. The usual disclaimer applies.

1. Introduction

Event history data have been increasingly used in recent years providing information both on the times over which individuals change from one discrete state to another and on the sequence of the states they occupy. These data are normally obtained through surveys carried out at some specific dates. As a consequence, the information on the length of spells is frequently incomplete and the data arising are said to be “censored” and/or ”truncated”.

Different kinds of censoring/truncation mechanisms exist in the framework of event history models. The most common of them - *right-censoring* - concerns missing information on the times and states occupied after the end of a given observation period and is routinely handled in event history analysis. Another censoring mechanism arises when information on the times and states occupied before the beginning of the observation period is missing. We generally refer to this mechanism as *left-censoring*. Attempts to correct for left-censoring in empirical applications are rare (e.g. Gritz, 1993; Rosholm, 2001; D'Addio and Rosholm, 2002), which is partially because it is a very complex issue, and partially because it has been the general perception that left-censored observations does not contain much information that can be exploited in empirical studies. The complexity is mainly a consequence of the fact that the entry rate into the initial state is unknown. Solutions to this problem have proceeded either conditionally on very restrictive assumptions or discarded all left-censored observations. However, in some situations the information embedded in the left-censored observations is crucial, e.g. when the observation period is short and the fraction of left-censored observations fairly large.

Left-censoring is a challenging issue, and the question of how to treat left-censoring arises often in social science disciplines that rely heavily on survey data (e.g. in the analysis of employment and unemployment duration, poverty duration, the duration of welfare dependence etc.). In addition, the problem of left-censoring occurs in many other disciplines, like epidemiology, where the beginning of the process is not known with certainty. For example, in studies disease duration - like cancers - the onset of the disease is often not known (at least not with certainty), see e.g. Andersen et al. (1993).

In this paper, we discuss ways in which we can best exploit the information contained in spells that are in progress when an observation period begins. We begin by providing a survey of censoring and truncation mechanisms in event history models, focusing on those affecting data from the left. Further, we will describe some approaches that have been suggested in the literature to deal with left-censoring. Our main contribution is the description of ways to use additional information to obtain more efficient parameter estimates using the left-censored informations, and particularly, the derivation of the associated likelihood expressions. In order to use the information efficiently, we often resort to the stationarity assumption. Hence, we provide a Hausman test for this assumption. The second part of the paper will briefly present some empirical examples which demonstrates the efficiency gains associated with the use of the information contained in the left-censored observations. In particular, we show how the use of some additional pieces of information allows us to obtain more efficient estimates of the parameters of interest. In doing this, we use the information reported in the waves 1990-1992 of the French Labour Force Surveys on young French individuals.

The paper is structured as follows. In the next section we characterise complete and incomplete data and provide a description of the most common sample designs and of the bias that can arise when spells in progress at the survey date are not duly accounted for. In section 3, we illustrate and derive the relevant expressions for the situations termed left-censoring and left-truncation, respectively. We derive the relevant likelihood expression when we condition on all the available information and construct the stationarity test. Section 4 is devoted to the empirical examples. In section 4.1 we concentrate on left-truncated schooling durations and in section 4.2 on left-censored unemployment durations. Some conclusions are drawn in section 5.

2. Characterisation of complete and incomplete data

2.1. Some definitions

We begin with a discussion of (left-) *censoring* and *truncation.* These concepts are often used with meanings that vary from one study to the other and represent therefore a source of confusion.

Some authors use the term *censoring* to refer to the situation when an individuals spell of interest ends before the beginning of the observation period and thus is not observed at all, see Keiding (1986) and Yamaguchi (1991). Conversely, *truncation* refers in this terminology to the case where a spell is in progress when the observation period begins (Yamaguchi, 1991). A distinction is made between left truncation with an unknown origin date and left truncation with a known origin date (Guo, 1993).

Other authors define *censoring* in duration and transition models as the lack of information due to the finiteness of the observation period; this implies that missing information on the right and the left is defined symmetrically, see e.g. Mayer and Tuma (1990) or Blossfeld and Rohwer (1995). A spell which is in progress at the beginning of the observation period, and for which we observe only the duration from that point in time, is referred to as left-censored. In this terminology, left *truncation* refers to spells that are in progress at the beginning of the observation period, the origin date of which is known

In the exposition we adopt the latter terminology. This means that we use the terms left *truncation* and left *censoring* to refer to situations for which the origin date of a spell in progress at the start of the observation period is known and unknown, respectively. In a similar manner, we define *right-censoring* as the situation arising when the complete duration is not observed due to the finiteness of the observation period.

2.2. Duration variables for spells in progress at the first survey date

Define τ to be calendar time (measured on the horizontal axis in figure 1) with τ_0 and τ_1 representing the beginning and the end of the observation period, respectively. For convenience we assume $\tau_0 = 0$. We denote with T_e the length of time from τ_0 to the beginning of the spell in which an individual was observed at calendar time τ_0, the elapsed duration (see for instance the case C in figure 1). Similarly, we denote with T_r the length of time from τ_0 to the time when the individual leaves the initial state for the first time, the remaining duration.[1]

Occupying a particular state at the beginning of the observation period implies having entered it at some previous date $-t_e$. An individual may at one time occupy one of J different states, we refer with $U(\tau)$ to the stochastic process in continuous time describing state occupancy.

[1] In most of the analytical part of the paper we will ignore the possibility of right-censoring. All the expression are straightforwardly modified to allow for right-censoring.

$$\begin{cases} U(\tau) = u_j & \Leftrightarrow \text{ the individual is in the } j'\text{th state at calendar time } \tau \\ U(\tau) \neq u_j & \Leftrightarrow \text{ the individual is not in the } j'\text{th state at calendar time } \tau \end{cases}$$

We can now formally define the elapsed and remaining durations, assuming that the state of interest is the state u_0

$$\begin{aligned} T_e &= -\sup\{\tau < 0 | U(\tau) \neq u_0\} \\ T_r &= \sup\{\tau > 0 | U(\tau) = u_0\} \end{aligned}$$

The full duration of a spell which is in progress at the survey date can be thus defined as $T = T_e + T_r$.[2]

2.3. Complete and incomplete data designs

We can distinguish among various situations regarding the observability of T_e and T_r. We represent them graphically in figure 1 below, where each line represents a spell experienced by some individual. The duration of the spell is measured by the length of the line ($T = T_e + T_r$). In particular, a solid line represents the observed part of the spell, while a dashed line represents the portion of the spell that is unobserved.

[*Figure* 1 to be inserted here]

We turn now to the analysis of the different situations illustrated in figure 1.

A) No censoring: The observation A is complete: the full length of the spell after the sampling date is known and no problem of either right or left censoring or truncation arises.

B) right-censoring: T_r is not observed (situations B_1 and B_2). Given that the observation period is finite, some events may be still in progress at its end and therefore are only

[2] Some authors use the terms "backward" and "forward" recurrence times to refer to the duration variables described above. In order to retrieve the useful distributions they use the properties of renewal processes, with the term renewal corresponding to the replacement of one person by his successor. See for example Cox (1962), Baydar and White (1988), and Lancaster (1990).

partially observed (right-censored). This form of missing data is routinely handled in event history analysis. Right-censored data may also occur due to the loss of follow-up, i.e. to nonresponse; in this situation right-censoring appears during the observation period (case B_2). If censoring is the result of a random process, the censored observation can be treated technically as a standard right-censored one.[3] If right-censoring is non-random (as censoring occurring because of attrition in a panel study), corrections should be introduced in order to avoid selectivity to bias the parameters of interest (see Rubin, 1987; Horowitz and Manski, 1996).

C) left-censoring: T_r is observed but T_e is not observed. This situation arises in many longitudinal surveys (as for instance in the French Labour Force Survey) where at each survey date individuals are asked to report the states they occupy at the date of the survey and only up to some time prior to that date. As an example consider the wave 1990 of the French Labour Force Survey. In January 1990 (the survey date), individuals were asked to declare their status on the labour market since January 1989. This means that only the state occupied at that date and those entered afterwards have been reported. Thus, owing to the lack of additional questions about the entry date in the state, the starting date of the spells that were in progress in January 1989 is unknown. Left-censoring is a part of the "initial conditions" problem in event history analysis (Heckman, 1981; Flinn and Heckman, 1982a,b; Ridder,1984; Heckman and Singer, 1984, 1986).

D) left-censoring and right-censoring: both T_e and T_r are unobserved. The observations arising in this case are both right and left-censored and occur, for instance, in panel studies in which labour market histories are recorded. Given a solution for left-censored observation, this case is routinely handled..

E) full right-censoring: Observation E is completely censored on the right. Entry and exit into the spells occurs after the observation period.

F) full left-censoring : Observation F is fully censored on the left, which means that the

[3] For instance in the analysis of marriage, the occurence of death by an accident may be regarded as independent of the hazard of getting married (Yamaguchi, 1991).

starting and ending dates are located before the beginning of the observation period. These spells, like those described in case E, are not observed.

G) Left-truncation: Both T_e and T_r are observed (case G_1). This happens when for each person the date of entry into the initial state is known and the follow-up lasts until the spell ends. Sometimes we have sufficient retrospective information about the time of entry into the initial state and thus we can reconstruct the complete spell of this individual. The case of incomplete observation of T_r due to right-censoring (case G_2) can be also arise. This situation is, once again, routinely handled given a solution for the case G_1.

H) Left-truncation and full right-censoring: T_e is observed and T_r is not observed at all (situation H in fig. 1). The observation is thus left-truncated and right-censored. Sampling from the stock of people officially registered as unemployed at a unique date and not conducting a follow-up study is an example of this truncation mechanism (see Nickell, 1979).

In this paper we are concerned mainly with the cases described in C, D, G_1 and G_2 above, i.e. with spells that are in progress at the beginning of the observation period. In general, one can assume that the state occupied at a certain date is the result of the entire history of the individual since he/she left the schooling system. The previous work on the initial conditions problem and on the distribution of duration data (Heckman, 1981; Flinn and Heckman, 1982a, b; Ridder, 1984; Heckman and Singer, 1984, 1986; Lancaster, 1990; Hamerle, 1991; Amemiya, 1999) provides an important framework within which various solutions to it can be better understood; we rely therefore on this literature in the exposition.

Data on individuals' behaviour over time can be collected essentially in two ways. When it is possible to observe individual processes from the very beginning (as in the study of the changes of labour market status of a group of school leavers) the sample likelihood is equal to the likelihood of the stochastic process describing the individual behaviour over time. Given that the beginning of the stochastic process is observed, the problem of initial conditions does not arise.

When such information is not available (as it is usually the case), data are collected by sampling stochastic processes that have been in progress for some time and by recording the history of the sample processes over a specified period. In this case information comes from two different sources, i.e. the evolution of the process and the observation period: both of them must be taken into account to build the appropriate likelihood that has to be based on the joint distribution of the past and on the future of the process as seen at the time of sampling (Ridder, 1984).

If the analyst is interested in the exit rate out of a particular state, the population can be sampled mainly in two ways. One way is to randomly sample the members of a population at a fixed point of time τ_0 (e.g. the stock of unemployed at τ_0). Another way is to sample the set of people when they enter (or leave) the state during an interval of time. We call the first scheme "*stock sampling*" (Lancaster and Chesher, 1981; Lancaster, 1990; De Toldi, Gouriéroux and Monfort, 1992). In this case one samples from a particular state and observes the length of time that individuals subsequently spend in that state.

The second scheme is called "*flow sampling*" (Lancaster and Chesher, 1981; Lancaster, 1990); the entrants to (or the leavers of) the state during a specified period are sampled. In this case one observes the starting time for spells of individuals who enter (leave) the state of interest during a certain period.

Spells depicted in cases C, D, and G in the above figure are stock-sampled. They share features of selective samples, that is, samples affected by some selection mechanism (see Ridder, 1984; Chesher and Lancaster, 1983). To see why this point is important consider the following: Let y be a scalar dependent variable and $\mathbf{x}$ a vector of covariates. Let $f(y, \mathbf{x}; \boldsymbol{\beta})$ be the joint population distribution of $(y, \mathbf{x})$, with $\boldsymbol{\beta}$ being a vector of parameters. In empirical investigations the analyst specifies a model for y that is usually described by the conditional distribution $f_1(y|\mathbf{x}, \boldsymbol{\beta}_1)$, with $\boldsymbol{\beta}_1$ being a subset of the vector $\boldsymbol{\beta}$. If $\mathbf{x}$ is exogenous to y, the population density $f(.)$ can be decomposed as follows (Ridder, 1984)

$$f(y, \mathbf{x}; \boldsymbol{\beta}) = f_1(y|\mathbf{x}; \boldsymbol{\beta}_1) f_2(\mathbf{x}; \boldsymbol{\beta}_2)$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have no elements in common and satisfy $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\}$ In other words $\mathbf{x}$ is ancillary for $\boldsymbol{\beta}_1$ (Cox and Hinkley, 1974; Ridder, 1984). This implies that inference on $\boldsymbol{\beta}_1$ can be based without loss of information on the conditional distribution $f_1(y|\mathbf{x}, \boldsymbol{\beta}_1)$.

In the case of selective samples (like stock-sampled duration data) where the selection rule is a consequence of a particular sample design, the sample density does not factorize, and the likelihood based on it depends on the generally unknown distribution of the explanatory variables. More precisely the distribution of explanatory variables become informative about the parameters of interest. Ignoring stock-sampling may result in biased estimates for the parameters of the duration distribution.

With reference to the cases C, D, G_1, G_2 and H illustrated in figure 1 above, we can say that the full length of a sampled spell will exceed the partial length measured by the survey, i.e. $T \geq T_r$. This phenomenon is called interruption-bias by Salant (1977).[4] Figure 1 also suggests that spells with longer than average full lengths are more likely to be in progress at the survey date, and this phenomenon is known as length-biased sampling (Salant, 1977; Cox, 1962).

The outlined concepts imply that a clear distinction must be drawn between the random variable "duration" that is sampled in existing surveys and the different concept that the job-search theorists label with the same name.

3. The treatment of initial conditions

The transition rate from state u_0 to state u_j, conditional on explanatory variables $\mathbf{x}$ (that we assume to be time-invariant) can be written as

$$h_j(t|\mathbf{x}) = \lim_{\Delta \to 0} \frac{\Pr\{t \leq T < t + \Delta, U(T) = u_j | T \geq t, \mathbf{x}\}}{\Delta}$$

where t denotes the time spent in state u_0, and $U(T)$ denotes the stochastic process described above, but here expressed in terms of the duration in the state u_0. The hazard rate is the sum of the transition intensities towards the $J-1$ attainable destinations

$$h(t|\mathbf{x}) = \sum_{j=1}^{J-1} h_j(t|\mathbf{x})$$

The distribution function of T, $F(t|\mathbf{x})$, writes

$$F(t|\mathbf{x}) = 1 - \exp\left\{-\int_0^t h(s|\mathbf{x})ds\right\}$$

and the survivor function is

$$S(t|\mathbf{x}) = 1 - F(t|\mathbf{x})$$

[4] What is observed is a part of an interrupted spell.

In addition, we define the destination-specific sub-density to be

$$f_j(t|\mathbf{x}) = h_j(t|\mathbf{x}) \cdot S(t|\mathbf{x})$$

3.1. The sample density of a stock sample

We will focus on the cases C, D, G_1 and G_2, in figure 1. The crucial point is that individuals are in a particular state at a specific sampling date. In order to derive the likelihood contributions for the spells sampled under this particular sampling scheme we need to find the joint conditional distribution of T_e and T_r where the word conditional refers to the presence of the individual in a particular state at the beginning of the observation period. In the following sections, we derive formally this joint density under alternatives assumptions concerning the observability of T_e.

We start the exposition with the case of left-truncation (cases G_1 and G_2). For this case, we observe both the relevant duration variables necessary to build the joint density of the initial spell. For the left-censoring cases C and D in figure 1, we need to derive the marginal densities for the observable parts of the spells from this conditional joint density.

3.1.1. Left-truncation

Being in state u_0 at the beginning of the observation periods is the condition for being sampled. Let E_1 correspond to the event "the individual enters the initial state at time $-t_e$ and leaves it at time t_r"

$$E_1 = \left\{U(-t_e^-) \neq u_0,\ \forall\tau \in [-t_e, t_r)\ : U(\tau) = u_0,\ U(t_r) \neq u_0\right\}$$

and let E_2 correspond to the event "the individual enters the initial state at time $-t_e$ and stays at least until t_r" (i.e. the spell is right-censored)

$$E_2 = \left\{U(-t_e^-) \neq u_0,\ \forall\tau \in [-t_e, t_r) : U(\tau) = u_0\right\}$$

say $E = \{E_1, E_{2.}\}$

Consider the case of an individual i being in the state u_0 at the sampling date $\tau_0 = 0$. The joint density of T_e and T_r conditional on selection, say $\Pr[E|u_0, \mathbf{x}])$, where u_0 is taken to be synonymous of the event $U(0) = u_0$, can be written as

$$\begin{aligned}\Pr\left[E|\mathbf{x},u_0\right] &= \frac{\Pr\left[E \cap u_0|\mathbf{x}\right]}{\Pr\left[u_0|\mathbf{x}\right]} \\ &= \frac{\Pr\left[E|\mathbf{x}\right]}{\Pr\left[u_0|\mathbf{x}\right]} \qquad (3.1)\end{aligned}$$

Denote now the conditional probabilities governing the stochastic process $U(\tau)$ by $\Pr\left\{U(\tau_b)|U(\tau_a)\right\}$, for all $\tau_a < \tau_b$. Owing to the characteristics of the process, we define a jump at time τ as $U(\tau^-) \neq U(\tau)$. The entry rate into the state u_0 is now defined by

$$e(-t_e|\mathbf{x}) \equiv \lim_{\Delta \to 0} \frac{\Pr\left[U(-t_e) = u_0|U(-t_e - \Delta) \neq u_0, \mathbf{x}\right]}{\Delta}$$

For E_1, the joint conditional density $\Pr\left[E_1|u_0, \mathbf{x}\right]$ can now be derived in the following way: Define

$$\begin{aligned}A(t_e, t_r|\mathbf{x}) &\equiv \Pr\left\{T_e > t_e, T_r \geq t_r|\mathbf{x}\right\} \\ &= \int_{t_e}^{\infty} e(-u|\mathbf{x}) S(t_r + u|\mathbf{x}) du \qquad (3.2)\end{aligned}$$

Evaluate $\Pr\left\{T_e = t_e, T_r = t_r|\mathbf{x}\right\}$ by taking the derivatives of (3.2) with respect to t_e and t_r,

$$\frac{\partial^2 A\left(t_e, t_r|\mathbf{x}\right)}{\partial t_r \partial t_e} = e\left(-t_e|\mathbf{x}\right) f\left(t_r + t_e|\mathbf{x}\right) \qquad (3.3)$$

Compute $P_0(\mathbf{x}) = \Pr\left[u_0|\mathbf{x}\right]$ by evaluating $A(0,0|\mathbf{x})$

$$\begin{aligned}P_0(\mathbf{x}) &= A(0,0|\mathbf{x}) \qquad (3.4) \\ &= \int_0^{\infty} e(-u|\mathbf{x}) S(u|\mathbf{x}) du\end{aligned}$$

which is the probability of being in the state u_0 at time $\tau_0 = 0$. Finally, compute

$$\Pr(E_1|\mathbf{x},u_0) = \frac{A(t_e, t_r|\mathbf{x})}{A(0,0|\mathbf{x})} \qquad (3.5)$$

The above expression writes as

$$g(t_e, t_r|\mathbf{x}, u_0) = \frac{e(-t_e|\mathbf{x}) f(t_e + t_r|\mathbf{x})}{\int_0^{\infty} e(-u|\mathbf{x}) S(u|\mathbf{x}) du} \qquad (3.6)$$

In the presence of multiple destinations, the density function $f(t_e + t_r|\mathbf{x})$ in (3.6) is replaced by $f_j(t_e + t_r|\mathbf{x})$, such that expression (3.6) now writes $f_j(t_r + t_e|\mathbf{x}) e(-t_e|\mathbf{x})$. The same holds for the expression in (3.4) where we have to sum over the $J - 1$ states attainable from u_0.

Substituting for the expressions found, the joint density of T_e, T_r, $U(t_r) = u_j$, conditional on covariates $\mathbf{x}$ and on u_0, writes

$$g(t_e, t_r, u_j|\mathbf{x}, u_0) = \frac{e(-t_e|\mathbf{x}) f_j(t_e + t_r|\mathbf{x})}{\int_0^\infty e(-u|\mathbf{x}) S(u|\mathbf{x}) du} \tag{3.7}$$

This density function represents the point of departure to build the likelihood contributions of both left-truncated and left-censored spells.

In the case of a right-censored and left-truncated spell, that is, event E_2 defined above, it is straightforward by going through the same steps to verify that

$$\Pr(E_2|\mathbf{x}, u_0) = \frac{e(-t_e|\mathbf{x}) S(t_e + t_r|\mathbf{x})}{\int_0^\infty e(-u|\mathbf{x}) S(u|\mathbf{x}) du}$$

The expression in (3.7) is, in its general form, intractable. To find a solution, knowledge of the entry rate, or some specific assumptions about it, is required.

3.1.2. Left-Censoring

In the cases C and D illustrated in figure 1 above, the elapsed duration is unknown.

In both situations, owing to the unobservability of T_e we need to integrate (3.7) over T_e in order to obtain the correct density function that writes

$$g_{T_r}(t_r|\mathbf{x}, u_0) = \frac{\int_0^\infty e(-u|\mathbf{x}) f_j(u + t_r|\mathbf{x}) du}{\int_0^\infty e(-u|\mathbf{x}) S(u|\mathbf{x}) du} \tag{3.8}$$

It follows that either knowledge of the entry rate or specific assumptions about it are necessary to make the density in (3.8) tractable. In the next section some of the solutions proposed in the literature to overcome this problem will be illustrated.

3.2. Some solutions to the left-censoring problem

3.2.1. The conditional likelihood approach as a solution to left-truncated data

When data about T_e and T_r are available, the conditional likelihood approach of Lancaster (1979) may be implemented. This methodology avoids the complications linked to the knowledge of the pre-observation period and efficiently uses the available information in the observation period. In particular, the explicit consideration of entry rates is unnecessary since this approach focuses on the distribution of duration conditional on the pre-interview duration t_e.

The likelihood construction is thus based on the conditional distribution of T_r given T_e, the vector of characteristics of the individuals, $\mathbf{x}$, and the event $U(0) = u_0$.

The joint density $g(t_e, t_r, u_j|\mathbf{x},u_0)$ can be decomposed (see Ridder, 1984; Goto, 1996) into a conditional and marginal component

$$g_c(t_r, u_j|t_e, \mathbf{x},u_0) = \frac{g(t_e, t_r, u_j|\mathbf{x},u_0)}{g_m(t_e|\mathbf{x},u_0)} \tag{3.9}$$

This implies to isolate the entire effect of the initial conditions within the marginal probability $g_m(t_e|\mathbf{x},u_0)$, allowing us to use the conditional density $g_c(t_r, u_j|t_e, \mathbf{x},u_0)$ alone to determine the relevant parameters. The marginal component $g_m(t_e|\mathbf{x}, u_0)$ is obtained from the joint conditional density in (3.7) by summing over all attainable states from the initial state u_0 and integrating over all possible values of the post-interview duration T_r

$$\begin{aligned} g_m(t_e|\mathbf{x},u_0) &= \sum_{j=1}^{J-1} \int_0^\infty g(t_e, v, u_j|\mathbf{x}, u_0) dv \\ &= \frac{e(-t_e|\mathbf{x})S(t_e|\mathbf{x})}{P_0(\mathbf{x})} \end{aligned}$$

Substituting for $g(.)$and $g_m(.)$ in (3.9) we get

$$g_c(t_r, u_j|t_e, \mathbf{x}, u_0) = \frac{f_j(t_e + t_r|\mathbf{x})}{S(t_e|\mathbf{x})} \tag{3.10}$$

We notice from the above expression that the information contained in t_e is merely used to eliminate the entry rate (see Ridder, 1984; Hamerle, 1991).

The conditional density $g_c(.)$ can be expressed in terms of the hazard and the destination specific transition rate,

$$g_c(t_r, u_j|t_e, \mathbf{x}, u_0) = h_j(t_e + t_r|\mathbf{x}) \exp\left[-\int_{t_e}^{t_e+t_r} h(u|\mathbf{x})du\right] \tag{3.11}$$

The maximum likelihood estimators proposed by Lancaster (1979) are based on this conditional density. Goto (1996) shows that this estimator coincides with the semi-parametric MLE of a model for left-truncated data, and that it achieves the semi-parametric efficiency bound.[5] As noticed for instance in Guo (1993), the conditional likelihood approach appears

[5] If we introduce a multiplicative unobserved heterogeneity component, ε, it can be seen that an assumption on the entry rate is necessary. In particular by assuming

$$e(-t_e|\mathbf{x}, \varepsilon) = k_1(\mathbf{x}, t_e)k_2(\varepsilon)$$

to be more efficient when compared to the assumption of stationarity (see below). Moreover, it allows to introduce time-varying covariates in the model. Indeed one only needs to know the value of the time-varying covariates during the observation period.

3.2.2. Stationarity assumption

When the elapsed duration, t_e, is not observed, an assumption that is commonly adopted is to consider the entry rate as constant over time. Indeed, if we consider the joint density in (3.7), we remark that this expression will not depend on the entry rate if stationarity is assumed. In this case, substituting for $e(-t_e|\mathbf{x}) = e(\mathbf{x})$ in (3.7), we get

$$\begin{aligned} g(t_e, t_r, u_j|\mathbf{x}, u_0) &= \frac{e(\mathbf{x}) f_j(t_e + t_r|\mathbf{x})}{e(\mathbf{x}) \int_0^\infty S(u|\mathbf{x}) du} \\ &= \frac{f_j(t_e + t_r|\mathbf{x})}{E[T|\mathbf{x}]} \end{aligned} \tag{3.12}$$

In the case of left-censoring, the stationarity assumption leads to the following expression obtained from (3.12) by integrating out T_e

$$g_{T_r}(t_r, u_j|\mathbf{x}, u_0) = \frac{\int_0^\infty f_j(u + t_r|\mathbf{x}) du}{E(T|\mathbf{x})} \tag{3.13}$$

This expression generalizes Ridder (1984) result for a single destination state in which case the previous expression (3.13) writes

$$g_{T_r}(t_r|\mathbf{x}, u_0) = \frac{S(t_r|\mathbf{x})}{E(T|\mathbf{x})} \tag{3.14}$$

This density corresponds to $f(t|\mathbf{x})$ only in the special case of an exponentially distributed duration, i.e. when the hazard is constant over time.[6]

we can write the conditional density as

$$g_c(t_r, u_j|t_e, \mathbf{x}, u_0) \propto \int_\epsilon f(t_e + t_r|\mathbf{x},\varepsilon) k_2(\varepsilon) v(\varepsilon) d\varepsilon \equiv \int_\epsilon f(t_e + t_r|\mathbf{x},\varepsilon) k^*(\varepsilon) d\varepsilon$$

where ϵ denotes the support of ε.

[6] Another approach to dealing with left-censored observations is that of Nickell (1979), which we will not pursue here.

3.3. Likelihood

Suppose we sample two types of spells; spells that are in progress at the beginning of the observation period (left-censored spells) and spells which begin during the observation period (fresh spells). These cases correspond to stock and flow samples, respectively, but here we consider the combined stock and flow sample of spells. Denote with n_1 the left-censored spells and with $n_2 = n - n_1$ the fresh spells. For fresh spells we observe both starting time τ and the duration t.

Let

$$P_1(\mathbf{x}) = \int_0^{\tau_1} e(u|\mathbf{x})du$$

denote the probability of observing a fresh spells.

Consider the following likelihood function

$$\mathcal{L} = \prod_{i=1}^{n_1} g_{T_r}(t_i|\mathbf{x}_i, u_0)\frac{P_0(\mathbf{x}_i)}{P_0(\mathbf{x}_i)+P_1(\mathbf{x}_i)}\prod_{i=1}^{n_2}\frac{e\left(\tau_i|\mathbf{x}_i\right)h(t_i|\mathbf{x}_i)S(t_i|\mathbf{x}_i)}{P_1(\mathbf{x}_i)}\frac{P_1(\mathbf{x}_i)}{P_0(\mathbf{x}_i)+P_1(\mathbf{x}_i)} \tag{3.15}$$

versus

$$\mathcal{L}^c = \prod_{i=1}^{n_1} g_{T_r}(t_i|\mathbf{x}, u_0)\prod_{i=1}^{n_2}\frac{e\left(\tau_i|\mathbf{x}_i\right)h(t_i|\mathbf{x}_i)S(t_i|\mathbf{x}_i)}{P_1(\mathbf{x}_i)} \tag{3.16}$$

The expression (3.15), suggested by Amemiya (1999), is the full likelihood of the sample of n individuals. It consists of separate expressions for stock and flow sampled observations, multiplied by the probabilities that an observed spell is either stock or flow sampled. The expression (3.16) is a conditional likelihood, in the sense that it is conditional on the type of spell observed. Amemiya (1999) shows that both estimators are consistent, but of course only estimators based on the full likelihood are efficient. He also demonstrates that the estimator based on the flow sample (that is, based on only the n_2 fresh spells) while being consistent is inefficient.

Making once more the assumption of a constant entry rate, we find that the expression (3.15) simplifies. By inserting previously derived expressions, we find

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{n_1}\frac{S(t_i|\mathbf{x}_i)}{E(T|\mathbf{x}_i)}\frac{e(\mathbf{x}_i)E[T|\mathbf{x}_i]}{e(\mathbf{x}_i)E[T|\mathbf{x}_i]+\tau_1 e(\mathbf{x}_i)}\prod_{i=1}^{n_2} e\left(\mathbf{x}_i\right)h(t_i|\mathbf{x}_i)S(t_i|\mathbf{x}_i)\frac{1}{e(\mathbf{x}_i)E[T|\mathbf{x}_i]+\tau_1 e(\mathbf{x}_i)}\\ &= \prod_{i=1}^{n} h(t_i|\mathbf{x}_i)^{d_i}S(t_i|\mathbf{x}_i)\frac{1}{E[T|\mathbf{x}_i]+\tau_1} \end{aligned} \tag{3.17}$$

where d_i is an indicator for a fresh spell; τ_1 denotes the length of the observation period, and t_i is the duration of a fresh spell (if $d_i = 1$) or the remaining duration (if $d_i = 0$). In the presence of right-censoring, an obvious modification of this expression is easily derived.

3.3.1. Improving efficiency further in the stationary model

Now, assume that we have additional knowledge regarding the elapsed duration. Specifically, we know for sure that $T_e \leq \bar{t}$. For example; $\bar{t}$ could be time since age 16, or time since labour market entry, or some other known limitation. If we study unemployment duration of youth, we may have the additional information that an individual has had three year of working experience before the current unemployment spell. This leads to a further reduction in the upper bound for the elapsed duration. There is potentially plenty of information that can be used to bound the elapsed duration in this way. The purpose of this 'bounding exercise' is twofold; first, it leads to an obvious gain in efficiency, and secondly, it may contribute to making the stationarity assumption less invalid (or perhaps even valid).

In order to derive the likelihood function in the case of an upper bounded elapsed duration, we define

$$\begin{aligned} A^*(t_e, t_r|\mathbf{x}) &= \Pr\left\{\bar{t} \geq T_e > t_e, T_r \geq t_r|\mathbf{x}\right\} \\ &= \int_{t_e}^{\bar{t}} e(-u|\mathbf{x})S(t_r + u|\mathbf{x})du \end{aligned} \tag{3.18}$$

We now have that

$$\Pr\left\{\bar{t} \geq T_e > t_e, T_r \geq t_r|\mathbf{x}, u_0, T_e \leq \bar{t}\right\} = \frac{A^*(t_e, t_r|\mathbf{x})}{\Pr\left(U(0) = u_0, T_e \leq \bar{t}|\mathbf{x}\right)} \tag{3.19}$$

In addition

$$\begin{aligned} P_0^*(\mathbf{x}) &= A^*(0, 0|\mathbf{x}) \\ &= \Pr\left(U(0) = u_0, T_e \leq \bar{t}|\mathbf{x}\right) \\ &= \int_0^{\bar{t}} S(u|\mathbf{x})e(-u|\mathbf{x})du \end{aligned} \tag{3.20}$$

Hence, we have that

$$\begin{aligned} g\left(t_e, t_r|\mathbf{x}, u_0, T_e \leq \bar{t}\right) &= \frac{\partial^2 A^*(t_e, t_r|\mathbf{x})}{\partial t_e \partial t_r} \frac{1}{\int_0^{\bar{t}} S(u|\mathbf{x})e(-u|\mathbf{x})du} \\ &= \frac{f\left(t_e + t_r|\mathbf{x}\right) e\left(-t_e|\mathbf{x}\right)}{\int_0^{\bar{t}} S(u|\mathbf{x})e(-u|\mathbf{x})du} \end{aligned} \tag{3.21}$$

Now, exploit once again the stationarity assumption to obtain

$$g\left(t_e, t_r|\mathbf{x}, u_0, T_e \leq \bar{t}\right) = \frac{f\left(t_e + t_r|\mathbf{x}\right)}{\int_0^{\bar{t}} S(u|\mathbf{x})du} \tag{3.22}$$

Finally, for left-censored spells, with the known maximum of the elapsed duration, we obtain

$$g_{T_r}(t_r|\mathbf{x}, u_0, T_e \leq \bar{t}) = \frac{S\left(t_r|\mathbf{x}\right) - S\left(t_r + \bar{t}|\mathbf{x}\right)}{\int_0^{\bar{t}} S(u|\mathbf{x})du} \tag{3.23}$$

For exponential and - more importantly - piecewise constant hazard distributions, (3.23) has simple analytical expressions, as we show in one of the empirical applications. Denoting the parameter vector to be estimated by ψ, the likelihood function becomes

$$\begin{aligned}\mathcal{L}(\psi) &= \prod_{i=1}^{n_1} g(t_{r,i}|\mathbf{x}_i, u_0, T_e \leq \bar{t})\frac{P_0^*(\mathbf{x}_i)}{P_0^*(\mathbf{x}_i) + P_1(\mathbf{x}_i)} \prod_{i=1}^{n_2} e\left(\tau_i|\mathbf{x}_i\right) h(t_i|\mathbf{x}_i) S(t_i|\mathbf{x}_i) \frac{1}{P_0^*(\mathbf{x}_i) + P_1(\mathbf{x}_i)} \\ &= \prod_{i=1}^{n_1} \frac{S\left(t_r|\mathbf{x}\right) - S\left(t_r + \bar{t}|\mathbf{x}\right)}{\int_0^{\bar{t}} S(u|\mathbf{x})du + \tau_1} \prod_{i=1}^{n_2} h(t_i|\mathbf{x}_i) S(t_i|\mathbf{x}_i) \frac{1}{\int_0^{\bar{t}} S(u|\mathbf{x})du + \tau_1} \end{aligned} \tag{3.24}$$

In the empirical application concerning unemployment spells, we will apply this likelihood specification.

3.3.2. Testing for stationarity

Since we have made the rather restrictive assumption of a constant entry rate, it seems appropriate to construct a test of its validity. An appropriate test is the Hausman test. Under the null hypothesis of stationarity, the likelihoods (3.15) and (3.24) lead to estimates that are consistent and efficient. If the null fails to hold, the likelihood based on the flow sample

$$\mathcal{L}^{flow}(\psi) = \prod_{i=1}^{n_2} e\left(\tau_i|\mathbf{x}_i\right) h(t_i|\mathbf{x}_i) S(t_i|\mathbf{x}_i) \frac{1}{P_1(\mathbf{x}_i)} \tag{3.25}$$

still leads to consistent, but inefficient estimates. Assuming that the entry rate and the hazard rate have no parameters in common, estimation may be based on maximization of

$$\mathcal{L}^{flow'}(\psi) = \prod_{i=1}^{n_2} h(t_i|\mathbf{x}_i) S(t_i|\mathbf{x}_i) \tag{3.26}$$

Denote the parameter vector estimated under the full likelihood by $\widehat{\psi}_{full}$ and the one obtained using only the flow sample by $\widehat{\psi}_{flow}$.

Hence, a Hausman test of the stationarity assumption may be based on

$$\mathcal{HS} = \left(\widehat{\psi}_{flow} - \widehat{\psi}_{full}\right)' \left(\widehat{V}_{flow} - \widehat{V}_{full}\right)^{-1} \left(\widehat{\psi}_{flow} - \widehat{\psi}_{full}\right) \tag{3.27}$$

which is asymptotically chi-squared with degrees of freedom equal to the number of parameters in ψ.

4. Empirical Applications

In this section we demonstrate some of the solutions presented above to left-censored and left-truncated single spell duration data. The first empirical application, in section 4.1, analyses left-truncated schooling durations of French young people, while the second looks at left-censored unemployment durations of the same population. In doing this we exploit the data extracted from the waves 1990-1992 of the French Labour Force Survey (FLFS).

The FLFS is a survey conducted annually by the INSEE. For the available data set, interviews were carried out on three dates, in January 1990, March 1991 and March 1992. The data we use in this study consists of 5824 young persons aged 18-29 in 1992. They are asked (in the 'Module Jeunes' 1992) to give more details about their occupational history and on family and individual status since they were 16 and until the beginning of the observation period. Moreover, the state they occupy at the date of the survey, and during each month of the previous year, is declared. This information allows us to construct the history of each individual since January 1989.

4.1. The analysis of left-truncated schooling durations

In this section we analyse schooling durations of the individuals who are in the schooling system at the beginning of the observation period (i.e. January 1989) and that are left-censored. Thanks to the available information, we can reconstruct for these spells their elapsed duration, and therefore study their duration in the conditional likelihood approach of Lancaster (1979).

4.1.1. The construction of the dataset

To analyse transitions out of the schooling system, we have selected all individuals who were in school in January 1989. This results in a sample of 2947 young individuals, 1460 men and 1487 women.

The structure of the data is such that we only know the length of stay of each individual since January 1989; no information is available before this date. However, we do have some indirect knowledge that we may use, and we can make a few assumptions. Assume that the schooling system cannot be re-entered once it has been left.[7] The minimum school-leaving age is 16 years. Consequently, the schooling duration since age 16 is the variable we are interested in. We assume further that individuals can not exit the education system during the academic year of their sixteenth birthday. We thus take the start of schooling duration to be the 1st of October in the year of the 16th birthday when born between January and September (we know the month and year of birth of each individual) and to be the 1st of October in the following year for those born between October and December. For an individual who was 16 in January 1986 and is observed in education at the beginning of the observation period, we take thus the 1st of October 1986 as the beginning of her schooling duration. This implies that t_e is equal to 27 months. Acknowledging the possibility of measurement error, we subsequently groups the data into annual durations.

4.1.2. The conditional likelihood approach for grouped duration data

As shown above, the conditional likelihood approach of Lancaster (1979) considers the distribution of the remaining duration conditional on the pre-interview duration, t_e, which is basically used to eliminate the entry rate from the likelihood.

In order to estimate the probability of leaving the schooling system of French young individuals, we use a model for grouped duration data, since that is what we have.

The underlying continuous durations are only observed in k disjoint time intervals $[0, t_1)$, $[t_1, t_2)$, $[t_2, t_3)$,...., $[t_{K-1}, \infty)$. The probability of exit in the k'th interval for person i,

[7] Re-entry is negligible as it is also shown in Magnac (2000).

conditional on survival until the beginning of the k'th interval, is

$$\begin{aligned} h_{ik}(\mathbf{x}) &\equiv \Pr\{T \in [t_{k-1}, t_k) | T \geq t_{k-1}, \mathbf{x}_i\} \\ &= 1 - \frac{S(t_k|\mathbf{x}_i)}{S(t_{k-1}|\mathbf{x}_i)} \end{aligned} \qquad (4.1)$$

Following Kiefer (1988a) and Meyer (1990,1995), we assume proportional hazards and rewrite (4.1) as

$$h_{ik}(\mathbf{x}) = 1 - \exp[-\exp(\mathbf{x}_i^{'}\boldsymbol{\beta} + \gamma_k)] \qquad (4.2)$$

with $\gamma_k = \ln \int_{t_{k-1}}^{t_k} \lambda_0(u)du$.Grouped duration data hazard models have an appealing relationship to binomial models as already noticed in Allison (1982), Kiefer (1988b, 1990) and Jenkins (1995). Define a new individual- and interval-specific indicator variable

$$y_{ij} = 1\left\{t_j - 1 \leq T < t_j\right\} \qquad (4.3)$$

Conditioning the likelihood on having survived up to t_e periods, taking logs and multiplying over the sample, consisting of n individuals, we obtain

$$\log \mathcal{L} = \sum_{i=1}^{n} \sum_{j=1}^{t_i} y_{ij} \log \left(\frac{h_{ij}(\mathbf{x})}{1 - h_{ij}(\mathbf{x})} \right) + \sum_{i=1}^{n} \sum_{j=t_e+1}^{t_i} \log\left(1 - h_{ij}(\mathbf{x})\right) \qquad (4.4)$$

where t_i denotes the sum of elapsed and remaining duration, see e.g. Jenkins (1995). Call this Model A. Model B neglects left-truncation and treats the observations as drawn from the inflow into schooling. This corresponds to the following likelihood

$$\log \mathcal{L} = \sum_{i=1}^{n} \sum_{j=1}^{t_i} y_{ij} \log \left(\frac{h_{ij}(\mathbf{x})}{1 - h_{ij}(\mathbf{x})} \right) + \sum_{i=1}^{n} \sum_{j=1}^{t_i} \log[1 - h_{ij}(\mathbf{x})] \qquad (4.5)$$

4.1.3. Results

The objective pursued in this section is to present some solutions to the initial conditions problem when single-spell data are available. Therefore, we consider only two explanatory variables. Descriptive statistics are in table 1. They are the nationality of the father of the individual and an indicator for being younger than 25 in 1992 - a cohort effect.

[$Table$ 1 to be inserted here]

In table 2 we compare the probabilities of leaving school estimated by the conditional likelihood approach (Model A) to those estimated by treating data as a flow sample (Model B) for men and women.[8].

[$Table$ 2 to be inserted here]

We can see that there exists differences in the estimated parameters on the covariates, not only in their values but also in the significance level. The age (cohort) effect in particular shows up to be different in the two specifications. In the model where the stock-sampling nature of the data is ignored, the cohort effect is such that the younger have a much higher exit rate from education. This is obviously because the sample of older individual consists only of those who are still in the schooling system at the time of entry into the sample, that is, the sub-sample among the older cohort which has stayed the longest in the schooling system. Correcting for this (Model A), we still find an age effect, but it is smaller. Both models show positive duration dependence, but the model that ignores left-truncation shows a remarkably different pattern over time. This is more easily observed in the following figures 2 and 3, where we have drawn the baseline hazards for men and women by putting the explanatory variables at their sample averages. The cause of the difference is well-known; by treating the total durations as a fresh sample, we have 'under-sampled' short durations. The conditional likelihood corrects for exactly this.

4.2. The analysis of left-censored unemployment spells

In this section, we want to analyse unemployment durations. Our sample consists partly of a stock sample, and partly of a flow sample. This survey design leads to a serious initial conditions problem owing to the unobservability of the beginning of the first spells for the stock sampled part of the data.

[8] The values of the log-likelihood are for the men equal to -2705.087 and -2885.347 in model A and B respectively. For women they are: -2443.385 and -2607.698 in model A and B respectively.

4.2.1. The construction of the data set and upper bounds on elapsed durations

The sample analysed is the same as above; the young individuals of the FLFS. 352 individuals are unemployed in January 1989, and 2080 flow into unemployment during the observation period. We have used the same type of information as we used above for schooling duration, to construct an upper bound on the elapsed unemployment duration: Individuals who are in the labour market at the beginning of the observation period left school before that date. We do not know when, but we know that they had to stay in school at least until they were 16. We can thus assume that they entered the labour market somewhere during the interval of time elapsed since that date until the beginning of the observation period. With these assumptions we can say that the maximum length of the unobserved pre-interview duration is the current age (measured with monthly precision) less 16 years. Denote this upper bound $\bar{t}_1$.

However, we know more; in the additional questionnaire "Module Jeunes" (conducted in 1992) the interviewed individuals were asked to declare the number of years spent in education since they were 16 until the beginning of the survey date, i.e. January 1989. Using this information, we can calculate the time of entry into the labour market. The time elapsed between this date and the beginning of the observation period is an upper bound on the elapsed duration, which is tighter than $\bar{t}_1$. Call this upper bound $\bar{t}_2$.

If we had access to more information, e.g. accumulated working experience, periods of maternity (and other types of) leave, military service, etc., we could narrow the interval for the elapsed duration even further.

Various explanatory variables have been used to assess their influence on the hazard rate. Specifically, we used the age of the individuals in 1992, their length of education, as well as indicators for the nationality of the father, being a women, having a technical education, belonging to a large family (more than two siblings), having had health problems in the childhood, having parents that are divorced, having a diseased father, and living in Paris. Most of them are frequently used in the study of unemployment determinants (see D'Addio, 1998; D'Addio, 2002). Descriptive statistics of the dependent variable, $\bar{t}_1$ and $\bar{t}_2$, and the explanatory variables used are presented in Table 3.

[*Table* 3 to be inserted here]

4.2.2. The likelihood function

We assume a proportional hazard specification, that is, we assume that the transition from unemployment to employment can be described by the hazard rate

$$h(t|\mathbf{x}) = \lambda(t) \exp [x\beta]$$

where $\lambda(.)$ is the baseline hazard. We make the further assumption that the baseline hazard is piecewise constant on each of K intervals with splitting times $t_0 = 0, t_1, t_2, ..., t_{K-1}, t_K = +\infty$. Define a function $M : \Re_+ \curvearrowright \{1, 2, ..., K\}$, which maps a duration, t, into one of the intervals defined by the splitting times above. We take the splitting times $t_1 = 1$, $t_2 = 2$,..., $t_{22} = 22$, and $t_{23} = +\infty$, that is, we assume that the hazard is constant after 22 months of unemployment (due to the small number of exits from unemployment after 22 months). The value of $\lambda(.)$ in the i'th interval is parameterized as $\exp [\gamma_i]$.

The likelihood functions needed for the estimation are given by expressions (3.15) and (3.24) for the full likelihood, without and with an upper bound on the elapsed duration, respectively, and (3.26) for the likelihood which is based solely on the flow part of the sample. The expressions in these functions are all fairly simple, using the specification set up above. The survival function is given by

$$\begin{aligned} S(t|\mathbf{x}) &= \exp \left[- \int_0^t h(s|\mathbf{x}) ds \right] \\ &= \exp[- \sum_{i=1}^{m(t)-1} h_i (\mathbf{x}) (t_i - t_{i-1}) - h_{m(t)}(\mathbf{x}) \left(t - t_{m(t)-1} \right) \qquad (4.6) \end{aligned}$$

where $h_i (\mathbf{x})$ denotes the value of the hazard rate in the i'th interval. The expected duration is given by

$$E(T|\mathbf{x}) = \sum_{i=1}^{K} \frac{1}{h_i(\mathbf{x})} \left[S (t_{i-1}|\mathbf{x}) - S (t_i|\mathbf{x}) \right] \qquad (4.7)$$

and the expression used when applying the upper bound $\int_0^{\bar{t}} S(u|\mathbf{x}) du$ may also be calculated easily

$$\int_0^{\bar{t}} S(u|\mathbf{x}) du = \sum_{i=1}^{m(\bar{t})-1} \frac{1}{h_i(\mathbf{x})} \left[S (t_{i-1}|\mathbf{x}) - S (t_i|\mathbf{x}) \right] + \frac{1}{h_{m(\bar{t})}(\mathbf{x})} \left[S(t_{m(\bar{t})-1}|\mathbf{x}) - S \left(\bar{t}|\mathbf{x} \right) \right] \qquad (4.8)$$

4.2.3. Results

In Table 4 we present the results of the estimation with four different models; model 1 corresponds to the model for the flow sample, that is, the likelihood (3.26), model 2 corresponds to the (Amemiya, 1999) likelihood function without an upper bound on the elapsed duration, e.g. (3.15) above, while models 3 and 4 both are using the model with an upper bound, the likelihood function (3.24). Model 3 uses the upper bound $\bar{t}_1$ (the least restrictive), while model 4 uses $\bar{t}_2$. In figure 4, we plot the baseline hazards corresponding to each of the four models, each evaluated at the average of $\exp[x\beta]$.

[$Table$ 4 to be inserted here]

[$Figure$ 4 to be inserted here]

The results show that there are some differences, particularly in the parameters of the baseline hazard around the 14th month. This is caused by a flaw in the data construction, which affects mainly those individuals who are stock-sampled; namely, in the reconstruction of the trajectories, the information about the status held in the month of February 1990 was missing. This problem was solved by imputing the information from the states occupied in January 1990. Consequently, there are no transitions in this interval (for the stock sampled individuals).

Another important thing to note is the uniform reduction in the standard errors when moving across the table from model 1 to model 4. It is also evident that the most important reduction in standard errors is in the inclusion of the stock-sample, that is, in going from model 1 to model 2. The inclusion of a tight upper bound on the elapsed duration does lead to a further reduction in the standard errors, but mostly so for the baseline parameters.

When we look at the coefficients on the explanatory variables, it is notable that some of those that were not significantly different from zero in the flow model are so in the model which includes the stock-sampled spells. Hence, the information contained in this relatively small stock sample is very useful. With access to a relatively larger stock sample or more information to tighten the upper bound on the elapsed duration, the information gain would obviously be much larger.

In Table 5, we present the Hausman test statistics for the models 2-4 against model 1. The Hausman test is performed for the full set of parameters as well as for the set of parameters on the covariates, only.

[*Table* 5 to be inserted here]

Note that stationarity is rejected in all cases but one, where the test statistic was negative (and hence set to zero). The rejection is much stronger when the baselines are included, due to the larger differences in these parameter values. Note that the tighter the upper bound on the elapsed duration, the stronger is our rejection of stationarity. This may be counterintuitive, since we would expect a tighter upper bound (which is closer to the sampling date) reduces the non–stationarity problem. However, the test is really not only for non–stationarity. It is also a test for unobservable variables affecting the probability of being observed in unemployment at the sampling date which are correlated with unobservables determining unemployment duration.

5. Conclusion

We have presented a survey of methods for dealing with left-censored and left-truncated duration data. Spells in progress at the moment at which the sample is drawn need to be adequately treated, since they are selective samples. In order to derive their likelihood contribution, the point of departure is the joint conditional density (3.8). Knowledge of two duration variables is necessary (i.e. the elapsed and remaining durations), as well as the entry rate into the state. However in many longitudinal surveys, the elapsed duration is not observed. We have derived methods which are useful for exploiting the available information as efficiently as possible. Information that can be used for determining the elapsed duration - either directly or by bounding it - is useful. Depending on the extent of the information available, we have demonstrated (in the conditional likelihood case) and developed methods to exploit this information, and we have illustrated the efficiency gain associated herewith. Finally, we have derived a test for the stationarity assumption, which was used in the empirical application, too.

References

[1] Allison, P.D. (1982), "Discrete-time Methods for the Analysis of Event Histories", in: S. Leinhardt, ed., Sociological Methodology, (Jossey-Bass Publishers, San Francisco), p. 61-97.

[2] Amemiya, T. (1999), "A note on Left Censoring", in *Hsiao, C. Pesaran, M. H.,*Lahiri, K:, Lee, L.-F., eds., *Analysis of Panels and Limited Dependent Variable Models.*, Cambridge University Press.

[3] Andersen, P. K., Borgan, Ø., Gill, R. D., Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer.

[4] Baydar, N., White, M. (1988), "A Method for analysing Backward Recurrence Time Data on Residential Mobility", *Sociological Methodology,* ed. by C.C. Clogg, American Sociological Association, Washington D.C., p. 105-135.

[5] Blossfeld, H.-P., Rohwer, G. (1995) *Techniques of Event History Modelling,* Lawrence Erlbaum Associates, Inc.

[6] Chesher, A.D., Lancaster, T.,(1983), "The Estimation of Models of Labour Market Behaviour", *Review of Economic Studies, L*, p. 609-624.

[7] Cox, D.R. (1962), *Renewal Theory,* Methuen, London.

[8] Cox, D.R., Hinkley, D.V. (1974), *Theoretical Statistics,* Chapman and Hall, London

[9] D'Addio, A.C. (1998), "Unemployment Duration of French Youth", *IRES*, *Discussion Paper n. 98-31,* Université Catholique de Louvain.

[10] D'Addio, A. C.(2002), *Mobility of Young People on the French Labour Market: Methodological Considerations and Empirical Analyses,* Collection des Thèse, n. 360, Université Catholique de Louvain, Ed. CIACO.

[11] D'Addio, A. C., Rosholm, M. (2002), "The Mobility of French Youth "in" and "out" of the Labour Market", Manuscript.

[12] De Toldi, M., Gouriéroux, C., Monfort, A. (1992), "On Seasonal Effects in Duration Models", *CREST, Document de Travail* n. 9216.

[13] Flinn, C.J.and Heckman, J.J. (1982a), "Models for the Analysis of Labour Market Dynamics", in *Advances in Econometrics*, vol. 1, eds. R. Basmann, G. Rhodes*,* JAI Press, p. 35-95.

[14] Flinn, C.J. and Heckman J.J. (1982b), "New Methods for analysing Event Histories", in *Sociological Methodology,* ed. by S. Leinhardt, Jossey-Bass Publishers, San Francisco, 99-140.

[15] Goto, F. (1996), "Achieving Semiparametric Efficiency Bounds in Left-Censored Duration Models", *Econometrica,* vol.64(2), p. 439-442.

[16] Gritz, M. (1993), "The Impact of Training on the Frequency and the Duration of Unemployment", *Journal of Econometrics,* n.57, p. 21-51.

[17] Guo, G. (1993), "Event History Analysis for Left-Truncated Data", in *Sociological Methodology* edited by P.V. Marsden, Washington DC, American Sociological Association, p. 219-243.

[18] Hamerle, A. (1991), "On the Treatment of Interrupted Spells and Initial Conditions in Even History Analysis", *Sociological Methods and Research*, vol. 19(3), p. 388-414.

[19] Heckman, J.J. (1981), "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating Discrete Time - Discrete Data Stochastic Process and Some Monte Carlo Evidence", in *Structural Analysis of Discrete Data with Econometric Applications,* eds. C. Manski and D. McFadden, M.I.T. Press, Cambridge, Mass.

[20] Heckman, J.J. and Singer, B. (1984), "Econometric Duration Analysis", *Journal of Econometrics,* n.24, p. 63-132.

[21] Heckman, J.J., Singer, B. (1986), "Econometric Analysis of Longitudinal Data", in *Handbook of Econometrics,* vol. III, eds. Z. Griliches and M.D. Intriligator, North-Holland, Amsterdam, p. 1689-1763.

[22] Heckman, J.J., Honoré, B. (1989), "The Identification of the Competing Risks Model", *Biometrika,* vol. 76(2), p. 325-330.

[23] Horowitz, J.L. Manski, C.F. (1996), "Censoring of outcomes and regressors to survey nonresponse: Identification and Estimation using Weights and Imputations", *Working Paper n. 9602,* University of Iowa.

[24] Jenkins, S.P. (1995), "Easy estimation methods for discrete-time duration models", *Oxford Bulletin of Economics and Statistics*, vol. 57 (1), p. 129-138.

[25] Keiding, N. (1986), "Non-Parametric Estimation Under Truncation", in *Encyclopedia of Statistical Sciences,* vol. 9, eds. S. Kotz and N.L. Johnson, Wiley, New York, p. 357-359.

[26] Kiefer, N.M., (1988a), "Economic Duration Data and Hazard Functions", *The Journal of Economic Literature*, vol. 26, p. 646-679.

[27] Kiefer, N.M., (1988b), "Analysis of Grouped Duration Data", in: N.U. Prabhu, ed., *Statistical Inference from Stochastic Processes*, Contemporary Mathematics, vol.80, Providence, p. 107-137.

[28] Kiefer, N.M. (1990), "Econometric Methods for Grouped Duration Data", in *Panel Data and Labour Market Studies*, eds J. Aartog, G. Ridder, J. Theewes, North-Holland, Amsterdam.

[29] Lancaster, T., (1979), "Econometric methods for the duration of unemployment", *Econometrica*, 47, p. 939-956.

[30] Lancaster, T., (1990), *The Econometric Analysis of Transition Data,* Cambridge University Press, Cambridge.

[31] Lancaster, T. and A. Chesher (1981), "Stock and Flow Sampling", *Economic Letters,* 12, p. 723-725.

[32] Magnac, T. (2000), "Subsidized training and youth employment histories" *The Economic Journal,* vol. 110 (466). 805-837.

[33] Mayer, K.U., Tuma, N.B. (1990), *Event History in Life Course Research,* University of Wiscounsin Press, Madison.

[34] Meyer, B.D. (1990), "Unemployment Insurance and Unemployment Spells", *Econometrica*, vol. 58 (4), p. 757-782.

[35] Meyer, B.D. (1995), "Semiparametric estimation of Hazard Models", mimeo.

[36] Nickell, S. (1979), "Estimating the Probability of leaving unemployment", *Econometrica*, 47, p. 1249-66.

[37] Ridder, G. (1984), "The distribution of Single Spell Duration Data", in *Studies in Labor Market Dynamics* eds. G.R. Neumann and N.C. Westergard-Nielsen. New-York: Springer, p. 45-73.

[38] Rosholm, M. (2001), "An Analysis of Labour Market Exclusion and (Re-) Inclusion" in *Transitions in the Labour Market,* Discussion Paper 332, IZA, Bonn.

[39] Rubin, D. (1987), *Multiple Imputation for NonResponse in Surveys*, New-York: John Wiley and Sons.

[40] Salant, S.W. (1977), "Search Theory and Duration Data: a Theory of Sorts", *Quarterly Journal of Economics,* vol.91(1), p. 39-57.

[41] Yamaguchi, K. (1991), *Event History Analysis,* Sage Publications, London.

6. Tables and figures

6.1. Tables

Table 1. Descriptive statistic, schooling durations

Variables	Men		Women	
	Mean	Std.dev.	Mean	Std.dev.
t_e (years)	1.25	1.84	1.26	1.89
t_r (years)	2.06	0.96	2.11	0.96
Age$\leq$ 25	0.95		0.96	
Father is French	0.61		0.61	
# observations	1460		1487	

Table 2. Estimation results for schooling durations

	Men		Women	
	Model A	Model B	Model A	Model B
γ_1	-3.40 (0.25)	-4.91 (0.23)	-3.21 (0.26)	-5.21 (0.25)
γ_2	-2.30 (0.24)	-3.68 (0.21)	-1.87 (0.25)	-3.76 (0.23)
γ_3	-2.18 (0.24)	-3.51 (0.22)	-1.70 (0.25)	-3.53 (0.23)
γ_4	-2.11 (0.24)	-3.32 (0.22)	-1.58 (0.25)	-3.32 (0.23)
γ_5	-2.03 (0.25)	-3.16 (0.22)	-1.59 (0.26)	-3.22 (0.24)
γ_6	-1.92 (0.24)	-2.93 (0.23)	-1.45 (0.26)	-2.98 (0.24)
γ_7	-1.28 (0.24)	-2.22 (0.23)	-0.97 (0.25)	-2.45 (0.24)
γ_8	-0.75 (0.21)	-1.37 (0.23)	-0.81 (0.24)	-2.06 (0.26)
γ_9	0.54 (0.24)	-0.60 (0.25)	-0.16 (0.22)	-0.49 (0.23)
γ_{10}	0.17 (0.25)	0.21 (0.25)	-0.19 (0.27)	-0.16 (0.27)
Age$\leq$ 25	1.68 (0.23)	2.75 (0.21)	1.19 (0.24)	2.74 (0.23)
French father	-0.17 (0.06)	-0.28 (0.06)	-0.12 (0.05)	-0.19 (0.05)

Note: Standard errors in parentheses.

Table 3. Descriptive statistics for unemployment durations.

	Mean	Std.dev.
Duration	6.82	6.61
t_1	6.16	2.42
t_2	3.82	2.71
Age in 1992	24.25	2.86
Woman	0.51	
Length of education	15.40	2.06
Technical education	0.69	
French father	0.62	
Large family	0.50	
Parents divorced	0.16	
Health problem	0.21	
Father dead	0.14	
Living in Paris	0.10	
# observations	2432	
# left-censored obs.	352	

Table 4:. Estimation results

Coefficients	**Flow model**	**Left-censoring Amemiya (1999)**	**Left- censoring using $\bar{t}_1$**	**Left-censoring using $\bar{t}_2$**
γ_1	-1.8908*	-1.8762*	-1.8914*	-1.9126*
	[0.0616]	[0.061]	[0.0608]	[0.0602]
γ_2	-1.9667*	-1.9603*	-1.9866*	-2.0147*
	[0.0699]	[0.0692]	[0.0685]	[0.0678]
γ_3	-1.9938*	-1.984*	-2.0194*	-2.0127*
	[0.0773]	[0.0769]	[0.076]	[0.0746]
γ_4	-2.0238*	-2.009*	-2.0455*	-2.041*
	[0.0871]	[0.0861]	[0.0847]	[0.0835]
γ_5	-2.1258*	-2.1174*	-2.1591*	-2.1587*
	[0.1001]	[0.0991]	[0.0974]	[0.0952]
γ_6	-2.1283*	-2.1154*	-2.1439*	-2.106*
	[0.1093]	[0.1082]	[0.106]	[0.102]
γ_7	-2.3837*	-2.3749*	-2.3926*	-2.3684*
	[0.1341]	[0.1331]	[0.1306]	[0.1284]
γ_8	-2.1677*	-2.167*	-2.1901*	-2.1996*
	[0.1335]	[0.1319]	[0.1288]	[0.1241]
γ_9	-2.3928*	-2.3891*	-2.3896*	-2.3377*
	[0.1612]	[0.1598]	[0.1562]	[0.1492]
γ_{10}	-2.5673*	-2.5748*	-2.5482*	-2.5428*
	[0.1899]	[0.1882]	[0.1815]	[0.1793]
γ_{11}	-2.842*	-2.8516*	-2.8108*	-2.7228*
	[0.231]	[0.2296]	[0.2249]	[0.2141]
γ_{12}	-2.2659*	-2.2985*	-2.2908*	-2.2153*
	[0.1866]	[0.1837]	[0.178]	[0.1727]
γ_{13}	-1.9074*	-1.9343*	-1.9545*	-1.783*
	[0.1869]	[0.182]	[0.1743]	[0.1633]
γ_{14}	-2.9324*	-3.0191*	-3.0703*	-5.2779*
	[0.3254]	[0.3222]	[0.3186]	[0.2337]
γ_{15}	-2.0846*	-1.8885*	-1.501*	-1.5112*
	[0.2327]	[0.2287]	[0.191]	[0.1885]
γ_{16}	-2.4453*	-2.2641*	-2.0581*	-1.9509*
	[0.3053]	[0.3]	[0.287]	[0.2706]
γ_{17}	-2.0948*	-1.9463*	-1.7835*	-1.7956*
	[0.2787]	[0.2742]	[0.2595]	[0.2531]

Table 4 (continued): Estimation results

Coefficients	**Flow model**	**Left-censoring Amemiya (1999)**	**Left- censoring using $\bar{t}_1$**	**Left-censoring using $\bar{t}_2$**
γ_{18}	-2.4382* [0.3484]	-2.305* [0.3454]	-2.1843* [0.3297]	-2.2* [0.3011]
γ_{19}	-2.706* [0.4232]	-2.5845* [0.4189]	-2.5169* [0.4045]	-2.294* [0.3604]
γ_{20}	-2.585* [0.4232]	-2.5107* [0.4181]	-2.5156* [0.4083]	-2.6536* [0.3809]
γ_{21}	-2.8565* [0.5099]	-2.7902* [0.5061]	-2.7506* [0.49]	-2.8498* [0.413]
γ_{22}	-2.2573* [0.4315]	-2.2029* [0.4226]	-2.187* [0.3902]	-2.3005* [0.4075]
γ_{23}	-2.3982* [0.2238]	-2.0939* [0.1254]	-2.0895* [0.141]	-2.0031* [0.1237]
Age in 1992	-0.0465 [0.1053]	-0.314* [0.084]	-0.318* [0.0827]	-0.3623* [0.0831]
Technical Education	0.0566 [0.0609]	-0.0066 [0.0481]	-0.0075 [0.0474]	0.0025 [0.047]
Father's nationality	0.1622* [0.0596]	0.1375* [0.048]	0.1413* [0.0472]	0.1179* [0.0467]
Large family	-0.0907 [0.0591]	-0.1309* [0.0478]	-0.1346* [0.0469]	-0.1464* [0.0461]
Education length	0.4972* [0.1344]	0.6238* [0.1106]	0.6331* [0.108]	0.6521* [0.1065]
Parents divorced	-0.0026 [0.0757]	-0.0268 [0.0586]	-0.0278 [0.0574]	0.0006 [0.0572]
Health problems during childhood	0.0658 [0.0671]	0.0693 [0.0533]	0.0696 [0.0523]	0.0527 [0.0511]
Sex	-0.2679* [0.0561]	-0.2704* [0.0452]	-0.2731* [0.0443]	-0.2778* [0.0436]
Father dead	-0.0963 [0.0888]	-0.0709 [0.0678]	-0.0689 [0.067]	-0.0488 [0.0673]
Living in Paris	0.1205 [0.0907]	0.046 [0.0718]	0.0455 [0.0706]	0.0445 [0.0704]
Log-likelihood	4650.84	14470.23	14467.09	14069.48
Mean expected duration	9.18	8.69	8.62	8.54
Mean duration	6.06	6.82	6.82	6.82
Number of spells	2080	2432	2432	2432

Table 5. Hausman tests

	Full parameter set	Only covariate parameter set
	DF=35	DF=12
Model 2 against 1	0	28.75
Model 3 against 1	93.93	27.64
Model 4 against 1	222.19	35.51

6.2. Figures

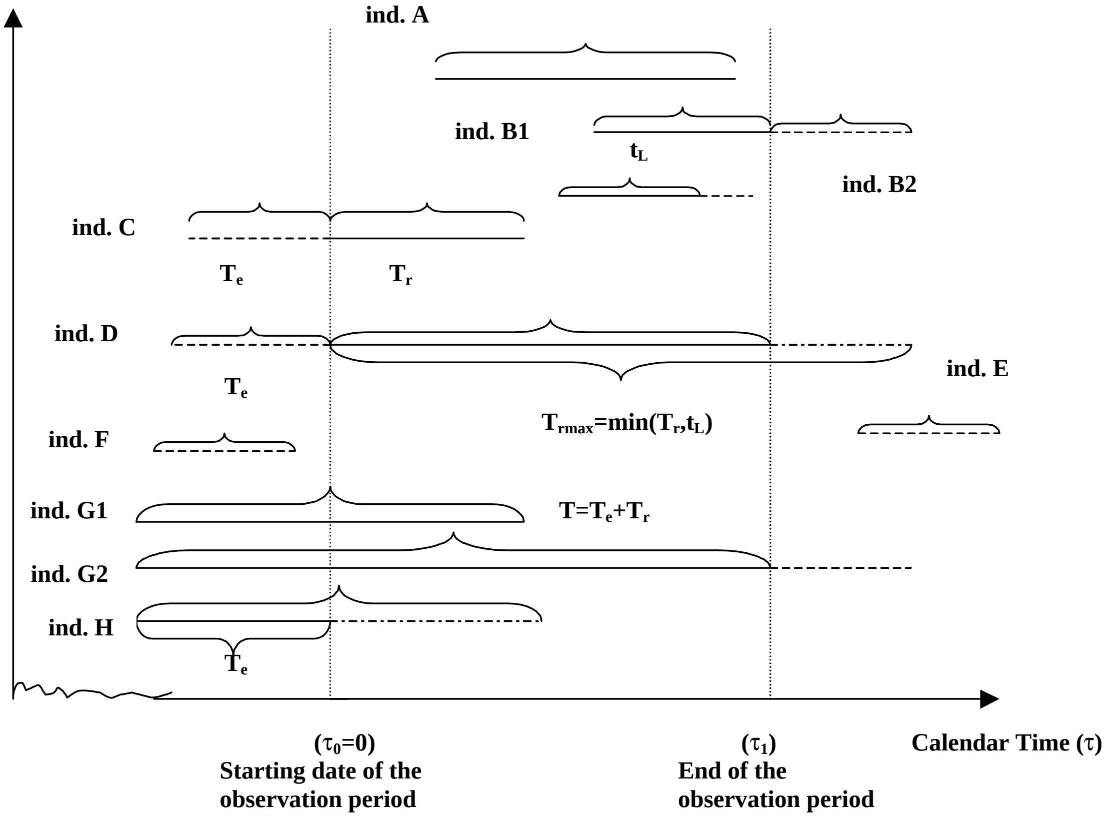

Figure 1: Different censoring mechanisms

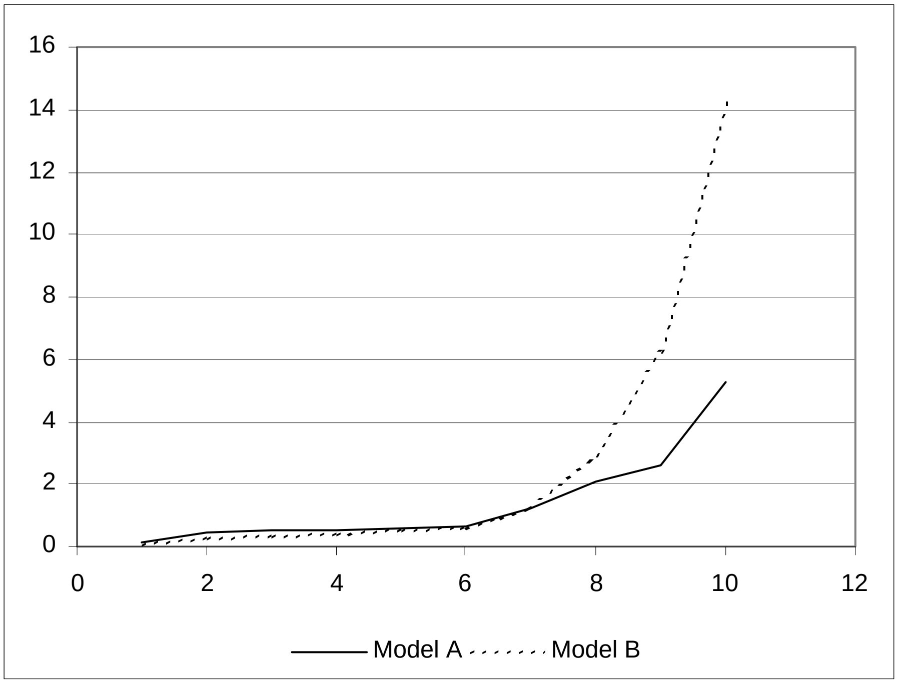

Figure 2: Baseline hazards, men

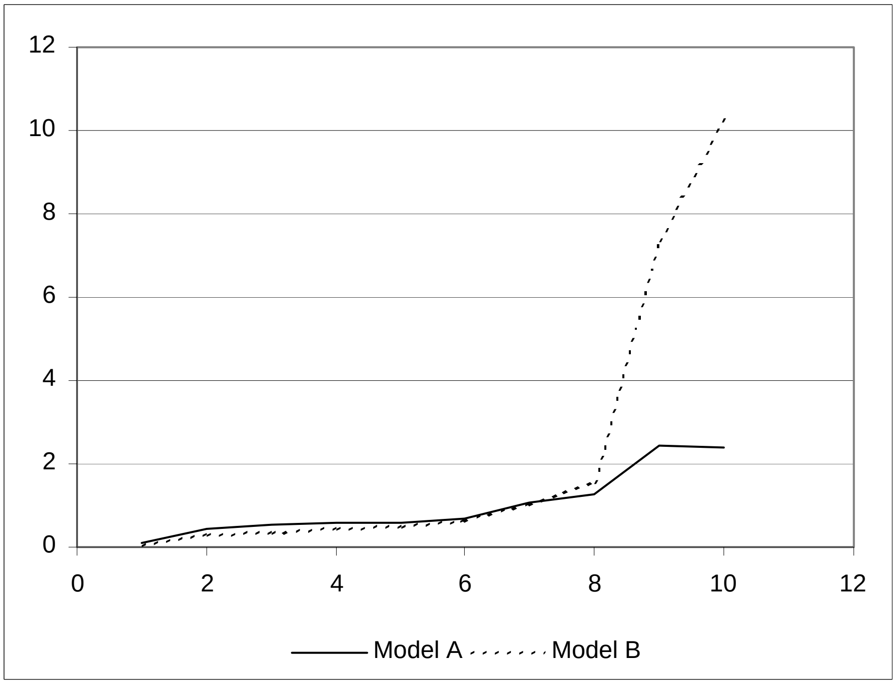

Figure 3: Baseline hazards, women

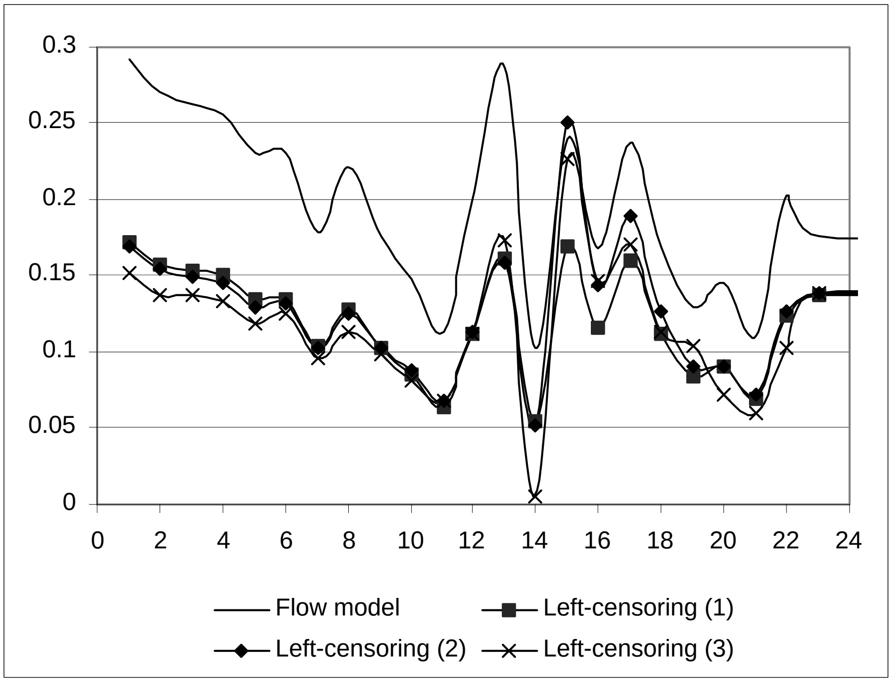

Figure 4: Baseline plots

Working Paper

2000-4 Bent Jesper Christensen and Morten Ø. Nielsen: Semiparametric Analysis of Stationary Fractional Cointegration and the Implied-Realized Volatility Relation in High-Frequency.

2001-5: Bo Sandemann Rasmussen: Efficiency Wages and the Long-Run Incidence of Progressive Taxation.

2001-6: Boriss Siliverstovs: Multicointegration in US consumption data.

2001-7: Jakob Roland Munch and Michael Svarer: Rent Control and Tenancy Duration.

2001-8: Morten Ø. Nielsen: Efficient Likelihood Inference in Non-stationary Univariate Models.

2001-9: Effrosyni Diamantoudi: Stable Cartels Revisited.

2001-16: Bjarne Brendstrup, Svend Hylleberg, Morten Nielsen, Lars Skipper and Lars Stentoft: Seasonality in Economic Models.

2001-17: Martin Paldam: The Economic Freedom of Asian Tigers - an essay on controversy.

2001-18: Celso Brunetti and Peter Lildholt: Range-based covariance estimation with a view to foreign exchange rates.

2002-1: Peter Jensen, Michael Rosholm and Mette Verner: A Comparison of Different Estimators for Panel Data Sample Selection Models.

2002-2: Torben M. Andersen: International Integration and the Welfare State.

2002-3: Bo Sandemann Rasmussen: Credibility, Cost of Reneging and the Choice of Fixed Exchange Rate Regime.

2002-4: Bo William Hansen and Lars Mayland Nielsen: Can Nominal Wage and Price Rigidities Be Equivalent Propagation Mechanisms? The Case of Open Economies.

2002-5: Anna Christina D'Addio and Michael Rosholm: Left-Censoring in Duration Data: Theory and Applications.