1 Chapter 3: Survival Analysis.

The analysis of duration data, also known as survival analysis, was developed to describe the timing of events such as when a person dies (thus the name) or when a machine breaks. These statistical techniques have become a subject of increasing interest in economics, especially labor economics. Numerous empirical papers have addressed such issues as unemployment duration (Lancaster 1979 and Nickell 1979), the effects of unemployment benefits on the spells of unemployment (Moffit 1985, Solon 1985, Meyer 1990, and McCall 1996), turnover (Burdett et al. 1985), occupational matching (McCall 1990), retirement (Diamond and Hausman 1984), strike length (Kennan 1985 and Gunderson and Melino 1990), and job search (Jovanovic 1984). Sometimes this type of technique is regarded as a reduced form for behavioral economic theories like the theory of job search or the theory of job matching. More appropriately, it is a flexible approximation of behavior or an informative method of describing the data.

Consider a random variable T which takes positive values and describes the length of time until an event of interest occurs, usually defined as exit time. Assume that the distribution of duration T can be specified by a distribution function $F(t)$, with associated density function $f(t)$. Other functions of interest associated with the duration process are the survivor function $S(t) = 1 - F(t)$ which represents the probability that a spell will last a time period t or longer, and the hazard function

$\lambda(t)$ which represents the instantaneous probability of the spell ending at t conditional on it not having ended prior to t. The hazard function can be written as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{\Delta F(t)}{\Delta t} S\left(t\right)^{-1} = \frac{f(t)}{S(t)}. \qquad (1)$$

From the previous definition, it is evident that, given a certain density, we can determine the associated hazard function; as we will show later on, the opposite is also true. The derivative of the hazard function $\partial\lambda(t)/\partial t$ is called the duration dependence. If $\partial\lambda(t)/\partial t > 0$, there is positive duration dependence, and if $\partial\lambda(t)/\partial t < 0$, there is negative duration dependence. Given the hazard rate for a certain duration process, it is easy to determine the associated survival function. Observe that, $\lambda(t) = -d \ln S(t)/dt$. Then, solving a simple differential equation, we obtain

$$\begin{aligned} S(t) &= \exp\left\{-\int_0^t \lambda(z)dz\right\} = \exp\left\{-\Lambda(t)\right\}, \qquad (2) \\ f\left(t\right) &= \lambda(t)S(t) = \lambda(t)\exp\left\{-\Lambda(t)\right\} \end{aligned}$$

where $\Lambda(t) = \int_0^t \lambda(z)dz$ is the integrated hazard function.

1.1 Simple Duration Models.

Consider a simple model of job search. Workers are infinitely lived, wealth maximizing agents and discount the future at a rate $(1+r)^{-1}$. At each moment, a worker can be employed or unemployed. Unemployed workers search actively for job offers. Offers are distributed according to the density $f(w)$ with finite mean. At each period, an unemployed worker incurs a cost of search c and receives a wage offer from one of many firms in the labor market. A wage offer represents a random draw from the time-invariant distribution of wages known to the worker. Once a job is accepted, employment lasts forever. The problem of the worker has the reservation property: there exists a value ξ, called the reservation wage, such that it is optimal for the worker to accept any wage offer if it is not smaller than ξ and to reject it otherwise. In addition, assume that at each point in time the instantaneous probability of receiving a job offer is λ_e. In this framework, if we denote the instantaneous probability of leaving the unemployment state as λ_e, this probability can be defined as the product of the instantaneous probability of receiving a job offer times the probability that the job offer is acceptable,

$$\lambda_e = \lambda \left(1 - F(\xi)\right) \tag{3}$$

with $F(\bullet)$ representing the distribution of wates. In this case, the integrated hazard has a simple expression

$$\Lambda(t) = \int_0^t \lambda_e(z) dz = \int_0^t \lambda_e dz = \lambda_e t$$

From equation (2),

$$S(t) = \exp\left\{-\lambda_e t\right\},$$

or

$$F(t) = 1 - \exp\{-\lambda_e t\}.$$

Observe that this also represents the exponential distribution function with mean $E(t) = \lambda_e^{-1}$. The following graph represents the shapes of the hazard and survival functions for the exponential distribution.

Figure: survival function for a distribution with

hazard $\lambda = 1$ and $\lambda = 2$.

It remains to show how the reservation wage is obtained. Each worker makes career decisions that maximize his/her intertemporal utility. For simplicity, assume that utility is only a function of income. The probability of receiving a wage offer in time interval Δt, for a certain individual, is $\lambda_e \Delta t + o(\Delta t)$ for. The probability of two or more job offers in interval Δt is negligible. Using Bellman's optimality principle for dynamic programming the intertemporal utility of the unemployment state, V_u, and the employment state, $V_e(w)$, are the unique solution to the following equations,

$$V_u = \frac{b\Delta t}{1+r\Delta t} + \frac{(1-\lambda_e \Delta t)}{1+r\Delta t} V_u^i + \frac{\lambda_e \Delta t}{1+r\Delta t} E_w \max\left\{V_u, V_e(w)\right\} + o(\Delta t) \tag{4}$$

$$V_e(w) = \frac{w\Delta t}{1+r\Delta t} + \frac{V_e(w)}{1+r\Delta t} + o(\Delta t) \tag{5}$$

The first term on the right of (4) is the discounted value of remaining unemployed in interval Δt, with b representing the value of leisure. The second term is the probability of not receiving an offer $(1-\lambda_u \Delta t)$ times the discounted value of search at the end of interval Δt. The third term is the probability of receiving a wage offer $(\lambda_e \Delta t)$ times the discounted value of the expected value of the maximum of the two options confronting the agent who receives a wage offer: to continue searching (with present value V_u) or to take the offer (with present value $V_e(w)$). The first term on the right of (5) is the discounted value of working, at a wage w, in interval Δt. The second term is the present value of employment, at wage w, $V_e(w)$.

Collecting terms in equations (4) and (5) and passing to the limit, we obtain

$$rV_u = b + \lambda_u E_w \max\left\{0, V_e(w) - V_u\right\} \tag{6}$$

$$rV_e(w) = w \tag{7}$$

where, $V_e(w)$ is strictly increasing with respect to w.[1] The reservation wage, denoted ξ, is the unique solution to the equation

$$V_u = V_e(\xi). \tag{8}$$

or

$$\xi = rV_u \tag{9}$$

From equations (6) and (9),

$$\xi = b + \lambda_u E_w \max\left\{0, \frac{w}{r} - \frac{\xi}{r}\right\}$$

and we obtain the familiar expression

$$\xi = b + \frac{\lambda_u}{r} \int_{\xi}^{+\infty} (w - \xi)\, dF. \tag{10}$$

In this framework ξ is constant and independend of the lenght of the unemployment spell. However, in many cases of practical interest the value of ξ may change with the lenght of the unemployment spell. In the theoretical framework developed above this scenario can be represented by assuming that the value of b varies over time. Consider then

$$\xi_t = b(t) + \frac{\lambda_u}{r} \int_{\xi_t}^{+\infty} (w - \xi_t)\, dF \tag{11}$$

[1]Intuitively, a higher wage means a higher utility of the working state.

by analogy with (10). Making use of the implicit function theorem,

$$\frac{\partial \xi_t}{\partial t} = \frac{\partial b(t)}{\partial t} - \frac{\lambda_u}{r}\left[(\xi_t - \xi_t) f(\xi_t) \frac{\partial \xi_t}{\partial t} + S(\xi_t) \frac{\partial \xi_t}{\partial t}\right]$$

or

$$\frac{\partial \xi_t}{\partial t} = \left[1 + \frac{\lambda_u}{r} S(\xi_t)\right]^{-1} \frac{\partial b(t)}{\partial t}. \tag{12}$$

Consider for example an unemployed worker receiving unemployment insurance, for a short period of time the worker receiving unemployment insurance will benefit from a higher level of income at the unemployment state. Eventualy, unemployment insurance payments are reduced or completely exhausted. In this case,

$$\frac{\partial b(t)}{\partial t} < 0 \longrightarrow \frac{\partial \xi_t}{\partial t} < 0$$

by (12). Furthermore, taking into account that $\lambda_e = \lambda(1 - F(\xi_t))$,

$$\frac{\lambda_e}{\partial t} > 0,$$

that is, the instantaneous probability of leaving the unemployment state increases as the unemployment benefits get exhausted. This feature of the U.S. unemployment system has been confirmed in empirical work (Meyer 1990). From an econometric perspective this emphasizes the need to account for duration models with flexible specifications for the hazard function. A certain distribution $F(t)$ is said to have positive duration dependence if the associated hazard is increasing over time and it is said to have negative duration dependence if it is decreasing over time.

The Weibull distribution allow for both types of dependence. It is probably for this reason that the Weibull distribution is so popular among applied economists. The Weibull distribution function has an expression of the form

$$\begin{aligned} F(t) &= 1 - \exp\{-\lambda t^{\alpha}\} \text{ and} \\ h(t) &= \alpha\lambda t^{\alpha-1}. \end{aligned}$$

Observe that $\partial h(t) / \partial t = \alpha(\alpha - 1)\lambda t^{\alpha-2}$ thus

α	$\partial h(t) / \partial t$	*duration dependence*
< 1	*negative*	*negative*
$= 0$	*zero*	*constant*
> 1	*positive*	*positive*

The next figure shows plots of the weibull hazard function for different values of the parameters.

Weibull hazard function.

(1) $\alpha = .5, \lambda = 2$, (2) $\alpha = 1.5, \lambda = 2$

(3) $\alpha = 1.5, \lambda = .5$, (4) $\alpha = .5, \lambda = .5$

Notice that, λ represents a [normalization?] parameter wile α represent a [?] parameter. The Weibull model is very popular among applied econometrician. Probably because of its hability to accomodate positive, negative and zero dependence.

Other type of duration model with simmilar characteristics is the log-logistic distribution. This model is characterized by a hazard function

$$h\left(t; \lambda, \alpha\right) = \frac{\lambda \alpha t^{\alpha - 1}}{1 + \lambda t^{\alpha}}.$$

with survivor function

$$S\left(t; \lambda, \alpha\right) = \left(1 + \lambda t^{\alpha}\right)^{-1}.$$

Compared with the Weibull, the log-logistic distribution allows different forms of duration dependence for different values of t, given $\left(\lambda, \alpha\right)$.

Other types of distributions used in duration analysis are the Gamma and the lognormal. The books by Kalbfleisch and Prentice (1980), Cox and Oakes (1984) or Lancaster (1990) provide additional examples.

1.1.1 Parametric estimation of duration models.

Consider a data set of the form $\{(t_i\}_{i=1}^{N}$ from a population of individuals where t_i represents the spell of time until an event of interest occurs for observation i. Let the distribution of duration be specified as $F(\bullet)$ with associated density function $f(\bullet)$. Cosider also the survivor function $S(\bullet) = 1- F(\bullet)$ and the hazard function $\lambda(\bullet)$. Furthermore, in most data sets we should account for the possibility of right censoring. A right censored observation is one where the end of the spell has not occurred by the end of the time of observation. For example, consider a sample of newly unemployed people observed between dates τ_1 and τ_2. Any unemployment spells that have not ended by τ_2 are right censored. For those spells, we do not know when, if ever, they would have ended; we know only that the spell had not ended by τ_2. An alternative cause of censoring would occur in unemployment insurance administrative records where the individual is observed until his UI benefits run out. When they run out, the researcher does not know whether the unemployment spell ended in a job or if just the UI coverage ended. This does not represent a problem as long as the censoring process is independent of the data generating process. Consider a sample of the form $\{(t_i, c_i)\}_{i=1}^{N}$, where c_i is an indicator of whether spell i is censored ($c_i = 1$ iff spell i is censored). In this case, the probability of an observation (t_i, c_i) will be $f(t_i) = \lambda(t_i) S(t_i)$ for $c_i = 0$ and $S(t_i)$ for $c_i = 1$. Thus, the log likelihood function is

$$\begin{aligned} LLF &= \sum_{i=1}^{N} (1 - c_i) \log f(t_i) + c_i \log S(t_i) \\ &= \sum_{i=1}^{N} (1 - c_i) \log \lambda(t_i) - \Lambda(t_i) . \end{aligned} \tag{13}$$

Given a certain parametric specification of the hazard, the density, the survival or the distribution function, the parametric log-likelihood function can be fully specified. In empirical applications, it is common to start with the specification of the hazard function. The parameter estimates can finally be obtained by classical maximum likelihood methohds. The general properties of MLE studied in previous chapters apply here.

1.1.2 Non-parametric estimation of duration models.

Consider a data set of the form $\{t_i\}_{i=1}^{N}$ from a population of individuals where t_i represents the spell of time until an event of interest occurs for observation i, or exit time. Define $t^{(1)} = \min\{t_i \mid i = 1, ..., N\}$ representing the first exit time in the sample. Simmilarly, define $t^{(2)}, t^{(3)}, ..$ and $t^{(M)}$, as the second, third and M-th exit time, respectively. Taking into account that different exits can occur at the same time $M \leq N$. In addition, define $n^{(1)}, n^{(2)}, ..., n^{(m)}$ as the number of exits that occur at each particular exit time. Using the analog principle, one can define an estimator for the probability of exit at time $t^{(j)}$ as

$$\hat{\lambda}_j = \frac{n_j}{N - \sum_{k=1}^{j-1} n_k}. \tag{14}$$

Similarly, the probability of survival at time t can be estimated by

$$\begin{aligned}\hat{S}(t) &= 1 - \frac{\sum_{\left\{k|t^{(k)}\leq t\right\}} n_k}{N} = \frac{N - \sum_{\left\{k|t^{(k)}\leq t\right\}} n_k}{N} = \\ & \frac{N-n_1}{N}\frac{N - \sum_{k=1,2} n_k}{N-n_1}\frac{N - \sum_{k=1,..,3} n_k}{N - \sum_{k=1,2} n_k}\cdots\frac{N - \sum_{k=1,..,h} n_k}{N - \sum_{k=1,...,(h-1)} n_k}\end{aligned}$$

or

$$\hat{S}(t) = \prod_{j=1}^{h}\left(1 - \hat{\lambda}_j\right) \tag{15}$$

where h is the largest k such that $t^{(k)} \leq t$.

Up to this point we have not accounted for the posibility of censoring. Define r_j as the number of observations at risk of exit an instant before $t^{(j)}$ and n_j as the number of exits at time $t^{(j)}$. Observations censored before time $t^{(j)}$ are not taken into account. Observe that, under random censoring the sample of n_j observations at risk of exit at time $t^{(j)}$ represent a random sample. Therefore, an estimator for the probability of exit at time can be defined as

$$\hat{\lambda}_j = \frac{n_j}{r_j}.$$

With this definition, ecuation (15) can be extended to samples with random censoring. The estimator in (15) is called the Kaplan-Meier survivor function estimator. This is a non-parametric estimator since no assumptions have been made about the distribution of $\tilde{t}$.

1.2 Accounting for Heterogeneity

Up to this point we have considered a simple duration process that does not depend on additional covariates. From an economic perspective, the main concern is usually to study the impact of key exogenous variables on the distribution of T. Consider a data set of the form $\{(t_i, X_i)\}_{i=1}^{N}$ from a population of individuals where t_i represents the spell of time until an event of interest occurs for observation i and X_i represent a vector of characteristics for i. The distribution of duration for agent i can be specified as $F(t \mid X_i)$ with associated density function $f(t \mid X_i)$. Similarly, we can define the survivor function $S(t \mid X_i) = 1- \ F(t \mid X_i)$ and the hazard function $\lambda(t \mid X_i)$. Furthermore, in most data sets we should account for the possibility of right censoring. As we have already mentioned in the previous section, this does not represent a problem as long as the censoring process is independent of the data generating process. Consider a sample of the form $\{t_i, c_i, X_i\}_{i=1}^{N}$, where c_i is an indicator of whether spell i is censored ($c_i = 1$ iff spell i is censored). In this case, the probability of an observation (t_i, c_i, X_i) will be $f\left(t_i \mid X_i\right) = \lambda\left(t_i \mid X_i\right) S\left(t_i \mid X_i\right)$ for $c_i = 0$ and $S\left(t_i \mid X_i\right)$ for $c_i = 1$. Thus, the log likelihood function is

$$LLF = \sum_{i=1}^{N} (1 - c_i) \log f\left(t_i \mid X_i\right) + c_i \log S\left(t_i \mid X_i\right)$$

$$= \sum_{i=1}^{N} (1 - c_i) \log \lambda (t_i \mid X_i) - \Lambda (t_i \mid X_i) .$$

In empirical applications, it is common to start with the specification of the hazard function. Equations (3.2) and (3.3) show how to use $\lambda (t \mid X)$ to specify the log likelihood function. Consider the following hazard function:

$$\lambda (t \mid X) = \lambda_0 (t) \lambda_1 (X, \beta)$$

where $\lambda_0 (t)$ is called the baseline hazard. If there is enough variation in the X variables and all of them are observable, $\lambda_0 (\bullet)$ and $\lambda_1 (\bullet)$ are identified up to a constant. This specification is known as the proportional hazard model. It includes most of the parametric models considered in empirical applications such as Lancaster (1979), Solon (1985), and Narendranathan, Nickell, and Stern (1985). In this case, the function $\Lambda (\bullet)$ can be written as

$$\Lambda (t \mid X) = \lambda_1 (X, \beta) \int_0^t \lambda_0 (s) \, ds.$$

Usually $\lambda_1 (X, \beta)$is modelled as $\exp \{X\beta\}$.[2] In this case we obtain

$$\lambda (t \mid X) = \lambda_0 (t) \exp \{X\beta\} .$$

If $\lambda_0 (t) = 1$, equation (3.6) becomes the hazard function associated with an exponential distribution with parameter $\exp \{X\beta\}$; $F (t) = 1 - \exp \{-\lambda_1 t\}$. If $\lambda_0 (t) = \alpha t^{\alpha - 1}$,

[2] In particular, this specification is equivalent to the one introduced by Cox (1972).

equation (3.6) becomes the hazard of a Weibull distribution; $F(t) = 1 - \exp\{-\lambda_1 t^{\alpha}\}$. The proportional hazard model does not arise from any theoretical economic model. Its popularity is probably due to the fact that the estimated parameters provide a straightforward interpretation: the estimated coefficients are the derivative of the log hazard with respect to the associated X variable.

Up to this point, we have assumed that all the variables that matter for the duration process are included in X and are observed by the econometrician. Even in early work (Lancaster 1979 and Lancaster and Nickell 1980), researchers were aware that ignoring the possibility of omitted variables in a duration model can heavily bias the included parameter estimates and lead to misleading conclusions. To illustrate this point, suppose that the proportional hazard model is the correct specification for the study of unemployment spells. In addition, assume that the econometrician does not observe all the variables affecting the duration of unemployment; i.e., $\lambda(t \mid X) = \exp\{X\beta + \varepsilon\}$ where ε measures the effect of omitted variables. For the sake of concreteness, assume the ε takes on two values, $\varepsilon^{+} > \varepsilon^{-}$, and that $\Pr[\varepsilon = \varepsilon^{+}] = \Pr[\varepsilon = \varepsilon^{-}] = 1/2$ at time zero. As the process evolves, people with high values of ε $(= \varepsilon^{+})$ will leave the sample faster than people with low values ε $(= \varepsilon^{-})$. This will change the relative proportions of people (with respect to ε) toward people with low values of ε. As time goes by, since the average value of ε is falling, the average hazard rate will fall. Thus, we will observe decreasing hazards, even after controlling for observed heterogeneity (the X's) even though the hazard for each particular worker is constant over time. Consequently, ignoring the problem of omitted variables will lead to negative duration dependence bias. Because of the nonlinearity of the model, it also leads to inconsistent estimates of all of the model's parameters (Flinn and Heckman 1982).

It is important to account for unobserved heterogeneity otherwise our results may lead to the wrong conclussions. For example, in the previous scenario we can wrongly conclude that the lenght of unemployment spell decreases the probability of employment. In the previous framework this is not a true statement. What is true there is that the unemployment process is correlated with a selection process. That is, more able workers find job after a short period of unemployment. The workers that remain unemployed after a long period of time are those with the lowest probability of finding a job. In this model duration dependence in the sample of unemployed workers is not the result of a "stigma" effect due to unemployment, but the result of a process in which more able workers find work at faster rates. This effect has important policy implications, the possible solution to long-term unemployment will be different if this is the result of a "stigma" effect as compared with this being the result of a lack of skills.

In order to be able to identify the effects of unobservables and observables, it is necessary to make explicit assumptions about the way in which they interact.[3] Assume then that the true specification of the hazard, for a certain agent i in the sample, is as follows:

$$\lambda(t \mid X_i, \nu_i) = \lambda(t \mid X_i)\, \nu_i = \lambda_0(t)\, \lambda_1(X_i, \beta)\, \nu_i$$

where X_i represents the set of observed variables, and $\nu_i = \exp\{\varepsilon_i\}$ summarizes the effect of any other variable that affects the duration process and is not observed (unobserved heterogeneity). Then the survival function is

$$\begin{aligned} S(t \mid X_i, \nu_i) &= \exp\left\{-\int_0^t \lambda(z \mid X_i, \nu_i)\, dz\right\} = \exp\{-\Lambda(t \mid X_i, \nu_i)\} \\ &= \exp\{-\Lambda(t \mid X_i)\, \nu_i\}, \\ f(t_i \mid X_i, \nu_i) &= \lambda(t \mid X_i, \nu_i)\, S(t \mid X_i, \nu_i). \end{aligned}$$

In this case, the LLF for the sample $\{t_i, c_i, X_i, \nu_i\}_{i=1}^N$ is

$$\sum_{i=1}^N (1 - c_i) \log f(t_i \mid X_i, \nu_i) + c_i \log S(t_i \mid X_i, \nu_i).$$

In principle, this expression cannot be used to estimate the model because the values of $\{\nu_i\}_{i=1}^N$ are not observed. This problem can be overcome if the distribution generating the unobserved heterogeneity is known. Then we can define a LLF based on the marginal probabilities once the unobserved heterogeneity has been integrated out. More precisely, assume that ν_i represents a particular realization from a distribution

[3] See Heckman (1991) for an iluminating review of this issue.

$G(\bullet)$ independent of X_i. It is usual in this case to choose the normalization $E(\nu) = 1$.[4] We can define

$$S(t \mid X) = \int S(t \mid X, \nu) dG(\nu) = \int \exp\{-\Lambda(t \mid X) v\} dG(\nu) = M_\nu [-\Lambda(t \mid X)]$$

and

$$\begin{aligned} f(t \quad \mid \quad X) &= \frac{-\partial S(t \mid X)}{\partial t} \\ &= \lambda(t \mid X) \int \nu \exp\{-\Lambda(t \mid X) v\} dG(\nu) = \lambda(t \mid X) M_\nu^{(1)} [-\Lambda(t \mid X)] \end{aligned}$$

where $M_\nu[\bullet]$ and $M_\nu^{(1)}[\bullet]$ represent the moment generating function of ν and its derivative, respectively.[5] We can now specify the LLF for the sample $\{t_i, c_i, X_i\}_{i=1}^N$ using the integrated, or marginal probabilities:

$$LLF = \sum_{i=1}^N (1 - c_i) \log f(t_i \mid X_i) + c_i \log S(t_i \mid X_i)$$

$$= \sum_{i=1}^N (1 - c_i) \left\{ \log \lambda(t \mid X) + \log M_\nu^{(1)} [-\Lambda(t \mid X)] \right\} + c_i \log M_\nu [-\Lambda(t \mid X)].$$

Lancaster (1979) lets

$$\lambda(t \mid X_i, \nu_i) = \alpha t^{\alpha - 1} \mu_i \nu_i$$

[4] Observe that this is not a restriction because a constant is included in $X_i \beta$, and any $E(\nu) \neq 1$ would be captured in the constant.

[5] For some distributions G, the moment generating function does not exist. More generally $M_\nu[-\Lambda(t \mid X)]$ is the Laplace transform of G.

where $\mu_i = \exp\{X_i\beta\}$ and the unobserved heterogeneity has a density $g(\nu) \propto \nu^{\sigma-1}\exp\{-\nu\sigma\}$. In this case, $M_\nu[z] = [1-(z/\sigma)]^{-\sigma}$ and we obtain:

$$S(t \mid X) = M_\nu[-\Lambda(t \mid X)] = \left(1 + \frac{\mu I(t)}{\sigma}\right)^{-\sigma},$$

where $\mu_i = \exp\{X_i\beta\}$ and $I(t) = \int_0^t \alpha t^{\alpha-1}dt$. From this expression, it is straightforward to obtain $f(t \mid X)$ and to construct the LLF of the sample. Using MLE, we can obtain consistent estimators of the parameters associated with the duration process (α, β, σ).

It is important to recognize the significance of controlling for unobserved heterogeneity. The true hazard for an individual with unobserved heterogeneity ν is

$$\lambda(t \mid X, \nu) = \alpha t^{\alpha-1}\mu\nu.$$

By controlling for unobserved heterogeneity, we consistently estimate the expected hazard

$$\lambda(t \mid X) = E_\nu\{\lambda(t \mid X, \nu)\} = \alpha t^{\alpha-1}\mu.$$

On the other hand, if we control just for the observed characteristics X, the survival function of the duration process is given in equation (3.14) with associated hazard

$$\lambda^*(t \mid X) = \frac{\partial \log S(t \mid X)}{\partial t} = \alpha t^{\alpha-1}\mu S(t \mid X)^{1/\sigma} = \lambda(t \mid X)\, S(t \mid X)^{1/\sigma}.$$

If we do not control for unobserved heterogeneity, we obtain a consistent estimator for $\lambda^*(\bullet)$ in equation (3.17). As t increases, the survival probability goes to zero and the hazard $\lambda^*(t \mid X)$ decreases over time with respect to the true hazard $\lambda(t \mid X)$; this confirms our previous intuition. This argument can be generalized easily by observing that

$$\lambda^*(t \mid X) = E\left(\lambda\left(t \mid X, \nu\right) \mid T > t\right) = \lambda\left(t \mid X\right) E\left(\nu \mid T > t\right)$$

where T represents the time of employment and $E\left(\nu \mid T > t\right)$ represents the expected value of the unobserved heterogeneity among the remaining unemployed workers at time t. Under the assumption that ν is independent of the observed heterogeneity, this expectation should be decreasing over time.

We have learned that it is important to take into account the possibility of unobserved heterogeneity. This is especially true in many applications of duration analysis to labor economics in which the concern has been to study the effect of unemployment benefits or the behavior of the unemployed workers over the spell of unemployment. In this sense, we have shown that by ignoring the existence of unobserved heterogeneity in the sample, we are potentially estimating more negative duration dependence than actually exists.

In the previous example, given data on spell duration and observed characteristics of the workers, it was possible to estimate the distribution of unobserved heterogeneity, the baseline hazard, and the effect of observed characteristics on the probability of leaving the unemployment state. In order to achieve that goal, it was necessary to make strong parametric assumptions about the form of the hazard function and the distribution of unobserved heterogeneity. In most cases, we may have little prior information about the correct distribution of unobserved heterogeneity, and we therefore may produce misleading results by misspecifying this distribution. It is then pertinent to ask to what extent the available data can provide a nonparametric identification of each one of the relevant functions separately. Lancaster and Nickell (1979) find that "it seems in practice very difficult to distinguish between the effects of heterogeneity and the effect of pure time variation in the hazard function." Elbers and Ridder (1982) show that, at least for the proportional hazard specification, it is possible to identify nonparametrically the distribution of unobserved heterogeneity independently of the other components of the duration process as long as there is enough variation in the observed characteristics. The intuition for this result is that changes in the duration variable and the covariates allow us to trace out the different components of the hazard. This result depends crucially on the form of the proportional hazard model, in particular, the separability of the hazard into one function of the duration, another of the covariates and the independence of the unobserved characteristics respect to the observed ones.[6] Gurmu, Rilstone and Stern (1996) present a similar result for the proportional hazard model allowing for the possibility of interactions between t and some of the covariates. The identification result of Elbers and Ridder (1982) and Heckman and Singer (1984a) opens the possibility of nonparametric estimation of the

[6] A similar result in the context of risk models can be found in Heckman and Honore(1989).

distribution of the unobserved heterogeneity.

1.2.1 Semiparametric estimation of Duration Models.

Authors like Lancaster (1979) and Heckman and Singer (1984b) have shown that ignoring unobserved heterogeneity can lead to biased estimates of the parameters of the hazard function. In addition, Heckman and Singer (1984b) have shown that different specifications for the distribution of the unobserved heterogeneity can lead to very different estimates. This finding led them to propose a flexible nonparametric method to control for unobserved heterogeneity. In this paper we pursue a heuristic description of this method. Readers interested in a more technical description should refer to the original paper. From equation (3.12), we obtain

$$LLF = \sum_{i=1}^{N}(1-c_i)\left\{\log\lambda\left(t \mid X\right) + \log M_{\nu}^{(1)}\left[-\Lambda(t \mid X)\right]\right\} + c_i \log M_{\nu}\left[-\Lambda(t \mid X)\right].$$

where

$$\begin{aligned} M_{\nu}\left[-\Lambda(t \mid X)\right] &= \int \exp\left\{-\Lambda(t \mid X)\nu\right\} dG\left(\nu\right) \\ M_{\nu}^{(1)}\left[-\Lambda(t \mid X)\right] &= \int \nu \exp\left\{-\Lambda(t \mid X)v\right\} dG\left(\nu\right). \end{aligned} \qquad (3.20)$$

The Heckman and Singer approach consists of approximating the unknown distribution of unobserved heterogeneity $G(\nu)$ with a discrete distribution with positive probability mass at a finite number of points. Consider then the set of points $\{\nu_1, \nu_2, ..., \nu_k\}$

with probability mass $\pi_j > 0$ associated to ν_j, $j = 1, ..., k$ and $\sum_{j=1}^{k} \pi_i = 1$. Substituting the discrete approximation of $G(\nu)$ in equation (3.20), we obtain

$$\begin{aligned} \hat{M}_\nu \left[-\Lambda(t \mid X)\right] &= \sum_{j=1}^{k} \pi_j \exp\left\{-\Lambda(t \mid X)\nu_j\right\} \\ \hat{M}_\nu^{(1)} \left[-\Lambda(t \mid X)\right] &= \sum_{j=1}^{k} \pi_j \nu_j \exp\left\{-\Lambda(t \mid X)\nu_j\right\} \end{aligned} \tag{3.21}$$

where $\hat{M}_\nu[\bullet]$ and $\hat{M}_\nu^{(1)}[\bullet]$ represent the approximations to the true functions, $M_\nu[\bullet]$ and $M_\nu^{(1)}[\bullet]$ respectively. Substituting of the approximations, $\hat{M}_\nu[\bullet]$ and $\hat{M}_\nu^{(1)}[\bullet]$, for the true functions, $M_\nu[\bullet]$ and $M_\nu^{(1)}(\bullet)$, in equation (3.19), we obtain an approximation for the LLF,

$$LLF_k = \sum_{i=1}^{N} (1 - c_i) \left\{ \log \lambda\left(t \mid X\right) + \log \hat{M}_\nu^{(1)} \left[-\Lambda(t \mid X)\right] \right\} + c_i \log \hat{M}_\nu \left[-\Lambda(t \mid X)\right]. \tag{3.22}$$

The method proceeds by obtaining estimators of $\{\nu_j, \pi_j\}_{j=1}^{k}$ and estimators of the parameters of the hazard function. These estimators will be the values that maximize equation (3.22). Observe that this problem is not a standard MLE problem because the number of parameters necessary to estimate the approximate distribution of unobserved heterogeneity, $2k$, can be in principle infinite and the asymptotics rely on $k \to \infty$. In practice, the estimation strategy consists of choosing the k for which LLF_k stops growing in k. Sometimes researchers use a criterion such as the Aikeke number; most of the time researchers increase k until the LLF stops growing (Card and Sullivan 1988, Gunderson and Melino 1990, and Gritz 1993) or fix k ahead of time (Heckman and Walker 1990, Behrmann, Sickles, and Taubman 1990, and Johnson and Ondrich 1990). One major criticism of this approach is the lack of asymptotic distribution theory for the parameters estimated . In practice, the results of the estimation will depend of the parameters of the unobserved heterogeneity.

Some authors, as for example Han and Hausman(1990) and Meyer (1990), have shown that a nonflexible specification of the baseline function $\lambda_0(t)$ can bias the estimates of the other parameters. They suggest to estimate $\lambda_0(t)$ semi-parametrically assuming that it can be represented as a step function,

$$\lambda_0(t) = \lambda_{0k} \text{ for } t \in [t_{k-1}, t_k), \text{ with } t_0 = 0.$$

This type of specification may be convinient whe the data provides information about duration in discrete time (for example years, months, weeks,...). Furthermore, these authors have argued that the biases in the proportional hazard model may be larger for misspecification of the baseline hazard than for misspecification of heterogeneity distribution. For a similar model, Sueyoshi (1992) presents Monte Carlo evidence indicating that estimates are sensitive to misspecification of the unobserved heterogeneity distribution. In addition, this type of misspecification yields biased estimates of the baseline function $\lambda_0(t)$. In conclusion, the literature on this subject shows that both types of misspecification are important and should be taken into account when estimating a proportional hazard model.[7]

1.3 Duration models in regression form.

Consider a duration model as described by the survival function

$$\begin{aligned} S_i(t) &= \exp\left[-\Lambda_i(t_i)\right], \\ where\ \Lambda_i(t) &= \Lambda(t \mid X_i, v_i) = \int_0^t \lambda(t \mid X_i, v_i)\, dz \end{aligned}$$

To obtain a linear model interpretation we consider the random variable ε_i defined by

$$\varepsilon_i = -\ln \Lambda_i(t_i).$$

To calculate the distribution of ε_i observe that

$$P(\varepsilon < E) = P(-\ln \Lambda_i(t_i) < E) = P(\Lambda_i(t_i) > \exp\{-E\})$$

and, taking into accout that $\Lambda_i(t_i)$ is an increasing function,

$$= \Pr\left(t_i > \Lambda_i^{-1}[\exp\{-E\}]\right) = S_i\left(\Lambda_i^{-1}[\exp\{-E\}]\right)$$

because t_i has an associated survival function $S_i(\bullet)$, by construction. Furthermore

$$\begin{aligned} S_i\left(\Lambda_i^{-1}[\exp\{-E\}]\right) &= \exp\left(-\Lambda_i\left(\Lambda_i^{-1}[\exp\{-E\}]\right)\right) \\ &= \exp\left[-\exp\{-E\}\right]. \end{aligned}$$

[7] Recent papers that use this approach are Holt, Merwin, and Stern (1996) and McCall (1996).

Therefore,

$$P\left(e<E\right)=\exp\left[-\exp\left\{-E\right\}\right],$$

which is the cumulative distribution function for the type 1 extreme value distribution.

In particular, the proportional hazard model has a specification of the form

$$\begin{aligned}\lambda_i\left(t\right) &= \lambda_0\left(t\right)\lambda_1\left(X_i,\beta\right)\nu_i \\ \Lambda_i\left(t_i\right) &= \Lambda_0\left(t\right)\lambda_1\left(X_i,\beta\right)\nu_i \text{ with } \Lambda_0\left(t\right)=\int_0^t\lambda_0\left(z\right)dz\end{aligned}$$

and can be specified in regression form as

$$-\ln\Lambda_0\left(t_i\right)=\ln\left(\lambda_1\left(X_i,\beta\right)\right)+\ln v+\varepsilon_i \tag{16}$$

with ε_i type 1 extreme value distribution and $\ln v$ of unknown distribution. In particular, if

$$\lambda_1\left(X_i,\beta\right)=\exp\left(X_i\beta\right)$$

the expression (16) adopts the form

$$-\ln\Lambda_0\left(t_i\right)=X_i\beta+\ln v+\varepsilon_i,$$

which resembles a tipical linear regression with residual equal to $\left(\ln v+\varepsilon_i\right)$. A generalization of the previous approach is to consider

$$-\ln\Lambda_0\left(t_i\right)=\ln\left(\lambda_1\left(X_i,\beta\right)\right)+w$$

with w representing rhe residual component. This is a more general case because no restrictions are imposed on w. In particular, we can still consider $w = (\ln v + \varepsilon_i)$. This new model is called the Generalized Acelerated Failure-Time Model (GARF), it was developed by Ridder (1990).

1.4 Multistate models.

We can think of several possible multivariate generalization of a duration process. A possible generalization will be a process generated by a vector of random variables $\left(\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_k\right)$ in which the duration of the process is defined as the first order statistic of the $\left\{\tilde{t}_j\right\}_{j=1}^{k}$, this is called a competing risk process. In this case, we observe

$$t = \min_{j=1,...,k} \left\{\tilde{t}_j\right\}$$

and all other durations are censored.

Another possible generalization will be a process in which, for each individual i in the sample, we observe several durations that may be related with each other $\left\{\left(\tilde{t}_{i1}, \tilde{t}_{i2}, ..., \tilde{t}_{ik}, ...\right)\right\}_{i=1}^{N}$, this process is called a multiple spells process. This are just two possibilities, but one can imagine many others, for example, we can think of different combinations of the previous ones.

We can find many applications of the multistate process to economic or demographic problems. For example, Flinn and Heckman (1982) study the employment-unemployment process in the labor market, following a parametric approach, McCall

(1996) studies the problem of job search distinguishing between part time versus full time jobs, Thomas (1996) considers the problem of Sectorial Movements in the labor market, using semiparametric techniques inspired in Meyer's work. Heckman and Walker (1990) (from now on HW) consider the problem of timing and spacing of births. Heckman and Walker (1990) notice that "the study of unobservables in multi-state models is in its infancy".

1.4.1 Competing risks.

The competing risk that I develop in this section is not the only possible model that can be developed in this context. However, this model is general enough to include most of the models that have been used in empirical work. In particular this section is based in work by Heckman and Honore (1989) and Sueyoshy (1992).

We consider a competing risks model, where failure can occur from any K distinct competing causes. Let T_j denote the $j-$th latent failure. We only observe the first realized value of duration time and the cause of failure. The observed random variables are given by

$$(T,\, V) = \left\{ \min_j(T_j),\, \arg\min_j(T_j) \right\},$$

where T is duration to the first failure and V is the cause of failure for a particular individual. For simplicity, we consider a competing risks model with two distinct possibilities of failure ($j = 1, 2$), with possibly right censored observations. Our results can easily be generalized to the multivariate competing risks model.

In the bivariate competing risks framework, the process of interest consists of three states $\{0, 1, 2\}$, where 0 represents the initial state and $\{1, 2\}$ represent two possible exit states. Assume that, for individuals $i = 1, 2, ..., N$, we observe data of the form $\{x_i, t_i, d_{1i}, d_{2i}, d_{ci}\}_{i=1}^{N}$, where x_i represents a vector of predetermined variables, $t_i = \min(t_{1i}, t_{2i}, t_{ci})$, with t_{1i} and t_{2i} representing the time of transition from state 0 to state 1 and 2 respectively, t_{ci} represents time of censoring, and $d_{ji} = I\left(t_{ji} = t_i\right)$, $j = 1, 2, c$, is an indicator variable. Since we just observe one of the t_{ji}'s, the other ones should be interpreted as "shadow" variables. We assume that the censoring variable is independent of (T_1, T_2)[8]. We also assume that, in addition to x_i, there are other characteristics of individual i that affect the process, and are not observed by the researcher. Let $\nu_i = (\nu_{1i}\ \nu_{2i})'$ denote the vector of unobserved heterogeneity components that differ across risks. Further, let $G(\nu_i)$ be the distribution function of the unobservables, assumed to be independent of x_i.

We specify the joint survivor function of T_1 and T_2 conditional on x and ν as:

$$S\left(t_1, t_2 \mid x, \nu; \theta\right) = \exp\left\{-\Lambda_1\left(t_1, x_1; \theta_1\right)\nu_1\right\}\exp\left\{-\Lambda_2\left(t_2, x_2; \theta_2\right)\nu_2\right\}, \qquad (17)$$

[8]Since exit and censoring are formally equivalent, independent censoring simply creates an additional dummy destination state; see,for example, Lancaster (1990, Chapter 5). In the bivariate competing risks model, we have two real exit states $\{1, 2\}$ and a dummy destination state c.

where θ is a parameter vector,

$$\Lambda_j(t_j \mid x_j; \theta_j) = \int_0^{t_j} \lambda_j(z \mid x_j; \theta_j) dz, \quad j = 1, 2, \tag{18}$$

and the i subscripts on exit times, the covariates, and the unobservables have been suppressed for notational convenience. The corresponding cause-specific hazard function is usually specified as

$$\lambda_j(t_j \mid x_j; \theta_j)\nu_j = \lambda_{0j}(t_j)\phi_j(x_j, \beta_j)\nu_j, \tag{19}$$

where θ_j now consists of the parameter vector β_j associated with the observed covariates and the parameters in the baseline hazard, $\lambda_{0j}(t_j)$. This specification corresponds to the proportional competing risks model.

In order to better understand specification (17) observe that

$$\begin{aligned} S(t_1, t_2 \mid x, \nu; \theta) &= \exp\{-\Lambda_1(t_1, x_1; \theta_1)\nu_1\} \exp\{-\Lambda_2(t_2, x_2; \theta_2)\nu_2\} \\ &= S(t_1 \mid x, \nu; \theta) S(t_2 \mid x, \nu; \theta) \end{aligned}$$

with

$$S(t_j \mid x, \nu; \theta) = \exp\{-\Lambda_j(t_j, x_j; \theta_j)\nu_j\} \;\; for\ j = 1, 2.$$

Thus, conditional on (x, ν), t_1 and t_2 are independent. Therefore, in this specification dependency between the different risks comes from observable and unobservable characteristics of the agents.

The specification of the survivor function given in (17) is justified for several reasons. We can introduce dependence between T_1 and T_2 by assuming that ν_1 and ν_2 are not necessarily independent. The specification is similar to the one analyzed in Heckman and Honore (1989), which guarantees identifiability of the model. As discussed later, the above specification can easily be modified to take account of discretely recorded data. Thus, our approach encompasses parametric competing risks models of Han and Hausman (1990) and Sueyoshi (1992) that are useful for data sampled at discrete intervals.

Identification. At this point some discussion of results of the identifiability of the competing risks model with correlated times of failure is warranted. In the competing risks model with no regressors, it may not be possible to discriminate between a dependent risks model and independent risk models giving rise to the same cause-specific hazard functions on the basis of data on T and V alone. We are not going to present here the formal proof of this result, due to Tsiatis (1975). However, the intuition behind this result can be illustrated as follows:

Consider a certain agent at risk of death from two different diseases. Assume that the random variable $\tilde{t}_d$ measures the time of survival from disease $d = 1, 2$. Assume that the true net probability of surviving up to age t in conditions when either disease one or two are the sole possible causes of death is characterized by the survival process

$$S\left(t_1,t_2\right) \neq S\left(t_1\right) S\left(t_2\right),$$

with associated density $f\left(t_1,t_2\right)$. This indicates that the probability of surviving disease one is not independent from the probability of surviving disease two. This function characterizes the true distribution of potential survival times (t_1,t_2) of individuals exposed to the competing risks of death from diseases one and two. From the data we observe

$$(T,\ d) = \left\{\min_{j=1,2}(T_j),\ \arg\min_{j=1,2}(T_j)\right\},$$

where T is duration to the first failure and d is the cause of failure. Therefore, from the data we observe $\left(t_1 \mid \tilde{t}_2 > \tilde{t}_1\right)$ and $\left(t_2 \mid \tilde{t}_1 > \tilde{t}_2\right)$. With this information we can infer

$$\begin{aligned} Q_1\left(t_1\right) &= S\left(t_1 \mid \tilde{t}_2 > \tilde{t}_1\right) = \int_{t_1}^{+\infty}\int_{z_1}^{+\infty} f\left(z_1,z_2\right)\ dz_2 dz_1 \\ \text{and } Q_2\left(t_2\right) &= S\left(t_2 \mid \tilde{t}_1 > \tilde{t}_2\right). \end{aligned}$$

Observe that $Q_1\left(t_1\right)$ is independent of t_2 and $Q_2\left(t_2\right)$ is independent of t_1. Therefore, the competing risk process characterized by $S\left(t_1,t_2\right)$ is observationally equivalent to the process characterized by

$$S^*\left(t_1,t_2\right) = Q_1\left(t_1\right) Q_2\left(t_2\right)$$

However, in this process the risk of death from desease one and two are independent. As a result, we are unable to distinguishing between $S\left(t_1,t_2\right)$ and $S^*\left(t_1,t_2\right)$ from the data, even if both are fundamentally different. In other words, the true duration process characterized by $S(t_1, t_2)$ is equivalent to a duration process characterized by $S^*(t_1, t_2)$, with independent risks.

As argued by Heckman and Honore (1989), in competing risks models without regressors it is necessary to make functional form assumptions about the joint distribution of failure times in order to identify the distribution. They show how the introduction of covariates helps to identify a large class of dependent competing risks models without invoking distributional assumptions. This remarkable result, which overturns the Cox-Tsiatis nonidentification theorem, is also relevant in establishing the identifiability of model (17).

Since ν_1 and ν_2 in (17) are unobserved, we need to integrate out the effects of the unobserved heterogeneity components. To do this, let θ now denote a vector of parameters associated with x, t, and distribution G. The ensuing mixture survivor function can be expressed as:

$$\begin{aligned} S(t_1, t_2 \mid x; \theta) &= \int \exp\{-\Lambda_1(t_1, x_1; \theta_1)\nu_1\} \exp\{-\Lambda_2(t_2, x_2, ; \theta_2)\nu_2\} \mathrm{d}G(\nu) \\ &= M\left[-\Lambda_1(t_1 \mid x_1; \theta_1),\ -\Lambda_2(t_2 \mid x_2; \theta_2)\right], \end{aligned} \tag{20}$$

where the integration is over the support of ν, assumed to be the positive real line, and $M(.)$ is the moment generating function (MGF) of ν evaluated at $(-\Lambda_1,\ -\Lambda_2)$. Alternatively, the mixture survivor function can also be written as

$$S(t_1, t_2 \mid x; \theta) = \varphi(\Lambda_1,\ \Lambda_2),$$

where $\varphi(\Lambda_1, \Lambda_2)$ is the Laplace transform of the bivariate mixing distribution $G(\nu)$.

The likelihood function can now be developed by assuming that M (or G) is known. Let $f(t_{1i}, t_{2i} \mid x_i; \theta)$ be the joint density function corresponding to the survivor function given in (20). Depending on which of the two real destination states we observe, let

$$q_{1i}(t_i; \theta) = f(t_{1i} = t_i, t_{2i} \geq t_i \mid x_i; \theta)$$

and

$$q_{2i}(t_i;\, \theta) = f(t_{1i} \geq t_i, t_{2i} = t_i \mid x_i; \theta)\,.$$

For instance,

$$f(t_1 = t_i, t_2 \geq t_i, \mid x_i; \theta) = \int_{t_i}^{+\infty} f(t_1 = t_i, t_2 \mid x_i; \theta)\, \mathrm{d}t_2.$$

Then, using Leibniz's rule to obtain the derivative of an integral[9], we get

$$\begin{aligned} q_{ji}(t_i; \theta) &= \left.\frac{-\partial S(t_{1i}, t_{2i} \mid x_i; \theta)}{\partial t_j}\right|_{t_{1i}=t_{2i}=t_i} \\ &= \lambda_j(t_i \mid x_{ji}; \theta_j) \times M_j^{(1)}\left[-\Lambda_1(t_i \mid x_{1i}; \theta_1),\, -\Lambda_2(t_i \mid x_{2i}; \theta_2)\right], \end{aligned} \tag{21}$$

for $j = 1, 2$, where $M_j^{(1)}[.]$ is the first order derivative of $M(.)$ with respect to the

[9] A similar argument can be provided by observing that,

$$f(t_1 = t_i, t_2 \geq t_i, \mid x_i; \theta) = \lim_{\Delta t \to 0} -\frac{S(t_i + \Delta t, t_i \mid x_i; \theta) - S(t_i, t_i \mid x_i; \theta)}{\Delta t}$$

$j-$th argument evaluated at t_i:

$$M_j^{(1)}[.] = \left.\frac{\partial M\left[-\Lambda_1(t_{1i} \mid x_{1i};\theta_1),\, -\Lambda_2(t_{2i} \mid x_{2i};\theta_2)\right]}{\partial \Lambda_j}\right|_{t_{1i}=t_{2i}=t_i} \tag{22}$$

Finally, since both t_{1i} and t_{2i} are censored for $j = c$, define

$$\begin{aligned} q_{ci}(t_i;\theta) &= S(t_i, t_i \mid x_i;\theta) \\ &= M\left[-\Lambda_1(t_i \mid x_{1i};\theta_1),\, -\Lambda_2(t_i \mid x_{2i};\theta_2)\right]. \end{aligned} \tag{23}$$

The single observation log-likelihood function can now be expressed as:

$$\begin{aligned} l_i(\theta) &= d_{1i}\log q_{1i}(t_i;\theta) + d_{2i}\log q_{2i}(t_i;\theta) + d_{ci}\log q_{ci}(t_i;\theta) \\ &= \textstyle\sum_{j=1}^{2} d_{ji}\left\{\log\lambda_j(t_i \mid x_{ji};\theta_j) + \log M_j^{(1)}\left[-\Lambda_1(t_i \mid x_{1i};\theta_1),\, -\Lambda_2(t_i \mid x_{2i};\theta_2)\right]\right\} \\ &\qquad + d_{ci}\log M\left[-\Lambda_1(t_i \mid x_{1i};\theta_1),\, -\Lambda_2(t_i \mid x_{2i};\theta_2)\right]. \end{aligned} \tag{24}$$

and the log-likelihood is

$$\mathcal{L}_0(\theta) = \sum_{i=1}^{N} l_i(\theta). \tag{25}$$

The estimator of θ can then be obtained by maximizing this log-likelihood function. This is an infeasible estimator because G and M are yet to be estimated. Therefore, G or M should be part of the MLE procedure.

1.4.2 Multiple Spells models.

We can find many applications of the multistate process to economic or demographic problems. For example, Flinn and Heckman (1982) study the employment-unemployment process in the labor market, following a parametric approach, Thomas (1996) considers the problem of Sectorial Movements in the labor market, using semiparametric techniques inspired in Meyer's work. Heckman and Walker (1990) (from now on HW) consider the problem of timing and spacing of births. All the papers previously cited have noticed the importance of unobserved heterogeneity in multistate processes.

Since accounting for unobserved heterogeneity has proved to be crucial in a duration process this concern can be extend to any generalization of a duration process. In addition, any multivariate generalization will most probably incorporate additional features, like correlation among different failure times, that will make this even a more important issue. Heckman and Walker (1990) notice the importance of accounting for unobserved heterogeneity, and in particular to account for the possibility of unobservables correlated over spells. They control for unobserved heterogeneity using Heckman-Singer's type techniques. In this section I will use the framework and notation of Heckman and Walker (1990) as close as possible. They notice that "the study of unobservables in multi-state models is in its infancy". The lack of an extensive literature in this subject will be reflected in this notes.

Assume that for each individual i in the sample we can observe up to $C(< \infty)$ events that occur at times $\tilde{t}_1, \tilde{t}_1 + \tilde{t}_2, \tilde{t}_1 + \tilde{t}_2 + \tilde{t}_3, ..., \sum_{j=1}^{C} \tilde{t}_j$, with $\tilde{t}_j$ a random variable representing the time interval between occurrence of events $(j-1)$ and j. Let $H_i(\tau)$

represent the relevant conditioning set at time τ,[10] t_{ij} represent the realization of $\tilde{t}_j$, and $\tau_i(j) = \sum_{l=1}^{j} t_{il}$ represent the time of occurrence of event j, for agent i. In order to simplify the notation, define $H_{ij} = H_i(\tau_i(j-1) + t_{ij})$, and $H_{iju} = H_i(\tau_i(j-1) + u)$. Also, I will use the notation H_i when the conditioning set is clear from the context.

Denote

$$g_j(t_{ij} \mid H_i(\tau_i(j-1) + t_{ij}), \nu_{ij}) \text{ and}$$

$$S_j(t_{ij} \mid H_i(\tau_i(j-1) + t_{ij}), \nu_{ij})$$

as the density and survival functions associated to event j, at time t_{ij}, conditional on $H_i(\tau_i(j-1) + t_{ij})$ and an unobserved heterogeneity term ν_{ij}. Let $\tilde{t}_i^c$ represent the time of censoring of observation i.

The contribution to the likelihood function of an individual with an observed history $\tilde{t}_{i1} = t_{i1}, \tilde{t}_{i2} =, ..., \tilde{t}_{ik_i} = t_{ik_i}$, and with spell censored at $\tilde{t}_i^c = t_{ik_i+1}^c$, is as follows,

$$\left\{ \prod_{j=1}^{k_i} g_j(t_{ij} \mid H_i(\tau_i(j-1) + t_{ij}), \nu_{ij}) \right\} \cdot S_{k_i+1}\left(t_{ik_i+1}^c \mid H_{ik_i+1}, \nu_{ik_i+1}\right) \tag{1}$$

Clearly, the previous expression cannot be used for estimation purposes, since the terms $\{\nu_{ij}\}$ are unknown. HW overcome this problem by using Heckman-Singer's

[10] That is $H_i(\tau)$ represents all information existing at τ relevant for the future behavior of the multispell process. For example, it may include specific characteristics of agent i in the sample.

type techniques.

The conditional hazard for individual i at duration $\tilde{t}_{ij}$ is defined to be

$$h_j\left(t_{ij} \mid H_i\left(\tau_i\left(j-1\right)+t_{ij}\right), \nu_{ij}\right) \tag{2}$$

HW consider the following specification for the hazard function:

$$h_j\left(t_{ij} \mid H_{ij}, \nu_{ij}\right) = h_j\left(t_{ij} \mid H_{ij}\right) \nu_{ij} \tag{3}$$

with

$$h_j\left(t_{ij} \mid H_{ij}\right) = \exp\left\{\gamma_{0j} + \sum_{m=1}^{m_j} \gamma_{mj}\left(\frac{t_{ij}^{\lambda_{mj}}-1}{\lambda_{mj}}\right) + Z_i \beta_j\right\} \tag{4}$$

More precisely, HW consider an additive unobserved heterogeneity term inside the exponential function, the multiplicative representation is clearly equivalent. In fact, the model of Heckman and Walker (1990) is a slight variation of a proportional hazard model, but I will not get into what these differences are. They define

$$S_j\left(t_{ij} \mid H_i, \nu_{ij}\right) = \tag{5}$$

$$p(j-1) + (1-p(j-1)) \exp\left[-\int_0^{t_{ij}} h_j\left(u \mid H_{iju}, \nu_{ij}\right) du\right]$$

where $p(j-1)$ is the probability that a woman with $j-1$ children is never at risk to have a jth birth[11]. Clearly, this specification is chosen in order to better fit the data. It is interesting to study how the work of BCR can be adapted to this scenario.

Lets start by defining some simplifying notation,

$$h_{ij} = h_j \left(t_{ij} \mid H_{ij} \right) \tag{6}$$

$$\Lambda_{ij} = \int_0^{t_{ij}} h_j \left(u \mid H_{iju} \right) du$$

Then,

$$\Lambda_j \left(t_{ij} \mid H_i, \nu_{ij} \right) = \int_0^{t_{ij}} h_j \left(u \mid H_{iju}, \nu_{ij} \right) du = \Lambda_{ij} \nu_{ij} \tag{7}$$

Note that in the framework defined by HW, h_{ij} is not exactly the hazard function as commonly defined. Formally, h_{ij} is the hazard function of event j conditional on the event $j-1$ having occurred. We can consider different formulations for ν_{ij}. In particular, we can consider

$$\nu_{ij} \text{ iid } F_{\omega_j} \left(\bullet \right) \ \forall \ i \in \{1, ...N\}, \ j \in \{1, ..., c\}. \tag{8}$$

Another possibility is

$$\nu_{ij} = \nu_i, \text{ with } \nu_i \text{ iid } F_\nu \left(\bullet \right) \ \forall i \in \{1, ...N\}. \tag{9}$$

[11] This is simmilar to the way in which the problem of excess of zeros is treated in the literature of count data (also called Hurdle regression models).

Also, it may be convinient in certain cases to consider

$$\nu_{ij} = \nu_i + \omega_{ij} \text{ with } \nu_i \text{ iid } F_{\nu}(\bullet) \ \ \forall i \in \{1, ...N\}, \ \omega_{ij} \text{ iid } F_{\omega_j}(\bullet) \ \ \forall j \in \{1, ..., c\} \quad (10)$$

In the first case, we model unobserved heterogeneity as a random state effect. That is, unobserved heterogeneity is state specific not individual specific. In the second case, we model unobserved heterogeneity as an individual random effect. The third case is a mixture of the previous two cases, and is the one that resembles more closely to the model specified by HW.

CASE 1: ν_{ij} iid $F_{\omega_j}(\bullet)$ $\forall$ $i \in \{1, ...N\}$, $j \in \{1, ..., c\}$

In this scenario we have,

$$S_j(t_{ij} \mid H_i, \nu_{ij}) = \exp[-\Lambda_{ij}\nu_{ij}] \quad (12)$$

and

$$g_j(t_{ij} \mid H_i, \nu_{ij}) = -\frac{\partial S_j(t_{ij} \mid H_i, \nu_{ij})}{\partial t} \quad (13)$$

$$= h_j(t_{ij} \mid H_{ij}) \nu_{ij} \exp[-\Lambda_{ij}\nu_{ij}]$$

The contribution to the likelihood function of an individual with history $\tilde{t}_{i1} = t_{i1}, \tilde{t}_{i2} =, ..., \tilde{t}_{ik_i} = t_{ik_i}$, and with spell censored at $t^c_{ik_i+1}$, is as follows,

$$l_i(\nu_i) = \prod_{j=1}^{k_i} g_j(t_{ij} \mid H_i, \nu_{ij}) \cdot S_{k_i+1}\left(t^c_{ik_i+1} \mid H_i, \nu_{ij}\right) \tag{14}$$

$$= \left\{ \prod_{j=1}^{k_i} h_j(t_{ij} \mid H_{ij})\, \nu_{ij} \exp\left[-\Lambda_{ij}\nu_{ij}\right] \right\} \exp\left[-\Lambda^c_{ik_i+1}\nu_{ij}\right]$$

then, integrating out the unobserved heterogeneity from $l_i(\nu_i)$ we obtain:

$$l_i = \int ..^{(k_i+1)}.. \int l_i(\nu)\, dF_{\omega_1}...dF_{\omega_{k_i+1}} = \tag{15}$$

$$= \left\{ \prod_{j=1}^{k_i} .h_j(t_{ij} \mid H_{ij})\, M^{(1)}_{\omega_j}(-\Lambda_{ij}) \right\} \cdot M_{\omega_{k_i+1}}\left(-\Lambda^c_{ik_i+1}\right)$$

However, unless the distribution of unobserved heterogeneity is known the values of $M_{\omega_j}(\bullet)$ and $M^{(1)}_{\omega_j}(\bullet)$ cannot be computed. As with the most simple duration models there are alternative ways to deal with this problem. One way is to assume parametric assumptions about the distribution of unobserved heterogeneity. Other possibility is to apply non-parametric techniques simmilar to these used in the single spell framework (Heckman and Walker, 1990).