Faculteit der Economische Wetenschappen en Econometrie
Vakgroep Econometrie

Serie Research Memoranda

Attrition in Longitudinal panel data, and the empirical analysis of dynamic labour market behaviour

Gerard J. van den Berg
Maarten Lindeboom
Geert Ridder

Research-Memorandum 1993-14
March 1993

vrije Universiteit *amsterdam*

ATTRITION IN LONGITUDINAL PANEL DATA, AND THE EMPIRICAL ANALYSIS OF DYNAMIC LABOUR MARKET BEHAVIOUR

GERARD J. VAN DEN BERG

Dept. of Econometrics, Free University Amsterdam, De Boelelaan 1105,
1081 HV Amsterdam, The Netherlands,
Tel. (+31)-20-5487063, Fax (+31)-20-6461449, E-mail vandenberg@sara.nl

MAARTEN LINDEBOOM

Dept. of Economics, Faculty of Law, Leiden University.

GEERT RIDDER

Dept. of Econometrics, Free University Amsterdam.

SUMMARY

In the empirical analysis of unemployment durations and job durations, it is generally assumed that the stochastic processes underlying labour market behaviour and the behaviour concerning participation in a panel survey are independent. However, there are reasons to believe that the probability of dropping out of the panel is related to the rate at which a (different) job is found. If there is such a relation, and if it is ignored, then the estimator of the rate at which individuals become employed or change jobs will generally be inconsistent. In this paper we analyze the relation between the duration spent in a particular labour market state and the duration of panel survey participation, by explicitly modelling and estimating the joint distribution of both durations. The emphasis will be on models allowing for stochastically related unobserved determinants of both types of duration. We estimate models both for unemployment durations and for job durations.

Date: February 24, 1993.
JEL classification: C41, C33, J64.
Keywords: attrition, panel data, duration models, unemployment.

1. Introduction

In the empirical analysis of individual unemployment durations and job durations, it is generally assumed that the stochastic processes underlying labour market behaviour and the behaviour concerning participation in a panel survey are independent. If this assumption is correct, then attrition from the panel before the duration is completed can be considered as independent right-censoring of the duration variable. Nevertheless, it seems plausible that panel survey participants who have a relatively high probability of finding a (different) job, also have a higher probability of dropping out of the panel (e.g., individuals may move to another town to work in a new job, and the agency running the survey may have trouble following them). If that is true then the commonly used procedure to estimate models for the duration spent in a particular state of the labour market (say unemployment) underestimates the rate at which individuals become employed or change jobs.

In Lillard (1989), a Weibull proportional hazard model for the duration of participation in the PSID panel survey is estimated. It appears that the exit rate out of the panel is significantly larger for individuals who expect to move in the near future. Although the model in Lillard (1989) is simple and not comparable to the models proposed in the present paper, this result supports the suspicion that the commonly used procedure to estimate duration models may produce biased results.

In this paper we will examine whether there is a relation between the duration spent in a labour market state and the duration of panel survey participation. In particular, we will estimate models for the joint distribution of these two duration variables. This means that we have to explicitly model the distribution of survey participation duration and its relation to the distribution of the duration spent in a particular state of the labour market. In accordance to the literature on duration analysis, we take a hazard rate approach when specifying the model. The hazard rate of the distribution of survey participation duration (or, equivalently, the exit rate out of the panel) can be interpreted as the rate at which contact between participating individuals and interviewers is lost. The duration of survey participation is treated as an absolutely continuous random variable. Of course, its realizations can only be observed to lie between two consecutive waves of the panel.

There are several ways to model dependence of the duration spent in a particular labour market state and the duration of survey participation. Here, the emphasis will be on models allowing for such dependence by way of stochastically related unobserved determinants of both types of duration. An advantage of such models is that they do not a priori restrict the sign of the dependence if a sufficiently flexible class of distributions is chosen for the unobserved determinants. Thus, such models can mimic other types of dependence between both durations.

The outline of the paper is as follows. In Section 2 we present the model. We derive the likelihood function taking into account that we have samples from the stocks of individuals in the labour market states of interest. We also examine in some detail the parameterization of the distribution of the unobserved heterogeneity terms. In Section 3 the results are presented. We check in a number of ways whether the results are sensitive with respect to the model specification.

Modelling the relationship between labour market duration variables and attrition by way of their unobserved determinants is in line with the popular modelling setup for sample selection introduced by Heckman (1979). In our application there could be a more direct relationship. It is conceivable that a positive fraction of the individuals who decide to start working in a new job, immediately leave the panel. In Section 4 we investigate this by constructing and estimating an alternative model. We find the phenomenon to be empirically unimportant. Section 5 concludes.

2. The joint distribution of spell length and observation period

2.1. *The model*

We are interested in estimating the distribution of the sojourn time t in a particular labour market state. If we follow a cohort of heterogeneous individuals, we can estimate the hazard rate of leaving the state, and relate this rate to observed and unobserved characteristics of the individuals. The observed characteristics are given by a vector of regressors x, and the unobserved characteristics are summarized by a scalar random variable v, with x and v independent. We assume that the hazard rate is of the Mixed Proportional Hazard (MPH) type,

$$(2.1) \qquad \theta(t|v,x) = \alpha t^{\alpha-1}.v.\exp(x_1'\beta_1)$$

Note that in (2.1) all explanatory variables are time-invariant and that the baseline hazard has a Weibull specification. The latter assumptions, which are adopted in the empirical analysis below, as well as all other main assumptions, will be listed together at the end of this subsection.

Because v is unobserved, empirical inference is based on the distribution with c.d.f. G(t|x) The hazard rate associated with G(t|x) is

$$(2.2) \qquad \theta(t|x) = \alpha t^{\alpha-1}.E(v|\geq t,x).\exp(x_1'\beta_1)$$

with E(v|≥t,x) denoting the mean of v among the survivors at t, i.e. conditional on the sojourn time exceeding t.

In practice we do not follow all individuals until after they have left the state of interest. As a consequence, some durations are censored. The length of the observation period may be determined in advance, even before the observation starts (Type I censoring), or it may be due to panel attrition. In any case, in most empirical applications it is assumed that this censoring can be treated as independent right-censoring. For this assumption to be valid, the observation process should not be selective, i.e. the observed hazard rate at t should not differ from the hazard rate specified in (2.2). In other words, the information that an individual is under observation just before t should not change our prediction of him leaving the state at t. Type I censoring satisfies this requirement. (See Aalen (1978), Williams & Lagakos (1977), and Lagakos (1979) for some general theory.)

Let a be the length of the period that a randomly chosen individual participates in the panel. This length is itself a duration. We can model it in the same way as t. The random variable u summarizes the unobserved heterogeneity in the distribution of a. The hazard of a conditional on x and u is denoted by $\zeta(a|u,x)$ and is assumed to be of the MPH type,

$$(2.3) \qquad \zeta(a|u,x) = \delta a^{\delta-1}.u.\exp(x_2'\beta_2)$$

Problems arise if a is related to the unobserved v. In that case, knowledge that $a \geq t$ is informative on v. Hence, it will alter E(v|≥t,x), and thus the hazard rate of t given x. For example, if the distribution of v given $a \geq t$ stochastically dominates the population distribution of v, and if this is ignored, then we obtain an overestimate of the conditional hazard $\theta(t|v,x)$.

We assume that t|x and a|x are related by way of their unobserved determinants being related. In other words, t|x and a|x are independent if and only if u and v are independent. If u and v are independent then we have an ordinary single-spell duration model for t in which a is the (independent right) censoring point. However, if u and v are dependent, then a is related to v, and inference on the distribution of t has to be based on the joint distribution of t,a|x.

Basically, we observe t iff $t \leq a$, and a iff $a < t$. Consequently, our model is a competing risks model in which we observe $\min(t,a)$, $I(t \leq a)$, and x, where I(E) is the indicator function of the event E. Recall that we have adopted MPH specifications for the hazards of $t|v,x$ and $a|u,x$. In such models, the joint distribution of u,v can be identified from observations on $\min(t,a)$, $I(t \leq a)$ and x, under general conditions (see Heckman and Honoré (1989)). The non-parametric identifiability of the joint distribution of u,v (and the baseline hazards) implies that we need not rely on arbitrary parametric or distributional assumptions. However, since the data sets we use are quite small, and since the nonparametric estimation theory of such models has not been established well yet, the application will be parametric. The parameters are $\alpha,\beta_1,\beta_2,\delta$ (see (2.1) and (2.3)) as well as the parameters of the joint distribution of u,v.

Below we summarize the assumptions we make. The last two assumptions will be discussed in the next subsections. Those two assumptions, as well as the assumption that we have Weibull baseline hazards, are not necessary for the type of empirical analysis carried out below (see Van den Berg, Lindeboom & Ridder (1991) for a more general setup). Rather, they are adopted because of computational and data limitations in our particular application, as will become clear in the sequel.

Assumption 1. The hazards $\theta(t|v,x)$ and $\zeta(a|u,x)$ are of the MPH type. The baseline hazards have Weibull specifications. There are no time-varying explanatory variables. (See (2.1) and (2.3).)

Assumption 2. $t|x$ and $a|x$ are dependent iff u and v are dependent.

Assumption 3. The inflow rate into the labour market state of interest is constant before the first interview and factorizes in v and x.

Assumption 4. The distribution of u,v is bivariate discrete with fixed numbers of points of support which have unknown locations and probability masses.

2.2. *Distribution of the endogenous variables in a stock sample*

In this subsection, the actual likelihood function is derived for the model, given the particular observation scheme (or sample setup) we use in the empirical analysis. To be able to do so, we first have to consider the joint distribution of the endogenous variables in such samples.

Our data on the durations spent in a particular state S of the labour market are based on a sample from the stock of individuals in that state, namely those who are participating in the first wave of the panel. As is well known, such samples are selective samples from the relevant population (see e.g. Ridder (1984)). There are two reasons for this. First, the probability that a particular spell is included is proportional to its length (length-biased sampling). Secondly, this probability depends on the rate at which spells start during the period preceding the sampling date (inflow rate dependence). We assume that the inflow rate at a given date, given v and x, does not depend on a or u. This makes sense since the inflow rate is a result of labour market behaviour. As may be clear intuitively, the selectivity induced by drawing from the stock does not affect the distributions of $a|u,x$ and u in the sample. Because of this, the standard results on stock samples in Ridder (1984) and Chesher & Lancaster (1983) can be straightforwardly extended to deal with the selectivity in the present context.

Let the duration variable r denote the time spent in state S between the moment at which the stock sample is drawn (which is the moment of the first interview) and the moment at which exit out of state S occurs. Subsequent waves (interviews) of the panel survey provide information on r. The first wave also gives information on the elapsed duration p in state S at the moment at which the stock sample is drawn. Let a subscript s of a density or probability denote the conditioning on presence in the stock of individuals in S. Let h be a generic symbol for a density. The argument of the density will make clear which variable is considered. However, we denote the density defined by the hazard $\theta(t|v,x)$ by $f(t|v,x)$ and the corresponding c.d.f. by $F(t|v,x)$. The likelihood function is based on the density $h_s(p,r,a|x)$ of the endogenous variables conditional on x and presence in the stock.

Using the literature mentioned above, it can be shown that $h_s(p,r,a|x)$ can be expressed as follows,

$$(2.4)\qquad h_s(p,r,a|x) = \sum_j \sum_i \frac{f(p+r|v_j,x)}{\int_p \overline{F}(p|v_j,x)\,dp} .h_s(a|u_i,x)\ .Pr_s(u=u_i,v=v_j)$$

$$\text{with } \overline{F}=1-F$$

in which (i) $h_s(a|u,x)$ is the density defined by the hazard $\zeta(a|u,x)$, (ii) the distribution of u,v in the stock is such that u and v in the stock are independent if and only if they are independent in the population, and (iii) u,v in the stock are independent of x (see Van den Berg, Lindeboom & Ridder (1991) for details and proofs). Note that the inflow rate does not enter the expression above. In equation (2.4), $h_s(p,r,a|x)$ follows from integrating $h_s(p,r|v,x).h_s(a|u,x)$ w.r.t. the distribution of u,v in the stock. So, in practice one can choose a parameterization for the distribution of u and v in the stock, and estimate the parameters of this distribution along with the parameters of $\theta(t|v,x)$ and $\zeta(a|u,x)$. In fact, this is the standard procedure for estimating reduced form duration models using stock samples in the presence of unobserved heterogeneity.

It should be noted that if the model is parameterized in a different way, or if we condition the likelihood on the realizations of p, then we do not need Assumptions 3 and 4 to get an estimable model. (details are in Van den Berg, Lindeboom & Ridder (1991)). This implies that by estimating the model in a number of different ways, it may be checked whether these conditions are satisfied. Note however that even with a Weibull baseline hazard in $\theta(t|v,x)$ the estimation is computationally demanding, since then the likelihood contribution contains a nonanalytical integral if both p and r are censored from above. Also note that the data on p provide information that is additional to the minimum amount of information needed for identification.

Construction of the likelihood function

Let the panel survey consist of j waves. Individuals who participate in the first wave are invited to participate in all subsequent waves. However, there is no return to the panel. The variable r may be censored either because the individual is still in state S at the date of the j^{th} interview or because the individual drops out of the panel before that date.

Consequently, we can distinguish between three different cases. Let τ_i $(i=1,..j)$ denote the length of the time period between the i^{th} and the first interview (so $\tau_1=0$). Case I is defined as the case in which r is observed exactly. Suppose $\tau_i \leq r < \tau_{i+1}$ $(i \in \{1,..j-1\})$. Then, if p is uncensored, the likelihood contribution $\mathcal{L}$ of the individual equals

$$(2.5) \qquad \mathcal{L} = \int_{\tau_{i+1}}^{\infty} h_s(p,r,a|x) \; da$$

If p is censored (i.e. if it is only known that p exceeds a certain value) then the r.h.s. of (2.5) has to be integrated over p accordingly.

Case II is the case in which we observe that the individual drops out of the panel before the date of the j^{th} interview and the individual was still in state S at the last wave at which he participated. Suppose the individual drops out of the panel between the i^{th} and the $(i+1)^{th}$ interview $(i \in \{1,..j-1\})$. Then, if p is uncensored,

$$(2.6) \qquad \mathcal{L} = \int_{\tau_i}^{\tau_{i+1}} \int_{\tau_i}^{\infty} h_s(p,r,a|x) \; dr \; da$$

Case III is the case in which we observe that the individual participates in all waves, and the individual is still in state S at the j^{th} interview (so $r>\tau_j$ and $a>\tau_j$). As a result, if p is uncensored,

$$(2.7) \qquad \mathcal{L} = \int_{\tau_j}^{\infty} \int_{\tau_j}^{\infty} h_s(p,r,a|x) \; dr \; da$$

2.3. *The parameterization of the distribution of unobserved heterogeneity*

In this subsection we discuss the parameterization of the joint distribution of u,v (see Assumption 4).

In the literature on unemployment durations, unobserved heterogeneity is often modelled by way of a discrete random variable (see e.g. Nickell (1979) and Ham & Rea (1987)). Usually, if more than two or three points of support are taken then the estimates of some of them coincide. Heckman & Singer (1984) show that in a class of mixed proportional hazard duration models the non-parametric maximum likelihood estimator of the heterogeneity distribution is a discrete distribution. However, the estimation procedure requires the number of points of support not to be fixed in advance, and estimation of standard errors is not straightforward. Moreover, the procedure is developed for situations in which censoring is independent. Nevertheless, this result illustrates the flexibility of discrete distributions in terms of the range of observed mixture duration distributions they can generate.

There is a large applied literature in which two duration variables are allowed to depend on each other by way of their unobserved explanatory variables u and v (see e.g. Devine & Kiefer (1990) for examples in which the duration variables denote durations spent in different labour market states). In most papers u and v are specified as $u=\exp(c_0.\omega)$ and $v=\exp(c_1.\omega)$, in which ω is a univariate random variable and c_0 and c_1 are parameters to be estimated. This restricts the way u and v are related. Lindeboom & Van den Berg (1994) show that in such models there may be insufficient flexibility to correctly estimate the dispersion as well as the relation of the duration variables. A genuine bivariate specification of the distribution of u,v seems preferable.

Butler, Anderson & Burkhauser (1989) estimate a model for retired individuals in which unobserved explanatory variables for the duration until return to work and the duration of life are allowed to be correlated. They assume that the heterogeneity terms follow a discrete bivariate distribution, like in the present paper. However, in the estimation procedure, the points of support for u and v are fixed in advance, whereas we will estimate these points along with the other parameters.

Van den Berg & Steerneman (1991) examine the range of values that the correlation of the duration variables can attain in bivariate MPH models, in general as well as for particular parametric families of the distribution of u,v. It turns out that when u and v have a bivariate discrete distribution with two or more points of support for each, and the locations of these points are not fixed in advance, then all possible values can be attained. On the other hand, when log u and log v have a bivariate normal distribution, or when they have a bivariate discrete distribution in which the locations of the points of support are fixed in advance, then the range of values that can be attained is smaller.

These results are taken to justify Assumption 4. In most of the empirical analysis below, we assume that u and v both have two points of support (u_1, u_2, v_1 and v_2). We will show that in that case it is relatively easy to interpret the estimation results and to test for independence. We take $u_1 \geq u_2 > 0$ and $v_1 \geq v_2 > 0$. The probabilities $Pr_s(u=u_i, v=v_j)$ are denoted as follows:

$p_1 = Pr_s(u=u_1, v=v_1)$ $\quad\quad$ $p_3 = Pr_s(u=u_1, v=v_2)$

$p_2 = Pr_s(u=u_2, v=v_1)$ $\quad\quad$ $p_4 = Pr_s(u=u_2, v=v_2) = 1-p_1-p_2-p_3$

The covariance of u and v in the stock can be written as

$$\text{(2.8)} \qquad \text{COV}(u,v) = (p_1p_4 - p_2p_3) \cdot (u_1 - u_2) \cdot (v_1 - v_2)$$

The support of the distribution of u,v in the stock equals the support in the population, but the probabilities associated with the points of support in the stock do not equal those in the population. However, it can be shown that the sign of the covariance of u and v in the stock equals the sign of it in the population. Further, u and v are independent if and only if COV(u,v)=0 (for details, see Van den Berg, Lindeboom & Ridder (1991)).

Since $r|x$ and $a|x$ are independent if and only if u and v are independent, it follows that $r|x$ and $a|x$ are independent if and only if $p_1p_4 = p_2p_3$ (conditional on $u_1 \neq u_2$ and $v_1 \neq v_2$). This makes it easy to test for independence between the duration in state S and the duration of participation in the panel survey. Moreover, there holds that $\text{COV}(r,a|x)$, COV(u,v), and the covariance $\text{COV}(t,a|x)$ of $t|x$ and $a|x$ in the population, always have the same sign (which, if $u_1 \neq u_2$ and $v_1 \neq v_2$, is the sign of $p_1p_4 - p_2p_3$).

In Section 3 we report estimates for three different model specifications. Model 3 is the general model. Model 1 is the model without unobserved heterogeneity, i.e. the model in which it is imposed that $u_1 = u_2$ and $v_1 = v_2$. In Model 2 we allow for unobserved heterogeneity in θ and ζ, but we impose that u and v are independent, i.e., we impose that $p_1p_4 - p_2p_3 = 0$ if $u_1 \neq u_2$ and $v_1 \neq v_2$. By comparing the results for Models 1 and 2 it can be tested whether there is unobserved heterogeneity in the exit rates θ and ζ. Note however that such a comparison is conditional on independence of u and v. By comparing the results for Models 2 and 3 it can be tested whether the unobserved heterogeneity terms u and v are dependent.

Because in Models 1 and 2 the likelihood factorizes in a part associated with unemployment durations and a part associated with the durations of survey participation, the test statistics for unobserved heterogeneity in θ and ζ are independent. The LR tests for H_0:v_1=v_2 and for H_0:u_1=u_2 are non-standard, because under the null hypothesis fewer parameters are identified than under the alternative. In the literature it is usually assumed that a test in which critical values of the χ_2^2 distribution are used is on the safe side.

Conditional on $u_1 \neq u_2$ and $v_1 \neq v_2$, testing for independence of u and v (or, equivalently, for COV(u,v)=0) in Model 3 means testing for $p_1p_4-p_2p_3=0$. Consequently, conditional on $u_1 \neq u_2$ and $v_1 \neq v_2$ and on $0<p_i<1$ for all i, the LR test for independence asymptotically has a χ_1^2 distribution under the null hypothesis.

We do not include constant terms in x_1 and x_2 (see equations (2.1) and (2.3)), since these would be undistinguishable from multiplicative constants in v and u, respectively. However, in Model 1 every individual has the same realization of v and u, and these will be represented by constant terms in x_1 and x_2, respectively (though they could as well be represented by v_1 and u_1 in a model in which it is imposed that p_1=1).

3. The data and the results

3.1. *Unemployment duration and attrition*

For the empirical analysis of the relation between the duration of unemployment or a job and the duration of survey participation, we use two datasets. In the present subsection we use data on unemployment durations from a panel survey conducted by the Netherlands Central Bureau of Statistics (CBS). In the next subsection we use data on job durations from a panel survey held by the Organization of Strategic Labor Market Research (OSA).

3.1.1. *The data*

As of April 1984, the Netherlands Central Bureau of Statistics conducts the Netherlands Socio-Economic Panel survey. Interviews are held twice a year. At every interview, respondents were asked to recall their labour market history for the past 6 months. At the first interview the observation period is extended to 12 months. We use the first four waves of the panel.

For our purposes we selected 223 men aged between 17 and 65, who reported that at the moment of the first interview (April 1984) their main activity was being unemployed and searching for work. Of the data on p, 64% is censored in the sense that it is only known that p exceeds a year. Further, Case I (see Subsection 2.2 for the definition) holds for 50% of the individuals, while Cases II and III hold for 28% and 22% of the individuals, respectively. Between the first and the second wave, 40 of the 223 individuals (18%) drop out of the panel. From the individuals in our sample who are unemployed at the date of the second (third) interview and who participate in that interview, 11% (14%) drop out of the panel before the third (fourth) interview.

In the present context, Assumption 3 in Section 2 states that the inflow rate into unemployment is constant before April 1984. One may question whether this assumption holds true. In the U.K. the inflow rate was fairly constant between 1967 and 1983 apart from an increase in 1979-1981 (see Pissarides (1986)). Reliable Dutch data are absent. However, in Van den Berg (1990) the sensitivity of unemployment duration models to changes in the time path of the inflow rate is checked using the same data set as in the present paper. It appears that the main results are insensitive to a priori reasonable changes. Finally, recall from Subsection 2.2 that we can test for the inflow rate assumption.

Concerning the parameterization of the model, we take $\delta=1$ in equation (2.3). Clearly, this is restrictive. The reason for assuming a constant exit rate here is that we do not have much information on the distribution of $a|u,x$: the sample is quite small and the data on a are grouped (we at most observe that a lies between two interviews).

Because we are dealing with a relatively small sample, we restrict ourselves to a small number of explanatory variables. It should be noted that the inclusion of additional variables in x_1 or x_2 (see equations (2.1) and (2.3)) did not alter the conclusions with respect to the sign and magnitude of the relationship between unemployment duration and the duration of panel survey participation.

The estimation method we have employed was ML using the BHHH algorithm. If an individual in the sample is in Case II or Case III and if p is censored for that individual, then the likelihood contribution contains an incomplete Gamma function. This function is calculated numerically using Gauss-Laguerre quadrature.

3.1.2. *Results*

Table 3.1 contains the parameter estimates. The unit time period is one week. For all three models, the estimates of β_1 and β_2 seem to be in accordance with intuition. Since here our primary interest is in the estimates of $h_s(u,v)$, we will not give a lengthy account of the results in parts (i) and (ii) of Table 3.1. Note that the magnitude of the parameter in β_2 associated with the constant term in x_2 (or, in Models 2 and 3, the mean of u), among other things reflects the efforts by the agency running the survey to follow respondents. Thus, this value probably depends very much on the survey setup.

Clearly, the negative duration dependence of θ in Model 1 has its counterpart in the unobserved heterogeneity in θ in Model 2. Note that in Model 2, the p_i are just the products of the probabilities of the associated realizations of u and v (e.g., $p_1 = P_s(u=u_1).P_s(v=v_1)$). There are two cross parameter restrictions on p_1, p_2, p_3 and p_4 in Model 2, a linear one ($p_1+p_2+p_3+p_4=1$) and a nonlinear one ($p_1p_4-p_2p_3=0$). By using the likelihood values reported in part (v) of Table 3.1, we can test whether Model 2 is a significant improvement over Model 1. It follows that $v_1=v_2$ is strongly rejected and that $u_1=u_2$ is weakly rejected.

Now let us compare the results for Model 2 to those for Model 3. The estimates of what normally are the parameters of interest (β_1 and α) are virtually identical in both models. Moreover, this also holds for the estimates of β_2, u_1, u_2, v_1 and v_2. The estimates of p_1, p_2, p_3 and p_4 in Model 3 differ only slightly from those in Model 2. As a result, there is a small negative correlation between u and v. From Table 3.1, CORR(u,v)=0 cannot be rejected. It follows that it cannot be rejected that r and *a* are independent (conditional on x).

If we integrate p out of the likelihood, or use the likelihood conditional on p to estimate the model, then this does not affect the results in any substantial way. This may be regarded as evidence that the constant inflow rate assumption is not violated, or, at least, that it is not a strong assumption.

When data on p and x_1 are available, then in principle these are sufficient to consistently estimate the parameters of the unemployment duration distribution. It might therefore be interesting to proceed along this way and compare the results to those in parts (i), (iii) and (iv) of Table 3.1. Unfortunately, it appeared that the information in the data on p is insufficient to disentangle the estimate of α from the estimates of v_1 and v_2 (when $v_1 \neq v_2$) or the constant in β_1 (when $v_1=v_2$).

A striking feature of the estimated $h_s(u,v)$ is that individuals who have

Table 3.1. Estimates for the unemployment duration model.

variable/parameter	model 1		model 2		model 3	
(i) *systematic part of the exit rate out of unemployment* $\exp(x_1'\beta_1)$						
level of education	0.25	(0.07)	0.58	(0.13)	0.58	(0.13)
log(age)	−0.13	(0.22)	−0.27	(0.46)	−0.29	(0.46)
log(#working+1)	0.51	(0.19)	1.21	(0.39)	1.21	(0.39)
Dutch nationality	0.31	(0.28)	0.37	(0.51)	0.37	(0.51)
constant	−3.38	(1.12)				
(ii) *systematic part of the exit rate out of the panel* $\exp(x_2'\beta_2)$						
level of education	−0.19	(0.14)	−0.35	(0.21)	−0.34	(0.21)
Dutch nationality	−0.72	(0.34)	−0.94	(0.41)	−0.94	(0.42)
constant	−4.05	(0.38)				
(iii) *duration dependence of the exit rate out of unemployment*						
α	0.74	(0.11)	1.67	(0.35)	1.66	(0.36)
(iv) *distribution* $h_s(u,v)$						
v_1.100			0.16	(0.38)	0.18	(0.43)
v_2.100			0.0091	(0.026)	0.010	(0.029)
u_1.100			197	(6203)	188	(6201)
u_2.100			2.05	(0.99)	2.02	(0.97)
p_1			0.02	(0.01)	0.00	(0.02)
p_2			0.20	(0.05)	0.21	(0.05)
p_3			0.06	(0.03)	0.07	(0.04)
p_4			0.72	(0.06)	0.72	(0.06)
(v) *other statistics*						
CORR(u,v)					−0.12	(0.19)
Log likelihood	−1164.7		−1159.6		−1159.4	
due to p,r	−994.7		−991.4			
due to *a*	−170.0		−168.2			

standard errors in parentheses.

u_1 as realization of u (7.6% of the stock of unemployed at the moment of the first interview), have a very high estimated exit rate out of the panel. For example, if such an individual has lower secondary education (education=2) and the Dutch nationality, then the estimated exit rate out of the panel $u.\zeta_0(x)$ equals 0.37, which implies that the expected duration of survey participation is as small as 2 weeks and 5 days. Moreover, the probability of exit out of the panel before the second interview (that is, before about 26 weeks) is as high as 0.9999. As a result, the model predicts that at the date of the second interview the remaining participants are virtually homogeneous with respect to u. If in reality the latter does not hold then our conclusions may be invalid. To investigate this, we estimated an extended model. In particular, we allowed u,v to have six points of support based on three possible realizations of u and two of v. It turned out that the results confirm the conclusions above (see Van den Berg, Lindeboom & Ridder (1991) for details).

3.2. *Job duration and attrition*

In this subsection we briefly report the results of applying the methods of this paper to panel survey data on job durations. Because these data have been discussed and analyzed extensively in Lindeboom & Theeuwes (1991) and Van den Berg (1992), the present exposition can be very brief. (Lindeboom & Theeuwes (1991) assume independent right-censoring; Van den Berg (1992) only uses the first spell of the survey.)

3.2.1. *The data*

The data are from three waves of a Dutch panel survey of individuals, conducted in April 1985, October 1986 and October 1988. From the first wave we selected 1726 respondents who were working in full time jobs at the date of interview. From these respondents, 1210 participated in the second wave, and only 835 participated in the third wave. Case I (see Subsection 2.2 for the definition) holds for 24% of the individuals, Case II for 50%, and Case III for 26%.

We assume absence of duration dependence in the job duration hazard $\theta(t|v,x)$ (so $\alpha=1$ in equation (2.1)). As noted in Section 2, this assumption is convenient because it precludes numerical integration in the likelihood. Previous studies such as Van den Berg (1992) and Lindeboom and Theeuwes (1991) found that this is a valid assumption for The Netherlands. There is no prior information on duration dependence of the exit rate out of the panel. Since the sample size is rather large, we try to allow for duration dependence of the exit rate out of the panel (see equation (2.3)).

3.2.2. *Results*

The results for the three estimated model versions are in Table 3.2. By comparing Models 1 and 2 we see that allowing for unobserved heterogeneity does not have large or significant effects. The estimates of v_1 and v_2 converged to the same value and hence the associated probabilities are not identified. As a result, the estimates for the parameters in θ in Model 2 are equal to those in Model 1. For the survey participation duration slight changes in the estimates are found in comparison to those for Model 1.

Now let us compare the results of Model 2 and 3. Model 3 shows large changes in comparison to the results of Model 2. The likelihood value changes considerable, and there are major shifts in the parameter estimates for ζ. The variables on age, gender, the education and other characteristics (breadwinner, nationality and region) have more pronounced effects. The estimate of the duration dependence parameter δ rises to 1.04, and is insignificantly different from one. There are also some large changes in the parameter estimates for $h_s(u,v)$.

The parameter estimates for the exit rate out of a job θ hardly change (although the precision improves). This may be due to the fact that the data contain on average a lot of uncensored and relatively long elapsed durations p. It may be that the information on θ that is contained in the data on p dominates other sources of information on θ.

The estimates on the heterogeneity distribution $h_s(u,v)$ suggest that for 10% of the sample relatively long job durations are combined with relatively short survey participation durations, whereas the reverse holds for the other 90% of the sample. Indeed, the estimates of CORR(u,v) and CORR(r,a|x) are negative, although insignificantly different from zero according to the Wald test. The LR test, on the other hand, rejects the hypothesis that CORR(u,v)=0, so there is some ambiguity. However, since the estimates of what normally are the parameters of interest (β_1) are virtually identical in model 1 and model 3, it does not matter in practice whether account is taken of the dependence of $r|x$ and $a|x$ or not.

Another test confirms our findings. We estimated β_1 using only data on p, in a model without unobserved heterogeneity but with a dummy in x_1 for respondents who leave the survey before the second wave. The results of this simple test again indicate a negative association (a significant coefficient was found). Leaving out the dummy variable, however, hardly changes the estimates of the remaining parameters in β_1.

Table 3.2. Estimates for the job duration model.

variable/parameter	model 1	model 2	model 3
(i) systematic part of the exit rate out of a job $exp(x_1'\beta_1)$			
15< age <26	1.83 (0.10)	-	1.87 (0.09)
25< age <36	1.17 (0.08)	-	1.19 (0.07)
35< age <51	0.43 (0.08)	-	0.44 (0.08)
Gender (male=1)	-0.04 (0.08)	-	-0.07 (0.05)
Prim. education	-0.28 (0.11)	-	-0.28 (0.08)
Ext.prim. (gen.)	-0.23 (0.13)	-	-0.23 (0.11)
Secondary (gen.)	0.11 (0.18)	-	0.11 (0.17)
Ext. prim (voc.)	-0.24 (0.10)	-	-0.25 (0.07)
Secondary (voc.)	-0.22 (0.09)	-	-0.23 (0.05)
Higher (non ac.)	0.02 (0.10)	-	0.01 (0.07)
Academic	0.31 (0.15)	-	0.31 (0.13)
Breadwinner	-0.16 (0.08)	-	-0.16 (0.05)
Dutch nat.	-0.00 (0.12)	-	0.01 (0.10)
Western region	0.04 (0.06)	-	0.04 (0.06)
Satisfied with job	-0.24 (0.08)	-	-0.26 (0.05)
Wage insufficient	-0.08 (0.07)	-	-0.07 (0.07)
Bad working cond.	-0.07 (0.05)	-	-0.07 (0.05)
Low skill job	0.26 (0.07)	-	0.25 (0.06)
Managerial	-0.06 (0.05)	-	-0.05 (0.05)
Constant	-2.73 (0.20)		
(ii) systematic part of the exit rate out of the panel $exp(x_2'\beta_2)$			
15< age <26	0.30 (0.14)	0.32 (0.16)	0.40 (0.16)
25< age <36	0.00 (0.11)	0.00 (0.12)	0.04 (0.14)
35< age <51	-0.10 (0.11)	-0.11 (0.12)	-0.08 (0.14)
gender (male=1)	0.14 (0.09)	0.14 (0.11)	0.20 (0.12)
Prim. education	0.17 (0.16)	0.19 (0.17)	0.21 (0.18)
Ext.prim. (gen.)	-0.09 (0.21)	-0.10 (0.22)	-0.06 (0.24)
Secondary (gen.)	0.42 (0.26)	0.42 (0.27)	0.55 (0.29)
Ext. prim (voc.)	0.11 (0.16)	0.10 (0.16)	0.18 (0.18)
Secondary (voc.)	0.11 (0.22)	0.10 (0.15)	0.20 (0.17)
Higher (non ac.)	0.07 (0.11)	0.07 (0.17)	0.16 (0.19)
Academic	0.27 (0.11)	0.26 (0.22)	0.38 (0.26)
Breadwinner	-0.45 (0.17)	-0.47 (0.13)	-0.55 (0.14)
Dutch nat.	-0.24 (0.24)	-0.27 (0.19)	-0.28 (0.17)
Western region	0.16 (0.04)	0.17 (0.09)	0.21 (0.10)
Constant	-1.14 (0.15)		

Table 3.2 (continued)

variable/parameter	model 1	model 2	model 3
(iii) duration dependence of exit rate out of the panel			
δ	0.83 (0.14)	0.87 (0.11)	1.04 (0.12)
(iv) distribution $h_s(u,v)$			
v_1.100			6.90 (1.20)
v_2.100			2.00 (1.46)
u_1.100		36.1 (15.0)	1872.0 (148320)
u_2.100		0.00 (0.09)	19.6 (6.90)
p_1		0.08 (0.03)	0.00 -
p_2		0.92 (0.03)	0.90 (0.04)
p_3			0.10 (0.04)
p_4			0.00 -
(v) other statistics			
CORR($r,a\|x$)			-0.13 (0.11)
Log Likelihood	-7786.13	-7786.06	-7777.58
due to p,r	6111.34		
due to a	1674.79	1674.73	

standard errors in parentheses

4. An alternative model

As noted in the introduction, one possible reason for a relation between the duration t spent in a state S (unemployment or being in a job), and the duration of panel survey participation a, is that individuals may drop out of the panel because they found a (different) job. If an individual moves to another address because of his new job, it may be hard for the agency conducting the survey to follow him. The kind of relation between t and a in such cases is different from the kind of relation modelled in Section 2. For example, here there is a positive probability that t and a are virtually equal, and there is a relation between $t|x$ and $a|x$ even if there are no unobserved explanatory variables. In this section we construct and estimate a model in line with the argument above.

For reasons of simplicity, we will for the moment abstract from unobserved explanatory variables v, and assume that (2.1) holds with $\alpha=1$. As a result, $p|x$ and $r|x$ are independent, and both have an exponential density with hazard $\exp(x_1'\beta_1)$. We now define the stochastic process characterizing attrition conditional on the stochastic process characterizing the labour market behaviour. In particular, we assume that exit out of the panel occurs at the rate $\exp(x_2'\beta_2)$ from the first interview onward. However, if, at the moment that the individual leaves state S, the individual still participates in the survey, then there is a probability $\pi\in[0,1]$ that exit from the panel occurs at that very moment. So, if it is only known that $a\geq r$, then, for given r, the probability that $a=r$ equals π. The variables r and a are independent if and only if $\pi=0$. It is clear that if, in this context, one would treat attrition as independent right-censoring of r, then the mean estimate of $\theta_0(x)$ would be underestimated if $\pi>0$. Note that we may allow π to depend on x. (It may be interesting to note that for $\zeta=0$ we get a model that is observationally equivalent to a so-called constant-product model proposed by Lagakos & Williams (1978).)

The conditional density $h_s(a|r,x)$ can be written as

$$h_s(a|r,x) = \zeta_0(x).e^{-\zeta_0(x).a}.I_{[0,r>}(a) \quad + \quad (1-\pi).\zeta_0(x).e^{-\zeta_0(x).a}.I_{<r,\infty>}(a)$$

(4.1)

$$P_s(a=r|r,x) = \pi.e^{-\zeta_0(x).r}$$

Clearly, it is not absolutely continuous. The density $h_s(r,a|x)$ follows from multiplication of (4.1) by the exponential density $h_s(r|x)$. From this it can be inferred that $P_s(r=a|x) = (\theta_0(x)/(\theta_0(x)+\zeta_0(x))).\pi$, so the probability that r and a occur simultaneously equals the probability that a does not occur before r, times the probability that a occurs when r occurs. From $h_s(\mathrm{r},a|\mathrm{x})$ it also follows that

$$(4.2) \qquad h_s(a|\mathrm{x}) = (1-\pi).\zeta_0(\mathrm{x}).\mathrm{e}^{-\zeta_0(\mathrm{x}).a} + \pi.(\theta_0(\mathrm{x})+\zeta_0(\mathrm{x})).\mathrm{e}^{-(\theta_0(\mathrm{x})+\zeta_0(\mathrm{x})).a}$$

which is a discrete mixture of two exponential densities. This implies that $h_s(a|\mathrm{x})$ displays negative duration dependence.

There holds that

$$(4.3) \qquad \mathrm{COV}(\mathrm{r},a|\mathrm{x}) = \frac{\pi}{(\theta_0(\mathrm{x})+\zeta_0(\mathrm{x}))^2}$$

so a negative relation between r and a is ruled out. Clearly, this means that the model is quite restrictive.

By multiplying $h_s(\mathrm{r},a|\mathrm{x})$ with the exponential density $h_s(\mathrm{p}|\mathrm{x})$ we obtain $h_s(\mathrm{p},\mathrm{r},a|\mathrm{x})$. The likelihood can be constructed from $h_s(\mathrm{p},\mathrm{r},a|\mathrm{x})$ along the lines of Subsection 2.2. The parameters to be estimated are π, β_1 and β_2. Estimation of the model on the unemployment duration data from Subsection 3.1 resulted in an estimate of π of 0.012 with standard error 0.14, so π is insignificantly different from zero. Estimation of the model on the job duration data from Subsection 3.2 forced the parameter π to the boundary (π=0) of the parameter space. This reconfirms the conclusions stated in Subsections 3.1 and 3.2. In light of these results, and in light of the restrictiveness of the model used here, we did not pursue more extensive analyses with this model.

5. Conclusion

In this paper we have analyzed the relation between individual labour market behaviour over time and the duration of participation in panel surveys. We used models which allow for dependence by way of stochastically related unobserved determinants of the duration of survey participation and the duration of being unemployed or the duration of being in a job.

We paid attention to the complications arising when the sample is from the stock of individuals in the labour market state of interest. We also showed that the family of bivariate discrete distributions with two points of support for both variables has some desirable properties as a model for the distribution of the unobserved components. These properties refer to flexibility as well as to the ease of interpretation of estimation results in terms of underlying population characteristics.

The empirical analysis suggests that unobserved explanatory variables for the duration of panel survey participation of an individual are not related to unobserved explanatory variables for the duration of unemployment of that individual. It seems therefore that survey participation duration is independent of unemployment duration. Consequently, in the empirical analysis of unemployment durations using panel data, spells that are incomplete due to attrition may be treated as spells that are subject to independent right–censoring. The results are confirmed by a number of sensitivity checks.

The empirical analysis using job spells showed a negative relation between job duration and the duration of survey participation. However, for the estimates of the exit rate out of a job it did not matter whether we took account of this or not, so for simplicity we might as well ignore it. It should be noted, however, that this result may be due to the abundant retrospective information on elapsed job durations that is present in the dataset used.

Since there could be a more direct relationship between the two types of duration (survey participation and labour market spell) than the relation by way of unobserved determinants, we also estimated an alternative model. In this model, a fraction of the individuals who start working in a new job leave the panel at the moment they start working in that job. However, we found this phenomenon to be empirically unimportant.

The models that are estimated are restrictive in the sense that they are heavily parameterized. A topic for further research would be to replicate the empirical analysis using more flexible specifications. In that case larger samples would be necessary, both in terms of numbers of respondents per wave and in terms of numbers of waves. Models allowing for non–monotonic duration dependence may be used to detect wave–specific effects in the exit out of the panel. Another topic for further research would be to estimate a model that simultaneously describes the durations of unemployment and employment, and the duration of panel survey participation. Such an analysis would integrate the two empirical analyses in this paper. Note that in that case we would have a three–dimensional distribution of unobserved heterogeneity. By using multi–state multi–spell data, the results may be less sensitive to the Mixed Proportional Hazard assumption.

ACKNOWLEDGEMENTS

The data were provided by The Netherlands Organization for Strategic Labor Market Research (OSA) and The Netherlands Central Bureau of Statistics (CBS). The empirical analysis based on the CBS data was carried out by Gerard J. van den Berg when he was at Tilburg University. We would like to thank Tony Lancaster and the referees for their constructive comments.

REFERENCES

Aalen, O.O. (1978), Nonparametric inference for a family of counting processes, *Annals of Statistics* 6, 701–726.

Butler, J.S., K.H. Anderson and R.V. Burkhauser (1989), Work and health after retirement, *Review of Economics and Statistics* 71, 46–53.

Chesher, A. and T. Lancaster (1983), The estimation of models of labour market behaviour, *Review of Economic Studies* 50, 609–624.

Devine, T. and N.M. Kiefer (1990), *Empirical labor economics: a search theory approach* (Oxford University Press, Oxford).

Ham, J.C. and S.A. Rea (1987), Unemployment insurance and male unemployment duration in Canada, *Journal of Labor Economics* 5, 325–353.

Heckman, J.J. (1979), Sample selection bias as a specification error, *Econometrica* 47, 153–161.

Heckman, J.J. and B.E. Honoré (1989), The identifiability of the competing risks model, *Biometrika* 76, 325–330.

Heckman, J.J. and B. Singer (1984), A method for minimizing the impact of distributional assumptions in econometric models for duration data, *Econometrica* 52, 271–320.

Lagakos, S.W. (1979), General right censoring and its impact on the analysis of survival data, *Biometrics* 35, 139–156.

Lagakos, S.W. and J.S. Williams (1978), Models for censored survival analysis: a cone class of variable–sum models, *Biometrika* 65, 181–189.

Lillard, L.A. (1989), Sample dynamics: some behavioral issues, in: D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, eds., *Panel surveys* (Wiley, New York).

Lindeboom, M. and J.J.M. Theeuwes (1991), Job duration in The Netherlands: the co–existence of high turnover and permanent job attachment, *Oxford Bulletin of Economics and Statistics* 53, 243–264.

Lindeboom. M. and G.J. van den Berg (1994), Heterogeneity in bivariate duration models: the importance of the mixing distribution, *Journal of the Royal Statistical Society Series B*, forthcoming.

Nickell, S.J. (1979), Estimating the probability of leaving unemployment, *Econometrica* 47, 1249–1266.

Pissarides, C.A. (1986), Unemployment and vacancies in Britain, *Economic Policy*, October.

Ridder, G. (1984), The distribution of single–spell duration data, in: G.R. Neumann and N. Westergård–Nielsen, eds., *Studies in labor market analysis* (Springer Verlag, Berlin).

Van den Berg, G.J. (1990), Search behaviour, transitions to nonparticipation and the duration of unemployment, *Economic Journal* 100, 842–865.

Van den Berg, G.J., M. Lindeboom and G. Ridder (1991), Attrition in longitudinal panel data, and the empirical analysis of dynamic labour market behaviour, Research Memorandum, Groningen University.

Van den Berg, G.J. and T. Steerneman (1991), The correlation of durations in multivariate hazard rate models, Research Memorandum, Groningen University.

Van den Berg, G.J. (1992), A structural dynamic analysis of job turnover and the costs associated with moving to another job, *Economic Journal* 102, 1116–1133.

Williams, J.S. and S.W. Lagakos (1977), Models for censored survival analysis: constant–sum and variable–sum models, *Biometrika* 64, 215–224.