CHAPTER 6: DURATION MODELING

Chandra R. Bhat
The University of Texas at Austin

1. INTRODUCTION

Hazard-based duration models represent a class of analytical methods which are appropriate for modeling data that have as their focus an end-of-duration occurrence, given that the duration has lasted to some specified time (Kiefer, 1988; Hensher and Mannering, 1994). This concept of conditional probability of termination of duration recognizes the dynamics of duration; i.e., it recognizes that the likelihood of ending the duration depends on the length of elapsed time since start of the duration.

Hazard-based models have been used extensively for several decades in biometrics and industrial engineering to examine issues such as life-expectancy after the onset of chronic diseases and the number of hours of failure of motorettes under various temperatures. Because of this initial association with time till failure (either of the human body functioning or of industrial components), hazard models have also been labeled as “failure-time models”. However, the label “duration models” more appropriately reflects the scope of application to any duration phenomenon.

Two important features characterize duration data. The first important feature is that the data may be censored in one form or the other. For example, consider survey data collected to examine the time duration to adopt telecommuting from when the option becomes available to an employee (Figure 1). Let data collection begin at calendar time A and end at calendar time C. Consider individual 1 in the figure for whom telecommuting is an available option prior to the start of data collection and who begins telecommuting at calendar time B. Then, the recorded duration to adoption for the individual is AB, while the actual duration is larger because of the availability of the telecommuting option prior to calendar time A. This type of censoring from the left is labeled as *left censoring*. On the other hand, consider individual 2 for whom telecommuting becomes an available option at time B and who adopts telecommuting after the termination of data collection. The recorded duration is BC, while the actual duration is longer. This type of censoring is labeled as *right censoring*. Of course, the duration for an individual can be both left- and right-censored, as is the case for individual 3 in Figure 1. The duration of individual 4 is uncensored.

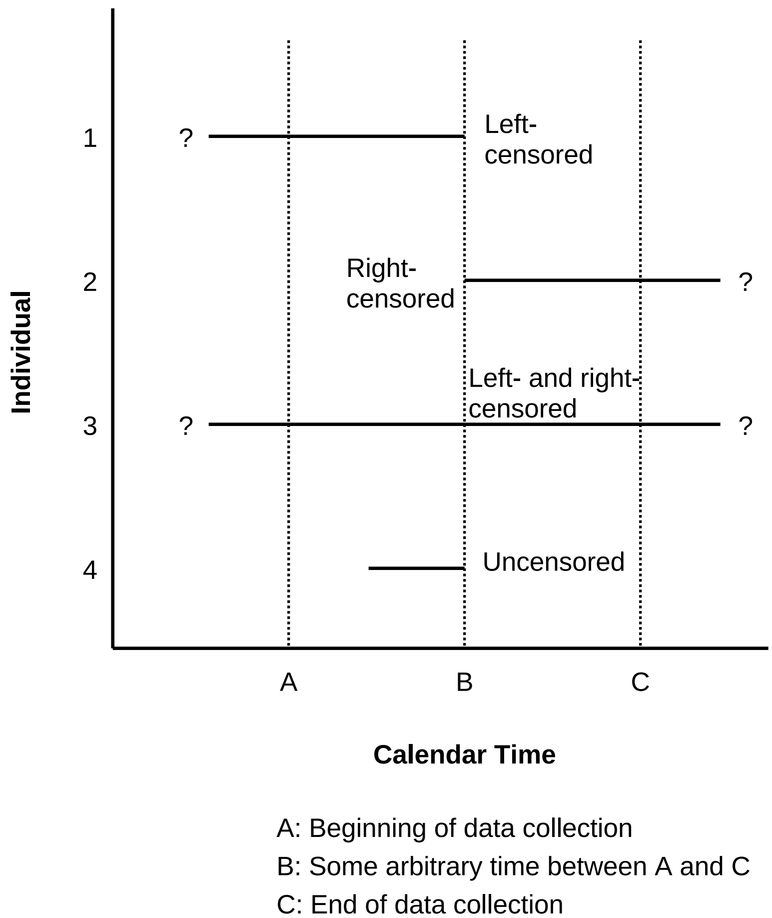

Figure 1. Censoring of Duration Data (modified slightly from Kiefer, 1998)

The second important characteristic of duration data is that exogenous determinants of the event times characterizing the data may change during the event spell. In the context of the telecommuting example, the location of a person's household (relative to his or her work location) may be an important determinant of telecommuting adoption. If the person changes home locations during the survey period, we have a time-varying exogenous variable.

The hazard-based approach to duration modeling can accommodate both of the distinguishing features of duration data; i.e., censoring and time-varying variables; in a relatively simple and flexible manner. On the other hand, accommodating censoring within the framework of traditional regression methods is quite cumbersome, and incorporating time-varying exogenous variables in a regression model is anything but straightforward.

In addition to the methodological issues discussed above, there are also intuitive and conceptual reasons for using hazard models to analyze duration data. Consider again that we are interested in examining the distribution across individuals of telecommuting adoption duration (measured as the number of weeks from when the option becomes available). Let our interest be in determining the probability that an individual will adopt telecommuting in 5 weeks. The traditional regression approach is to specify a probability distribution for the duration time and fit it using data. The hazard approach, however, determines the probability of the outcome as a sequence of simpler conditional events. Thus, a theoretical model we might specify is that the individual re-evaluates the telecommuting option every week and has a probability λ of deciding to adopt telecommuting each week. Then the probability of the individual adopting telecommuting in exactly 5 weeks is simply $(1-\lambda)^4 \times \lambda$. (Note that λ is essentially the hazard rate for termination of the non-adoption period). Of course, the assumption of a constant λ is rather restrictive; the probability of adoption might increase (say, because of a "snowballing" effect as information on the option and its advantages diffuses among people) or decrease (say, due to "inertial" effects) as the number of weeks increases. Thus, the "snowballing" or "inertial" dynamics of the duration process suggest that we specify our model in terms of conditional sequential probabilities rather than in terms of an unconditional direct probability distribution. More generally, the hazard-based approach is a convenient way to interpret duration data the generation of which is fundamentally and intuitively associated with a dynamic sequence of conditional probabilities.

As indicated by Kiefer (1988), for any specification in terms of a hazard function, there is an exact mathematical equivalent in terms of an unconditional probability distribution. The question that may arise is then why not specify a probability distribution, estimate the parameters of this distribution, and then obtain the estimates of the implied conditional probabilities (or hazard rates)? While this can be done, it is preferable to focus directly on the implied conditional probabilities (*i.e.*, the hazard rates) because the duration process may dictate a particular behavior regarding the hazard which can be imposed by employing an appropriate distribution for the hazard. On the other hand, directly specifying a particular probability distribution for durations in a regression model may not immediately translate into a simple or interpretable implied hazard distribution. For example, the normal and log-normal distributions used in regression methods imply complex, difficult to interpret, hazards that do not even subsume the simple constant hazard rate as a special case. To summarize, using a hazard-based approach to modeling duration processes has both methodological and conceptual advantages over the more traditional regression methods.

In this chapter the methodological issues related to specifying and estimating duration models are reviewed. Sections 2–4 focus on three important structural issues in a hazard model for a simple unidimensional duration process: (i) specifying the hazard function and its distribution (Section 2); (ii) accommodating the effect of external covariates (Section 3); and (iii) incorporating the effect of unobserved heterogeneity (Section 4). Sections 5–7 deal with the estimation procedure for duration models, miscellaneous advanced topics related to duration processes, and recent transport applications of duration models, respectively.

2. THE HAZARD FUNCTION AND ITS DISTRIBUTION

Let T be a non-negative random variable representing the duration time of an individual (for simplicity, the index for the individual is not used in this presentation). T may be continuous or discrete. However, discrete T can be accommodated by considering the discretization as a result of grouping of continuous time into several discrete intervals (see later). Therefore, the focus here is on continuous T only.

The hazard at time u on the continuous time-scale, $\lambda(u)$, is defined as the instantaneous probability that the duration under study will end in an infinitesimally small time period h after time u, given that the duration has not elapsed until time u (this is a continuous-time equivalent of the discrete conditional probabilities discussed in the example given above of telecommuting adoption). A precise mathematical definition for the hazard in terms of probabilities is

$$\lambda(u) = \lim_{h \to 0^+} \frac{\Pr(u \le T < u + h \mid T > u)}{h}. \tag{1}$$

This mathematical definition immediately makes it possible to relate the hazard to the density function $f(.)$ and cumulative distribution function $F(.)$ for T. Specifically, since the probability of the duration terminating in an infinitesimally small time period h after time u is simply $f(u)*h$, and the probability that the duration does not elapse before time u is $1\text{-}F(u)$, the hazard rate can be written as

$$\lambda(u) = \frac{f(u)}{[1-F(u)]} = \frac{f(u)}{S(u)} = \frac{dF/du}{S(u)} = \frac{-dS/du}{S(u)} = \frac{-d\ln S(u)}{du}, \tag{2}$$

where $S(u)$ is a convenient notational device which we will refer to as the endurance probability and which represents the probability that the duration did not end prior to u (*i.e.*, that the duration "endured" until time u). The duration literature has referred to $S(u)$ as the "survivor probability", because of the initial close association of duration models to failure time in biometrics and industrial engineering. However, the author prefers the term "endurance probability" which reflects the more universal applicability of duration models.

The shape of the hazard function has important implications for duration dynamics. One may adopt a parametric shape or a non-parametric shape. These two possibilities are discussed below.

2.1. Parametric Hazard

In the telecommuting adoption example discussed earlier, a constant hazard was assumed. The continuous-time equivalent for this is $\lambda(u) = \sigma$ for all u, where σ is the constant hazard rate. This is the simplest distributional assumption for the hazard and implies that there is no duration dependence or duration dynamics; the conditional exit probability from the duration is not related to the time elapsed since start of the duration. The constant-hazard assumption corresponds to an exponential distribution for the duration distribution.

The constant-hazard assumption may be very restrictive since it does not allow "snowballing" or "inertial" effects. A generalization of the constant-hazard assumption is a two-parameter hazard function, which results in a Weibull distribution for the duration data. The hazard rate in this case allows for monotonically increasing or decreasing duration dependence and is given by $\lambda(u) = \sigma\,\alpha\,(\sigma\,u)^{\alpha-1}$, $\sigma > 0$, $\alpha > 0$. The form of the duration dependence is based on the parameter α. If $\alpha > 1$, then there is positive duration dependence (implying a "snowballing" effect, where the longer the time has elapsed since start of the duration, the more likely it is to exit the duration soon). If $\alpha < 1$, there is negative duration dependence (implying an "inertial" effect, where the longer the time has elapsed since start of the duration, the less likely it is to exit the duration soon). If $\alpha = 0$, there is no duration dependence (which is the exponential case).

The Weibull distribution allows only monotonically increasing or decreasing hazard duration dependence. A distribution that permits a non-monotonic hazard form is the log-logistic distribution. The hazard function in this case is given by

$$\lambda(u) = \frac{\sigma\,\alpha\,(\sigma\,u)^{\alpha-1}}{1+(\sigma\,u)^{\alpha}}. \tag{3}$$

If $\alpha < 1$, the hazard is monotonic decreasing from infinity; if $\alpha = 1$, the hazard is monotonic decreasing from σ; if $\alpha > 1$, the hazard takes a non-monotonic shape increasing from zero to a maximum of $u = [(\alpha-1)^{1/\alpha}]/\sigma$, and decreasing thereafter.

The reader is referred to Hensher and Mannering (1994) for diagrammatic representations of the hazard functions corresponding to the exponential, Weibull, and log-logistic duration distributions. Several other parametric distributions may also be adopted for the duration distribution, including the Gompertz, log-normal, gamma, generalized gamma, and generalized *F* distributions. Alternatively, one can adopt a general non-negative function for the hazard, such as a Box-Cox formulation, which nests the commonly used parametric hazard functions. The Box-Cox formulation takes the following form

$$\lambda(u) = \exp\left[\alpha_0 + \sum_{k=1}^{K} \alpha_k \left(\frac{u^{\gamma_k} - 1}{\gamma_k}\right)\right], \tag{4}$$

where α_0, α_k, and γ_k ($k = 1, 2, \ldots, K$) are parameters to be estimated. If $\alpha_k = 0\ \forall\ k$, then we have the constant-hazard function (corresponding to the exponential distribution). If $\alpha_k = 0$ for (k = 2, 3, …, K), $\alpha_1 \neq 0$, and $\gamma_1 \to 0$, we have the hazard corresponding to a Weibull duration distribution if we reparameterize as follows: $\alpha_1 = (\alpha - 1)$ and $\alpha_0 = \ln(\alpha\sigma^{\alpha})$.

2.2. Non-Parametric Hazard

The distributions for the hazard discussed above are fully parametric. In some cases, a particular parametric distributional form may be appropriate on theoretical grounds. However, a problem with the parametric approach is that it inconsistently estimates the baseline hazard when the assumed parametric form is incorrect (Meyer, 1990). Also, there may be little theoretical support for a parametric shape in several instances. In such cases, one might consider using a nonparametric baseline hazard. The advantage of using a nonparametric form is that, even when a particular parametric form is appropriate, the resulting estimates are consistent and the loss of efficiency (resulting from disregarding information about the distribution of the hazard) may not be substantial (Meyer, 1987).

A nonparametric approach to estimating the hazard distribution was originally proposed by Prentice and Gloeckler (1978), and later extended by Meyer (1987) and Han and Hausman (1990). (Another approach, which does not require parametric hazard-distribution restrictions, is the partial likelihood framework suggested by Cox (1972); however, the Cox approach only estimates the covariate effects and does not estimate the hazard distribution itself).

In the Han and Hausman nonparametric approach, the duration scale is split into several smaller discrete periods (these discrete periods may be as small as needed, though each discrete period should have two or more duration completions). Note that this discretization of the time-scale is not inconsistent with an underlying continuous process for the duration data. The discretization may be viewed as a result of small measurement error in observing continuous data or a result of rounding off in the reporting of duration times (e.g., rounding to the nearest 5 minutes in reporting activity duration or travel-time duration). Assuming a constant hazard (i.e., an exponential duration distribution) within each discrete period, one can then estimate the continuous-time step-function hazard shape. Under the special situation where the hazard model does not include any exogenous variables, the above nonparametric baseline is equivalent to the sample hazard (also, referred to as the Kaplan-Meier hazard estimate).

The parametric baseline shapes can be empirically tested against the nonparametric shape in the following manner:

(1) Assume a parametric shape and estimate a corresponding "nonparametric" model with the discrete period hazards being constrained to be equal to the value implied by the parametric shape at the mid-points of the discrete intervals.

(2) Compare the fit of the parametric and nonparametric models using a log (likelihood) ratio test with the number of restrictions imposed on the nonparametric model being the number of discrete periods minus the number of parameters characterizing the parametric distribution shape.

It is important to note that, in making this test, the continuous parametric hazard distribution is being replaced by a step-function hazard in which the hazard is specified to be constant within discrete periods but maintains the overall parametric shape across discrete periods.

3. EFFECT OF EXTERNAL CO-VARIATES

In the previous section, the hazard function and its distribution were discussed. In this section, a second structural issue associated with hazard models is considered, i.e., the incorporation of the effect of exogenous variables (or external covariates). Two parametric forms are usually employed to accommodate the effect of external covariates on the hazard at any time u: the proportional hazards form and the accelerated lifetime form. These two forms are discussed in the subsequent two sections. Section 3.3 briefly discusses more general forms for incorporating the effect of external covariates. In the ensuing discussion, time-invariant covariates are assumed.

3.1. The Proportional Hazard Form

The proportional hazard (PH) form specifies the effect of external covariates to be multiplicative on an underlying hazard function:

$$\lambda(u,x,\beta,\lambda_0) = \lambda_0\,\phi(x,\beta)\,, \tag{5}$$

where λ_0 is a baseline hazard, x is a vector of explanatory variables, and β is a corresponding vector of coefficients to be estimated. In the PH model, the effect of external covariates is to shift the entire hazard function profile up or down; the hazard function profile itself remains the same for every individual.

The typical specification used for $\phi(x,\beta)$ in equation (5) is $\phi(x,\beta) = e^{-\beta' x}$. This specification is convenient since it guarantees the positivity of the hazard function without placing constraints on the signs of the elements of the β vector. The PH model with $\phi(x,\beta) = e^{-\beta' x}$ allows a convenient interpretation as a linear model. To explicate this, consider the following equation:

$$\ln \Lambda_0(u) = \ln \int_0^u \lambda_0(w)dw = u^* = \beta' x + \varepsilon\,, \tag{6}$$

where $\Lambda_0(u)$ is the integrated baseline hazard. From the above equation, we can write

$$\begin{aligned}\text{Prob}(\varepsilon < z) &= \text{Prob}[\ln \Lambda_0(u) - \beta' x < z] \\ &= \text{Prob}\{u < \Lambda_0^{-1}[\exp(\beta' x + z)]\} \\ &= 1 - \text{Prob}\{u > \Lambda_0^{-1}[\exp(\beta' x + z)]\}.\end{aligned} \tag{7}$$

Also, from equation (2), the endurance function may be written as $S(u) = \exp[-\Lambda(u)]$. The above probability is then

$$\begin{aligned}\text{Prob}(\varepsilon < z) &= 1 - \exp\left(-\Lambda_0\{\Lambda_0^{-1}[\exp(\beta' x + z)]\}\exp(-\beta' x)\right) \\ &= 1 - \exp[-\exp(z)].\end{aligned} \tag{8}$$

Thus, the PH model with $\phi(x, \beta) = \exp(-\beta' x)$ is a linear model, $\ln \Lambda_0(u) = \beta' x + \varepsilon$, with the logarithm of the integrated hazard being the dependent variable and the random term ε taking an extreme value form, with distribution function given by

$$\text{Prob}(\varepsilon < z) = G(z) = 1 - \exp[-\exp(z)]. \tag{9}$$

Of course, the linear model interpretation does not imply that the PH model can be estimated using a linear regression approach because the dependent variable, in general, is unobserved and involves parameters which themselves have to be estimated. But the interpretation is particularly useful when a nonparametric hazard distribution is used (see Section 5.2). Also, in the special case when the Weibull distribution or the exponential distribution is used for the duration process, the dependent variable becomes the logarithm of duration time. In the exponential case, the integrated hazard is λu and the corresponding log-linear model for duration time is $\ln u = \delta + \beta' x + \varepsilon$, where $\delta = -\ln(\lambda)$. For the Weibull case, the integrated hazard is $(\lambda u)^{\alpha}$, so the corresponding log-linear model for duration time is $\ln u = \delta + \beta^* x + \varepsilon^*$, where $\delta = -\ln \lambda$, $\beta^* = \beta / \alpha$, and $\varepsilon^* = \varepsilon / \alpha$. In these two cases, the PH model may be estimated using a least-squares regression approach if there is no censoring of data. Of course, the error term in these regressions is non-normal, so test statistics are appropriate only asymptotically and a correction will have to be made to the intercept term to accommodate the non-zero mean nature of the extreme value error form.

The coefficients of the covariates can be interpreted in a rather straightforward fashion in the PH model of equation (5) when the specification $\phi(x,\beta) = e^{-\beta' x}$ is used. If β_j is positive, it implies that an increase in the corresponding covariate decreases the hazard rate (*i.e.*, increases the duration). With regard to the magnitude of the covariate effects, when the *j*th covariate increases by one unit, the hazard changes by $\{\exp(-\beta_j) - 1\} \times 100\%$.

3.2. The Accelerated Lifetime Form

The second parametric form for accommodating the effect of covariates, the accelerated lifetime form, assumes that the covariates rescale time directly. That is, the probability that the duration will endure beyond time *u* is given by the baseline endurance probability computed at a rescaled (by external covariates) time value:

$$S(u,x,\beta) = S_0\,[u\,\phi(x,\beta)]. \tag{10}$$

In this model, the effect of the covariates is to alter the rate at which an individual proceeds along the time axis. Thus, the role of the covariates is to accelerate (or decelerate) the termination of the duration period.

The typical specification used for $\phi(x,\beta)$ in equation (10) is $\phi(x,\beta) = \exp(-\beta' x)$. With this specification, the accelerated lifetime formulation can be viewed as a log-linear regression of duration on the external covariates. To see this, let $\ln(u) = \beta' x + \varepsilon$. Then, we can write

$$\begin{aligned} \Pr(\varepsilon < z) &= \Pr[\ln u - \beta' x < z] \\ &= \Pr\{u < \exp(\beta' x + z)]\} \\ &= 1 - \Pr\{u > \exp(\beta' x + z)\}. \end{aligned} \tag{11}$$

Next, from the survivor function specification in the accelerated lifetime model, we can write the above probability as

$$\begin{aligned} \Pr(\varepsilon < z) &= 1 - S_0[\{\exp(\beta' x + z)\} \cdot \exp(-\beta' x)] \\ &= 1 - S_0[\exp(z)] \\ &= F_0[\exp(z)]. \end{aligned} \tag{12}$$

Thus, the accelerated lifetime model with $\phi(x,\beta) = \exp(-\beta' x)$ is a log-linear model, $\ln(t) = \beta' x + \varepsilon$, with the density for the error term being $f_0[\exp(\varepsilon)]\exp(\varepsilon)$, where the specification of f_0 depends on the assumed distribution for the survivor function S_0. In the absence of censoring, therefore, the accelerated lifetime specification can be estimated directly using the least-squares technique (Note also that the PH model with exponential or Weibull distributions can be interpreted as accelerated lifetime models, since they can be written in a log-linear form).

The linear model representation of the accelerated lifetime model provides a convenient interpretation of the coefficients of the covariates; a one unit increase in the *j*th explanatory variable results in an increase in the duration time by β_j percent.

The PH and the accelerated lifetime models have seen widespread use. Of the two, the PH model is more commonly used. The PH formulation is also more easily extended to accommodate nonparametric baseline methods and can incorporate unobserved heterogeneity (unobserved heterogeneity cannot be incorporated in an accelerated lifetime model due to identification problems).

3.3. General Form

The PH and the accelerated lifetime models are rather restrictive in specifying the effect of covariates over time. The PH model assumes that the effect of covariates is to change the baseline hazard by a constant factor that is independent of duration. The accelerated lifetime model allows time-varying effects, but specifies the time-varying effects to be monotonic and smooth in the time domain. In some situations, the use of more general time-varying covariate effects may be preferable. For example, in a model of departure time from home for recreational trips, the effect of children on the termination of home-stay duration may be much more "accelerated" during the evening period than in earlier periods of the day, because the evening period is most convenient (from schedule considerations) for joint-activity participation with children. This sudden non-monotonic acceleration during a specific period of the day cannot be captured by the PH or the accelerated lifetime model.

An alternative is to accommodate more flexible interaction terms of the covariates and time: $\lambda(u) = \lambda_0 \exp[g(u,x,\beta)]$, where the function g can be as general as desired. An important issue, however, in specifying general forms is that interpretation can become difficult; the analyst would do well to retain a simple specification for g that captures the salient interaction patterns for the duration process under study. For example, one possibility in the context of the departure time example discussed earlier is to estimate separate effects of covariates for each of a few discrete periods within the entire time domain.

4. UNOBSERVED HETEROGENEITY

The third important structural issue in specifying a hazard duration model is unobserved heterogeneity. Unobserved heterogeneity arises when unobserved factors (*i.e.*, those not captured by the covariate effects) influence durations. It is now well-established that failure to control for unobserved heterogeneity can produce severe bias in the nature of duration dependence and the estimates of the covariate effects (Heckman and Singer, 1984). Specifically, failure to incorporate heterogeneity appears to lead to a downward biased estimate of duration dependence and a bias toward zero for the effect of external covariates.

The standard procedure used to control for unobserved heterogeneity is the random effects estimator (see Flinn and Heckman, 1982). In the PH specification, heterogeneity is introduced as follows:

$$\lambda(u) = \lambda_0(u)\exp(-\beta' x + w), \tag{13}$$

where w represents unobserved heterogeneity. This formulation involves specification of a distribution for w across individuals in the population. Two general approaches may be used to specify the distribution of unobserved heterogeneity: one is to use a parametric distribution, and the second is to adopt a nonparametric heterogeneity specification.

Most earlier research used a parametric form to control for unobserved heterogeneity. The problem with the parametric approach is that there is seldom any justification for choosing a particular distribution. Furthermore, the consequence of a choice of an incorrect distribution on the consistency of the model estimates can be severe (see Heckman and Singer, 1984). An alternative, more general, approach to specifying the distribution of unobserved heterogeneity is to use a nonparametric representation for the distribution and to estimate the distribution empirically from the data.

5. MODEL ESTIMATION

The estimation of duration models is typically based on the maximum likelihood approach. Here this approach is discussed separately for parametric and nonparametric hazard distributions. The index *i* is used for individuals and each individual's spell duration is assumed to be independent of those of others.

5.1. Parametric Hazard Distribution

For a parametric hazard distribution, the maximum likelihood function can be written in terms of the implied duration density function (in the absence of censoring) as follows:

$$‹\,(\theta,\beta) = \prod_i f\,(u_i,\theta,x_i,\beta)\,, \tag{14}$$

where θ is the vector of parameters characterizing the assumed parametric hazard (or duration) form.

In the presence of right censoring, a dummy variable δ_i is defined that assumes the value 1 if the *i*th individual's spell is censored, and 0 otherwise. The only information for censored observations is that the duration lasted at least until the observed time for that individual. Thus, the contribution for censored observations is the endurance probability at the censored time. Consequently, the likelihood function in the presence of right censoring may be written as

$$‹\,(\theta,\beta) = \prod_i \left\{ [\,f(u_i,\theta,x_i,\beta)]^{(1-\delta_i)}[S(u_i,\theta,x_i,\beta)]^{\delta_i} \right\}. \tag{15}$$

The above likelihood function may be rewritten in terms of the hazard and endurance functions by using equation (2):

$$‹\,(\theta,\beta) = \prod_i \left\{ [\lambda\,(u_i,\theta,x_i,\beta)]^{(1-\delta_i)}\;[S\,(u_i,\theta,x_i,\beta)]^{\delta_i} \right\}. \tag{16}$$

The expressions above assume random (or independent) censoring; i.e., censoring does not provide any information about the level of the hazard for duration termination.

In the presence of unobserved heterogeneity, the likelihood function for each individual can be developed conditional on the parameters η characterizing the heterogeneity distribution function

$J(.)$. To obtain the unconditional (on η) likelihood function, the conditional function is integrated over the heterogeneity density distribution:

$$\mathcal{L}(\theta,\beta) = \prod_{i} \int_{H} \mathcal{L}_i(\theta,\beta,\eta)\,dJ(\eta)\,, \tag{17}$$

Where H is the range of η. Of course, to complete the specification of the likelihood function, the form of the heterogeneity distribution has to be specified.

As discussed in Section 4, one approach to specifying the heterogeneity distribution is to assume a certain parametric probability distribution for $J(.)$, such as a gamma or a normal distribution. The problem with this parametric approach is that there is seldom any justification for choosing a particular distribution. The second, nonparametric, approach to specifying the distribution of unobserved heterogeneity estimates the heterogeneity distribution empirically from the data. This is achieved by approximating the underlying unknown heterogeneity distribution by a finite number of support points and estimating the location and associated probability masses of these support points.

5.2. Non-Parametric Hazard Distribution

The use of a nonparametric hazard requires grouping of the continuous-time duration into discrete categories. The discretization may be viewed as a result of small measurement error in observing continuous data, as a result of rounding off in the reporting of duration times, or a natural consequence of the discrete times in which data are collected.

Let the discrete time intervals be represented by an index k (k = 1, 2, 3, ..., K) with $k = 1$ if $u \in [0, u^1]$, $k = 2$ if $u \in [u^1, u^2]$, ..., $k = K$ if $u \in [u^{K-1}, \infty]$. Let t_i represent the discrete period of duration termination for individual i (thus, $t_i = k$ if the shopping duration of individual i ends in discrete period k). The objective of the duration model is to estimate the temporal dynamics in activity duration and the effect of covariates (or exogenous variables) on the continuous activity duration time.

The subsequent discussion is based on a PH model (a non-parametric hazard is difficult to incorporate within an accelerated lifetime model). The linear model interpretation is used for the PH model since it is an easier starting point for the nonparametric hazard estimation:

$$\ln \Lambda_0 (u_i) = \beta' x_i + \varepsilon_i \text{ , where } \Pr(\varepsilon_i < z) = G(z) = 1 - \exp[-\exp(z)]. \tag{18}$$

The dependent variable in the above equation is a continuous *unobserved* variable. However, we do observe the discrete time-period, t_i, in which individual *i* ends his or her duration. Defining u^k as the continuous-time value representing the upper bound of discrete time period *k*, we can write:

$$\begin{aligned}\text{Prob}[t_i = k] &= \text{Prob}[u^{k-1} < T_i \le u^k] \\ &= \text{Prob}[\ln \Lambda_0 (u^{k-1}) < \ln \Lambda_0 (T_i) \le \ln \Lambda_0 (u^k)] \\ &= G(\psi_k - \beta' x_i) - G(\psi_{k-1} - \beta' x_i)\end{aligned} \tag{19}$$

from equation (18), where $\psi_k = \ln \Lambda_0 (u^k)$. The parameters to be estimated in the nonparametric baseline model are the (K – 1) ψ parameters ($\psi_0 = -\infty$ and $\psi_K = +\infty$) and the vector β. Defining a set of dummy variables

$$M_{ik} = \begin{cases} 1 & \text{if failure occurs in period } k \text{ for individual } i \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

$$(i = 1, 2, \ldots, N;\ k = 1, 2, \ldots, K),$$

the likelihood function for the estimation of these parameters takes the familiar ordered discrete choice form

$$‹ = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[G(\psi_k - \beta' x_i) - G(\psi_{k-1} - \beta' x_i) \right]^{M_{ik}} . \tag{21}$$

Right censoring can be accommodated in the usual way by including a term which specifies the probability of not failing at the time the observation is censored.

The continuous-time baseline hazard function in the nonparametric baseline model is estimated by assuming that the hazard remains constant within each time period *k*; i.e., $\lambda_0(u) = \lambda_0(k)$ for all $u \in \{u^{k-1}, u^k\}$. Then, we can write:

$$\lambda_0 (k) = \frac{\exp(\psi_k) - \exp(\psi_{k\text{-}1})}{\Delta u^k} \text{ , } k = 1, 2, \ldots, K-1 \text{ ,} \tag{22}$$

where Δu^k is the length of the time interval *k*.

The discussion above does not consider unobserved heterogeneity. In the presence of unobserved heterogeneity, the appropriate linear model interpretation of the PH model takes the form

$$\ln \Lambda_0(u_i) = \beta' x_i + \varepsilon_i + w_i \,, \tag{23}$$

where w_i is the unobserved heterogeneity component. We can then write the probability of an individual's duration ending in the period *k*, conditional on the unobserved heterogeneity term, as

$$\text{Prob}\,[t_i = k \mid w_i] = G(\psi_k - \beta' x_i + w_i) - G(\psi_{k-1} - \beta' x_i + w_i)\,. \tag{24}$$

To continue the development, an assumption needs to be made regarding the distributional form for w_i. This assumed distributional form may be one of several parametric forms or a nonparametric form. We next consider a gamma parametric mixing distribution (since it results in a convenient closed-form solution) and a more flexible non-parametric shape.

For the gamma mixing distribution, consider equation (24) and rewrite it using equations (18) and (19):

$$\text{Prob}\,[t_i = k \mid w_i] = \exp[-\{I_{i,k-1}\exp(w_i)\}] - \exp[-\{I_{i,k}\exp(w_i)\}]\,, \tag{25}$$

where $I_{ik} = \Lambda_0(u^k)\exp(-\beta' x_i)$. Assuming that $v_i [= \exp(w_i)]$ is distributed as a gamma random variable with a mean of 1 (a normalization) and variance σ^2, the unconditional probability of the spell terminating in the discrete-time period *k* can be expressed as

$$\text{Prob}\,[t_1 = k] = \int_0^{\infty} \left(\exp[-\{I_{i,k-1} v_i\}] - \exp[-\{I_{i,k} v_i\}]\right) f(v_i)\, dv_i \tag{26}$$

Using the moment-generating function properties of the gamma distribution (see Johnson and Kotz, 1970), the expression above reduces to

$$\text{Prob}\,[t_i = k] = [1 + \sigma^2 I_{i,k-1}]^{-\sigma^{-2}} - [1 + \sigma^2 I_{i,k}]^{-\sigma^{-2}}\,, \tag{27}$$

and the likelihood function for the estimation of the (K-1) integrated hazard elements $\Lambda_0(T^k)$, the vector β, and the variance σ^2 of the gamma mixing distribution is

$$\text{‹} = \prod_{i=1}^{N} \prod_{k=1}^{K} \left\{ [1+\sigma^2 I_{i,k-1}]^{-\sigma^{-2}} - [1+\sigma^2 I_{i,k}]^{-\sigma^{-2}} \right\}^{M_{ik}} \tag{28}$$

For a nonparametric heterogeneity distribution, reconsider equation (23) and approximate the distribution of w_i by a discrete distribution with a finite number of support points (say, S). Let the location of each support point ($s = 1, 2, ..., S$) be represented by l_s and let the probability mass at l_s be π_s. Then, the unconditional probability of an individual i terminating his or her duration in period k is

$$\text{Prob}\,[t_i = k] = \sum_{s=1}^{S} \left\{ \left[\text{G}(\delta_\text{k} - \beta' x_i + l_s) - \text{G}(\delta_\text{k-1} - \beta' x_i + l_s) \right] \pi_s \right\}. \tag{29}$$

The sample likelihood function for estimation of the location and probability masses associated with each of the S support points, and the parameters associated with the baseline hazard and covariate effects, can be derived in a straightforward manner as

$$\text{‹} = \prod_{i=1}^{N} \left\{ \sum_{s=1}^{S} \left[\left\{ \prod_{k=1}^{K} \left[G(\delta_k - \beta' x_i + l_s) - G(\delta_{k-1} - \beta' x_i + l_s) \right]^{M_{ik}} \right\} \pi_s \right] \right\}. \tag{30}$$

Since we already have a full set of (K–1) constants represented in the baseline hazard, we impose the normalization that

$$E(w_i) = \sum_{s=1}^{S} \pi_s \, l_s = 0 \tag{31}$$

The estimation procedure can be formulated such that the cumulative mass over all support points sum to one.

One critical quantity in empirical estimation of the nonparametric distribution of unobserved heterogeneity is the number of support points, S, required to approximate the underlying distribution. This number can be determined by using a stopping-rule procedure based on the Bayesian information criterion, which is defined as follows:

$$BIC = -\ln(\text{‹}) + 0.5 \cdot R \cdot \ln(N) \tag{32}$$

where the first term on the right-hand side is the log (likelihood) value at convergence, R is the number of parameters estimated, and N is the number of observations. As support points are added, the BIC value keeps declining till a point is reached where addition of the next support point results in an increase in the BIC value. Estimation is terminated at this point and the number of support points corresponding to the lowest value of BIC is considered the appropriate number for S.

6. MISCELLANEOUS OTHER TOPICS

In this section other methodological topics are briefly discussed, including left censoring, time-varying covariates, multiple spells, multiple-duration processes, and simultaneous-duration processes.

6.1. Left Censoring

Left censoring occurs when a duration spell has already been in progress for sometime before duration data begins to be collected. One approach to accommodate left censoring is to jointly model the probability that a duration spell has begun before data collection by using a binary choice model along with the actual duration model. This is a self-selection model and can be estimated with specialized econometric software.

6.2. Time-Varying Covariates

Time-varying covariates occur in the modeling of many duration processes and can be incorporated in a straightforward fashion. The hazard, survivor, and density functions at any time u will depend on the entire time path of the time-varying regressors until time u. The maximum likelihood functions will need to be modified to accommodate this effect. In practice, regressors may change only a few times over the range of duration time, and this can be used to simplify the estimation. For the nonparametric hazard, the time-varying covariates have to be assumed to be constant for each discrete period. To summarize, there are no substantial conceptual or computational issues arising from the introduction of time-varying covariates. However, interpretation can become tricky, since the effects of duration dependence and the effect of trending regressors is difficult to disentangle.

6.3. Multiple Spells

Multiple spells occur when the same individual is observed in more than one episode of the duration process. This occurs when data on event histories are available. For example, Hensher (1994) considers the timing of change for automobile transactions (i.e., whether a household keeps the same car as in the year before, replaces the car with another used one, or replaces the car with a new one) over a 12-year period. In his analysis, the data includes multiple transactions of the same household. Another example of multiple spells in a transportation context arises in the modeling of home-stay duration of individuals during a day; there can be multiple home-stay duration spells of the same individual. In the presence of multiple spells, three issues arise. First, there may be lagged duration dependence, where the durations of earlier spells may have an influence on later spells. Second, there may be occurrence dependence where the number of earlier spells may have a bearing on the length of later duration spells. Third, there may be unobserved heterogeneity specific to all spells of the same individual (e.g., all home-stay durations of a particular individual may be shorter than those of other observationally equivalent individuals). Accommodating all the three effects at the same time is possible, though interpretation can become difficult and estimation can become unstable. The reader is referred to Mealli and Pudney (1996) for a detailed discussion.

6.4. Multiple Duration Processes

The discussion thus far has focused on the case where durations end as a result of a single event. For example, home-stay duration ends when an individual leaves home to participate in an activity. A limited number of studies have been directed toward modeling the more interesting and realistic situation of multiple-duration-ending outcomes. For example, home stay duration may be terminated because of participation in shopping activity, social activity, or personal business.

Previous research on multiple-duration-ending outcomes (*i.e.*, competing risks) have extended the univariate PH model to the case of two competing risks in one of three ways:

(1) The first method assumes independence between the two risks (see Gilbert, 1992). Under such an assumption, estimation proceeds by estimating a separate univariate hazard model for each risk. Unfortunately, the assumption of independence is untenable in most situations and, at the least, should be tested.

(2) The second method generates a dependence between the two risks by specifying a bivariate parametric distribution for the underlying durations directly (see Diamond and Hausman, 1985).

(3) The third method accommodates interdependence between the competing risks by allowing the unobserved components affecting the underlying durations to be correlated (Cox and Oakes, 1984, page 159-161; Han and Hausman, 1990).

A shortcoming of the competing-risk methods discussed above is that they tie the exit state of duration very tightly with the length of duration. The exit state of duration is not explicitly modeled in these methods; it is characterized implicitly by the minimum competing duration spell. Such a specification is restrictive, since it assumes that the exit state of duration is unaffected by variables other than those influencing the duration spells and implicitly determines the effects of exogenous variables on exit-state status from the coefficients in the duration hazard models.

Bhat (1996a) considers a generalization of the Han and Hausman competing-risk specification where the exit state is modeled explicitly and jointly with duration models for each potential exit state. Bhat's model is a generalized multiple-durations model, where the durations can be characterized either by multiple entrance states or by multiple exit states, or by a combination of entrance and exit states.

6.5. Simultaneous Duration Processes

In contrast to multiple-duration processes, where the duration episode can end because of one of multiple outcomes, a simultaneous-duration process refers to multiple-duration processes that are structurally interrelated. For example, Lillard (1993) jointly modeled marital duration and the timing of marital conceptions, because these two are likely to be endogenous to each other. Thus, the risk of dissolution of a marriage is likely to be a function of the presence of children in the marriage (which is determined by the timing of marital conception). Of course, as long as the marriage continues, there is the "hazard" of another conception. In a transportation context, the travel-time duration to an activity and the activity duration may be inter-related. The methodology to accommodate simultaneous-duration processes is straightforward, though cumbersome. There have been very few empirical applications of simultaneous-duration processes.

7. CONCLUSIONS AND TRANSPORT APPLICATIONS

Hazard-based duration modeling represents a promising approach for examining duration processes in which understanding and accommodating temporal dynamics is important. At the same time, hazard models are sufficiently flexible to handle censoring, time-varying covariates, and unobserved heterogeneity.

There are several potential areas of application of duration models in the transportation field. These include the analysis of delays in traffic engineering (e.g., at signalized intersections, at stop-sign controlled intersections, at toll booths), accident analysis (*i.e.*, the personal or environmental factors that affect the hazard of being involved in an accident), incident-duration analysis (e.g., time to detect an incident, time to respond to an incident, time to clear an incident, time for normalcy to return), time for adoption of new technologies or new employment arrangements (electric vehicles, in-vehicle navigation systems, telecommuting, for example), temporal aspects of activity participation (e.g., duration of an activity, travel time to an activity, home-stay duration between activities, time between participating in the same type of activity), and panel-data related durations (e.g., drop-off rates in panel surveys, time between automobile transactions, time between taking vacations involving intercity travel).

In contrast to the large number of potential applications of duration models in the transport field, there have been surprisingly very few actual applications. Hensher and Mannering (1994) provide a comprehensive review of applications before 1994, and so attention is confined here to applications past 1993. In this post-1993 period, almost all applications have been in the area of activity and travel-behavior analysis, as discussed below.

Mannering *et al*. (1994a) analyzed home-stay duration between successive participations in out-of-home activity episodes. A Cox partial likelihood approach was used, which does not require specification of a parametric hazard shape, but which also does not estimate the hazard distribution. A PH form was used to accommodate the effect of external covariates. Unobserved heterogeneity is not considered. The authors use 2-day activity diaries and were therefore able to accommodate state-dependence effects by modeling the second-day home-stay durations as a function of the number and length of home-stay durations on the first day.

In another study, Mannering *et al*. (1994b) examined the length of traffic delay needed to induce a commuter to change routes. A Weibull duration distribution is used in combination with a

PH functional form. This study accommodates unobserved heterogeneity using a gamma distribution, which resulted in a closed-form expression for the likelihood function.

Ettema (1995) formulate a competing-risk model to model purpose-specific activity durations, with the termination states being any one of several activity types such as in-home leisure, work or education, shopping, *etc*. Several parametric hazard shapes were tested. An accelerated lifetime formulation was employed to include the effect of covariates so that the covariates rescale time directly. The use of such a formulation does not allow unobserved heterogeneity. The authors assume independence among the competing risks.

Neimeier and Morita (1996) estimated models for the duration of out-of-home activity episodes. They focus on examining how durations for household and family support shopping, personal business, and free time vary between men and women. As in Mannering *et al*. (1994a), a Cox partial likelihood approach was combined with a PH form to accommodate the effect of external covariates. Unobserved heterogeneity is not considered.

Wang (1996) used a Weibull duration distribution to model the start times during the day of a variety of different activities such as breakfast, lunch, and shopping. He used a PH form and did not accommodate unobserved heterogeneity. Kitamura *et al.* (1997) used a similar model structure to examine duration times for out-of-home activities.

Bhat (1996b) combines a nonparametric baseline hazard (based on the Han and Hausman approach) and a nonparametric unobserved heterogeneity specification with a PH form in his examination of shopping duration during the evening commute. His study indicates that, at least in the context of the empirical analysis of the paper, the nonparametric baseline-nonparametric unobserved heterogeneity specification is preferable to other parametric specifications for the baseline or for heterogeneity or both.

In another study, Bhat (1996a) considered a generalization of the Han and Hausman competing-risk specification in which he modeled activity-type choice and activity duration during the evening commute. The formulation was based on a nonparametric baseline hazard and a PH specification for the duration spells of each activity type. Unobserved heterogeneity was not considered.

Recently, Misra and Bhat (1999) used a nonparametric baseline approach combined with a PH form to examine the socio-demographic factors influencing the home-stay duration prior to the first out-of-home sojourn for nonworkers. The authors adopted a gamma distribution for unobserved heterogeneity, because it results in an analytically convenient closed-form likelihood function during estimation.

It is clear from the above review that there have been few applications of duration models in the transport field. In their review paper of duration models in 1994, Hensher and Mannering also pointed to this lack of use of hazard-based methods. The hope is that, by laying bare the simple underlying technical concepts involved in the formulation of duration models, the present chapter will promote the use of duration models in the years to come.

Acknowledgments

This research was supported by National Science Foundation grants DMS 9208758 and DMS 9313013 to the National Institute of Statistical Sciences (NISS). The author would like to thank Jennifer Steed for her help in searching the transport literature for applications of duration models and Lisa Weyant for help with the illustration in Figure 1.

REFERENCES

Bhat, C.R. (1996a). A hazard-based duration model of shopping activity with nonparametric baseline specification and nonparametric control for unobserved heterogeneity, *Transportation Research*, 30B, 3, 189-207.

Bhat, C.R. (1996b). A generalized multiple durations proportional hazard model with an application to activity participation behavior during the work commute, *Transportation Research*, 30B, 432-452.

Cox, D.R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society*, B, 26, 186-220.

Cox, D.R. and D. Oakes (1984). *Analysis of Survival Data*, Chapman and Hall, London.

Diamond, P. and J. Hausman (1984). The retirement and unemployment behavior of older men, in H. Aaron and G. Burtless (eds.), *Retirement and Economic Behavior*, Brookings Institute, Washington, D.C.

Ettema, (1995), Competing risk hazard model of activity choice, timing, sequencing, and duration, *Transportation Research Record*, 1493, 101-109.

Flinn, C. and J. Heckman (1982). New methods for analyzing structural models of labor force dynamics, *Journal of Econometrics*, 18, 115-168.

Gilbert, C.C.M. (1992). A duration model of automobile ownership, *Transportation Research*, 26B, 2, 97-114.

Han, A. and J.A. Hausman (1990). Flexible parametric estimation of duration and competing risk models, *Journal of Applied Econometrics*, 5, 1-28.

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

Hensher, D.A. and F.L. Mannering (1994). Hazard-based duration models and their application to transport analysis, *Transport Reviews*, 14, 1, 63-82.

Hensher, D.A. (1994). The timing of change for automobile transactions: a competing risk multispell specification, Presented at the Seventh International Conference on Travel Behavior, Chile, June.

Johnson, N. and S. Kotz (1970) *Distributions in Statistics: Continuous Univariate Distributions*, John Wiley, New York, Chapter 21.

Kiefer, N.M. (1988). Economic duration data and hazard functions, *Journal of Economic Literature*, 27, June, 646-679.

Kitamura, R., C. Chen and R. Pendyala (1997). Generation of synthetic daily activity-travel patterns, *Transportation Research Record*, 1607, 154-162.

Lillard, L.A. (1993). Simultaneous equations for hazards: marriage duration and fertility timing, *Journal of Econometrics*, 56, 189-217.

Mannering, F., E. Murakami and S-G. Kim (1994a). Temporal stability of travelers' activity choice and home-stay duration: some empirical evidence, *Transportation*, 21, 371-392.

Mannering, F., S-G. Kim, W. Barfield and L. Ng (1994b). Statistical analysis of commuters' route, mode, and departure time flexibility, *Transportation Research*, 2C, 35-47.

Mealli, F. and S. Pudney (1996). Occupational pensions and job mobility in Britain: Estimation of a random-effects competing risks model, *Journal of Applied Econometrics*, 11, 293-320.

Meyer, B.D. (1987). Semiparametric estimation of duration models, Ph.D. Thesis, MIT, Cambridge, Massachusetts.

Meyer, B.D. (1990). Unemployment insurance and unemployment spells, *Econometrica*, 58, 4, 757-782.

Misra, R. and C.R. Bhat (1999). A representation and analysis framework for non-worker activity-travel patterns, Technical Report, Department of Civil Engineering, University of Texas at Austin.

Niemeier, D.A. and J.G. Morita (1996). Duration of trip-making activities by men and women, *Transportation*, 23, 353-371.

Prentice, R. and L. Gloeckler (1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics*, 34, 57-67.

Wang, J.J. (1996). Timing utility of daily activities and its impact on travel, *Transportation Research*, 30A, 189-206.