A Note on Stock Sampling and Maximum Duration

Govert E. Bijwaard*
Econometric Institute
Erasmus University Rotterdam

Econometric Institute Report EI 2006-22

*Erasmus University Econometric Institute, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands; Phone: (+31) 10 40 81424; Fax: (+31) 10 40 89162; E-mail: bijwaard@few.eur.nl This research is financially supported by the Netherlands Organisation for Scientific Research (NWO) nr. 451-04-011

A Note on Stock Sampling and Maximum Duration

Abstract

An issue hardly ever mentioned in the analysis of labour market transitions is that for some individuals labour market transitions occur at a very low rate. Therefore, these individuals might stay on disability benefits or in domestic care till they reach the retirement age of 65. This implies that the duration on disability and of non-participating women has a upper bound of the time till retirement.

Despite the growing availability of panel data on labour market transitions many household surveys are still based on stock based sampling. In this paper estimation of a duration model in which a positive fraction of individuals reaches a maximum duration is derived for stock sampled data. A mixed proportional hazard model with a piecewise constant baseline hazard leads to a relatively simple closed-form expression in the log likelihood. Discrete unobserved heterogeneity is assumed. Non-constant entry rates into the labour market state are allowed for by assuming a yearly fluctuating rate.

JEL classification: C41; J64

Key words: Maximum duration; Stock sampling; Mixed proportional hazard; mover-stayer model

1 Introduction

In most Western countries unemployment benefits and disability benefits cease after retirement. In fact, everybody leaves the potential workforce when reaching the retirement age. An issue hardly ever mentioned in the analysis of labour market transitions is that for some individuals labour market transitions occur at a very low rate. Therefore, these individuals might stay on disability benefits or in domestic care till they reach the retirement age of 65. This implies that the duration on disability and of non-participating women has a upper bound of the time till retirement.

Despite the growing availability of panel data on labour market transitions many household surveys are still based on stock based sampling. In these surveys the data are obtained from the stock of people in a particular labour market state. The kind of problems generated with such sampling scheme are discussed in Nickell (1979), Ridder (1984), Heckman and Singer (1984a), Lancaster (1990) and Murphy (1996). The main problem is that sampling from the stock results in length-biased sampling, because the stock contains more longer durations in a particular state than the inflow to that state. Murphy (1996) showed that a mixed proportional hazard specification with a piecewise constant baseline hazard and a gamma unobserved heterogeneity distribution lead to a rather simple likelihood specification.

In this paper the model of Murphy (1996) is extended to account for the possibility that some individuals have a positive probability to reach a maximum duration, like the time till retirement. The gamma unobserved heterogeneity assumption is also relaxed. Finally, non-constant entry rates into the labour market state is allowed for by assuming a yearly fluctuating rate.

2 Maximum duration in a duration model

In duration analysis the hazard rate or intensity is usually modelled. A common way to accommodate the presence of observed characteristics is to specify a proportional intensity model

$$\lambda(t|x) = \lambda_0(t)e^{\beta' x_i(t)},$$

where $\lambda_0(t)$ represents the baseline hazard, that is, the duration dependence of the intensity common to all individuals. The covariates affect the intensity proportionally and the time-varying variables are external variables that change independent of the employment state, such as the age of a disabled individual.

Suppose that the duration of each individual has an upper bound of $\bar{t}_i$. An example of such a maximum duration is the time left till retirement of an individual on disability benefits. If a non-zero, albeit unknown, percentage p of the individuals reach the upper bound, the survival conditional on not have reached the maximum duration is (for $0 \leq t \leq \bar{t}_i$)

$$S(t|\bar{t}_i, x_i) = (1-p)\frac{\exp\bigl(-\int_0^t \lambda_0(s)e^{\beta' x_i(s)}\,ds\bigr) - \exp\bigl(-\int_0^{\bar{t}_i} \lambda_0(s)e^{\beta' x_i(s)}\,ds\bigr)}{1-\exp\bigl(-\int_0^{\bar{t}_i} \lambda_0(s)e^{\beta' x_i(s)}\,ds\bigr)} + p$$

Thus a fraction of $1-p$ the individuals will make a transition before the maximum duration is reached. This is an extension of a so-called mover-stayer model that accounts for an upper bound in the duration. This approach was developed by Boag (1949) and applied to model the recidivism of criminals (Schmidt and Witte 1989) and labor market transitions (Dunsmuir et al. 1989).[1] It assumes that a latent group of individuals have a zero probability

[1] Schmidt and Witte (1989) use the term 'split-population'model. In the biomedical literature the mover-stayer model is known as cure-model. Maller and Zhou (1996) discuss the implications of such models.

to make a transition, the *stayers.* Here the stayers are those individuals that reach the maximum duration.

If we do not account for possible missing variables, the parameter estimator may be biased. It may lead to spurious negative duration dependence. Therefore, it is important to allow for individual-specific unobserved heterogeneity in the model. The conventional way to capture this effect is to include a multiplicative random variable in the hazard to get a mixed proportional hazard model with hazard

$$\lambda(t|v_i, x_i) = v_i \lambda_0(t) \exp\bigl(\beta' x_i(t)\bigr), \tag{1}$$

where the $v_i > 0$ are i.i.d. random variables with (mixture) distribution function $G(v)$. The Gamma distribution with mean one and variance σ^2 is most often chosen to represent the unobserved heterogeneity. However, if the underlying distribution of the unobserved heterogeneity is not a gamma distribution the results may be biased. Any other mixture distribution, like the normal, or log normal distribution, have the same problem. More robust, and very flexible, is to assume that the mixture distribution can be approximated by a finite discrete mixture, see Heckman and Singer (1984b). For a discrete mixture model, there are a finite number of values or classes, v_l $(l = 1, \ldots, L)$, each having probability q_l $(l = 1, \ldots, L)$ in the population, where $\sum q_l = 1$. The observed survival function $\tilde{S}(t|\bar{t}_i, x_i)$ is obtained by integrating out v.

It is important to point out that the presence of stayers is compatible with a discrete mixture duration model. Heckman and Walker (1987) recognize that some specifications of the latent intensity can deliver stayers, like for one particular $l' : v_{l'} = 0$ with $q_{l'} > 0$. The close link between mover-stayer models and a discrete mixture model implies that the two can easily be combined.

3 Stock sampling

If we sample from a stock of individuals at time 0 (in calendar time) in a particular state, e.g. from the stock of people on disability benefits, and observe the elapsed time e in that state (together with some covariates), then the distribution of the observations e is a conditional distribution. The condition is the presence of a particular individual in the stock. First, consider the case of stock sampling with no unobserved heterogeneity. I follow Lancaster (1990) to derive the conditional distribution.

Abstracting from early retirement schemes the time till retirement, $\bar{t}_a$, depends only on the age, a, of the individual. Let $r(-e|a, x_i)$ be the entry rate, the probability to enter the state during $[e, e + de)$ in the past given x_i and age a. Denote by $S_{-y}(t|\bar{t}_a, x_i)$ the survival up to $t \leq \bar{t}_a$ of an individual of age a and with covariates x_i entering the state at y time ago. Then, the size of the observed stock at time 0 of individuals who entered at age a with observed covariates x_i is

$$\int_0^{\bar{t}_i} r(-s|a, x_i) S_{-s}(s|\bar{t}_a, x_i)\, ds \tag{2}$$

Of this total number of individuals, the number of individuals who have been in the state for at least a period of length t is

$$\int_t^{\bar{t}_i} r(-s|\bar{t}_i, x_i) S_{-s}(s|\bar{t}_i, x_i)\, ds \tag{3}$$

Thus the proportion of individuals in the population from which we sample who have been in the state for at least t time units is the ratio of (2) to (3). If we assume (i) there are no calendar time (business cycle) effects on the survival, that is the survival is independent of the entry time $S_{-s}(s|\bar{t}_i, x_i) = S(s|\bar{t}_i, x_i)$, (ii) a (Mixed) proportional hazard model as in (1), (iii) the age of the individual only enters the hazard through $\exp\big(\beta' x_i(t)\big)$ and, (iv) following

Nickell (1979), the entry rate is separable $r(-s|a, x_i) = r_1(-s)r_2(x_i)r_3(a)$, then the density of the elapsed duration is (in a PH model)

$$h(e|x_i, a) = \frac{r_1(-e)\Big[(1-p)e^{-\Lambda(e|x_i)} + \big(p - e^{-\Lambda(\bar{t}_a|x_i)}\big)\Big]}{(1-p)\int_0^{\bar{t}_a} r_1(-\tau)e^{-\Lambda(\tau|x_i)}\,d\tau + \big(p - e^{-\Lambda(\bar{t}_a|x_i)}\big)\int_0^{\bar{t}_a} r_1(-\tau)\,d\tau}$$

where $\Lambda(e|x_i) = \int_0^e \lambda_0(s)\exp\big(\beta' x_i(s)\big)\,ds$, the integrated hazard. In an MPH model with unobserved heterogeneity the terms $e^{-\Lambda(e|x_i)}$ and $e^{-\Lambda(\bar{t}_a|x_i)}$ in the density are replaced by $\int_0^\infty e^{-v\Lambda(e|x_i)}dG(v)$ and $\int_0^\infty e^{-v\Lambda(\bar{t}_a|x_i)}dG(v)$ respectively.

In practice it is hard to find a closed form solution to integrals in the density. For example, the commonly applied Mixed proportional hazard model with Weibull baseline hazard and unit-mean gamma distributed unobserved heterogeneity lead to intractable integrals. Although these integrals may be approximated, the Weibull baseline and gamma unobserved heterogeneity are also very restrictive.

4 Piecewise constant entry rates and baseline hazards

A reasonable assumption for entry into disability benefits is that the entry rate is constant on (yearly) intervals. A very flexible and tractable assumption is to use a piecewise constant baseline hazard. Then with a discrete unobserved heterogeneity we have a closed form expression for the density of the elapsed duration, from which we can easily derive a maximum likelihood estimator for the parameters of the model.

Suppose the baseline hazard is constant on M intervals. Let the intervals $I_m(t) = I(d_{m-1} \leq t < d_m)$ for $m = 1, \ldots, M$ with $d_0 = 0$ and $d_M = \max_i\{\bar{t}_i\}$ be the intervals on which we define the baseline hazard. Then, the baseline hazard is $\lambda_0(t) = \sum_{m=1}^{M} e^{\alpha_m} I_m(t)$. Suppose the time-varying covariates may only change on the same intervals then the integrated hazard is

$$\Lambda(t|x) = \sum_{m=1}^{M} \Big[J_m(t)(d_m - d_{m-1}) + I_m(t)(t - d_{m-1}) \Big] \exp(\alpha_m + x(d_{m-1})\beta)$$

where $J_m(t) = I(t > d_m)$. When the unobserved heterogeneity distribution is discrete the density becomes

$$\begin{aligned} h(e|x_i, a) \;\; = \;\; & \Bigg[r_1(-e) \Big[(1-p) \sum_{l=1}^{L} q_l e^{-v_l \Lambda(e|x_i)} + \big(p - \sum_{l=1}^{L} q_l e^{-v_l \Lambda(\bar{t}_a|x_i)} \big) \Big] \Bigg] \\ & \div \Bigg[(1-p) \int_0^{\bar{t}_a} r_1(-\tau) \sum_{l=1}^{L} q_l e^{-v_l \Lambda(\tau|x_i)} \, d\tau \\ & \quad + \big(p - \sum_{l=1}^{L} q_l e^{-v_l \Lambda(\bar{t}_a|x_i)} \big) \int_0^{\bar{t}_a} r_1(-\tau) \, d\tau \Bigg] \end{aligned}$$

If we assume a piecewise constant hazard and that the covariates and the entry rate may only change on the same intervals, the density involves the summation of M integrals[2]

$$\begin{aligned} & r(-d_m) \int_{d_{m-1}}^{d_m} \sum_{l=1}^{L} q_l \exp\big(-v_l \Lambda(\tau|x_i) \big) \, d\tau \\ & = \sum_{l=1}^{L} q_l \exp\Big(-\alpha_m - x_i(d_{m-1})\beta - v_l - e^{v_l} \sum_{j=1}^{m-1} (d_j - d_{j-1}) e^{\alpha_j + x(d_j)\beta} \Big) \\ & \quad - \sum_{l=1}^{L} q_l \exp\Big(-\alpha_m - x_i(d_{m-1})\beta - v_l - e^{v_l} \sum_{j=1}^{m} (d_j - d_{j-1}) e^{\alpha_j + x(d_j)\beta} \Big) \end{aligned}$$

The latent probability of reaching the maximum duration lays between zero and one and can be modelled in a logit form $p = 1/(1 + e^{\gamma})$ or in a log-log form $p = \exp(-e^{-\gamma})$. This proportion of survival till retirement can also depend on observed characteristics of the individuals.

[2] This is for notational convenience. If either the covariates or the entry rate or both changes at different points, we can just add additional change points.

5 Conclusion

When we have stock sampled duration data for durations that reach with a positive probability a maximum, like the duration on disability benefits, a mixed proportional hazard model with a piecewise constant baseline hazard leads to a relatively simple closed-form expression in the log likelihood. Bijwaard and Veenman (2006) apply these results to model data for four different ethnic groups in The Netherlands on the duration on unemployment benefits, on disability benefits and of women in domestic care.

References

Adddison, J. T. and P. Portugal (2003). Unemployment duration: Competing and defective risks. *Journal of Human Resources 38*, 156–191.

Bijwaard, G. E. and J. Veenman (2006). Unequal chances on the transitional labour market, The case of the Netherlands. Working paper, Erasmus University Rotterdam.

Boag, J. W. (1949). Maximum likelihood estimation of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society: Series B 11*, 15–44.

Dunsmuir, W., R. Tweedie, L. Flack, and K. Mengersen (1989). Modelling of transitions between employment states for young Australians. *Australian Journal of Statistics 31A*, 165–196.

Heckman, J. J. and B. Singer (1984a). Econometric duration analysis. *Journal of Econometrics 24*, 63–132.

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

Heckman, J. J. and J. R. Walker (1987). Using goodness of fit and other criteria to choose among competing duration models: A case study of Hutterite data. In C. C. Clogg (Ed.), *Sociological Methodology (1987)*, pp. 247–308. American Sociological Association: Washington DC.

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

Maller, R. A. and X. Zhou (1996). *Survival Analysis with Long-term Survivors*. Chichester: John Wiley.

Murphy, A. (1996). A piecewise–constant hazard-rate model for the duration of unemployment in single-interview samples of the stock of unemployed. *Economics Letters 51*, 177–183.

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

Ridder, G. (1984). The distribution of single–spell duration data. In G. R. Neumann and N. C. Westergaard-Nielsen (Eds.), *Studies in Labor Market Dynamics*, pp. 45–73. Springer–Verlag.

Schmidt, P. and A. D. Witte (1989). Predicting criminal recidivism using 'split population' survival time models. *Journal of Econometrics 40*, 141–159.