A Class of Binary Response Models for Grouped Duration Data

Author(s): Glenn T. Sueyoshi
Source: *Journal of Applied Econometrics*, Vol. 10, No. 4 (Oct. - Dec., 1995), pp. 411-431
Published by: John Wiley & Sons
Stable URL: http://www.jstor.org/stable/2285055
Accessed: 22/03/2010 05:48

JOURNAL OF APPLIED ECONOMETRICS, VOL. 10, 411–431 (1995)

A CLASS OF BINARY RESPONSE MODELS FOR GROUPED DURATION DATA

GLENN T. SUEYOSHI
Department of Economics, D008, University of California, San Diego, La Jolla, CA 92093-0508, USA

SUMMARY

This paper explores the relationship between conventional models for binary response such as the probit and logit, and the proportional hazard (PH) and related specifications for grouped duration data. I outline a general class of hazard models for grouped duration data based upon the choice of period-specific distribution functions, facilitating a thorough analysis of the implications of various specifications and consideration of various issues of model identification. This class of models nests, among others, the proportional hazard, probit, and logit specifications for interval survival. I consider the implications of various specifications for hazard behaviour, focusing on familiar specifications. While the specifications will generally yield results that are quite similar along a number of dimensions, there are significant differences. The probit model generates non-proportional effects of variables on the discrete hazard, while the logit and PH tend to show only slight non-proportionality. Furthermore, while the effects of variables on the derivatives are considerably larger for the probit specification, the time-pattern of the probit effects is relatively insensitive to changes in explanatory variables. I illustrate these issues by providing an example taken from Katz's (1986) unemployment data from the Panel Study of Income Dynamics.

1. INTRODUCTION

One strand of econometric duration research has emphasized the close relationship between the standard maximum likelihood estimator for the grouped data, proportional hazard (PH) model, and traditional binary outcome specifications (Kiefer, 1988a; Sueyoshi, 1991). This literature argues that at a fundamental level, the likelihood of a particular observation on a grouped PH duration is simply the probability of observing a series of binary outcomes, with the probabilities for each trial given by the extreme value cumulative distribution function evaluated at an aggregator of period specific data and parameters.[1]

The equivalence of the PH specification and a binary extreme value model suggests that it may be profitable to explore further the relationship between commonly employed hazard specifications and familiar models for binary data such as the probit and the logit. Given the considerable accumulated experience of empirical researchers in estimating and interpreting binary response models, probit or logit duration specifications may be a natural approach for analysing grouped duration data. Indeed, the PH model is sometimes advanced as a more sophisticated alternative to a naive approach of estimating a probit-type model with observations pooled across discrete durations (Diamond and Hausman, 1984). The PH model is perceived as an attractive alternative since it allows for a simple, easily estimable form of time-variation and

[1] The presence of unobserved, individual specific components in the hazard complicates matters slightly, generating a panel random effects specification. See Sueyoshi (1994) for details.

CCC 0883–7252/95/040411–21

Received November 1991
Revised July 1994

covariate-dependence in the conditional survival probabilities. Thus, a generalization of the naive pooled probit or logit specification would address an obvious gap in the literature and have the added benefit of allowing a class of flexible duration models to be estimated using conventionally available econometric software.[2]

In this paper I examine more closely the relationship between grouped hazard specifications and binary response models. I link the two specifications by outlining a general class of models for discrete conditional survivors that is based upon cumulative distribution functions. I first derive the class of parametric, continuous-time hazard specifications which generates this family of conditional survivor functions. Next, I discuss a number of specific cases, including, but not limited to the extreme value (PH), logit and probit specifications, and use the results to discuss important issues of model identification in the grouped duration setting. These results provide an explicit relationship between hazard and sequential binary outcome models, allowing me to identify the implicit restrictions on hazard behaviour associated with simple logit and probit specifications, and to relax those restrictions in practical ways. The framework of analysis extends readily to other specifications for conditional survivor functions.

The comparison of various duration models outlined in this paper provides a number of insights into the restrictions on hazard behaviour imposed by alternative specifications. For example, examination of the PH and the extended probit and logit specifications suggests that results from the logit and PH specifications will be quite similar. In contrast, estimates from a probit-type group duration model should depart significantly from both of these specifications, exhibiting covariate effects that are decidedly non-proportional. Furthermore, the form of the probit non-proportionality is relatively insensitive to changes in explanatory variables. More generally, as with standard binary response models, it appears that the choice of distributional form for the grouped duration model is not innocuous, and in a duration context will have important implications for the effects of covariates on exit probabilities.

In Section 2, I review results for duration models and outline the structure of the sequential binary outcome approach to analyzing discrete duration data. Section 3 describes a class of hazard specifications which generates binary outcome models with specified probability functions and provides familiar examples based upon PH, as well as models with log-logistic and log-normal durations. In Section 4, I explore the different implications of various assumptions about the form of the binary response model for hazard behaviour. Section 5 contains an empirical example using unemployment duration data from the Panel Study of Income Dynamics (PSID).

2. BASIC DURATION RESULTS

I begin by briefly reviewing relevant results for duration models. Let T be a positive, continuous random variable for the time to exit from a given state. The hazard function at time t is defined as the conditional probability of exiting the state, given survival up to time t:

$$\lambda(t) = \lim_{\mathrm{d}t \downarrow 0} \frac{\mathrm{Prob}(t \leqslant T < t + \mathrm{d}t \mid T \geqslant t)}{\mathrm{d}t} \tag{1}$$

I allow for the influence of observable individual heterogeneity on the hazard rates through observable covariates, or regressor variables, X. More specifically, the hazard λ is allowed to differ across individuals through a parameterization which depends upon observable variables X

[2] See for example, Narendranathan and Stewart (1993), who estimate duration models that follow this approach.

and parameters β; this more general hazard function is denoted $\lambda(t, X, \beta)$. By standard arguments, the survivor function associated with this hazard specification is given by $S(t, X, \beta) = \text{Prob}(T \geq t) = \exp(-\int_0^t \lambda(s, X, \beta)\, ds)$, and the corresponding probability density for T is $f(t, X, \beta) = \lambda(t, X, \beta) S(t, X, \beta)$. Likelihood contributions for a sample of individuals with observable and censored continuous failure times are based upon the specification for S and f.[3]

Conventional economic data provide observations on failure times which are aggregated to form discrete intervals. Thus, one typically observes unemployment spell and union strike durations in weeks, or job tenure in years, rather than as continuous realizations of T. Discrete failure data of this form have been termed grouped duration data (Kiefer, 1988a) and are easily handled by describing a mapping from a continuous-time specification to the discrete observations.[4]

Consider the survivor function evaluated at a set of arbitrarily chosen durations t_j for $j = 1, \ldots, J$. These durations will typically correspond to survey design points; for example, the t_j may refer to the weeks of unemployment or the months of a work stoppage observed in the sample. Following convention, I divide the time until period t_j into j half-open intervals, with bounds given by a fixed set of durations $t_1, t_2, \ldots, t_j$ with $t_0 = 0$. Survival to time t_j is the same as surviving each of the intervals $[t_{k-1}, t_k)$ for $k = 1, \ldots, j$, so the overall survivor function may be expressed in terms of interval specific, conditional survivor functions α defined by

$$\alpha_k(X, \beta) = S(t_k, X, \beta \mid T \geq t_{k-1}) = \exp\left(-\int_{t_{k-1}}^{t_k} \lambda(s, X, \beta)\, ds\right) \tag{2}$$

By the definition of conditional probabilities, the survivor function at an arbitrary t_j may be written as $S(t_j, X, \beta) = \prod_{k=1}^{j} \alpha_k(X, \beta)$.

Note further that likelihood contributions for grouped duration data are based solely upon the survivor functions $S(t)$ evaluated at the periods of interest, and hence upon the α evaluated at various t_j. For example, the probability of an observed exit in the jth interval is given by: $S(t_{j-1}, X, \beta) - S(t_j, X, \beta) = (1 - \alpha_j(X, \beta)) \prod_{k=1}^{j-1} \alpha_k(X, \beta)$, which is the probability of surviving the first $j-1$ intervals, but not surviving the jth. If individual i's duration data takes the form (Y_i, δ_i, X_i) where Y_i represents the interval associated with the observed grouped duration, δ_i is a $(0, 1 = \text{censored})$ right censoring indicator, and X_i is the vector of explanatory variables, the log likelihood function for the N^* individuals in the sample may be written

$$\log L(\theta) = \sum_{i=1}^{N^*} \log\left\{(1 - \alpha_{Y_i}(X_i, \theta))^{1-\delta_i} \prod_{k=1}^{Y_i - 1} \alpha_k(X_i, \theta)\right\} \tag{3}$$

where θ contains β and any extra parameters of interest.

An equivalent specification is described by Kiefer (1988a). Rearranging equation (3), the sample likelihood function may be written as the likelihood for non-identical Bernoulli trials taken over all $N = \sum_{i=1}^{N^*} Y_i$ individual-period combinations,

$$\log L(\theta) = \sum_{n=1}^{N} \{d_n \log \alpha_{j_n}(X_n, \theta) + (1 - d_n) \log(1 - \alpha_{j_n}(X_n, \theta))\} \tag{4}$$

[3] See also the general surveys provided by Lancaster (1990), Kalbfleisch and Prentice (1980), and Kiefer (1988b).

[4] Both Heckman and Singer (1984) and Lancaster argue that it is desirable to work in continuous time and translate to discrete as necessary. In particular, this approach makes explicit the assumptions of stationarity implicit in identification of the marginal effects of explanatory variables. See the above references for further discussion. This paper works in the other direction, deriving the class of continuous-time hazard functions associated with a particular discrete hazard specification, and discussing associated identification issues.

where j_n is the interval associated with the nth individual-period trial, d_n is a binary indicator representing survival of the interval, and X_n is the corresponding vector of covariates. This specification differs from the standard binary response likelihood only in that the usual likelihoods based upon the logistic or normal cumulative distribution functions are replaced by the α response probabilities which depend upon integrated hazard components.

There are three aspects of this representation of the grouped duration specification that deserve emphasis. First, given data on grouped durations, the probability model for the observed data is completely specified through parameterization of the α functions. Thus, any issues associated with duration dependence and the effects of covariates on hazards are embodied in the specification for the functional form of the α, as well as the ways that these conditional survivor functions vary across time. This result has important implications for the identification of duration specifications in grouped settings. Second, since the specification of the α characterizes the observable duration process, any features of the underlying probability model not derived directly from the α are not identified from grouped duration data. Lastly, the apparent similarity of the specification above and a sequential logit or probit model highlights the close relationship between duration models and more familiar models for discrete choice. For example, a conventional binary choice specification which pools observations across time-intervals is equivalent to choosing time-constant α based upon a cumulative distribution function such as the normal or logistic. The stationarity assumption placed upon the coefficients in the pooled specification is therefore associated with particular restrictions upon the underlying hazard process. Given the framework outlined above, it is easy to analyse characteristics of the hazard specifications implied by standard binary choice models, and to build specifications which admit various forms of duration dependence and covariate effects through flexibility in the specification of the α and the associated coefficients.

3. HAZARD SPECIFICATION

3.1. A Class of Parametric Hazards

I consider a class of continuous time hazard models which is a natural extension of the semi-parametric baseline hazard models considered by Prentice and Gloeckler (1978), Meyer (1986), Kiefer (1988a), Han and Hausman (1990), and Sueyoshi (1992, 1991). Translating the continuous time specifications of this section to the grouped duration data found in applied settings raises a number of important identification issues which I defer to Section 3.3. In this extended specification, the hazard is expressed as a set of interval specific functions based upon density and cumulative distribution functions. This class nests the PH, sequential logit and probit models outlined above, as well as other standard binary outcome specifications. Moreover, it allows for a generalization of the simple pooled binary choice specification by allowing for duration dependence and non-proportional effects of covariates.

Suppose for the remainder of the discussion and without loss of generality that a duration of interest t is in the jth interval so that it satisfies, $t_{j-1} \leq t < t_j$. For notational convenience, define the time-varying index function, $Z_j(t) = X\beta + h_j(t)$, (the dependence of Z upon X and β is suppressed), and consider a hazard specification of the form

$$\lambda_j(t, X, \beta) = h_j'(t) \left\{ \frac{f_j(Z_j(t))}{1 - F_j(Z_j(t))} \right\} \tag{5}$$

where j indexes the interval of interest, and where f_j and F_j are the density and cumulative distribution functions for an arbitrary continuous random variable.[5] The expression on the right is simply the negative of the derivative of $\log\{1 - F_j(Z_j(t))\}$. The flexibility in the model arises from the choices of f_j and F_j and the choice of h_j across intervals j. The full specification of the hazard is therefore dependent upon the pair $\{F_j, h_j\}$ for all relevant intervals.

In order to satisfy the positivity restrictions imposed by hazard specifications and the corner conditions for the conditional survivor functions, the continuous and everywhere on the domain $[t_{j-1}, t_j)$ differentiable functions h_j must, for all j, satisfy $\lim_{s\to t_{j-1}} h_j(s) = -\infty$ and $h_j'(s) \geq 0$ for all $s \in [t_{j-1}, t_j)$. The first condition is required for the conditional survivor to have the desired form and for it to approach 1 at the beginning of the interval; the latter condition is obviously required for non-negativity of the hazard. Note that these restrictions do not in themselves impose important shape restrictions upon the hazards because of the flexibility I allow in the choice of F, and the form of h_j for $t > t_{j-1}$. For simplicity, I will assume that the distribution function F is constant across intervals, as are the β, so that the only time-dependence in the hazards is derived from the successive specifications of h_k. These restrictions may be relaxed with little substantive effect on the arguments below.

After substituting in the hazard and performing the required integration, the conditional interval survivor function for durations in the jth interval is given by $\alpha_j(t, X, \beta) = 1 - F(Z_j(t))$.[6] Thus, by standard analysis, the survival probabilities for the kth interval may be viewed as the probability that a random variable exceeds the aggregator, $\Pr(\varepsilon > Z_j(t_j))$, where ε has a cumulative distribution function F.

The overall survivor may be written as $S_T(t, X, \beta) = \alpha_j(t, X, \beta) \prod_{k=1}^{j-1} \alpha_k(t_k, X, \beta)$, with corresponding density function, $f_T(t, X, \beta) = h_j'(t) f(Z_j(t)) \prod_{k=1}^{j-1} \alpha_k(t_k, X, \beta)$. More importantly for practical purposes, the unconditional probability of a failure in the jth discrete interval is given by the simple expression

$$\begin{aligned}\text{Prob}(t_{j-1} \leq T < t_j | X, \beta) &= S(t_{j-1}, X, \beta) - S(t_j, X, \beta) \\ &= F(Z_j(t_j)) \prod_{k=1}^{j-1} \{1 - F(Z_k(t_k))\} \qquad (6)\end{aligned}$$

which is identical to the likelihood associated with a series of binary outcomes given in equation (3). A grouped duration, hazard specification of the form (5) is therefore equivalent to a sequential discrete choice model with error distribution functions F, and period-specific aggregator functions $Z_j(t_j)$.

Because of the importance of the proportional hazards model in applied research, it is worth comparing the present continuous time specification to a conventional continuous time proportional hazards specification. The derivative of the log hazard at t in interval j with respect to the mth covariate is

$$\frac{\partial \log \lambda(t, X, \beta)}{\partial X_m} = \beta_m \left\{ \frac{f'(Z_j(t))}{f(Z_j(t))} + \frac{f(Z_j(t))}{1 - F(Z_j(t))} \right\} \qquad (7)$$

[5] In the general case, the F_j functions may vary across intervals, but in applied work, F_j will most commonly be assumed to be the same for each interval. I choose $Z_j(t)$ to be a linear aggregator of parameters and explanatory variables to correspond to standard treatments of logit and probit estimation, but as in all such models, other single or multiple index specifications may be employed with no substantive effect on the results in this paper.

[6] Note that in the current continuous duration setting, these functions differ in notation from the α functions given above in their dependence on an additional argument t. For purposes of correspondence, one may view the earlier α conditional survivor functions as implicitly evaluated at $t = t_j$.

Proportionality occurs if and only if the term in parentheses on the right is a constant so that the log derivative is independent of the period of observation.[7] Obviously, this restriction holds for only a limited subset of all valid choices for F and h, with various choices generating differing degrees of departure from the proportionality assumption.

3.2. Examples of Hazard Assumptions

In this section, I examine several leading models from the literature on binary outcomes and examine their relationship to the class of continuous-time hazards described above.

3.2.1. Proportional Hazards

Suppose first that for each interval, F is the cumulative distribution function for a Type-I extreme value random variable:

$$F(z) = 1 - \exp\{-\exp(z)\} \tag{8}$$

$$f(z) = \exp\{z - \exp(z)\} \tag{9}$$

for $-\infty < z < \infty$. While the time-function for the jth interval, h_j, is not identified from grouped duration data and in general may take a variety of forms, for illustration purposes it will be useful to choose h_j to satisfy

$$h_j(t) = \log \int_{t_{j-1}}^{t} \lambda_0(s)\,\mathrm{d}s \tag{10}$$

for some arbitrary non-negative function, $\lambda_0(s)$: $\mathrm{R}^+ \rightarrow \mathrm{R}^+$, so that

$$h_j'(t) = \frac{\lambda_0(t)}{\int_{t_{j-1}}^{t} \lambda_0(s)\,\mathrm{d}s} \tag{11}$$

Note that the within-interval dynamics of the hazard depend upon the shape assumptions embodied in the individual h_j, while the overall specification for duration dependence also depends upon the variation in the h across intervals.

By virtue of equation (5), the typical hazard in the jth interval takes the form

$$\lambda_j(t, X, \beta) = \frac{\lambda_0(t)}{\int_{t_{j-1}}^{t} \lambda_0(s)\,\mathrm{d}s} \exp\{Z_j(t)\} = \lambda_0(t)\exp(X\beta) \tag{12}$$

with additive separability of the log hazard in t and X, yielding a traditional proportional hazards specification.

Consider the special case where, for t in the jth interval, $\lambda_0(t) = C_j$, where C_j is a constant. It follows that $h_j(t) = \log(C_j \cdot (t - t_{j-1}))$ and the hazard in the interval is exponential with a constant, individual specific hazard; $\lambda_j(t, X, \beta) = C_j \exp(X\beta)$ and $\dot{\lambda}_j(t, X, \beta) = 0$ for $t \in [t_{j-1}, t_j)$. Note that even in this simple interval exponential form, with F the extreme value CDF, the interval hazard specification offers a generalization of the traditional exponential hazard since different constants C may be chosen for different intervals, generating an interval constant hazard shape as described by Lancaster (1990, p. 43). In this specification, durations will be

[7] Lancaster provides the additional caveat that models with time-varying X are not, strictly speaking, proportional hazard since changes in the X and the hazard may occur only locally.

exponential within intervals but the level of this constant hazard will change across intervals. Alternatively, more general time-variation may be generated through greater flexibility in the choice in the h_j across intervals, generating an interval proportional hazard model (Sueyoshi, 1991). Different proportional hazards models may be derived by specifying $\lambda_0(t)$ and evaluating $h_j(t)$ accordingly. For example, if $h_j(t) = \log(t^\delta - t_{j-1}^\delta)$ the specification is Weibull; $h_j(t) = \log(\delta/\gamma) + \log(e^{\gamma(t-1)} - e^{\gamma(t_{j-1}-1)})$ generates a Gompertz interval hazard.

Not surprisingly, the force of the proportionality assumption is the property that the log derivatives of the hazards associated with respect to explanatory variables are time-invariant; $\partial \log \lambda_j(t, X, \beta)/\partial X_m = \beta_m$ for all intervals and durations, and that the overall survivor and density functions for durations are also extreme value and thus PH:

$$
\begin{aligned}
S_T(t, X, \beta) &= \exp\{-\exp(Z_j(t))\} \prod_{k=1}^{j-1} \exp\{-\exp(Z_k(t_k))\} \\
&= \exp\left\{-\exp(X\beta + \log \int_0^t \lambda_0(s)\, ds\right\}
\end{aligned}
\tag{13}
$$

The overall density function for the duration is defined accordingly. Note, however, that the form of the interval proportional hazard is generally not preserved in the overall model. If, for example, each of the $h_j(t)$ is chosen to generate an interval Weibull hazard, the overall duration density is also Weibull if and only if the δ are the same across intervals. It is easy to verify that the requirement of parameter stationarity also holds for the exponential and Gompertz distributions.

3.2.2. *Log-logistic interval hazards*

Given the familiarity and computational simplicity of the logit binary response specification, an obvious alternative to the PH specification is to choose the F to be a logistic CDF. The logit specification is based upon the cumulative distribution function and density:

$$F(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{14}$$

$$F(z) = \frac{\exp(z)}{(1 + \exp(z))^2} \tag{15}$$

for $-\infty < z < \infty$. A simple analogue of the flexible form PH model might then involve estimating a pooled logit with period-specific constant terms to allow for time-effects.

Once again, so long as the regularity conditions are satisfied, h is not identified from grouped data and may be chosen in quite general ways. Nevertheless, one particular specification may be of interest. Suppose that the h function takes the within-interval log-linear form

$$h_j(t) = \delta \log(t - t_{j-1}) + C \tag{16}$$

for $\delta > 0$ and some constant C. Note that this is a 'local-memory' specification for h_j in the sense that the time function depends upon the elapsed interval duration, rather than the elapsed total duration. It follows from equation (5) that the hazard is given by

$$\lambda_j(t, X, \beta, \delta) = \frac{\delta}{t - t_{j-1}} \left\{ \frac{\exp(X\beta + \delta \log(t - t_{j-1}) + C)}{1 + \exp(X\beta + \delta \log(t - t_{j-1}) + C)} \right\}, \tag{17}$$

which is also the hazard for interval durations which follow the log-logistic distribution with mean $\mu = -\delta^{-1}(X\beta + C)$, and variance $\sigma^2 = \pi^2/3\delta^2$, (Lancaster, 1990, p. 44).[8]

The overall survivor function is given by the product of the conditional survivors,

$$S_T(t, X, \beta) = \frac{1}{1 + \exp(Z_j(t))} \prod_{k=1}^{j-1} \{1 + \exp(Z_k(t_k))\}^{-1} \tag{18}$$

and the underlying probability density for durations is

$$f_T(t, X, \beta) = \frac{\exp(Z_j(t))}{1 + \exp(Z_j(t))} \prod_{k=1}^{j-1} \{1 + \exp(Z_k(t_k))\}^{-1} \tag{19}$$

In contrast to the PH specification above, even when the h_j functions are the same across intervals, the overall survivor is not of the same form as the interval specific survivor functions and thus the overall duration density is not log-logistic. As a result, the duration behaviour implied by the sequential logit interpretation of survival imposes more complicated time-interactions than the PH model.

Evaluating the effects of explanatory variables on the hazard rate is also more involved than in the PH specification. Substituting into equation (7) yields

$$\frac{\partial \log \lambda(t, X, \beta, \delta)}{\partial X_m} = \beta_m \left\{1 - \lambda(t, X, \beta, \delta) \frac{(t - t_{j-1})}{\delta}\right\} \tag{20}$$

so that the effect of covariate m on the hazard at time t involves β_m, but weighted by a time-dependent term that depends upon 'elapsed interval duration' as well as the hazard level. Examination of the parameter β_m alone overstates the effect of the covariate within an interval, with the deviation between β and the true effect increasing with the hazard level and duration.

This specification of the hazard process provides a practical framework for application of the traditional binary logistic specification to duration data. The framework provides an explicit linkage between the binary specification and the underlying hazard, allowing one easily to assess the implications and assumptions of a given specification. For example, if the underlying interval durations are assumed to be distributed as a log-logistic as above, the grouped data maximum likelihood specification involves estimating a pooled logit specification with period specific constant terms. This is a simple generalization of the naive specification in which one estimates a pooled logit with a single constant term. If, however, the h functions are assumed to be stationary and constant with observations on equal length intervals, the pooled specification is optimal.

3.2.3. Log-normal interval hazards

Suppose that one chooses F to be the standard normal distribution function Φ, and estimates the binary outcome model as though it were a standardized probit. Then,

$$F(z) = \Phi(z) \tag{21}$$

$$f(z) = \phi(z) \tag{22}$$

[8] It is easy to verify that the log-logistic hazard may take a variety of shapes depending upon the choice of the parameter δ. The time-derivative of λ, $\dot{\lambda}(t) = \lambda(t)\{(\delta - 1)/(t - t_j) - \lambda(t)\}$. For $\delta < 1$, the hazard is monotonically decreasing; for $\delta = 1$, it is monotonically decreasing with a different origin; for $\delta = 1$ the hazard possesses an inverted U-shape and attains a maximum at $t = t_j + \{(\delta - 1)/\exp(X\beta + C)\}^{1/\delta}$.

As above, there are a number of specifications possible for h_j, but a familiar choice is given by $h_j(t) = \log(t - t_{j-1}) + C$.[9] The resulting hazard specification

$$\lambda(t, X, \beta, \delta) = \frac{1}{t - t_{j-1}} \left\{ \frac{\phi(X\beta - \log(t - t_{j-1}) + C)}{1 - \Phi(X\beta - \log(t - t_{j-1}) + C)} \right\} \tag{23}$$

is recognizable as the hazard for a log-normal interval duration model. The overall survivor function is

$$S_T(t, X, \beta) = \{1 - \Phi(Z_j(t))\} \prod_{k=1}^{j-1} \{1 - \Phi(Z_k(t_k))\} \tag{24}$$

and the underlying probability density for durations is given by

$$f_T(t, X, \beta) = \frac{1}{t - t_{j-1}} \phi(Z_j(t)) \prod_{k=1}^{j-1} \{1 - \Phi(Z_k(t_k))\} \tag{25}$$

which is proportional to the log-normal density within intervals, but is not log-normal across intervals.

Using equation (7) to evaluate the derivative the log hazard with respect to an arbitrary covariate at t,

$$\frac{\partial \log \lambda(t)}{\partial X_m} = \beta_m \left\{ -Z_j(t) + \frac{\phi(Z_j(t))}{1 - \Phi(Z_j(t))} \right\} \tag{26}$$

so that the effect of covariate m on the hazard at time t is multiplicative in β_m, but with the proportionality term depending upon the X, β and t through $Z_j(t)$ and through the Mills-ratio evaluated at $Z_j(t)$.

3.3. Grouped Data and Identification

In the preceding discussion, I outlined a class of continuous-time, parametric hazards which generates binary outcome models for conditional survival. For purposes of illustration, I also provided examples where the $h_j(t)$ functions are assumed to be log-linear in the interval duration so that the interval durations are distributed as extreme value, log-logistic or log-normal, yielding exponential, logistic and normal cumulative distribution functions, respectively. While the log-linearity assumption is sufficient for interval survivors to follow the desired distributions, that assumption is not required for the result. Indeed, any function satisfying the regularity conditions for h outlined above will generate the desired binary response model. For example, one alternative to the log-logistic is to maintain the assumption that F is logistic, but that $h_j(t) = -\delta/(t - t_{j-1})$, $h_j'(t) = \delta/(t - t_{j-1})^2$ for $\delta > 0$. Similarly, one can 'mix' models by assuming, for example, a logistic F but choosing $h_j(t)$ to follow the Weibull or Gompertz forms for h_j outlined above. While the underlying interval durations are not log-logistic, each of these specifications also generates a logit interval survival model; in each case, however, the period-specific constant will possess a different interpretation.

[9] As with all probit models, there is an implicit normalization with respect to the error variance. The force of the arguments is unchanged if every term is standardized by the standard deviation σ. Thus, $h_j(t)$ may be replaced by $\log(t - t_{j-1})/\sigma$, and Φ would be evaluated at $(X\beta + h_j(t))/\sigma$. In this specification, $1/\sigma$ replaces the δ in the log-logistic and is only identified given data with varying interval lengths.

The flexibility allowed in the choice of h is related to more general issues of identifiability in grouped duration settings. To better focus ideas, suppose as above that grouped durations are available on a sample of individuals and that a representative individual i's duration data takes the form described in Section 1. Likelihood contributions are then based upon functions of the aggregators Z_j evaluated at durations t defined by the design points $t_1, \ldots, t_j$. In the context of the continuous time specifications above, the survivor and the Z_j functions are evaluated only at the endpoints $t = t_j$ and never for other t in the interval (t_{j-1}, t_j). Thus, there is no sample information available to identify the behaviour of h within a given j interval.

It follows that assumptions about the within-interval behaviour of the hazard are not in general identified non-parametrically, are not testable using grouped data, and that identification of the entire hazard shape requires some parametric smoothness applied to the h_j functions. Note further that even if one is willing to make some parametric assumptions about within-interval behaviour of h_j, the shape of the hazard may still not be identified solely from the observable data. If, for example, the intervals are of constant length, even the strong assumption that the h_j are identical parametric functions is not sufficient to pin down the within-interval dynamics.[10] Alternatively, if the h_j are assumed to be identical parametric functions of elapsed interval duration and the intervals are of variable length, the variable interval length and assumption of identical h_j functions allows one to identify *parts* of the within-interval hazard shape.

More generally, the current discussion suggests that while the concept of hazard proportionality is of theoretical interest, a continuing focus on properties of the continuous time hazard is misplaced. As noted above, only characteristics associated with the conditional survivor probabilities α are identified from the data. Thus, absent non-testable restrictions on the hazard shape, all that can be identified from grouped data are the $h_j(t_j)$.[11] While specifications of the continuous time hazard function of the general form (5) are of academic interest and facilitate comparisons with familiar proportional hazards specifications such as the Weibull, any assumptions in those specifications that go beyond the choice of α are fundamentally untestable in grouped data settings.

Given the impossibility of identifying within-period dynamics from grouped duration data, it may instead be desirable to consider properties of the hazard which do not rely upon untestable restrictions on within-interval hazard shapes. Evaluation of alternative duration specifications should instead focus on the implications of alternative assumptions about the discrete-time hazard specification, $\Psi_j(X, \beta) = 1 - \alpha_j(X, \beta) = F(Z_j(t_j))$. Indeed, the force of alternative assumptions for a discrete time hazard process lies in differences between the predictive ability of the models and between the estimated duration and covariate effects associated with alternative specifications for the α_j.

4. A COMPARISON OF MODELS

In this section I provide a simple framework for consideration of the implications for hazard behaviour of different specifications for the α. There are two primary points of comparison.

[10] To take a simple example, if the h_j functions are identically log-linear in the elapsed interval duration, interval-specific constants estimated from the binary outcome model provide estimates of $\delta \log(t_j - t_{j-1}) + C$. If there is variation in the interval length, the parameters C and δ are identified (from a regression of the period-specific constants on the $\log(t_j - t_{j-1})$); with common interval lengths, δ and C are not separately identified. Note further that in the special case where the intervals are of constant unit length, C is identified, but δ is not.

[11] This point has been emphasized by Kiefer in the context of PH specifications and is equally valid in the more general current discussion.

First, alternative specifications may imply different predicted probabilities of survival for an individual with given characteristics. Practical experience with discrete choice models suggests, however, that the predicted probabilities, and hence the goodness-of-fits for the models, will generally be quite similar over a wide range of specifications for F.

Apart from goodness-of-fit, applied researchers are most often interested in the effects of changes in the explanatory variables upon hazard rates. Covariates may be thought of as having two primary effects on exit hazards: (1) they can alter the hazard at a given point in time, and (2) they may change the nature of time-dependence in the hazard. By their nature, proportional hazards models concern themselves with (1) and ignore (2) since they place restrictions on the effects of variables upon the hazard rate. In contrast, more general models allow for interaction between the time-dependence in the hazards and the explanatory variables. Thus, of particular interest is a comparison of the extent to which alternative specifications for α with similar goodness-of-fit imply different effects of changing covariates.

For continuous time hazard specifications, a natural measure of the interaction between time-dependence and the explanatory variable is derived by examining the derivative of the log hazard function at various durations. This derivative function is, however, not fully identified from grouped data. A natural alternative for discrete hazards is to consider the derivative of the log discrete hazard with respect to an arbitrary covariate,

$$\frac{\partial \log \Psi(Z_j(t_j))}{\partial X_m} = \beta_m \frac{f(Z_j(t_j))}{F(Z_j(t_j))} \tag{27}$$

which yields a discrete analogue to equation (7). The expression, which should be familiar to those who work with binary response models, provides a measure the proportionate change in the discrete hazard resulting from changes in the explanatory variables.

In continuous time, the assumption of proportionality would ensure that the derivatives of $\log \lambda$ are constant across durations; in the most common specification where the aggregator function is $\exp(X\beta)$, the log derivatives are also independent of the X. These results do not obtain for other distributional assumptions, nor do they necessarily follow in the discrete setting even if the model is PH, since $f(z)/F(z)$ is likely to differ slightly across durations. More generally, the two relevant issues concern whether the logarithmic derivatives of F are roughly constant across time (discrete hazard proportionality), and whether individual characteristics X alter the nature of any observed time-dependence.

To examine these questions, I first consider the behaviour of the derivative of the log discrete hazard function at comparable points. Intuitively, maximum likelihood estimation of binary response models will roughly fit the empirical exit probabilities under the assumed distributional shape.[12] Thus, for a given individual with exit probability p, each model will be evaluated at roughly the inverse of the cumulative distribution function at that point, $z = F^{-1}(p)$. Of primary interest is the degree to which the log derivative $f(z)/F(z)$ varies across specifications, holding the predicted probability $p = F(z)$ constant.

The relationship between models will obviously depend upon the shapes of the cumulative distribution functions and the log derivatives of the cumulatives evaluated at various points. To ground the discussion, I will analyse the differences between the three familiar distributions discussed above: extreme value, normal and logistic. As is well known, the normal and logistic

[12] More formally, White (1982) demonstrates that under reasonable regularity conditions, the maximum likelihood estimator is a strongly consistent estimator for the parameter vector which minimizes the Kullback–Leibler Information. In essence, ML minimizes the distance between the expected true log-likelihood and the expected assumed log-likelihood, with expectations taken with respect to the true distribution.

distribution functions are quite similar, with the density functions differing primarily in the tails. The extreme value distribution is skewed, with the cumulative distribution increasing more rapidly in the left tail than is the case for both the normal and logistic.[13] One might therefore expect that the normal and logistic models will generate estimates that are roughly the same, while the extreme value results will differ due to the asymmetry of the density function.

Figure 1 plots the derivatives of the log cumulative distribution function, $f(z)/F(z)$ for the three specifications against various values of $F(z)$. It is immediately apparent that the log derivatives differ substantively between the three distributions at most cumulative probabilities, with the probit diverging significantly at low probability values, and the logit exhibiting smaller values at higher probabilities. For example, at $F(z)=0{\cdot}10$, the probit derivative is 1·76 as compared with the 0·95 and 0·90 values for the proportional hazard model and logit specification, respectively. This divergence implies that the models will provide very different predictions of the proportionate change in the discrete hazard resulting from changes in the explanatory variables. In particular, for an individual with an exit probability less than 0·5, the probit will show significantly greater sensitivity to explanatory variables than will the other specifications. Thus, the conventional wisdom regarding the similarity of probit and logit models does not extend to evaluation of the proportionate changes in the discrete hazard, nor does it apply to the probit and PH specifications.

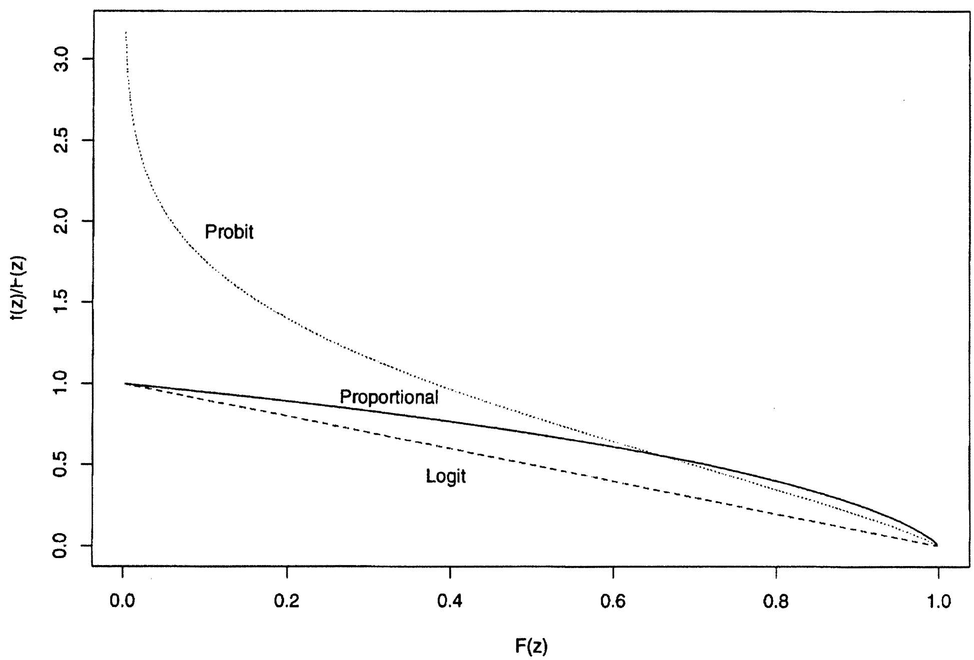

Figure 1. Derivatives of the log hazard function, $f(z)/F(z)$, evaluated at values of the hazard probabilities $F(z)$; various specifications for the discrete hazard function F

[13] In their common forms, the distributions have means 0, 0 and $\delta=-0{\cdot}5772$, where δ is Euler's constant, and variances 1, $\pi^2/3$ and $\pi^2/6$, respectively. For further details and results, see Johnson and Kotz (1970), Nelson (1982), and Lancaster (1990).

The choice of F generates other implications for the predicted influence of X. Bear in mind that the actual z will depend upon characteristics X, parameters β and duration through the h function. Since variation in the hazard for a given specification depends upon changes in the z for a given (generally) nonlinear function F, the hazards will vary across both time, as a result of variation in the period-specific constants, and across individuals, as a result of variation in the X.

To examine the importance of this variation, I plot the log derivative $f(z)/F(z)$ against standardized values of z to assess the sensitivity to changes in z (Figure 2). There are two aspects of Figure 2 that are of primary importance. First, the non-proportionality in the discrete hazard results from a non-zero slope for the curve. For an individual with time-constant covariates, time-variation in the derivative of the log discrete hazard occurs solely as the period-specific constants alter the evaluation point z around $X\hat{\beta}$; with a perfectly flat curve, the effects are constant at all durations. From Figure 2 it is apparent that the slopes for the three specifications differ considerably at various values of z. The slope of the probit curve is considerably larger (in absolute value) than the corresponding logit and PH curves for small values of z, while the proportional and logistic models diverge for values of z in the centre of the distributions. For z in the lower tail, the slope of the PH curve is approximately zero; the logistic exhibits greater variability, but the curve is still relatively flat. As a result, for duration data exhibiting low to moderate exit rates, the binary response hazard models employing a probit specification will tend to depart from proportionality far more than logistic models, which in turn will be slightly less proportional than the extreme value specification.

Second, estimates of the effects of variables on the time-pattern of the percentage changes in the hazard will depend upon the second derivatives of the curves in Figure 2. As noted above, a

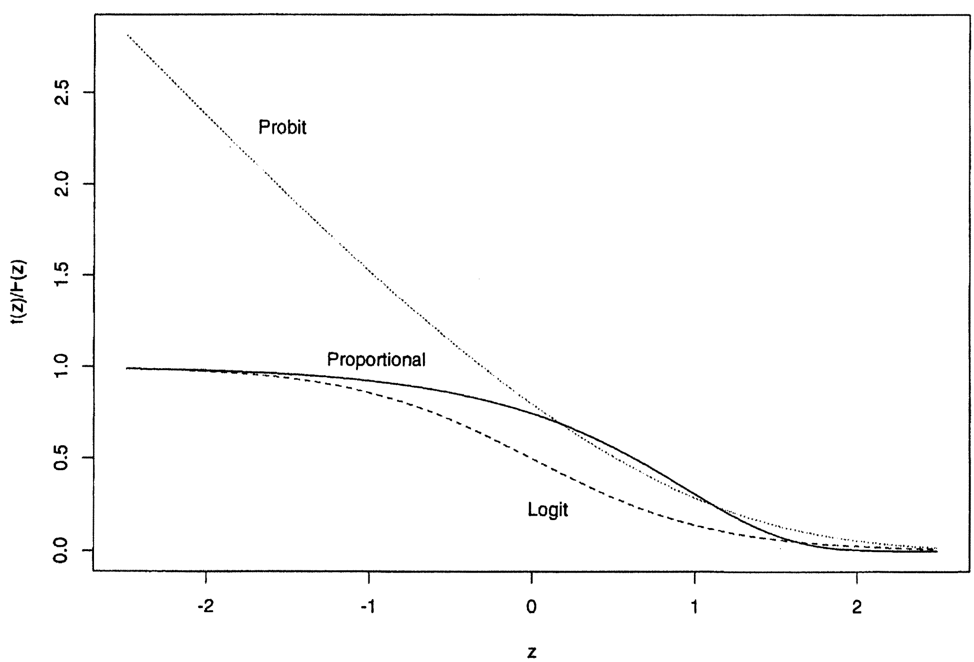

Figure 2. Derivatives of the log hazard function, $f(z)/F(z)$, for various standardized values for z; various specifications for the discrete hazard function F

particular time-pattern for the proportionate impact of covariates is generated by evaluating the derivative functions at a given $X\beta$, and allowing the period-specific constant terms to vary. In general, evaluating these expressions at different X will result in different patterns of time-variation. In this regard, note that the probit specification curve is close to linear for moderate z. Thus, while the log derivatives for the probit specification will have non-proportional effects, the pattern of this time-variation should be approximately the same for various values of X. From Figure 2, it is apparent that individuals with large $X\beta$ will exhibit smaller proportionate changes in the discrete hazard than those with small $X\beta$, but the difference between the two should be roughly constant for various durations. In contrast, the relatively small slopes and greater amounts of curvature in the proportional and logit curves imply that variation in the X will generate only slight changes in the log derivative $f(z)/F(z)$, but that variation in X will alter the time-pattern of these changes.

5. AN EMPIRICAL EXAMPLE

To illustrate a number of the specification issues discussed above, I estimate duration models for unemployment data taken from the Panel Study of Income Dynamics (PSID). I focus on the question of whether probit, logit, and PH models estimated on the same data generate different implications for the hazard of exit from unemployment and the effects of explanatory variables on these conditional exit probabilities.

I estimate relatively simple sequential binary outcome models where duration dependence is built into the specification through a period-specific constant. The data used in this analysis are derived from the 1980 and 1981 PSID (Waves 14 and 15) and consist of observations on the duration of unemployment spells for 1055 individuals.[14] In addition to information on spell duration, the data indicate whether a spell ended via a new job, recall, or by censoring. Accompanying the unemployment spell information are a variety of demographic and economic characteristics.[15] The variables used in this analysis are described briefly and descriptive statistics presented in Table I.

In Table II, I present parameter estimates from three binary outcome specifications of the recall hazard. The estimates are derived from an independent competing risk specification where exit from unemployment via new job is treated as a censored outcome.[16] The first two columns contain parameter estimates and asymptotic standard errors for the PH model and correspond to the model of Table 3 of Sueyoshi (1991). The latter models consist of estimates derived from sequential probit and logit models, where each model contains period-specific constant terms that are designed to capture duration dependence in the hazard process. The models may alternatively be viewed as a simple pooled binary response model with period-specific constants. Approximately 13,300 individual-period responses are used to estimate the 57 parameters in each model.

The fit of the models as measured by a simple likelihood criterion is almost identical across specifications. While the PH specification fits marginally better than the logit specification,

[14] Katz (1986) describes the construction of the data set and characteristics of the sample in greater detail. The data correspond to a subset of observations from the PSID, and excludes spells initiated by plant closing.

[15] More extensive analyses of these data may be found in papers by Katz (1986), Han and Hausman (1990), and Sueyoshi (1991). In particular, I defer consideration of the possibility of time-variation in the coefficients to future analysis.

[16] The maximum likelihood estimation of the sequential binary outcome models employed the Gauss–Newton algorithm in an iterated, weighted nonlinear least squares procedure. Each model is estimated for 40 weeks with durations beyond that period artificially right-censored. Corresponding estimates for the new job hazard are not presented, but are available upon request.

Table I. Variable names and descriptive statistics, Katz (1986) sample of 1055 spells from the PSID

Variable	Description	Mean	Std. dev.
Demographic and economic variables			
Age	Age in previous year (year of spell)	33·154	10·607
Female	Indicator for female	0·167	0·373
Schooling	Number of years schooling	11·341	2·170
Non-white	Indicator for non-white	0·506	0·500
Dependents	Number of dependents in households	3·038	1·639
UI eligible	Indicator for individual receiving unemployment insurance during spell	0·636	0·481
Married	Indicator for marital status in previous year	0·632	0·482
Area unemploy.	County unemployment rate	7·701	2·551
Wife works	Indicator for wife works	0·342	0·475
Homeowner	Indicator for homeownership	0·439	0·496
Occupation dummies			
Labour	Indicator for labourer or operative	0·508	0·500
Craft	Indicator for craft	0·223	0·416
Clerical	Indicator for clerical, services, sales	0·186	0·389
Professional	Indicator for professional and technical	0·039	0·193
Manager	Indicator for manager	0·045	0·206
Industry dummies			
Metals	Indicator for metals	0·058	0·234
Transp. equip.	Indicator for transportation equipment	0·118	0·322
Other durables	Indicator for other durable goods manufacturing (excluding metals)	0·123	0·329
Non-durables	Indicator for non-durable goods manufacturing	0·133	0·339
Trade	Indicator for wholesale and retail trade	0·103	0·305
Transportation	Indicator for transportation and public utilities	0·080	0·271
Mining	Indicator for mining and agriculture	0·034	0·182
Services	Indicator for services	0·172	0·377
Construction	Indicator for construction	0·180	0·384

which in turn fits better than the probit model, the likelihood differences are not striking. It would therefore appear to be premature to base strong conclusions on the proportional hazard results in preference to those from the other specifications. This close equivalence in fit raises the question of whether predictions of the effects of variables on hazard rates derived from these models are identical so that the specification differences are unimportant. The remainder of the analysis will focus on the influence of being *NON-WHITE* on the recall hazard. While many of the other parameters associated with individual characteristics are imprecisely measured, the negative impact of *NON-WHITE* is pronounced, and statistically significant at conventional significance levels.

As is usually the case with models involving binary responses, interpretation of the coefficients is made difficult by the inherent non-linearity of the model. Since the first specification is PH, if this were a continuous model of durations, the β coefficients could be interpreted as the derivative of $\log \lambda$. It is easy to see that this is not the case for the continuous probit and logit specifications since their log derivatives contain both X and t components. The conventional approach in the literature for handling problems of this nature has been to derive

Table II. Selected parameter estimates for binary outcome hazard models of independent recall hazard; various specifications for the discrete hazard function F

	Proportional[a]		Probit		Logit	
Variable	Est.	Std. err.[b]	Est.	Std. err.	Est.	Std. err.
Age	0·014	0·004	0·006	0·002	0·015	0·005
Female	−0·016	0·157	0·006	0·075	−0·014	0·163
Schooling	−0·029	0·022	−0·015	0·012	−0·030	0·024
Non-white	−0·240	0·098	−0·116	0·049	−0·251	0·102
No. of dependents	−0·002	0·030	−0·005	0·015	−0·004	0·031
UI receipt	−0·188	0·097	−0·085	0·048	−0·195	0·102
Married	0·045	0·148	0·037	0·073	0·051	0·154
Area unemploy.	−0·007	0·017	−0·003	0·008	−0·007	0·017
Wife works	0·127	0·101	0·065	0·052	0·135	0·107
Homeowner	0·391	0·099	0·206	0·050	0·411	0·104
Industry indicators	Yes		Yes		Yes	
Occup. indicators	Yes		Yes		Yes	
Log likelihood	−2167·89		−2170·57		−2168·37	

[a] This specification corresponds to the proportional hazard model estimated in Table 3 of Sueyoshi (1991). For each of the three models, estimates of the 57 parameters are derived from 1055 individual observations which correspond to 13,246 individual-interval binary trials. [b] Asymptotic standard errors.

implications for the model evaluated at mean values for the explanatory variables. I modify and extend this approach to consider the behaviour of the model evaluated at quartiles of the aggregator.

The discrete hazard for each period is computed by evaluating the appropriate cumulative distribution function at the sum of an aggregator representing $X\hat{\beta}$ and the period specific intercept term. Figure 3 depicts the estimates of the discrete recall hazard function evaluated at the lower quartile of the sample.[17] Not surprisingly, all three models appear to track closely the same discrete hazard shapes and depict the familiar pattern of declining recall exit hazard. For these data, there are only minor differences between the specifications in terms of predicting discrete hazard rates. There is also little difference if the predicted hazards are disaggregated by race.

Next, I analyse the change in the log discrete hazard associated with the *NON-WHITE* indicator variable. If this were a continuous covariate, one could simply evaluate the term $\beta f(z)/F(z)$ at various z representing the different periods of interest. However, since the variable in question is a binary indicator, I instead compute the discrete change in log F at various durations when the indicator for *NON-WHITE* is first set to 0 and then to 1, and evaluating the proportionate change in F: $(F(1)/F(0) - 1)$. The discrete log derivatives depicted in Figure 4 are computed at the sum of the lower quartile value of the $X\hat{\beta}$ and the estimated $h_j(t_j)$ for each period.

[17] The quartiles of the aggregator are computed for the entire sample, so that they do not reflect changes in the population of interest as individuals exit. Since larger aggregator values are positively associated with exit from unemployment, the quartile estimates will systematically overstate the corresponding values for the population at risk in each period. The lower quartiles of the aggregator are −0·468, −0·196, and −0·481 for the proportional, probit, and logit models, respectively. The corresponding upper quartiles are 0·486, 0·201 and 0·497. The analysis below was repeated for other quantiles. The quantitative results differ slightly, but not substantively.

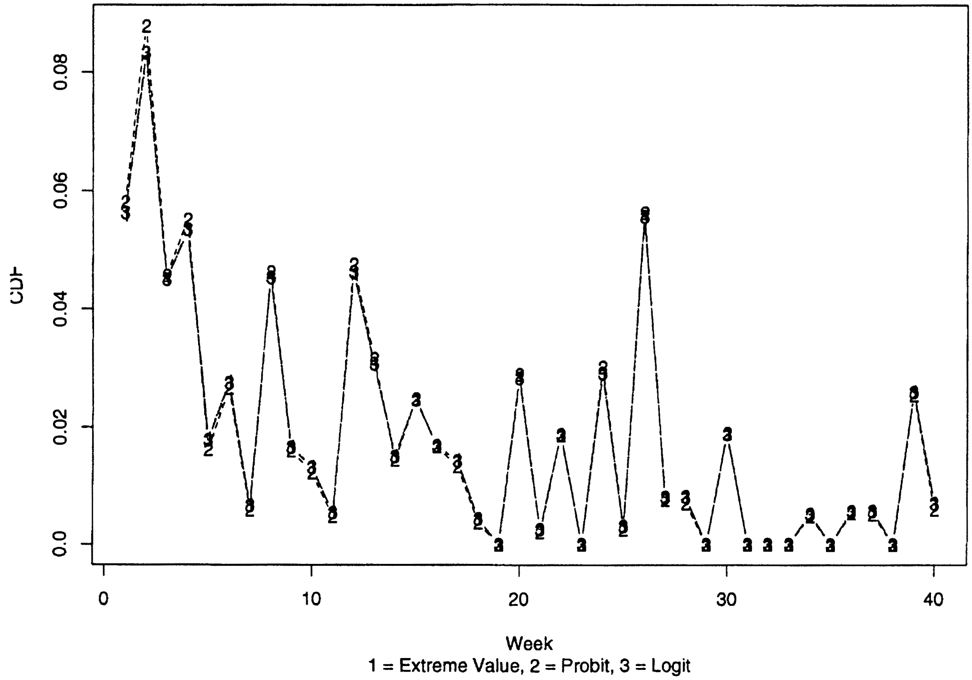

Figure 3. Estimates of the discrete recall hazard, $F(\hat{\gamma}_j + X\hat{\beta})$, with the aggregator evaluated at the lower quartile of the $X\hat{\beta}$ and an estimate of the period specific constant γ_j; various specifications of the discrete hazard function F

The differences across specifications are striking. While the PH and logit models exhibit only minor time-variability for the change in the log discrete hazard, the probit model predicts large, decidedly non-proportional effects of the *NON-WHITE* indicator.[18] For example, the marginal effect of *NON-WHITE* on the hazard reduces the discrete unemployment recall hazard by 20% at a duration of 2 weeks, and by 33% at 21 weeks. The corresponding effects for the PH specification are 22% and 24%. Furthermore, the pattern of the variation clearly mirrors the variation in the underlying level of the log discrete hazard. As expected given the shapes of the underlying distribution functions, the logit model exhibits somewhat greater non-proportionality than the discrete proportional hazard model, but considerably less variability than the probit specification (Table III). While the mean proportionate change is generally comparable across all three specifications (though somewhat larger for the probit), the across-time variability as measured by both the standard deviation and inter-quartile range is 10–20 times greater for the probit than the other two specifications, with the logit model again acting as the intermediate specification.

The second prediction of the framework above is that the time-pattern of the changes in the log discrete hazard for the probit model should not be sensitive to changes in the X. In Figure 5 and Table IV, I examine this characteristic by computing the change in the log discrete hazard

[18] Since the derivatives for the probit model exhibit so much time-variation, to aid in visual interpretation of Figure 4, I also plot a line representing a smooth of the probit derivatives using locally weighted regression (Chambers *et al.*, 1983).

Table III. Descriptive statistics for the proportionate change in the discrete recall hazard associated with the *NON-WHITE* indicator; various specifications for the discrete hazard function F

	Proportional	Probit	Logit
Mean[a]	−0·2111	−0·2561	−0·2177
Std. dev.	0·0017	0·0311	0·0035
Minimum	−0·2129	−0·3117	−0·2214
Maximum	−0·2060	−0·1949	−0·2072
Q3 − Q1	0·0020	0·0501	0·0040

[a] The proportionate change is computed as $F(1)/F(0) - 1$, where $F(1)$ is the discrete hazard with the *NON-WHITE* indicator set to 1, and $F(0)$ is the discrete hazard with the indicator set to 0. The $X\hat{\beta}$ are set to their lower quartile values and the period specific constants are used to estimate $h_j(t_j)$. The descriptive statistics are computed for the 32 periods with non-zero discrete hazard estimates.

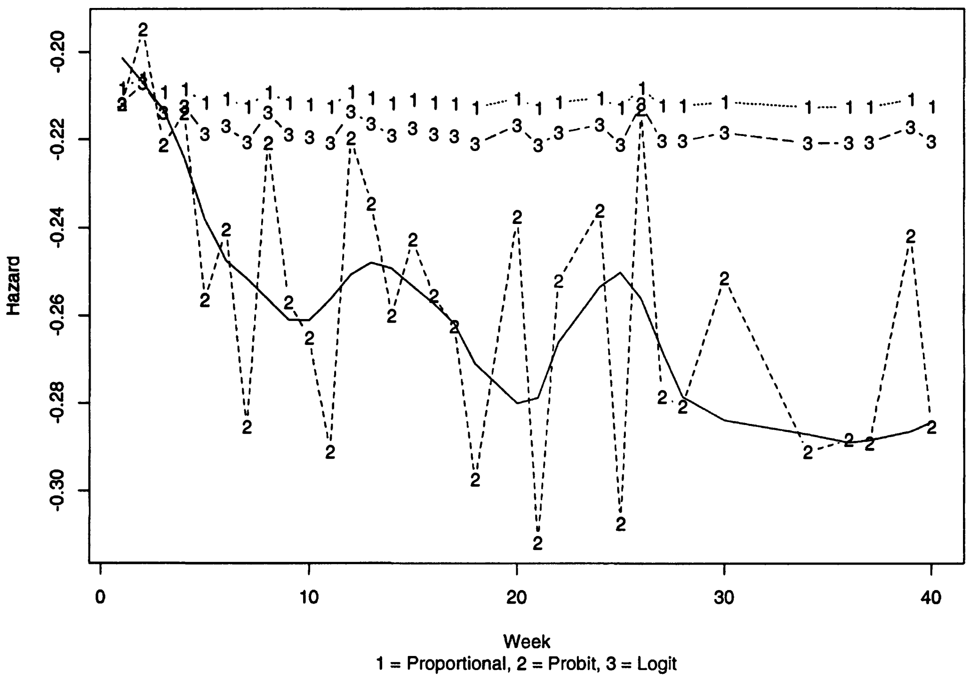

Figure 4. Proportionate change in the discrete recall hazard associated with the *NON-WHITE* indicator variable, $F(X\hat{\beta} + \hat{\gamma}_j)$, with the aggregator evaluated at lower quartile of $X\hat{\beta}$ and an estimate of the period specific constant γ_j. The discrete change in the hazard is computed as $F(1)/F(0) - 1$, where $F(1)$ is the hazard with the *NON-WHITE* indicator set to 1, and $F(0)$ is the hazard with the indicator set to 0

evaluated at both the upper and lower quartiles of the $X\hat{\beta}$ and computing the difference at each duration. As predicted, the magnitude of the change in the hazard associated with the indicator variable differs greatly across quartiles for the probit model. The probit specification indicates that high quartile individuals possess, on average, derivatives that are roughly 3 percentage points higher than the derivatives for individuals with the low quartile value for $X\hat{\beta}$. However, as confirmed by the flat shape of the probit graph, there is little variation in the pattern of the changes across time so that the non-proportional time-pattern observed in Figure 4 is not altered

Table IV. Descriptive statistics for the difference between the upper and lower quartile estimates of the proportionate effect of race upon the discrete recall hazard of exit from unemployment; various specifications for the discrete hazard function F

	Proportional	Probit	Logit
Mean[a]	0·0024	0·0295	0·0048
Std. dev.	0·0020	0·0000	0·0039
Minimum	0·0003	0·0286	0·0005
Maximum	0·0084	0·0300	0·0160
Q3 – Q1	0·0023	0·0007	0·0048

[a] The quartile differences are computed as $D(u) - D(l)$ where $D(u) = F(1, u)/F(0, u) - 1$ is the proportionate change in the discrete hazard associatedwith the *NON-WHITE* indicator, evaluated at the upper quartile, and $D(l)$ is the proportionate change evaluated at the lower quartile. The descriptive statistics are computed for the 32 periods with non-zero discrete hazard estimates.

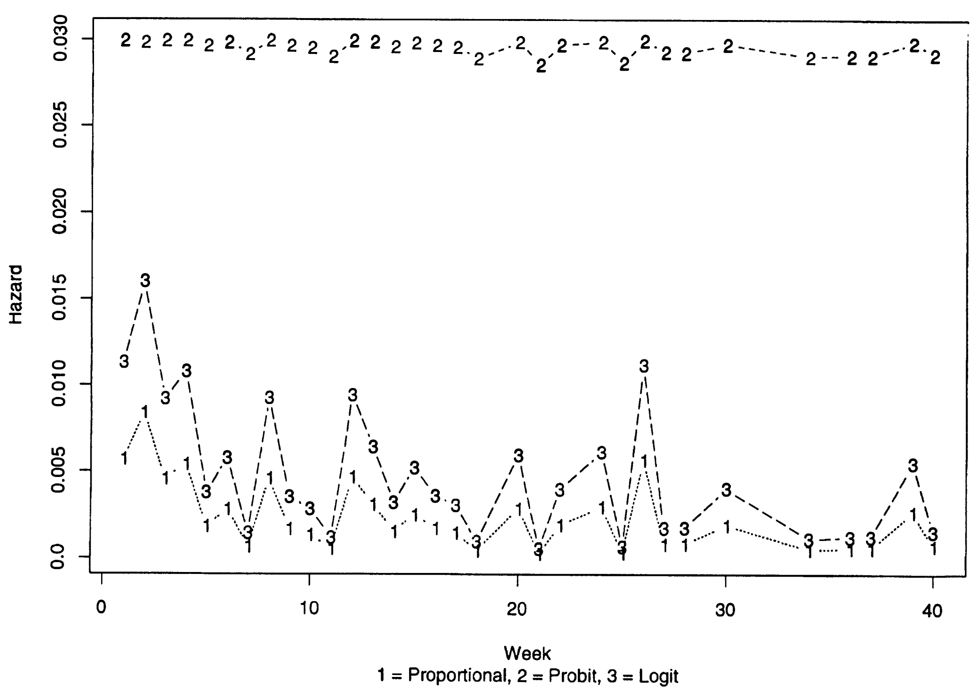

Figure 5. Difference between the upper and lower quartile estimates of the proportionate effect of race upon the discrete recall hazard of exit from unemployment; various specifications for the discrete hazard function F. The quartile differences are computed as $D(u) - D(l)$ where $D(u) = F(1, u)/F(0, u) - 1$ is the proportionate change in the discrete hazard associated with the *NON-WHITE* indicator, evaluated at the upper quartile, and $D(l)$ is proportionate change evaluated at the lower quartile

by changes in X. In contrast, the logit and proportional hazards models exhibit considerably smaller differences in the log discrete hazard across quartiles, but with greater changes in the time-variation. These basic results are reflected in the descriptive statistics presented in Table IV which show a standard deviation for the probit that is zero to four decimal places, but considerably larger for the alternative specifications.

6. CONCLUSION

In this paper I provide an explicit link between the estimation of binary response models for survival in a duration model setting, and the underlying hazards which generate responses with that structure. I show that the class of hazards which generates these models involves a set of interval-specific cumulative distribution functions F and within-interval functions of time h. This class of models is easy to specify, nests the proportional hazards, and probit and logit specifications, and allows for the familiar log-logistic and log-normal interval durations. Furthermore, the analysis provides a framework for a thorough analysis of the implications of various specifications for the interval survivor functions, as well as issues of identification of the within-interval hazard dynamics.

Along these lines, I show that the natural analogues of the PH model involve estimating pooled logit or probit models with period specific constant terms. These specifications have particular implications for duration behaviour, with the probit model, in particular, deviating from proportionality, but imposing approximately proportional effects of explanatory variables on the derivatives of the log discrete hazard. This result suggests that some care should be taken to investigate the assumptions embodied in a particular specification of the conditional exit probabilities. Plotting the derivative of the log discrete hazard against standardized values and examining the first and second derivatives should provide a useful guide to evaluating the implicit assumptions underlying the assumed specification.

The framework suggest that one may further generalize binary response models for duration analysis by allowing for interval-specific time-varying coefficients. In the most general form, this specification involves estimating a sequential binary response model with a full set of parameters for each period, but more restrictive forms of variation are allowed. For a PH model, the time-variation in coefficients constitutes a form of non-proportionality. For other choices of F, time-varying coefficients merely provide extra time-variation in the influence of X upon the log discrete hazards. In all cases, however, an alternative hypothesis involving full or limited time-variation in the β coefficients is readily testable using auxiliary regression techniques (Sueyoshi, 1991; Engle, 1984; Davidson and MacKinnon, 1984, 1990) to evaluate the Lagrange Multiplier tests. These tests require only information available under the null hypothesis, and are easy to compute using conventional econometric software. The auxiliary regressions will often provide useful information about the nature of departures from the null.

Finally, the close correspondence between grouped duration and binary response models suggests that there should be considerable gain to applying results from the discrete choice literature to the analysis of grouped duration models. The analysis in this paper suggests that much of the extensive and growing research on binary response specifications is directly applicable to a general class of hazard models.

REFERENCES

Chambers, J. M., W. S. Cleveland, B. Kleiner and P. A. Tukey (1983), *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole, Pacific Grove, CA.

Davidson, R. and J. G. MacKinnon (1984), 'Convenient specification tests for logit and probit models', *Journal of Econometrics*, **25**, 241–262.

Davidson, R. and J. G. MacKinnon (1990), 'Specification tests based on artificial regression', *Journal of the American Statistical Association*, **85**(409), 220–227.

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

Engle, R. F. (1984), 'Wald, likelihood ratio, and lagrange multiplier tests in econometrics', in Z. Griliches and M. D. Intriligator (eds), *Handbook of Econometrics*, Vol. 2 of *Handbook of Econometrics*, Chapter 13, 775–826, Elsevier Science, New York.
Han, A. and J. A. Hausman (1990), 'Flexible parametric estimation of duration and competing risk models', *Journal of Applied Econometrics*, **5**(1), 1–28.
Heckman, J. J. and B. Singer (1984), 'Econometric duration analysis', *Journal of Econometrics*, **24**(1/2), 63–132.
Johnson, N. L. and S. Kotz (1970), *Distributions in Statistics: Continuous Univariate Distributions*, Houghton Mifflin, Boston.
Kalbfleisch, J. D. and R. L. Prentice (1980), *The Statistical Analysis of Failure Time Data*, John Wiley, New York.
Katz, L. (1986), 'Layoffs, recall and the duration of unemployment', Technical Report 1825, National Bureau of Economic Research.
Kiefer, N. M. (1988a), 'Analysis of grouped duration data', *Contemporary Mathematics*, **80**.
Kiefer, N. M. (1988b), 'Economic duration data and hazard functions', *Journal of Economic Literature*, **26**(2), 646–679.
Lancaster, T. (1990), *The Econometric Analysis of Transition Data*, Cambridge University Press, Cambridge.
Meyer, B. D. (1986), 'Semi-parametric estimation of duration models', MIT working paper.
Narendranathan, W. and M. B. Stewart (1993), 'How does the benefit effect vary as unemployment spells lengthen?' *Journal of Applied Econometrics*, **8**(4), 361–381.
Nelson, W. (1982), *Applied Life Data Analysis*, John Wiley, New York.
Prentice, R. L. and L. A. Gloeckler (1978), 'Regression analysis of grouped survival data with application to breast cancer data', *Biometrics*, **34**, 57–67.
Sueyoshi, G. T. (1991), 'Evaluating simple alternatives to the proportional hazards model: unemployment insurance and the duration of unemployment', University of California, San Diego. February.
Sueyoshi, G. T. (1992), 'Semi-parametric proportional hazards estimation of competing risks models with time-varying covariates', *Journal of Econometrics*, **51**(1–2), 25–58.
Sueyoshi, G. T. (1994), 'Semiparametric estimation of generalized accelerated failure time models with grouped data', University of California, San Diego, May.
White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica*, **50**(1), 1–23.