The analysis of longitudinal studies

Gad Nathan
Hebrew University of Jerusalem

1. Introduction

Recently, economists, sociologists and other social scientists have become increasingly interested in longitudinal studies and their analysis, in their attempts to understand the dynamics of economic and social processes. Previously, single time cross-sectional surveys and their analysis formed the primary basis for empirical investigations in the social sciences. Longitudinal studies, in which the same units are investigated on several occasions, over extensive periods of time, are expensive undertakings and are complex operationally and methodologically. Even when longitudinal data are theoretically available, e.g., from panel surveys, such as the Labour Force Survey (LFS) in the UK and the Current Population Survey in the US, technical and methodological problems considerably reduce their usefulness for longitudinal analysis. In addition, their overall time span for a single unit is relatively short (about two years) and their panel design is primarily geared to increasing the efficiency of cross-sectional estimates, rather than that of the analysis of gross flows or developments over time.

However, in recent years, longitudinal studies have become of prime importance as a basis of empirical research in the social sciences. They are now being used increasingly for longitudinal analysis and in many cases longitudinal surveys are carefully designed in order to be able to derive from them sophisticated analyses of the long-term dynamics of social and economic processes. Thus the British Household Panel Survey, carried out on a continuous basis since 1991 by the ESRC Research Centre on Micro-Social change at Essex University, has as its aims “to further understanding of social and economic change at the individual and household level in Britain, and to identify, model and forecast such changes and their causes and consequences in relation to a range of socio-economic variables” – University of Essex (2006). For some insight on the various analytical uses of longitudinal analyses, in the social sciences see Solon (1989) and Heckman and Robb (1989).

In this chapter we consider model-based methods of analysis appropriate to the analysis of the dynamic aspects of longitudinal studies. It should be noted that the terms “longitudinal studies” and “panel surveys” are often used interchangeably (sometimes due to differences in US and British usage). In this chapter we consider longitudinal analysis as relating to any data collected for the same units over a series of time points (or even continuously), usually over a considerable length of time. Although the emphasis is on the analysis of data from sample surveys, we shall also consider methods of analysis developed or used for other types of data (e.g.,
administrative data, census data or experimental data), in as far as they can be applied also for the analysis of survey data. Other chapters in this volume deal with various related aspects of longitudinal studies and panel surveys. Thus, Kalton (chapter 5) deals with the design and analysis of panel surveys, primarily those with short periods of repetition and rotating panels, and their use for cross-sectional estimation and for estimation of change, under a design based approach. The chapter also deals with special issues of panel surveys, such as the effects of changing modes of collection and weighting adjustments for attrition and wave nonresponse. Steel (chapter 32) considers inference over time on the basis of repeated survey data, primarily under a design-based approach and time series methods. Finally Singh (chapter 37) describes some approaches and recent developments in the analysis of longitudinal categorical data, under a model-based approach, as well as joint modeling for cross-sectional and longitudinal analysis of categorical data.

In the following section we consider the various types of longitudinal studies and the problems they pose. Section 3 reviews general and specific models used for the analysis of longitudinal data, primarily those applicable to survey data. In section 4 we consider some model-based methods for the treatment of wave nonresponse, attrition, and misclassification errors. Finally, section 5 deals with the effects of complex and informative sample design on longitudinal analysis and their treatment for purposes of analysis.

2. Types and problems of longitudinal studies

Longitudinal data can be derived from a wide range of sources and by a variety of collection methods. In the following we shall examine the main modes by which longitudinal data are obtained and examine the possible ramifications of the methods used on the way in which the data can be analyzed.

The prevalent method of collection of data for a longitudinal survey is *prospective* measurement, in which a sample of respondents is followed forwards in time and data is collected on their current situation at a series of points in time. This is the method used in the important set of *birth cohort studies*, such as the UK 1970 British Cohort Study (BCS70). This was based on a sample of all births in Great Britain occurring during one week in 1970, followed up in four subsequent sweeps over a period of 20 years, in order to study the medical, physical, educational and social development of the cohort and to investigate the forces and patterns that shape their lives – Butler, Desotidou and Shepherd (1997). This pioneering project has now been replaced, since 2000, by the similarly designed Millennium Cohort Study (MCS) – Smith and Joshi (2002). The major disadvantage of the prospective method of collection is the long lead time required until a sufficient body of longitudinal data is available for analysis. Other practical problems involve the difficulties of follow-up for dynamic populations and the inherent cumulative attrition, with its implications for nonresponse bias – see Nathan (1999).

The *retrospective* method of measurement collects information at the current time on past events, based on recollection or records. This is the widely used method for case control studies, in which subjects are recruited according to their disease status and their past exposure to risk factors is examined. However it has been increasingly used also in longitudinal sample surveys, often in conjunction with prospective
measurement. Thus, in the sixth wave of the British Household Panel Survey of 1996 all respondents were asked on their family structure during childhood (i.e., whether they lived with one or both parents or other family members during childhood) – Francesconi (2005). While the retrospective method of collection overcomes the problem of attrition, it is associated with possible acute effects of response error, when it relies on memory. Several studies have indicated the serious problems associated with memory effects in retrospective studies. For instance, Kazemian and Farrington (2005) compare the validity of retrospective reports with that of prospective reports and official records on the age of onset for criminal offences and find that retrospective reporting is unsuitable for a wide range of research questions. Similarly Smith and Thomas (2003) find, by test-retest reliability methods, that the quality of long-term recall reports on migration histories may be poor, thought they do propose steps which can be taken to improve it.

Although the prospective and retrospective methods differ with respect to their non-sampling errors, standard methods for their longitudinal analysis are basically the same. Thus, a classic result of Prentice and Pyke (1979) shows that prospective and retrospective logistic models applied to case control data give equivalent results. Similar results for Bayesian analysis are obtained by Seaman and Richardson (2004). However, when observations are clustered, such as in case–control studies with covariate variables obtained from family members, Neuhaus, Scott, and Wild (2002) show that in some cases the prospective and retrospective analyses differ.

An important category of longitudinal studies is that of observational studies. These could be retrospective observations on covariate data, obtained from historical administrative or medical records, in order to supplement data obtained from a prospective longitudinal sample survey. For instance, in the Millennium Cohort Study, mentioned previously, interview reports on children were linked to birth register and hospital maternity records (after requesting informed consent - obtained for 92% of respondents), in order to investigate relationships between current status and birth data – see Tate et al. (2006). In other cases the observational study is carried prospectively on a sample of patients, for which case control studies are difficult or impossible to implement. In particular, this has been the favoured research mode for a wide range of longitudinal studies of the effects of therapeutic interventions and environmental factors on the progression of HIV infection, in which the natural histories of cohorts of those infected are observed over a length of time – see, for instance, Ko, Hogan and Mayer (2003). Since treatments in these studies are not randomly assigned and in order to overcome the problem of the confounding of treatment-response relationships by time-varying variables, specialized methods of analysis, such as those based on marginal structural models, intensity scores and inverse probability weighting are required – see, for instance, Gill and Robins (2001), Brumback et al. (2003) and Hogan and Lee (2004).

Another form of longitudinal study in which the effects of different treatments are studied is that of intervention studies. In this type of study an intervention for a medical or social process is initiated after the start of the longitudinal data collection, and subjects are selected for the new intervention on the basis of their previous measurements. Thus, in a study described by Lin and Hughes (1997) historical data on a marker for disease progression defines whether subjects are chosen to receive a new treatment. A similar longitudinal intervention study in the economic area is
described by Heckman and Robb (1985). They consider the analysis of the effect of training on earnings when enrollment into training is the outcome of a non-random selection process.

3. General models for analysis of longitudinal data

3.1. Repeated measures models and general estimating equations

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. To set ideas, consider the following models:

Let \( y_i \) be the value of the outcome random variable and \( x_i \) be a \( p \times 1 \) vector of fixed covariates for subject \( i \) at time \( t \). Let \( z_i \) be a \( q \times 1 \) vector of covariates associated with the random effect vector \( b_i \). Let \( u_i = E(y_i | b_i) \) and \( \mu_i = E(y_i) \) be the conditional and unconditional expectations of the outcome variable, respectively. Then the mixed GLM, under the subject-specific approach, is defined by:

\[
\begin{align*}
    h(u_i) &= x_i' \beta + z_i' b_i; \\
    \text{var}(y_i | b_i) &= g(u_i) \cdot \phi,
\end{align*}
\]  

(3.1)

where \( b_i \) are independently distributed \( F \), a mixture distribution, and the functions \( h \) and \( g \) are the link and variance functions, respectively. Under the population-average approach, the marginal expectation is similarly modeled by:

\[
\begin{align*}
    h^*(\mu_i) &= x_i' \beta^*; \\
    \text{var}(y_i) &= g^*(\mu_i) \cdot \phi
\end{align*}
\]

(3.2)

Estimation of the model parameters is obtained, under both models, by solving appropriate generalized estimation equations – see details in Zeger, Liang and Albert (1988), which includes an example of the analysis of longitudinal data from the Harvard Study of Air Pollution. Further extensions of these models to Markov transition models and examples of their application, primarily in the health sciences, may be found in Diggle, Liang, and Zeger (1994).

3.2. Multi-level models
Frequently longitudinal sample surveys deal with hierarchical populations, such as individuals within households or employees within establishments, for which multi-level modeling is appropriate. On the other hand, Goldstein, Healy and Rasbash (1994) consider the analysis of repeated measurements using a two-level hierarchical model, with individuals as second levels and the repeated measurements as the first levels. Thus, denoting $y_{it}$ as above, they propose the following two-level model:

$$y_{it} = \sum_{k} x_{it} \beta_k + \sum_{p} z_{pit} \epsilon_{2i}^{p} + \sum_{m} z_{mit} \epsilon_{1im0} ,$$

where the first term denotes fixed effects and the last two terms denote random effects at the higher level (individuals) and at the lower level (measurements), respectively. Assuming multivariate normality and standard assumptions on covariances, they obtain maximum likelihood estimates of the parameters by the use of an Iterative Generalized Least Squares algorithm. The results are extended to discrete first order and second-order autoregressive time models and to continuous time models. A small dataset of nine height measurements for each of a sample of boys over five years, with age (and its exponents) as the fixed effect, provides an example of the analysis.

Skinner and Holmes (2003) consider a random effects model, in which permanent individual random effects, $u_{it}$, are the higher level effects and transitory random effects, $\nu_{it}$, are the lower level effects, which may be correlated over time. As an example they set up a basic hierarchical model for the log earnings, $y_{it}$ of individual $i$ at wave $t$, from data from the British Household Panel Survey (BHPS), as:

$$y_{it} = \beta + u_{it} = 1, \ldots, T$$

where the transitory random effects follow a first-order autoregressive model AR(1):

$$\nu_{it} = \rho \nu_{i,t-1} + \epsilon_{it} , \quad t=1,\ldots,T$$

The random variables $u_{i}$ and $\epsilon_{it}$ are assumed to be mutually independent with $E(u_{i}) = E(\epsilon_{it}) = 0$ and $\text{var}(u_{i}) = \sigma_u^2$; $\text{var}(\epsilon_{it}) = \sigma_e^2$.

Fitting of the models and estimation of the parameters can be carried out by two alternative methods. The first is a covariance structure approach, in which the observations on the $T$ waves are treated as a multivariate outcome with individuals as 'single level' units. The second approach treats the data as hierarchical, with the lower level units as the waves, $t=1,\ldots,T$, and the higher level units as the individuals, $i$.

Feder, Nathan and Pfeffermann (2000) consider a model that encompasses both the hierarchical nature of many human populations and the time series relationships between repeated measurements and random effects of higher-level
groups. Higher-level groups will be called 'households' and lower level units 'individuals'. The proposed model combines standard multi-level mixed linear models (Goldstein, 1986, 1995), operating at given points in time by a state-space model that represents the time series relationships of the random group effects and the individual measurements. Basic notation and assumptions are as follows:

Let \( y_{ht} \) define the value of the response variable at time \( t = 1, \ldots, T \), for individual \( j = 1, \ldots, n_h \), belonging to household \( h = 1, \ldots, N \). The measurements \( y_{ht} \) are assumed to follow the hierarchical two level linear model:

\[
\begin{align*}
\begin{align*}
    y_{ht} &= x_{ht}^\prime b_t + z_{ht}^\prime v_t + z_{ht}^\prime u_{ht} + e_{ht},
\end{align*}
\end{align*}
\]

(3.5)

where \( x_{ht} \) is a \( p \)-dimensional vector of individual level explanatory variables values; \( z_{ht} \) is a \( q \)-dimensional vector of household level explanatory variables; \( b_t \) and \( v_t \) are fixed vector coefficients of appropriate orders; \( u_{ht} \) is a \((q \times 1)\) vector of household level random effects and \( e_{ht} \) is an individual level random error. The random household effects represent specific household characteristics not represented by the fixed effects.

The individual and household level random errors are assumed to follow independent first order autoregressive models,

\[
\begin{align*}
\begin{align*}
    u_{ht} &= A u_{ht-1} + d_u; & d_u &\sim N(0, D),
\end{align*}
\end{align*}
\]

(3.6)

\[
\begin{align*}
\begin{align*}
    e_{ht} &= \rho e_{ht-1} + e_{ht}; & e_{ht} &\sim N(0, \sigma_e^2),
\end{align*}
\end{align*}
\]

(3.7)

For convenience, \( A \) and \( D \) are assumed to be diagonal, implying independence of the random group level effects. We also assume \(|A| < 1\) and \(|\rho| < 1\) to ensure stationarity. It should be noted that more elaborate models can be considered, provided one has a sufficiently large number of observations per unit. It follows from (3.6) and (3.7) that for a given time \( t \), the marginal distributions are:

\[
\begin{align*}
\begin{align*}
    u_{ht} &\sim N(0, D^\prime); & D^\prime &= (I - A^\prime)^-1 D,
\end{align*}
\end{align*}
\]

(3.8)

\[
\begin{align*}
\begin{align*}
    e_{ht} &\sim N(0, \sigma_e^2); & \sigma_e^2 &= (1 - \rho^2)^{-1} \sigma_e^2.
\end{align*}
\end{align*}
\]

(3.9)

Thus, the models operating at various time points are standard multilevel models with fixed variances for the random first and second level effects.

Although the likelihood of this model is easily constructed by employing the time series properties of the combined model, the large number of parameters to be estimated results in unstable estimates, if direct maximization of the likelihood is employed. Rather a two-stage estimation procedure is proposed. At the first stage, a separate two-level model is fitted for each time point, yielding estimates for the
fixed effects and for the variances. At the second stage, the time series likelihood is maximized to yield estimates of the time series model parameters.

The methods are illustrated by a simulation study and an empirical application to data from the Israeli Labour Force Survey, with weekly hours worked as the dependent variable, while years of education and gender serve as individual level explanatory variables and the number of employed persons in the household as a household level explanatory variable.

3.3. Other methods of analysis

Path analysis 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. Its generalization to modelling longitudinal data has been primarily by means of Graphical Chain Modelling (GCM) and Structural Equation Modelling (SEM). Both approaches provide pictorial representations of the association between variables which are ordered, usually temporally, with the aim of identifying the direct and indirect effects of one variable on another. While the GCM approach builds up a model for the complete system by fitting a sequence of sub-models, the SEM approach specifies a single model for the complete system of variables being studied.

The GCM approach is based on the construction of a causal diagram which represents the investigator’s understanding of the major causal influences among the measurable quantities involved. A basic conditional independence graph is constructed in order to characterize the conditional independence structure of the data. Each vertex of the graph represents a variable and two vertices are connected if there is a direct association between the variables, whereas unconnected vertices represent variables that are conditionally independent, given all the other variables. The graphs may be used to formulate research hypotheses about indirect relations in an association structure, under the assumption that the set of direct relations is sufficient to understand all associations in the system and that it cannot be further reduced without destroying such association. For further details on how graphical chain models help to identify analogies and equivalences between different models and to provide a unifying concept for many statistical techniques employed in the analysis of longitudinal data, see Wermuth and Lauritzen (1990). For an interesting example of the application of graphical chain modelling to the study of the determinants of neonatal and post-neonatal mortality in Malaysia, see Mohamed, Diamond and Smith (1998). The method allows both the examination of the effects of direct association of each determinant on mortality and the pathways by which intermediate socio-economic determinants affect mortality.

The SEM approach extends standard regression models to include multiple outcomes, sometimes called endogenous variables, and unobservable latent variables. The basic structural model is a set of regression equations relating each endogenous variable with other endogenous variables and with exogenous variables or covariates. A second component of the SEM is a measurement model, which relates observed study variables to unobservable underlying constructs, represented by one or more latent variables. For a thorough review of the SEM field, its relationship to latent variabel models for multivariate outcomes and to
measurement theory, as well as applications to environmental epidemiology, see Sánchez et al. (2005). An interesting application of structural equation modelling to longitudinal data from the UK National Child Development Study, which studies simultaneously six different pathways hypothesized to link education and health to other variables, is given by Chandola, et al. (2006). They find by applying SEM methods that the association can be explained by a combination of mechanisms, such as adolescent and adult health behaviours and adult and parental social class.

Among a variety of other models used for the analysis of longitudinal data, the role of Antedependence Models in dealing with nonstationarity deserves special attention. The idea of antedependence, first formulated by Gabriel (1962), relates to a set of ordered variables, such as longitudinal observations, which are defined as being $s$-th order antedependent if each variable, given at least $s$ immediate antecedent variables, is independent of all other preceding variables. Núñez-Antón and Zimmermann (2000) consider unstructured and structured antedependence models for longitudinal data. The unstructured normal model is defined by:

$$y_j = x_j' \beta + \varepsilon_j \text{ and } y_j = x_j' \beta + \sum_{i=1}^{j} \phi_{jk} (y_{j-i} - x_{j-i}' \beta) + \varepsilon_j; \quad (j = 2, \ldots, n),$$

where $s^* = \min(s, j-1)$, $\varepsilon_j$ are independent normal random variables with mean zero and possibly time-dependent variances, $\sigma_j^2 > 0$, and the $\phi_{jk}$'s are unrestricted parameters. The model is unstructured in the sense that the $(s+1)(2n-s)/2$ parameters $\{\phi_{jk}\}$ and $\{\sigma_j^2\}$ cannot be expressed as functions of a smaller set of parameters.

Structured antedependence models follow the same basic model above, but relationships are assumed between the parameters, resulting in more parsimonious models. An example is a model in which correlations over the same time lags are equal and are just monotonic functions of the time lag. Núñez-Antón and Zimmermann (2000) use several empirical data sets to compare structured and unstructured antedependence models with unstructured covariance models, ARIMA models and random coefficient models.

4. Treatment of nonresponse

The problems posed by nonresponse in longitudinal surveys have much in common with those occurring in cross-sectional surveys. However, there are some special aspects of nonresponse in longitudinal data, which must be considered. On the one hand, the fact that the same individuals or households are repeatedly requested to provide information on repeated occasions obviously leads to attrition and wave nonresponse, due to fatigue and to difficulties in tracing sample units which are often highly mobile, see for example Nathan (1999). On the other hand, the existence of observations for some points in time for the same unit suggests that this information can assist in dealing efficiently with the effects of nonresponse, by considering plausible relationships over time between individual measurements. In the following, we focus on the treatment of missing data resulting from wave nonresponse, where data are available for some points in time and missing for others, rather than complete nonresponse, which can be dealt with similarly to the ways use for dealing with nonresponse in cross-sectional surveys, see for instance Little (1995). Different patterns of wave nonresponse to be considered are attrition (no observations from
some time point onwards), missing for a single time or for a continuous period and intermittent dropout. The relationships between the missing data mechanism and the missing and observed data need to be specified. An important distinction is between the mechanisms of missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) or informative missingness - Little and Rubin (1987).

The design-based treatment of wave nonresponse in panel surveys has been addressed in papers by Kalton (1986) and Lepkowski (1989) and by Kalton in Chapter 5 of the present volume. The methods proposed use imputation and weighting based on regression models, incorporating known auxiliary data, including response to other waves, and taking into account cross-sectional and longitudinal interrelationships.

Model-based treatment of nonresponse in longitudinal data has been considered primarily in the context of experimental sciences applications. Thus Diggle and Kenward (1994) propose a modeling framework for longitudinal data with informative dropouts, which explicitly considers MCAR and MAR dropout as sub-models. Under a general multivariate normal model, they specify a logistic regression model for the dropout process, which allows dependence of the drop-out probability on missing observations. The model parameters are estimated by maximum likelihood and examples are given of applications to data from milk protein trials, for milk yields and from multicentre clinical trials in the study of depression. Rotnizky, Robins and Scharfstein (1998) consider the use of semiparametric regression for the treatment of informative nonresponse. They propose a class of augmented inverse probability of response weighted estimators, which are consistent and asymptotically normal under parametric modeling of the response probabilities. Their estimation procedure can be viewed as an extension of the GEE method that allows for informative nonresponse.

In the sample survey context, Skinner and Holmes (2003) consider the effects of nonresponse in a longitudinal survey, under the models described in section 3 (eqns. 3.3-3.4), by modifying their estimator of covariance so that it is based on all `attrition samples, s, those responding until and including time t. However this does not deal with the problem of informative nonresponse. Miller et al. (2001) develop a method for analyzing categorical outcomes obtained from longitudinal survey samples, with outcomes subject to multiple-cause nonresponse, within the framework of weighted generalized estimating equations. They assume a model that combines different multivariate link functions to permit fitting Markov models to an outcome with categories represented by a mixture of ordinal success states and multiple failure states. They extend the missing data approach of Rotnizky, Robins and Scharfstein (1998) to the use of multiple-logit models, in order to model the probability of multiple reasons for missing success or failure outcomes, assuming that the probability of nonresponse depends only on observed responses and covariates specified in the missing data. Taylor series and jackknife variance estimators are developed for parameters estimated from these models and are presented within the context of adjusting for survey considerations and multiple-cause nonresponse. The results are applied to disability data obtained from the US Longitudinal Study of Aging (LSOA). A similar approach is proposed by Gong, Little, and Ragmunathan (2003)

Comment: Graham – please comment.
Pfeffermann and Nathan (2001) use the time series structures with hierarchical modeling, described in section 3 (eqns. 3.5-3.9), to deal with informative nonresponse in longitudinal surveys. They consider two methods based on these models, an augmented regression method and one based on a state-space model. The augmented regression prediction extends standard regression prediction by adding a correction term that takes into account the existing correlations between the observed and the missing data, so that imputation of missing data is based on all observations for all the time periods. The state-space method is based on the formulation of the model (3.5)-(3.9) in a state-space form, with appropriate observation and transition equations. Under the model (with known parameters), the random components can be predicted, either by application of the Kalman filter, if only current and past observations are available, or by an appropriate smoothing filter, if data for subsequent time periods are known. Estimation of the unknown model parameters is obtained by iterative generalized least squares for the augmented regression prediction and by the method of scoring for the state-space method. A simulation study and an empirical example, based on Israeli Labor Force data, compare the performances of the proposed methods favorably with those of conventional imputation methods, such as mean imputation (within homogenous groups), nearest neighbor imputation and simple regression imputation.

5. The effects of informative sample design on longitudinal analysis

Standard analysis of longitudinal survey data often fails to account for the complex nature of the sampling design, such as the use of unequal selection probabilities, clustering, post-stratification and other kinds of weighting used for the treatment of nonresponse. Thus it does not incorporate all the design variables in the analysis model, either because there may be too many of them or because they are not of substantive interest. However, if the sampling design is informative, in the sense that the outcome variable is correlated with design variables not included in the model, even after conditioning on the model covariates, standard estimates of the model parameters can be severely biased, leading possibly to false inference.

Eideh and Nathan (2006) fit time series models and, in particular, an autoregressive model of order one for longitudinal survey data, when the sampling design is informative. This is done by extending recent work defining the sample distribution under informative sampling for cross-sectional data to longitudinal data. Sample 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 to 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 repeated survey data refers to the superpopulation model. When the sample selection probabilities depend on the values of the model response variable, 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 (model) distribution under informative sampling that consists of approximating the parametric distribution of the sample measurements for given population distributions and first-order sample selection probabilities. The (marginal) sample distribution is defined as the conditional distribution, given that unit \( i \) is in the
sample, i.e., as the conditional distribution of the observed data. By application of Bayes’ theorem, they obtain the following relationship between the sample distribution, \( f_s(y_i | \theta) \), and the population distribution, \( f_p(y_i | \theta) \):

\[
f_s(y_i | \theta) = \frac{E_p(\pi_i | \theta, y_i)}{E_p(\pi_i | \theta)} f_p(y_i | \theta),
\]

where \( \pi_i = \Pr(i \in s | y_i) \) is the inclusion probability, assumed to depend on \( y_i \). Under informative sampling, i.e., when \( E_p(\pi_i | \theta, y_i) \neq E_p(\pi_i | \theta) \), this distribution is different from the corresponding population distribution. For further details see also, Pfeffermann and Sverchkov (Chapter 38). The extension of these results to study the case of longitudinal panel sample observations under informative sampling is carried out by Eideh and Nathan (2006), as follows.

Under previously defined notation, assume that the observed measurements, \( y_{it} \), follow the first order AR model; that is:

\[
y_{it} - \mu = \phi(y_{i,t-1} - \mu) + \epsilon_{it}, \quad i = 1,...,N; t = 1,...,T,
\]

where the errors \{\( \epsilon_{it} \)\} are normally distributed with zero mean and variance \( \sigma^2 \), and \( |\phi|<1 \), and that the errors \( \epsilon_{it} \) pertaining to the same subject are independent. The sample is assumed to be a panel sample selected at time \( t = 1 \) and all units remain in the sample till time \( t = T \). Then, it is intuitively reasonable to assume that the first order inclusion probabilities depend on the population values of the response variable at the first occasion only, \( y_{i1} \), and, possibly, on the values of a design variable, \( z_i \), used for the sample selection, but not included in the working model under consideration, that is, \( \pi_i = \Pr(i \in s | y_{i1}, z_i), i = 1,...,N \) depend only on \( y_{i1} \) and \( z_i \). Pfeffermann, Krieger, and Rinott (1998) propose two alternative approximation models for this population conditional expectation;

(a) **Exponential inclusion probability model:** \( E_p(\pi_i | y_{i1}) = \exp(a_0 + a_1y_{i1}) \).

Under this model, the sample log likelihood function is given by:

\[
l_s(\mu, \phi, \sigma^2, a_1) = -\frac{nT}{2} \log(\sigma^2) + \frac{n}{2} \log(1 - \varphi^2) - \frac{1 - \varphi^2}{\sigma^2} \sum_{i=1}^n y_{i1} - \mu - \frac{a_1 \sigma^2}{1 - \varphi^2} \]

\[
+ \sum_{i=1}^n \left[ -\frac{1}{2\sigma^2} \sum_{r=2}^T \{y_{ir} - \mu - \phi(y_{i,r-1} - \mu)\}^2 \right]
\]

This implies that the sample pdf belongs to the same family as the population pdf but differs only in the mean of \( y_{i1} \), which changes from \( \mu \) to \( \mu + a_1 \sigma^2 / (1 - \varphi^2) \).

(b) **Linear inclusion probability model:** \( E_p(\pi_i | y_{i1}) = b_0 + b_1 y_{i1} \).
Under this approximation the sample log likelihood function is given by:

\[
l_i(\mu, \varphi, \sigma^2, b_0, b_1) = -n \log(b_0 + b_1 \mu) - \frac{nT}{2} \log(\sigma^2) + \frac{\mu}{2} \log(1 - \varphi^2) - \frac{1 - \varphi^2}{\sigma^2} \sum_{i=1}^{T} (y_{ni} - \mu)^2
\]

\[
+ \sum_{i=1}^{T} \left[ -\frac{1}{2\sigma^2} \sum_{i=2}^{T} (y_{ni} - \mu - \varphi(y_{ni-1} - \mu))^2 \right]
\]

Thus, in this case, the effect of informativeness on the pdf is by a multiplicative factor.

The methods of estimation proposed for cross-sectional data, such as the two-step method (Pfeffermann, Krieger and Rinott, 1998, and Pfeffermann and Sverchkov, 1999) and the pseudo likelihood method (Binder, 1996, and Skinner, 1989) can be extended to longitudinal survey data. Eideh and Nathan (2006) propose a two-step method of estimation, and two versions of the pseudo likelihood method for the estimation of the unknown parameters of the above population models. A simulation study shows that the estimators based on the sample distribution differ from those obtained under the assumption that the sample design is not informative, but that their performances are relatively robust to the choice of model.

Finally the above results can be extended to incorporate both the effects of informative sample design and those of informative nonresponse. This can be done by assuming a model for the response propensity, such as the logistic model for informative dropout proposed by Diggle and Kenward (1994). By considering both the dropout process and the informative sample design, the joint sample distribution for the incomplete sequence of observations can be obtained and its parameters can be estimated by the methods proposed in Eideh and Nathan (2006).

6. References


