Joint Treatment of Nonignorable Dropout and Informative Sampling for Longitudinal Survey Data

Abdulhakeem A.H. Eideh           Gad Nathan
Alquds University, Palestine      Hebrew University of Jerusalem, Israel
E-mail: msabdul@palnet.com        E-mail: gad@huji.ac.il

Abstract

In this paper, we study, within a modeling framework, the joint treatment of nonignorable dropout and informative sampling for longitudinal survey data, by specifying the probability distribution of the observed measurements when the sampling design is informative. The sample distribution of the observed measurements model is extracted from the population distribution model, assumed to be multivariate normal. The sample distribution is derived first by identifying and estimating the conditional expectations of first order sample inclusion probabilities, (assuming complete response at the first time period), given the study variable, based on a variety of models, such as linear, exponential, logit and probit. Next, we consider a logistic model for the informative dropout process. The proposed method combines two methodologies used in the analysis of sample surveys: for the treatment of informative sampling and informative dropout. One incorporates the dependence of the first order inclusion probabilities at the initial time period on the study variable, see Eideh and Nathan (2006), while the other incorporates the dependence of the probability of nonresponse on unobserved or missing observations, see Diggle and Kenward (1994). An empirical example based on British Labour Force Survey data illustrates the methods proposed.

1. Introduction

Data collected by sample surveys, and in particular by longitudinal surveys, are used extensively to make inferences on assumed population models. Often, survey design features (clustering, stratification, unequal probability selection, etc.) are ignored and the longitudinal sample data are then analyzed using classical methods based on simple random sampling. This approach can, however, lead to erroneous inference because of sample selection bias implied by informative sampling. To overcome the difficulties associated with the use of classical inference procedures for cross sectional survey data, Pfeffermann, Krieger and Rinott (1998) proposed the use of the sample distribution induced by the assumed population models, under informative sampling, and developed expressions for its calculation. Similarly, Eideh and Nathan (2006) fitted time series models for longitudinal survey data under informative sampling.

In addition to the effect of complex sample design, one of the major problems in the analysis of longitudinal data is that of missing values. In longitudinal surveys we intend
to take a predetermined sequence of measurement on a sample of units. Missing values occur when measurements are unavailable for one or more time points, either intermittently or from some point onwards (attrition).

The literature dealing with the treatment of longitudinal data considers three major areas of research:

(1) Analysis of complete non-survey longitudinal data (without nonresponse).


(2) Treatment of nonresponse in longitudinal data in the non-survey context.


(3) Treatment of effects of complex sample design and of nonresponse in longitudinal surveys.

Some recent work has considered the use of the sample distribution under informative sampling. Longitudinal survey data may be viewed as the outcome of two processes: the process that generates the values of units in the finite population, often referred as the superpopulation model, and the process of selecting the sample units from the finite population, known as the sample selection mechanism. Analytic inference from longitudinal survey data refers to the superpopulation model. When the sample selection
probabilities depend on the values of the model response variable at the first time period, even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process. Pfeffermann, Krieger and Rinott (1998) propose a general method of inference on the population distribution (model) under informative sampling that consists of approximating the parametric distribution of the sample measurements. The sample distribution is defined as the distribution of the sample measurements given the selected sample. Under informative sampling, this distribution is different from the corresponding population distribution, although for several examples the two distributions are shown to be in the same family and only differ in some or all the parameters. The authors discuss and illustrate a general approach to the approximation of the marginal sample distribution for a given population distributions and of the first order sample selection probabilities.


The joint treatment of the effects of complex sample design and of nonresponse in longitudinal surveys has been considered by several authors. Feder, Nathan and Pfeffermann (2000) develop models and methods of estimation for longitudinal analysis of hierarchically structured data, taking unequal sample selection probabilities into account. The main feature of the proposed approach is that the model is fitted at the individual level but contains common higher-level random effects that change stochastically over time. The model allows the prediction of higher and lower level random effects using data for all the time points with observations. This should enhance model-based inference from complex survey data. The authors introduced a two-stage procedure for estimation of the parameters of the model proposed. At the first stage, a separate two-level model is fitted for each time point, thus yielding estimators for the fixed effects and the variances. At the second stage, the time series likelihood is maximized only with respect to the time series model parameters. This two-stage procedure has the further advantage of permitting appropriate first and second level weighting to account for possible informative sampling effects. Pfeffermann and Nathan (2001) use time series structures with hierarchical modeling for imputation for wave nonresponse. Skinner and Holmes (2003) consider a model for longitudinal observations that consist of a permanent random effect at the individual level and autocorrelated transitory random effects corresponding to different waves of investigation. Lawless (2003) uses an event history approach for the analysis of longitudinal survey data.

Eideh and Nathan (2005) fit time series models for longitudinal survey data under informative sampling via the sample likelihood approach and pseudo maximum likelihood methods and introduce a new test of sampling ignorability based on the Kullback-Leibler information measure.
None of the above studies consider simultaneously the problem of informative sampling and the problem of informative dropout when analyzing longitudinal survey data. In this paper, we study, within a modeling framework, the joint treatment of nonignorable dropout and informative sampling for longitudinal survey data, by specifying the probability distribution of the observed measurements when the sampling design is informative. This is the most general situation in longitudinal surveys and other combinations of sampling informativeness and response informativeness can be considered as special cases. The sample distribution of the observed measurements model is extracted from the population distribution model, such as the multivariate normal distribution. The sample distribution is derived first by identifying and estimating the conditional expectations of first order (complete response) sample inclusion probabilities, given the study variable, based on a variety of models, such as linear, exponential, logit and probit. Next, we consider a logistic model for the informative dropout process. The proposed method combines two methodologies used in the analysis of sample surveys for the treatment of informative sampling and informative dropout. One incorporates the dependence of the first order inclusion probabilities on the study variable, see Pfeffermann, Krieger and Rinott (1998), while the other incorporates the dependence of the probability of nonresponse on unobserved or missing observations, see Diggle and Kenward (1994). This is possible in longitudinal surveys by using models based on observations in previous rounds.

The main purpose here is to consider how to account for the joint effects of informative sampling designs and of informative dropout in fitting general linear models for longitudinal survey data with correlated errors.

### 2. Population Model

Let $y_{it}$, $i = 1, \ldots, N; t = 1, \ldots, T$ be the measurement on the $ith$ subject at time $t = 1, \ldots, T$. Associated with each $y_{it}$ are the (known) values $x_{ik}, k = 1, \ldots, p$, of $p$ explanatory variables. We assume that the $y_{it}$ follow the regression model:

$$y_{it} = \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_{it}$$  \hspace{1cm} (1)

where the $\epsilon_{it}$ are a random sequence of length $T$ associated with each of the $N$ subjects. In our context, the longitudinal structure of the data means that we expect the $\epsilon_{it}$ to be correlated within subjects.

Let $y_i = (y_{i1}, \ldots, y_{iT})'$ be the column vector of size $T$ representing the measurements of subject $i$ over the $T$ time periods, and let $y = (y_1, \ldots, y_N)'$ be the complete set of $M = NT$ values from the $N$ subjects. The general linear model for longitudinal survey data treats $y$ as a multivariate normal random vector, that is:

$$y \sim MVN(x\beta, V^*)$$

$$V^* = \sigma^2 V_0$$  \hspace{1cm} (2)
where \( \mathbf{x} \) is the \( NT \times p \) matrix of explanatory variables, \( \mathbf{\beta} \) is a column vector of \( p \) unknown parameters, and \( \mathbf{V}^* = \sigma^2 \mathbf{V}_0 \) is a block-diagonal matrix with nonzero \( T \times T \) blocks \( \mathbf{V}^* = \sigma^2 \mathbf{V}_0 \), each representing the variance matrix for the vector of \( T \) measurements on a single unit; see Diggle, Liang and Zeger (1994).

Under the model defined by (2) and using the properties of multivariate normal distributions, we can conclude that: the random vectors \( \mathbf{y}_i, i = 1,...,N \) have independent multivariate normal distributions, that is

\[
\mathbf{y}_i \sim \text{MVN}(\mathbf{x}_i \mathbf{\beta}, \mathbf{V}^*)
\]

where \( \mathbf{x}_i \) is the matrix of size \( T \) by \( p \) of explanatory variables for subject \( i \), and \( \mathbf{V}^* \) has \( (jk)h \) element, \( v_{jk} = \text{cov}_p (y_{ij}, y_{ik}), j, k = 1,...,T \).

3. Sampling Design and Sample Distribution

We assume a single-stage informative sampling design, where the sample is a panel sample selected at time \( t = 1 \) and all units remain in the sample till time \( t = T \). Examples of longitudinal surveys, some of which are based on complex sample designs, and of the issues involved in their design and analysis can be found in Friedlander et. al. (2002), Herriot and Kasprzyk (1984), and Nathan (1999). In many of the cases described in these papers, a sample is selected for the first round and continues to serve for several rounds). Then it is intuitively reasonable to assume that the first order inclusion probabilities, \( \pi_i \), depend on the population values of the response variable at the first occasion only, the values \( y_{i1} \), and the values of known design variable, \( \mathbf{z} = \{z_1,...,z_N\} \), used for the sample selection, but not included in the working model under consideration. Then we have the following results.

**Theorem 1.**

Let \( \mathbf{y}_i \sim f_p (\mathbf{y}_i | \mathbf{x}_i, \mathbf{\theta}) \) be the population distribution of \( \mathbf{y}_i \) given \( \mathbf{x}_i \). If we assume that \( \pi_i \) depends only on \( y_{i1} \), on \( \mathbf{x}_i = (x_{i1},...,x_{ip})' \) and on \( \mathbf{z}_i \), then the sample distribution of \( \mathbf{y}_i \) given \( \mathbf{X}_i \) is given by:

\[
f_s (\mathbf{y}_i | \mathbf{x}_i, \mathbf{\theta}, \gamma) = f_s (\mathbf{y}_{i1} | \mathbf{x}_{i1}, \mathbf{\theta}, \gamma) f_p (y_{i2}, y_{i3},...,y_{iT} | y_{i1}, \mathbf{x}_i; \mathbf{\theta})
\]

where

\[
f_s (\mathbf{y}_{i1} | \mathbf{x}_{i1}, \mathbf{\theta}, \gamma) = \frac{E_p (\pi_i | y_{i1}, \mathbf{x}_{i1}; \gamma)}{E_p (\pi_i | \mathbf{x}_{i1}, \mathbf{\theta}, \gamma)} f_p (y_{i1} | \mathbf{x}_{i1}, \mathbf{\theta})
\]

is the sample distribution of \( y_{i1} \) given \( \mathbf{x}_{i1} \) and
\[
E_p(\pi_i \mid x_i, 0, \gamma) = \int E_p(\pi_i \mid y_{ii}, x_{ii}, \gamma)f_p(y_{ii} \mid x_{ii}, 0)dy_{ii}
\]

Also \( \gamma \) is the parameter indexing \( E_p(\pi_i \mid y_{ii}, x_{ii}; \gamma) \).

**Proof:** See Eideh and Nathan (2005).

**Comment 1:** Note that \( E_p(\pi_i \mid y_i) = E_{z_i}E_p(\pi_i \mid y_i, z_i) \), so that \( z_i \) is integrated out in (5).

**Comment 2:** Note that the sample and the population distribution of longitudinal observations are, in general different. In particular in (4) the difference between the population and sample distributions is due to the difference in the distribution of \( y_{ii} \) only. The second product term in (4) does not change. This is because the selected sample is a panel sample selected at time \( t = 1 \) and all units remain in the sample till time \( t = T \).

**Comment 3:** According to (4), we can see that, for a given population distribution, the sample distribution is completely determined by the specification of the population conditional expectations of sample inclusion probabilities, \( E_p(\pi_i \mid y_{ii}, x_{ii}; \gamma) \).

Assuming independence of the population measurements, Pfeffermann *et al.* (1998) establish an asymptotic independence of the sample measurements with respect to the sample distribution under commonly used sampling schemes for selection with unequal probabilities. Thus the use of the sample distribution permits the use of standard efficient inference procedures as likelihood based inference.

Let \( E_p \) and \( E_s \) denote expectations under the population and sample pdf’s, respectively, and let \( w_i = \pi_i^{-1} \) define the sampling weight of unit \( i \). Pfeffermann and Sverchkov (1999) show the following relationships:

\[
E_p(y_i \mid x_i) = \frac{E_s(w_i y_i \mid x_i)}{E_s(w_i \mid x_i)}, \quad E_p(\pi_i \mid x_i) = \frac{1}{E_s(w_i \mid x_i)} \quad E_p(\pi_i) = \frac{1}{E_s(w_i)} \quad \text{and} \quad E_s(y_i) = \frac{E_s(w_i y_i)}{E_s(w_i)}
\]

Consider the following approximation models for the population conditional expectation; \( E_p(\pi_i \mid y_{ii}, x_{ii}; \gamma) \), Pfeffermann, Krieger, and Rinott (1998), and Skinner (1994).

(a) **Exponential Inclusion Probability Model:**

\[
E_p(\pi_i \mid y_{ii}, x_{ii}) = \exp(a_0 + a_0 y_{ii} + a_1 x_{i11} + a_2 x_{i12} + \ldots + a_p x_{i1p})
\]

In this case, we can show that the sample distribution of \( y_i \) given \( x_i \) is given by:
\[
f_{s}(y_i | x_i, \theta^*) = \frac{\exp(a_0 y_{i1}) f_p(y_i | x_i, \theta^*)}{M_p(a_0, \theta)} f_p(y_{i2}, y_{i3}, \ldots, y_{it} | y_{i1}, x_i, \theta^*), \tag{8}
\]

where \( \theta^* = (0, a_0), \ a_0^*, a_0, a_1, \ldots, a_p \) are informativeness parameters to be estimated from the sample and \( M_p(a_0, \theta) = E_p[\exp(a_0 y_{i1})] \) is the moment generating function of the population distribution of \( y_{i1} \) given \( x_{i1} \).

(b) Linear Inclusion Probability Model:

\[
E_p(\pi_i | y_{i1}, x_i) = b_0^* + b_{01} y_{i1} + b_{11} x_{i11} + b_{21} x_{i12} + \ldots + b_{p1} x_{i1p} \tag{9}
\]

In this case the sample distribution of \( y_i \) is given by:

\[
f_i(y_i | x_i, \theta^*) = \frac{(b_0^* + b_{01} y_{i1} + b_{11} x_{i11} + b_{21} x_{i12} + \ldots + b_{p1} x_{i1p}) f_p(y_i | x_i, \theta^*)}{(b_0^* + b_{01} E_p(y_{i1}) + b_{11} x_{i11} + b_{21} x_{i12} + \ldots + b_{p1} x_{i1p})} \tag{10}
\]

where \( \theta^* = (0, b_0^*, b_{01}, b_{11}, b_{21}, \ldots, b_{p1}) \), and \( b_0^*, b_{01}, b_{11}, b_{21}, \ldots, b_{p1} \) are informativeness parameters to be estimated from the sample.

Eideh (2003) and Eideh and Nathan (2003) consider additional approximations for the population conditional expectation, namely, the logit and probit models:

(c) Logit Model:

\[
E_p(\pi_i | y_{i1}, x_i) = \frac{\exp(c_{01} + c_{01} y_{i1} + c_{11} x_{i11} + c_{21} x_{i12} + \ldots + c_{p1} x_{i1p})}{1 + \exp(c_{01} + c_{01} y_{i1} + c_{11} x_{i11} + c_{21} x_{i12} + \ldots + c_{p1} x_{i1p})} \tag{11}
\]

Under this approximation, the sample pdf of \( y_i \) is:

\[
f_i(y_i | x_i, \theta^*) = \frac{\exp(c_{01} + c_{01} y_{i1})}{(1 + \exp(c_{01} + c_{01} y_{i1})) E_p(\exp(c_{01} + c_{01} y_{i1}))/\exp(c_{01} + c_{01} y_{i1}))} \tag{12}
\]

where \( \theta^* = (0, c_{01}^*, c_{01}, c_{11}, \ldots, c_{p1}) \), and \( c_{01}, c_{01}, c_{11}, \ldots, c_{p1} \) are informativeness parameters to be estimated from the sample.

Note that an explicit expression for

\[
E_p\left[ \frac{\exp(c_{01} + c_{01} y_{i1} + c_{11} x_{i11} + c_{21} x_{i12} + \ldots + c_{p1} x_{i1p})}{[1 + \exp(c_{01} + c_{01} y_{i1} + c_{11} x_{i11} + c_{21} x_{i12} + \ldots + c_{p1} x_{i1p})]} \right] \tag{13}
\]

cannot be obtained, but Taylor series approximation can be used. See Eideh (2002) and Nathan and Eideh (2003).

(d) Probit Model:

\[
E_p(\pi_i | y_{i1}, x_i) = \Phi(d_{01}^* + d_{01} y_{i1} + d_{11} x_{i11} + d_{21} x_{i12} + \ldots + d_{p1} x_{i1p}) \tag{14}
\]
where \( \Phi \) denotes the cumulative distribution function of standard normal distribution.

Under this approximation and similar to (c) we can show that, the sample pdf of \( y_i \) is:

\[
 f_s(y|x, \theta^*) = \frac{\Phi(d_0^* + d_0y_{ij} + d_1x_{i1} + d_2x_{i2} + \ldots + d_px_{ip})}{\Phi(d_0^* + d_0y_{ij} + d_1x_{i1} + d_2x_{i2} + \ldots + d_px_{ip})} * \left( \frac{f_p(y_{i1}|x_{i1}, \theta)f_p(y_{i2}, y_{i3}, \ldots, y_{iT}|y_{i1}, x_i, \theta)}{f_p(y_{i2}, y_{i3}, \ldots, y_{iT})} \right)
\]

where \( \theta^* = (\theta, d_0^*, d_0, d_1, \ldots, d_p) \), and \( d_0^*, d_0, d_1, \ldots, d_p \) are informativeness parameters to be estimated from the sample.

Note that no explicit expression for \( E_p[\Phi(d_0 + d_0^*y_{ij} + d_1x_{i1} + d_2x_{i2} + \ldots + d_px_{ip})] \) is available, but Taylor series approximation can be used. See Eideh (2002) and Nathan and Eideh (2003).

Note that if \( y_{ij} \) is regarded as fixed in longitudinal surveys, then the population and sample conditional pdf's of \((y_{i2}, \ldots, y_{iT})|y_{i1}\) coincide. Hence there is no problem of informativeness.

**Theorem 2.**

We assume the multivariate normal population distribution of \( y_i \) given \( x_i \), defined by equation (3). Let \( y_{i,T-1} = (y_{i2}, y_{i3}, \ldots, y_{iT})' \).

(a) Under the exponential inclusion probability model given by equation (7), we have:

\[
 y_{ij} | x_{i1}, \theta^* \sim \mathcal{N}(x_i'\beta + a_0v_{ij1}^*, v_{ij1}^*), \quad \theta^* = (\theta, a_0) = (\beta, \sigma^2, v_{jk}, a_0), \quad j, k = 1, \ldots, T
\]

and

\[
 y_{i2}, y_{i3}, \ldots, y_{iT} | y_{i1}, x_i; \theta^* \sim \mathcal{N}(\mu_{i,T-1}, V_{0,T-1}^*)
\]

where the mean vector is:

\[
 \mu_{i,T-1} = E_p[y_{i,T-1} | y_{i1}, x_i] = \left[ x_{i2}'\beta + \frac{v_{11}^*}{v_{11}} (y_{i1} - x_{i1}'\beta), \ldots, x_{iT}'\beta + \frac{v_{T1}^*}{v_{11}} (y_{i1} - x_{i1}'\beta) \right]
\]

and the covariance matrix is:

\[
 Cov_p\left[(y_{i2}, y_{i3}, \ldots, y_{iT})' | y_{i1}, x_i\right] = Cov_p[y_{i,T-1} | y_{i1}, x_i] = V_{0,T-1}^*
\]

with general term: \( v_{tt'}^* = v_{tt'+1}^* - \frac{v_{tt'+1}^* v_{tt+1}^*}{v_{11}^*} \); \( t, t' = 1, \ldots, T - 1 \)

That is,
Under the linear inclusion probability model defined by equation (9), we have:

$$f_s(y_i | x_i, \theta, \alpha_0) = \frac{1}{\sqrt{2\pi \nu_{i1}^*}} \exp \left[ -\frac{1}{2\nu_{i1}^*} (y_i - x_i' \beta - \alpha_0 \nu_{i1}^*)^2 \right] \cdot \frac{1}{(2\pi)^{r-1}|\nu_{0,r-1}^*|^{1/2}} \exp \left[ -\frac{1}{2} (y_{i,r-1} - \mu_{r-1})'(\nu_{0,r-1}^*)^{-1}(y_{i,r-1} - \mu_{r-1}) \right]$$

(b) Under the linear inclusion probability model defined by equation (9), we have:

$$f_s(y_i | x_i, \theta, b_0^*, b_0, b_1, ..., b_p) = \frac{\left( b_0^* + b_0 y_i + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{i,p} \right) f_p(y_i | \theta)}{\left( b_0^* + b_0 (x_i' \beta) + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{i,p} \right)} \cdot \frac{1}{\sqrt{2\pi \nu_{i1}^*}} \exp \left[ -\frac{1}{2\nu_{i1}^*} (y_i - x_i' \beta)^2 \right] \cdot \frac{1}{(2\pi)^{r-1}|\nu_{0,r-1}^*|^{1/2}} \exp \left[ -\frac{1}{2} (y_{i,r-1} - \mu_{r-1})'(\nu_{0,r-1}^*)^{-1}(y_{i,r-1} - \mu_{r-1}) \right]$$

Similar results are obtained for the logit and probit inclusion probability models.

4. Covariance Structure of $y_i$

It is useful at this stage to consider what form the nonzero blocks $V_{ij}$ might take. We consider three cases: the exponential correlation model, the uniform correlation model; see Diggle, Liang and Zeger (1994), and the random effect model; see Skinner and Holmes (2003).

Case1: The exponential correlation model

In this model, the $(t,s)-\text{th}$ element of $V_0^*$ has the form:

$$v_{ts} = \text{cov}_s(y_t, y_s) = \sigma^2 \rho^{t-s}, t, s = 1, ..., T$$

(20)

Note that the correlation $v_{ts}/\sigma^2$, between a pair of measurements on the same unit decays toward zero as the time separation between the measurements increases.

A justification of the exponential correlation model is the following. Let

$$y_t = \mu_t + w_t$$

$$w_t = \rho w_{t-1} + z_t, i = 1, ..., N; t = 1, ..., T$$

where $\mu_t = E_p(y_t)$, the $z_t \sim N(0, \sigma^2(1 - \rho^2))$ are mutually independent random variables, so that $\text{var}_p(w_t) = \sigma^2$.

Case2: The uniform correlation model
In this model, we assume that there is a positive correlation, $\rho$, between any two measurements on the same subject. So that the $(t,s)$-th element of $V_0^*$ has the form:

$$v_{ts} = \text{cov}_p(v_{ts}, v_{st}) = \sigma^2 \rho, t \neq s = 1,...,T ; \quad v_{tt} = \sigma^2, t = 1,...,T$$

(22)

A justification of (22) is the following. Let

$$y_{it} = \mu_{it} + u_{it} + z_{it}, i = 1,...,N; t = 1,...,T$$

(23)

where $\mu_{it} = E_p(y_{it})$, the $u_{it} \sim N(0,\tau^2)$ are mutually independent random variables, the $z_{it} \sim N(0,\sigma^2)$ are mutually independent random variables, and the $u_{it}$ and $z_{it}$ are independent of one another. Then, the covariance structure of the data corresponds to (22) with $\rho = \tau^2/\nu^2 + \tau^2$ and $\sigma^2 = \nu^2 + \tau^2$.

Case 3. Random Effects Models

Random effects models have a number of important uses in the analysis of longitudinal survey data, consisting of repeated measurements on the same variables. Skinner and Holmes (2003) consider a model for longitudinal observations that consists of a ‘permanent’ random effect at the subject level and autocorrelated transitory random effects corresponding to different waves of investigation. The authors study two broad approaches to fitting these models under a complex sample design and ‘noninformative’ attritions.

In this subsection our aim is to consider how to take account of informative sampling in the estimation of the unknown model parameters, using the sample distribution approach.

We adopt the model given by Skinner and Holmes (2003) such that:

$$y_{it} = \beta_i + u_{it} + v_{it}$$

(24)

$$v_{it} = \rho v_{i(t-1)} + \epsilon_{it}$$

(25)

where the unknown parameter $\beta_i = E_p(y_{it})$, the $u_{it} \sim N(0,\sigma^2)$ are mutually independent random variables, the $\epsilon_{it} \sim N(0,\sigma^2_{\epsilon})$ are mutually independent random variables, the $\epsilon_{it}$ and $u_{it}$ are independent of one another, and $v_{it}$ is a stationary process with variance $\sigma^2 = \sigma^2_{\epsilon}/(1-\rho^2)$. If $\rho = 0$ the model is known as the variance components model, otherwise it is a stationary AR (1) with transitory effects.

Under this model the multivariate outcomes $\mathbf{y}_i = (y_{i1},...,y_{iT})', i = 1,...,N$ are independent with mean vector and covariance matrix given respectively by:

$$E_p(\mathbf{y}_i) = (\beta_1,...,\beta_T) = \mathbf{\mu}$$

(26)
\[ \text{cov}_p(y_i) = \sigma^2 J_T + \sigma^2 V_T = \sum \]

where \( J_T \) denotes the \( T \times T \) matrix all of whose elements are one and the \((tt')\)th element of \( V_T \) is \( \rho^{\left| t - t' \right|}; t, t' = \ldots, T \).

Under each of the three correlation structures above and under the different specifications of the population conditional expectation of the inclusion probabilities considered in Section 3, the sample pdf of \( y_i \) can be obtained by suitable substitution.

5. Sample Distribution under Informative Sampling and Informative Dropout

Missing values arise in the analysis of longitudinal data whenever one or more of the sequence of measurements from units within the study is incomplete, in the sense that intended measurements are not taken, or lost, or otherwise unavailable. For example, firms are born and die, plants open and close, individuals enter the survey and exit and animals may die during the course of the experiment. We follow much of the literature on the treatment of missing value in longitudinal data in restricting ourselves to dropout (or attrition); that is to patterns in which missing values are only followed by missing values.

Suppose we intend to take a sequence of measurements, \( y_{iT}, \ldots, y_{iT} \), on the \( i \)th sampled unit. Missing values are defined as dropout if whenever \( y_{ij} \) is missing, so are \( y_{ik} \) for all \( k \geq j \). One important issue which then arises is whether the dropout process is related to the measurement process itself. Following the terminology in Rubin (1976) and Little and Rubin (1987), a useful classification of dropout processes is:

1. Completely random dropout (CRD): the dropout process and measurement processes are independent, that is, the missingness is independent of both observed and unobserved data.

2. Random dropout (RD): the dropout process depends only on the observed measurements, that is, those preceding dropout.

3. Informative dropout (ID): if the dropout process depends both on the observed and on the unobserved measurements, that is, those that would have been observed if the unit had not dropped out.

Following Diggle and Kenward (1994), assume that a complete set of measurements on a sample unit \( i = 1, \ldots, n \) could potentially be taken at all times: \( t = 1, \ldots, T \). Let \( y^*_i = (y^*_{i1}, y^*_{i2}, \ldots, y^*_{iT}) \) denote the complete vector of intended measurements, and \( y_i = (y_{i1}, y_{i2}, \ldots, y_{idi}) \) denote the vector of observed measurements, where \( d_i \) denotes the time of drop-out, with \( d_i = T + 1 \) if no drop-out occurs for unit \( i \). We assume that \( y^*_i \) and \( y_i \) coincide for all time periods during which the \( i \)th unit remains in the study, that is, \( y^*_i = y_i \) if \( t < d_i \).
We define $D_i$ as the random variable, $2 \leq D_i \leq T + 1$, which takes the value of the dropout time, $d_i$ of the $i$th unit, $i = 1, 2, \ldots, n$.

Let $H_{id} = \{y_{i1}, y_{i2}, \ldots, y_{i, t-1}, x_i\}$. Then under the exponential inclusion probability model (7), according to Theorem 2 (a), the sample pdf of the complete series, $y_i^* = (y_{i1}^*, y_{i2}^*, \ldots, y_{iT}^*)'$, is multivariate normal with pdf $f_{y_i^*}(y_i^* | x_i, a_0, \beta, V_0)$ given by (18). Similarly under the linear inclusion probability model (9) according to Theorem 2 (b) the sample pdf is a mixture of multivariate normal and weighted multivariate normal distributions, $f_{y_i^*}(y_i^* | x_i, b, \beta, V_0), b = (b_0, b_1, \ldots, b_p)$, given by (19).

The general model for the dropout process assumes that the probability of dropout at time $t = d_i$ depends on $H_{id}$ and $y_{id}^*$. Then for $d_i \leq T$, we need to specify a model for dropout process of the form:

$$
\Pr(D_i = d_i | H_{id}) = P_{d_i}(H_{id}, y_{id}^*; \varphi) \tag{28}
$$

where $\varphi$ is a vector of unknown parameters.

This modeling of dropout allows the dropout probability to depend on the observed measurement history $H_{id}$ and the unobserved value $y_{id}^*$ as well as on a set of parameters $\varphi$. Diggle and Kenward (1994) propose the following logistic model for informative dropout process with dropout at time $d_i$:

$$
\Pr(D_i = d_i | H_{id}) = P_{d_i}(H_{id}, y_{id}^*; \varphi) = \frac{\exp(\phi_1 y_{i1} + \ldots + \phi_{d_i-1} y_{i,d_i-1} + \phi_{d_i} y_{id}^*)}{1 + \exp(\phi_1 y_{i1} + \ldots + \phi_{d_i-1} y_{i,d_i-1} + \phi_{d_i} y_{id}^*)} \tag{29a}
$$

or

$$
\text{logit} P_{d_i}(H_{id}, y_{id}^*; \varphi) = \log \frac{P_{d_i}(H_{id}, y_{id}^*; \varphi)}{1 - P_{d_i}(H_{id}, y_{id}^*; \varphi)} = \phi_1 y_{i1} + \ldots + \phi_{d_i-1} y_{i,d_i-1} + \phi_{d_i} y_{id}^* \tag{29b}
$$

Once a model for the dropout process has been specified, we can derive the joint distribution of the observed random vector $y_i$, under the assumed multivariate normal sample distribution of $y_i^*$, via the sequence of conditional sample distributions of $y_i^*$, given $H_{id}$. Let $f_{y_i^*}(y_i^* | H_{id}, a^*, \beta, V_0)$ denote the conditional sample pdf of $y_i^*$ given $H_{id}$ where $a^*$ may be $a_0$ or $b$ in Theorem 2. Let $f_{y_i^*}(y_i^* | H_{id}, a^*, \beta, V_0, \varphi)$ the conditional sample pdf of $y_i^*$ given $H_{id}$.

Now, according to Theorem 2 and Diggle and Kenward (1994), we can show that the joint sample distribution of the vector of the complete sequence of measurements $y_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ is given by:
\[ f_s(y_i | x_i) = f_s^*(y_i | x_i) \prod_{t=2}^{T} [1 - P_t(H_{it}, y_{it}; \phi)] \]  

where \( f_s^*(y_i | x_i) = f_s^*(y_i | x_i)f_p^*(y_{i1}, y_{i2}, \ldots, y_{iT} | y_{i1}, x_i) \); see Theorem 2.

For an incomplete sequence \( y_j = (y_{i1}, y_{i2}, \ldots, y_{id_{-1}}) \) with dropout at time \( d_i \), the joint sample distribution is given by:

\[ f_s(y_j | x_j) = f_{s,ld_{-1}}^*(y_j | x_j) \prod_{t=2}^{d_{-1}} [1 - P_t(H_{it}, y_{it}; \phi)] P(D_i = d_i | H_{id_{-1}}) \]  

where \( f_{s,ld_{-1}}^*(y_j | x_j) = f_s^*(y_j | x_j)f_p^*(y_{i2}, \ldots, y_{id_{-1}} | y_{i1}, x_i) \), see Theorem 2, and

\[ P(D_i = d_i | H_{id_{-1}}) = \int P_i(H_{it}, y_{it}; \phi)f_p^*(y_{i1} | H_{it}, \beta, V_0) dy_{i1} \]  

Comment: Note that (31) and (32) take into account the effect of informative sampling and informative dropout.

6. Sample Likelihood and Estimation

In this section we extend the methods of estimation for the analysis of longitudinal survey data under informative sampling, see Eideh and Nathan (2005), to take into account the effects of attrition or dropout, according to the model proposed by Diggle and Kenward (1994). We propose two alternative methods, based on the sample distribution of the observed measurements in the presence of informative sampling and informative dropout, based on the results of the previous section.

6.1. Two Step Estimation

The two sets of parameters in (31), which need to be estimated are those on which the population distribution depends, \( \theta = (\beta, \sigma^2, \nu_{jk}, j,k = 1, \ldots, T) \), and the parameters on which the sample distribution of observed measurements depends, \( \theta^* = (\theta, a, \phi) \), where \( a \) is the parameter indexing the informative sampling process and \( \phi \) is the parameter indexing the dropout process; see equations (18) and (31). Thus the parameters of the sample distribution of the observed measurements, \( \theta^* \) include the parameters of the sample and population distributions. The parameters of the population distribution can be estimated using the sample distribution of observed measurements and using a two-step method. At the first step the parameters of the sample inclusion probability are estimated and the remaining parameters are estimated by minimizing the log likelihood with the first step estimates substituted for the true parameters. As pointed out by Pfeffermann, Krieger and Rinott (1998), the use of this two-step procedure becomes necessary when the conditional expectation of the first order sample selection probabilities is exponential, since in this case there is a problem of identifiability.
In practice, the conditional expectations of the sample inclusion probabilities, $E_Y(\pi_i | y_{ii})$, are not known and usually the only data available to the analyst for the first time period are $\{y_{ij}, w_i; i \in s\}$, where $w_i = 1/\pi_i$ are the sample weights. The estimation of $E_Y(\pi_i | y_{ii})$, using only the sample data can be based on the following relationship, due to Pfeffermann and Sverchkov (1999):

$$E(y_i | y_{ii}) = \frac{1}{E_Y(\pi_i | y_{ii})}$$

(33)

The prominent feature of this relationship is that the expectation of the population conditional sample inclusion probabilities can be identified and estimated from the sample data.

Thus for the exponential inclusion probability model (7), the two-step estimation proceeds as follows:

**Step-one:** Estimation of $a_0$:

According to (33) we estimate $a_0$ by regressing $-\log(w_i)$ against $y_{ij}, x_{i11}, ..., x_{ip}, i \in s$.

**Step-two:** Substituting the ordinary least squares estimator, $\tilde{a}_0$ of $a_0$ in the sample distribution of observed measurements. The contribution to the log-likelihood function for the observed measurements of the $i$th sampled unit can be written as:

$$L_i(\beta, \sigma^2, v_{jk}, \phi) = \log f(y_i | x_i, \beta, \sigma^2, v_{jk}, \tilde{a}_0)$$

$$= \log \int f_{x,d} x_{i,d} | x_i, \beta, \sigma^2, v_{jk}, \tilde{a}_0 + \log \prod_{t=2}^{d} \left[1 - P(H_{id}, y_{it}, \phi)\right]$$

$$+ \log P(D_i = d_i | H_{id}, \beta, \sigma^2, v_{jk}, \phi)$$

(34)

Thus the full sample log-likelihood function for the observed measurements to be maximized with respect to $(\beta, \sigma^2, v_{jk}, \phi)$ is given by:

$$L_{v_k}(\beta, \sigma^2, v_{jk}, \phi) = \sum_{i=1}^{n} L_i(\beta, \sigma^2, v_{jk}, \phi)$$

$$= L_1(\beta, \sigma^2, v_{jk}, \tilde{a}_0) + L_2(\phi) + L_3(\beta, \sigma^2, v_{jk}, \phi)$$

(35)

where

$$L_1(\beta, \sigma^2, \psi, \tilde{a}_0) = \sum_{i=1}^{n} \log f_{x,i} x_{i} | x_i, \beta, \sigma^2, v_{jk}, \tilde{a}_0$$

$$= \sum_{i=1}^{n} \log f_{y,i} y_{ii} | x_i, \beta, \sigma^2, v_{jk}, \tilde{a}_0$$

$$+ \sum_{i=1}^{n} \log f_{y,i} (y_{i2}, ..., y_{ip} | y_{ii}, x_i, \beta, \sigma^2, v_{jk})$$

(36a)
\[
L_2(\varphi) = \sum_{i=1}^{n} \sum_{t=2}^{d-1} \log \left[ 1 - P_i(H_{it}, y_{it}; \varphi) \right] 
\]
(36b)

\[
L_3(\beta, \sigma^2, v_{jk}, \varphi) = \sum_{i, t \in T} \log P(D_i = d_i | H_{id}, \beta, \sigma^2, v_{jk}, \varphi) 
\]
(36c)

The explicit form of \( L_i(\beta, \sigma^2, v_{jk}; \varphi) \) is obtained from equations (16)-(18) by setting \( T = d_i - 1 \). \( L_2(\varphi) \) is determined by equation (32). The last term \( L_3(\beta, \sigma^2, v_{jk}, \varphi) \) is determined by equation (31) which requires the distribution \( f_{\varphi}^*(y_i | H_{it}, \beta, V_0) \).

Now, under the assumption that \( y_i \sim MN(x, \beta, V^+) \) and using the properties of the multivariate normal distribution, see Johnson and Wichern (1998, p. 170), we have that the conditional distribution of \( y_{id} \) given \( H_{id} = \{y_{i1}, ..., y_{i(d_i-1)}\} \) is univariate normal with mean

\[
E(y_{id} | H_{id}) = E(y_{id}) + \left( g^{(d_i-1)^T} \right)^{(-1)} (y_{id}^{(d_i-1)} - \mu^{(d_i-1)}) 
\]
(37)

and variance

\[
Var(y_{id} | H_{id}) = v^{d_i,d_i} - \left( g^{(d_i-1)^T} \right)^{(-1)} \left( g^{(d_i-1)} \right) 
\]
(38)

where

\[
E(y_{id}) = x_{id} \beta; \quad g^{(d_i-1)} = (g_{i1}^{(d_i-1)}, g_{i2}^{(d_i-1)}, ..., g_{id_{i-1}}^{(d_i-1)}); \quad g_{ij}^{(d_i-1)} = Cov(y_j, y_{id}), j = 1, ..., d_i \\
\mu^{(d_i-1)} = (E(y_{i1}), E(y_{i2}), ..., E(y_{id_{i-1}}))^T; \quad V^{(d_i-1)} = Var(y_{i}^{(d_i-1)}) = Cov(y_i^{(d_i-1)}, y_i^{(d_i-1)}) \\
V^{(d_i-1)} = Var(y_{i}^{(d_i-1)}) = Cov(y_i^{(d_i-1)}, y_i^{(d_i-1)}) 
\]

Similar results can be obtained under the linear inclusion probability model (9).

6.2. Pseudo Likelihood Approach

This approach is based on solving the estimated census maximum likelihood equations. The census maximum likelihood estimator of \( \theta = (\beta, \sigma^2, v_{jk}) \) solves the census likelihood equations, which in our case are:

\[
U(\theta) = \frac{\partial}{\partial \theta} \log L_j(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log L_{ip}(\beta, \sigma^2, v_{jk}) 
\]
(39)

The pseudo maximal likelihood (PML) estimator is defined as the solution of \( \hat{U}(\theta) = 0 \) where \( \hat{U}(\theta) \) is a sample estimator of the census log-likelihood, \( U(\theta) \). See Binder (1983).
We consider the following estimator for estimating \( L_1(\beta, \sigma^2, v_{jk}) \) from the sample data:

\[
L_{w_1}(\beta, \sigma^2, v_{jk}) = \sum_{i=1}^{n} w_i \log f(y_{il} | x_{il}, \beta, \sigma^2, v_{11}) + \sum_{i=1}^{n} w_i^* \log f(y_{il} | x_{il}, \beta, \sigma^2, v_{jk})
\]

where we can take \( w_i^* = w_i \) or \( w_i^* = N/n \) or \( w_i^* = w_i \).

Thus the pseudo maximal likelihood (PML) estimator of \( \theta = (\beta, \sigma^2, v_{jk}) \) is the solution of

\[
\frac{\partial}{\partial \theta} \hat{U}_w(\beta, \sigma^2, v_{jk}, \phi) = \frac{\partial}{\partial \theta} [L_{w_1}(\beta, \sigma^2, v_{jk}) + L_2(\phi) + L_3(\beta, \sigma^2, v_{jk}, \phi)]
\]

For more details, see Eideh and Nathan (2005).

7. **Empirical Example – British Labour Force Survey**

The British Labour Force Survey (LFS) is a household survey, gathering information on a wide range of labour force characteristics and related topics. Since 1992 it has been conducted on a quarterly basis, with each sample household retained for five consecutive quarters, and a fifth of the sample replaced each quarter. The survey was designed to produce cross-sectional data, but in recent years it has been recognized that linking together data on each individual across quarters would produce a rich source of longitudinal database that could be exploited to give estimates of gross change over time – see e.g., Tate (1999). Labour force gross flows are typically defined as transitions over time between the major labour force states, employed, unemployed, and not in labour force (or economically inactive). Quarter to quarter gross flows show how individuals or persons with each labour force state or classification in one quarter are classified in the following quarter. Gross flows provide estimates of the number of individuals who went from employed in one quarter to employed in the next quarter, employed to unemployed, employed to not in labour force, and so forth. Estimates of labour force flows are useful for answering questions such as: (1) how much of the net increase in unemployment is due to individuals losing jobs and how much is due to individuals formerly not in the labour force starting to look for jobs; (2) how many unemployed individuals become discouraged and leave the labour force? A number of problems are associated with the estimation of gross flows. Some of these problems are (1) quarter to quarter nonresponse; (2) measurement errors or response errors; (3) sample rotation; and (4) complex sample design effects. In this numerical example we consider only the quarter-to-quarter nonresponse. The problem of handling quarter-to-quarter nonresponse was discussed and studied by Clark and Chambers (1998), Clarke and Tate (1999).

In order to accommodate the differences between the assumptions of sections 2-7 and those required for the present application, primarily due to the fact the LFS data relate to categorical rather than to continuous variables, the following modifications were made:
Let $N_h = [N_h(1,1), \ldots, N_h(3,3)]$ 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 $n_h = [n_h(1,1), \ldots, n_h(3,3)]$. 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 $\omega = [\omega(1,1), \ldots, \omega(3,3)]$, of which 8 are free.

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 response pattern of household $h$ by $R_h = (R_{h1}, R_{h2})$ 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 $r_h = (r_{h1}, r_{h2})$. Let $S_{uv} = \{h : r_h = (u,v); u,v = 0,1\}$, so that $S_{11} = \{h : r_h = (1,1)\}$ denotes the subset of households with response 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 who responded in both quarters.

The estimates of labour force gross flows are shown in Table 1. Following are details on the methods of estimation used:

(1) **Unweighted method:**

The first column of Table 1 gives estimates from the unweighted data, obtained by maximizing the likelihood:

$$l(\omega) = n_+ (1,1) \log \omega(1,1) + n_+ (1,2) \log \omega(1,2) + \ldots + n_+ (3,3) \log \omega(3,3)$$  \hspace{1cm} (42)

subject to the constraint $\sum_{a,b} \omega(a,b) = 1$, where $n_+(a,b) = \sum_{h=1}^{m} n_h(a,b)$.

(2) **Weighted methods:**

The second column of Table 1 gives estimates from the weighted data at the household level, computed as:

$$\hat{\omega}(a,b) = \frac{\sum_h d_h n_h(a,b)}{\sum_{h=1}^{m} d_h}; a,b = 1,2,3$$  \hspace{1cm} (43)

(3) **The sample likelihood method:**
The sample likelihood was derived under the assumptions of the exponential and the linear models for the household weights as a function of the labour force flow frequencies, on the basis of the relationships between the population likelihood and that of the respondent sample.

(a) Exponential model (SMLE):

The third column in Table 1 gives the estimates based on the sample log-likelihood under the exponential model:

$$E(d_h^{-1}|N_h = n_h, \alpha) = \exp(\alpha' n_h) = \exp\left(\alpha(0,0)n_h(0,0) + \sum_{a,b} \alpha(a,b)n_h(a,b)\right),$$  \hspace{1cm} (44)

where $n_h(0,0) = 1$, $\alpha = (\alpha(0,0), \alpha(1,1), ..., \alpha(3,3))$ and $d_h = \frac{1}{n_h} \sum_{i=1}^{n_h} d_{hi}$, $d_{hi}$ represents the longitudinal weight within household $h$ for individual $i$ and $n_h$ is the number of individuals in household $h$.

They are obtained by maximizing:

$$l_r (\omega) = n_+(1,1) \log \omega(1,1) + n_+(1,2) \log \omega(1,2) + ... + n_+(3,3) \log \omega(3,3)$$
$$- n_+ \log(\omega(1,1) \exp(\bar{\alpha}(1,1)) + \omega(1,2) \exp(\bar{\alpha}(1,2)) + ... + \omega(3,3) \exp(\bar{\alpha}(3,3)))$$  \hspace{1cm} (45)

subject to the constraint $\sum_{a,b} \omega(a,b) = 1$, where $n_+(a,b) = \sum_{h=1}^{m} n_h(a,b); a,b = 1,2,3$, $n_+ = \sum_{h=1}^{m} n_h(a,b)$ and $\bar{\alpha}(a,b)$ are the estimated coefficients.

(4) Sample likelihood based on estimates of propensity scores:

Response propensity scores (3) were estimated by two alternative methods, based on sets of explanatory variables. Heckman (1976, 1979) model used to deal with specification bias was used to estimate the probability of nonresponse. The reciprocals of the estimated propensities are used instead of the household weights for the derivation of the sample likelihood function. Maximum likelihood estimates of the labour force flows were then obtained, under the exponential and linear models.

Estimating the Probability of Response - Model-Based Approach:

Under the assumption that the response set is $S_{11}$ and the nonresponse set is $S_{10} \cup S_{01}$, let the response indicator variable be redefined by:

$$R_h^* = 1,$$ if household $h$ responds in both quarters, that is if $h \in S_{11}$
Let $x_h$ be a vector of covariate or auxiliary information at household level, with data at the individual summarized at the household level, which is available for all units (respondent=$S_{11}$ or nonrespondent=$S_{10} \cup S_{01}$), for example, tenure, age, sex, type of household, type of family unit, composition of household, own or rent accommodation, social class, time lived at this address, highest qualification, number of dependent children in household under age 19, number of dependent children in family aged 5-9, type of family unit, age youngest child in household under 19, and region of usual residence. Let $\beta$ be a vector of unknown coefficients including the constant term.

The probability of response or the propensity score for household $h$ is given by:

$$
\psi_h = P(R^*_h = 1|x_h), \quad h = 1, \ldots, n^*
$$

In order to estimate $\psi_h$ we adopt the Heckman Model:

In order to apply the Heckman method (1976, 1979) for dealing with nonresponse for LFS data, let $y_h(a,b)$ be the number of individuals who have labour force flow $(a,b)$ between the two quarters. Consider a random sample of $n^*$ households. Equations for household $h$ are given by:

$$
y_{1h}(a,b) = \beta_1 x_{1h} + u_{1h}
$$

$$
y_{2h}(a,b) = \beta_2 x_{2h} + u_{2h}
$$

where $h = 1, \ldots, n^*$, $x_{ih}, (i = 1,2)$ are vectors of covariates or auxiliary information at household level, which is available for all units (respondent=$S_{11}$ or nonrespondent=$S_{10} \cup S_{01}$), $\beta_i$ are vectors of unknown coefficients including the constant term, and

$$
E(u_{ih}) = 0, E(u_{ih}u_{i'h'}) = \sigma_{i'i'}^* \text{ for } h = h'\text{ and } 0 \text{ for } h \neq h'; i, i' = 1, 2.
$$

We are interested in estimating equation (47a) but data are missing on $y_{1h}$ for $h \in S_{10} \cup S_{01}$.

The population regression for equation (47a) may be written as:

$$
E_p(y_{1h}(a,b)) = \beta_1 x_{1h}, \ h = 1, \ldots, n^*
$$

The regression equation for the subsample of available data, that is $S_{11}$, is:
\[ E(y_{1h}(a,b)|x_{1h}, \text{sample selection rule}) = \beta_1 x_{1h} + E(u_{1h}|\text{sample selection rule}) \]  

(49)

\( h = 1, \ldots, n^* \), where the convention is adopted that the first \( n_{11} < n^* \) households have data available on \( y_{1h}(a,b) \).

If the conditional expectation of \( u_{1h} \) in (49) is zero, the regression function for the selected subsample is the same as the population regression function. Least squares estimators may then be used to estimate \( \beta_1 \) on the selected subsample, \( S_{11} \), and the only cost of having an incomplete sample is a loss in efficiency.

In the general case, the sample selection rule that determines the availability of data has more serious consequences. Suppose that data are available on \( y_{1h} \) if \( y_{2h} \geq 0 \) while if \( y_{2h} < 0 \), there are no observations on \( y_{1h} \). The choice of zero as a threshold involves an inessential normalization.

In the general case, we assume:

\[ E(u_{1h}|x_{1h}, \text{sample selection rule}) = E(u_{1h}|x_{1h}, y_{2h} \geq 0) = E(u_{1h}|x_{1h}, u_{2h} \geq -\beta_2 x_{2h}) \]  

(50)

In the case of independence of \( u_{1h} \) and \( u_{2h} \), so that the data on \( y_{1h} \) are missing at random, the conditional mean of \( u_{1h} \) in (49) is zero. In the general case, it is nonzero, so that the data on \( y_{1h} \) are not missing at random, and the subsample regression function is:

\[ E(y_{1h}|x_{1h}, \text{sample selection rule}) = \beta_1 x_{1h} + E(u_{1h}|x_{1h}, u_{2h} \geq -\beta_2 x_{2h}) \]  

(51)

The selected sample regression function depends on \( x_{1h} \) and \( x_{2h} \). Regression estimators of the parameters of equation (47a) based on the selected sample only omit the final term of equation (51) as a regressor, so that the bias that results from using nonrandomly selected samples to estimate behavioral relationships is seen to arise from the ordinary problem of omitted variables.

In order to compute the conditional expectation, \( E(u_{1h}|x_{1h}, u_{2h} \geq -\beta_2 x_{2h}) \), in equation (51), we assume that \( (u_{1h}, u_{2h}) \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right) \). Thus we have:

\[ E(u_{1h}|x_{1h}, u_{2h} \geq -\beta_2 x_{2h}) = \frac{\sigma_{12}}{\sqrt{\sigma_{22}}} \lambda_h \]  

(52a)

\[ E(u_{2h}|x_{1h}, u_{2h} \geq -\beta_2 x_{2h}) = \frac{\sigma_{22}}{\sqrt{\sigma_{22}}} \lambda_h \]  

(52b)

where:
\begin{align*}
\lambda_h &= \frac{\phi(z_h)}{\Phi(-z_h)} = \frac{\phi(z_h)}{1 - \Phi(z_h)}, \quad z_h = -\frac{\beta_2 x_{2h}}{(\sigma_{22})^{0.5}}
\end{align*}

and \( \phi \) and \( \Phi \) are, respectively, the density and distribution function for a standard normal variable.

Thus the conditional regression functions for the selected samples can be written as:

\begin{align*}
E(y_{1h} | x_{1h}, y_{2h} \geq 0) &= \beta_1 x_{1h} + \frac{\sigma_{12}}{(\sigma_{22})^{0.5}} \lambda_h \quad (53a) \\
E(y_{2h} | x_{2h}, y_{2h} \geq 0) &= \beta_2 x_{2h} + \frac{\sigma_{22}}{(\sigma_{22})^{0.5}} \lambda_h \quad (53b)
\end{align*}

Now using the Heckman method we can estimate the probability of response of household:

\begin{equation}
\Pr(y_{2h} \geq 0) = \Pr(u_{2h} \geq -\beta_2 x_{2h}) = \Phi\left( -\frac{\beta_2 x_{2h}}{(\sigma_{22})^{0.5}} \right) = 1 - \Phi\left( \frac{\beta_2 x_{2h}}{(\sigma_{22})^{0.5}} \right) \quad (54)
\end{equation}

using probit analysis for the full sample, \( S^* = S_{11} \cup S_{10} \cup S_{01} \).

Here we assume that the response variable is \( r = (1,1,...,1,0,0,...,0) \), where the number of ones is \( n_{11} \) and number of zeros is \( n_{01} + n_{00} \). Also \( x_{2h} \) represents the auxiliary variables of interest for the full sample, \( S^* = S_{11} \cup S_{10} \cup S_{01} \). After that we estimate the probability of response of household \( h \) as follows:

\begin{equation}
\hat{\psi}_h = \hat{\Pr}(y_{2h} \geq 0) = \Phi\left( \hat{\beta}_2^* x_{2h} \right), \quad \hat{\beta}_2^* = -\frac{\hat{\beta}_2}{(\hat{\sigma}_{22})^{0.5}}, \quad \text{for } h = 1,...,n_{11}, \text{ that is, for } h \in S_{11} \quad (55)
\end{equation}

Having estimated the probability of response of household \( h \), \( \hat{\psi}_h \), using the Heckman model – equation (54), we can apply the response sample likelihood introduced in (45) by taking \( \hat{d}_h = \hat{\psi}_h^{-1} \). In our case the covariates used at household levels are:

1. Mean of individual calibrated weights – these weights are functions of age, sex, education level, social class, tenure.
2. Number of dependent children in family under age two.

The results obtained are summarized in the following table.
Table 1: Gross Flows Estimates (percentages)

<table>
<thead>
<tr>
<th>Flow</th>
<th>Unweighted</th>
<th>Weighted Exponential-SMLE</th>
<th>Heckman-SMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>EE</td>
<td>70.62</td>
<td>69.32</td>
<td>69.78</td>
</tr>
<tr>
<td>EU</td>
<td>1.08</td>
<td>1.25</td>
<td>1.17</td>
</tr>
<tr>
<td>EN</td>
<td>1.53</td>
<td>1.71</td>
<td>1.61</td>
</tr>
<tr>
<td>UE</td>
<td>1.61</td>
<td>1.62</td>
<td>1.61</td>
</tr>
<tr>
<td>UU</td>
<td>3.78</td>
<td>4.41</td>
<td>4.16</td>
</tr>
<tr>
<td>UN</td>
<td>1.00</td>
<td>1.12</td>
<td>1.07</td>
</tr>
<tr>
<td>NE</td>
<td>1.40</td>
<td>1.29</td>
<td>1.35</td>
</tr>
<tr>
<td>NU</td>
<td>1.06</td>
<td>1.14</td>
<td>1.12</td>
</tr>
<tr>
<td>NN</td>
<td>17.92</td>
<td>18.12</td>
<td>18.13</td>
</tr>
</tbody>
</table>

The main findings from Table 1 are:

1. There are small differences between unweighted and weighted gross flow estimates, both.

2. There are only small differences between gross flow estimates based on household level weighting and those obtained based on sample likelihoods, under 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, marital 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.

3. Both the household level weighting and sample likelihood procedures for estimating the labour force gross flows seem to reduce at least part of the effects of nonresponse, compared to the unweighted method. Based on their simulation study, Clarke and Tate (2003), recommend similarly that weighting should be used to produce flows estimates that offer a considerable improvement in bias over unadjusted estimates. Although the simple sample likelihood estimates cannot be shown to be better than the weighted estimators, their similarity to the weighted estimates indicate that they are an improvement over the unweighted estimates.

4. Using propensity response scores based on the Heckman model accounts partially for informative nonresponse.

8. Conclusion

In the empirical result, we introduce alternative methods of obtaining weighted estimates of gross flows, taking into account informative nonresponse. The first method is based on extracting the response labour force sample likelihood as a function of the population
labour force likelihood and of the response probabilities, which are obtained using two methods: -based on the reciprocals of the adjusted calibrated weights, and based on the Heckman model. The new methods are model based while the classical method is based on the adjusted weights. Thus we think that the new methods are more efficient than the weighted method, although no hard evidence for this is available. However the two methods, sample likelihood and weighting, give approximately the same estimates of labour force gross flows when the propensity scores are based on the reciprocals of the adjusted calibrated weights, while the sample likelihood method using the Heckman model for constructing propensity score seems to give better estimates than the other methods because it accounts partially for informative nonresponse.

Initially we considered that the estimates of gross flows based on the response sample likelihood might explain the nonignoreable nonresponse. The similarity of the results of the weighted and response likelihood methods is not surprising, since 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 contain 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.

Overall, we recommend constructing propensity scores base on the Heckman model because this method accounts partially for informative nonresponse. The results obtained for labour force gross estimates are closer to the results based on labour force flow model and nonignoreable nonresponse model; see Clarke and Tate (2002).

References


