Survey Methodology, December 1998
Vol. 24, No. 2, pp. 123-129
Statistics Canada

Estimating Labour Force Gross Flows From Surveys Subject to Household-level Nonignorable Nonresponse

PAUL S. CLARKE and RAY L. CHAMBERS[1]

ABSTRACT

Measurement of gross flows in labour force status is an important objective of the continuing labour force surveys carried out by many national statistics agencies. However, it is well known that estimation of these flows can be complicated by nonresponse, measurement errors, sample rotation and complex design effects. Motivated by nonresponse patterns in household-based surveys, this paper focuses on estimation of labour force gross flows, while simultaneously adjusting for nonignorable nonresponse. Previous model-based approaches to gross flows estimation have assumed nonresponse to be an individual-level process. We propose a class of models that allow for nonignorable household-level nonresponse. A simulation study is used to show, that individual-level labour force gross flows estimates from household-based survey data, may be biased and that estimates using household-level models can offer a reduction in this bias.

KEY WORDS: Gross flows; Household-based surveys; Nonignorable nonresponse.

1. INTRODUCTION

Labour force gross flows are typically defined as transitions over time between the three major labour force states, employed, unemployed and economically inactive. Gross flows estimates are an important tool in the study of labour force dynamics (for example, see Vanski 1985). Large-scale on-going surveys such as the British Labour Force Survey and the U.S. Current Population Survey, provide data for gross flows estimation. However, nonresponse, measurement error, sample rotation and complex design effects, affect gross flows estimation from these surveys. A discussion of these and other factors affecting gross flows estimation, is given in Hogue (1985). Here we focus on the problem of nonresponse.

We assume that a nonresponse mechanism leads to the observed data being incomplete. If the probability of not responding depends on the missing data, then the nonresponse mechanism is nonignorable (Rubin 1976). The model-based approach to analysing incomplete survey data, is detailed in Little (1982). Model-based approaches to the estimation of labour force gross flows, involve modelling both the labour force flows and the nonresponse mechanism, and simultaneously fitting both models to the incomplete data. Examples of such models are given in Stasny and Fienberg (1985), Stasny (1986) and, for nonignorable nonresponse, in Little (1985). We call these individual-level models, because individuals are modelled as responding or not responding, independently of other sampled individuals.

Both the Labour Force Survey and the Current Population Survey, are examples of household-based surveys, that is, surveys based on a random sample of households, rather than individuals. Household-based surveys can lead to correlated nonresponse behaviour within households. For example, in the Current Population Survey, a single household member (usually the head-of-household) acts as a proxy for the other household members; thus, if the chosen household member is a nonrespondent, so are other household members. It follows that, due to correlated within-household nonresponse behaviour, individual-level nonresponse models are unsuitable for the estimation of labour force gross flows, using household-based survey data.

In this paper, we propose a class of models for individual-level labour force flows, and household-level nonresponse, that account for correlated within-household nonresponse behaviour. A number of plausible nonresponse models that are estimable from the observed data, both ignorable and nonignorable, are also presented. We then simulate household-based survey data, using these household-level models, to demonstrate the potential utility of our approach: first, individual-level labour force gross flows estimates are shown to be biased, when fitted to household-based survey data; and second, the bias of individual-level and household-level gross flows estimates are compared, to show the advantages of fitting household-level models to household-based survey data. To conclude, we summarise the findings of our simulation studies and discuss ideas for further research in this area.

2. A MODEL FOR HOUSEHOLD-LEVEL NONRESPONSE

2.1 The Data

A gross flow is the probability or frequency of individuals in the population, making a state transition between two points in time, t_1 and $t_2 (t_1 < t_2)$. Labour force gross flows refer to transitions between the three main

[1] Paul S. Clarke and Ray L. Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom.

labour force states: 1 = 'employed', 2 = 'unemployed' and 3 = 'not in labour force', where the last category refers to economically inactive individuals, such as retired individuals and students. Let S denote a simple random sample of households, indexed by h. Within household h, there are n_h eligible individuals, of which $n_h(ab)$ have labour force flow (a, b) between t_1 and t_2, where $\sum_{a,b} n_h(ab) = n_h$, and $a, b = 1, 2, 3$. We refer to $\{n_h(ab)\}$ as the complete data, that is, the frequencies that would be observed in the absence of nonresponse.

Table 1 shows the complete labour force flows data for household h as a 3 x 3 contingency table. If h responds at both times, the observed data are the cells of this 2-way table. However, if the household does not respond at t_1 or t_2, the observed data correspond to the margins of the table: $n_h(1+), n_h(2+), n_h(3+)$ are the observed data if h responds at t_1, but does not respond at t_2; and $n_h(+1)$, $n_h(+2), n_h(+3)$ are the observed data if h responds at t_2 but does not respond at t_1. (An index replaced by ' + ' denotes summation over all levels of that index.) Furthermore, if h does not respond at both t_1 and t_2, the observed data is the household size, n_h, which we take to be known and fixed between t_1 and t_2.

Table 1
Complete Labour Force Flows Data for Household h

Status		t_2			
		1	2	3	
	1	$n_h(11)$	$n_h(12)$	$n_h(13)$	$n_h(1+)$
t_1	2	$n_h(21)$	$n_h(22)$	$n_h(23)$	$n_h(2+)$
	3	$n_h(31)$	$n_h(32)$	$n_h(33)$	$n_h(3+)$
		$n_h(+1)$	$n_h(+2)$	$n_h(+3)$	n_h

2.2 Model Specification

It is inappropriate to treat the nonresponse behaviour of individuals within a household as independent, in household-based surveys. In the Labour Force Survey, for example, one eligible household member determines whether the household can be interviewed. Therefore, if no eligible individual can be contacted, each household individual is a nonrespondent. To construct a model for household-level nonresponse, we take the ideas behind individual-level nonresponse and extend them to the household, by considering a household to be an entity with its own nonresponse flow between t_1 and t_2. To allow for nonignorable nonresponse, the probability of a household nonresponse flow is modelled as a function of its individual labour force flows, as shall now be described.

Let $N_h = (N_h(11), N_h(21), ..., N_h(33))$ be the random vector of labour force flows frequencies for household h, where $N_h(ab)$ is the random variable, whose outcome corresponds to the number of individuals with labour force flow $(a, b), a, b = 1, 2, 3$. Further, denote the random vector for the nonresponse flow of household h by $\boldsymbol{R}_h = (R_{h1}, R_{h2})$, where

$$R_{hj} = \begin{cases} 1, & \text{if household responds at } t_j \\ 0, & \text{otherwise} \end{cases}$$

is the nonresponse status random variable for h at $t_j, j = 1, 2$. The realisations of these random quantities are denoted by $\boldsymbol{n}_h$ and $\boldsymbol{r}_h$. We now assume that $\boldsymbol{n}_h$ and $\boldsymbol{r}_h$ are known, and write the joint probability of $\boldsymbol{N}_h$ and $\boldsymbol{R}_h$ as

$$\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h, \boldsymbol{R}_h = \boldsymbol{r}_h) = \Pr(\boldsymbol{N}_h = \boldsymbol{n}_h) \Pr(\boldsymbol{R}_h = \boldsymbol{r}_h \mid \boldsymbol{N}_h = \boldsymbol{n}_h),$$

where $\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h)$ is the labour force flows model, and $\Pr(\boldsymbol{R}_h = \boldsymbol{r}_h \mid \boldsymbol{N}_h = \boldsymbol{n}_h)$ is called the nonresponse flows model.

The labour force flows model is taken to be multinomial, with probability function

$$\Pr(\boldsymbol{N}_h = \boldsymbol{n}_h; \boldsymbol{\omega}) = n_h! \prod_{a,b} \frac{\omega(ab)^{n_h(ab)}}{n_h(ab)!}, \qquad (1)$$

where $\omega(ab) > 0$ is the probability of an individual having labour force flow (a, b) and $\sum_{a,b} \omega(ab) = 1$. The vector of labour force flows parameters is denoted by $\boldsymbol{\omega} = (\omega(11), \omega(21), ..., \omega(33))$, of which 8 are free. The assumption of multinomial sampling in (1), implies that individuals' labour force flows behaviour, is independent within households, and that households are homogeneous with respect to their labour force flows behaviour. These assumptions are unrealistic, but (1) can easily be extended to a more realistic model for the labour force flows, as we discuss in Section 4.

The probability of household h having nonresponse flow (u, v), is taken to be

$$\begin{aligned} \pi(uv \mid \boldsymbol{n}_h) &= \Pr(\boldsymbol{R}_h = (u, v) \mid \boldsymbol{N}_h = \boldsymbol{n}_h; \boldsymbol{\psi}) \\ &= \frac{1}{n_h} \sum_{a,b} n_h(ab) \psi(uv \mid ab), \end{aligned} \qquad (2)$$

for $u, v = 0, 1$, namely, a weighted average of the nonresponse model parameters. By setting $n_h = 1$, it can be seen that $\psi(uv \mid ab) > 0$ is the probability of a household of size one (*i.e.*, an individual) having nonresponse flow (u, v), given it has labour force flow (a, b). Thus, $\sum_{u,v} \psi(uv \mid ab) = 1$ and $\boldsymbol{\psi} = (\psi(11 \mid 11), \psi(01 \mid 11), ..., \psi(00 \mid 33))$ is the vector of nonresponse parameters, of which 27 are free.

Before defining the likelihood function for the complete data, partition S into 4 mutually exclusive and exhaustive subsets

$$S = S_{11} \cup S_{01} \cup S_{10} \cup S_{00},$$

where $S_{uv} = \{h : \boldsymbol{r}_h = (u, v)\}$ is the subset of households with nonresponse flow (u, v). Thus, since S is a simple random sample of households, the likelihood function for the complete data is

$$L(\omega, \psi; \{\boldsymbol{n}_h, \boldsymbol{r}_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} L_h(\omega, \psi; \boldsymbol{n}_h, (u, v)), \quad (3)$$

where $L_h(\omega, \psi; \boldsymbol{n}_h, (u, v))$ is the contribution of household $h \in S_{uv}$ to the likelihood, the product of (1) and (2).

2.3 Model Fitting

2.3.1 Maximum Likelihood Estimation

Since the complete data are unavailable, (3) must be modified to give the likelihood based on the observed data. Denote the observed data by $\{\boldsymbol{n}_h^*\}$. As discussed in Section 2.1, the observed data for households that respond at t_1 and t_2, is the full cross-classification in Table 1, namely, $\boldsymbol{n}_h^* = \boldsymbol{n}_h$. Similarly, if $h \in S_{10}$ then $\boldsymbol{n}_h^* = (n_h(1+), n_h(2+), n_h(3+))$; if $h \in S_{01}$ then $\boldsymbol{n}_h^* = (n_h(+1), n_h(+2), n_h(+3))$; and if $h \in S_{00}$, then $\boldsymbol{n}_h^* = n_h$.

The contribution of household $h \in S_{uv}$ to the observed data likelihood, is obtained by summing $L_h(\omega, \psi; \boldsymbol{n}_h, (u, v))$ over all possible values that the full 3×3 cross-classification of labour force flows can take, given the observed margin. Representing this set of tables by $\boldsymbol{n}_h : \boldsymbol{n}_h^*$, the observed data likelihood for S is

$$L(\omega, \psi; \{\boldsymbol{n}_h^*, \boldsymbol{r}_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} \sum_{\boldsymbol{n}_h : \boldsymbol{n}_h^*} L_h(\omega, \psi; \boldsymbol{n}_h, (u, v)). \quad (4)$$

Model fitting requires calculating (4) at each stage of an iterative optimization process. This is computationally intensive, because the complete data likelihood function must be summed explicitly over the missing data. For example, the observed data for $h \in S_{10}$ is $\boldsymbol{n}_h^* = (n_h(1+), n_h(2+), n_h(3+))$ and the likelihood contribution of this household to the observed data likelihood is

$$\sum_{\boldsymbol{n}_h : \boldsymbol{n}_h^*} L_h(\omega, \psi; \boldsymbol{n}_h, (1, 0)).$$

To explicitly calculate this contribution, each 3×3 complete data table $\boldsymbol{n}_h$ for fixed $\boldsymbol{n}_h^*$ is generated and $L_h(\omega, \psi; \boldsymbol{n}_h, (1, 0))$ evaluated for each. For household size $n_h = 5$, there are at least 21 and at most 108 possible tables, depending on the values in the fixed margin; for $n_h = 15$, a very large household size, the respective numbers are 136 and 9,261. A similar procedure is used for $h \in S_{01}$, except here $\boldsymbol{n}_h^* = (n_h(+1), n_h(+2), n_h(+3))$ is the fixed margin. If $h \in S_{00}$, then no data about labour force status are observed, only the household size n_h. So each 3×3 table with total n_h must be generated, and the likelihood function calculated for each: for $n_h = 5$ there are 1,287 tables and for $n_h = 15$ there are 490,314. It is not infeasible, in terms of computer run-time, to calculate such sums directly. The number of explicit calculations can be reduced, by recognising that each household is defined only by its observed labour force flows frequencies and nonresponse flow. Thus, summation over the missing data need only be performed once for a household with a particular nonresponse flow and labour force flows frequencies; the contribution of this household to the likelihood is then raised to the power of the number of similarly defined households in S.

2.3.2 Parameter Estimability

If we fix $n_h = 1$ for all h, the complete data have no household structure, and form a 4-way table cross-classified by labour force status and nonresponse status at t_1 and t_2. The observed data log-likelihood (4) is now equivalent to that of the individual-level models in Stasny and Fienberg (1985), Little (1985) and Stasny (1986). For these models, estimability requires that the number of model parameters does not exceed 15 (one for each observed table cell, less one for the multinomial sampling constraint). Hence, (ω, ψ) are inestimable because there are $8 + 27 = 35$ free parameters. Since interest is focused on the labour force gross flows probabilities, ω, it is neccessary to constrain ψ to ensure estimability.

When $n_h > 1$, determining parameter estimability is more difficult, because (4) has a complicated closed-form expression. Fitzmaurice, Laird and Zahner (1996) use a numerical method to determine estimability, that involves showing that the information matrix is non-singular in the neighbourhood of the maximum likelihood estimate. However, not only is this impractical for problems of a high dimension, but evaluating the information matrix for the household-level model, is particularly difficult in this case. Instead, we adopt a pragmatic approach for determining parameter estimability: first, we restrict attention to models that satisfy the necessary condition for estimability when $n_h = 1$; and second, different starting values are used to for each fit. If the different starting values reveal a non-unique maximum likelihood estimate, or any parameter estimate is unchanged from its starting value then the model parameters are taken to be inestimable.

2.4 Nonresponse Models

To enable parameter estimates to be obtained from the observed data, $\boldsymbol{\theta}$ and ψ must be constrained in accordance with assumptions about the nonresponse mechanism. The nonresponse parameters are interpreted as individual nonresponse probabilities, but within the household framework established thus far, it is inappropriate to talk about individuals not responding. However, in reality, it is individuals within households that determine a household's nonresponse flow, not the household itself. Therefore, constraints are placed on the nonresponse parameters at the individual level, that apply at the household level through the functional dependence of $\pi(uv \mid \boldsymbol{n}_h)$ on ψ in (2). For example, if the nonresponse parameters are constrained such that $\psi(uv \mid ab) = \psi(uv)$ for all a, b, then the household nonresponse mechanism is ignorable, because household nonresponse flows are independent of the labour force flows.

We now present four models for the nonresponse mechanism, two of which are ignorable, and two nonignorable.

- Ignorable models.
  - Model I_A: Constant nonresponse probability,

    $$\psi(uv \mid ab) = \lambda^{1-u}(1-\lambda)^u \times \lambda^{1-v}(1-\lambda)^v,$$

    which has 1 parameter, λ, the probability of an individual not responding;
  - Model I_B: Independent of labour force status, but different nonresponse probabilities, at t_1 and t_2,

    $$\psi(uv \mid ab) = \lambda^{1-u}(1-\lambda)^u \times \theta^{1-v}(1-\theta)^v,$$

    which has 2 parameters, λ, θ, the probabilities of nonresponse at t_1 and t_2, respectively.
- Nonignorable models.
  - Model N_A: The nonresponse distributions at t_1 and t_2 are independent but depend on labour force status at t_1 and t_2, respectively,

    $$\psi(uv \mid ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \theta(b)^{1-v}(1-\theta(b))^v$$

    which has 6 parameters, $\boldsymbol{\lambda} = (\lambda(1), \lambda(2), \lambda(3))$ and $\boldsymbol{\theta} = (\theta(1), \theta(2), \theta(3))$, where $\lambda(a)$ is the probability of not responding at t_1, given labour force status a at t_1, and $\theta(b)$ that at t_2, given labour force status b at t_2;
  - Model N_B: The nonresponse distributions at t_1 and t_2 depend on labour force status at t_1 and t_2 respectively, *i.e.*, a first-order Markov process. Unlike N_A, the nonresponse distributions at t_1 and t_2 are dependent: if the nonresponse status at t_1 is 1, then the nonresponse distribution at t_2 is the same as at t_1; but if the nonresponse status at t_1 is 0, the nonresponse distributions are distinct,

    $$\psi(uv \mid ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \begin{cases} \lambda(b)^{1-v}(1-\lambda(b))^v, & \text{if} \quad u = 1, \\ \theta(b)^{1-v}\,(1-\theta(b))^v, & \text{if} \quad u = 0, \end{cases}$$

for $a, b = 1, 2, 3$ and $u, v = 0, 1$. Under model I_A, there are a total of $8 + 1 = 9$ free parameters, satisfying the necessary condition for estimability of an individual-level model. Models I_B, N_A and N_B have 10, 14 and 14 free parameters, respectively, and so also satisfy the necessary condition for estimability.

3. SIMULATION STUDY

3.1 Simulation Procedure

We used a simulation study to investigate the consequences of failing to account for the household structure of household-based survey data, and to compare labour force gross flows estimates for individual-level and household-level models. For this purpose, household-based survey data was generated using Monte Carlo sampling. Each sample data set consisted of 10,000 individuals arranged into households of size $n_h = k$ for all h. Within each household, labour force flows were generated from (1), and the nonresponse flow was generated from (2), under one of models N_A or N_B. The data were made incomplete by collapsing each complete labour force flows data table, to be consistent with the household nonresponse flow. In total, 1,000 independent data sets were generated in this way.

The population parameters used to generate the labour force flows are shown in the following table:

			b	
	$\omega(ab)$	1	2	3
a	1	0.43	0.245	0.035
	2	0.02	0.160	0.01
	3	0.015	0.035	0.05

This is clearly a population in recession, since the probability of moving from being employed to unemployed is very large ($\omega(12) = 0.245$). Under models N_A and N_B, the population parameters are

		i	
	1	2	3
$\lambda(i)$	0.2	0.8	0.5
$\theta(i)$	0.5	0.2	0.8

It should be noted that these parameter values do not represent realistic nonresponse flows behaviour, they were chosen for the purpose of illustrating this methodology. However, this does not affect the general conclusions of the paper, which are also relevant for realistic values of the true nonresponse probabilities.

3.2 Simulation Results

Estimates for individual-level models are obtained by fitting (4) with $n_h = 1$ to each incomplete data set. Figure 1 summarises the sampling distributions of the individual-level maximum likelihood estimate of $\omega(12)$, $\hat{\omega}(12)$, for nonresponse models I_A, I_B, N_A and N_B (estimates for ignorable models I_A and I_B are included together, because both yield the same estimates of the labour force flows). The vertical lines represent the intervals between the 2.5-percentile and the 97.5-percentile of each estimate's sampling distribution, and the bold point represents its median. There are three distributions obtained for each individual-level estimate: the left-most distribution is that when the household size is $k = 1$, *i.e.*, the simulated data have no household structure; and reading from left to right, the next two distributions are those obtained when the household size is $k = 2$ and $k = 5$, respectively. The solid horizontal line denotes the true flow probability, $\omega(12) = 0.0245$. The behaviour of the sampling distribution of $\hat{\omega}(12)$ in this study, reflects that of the other labour force gross flows estimates.

Figure 1a summarises the sampling distributions when N_A is the true model. If the fitted individual-level model is I_A, I_B or N_B, the labour force gross flows estimates have large biases, whatever the household size. As would be expected, the median estimate for correct model N_A, is unbiased if $k = 1$ and a small bias is apparent for $k = 2$ and $k = 5$ (although this bias is smaller for $k = 5$ than $k = 2$). Bias reduction with increasing k is also apparent for individual-level estimates I_A, I_B and N_B. This behaviour is unexpected, since it seems natural to expect the bias of the individual-level estimates, to increase with the household size. The results are slightly different in Figure 1b when N_B is true. Here the estimate for individual-level model N_B, becomes more biased as k increases, but the bias decreases for mis-specified individual-level models I_A, I_B and N_A. Furthermore, the misspecified estimates for I_A and I_B have a small bias, when compared to those for misspecified model N_A. These results are discussed in Section 3.3.

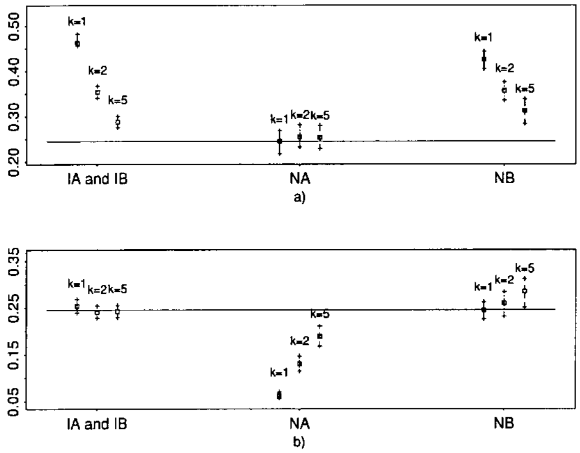

Figure 1. Sampling Distribution of $\hat{\omega}(12)$ for Individual-Level Models I_A, I_B, N_A and N_B When the True Nonresponse Model is a) N_A and b) N_B and the Household Size is $k = 1, 2, 5$.

A comparison of the median estimates of $\omega(12)$ for the fitted individual-level and household-level models when N_B is true, is presented in Figure 2. There are four sampling distributions associated with each model: the first two represent those from fitting an individual-level nonresponse model, and a household-level nonresponse model, when the household size is $k = 2$; and similarly, the next two distributions are those when the household size is 5.

For a particular pair of individual-level and household-level sampling distributions, it can be seen that the household-level estimate is less biased than its equivalent individual-level estimate, and the spread of each household-level sampling distribution, is narrower. The exception to this, is when fitting model I_A, where the household-level and individual-level distributions are identical. This equality occurs because the observed data likelihood for the individual-level and household-level models, are equivalent when the nonresponse model is ignorable. Another feature is that, if the nonresponse model is correctly specified, the household-level estimates are unbiased.

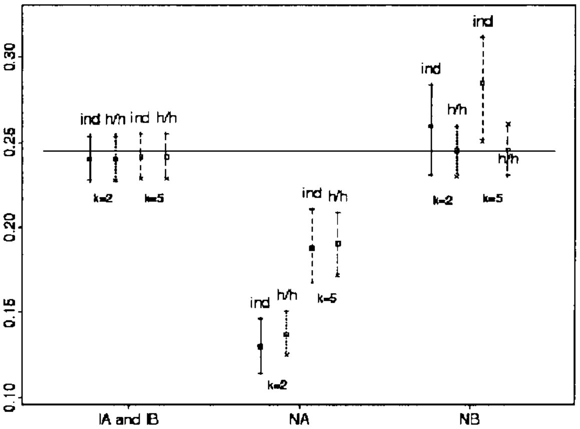

Figure 2. Sampling Distributions of $\hat{\omega}$ (12) for Individual-Level and Household-Level ModelsI_A, I_B, N_A and N_B When the True Nonresponse Model is N_B and the Household Size is $k = 2, 5$.

3.3 Summary

The estimates of the labour force gross flows under individual-level models, are never less biased than those of household-level models, when fitted to household-based survey data in our study. It should be noted, that if the true model is ignorable, it is unnecessary to utilise a household-level nonresponse model, because the individual-level and household-level models are equivalent. For example, if I_A is true, (2) reduces to $\lambda^{u+v}(1 - \lambda)^{1-u-v}$, and (4) factorizes into two components, dependent on ω only and λ only; the factor dependent on ω can be shown to be equivalent to that for the individual-level model, and thus the labour force flows estimates are the same.

It appears, as the household size increases, that the bias of the labour force flows estimates decreases, if the true model is nonignorable. In fact, this result arises because we use (1) to generate the labour force flows, and not because the model estimates are unbiased for large n_h. To see why, consider the household formation process, used to generate each Monte Carlo sample: as n_h increases, each household frequency tends to the same value, *i.e.*, $n_h(ab)$ converges to $n_h\omega(ab)$; hence,

$$\pi(uv \mid \boldsymbol{n}_h) \to \frac{1}{n_h}\sum_{a,b} n_h\omega(a,b)\psi\ (uv \mid ab)$$

$$= \sum_{a,b} \omega(ab)\psi(uv \mid ab),$$

which is independent of $\boldsymbol{n}_h$, that is, the simulated household nonresponse mechanism is ignorable. Therefore, the labour force flows estimates are unbiased, because fitting the nonignorable models to the simulated data, yields parameter estimates that are consistent with ignorable nonresponse. To generate nonignorable household-level nonresponse, it is necessary to prevent $n_h(ab) \to n_h\omega(ab)$, by extending (1), to allow for differential labour force flows between households. Such extensions to the labour force flows model are discussed in Section 4.

Figure 1b) shows two anomalous results that contradict the above explanation, when N_B is the true model. First, the bias of individual-level model N_B's estimate, increases as n_h increases. However, further simulations with household size $n_h = 10$, revealed that the individual-level estimate bias is zero. Thus, asymptotic ignorable nonresponse is also evident when N_B is true, but n_h must be large before its effect becomes apparent for individual-level model N_B. Second, the bias of the ignorable individual-level model estimates is small, almost zero, when N_B is true. This small bias reduces even further as n_h increases, in line with asymptotic ignorability, but we have yet to arrive at a satisfactory explanation as to why the ignorable models perform so well in this situation. Further study is necessary to investigate this finding.

4. DISCUSSION

In Sections 3 and 4, it is demonstrated by means of a simulation based study, that modelling household-level nonignorable nonresponse, when estimating labour force gross flows from household-based surveys, leads to reduced bias in the flows estimates, compared to those from individual-level models. If the nonresponse model is ignorable, it is unnecessary to use household-level models, because the individual-level and household-level models are equivalent. Furthermore, it is shown that controlling for household-level nonresponse does not necessarily remove all bias from the estimates of the labour force flows. Correct specification of the nonresponse model is still seen to be imperative, although taking the household structure of the data into account, may lead to a refinement of the flows estimates if the nonresponse model is misspecified. In particular, we show that household-level estimates are less biased than their equivalent individual-level estimates.

Our nonresponse model is an extension of the idea that nonresponse can depend upon the characteristics of a unit, in this case, the labour force flows of household members. Nonresponse in household-based surveys can occur for more than one reason, *e.g.*, refusal, non-contact, moving house or sample rotation. The current model can easily be extended to model more complex nonresponse patterns, by specifying the nonresponse indicator as a polytomous variable, and parameterizing the nonresponse model in accordance with the complex nonresponse patterns. It should also be noted, that we do not assume that the household-level model is an accurate representation of household nonresponse behaviour; rather, we assume that the household-level model, offers an approximation of within-household nonresponse dynamics.

An important problem, highlighted by the results from the simulation study, is our assumption that individual labour force flows behaviour is homogeneous within households. Clearly, this is an unrealistic assumption. The model is easily extended, by specifying the labour force flows and nonresponse flows probabilities, as regression models to accommodate individual-level, household-level, or higher level covariate information. For example, the labour force flows probabilities could be specified as a multinomial-logistic regression:

$$\log\left(\frac{\omega_{hi}(ab)}{\omega_{hi}(11)}\right) = \beta_0^{(ab)} + \boldsymbol{\beta}_1^{(ab)}\boldsymbol{x}_{hi}^T,$$

where $\omega_{hi}(ab)$ denotes the probability of individual i in household h, making labour force flow (a, b), $\boldsymbol{x}_{hi}$ is a (row) vector of covariates, and $(\beta_0^{(ab)}, \beta_1^{(ab)})$ are the regression coefficients for multinomial-logit (a, b). However, fitting these models requires conditional independence assumptions to be made, about the relationship between the distributions of the covariates, the labour force flows and the nonresponse flows, because the covariate information may be missing for nonresponding households. An alternative solution, is to allow for heterogeneous between household labour force flows, using random effects, by making assumptions about the distribution of between household differences. Fitting these models is also complicated and would require, for example, a Markov chain Monte Carlo procedure to perform the necessary integration. If S is not a simple random sample, auxiliary design variables can be incorporated into the fitting process, using the regression framework just described.

ACKNOWLEDGEMENTS

The work of Paul Clarke on this paper was funded by an Economic and Social Research Council studentship (award no. R00429614273); the work of Ray Chambers was funded by the contract between the Office for National

Statistics and the University of Southampton for the provision of research services in statistical methodology. Both authors would like to thank the referees, whose comments and practical advice helped make the final version of the manuscript considerably more readable.

REFERENCES

FITZMAURICE, G.M., LAIRD, N.M., and ZAHNER, G.E.P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91, 99-107.

HOGUE, C.R. (1985). History of the problems encountered in estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 1-8.

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistical Association*, 15, 1-15.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

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

STASNY, E.A., and FIENBERG, S.E. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 25-39.

VANSKI, J.E. (1985). Uses of gross change data in assessing demographic labor market dynamics. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 9-12.