ELSEVIER

Journal of Econometrics 66 (1995) 325–348

JOURNAL OF Econometrics

Optimal stock/flow panels

Tony Lancaster*[,a], Guido Imbens[b]

[a]*Department of Economics, Brown University, Providence, RI 02912, USA*
[b]*Department of Economics, Harvard University, Boston, MA 02163, USA*

(Received March 1992; final version received July 1994)

Abstract

A stock/flow panel is a way of sampling a population of agents moving through a collection of discrete states. The scheme is to form separate samples of the residents of each state – the stocks – and of those moving between states – the flows. We calculate optimal stock/flow sampling schemes and provide efficient estimators of the transition intensities in the particular case of an alternating Poisson process. We also compute the efficiency gains compared to randomly sampled panels.

Key words: Endogenous sampling; Choice-based sampling; Panel; Cross-section
JEL classification: C33; C35; C41; C42

1. Introduction

Selection of people to form a panel can take place by randomly or exogenously sampling a population, or it can be done by sampling people in a way that depends upon the endogenous variables of the problem. In this paper we show that major gains can be obtained by balanced endogenous sampling. The same precision of estimation can be obtained from an endogenous sample observed for a short time as from a randomly selected panel observed for a long time.

* Corresponding author.

The authors acknowledge the support of the Institute for Research on Poverty, University of Wisconsin, Madison, and the comments of seminar participants at the Universities of Wisconsin, Groningen, Texas, Bristol, Harvard, and Tilburg. Some results of the present paper were circulated in IRP discussion paper no. 917–90, July 1990, entitled *Investigating Homelessness – A Renewal Theory Approach.*

SSDI 0304076940162O F

Endogenous panel selection is likely to lead to savings in sampling costs and to a reduction in biases due to nonrandom attrition.

These gains are made possible by the application of the recently developed method of moments procedure for efficient estimation from endogenous samples (Imbens, 1992; Lancaster and Imbens, 1991). This procedure solves the problem of simple, efficient estimation of parametric models using endogenous samples, a problem previously considered by Manski and McFadden (1981), Cosslett (1981), Hsieh, Manski, and McFadden (1986), and Hausman and Wise (1981).[1]

The question of panel design has been a major focus of research in the econometrics of discrete state, continuous time stochastic processes. The approach to the design problem in that literature has, however, typically been passive, emphasising the critical importance of taking account of the sampling scheme in the construction of the likelihood, particularly in models with neglected heterogeneity. Heckman and Singer's work, for example their paper of 1985, is perhaps the most outstanding example, and the effect of the sampling scheme on the likelihood is a major theme of Lancaster (1990). This approach is natural in a world in which most data used by econometricians are gathered by other people, for other purposes, and often in peculiarly complicated ways. In contrast the present paper focuses on the question of *optimal* (or at least pretty good) design – how should an econometrician gather his own sample?

The endogenous variable in the model studied in this paper is a two-state, alternating, continuous time stochastic process. The endogenous sampling scheme is one in which people are selected on the basis of the state, or sequence of states, that they occupy. In particular, four groups of the population are identified; the residents of each state at a point in time, and the people who change states in one direction or the other during a short interval of time. These groups form an endogenous stratification of the population and samples from each stratum are selected with predetermined probabilities. This is called a stock/flow sampling scheme. Some theory of stock/flow sampling was developed in Chesher and Lancaster (1983). Ridder, in his University of Amsterdam dissertation (1987), devised and implemented a stock/flow scheme for sampling the Dutch labour market. He also adapted the choice-based sampling technique of Manski and Lerman (1987) to provide a consistent estimator of the parameters of a transition model.

In order to realise the gains from endogenous sampling it is necessary for the investigator to supply a model for the conditional state occupancy probabilities

[1] The problem has also been studied in the biometric, technometric, and general statistical literature, usually under the name *case/control sampling*, by, for example, Prentice and Breslow (1978).

given the covariates. We solve this problem in the present work by ruling out unmeasured heterogeneity and time-varying covariates and by supposing that the process from which we sample is in stochastic equilibrium. In consequence, our paper represents a theoretical exploration of the potential gains from endogenous sampling and is not a template for immediate general application. The surprising magnitude of the gains revealed by our calculations indicate that research to make the procedure applicable to more general models may have large returns.

The research reported in this paper originated in our attempt to understand the sampling scheme used in a survey of homelessness among women and children in New York city.[2] This phenomenon, from the simplest statistical point of view, is a two-state alternating process – homeless and housed. The data for the survey comprised two distinct samples. One was a sample of people who were not homeless; this is a sample from the population of people who, at a particular point of time, occupied the state 'housed'. The other was a sample of women who, during a particular week, entered the city's emergency shelter system; this is a sample from the population of those who moved from one state to the other in that week. The former is a stock sample; the latter is a flow sample in the terminology of this paper.

In Section 2 of the paper we define a stock/flow sampling scheme. Section 3 describes the stochastic process that we shall assume in the theoretical calculations that follow. The information provided by a stock/flow panel can be decomposed into that provided by the cross-section and that provided by the panel. In Sections 4 and 5 of the paper we calculate these two components of the total information in the stock/flow panel and in Section 6 we calculate the total information. In Section 7 we report some calculations, both theoretical and empirical, of the information in cross-section and panel under alternative sample designs. Section 8 summarises the conclusions to be drawn and comments on their implications.

2. Stock/flow sampling

Consider a large population of people such that associated with each member is a realisation of a continuous time, discrete state, alternating process, and a time-invariant covariate vector, x. We identify the states by the labels 1 and 2. The objects of inference are the two transition intensities from state 1 to state 2, and from state 2 to state 1. In the next section we shall specify the process and functional forms for these transition intensities.

[2] Knickman, Weitzman, and Marcus (1989).

Let us identify four subgroups of the population by an indicator j such that

$$j = \begin{cases} 01 & \text{if a person moves from state 2 to state 1, at least once, in } t_0,\ t_0+\delta, \\ 02 & \text{if a person moves from state 1 to state 2, at least once, in } t_0,\ t_0+\delta, \\ 1 & \text{if a person resides in state 1 at } t_1, \\ 2 & \text{if a person resides in state 2 at } t_2, \end{cases}$$

where the times $t_j, j = 0, 1, 2$, and the interval δ are chosen by the sampler. We shall describe a sampling scheme in which the investigator selects at random from within each of the strata defined above. A sample from either stratum 1 or stratum 2 is a *stock sample* since it selects from the stock of members of each state at particular points of time. Examples might be samples of unemployed people or samples of homeless people. A sample from strata 01 or 02 is a *flow sample* since it selects from the set of people who flow between states during an interval. Examples might be samples of those registering as unemployed or entering shelters for the homeless on a particular day.

We define a stock/flow sample in the following way. Let $H_j, j = 01, 02, 1, 2$, be four probabilities summing to one. The sampler carries out N independent multinomial trials with these probabilities. If event j occurs, an individual is selected at random from subgroup j. The outcome of this sampling process is N observations of which N_j come from group j. These N_j are random subsample sizes (with expectations NH_j).[3]

When a person is selected we observe not merely the subgroup of which she is a member but also (a) a time-invariant covariate vector x and (b) her state biography for s periods immediately following t_1, t_2, or $t_0 + \delta$ as the case may be. By state biography we mean the times and directions of all changes of state. For example, we might select an unemployed person, observe that she is 28 years old and that in the 12 weeks after selection she resumed employment after 3 weeks and was still employed 9 weeks later.

A sampling scheme is a description of the way in which the population is to be sampled and a list of the data to be gathered from each sample member. In order to clarify further the nature of a stock/flow scheme we shall develop the joint distribution of the data to be gathered in such a scheme and compare it with the distributions for some other reasonable sampling schemes.

Let us denote by t the vector giving the times and directions of all changes of state during the observation period of length s after t_1, t_2, or $t_0 + \delta$, as the case may be. The symbol t stands for what we are calling the state biography. The

[3] This is a technically convenient model since it will lead to i.i.d. data. The asymptotic results given below do not depend on this assumption and would also apply if, for example, the N_j had been chosen by some deterministic mechanism.

total data available for a person selected according to a stock/flow scheme are j, the stratum from which she was selected, x the covariate, and t the state biography. The likelihood contribution of a single person is

$$p(j,x,t) = p(j)\,p(x\,|\,j)\,p(t\,|\,j,x). \tag{1}$$

In the stock/flow scheme $p(j)$ represents the multinomial mechanism with probabilities H_j by which the population is sampled; $p(x\,|\,j)$ represents the distributions of the covariate in the various sampling strata; finally, $p(t\,|\,j,x)$ represents the distribution of the state biography given the covariate and the stratum from which she was selected. Information about the transition intensities derives from two components of this distribution. The first is $p(x\,|\,j)$. The information here is analogous to that available from choice-based samples which also derives from the way in which the covariate distribution shifts between sampling strata. All the information in $p(x\,|\,j)$ can be readily extracted – at least asymptotically – since Imbens (1992) has provided a computationally simple semiparametrically efficient estimator from data obtained by choice-based sampling. The inverse of the asymptotic covariance matrix of this estimator provides a measure of the information in $p(x\,|\,j)$. We shall refer to this as *the cross-section information*. This is a slightly misleading phrase since two of the strata j refer to events happening over a finite interval of time, albeit one whose length, δ, is small relative to the mean time between transitions.

The second informative component of the likelihood is $p(t\,|\,x,j)$. This is a more conventional object, being just the likelihood for a panel continuously observed for s periods, conditional on j, which determines the initial state, and upon the covariate. We shall refer to the information arising from $p(t\,|\,x,j)$ as *the panel information*. In view of the structure of (1) the total information about the transition intensities is the sum of that provided by the cross-section and the panel. In Section 4 and later we shall evaluate these components of information.

The class of stock/flow schemes can be compared with a more conventional scheme in which we randomly sample the population at the fixed time t_0, observe their state at that time, say i, their covariate x, and their subsequent state biography, t. The likelihood is

$$p(i,x,t) = p(x)\,p(i\,|\,x)\,p(t\,|\,i,x). \tag{2}$$

The covariate distribution $p(x)$ provides no information about the transition intensities – x is exogenous in this scheme. The distribution of the state occupied at t_0, $p(i\,|\,x)$, provides no information about the two transition intensities separately. It is, with two states, a binary choice model from which at most one coefficient vector can be identified. The panel component $p(t\,|\,i,x)$ provides information about both transition intensities, and $p(i\,|\,x)$ augments such information.

A third scheme is one in which the population is stratified solely by the state occupied at t_0 and random samples are taken from within these two mutually

exclusive and exhaustive groups. This is just a stock/flow scheme with no flow, that is, H_{01} and H_{02} are zero. Here the cross-section likelihood is just a binary, but choice-based, likelihood. It can, by itself, provide no information about the two separate transition intensities for the reason given in the previous paragraph.

In order to obtain explicit and readily comprehensible answers we shall adopt the simplest interesting model – the equilibrium alternating Poisson process. The ultimate objective of the work is to obtain guidance about good ways of sampling two-state processes generated by econometrically reasonable models, which may be more complex than the equilibrium alternating Poisson.

3. A parametric model

Let the sequence of states occupied by someone whose covariate is x be a realisation of an equilibrium alternating Poisson process (APP)[4] and assume that realisations for distinct individuals are stochastically independent. The lengths of visits to state i are, therefore, exponentially distributed with means, say, $\mu_i(x)$, $i = 1,2$, and these variates are mutually stochastically independent. Let $\lambda_i = 1/\mu_i$, $i = 1,2$. We shall adopt the parametric model and notation

$$\mu_i(x) = \exp\{\beta_{i0} + \beta_{i1}x_1\}, \qquad i = 1,2,$$

$$\mu(x) = \mu_1(x) + \mu_2(x), \qquad \lambda(x) = \lambda_1(x) + \lambda_2(x). \tag{3}$$

The probability that an individual randomly selected from among those with covariate x occupies state i at any fixed date is

$$P_i(x) = \mu_i(x)/\mu(x), \qquad i = 1,2. \tag{4}$$

This is a consequence of the assumption that the processes are in stochastic equilibrium.

The probability that a randomly selected x individual moves from state 1 to state 2 in any interval of length δ is $\delta/\mu(x) + O(\delta^2)$ as $\delta \to 0$. This is also the probability that she moves from 2 to state 1. We shall write

$$P_{0i}(x,\delta) = \kappa_i(x,\delta)/\mu(x), \qquad i = 1,2,$$

where, for every x, $\kappa_i(x,\delta) = \delta + O(\delta^2)$ as $\delta \to 0$. These are the conditional probabilities of observing a randomly selected x person making each of the two possible changes of state in an interval of length δ. These transition probabilities

[4] The equilibrium APP is an alternating renewal process. Cox (1962) is the standard reference on Renewal Theory.

depend on the length of the sampling interval, δ. At a later stage in the analysis we shall let δ approach zero. We therefore also define

$$\bar{P}_{0i}(x) = \lim_{\delta \to 0} (1/\delta)(\kappa_i(x,\delta)/\mu(x)) = 1/\mu(x), \qquad i = 1,2.$$

These are the instantaneous probabilities of movement per unit time period – the transition intensities – for people with covariate x. An analysis in which $\delta \to 0$ is intended to capture the idea that the interval during which the flow sample is gathered is small compared to the mean lengths of stay in each state. The reasonableness of this specification depends upon the context of application. For example, people don't normally register as unemployed twice on the same day. Nor do they begin new jobs twice a day.

Since the limiting conditional flow probabilities for each direction of movement are identical, the two flows can, for part of our analysis, be treated as one stratum – *the flow*. We define

$$P_0(x,\delta) = \kappa(x,\delta)/\mu(x), \qquad \kappa(x,\delta) = 2\delta + O(\delta^2),$$
$$\bar{P}_0(x) = \lim_{\delta \to 0} (1/\delta)(\kappa(x,\delta)/\mu(x)) = 2/\mu(x). \tag{5}$$

The distribution of the covariate over the population is denoted $f(x)$. The marginal probabilities that randomly selected individuals will occupy each of the three groups are

$$Q_i = \int P_i(x) f(x)\,dx, \qquad i = 1,2,$$
$$Q_0(\delta) = \int P_0(x,\delta) f(x)\,dx, \tag{6}$$
$$\bar{Q}_0 = \lim_{\delta \to 0} Q_0(\delta)/\delta.$$

Note that the Q_i do not sum to 1 since the subgroups are, in general, neither mutually exclusive nor exhaustive.

The numbers Q_i can be known to the investigator, for example from census data. In the unemployment application Q_1 and Q_2 are the unemployment rate and its complement. $\bar{Q}_0$ is the instantaneous rate of movement per unit time period into and out of employment. In what follows the analysis is slightly different depending on whether or not the Q_i are assumed known to the econometrician. We shall assume for the most part that they are known, though we shall comment on the analysis in the contrary case and in our empirical calculations we shall report results under both hypotheses.

We now have a model in which the transition intensities are specified parametrically by β_1 and β_2. The single-agent likelihood, Eq. (1), is

$$p(j,x,t) = p(j)\,p(x\,|\,j)\,p(t\,|\,j,x).$$

The assumption of an alternating Poisson process and the specific forms, Eq. (3), determine $p(t|j,x)$ as a function of β_1, β_2 for given data. These assumptions do not determine $p(x|j)$ as a function of a finite parameter vector since this distribution involves the unknown population covariate distribution $f(x)$ as we shall show explicitly in the next section. Inference about β_1, β_2 from $p(x|j)$, which is a choice-based sampling likelihood, is a *semiparametric* problem but, as we remarked in the last section, a simple semiparametrically efficient estimator is known. So the specifications of this section suffice to enable us to proceed with efficient inference about the parameters of the two transition intensities.

The total single-agent likelihood, Eq. (1), is the product of

$$\mathscr{L}_1 = p(j,x) \quad \text{and} \quad \mathscr{L}_2 = p(t|j,x),$$

the cross-section and panel likelihoods. In the next three sections we shall inspect the information content of each component and of their product.

4. The cross-section information

This is the information in $p(j,x)$. The marginal distribution of the subgroup membership indicator is multinomial with probabilities H_j, with $H_0 = H_{01} + H_{02}$, by the definition of the stock/flow sampling scheme.[5] The conditional density of x given j can be found from $P_j(x)$, Q_j, and $f(x)$ by the law of conditional probability. Hence

$$p(j,x) = g(j)g(x|j) = H_j \frac{P_j(x)f(x)}{Q_j}. \tag{7}$$

In this expression $p(0,x)$ depends on δ, the flow sampling interval, since both $P_0(x,\delta)$ and $Q_0(\delta)$ do. We shall complete the parametric specification of the model by dividing numerator and denominator in $p(0,x)$ by δ and allowing δ to go to zero. Since both $P_0(x,\delta)/\delta$ and $Q_0(\delta)/\delta$ approach nonzero limits $\bar{P}_0(x)$ and $\bar{Q}_0$ where the latter is the expectation of the former with respect to $f(x)$ the resulting limit provides a proper joint probability distribution, $p(j,x)$. We shall call it the *equilibrium stock/flow sampling cross-section likelihood*. It is meant to be an appropriate likelihood when the flow sampling interval is small compared to the mean lengths of stay in each state.

In order to avoid a proliferation of notation we shall continue to describe the likelihood by the notation of (7). The ratio $P_0(x)/Q_0$ is to be interpreted as

$$\bar{P}_0(x)/\bar{Q}_0 = \mu(x)^{-1}/\int \mu(z)^{-1} f(z)\,dz.$$

[5] We are now pooling the two flow strata since they have identical probabilities.

The joint density for a single observation is therefore

$$\mathscr{L}_1(\beta) = \prod_{j=0}^{2} (H_j P_j(x)/Q_j)^{y_j} f(x), \tag{8}$$

where the y_j are binary indicators of group membership. This is just (7) written in indicator notation.

To develop the efficient GMM estimator of β_1, β_2 we require the marginal distribution of x and the conditional distribution of the strata given x *that are induced by the choice-based sampling scheme.* The efficient moment vector has elements which are the scores from the conditional likelihood induced by the sampling scheme and a pair of moments which extract the information provided by the covariate distribution induced by the scheme.

The marginal distribution of x induced by stock/flow sampling is found by summing (7) over j and is

$$g(x) = S(x)f(x) \quad \text{where} \quad S(x) = \sum_{j=0}^{2} H_j P_j(x)/Q_j. \tag{9}$$

Consequently the conditional distribution of j given x is

$$p(j \mid x) = \prod_{j=0}^{2} R_j^{y_j} \quad \text{where} \quad R_j = H_j P_j(x)/Q_j S(x). \tag{10}$$

Using the functional forms specified in (3) we find

$$R_j = \frac{\exp\{\theta_{j0} + \theta_{j1}' x_1\}}{1 + \exp\{\theta_{10} + \theta_{11}' x_1\} + \exp\{\theta_{20} + \theta_{21}' x_1\}}, \qquad j = 1, 2. \tag{11}$$

Here $R_0 = 1 - R_1 - R_2$ and

$$\theta_{j0} = \log(H_j \bar{Q}_0 / H_0 Q_j) + \beta_{j0}, \qquad \theta_{j1} = \beta_{j1}. \tag{12}$$

It follows that the likelihood based on the conditional distribution (10) is that for a multinomial logit model with cell probabilities R_j. The coefficients of the covariates in x_1 are the β_{j1} while the intercept terms involve both the β_{j0} and the H_j, Q_j. The intercept parameters, β_{j0}, are not identifiable from the cross-sectional likelihood without knowledge of the $\{Q_j\}$. Let us therefore proceed on the assumption that the $\{Q_j\}$ are known. We shall comment later on the effect of relaxing this assumption.

The algebra of estimation is most conveniently developed using matrix notation in which

$$p = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}, \quad r = \begin{pmatrix} R_1 \\ R_2 \end{pmatrix}, \quad h = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix}, \quad q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

and

$$P = \operatorname{diag}\{P_m\}, \qquad \bar{P} = \begin{pmatrix} P_1(1-P_1) & -P_1P_2 \\ -P_1P_2 & P_2(1-P_2) \end{pmatrix} = P - pp',$$

with analogous definitions for H, $\bar{H}$, Q, $\bar{Q}$, and R, $\bar{R}$. The conditional likelihood scores for β are

$$\psi_1 = (I \otimes x)(y - r), \tag{13}$$

and the covariate moments are

$$\psi_3 = r - h, \tag{14}$$

where $x' = (1\ x_1')$.

An efficient estimator of β is the value $\hat{\beta}$ solving $\bar{\psi}(\hat{\beta}, \hat{h}) = 0$ where $\bar{\psi} = (\bar{\psi}_1, \bar{\psi}_3)$ and the overbar indicates a sample average. Note that the parameter h is also estimated. The asymptotic covariance matrix of $\sqrt{N}(\hat{\beta} - \beta)$, $\sqrt{N}(\hat{h} - h)$ has the form $\Gamma^{-1}\Delta\Gamma^{-1\prime}$ where Δ is the moment covariance matrix and $\Gamma = \mathcal{E}(\partial\psi/\partial\beta, h)$. These matrices take the following forms:

$$\Delta = \mathcal{E}(\psi\psi') = \begin{pmatrix} \Delta_{11} & 0 \\ 0 & \Delta_{33} \end{pmatrix} = \begin{pmatrix} \mathcal{E}(\bar{R} \otimes xx') & 0 \\ 0 & \bar{H} - \mathcal{E}(\bar{R}) \end{pmatrix},$$

$$\Gamma = \mathcal{E}\begin{pmatrix} \partial\psi_1/\partial\beta & \partial\psi_1/\partial h \\ \partial\psi_3/\partial\beta & \partial\psi_3/\partial h \end{pmatrix} = \begin{pmatrix} -\mathcal{E}(\bar{R} \otimes xx') & -\mathcal{E}(\overline{RH^{-1}} \otimes x) \\ \mathcal{E}(\bar{R} \otimes x') & -(\bar{H} - \mathcal{E}(\bar{R}))\bar{H}^{-1} \end{pmatrix}. \tag{15}$$

Carrying out the multiplication we find that $\hat{\beta}$ and $\hat{h}$ are asymptotically uncorrelated with

$$\operatorname{var}\sqrt{N}(\hat{\beta} - \beta) = \Delta_{11}^{-1} - \bar{H}^{-1} \otimes J,$$

$$\operatorname{var}\sqrt{N}(\hat{h} - h) = \bar{H}. \tag{16}$$

Here, $J = jj'$ and $j' = (1\ 0\ \ldots\ 0)$ of order $1 \times K + 1$ where K is the dimension of β.[6]

The covariance matrix clearly depends upon the true value of β, the parameter being estimated and therefore so does the optimal design – choice of the H_j. To calculate the information content of cross-section and panel we shall use two strategies. First we shall evaluate the information and optimal design at a specific point in the parameter space, namely that in which the covariate has no effect

[6] Since $\hat{\beta}$ and $\hat{h}$ are asymptotically uncorrelated, an estimator of β with the same variance as (16) could also have been by minimising an appropriate quadratic form in $\bar{\psi}$ only with respect to β, using the known value of h in forming the moments.

on the transition intensities. Second we shall, in Section 7, estimate these information components using real data.

In that part of the natural β parameter space in which $\beta_{11} = \beta_{12} = 0$, so that the covariates have no effect on the process, $\bar{R}$ becomes equal to $\bar{H}$ and so nonstochastic. The conditional likelihood information matrix becomes $\Delta_{11} = \bar{H} \otimes \Sigma$ for $\Sigma = \mathscr{E}(xx')$, and

$$\operatorname{var} \sqrt{N}(\hat{\beta} - \beta) = \bar{H}^{-1} \otimes (\Sigma^{-1} - J).$$

This is a singular matrix, since $\Sigma^{-1} - J$ is, but the submatrix refering to the slope coefficients (of x_1) is

$$V_1 = \bar{H}^{-1} \otimes \Sigma^{22}, \tag{17}$$

where Σ^{22} is the $K \times K$ lower right submatrix of Σ^{-1}. The matrix V_1 is nonsingular, assuming the covariates are linearly independent, and it has determinant

$$|V_1| = |\bar{H}|^{-K} |\Sigma^{22}|^2.$$

But $|\bar{H}| = H_0 H_1 H_2$ which is maximised at $H_0 = H_1 = H_2 = \frac{1}{3}$, the equal shares solution. We conclude that when the true covariate effects are zero the stock/flow sampling scheme which maximises the cross-sectional information in the sense of minimising the generalised variance of the slope coefficient estimators is to take equal numbers from the flow and from both stocks.[7]

For future use we note that the value of V_1 at the optimal design is

$$V_1 = \begin{pmatrix} 6 & 3 \\ 3 & 6 \end{pmatrix} \otimes \Sigma^{22}. \tag{18}$$

5. The panel information

That part of the likelihood that derives from the panel information is conditional on the stock/flow sampling stratum, $j = 01, 02, 1$, or 2. Given this stratum the initial state for the panel is determined. Strata 01 and 1 imply initial state 1; strata 02 and 2 imply initial state 2. Since the probabilities of the strata given the initial state do not involve the parameter β, the panel likelihood contribution can be written as the distribution of the panel information given the initial state.

[7] Stock/flow sampling is, as we have emphasized, a type of choice-based sampling. The optimality of equal shares at zero slopes for the choice-based sampled multinomial logit model has been pointed out by Scott and Wild (1986). Actually this result is true for a much broader class of models than the logit as is shown in Lancaster and Imbens (1991).

For a single observation the likelihood for the panel data is the density function for an alternating Poisson process observed for s periods conditional on both the covariate x and the initial state i. This distribution is

$$p(t \mid i, x) = \mu_1^{-d_1} \mu_2^{-d_2} \exp\{ -\mu_1^{-1} T_1 - \mu_2^{-1} T_2 \}. \tag{19}$$

Here

$$\begin{aligned} d_1 &= \text{number of } 1 \to 2 \text{ transitions}, \\ d_2 &= \text{number of } 2 \to 1 \text{ transitions}, \end{aligned} \tag{20}$$

and T_s is the total time spent in state $s, j = 1, 2$.[8] The log-likelihood is therefore

$$L_2(\beta) = -d_1 x'\beta_1 - d_2 x'\beta_2 - T_1 e^{-x'\beta_1} - T_2 e^{-x'\beta_2}. \tag{21}$$

Let $w_s = T_s e^{-x'\beta_s}$ for $s = 1, 2$ and set $w' = (w_1, w_2)$, $d' = (d_1, d_2)$. Then the score vector is

$$\psi_4 = (w - d) \otimes x. \tag{22}$$

The zero mean property of scores tells us that

$$E(T_j e^{-x'\beta_j}) = E(d_j), \qquad j = 1, 2. \tag{23}$$

A second differentiation gives the Hessian as the block-diagonal matrix

$$\frac{\partial^2 L_2}{\partial \beta \partial \beta'} = -W \otimes xx', \tag{24}$$

where $W = \text{diag}\{w_s\}$. The information matrix is therefore

$$\mathcal{I}_2 = \mathcal{E}(D \otimes xx' \mid i, x), \tag{25}$$

where $D = \text{diag}\{d_s\}$.

A comparison of the information in the panel and in the cross-section may be made in that part of the parameter space in which the covariates have no effect on the process – the zero slopes set. In this set the information becomes

$$\mathcal{I}_2 = \mathcal{E}(D \mid i) \otimes xx'.$$

To make the comparison we need the information unconditional on initial state and covariate. This gives

$$\mathcal{I}_2 = \bar{D} \otimes \Sigma. \tag{26}$$

[8] (19) is just the density function for a right censored sequence of alternating indepenent exponential distributions.

Here,

$$\bar{D} = \mathscr{E}(D) = \begin{pmatrix} \mathscr{E}(d_1) & 0 \\ 0 & \mathscr{E}(d_2) \end{pmatrix}.$$

The slope covariance matrix is the appropriate submatrix of the inverse information, namely

$$V_2 = \bar{D}^{-1} \otimes \Sigma^{22}, \tag{27}$$

which may be compared to the corresponding result for the cross-sectional information, (17), which was $V_1 = \bar{H}^{-1} \otimes \Sigma^{22}$.

In the last section we were able to determine an optimal stock/flow sampling scheme using only the cross-sectional information in the special case in which the true covariate effects were zero. Let us examine this question for the panel likelihood under the same hypothesis. In order to make this calculation we need to find $\mathscr{E}(d_1)$ and $\mathscr{E}(d_2)$, the expected numbers of transitions in an interval of length s.

The expected numbers of $1 \to 2$ and $2 \to 1$ transitions in an interval of length s, which are the diagonal elements of $\bar{D}$, depend upon the probability distribution of the initial states. These in turn depend upon the probabilities H_j. For example, an observation from subgroup $j = 2$ implies that state 2 was the initial state. We find $\mathscr{E}(d_j)$ by first conditioning on the initial state and then averaging over the distribution of initial states implied by the stock/flow sampling scheme. We find[9]

$$\begin{aligned} E(d_1 \mid i = 1) &= \omega(s) + (\lambda_1/\lambda)(1 - \mathrm{e}^{-\lambda s}), \\ E(d_1 \mid i = 2) &= \omega(s), \\ E(d_2 \mid i = 1) &= \omega(s), \\ E(d_2 \mid i = 2) &= \omega(s) + (\lambda_2/\lambda)(1 - \mathrm{e}^{-\lambda s}). \end{aligned} \tag{28}$$

Here, i indicates the initial state, $\lambda_j = \mu_j^{-1}$ is the transition intensity out of state j, and

$$\omega(s) = (\lambda_1\lambda_2/\lambda^2)[\mathrm{e}^{-\lambda s} + \lambda s - 1], \qquad \lambda = \lambda_1 + \lambda_2.$$

Stock/flow sampling strata 01 and 1 imply initial state 1, strata 02 and 2 imply initial state 2. Thus the distribution of initial states is

$$P(i = 1) = H_1 + H_{01}, \qquad P(i = 2) = H_2 + H_{02}.$$

[9] Appendix 1 describes these calculations.

In consequence, the unconditional expected transition counts, under zero slopes, are

$$E(d_1) = \omega(s) + (H_1 + H_{01})(\lambda_1/\lambda)(1 - e^{-\lambda s}),$$
$$E(d_2) = \omega(s) + (H_2 + H_{02})(\lambda_2/\lambda)(1 - e^{-\lambda s}). \tag{29}$$

As $s \to \infty$ and the panel becomes longer the dependence of the $E(d_j)$ on the probability distribution on the initial states, and hence on the H_j, becomes negligible. This is because $\omega(s)$ becomes linear in s while the component involving H_j becomes constant. So for long panels it does not matter what stock/flow sampling scheme was used. On the other hand for small s – short panels – $\omega(s)$ is $O(s^2)$ as $s \to 0$ and

$$E(d_1) = p\lambda_1 s + O(s^2),$$
$$E(d_2) = (1 - p)\lambda_2 s + O(s^2),$$

where $p = H_1 + H_{01}$. Thus, for small s, $|\bar{D}| \propto p(1 - p)$ and optimal stock/flow sampling is any choice of the H_j such that $H_1 + H_{01} = 0.5$. An equal shares stock/flow scheme $H_j = \frac{1}{3}$, $j = 1, 2$, and $H_{01} = H_{02} = \frac{1}{6}$ satisfies this condition.

We conclude that for short panels an equal shares stock/flow scheme is optimal when the regression slopes are at, or close to, zero. How short is short? And how close is close? In Section 7 we report some calculations on real data (though with a rather unreal model).

6. The total information

We may pool the panel and cross-section information in a method of moments procedure which uses the cross-section moments, ψ_1 and ψ_3, of (13) and (14) together with the panel likelihood scores, (22), as a third moment set, ψ_4. Let $\psi = (\psi_1, \psi_3, \psi_4)$. The covariance matrix of ψ is block-diagonal because ψ_4 has mean zero conditional on j, x while ψ_1, ψ_3 depend only upon j, x. The efficient pooled estimator has asymptotic covariance matrix

$$V = (\Delta_{11} + \Gamma_2' \Delta_{33}^{-1} \Gamma_2 + \mathcal{I}_2)^{-1}, \tag{30}$$

where $\mathcal{I}_2$ is the panel information matrix for β, (26) and $\Gamma_2 = \mathcal{E}(\bar{R} \otimes x')$. Inverting and taking the limit as the slope coefficients approach zero gives the zero slope covariance matrix,

$$V = (\bar{H} + \bar{D})^{-1} \otimes (\Sigma^{-1} - J). \tag{31}$$

The slope estimator covariance matrix is

$$V_1 = (\bar{H} + \bar{D})^{-1} \otimes \Sigma^{22}, \tag{32}$$

which may be compared with (17) and (26).

We may compare these results with the precision of random cross-section sampling, i.e., a random sample of the population at an instant of time taken without regard to the state people occupy. Under this scheme the distribution of initial states is

$$P(i=1)=\lambda_2/\lambda, \qquad P(i=2)=\lambda_1/\lambda. \tag{33}$$

Substituting these into (5.11) in place of $H_1 + H_{01}$ and $H_2 + H_{02}$ gives

$$\mathscr{E}(d_1)=\mathscr{E}(d_2)=(\lambda_1\lambda_2/\lambda)s=s/\mu. \tag{34}$$

The likelihood for a random panel is the product of the likelihood given the initial state, (19), and the probability distribution of the initial state, given x. Under random sampling x is exogenous and its distribution uninformative about β. A routine calculation then provides the information matrix in the zero slopes set as

$$\mathscr{I}_R=(\bar{H}_R+\bar{D}_R)^{-1}\otimes\Sigma, \tag{35}$$

where

$$\bar{H}_R=\frac{1}{\lambda^2}\begin{pmatrix}\lambda_1\lambda_2 & -\lambda_1\lambda_2\\ -\lambda_1\lambda_2 & \lambda_1\lambda_2\end{pmatrix},$$

and the diagonal elements of $\bar{D}_R$ are both equal to s/μ. The matrix $\bar{H}_R$ is got from $\bar{H}$ by putting $H_1=\lambda_2/\lambda=\mu_1/\mu$, $H_2=\lambda_1/\lambda=\mu_2/\mu$, $H_0=0$. This reflects the fact that a random sample is equivalent to a stock/flow sample in which no observations are taken from the flow and the stocks are sampled in the same proportions as those stocks occupy in the population. The singularity of $\bar{H}_R$ reflects the fact that when no observations are taken from the flow, the model reduces to a *binary* logit from which two coefficient vectors, β_1 and β_2, cannot be identified.

The slope estimator covariance matrix at zero slopes is

$$V_R=(\bar{H}_R+\bar{D}_R)^{-1}\otimes\Sigma^{22}. \tag{36}$$

This is to be compared with the corresponding matrix for general stock/flow sampling, (32). We shall report such a comparison in the next section.

7. Some numerical calculations

We shall use some data from the Dutch labor market[10] in order to get an idea of the practical relevance of our results. A sample of 372 men were observed over

[10] Appendix 2 gives further details about the data.

Table 1
Numbers of transitions in each direction in seven years

Transitions	Frequency $1 \rightarrow 2$	Frequency $2 \rightarrow 1$
0	338	268
1	29	98
2	4	5
3	1	1
Total	372	372

the 84 months from January 1977 to December 1983. They were observed to move between the two states *not employed*, state 1, and *employed*, state 2. A single covariate, age − 35 years, was also observed. This was approximately uniformly distributed from − 15 to 15 with variance $\sigma^2 = 75$. The matrix Σ was therefore

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 75 \end{pmatrix}.$$

Modelling the process as a conditional alternating Poisson process the ML estimates of the regression functions were

$$\mu_1(x) = \exp\{4.0 + 0.043x\}, \qquad \mu_2(x) = \exp\{5.6 + 0.030x\}. \tag{37}$$

Thus $\beta_{10} = 4.0$, $\beta_{11} = 0.043$, etc. The numbers of transitions in this sample had the frequency distribution given in Table 1.

Thus 338 men had no transitions into employment; 268 had no transitions out of employment, etc. A total of 138 transitions were observed in the 84 months, or about 1 transition for every 3 men. This is therefore a group displaying little movement between states. The average number of $1 \rightarrow 2$ transitions is 0.017; the average number of $2 \rightarrow 1$ transitions is 0.298.

We shall report two sets of calculations.

7.1. Random and equal shares stock/flow sampling at zero slopes

In this comparison we shall compare the covariance matrices (32) and (34). For the μ_j we set the slopes in (37) equal to zero giving

$$\mu_1 = \exp\{4\}, \qquad \mu_2 = \exp\{5.6\},$$

with λ_1 and λ_2 the reciprocals of these. One way of comparing these matrices is by calculating the square root of the determinant of $\bar{H} + \bar{D}$ and comparing it to the square root of $|\bar{H}_R + \bar{D}_R|$. These quantities are proportional to the square

Table 2
The information in random and equal shares panels

Panel length in months	Equal shares panel	Random panel
1	0.20	0.03
12	0.26	0.11
84	0.55	0.37

roots of the inverse generalised variances for the covariate effects. The comparison is given in Table 2.

These people, who move between states very infrequently, supply virtually no information about the transition intensities when randomly sampled and observed for only a month. In contrast, there is a good deal of information in the cross-section which can be extracted with a stock/flow design. This general pattern is confirmed in the second set of comparisons which allow for the effect of the covariate.

7.2. Simulation of alternative sampling schemes

In this calculation we simulate 50,000 realisations of an alternating Poisson process with parameter values as in (37). We then sample these realisations (a) according to various stock/flow sampling schemes and (b), for comparison, randomly. Averaging over repeated samples we calculate the covariance matrices of $\sqrt{N}(\hat{\beta}-\beta)$. We use three different panel lengths, s, namely 1 month, 2 months, and 84 months. We use two different stock/flow schemes in both of which there are four subgroups or sampling strata, namely

$$\begin{aligned} &01 \quad \text{the flow } 2 \to 1 \text{ during } \delta, \\ &02 \quad \text{the flow } 1 \to 2 \text{ during } \delta, \\ &1 \quad \text{the residents of state 1,} \\ &2 \quad \text{the residents of state 2.} \end{aligned} \tag{38}$$

These groups are sampled with two sets of probabilities; see Table 3. Under random sampling the residents of state 1 are sampled with probability equal to the fraction of the population resident in state 1. This is calculated to be 0.17. Similarly for state 2. The equal shares stock/flow scheme is optimal under zero slopes and when the panel observation is brief.

We shall report asymptotic variances of coefficient estimates under both sampling schemes. In our theoretical development we have assumed that the marginal stratum probabilities, the $\{Q_j\}$, are known. This is because the inter-

Table 3
Two stock/flow sampling schemes

Stratum	Random	Equal shares
01	0.00	1/6
02	0.00	1/6
1	0.17	1/3
2	0.83	1/3

Table 4
Asymptotic variances of coefficient estimates under alternative sampling schemes and observation periods

Scheme	Panel length	Q's known	β_{10}	β_{11}	β_{20}	β_{21}
Random	84	No	3.03	0.04	2.99	0.04
	12	No	16.14	0.21	16.14	0.21
	12	Yes	1.34	0.21	1.16	0.21
	1	Yes	14.58	2.47	14.37	2.47
Equal shares	12	No	10.31	0.05	40.70	0.07
	12	Yes	0.23	0.05	0.19	0.07
	1	Yes	0.26	0.08	0.24	0.08

cept parameters β_{10}, β_{20} are unidentifiable, when regression effects are zero, from a stock/flow cross-section without such prior information. The slope parameters β_{11}, β_{21} are always identifiable from a stock/flow sample as long as some observations are taken from both stocks and the flow and the regressors are linearly independent. Panel data allows identifiability of all parameters, so this is an essentially minor complication. In our calculations we have estimated the parameters both under the assumption that the Q's are known and under the assumption that they are unknown. The nonidentifiability of the intercepts, under zero slopes, from a cross-section can be expected to show up in a high variance for these parameters when the panel length, s, is short.

The results are reported in Table 4 and in Figs. 1 and 2. The intercepts are indeed poorly determined from short panel data when the Q's are not known as compared to when they are known. By, contrast, the presence or absence of knowledge of the Q's has no effect on the accuracy of the slope estimates.

Turning to the numerical values of the variances given in the table we can make one interesting observation. Consider the variances of the slope coefficient estimates for the two stock flow schemes which are 0.08 for the 1-month scheme and 0.07 for the 12-month scheme. If we refer back to Section 4, we gave there the covariance matrix of the equal shares stock/flow sampling scheme with no

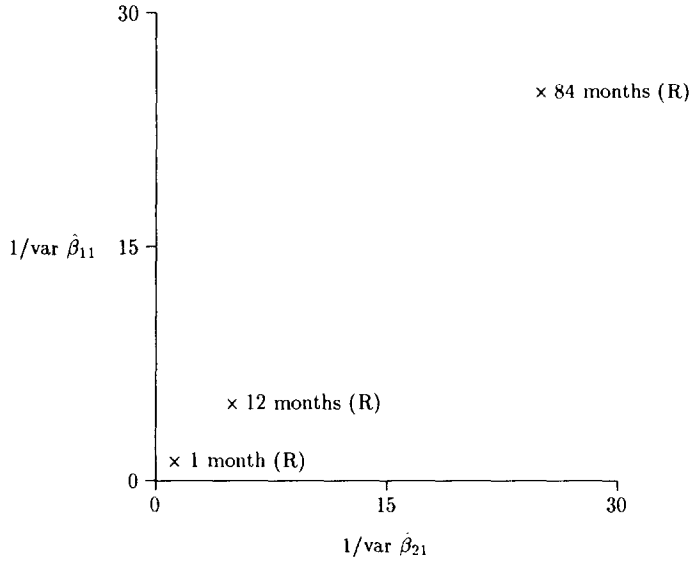

Fig. 1. Random panels of different lengths.

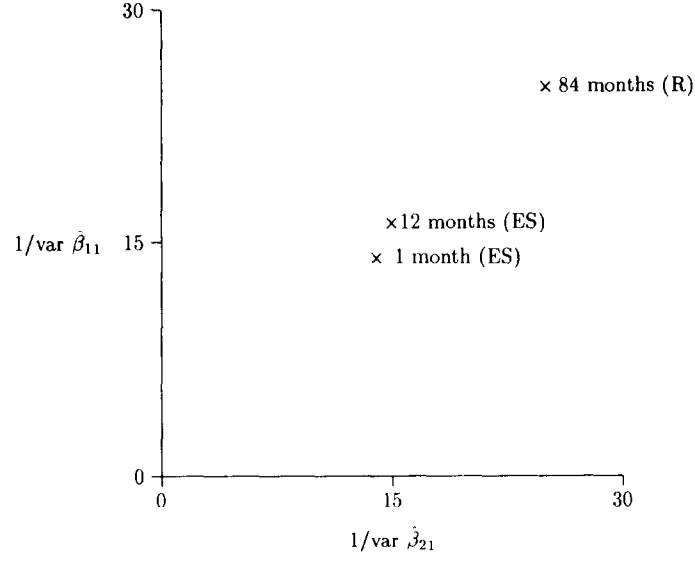

Fig. 2. Stock/flow and random panels of different lengths.

biographical data when the true slopes were zero. In particular the slope variances were equal and given by $3 \times 2 = 6$ times the lower right element of Σ^{-1}. But Σ is diagonal with $\sigma_x^2 = 75$. Hence with no biographical information and *zero true slopes* we should find the variance of the slope estimators to be both equal to $6/75 = 0.08$. This is precisely the value of these variances under equal shares stock/flow sampling with one months observation. Hence we can argue that (a) 1-month observation virtually amounts to not having biographical data at all, but, more importantly, (b) zero slope variances can be a rather good guide to the asymptotic variances when the true slopes are not zero.[11]

Comparison of the precision of estimators from an equal shares stock/flow panel with those from a random panel is best done graphically. In Figs. 1 and 2 the axes measure the precision – one over the variance – of the two slope coefficients estimates $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$, or rather of $\sqrt{N}(\hat{\beta}_{j1} - \beta_{j1})$. In each figure, the further a point is from the origin the more precise the estimate. If two points fall on the 45° line through the origin and point A is twice as far from the origin as point B, then the precision of estimator A could be achieved with just twice as many observations using the estimator associated with point B. The number next to each point is the length of the panel in months. R stands for cross-section random sampling and ES for equal shares stock/flow sampling.

Fig. 1 compares random sampling with biographical information for different lengths of time. A lot of precision is lost if we randomly sample and observe only for 12 months instead of for 84. Since the point labelled 84 is about four times further from the origin as the point labeled 12, we see that observing for seven times as long is equivalent to having four times as many observations.

Fig. 2 compares equal shares panels of 1 and 12 months with a random panel observed for 84 months. Compared with Fig. 1 we see a dramatic improvement in the precision of the shorter observations schemes. From a comparison of the points labelled 1 and 84 we see that the same precision of estimation could be achieved with (a) N randomly sampled people observed for 7 years or (b) $2N$ stock/flow sampled people observed for 1 month. This is broadly consistent with the figures reported in the last subsection and is a quite remarkable result. The explanation is that these people move only slowly between states so that if, as with random sampling, a relatively large proportion of the information comes from the biographies, long biographies are required before much information is obtained. In a more volatile population biographical information may be expected to be more informative as compared to the information in the covariate distribution, i.e., in the cross-section.

[11] This may be another reflection of the observation that equal shares sampling, which is optimal for zero slopes, is often nearly optimal when slopes are not zero.

8. Conclusions

When people move only slowly from state to state, they must be observed for a long time before a panel yields much information about the way in which covariates modify the speeds of transition. The time series dimension of the data is relatively uninformative. The variation in the covariates between people observed in different states at a moment in time or flowing between states in a short interval can, by contrast, yield a good deal of information about these effects. Doubtless this qualitative result has been long appreciated but this paper shows, using real labor market data, that the quantitative difference can be startlingly large. A randomly selected panel observed for 7 years provides about as much information as a judiciously designed endogenous cross-section observed for less than 1 year.

A cross-sectional sample of people selected according to the states they occupy is a choice-based (endogenous) sample. In this paper we have analysed a particular class of endogenous sampling scheme, the stock/flow sample. The information about covariate effects in such a sample can be obtained rather simply using the recently developed efficient method of moments estimator for endogenous samples. This is a way of making inferences from the distribution of the covariates given the states without adopting a parametric model for the population covariate distribution.

To extract the information in the cross-sectional covariate distribution requires that the investigator supply a model for the conditional probabilities of state occupancy and of appearance in the flow, given the covariates. That is, he must supply a model explaining why a person with covariate x should be in state j at the sampling date, or why she should be moving from i to j during a short observational interval. We 'solved' this problem in the present paper by the assumption that the stochastic process we were observing was in stochastic equilibrium. Actually this solution leads to a model that is parametrically overidentified. The cross-sectional likelihood is a multinomial logit model with four cells, when the flow is separated into its components, but only two functionally independent parameter vectors. It is thus possible to develop model specification tests. One way to do this might be to allow the two conditional flow probabilities to have different parameter vectors and to test for their equality. It would also be possible to develop tests for neglected heterogeneity, and, indeed, to enlarge the model to allow for such effects.

It would also be of interest to examine the value and use of repeated, balanced, stock/flow cross-sections as an alternative to a single stock/flow panel. There are numerous other possibilities, including relaxing our functional form restrictions. We have yet to examine these options.

The model that we have examined makes a number of simplifying assumptions. Even though these are testable, as we have just pointed out, we take the view that this particular model is not one that an investigator would want to

entertain seriously in a practical study, except perhaps as a preliminary calculation. Moreover our conclusions about optimal design are limited in that they apply, strictly, only to a particular region of the parameter space. And, of course, the optimality criterion, generalised variance, ignores the crucial issue of the costs of alternative designs.

The aim of this paper has been to define a stock/flow panel, to show how recent work of inference from choice-based samples solves the problem of inference from stock/flow panels, and to examine both theoretically and with real data the consequences of alternative designs.[12] Our results show that major gains can, in principle, be made by gathering panels from balanced stock/flow panels and that further research into the use of such panels with other models may well be fruitful.

Appendix 1: Expected numbers of transitions in the APP

These expectations follow from inversion of the Laplace transform of the renewal function $H(t) = E(N_t)$ which gives the expected numbers of events in a renewal process in an interval of length t. The alternating Poisson process starting in state 1 for which events are transitions to state 1 is an ordinary renewal process whose interevent distribution is the convolution of two exponential distributions, $g_1(s)$, $g_2(s)$, with means μ_1, μ_2. The Laplace transform of the renewal function is[13]

$$H_1^*(s) = \frac{1}{s}\frac{g_1^*(s)g_2^*(s)}{1 - g_1^*(s)g_2^*(s)}. \tag{A.1}$$

where $*$ indicates a Laplace transform. The alternating Poisson process starting in state 1 for which events are entrances to state 2 is a modified renewal process for which the waiting time to the first event has density $g_1(s)$ and whose subsequent interevent times have density the convolution of g_1, g_2. The Laplace transform of the renewal function is

$$H_2^*(s) = \frac{1}{s}\frac{g_1^*(s)}{1 - g_1^*(s)g_2^*(s)}. \tag{A.2}$$

[12] In its methodological approach our work is similar to that of de Stavola (1986) who studied the effect on parameter estimates of different ways of sampling an alternating process. She also studied a very simple process but her general conclusions, like ours, are likely to give useful guidance in more complex contexts.

[13] Cox (1962).

Since the Laplace transform of the exponential density function with parameter λ is $g^*(s) = \lambda/(\lambda + s)$ explicit forms for (A.1) and (A.2) are

$$H_1^*(s) = \frac{\lambda_1\lambda_2}{s^2(s+\lambda)}, \qquad H_2^*(s) = \frac{\lambda_1\lambda_2}{s^2(s+\lambda)} + \frac{\lambda_1}{s+\lambda}. \tag{A.3}$$

A partial fraction expansion of the first factor in these expressions gives

$$\frac{\lambda_1\lambda_2}{s^2(s+\lambda)} = \frac{\lambda_1\lambda_2}{\lambda}\left[-\frac{1}{s\lambda} + \frac{1}{s^2} + \frac{1}{\lambda(s+\lambda)}\right].$$

Using this expansion and a table of Laplace transforms gives the results used in the body of the paper.

Appendix 2: Data details

The data used in Section 4 are from the ORIN data set and form a random sample of size 372 from that part of the male population that was between 23 and 53 years of age in 1977. Their labour market histories have been recorded for the 84 months between January 1977 and December 1983.

The standard errors of the coefficient estimates that have been used as the θ^*'s in Section 4 are

$$\beta_{10} = 0.11, \quad \beta_{11} = 0.013, \quad \beta_{20} = 0.10, \quad \beta_{21} = 0.012.$$

The mean age was 34.9 years. The average total time spent in state 1 – not employed – was 15.7 months, so the average time spent employed was 68.3 months. The marginal probability of being in state 1, Q_1 was calculated to be 0.17. The marginal probability of a transition per unit time period, Q_3, was 0.0031. A typical individual with age 35 would be expected to complete a cycle through the states in about 325 months.

References

Chesher, A.D. and T. Lancaster 1983, The estimation of models of labour market behaviour, Review of Economic Studies 50, 609–624.

Cosslett, S.R., 1981, Efficient estimation of discrete choice models, in: Manski and McFadden (1981a).

Cox, D.R., 1962, Renewal theory (Chapman and Hall, London).

de Stavola, B.L., 1986, Sampling designs for short panel data, Econometrica 54, 415–424.

Hausman, J.A. and D. Wise, 1981, Stratification on endogenous variables and estimation: The Gary income maintenance experiment, in: Manski and McFadden (1981a).

Heckman, J.J. and B. Singer, 1985, Social science duration analysis, in: J.J. Heckman and B. Singer, eds., Longitudinal analysis of labor market data (Cambridge University Press, Cambridge).

Hsieh, D.A., C.F. Manski, and D. McFadden, 1985, Estimation of response probabilities from augmented retrospective observations, Journal of the American Statistical Association 80, 391, 651–662.

Imbens, G.W., 1992, An efficient method of moments estimator for discrete choice models with choice-based sampling, Econometrica 60, 1187–1214.

Imbens, G.W. and T. Lancaster, 1990, Efficient estimation and stratified sampling, Mimeo. (Department of Economics, Harvard University, Cambridge, MA).

Kinckman, J., E. Weitzman, and E. Marcus, 1989, A study of homeless families in New York City (Department of Social Administration, New York University, New York, NY).

Lancaster, T., 1990, The econometric analysis of transition data (Cambridge University Press, Cambridge).

Lancaster, T. and G.W. Imbens, 1991, Choice-based sampling: Inference and optimality, Mimeo. (Department of Economics, Brown University, Providence, RI).

Manski, C.F. and S. Lerman, 1977, Estimation of choice probabilities from choice-based samples, Econometrica 45, 1977–1988.

Manski, C.F. and D. McFadden, eds., 1981a, Structural analysis of discrete data with econometric applications (MIT Press, Cambridge, MA).

Manski, C.F. and D. McFadden, 1981b, Alternative estimators and sample designs for discrete choice analysis, in: Manksi and McFadden (1981a).

Prentice, R.L. and N.E. Breslow, 1978, Retrospective studies and failure time data, Biometrika 65, 153–158.

Rao, C.R., 1973, Linear statistical inference, 2nd ed. (Wiley, New York, NY).

Ridder, G., 1987, Life cycle patterns in labor market experinece, Ph.D. dissertation (University of Amsterdam, Amsterdam).

Scott, A.J. and C.J. Wild, 1986, Fitting logistic models under case control or choice-based sampling, Journal of the Royal Statistical Society B 48, 170–182.