Institute of Transport Studies
Graduate School of Business
The University of Sydney
NSW 2006

ITS-WP94-1

Panel Data, Event Histories and Dynamic Choice Modelling: its Usefulness in Tourism Research

David A. Hensher

January 1, 1994

This paper was presented at the National Conference on Forecasting in Tourism, Canberra, November 25-26, 1993

TITLE: Panel Data, Event Histories and Dynamic Choice Modelling: its Usefulness in Tourism Research

ABSTRACT:

In recognition of the dynamic nature of many consumer decisions, household-based panel studies have been designed and implemented in a number of non-tourism contexts. Such panels involve the monitoring of individual and household behaviour over a period of time in order to gain an understanding of behavioural change as a basis for predicting change. Panel surveys can be of value in tourism contexts. They provide an opportunity to investigate a richer set of tourism behaviours than is possible with a single cross-section. In this paper we discuss the key advantages of a panel approach to studying tourism behaviour. We also outline a number of modelling frameworks which provide suitable approaches to measuring and predicting the sensitivity of choice behaviour to changes in the wider set of factors influencing behavioural response.

.

AUTHORS: David A. Hensher
Professor of Management

CONTACT: Institute of Transport Studies
Graduate School of Business
University of Sydney NSW 2006
Australia

Telephone: + 61 2 550 8631
Facsimile: + 61 2 550 4013
e-mail: davidh@gsb.su.oz.au

DATE: January 1, 1994

1. Introduction

Individual and group choice behaviour in tourism activities in its most general form is best considered as a process in continual flux, with a history and path through time. Although there might be strong agreement with this position, the empirical approach dominating the practice of tourism demand is essentially cross-sectional, based on comparative statics and the assumption that the theoretical construct of *equilibrium* was observable in the real world. In recognition of the limited perspective of a single cross-section, there have been a number of efforts outside of the tourism context to develop a longitudinal orientation to the study of individual and group behaviour. Example panels in transportation, a complementary discipline to tourism, are growing rapidly (Raimond and Hensher 1992).

There are many longitudinal approaches, with a panel of individuals and households *followed* through time being one of the richer specifications. At one extreme we have an event-history panel in which all activities are recorded in continuous time. At the other extreme we have a repeated-cross section (RCS) of individuals or groups which represent a new sample of individuals or groups drawn from the same closed population at discrete time points. The domestic tourism monitor and the international visitors survey are examples of RCS's. A traditional panel is typically a repeated set of observations on a sample of the *same* individuals or groups recorded at discrete time periods, and as such may not necessarily be as rich as a continuous event history panel.

Within the full spectrum of longitudinal data strategies we have the possibility, to varying degrees, of identifying a number of dynamic processes which translate into well-defined sources of variation in behaviour. The literature has tended to classify these sources of behavioural variation into (i) inter temporal relationships exhibited in choice behaviour (statistically referred to as true state dependence), and (ii) variation in behaviour which exists between individuals or groups in the sampled population due to exogenous influences, some of which are not observed by the analyst. The explicit account of the unobserved sources of influence on choice behaviour is referred to as the heterogeneity problem. The set of unobserved sources of influence on behaviour may exist within the sampled time period of a panel and the periods prior to and subsequent to the sample period, and usually have an important influence on behaviour. The presence of heterogeneity prior to the sampled period is known as the initial conditions problem. Failure to allow for the presence of unobserved influences on choice behaviour may lead to spurious inferences about true state dependence. An example of true state dependence is vacation destination being influenced by previous vacation destination. An example of heterogeneity is vacation destination being influenced by a life style propensity which is constant and exists over time for each individual, which is not measured by a set of observed variables.

This paper reviews the appeal of a longitudinal approach to studying the demand for tourism, with a particular emphasis on a panel approach in both discrete and continuous time. We begin with an overview of the advantages of a panel-based longitudinal strategy, followed by a synthesis of some illustrative modelling approaches capable of capturing various dynamic aspects of choice behaviour. This paper is primarily methodological.

2. The Panel-Based Longitudinal Strategy

The single cross-section (SCS) and the RCS currently dominate empirical studies of tourism demand. Traditional monthly, quarterly or annual time series data are also an important source of trend data, although limited in any serious study of tourist behaviour. There are two reasons for this situation. First, the relative ease and cost involved in data collection and statistical analysis. Second, the historical emphasis on long-range forecasting (i.e. beyond 10 years) where the interest is on projections at a very aggregate level, (e.g. total inbound tourism in NSW), where the driving force is forecasts of the distributions of macro economic exogenous factors such as gross domestic product, exchange rates, and past trends. As the tourism research industry becomes dissatisfied with the dimensionality of existing data strategies as an adequate vehicle for studying the microeconomics of tourism behaviour, alternative data paradigms take on appeal. One paradigm which has been neglected in tourism behavioural studies is the panel. In its continuous time format, the panel offers a framework for capturing the event history of a sample of individuals and groups resident in Australia and overseas. By recording the units event history, a pattern of behaviour and behavioural *change* emerges which can be subject to formal statistical analysis to establish the underlying influences on such behaviour at a point in time. A panel based on the recording of events in discrete time can also provide rich behavioural insights although the temporal linkages are not quite as precise as those in continuous time. A longitudinal panel-based strategy thus can add significant power to our understanding of behaviour and especially to the timing and extent of behavioural change in response to an intervention by government or some other source of influence on the exogenous variables explaining choice behaviour.

There are a number of panel survey designs. A summary of the objectives and designs of longitudinal surveys is given in Table 1. Panel surveys that are based on elements from population subgroups that have shared a similar experience (i.e. travel to Hawaii in a given month) are called cohort studies. A panel survey in which sample elements are kept in the panel for only a portion of the duration of a survey is called a rotating panel survey. A split panel survey is a combination of a panel and a repeated or rotating panel. An important distinction must be made between a panel survey and panel data. The latter consist of continuous or periodic measurements of items of interest. In theory it is possible to obtain reliable panel data from a survey conducted at a single point in time; but measurement considerations usually dictate the use of a panel survey design rather than retrospective questioning to collect reliable panel data on individuals or groups (Hensher and LePlastrier 1985). In large part, the analytic advantages derived from panel survey designs are synonymous with those inherent in reliable panel data.

Table 1. Objectives and designs of longitudinal surveys

Objectives	Relevant Design	Examples
A. Estimation of population parameters at time points or during time periods within which changes are assumed negligible	RPS MPS, (PS*) RCS	Monthly trip rate for a particular month Annual kilometres of vehicle use
B. Estimation of mean values of population parameters across a period of time	RPS MPS, (PS*) RCS	Average monthly trip rate Average annual vehicle kilometres
C. Measurement of change at an aggregate level (net change)	PS RPS MPS, (RCS+)	Change in monthly trip rate between 2 periods of time
D. Aggregation of data over time for individuals	PS MPS	Weekly trip rates cumulated to quarterly trip rates
E. Cumulating samples over time	RPS MPS RCS, (PS#)	
F. Measurement of components of individual change	PS MPS	Differences in levels of vehicle use at the individual level at two time periods (gross change)
G. Measurement of frequency, timing, duration of events occurring within given period	PS RPS MPS, (RCS**)	Time length between vacation destination visits for an individual The proportion of people visiting each destination fora particular time period

Notes: PS = panel survey; RPS = rotating panel survey; MPS = mixed panel survey; RCS = repeated cross sections.
* A PS can satisfy this objective if representativeness is maintained.
+ An RCS cannot decompose net change into changes in individual values and changes in population composition (i.e. intra-individual v. inter-individual effects).
# A PS can cumulate samples over time where the characteristics of relevance are not static and of low incidence.
** An RCS is subject to considerable recall error in this context.

To illustrate the context in which a panel design has appeal, it is useful to use the concept of a time-line to depict a set of tourism decisions made over time (Figure 1). If a panel in continuous time had been maintained over the last 5 years, we would be able to pinpoint the exact time at which each individual in a sampled household undertook domestic and international vacations, which household members were in the travelling group, the airline and ticket class, and person-specific constraints which are relevant to the vacation decision. The timeline highlights the importance of the temporal profile of vacations. We could identify by a set of questions the influences on the decision to select each vacation destination and what alternative destinations were considered. Some of these influences could be related to the other endogenous choice at an earlier or later time point. The richness of the data give us more statistical leverage in sorting out causal patterns because they enable the analyst to separate the effects of persistent interpersonal differences from real inter-temporal relationships (Duncan and Kalton 1985). The time-line data can be captured in continuous time and analysed using duration models (see below and Hensher and Mannering 1993).

3. Taking a Closer Look at Some Specific Issues

3.1 Issue 1: Capturing the Important Dynamic Elements of Behaviour

The individual histories of tourists are important because tourism behaviour is a dynamic phenomenon. There is a relationship between current behaviour and the behaviour and circumstances in earlier periods (as highlighted in Figure 1). Some changes occur only after a time lag. It is often important to know the time gap between successive vacations; it is also important to know how long it takes for an individual to respond to an exogenous shock such as a major increase in air fares, the opening of a ‘new’ destination (e.g. Club Med at Lindeman Island), or the general improvement in the state of the local economy. Panel analysis should lead to improved explanations and forecasts of tourism behaviour. Specifically it should be possible to improve explanatory power relating to the dynamics of tourism mobility patterns, and changes in choice and duration of visits to destinations.

This knowledge is important not only in periods of sustained economic growth but also in periods of transition from growth to recession. In periods of recession the uncertainties about tourism behaviour can be expected to grow, making forecasts more difficult at a time when it is probably more important. For example, developing a new airport or vacation destination facility too early is much less of a problem when the sustained growth will catch up with it a few years later, than when the investment is found not to be needed at all.

Individuals observed in a particular state at a point in time typically have a much higher than average probability of being in that state at a subsequent period of time. Does the association between states arise because there is something about the state itself that changes the likelihood of remaining in that state (i.e. true state dependence) or because some unmeasured and unchanging set of personal characteristics of individuals causes observed states to persist (i.e. heterogeneity)? A panel has the required leverage to separate out and evaluate the role of each type of influence on behaviour, together with the myopic sources of influence on tourism behaviour. Within-individual and between-individual sources of variation in behaviour over time identifiable with a panel can be properly addressed, giving a statistical advantage to estimating aggregate changes in comparison to a RCS: sampling errors of the difference between two points in time are substantially lower when the measurements are obtained from a panel than a RCS because one element of the sampling error is eliminated.

3.2 Issue 2: Keeping the Panel Up-to-Date as it Ages

The dynamics of behaviour are inherently individual-specific, whether one is undertaking tourism activity individually or as a group. The notion of a group as the “unit” is not very useful; rather it is more appropriate to see participation in a group as a constraint or contextual effect on tourism activity. Nevertheless a dominating theme in tourism studies is data collected on the activities of families. This requires careful consideration as families "age" and change their composition.

A panel of individuals can begin with a sample of individuals from a population. The sampling procedures may result in a known selection probability for each individual, and the inverse of this probability can be used as a weight to produce estimates of population parameters from the sample. When followed over time, individuals may *leave* the population through death, emigration, quitting (if the panel is limited to an eligible subset of the population). Without updating the panel membership over time, the original sample, appropriately reweighted for differential non-response or with statistical adjustment for sample selection bias, may remain representative with respect to the *original* population from which they were drawn, but not with respect to changes that have taken place in the population since the commencement of the panel.

Panel studies of groups such as households or families must cope with the dynamic aspects of the population of groups. New families are established or *born* into the population when children leave their parental homes or when a married couple splits into two families through separation or divorce, or through immigration. If a panel is to maintain a representative sample of both individuals and families in the population then there must be a replacement mechanism that allows families and individuals to enter the sample with known selection probabilities. In the first year of a panel in which a probability sample of households is drawn, the individual living in the household has the same selection probability as the household. Over time individuals retain their selection probabilities (with possible adjustment for differential *individual* response). Individuals enter the population through birth and immigration; they leave the population through death and emigration. Children leaving the parental homes of a representative sample of households are themselves a representative sample of children leaving the homes of households in the population, and so an ongoing panel study which includes the newly formed households is an aging panel preserving its representative sample through time. Where non-sample adults appear in the "new" households formed from a person in an originally sampled households, adjustment must be made to avoid giving the new household a double chance of falling in the sample.

3.3 Issue 3: An Appealing Analysis Framework

A key objective of modelling is to accurately measure the sensitivity of choice behaviour to changes in relevant explanatory variables. In classifying the types of variables, the two key considerations are (i) to what extent do they vary over time and (ii) to what extent can they be observed? To give some substance to this distinction, which has some element of specific context, we can categorise variables into six types as follows:

	Non-varying	Infrequently varying	Frequently varying
Observed	sex age education level	household type employment status income level car ownership workplace location residential location	travel times travel costs activity choice destination choice mode choice vacation package choice
Unobserved	personality handicaps	tastes attitudes social constraints measurement error	experiences/learning situation constraints

The variables increase in time variability as we move from the left to the right. Age is in the time-invariant category because it is perfectly correlated with time. For most individuals, education level does not change after reaching adulthood. The variables in the next column change over time, but not very frequently. In typical panels with waves up to one year apart, these variables remain unchanged between waves for most respondents. Measurement of their within-individual effects thus requires surveys with a large number of respondents and/or a large number of waves. Household-based variables such as life cycle and car ownership may change more frequently than person-specific variables, because they can be influenced by more than one individual.

The variables in the last column include a number of travel and location decisions which are often modelled in tourism studies. They are subject to a fair amount of variability; panels which record single trips or activity diaries are likely to capture this variation. Alternatively, if individuals are asked about their repetitive choice behaviour (e.g. how often do you go on holiday per annum), then behaviour may appear to be more stable over time. Variables which are not observed in most survey contexts include individual-specific traits such as personality, tastes, attitudes and experiences, all of which define in generic terms major dimensions of impact on behaviour. As rather vague but important terms they are used to separate traits which tend to be stable over time, such as risk aversion, introversion etc. from biases or attitudes which might shift gradually over time as a result of experiences, information, social pressures etc.

Constraints on choice also often remain unobserved. Individual-specific constraints may be of a fairly permanent nature, such as physical handicaps or phobias concerning certain types of tourism activity. Social constraints resulting from arrangements in the household will tend to change over time, although perhaps not very often. Many other types of constraints are likely to vary from time to time such as sickness, airport strikes, bad weather etc.

Measurement error relates to unobserved variations in the observed variables. The more frequently a variable changes, the more likely it is that the measured value will be inaccurate for explaining a specific instance of choice. Travel times and prices of travel to destinations are important examples; a network-based approximation to travel times and prices include spatial as well as temporal measurement error. One can pose survey questions about certain unobserved effects such as attitudes and situational constraints. While such explanatory variables may provide useful insights in an explanatory sense, they are generally not useful for prediction over time, since one cannot predict how these variables will change. It is often preferable to use such variables as indicators of latent characteristics which can in turn be related to demographic and socioeconomic factors. There are a number of ways of representing these unobserved effects:

(i) in alternative-specific constants representing the net utility, or systematic bias for each alternative resulting from all unobserved effects. The strength of these biases is assumed to be the same for all individuals in a sample.

(ii) in the residual random error term representing the non-systematic utility component for each alternative resulting from all unobserved factors. The variance of the random component is usually assumed to be identically and independently distributed (iid) across all alternatives and individuals (i.e. the multinomial logit model). Alternative model specifications such as nested logit and simulated moments multinomial probit relax these assumptions to varying degrees.

Using panel data, within-person changes can be identified separately from across-person variation which may be correlated with heterogeneity. For panel data however, the iid assumption may be violated if serial correlation is present in the random error terms for choices made by the same individuals in different periods. Concepts such as habit or inertia are indications of serial correlation; individuals are often likely to repeat the choices which they made in previous periods, for reasons which cannot be related to the observed explanatory variables.

It is not obvious whether unexplained stability in behaviour is due to heterogeneity (missing variables) or due to true state dependence. In contrast to heterogeneity which reflects individual-specific differences, state dependence reflects structural resistance to change due to transactions costs, imperfect information, time lags etc. Fairly complex model specifications are necessary to disentangle these effects; although some approximations procedures can be developed to test whether such effects are present, as well as being useful in predicting behaviour. We discuss some modelling strategies in Section 5.

3.4 Synthetic households as a useful way of forecasting demand

The sample of individuals or households used in model estimation are obtained from a sampled population. In application, we are interested in predicting demand for the population as a whole from which the sample is drawn. The identification of a set of representative household types, called synthetic households which can be carried forward in time with differential population weights provides an appealing way of undertaking forecasts of behaviour.

A synthetic household is a household with particular socioeconomic characteristics and a population weight to indicate the incidence of such a household in the base and future populations. The characteristics of each household will not change over time; the weight will vary over time to represent the changing incidence of each household type. Synthetic households have to be defined on a number of socioeconomic dimensions. The selection is constrained by the availability of population-wide data and the reliability of these base year dimensions as projection criteria in *carrying* synthetic households through time (i.e. the reliability of the population weights through time). The census is an appealing data source for constructing synthetic domestic households in the base year. To manage the design of such households, it is convenient to distinguish a set of core socioeconomic criteria and a set of non-core criteria. Core socio-demographic variables might be stage in the life cycle and household gross income. They are used to specify a multi-way contingency table, and in the forecast years when only marginal distributions of each variable is known, used to generate cell numbers using iterative proportional fitting, given projected marginal distributions. Within each of the core cells, distributions might be established for a set of non-core socio-economic variables which are explanatory variables in the tourist choice models. It is essential that the core and non-core socioeconomic variables are limited to the set of explanatory variables found to have a significant influence on the modelled choices, because it is through this linkage that the model system can be effectively implemented.

3.5 The Appeal of Panels

What then are the appealing features of a panel data strategy? Panel surveys in their classical form have as their primary objective the study of components of change over time at the sample unit level. By monitoring the same sample units throughout the whole duration of the survey it is possible to measure (i) the average change in behaviour for each individual, (ii) gross change between points in time and (iii) the stability of behaviour for an individual. Furthermore, the panel, subject to representativeness at each point in time can satisfy the conventional single cross section and repeated cross-section objectives. Given the importance of this additional *information power* in the panel data, it is useful to take a closer look at the ideas of stability and change before discussing modelling strategies.

4. Stability and Change

To motivate the discussion on stability and change, we will use the example of annual visits to primary domestic destinations for vacation (V) for a particular life cycle/lifestyle segment of the population, defined for *w* waves (*w* = 1,...,W). At the household level, q, the study of change involves simple differencing for any combination of waves:

$$\Delta V_q = V_{qt} - V_{q,t-w} \qquad (1)$$

The population specification for this is:

$$\boxed{\text{cov}(V_{t-w}, V_t) = \text{var}(V_{t-w}) + \text{cov}(V_{t-w}, \Delta V)} \qquad (2)$$

The correlation between V_{t-w} and V_t can be decomposed into a measure for the *lack of change* (or stability index): var(V_{t-w}), and the *causal* influence of V_{t-w} on the change component of V_t: $\text{cov}(V_{t-w}, \Delta V)$ (Kessler and Greenberg, 1981). The greater the contribution of var (V_{t-w}), the less annual visit rates change over time for the segment.

The covariance between the initial level of V and the change in V ($\kappa_1 - 1 = \kappa_1^*$) can be found empirically by transforming the partial regression coefficient on V:

$$\boxed{V_t = \kappa_0 + (\kappa_1^* + 1) V_{t-w} + \sum_{k=1}^{K} \kappa_k S_k + \xi} \qquad (3)$$

where S_k is the kth socioeconomic effect (k = 1,...,K) and ξ is the residual disturbance, with the property N(0, σ_ξ^2). Note that the linear difference equations (i.e. $V_t - V_{t-w}$ as the left-hand side variable) and equations with static variable definitions (as in equation 3) are mathematically equivalent; the only difference is in the interpretation of the particular parameter estimates. Thus in the absence of change, the coefficient of V_{t-w} would be unity. Since κ^* represents the influence of V_{t-w} on a change in V_t, deducting unity from the empirical parameter estimate gives the effect of change. This result is derived from a model of the form:

$$V_t = \kappa_0 + \kappa_1 V_{t-w} + \sum_{k=1}^{K} \kappa_k S_k + \xi; \quad \text{or}$$

$$\boxed{V_t - V_{t-w} = \kappa_0 + (\kappa_1 - 1) V_{t-w} + \sum_{k=1}^{K} \kappa_k S_k + \xi,}$$

$$= \kappa_0 + \kappa_1^* V_{t-w} + \ldots\ldots + \xi \qquad (4)$$

where κ_1^* is the regression of ΔV on V_{t-w} when S_k is controlled. The formulae above are suitable for simple causal analysis of stability and change. We can build on this by using a set of *descriptive* formulae and by further dissecting the components of change. The amount of change in the population of I individuals between time t-w and t is:

$$C^2 = \frac{\sum_{i=1}^{I} \left(V_{i,t} - V_{i,t-w}\right)^2}{I} \tag{5}$$

which can be expanded out and rearranged to give:

$$C^2 = \mathrm{var}(V_t - V_{t-w}) + (\overline{V}_t - \overline{V}_{t-w})^2 \tag{6}$$

The terms on the right identify respectively how much of the change affects all households equally and how much is change relative to other households. Empirical calculation of C^2 is straightforward, with:

$$\mathrm{var}(V_t - V_{t-w}) = \mathrm{var}(V_t) + \mathrm{var}(V_{t-w}) - 2(\rho_{V_{t-w}V_t} s_{V_{t-w}} s_{V_t}), \tag{7}$$

where ρ is the partial correlation between the values of a variable in adjacent waves and s is the standard deviation. Finally, one may want to distinguish changes in vacation visit rates that are, and are not, predicted by a previous level. To do this we define:

$$\mathrm{var}(V_t) = \mathrm{var}(V_{t-w}) + [\kappa_Z s_{t-w}]^2 + 2\kappa_Z s_{t-w}^2 + \mathrm{var}(\xi), \tag{8}$$

where κ_Z is the unstandardised regression coefficient of ΔV on V_{t-w}. The first term in equation (8) is the component of V_t due solely to the lack of change in V over time, the second term is the component uniquely due to the structural relationship between V_t and V_{t-w}, the third term accounts for interaction (that is a non-uniqueness effect, households with different initial values of V_{t-w} will change by different amounts), and the last element is the residual change unrelated to V_{t-w}.

Thus far we have implicitly assumed that the errors in equation (3) are serially uncorrelated. Since the magnitude of κ^* is important in the assessment of the relative impact of stability and change, it is necessary to *correct* for serial correlation due to the lagged endogenous variable V_{t-w}. When there are a large number of time periods (for example time series data), a correlogram can be used to determine the form of the serially correlated errors, whereas this is not possible when one is using a panel which is typically limited to a few waves. Instead one would then use rough empiricism and/or *a priori* reasoning to specify the form of the serial correlation (that is, to specify the nature of the off-diagonal elements in the error variance-covariance matrix).

The alternative assumptions on the form of the serial correlation (first order autoregressive, etc), about the nature of omitted variables or the appropriate nature of instrumental variables, as ways of accommodating serial correlation are, in the panel context, not well defined. Generalised least squares estimation is capable of yielding estimates less efficient than those yielded by ordinary least squares estimation of untransformed variables. This is because the estimated value of the correlation between the error terms (in two waves) is so different from the unknown true value (see Rao and Miller 1971, 71-74).

A first order autoregressive structure is commonly assumed (equation 9) as an approximation enabling us to investigate all non-initial waves with at least one lag; e_t is the nonautoregressive error component. Using Hatanaka's two-step estimator, the estimates of κ_0, (κ^*+1), κ_k and ρ are efficient and equivalent to maximum likelihood estimates (Hatanaka 1974). This procedure is fully implemented in LIMDEP (Econometric Software 1993).

$$\xi_t = \rho\xi_{t-w} + e_t \tag{9}$$

5. Modelling Strategies

5.1 Event History Analysis in Continuous Time: Application of Duration Models

To illustrate the dimensions of the problems presented when one chooses to analyse duration data, the example of the introduction of a new vacation package is used. In this case, the analyst would not necessarily be interested in the "equilibrium" state of vacation package acceptance, which is the state that is presumably being captured when one undertakes the estimation of standard logit-based choice models. That is, standard logit-based analyses of vacation choice probabilities assume an instantaneous adjustment to price, quality, and other factors upon which vacation packages are compared. However, there may be a strong interest in looking at the rate at which individuals initially try the new package on their way to establishing an equilibrium state, because slow acceptance may create political and financial pressures that could affect package viability. Analysis of this acceptance rate is a classic application of duration data, where duration in this case is defined as the time between the introduction of the new vacation package and the time individuals first try the new package.

Structurally, the data needed to model this duration problem is illustrated by the example provided in Figure 2. In this figure, five individuals are sampled to obtain information on their trying a new vacation package. Information on the package choices of these individuals is collected over some period of time until the survey is terminated at time C. At time C, there will likely be a group of individuals (e.g. individuals represented by person 2 in Figure 2) that either; a) will never try the new package, or b) will eventually try the new package, but just have not done so up to time C. The duration spells of these individuals will be censored since they are not observed trying the new vacation package. This type of censoring is referred to as right-censoring.

Another type of censoring could arise if the survey was begun some time after the new package was introduced. In Figure 2, for example, if the survey was started at time B, it may be difficult to determine when an individual (such as individual 3) was first exposed to the new vacation option. Being unable to determine when duration's begin is referred to as left-censoring. Left-censoring poses the additional problem of not knowing the value of the determinants of duration (e.g. income, household size, attitudes) at the beginning of the duration period. Left-censoring can be avoided, in this case, by beginning the survey when the vacation package is first introduced (time A in Figure 2). This will ensure full knowledge of the lengths of durations as well as possible determinants of durations. The duration model in its statistical form is referred to as a hazard function.

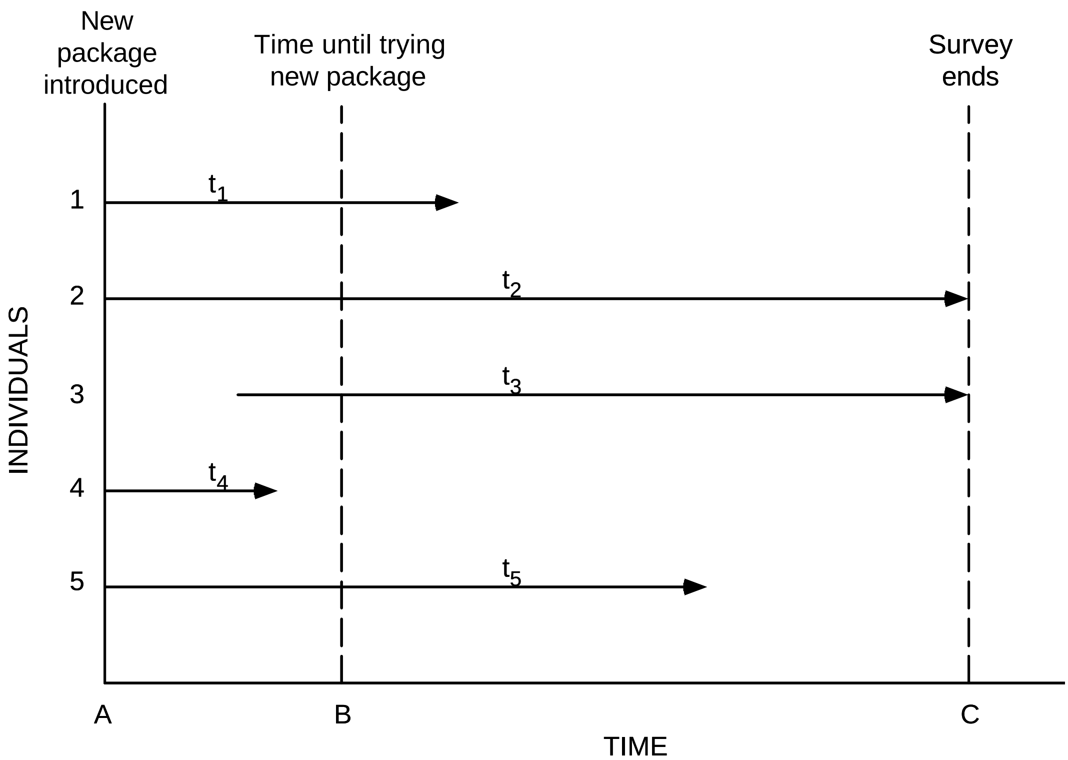

Figure 2: Example of Duration Data

Mathematically, the hazard function can be expressed in terms of a cumulative distribution function, F(t), and a corresponding density function, f(t). The cumulative distribution is written as,

$$F(t) = \Pr[T < t] \tag{10}$$

where Pr denotes the probability, T is a random time variable, and t is some specified time. In the case of the time until the acceptance of a vacation package, equation 10 gives the probability of trying the new package before some transpired time, t. The corresponding density function (the first derivative of the cumulative distribution with respect to time) is,

$$f(t) = dF(t)/dt \tag{11}$$

and the hazard function is,

$$h(t) = f(t)/[1 - F(t)] \tag{12}$$

where h(t) is the conditional probability that an event will occur between time t and t+dt given that the event has not occurred up to time t. In words, the hazard, h(t), gives the rate at which events (such as trying a new vacation package) are occurring at time t, given that the event has not occurred up to time t.

Another important construct in hazard-based models is the *survivor function*. The survivor function gives the probability that a duration will be greater than or equal to some specified time t. That is, the probability that an individual remains in the state ("survives") until time t. The survivor function is written as,

$$S(t) = Pr[T \geq t] \tag{13}$$

and therefore is related to the cumulative distribution function by,

$$S(t) = 1 - F(T) \tag{14}$$

and to the hazard function by,

$$h(t) = f(t)/S(t) \tag{15}$$

Graphically, hazard, density, cumulative distribution and survivor functions are illustrated in Figure 3. This figure provides a visual perspective of the equations presented above. Turning specifically to the hazard function, its slope has important implications. The probability of ending a duration may be dependent on the length of the duration. This is referred to as duration dependence and the first derivative of the hazard function with respect to time (i.e. the slope of the hazard function) provides this information.

To illustrate this, consider the four hazard functions shown in Figure 4. In this figure, the first hazard function, $h_1(t)$, has $dh_1(t)/dt > 0$ for all t. This is a hazard that is monotonically increasing in duration implying that the longer individuals go without exiting a duration, the more likely they are to exit soon. The second hazard function has $dh_2(t)/dt < 0$ for all t and is monotonically decreasing in duration. This implies the longer individuals go without exiting a duration the less likely they are to exit soon. The third hazard function has $dh_3(t)/dt = 0$ which means that exit probabilities are independent of duration and no duration dependence exists. Finally, the fourth hazard function is non-monotonic and has $dh_4(t)/dt > 0$

and $dh_4(t)/dt < 0$ depending on the length of duration t. In this case the exit probabilities increase or decrease in duration.

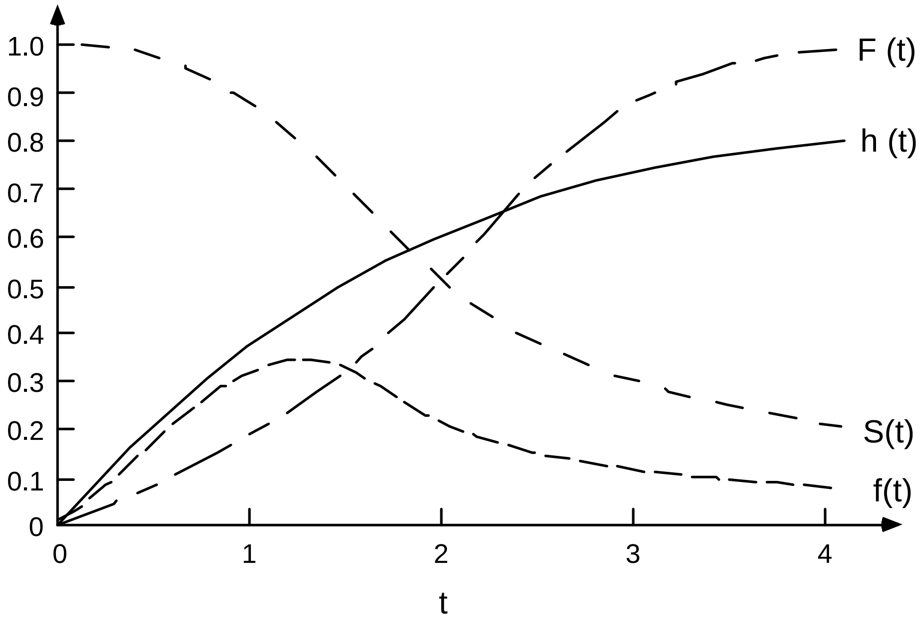

Figure 3: Illustration of Hazard (h(t)), Density (f(t)), Cumulative Distribution (F(t)) and Survivor Functions (S(t))

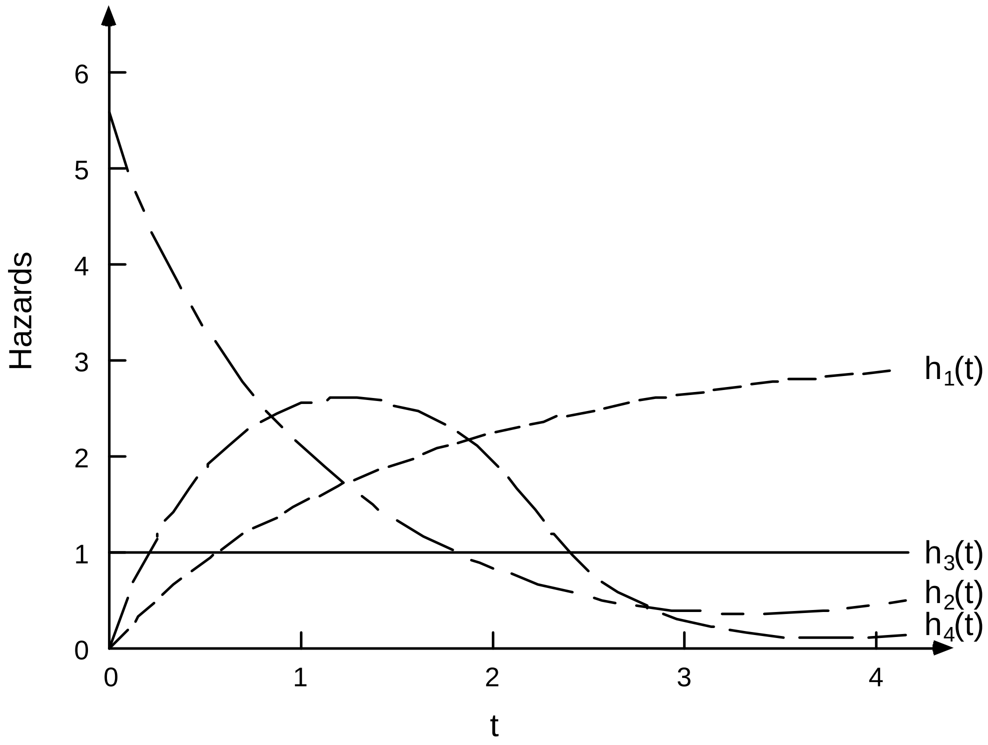

Figure 4: Illustration of Four Alternative Hazard Functions

Information relating to duration dependence, as derived from the first derivative of the hazard function with respect to time, can provide important insights into the duration process being modelled. However, there are clearly important determinants of duration (e.g. socioeconomic characteristics) that must be accounted for in the modelling approach as well. These determinants, or covariates, are included in hazard-based models using two alternative methods; *proportional hazards* and *accelerated lifetime*.

Proportional hazards models operate on the assumption that covariates (i.e. factors that affect duration) act multiplicatively on some underlying hazard function. The proportionality is due to the decomposition of the hazard rate into one term dependent upon time, and another dependent only on the covariates. This concept is illustrated in Figure 5.

In this figure, the underlying (or baseline) hazard function is denoted $h_o(t)$, and this is the hazard function assuming all elements of the covariate vector, Z, are zero. The manner in which covariates are assumed to act on the baseline hazard is usually specified as the function $\exp(\beta Z)$, where β is a vector of estimable parameters. Therefore the hazard rate with covariates, $h(t|z)$, is given by the equation (as shown in Figure 5),

$$h(t|Z) = h_o(t)\exp(\beta Z) \tag{16}$$

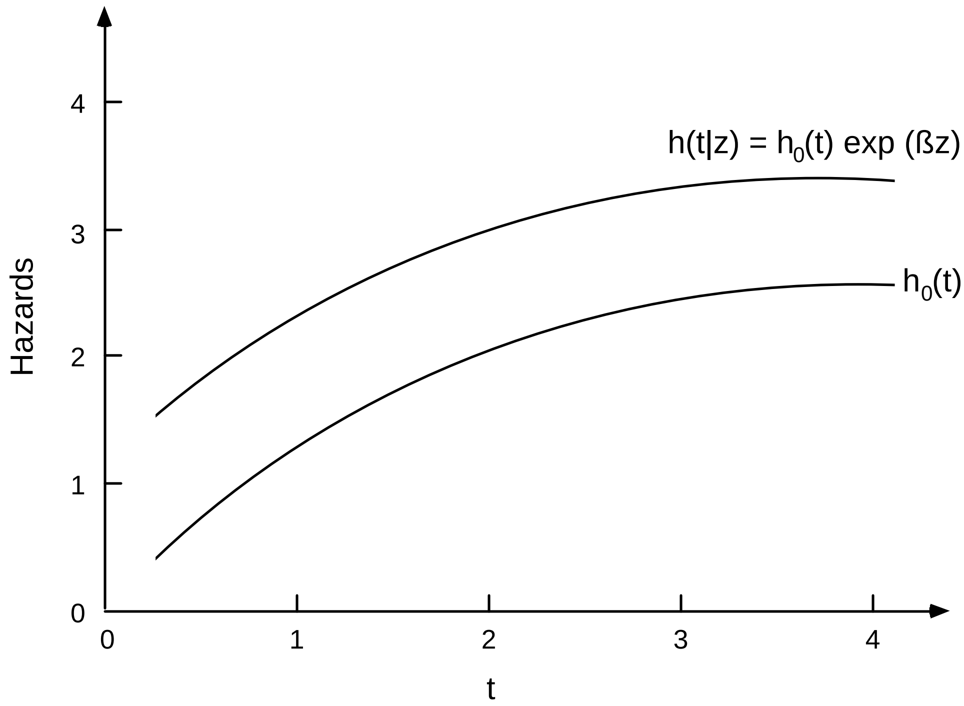

Figure 5: Illustration the Proportional Hazards Model

Proportional hazards models have enjoyed considerable popularity in a variety of fields (see Fleming and Harrington, 1990). These models can easily handle right-censored data and they provide a nice interpretation of estimated parameters (i.e. simple multiplicative effect on the underlying hazard). The assumption of proportionality however limits the application set. For example, if a covariate is car ownership (zero, non-zero), the quotient of the hazard rate of owners and non-owners should not vary over time. This restriction can be relaxed to an extent by introducing class-specific hazard rates: $h_c(t|z) = h_{oc}(t)\ \exp(\beta z)$ where c = 1, ...,C classes.

An alternative approach of incorporating covariates in hazard-based models is the accelerated lifetime model. This model assumes that the covariates rescale time directly (i.e. accelerate time) in a baseline survivor function which is the survivor function when all covariates are zero. Assuming that the covariates act in the form $\exp(\beta Z)$, as was the case for the proportional hazards model, the accelerated lifetime model can be written as,

$$S(t|Z) = S_o[t\exp(\beta Z)] \tag{17}$$

and it follows that this model can be written in terms of hazard functions as,

$$h(t|Z) = h_o[t\exp(\beta Z)]\exp(\beta Z) \tag{18}$$

Accelerated lifetime models have, along with proportional hazards models, enjoyed wide-spread use (see Kalbfleisch and Prentice, 1980). The selection of accelerated lifetime or proportional hazards models is often determined on the basis of distributional assumptions (i.e. the assumed distribution of durations).

Two general approaches to implementing hazard-based model are possible. One is to assume a distribution of duration (e.g. Weibull, exponential, etc.) and the other is to apply a generalised approach that does not require a distributional assumption. The former approach is called "fully parametric" because a distributional assumption is being made for the hazard along with an assumption on the functional form specifying how covariates interact in the model (i.e. the $\exp(\beta Z)$ used in the previous section). The latter approach is semi-parametric because only the covariate functional form is specified.

Fully parametric models can be estimated in proportional hazards or accelerated lifetime forms, and a variety of duration-distribution alternatives are available including gamma, exponential, Weibull, log-logistic, and log-normal. The choice of any one of these alternatives can be justified on theoretical grounds, and each has important implications relating to the shape of their underlying hazard functions. Three common distributions; exponential, Weibull, and log-logistic, are summarised below.

The exponential distribution is the simplest to apply and interpret. With parameter $\lambda > 0$, the exponential density function is,

$$f(t) = \lambda\exp(-\lambda t) \tag{19}$$

with hazard,

$$h(t) = \lambda \tag{20}$$

Equation 20 implies that this distribution's hazard is constant and thus the probability of exiting a duration is independent of the length of time of the duration. This is a fairly restrictive assumption because the exponential distribution does not allow any sort of duration dependence to be captured.

The Weibull distribution is a more generalised form of the exponential in that it allows for positive duration dependence (hazard is monotonic increasing in duration), negative duration dependence (hazard is monotonic decreasing in duration) or no duration dependence (hazard is constant in duration). With parameters $\lambda > 0$ and $P > 0$, the Weibull distribution has density function,

$$f(t) = \lambda P(\lambda t)^{P-1}\exp[-(\lambda t)^P] \tag{21}$$

with hazard,

$$h(t) = \lambda P(\lambda t)^{P-1} \tag{22}$$

In Equation 22, if the Weibull parameter P is greater than one, the hazard is monotone increasing in duration, if P is less than one it is monotone decreasing in duration, and if P equals one, the hazard is constant in duration and reduces to the exponential distribution's hazard (i.e. $h(t) = \lambda$). Since the Weibull distribution is a generalised form of the exponential distribution it provides a more flexible means of capturing duration dependence, but it is still limited due to the monotonicity restriction that it places on the hazard. In many applications, a non-monotonic hazard may be theoretically justified.

The log-logistic distribution allows for non-monotonic hazard functions and is often used as an approximation of the more computationally cumbersome log-normal distribution. The log-logistic, with parameters $\lambda > 0$ and $P > 0$ has the density function,

$$f(t) = \lambda P(\lambda t)^{P-1}[1+(\lambda t)^P]^{-2} \tag{23}$$

and hazard function,

$$h(t) = [\lambda P(\lambda t)^{P-1}]/[1+(\lambda t)^P] \tag{24}$$

Note that the log-logistics hazard is identical to the Weibull's except for the denominator. Equation 24 shows that if $P < 1$, the hazard is monotone decreasing, if $P = 1$, the hazard is monotone decreasing from parameter λ, and if $P > 1$, the hazard increases from zero to a maximum at time $t = [(P-1)^{1/P}]/\lambda$ and decreases toward zero thereafter.

Figure 6 shows a comparison of the hazards of the three distributions discussed. In this figure an exponential distribution is presented along with monotone increasing and decreasing Weibull distributions, and a non-monotonic log-logistic distribution. The selection of a distribution is in part guided by reasonable hypotheses on behavioural response over time. For example, in the case of a new vacation package, individuals who are eager to choose it but then lose interest might be represented by the log-logistic; those unaffected by advertising and word-of-mouth might be represented by the exponential distribution.

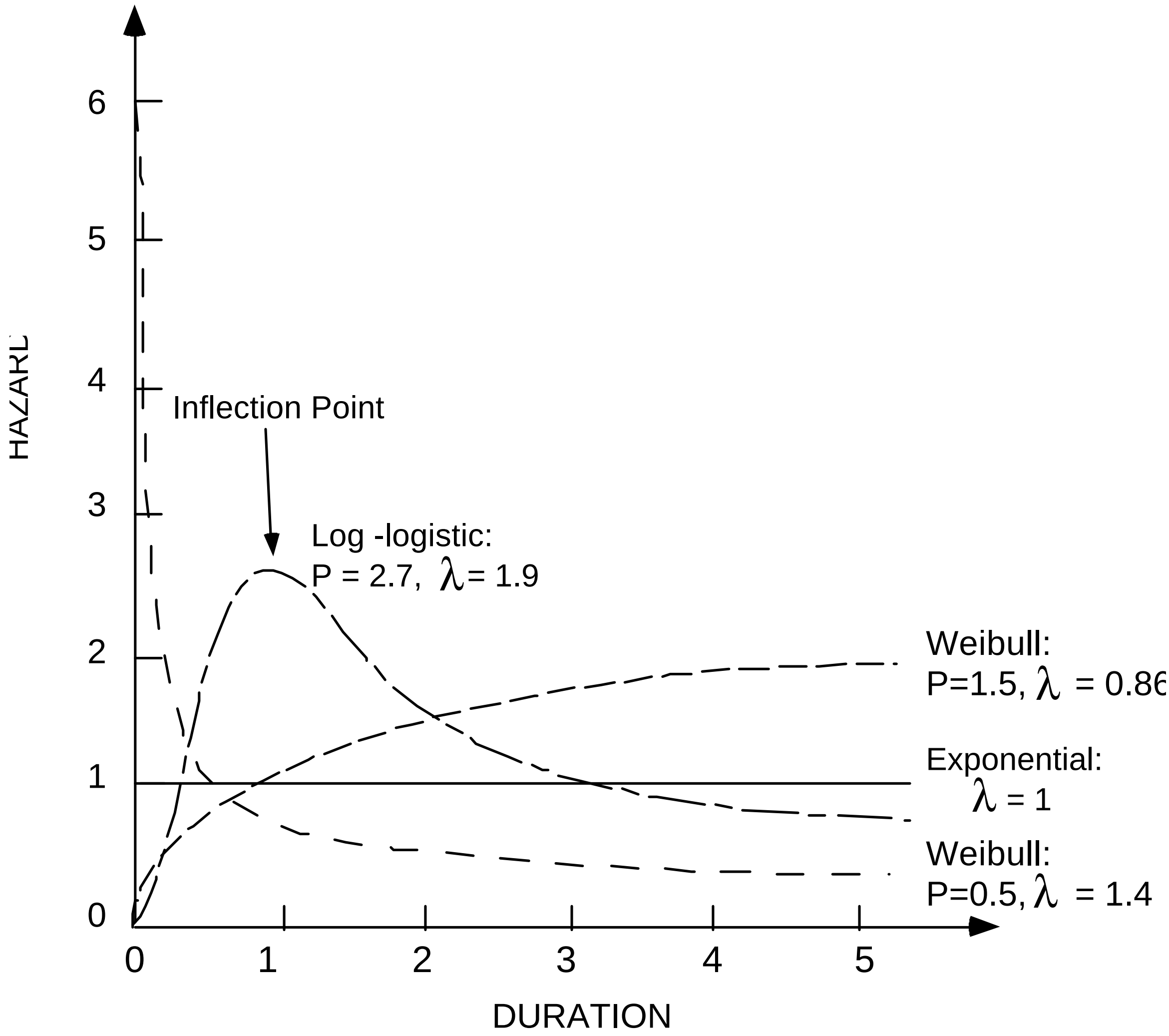

Figure 6: Hazard Function Distributions (Kiefer, 1988)

The alternative to assuming a distribution of the hazard is to use a non-parametric approach for modelling the hazard. This is convenient when little or no knowledge of the functional form of the hazard is available. Such an approach was developed by Cox (1972) and is based on the proportional hazards approach. The Cox proportional hazards model is semi-parametric because exp(βZ) is used as the functional form of the covariates. The model is based on the ratio of hazards so that the probability of an individual, i, exiting a duration at time t_i, given that at least one individual exits at time t_i, is given as,

$$\exp(\beta Z_i) / \sum_{j \in R_i} \exp(\beta Z_j) \qquad (25)$$

where R_i denotes the set of individuals with durations greater than or equal to t_i.

The Cox proportional hazard model has been used in a number of fields (see Fleming and Harrington, 1990, Breslow, 1974, Elandt-Johnson and Johnson, 1983). Some caution should be exercised when applying semi-parametric models. If the hazard is generated from a known distribution, and a Cox model is applied, statistical efficiency will be lost since information regarding the hazard's distribution is not being used. This could result in less precise coefficient estimates as reflected by their higher standard errors. Although this efficiency matter is of some concern, several studies (e.g. Efron, 1977; Oak, 1977) have found the asymptotic variance matrix of Cox model estimators to be close to those generated from fully parametric hazards models. Thus, in most cases, Cox models can be applied without serious efficiency losses.

5.1.1 Heterogeneity and State Dependence

Two issues of importance in panel studies are heterogeneity and state dependence. The assumption implicitly made in proportional hazards models is that the survivor function is homogeneous over the population being studied. As such, all of the variation in durations is assumed to be captured by the covariate vector Z. A problem arises when some unobserved factors (i.e. not included in Z) influence durations. This is called unobserved heterogeneity and can result in a major specification error that can lead one to draw erroneous inferences on the shape of the hazard function and covariate coefficient estimates (see Heckman and Singer, 1984). Ignoring heterogeneity is the equivalent to leaving out an important covariate in the $\exp(\beta Z)$ function. Fortunately, a number of corrections have been developed to explicitly account for heterogeneity. The most common is to include a heterogeneity term that is designed to capture unobserved effects across the population, and work with the conditional duration density function. With a heterogeneity term, v, having a distribution over the population, g(v), and with a conditional duration density function, f(t|v), the unconditional duration density function can be determined from,

$$f(t) = \vee f(t|v)g(v)dv \qquad (26)$$

With this formulation, hazard models can be derived using procedures identical to those used in the derivation of the non-heterogeneity hazards models.

The problem in operationalizing such an heterogeneity model is that a distribution of heterogeneity in the population must be selected. There is seldom any theoretical justification for selecting one distribution over another, and the economics and marketing literature is strewn with papers that have used a wide-variety of heterogeneity distributions, the most popular of which is the gamma distribution (Econometric Software, 1993). The selection of a heterogeneity distribution must not be taken lightly. The consequences of incorrectly specifying g(v) are potentially severe and can result in inconsistent estimates as demonstrated both theoretically and empirically by Heckman and Singer (1984). Fortunately, from the perspective of choosing among many possible distributions, it has been shown (Kiefer, 1988) that if a correctly specified duration distribution is used, the coefficient estimate results are not highly sensitive to alternate distributional assumptions of heterogeneity. To avoid concern about heterogeneity assumptions entirely, Heckman and Singer (1984) propose a non-parametric representation of heterogeneity that requires no prior parametric assumptions. Their method has been successfully applied and appropriate software is available (see Econometric Software, 1993).

State dependence in duration models considers the effect that past duration experiences have on current durations. Such dependence can capture important habitual behaviour effects that can be strong indicators of the length of durations. Heckman and Borjas (1980) provide an extensive discussion of state dependence issues in hazard-based models.

In most models of duration, three types of state dependence can exist; duration dependence, occurrence dependence, and lagged duration dependence. Duration dependence simply focuses on the conditional probability of a duration ending soon, given that it has lasted until some known time. This type of state dependence is captured in the shape of the hazard function (see Figure 5). For example, a monotone increasing hazard (Weibull with P = 1.5 and λ= 0.86 as shown in Figure 5) has positive duration dependence since the longer the individual's duration, the more likely the duration will end soon. Most hazard models (with the notable exception of the exponential distribution) implicitly embody some form of duration dependence.

Occurrence dependence captures the effect that the number of previous durations has on the current duration. For example, individuals that have delayed their departure from work to home to avoid traffic congestion four times during the past week may have different current-day departure-delay durations than individuals that have delayed only once in the past week. The four-delay individuals may have longer or shorter current-day delay durations because they are more experienced delaying and perhaps have a better notion of when to leave to optimise their avoidance of traffic congestion. Occurrence duration is accounted for by including the number of previous duration occurrences in the covariate vector Z.

Finally, lagged duration dependence captures the effect that the lengths of previous durations have on current duration. Returning to the example of delaying departure from work, an individual who has delayed a specified amount of time on a preceding day may have developed a habitual pattern that would make previous-day delay duration a good predictor of current-day delay duration. Again, this type of state dependence is accounted for by including lagged durations in the covariate vector Z.

Great caution must be exercised when including and interpreting state dependence. The common problem is that unobserved effects (heterogeneity) remain in the model and are "picked up" in the coefficients of the state variables included in the covariate vector Z. For example, suppose that income is an important determinant of the length of time that individuals delay their departure from work, but a duration model is estimated without income (i.e. income becomes the equivalent of an unobserved effect). If a lagged duration variable is included in the model, its estimated coefficient will be capturing lagged duration effects as well as residual income effects because income also determines, and is therefore correlated with, lagged duration dependence. In the presence of such heterogeneity, inferences drawn on state dependence could be erroneous because non-state effects are being captured.

Elbers and Ridder (1982) have shown that if heterogeneity is properly accounted for, duration dependence (i.e. the dependence captured by the shape of the hazard function) can be accurately captured. However, there are really two types of heterogeneity. One is "pure" heterogeneity which refers to unobserved factors that are not influenced by previous duration involvement (as discussed above in the heterogeneity portion of the paper). The second is "state dependent" heterogeneity and refers to unobserved factors that are influenced by an individual's previous duration involvement. This second type of heterogeneity is extremely difficult to distinguish from occurrence and lagged duration dependence even if heterogeneity is explicitly accounted for as shown in Equation 17, because such corrective methods typically capture "pure" but not "state dependent" heterogeneity (see Heckman and Borjas, 1980). One relatively simple solution to this problem is to instrument state variables by regressing them against exogenous covariates and using regression-predicted values as variables in the duration model.

5.2 Event History Analysis in Discrete Time: Application of Discrete Choice Models

More typically, panel data is collected as a series of discrete snapshots at regular-spaced time intervals without any event-history data in continuous time or capable of being translated into continuous time. Furthermore, many tourism demand interests centre on individuals or groups making discrete choices. To illustrate a set of useful analytical tools, we will use an example of the choice between destinations (m = 1,...,M), airline (b = 1,...,B) and season (s = 1,...,S). The aim of the tourism choice model system is to replicate the choice process of individuals. This is achieved by constructing a discrete-choice model such as multinomial logit or nested logit. In deriving an empirically estimable model we distinguish between the observable and unobservable components of the utility function by specifying:

$$V_{aht} = \overline{V}_{aht} + e_{aht} \tag{27}$$

where $\overline{V}_{aht}$ is the observable or representative component of the utility function associated with the vacation package a for individual h at time t, and e_{aht} is an heuristically treated error term. Given a functional form for $\overline{V}_{aht}$, and that the e_{aht} are independently and identically distributed (iid) extreme value type I (EVI), the choice process delineated by equation (27) could be estimated by multinomial logit (MNL). Our a priori suspicions are, however, that the e_{aht} will not be iid; instead varying systematically across different vacation packages. Under these circumstances the MNL model will yield biased estimates.

As a statistical response to our suspicion, rather than approximating V_{aht} directly, we have chosen to decompose the vacation package into destination choices, airline choices and choice of time of year.

The e_{aht} are allowed to be correlated across packages with different destination, airline and season mixes, but are assumed to be iid EVI for bundles with the same destination, airline and season mixes. The resultant nested logit model is:

$$P_{(sbm)ht} = \frac{\exp\left\{\overline{V}_{(sbm)ht}\right\}\left(\sum\limits_{m'\in M}\exp\left\{\overline{V}_{(sbm')ht}\right\}\right)^{\left(\frac{1}{\left\{1-\rho^{*}\varepsilon_{mt}\varepsilon_{bt}\right\}}\right)}}{\sum\limits_{m'\in M}\exp\left\{\overline{V}_{(sbm')ht}\right\}\sum\limits_{b'\in B}\left(\sum\limits_{m'\in M}\exp\left\{\overline{V}_{(sb'm')ht}\right\}\right)^{\left(\frac{1}{\left\{1-\rho^{*}\varepsilon_{mt}\varepsilon_{bt}\right\}}\right)}} \cdot$$

$$\frac{\sum\limits_{b'\in B}\left(\sum\limits_{m'\in M}\exp\left\{\overline{V}_{(sb'm')ht}\right\}\right)^{\frac{\left(1-\rho^{*}_{\varepsilon_{bt}\varepsilon_{st}}\right)}{\left(1-\rho^{*}_{\varepsilon_{mt}\varepsilon_{bt}}\right)}}}{\sum\limits_{s'\in S}\left\{\sum\limits_{b'\in B}\left(\sum\limits_{m'\in M}\exp\left\{\overline{V}_{(s'b'm')ht}\right\}\right)^{\frac{\left(1-\rho^{*}_{\varepsilon_{bt}\varepsilon_{st}}\right)}{\left(1-\rho^{*}_{\varepsilon_{mt}\varepsilon_{bt}}\right)}}\right\}} \qquad (28)$$

The log of the denominator of the first term in (28) is the expected maximum utility (or inclusive value) associated with the choice process for the lower level choice (e.g choice of destination). The log of the denominator of the middle term is the inclusive value associated with the season choice. The ρ^{*}_{ab} are a measure of the correlation between a and b. The joint probability of choosing a destination, airline and season mix is the product of two conditional probabilities and a marginal probability. Equation (28) may be simplified with the argument (1 - $\rho^{*}\varepsilon_{bt}\,\varepsilon_{st}$) set to unity for identification for the vth destination, mth airline and nth season:

$$P_{vmn} = P_{v|mn} \cdot P_{m|n} \cdot P_{n} \qquad (29)$$

$$P_{v|mn} = \frac{\exp\left\{\overline{V}\left(\beta' W_{v|mn}\right)\right\}}{\sum\limits_{v'\in V}\exp\left\{\overline{V}\left(\beta' W_{v'|mn}\right)\right\}} \qquad (30)$$

$$P_{m|n} = \frac{\exp\left\{\kappa' Y_{mn} + \lambda I_{mn}\right\}}{\sum\limits_{m'\in M}\exp\left\{\kappa' Y_{m'n} + \lambda I_{m'n}\right\}} \qquad (31)$$

$$P_{n} = \frac{\exp\left\{\gamma' Z_{n} + \tau J_{n}\right\}}{\sum\limits_{n'\in N}\exp\left\{\gamma' Z_{n'} + \tau J_{n'}\right\}} \qquad (32)$$

$$I_{mn} = \log\sum_{v'\in V}\exp\left\{\overline{V}\left(\beta' W_{v'mn}\right)\right\} \qquad (33)$$

$$J_n = \log \sum_{m' \in M} \exp\left\{\kappa' Y_{m'n} + \lambda I_{m'n}\right\} \tag{34}$$

A global sufficiency condition for a static nested logit model to be consistent with individual utility maximisation is that the parameters of inclusive value (i.e. l, t) be in the unit interval (McFadden 1984). In developing a functional form for the $\overline{V}_{qht}$ we assume that the forward and backward dynamic conditioning can be captured by two effects: *an expectations effect* and *an experience effect* To simplify notation, define vectors, G_{qst}, q = m, b, s, containing all attributes relevant to vacation package choice q. The expectations effect for attribute g_{iqht} is given by:

$$g^*_{iqht}(\theta) = \sum_{r=0}^{t-1} \theta^r g_{iqh,t-r} \tag{35}$$

where q is a fixed (estimable) parameter. The experience effect is designed to capture the influence of habit (occupation of previous states) and for attribute g_{iqht} is given by:

$$\overline{g}_{iqht}(\theta) = \sum_{r=0}^{t-1} \theta^r \left[\left|g_{iq^ch,t-r-1} - g_{iqh,t}\right|\right] \tag{36}$$

where g_{iq^cht} is the level of attribute g_i for the chosen alternative [c] in period t. q is constrained to take the same value for the expectations and experience effects and across attributes. The influence exerted by attribute levels pertinent to individual or group h prior to the commencement of the panel can be summarised in an initial conditions term:

$$\eta_{qho} = \lambda \sum_{r=0}^{\infty} \theta^r g_{qh,-r} = E_{qh}\left(\overline{V}_{qho}\right) \tag{37}$$

The utility function for choice q can now be written as:

$$V_{qht} = \overline{V}_{qht}\left(G^*_{qht},\ \widetilde{G}_{qht},\ h_{qho}\right) + \omega_{qht} \tag{38}$$

where $\omega_{qht} = \mu_{qht}/(1-\theta L)$, L being the lagged operator (Amemiya 1985) and μ a disturbance term.

Equation (38) is a particular solution of the constrained maximisation of a direct utility function which includes state variables describing the dependence of the current choice behaviour on past behaviour. Specifically, these state variables represent stocks of vacations and stocks of habits, including the cumulative effect on the present choice of the most recent continuous experience in a state as well as habit persistence (Hensher and Wrigley 1986). The state variable effect can be captured in the notion of rational habit formation, and that the *cost of the initial stock of habits* can be measured by lagged indices (given in the form of (35), (36) and (37)).

We have postulated the presence of two types of observable dynamic effects, an expectations effect (equation 35), and an experience effect (equation 36). The expectations effect measures the influence that present and past attribute values, and future expectations of these values, exerts on current choice behaviour. This effect accounts for the discounted expected utilities over time associated with a choice outcome. This is defined over the domain of each alternative in the choice set with the postulated exogenous influences being discounted by the q^r functional form. For example, if one undertook a vacation in wave t, then we assume that a household makes the choice aided by the knowledge of the vacation opportunity profile through time. Financial variables are good examples of dimensions displaying an expectations impact on choice. The experience effect accounts for the influence of previous behaviour on the current choice. For example, if the quality of service of an alternative airline in the current choice set is the same as that of the *chosen* airline in a previous period, one would expect, ceteris paribus, an increased probability of selecting that alternative in the current period. The probability would increase as the equivalence is repeated back through the past.

There are two ways we might estimate equation (38). One procedure is to use only a single period's data on the *choice* variable. With this approach serial correlation from any source does not arise, so that all mean parameter estimates and their standard errors are consistently estimated, conditional on q . Another procedure is to pool the data so that the choice variable applies to all periods. Under this approach the (uncorrected) disturbances will be heteroscedastic, since the unobserved component contains an unknown which varies across time (i.e. q^t), resulting in inconsistent parameter estimates. A way to resolve this situation, is to assume that $h_{qho} = h_{ho}$, that is, that expectations and experiences do not depend on the state occupied by the household prior to the panel period. A superior resolution is to correct for the heteroskedasticity by applying a correction weight to $\overline{V}_{qht}$ equal to $\theta^{2t}\left\{\sigma^2_{\eta_{qho}}/\sigma^2_{\omega_{qht}} + 1\right\}^{0.5}$ where s_e^2 is the variance of ε.

Models for choices in each wave and choices in all waves can be estimated, the latter with the application of the correction weight in order to give consistent mean estimates. The empirical weights for each wave are derived from the parameter estimates of one explanatory variable in the wave-specific models, so that the error variances are scaled by a constant proportion of these parameters. The unobserved heterogeneity in the sampled population which predates the observational period can be captured in a series of proxy variables, such as duration history dummy variables for categories of time periods each individual has undertook a vacation of a particular type. The empirical forms of equation (38) are given in equations (39) and (40).

$$\overline{V}_{mqt} = \kappa_0 + \kappa_1 \sum_{r=0}^{t-1} \theta^r \tau_0 c_{m_{h,t-r}} + \kappa_1 \sum_{r=0}^{t-1} \theta^r \tau_1 Y_h c_{m_{h,t-r}} - \kappa_1 \sum_{r=0}^{t-1} \theta^r F_{m_{h,t-r}}$$
$$+ \kappa_2 \sum_{r=0}^{t-1} \left(\theta^r b_{m_{hl,t-r}}\right) + \kappa_3 \sum_{r=0}^{t-1} \left(\theta^r b_{mkK,t-r}\right)$$
$$+ \sum_{k}^{K} \sum_{r=0}^{t-1} \kappa_{k+2} \theta^r \left[\left| b_{m_c hk,t-r-1} - b_{m hk,t} \right| \right] \qquad (39)$$

and

$$\overline{V}_{(sb)qt} = \xi_0 + \sum_{n}^{N} \sum_{r=0}^{t-1} \xi_n \left(\theta^r w_{hn,t-r}\right) + \sum_{k}^{K} \sum_{r=0}^{t-1} \xi_{N+k} \theta^r \left[\left| b_{(sb)c^{hk,t-r-1}} - b_{(sb)hk,t} \right| \right]$$
$$+ \xi_{N+K+1} (IC)_{h0} + \xi_{N+K+2} (IV)_{(sb)h} \qquad (40)$$

In addition to terms already defined, τ is an unknown consumer discount rate, as a function of individual or group income, F_{iq} is annual vacation cost, the b vector is a set of other exogenous variables, IC_{q0} is a proxy variable for the initial conditions, IV_{iq} is the inclusive value associated with the lower level season choice model, and κ and ξ are unknown parameter vectors. This model system has been implemented in another context, fully detailed in Hensher et al. (1992).

6. Conclusions

This paper has presented a number of themes in support of the development of an event-history based panel data strategy as a rich empirical basis for understanding and predicting the demand for tourism activities. The methods outlined are now popular in complementary literatures, especially the travel behaviour and demand literature. They deserve more serious consideration by tourism researchers. As Australia's major industry, we should recognise the benefits of more insightful investigations into the motivating forces influencing the size and composition of the domestic and international tourism sectors. The forecasting techniques discussed herein are an important aid in this process of investigation.

7. References

Amemiya, T. (1985) *Advanced Econometrics*, Basil Blackwell, Oxford.

Cox, D.R., (1972) *Regression Models and Life Tables.* Journal of the Royal Statistical Society, B34, 187-220.

Duncan, G.J. and Kalton, G. (1985) *Issues of design and analysis of surveys across time*, paper to Centenary Session of the International Statistical Institute, Amsterdam.

Breslow, N.E., (1974) *Covariance analysis of censored survival data.* Biometrics, 30, 89-99.

Econometric Software (1993) *LIMDEP 6.0*, Econometric Software, New York and Sydney.

Efron, B., (1977) *Efficiency of Cox's likelihood function for censored data.* Journal of the American Statistical Association, 72, 557-565.

Elandt-Johnson, R. and Johnson, N.L., (1982) *Survival Models and Data Analysis.* John Wiley and Sons, New York, New York.

Elbers, C. and Ridder, G. (1982) *True and spurious duration dependence: The Identifiability of the Proportional Hazard Model.* Review of Economic Studies, 44, 403-409.

Fleming, T. and Harrington, D., (1990) *Counting Processes and Survival Analysis.* John Wiley and Sons, New York, 1990.

Golob, J.M., Schreurs, L.J.M. and Smit, J.G. (1985) *The design and policy applications of a panel for studying changes in mobility over time*, in Behavioural Research for Transport Policy (pp. 81-95), VNU Press, Utrecht, The Netherlands.

Goodwin, P.B., Dix, M.C. and Layzell, A.D. (1987), *The case for heterodoxy in longitudinal analysis*, Transportation Research A, 21A(4/5), 363-376.

Goodwin, P.B. (1989) *Family changes and public transport use 1984-1987: A dynamic analysis using panel data*, Transportation, 16, 121-154.
Greene, W. (1992) *LIMDEP 6.0*, Econometric Software, New York and Sydney.

Guy, C., Wrigley, N., O'Brien, L. and Hiscocks, G. (1983) *The Cardiff Consumer Panel: A Report on the Methodology. Papers in Planning Research 68*, Department of Town Planning, University of Wales Institute of Science and Technology, Cardiff.

Hatanaka, (1974) *An efficient two-step estimator for the dynamic adjustment model with autocorrelated errors*, Journal of Econometrics, 10, 199-220.

Heckman, J. and Borjas, G. (1980) *Does unemployment cause future unemployment? definitions, questions and answers from a continuous time model of heterogeneity and state dependence.* Econometrica, 47, 247-283.

Heckman, J. and Singer, B. (1984) *A method for minimizing the impact of distributional assumptions in econometric models for duration data.* Econometrica, 52(2), 271-320.

Hensher, D.A. (1985) *Longitudinal surveys in transport: an assessment*, in Ampt, E.S., Richardson, A.J. and Brog, W. (eds.) New Survey Methods in Transport, VNU Science Press, Utrecht, 77-98.

Hensher, D.A. and Le Plastrier, V. (1985) *Towards a dynamic discrete-choice model of household* *automobile fleet size and composition*, Transportation Research B, 19B (6), 481-495.

Hensher, D.A. and Mannering, F. (1993) *Hazard based duration models and their application to transportation analysis*, Transport Reviews (in press).

Hensher, D.A. and Wrigley, N. (1986) *Statistical modelling of discrete choice in discrete time with panel data*, in Department of Public Works, Behavioural Research for Transport Policy, Utrecht: VNU Science Press BV, 97-118.

Hensher, D.A. and Raimond, T. (1992) *Behavioural panel surveys: an assessment of alternative specifications for studying the timing of change and adaptation*, paper to the First U.S. Conference on Panels for Transportation Planning, October 25-27, Lake Arrowhead, California.

Hensher, D.A. , Smith, N.C., Milthorpe, F.W. and Barnard, P.O. (1992) *Dimensions of Automobile Demand: A Longitudinal Study of Household Automobile Ownership and Use,* North-Holland, Amsterdam.

Kalbfleisch, J. and Prentice, R. (1980) *The statistical analysis of failure time data*, John Wiley & Sons, New York, New York.

Kessler, R.C. and Greenberg , D.F. (1981) *Linear Panel Analysis*, Academic Press, New York.

Kiefer, N. (1988) *Economic duration data and hazard functions.* Journal of Economic Literature, 26, 646-679.

Kitamura, R. (1987) A panel analysis of household car ownership and mobility, of Proceeding of Japan Society of Civil Engineers 338/IV-7 (Infrastructure Planning and Management), 13-37.

Kitamura, R. (1990) *Panel analysis in transportation planning: an overview*, Transportation Research , 21A(6), 401-415.

McFadden, D. (1984) *Econometric analysis of qualitative response models*, in Griliches, Z. and Intriligator, M.D. (eds.) Handbook of Econometrics, Vol. 11, Elsevier Science Publishers BV, 1395-1457.

Oak, D. (1977) *The asymptotic information in censored survival data.* Biometrika, 64, 441-448.

Raimond, T. and Hensher, D.A. (1992)*Panel Surveys and Other Longitudinal Techniques: An Annotated Bibliographic Review*, Institute of Transport Studies Working Paper No. 92-19, The University of Sydney.

Rao, P. and Miller, R.L. (1971) *Applied Econometrics*, Wadsworth Publishing Company, Belmont.