A Hazard-Based Duration Model of Shopping Activity with Nonparametric Baseline Specification and Nonparametric Control for Unobserved Heterogeneity

Chandra R. Bhat

University of Massachusetts at Amherst

Abstract

Activity duration is an important component of the activity participation behavior of individuals, and therefore, an important determinant of individual travel behavior. In this paper, we examine the factors affecting shopping activity duration during the return home from work and develop a comprehensive methodological framework to estimate a stochastic hazard-based duration model from grouped (interval-level) failure data. The framework accommodates a nonparametric baseline hazard distribution and allows for nonparametric control of unobserved heterogeneity, while incorporating the effects of covariates. The framework also facilitates statistical testing of alternative parametric assumptions on the baseline hazard distribution and on the unobserved heterogeneity distribution. Our empirical results indicate significant effects of unobserved heterogeneity on shopping activity duration of individuals. Further, we find that parametric forms for the baseline hazard and unobserved heterogeneity distributions are inadequate, and are likely to lead to substantial biases in covariate effects and hazard dynamics. The empirical results also provide insights into the determinants of shopping activity duration during the commute trip.

1. Introduction

The need to view travel within the context of participation in activities has led to a human activity representation of travel in which travel forms part of a continuous pattern of daily (or some other unit of time, such as weekly) individual activity behavior (Bhat and Koppelman, 1993). The basic concept of this activity-based approach to travel is that travel is the means by which people change locations to participate in activities distributed in space.

The operationalization of the human activity approach requires the modeling of daily (or weekly) activity patterns. Such a modeling task is complex because of the many dimensions comprising an individual's activity pattern. These dimensions include the timing of activity, duration of activity, location of activity, mode of travel activity, and activity sequencing (*i.e.*, sequence of activity choices over time). Many research efforts have focused on one or more of these dimensions with the idea that a better understanding of the different individual dimensions will not only facilitate our efforts toward developing a comprehensive full-scale model of activity patterns, but will also provide useful insights into the nature of the impact of socio-demographic variables and time-space constraints on individual dimensions of activity behavior. For example, Damm (1980) has examined the timing of non-work activity. Van der Hoorn (1983) and Dunn and Wrigley (1985) have developed models for the choice of location of non-work activity participation. Uncles (1987) has examined the choice of travel mode for participation in shopping activity. The work of Kitamura and Kermanshah (1983), Golob (1986), and Nishii *et al.* (1988) has focused on understanding the mechanism by which individual non-work activities are chosen for participation and sequenced. Mannering and his colleagues (Mannering *et al.*, 1992; Kim and Mannering, 1992; Hamed and Mannering, 1993) have focused on non-work activity type choice and non-work activity duration (including home-stay duration and activity duration).

This study attempts to contribute to the literature on activity-based analysis by modeling shopping activity duration during the return home from work. Recent studies have indicated the increasing trend in non-work activity stops during the work tour. Gordon *et al.* (1988) report, based on their analysis of the 1990 US National Personal Transportation Survey (NPTS), that non-work travel is the major cause of the evening peak-period congestion and accounts for more than two-thirds of all evening peak-period trips. In an analysis of nonwork trips in the northern Virginia suburbs of the Washington, D.C. metropolitan area, Lockwood and Demetsky (1994) find that a significant number of individuals make one or more nonwork activity stops during their commute. Among such individuals, they observe that most make a single shopping activity stop during the work-to-home commute. These studies emphasize the importance of focusing on the activity pattern during the work commute in general, and on the dimensions characterizing shopping activity participation during the work-to-home commute in particular. Our focus on shopping activity duration is a reflection of such a need to better understand the individual dimensions of shopping activity behavior during the work-to-home commute. Among other things, the duration of shopping activity during the commute trip is an important determinant of the timing of travel and, therefore, of peak-period congestion.

We examine shopping activity duration using a hazard-based duration model in this paper. Hazard-based duration models are ideally suited to modeling duration data. Such models focus on an end-of-duration occurrence (such as end of shopping activity participation) given that the duration has lasted to some specified time (Kiefer, 1988; Hensher and Mannering, 1994). This concept of conditional probability of "failure" or termination of activity duration recognizes the dynamics of duration; that is, it recognizes that the likelihood of ending a shopping activity participation depends on the length of elapsed time since start of the activity.

In the next section, we provide an overview of duration models and highlight the important characteristics of the model developed in this paper. Section 3 develops the model structure and presents the estimation procedure. Section 4 discusses the data and the results of the empirical analysis. The final section provides a summary and identifies directions for future research.

2. Overview of Duration Models

Hazard-based duration models, which had their roots in biometrics and industrial engineering, are being increasingly used to model duration time in the fields of economics, transportation, and marketing (see Kiefer, 1988, Hensher and Mannering, 1994, and Jain and Vilcassim, 1991 for a review of the applications of duration models in economics, transportation, and marketing, respectively). To include an examination of covariates which affect duration time, most studies use a proportional hazard model which operates on the assumption that covariates act multiplicatively on some underlying or baseline hazard.

Two important specification issues in the proportional hazard model are a) the distributional assumptions regarding duration (equivalently, the distributional assumptions regarding the baseline hazard) and b) the assumptions about unobserved heterogeneity (*i.e.*, unobserved differences in duration across people). We discuss each of these issues in the following two sections.

2.1. Baseline hazard distribution

The distribution of the hazard may be assumed to be one of many parametric forms or may be assumed to be nonparametric. A serious problem with the parametric approach is that it inconsistently estimates the baseline hazard and the covariate effects when the assumed parametric form is incorrect (Meyer, 1990). In general, there is little theoretical support for any particular parametric shape. On the other hand, even if one uses a nonparametric baseline hazard when a particular parametric form is appropriate, the resulting estimates are consistent and the loss of efficiency (resulting from disregarding information about the hazard's distribution) may not be very substantial (Meyer, 1987). This strongly suggests the use (or at the least testing) of a nonparametric baseline in the estimation of duration models. However, most studies of duration to date have made an *a priori* assumption of a parametric hazard (some studies in the marketing literature have used general parametric forms which nest the more frequently used weibull, exponential and Gompertz distributions; see Jain and Vilcassim, 1991; Vilcassim and Jain, 1991).

Within the nonparametric approach, one may use the partial likelihood framework suggested by Cox (1972) which estimates the covariate effects but not the baseline hazard, or the approach suggested by Han and Hausman (1990) which estimates both the covariate effects and the baseline hazard parameters (also sometimes referred to as the incidental or nuisance parameters) simultaneously (the Han and Hausman approach is an alternative formulation of the approach originally proposed by Prentice and Gloeckler, 1978 and extended by Meyer, 1987). Between these approaches, the Han and Hausman (HH) approach has many advantages. First, in many studies, the dynamics of duration is itself of direct interest; the Cox approach, however, conditions out the nuisance parameters. Second, the Cox approach becomes cumbersome in the presence of many tied failure times (Kalbfleisch and Prentice, 198, page 101). Third, unobservable heterogeneity (which we discuss in the next section) cannot be accommodated within the Cox partial likelihood framework without the presence of multiple integrals of the same order as the number of observations in the risk set at each time period (see Han and Hausman, 1991 for a more detailed discussion). Estimation in the presence of such large orders of integration is impractical even with recent advances in the computation of multidimensional integrals.

It is clear that the nonparametric approaches may be more appropriate than the parametric approach in many situations and that within the nonparametric methods the HH approach is preferable to the Cox approach. In addition, the HH approach is the only appropriate method when duration models are to be estimated from interval-level data arising from the grouping of underlying continuous duration times. The parametric and Cox approaches use density function terms in their respective likelihood functions which are appropriate only for estimation from continuous duration data. If they are used to model grouped (or interval-level) duration data, the resulting estimates would generally be inconsistent (Prentice and Gloeckler, 1978). In the activity duration data used in this paper, most of the activity duration times are integral multiples of five minutes; almost 95% of all duration times end at a time which is an integral multiple of five minutes (*e.g.*, 5,10,15,30,60 minutes, *etc.*), leading to substantial number of ties at these times. For example, 45 individuals report terminating their activity at 5 minutes, 53 report termination at 10 minutes, 37 at 15 minutes, and so on, with extremely few reporting termination in between. This indicates that respondents tend to report the timing of their activities by rounding-off to the nearest five minute interval. Therefore, the activity duration data should be treated as interval-level data and we must use a discrete model (of the Han and Hausman type) that retains an interpretation as an incompletely observed continuous time hazard model.

The activity duration modeling efforts mentioned earlier (including another recent effort by Niemier and Morita, 1994) have ignored the interval-level nature of activity duration data and the resulting presence of many tied failure times. They have employed either the weibull parametric baseline hazard approach or the Cox nonparametric baseline hazard partial likelihood approach, both of which have conceptual deficiencies (as discussed above). More generally, all transportation-related applications of duration models have used either a parametric baseline hazard or the Cox nonparametric partial likelihood approach (Hensher and Mannering, 1994). No transportation or marketing-related study, at least to the author's knowledge, has estimated a non-parametric baseline hazard along with the estimates of the covariate effects. Even in the economics field, there have been only a handful of applications which have estimated a nonparametric hazard along with covariate effects (Han and Hausman, 1990; Meyer, 1990).

2.2. Unobserved heterogeneity

Unobserved heterogeneity arises when unobserved factors (*i.e.*, those not captured by the covariate effects) influence durations. It is well-established now 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; Lancaster, 1985).

The standard procedure used to control for unobserved heterogeneity is the random effects estimator (see Flinn and Heckman, 1982). This involves specification of a distribution for the unobserved heterogeneity (across individuals) in the population. Two general approaches may be used to specify the distribution of unobserved heterogeneity. One approach is to use a parametric distribution such as a gamma distribution or a normal distribution (most earlier research has used a gamma distribution). The problem with this parametric approach is that there is seldom any justification for choosing a particular distribution; further, the consequence of a choice of an incorrect distribution on the consistency of the model estimates can be severe (see Heckman and Singer, 1984). A second 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. 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. The nonparametric approach enables consistent estimation since it does not impose a prior probability distribution.

Application of duration models in the transportation field have, in most part, ignored unobserved heterogeneity. Those which have attempted to control for it have used a parametric distribution (see Hensher and Mannering, 1994). Researchers in the marketing and economics fields have paid more attention to unobserved heterogeneity. However, even in these fields, most applications have employed a parametric heterogeneity specification (see Gupta, 1991, Manston *et al.*, 1986, Meyer, 1990, Han and Hausman, 1990, all of whom use a gamma distribution). Very few studies have adopted a nonparametric heterogeneity distribution (see Jain and Vilcassim, 1991 and Vilcassim and Jain, 1991).

2.3. Specification of baseline hazard and unobserved heterogeneity in current paper

In this paper, we use a nonparametric baseline hazard based on the Han and Hausman (1990) approach and a nonparametric unobserved heterogeneity specification based on the Heckman and Singer approach (1984). The author is not aware of any previous study which has used such a nonparametric specification for both the baseline hazard as well as the heterogeneity distribution. Meyer (1990) and Han and Hausman (1990) use a nonparametric baseline hazard specification and indicate that one could use a nonparametric unobserved heterogeneity distribution, but use a gamma distribution during estimation since it provides a convenient closed-form expression for the likelihood. They also suggest that the choice of the heterogeneity distribution may be unimportant when the baseline hazard is nonparametrically estimated and speculate that the finding of Heckman and Singer (1984) that parametric heterogeneity approaches provide inconsistent covariate effects is due to their assumption of a parametric baseline hazard. Jain and Vilcassim (1991), Vilcassim and Jain (1991), Hensher (1994), and Heckman and Singer (1984), on the other hand, use a parametric baseline hazard with a nonparametric heterogeneity distribution. Their results indicate that the covariate effects are sensitive to the specification of the heterogeneity distribution and that a nonparametric heterogeneity distribution is the best in terms of overall fit and reasonableness of covariate effects.

By allowing a nonparametric distribution for both the baseline hazard and unobserved heterogeneity, our paper sheds light on the importance of allowing a nonparametric specification for the baseline hazard, for unobserved heterogeneity, and for both of these (as indicated by Hensher and Mannering, 1994, "There is considerable debate in the literature as to whether the baseline hazard or the mixture distribution should be nonparametric"; our paper attempts to investigate this important issue). The paper estimates and compares six different models with different specifications for the hazard and heterogeneity distributions. The first two models do not accommodate heterogeneity, the second two use a gamma distribution for heterogeneity, and the final two estimate nonparametric heterogeneity distributions. In each of these three classes of models, the first model specifies a weibull baseline and the second uses a nonparametric baseline (we use a weibull form for the parametric baseline hazard and a gamma distribution for parametric heterogeneity, since these have been the most commonly used parametric specifications).

3. Model Structure and Estimation

Let T_i represent the continuous activity duration time for individual i (we consider the time unit of the continuous scale to be in minutes). This continuous duration is not observed; instead, we only observe discrete time intervals in which failure (*i.e.*, end of participation in shopping activity) occurs. Let u represent some specified time on the continuous time scale and 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 $u \in [u^{K\text{-}1},\infty]$. Let t_i represent the discrete period of failure 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 (that is, how the elapsed time since start of the shopping activity impacts the future termination of the activity) and the effect of covariates (or exogenous variables) on the continuous activity duration time. We assume that the covariates do not change with time (this is a reasonable assumption in the short-term context of examining shopping activity duration on the return home from work; see Sueyoshi, 1992 for an extension to the case of time-varying covariates).

The modeling methodology is discussed in the subsequent four sections. Section 3.1 presents the models (with weibull and nonparametric baseline hazard distributions) assuming no heterogeneity. Section 3.2 focuses on the case when unobserved heterogeneity is incorporated, but is assumed to take a gamma parametric form. Section 3.3 considers the models with a nonparametric form of heterogeneity. Finally, section 3.4 discusses the procedure to estimate the baseline hazard function from the model estimates.

3.1. Models with no heterogeneity

The hazard for individual *i* at some specified time u on the continuous-time scale, $\lambda_i(u)$, is defined using the proportional hazard specification as (see Kiefer, 1988):

$$\lambda_i(u) = \lim_{\delta\to 0^+} \frac{\text{prob}[u+\delta > T_i \geq u \mid T_i \geq u]}{\delta} = \lambda_0(u)\exp(-\beta' x_i) \qquad (1)$$

where $\lambda_0(u)$ is the baseline hazard (to be estimated) at time *u*, x_i is a column vector of covariates for individual *i*, and β is a column vector of parameters (to be estimated). It is easy to show that equation (1) can be written in the equivalent form,

$$\ln \Lambda_0(T_i) = \ln \int_0^{T_i} \lambda_0(u)du = \beta' x_i + \varepsilon_i \qquad (2)$$

where ε_i takes an extreme value form with distribution function given by:

$$\text{prob}(\varepsilon_i < z) = G(z) = 1 - \exp[-\exp(z)]. \qquad (3)$$

The dependent variable in equation (2) is a continuous unobserved variable. However, we do observe the discrete time period, t_i, in which individual *i* ends her/his shopping participation. Defining u^k as the continuous time value representing the upper bound of discrete time period *k*, we can write:

$$\begin{aligned} \text{prob}[t_i = k] &= \text{prob}[u^{k-1} < T_i \leq u^k] = \text{prob}[\ln \Lambda_0(u^{k-1}) < \ln \Lambda_0(T_i) \leq \ln \Lambda_0(u^k)] \\ &= G(\delta_k - \beta' x_i) - G(\delta_{k-1} - \beta' x_i) \text{ from (2) and (3), where } \delta_k = \ln \Lambda_0(u^k). \end{aligned} \qquad (4)$$

The parameters to be estimated in the nonparametric baseline model are the (K-1) δ parameters ($\delta_0 = -\infty$ and $\delta_K = +\infty$) and the vector β. Defining a set of dummy variables

$$M_{ik} = \begin{cases} 1 \text{ if failure occurs in period } k \text{ for individual } i \\ 0 \text{ otherwise} \end{cases} \qquad (i = 1,2,...N,\ k = 1,2,...K) \tag{5}$$

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

$$\mathfrak{L} = \prod_{i=1}^{N}\prod_{k=1}^{K}[G(\delta_k - \beta' x_i) - G(\delta_{k-1} - \beta' x_i)]^{M_{ik}}\ . \tag{6}$$

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. In the case of shopping activity duration, there is no right censoring because all individuals end their shopping participation.

The discrete model discussed above is the uniquely appropriate one for analysis using grouped (interval-level) data based on the continuous proportional hazard model of equation (1).

A weibull assumption about the baseline hazard essentially places restrictions on the δ parameters in the nonparametric model. Specifically, theδ parameters are now characterized by the two weibull parameters as follows:

$$\delta_k = p\ln(\alpha u^k);\, p, \alpha > 0 \tag{7}$$

where u^k is as defined earlier. The likelihood function of equation (6) is then maximized in the usual way after imposing the constraints in equation (7) to obtain estimates of the weibull parameters and the β vector. Standard likelihood ratio tests can be used to compare the weibull baseline hazard assumption against the nonparametric baseline hazard.

3.2. Models with gamma heterogeneity

We introduce unobserved heterogeneity into the model specification by assuming a multiplicative form:

$$\lambda_i(u) = \lambda_0(u)\exp(-\beta' x_i + w_i) \tag{8}$$

where w_i represents the unobserved heterogeneity. Then, as in equation (4),

$$\begin{aligned} \text{prob}[t_i = k \mid w_i] &= G(\delta_k - \beta' x_i + w_i) - G(\delta_{k-1} - \beta' x_i + w_i) \\ &= \exp[\text{-}\{I_{i,k-1}\exp(w_i)\}] - \exp[-\{I_{i,k}\exp(w_i)\}] \end{aligned} \tag{9}$$

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 one (a normalization) and variance σ^2, the unconditional probability of failure in the discrete time period *k* can be expressed as:

$$\text{prob}[t_i = k] = \int_0^\infty \left(\exp[-\{I_{i,k-1}v_i\}] - \exp[-\{I_{i,k}v_i\}]\right) f(v_i)\,dv_i\,. \tag{10}$$

Using the moment-generating function properties of the gamma distribution (see Johnson and Kotz, 1978), the expression above reduces to:

$$\text{prob}[t_i = k] = [1 + \sigma^2 I_{i,k-1}]^{\text{-}\sigma^{-2}} - [1 + \sigma^2 I_{i,k}]^{\text{-}\sigma^{-2}} \tag{11}$$

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

$$\mathfrak{L} = \prod_{i=1}^{N}\prod_{k=1}^{K}\left\{\left[1 + \sigma^2 I_{i,k-1}\right]^{-\sigma^{-2}} - \left[1 + \sigma^2 I_{i,k}\right]^{-\sigma^{-2}}\right\}^{M_{ik}}. \tag{12}$$

In the case of a weibull baseline hazard model with gamma heterogeneity, the likelihood function (12) is maximized after imposing the constraints in equation (7).

The models with and without gamma heterogeneity for the weibull and nonparametric baseline hazard models can be compared using standard likelihood ratio tests. This is because it can be shown (the proof is available from the author) that as $\sigma^2 \to 0$ in equation (12), the likelihood function collapses to that in equation (6). Alternatively, one can use the asymptotic *t*-test to examine if the variance of the gamma distribution is significantly different from zero.

3.3. Models with nonparametric unobserved heterogeneity

To formulate the model with nonparametric unobserved heterogeneity, we start again with equation (4) where w_i is now nonparametrically distributed. Then, as earlier, we can write

$$\text{prob}[t_i = k \mid w_i] = G(\delta_k - \beta' x_i + w_i) - G(\delta_{k-1} - \beta' x_i + w_i)\,. \tag{13}$$

We now 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* failing in period *t* is

$$\text{prob}[t_i = k] = \sum_{s=1}^{S} \{[G(\delta_k - \beta' x_i + l_s) - G(\delta_{k-1} - \beta' x_i + l_s)]\pi_s\} \tag{14}$$

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:

$$\mathfrak{L} = \prod_{i=1}^{N} \left\{ \sum_{s=1}^{S} \left[\left\{ \prod_{k=1}^{K} [G(\delta_k - \beta' x_i + l_s) - G(\delta_{k-1} - \beta' x_i + l_s)]^{M_{ik}} \right\} \pi_s \right] \right\}\,. \tag{15}$$

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

$$E(w_i) = \sum_{s=1}^{S} \pi_s l_s = 0\,. \tag{16}$$

Our estimation procedure ensures that the cumulative mass over all support points sum to one.

One critical quantity in empirical estimation of the distribution of unobserved heterogeneity is the number of support points, *S*, required to approximate the underlying distribution. This number is determined by using a stopping-rule procedure based on the Bayesian information criterion (see Allenby, 1990), which is defined as follows:

$$BIC = -\ln(\mathcal{L}) + 0.5 \cdot R \cdot \ln(N) \tag{17}$$

where the first term on the right 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*.

The weibull baseline model with nonparametric heterogeneity is estimated in the usual way by maximizing the likelihood function in (15) after imposing the constraints in equation (7).

The models with nonparametric heterogeneity and no heterogeneity for each type of baseline hazard (*i.e.*, weibull and nonparametric) can be compared using a likelihood ratio test. The degrees of freedom for this test is equal to 2 (*S*-1) where *S* is the number of support points in the nonparametric heterogeneity model. The models with nonparametric and gamma heterogeneity are non-nested, but can be compared using a non-nested test.

3.4. Baseline hazard estimation

Two different time scales may be used for the estimation of the baseline hazard. The first time scale is the discrete time periods in which failure data is grouped. The second time scale is the continuous time scale. In either case, the baseline hazard estimation is consistent with the continuous-time proportional hazard model of equation 1 (and its counterparts for the gamma and nonparametric unobserved heterogeneity distribution). Adopting the discrete time period does not require any additional assumptions other than the functional form of the continuous-time proportional hazard model; on the other hand, one must make an additional assumption about the within-period dynamics of the shopping duration hazard for estimation of a continuous-time baseline hazard from the nonparametric baseline models (see Sueyoshi, 1992). In this paper, we specify a constant hazard rate within the discrete intervals (or periods) to estimate a continuous-time baseline hazard. We now discuss the baseline hazard estimation for both the discrete time period case and for the continuous-time case.

If interest centers around the prediction of duration times at the discrete-period level, then we can compute the discrete-period baseline hazard for period *k*, $\lambda_0^*(k)$, using

$$\lambda_0^*(k) = \text{prob}\{u \in [u^{k-1}, u^k] \mid u \geq u^{k-i}\} \frac{prob\{t = k\}}{prob\{t > k-1\}} = \frac{G(\delta_k) - G(\delta_{k-1})}{1 - G(\delta_{k-1})} \qquad (18)$$

where the $\delta_k{}'s$ are estimated directly for the models with no heterogeneity and nonparametric heterogeneity, and are estimated from the estimates of $\Lambda_0(u^k)$ for the models with gamma heterogeneity. The continuous-time proportional hazard assumption of equation (1) for the (weibull and nonparametric baseline) models without heterogeneity and of equation (8) for the models with heterogeneity translate to the following relationships between the discrete-period baseline hazard and the discrete-period hazard at time *k* for individual *i*, $\lambda_1^*(k)$:

$$\begin{aligned} \lambda_i^*(k) &= 1 - [1 - \lambda_0^*(k)]^{\exp(-\beta' x_i)} && \text{for the models with no heterogeneity} \\ &= 1 - \left[1 - \frac{\ln[1 - \lambda_0^*(k)]\exp(-\beta' x_i)}{\alpha}\right]^{-\alpha} && \text{for the models with gamma heterogeneity} \qquad (19) \\ &= \sum_{s=1}^{S} \left[\left\{1 - [1 - \lambda_0^*(k)]^{\exp(-\beta' x_i + l_s)}\right\}\pi_s\right] && \text{for the models with nonparametric heterogeneity.} \end{aligned}$$

The baseline hazard function for the case of a continuous-time scale is obtained directly from the estimates of α and P in the weibull baseline models as $\hat{\lambda}_0(u) = \hat{\alpha}\hat{P}(\hat{\alpha}u)^{\hat{P}-1}$. The continuous-time baseline hazard function in the nonparametric baseline models is estimated by assuming that the hazard remains constant within each time period *k*; that is, $\lambda_0(u) = \lambda_0(k)$ for all $u \in \{u^{k-1}, u^k\}$. Then, we can write:

$$\hat{\lambda}_0(k) = \frac{\exp(\hat{\delta}_k) - \exp(\hat{\delta}_{k-1})}{\Delta u^k} = -\frac{\ln[1 - \hat{\lambda}_0^*(k)]}{\Delta u^k},\ k = 1, 2, ..., K-1 \tag{20}$$

where Δu^k is the length of the time interval *k* and $\hat{\lambda}_0^*(k)$ is the estimate of the discrete-period hazard in period *k* computed from equation (19).

4. Data and Empirical Results

The data source used in the present study is a household activity survey conducted in April of 1991 by the Central Transportation Planning Staff (CTPS) in the Boston Metropolitan region. The survey collected data on socio-demographic characteristics of the household and each individual in the household. The survey also included a one-day (mid-week working day) activity diary to be filled out by all members of the household above five years of age. Each activity was described by: (a) start time, (b) stop time, (c) location of activity participation, (d) travel time from previous activity, (e) travel mode to activity location, and (f) activity type.

The sample for the current analysis comprises 355 employed adult individuals who made a work-trip on the diary day and who made a single shopping activity stop during the return home from work. Table 1 provides descriptive information on shopping activity duration. The column labeled "Failures" indicates the number of individuals whose activity participations end in discrete time period *k*. The column titled "No. at Risk" provides information on the number of individuals who are "at risk" of termination of their activity participation in period *k*; that is, it is the number of individuals whose activity durations have not ended at the beginning of period *k*. The discrete-period sample hazard rates associated with each period is computed using the Kaplan-Meier (KM) nonparametric estimator as the number of individuals who end their activity participations in period k divided by the risk set in period *k*.

The choice of variables for potential inclusion in the model was guided by previous theoretical and empirical work on activity modeling (see Chapin 1974, Steinberg *et al*. 1980, and other studies reviewed in the first section of the paper) and intuitive arguments regarding the effect of exogenous variables on shopping activity duration. Table 2 provides a list of exogenous variables used in the model and their definitions. An important specification improvement in the current paper over earlier activity duration modeling efforts (*e.g.*, Niemeier and Morita, 1994; Hamed and Mannering, 1993; and Damm, 1980) is the consideration of interactions within individuals in a household. Specifically, we include variables which are likely to be important determinants of the allocation of responsibility for shopping activity between an individual and her/his spouse. These variables are spouse's employment status, spouse's work duration, spouse's travel time to work and spouse's mode of travel to work.

The summary statistics for the six different models discussed in section 2.3 are presented in Table 3. The value of the log-likelihood function with baseline parameters only corresponds to the case when no covariates are included and when unobserved heterogeneity is ignored (that is, when only sample information on temporal dynamics is used in the hazard model). The value of the log-likelihood function at zero is the same for all models and corresponds to the case when the hazard is constant over time (that is, when no sample information on temporal dynamics is used). The final row presents the adjusted likelihood ratio index defined as:

$$\overline{p}^2 = 1 - \frac{\text{log likelihood at convergance - number of parameters estimated in the model}}{\text{log likelihood at zero}}. \quad (21)$$

We now discuss the estimation results in detail by first focusing on the baseline hazard, next on unobserved heterogeneity, and finally on the covariate effects.

4.1. Baseline hazard

Our discussion of the baseline hazard will be based on the continuous-time scale. The substantive conclusions from the discrete-period baseline hazard estimates are similar to those from the continuous-time scale.

The weibull baseline hazard function is characterized by the parameters α and *P*. The baseline hazard functions corresponding to the weibull baseline models are plotted in the first column of Fig. 1. Ignoring unobserved heterogeneity leads to a hazard function that is near-exponential with a constant hazard of about 0.0095. This implies that there is no duration dependence; in other words, the conditional probability of ending the shopping activity participation is the same regardless of how long an individual has been shopping. However, controlling for unobserved heterogeneity, the baseline hazard has a significant positive duration dependence. Thus, the longer an individual has been shopping, the more likely that the participation will terminate. Between the weibull models with gamma and nonparametric heterogeneity, the model with gamma heterogeneity has a steeper positive duration dependence.

A disadvantage of the weibull specification is that it specifies a monotonicity restriction on the hazard as well as a particular parametric form of duration dependence. Using a nonparametric approach to estimate the baseline hazard function avoids these problems. The nonparametric estimates of the baseline hazard are plotted in the second column of Fig. 1. As in the weibull baseline case, assumptions regarding heterogeneity have an effect on the nonparametric baseline hazard estimates. The models with heterogeneity have substantially higher baseline estimates compared to the model with no heterogeneity (in fact, the baseline hazard when heterogeneity is ignored is relatively flat compared to the case when heterogeneity is incorporated). These results again conform to Kiefer's general result that ignoring heterogeneity leads to a downward biased estimate of duration dependence. Between the two models with heterogeneity, the nonparametric heterogeneity model tends to have slightly lower values of the hazard till about 110 minutes after which the hazard increases dramatically relative to the model with gamma heterogeneity (the hazard rate rises sharply to 9.79 for all *T* between 132.5 minutes and 152.5 minutes and to 17.10 for all *T* between 152.5 minutes and 212.5 minutes; these estimates are not shown on the plot because of their high values relative to the scale adopted). The results clearly show the serious biases in the baseline hazard shape that could result from ignoring heterogeneity or using an inappropriate parametric heterogeneity specification, even if a nonparametric baseline hazard is used (we test the different heterogeneity specifications using formal statistics in the next section).

We now turn to a comparison of the weibull baseline and nonparametric baseline shapes. The nonparametric baseline estimates indicate substantial variation in hazard rates over time. For all three heterogeneity specifications, there is clear evidence that the baseline hazard is non-monotonic; in particular, there are periods of both increases and decreases. The spikes in the nonparametric baseline correspond approximately to durations of 30, 45, 60 and 90 minutes. Beyond a duration of about 2 hours, the nonparametric hazard increases considerably in the gamma and nonparametric heterogeneity cases. The weibull baseline underestimates the hazard rather substantially at almost all values of shopping activity duration for the gamma and nonparametric heterogeneity specifications.

Likelihood ratio tests between the weibull and nonparametric baseline models for each of the unobserved heterogeneity specifications clearly reject the null hypothesis of a weibull baseline. The results also indicate that accommodating heterogeneity does not alleviate the problems caused by assumption of an incorrect hazard distribution.

4.2. Unobserved heterogeneity

In this section, we compare the models with different heterogeneity specifications within the weibull baseline specification and the nonparametric baseline specification.

The model with gamma heterogeneity can be compared with the model with no heterogeneity for each of the two baseline specifications by examining the significance of the variance parameter σ^2.This parameter is estimated to be 0.732 (with a standard error of 0.14) in the weibull baseline case and 0.950 (with a standard error of 0.29) in the nonparametric baseline case. Thus, for both the weibull and nonparametric baseline specifications, this parameter is significantly different from zero rejecting the null of no heterogeneity.

In estimating a nonparametric form for heterogeneity, we found that two support points were sufficient to approximate the underlying distribution for the weibull baseline. For the nonparametric baseline, three support points were needed. The estimated support points for the weibull baseline case were -1.59 and 0.59 with associated probability masses of 0.34 and 0.66, respectively; likewise, the support points for the nonparametric baseline case were -7.46, -0.99, and 0.75 with associated probability masses of 0.04, 0.24, and 0.72, respectively. These distributions indicate asymmetry with respect to the expected value of zero. The variance of the nonparametric heterogeneity distribution in the weibull baseline case is 0.68; interestingly, this is close to the variance obtained (= 0.732) using a gamma distribution in the weibull baseline case.

The variance of the nonparametric heterogeneity distribution in the nonparametric baseline specification is 2.88, which is much larger than the corresponding gamma heterogeneity variance of 0.95. These findings suggest that the use of a parametric baseline hazard or a parametric heterogeneity distribution or both tends to underestimate unobserved heterogeneity.

The nonparametric heterogeneity specification can be formally compared with the no-heterogeneity specification using likelihood ratio tests. These tests reject the model with no heterogeneity for both the weibull and nonparametric baseline specifications. The gamma heterogeneity and the nonparametric heterogeneity specifications cannot be compared using standard likelihood ratio tests since they are mutually non-nested. Ben-Akiva and Lerman (1985) suggest the use of a non-nested test for discrete-choice models based on the likelihood ratio index (since our formulation of the duration models takes a discrete-choice form, this test is appropriate here). The test statistic for this non-nested test is:

$$\text{prob}(\bar{p}_2^2 - \bar{p}_1^2 > \tau) \leq \Phi\{-[-2\tau \ln\mathcal{L}(0) + (\theta_2 - \theta_1)]^{0.5}\}, \tau > 0 \qquad (22)$$

where $\bar{p}_1^2$ and $\bar{p}_2^2$ are the adjusted likelihood ratio index values for each of the two models under consideration, θ_1 and θ_2 are the number of parameters estimated in the two models and Φ represents the cumulative standard normal distribution. In our case, the adjusted likelihood ratio index values are 0.0997 and 0.0999 for the weibull baseline with gamma and nonparametric heterogeneity, respectively, and 0.1209 and 0.1233 for the nonparametric baseline with gamma and nonparametric heterogeneity, respectively. Application of equation (22) indicates that the probability that a difference of 0.0002 in $\bar{p}^2$ (between the nonparametric and gamma heterogeneity specifications) for the weibull baseline could have occurred by chance is less than $\Phi(-1.16) = 0.123$. The corresponding figure for the difference of 0.0024 in $\bar{p}^2$ for the nonparametric baseline is $\Phi(-2.83) = 0.002$. Thus, we can conclude that the nonparametric heterogeneity specification is the preferred specification for the nonparametric baseline. However, in the case of the weibull baseline, the nonparametric heterogeneity specification is not significantly better than the gamma specification (at the 0.1 significance level).

The finding above that the nonparametric heterogeneity specification is the appropriate one for the nonparametric baseline specification, combined with the finding in the previous section that the nonparametric specification is the most suitable one for the baseline hazard, indicates that, at least in the context of the current empirical analysis, the nonparametric baseline-nonparametric unobserved heterogeneity specification is preferable to other specifications. This result is important. It is contrary to the commonly held view that the choice of the mixing distribution may not be important if the baseline hazard is nonparametrically specified (see Meyer, 1990; Han and Hausman, 1990; Manston *et al.*, 1986). The results suggest that specifying a nonparametric baseline and a nonparametric unobserved heterogeneity distribution may both be important and cautions against resorting to parametric heterogeneity specifications by appealing to the nonparametric specification of the baseline hazard.

4.3. Covariate Effects

In this section, we discuss the effects of covariates on the duration hazard. It should be observed that a positive coefficient on a covariate implies that the covariate increases shopping activity duration (equation 2) or, equivalently, the covariate lowers the hazard rate (equation 1).

Table 4 shows the estimated covariate effects for the alternative model specifications. We include four sets of covariates: Individual's work characteristics, Spouse's work characteristics, mode to work of individual and spouse, and socio-demographic variables. The different model specifications provide similar results with regard to the sign and significance of the covariate effects. For example, work duration has a positive and significant effect on the hazard; departure from work before 4 pm has a negative effect; travel time to work is insignificant in all models. We also find consistency in the direction of the effect of the other variables. However, the sensitivity of the hazard to the covariates is influenced by the assumptions made about the baseline hazard and unobserved heterogeneity. For instance, the duration hazard is much less sensitive to the work duration of individuals in the nonparametric (NP) baseline - NP heterogeneity specification than in other specifications. The covariate "Departure from work before 4 pm" has a higher impact on the duration hazard in the NP baseline models with gamma and NP heterogeneity. The effects of the spouse's work duration and travel time to work variables are different among the alternative heterogeneity specifications. In particular, the NP baseline-NP heterogeneity specification indicates that the effect of spouse's work duration on an individual's hazard is higher when the individual is a female (spouse is a male) than when the individual is a male (spouse is female). The magnitudes of these effects are reversed in all other specifications. A similar result holds for the spouse's travel time to work variable.

The results discussed above suggest that parametric assumptions about the baseline hazard distribution and unobserved heterogeneity distribution have biased the covariate estimates. As noted earlier, the nonparametric baseline-nonparametric unobserved heterogeneity specification (Model 6) is the most preferred specification and rejects all other model specifications in formal statistical tests. We now examine and discuss the effect of covariates in greater detail, focusing on this nonparametric baseline-nonparametric unobserved heterogeneity specification.

Among the variables representing an individual's work characteristics, work duration and departure from work before 4 pm have significant impacts on the shopping activity duration hazard. The effect of these variables is as expected. A longer duration at work leaves less time for participation in other activities after work and, therefore, decreases shopping activity duration (increases the activity duration hazard) after work. Departure from work before 4:00 pm provides more time and opportunity to participate in shopping activity after work and therefore increases shopping duration. The effect of travel time to work is not significant.

Three variables; employment status of spouse, work duration of spouse, and travel time to work of spouse; are included to represent the impact of spouse's work characteristics on an individual's shopping duration (these variables are relevant only for married individuals). The empirical results show that if an individual's spouse is employed, the individual's shopping duration increases. Spouse's employment is likely to lead to a more equitable allocation of shopping activities between the individual and her/his spouse, thereby increasing the involvement of the individual in shopping. Further, the individual's shopping duration tends to increase as her/his spouse's travel time to work and work duration increase. This is an expected result; as the work commitments of an individual's spouse increases, constraints on the spouse's time also increase, and the individual contributes greater amounts of her/his time to household maintenance activities such as shopping. The results, however, indicate differences in this contribution based on the sex of the individual and spouse. The husband's duration hazard is less sensitive to his wife's travel time and work duration than is the wife's duration hazard to her husband's travel time and work duration. It is particularly interesting to note that the husband's hazard is about equally sensitive to his work hours (see coefficient on "work duration" under individual's work characteristics) and his spouse's work hours (see coefficient on "female spouse" under spouse's work duration), while the wife's duration hazard is more sensitive to her husband's work hours than to hers. We did not find any differences in the duration hazard between males and females who are unmarried or who are married, but whose spouses do not work.

The mode to work variables indicate that individuals who drive alone to work have shorter shopping activity durations compared to individuals who rideshare or use transit. However, if an individual drives alone and her/his spouse rideshares or uses transit (relevant only for married individuals with an employed spouse), then there is an additional positive effect on the individual's shopping duration. The net effect for such individuals is obtained by adding the coefficients on the two mode-related variables. This net effect is positive (0.961-0.817 = 0.144), but insignificant.

Among the socio-demographic variables, the only significant one is the "returning young adult" variable. The parameter on this variable indicates that returning young adults tend to participate for longer periods in shopping activity during the return home from work compared to other individuals.

5. Conclusions and Directions for Future Research

This paper provides a unified methodological framework to estimate a proportional hazard duration model with a nonparametric baseline hazard distribution and a non-parametric unobserved heterogeneity distribution from grouped (interval-level) data, while incorporating the effects of covariates. The framework facilitates the comparison of alternative parametric specifications of the baseline hazard and of unobserved heterogeneity against the nonparametric specification using statistical tests. The methodological framework is applied to an analysis of individual shopping activity duration during the return home from work. Such an analysis contributes to a better understanding of the activity pattern of individuals during the work-to-home trip.

A number of important findings emerge from our empirical analysis. First, the results indicate that parametric baseline forms may provide biased estimates of duration dependence. In the current analysis, we find that the duration dependence is considerably underestimated when a weibull baseline form is used (as a matter of future research, it would be useful to examine if a similar result holds in other empirical applications). Second, the results suggest that accommodating heterogeneity (in either a parametric or a nonparametric form) does not alleviate the bias in duration dynamics resulting from an incorrect assumption of the hazard distribution. Third, we find that ignoring heterogeneity results in a substantial underestimation of duration dependence (an observation also made by Kiefer, 1988). Fourth, our results indicate the persistence of bias in duration dependence if a parametric form of heterogeneity is used (this result is consistent with the results of Heckman and Singer, 1985). This holds true even if the baseline hazard is nonparametrically specified. Thus, the result suggests that the form of the unobserved heterogeneity distribution may be important even when a flexible nonparametric baseline specification is used and cautions against ignoring heterogeneity or resorting to a parametric heterogeneity specification by appealing to the nonparametric baseline specification. Fifth, we find that the use of a weibull baseline hazard with a nonparametric heterogeneity distribution, or a gamma heterogeneity distribution with a nonparametric baseline hazard, or a weibull hazard with a gamma heterogeneity distribution tends to underestimate unobserved heterogeneity. This finding cautions against inferring about the presence and magnitude of unobserved heterogeneity if a parametric baseline form and/or a parametric heterogeneity distribution is used. Finally, we find that there are differences in covariate effects based on the specification of the baseline hazard and heterogeneity distributions. It should be emphasized that all the foregoing findings are specific to the empirical analysis in this paper; the results may be different in other data samples. Also, there is the possibility that the findings are a consequence of the variable specification adopted. However, a general result is that it is always preferable (unless there is a strong theoretical basis otherwise) to estimate a nonparametric baseline-nonparametric heterogeneity specification and test parametric specifications as special cases of this more general formulation rather than make an *a priori* assumption of a particular parametric form).

The results of this research also provide insights into the determinants of shopping activity duration during the commute trip. An individual's shopping duration is affected by the work characteristics of the individual, the work characteristics of the individual's spouse (if individual is married and spouse is employed), the mode to work used by the individual and her/his spouse (if individual is married), and whether the individual is a returning young adult or not. Our analysis shows the important influence of the spouse's work characteristics and the mode to work used by the individual *vis-a-vis* the mode used by the individual's spouse.

A number of future research directions may be identified based on this research. An important extension of the current model is to analyze the choice of the decision to participate in shopping activity along with a hazard model of shopping duration to accommodate any sample selection in activity duration based on the choice to participate in shopping activity. More generally, it would be useful to examine the choice of participation in different activities (shopping, recreation, social, or go home directly from work) using a discrete-choice model along with the hazard duration model of participation in the different out-of-home activities. Other extensions include consideration of other dimensions of activity participation, modeling the activity behavior during the commute trip as a component of a broader daily activity participation decision-making process of individuals, and accommodating inter-personal interactions among household members in activity decisions (for example, work and shopping participation and duration decisions of an individual and her/his spouse may be jointly determined).

Acknowledgement

The author would like to thank Mr. Ian Harrington at the Central Planning Transportation Staff Office in Boston for making the Boston household activity survey data available for this research. Prof. Frank Haight and three anonymous referees provided useful comments on an earlier draft of the paper.

References

Allenby, G.M. (1990). Hypothesis testing with scanner data: the advantage of bayesian methods, Journal of Marketing Research, 27-379-389.

Ben-Akiva, M. and S.R. Lerman (1985). Discrete Choice Analysis: Theory and Application to Travel demand, The MIT Press, Cambridge.

Bhat, C.R. and F.S. Koppelman (1993). A conceptual framework of individual activity program generation, Transportation Research, 27A, 6, 433-446.

Chapin, F.S. (1974). Human Activity Patterns in the City, John Wiley & Sons, New York.

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

Damm, D. (1980). Interdependencies in activity behavior, Transportation Research Record, 750, 33-40.

Dunn, R. and N. Wrigley (1985). Beta-logistic models of urban shopping center choice, Geographic Analysis, 17, 95-113.

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

Golob, T.F. (1986). A non-linear canonical correlation analysis of weekly chaining behavior, Transportation Research, 20A, 385-399.

Gordon, P., A. Kumar, and H.W. Richardson (1988). Beyond the journey to work, Transportation Research, 22A, 6, 419-426.

Gupta, (1991). Stochastic models of interpurchase time with time-dependent covariates, Journal of Marketing Research, 28, 1-15.

Hamed, M.M and F.L. Mannering (1993). Modeling travelers' postwork activity involvement: toward a new methodology, Transportation Science, 27, 4, 381-394.

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.

Hoorn, T. van der (1983). Development of an activity model using a one-week activity-diary data base, in S. Carpenter & P. Jones (eds.), Recent Advances in Travel Demand Analysis, 335-349, Gower, Aldershot, England.

Jain, D.C. and N.J. Vilcassim (1991). Investigating household purchase timing decisions: a conditional hazard function approach, Marketing Science, 10, 1, 1-23.

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

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

Kim, S.G. and F. Mannering (1992). Panel data and activity duration models: econometric alternatives and applications, Paper prepared for the First US Conference on Panels for Transportation Planning, Lake Arrowhead, California.

Kitamura, R. and M. Kermanshah (1983). Identifying time and history dependencies of activity choice, Transportation Research Record, 944, 22-30.

Lancaster, T. (1985). Generalized residuals and heterogenous duration models with applications to the weibull model, Journal of Econometrics, 28, 1, 155-169.

Lockwood, P.B. and M.J. Demetsky (1994). Nonwork travel - a study of changing behavior, presented at the 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., January.

Mannering, F., E. Murakami and S.G. Kim (1992). Models of traveler's activity choice and home-stay duration: analysis of functional form and temporal stability, submitted to Transportation.

Manston, K.G., E. Stallard, and J.W. Vaupel (1986). Alternative models for the heterogeneity of mortality risks among the aged, Journal of the American Statistical Association, 81, 395, 635-644.

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.

Niemeier, D.A. and J. Morita (1994). Duration of trip-making activities by men and women: a survival analysis, presented at the 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., January.

Nishii, K., K. Kondo and R. Kitamura (1988). An empirical analysis of trip chaining behavior, presented at the 67th Annual Meeting of the Transportation Research Board, Washington, D.C., January.

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

Steinberg, D., P.M. Allaman and F.C. Dunbar (1980). The allocation of individual and household activity time and its impact on travel behavior, unpublished report, Charles River Associates.

Sueyoshi, G.T. (1992). Semiparametric proportional hazards estimation of competing risks models with time varying covariates, Journal of Econometrics, 51, 25-58.

Uncles, M.D. (1987). A beta-logistic model of mode choice: goodness of fit and intertemporal dependence, Transportation Research, 21B, 3, 195-205.

Vilcassim, N.J. and D.C. Jain (1991). Modeling purchase-timing and brand-switching behavior incorporating explanatory variables and unobserved heterogeneity, Journal of Marketing Research, 28, 29-41.

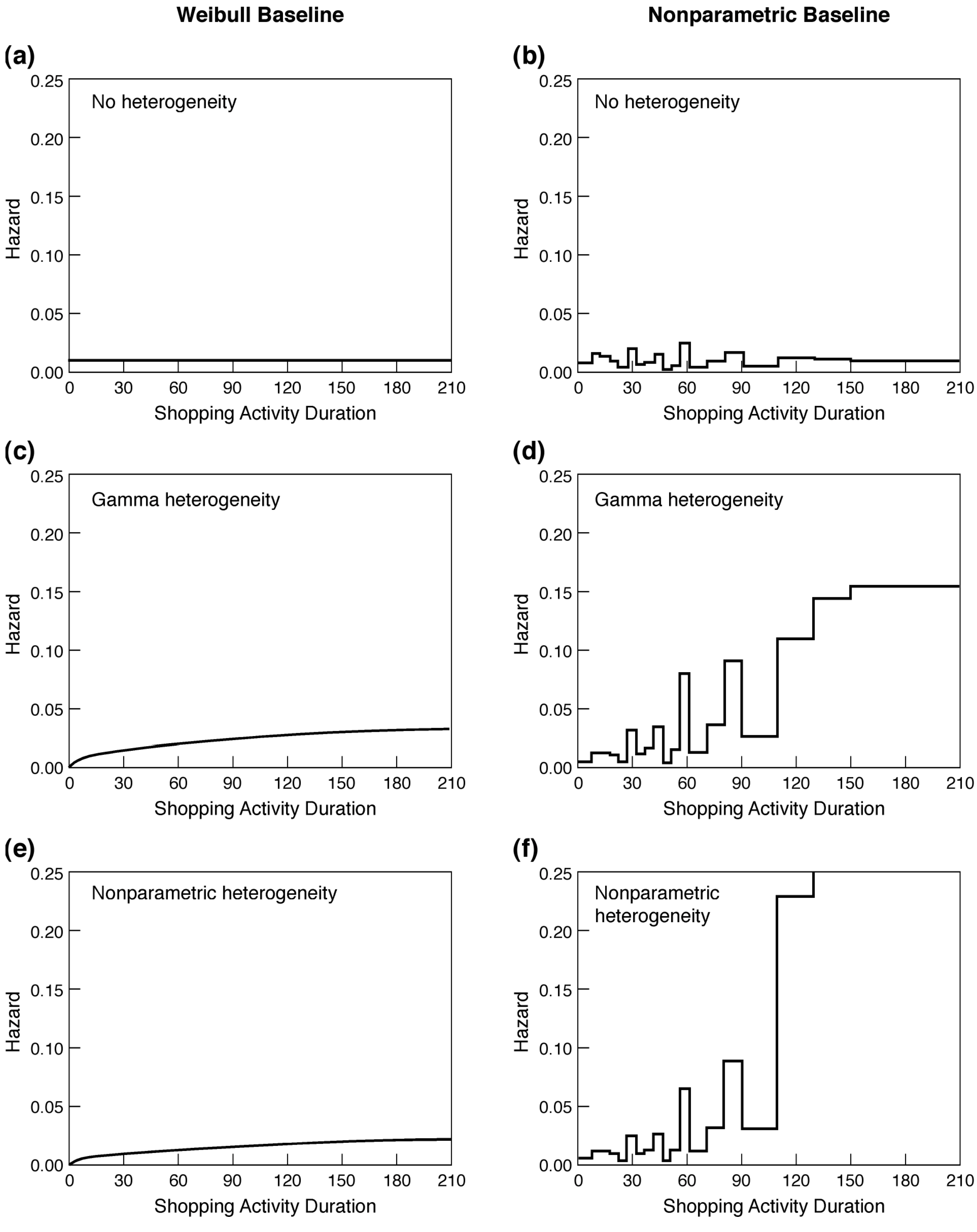

Figure 1. Baseline hazard functions.

Table 1. Shopping Activity Durations and the Discrete Period Sample Hazard

Period k	Time interval (mins.)	Failures F_k [1]	No. at Risk R_k [2]	Discrete-period hazard $H_k = F_k/R_k$	Std. error of H_k
1	0.0 - 7.5	64	355	0.180	0.020
2	7.5 - 12.5	59	291	0.203	0.024
3	12.5 - 17.5	38	232	0.164	0.024
4	17.5 - 22.5	22	194	0.113	0.023
5	22.5 - 27.5	9	172	0.052	0.017
6	27.5 - 32.5	35	163	0.215	0.032
7	32.5 - 37.5	10	128	0.078	0.024
8	37.5 - 42.5	11	118	0.093	0.027
9	42.5 - 47.5	17	107	0.159	0.035
10	47.5 - 52.5	2	90	0.022	0.015
11	52.5 - 57.5	6	88	0.068	0.027
12	57.5 - 62.5	20	82	0.244	0.048
13	62.5 - 72.5	5	62	0.081	0.035
14	72.5 - 82.5	10	57	0.175	0.050
15	82.5 - 92.5	14	47	0.298	0.067
16	92.5 - 112.5	5	33	0.152	0.062
17	112.5 - 132.5	11	28	0.393	0.092
18	132.5 - 152.5	6	17	0.353	0.116
19	152.5 - 212.5	6	11	0.546	0.150
20	> 212.5	5	5	1.000	-

[1] Failures, F_k, represents the number of individuals whose shopping participation end in period k.

[2] The number at risk, R_k, is the number of individuals who are "at risk" of terminating their shopping participation in period k; alternatively, it is the number of shopping spells which have "survived" till the beginning of period k.

Table 2. List of Exogenous Variables in Model

Variable	Definition
Work duration	Time between arrival at work in the morning to departure from work at the evening (in minutes)
Travel time to work	Travel time to work from home if individual does not make any diversions during the commute (in minutes)
Departure from work before 4 pm	1 if individual departs from work before 4 pm, 0 otherwise
Spouse's employment status	1 if individual is married and individual's spouse is employed, 0 otherwise
Work duration of female spouse	Work duration of spouse if spouse is a female (individual is a male), 0 for unmarried individuals (UI), married individuals with an unemployed spouse (MU), and for female married individuals with an employed spouse (FE).
Work duration of male spouse	Work duration of spouse if spouse is a male (individual is a female), 0 for UI, MU, and for male married individuals with an employed spouse (ME)
Travel time to work of female spouse	Travel time to work from home (without diversions) of spouse (in minutes) if spouse is a female (individual is a male), 0 for UI, MU, and FE
Travel time to work of male spouse	Travel time to work from home (without diversions) of spouse if spouse is a male (individual is a female), 0 for UI, MU, and ME
Individual drives alone to work	1 if individual drives alone to work, 0 otherwise
Individual drives alone and spouse rideshares [3]	1 if individual drives alone to work and spouse rideshares to work, 0 for UI, MU, and for ME/FE who drive alone and whose spouse also drives alone
Returning young adult	1 if individual is an employed adult living with one or both parents, 0 otherwise

[3] We use the term "ridesharing" in a broad sense to include all non-drive alone modes of travel such as carpooling, vanpooling, and using transit.

Table 3. Summary Statistics for the Hazard Models

Summary Statistic	Weibull Baseline Models			Nonparametric Baseline Models		
	No Heterogeneity	Gamma Heterogeneity	Nonparametric Heterogeneity	No Heterogeneity	Gamma Heterogeneity	Nonparametric Heterogeneity
Number of baseline parameters	2	2	2	19	19	19
Number of unobserved heterogeneity parameters	0	1	2	0	1	4
Number of covariates	11	11	11	11	11	11
Total number of estimated parameters	13	14	15	30	31	34
Log-likelihood at convergence	-931.85	-927.02	-925.88	-891.25	-887.94	-882.37
Log-likelihood with baseline parameters only	-968.58	-968.58	-968.58	-925.80	-925.80	-925.80
Log-likelihood at zero	-1045.27	-1045.27	-1045.27	-1045.27	-1045.27	-1045.27
Adjusted likelihood ratio index	0.0961	0.0997	0.0999	0.1186	0.1209	0.1233

Table 4. Estimated Covariate Effects

Variable	Weibull Baseline						Nonparametric (NP) baseline					
	No heter.		Gamma heter.		NP heter.		No heter.		Gamma heter.		NP heter.	
	Parm.	t-stat.	Parm.	t-stat.	Parm.	t-stat.	Parm.	t-stat.	Parm.	t-stat.	Parm.	t-stat.
Work characteristics												
Work duration x 10^{-2}	-0.153	-2.57	-0.179	-2.06	-0.229	-2.77	-0.150	-2.52	-0.209	-1.95	-0.095	-1.82
Travel time to work x 10^{-1}	-0.042	-1.28	-0.019	-0.52	-0.039	-0.96	-0.037	-1.20	-0.012	-0.23	0.003	0.10
Departure from work before 4 pm	0.330	2.33	0.532	2.71	0.430	2.44	0.323	2.36	0.601	2.50	0.605	3.31
Spouse's work characteristics												
Employment status	0.797	2.54	1.175	2.63	0.969	2.47	0.801	2.55	1.281	2.42	1.204	2.50
Work duration x 10^{-2}												
Female spouse	0.209	2.88	0.186	1.88	0.226	2.48	0.207	2.89	0.189	1.72	0.095	1.34
Male spouse	0.098	1.80	0.162	2.05	0.134	1.90	0.101	1.87	0.187	1.93	0.176	2.05
Travel time to work x 10^{-1}												
Female spouse	0.175	2.80	0.155	2.20	0.181	2.19	0.178	3.02	0.164	2.04	0.079	1.71
Male spouse	0.070	2.86	0.122	3.02	0.134	3.74	0.067	2.76	0.149	2.10	0.164	5.07
Mode to work												
Individual drives alone	-0.562	-4.32	-0.758	-3.68	-0.800	-5.10	-0.536	-4.33	-0.840	-3.31	-0.817	-4.20
Individual drives alone and spouse rideshares	0.666	2.33	0.862	3.08	0.804	4.18	0.654	3.20	0.928	2.87	0.961	3.61
Socio-demographic characteristics												
Returning young adult	0.975	2.86	1.268	2.68	1.518	3.07	0.926	2.88	1.518	2.31	1.465	2.88