# Econometric Issues in the Analysis of Linked Longitudinal Employer-Worker Data.

Andrew K. G. Hildreth
University of California-Berkeley

Stephen E. Pudney
University of Leicester

January 28, 2000

**Abstract**

Using matched employer-worker data for Britain, taken from Government surveys, we examine the implications of differential sampling rates over time in its effects on the consistency of estimates in an econometric model of turnover and job separations. The problem for matching occurs when the Govenment survey changes the sampling ratios within strata. This affects both the probability of observing a particular employer from one year to the next, but also whether a match between a worker and employer is observed in the matched data over time. We present methods for deriving the conditioning probabilities for inclusion in a random effects model estimated by maximum likelihood.

## 1 Introduction

## 2 A model of the sampling / linking process

### 2.1 The cross-section sample distribution

Define the following notation, relating to the first of our survey years (1994), indicated by the subscript 1. At this first wave, we have three sources of information:

($i$) NES information on individuals who are in employment, denoted $w_1$;

($ii$) information on individuals' current and past experience of unemployment, denoted $u_1$;

($iii$) ACOP information on the characteristics of the employer, denoted $f_1, s_1$, where $s_1$ is firm size measured by employment.

At the second wave of observation, three new sets of information, $w_2$, $u_2$ and $f_2$ are produced from the same sources, with $u_2$ giving retrospective information on any unemployment occurring between the two waves. In general, the observed data is not complete: if an individual is not in work at wave $t$ then $w_t$ and $\{f_t, s_t\}$ will not exist; and even where an individual is observed in employment,

his or her employer may not be captured by ACOP, and thus $\{f_t, s_t\}$ would be unobserved.

At each wave, an individual must be in one of four possible observational states: employed, with ACOP information; employed, without ACOP information; unemployed; and out of the labour force. Combining the two waves, there are 16 observational regimes, corresponding to all possible combinations of the 4 possible states at waves 1 and 2. We are interested in modelling the processes governing job loss, job change and unemployment duration, together with their consequences for subsequent labour market experience. For this, we certainly need to know something about the characteristics of the initial job match, so that we can investigate the influence of employer, worker and job characteristics on the probability of separation. Thus only 4 of the 16 observational regimes are directly useful to us: those where the individual is employed by an ACOP-observed firm at wave 1. For the rest of this paper, we will condition on this regime by discarding all observations where the individual is not with an ACOP-observed employer at wave 1.[1]

Our task is now to derive the sample distribution of the remaining observations, to provide a basis for drawing inferences about the processes of job loss and unemployment duration. From the viewpoint of the theoretical statistician, there is a serious problem to be overcome in analysing the dataset that results from the firm-worker matching process described in section 2. The techniques customarily used to model survey data are based on the assumption of an underlying continuous distribution. For example, probit or Tobit analysis requires that behavioural disturbances are drawings from a normal distribution; logit analysis is based on the logistic distribution. But a distribution can only be continuous if the population it describes is infinite. Since the total number of firms and workers in existence at any time is very large, one usually feels safe in using these infinite-population methods for analysing company or worker cross-sections. However, with matched data, this is more problematic. If there are infinite numbers of firms and workers, then the probability of observing even a single firm-worker match is essentially zero in any finite sample, unless there exist firms of infinite size. There is a further problem in the case of the ACOP, since the survey design requires that every (manufacturing) firm with over 250 employees is sampled. In an infinite population of firms, this would generally imply an infinite sample of large firms.

Fortunately, a suitable theoretical framework is available for situations like this. Superpopulation theory (Cassel *et. al.*, 1977; Pudney, 1989) allows us to work with a finite population, by postulating the existence of an underlying infinite superpopulation from which the actual finite population is assumed to have been drawn at random by "nature". Essentially, the superpopulation describes the set of possible forms the actual finite population might have taken. The objective of our statistical analysis is then to estimate fundamental relationships present in the superpopulation - in other words the (random) processes that

[1] Note that this restriction of the sample prevents us from isolating worker-specific unobservables that might influence the initial state. However, a long run of repeat observations is not available to us here, so this does not greatly reduce the scope of the analysis.

govern the nature of the actual population we see around us. Since the superpopulation is infinite, it is admissible to estimate these statistical relationships using techniques which assume continuous distributions. However, in doing so, it is necessary to take proper account of the fact that the process generating our data has two stages - a draw (made by "nature") from the superpopulation, followed by a second draw (made by the survey designer) from the finite population.

At the time of wave 1, the superpopulation consists of an infinite set of elements, each corresponding to a manufacturing firm and its workforce. For the moment, we suppress the subscript 1 which identifies wave 1 variables, and also ignore the unemployment information, $u_1$.[2] The size of the firm's workforce is $s$, and the information set describing the firm and its $s$ workers is denoted $X = \{s, f, w^1 ... w^s\}$. In the superpopulation, the distribution of these firm/workforce clusters is described by a probability density/mass function (pdf) $g(X)$. Our sample falls into two parts: an exhaustive sample of firms in the section of the population for which $s \geq 250$; and a random sample (without replacement) from the section of the population for which $s < 250$. The total sample size and the numbers of observations from these two parts of the population are $n = n^* + n^{**}$. We consider the two parts of the sample separately. Henceforth, we use the symbol $g(\,.\,)$ as generic notation for any distribution that describes the superpopulation; the symbol $h(\,.\,)$ is used to represent sample distributions.

*Large firms*

Let the size of the actual population be $N$ firms, and let $P = G_s(250)$ be the frequency of large firms in the superpopulation, where $G_s$ is the cumulative distribution function (cdf) of firm size in the superpopulation. The number of large firms in the actual population, $N^*$, is therefore random and distributed as binomial $(N, P)$. Exhaustive sampling implies that the number of sampled large firms, $n^*$, is exactly equal to $N^*$. Conditional on this number, we can regard the sampled large firms as a simple random sample of size $n^*$, drawn directly from the superpopulation. Thus, the joint distribution of all potentially observable information relating to the sampled large firms is:

$$h(n^*, X_1 ... X_{n^*}) = \binom{n^*}{N} P^{n^*} (1-P)^{N-n^*} \prod_{j=1}^{n^*} g(X_j | s_j \geq 250) \qquad (1)$$

However, in general the collection of variables $X$ is not fully observable, since the NES is a 1 in (approximately) 100 random sample of NI numbers. Thus, any particular worker has a known probability $\rho \approx 0.01$ of being observed in the NES. Conditional on the size of the firm, $s$, the number of workers captured by the NES $(q)$ therefore has a binomial $(s, \rho)$ distribution. Letting the symbol

[2] For the purposes of cross-section analysis of wave 1 data, the unemployment information $u_1$ can be treated in exactly the same way as the NES information $w_1$, so the omission of $u_1$ entails no loss of generality.

$\widetilde{X}$ denote the part of $X$ that is observed, the joint sample distribution for large firms is therefore:

$$\begin{aligned} h(n^*, \widetilde{X}_1 ... \widetilde{X}_{n^*}) &= \binom{n^*}{N} P^{n^*} (1-P)^{N-n^*} \\ \times \prod_{j=1}^{n^*} g(f_j | s_j &\geq 250) \binom{q_j}{s_j} \rho^{q_j} (1-\rho)^{s_j - q_j} \prod_{i=1}^{q_j} g(w_j^i | f_j) \end{aligned} \tag{2}$$

*Small firms*

The remainder of the sample is a set of $n^{**}$ small firms, where $n^{**}$ is a fixed number chosen as part of the survey design. Under the superpopulation approach, observations on small firms ($s < 250$) are viewed as being generated as a random sample (without replacement) drawn from a random sample drawn from the relevant part of an underlying infinite population. But a random sample of a random sample is itself a random sample, so the joint distribution of information relating to sampled small firms is:

$$h(X_{n^*+1} ... X_n) = \prod_{j=n^*+1}^{n} g(X_j | s_j < 250) \tag{3}$$

Taking account of the random sampling of workers within firms as we did before, the full sample distribution for all observed firms and workers is:

$$\begin{aligned} h(n^*, \widetilde{X}_1 ... \widetilde{X}_n) &= \binom{n^*}{N} (1-P)^{N-n} \\ &\times \prod_{j=1}^{n^*} g(f_j, s_j) \binom{q_j}{s_j} \rho^{q_j} (1-\rho)^{s_j - q_j} \prod_{i=1}^{q_j} g(w_j^i | f_j, s_j) \end{aligned} \tag{4}$$

where we have used the relationships $g(f, s | s \geq 250) = g(f, s)/P$ and $g(f, s | s < 250) = g(f, s)/(1-P)$.

## 2.2 Population dynamics

We now need to extend this distribution to cover wave 2 information, so that labour market transitions and unemployment durations can be modelled. This brings in the possibility of moves by workers into the non-manufacturing sector, which is not sampled by the ACOP. To take account of this, we introduce a new variable at wave 2, a binary indicator $m$ equal to 1 if the firm is in the manufacturing sector, and 0 otherwise. We do not allow the possibility that firms change sector. We now need to specify the dynamic processes governing the development of firms and their workforces. Define the following components.

### 2.2.1 Separation probabilities

The following four probabilities cover all possible eventualities for the labour force status of an individual during the time interval between the dates of the two waves. They are: continuous employment with the same firm throughout; a change to a new employer, with no intervening spell of unemployment; a move into unemployment (whether still in progress or not at the time of wave 2); and a move out of the labour force.

$$\begin{aligned}
&\Pr(\text{no job change}|f_1, w_1) = Q_0(f_1, s_1, w_1) \\
&\Pr(\text{new employer with characteristics } f_2, s_2, m_2|f_1, s_1, w_1) \\
&\qquad\qquad = Q_1(f_2, s_2, m_2|f_1, s_1, w_1) \\
&\Pr(\text{move into unemployment}|f_1, s_1, w_1) = Q_2(f_1, s_1, w_1) \\
&\Pr(\text{move out of labour force}|f_1, s_1, w_1) = Q_3(f_1, s_1, w_1)
\end{aligned} \tag{5}$$

Note that, since there is a finite population of firms, $Q_1$ is a discrete probability with respect to firm identities, and should be interpreted as being conditional on the characteristics of the set of firms in existence at wave 2. $Q_1$ governs transitions to employers outside the manufacturing sector (who cannot be observed in ACOP), as well as transitions within manufacturing; the variable $m_2$ distinguishes these cases.

### 2.2.2 Firm development

The sampling process and behavioural processes like wage determination are linked to the characteristics of the firm. Thus we need to model the evolution of those characteristics. We write the pdf and cdf of possible outcomes for the firm, conditional on its wave 1 characteristics, as $g(f_2, s_2|f_1, s_1)$ and $G(f_2, s_2|f_1, s_1)$ respectively. Note that the possible outcomes could in principle include firm death, merger, etc., and firm birth should also be captured here. In general, therefore, this will be a complex multi-part distribution with both discrete and continuous elements. However, to avoid undue complication, we assume here that there are no firm births or deaths.

### 2.2.3 Unemployment duration and post-unemployment destination

For a worker who leaves employment to start a spell of unemployment, we need an assumption about the distribution of unemployment duration, and the probabilities of ending the spell by a move to a particular employer. The function specified here is a probability density with respect to unemployment duration $u_2$, but a discrete probability with respect to firm identities, conditional on the characteristics of the set of firms in existence at wave 2.

$$\Pr(\text{exit to job in firm } f_2, s_2, m_2 \text{ after duration } u_2 \mid u_1, f_1, s_1, w_1)$$

$$= g(u_2, f_2, s_2, m_2 | u_1, f_1, s_1, w_1) \tag{6}$$

The probability that the unemployment spell will still be in progress at wave 2 conditional on the job loss occurring a fraction $t$ of a year after wave 1 is:

$$\Pr(u_2 > 1 - t | u_1, f_1, s_1, w_1) = 1 - G_{u_2}(1 - t | u_1, f_1, s_1, w_1) \tag{7}$$

We treat the timing of the unemployment spell, $t$, as exogenous and thus do not explicitly incorporate its distribution. We also assume that there are no multiple unemployment spells between waves, or, if there are, that $u_2$ represents their total length. For simplicity, transitions from unemployment out of the labour force are also ignored.

#### 2.2.4 Determination of earnings and other job characteristics

We allow for a pair of wage relationships here, with different processes governing the initial earnings of a new recruit and an incumbent employee. Any unemployment experienced between waves 1 and 2 may affect the starting wage of a new recruit.

$$\text{incumbent workers:} \qquad g^I(w_2 | w_1, f_1, s_1, f_2, s_2, m_2) \tag{8}$$

$$\text{new recruits:} \qquad g^N(w_2 | w_1, f_1, s_1, f_2, s_2, u_2, m_2) \tag{9}$$

### 2.3 The 2-wave sample distribution

To derive the full two-wave sample distribution, we start by looking at the possible outcomes for each of the firms sampled at wave 1. There are two possibilities: the firm may be observed at wave 2 or not, depending on its size and the sampling draw. To bring these outcomes into the distribution (4), the term for firm $j$ must be updated by multiplying by a conditional probability term. Define $K$ to be the sampling fraction among small firms at wave 2.[3] Then the extra multiplicative term to be included in the distribution (4) is one of the following two forms:

*Firm observed in ACOP:*

$$\lambda_j = \Psi_j^i \; g(f_{j2}^i, s_{j2}^i | f_{j1}, s_{j1}) \tag{10}$$

[3] In other words, $K$ is the ratio of $n^{**}$ to the number of small firms in existence at wave 2. Strictly speaking, the latter is random under our assumptions, but nothing is lost by assuming $K$ to be a constant, as we do henceforth.

*Firm not observed in ACOP:*

$$\lambda_j = (1-K)\ g(f^i_{j2}, s^i_{j2}|f_{j1}, s_{j1}) \tag{11}$$

where $\Psi^i_j$ is the following function of $s^i_{j2}$:

$$\begin{aligned} \Psi^i_j &= 1 && s^i_{j2} \geq 250 \\ &= K && s^i_{j2} < 250 \end{aligned} \tag{12}$$

We then need to update each of the workers observed at wave 1. Take the $i$th worker observed for firm $j$ and multiply by a conditional probability term, $\Phi^i_j$, which takes different forms for the following seven different types of outcome that could occur for the worker. Denote these possible forms for the $i$th observed worker in firm $j$ by $\Phi^i_{j1}...\Phi^i_{j7}$, where:

*Continuous employment with one firm*

$$\Phi^i_{j1} = Q_0(f_{j1}, s_{j1}, w^i_{j1}) g^I(w^i_{j2}|w^i_{j1}, f_{j1}, s_{j1}, f^i_{j2}, s^i_{j2}) \tag{13}$$

*Job change within manufacturing, no intervening unemployment*

$$\begin{aligned} \Phi^i_{j2} &= Q_1(f^i_{j2}, s^i_{j2}, m_{j2}=1|f_{j1}, s_{j1}, w^i_{j1}) \\ &\quad \times g^N(w^i_{j2}|w^i_{j1}, f_{j1}, s_{j1}, f^i_{j2}, s^i_{j2}, m_{j2}=1, u_{j2}=0) \end{aligned} \tag{14}$$

*Change to job outside manufacturing, no intervening unemployment*

$$\begin{aligned} \Phi^i_{j3} &= Q_1(f^i_{j2}, s^i_{j2}, m_{j2}=0|f_{j1}, s_{j1}, w^i_{j1}) \\ &\quad \times g^N(w^i_{j2}|w^i_{j1}, f_{j1}, s_{j1}, f^i_{j2}, s^i_{j2}, m_{j2}=0, u_{j2}=0) \end{aligned} \tag{15}$$

*Move into unemployment, spell still in progress at wave 2*

$$\Phi^i_{j4} = Q_2(f_{j1}, s_{j1}, w^i_{j1})[1 - G_{u_2}(1 - t^i_j|u^i_{j1}, f_{j1}, s_{j1}, w^i_{j1})] \tag{16}$$

*Move into unemployment, but spell ends with job in manufacturing*

$$\begin{aligned} \Phi^i_{j5} &= Q_2(f_{j1}, s_{j1}, w^i_{j1}) g(u^i_{j2}, f^i_{j2}, s^i_{j2}, m_{j2}=1|u^i_{j1}, f_{j1}, s_{j1}, w^i_{j1}) \\ &\quad \times g^N(w^i_{j2}|w^i_{j1}, f_{j1}, s_{j1}, f^i_{j2}, s^i_{j2}, m_{j2}=1, u^i_{j2}) \end{aligned} \tag{17}$$

*Move into unemployment, but spell ends with job outside manufacturing*

$$\begin{aligned} \Phi^i_{j6} &= Q_2(f_{j1}, s_{j1}, w^i_{j1}) g(u^i_{j2}, f^i_{j2}, s^i_{j2}, m_{j2}=0|u^i_{j1}, f_{j1}, s_{j1}, w^i_{j1}) \\ &\quad \times g^N(w^i_{j2}|w^i_{j1}, f_{j1}, s_{j1}, f^i_{j2}, s^i_{j2}, m_{j2}=0, u^i_{j2}) \end{aligned} \tag{18}$$

*Exit from labour force*

$$\Phi^i_{j7} = Q_3(f_{j1}, s_{j1}, w^i_{j1}) \tag{19}$$

If we were able to observe every employer in wave 2 of ACOP, there would be no further difficulty: the 2-wave sample distribution would be the sample distribution for the wave 1 observations (given by equation (4)) multiplied by the conditional probability term $\lambda_j$ for each firm, and by another term $\Phi^i_j$ for each worker, where $\Phi^i_j$ is defined as follows.

$$\Phi^i_j = \sum_{k=1}^{7} \varsigma^i_{jk} \Phi^i_{jk} \tag{20}$$

where $\varsigma^i_{jk} = 1$ if outcome $k$ is observed, and $\varsigma^i_{jk} = 0$ otherwise.

However, there is an observational problem. These seven outcomes are not directly observable. Instead, there are ten possible types of observation, each identified by a binary variable, $\xi^i_{jk}$, $k = 1...10$, where these ten observable outcomes are as set out in the following table.

**Table 2** Observable transitions between waves

| $k$ | **Observational regime** | **Probability element** | **Sample numbers** |
|---|---|---|---|
| 1 | Continuous employment, with observed employer | $\Phi^i_{j1}\Psi^i_j$ | 7394 |
| 2 | Change to new observed employer, no unemployment | $\Phi^i_{j2}\Psi^i_j$ | 2190 |
| 3 | Change to same or new observed employer, with unemployment | $\Phi^i_{j5}\Psi^i_j$ | 862 |
| 4 | Change to non-manufacturing employer, no unemployment | $\Phi^i_{j3}$ | 508 |
| 5 | Change to unobserved manufacturing employer, no unemployment | $\Phi^i_{j2}(1-\Psi^i_j)$ | 1051 |
| 6 | Employed in manufacturing, no employer observed at wave 2 | $\left(\Phi^i_{j1}+\Phi^i_{j2}\right) \times (1-\Psi^i_j)$ | 1470 |
| 7 | Change to non-manufacturing employer, with unemployment | $\Phi^i_{j6}$ | 8 |
| 8 | Change to unobserved manufacturing employer, with unemployment | $\Phi^i_{j5}(1-\Psi^i_j)$ | 43 |
| 9 | Unemployment spell in progress | $\Phi^i_{j4}$ | 160 |
| 10 | Out of the labour force | $\Phi^i_{j7}$ | |

The appropriate wave 2 probability element is then:

$$\begin{aligned}\Lambda^i_j &= \xi^i_{j1}\Phi^i_{j1}\Psi^i_j + \xi^i_{j2}\Phi^i_{j2}\Psi^i_j + \xi^i_{j3}\Phi^i_{j5}\Psi^i_j + \xi^i_{j4}\Phi^i_{j3} \\ &\quad + \xi^i_{j5}\Phi^i_{j2}(1-\Psi^i_j) + \xi^i_{j6}(\Phi^i_{j1}+\Phi^i_{j2})(1-\Psi^i_j) \\ &\quad + \xi^i_{j7}\Phi^i_{j6} + \xi^i_{j8}\Phi^i_{j5}(1-\Psi^i_j) + \xi^i_{j9}\Phi^i_{j4} + \xi^i_{j10}\Phi^i_{j7} \end{aligned} \tag{21}$$

Now define the following two sets of workers:

$$\begin{aligned}\Omega^0 &= \{i,j \mid f^i_{j2}, s^i_{j2} \text{ unobserved}\} \\ \Omega^1 &= \{i,j \mid f^i_{j2}, s^i_{j2} \text{ observed}\}\end{aligned} \tag{22}$$

In observational regimes 4...8, the wave 2 employer is not observed in ACOP, and therefore the unobserved firm characteristics $f^i_{j2}, s^i_{j2}$ are unobserved and must be integrated out of the sample distribution. The resulting rather cumbersome expression is as follows:

$$\begin{aligned} h(\,sample\,) &= \binom{n^*}{N}(1-P)^{N-n} \prod_{j=1}^{n}\binom{q_j}{s_{j1}}\rho^{q_j}(1-\rho)^{s_{j1}-q_j}\, g(f_{j1},s_{j1}) \\ &\quad \times \prod_{i,j\in\Omega^1} g(w^i_{j1}|f_{j1},s_{j1})\ \Lambda^i_j\, g(f^i_{j2},s^i_{j2}|f_{j1},s_{j1}) \\ &\quad \times \prod_{i,j\in\Omega^0} g(w^i_{j1}|f_{j1},s_{j1}) \int \Lambda^i_j\, dG(f^i_{j2},s^i_{j2}|f_{j1},s_{j1}) \end{aligned} \tag{23}$$

# 3 Implications for econometric analysis

The rather complicated sampling distributions derived in the previous section have some important implications for the use of standard econometric techniques with linked NES/ACOP data. To draw out these implications, we consider in this section some purposes for which the dataset might be used, and indicate which of these might be affected by bias induced by the complex sampling process. We begin with the problem of estimating a simple earnings regression.

## 3.1 Estimating a cross-section earnings equation

For some practical purposes, we may be interested in the distribution of one or more worker variables conditional on the characteristics of the firm and remaining characteristics of the worker. This is so, for example, if we use the data to estimate an earnings regression. To find this conditional distribution, we divide the worker variables $w$ into an endogenous variable $w^*$ (earnings) and

the remaining conditioning variables $w^{**}$ (age, gender, etc.). We then need to integrate out the endogenous worker variable $w^*$ to derive the marginal distribution of the conditioning variables $\{f, s, w^{**}, q, n^*\}$, and then divide the full sample distribution by this marginal. When this is done, many terms in the sample distribution (4) cancel, and the result is the following conditional sample distribution:

$$\begin{aligned} h(w^*|f,s,q,n^*,w^{**}) &= \prod_{j=1}^{n}\prod_{i=1}^{q_j}\frac{g(w_j^i|f_j,s_j)}{g(w_j^{**i}|f_j,s_j)} \\ &= \prod_{j=1}^{n}\prod_{i=1}^{q_j} g(w_j^{*i}|f_j,s_j,w_j^{**i}) \end{aligned} \tag{24}$$

The important result here is that this is essentially identical to the distribution of $w^*$ conditional on $\{f, s, w^{**}\}$ in the superpopulation. Thus, the usual type of sample analysis of earnings conditional on firm and worker characteristics will give valid inferences about the underlying (super)population.

### 3.2 Estimating a cross-section model of firm variables

The same general conclusion is true if we derive the distribution of firm characteristics $f$ conditional on firm size $s$. However, any use of the ACOP data to analyse firm size, either unconditionally or conditional on other firm characteristics (or observed worker characteristics) will produce biased results unless we make some allowance for the non-uniform ACOP sampling rates. For example, the distribution of the $s_j$ conditional on the $f_j$ is:

$$\begin{aligned} h(s_1 \ldots s_n|f_1 \ldots f_n, n^*) &= \prod_{j=1}^{n}\frac{g(s_j|f_j)}{G_{s|f}(250)^{(1-\xi_j)}[1-G_{s|f}(250)]^{\xi_j}} \\ &\neq \prod_{j=1}^{n} g(s_j|f_j) \end{aligned} \tag{25}$$

where $\xi_j = 1$ if $s_j \geq 250$ and $\xi_j = 0$ otherwise. Thus, conventional unweighted sample-based models of firm size would give biased inferences about the (super)population distribution $g(s_j|f_j)$, and bias-corrected methods such as weighted ML or truncated regression are appropriate.

### 3.3 Estimating a 2-wave model of labour turnover and unemployment duration

To estimate the processes governing job separations and the durations of consequent unemployment spells, we need to use the 2-wave panel to give repeated

observations on workers and firms. Our aim here is to estimate the four separation probabilites $Q_0(.)...Q_3(.)$ and the distribution of unemployment spell lengths, $g(u_2|u_1, f_1, s_1)$. In practice, job separations and unemployment durations might be estimated separately, but the considerations involved are the same, and we deal with them together. The relevant conditional sample distribution here is:

$$h(\varsigma_1...\varsigma_7, u_2|f_1, s_1, w_1, u_1) = \frac{\int dH(\,sample\,)}{h(n^*, \widetilde{X}_1...\widetilde{X}_n)} \tag{26}$$

where the seven binary variables $\varsigma_k$ indicate the type of transition that occurs and $u_2$ is the duration of unemployment. The Stieltjes-Lebesgue integral in the numerator of (26) is with respect to the irrelevant variables $(f_2, s_2, m_2, w_2)$, and $H(.)$ is the cdf corresponding to the sample distribution (23). Performing the division and integration in (26):

$$h(\varsigma_1...\varsigma_7, u_2|f_1, s_1, w_1, u_1) = \prod_{j=1}^{n} \int \left[ \prod_{i=1}^{q_j} \int \Lambda_j^i \; dw_{j2}^i \right] \; dG(f_{j2}^i, s_{j2}^i | f_{j1}, s_{j1}) \tag{27}$$

However, there is an important *caveat* here. We have assumed that the variables $f_2$ and $s_2$ (and other unobservable firm characteristics) do not influence the separation probabilities. This may be unrealistic: if there are firm-specific factors (such as trading conditions) which influence the nature and size of the firm and also affect separation probabilities, then the simplicity of the estimation process will be lost. An extreme example of this problem is the case of a firm that is forced into liquidation: its size $s_2$ will fall to zero, and for all of its workers the probability of continuous employment with the firm must also be zero. In this more general setting, consistent estimation will require us to take account of the factors which are common to both separation probabilities and the state of the firm at wave 2.

### 3.4 Analysing the characteristics of new jobs

The major source of estimation difficulties with the linked NES/ACOP data is the non-observability of firm characteristics at wave 2 for a large majority of small firms.

## 4 A model of unemployment incidence and duration

**** Bit here on types of modelling that could be attempted.

### 4.1 The incidence of job loss and job change

**** A simple multinomial logit for 4 outcomes:
(1) no change of job
(2) job loss (i.e. transition to unemployment)
(3) transition to new employer
(4) transition to non-employment (no registered unemployment)

### 4.2 Unemployment duration

**** A simple model of unemployment duration, allowing for right-censoring

## 5 Conclusions

This paper has a number of objectives. We have described the construction of a new dataset formed from the British New Earnings Survey (NES) of employees and the Annual Census of Production (ACOP), covering manufacturing firms, with national insurance records used to provide additional information on periods of unemployment. This linked dataset is in effect a panel, with two waves in the years 1994 and 1995.

Secondly, we have looked at the feasibility of using this linked panel as a basis for various types of econometric analysis, especially the modelling of worker-employer separations, unemployment duration, and post-unemployment experience. We find that the relatively small number of firm-worker matches, particularly for workers with employers observed in ACOP at both waves, makes this kind of analysis difficult, unless a way can be found of exploiting the large number of partial observations that occur.

Thirdly, using a theoretical foundation in superpopulation sampling theory, we have considered the methodological problems raised by the incomplete coverage and non-uniform sampling design of the ACOP. We have established a number of results, which can be summarised in the following series of propositions.

**Proposition 1** For the simplest purposes of estimating a cross-section relationship such as an earnings equation relating the level of pay to firm and worker characteristics, conventional methods such as multiple regression will not be biased by the NES/ACOP sampling scheme, provided the estimated relationship is interpreted as holding only for jobs in manufacturing industry. However, any model with firm size as an endogenous variable will in general be affected by sample selection bias as a result of the non-uniform ACOP sampling rate.

**Proposition 2** For the purpose of estimating a model of firings and quits, using sample information from waves 1 and 2, the position is less straightforward. Provided there are no unobserved firm characteristics that influence *both* separation probabilities *and* observable firm characteristics (such as firm size),

conventional discrete-response models of separations and models of unemployment durations will be unaffected by sample selection biases. However, it is likely that such unobservables will exist (for example, poor trading conditions may underlie both raised separation probabilities for individual workers and a reduction in firm size), and then more complex estimators are called for.

**Proposition 3** For the purpose of estimating a model of the type of job that workers go into after separating from an employer, the NES/ACOP sampling scheme generates major complications, and estimation becomes considerably more difficult. However, we have suggested a possible avenue of progress.

# References


[1] Abowd, J. M., Kramarz, F., and Margolis, J. N. (1997), "High Wage Workers and High Wage Firms", *Econometrica*, forthcoming.

[2] Cassel, C., Särndal, C. E. and Wretman, J. H. (1977), *Foundations of Inference in Survey Sampling.* New York: Wiley.

[3] Coles, M. G., (1997), "Equilibrium Wage Dispersion, Firm Size and Growth", mimeo, Department of Economics, University of Essex.

[4] Gregory, M. and Jukes, R. (1997). "The Effects of Unemployment on Subsequent Earnings: A Study of British Men 1984-94", *Research Report to the Department of Employment, London.*

[5] Hildreth, A. and Pudney, S. E. (1996), "Employers, Workers and Unions: An Analysis of a Firm-Worker Panel with Endogenous Sampling, Attrition and Missing Data." University of Leicester Discussion Paper in Economics no. 96/15.

[6] Pudney, S. E. (1989). *Modelling Individual Choice. The Econometrics of Corners, Kinks and Holes.* Oxford: Blackwell.