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

Abdulhakeem A.H. Eideh
Al-Quds University, Palestine
E-mail: msabdul@palnet.com

Gad Nathan
Hebrew University of Jerusalem, Israel
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 of 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 data from the British Household Panel Survey 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 - the sample selection probabilities depend on the values of the model outcome variable (or the model outcome variable is correlated with design variables not included in the model). For example, if the sample design is clustered, with PSU's selected with probabilities proportional to size (e.g., size of locality) and the dependent variable (e.g., income) is related to the size of the locality, ignoring the effects of this dependence can cause bias in the estimation of regressions coefficients. In theory, the effect of the sample selection can be controlled for by including among the model all the design variables used for the sample selection. However, this possibility is often not operational because there may be too many of them or because they are not of substantive interest. Initial non-response in a longitudinal survey may also be considered as a form of 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 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).*

The predominant method of analysis for longitudinal data has long been based on the application of generalized linear models (GLMs) – McCullagh and Nelder (1999) – to repeated measures and the use of generalized estimating equations (GEEs) to estimate the model parameters – see for instance Diggle, Liang, and Zeger (1994). The generalized linear model describes the conditional distribution of the outcome, given its past, where the distribution parameters may vary across time and across subjects as a stochastic process, according to a mixing distribution. Two different approaches to longitudinal analysis are dealt with by means of similar GLMs. In the "subject-specific" approach, sometimes referred to as the *random effects model*, the heterogeneity between subjects is explicitly modelled, while in the "population-averaged" approach, sometimes referred to as the *marginal model*, the average response is modeled as a function of the covariates, without explicitly accounting for subject heterogeneity.

Frequently longitudinal sample surveys deal with hierarchical populations, such as individuals within households or employees within establishments, for which multi-level modeling is appropriate. In another approach, Goldstein, Healy and Rasbash (1994) consider the analysis of repeated measurements using a two-level hierarchical model, with individuals as second level and the repeated measurements as the first level.

Path analysis, which has long been a preferred method of modelling complex relationships between large numbers of variables in cross-sectional analysis of structured data sets in the social sciences, has been generalized to modelling longitudinal data primarily by means of *Graphical Chain Modelling* (GCM) and *Structural Equation Modelling* (SEM) - see Wermuth and Lauritzen (1990) and Mohamed, Diamond and Smith (1998).

Other models used for the anlaysis of longitudinal data, include *Antedependence Models* designed to deal with nonstationarily. Zimmerman and Nunez-Anton (2000) propose structured and unstructured antedependence models for longitudinal data, primarily in the context of growth analysis.

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

The analysis of longitudinal data with nonignorable missing values has received serious attention in the last 20 years. For example, Little (1993) explores pattern-mixture models and pattern-set mixture models for multivariate incomplete data. Diggle and Kenward (1994) propose a likelihood-based method for longitudinal data subject to informative and monotone missingness. Little (1995) discusses methods that simultaneously model the data and the drop-out process via two broad classes of models – random-coefficient selection models and random-coefficient pattern-mixture models, based on likelihood inference methods. Little and Wang (1996) fit pattern mixture models for multivariate incomplete data with covariates, via maximum likelihood and Bayesian methods, using the EM algorithm. Troxel, Harrington and Lipsitz (1998) use full likelihood methods to analyze continuous longitudinal data with non-ignorable (informative) missing values and non-monotone patterns. Lipsitz, Ibrahim and Molenberghs (2000) use a Box-Cox transformation for the analysis of longitudinal data with incomplete responses, while Ibrahim, Chen and Lipsitz (2001) estimate the parameters in the generalized linear mixed models with nonignorable missing response data and with non-monotone patterns of missing data in the response variable. For further discussion see Schafer (1995) and Little and Rubin (2002).

(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, which 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 belong to the same family and only differ in some or all the parameters. Several 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. Pfeffermann and Sverchkov (1999) propose two new classes (parametric and semi-parametric) of estimators for regression models fitted to survey data. Sverchkov and Pfeffermann (2004) use the sample distribution for predicting finite population totals under informative sampling. Pfeffermann, Moura and Silva (2001) propose a model-dependent approach for two-level modeling that accounts for informative sampling. Pfeffermann and Sverchkov (2003) consider four different approaches to defining parameter-estimating equations for generalized linear models, under informative probability sampling design, utilizing the sample distribution. Chambers, Dorfman and Sverchkov (2003) describe a framework for applying a common exploratory data analysis procedure, nonparametric regression, to sample survey data. Eideh (2007) use the sample distribution for deriving best linear unbiased predictors of the area means for areas in the sample and for areas not in the sample. For more discussion; see Eideh and Nathan (2003, 2006) and Nathan and Eideh (2004).

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 introduce 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 for 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 consists of a permanent random effect at the individual level and autocorrelated transitory random effects corresponding to different waves of investigation. Eideh and Nathan (2006) 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 $y_{it}, i = 1,...,N; t = 1,...,T$ be the measurement on the *i-th* subject at time $t = 1,...,T$. Associated with each y_{it} are the (known) values, $x_{itk}, k = 1,...p$, of p explanatory variables. We assume that the y_{it} follow the regression model:

$$y_{it} = \beta_1 x_{it1} + ... + \beta_p x_{itp} + \varepsilon_{it}, \tag{2.1}$$

where the ε_{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 ε_{it} to be correlated within subjects.

Let $\mathbf{y}_i = (y_{i1},...,y_{iT})'$, $\mathbf{x}_{it} = (x_{it1},...,x_{itp})'$ and let $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ be the vector of unknown regression coefficients. The general linear model for longitudinal survey data treats the random vectors $\mathbf{y}_i, i = 1,..., N$ as independent multivariate normal variables, that is

$$\mathbf{y}_i \underset{p}{\sim} MVN(\mathbf{x}_i'\boldsymbol{\beta}, \mathbf{V}), \tag{2.2}$$

where $\mathbf{x}_i$ is the matrix of size T by p of explanatory variables for subject i, and $\mathbf{V}$ has $(jk)-th$ element, $v_{jk} = \operatorname{cov}_p(y_{ij}, y_{ik}), j,k = 1,...,T$; see Diggle, Liang and Zeger (1994).

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, π_i, depend on the population values of the response variable, y_{i1}, at the first occasion only, and on $\mathbf{x}_{i1} = (x_{i11},\ldots,x_{i1p})'$.

Theorem 1.

Let $\mathbf{y}_i \underset{p}{\sim} f_p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})$ be the population distribution of $\mathbf{y}_i$ given $\mathbf{x}_i$ and $\mathbf{y}_{i,T-1} = (y_{i2}, y_{i3},\ldots, y_{iT})'$. If we assume that π_i depends only on y_{i1} and on $\mathbf{x}_{i1}$, then the (marginal) sample distribution of $\mathbf{y}_i$ given $\mathbf{x}_i$ is given by:

$$f_s(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) = f_s(y_{i1}|\mathbf{x}_{i1}, \boldsymbol{\theta}) f_p(\mathbf{y}_{i,T-1}|y_{i1}, \mathbf{x}_i; \boldsymbol{\theta}), \tag{3.1}$$

where

$$f_s\left(y_{i1}\left|\mathbf{x}_{i1},\boldsymbol{\theta}\right.\right)=\frac{E_p\left(\pi_i\left|y_{i1},\mathbf{x}_{i1}\right.\right)}{E_p\left(\pi_i\left|\mathbf{x}_{i1},\boldsymbol{\theta}\right.\right)}f_p\left(y_{i1}\left|\mathbf{x}_{i1},\boldsymbol{\theta}\right.\right) \tag{3.2}$$

is the sample distribution of y_{i1} given $\mathbf{x}_{i1}$ and

$$E_p\left(\pi_i\mid\mathbf{x}_{i1},\boldsymbol{\theta}\right)=\int E_p\left(\pi_i\mid y_{i1},\mathbf{x}_{i1}\right)f_p\left(y_{i1}\mid\mathbf{x}_{i1},\boldsymbol{\theta}\right)dy_{i1} \tag{3.3}$$

Proof: See Eideh and Nathan (2006).

Assuming independence of the population measurements, Pfeffermann, Krieger and Rinott. (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.

Note that, for a given population distribution, $f_p\left(y_{i1}\left|\mathbf{x}_{i1},\boldsymbol{\theta}\right.\right)$, the sample distribution, $f_s\left(y_{i1}\left|\mathbf{x}_{i1},\boldsymbol{\theta},\boldsymbol{\gamma}\right.\right)$, in equation (3.2) is completely determined by $E_p\left(\pi_i\mid y_{i1},\mathbf{x}_{i1};\boldsymbol{\gamma}\right)$.

Consider the following approximation models for the population conditional expectation; $E_p\left(\pi_i\mid y_{i1},\mathbf{x}_{i1};\boldsymbol{\gamma}\right)$, proposed by Pfeffermann, Krieger, and Rinott (1998), and Skinner (1994):

(a) *Exponential Inclusion Probability Model*:

$$E_p\left(\pi_i\left|y_{i1},\mathbf{x}_{i1}\right.\right)=\exp\left(a_0^*+a_0y_{i1}+a_1x_{i11}+a_2x_{i12}+\ldots+a_px_{i1p}\right) \tag{3.4}$$

(b) *Linear Inclusion Probability Model*:

$$E_p\left(\pi_i\left|y_{i1},\mathbf{x}_{i1}\right.\right)=b_0^*+b_0y_{i1}+b_1x_{i11}+b_2x_{i12}+\ldots+b_px_{i1p} \tag{3.5}$$

Nathan and Eideh (2004) consider additional approximations for the population conditional expectation, namely the logit and probit models.

Theorem 2.

We assume the multivariate normal population distribution of $\mathbf{y}_i$ given $\mathbf{x}_i$, defined by equation (2.2). Then under the exponential inclusion probability model given by equation (3.4), it can be shown that the sample distribution of $\mathbf{y}_i$, given $\mathbf{x}_i,\boldsymbol{\theta},a_0$, is multivariate normal:

$$\mathbf{y}_i\left|\mathbf{x}_i,\boldsymbol{\theta},a_0\right.\underset{S}{\sim}MVN\left(\boldsymbol{\mu}^*,\mathbf{V}\right) \tag{3.6}$$

where $\boldsymbol{\mu}^* = \left(\mathbf{x}_{i1}'\boldsymbol{\beta} - a_0 v_{11}, \mathbf{x}_{i2}'\boldsymbol{\beta}, \ldots, \mathbf{x}_{iT}'\boldsymbol{\beta}\right)'$. Thus, only the mean of the sample distribution (3.6) differs from that of the population distribution (2.2), whereas the variance matrix, **V,** remains the same. The distributions coincide when $a_0 = 0$, that is when the sampling design is noninformative.
Similar results are obtained for the linear, logit and probit inclusion probability models.

4. 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, $y_{i1},\ldots,y_{iT}$, on the *i*th sampled unit. Missing values are defined as dropout if whenever y_{i_j} is missing, so are y_{i_k} 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 (2002), 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): 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 $\mathbf{y}_i^* = \left(y_{i1}^*, y_{i2}^*, \ldots, y_{iT}^*\right)'$ denote the complete vector of intended measurements, and $\mathbf{y}_i = \left(y_{i1}, y_{i2}, \ldots, y_{i,d_i-1}\right)'$ 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 $\mathbf{y}_i^*$ and $\mathbf{y}_i$ coincide for all time periods during which the *i*th unit remains in the study, that is, $y_{it} = y_{it}^*$ 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 $H_{it} = \{y_{i1}, y_{i2}, ..., y_{i,t-1}, \mathbf{x}_i\}$. Then under the exponential inclusion probability model (3.4), according to Theorem 2, the sample pdf of the complete series, $\mathbf{y}_i^* = (y_{i1}^*, y_{i2}^*, ..., y_{iT}^*)'$, is multivariate normal as defined by (3.6).

The general model for the dropout process assumes that the probability of dropout at time $t = d_i$ depends on H_{id_i} and on $y_{id_i}^*$. Then for $d_i \leq T$, Diggle and Kenward (1994) propose the following *logistic model* for informative dropout process with dropout at time d_i:

$$\log\mathrm{it}\left[P_{d_i}\left(H_{id_i}, y_{id_i}^*; \boldsymbol{\varphi}\right)\right] = \log\frac{P_{d_i}\left(H_{id_i}, y_{id_i}^*; \boldsymbol{\varphi}\right)}{1 - P_{d_i}\left(H_{id_i}, y_{id_i}^*; \boldsymbol{\varphi}\right)} = \phi_1 y_{i1} + ... + \phi_{d_i-1} y_{i,d_i-1} + \phi_{d_i} y_{id_i}^*, \quad (4.1)$$

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

Once a model for the dropout process has been specified, we can derive the joint distribution of the observed random vector $\mathbf{y}_i$, under the assumed multivariate normal sample distribution of $\mathbf{y}_i^*$, via the sequence of conditional sample pdf's of y_{it} given H_{it}, $f_{ts}\left(y \middle| H_{it}, \mathbf{a}^*, \boldsymbol{\beta}, \mathbf{V}_0, \boldsymbol{\varphi}\right)$ and the sequence of conditional sample pdf's of y_{it}^* given H_{it}, $f_{ts}^*\left(y \middle| H_{it}, \mathbf{a}^*, \boldsymbol{\beta}, \mathbf{V}_0\right)$.

For an incomplete sequence $\mathbf{y}_i = \left(y_{i1}, y_{i2}, ..., y_{i,d_i-1}\right)'$ with dropout at time d_i, the joint sample distribution is given by:

$$f_s\left(\mathbf{y}_i \middle| \mathbf{x}_i\right) = f_{s,d_i-1}^*\left\{\mathbf{y}_{i,d_i-1} \middle| \mathbf{x}_i\right\} \; * \left\{\prod_{t=2}^{d_i-1}\left[1 - P_t\left(H_{it}, y_{it}; \boldsymbol{\varphi}\right)\right]\right\} P\left(D_i = d_i \middle| H_{id_i}\right) \qquad (4.2)$$

where $f_{s,d_i-1}^*\left\{\mathbf{y}_{i,d_i-1} \middle| \mathbf{x}_i\right\} = f_s^*\left(y_{i1} \middle| \mathbf{x}_{i1}\right) f_p^*\left(y_{i2}, ..., y_{i,d_i-1} \middle| y_{i1}, \mathbf{x}_i\right)$, see Theorem 2 , and

$$P\left(D_i = d_i \middle| H_{id_i}\right) = \int P_t\left(H_{it}, y_{it}; \boldsymbol{\varphi}\right) f_{tp}^*\left(y_{it} \middle| H_{it}, \boldsymbol{\beta}, \mathbf{V}_0\right) dy_{it} \qquad (4.3)$$

Comment: Note that (4.2) and (4.3) take into account the effect of informative sampling and informative dropout.

5. 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 (2006), 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 results of the previous section on the sample distribution of the observed measurements, in the presence of informative sampling and informative dropout.

5.1. *Two Step Estimation*

The two sets of parameters in (4.2), which need to be estimated are those on which the population distribution depends, $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \sigma^2, v_{jk}\right), j.k = 1,...,T$, and the parameters on which the sample distribution of observed measurements depends, $\boldsymbol{\theta}^* = \left(\boldsymbol{\theta}, a_0, \boldsymbol{\varphi}\right)$, where a_0 is the parameter indexing the informative sampling process and $\boldsymbol{\varphi}$ is the parameter indexing the dropout process; see equations (3.4) and (4.2). Thus the parameters of the sample distribution of the observed measurements, **θ***, 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.

Step-one: Estimation of a_0:

According to Pfeffermann and Sverchkov (1999), the following relationship holds:

$$E_s\left(w_i \middle| y_{i1}\right) = \frac{1}{E_p\left(\pi_i \middle| y_{i1}\right)}. \tag{5.1}$$

Thus, we can estimate a_0 by regressing $-\log(w_i)$ against $y_{i1}, x_{i11}, ..., x_{i1p}$, $i \in s$.

Step-two: Substitution of 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:

$$\begin{aligned} L_i\left(\boldsymbol{\beta}, \sigma^2, v_{jk}, \boldsymbol{\varphi}\right) &= \log f_s\left(\mathbf{y}_i \middle| \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, v_{jk}, \tilde{a}_0\right) \\ &= \log f^*_{s,d_i-1}\left\{\mathbf{y}_{i,d_i-1} \middle| \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, v_{jk}, \tilde{a}_0\right\} + \log\left\{\prod_{t=2}^{d_i-1}\left[1 - P_t\left(H_{it}, y_{it}; \boldsymbol{\varphi}\right)\right]\right\} \quad (5.2) \\ &+ \log P\left(D_i = d_i \middle| H_{id_i}, \boldsymbol{\beta}, \sigma^2, v_{jk}, \boldsymbol{\varphi}\right) \end{aligned}$$

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

$$L_{rs}\left(\boldsymbol{\beta}, \sigma^2, v_{jk}, \boldsymbol{\varphi}\right) = \sum_{i=1}^{n} L_i\left(\boldsymbol{\beta}, \sigma^2, v_{jk}, \boldsymbol{\varphi}\right) = L_1\left(\boldsymbol{\beta}, \sigma^2, v_{jk} \middle| \tilde{a}_0\right) + L_2\left(\boldsymbol{\varphi}\right) + L_3\left(\boldsymbol{\beta}, \sigma^2, v_{jk}, \boldsymbol{\varphi}\right), \tag{5.3}$$

where the three components of (5.3) correspond to the three terms of (5.2). The explicit form of $L_1\left(\boldsymbol{\beta}, \sigma^2, v_{jk} \middle| \tilde{a}_0\right)$ is obtained from equation (3.6), replacing *T* by $d_i - 1$. $L_2(\boldsymbol{\varphi})$ is determined by equation (4.1) The last term $L_3\left(\boldsymbol{\beta}, \sigma^2, v_{jk}, \boldsymbol{\varphi}\right)$ is determined by equation (4.2) which requires the distribution $f^*_{tp}\left(y_{it} \middle| H_{it}, \boldsymbol{\beta}, \mathbf{V}_0\right)$.

5.2. *Pseudo Likelihood Approach*

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

$$U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L_p(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}} \log L_{ip}\left(\boldsymbol{\beta}, \sigma^2, v_{jk}\right) \tag{5.4}$$

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

$$\begin{aligned} \frac{\partial L_{ws}\left(\boldsymbol{\beta}, \sigma^2, v_{jk}\right)}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} w_{i1} \frac{\partial \log f_p^*\left(y_{i1} \middle| \mathbf{x}_{i1}, \boldsymbol{\beta}, \sigma^2, v_{11}\right)}{\partial \boldsymbol{\theta}} + \\ \sum_{i=1}^{n} w_i^* \frac{\partial f_p^*\left(y_{i2}, y_{i3}, \ldots, y_{iT} \middle| y_{i1}, \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2, v_{jk}\right)}{\partial \boldsymbol{\theta}}, \end{aligned} \tag{5.5}$$

where we can take $w_i^* = w_i$ or $w_i^* = N/n$.

For more details, see Binder (1983) and Eideh and Nathan (2006).

6. 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 data, 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 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 Clarke and Chambers (1998), Clarke and Tate (1999).

In order to accommodate the differences between the assumptions of sections 2-5 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 $n_h(a,b)$ be the number of individuals with labour force flow (a,b), $a,b=1,2,3$ in household h and let $\omega(a,b)>0$ be the probability of an individual having labour force flow(a,b).

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. Let S_{11} denote the subset of households who responded in both quarters, i.e., the subset representing the longitudinal linked data on the same persons.

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 second column of Table 1 gives estimates from the unweighted data, obtained by maximizing the simple likelihood,

$$\hat{\omega}_U\left(a,b\right)=\frac{\sum_{h\in S_{11}} n_h\left(a,b\right)}{n_{11}};\ a,b=1,2,3, \qquad (6.1)$$

where $n_{11}=\sum_{h\in S_{11}}\sum_{a,b} n_h(a,b)$ is the total number of persons in the households of subset S_{11}

(2) Weighted method:

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

$$\hat{\omega}_h\left(a,b\right)=\frac{\sum_{h\in S_{11}} w_h n_h\left(a,b\right)}{\sum_{h\in S_{11}} w_h};\ a,b=1,2,3, \qquad (6.2)$$

where w_h is the longitudinal weight of household h.

(3) The sample likelihood method:

The sample likelihood was derived under the assumptions of the exponential model, for the household weights as a function of the labour force flow frequencies, defined by equation (3.4) and on the basis of the relationships between the population likelihood and that of the respondent sample, defined by equations (3.6) and (4.2).

The forth column in Table 1 gives the estimates (SMLE), based on the sample log-likelihood under the exponential model, which in this case is given by:

$$E\left(w_h^{-1} \left| \left\{n_h\left(a,b\right)\right\}, \left\{\alpha_h\left(a,b\right)\right\}\right.\right) = \exp\left(\alpha\left(0,0\right) + \sum_{a,b} \alpha\left(a,b\right) n_h\left(a,b\right)\right), \qquad (6.3)$$

where $\alpha\left(0,0\right), \alpha\left(1,1\right), ..., \alpha\left(3,3\right)$ are parameters to be estimated.

Table 1: Gross Flows Estimates (percentages)

Flow	Unweighted	Weighted	Exponential-SMLE
EE	70.62	69.32	69.78
EU	1.08	1.25	1.17
EN	1.53	1.71	1.61
UE	1.61	1.62	1.61
UU	3.78	4.41	4.16
UN	1.00	1.12	1.07
NE	1.40	1.29	1.35
NU	1.06	1.14	1.12
NN	17.92	18.12	18.13

The main findings from Table 1 are:

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

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, 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.

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 (2002), recommend similarly that weighting should be used to produce flows estimates that offer a considerable improvement in bias over unadjusted estimates. Although the

sample likelihood estimates cannot be shown to be better than the weighted estimators, their similarity to the weighted estimates indicate that they also are an improvement over the unweighted estimates.

7. Conclusions

In the empirical result, we introduce alternative method of obtaining weighted estimates of gross flows, taking into account informative nonresponse. The 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-based on the reciprocals of the adjusted calibrated weights. The proposed method is model based while the classical method is based on the adjusted weights. Thus we think that the new method is 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.

Initially we considered that the estimates of gross flows based on the response sample likelihood might explain the nonignorable 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 proposed method based on the response likelihood may give similar results.

8. References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51**, 279-292.

Chambers, R. L, Dorfman, A. and Sverchkov, M. (2003). Nonparametric regression with complex survey data. In: *Analysis of Survey Data*, Eds. R. Chambers and C. Skinner. New York: Wiley, pp. 151-173.

Clarke, P.S and Chambers, R.L (1998). Estimating labour force gross flows from surveys 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 for National Statistics.

Clarke, P.S and Tate, P.F. (2002). An application of non–ignorable non–response models for gross flows estimation in the British Labour Force Survey. *Australian and New Zealand Journal of Statistics* **4**, 413-425.

Diggle, P., and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis (Disc: p73-93). *Applied Statistics* **43**, 49-73.

Diggle, P. J., Liang, K. Y, and Zeger, S. L. (1994). *Analysis of Longitudinal Data*, Oxford: Science Publication.

Eideh, A.H. and Nathan, G. (2003). Model-based analysis of Labour Force Survey gross flow data under informative nonresponse. *Contributed Paper for the 54th Biennial Session of the International Statistical Institute*, August 13-20, Berlin, Germany.

Eideh, A. H. and Nathan, G. (2006). Fitting time series models for longitudinal survey data under informative sampling. *Journal of Statistical Planning and Inference* **136,** 3052-3069.

Feder, M.F, Nathan, G., and Pfeffermann, D. (2000). Time series multilevel modeling of Complex Survey Longitudinal Data. *Survey Methodology* **26**, 53-65.

Friedlander, D., Okun, B.S., Eisenbach, Z. and Elmakias, L.L. (2002). Immigration, social change and assimilation: educational attainment among Jewish ethnic groups in Israel. Research Reports, Jerusalem: *Israel Central Bureau of Statistics*.

Goldstein, H., Healy, M. J. R. and Rasbash, J. (1994) Multilevel time series models with applications to repeated measures data. *Statistics in Medicine* **13**, 1643-1655.

Herriot, R.A., and Kasprzyk, D. (1984). The survey of income and program participation. *American Statistical Association, Proceedings of the Social Statistics Section*, pp.107-116.

Ibrahim, J. G. , Chen, M. H. , and Lipsitz, S. R. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88**, 551-564.

Lipsitz, S. R., Ibrahim, J., and Molenberghs, G. (2000). Using a Box-Cox transformation in the analysis of longitudinal data With incomplete responses. *Applied Statistics* **439,** pp 287-296.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88** , 125-134.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90** , 1112-1121.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, New York: Chichester.

Little, R. J. A., and Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates, *Biometrics* **52** , 98-111

McCullagh, P. and Nelder, J. A. (1999). *Generalized linear models*. London: Chapman & Hall.

Nathan, G. (1999). A review of sample attrition and representativeness in three longitudinal surveys. *GSS methodology Series No. 13*. London: Office of National Statistics.

Mohamed, W. N., Diamond, I. and Smith, P. W. F. (1998). The determinants of infant mortality in Malaysia: A graphical chain modelling approach. *Journal of the Royal Statistical Society* **A161**, 349-366.

Nathan, G. and Eideh, A. H. (2004). L'analyse des données issues des enquêtes longitudinales sous un plan de sondage informatif. In**:** *Échantillonage et Méthodes d'Enquêtes*. Ed. P. Ardilly. Paris: Dunod., pp 227-240.

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

Pfeffermann, D., Moura, F. A. S. and Silva, N.P.L. (2001). Multilevel modeling under informative sampling. *Proceedings of the Invited Paper Sessions for the 53rd session of the International Statistical Institute*, pp. 505-532.

Pfeffermann, D. and Nathan, G. (2001). Imputation for wave nonresponse – existing methods and a time series approach. In: *Survey Nonresponse*. Chap. 28. Eds. R. Groves, D. Dillman, J. Eltinge, and R. Little. New-York: Wiley, pp. 417-429.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya* **61**, Ser. B, 66-186.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In: *Analysis of Survey Data*. Eds. R. Chambers and C. J. Skinner. New York: Wiley, pp. 175-195.

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

Schafer, J .L. (1995). *Analysis of incomplete multivariate data by simulation*. London: Chapman & Hall

Skinner, C. J. (1994). Sample models and weights**.** *American Statistical Association, Proceedings of the Section on Survey Research Methods,* 133-142.

Skinner, C.J., and Holmes, D. (2003). Random Effects Models for Longitudinal Data. In: *Analysis of Survey Data*. Eds. R. Chambers and C. Skinner. New York: Wiley, pp. 175-195.

Sverchkov, M. and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology* **30,** 79-92.

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

Troxel, A. B., Harrington, D. P., and Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics* **47,** 425-438.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society **B*** **52**, 21-50.

Zimmerman, D.L., and Nunez-Anton, V. (2000). Modeling nonstationary longitudinal data. *Biometrics* **56,** 699-705