Contributed Paper for the 54th Biennial Session of the International Statistical Institute, August 13-20, Berlin, Germany.

Model-Based Analysis of Labour Force Survey Gross Flow Data Under Informative Nonresponse

Gad Nathan
Hebrew University of Jerusalem, Israel
gad@huji.ac.il

Abdulhakeem Eideh
Alquds University, Palestinian Territories
msabdul@ppu.com and msabdul@palnet.com

Abstract

In this article we introduce new method of obtaining weighted estimates of gross flows, taking into account informative nonresponse. This method based on extracting the response labour force model as a function of the population labour force model and of the response probabilities, which are obtained as reciprocals of the adjusted calibrated weights. The new method is model based while the classical method is based on the adjusted weights. We think that the first method is more efficient than the weighted method. However the two methods, sample likelihood and weighting, give approximately the same estimates of labour force gross flows. Also we consider exponential and linear models to explain the variations in the calibrated weights under household level and from the exponential model we conclude that the unemployed persons at both quarters are under-represented in the labour force survey sample. This is similar to the result obtained by Clarke and Tate (1998) by comparing the estimates of gross flows under ignorable and nonignorable models which is very sophisticated model. The interesting result is that if we have sample data that contains the response variable and the sampling weights and for nonresponse the calibrated adjusted weights, then basing inference using classical weighted method and the new method based on the response likelihood may give similar results.

Key words: Informative nonresponse, logistic regression, sample likelihood method, weighted estimates.

1. Introduction

Labour Force Surveys are carried out in many countries. These surveys were often designed originally for cross-sectional analysis of households and individual data, so as to study labour force and other socio-economic characteristics on a current basis. Complex rotating sampling schemes have after been introduced in order to improve comparisons over time. For example, the quarterly Israel Labour Force Survey (LFS) employs a rotating panel sampling scheme whereby each unit in the sample is interviewed for two consecutive quarters; is left out of the sample for the next two quarters and then interviewed again for two more consecutive quarters.
Another example of LFS is the British Labour Force Survey, which is a household survey carried out by the Office for National Statistics (ONS) that gathers information on a wide range of labour force characteristics and other related topics. Since 1992,

the LFS has been conducted on a quarterly basis using a rotating sample design, where each household is retained in the sample for five consecutive quarters and then replaced. Thus, the 60000 sampled households from each quarter (around 120000 individuals) can be divided into five 'waves', with wave one households appearing in the sample for the first time, wave two household appearing for the second time, and so on until wave five. The survey is designed to produce cross-sectional data, but in recent years it has been recognized that linking together data on each individual across quarters could produce a rich source of longitudinal data, the uses of which include estimation of labour force gross flows. The process of producing linked data sets by linking records from samples at two quarters is relatively straightforward; the linked sample is simply the subset of individuals who responded in both quarters.

In the area of nonresponse several papers relate to the problem of informative (nonignorable) nonresponse in estimating labour force gross flows. Stasny (1986) considers the problem of using categorical data from a panel survey in which there is non-random nonresponse to estimate gross flows. The methods are illustrated for the case of estimating gross flows in labour force participation using the data from the Canadian Labour Force Survey. Three models are proposed that allow nonresponse to be related to employment classification, time, or to both employment classification and time. Maximum likelihood estimation is used to fit the models to a single panel of Labour Force Survey data. Clark and Chambers (1998) model the probability of a household nonresponse pattern as a weighted average of individual nonresponse probabilities and the household flow probabilities as multinomial, based on individual flow probabilities. By modeling both the labour force flow frequencies and the nonresponse, they simultaneously fit the joint models to incomplete data. Clarke and Tate (1998) compare the use of calibration weights developed based on adding tenure as a weighting variable to age and region with the model-based analysis developed by Clark and Chambers (1998). The authors conclude that there is a significant difference between unweighted, weighted and model-based estimation.

One of the problems associated with the estimation of gross flows is quarter-to-quarter or month-to-month nonresponse. This problem was discussed by Stasny (1986), Clarke and Chambers (1998), Clarke and Tate (1998) and Tate (1999). The authors assume that individuals' labour force flows behaviour, is independent within households, and that households are homogeneous with respect to their labour force flows behaviour. Similarly, individual conditional response probabilities, given the labour force flows data, are assumed homogeneous within and between households. As they point out these assumptions are clearly unrealistic. The aims of this article are as follows:

1. Extend the model of Clark and Chambers (1998) by specifying the labour force flows and nonresponse probabilities as regression models to accommodate individual level variables such as age, gender, and education, and household level variables such as region and tenure. One possible solution for this is to consider a logistic regression model for the labour force flow probabilities, possibly using random effects models.
2. Use the model-based approach to estimate the labour force gross flows by incorporating the model for the labour force flow frequencies or probabilities and the model for nonresponse. This is done on the basis of the relationships between the population likelihood and that of the respondent sample. This can improve the estimates, as shown by comparing the proposed estimates of labour force gross flows and their accuracy with those obtained by weighting (Clark and Tate, 1999, 2003) and to those obtained by simple modeling (Clark and Chambers, 1998).

In Section 2 a simple labour force flow model. In Section 3 we introduce the labour force flow model under informative nonresponse. In Section 4 we model the conditional expectation of response probabilities and derive the response sample likelihood function. In Section 5 we develop a new method of estimation of labour force gross flows. In Section 6 we specify the labour force flows and nonresponse probabilities as regression models. In section 7 we compare the unweighted, weighted, sample maximum likelihood and binomial logistic regression model approaches for estimating the labour force gross flows using a real data set previously analyzed by the Office for National Statistics. Finally Section 8 is devoted to the conclusions and future work.

2. A Simple Labour Force Flow Model

A gross flow is the probability or frequency of individuals in the population, making a state transition between two quarters, say $Q1$ and $Q2$ $(Q1 < Q2)$. Labour force gross flows refer to transitions between the three main labour force states: 1=employed, 2=unemployed and 3=not in labour force.

Let S denote the hypothetical complete sample of households, indexed by h, and assume that S is a simple random sample of households. Within household h, let n_h be the total number of eligible individuals, of which $n_h(a,b)$ have the labour force flow (a,b) between $Q1$ and $Q2$ where $\sum_{a,b} n_h(a,b) = n_h$ and $a,b = 1,2,3$.

Table 1 shows the complete labour force flows data for household h as a 3x3 contingency table. If household h responds at both quarters, the observed data are the cells of this two-way table. However, if household h does not response at quarters $Q1$ or $Q2$, the observed data correspond to the appropriate margin of the table: $n_h(a,+) = \sum_{b=1}^{3} n_h(a,b), a = 1,2,3$ are the observed data if household h responds at quarter $Q1$ but does not respond at quarter $Q2$; and $n_h(+,b) = \sum_{b=1}^{3} n_h(a,b), b = 1,2,3$ are the observed data if household h responds at quarter $Q2$ but does not respond at quarter $Q1$. Furthermore, if household h does not respond at either of the quarters $Q1$ and $Q2$, the observed data is the household size, n_h, which we take to be known and fixed between $Q1$ and $Q2$. For more discussion; see Clarke and Chambers (1998).

Table 1
Complete Labour Force Flow Data for Household h

	$Q2$				
$Q1$	Status	1	2	3	Total
	1	$n_h(1,1)$	$n_h(1,2)$	$n_h(1,3)$	$n_h(1,+)$
	2	$n_h(2,1)$	$n_h(2,2)$	$n_h(2,3)$	$n_h(2,+)$
	3	$n_h(3,1)$	$n_h(3,2)$	$n_h(3,3)$	$n_h(3,+)$

	Total	$n_h(+,1)$	$n_h(+,2)$	$n_h(+,3)$	n_h

Let $\mathbf{N}_h = [N_h(1,1),...,N_h(3,3)]$ be the vector random variable of labour force flows frequencies for household h, where $N_h(a,b)$ is the random variable whose outcome corresponds to the number of individuals, $n_h(a,b)$, with labour force flow (a,b), $a,b = 1,2,3$. The realization of this random vector is denoted by $\mathbf{n}_h = [n_h(1,1),...,n_h(3,3)]$.

Let $\omega(a,b) > 0$ be the probability of an individual having labour force flow (a,b), where $\sum_{a,b} \omega(a,b) = 1$. The vector of labour force flow probabilities is denoted by $\boldsymbol{\omega} = [\omega(1,1),...,\omega(3,3)]$, of which 8 are free.

The simple labour force flows model for $\mathbf{N}_h = \mathbf{n}_h$ is taken to be multinomial, with probability function:

$$\Pr(\mathbf{N}_h = \mathbf{n}_h | \boldsymbol{\omega}) = n_h! \prod_{a,b} \frac{(\omega(a,b))^{n_h(a,b)}}{n_h(a,b)!}. \quad (1)$$

Model (1) assumes that individuals' labour force flows behaviour, is independent within household, and that household are homogeneous with respect to their labour force flows behaviour. These assumptions are unrealistic, but equation (1) can be extended to a more realistic model for the labour force flows, as we discuss in Sections 5 and 6.

3. The Labour Force Flow Model under Informative Nonresponse

One of the problems associated with the estimation of labour force gross flows is quarter-to-quarter nonresponse. The problem of handling quarter-to-quarter nonresponse was studied by Clark and Chambers (1998) and Clark and Tate, (1999). The authors developed two methods for taking into account the problem of nonresponse - that of weighting and that of model-based adjustments for nonresponse bias by modeling the probability of a household nonresponse pattern as a weighted average of individual nonresponse probabilities and the household flow probabilities as multinomial, based on individual flow probabilities. By modeling both the labour force flow frequencies and the nonresponse, they simultaneously fit the joint models to incomplete data. A new model-based approach, for dealing with the problem of nonresponse, is introduced in the following, based on the concept of informative sampling, as developed by Pfeffermann, Krieger and Rinott (1998).

We assume that nonresponse is of whole households so that responses for all individuals are obtained if the household responds and none are obtained if the household fails to respond. This closely approximates the situation in most household labour force surveys. Denote the random vector for the nonresponse pattern of household h by $R_{hj} = 1$ if household h responds at quarter Q_j, and $R_{hj} = 0$, otherwise, where $j = 1,2$. The realization of this random quantity is denoted by $\mathbf{r}_h = (r_{h1}, r_{h2})$. Let $S_{uv} = \{h : \mathbf{r}_h = (u,v)\}; u,v = 0,1$, so that $S_{11} = \{h : \mathbf{r}_h = (1,1)\}$ denotes the subset of households with nonresponse pattern $(1,1)$, which is a subset of the complete sample

$S = S_{11} \cup S_{10} \cup S_{01} \cup S_{00}$. This subset, S_{11}, represents the longitudinal linked data on the same persons across two quarters.

The labour force flow model for household h with nonresponse pattern (u,v) is defined by:

$$P_{S_{uv}}\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega},\boldsymbol{\alpha}\right) = P_r\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega},\boldsymbol{\alpha}, h \in S_{uv}\right), \tag{2}$$

where $\boldsymbol{\alpha}$ is a vector of unknown parameters.

By application of Bayes theorem, (2) can be written as:

$$P_{S_{uv}}\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega},\boldsymbol{\alpha}\right) = \frac{\Pr\left(h \in S_{uv}|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right)}{\Pr\left(h \in S_{uv}|\boldsymbol{\omega},\boldsymbol{\alpha}\right)}\Pr\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega}\right). \tag{3}$$

Note that unless $\Pr\left(h \in S_{uv}|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right) = \Pr\left(h \in S_{uv}|\boldsymbol{\omega},\boldsymbol{\alpha}\right)$ for all $\mathbf{n}_h$, the labour force flows model and the respondent model are different, in which case the nonresponse mechanism is informative.

In particular, the labour force flow model for household h with nonresponse pattern $(1,1)$, that is for households who responded at both quarters is given by:

$$P_{S_{11}}\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega},\boldsymbol{\alpha}\right) = \frac{\Pr\left(h \in S_{11}|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right)}{\Pr\left(h \in S_{11}|\boldsymbol{\omega},\boldsymbol{\alpha}\right)}\Pr\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega}\right). \tag{4}$$

Model (3) incorporates the population (or complete sample) model for the labour force flow frequencies and the nonresponse effects as a function of the labour force flow. This is done by modeling the population distribution of labour force flow frequencies to take into account informative nonresponse. The modification is based on the ratio of the conditional probability of response given the true labour force flow to the unconditional probability. Nonresponse of this kind is an example of nonignorable (informative) nonresponse and its presence would imply that estimates of the important measures of labour force gross flows could be biased even after the application of a weighting process for adjustment of nonresponse.

In order to apply (4), it is necessary to model the probabilities, $\Pr\left(h \in S_{11}|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right)$. Let $\mathbf{z}_h$ be an observable auxiliary variable, considered as random, which is not included in the working model for the labour force flows.

Let $\phi_h = \Pr\left(h \in S_{11}|\mathbf{N}_h = \mathbf{n}_h,\mathbf{z}_h\right)$. Then according to Pfeffermann and Sverchkov (1999) the following relationships hold:

$$P_{S_{11}}\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega},\boldsymbol{\alpha}\right) = \frac{E\left(\phi_h|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right)}{E\left(\phi_h|\boldsymbol{\omega},\boldsymbol{\alpha}\right)}\Pr\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega}\right), \tag{5a}$$

$$E\left(\phi_h|\boldsymbol{\omega},\boldsymbol{\alpha}\right) = \sum_{a,b} E\left(\phi_h|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right)\Pr\left(\mathbf{N}_h = \mathbf{n}_h|\boldsymbol{\omega}\right) \tag{5b}$$

and

$$E\left(\phi_h|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right) = \frac{1}{E_{S_{11}}\left(\phi_h^{-1}|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right)} \text{ and } E\left(\phi_h|\boldsymbol{\omega},\boldsymbol{\alpha}\right) = \frac{1}{E_{S_{11}}\left(\phi_h^{-1}|\boldsymbol{\omega},\boldsymbol{\alpha}\right)}. \tag{5c}$$

It should be emphasized that, according to (5a), the labour force flows model of the linked data is completely determined by the specification of the conditional probabilities, $\Pr\left(h \in S_{11}|\mathbf{N}_h = \mathbf{n}_h,\boldsymbol{\alpha}\right)$. However the respondent probabilities themselves are not observable. In order to estimate the unknown parameters, we replace ϕ_h^{-1} by d_h - the longitudinal calibrated weights of household h, under the assumption

that they represent the reciprocals of the response probabilities. For more discussion of calibrated weights; see Clarke and Tate (1998). The question that arises here is how to compute the longitudinal weight of a household, when we only know the individuals' longitudinal weights. One way to overcome this problem is to take the average of the longitudinal weights of the persons in that household.

According to (5a), the labour force flow model of the linked data is completely determined by the specification of the conditional expectation $E\left(\phi_h|\mathbf{N}_h=\mathbf{n}_h,\boldsymbol{\alpha}\right)$, which can be identified and estimated based on the longitudinal weights and using the relationship (5c). This allows us to construct the respondent likelihood based on (5a) and apply standard inference procedures. In particular, since the parameters of the respondent model (5a) include the parameters of the labour force flow model, these can be estimated by applying maximum likelihood or other methods to the linked data, employing their respondent model.

4. Modeling the Conditional Expectation of Response Probabilities

The specification of the respondent model of the labour force flows given in (5a) depends on the identification of the conditional expectation $E\left(\phi_h|\mathbf{N}_h=\mathbf{n}_h,\boldsymbol{\alpha}\right)$. Let us consider the following approximations:

1. Exponential model:

$$E\left(\pi_h|\mathbf{N}_h=\mathbf{n}_h,\boldsymbol{\alpha}\right)=\exp\left(\boldsymbol{\alpha}'\mathbf{n}_h\right)=\exp\left(\alpha(0,0)n_h(0,0)+\sum_{a,b}\alpha(a,b)n_h(a,b)\right), \qquad (6a)$$

where $n_h(0,0)=1, \boldsymbol{\alpha}=\left(\alpha(0,0),\alpha(1,1),...,\alpha(3,3)\right)$.

Under this approximation and using (1) and (5b) we have:

$$\begin{aligned} E\left(\pi_h|\boldsymbol{\alpha},\boldsymbol{\omega}\right) &= E\left(E\left(\pi_h|\mathbf{N}_h,\boldsymbol{\alpha}\right)\right) \\ &= \sum_{a,b}\left(\left(\sum_{a,b}\exp\left(\alpha(0,0)+\alpha(a,b)n_h(a,b)\right)\right)*n_h!\prod_{a,b}\frac{\left(\omega(a,b)\right)^{n_h(a,b)}}{n_h(a,b)!}\right) \qquad (6b) \\ &= \exp\left(\alpha(0,0)\right)\left(\omega(1,1)\exp\left(\alpha(1,1)\right)+...+\omega(3,3)\exp\left(\alpha(3,3)\right)\right)^{n_h}. \end{aligned}$$

Now substituting (6) in (5a), we can show that:

$$P_{S_{11}}\left(\mathbf{N}_h=\mathbf{n}_h|\boldsymbol{\omega}\right)=n_h!\prod_{a,b}\frac{\left(\omega^*(a,b)\right)^{n_h(a,b)}}{n_h(a,b)!}, \qquad (7a)$$

where

$$\omega^*(a,b)=\frac{\omega(a,b)\exp\left(\alpha(a,b)\right)}{\sum_{a,b}\omega(a,b)\exp\left(\alpha(a,b)\right)}; a,b=1,2,3. \qquad (7b)$$

Thus the respondent and population labour force flow models follow a multinomial distribution but the transitions probability, $\omega(a,b)$, under the population model, changes, under the respondent model, to $\omega^*(a,b)$, defined in (7b).

Note that if $\alpha(a,b)=0$ for all (a,b), that is, the nonresponse mechanism is ignorable (noninformative) then the population and respondent labour force flow models are the same.

An alternative approximation of (4.2) is:

2. Linear model:

$$E\left(\pi_h \middle| \mathbf{N}_h = \mathbf{n}_h, \boldsymbol{\alpha}\right) = \boldsymbol{\alpha}'\mathbf{n}_h = \alpha(0,0)n_h(0,0) + \sum_{a,b}\alpha(a,b)n_h(a,b). \tag{8}$$

Under this approximation, we have:

$$E\left(\pi_h \middle| \mathbf{N}_h = \mathbf{n}_h, \boldsymbol{\alpha}\right) = \sum_{a,b}\left(\left(\sum_{a,b}\left(\alpha(0,0) + \alpha(a,b)n_h(a,b)\right)\right) * n_h! \prod_{a,b}\frac{\left(\omega(a,b)\right)^{n_h(a,b)}}{n_h(a,b)!}\right) = \alpha(0,0) + n_h\sum_{a,b}\alpha(a,b)\omega(a,b). \tag{9}$$

Now substituting (8) and (9) in (5a), we obtain:

$$P_{S_{11}}\left(\mathbf{N}_h = \mathbf{n}_h \middle| \boldsymbol{\omega}\right) = \frac{\alpha(0,0) + \sum_{a,b}\alpha(a,b)n_h(a,b)}{\alpha(0,0) + n_h\sum_{a,b}\alpha(a,b)\omega(a,b)} * n_h! \prod_{a,b}\frac{\left(\omega(a,b)\right)^{n_h(a,b)}}{n_h(a,b)!}. \tag{10}$$

Again, if $\alpha(a,b) = 0$ for all (a,b), that is, the nonresponse mechanism is ignorable (noninformative) then the population and respondent labour force flow models are the same.

5. Estimation

One of the main purposes of the labour force survey is the estimation of the labour force gross flows. These estimates are an important tool in the study of labour force dynamics. We consider three methods of estimation: the two-step method (see Pfeffermann, Krieger, and Rinott (1998)), which takes into account the informativeness of response, the unweighted method which does not account for response informativeness, and the weighted method which deals with nonresponse problem only via the calibration weights.

Two-step method:
Since the number of parameters indexing the respondent labour force flow model is large, (there are 10 informative parameters $\alpha(0,0), \alpha(a,b), a,b = 1,2,3$, and 8 transition probabilities $\omega(a,b), a,b = 1,2,3$ and $\sum_{a,b}\omega(a,b) = 1$), according to Pfeffermann, Krieger, and Rinott (1998) and because of problems of identifiability, under the exponential model, it is often computationally easier to estimate these parameters in two steps:

First-step: the coefficients $\{\alpha(0,0), \alpha(a,b)\}$ are estimated from the observed calibrated longitudinal weights $\{d_h\}$, employing the relationships (5c) and (7a) or (5c) and (10).

Second-step: the estimates of the informative parameters obtained in the first step are substituted in (7a) or (10), and then the parameters indexing the labour force flow model, $\{\omega(a,b)\}$, are estimated by the maximum likelihood procedure.

Estimation of $\{\alpha(0,0), \alpha(a,b), a,b = 1,2,3\}$ under the exponential model:
In this case, using the relationship (5c), we have:

$$E_{S_{11}}\left(d_h \middle| \mathbf{N}_h = \mathbf{n}_h, \boldsymbol{\alpha}\right) = \exp\left(-\boldsymbol{\alpha}'\mathbf{n}_h\right) = \exp\left(-\left(\alpha(0,0) + \sum_{a,b}\alpha(a,b)n_h(a,b)\right)\right). \quad (11)$$

The least squares estimate of $\boldsymbol{\alpha}$ can be obtained by using nonlinear regression where d is treated as the response variable and $n(a,b)$ as the explanatory variables. Alternatively we can use the approximation, $E(\ln(X)) \approx \ln E(X)$. Under this approximation we have:

$$E_{S_{11}}\left(\log(d_h) \middle| \mathbf{N}_h = \mathbf{n}_h, \boldsymbol{\alpha}\right) = \left(-\boldsymbol{\alpha}'\mathbf{n}_h\right) = -\alpha(0,0) - \sum_{a,b}\alpha(a,b)n_h(a,b). \quad (12)$$

Thus by regressing $-D_h = -\ln(d_h)$ on $\mathbf{n}_h = (1, n_h(1,1),..., n_h(3,3)), h = 1,..., m$, where m is the number of households in the linked sample data, S_{11}, the ordinary least squares estimate of $\boldsymbol{\alpha}$ is given by:

$$\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}(0,0), \tilde{\alpha}_1(1,1),..., \tilde{\alpha}(3,3)) = -(\mathbf{n}'\mathbf{n})^{-1}(\mathbf{n}'\mathbf{D}), \quad (13)$$

where:

$$\mathbf{n} = \begin{bmatrix} 1 & n_1(1,1) & ... & n_1(3,3) \\ ... & ... & ... & ... \\ 1 & n_m(1,1) & ... & n_m(3,3) \end{bmatrix} \text{ and } \mathbf{D} = (\ln(d_1),..., \ln(d_m)). \quad (14)$$

Having estimated the informativeness response parameters, the resulting sample log-likelihood of the multinomial probability density function (pdf in (7a) is given by:

$$l_r(\boldsymbol{\omega}) = \sum_{h=1}^{m}\left(\sum_{a,b} n_h(a,b)\ln(\tilde{\omega}^*(a,b))\right) + c, \quad (15)$$

where c is a constant that does not depend on $\tilde{\omega}^*(a,b)$, which is defined as:

$$\tilde{\omega}^*(a,b) = \frac{\omega(a,b)}{\sum_{a,b}\omega(a,b)\exp(\tilde{\alpha}(a,b))}, \text{ a, b} = 1,2,3.$$

Note that (6.15) can be written as:

$$l_r(\boldsymbol{\omega}) = n_+(1,1)\log\omega(1,1) + n_+(1,2)\log\omega(1,2) + ... + n_+(3,3)\log\omega(3,3) - n_+\log(\omega(1,1)\exp(\tilde{\alpha}(1,1)) + \omega(1,2)\exp(\tilde{\alpha}(1,2)) + ... + \omega(3,3)\exp(\tilde{\alpha}(3,3))), \quad (16)$$

where $n_+(a,b) = \sum_{h=1}^{m} n_h(a,b); a,b = 1,2,3$ and $n_+ = \sum_{h=1}^{n}\sum_{a,b} n_h(a,b)$.

Now the maximum likelihood estimates of $\omega(a,b)$ are the values which maximize $l_r(\boldsymbol{\omega})$ subject to the constraint $\sum_{a,b}\omega(a,b) = 1$.

To get these ML estimators $\omega(a,b)$ we differentiate (15) with respect to $\tilde{\omega}^*$ subject to the constraint $\sum_{a,b}\tilde{\omega}^*(a,b) = 1$ and setting $\frac{\partial l(\tilde{\omega}^*)}{\partial\tilde{\omega}^*} = 0$, we can show that the ML estimators of $\tilde{\omega}^*(a,b)$ are given by:

$$\hat{\tilde{\omega}}^{*}(a,b)=\frac{\hat{\omega}(a,b)\exp\tilde{\alpha}(a,b)}{\sum_{a,b}\hat{\omega}(a,b)\exp\tilde{\alpha}(a,b)}=\frac{\sum_{h=1}^{m}n_h(a,b)}{\sum_{h=1}^{m}n_h};\ a,b=1,2,3. \tag{17}$$

Thus we can show that the ML estimators of $\omega(a,b)$ are the solutions of the following system of equations:

$$\begin{aligned}&\left(\hat{\tilde{\omega}}^{*}(1,1)(\exp(\tilde{\alpha})(1,1)-\exp(\tilde{\alpha})(3,3))-\exp(\tilde{\alpha})(1,1)\right)\hat{\omega}(1,1)+\hat{\tilde{\omega}}^{*}(1,1)(\exp(\tilde{\alpha})(1,2)-\exp(\tilde{\alpha})(3,3))\\&*\hat{\omega}(1,2)+\hat{\tilde{\omega}}^{*}(1,1)(\exp(\tilde{\alpha})(3,2)-\exp(\tilde{\alpha})(3,3))\hat{\omega}(3,2)=-\hat{\tilde{\omega}}^{*}(1,1)\exp(\tilde{\alpha})(3,3)\\&\qquad\qquad\cdots\\&\hat{\tilde{\omega}}^{*}(3,2)(\exp(\tilde{\alpha})(1,1)-\exp(\tilde{\alpha})(3,3))\hat{\omega}(1,1)+\hat{\tilde{\omega}}^{*}(3,2)((\exp(\tilde{\alpha})(1,2)-\exp(\tilde{\alpha})(3,3))\hat{\omega}(1,2))\hat{\omega}(1,2)\\&+\left(\hat{\tilde{\omega}}^{*}(3,2)(\exp(\tilde{\alpha})(3,2)-\exp(\tilde{\alpha})(3,3))-\exp(\tilde{\alpha})(3,2)\right)\hat{\omega}(3,2)=-\hat{\tilde{\omega}}^{*}(3,2)\exp(\tilde{\alpha})(3,3),\end{aligned} \tag{18}$$

subject to $\sum_{a,b}\hat{\omega}(a,b)=1$.

Estimation of $\{\alpha(0,0),\alpha(a,b),a,b=1,2,3\}$ under the linear model:

In this case, using the relationship (5c), we have:

$$\begin{aligned}E_{S_{11}}\left(d_h|\mathbf{N}_h=\mathbf{n}_h,\boldsymbol{\alpha}\right)&=\left(\alpha(0,0)n_h(0,0)+\boldsymbol{\alpha}'\mathbf{n}_h\right)^{-1}\\&=\left(\alpha(0,0)n_h(0,0)+\sum_{a,b}\alpha(a,b)n_h(a,b)\right)^{-1}.\end{aligned} \tag{19}$$

The estimate of $\{\alpha(0,0),\alpha(a,b),a,b=1,2,3\}$ can be obtained by fitting a non-linear regression model. Alternatively we can use the approximation, $E\left(\frac{1}{X}\right)=\frac{1}{E(X)}$, and regress $D_h=\frac{1}{d_h}$ against the repressor variables in $\mathbf{n}_h$. So that in this case the OLS estimator of $\boldsymbol{\alpha}$ is similar to (14), that is: $\tilde{\boldsymbol{\alpha}}=(\tilde{\alpha}(0,0),\tilde{\alpha}_1(1,1),...,\tilde{\alpha}(3,3))=(\mathbf{n}'\mathbf{n})^{-1}(\mathbf{n}'\mathbf{D})$, where $\mathbf{D}=\left(\frac{1}{d_1},...,\frac{1}{d_m}\right)$.

Having estimated the informative parameters, they can be substituted in (10), so that the log-likelihood of the resulting sample probability density function (pdf) given in (10) is given by:

$$\begin{aligned}l_r(\boldsymbol{\omega})&=\sum_{h=1}^{m}\log\left(P_{S_{11}}\left(\mathbf{N}_h=\mathbf{n}_h|\boldsymbol{\omega}\right)\right)\\&=\sum_{a,b}n_+(a,b)\log(\omega(a,b))-\sum_{h=1}^{m}\log\left(\sum_{a,b}(\tilde{\alpha}(0,0)+\tilde{\alpha}(a,b)n_h(\omega(a,b)))\right)+c.\end{aligned} \tag{20}$$

This likelihood function can be written as:

$$l_r(\boldsymbol{\omega})=\sum_{a,b}n_+(a,b)\log(\omega(a,b))-m\log\tilde{\alpha}(0,0)-\log\sum_{h=1}^{m}\left(1+n_h(\omega(a,b))\frac{\tilde{\alpha}(a,b)}{\tilde{\alpha}(0,0)}\right) \tag{21}$$

where c is a constant that does not depend on $\omega(a,b)$.

Now the maximum likelihood estimates of $\omega(a,b)$ are the values, which maximize $l_r(\boldsymbol{\omega})$ subject to the constraint $\sum_{a,b}\omega(ab)=1$. In this case, closed forms of the estimates cannot be obtained, so that numerical optimization has to be used.

Unweighted method:

Here we assume the nonresponse mechanism is ignorable, so that the maximum likelihood estimates of the labour force flow transition probabilities, $\omega(a,b);a,b=1,2,3$ are the values, which maximize $l(\boldsymbol{\omega})$:

$$l(\boldsymbol{\omega})=n_{+}(1,1)\log\omega(1,1)+n_{+}(1,2)\log\omega(1,2)+...+n_{+}(3,3)\log\omega(3,3), \qquad (22)$$

subject to the constraint $\sum_{a,b}\omega(a,b)=1$.

We can show that the maximum likelihood estimates of $\omega(a,b)$ are given by:

$$\hat{\omega}(a,b)=\frac{\sum_{h}n_h(a,b)}{\sum_{h=1}^{m}n_h};\ a,b=1,2,3. \qquad (23)$$

Weighted method:

The weighted estimates of the labour force flows transition probabilities $\omega(a,b);a,b=1,2,3$ can be based on household weights or on individual weights:

1. Based on household level weights:

$$\hat{\omega}(a,b)=\frac{\sum_{h}d_h n_h(a,b)}{\sum_{h=1}^{m}d_h};\ a,b=1,2,3. \qquad (24)$$

2. Based on individual level weights:

$$\hat{\omega}(a,b)=\frac{\sum_{h=1}^{m}\sum_{i=1}^{n_h}d_{hi}n_{hi}(a,b)}{\sum_{h=1}^{m}\sum_{i=1}^{n_h}d_{hi}};\ a,b=1,2,3. \qquad (25)$$

where d_{hi} is the longitudinal weight for person i in household h and $n_{hi}(a,b)$ is the flow of person i in household h.

6. Multinomial Logistic Regression for the Labour Force Flows

In the previous sections we studied and discussed the labour force flows model under the assumption that individual labour force flows are homogenous and independent within household. Similarly, individual conditional response probabilities, given the flow, are assumed homogeneous within and between households. These assumptions are clearly unrealistic. In this section we extend the model of Clark and Chambers (1998) by specifying the labour force flows and nonresponse probabilities as regression models to accommodate individual level variables such as age, gender, and education, and household level variables such as region and tenure. One possible solution for this is to consider a logistic regression model for the labour force flow probabilities, possibly using random effects models.

Let $\mathbf{x}_{hi}$ be a vector of covariate information, at individual or household level, which is available for all units (respondent or nonrespondent), for example, tenure, age-sex and education. Let $\boldsymbol{\beta}^{(a,b)}$ be a vector of unknown coefficients including the constant term, and let $\omega_{hi}(a,b)$ denote the probability of individual i in household h, having labour force flow (a,b). The multinomial logistic regression is given by:

$$\log\left(\frac{\omega_{hi}(a,b)}{\omega_{hi}(1,1)}\right) = \boldsymbol{\beta}^{(a,b)}\mathbf{x}'_{hi}. \qquad (26)$$

To fit this model we do the following:

1. Estimate the regression coefficients, $\boldsymbol{\beta}^{(a,b)}$, on the basis of respondent households, that is based on S_{11}, by assuming that the nonresponse mechanism is ignorable or noninformative.
2. Estimate the individual flows probabilities, $\omega_{hi}(a,b)$, from the multinomial logistic model (26).

Having estimated the regression coefficients and the individual flows probabilities; we can derive the household flows probabilities from the individual flows probabilities as follows:

Let $\delta_{hi}(a,b)$ be the indicator variable, which takes the value 1 if and only if individual i in household h has flow (a,b). Thus:

$$\sum_{a,b}\delta_{hi}(a,b)=1;\ \sum_{i=1}^{n_h}\delta_{hi}(a,b)=n_h(a,b);\ \sum_{a,b}\sum_{i=1}^{n_h}\delta_{hi}(a,b)=\sum_{a,b}n_h(a,b)=n_h. \qquad (27)$$

Let $\boldsymbol{\delta}_h=\left(\delta_{h1}(1,1),\ldots,\delta_{h1}(3,3),\ldots,\delta_{hn}(1,1),\ldots,\delta_{hn_h}(3,3)\right)$ be the partition vector of all the indicator variables, $\delta_{hi}(a,b)$, for household *h*.

For given $\mathbf{n}_h=(n_h(1,1),\ldots,n_h(3,3))$, let Δ_h be the set of all the possible partitions, $\boldsymbol{\delta}_h$, such that $\sum_{a,b}\delta_{hi}(a,b)=1$ and $\sum_{i=1}^{n_h}\delta_{hi}(a,b)=n_h(a,b)$.

For example if $n_h=2$ and $\mathbf{n}_h=(1,1,0\ldots,0)$, i.e. $n_h(1,1)=n_h(1,2)=1$ (and all other values of $n_h(a,b)$ are zero), then Δ_h has two partitions: (1,0,0,0,0,0;0,1,0,0,0,0) and (0,1,0,0,0,0;1,0,0,0,0,0), i.e. either $n_{h1}(1,1)=n_{h2}(1,2)=1$ (and all other values are zero) or $n_{h1}(1,2)=n_{h2}(1,1)=1$ (and all other values are zero). Similarly if $\mathbf{n}_h=(2,0,0\ldots,0)$, then Δ_h has just one partition: (1,0,0,0,0,0;1,0,0,0,0,0).

The household flow probability can then be computed as follows:

$$\Pr(\mathbf{N}_h=\mathbf{n}_h|\boldsymbol{\omega})=\sum_{\boldsymbol{\delta}\in\Delta_h}\prod_{a,b}\prod_{i=1}^{n_h}(\omega_{hi}(a,b))^{\delta_{hi}(a,b)}. \qquad (28)$$

For example if $n_h=2$ and $\mathbf{n}_h=(1,1,0\ldots,0)$, then:

$$\Pr(\mathbf{N}_h=\mathbf{n}_h|\boldsymbol{\omega})=\omega_1(1,1)\omega_2(1,2)+\omega_1(1,2)\omega_2(1,1). \qquad (29a)$$

Similarly if $\mathbf{n}_h=(2,0,0\ldots,0)$, then:

$$\Pr(\mathbf{N}_h=\mathbf{n}_h|\boldsymbol{\omega})=\omega_1(1,1)\omega_2(1,2). \qquad (29b)$$

But in our case, for the labour force survey data, the $\mathbf{n}_h$ is known, so that the household flow probabilities can then be computed as follows:

$$\Pr\left(\mathbf{N}_h = \mathbf{n}_h \middle| \boldsymbol{\omega}\right) = \prod_{a,b}\prod_{i=1}^{n_h}\left(\omega_{hi}(a,b)\right)^{\delta_{hi}(a,b)}. \quad (30)$$

This can be computed only if $\omega_{hi}(a,b)$ are known – see equation (6.26).

For example if $n_h = 2$ and $\mathbf{n}_h = (1,1,0\ldots,0)$, then:

$$\Pr\left(\mathbf{N}_h = \mathbf{n}_h \middle| \boldsymbol{\omega}\right) = \omega_1(1,1)\omega_2(1,2). \quad (31)$$

Similarly if $\mathbf{n}_h = (2,0,0\ldots,0)$, then:

$$\Pr\left(\mathbf{N}_h = \mathbf{n}_h \middle| \boldsymbol{\omega}\right) = \omega_1(1,1)\omega_2(1,2). \quad (32)$$

Note that the (30) can be used only to compute the household flow probability.

7. Real Data

1. Creation of household weight:

In order to compare the unweighted, weighted, sample maximum likelihood and binomial logistic regression model approaches for estimating the LFS gross flows, as considered in the previous sections, we use a real data set previously analyzed by the Office for National Statistics (ONS). This data set is the 1995 Summer-Autumn LFS linked longitudinal data file. The data is on an individual level, so that in order to apply the proposed methods, which are household-based, we had to create and identify a household record based on the individual level data. The file contains individual calibration weights - see Clarke and Tate (1999) for details of the construction of these weights. For our purposes a household weight is required. On the basis of some explorative analysis (see below) we decided to use their mean as an approximate measure of the weight of the household.

In order to create a good measure for household weight we examined the individual longitudinal weights within household. The results are summarized in the following table, which contains the mean and the variance of the minimum, maximum and average of the individual longitudinal weights within household

Table 2

	Minimum	Maximum	Average
Mean	547.86	600.25	571.90
Standard Deviation	92.28	245.22	93.18

The histograms of the values of the minimum and maximum of the individual longitudinal weights within household shows that these measures produce distributions that are skewed to the right, while the histograms of the values of the mean of the individual longitudinal weights within household shows that these measures produce approximately bell shaped distributions. Based on this descriptive study we adopted the mean as an approximate measure of weight of the household.

2. Fitting models for the household weights:

2.1 Exponential model:

Here we fit the exponential model for the logs of the mean household weights, $d_h = \frac{1}{n_h}\sum_{i=1}^{n_h} d_{hi}$, as a response variable, where d_{hi} represents the longitudinal weight within household h for individual i and n_h is the number of individuals in household h. The explanatory variables are the labour force flow frequencies for

households which respond in the two quarters - $n_h(a,b)$ - the number of individuals who have labour force flow (a,b) between the two quarters.

The resulting fitted model is given by:

$$\begin{aligned}\log(d_h) = 6.3494 &- 0.0182n_h(1,1) + 0.075n_h(1,2) + 0.0422n_h(1,3) \\ &- 0.008n_h(2,1) + 0.0897n_h(2,2) + 0.0538n_h(2,3) \\ &- 0.0415n_h(3,1) + 0.0507n_h(3,2) + 0.0061n_h(3,3).\end{aligned} \tag{33}$$

The residual standard error is 0.1445 and the p-value of the F-statistic is very close to zero.

2.2 Linear model:

Here we fit the reciprocals of the household weights as a linear function of the labour force flow frequencies. In this case the fitted model is given by:

$$\begin{aligned}\frac{1}{d_h} = 0.00178 &+ 0.00003n_h(1,1) - 0.00013n_h(1,2) - 0.00008n_h(1,3) \\ &+ 0.00001n_h(2,1) - 0.00015n_h(2,2) - 0.00009n_h(2,3) \\ &+ 0.00007n_h(3,1) - 0.00090.0507n_h(3,2) - 0.000021n_h(3,3).\end{aligned} \tag{34}$$

The residual standard error is 0.00024 and the p-value of the F statistic is again very close to zero.

Some conclusions on the characteristics of the nonresponse can be obtained by examining the estimated regression coefficients from these fitted models. For instance, for the exponential model, the regression coefficients for individuals in a household whose labour force status is unemployed in the second quarter (b=2) are higher than for persons whose labour force status is employed (b=1) or not in labour force (b=3) in the second quarter, irrespective of their status in the first quarter. The converse holds for the linear model, since the dependent variable is the reciprocal of the weight. In both cases this implies that persons whose labour force status is unemployed in the second quarter have higher predicted weights than other persons and the highest predicted weights occurs for persons whose labour force status in both the two quarters is unemployed. We conclude that the least represented persons in the sample are the persons who were unemployed in both quarters. Conversely for persons whose labour force status at the second quarter is employed. This is similar to the result obtained by Clarke and Tate (1998) by comparing the estimates of gross flows under ignorable and the more sophisticated nonignorable models.

3. Gross flows estimates:

By maximizing the likelihood functions under the exponential and the linear models, estimates (or predictions) of the gross flows are obtained. These are compared with the classical maximum likelihood, binomial logistic regression and the weighted estimates, weighted at the individual and at the household levels. The estimates of labour force gross flows are shown in Table 3. Following are details on the methods of estimation used:

3.1 Unweighted method: The first column gives estimates from the unweighted data obtained maximizing equation (22).

3.2 Weighted methods:

3.2.1 At the individual level: The second column gives estimates from the weighted data at the individual level using equation (25).

3.2.2 At the household level: The third column gives estimates from the weighted data at the household level, computed using equation (24).

3.3 Sample likelihood method:

3.3.1 Exponential model:
The fourth column gives the estimates based on the sample log-likelihood under the exponential model. They are obtained by maximizing equation (16).

3.3.2 Linear model:
The fifth column gives the estimates based on the sample log-likelihood under the linear model. They are obtained by maximizing equation (21).

3.4 Binomial logistic regression:
The last column gives the estimates from the binomial logistic regression. They are computed as follows:

The binomial logistic regression model for labour force flow (a,b) is given by:

$$\log\left(\frac{\omega_h(a,b)}{1-\omega_h(a,b)}\right) = \beta_0 + \beta_1 n_h + \beta_2 d_h, h = 1,...,m, \quad (35)$$

where $\beta_0, \beta_1, \beta_2$ are unknown parameters, n_h is the household size, d_h is the mean household weight and $\omega_h(a,b)$ denote the probability of household h, having labour force flow (a,b).

To fit this model we do the following:

1. Estimate the regression coefficients, $\beta_0, \beta_1, \beta_2$, on the basis of respondent households, that is based on S_{11}, by assuming that the nonresponse mechanism is ignorable or noninformative.

2. Estimate the household flow probabilities, $\omega_h(a,b)$, from the logistic model (35) as:

$$\hat{\omega}_h(a,b) = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 n_h + \hat{\beta}_2 d_h\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 n_h + \hat{\beta}_2 d_h\right)}. \quad (36)$$

3. Estimate the overall labour force flow probability as:

$$\hat{\omega}(a,b) = \frac{1}{m}\sum_{h=1}^{m}\hat{\omega}_h(a,b), \quad (37)$$

where m is the number of households in S_{11}.

Table 3
Gross Flows Estimates (Percentages)

Status	Method of Estimation					
Flow	Unweighted	Weighted Individual	Weighted Household	SMLE-Exponential	SMLE-Linear	Logistic Regression
EE	70.62	70.51	69.32	69.78	69.80	68.69
EU	1.08	1.22	1.25	1.17	1.16	1.04
EN	1.53	1.77	1.71	1.61	1.61	1.39
UE	1.61	1.76	1.62	1.61	1.61	1.50
UU	3.78	3.90	4.41	4.16	4.15	4.22
UN	1.00	1.09	1.12	1.07	1.06	1.03
NE	1.40	1.52	1.29	1.35	1.35	1.25
NU	1.06	1.12	1.14	1.12	1.11	1.12
NN	17.92	17.12	18.12	18.13	18.15	19.76

The main findings from Table 3 are as follows:

1. The gross flows estimates based on the sample likelihood under the exponential and linear models (last two columns) give approximately the same results. Thus modelling

the calibrated weights at the household level as an exponential or a linear function of labour force frequencies does not affect the point estimates of gross flows.
2. There are small differences between unweighted and weighted gross flow estimates, both when using individual and household weights. The differences between estimates based on individual weights and the unweighted estimates are smaller than those between estimates based on household weights and unweighted estimates. This may be explained by the fact that nonresponse for the LFS is, in general, at the household level and not at the individual level, due to the high prevalence of use of proxy respondents.
3. There are only small differences between gross flow estimates based on household level weighting and those obtained based on sample likelihoods, under both the linear and the exponential models. The household level weighted estimates use the calibrated longitudinal weights, while the sample likelihood method uses the predicted weights based on modeling. Also the calibrated weights as constructed by the ONS are functions of auxiliary variables, like age, tenure, martial status and do not depend on the labour force frequencies. Thus these calibrated weights might be considered as ignorable because they depend only on auxiliary variables and do not depend on the labour force status. The fact that the differences between them are small implies that the estimates based on the sample likelihoods are basically just reconstructing the present weights (possibly with some smoothing) and may not reflect the full effects of informative nonresponse.
4. Both the household level weighting and sample likelihood procedures for estimating the labour force gross flows seem to reduce at least part of the nonresponse, compared to the unweighted method. Because based on simulation study by Clarke and Tate (2003), the authors recommended that the weighting could be used to produce flows estimates that offer a considerable improvement in bias over unadjusted estimates.
5. The labour force gross flows estimates under the binomial logistic regression differ quite considerably from the estimates based only on the weights. This implies that the use of additional covariate information may reduce nonresponse biases. Also the use of the predicted household weights based on the exponential and on the linear models, instead of the adjusted calibrated weights - d_h , produces the same results.

8. Conclusions and Future Work

In this article we introduce alternative methods of obtaining weighted estimates of gross flows, taking into account informative nonresponse. The first method based on extracting the response labour force model as a function of the population labour force model and of the response probabilities, which are obtained as reciprocals of the adjusted calibrated weights. The second method is based on the binomial logistic regression. The new methods are model based while the classical method is based on the adjusted weights. We think that the first method is more efficient than the weighted method. However the two methods, sample likelihood and weighting, give approximately the same estimates of labour force gross flows. Also we consider exponential and linear models to explain the variations in the calibrated weights under household level and from the exponential model we conclude that the unemployed persons at both quarters are under-represented in the labour force survey sample. This is similar to the result obtained by Clarke and Tate (1998) by comparing the estimates of gross flows under ignorable and nonignorable models.

Initially we considered that the estimates of gross flows based on the response labour force likelihood may explain the nonignorable nonresponse. We are not surprised at the similarity of the results of the weighted and response likelihood methods because the calibrated weights used in both methods are only a function of auxiliary variables and do not depend on the labour force status. The interesting result is that if we have sample data that contains the response variable and the sampling weights and for nonresponse the calibrated adjusted weights, then basing inference using classical weighted method and the new method based on the response likelihood may give similar results.

The alternative methods of estimating gross flows, based on modeling only adjusted calibrated weights give approximately the same estimates of labour force gross flows as those based only on the calibrated weights themselves. The addition of other covariates incorporated in the logistic model provides different values of the estimates, which may reflect better the effects of informative nonresponse. We propose to carry out additional experiments with alternative sets of covariates, in order to find better models (i.e., with better fit). However we still do not have well-founded evidence that these estimates have smaller bias or smaller mean square error, unless we assume that the model used is the correct one. No validation of the correctness of the model is possible without some additional information on the nonresponse. Further data on the cases that responded in at least one of the months for which they were sampled but did not respond in one or both of the periods compared has been requested. Based on this data, which we shall assume to represent also the total nonresponse, we propose to fit the Heckman model. This basically assumes that response is determined by the threshold of an unknown response variable, which is a linear function of known auxiliary variables. The regression residuals of the response variable are assumed to be correlated with those of the regression for the variable of interest, thereby modelling informative nonresponse. Estimation of the unknown model parameters is based on data for the auxiliary variables for both respondents and non-respondents and on the values of the variable of interest. The models then will provide both estimates of nonresponse probabilities as a function of the covariates and estimates of the labour force flows, taking into account informative nonresponse. These estimates will then be compared with those obtained from models without information on the nonresponse.

References

Clarke, P.S and Chambers, R.L. (1998). Estimating labour force gross flows from survey subject to household-level nonignorable nonresponse. *Survey Methodology*, 24, 123-129.

Clarke, P.S and Tate, P.F. (1999). *Methodological issues in the production and analysis of longitudinal data from the Labour Force Survey*. GSS Methodology Series No. 17. London: Office of National Statistics.

Clarke, P.S and Tate, P.F. (2003). Weighting versus modeling in adjusting for nonresponse in the British Labour Force Survey: an application to gross flows estimation. *Australian and New Zealand Journal of Statistics* 45. (To appear).

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47: 153-161.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya***,** 61, Ser. B, Pt. 1: 166-186.

Pfeffermann, D., Krieger, A.M, and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica* 8: 1087-1114.

Stasny, E.A. (1986). Estimation gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association* 81: 42-47.

Statistical Sciences, (1990). *S-Plus Reference Manual*, Seattle: Statistical Sciences.

Tate, P.F. (1999). Utilizing longitudinal linked data from the British Labour Force Survey. *Survey Methodology* 25: 99-103.