University of Wisconsin-Madison

# IRP Discussion Papers

Tony Lancaster
Guido Imbens

# INVESTIGATING HOMELESSNESS: A RENEWAL THEORY APPROACH

DP # 917-90

Institute for Research on Poverty
Discussion Paper no. 917-90

**Investigating Homelessness:
A Renewal Theory Approach**

Tony Lancaster
Department of Economics
Brown University

Guido Imbens
Brown University
and
Tilburg University

July 1990

This paper was prepared for the Institute for Research on Poverty, University of Wisconsin-Madison, under its Small Grants program, sponsored by the U.S. Department of Health and Human Services. Any opinions expressed in this paper are solely those of the authors.

**Abstract**

This paper deals with statistical methods for analyzing samples of data in which individuals move alternately between two discrete states. The intended application is to people suffering repeated spells of homelessness. We model the history of an individual as an alternating renewal process and then use this stochastic framework to pose and answer a number of questions. We show how to efficiently analyze data gathered by randomly sampling the occupants of each state and by sampling people changing state--becoming homeless, for example. We offer a partial solution to the question of how best to design a survey of this state. Finally, we study how to handle data giving the state history of individuals over a period of time. In particular we show how an optimally designed sample of people observed for a brief period of time can be as informative as a random sample observed for a prolonged period. This latter result appears to promise significant economies in the design and analysis of surveys of the homeless and allied phenomena.

## 1. Introduction.

Homelessness is surely a phenomenon worthy of serious scientific study. This is particularly so when those involved are children. Of the methods required for such a study, statistical analysis is central. The work reported here is a contribution to the development of effective methods for the statistical analysis of homelessness. [1]

Viewed purely as a statistical phenomenon, homelessness has three outstanding characteristics. It is binary; one is either homeless or one is not. It is rare; most people are never homeless. It is of variable duration; some spells of homelessness are brief, others prolonged. These characteristics have implications both for the design of surveys of homeless, or potentially homeless, people, and for the statistical analysis of the results. It is these implications that we propose to explore.

The present work originated out of a consideration of a survey of homeless women and children in New York city.[2] This survey gathered data on about twelve hundred women in the city who were pregnant or with children. About half of these women were obtained by sampling applicants for shelter at the city's Emergency Assistance Units. The remainder were sampled from the population of welfare recipients in the city. The information obtained from each woman consisted of a list of personal characteristics, which we shall refer to as *covariates*. These were supplemented by taking from administrative records a partial history of each woman's previous shelter visits, if any. A number of questions naturally arise out this survey. What is a sensible family of models with which to analyse the data ? Was it a good idea to sample women in this way and are there better schemes ? Could one have done better by using different proportions of women from each of the two sampling modes ?

The main, but not exclusive, focus of the present paper is on the analysis of the covariate data. We consider a family of models based on a view of homelessness as an alternating renewal process. The elements of the theory of such processes are sketched in section 2 of the paper. The principal advantage of this point of view is that it enables one to gain a clear understanding of a sampling scheme, such as that used in the New York study, in which some people are sampled while they are becoming homeless and others

[1]The authors of this paper wish to acknowledge the comments of C.F. Manski, discussant when the paper was presented at an IRP Workshop, and Renya Reed.

[2]Knickman et al. (1989).

are sampled when they are not homeless. The principal disadvantage is that it is not an approach which readily allows the investigator to model the effect of covariates which are changing over time. This latter is, however, a problem the solution to which is not well understood anywhere in the literature at the time of writing. Moreover, the model is rather sensitive to misspecification due to neglected heterogeneity. This exploration of the renewal model finds its main justification in the way it facilitates the study of the question of how to interpret and design sampling schemes for a two state process. A more flexible class of models is probably necessary for applied work designed to get accurate estimates of policy relevant parameters.

In section 3 of the paper we apply the alternating renewal model to the problem of analysing the covariate information. Here we point out the connection between the analysis of covariate data gained by sampling an alternating renewal process and the problem of inference from choice or response based samples — called case/control samples in the biometric literature. In this section we identify a simple procedure for efficient estimation for a class of ways of sampling an alternating renewal process and give a result showing the equivalence of two apparently different estimators. We also consider the question of how best to sample such a process and we are able to add a little to the literature on this difficult question. Finally in this section we consider the question of how estimators are affected when the simplifying assumptions of the renewal model are violated in what seems a plausible way.

In section 4 we examine some of the preceding questions when the data include not only the covariate information but also give sections of the state history — the times of entrances to and exits from each state. Here we use the numerical values of parameters and actual covariate distributions taken from an alternating model fitted to some Dutch labor market data. The principal issue examined here is the potential gains from sampling these data in various 'response–based' ways.

It should be pointed out that the present enquiry is motivated by the problems posed by a rather special type of homelessness, that of women who move in and out of emergency shelter. Such women are readily counted and listed. It is therefore relatively easy to devise and analyse alternative schemes for sampling them. It may well be that for most homelessness, that of men or youths or childless women, the principal statistical problem is is precisely that of counting and listing the homeless. We do not address this question and we concede that in consequence our work may in fact may reasonably be regarded as missing the main statistical point, so far as much homelessness is concerned.

## 2. Renewal Processes.

**2.1 Basic Theory.** An *ordinary renewal process* is a sequence of independent and identically distributed non-negative random variables, $X_1, X_2, \ldots$ These quantities may be interpreted as the times between events, an event, or renewal, occurring at times $X_1, X_1 + X_2, X_1 + X_2 + X_3, \ldots$ Associated with any such process there are two sequences of random variables

((1)) $$S_r = X_1 + X_2 + X_3 + \ldots + X_r, \qquad r = 1, 2, \ldots$$

$N_t$, defined as the number of renewals occurring in $(0, t)$.

$S_r$ is the time until the $r$'th renewal.

The common distribution of the $\{X_s\}$ will be taken, in what follows, to be absolutely continuous with probability density function $f(x)$.

A *modified renewal process* is defined as above except that the density of the first time, $X_1$, namely $f_1(x)$, may be different from that of the others, $f(x)$.

An *alternating renewal process* (ARP) is a sequence of independent and identically distributed pairs of non-negative random variables $T_1, S_1, T_2, S_2, \ldots$ in which all $T$'s and $S$'s are stochastically independent. $T_j$ can be interpreted as the time spent in the $j$'th visit to state 1, and $S_j$ as the time spent in the $j$'th visit to state 2. The process is alternately in state 1 then in state 2. These times have absolutely continuous distributions with density functions $g_1(t), g_2(s)$ whose means are finite.

Associated with an ARP which starts in state 1 are four sequences of random variables,

(2) $$S_{1r} = T_1 + S_1 + T_2 + S_2 + \ldots + T_r + S_r, \qquad r = 1, 2, \ldots$$

(3) $$S_{2r} = T_1 + S_1 + T_2 + S_2 + T_3 + \ldots + S_{r-1} + T_r, \qquad r = 1, 2, \ldots$$

$N_{1t}$, defined as the number of entrances to state 1 in $(0, t)$, and

$N_{2t}$, defined as the number of entrances to state 2 in $(0, t)$.

$S_{1r}$ is the time until the $r$'th entrance to state 1; $S_{2r}$ is the time until the $r$'th entrance to state 2. Note that

(3) $$N_{1t} = \begin{cases} N_{2t} & \text{if the process is in state 1 at } t \\ N_{2t} - 1 & \text{otherwise.} \end{cases}$$

This is because, for the process to be in state 1 at $t$, there must have been as many entrances to state 1 as exits from it (entrances to 2).

Putting $X_r = T_r + S_r$, then $\{X_r\}$ is an ordinary renewal process with density function $f(x)$ given by the convolution of $g_1(.)$ and $g_2(.)$. Similarly, putting $X_1 = T_1, X_r = S_{r-1} + T_r$, $r > 1$, then $\{X_r\}$ is a modified renewal process with $f_1(x) = g_1(x); f(x) = g_1(x) * g_2(x)$.

Let $k_r(\mathrm{x})$ be the probability density function, (pdf), of $S_r$, $r =, 1, 2, \ldots$ and let an asterisk above any function denote its Laplace transform. Then from (1), in view of the independence of the $\{X_r\}$,

$$k_r^*(s) = f^*(s)^r, \tag{4}$$

for the ordinary renewal process;

$$k_r^*(s) = f_1^*(s) f^*(s)^{r-1}, \tag{5}$$

for the modified renewal process;

$$k_r^*(s) = g_1^*(s)^r g_2^*(s)^r \tag{6}$$

for the ARP starting in state 1 whose events are entrances to state 1, and

$$k_r^*(s) = g_1^*(s)^r g_2^*(s)^{r-1} \tag{7}$$

for the ARP starting in state 1 whose events are entrances to state 2.

The *renewal function* is $H(t) = E[N_t]$, the expected number of renewals in $(0, t)$. If $K_r(t)$ is the distribution function associated with $k_r(t)$, we have

$$\begin{aligned} H(t) &= \sum_{r=0}^{\infty} r P(N_t = r) \\ &= \sum_{r=0}^{\infty} r(K_r(t) - K_{r+1}(t)) \\ &= \sum_{r=1}^{\infty} K_r(t). \qquad (8) \end{aligned}$$

The second line follows because $N_t < r$ if and only if $S_r > t$ so that

$$\begin{aligned} P(N_t < r) &= 1 - K_r(t), \\ P(N_t < r+1) &= 1 - K_{r+1}(t), \\ P(N_t = r) &= K_r(t) - K_{r+1}(t). \end{aligned}$$

It follows from (8) that the Laplace transform of $H(t)$ is

$$H^*(s) = (1/s)\sum_{r=1}^{\infty} k_r^*(s) \tag{(9)}$$

since $K_r^*(s) = (1/s)k_r^*(s)$ is the relation between the the Laplace transform of density and distribution functions.

Concentrating on the ARP we shall derive the renewal function for entrances to state 1 for a process which starts in state 1, which we shall denote by. $H_1(t)$. Using (6) we have

$$\begin{aligned} H_1^*(s) &= (1/s)\sum_{r=1}^{\infty}[g_1^*(s)g_2^*(s)]^r \\ &= \frac{1}{s}\frac{g_1^*(s)g_2^*(s)}{1 - g_1^*(s)g_2^*(s)} \end{aligned} \tag{10}$$

after summing the geometric progression. Similarly, the renewal function for entrances to state 2 for a process which starts in state 1 has Laplace transform, using (7),

$$H_2^*(s) = \frac{1}{s}\frac{g_1^*(s)}{1 - g_1^*(s)g_2^*(s)} \tag{11}$$

$H_1(t)$ and $H_2(t)$, which can be found by inversion of (10) and (11), give the expected number of transitions into state 1 and 2 for a process which starts in state 1. A related important probability is that of the event that state 1 is occupied at any time $t$. This may be found by the following simple argument. As we saw above, for a process that starts in state 1, that state is occupied at $t$ if and only if $N_{1t} = N_{2t}$. Let $\delta = N_{1t} - N_{2t} + 1$. Then

$\delta = 1$ if state 1 is occupied at $t$ and is zero otherwise. Thus

$$
\begin{aligned}
P(\delta = 1) &= P(\text{ state 1 occupied at } t \text{ given it was occupied at } 0) = \pi_1(t) \\
&= E(\delta) \\
&= E(N_{1t}) - E(N_{2t}) + 1 \\
&= H_1(t) - H_2(t) + 1. \qquad (12)
\end{aligned}
$$

Thus

$$\pi_1^*(s) = H_1^*(s) - H_2^*(s) + 1/s.$$

While $H_1(t)$ and $H_2(t)$ may be calculated exactly, in principle, once the density functions $g_1$ and $g_2$ are specified, the behaviour of these renewal functions for large values of $t$ is of interest. Their behaviour for large $t$ is determined by the behaviour of their transforms for $s$ close to zero.[3] Since the Laplace transform is a moment generating function the sign of whose argument is reversed we can write $g_j^*(s) = 1 - s\mu_j + O(s^2)$ as $s \to 0$, where $\mu_j$ is the (finite) expectation $T$, $(j = 1)$ or of $S$, $(j = 2)$, we find from (10)

$$H_1^*(s) = \frac{1}{s^2\mu} - \frac{1}{s}\frac{\mu_1\mu_2}{\mu^2} + o(1/s) \quad \text{as } s \to 0; \qquad (13)$$

$$H_2^*(s) = \frac{1}{s^2\mu} + \frac{1}{s}\frac{\mu_2^2}{\mu^2} + o(1/s) \quad \text{as } s \to 0 \qquad (14)$$

$$\pi_1^*(s) = +\frac{1}{s}\frac{\mu_1}{\mu} + o(1/s) \quad \text{as } s \to 0. \qquad (15)$$

Here $\mu = \mu_1 + \mu_2$. Formally inverting these expressions gives

$$H_1(t) = \frac{t}{\mu} - \frac{\mu_1\mu_2}{\mu^2} + o(1) \quad \text{as } t \to \infty. \qquad (16)$$

$$H_2(t) = \frac{t}{\mu} + \frac{\mu_2^2}{\mu^2} + o(1) \quad \text{as } t \to \infty. \qquad (17)$$

$$\lim_{t\to\infty} \pi_1(t) = \frac{\mu_1}{\mu} \qquad (18)$$

---

[3]Proofs involve Tauberian theorems; see Feller (1966), chapter 8.

**2.2 An ARP Model for Homelessness.** Imagine a large population of constant size, the members of which are at risk of homelessness. With each member of this population is associated a vector of time–invariant characteristics, or covariates, $\mathbf{x}$. We proceed by considering sub–populations homogenous with respect to $\mathbf{x}$, i.e. by arguing conditionally on $\mathbf{x}$. Let state 1 represent being housed and let state 2 represent being homeless. We suppose the experience of each member of such a sub–population is a realisation of an alternating renewal process with $g_j(t) = g_j(t; \mathbf{x})$ for $j = 1, 2$. Members of different sub–populations follow alternating renewal processes involving different distributions of the lengths of stay in each state. We assume the processes for different people are stochastically independent.

In terms of the three statistical features of homelessness described in section 1, this model is binary — there are only two states; homelessness may be rare — state 2 may be rarely visited; and, unless $g_2$ is degenerate, visits to state 2 last for varying lengths of time.

This is a *model* for the phenomenon, i.e. it simplifies reality, and it does so in two main ways. The first is that it assumes that the lengths of visits to each state are stochastically independent, given the covariate vector. How serious this is presumably depends on how completely the covariate vector $\mathbf{x}$ captures the systematic determinants of the average lengths of stay in each state. We shall show later that the omission of autocorrelated determinants of the expected stay in each state is likely to bias estimates of the effect of $\mathbf{x}$ on these means.

The second, and perhaps more important, simplification is that it allows only for time–invariant covariates $\mathbf{x}$. For example, the speed at which women leave the New York city shelter system is known to depend in part upon the number of children for which they are responsible. This is not time–invariant. On the other hand, this model is going to be applied to data in which the main dimension of variation is cross–sectional, between people, and not inter–temporal. We shall in effect be dealing with short time series for many people. The effect of differing numbers of children upon the length of stay in shelter will be largely determined by comparing the experiences of different women — with differing numbers of children — over short time periods, and not by comparing the experiences of the same woman as her family size changes. It may not be unreasonable to suppose that most relevant, time–varying, covariates are roughly constant over the time intervals for which data is likely to be available.

We may apply the results of the preceding section to deduce relevant probabilities and expectations of homelessness. For example the expected number of episodes of homelessness — visits to state 2 — between times $t_1$ and $t_2$ for a person who is initially housed and whose covariate vector is $\mathbf{x}$ is

$$E(N_{2t_2} - N_{2t_1}) = H_2(t_2;\mathbf{x}) - H_2(t_1;\mathbf{x}) \tag{19}$$

The probability that such a person is homeless at time $t$ is, from (12),

$$\pi_2(t;\mathbf{x}) = H_2(t;\mathbf{x}) - H_1(t;\mathbf{x}), \tag{20}$$

which is one minus the probability that she is housed.

The exact form of these expressions will depend upon the form of the probability density functions for the lengths of stay in each state, $g_1(.), g_2(.)$. In the particular case in which these distributions are Exponential these expressions are reasonably simple. If we suppress the dependence of the means on $\mathbf{x}$ for notational simplicity, and let $\lambda = \mu/\mu_1\mu_2$, we find that

$$\begin{aligned} H_1(t;\mathbf{x}) &= \frac{t}{\mu} - \frac{\mu_1\mu_2}{\mu^2}\{1 - e^{-\lambda t}\}, \\ H_2(t;\mathbf{x}) &= \frac{t}{\mu} + \left(\frac{\mu_2}{\mu}\right)^2 \{1 - e^{-\lambda t}\}. \end{aligned} \tag{21}$$

These give

$$E(N_{2t_2} - N_{2t_1}|\mathbf{x}) = \frac{t_2 - t_1}{\mu} + \left(\frac{\mu_2}{\mu}\right)^2 \{e^{-\lambda t_1} - e^{-\lambda t_2}\}, \tag{23}$$

$$\pi_2(t|\mathbf{x}) = \frac{\mu_2}{\mu}\{1 - e^{-\lambda t}\}. \tag{24}$$

For $t_2$ close to $t_1$, i.e. over a short time interval, (23) is approximately

$$E(N_{2t_2} - N_{2t_1}|\mathbf{x}) \sim \frac{t_2 - t_1}{\mu}\{1 + \frac{\mu_2}{\mu_1}e^{-\lambda_1 t}\}. \tag{25}$$

And when $t_1$ is remote from the time origin

$$E(N_{2t_2} - N_{2t_1}|\mathbf{x}) \sim \frac{t_2 - t_1}{\mu}, \tag{26}$$

and

$$\pi_2(t|\mathbf{x}) \sim \frac{\mu_2}{\mu}. \tag{27}$$

These expressions, (21) to (27) are given to exemplify the particular forms that can be taken on by the expressions given earlier. In what follows we do not assume that durations are in fact Exponential, which seems a most unlikely assumption. The form (27), which we have previously given in (16) is correct quite generally as $t_1 \to \infty$ — for the Exponential case this approach is exponentially fast. The form (26) is also generally correct, neglecting terms of $O(t_2 - t_1)^2$, for large $t_1$.

We see that the expected number of episodes of homelessness in any interval is approximately proportional to the length of the interval and inversely proportional to the mean time between such episodes. Entrances to homelessness occur, on average, every $\mu$ days, so it is plausible that the chance of observing such an entrance on any one day is $1/\mu$. The probability of being found homeless is approximately the ratio of the mean duration of homelessness to the mean time between spells of homelessness.

## 3. Sampling Schemes and their Likelihoods.

Let us maintain the accuracy of the ARP model and assume that the object of the investigator is to study the effect of the covariates on the process. There are various probabilities that may be of interest and it is important to distinguish between them. One possible object of interest is the probability that a person with covariate $\mathbf{x}$ is homeless at a point of time. This is $\pi_2(t;\mathbf{x})$, and it is approximately equal to $\mu_2(\mathbf{x})/\mu(\mathbf{x})$ if $t$ can be assumed large. Its relative frequency interpretation is the proportion of people of that $\mathbf{x}$ class who are homeless after $t$ years have elapsed since the start of the process. Another object of interest is the probability that a person with covariate $\mathbf{x}$ will become homeless during an interval of time from say $t_1$ to $t_2$. This is $P(N_{2t_2} - N_{2t_1} > 0|\mathbf{x})$, which is approximately equal to $(t_2 - t_1)/\mu(\mathbf{x})$ if the interval is short and $t_1$ large. Yet another quantity of potential interest is the expected total time spent homeless during an interval by someone whose covariate vector is $\mathbf{x}$.

Now consider sampling a large population of people each of whose experiences are realisations of ARP processes appropriate to their particular covariate vector. We might randomly sample the population so that each member has the same chance of inclusion as every other. The difficulty with this is that if homelessness is rare, for those types of people who comprise the bulk of the population, even a large random sample may contain few, if any, who experience this event.

It will therefore be desirable to construct a sampling scheme that is guaranteed to contain a substantial fraction of people with experience of homelessness. One way of doing is to sample from the subset of people who occupy state 2 — who are homeless — at a moment in time. This is called, for obvious reasons, a *stock sample*. Another way of achieving this end is to sample from the subset of people who, during an interval of time, are observed to enter state 2 — to become homeless. This is a *flow sample*. A total sample could thus be comprised of separate random samples of the sub-populations who (1) occupy state 1 at a point in time; (2) who occupy state 2 at a point in time; and (3) who move from state 1 to state 2 (or, alternatively, from 2 to 1) during an interval of time. Such a total sample is called a stock/flow sampling scheme and the likelihood to which it leads is given in the next sub-section.

The sampling scheme used in the New York city study could be interpreted as a stock/flow scheme, although in fact it departed in its details from the simplest version of such a scheme. Women were excluded from the sample if they had had a very recent stay in shelter; and there was, of course, a fraction of non-respondents.

**3.2 Stock and Flow Samples.** Let us label the subgroups of the population by an indicator $j$ such that

$$j = \begin{cases} 0 \text{ if a person enters state 2 in } t_0, t_0 + \Delta, \\ 1 \text{ if a person resides in state 1 at } t_1, \\ 2 \text{ if a person resides in state 2 at } t_2; \end{cases}$$

where the $t_j$, $j = 0, 1, 2$, and $\Delta$ are chosen by the sampler. It will be convenient also to define binary indicators of group membership

$$y_i = \begin{cases} 1 \text{ if } j = i \\ 0 \text{ otherwise} \end{cases} \qquad j = 1, 2.$$

Let $P_{j\mathbf{x}} = P(j|\mathbf{x})$ $j = 0, 1, 2$. These give the fractions of people with covariate $\mathbf{x}$ who fall into each of these three groups. Note that these groups are neither mutually exclusive nor exhaustive. The probabilities $\{P_{j\mathbf{x}}\}$ may be deduced from the ARP model and we shall give explicit forms later.

The marginal probability mass function of the covariates, $\mathbf{x}$, over the population will be assumed to have finite support $X$, with probability $p_{\mathbf{x}}$ attached to the point $\mathbf{x}$. The fractions of the population who fall into each group are $q_j = \sum_{\mathbf{x}} P_{j\mathbf{x}} p_{\mathbf{x}}$, $j = 0, 1, 2$ Finally, the fraction of the population in group $j$ who have covariate $\mathbf{x}$ is $Q_{\mathbf{x}j}$, $\mathbf{x} \in X$. This gives the conditional distribution of $\mathbf{x}$ given $j$.

The object of the investigation is to study the effect of the covariates on the renewal process. The information for this is provided by the $\{P_{j\mathbf{x}}\}$. To determine what can, in principle, be learned from a stock/flow sampling scheme we must determine whether the $P_{j\mathbf{x}}$ are identifiable in such a scheme and then find out what can be deduced from this knowledge.

The sampling scheme is observe a total of $N$ people, where each sampled person is chosen from group $j$ with probability $h_j$, $\sum h_j = 1$. The vector $h = (h_0, h_1, h_2)$ is chosen by the investigator and we shall make some remarks in a subsequent section on how $h$ should be chosen. Once a person is sampled we observe her covariate vector.[4]

In an indefinitely large sample we could learn $Q_{\mathbf{x}j}$, $j = 0, 12$, as long as each $h_j$ was positive. And by the law of conditional probability,

$$P_{j\mathbf{x}} = \frac{Q_{\mathbf{x}j} q_j}{p_{\mathbf{x}}}. \tag{28}$$

It follows that the ratios of the $\{P_{j\mathbf{x}}\}$ are identifiable from knowledge of the $\{Q_{\mathbf{x}j}\}$ and of the $\{q_j\}$. The $\{q_j\}$ may be known. For example in New York city the whole population may be taken as the women with children in receipt of social welfare; state 2 may be identified as 'in a city shelter'; and state 0 as 'entering shelter in a particular week'. Administrative procedures count the people in each group.[5] If the $q_j$ are not known, they may be estimated from an auxiliary sample of size, say, $n$ in which only group membership is observed. The argument above shows that the ratios of the $\{P_{j\mathbf{x}}\}$ are non-parametrically identified as $N$ and $n \to \infty$. We shall show later that this, together with a stationarity assumption, implies that the mean durations are non-parametrically identified as functions of the covariates.

We shall now describe efficient and computationally simple methods of parametric inference from stock/flow samples, concentrating on the case in which the only information gathered from a sampled person is her covariate vector together, of course, with the identity of the group from which she was taken. The argument here is based, in part, on recent work by Imbens (1990), which in turn follows earlier work by Manski, McFadden and Cosslett (1981). We shall concentrate on inference in the case in which the $q_j$ are assumed known

[4] We may also observe some aspect of each person's state biography, for example by asking for retrospective information about states occupied in the past. We shall assume for the moment that no such additional information is gathered. In section 4 we shall look at the case in which the we observe both the covariate vector and a section of state biography.

[5] A similar example arises in studies of unemployment where state 2 may be 'registered as unemployed'; state 0 as registering as unemployed during a particular week'; and state 1 as 'in the labor force but not unemployed'. Many countries keep count of the total in each category.

to the investigator prior to sampling but the distribution of the covariate vector, $p_{\mathbf{x}}$, is not. We suppose that the $P_{j\mathbf{x}}$ are specified up to a $2K$ dimensional parameter vector $\theta \in \Theta \subset R^{2K}$.

Consider a single woman. She is drawn from group $j$ with probability $h_j$, and given that she is from group $j$ her covariate vector has probability $Q_{\mathbf{x}j}$. Hence the likelihood contribution is

$$p(j,\mathbf{x}) = h_j Q_{\mathbf{x}j} = h_j P_{j\mathbf{x}} p_{\mathbf{x}}/q_j = w_j P_{j\mathbf{x}} p_{\mathbf{x}},$$

where $w_j = h_j/q_j$. This is the joint distribution of the covariate vector $\mathbf{x}$ and random group label $j$. The associated marginal and conditional distributions are

$$\begin{aligned} p(j) &= \sum_{\mathbf{x}} p(j,\mathbf{x}) = w_j \sum_{\mathbf{x}} P_{j\mathbf{x}} p_{\mathbf{x}} = h_j \\ p(\mathbf{x}) &= \sum_{j} p(j,\mathbf{x}) = p_{\mathbf{x}} \sum_{j} w_j P_{j\mathbf{x}} = r_{\mathbf{x}} \\ p(j|\mathbf{x}) &= \frac{p(j,\mathbf{x})}{p(\mathbf{x})} = \frac{w_j P_{j\mathbf{x}}}{\sum_k w_k P_{k\mathbf{x}}} = R_{j\mathbf{x}}. \end{aligned} \tag{29}$$

$r_{\mathbf{x}}$ is the distribution of $\mathbf{x}$ induced by the stock/flow sampling scheme and it generally differs from $p_{\mathbf{x}}$ which is the distribution of the covariate vector in the population. Similarly, $R_{j\mathbf{x}}$ is the conditional probability of group membership induced by the stock/flow sampling scheme and generally differs from $P_{j\mathbf{x}}$ which is the probability that a person randomly sampled from the population of people with covariate $\mathbf{x}$ is in group $j$.

Efficient and simple parametric inference for this problem can be approached via the method of moments. We shall describe a set of moments or orthogonality conditions which, together, exhaust the information in the sample about the true value of $\theta$ which we shall label $\theta^*$. In making these calculations it is convenient to regard the vector $h$ as an additional unknown parameter, even though it is in fact known to the sampler. We distinguish between the vector $h$ which will be an argument to various functions and the true (known) value $h^*$. Put $\gamma = (\theta\, h)$ and $\gamma^* = (\theta^*\, h^*)$.

The starting point for inference will be the likelihood formed from the conditional distribution of $j$ given $\mathbf{x}$; $p(j|\mathbf{x})$ given in (29). For a single observation[6] the log likelihood

[6]We shall always assume that random variables corresponding to different agents are

is

$$L(\theta, h) = \sum_{j=0}^{2} y_j \log R_{j\mathbf{x}}(\theta, h). \tag{30}$$

The $2K$ $\theta$ scores from this log likelihood are assembled in the vector $\psi_1$,

$$\psi_1(\theta, h) = \frac{\partial L}{\partial \theta} = \sum_{j=0}^{2} y_j \frac{\partial \log R_{j\mathbf{x}}}{\partial \theta}. \tag{31}$$

This expression, being a score function, has mean zero at $\theta^*, h^*$.

The next set of moments exploit the fact that, by design, $E(y_j) = h_j^*$. Thus we let the 2 vector $\psi_2$ be

$$\psi_2(h) = \begin{pmatrix} h_1 - y_1 \\ h_2 - y_2 \end{pmatrix} \tag{32}$$

Again, $\psi_2$ has mean zero at $h^*$.

The final set of moments exploit the fact that the marginal probabilities $q_j$ are the expectations of the conditional probabilities $P_{j\mathbf{x}}$ with respect to the population distribution of the covariate. Thus

$$q_j = \sum_{\mathbf{x}} P_{j\mathbf{x}}(\theta^*) p_{\mathbf{x}},$$

which may be written as

$$q_j = \frac{q_j}{h_j^*} \sum_{\mathbf{x}} \frac{w_j^* P_{j\mathbf{x}}(\theta^*)}{\sum_k w_k^* P_{k\mathbf{x}}(\theta^*)} p_{\mathbf{x}} \sum_k w_k^* P_{k\mathbf{x}}(\theta^*), \tag{33}$$

where $w_j^* = h_j^*/q_j$. Equation (33) may be written

$$h_j^* = \sum_{\mathbf{x}} r_{\mathbf{x}}^* R_{j\mathbf{x}}^*,$$

---

stochastically independent so that log likelihoods are sums of their single observation counterparts. This rules out aggregate shocks. Whether this is a serious qualification to our argument is unclear to us.

where the $*$ indicates that $\gamma$ is evaluated at $\gamma^*$. Thus the vector $\psi_3$ defined by

$$\psi_3(\theta, h) = \begin{pmatrix} h_1 - R_{1\mathbf{x}}(\theta, h) \\ h_2 - R_{2\mathbf{x}}(\theta, h) \end{pmatrix} \tag{34}$$

has mean zero at $\gamma^*$. This is our third set of moment conditions. Note that the distribution with respect to which the expectations are taken is that governing the data, $p(j, \mathbf{x})$, given by (29).

Now let $\psi = (\psi_1\ \psi_2\ \psi_3)$; let $W = E(\psi\psi')$; and let $\Gamma = E(\partial\psi/\partial\gamma)$. Assume that $W$ is non–singular. If it is not some moments are redundant and may be deleted. Let $\overline{\psi}$ be the sample mean value of $\psi$. Then

$$\hat{\gamma} = \text{argmin}_{\theta,\, h}\, \overline{\psi}'(\gamma) W^{-1} \overline{\psi}(\gamma) \tag{35}$$

is the generalised method of moments (GMM) estimator of $\theta^*$, $h^*$. (Imbens (1990)). It is such that $\sqrt{N}(\hat{\gamma} - \gamma^*) \to N(0, (\Gamma' W^{-1} \Gamma)^{-1})$ and $\hat{\gamma}$ is fully asymptotically efficient.[7]

Two other proposed estimators for the choice based sampling framework can also be readily expressed in this method of moments framework. Firstly, the estimator which solves

$$\tilde{\theta} = \text{argmin}_{\theta}\, \overline{\psi}_1(\theta,\, h^*) W_1^{-1} \overline{\psi}_1(\theta,\, h^*), \tag{35}$$

i.e. which solves the equation $\overline{\psi}_1(\theta,\, h^*) = 0$ is the conditional maximum likelihood estimator, CML, of Manski and McFadden (1981). Here $W_1$ is the covariance matrix of $\psi_1$.

Secondly, the estimator which solves

$$\bar{\theta} = \text{argmin}_{\theta,\, h}\, [\overline{\psi}_1(\theta,\, h)\, \overline{\psi}_2(h)]' W_2^{-1} [\overline{\psi}_1(\theta,\, h)\, \overline{\psi}_2(h)], \tag{36}$$

[7]There is a version of this procedure when some or all of the $\{q_j\}$ are unknown. And it may also be readily extended to the case in which supplementary information about the $\{q_j\}$ is available from a random sample of the $y$'s. This information yields an additional moment equation.

i.e. which solves $\overline{\psi}_1(\theta, \bar{y}) = 0$ is the modification to the CML estimator suggested by Cosslett (1981). In this procedure the true sampling probabilities $\{h_j^*\}$ are replaced in the conditional likelihood by the fractions of observations falling into each group, $\bar{y}_j = N^{-1}\sum_n y_{jn}$.[8]

It remains to consider the choice of functional form for $P_{j\mathbf{x}}(\theta)$. The simplest choice, and the only one which avoids having to specify an origin for the ARP,[9] is to use the forms which apply when $t \to \infty$. These are,

$$
\begin{aligned}
P_{0\mathbf{x}} &= \Delta/\mu(\mathbf{x};\theta) \\
P_{1\mathbf{x}} &= \mu_1(\mathbf{x};\theta)/\mu(\mathbf{x};\theta) \\
P_{2\mathbf{x}} &= \mu_2(\mathbf{x};\theta)/\mu(\mathbf{x};\theta).
\end{aligned}
\tag{37}
$$

Clearly the $\mu_j$ are identifiable from the ratios of the $\{P_{j\mathbf{x}}\}$ and are thus, by our previous argument, identified. Moreover, only the means of the distributions $g_1$, $g_2$ are identified from the likelhood we have described. To identify other aspects of these distributions we must observe more about the process than the covariate vectors of stock/flow sampled individuals. Notice that to identify both the means it is essential to have observations from all three groups. The New York city study mentioned earlier drew observations only from those entering shelter — group 0 — and from those out of shelter — group 1 , from which only $\mu_1$ may be determined.[10] Samples from the stocks alone identify only the ratio of the mean stays in each state. A flow sample is necessary to separate these means. The

[8] The problem (36) reduces to solving $\overline{\psi}_1(\theta, \bar{y}) = 0$ because $\overline{\psi}_2(h) = 0$ is solved by $h = \bar{y}$.

[9] It is possible to calculate the probabilities $\{P_{j\mathbf{x}}\}$ exactly for any choice of density functions $g_j(.)$ by numerical inversion of the relevant Laplace transforms. We do not do this because we wish to focus on theoretical issue for which the fine detail of the specification of the $\{P_{j\mathbf{x}}\}$ has little relevance. Taking $t$ large also lets us avoid specifying a time origin for the process, to do which involves formulating a more detailed behavioural model than is required for our present purposes.

[10] They also obtained supplementary observations — from administrative records — of the previous shelter stays of all sampled individuals. This information would help to identify $\mu_2$.

intution here is that to identify the mean stays in each state we must have a realisation of the process, i.e. a section of the state biography. A flow sample is just such a section and presumably the shortest section of biography that would yield identification.

Under this specification, the conditional probabilities of the groups that are induced by stock/flow sampling take the form

$$R_{j\mathbf{x}}(\theta,\, h) = \frac{w_j\mu_j}{w_0 + w_1\mu_1 + w_2\mu_2} \quad j = 1, 2. \tag{38}$$

This is a model with a trinomial Logit structure. It has the standard Logit form if

$$\mu_j = \exp\{\theta_{j0} + \theta_{j1}'\mathbf{x}\}, \quad j = 1, 2. \tag{39}$$

This is the functional form we shall assume in what follows, although some results apply to more general forms for the $\{\mu_j\}$.

We can now examine the proposed method of moments estimator in the light of these functional forms, since they enable us to get explicit expressions for the conditional likelihood scores which are the elements of $\psi_1$. On differentiation we find

$$\frac{\partial L}{\partial \theta} = (R_{j\mathbf{x}}(\theta,\, h) - y_j)x_l, \quad \text{if } \theta \text{ is the coefficient of } x_l \text{ in } \mu_j. \tag{40}$$

In particular, the scores correspoding to the intercepts are

$$R_{1\mathbf{x}} - y_1; \quad R_{2\mathbf{x}} - y_2. \tag{41}$$

Comparing these expressions with the elements of $\psi_2$ and $\psi_3$ given by (32) and (34) we see that these scores are linear combinations of the elements of $\psi_2$ and $\psi_3$, e.g. $R_{1\mathbf{x}}(\theta,\, h) - y_1 = h_1 - y_1 - (h_1 - R_{1\mathbf{x}}(\theta,\, h)$. Thus the elements of $\psi$ are linearly dependent so that the modified CML estimator, $\overline{\theta}$, which equates $\overline{\psi}_1$ and $\overline{\psi}_2$ to zero will necessarily equate $\overline{\psi}_3$ to zero. Thus $\overline{\theta}$ is fully efficient in this model.[11]

[11]This result clearly only depends on the way in which the intercepts enter the $\mu$'s — multiplicatively — and not on the rest of the parametrisation of the $\mu_j$. Essentially this result has been previously given by Manski and McFadden (1981).

We may confine attention, with this parametrisation, to the modified CML estimator, $\bar{\theta}$. Even here further simplification is possible. In particular let us compare the CML estimator, $\tilde{\theta}$, to $\bar{\theta}$. The conditional choice probabilities induced by stock/flow sampling may be rewritten, exploiting the functional form (39), as

$$R_{j\mathbf{x}}(\theta,\, h) = \frac{\exp\{\theta_{j0} + \theta'_{j1}\mathbf{x} + a_j\}}{1 + \exp\{\theta_{10} + \theta'_{11}\mathbf{x} + a_1\} + \exp\{\theta_{20} + \theta'_{21}\mathbf{x} + a_2\}}, \tag{42}$$

where $a_j = \log[w_j/w_0]$. Similarly,

$$R_{j\mathbf{x}}(\theta,\, \bar{y}) = \frac{\exp\{\theta_{j0} + \theta'_{j1}\mathbf{x} + b_j\}}{1 + \exp\{\theta_{10} + \theta'_{11}\mathbf{x} + b_1\} + \exp\{\theta_{20} + \theta'_{21}\mathbf{x} + b_2\}}, \tag{42}$$

where $b_j = \log[\overline{w}_j/\overline{w}_0]$, and $\overline{w}_j = \overline{y}_j/q_j$.

Now let $\tilde{\theta}_{11}$ and $\tilde{\theta}_{21}$ be the slope components of the CML estimator which solves $\overline{\psi}_1(\theta^*,\, h) = 0$. Then we must have

$$\sum_{n=1}^{N} \left[ \frac{\exp\{\tilde{\theta}_{j0} + \tilde{\theta}'_{j1}\mathbf{x}_n + a_j\}}{1 + \exp\{\tilde{\theta}_{10} + \tilde{\theta}'_{11}\mathbf{x}_n + a_1\} + \exp\{\tilde{\theta}_{20} + \tilde{\theta}'_{21}\mathbf{x}_n + a_2\}} - y_n \right] x_{ln} = 0, \tag{43}$$

for $l = 1, 2, \ldots K, j = 1, 2$. But if we substitute $\tilde{\theta}_{11}$ and $\tilde{\theta}_{21}$ into the corresponding expressions for $\overline{\psi}_1(\theta,\, \bar{y})$ we find

$$\sum_{n=1}^{N} \left[ \frac{\exp\{\hat{\theta}_{j0} + \tilde{\theta}'_{j1}\mathbf{x}_n + b_j\}}{1 + \exp\{\hat{\theta}_{10} + \tilde{\theta}'_{11}\mathbf{x}_n + b_1\} + \exp\{\hat{\theta}_{20} + \tilde{\theta}'_{21}\mathbf{x}_n + b_2\}} - y_n \right] x_{ln}, \tag{44}$$

which equal zero for

$$\hat{\theta}_{j0} + b_j = \tilde{\theta}_{j0} + a_j, \quad j = 1, 2.$$

Hence the slope coefficients in the CML estimator and its modified version are identical; these estimators differ only in the intercepts. This fact will be exploited in the next section, in which look at the question of how many observations to select from the flow and from each stock This is the question of optimal sampling fractions.

### 3.3 Optimal Sampling Fractions.

An important problem is how best to choose the numbers of observations, $m_i$, from each group. This admits of no general solution because the optimal numbers of observations depend upon the true, but unknown, value of $\theta$. As Hsieh, Manksi and McFadden remark, 'This fact severely limits the guidance we can give to a researcher attempting to select an efficient sample design'.[12] Numerical calculations reported by Cosslett (1981) indicate that the optimal design is often not far from that of equal shares, in which the same number of observations is taken from each group.

There is, however, one important case in which the optimal design can be determined analytically, a case which helps to explain Cosslett's results. Consider the ARP model with specification (39) for the $\mu_j$. For this model we know that the CMLE is fully efficient for the slope coefficients $\theta_{j1}$ so let us examine its covariance matrix and see how this depends upon the choice of $\{h_j\}$. Differentiating $\psi_1(\theta, h)$ with respect to $\theta$ and taking expectations gives for the information matrix for a single observation

$$\mathcal{I}(\theta^*, h) = E_{r^*_{\mathbf{x}}}(R^*_{\mathbf{x}} \otimes \mathbf{x}\mathbf{x}') \tag{45}$$

where

$$\mathbf{R}^*_{\mathbf{x}} = \begin{pmatrix} R^*_{1\mathbf{x}}(1 - R^*_{1\mathbf{x}}) & -R^*_{1\mathbf{x}}R^*_{2\mathbf{x}} \\ -R^*_{1\mathbf{x}}R^*_{2\mathbf{x}} & R^*_{2\mathbf{x}}(1 - R^*_{2\mathbf{x}}) \end{pmatrix}, \tag{46}$$

and $\otimes$ denotes the Kronecker product. The matrix $\mathbf{x}\mathbf{x}'$ is of order $K \times K$ and the vector $\mathbf{x}$ includes a unit for the intercepts.

Now suppose we consider the particular case in which $\theta^*_{j1} = 0$, $j = 1, 2$, so that the $P_{j\mathbf{x}}$ do not in fact depend upon the covariates, although the investigator is unaware of this. Then $\mathcal{I}$ simplifies considerably, and in particular $P^*_{j\mathbf{x}} = q_j$ and $R^*_{j\mathbf{x}} = h_j$. Then we readily find that $\mathcal{I}$ may be written in the form

$$\mathcal{I}(\theta^*, h) = H \otimes \Sigma, \tag{47}$$

[12] HMF page 659.

where

$$H = \begin{pmatrix} h_1(1-h_1) & -h_1h_2 \\ -h_1h_2 & h_2(1-h_2) \end{pmatrix}$$

and

$$\Sigma_{lm} = E_p(x_l x_m), \quad l, m = 1, 2, \dots K.$$

The upper left element of $H$ refers to the parameters of $\mu_1$ and the lower right to those of $\mu_2$. The inverse of $\mathcal{I}$ is the asymptotic covariance matrix of $\sqrt{N}(\tilde{\theta} - \theta^*)$ which is

$$V = H^{-1} \otimes \Sigma^{-1}. \tag{48}$$

But the fully efficient modified CML estimator for the slope coefficients $\{\theta_{j1}\}$ is identical to the slope components of $\tilde{\theta}$ so the asymptotic covariance matrix of $\overline{\theta}_{j1}$ can be found by deleting from (48) the two rows and columns corresponding to the intercept estimates. This gives

$$V_1 = H^{-1} \otimes \Sigma_1^{-1} \tag{49}$$

where $\Sigma_1$ is $\Sigma$ after deletion of the row and column corresponding to the unit element in $\mathbf{x}$. If we now ask for the choice of $h$ which will, in some sense, minimise (49), a natural criterion is $|V_1|$, the generalised variance. This is

$$|V_1(h)| = |H|^{-(K-1)}|\Sigma_1|^{-2}$$

from a well known property of the determinant of Kronecker products.[13] But

$$|H| = h_0 h_1 h_2$$

which is maximised over $h_0 + h_1 + h_2 = 1$ by $h_j = 1/3$, $j = 0, 1, 2$. It follows that the generalised variance is minimised by an equal shares allocation of observations — take equal numbers from the flow and from both stocks when the effect of the regressors on the means is believed to be 'small'. In particular, equal shares would be sensible when the

[13] Magnus and Neudecker (1989), page 29.

main object is to test whether the regressors have any effect. This result presumably helps to explain the prominence of the equal shares allocation in Cosslett's numerical results referred to earlier. Note that the form (47) is the leading term in a Taylor series expansion of the information matrix about $\theta^*_{j1} = 0$ so that equal shares will be nearly optimal for small departures of the slopes from zero.[14]

It will be interest for the material of section 4 to give the explicit form of the information matrix under equal shares sampling. Substituting $h_j = 1/3$ into (47) gives

$$\mathcal{I} = \frac{1}{9}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \otimes \Sigma,$$

so that the asymptotic covariance matrix of $\sqrt{T}(\tilde{\theta} - \theta^*)$ will be

$$3\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \otimes \Sigma^{-1}.$$

We shall see in section 4 that this expression is surprisingly accurate even when the slopes are not zero. Note that it is independent of the values of the intercepts, $\{\theta^*_{j0}\}$.

An interesting comparison is between the covariance matrix under equal shares sampling and that under a sampling scheme in which each group is sampled in proportion to its frequency in the population, which we might call random sampling. Under random sampling we must have $h_j = q_j / \sum q_i$, since the $h$'s must sum to one, although the $q$'s will not. Substituting this form into (47) we find

$$\frac{\sum q_i}{q_0 q_1 q_2}\begin{pmatrix} q_2(q_0 + q_1) & q_1 q_2 \\ q_1 q_2 & q_1(q_0 + q_2) \end{pmatrix} \otimes \Sigma^{-1}$$

for the asymptotic covariance matrix of the CMLE under 'random' sampling. The ratio of the generalised variance of the coefficient estimators under equal shares sampling to that

[14] It should be emphasised that our optimality result is not universal. It refers to estimation of $\theta$, within the class of stock/flow schemes, and with a loss function given by the generalised variance. Other sampling schemes may be optimal when the objects of estimation are, for example, covariate mean stock or flow probabilities, or when a different loss function is appropriate.

under random sampling is equal to

$$\frac{q_0 q_1 q_2}{\bar{q}^3}$$

which is one when all the $q$'s are equal (to 0.5) and is otherwise less than one by the arithmetic/geometric mean inequality.

There is one other result on optimal allocation that might be mentioned here. This is that equal shares is also optimal when there is but a single (real) regressor whose distribution has two points of support and the model is, therefore, saturated.

### 3.5. The Effect of Neglected Heterogeneity.

The renewal model assumes independence of the durations of successive spells, conditional on the measured covariate $\mathbf{x}$. Perhaps the most natural way to relax this assumption is to suppose that there may be persistent, individual specific, determinants of the duration of stay that have not been measured by the econometrician. Such effects would induce a correlation between the durations of stay. Suppose that, conditional on $\mathbf{x}$ and some time–invariant scalar $u_j$, the distribution of the lengths of stays in state $j$ is $g_j(t; \mathbf{x}, u_j)$, $j = 1, 2$. The interpretation of $u_j$ is that it captures the effect of the unmeasured persistent determinants on the durations of stay in each state. Let the corresponding means be $\mu_j(\mathbf{x}, u_j)$ and let $\mathbf{u} = (u_1, u_2)$. The vector $\mathbf{u}$ has some joint distribution over the population and is distributed independently of $\mathbf{x}$.

Suppose now that the ARP model is satisfied *for populations homogeneous with respect to both* $\mathbf{x}$ *and* $\mathbf{u}$. Then the algebra of section 2.1 goes through conditionally on $\mathbf{x}$ and $\mathbf{u}$ and the probability that a person randomly selected from a sub–population homogeneous with respect to $\mathbf{x}$ and $\mathbf{u}$ will be found to be in state $j$ is, for large $t$,

$$P(\text{in state } j|\mathbf{x}, \mathbf{u}) = \frac{\mu_j(\mathbf{x}, \mathbf{u})}{\mu(\mathbf{x}, \mathbf{u})}, \quad j = 1, 2. \tag{50}$$

However, if all the econometrician can be observe is $\mathbf{x}$, the relevant probability is that unconditional on $\mathbf{u}$, namely,

$$P(\text{in state } j|\mathbf{x}) = E_{\mathbf{u}}\left[\frac{\mu_j(\mathbf{x}, \mathbf{u})}{\mu(\mathbf{x}, \mathbf{u})}\right] = P^*_{j\mathbf{x}}, \quad j = 1, 2. \tag{51}$$

Similarly, the probability of an entry to state 2 in an interval of length $\Delta$ is

$$P(\text{entry to state 2 in } t, t+\Delta|\mathbf{x}) = E_{\mathbf{u}}\left[\frac{\Delta}{\mu(\mathbf{x},\mathbf{u})}\right] = P^{*}_{0\mathbf{x}}. \tag{52}$$

Notice that (51) and (52) are expectations of ratios, not ratios of expectations. Therein lies the source of the difficulty.

We shall now use these expressions to examine the specification error that is made when an investigator assumes that the data is generated by an ARP model conditional on $\mathbf{x}$ when in fact the correct model is given by the probabilities (51) and (52). Such an investigator might adopt the specification

$$\mu_j(\theta) = \exp\{\theta_{j0} + \theta'_{j1}\mathbf{x}\}, \quad j = 1, 2, \tag{53}$$

considered earlier. Suppose further that the means conditional on both $\mathbf{x}$ and $\mathbf{u}$ are

$$\mu_j(\theta^*) = \exp\{\theta^*_{j0} + \theta^{*\prime}_{j1}\mathbf{x} + u_j\}, \quad j = 1, 2. \tag{54}$$

This investigator would naturally use the CMLE to estimate the $\theta$'s, since this would be asymptotically efficient on his (false) assumptions. He would therefore maximise

$$L(\theta) = N^{-1}\sum_{n=1}^{N}\sum_{j=0}^{2} y_{jn}\log R_{j\mathbf{x}_n}(\theta) \tag{55}$$

where

$$R_{j\mathbf{x}}(\theta) = \frac{P_{j\mathbf{x}}(\theta)w_j}{\sum_k P_{k\mathbf{x}}(\theta)w_k}.$$

In a large sample, as $N \to \infty$, $L(\theta)$ will converge uniformly in $\theta$, under suitable regularity conditions, to

$$\overline{L}(\theta) = \sum_{j=0}^{2} E_{r^*_{\mathbf{x}}} R^*_{j\mathbf{x}}\log R_{j\mathbf{x}}(\theta). \tag{56}$$

Here,

$$R^*_{j\mathbf{x}} = \frac{P^*_{j\mathbf{x}}w_j}{\sum_k P^*_{k\mathbf{x}}w_k}, \quad j = 0, 1, 2,$$

and

$$r^*_{\mathbf{x}} = p_{\mathbf{x}} \sum_j P^*_{j\mathbf{x}} w_j. \tag{57}$$

Here $r^*_{\mathbf{x}}$ is the true distribution of $\mathbf{x}$ induced by stock/flow sampling.

Under suitable regularity conditions the CML estimator will converge to the unique maximiser of $\overline{L}(\theta)$. It will, therefore, satisfy the equations

$$\sum_{\mathbf{x}} r^*_{\mathbf{x}} \left[R^*_{1\mathbf{x}} - R_{1\mathbf{x}}(\theta)\right] x_j = 0 \quad j = 1, 2, \ldots K \tag{58}$$

$$\sum_{\mathbf{x}} r^*_{\mathbf{x}} \left[R^*_{2\mathbf{x}} - R_{2\mathbf{x}}(\theta)\right] x_j = 0 \quad j = 1, 2, \ldots K \tag{59}$$

Let us then consider the relation between the solution of these equations, $\theta$, and the true parameters $\theta^*$. We can make some analytical progress and then we shall resort to computation. Assume a single real regressor for simplicity. In the first place note that if $\theta^*_{11} = \theta^*_{21} = 0$, then the $R^*_{j\mathbf{x}}$ do not depend upon $\mathbf{x}$. Hence equations (58, 59) can be solved by $\theta_{11} = \theta_{21} = 0$, since this will equate $R^*_{j\mathbf{x}}$ and $R_{j\mathbf{x}}$ for every $\mathbf{x}$. Thus when $\mathbf{x}$ really has no effect on the durations of stay in each state the ML estimator will tell us so.

Secondly, consider the simpler problem of binary choice and random sampling, when the equations analogous to (58, 59) are

$$\sum_{\mathbf{x}} p_{\mathbf{x}} \left[P^*_{\mathbf{x}} - P_{\mathbf{x}}(\theta)\right] x_j = 0, \tag{60}$$

where

$$P^*_{\mathbf{x}} = E_u \left[\frac{1}{1 + \exp\{u + \theta^{*\prime}_1 \mathbf{x}\}}\right]; \quad P_{\mathbf{x}}(\theta) = \left[\frac{1}{1 + \exp\{\theta_0 + \theta_1 \mathbf{x}\}}\right],$$

and $\mathbf{x} = (1, x)$. Suppose that $p_{\mathbf{x}}$ has two points of support, say at $x = x_1, x_0$. Then the equations (60) imply that

$$P^*_{x_0} = P_{x_0}(\theta); \quad P^*_{x_1} = P_{x_1}(\theta).$$

This is approximately equivalent to equating both $P_x^* - P_x(\theta)$ and its derivative to zero at $x = x_0 = 0$, say. This implies that

$$E\left[\frac{1}{1+e^u}\right] = \frac{1}{1+e^{\theta_0}}; \quad -\theta_1^* E\left[\frac{e^u}{(1+e^u)^2}\right] = -\theta_1 \frac{e^{\theta_0}}{(1+e^{\theta_0})^2}.$$

Putting $v = 1/(1+e^u)$ and solving for $\theta_1/\theta_1^*$ gives

$$\frac{\theta_1}{\theta_1^*} = \frac{Ev(1-v)}{E(v)(1-E(v))}.$$

Since $g(u) = u(1-u)$ is a concave non-negative function of $u$ it follows from Jensen's inequality that $0 \leq \theta_1/\theta_1^* \leq 1$. Only if there is no neglected heterogeneity, and $u$ has a degenerate distribution, does $\theta_1 = \theta_1^*$. Thus in this case the effect of neglected heterogeneity is to attenuate the estimated effect of a regressor on the event probability.

This heuristic argument does not prove that attenuation will be the result for more general regressor distributions. Still less does it prove this for the more complicated stock/flow sampled ARP. It does, however point in that direction.

The tables below report a selection of the results of some calculations designed to measure the effect of neglected heterogeneity on estimates of the effects of the covariates on the renewal process. While the slope coefficients $\{\theta_{1j}\}$ are the most obvious measures of these effects, in practice an investigator is likely to want to calculate the effects of the covariates on more readily interpretable quantities. In particular he may wish to calculate the way in which[15] $x$ affects the probability that a woman will be homeless. Since this probability is $\mu_2(x)/\mu(x)$, this measure might be $\partial \log[\mu_2(x)/\mu(x)]/\partial x$. The correct measure of this elasticity at the point $x$ is

$$\epsilon = \frac{\partial}{\partial x} \log E_u \left[\frac{\mu_2(x, \theta^*)}{\mu_1(x, \theta^*) + \mu_2(x, \theta^*)}\right] \tag{61}$$

where $\mu_j(x, \theta^*)$ are given by (54). An investigator who ignored unmeaured heterogeneity whould, however, calculate

$$e = \frac{\partial}{\partial x} \log \left[\frac{\mu_2(x, \theta)}{\mu_1(x, \theta) + \mu_2(x, \theta)}\right] \tag{62}$$

[15]We shall assume a single real regressor for simplicity.

where the $\mu_j(x, \theta)$ are given by (53) and $\theta$ is the CML estimate which, in a large sample, is close to the solution of (58) and (59).

We have, therefore, solved equations (58) and (59) for various choices of the joint distribution of $u_1$ and $u_2$; of the distribution of $x$, $p_x$; and for various sampling schemes $h$. Given the solution $\theta$ we calculated $e$, (62), and compared it to $\epsilon$, (61). The value of $x$ at which the elasticities were calculated was the population mean of $p_x$.

The tables below give a selection of the results. We varied the distribution of $x$ between one which was symmetrical and one which was highly skewed. We assigned equal variances to $u_1$ and $u_2$ and measured the magnitude of these variances by the ratio var $u/(\text{var } u + \text{var } \theta x)$, which we denote by $1 - r^2$. The larger this number the more serious is the omitted heterogeneity. And we chose to compare the equal shares sampling scheme with that which we termed random sampling in the last section. We also include the true probability of being in state 2, denoted by $p^*$.

True and Estimated Elasticities
Equal Shares Sampling

| $p_x$ | $1 - r^2$ | $p^*$ | $\epsilon$ | $e$ |
|---|---|---|---|---|
| Sym. | .03 | .57 | $-.81$ | $-.83$ |
| Skew | .07 | .99 | $-.012$ | $-.010$ |
| Sym. | .34 | .53 | $-.52$ | $-.61$ |
| Skew | .56 | .91 | $-.13$ | $-.03$ |

True and Estimated Elasticities
Random Sampling

| $p_x$ | $1-r^2$ | $p^*$ | $\epsilon$ | $e$ |
|---|---|---|---|---|
| Sym. | .03 | .57 | $-.81$ | $-.82$ |
| Skew | .07 | .99 | $-.011$ | $-.012$ |
| Sym. | .34 | .53 | $-.52$ | $-.59$ |
| Skew | .56 | .91 | $-.13$ | $-.04$ |

The conclusions to be drawn from this evidence appear to be as follows. Neglected heterogeneity does bias the estimated elasticities. There is little difference in the bias between random and equal shares sampling. The bias[16] is generally not large except when there is both large heterogeneity omitted and the state probabilities are highly unequal.

We have also considered the bias in the estimates of the slope coefficients. The biases are always in the direction of zero, confirming the attenuation suggested by the analysis given above. These biases, in proportionate terms, are much more severe than those shown in the table above. If the primary interest of an investigator is in the $\theta$ coefficients themselves then it will be important to make some allowance for the possibility of neglected heterogeneity. The distribution of the $u$'s is clearly non–parametrically unidentified from the distribution only of the covariate vector. It will be necessary to observe sections of the state biography before attempting to estimate a model which allows for unmeasured heterogeneity. In the next section we shall report some results on inference from stock/flow sampled populations when the data also includes state biographies.

[16] Strictly, the inconsistency

**4. Optimal Sampling with Biographical Data.**

We now turn to an examination of the question of optimal sampling of a population moving alternately between two states for which we observe both the covariate vector and a section of the state biography. We shall define a collection of strata of the population and consider sampling schemes which involve randomly sampling individuals from within each stratum. The strata we shall consider are the stocks and flows of section 3 except that we shall distinguish between the flows in two directions, $1 \to 2$ and $2 \to 1$, and we shall also consider the whole population as a stratum. This gives us five strata as opposed to the three of section 3. They will be labelled by $S$ and we shall define corresponding random binary indicators of stratum membership.

$$S = \begin{cases} 0 & \text{the whole population} \\ 1 & \text{residents of state 1 at } t \\ 2 & \text{residents of state 2 at } t \\ 3 & \text{entrants to state 2 in } t, t+\Delta \\ 4 & \text{entrants to state 1 in } t, t+\Delta \end{cases} \tag{63}$$

$$s_j = \begin{cases} 1 & \text{if } S = j \\ 0 & \text{otherwise} \end{cases} \qquad j = 0,1,2,3,4.$$

The probabilities of stratum membership, conditional on the covariate vector, $\mathbf{x}$, are $P_{S\mathbf{x}} = P(S|\mathbf{x}), S = 0, \ldots 4$. These probabilities are specified, as in section 3, to take the form

$$\begin{aligned} P_{S\mathbf{x}} &= \frac{\mu_S(\mathbf{x};\theta)}{\mu(\mathbf{x};\theta)}, \quad S = 1,2, \\ P_{S\mathbf{x}} &= \frac{\Delta}{\mu(\mathbf{x};\theta)}, \quad S = 3,4. \end{aligned} \tag{64}$$

Of course $P_{0\mathbf{x}} = 1$. The corresponding marginal probabilities are the expectations of the $P_{S\mathbf{x}}$ with respect to $p_{\mathbf{x}}$, the distribution of the covariate vector over the population. They are, as before, denoted by $\{q_S\}$, and $q_S = \sum_{\mathbf{x}} P_{S\mathbf{x}} p_{\mathbf{x}}$. Of course $q_0 = 1$.

We shall also observe for each sampled person her state biography over $T$ months. This biography is completely described by (a) the initial state $i$, and (b) a vector $t = (t_1, t_2, \ldots)$

of times between changes of state. A binary indicator of the initial state will be

$$y = \begin{cases} 1 \text{ if } i = 1 \\ 0 \text{ if } i = 2 \end{cases}$$

In order to handle observations on the state biography $(i, t)$ we can no longer be content to specify only the mean stays in each state but must, at least for fully parametric inference, specify the density functions $g_1(t|\mathbf{x}), g_2(t|\mathbf{x})$ that govern the duration of each type of stay. Since our present purpose is purely theoretical we shall choose the simplest model in which $g_1$ and $g_2$ are Exponential distributions. Thus

$$g_j(t|\mathbf{x}) = \mu_j^{-1}(\mathbf{x};\theta)\exp\{-t/\mu_j(\mathbf{x};\theta)\}, \quad j = 1,2.$$

This implies that our model is an alternating Poisson process (APP). For such a process the probability of state biography $t$ conditional on initial state $i$ and covariate $\mathbf{x}$ is

$$p(t|i,\mathbf{x}) = \mu_1^{-d_1}\mu_2^{-d_2}\exp\{-T_1/\mu_1 - T_2/\mu_2\}, \tag{65}$$

where

$$\begin{aligned} d_1 &= \#\text{ transitions from 1 to 2} \\ d_2 &= \#\text{ transitions from 2 to 1} \\ T_j &= \text{total time spent in state } j, \quad j = 1,2. \end{aligned}$$

Assuming that the process is in equilibrium the probabilities of the initial states conditional on $\mathbf{x}$ are

$$p(i|\mathbf{x}) = \frac{\mu_i(\mathbf{x};\theta)}{\mu(\mathbf{x};\theta)}, \quad i = 1,2 \tag{66}$$

The sampling scheme is the natural extension of that used in section 3 in which we take a sample of size $N$ by selecting stratum $S$ with probability $h_s$, $\sum h_s = 1$, and then randomly selecting a person from within the chosen stratum. The vector $h = (h_0, \ldots h_4)$

is chosen by the investigator and we aim to study how the choice of $h$ affects the precision of inference about $\theta$.

In order to do this we define the generalised method of moments estimator $\hat{\theta}$ by a straightforward extension of the method of the last section, and write down the asymptotic covariance matrix of $\sqrt{N}(\hat{\theta} - \theta^*)$. This covariance matrix will be computed numerically from simulated APP data using parameter values estimated from real observations. We shall study the diagonal elements of this matrix for various choices of sampling scheme, $h$, and length of biography, $T$.

To simulate data we need to select a particular alternating Poisson process and to do this we fitted such a model to a sample of 372 Dutch males observed over the 84 months from January 1977 to December 1983. The scalar covariate was initial age minus 35 years which is approximately uniformly distributed from -15 to 15. The states were (1) not employed and (2) employed. The estimated regression functions were

$$\mu_1(x) = \exp\{4.0 + 0.043x\}$$
$$\mu_2(x) = \exp\{5.6 + 0.030x\}$$

Thus $\theta^*_{10} = 4.0, \theta^*_{11} = 0.043$, etc. The numbers of transitions in this sample had the following frequency distributions.

| #Transitions | Frequency $1 \to 2$ | Frequency $2 \to 1$ |
|---|---|---|
| 0 | 338 | 268 |
| 1 | 29 | 98 |
| 2 | 4 | 5 |
| 3 | 1 | 1 |

Thus 338 men had no transitions into employment; 268 had no transitions out of employment; etc. A total of 151 transitions were observed in the 7 years, or about 1 transition for every 2 men. This is therefore a group displaying little movement between states.[17] It

[17]Further information about the sample and data base is given in the appendix.

is arguable that sluggish movement such as this is not dissimilar to what one might find in studying homelessness.

We should emphasise that we do not claim that an alternating Poisson process is a good description of these data — it is almost certainly not. Our intention is to use the data to get a feel for some reasonable values for the parameters, and we regard fitting an APP as somewhat analogous to fitting a straight line through an almost surely curved scatter of points.

To construct the moment equations we need the likelihood of a single observation generated in the following steps: (1) select stratum $s$ with probability $h_s$; (2) select an individual from that stratum; (3) observe her covariate vector; and (4) observe her transition history for $T$ months from the date of sampling. In the case of strata other than the whole population — stratum zero — the stratum choice determines the initial state $i$. For example, stratum 3 implies initial state 2. The joint probability of stratum indicator $s$, initial state indicator $y$, biography $t$ and covariate $x$ is $p(s,y,t,x)$ which we factor as

$$p(s,y,t,x) = p(t|s,y,x)p(s,y,x). \tag{67}$$

Since the biography depends on the stratum only via the initial state $p(t|s,y,x) = p(t|y,x)$, which is (65). The second component of (67) is

$$\begin{aligned} p(s,y,x) &= h_s p(x|s)p(y|x,s) \\ &= (h_s/q_s)p(s|x)p_x p(y|x,s) \\ &= w_s p(s|x)p_x p(y|x,s). \end{aligned} \tag{68}$$

Here $p(y|x,0) = \mu_1^y \mu_2^{1-y}/\mu$, while for $s = 1,2,3,4$, $p(y|x,s)$ is either 1 or 0 depending on which initial state is implied by the stratum selected. Also the $p(s|x)$ are given by (64). Hence, for feasible combinations of stratum and initial state,

$$p(s,y,x) = \left[w_0 \frac{\mu_1^y \mu_2^{1-y}}{\mu}\right]^{s_0} \left[w_1 \frac{\mu_1}{\mu}\right]^{s_1} \left[w_2 \frac{\mu_2}{\mu}\right]^{s_2} \left[w_3 \frac{\Delta}{\mu}\right]^{s_3} \left[w_4 \frac{\Delta}{\mu}\right]^{s_4} p_x. \tag{69}$$

The marginal distribution of $x$ (induced by stock/flow sampling) is got by summing over $s, y$ to give

$$r_x = p(x) = p_x \sum_{k=0}^{4} w_k P_{kx} \tag{70}$$

as in (29). Note that in a random sample from the whole population, where $h_0 = 1, r_x$ reduces to $p_x w_0 P_{0x} = p_x$, whole population distribution of $x$, as it should. Otherwise $r_x$ differs from $p_x$.

We now form the first moment conditions by considering the conditional distribution of the data given the covariate vector. This is

$$\begin{aligned} p(s,y,t|x) &= p(s,y,t,x)/r_x \\ &= \frac{p(s,y,x)}{r_x} p(t|y,x). \end{aligned} \tag{71}$$

The first factor here is (69) divided by (70). The second factor is (65). Notice that, as in section 3, $p_x$ has cancelled from this conditional distribution. The first set of moments are the $\theta$ scores from this conditional distribution.

$$\psi_1(\theta, h, q) = \frac{\partial}{\partial\theta} \log p(s,y,t|x). \tag{73}$$

The second set of moments recognise the fact that $E(s_j) = h_j, j = 0,1,2,3,4$. Hence

$$\psi_2(h) = \begin{pmatrix} h_1 - s_1 \\ h_2 - s_2 \\ h_3 - s_3 \\ h_4 - s_4 \end{pmatrix} \tag{74}$$

Note that $\psi_2$ has only four elements since the $h$'s sum to one.

The final set of moments expresses the fact that the $q$'s are the means of the conditional stratum probabilities. Thus

$$\begin{aligned} q_j &= \sum_x P_{jx} p_x \\ &= \frac{q_j}{h_j} \sum_x \frac{w_j P_{jx}(\theta^*)}{\sum_k w_k P_{kx}(\theta^*)} p_x \sum_k w_k P_{kx}(\theta^*), \quad j = 1,3. \end{aligned} \tag{75}$$

Thus the moments are

$$\psi_3(\theta, h, q) = \begin{pmatrix} h_1 - R_{1x}(\theta) \\ h_3 - R_{3x}(\theta) \end{pmatrix} \tag{76}$$

where

$$R_{j\mathbf{x}} = w_j P_{j\mathbf{x}} / \sum_k w_k P_{k\mathbf{x}}.$$

Since $q_0 = 1, q_2 = 1 - q_1, q_4 = q_3$, only two such conditions are required.

If $\Delta$ is the covariance matrix of $\psi = (\psi_1, \psi_2, \psi_3)$ the method of moments estimator of $\theta^*$ is the $\theta$ solution of the problem[18]

$$min_{\theta,h,q}\, \overline{\psi}' \Delta^{-1} \overline{\psi}.$$

Note that if $q$ is known the minimisation is only with respect to $\theta, h$. The asymptotic covariance matrix of $\sqrt{N}(\hat{\gamma} - \gamma^*)$ is given by

$$V = (\Gamma' \Delta^{-1} \Gamma)^{-1},$$

where $\Gamma = E(\partial\psi / \partial\gamma)$. It is $V$ that we evaluate numerically for alternative choices of $T$ and $h$. We consider specifically three cases

$$\begin{array}{lll} 1. & T = 12; & q_1,\, q_3 \text{ unknown.} \\ 2. & T = 12; & q_1,\, q_3 \text{ known.} \\ 3. & T = 1; & q_1,\, q_3 \text{ known.} \end{array} \tag{77}$$

We shall compare $V$ for these three cases with the reference case of simple random sampling and observation of the biography for 84 months. We shall refer to this as case 0. The sampling schemes for which we shall report results are

1. $h = (1, 0, 0, 0, 0)$;
2. $h = (0, 0.25, 0.25, 0.25, 0.25)$;
3. $h = (0, 0.33, 0.33, 0.17, 0.17)$.

[18] $\overline{\psi}$ contains the sample means of the moment conditions.

The first is simple random sampling while the second and third are two types of equal shares stock/flow sampling. Case 3, which takes equal numbers from the flow and both stocks, matches the definition of equal shares used in section 3.

The matrix $V$ was calculated by simulating 50,000 realisations of an equilibrium alternating Poisson process with parameter values given above and with $x$ uniform in the population over -15 to 15. This implies that the covariance matrix $\Sigma$ defined in section 3 is diagonal with $\sigma_x^2 = 75$. These realisations were used to construct $\psi$ and hence to calculate $\Delta$ and $\Gamma$, leading to an estimate of $V$. The results are given in the table below and diagrammatically in figures 1 to 3.

**Asymptotic variances of coefficient estimates under alternative sampling schemes and observation periods.**

| $h_1$ | $h_2$ | $h_3$ | $h_4$ | $h_5$ | $T$ | $q_1$ | $q_3$ | $\theta_{10}$ | $\theta_{11}$ | $\theta_{20}$ | $\theta_{21}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 4.000 | 0.043 | 5.600 | 0.030 |
| 1 | 0 | 0 | 0 | 0 | 84 | | | 3.03 | 0.04 | 2.99 | 0.04 |
| | | | | | 12 | | | 16.14 | 0.21 | 16.14 | 0.21 |
| | | | | | 12 | $\times$ | $\times$ | 1.34 | 0.21 | 1.16 | 0.21 |
| | | | | | 1 | $\times$ | $\times$ | 14.58 | 2.47 | 14.37 | 2.47 |
| 0 | 1/4 | 1/4 | 1/4 | 1/4 | 12 | | | 10.58 | 0.05 | 38.65 | 0.07 |
| | | | | | 12 | $\times$ | $\times$ | 0.18 | 0.05 | 0.13 | 0.07 |
| | | | | | 1 | $\times$ | $\times$ | 0.20 | 0.08 | 0.17 | 0.08 |
| 0 | 1/3 | 1/3 | 1/6 | 1/6 | 12 | | | 10.31 | 0.05 | 40.70 | 0.07 |
| | | | | | 12 | $\times$ | $\times$ | 0.23 | 0.05 | 0.19 | 0.07 |
| | | | | | 1 | $\times$ | $\times$ | 0.26 | 0.08 | 0.24 | 0.08 |

Notes: The first five columns give the sampling probabilities; $T$ is the length of the state biography that is observed; $q_1$ and $q_3$ are the marginal probabilities of strata 1 and 3, a $\times$ indicates they are known; the remaining columns give the asymptotic variances of the method of moments estimates with the true values at the top.

In the figures the axes measure the precision — one over the variance – of the two slope coefficients estimates $\hat{\theta}_{11}$ and $\hat{\theta}_{21}$, or rather of $\sqrt{N}(\hat{\theta}_{j1} - \theta^*)$. The points marked $0, 1, 2, 3$

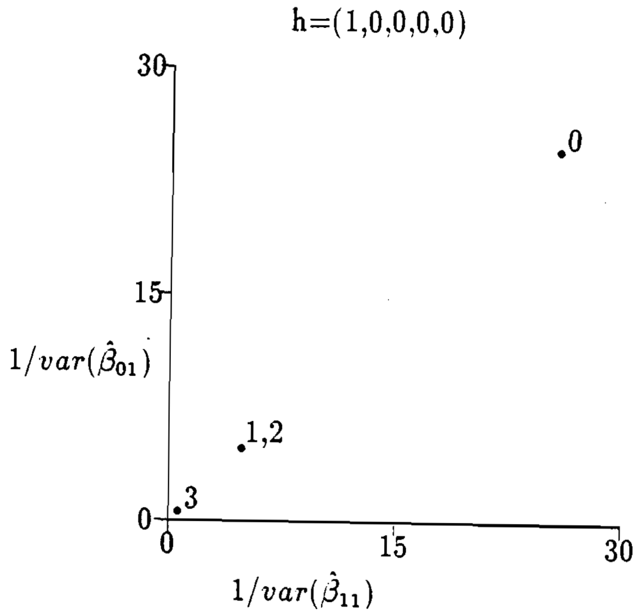
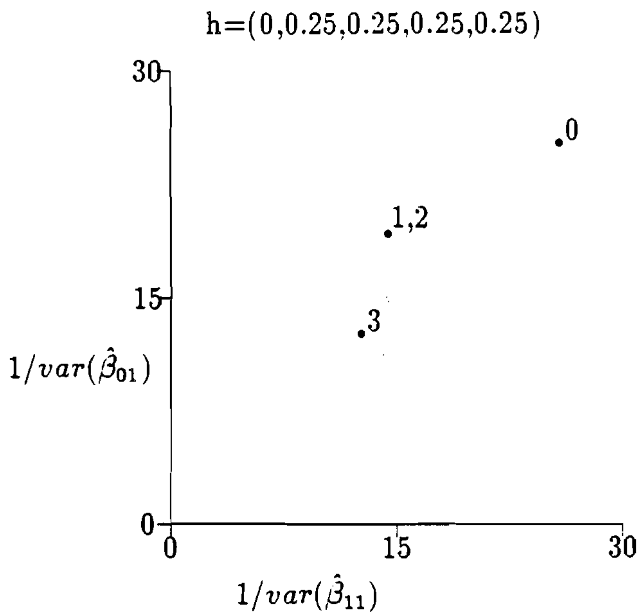



Figure 1


h=(0,0.25,0.25,0.25,0.25)
30
15
0
0
15
30
0
1,2
3
$1/var(\hat{\beta}_{01})$
$1/var(\hat{\beta}_{11})$

2

Figure 2

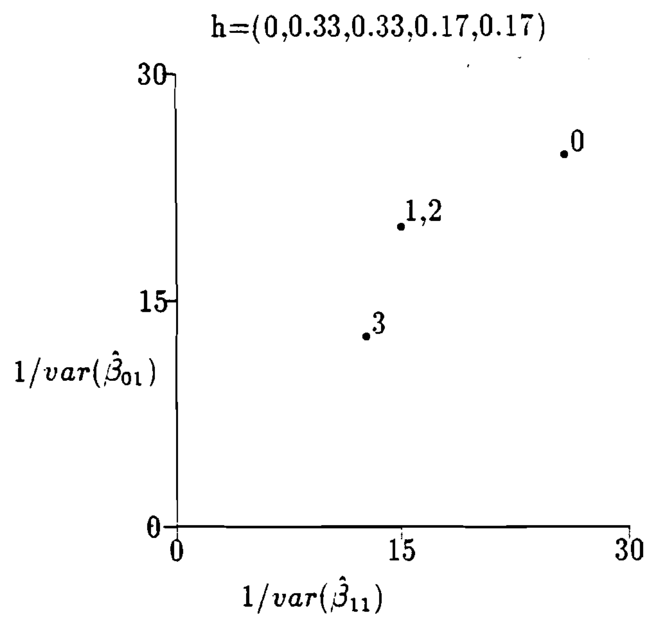

h=(0,0.33,0.33,0.17,0.17)
30
15
0
0
15
30
0
1,2
3
$1/var(\hat{\beta}_{01})$
$1/var(\hat{\beta}_{11})$

Figure 3

3

plot these vectors for the three cases given in (77) together with the reference case 0. Each figure refers to a different $h$ vector. The further a point is from the origin the more precise the estimate. If two points fall on the 45 degree 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.

Figure 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 0 is about four times further from the origin as the (coincident) points labeled 1 and 2 we see that observing for seven times as long is equivalent to having four times as many observations. Since points 1 and 2 coincide, knowledge of the $q$'s has virtually no effect on the precision of the slope estimates when observation is for only twelve months.

Figures 2 and 3 depict pure stock/flow sampling schemes, together with the reference case. Comparing these with figure 1 we see a dramatic improvement in the precision of the shorter observations schemes. Comparing point 3 to point 0 we see that the same precision 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 a quite remarkable result. Inevitably, however, the euphoria must be qualified. These people move only slowly between states so that if, as with random sampling, all 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.

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 one month scheme and 0.07 for the 12 month scheme. If we refer back to section 3.3 we gave there the covariance matrix of the equal shares stock/flow sampling scheme with no biographical data when the true slopes were zero. In particular the slope variances were equal and given by $3\times2 = 6$ times the lower right element of $\Sigma^{-1}$. But $\Sigma$ 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 month's observation. Hence we can argue that (a) one month's 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 even when the true slopes are not zero.[19]

[19]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.

## 5. Summary and Conclusions.

We have examined some statistical consequences of viewing homelessness and similar phenomena as an alternating renewal process conditional on some time invariant covariate vector. We assumed that an investigator, perhaps for predictive purposes, wishes to calculate the effect of the covariates on various aspects of the process. Our main results are as follows.

1. To gather data to provide the basis for estimates of these effects one can do significantly better than randomly sampling the population at risk.

2. A useful sampling scheme is to randomly sample the occupants of each state and those who move from state to state. This is a stock/flow sampling scheme.

3. Such a scheme is a dynamic version of the choice–based sampling scheme discussed in the econometrics of discrete choice.

4. We have given an asymptotically efficient and computationally simple estimator for both the choice–based sampling scheme and its stock/flow extension. Depending on the parametrisation used this procedure can reduce to a multinomial Logit calculation.

5. A stock/flow scheme which takes equal numbers from each stock and the flow is optimal when regression effects are zero and can be expected to be not far from optimal, usually, otherwise.

6. When individuals are followed through time in order to observe their state transitions an efficient stock/flow sample with a brief period of observation can yield as much information as a random sample followed for much longer.

7. When the correct model is an alternating renewal process disturbed by unobserved inter–personal heterogeneity the method of moments estimator yields estimates which are typically attenuated, and this attenuation can be severe. However, estimates of elasticities of state occupancy probabilities with respect to covariates are, apparently, less severely affected by this misspecification.

It seems reasonable to conclude from these results that in designing surveys of the homeless it would be sensible to consider a balanced stock/flow scheme, and to make inferences from it with the method of moments procedure that we have described.[20]

[20]We have not mentioned the costs of sampling individuals in different ways and these

## Appendix

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 $\theta^*$'s in section 4 are

| | |
|---|---|
| $\theta_{10}$ | 0.11 |
| $\theta_{11}$ | 0.013 |
| $\theta_{20}$ | 0.10 |
| $\theta_{21}$ | 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.

would of course have to be taken into account in computing an optimal design.

**References**

Cosslett, S. R. 1981. "Efficient Estimation of Discrete Choice Models." In Manski and McFadden (1981).

Cox, D. R. 1962. Renewal Theory. London: Methuen.

Feller, W. 1966. An Introduction to Probability Theory and Its Applications. New York: John Wiley.

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.

Imbens, G. W. 1990. "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling." Discussion Paper 9009, Center for Economic Research, Tilburg University.

Knickman, J., Weitzman, and E. Marcus. 1989. "A Study of Homeless Families in New York City." Department of Social Administration, New York University.

Magnus, J. R. and H. Neudecker. 1989. Matrix Differential Calculus. New York: John Wiley.

Manski, C. F. and D. McFadden, eds. 1981. Structural Analysis of Discrete Data with Econometric Applications. Cambridge, Mass.: MIT Press.

Manski, C. F. and D. McFadden. 1981. "Alternative Estimators and Sample Designs for Discrete Choice Analysis." In Manski and McFadden (1981).