Estimation for Two-Phase Panel Surveys

Jason Legg, Wayne A. Fuller, and Sarah M. Nusser
Center for Survey Statistics and Methodology
Iowa State University

Abstract

Two-phase panel surveys are conducted to study trends over time. The outcome of these studies is often a dataset containing characteristic values and weights for a set of observations. Replication weights are often included in the dataset to allow variance estimation for nonlinear functions. We propose a general linear model as a basis for estimating means and totals at each survey time point. A consistent replication variance estimator is provided as well as a central limit theorem for use in constructing confidence intervals for functions of means or totals.

KEY WORDS: longitudinal, panel survey, two-phase sampling

Introduction

Longitudinal surveys where some, or all, of the units are revisited are used to provide estimates of change. Correlations from repeated observations on the same unit may be used to form estimators that are superior to cross-sectional estimators. We consider estimation of time point means from longitudinal surveys composed of a fixed number panels, where units in a panel are observed with the same observation pattern. Such surveys can be represented as two-phase samples.

In two-phase sampling, a large first-phase sample, A_1, is selected using a design $p_1(\bullet)$ with inclusion probabilities $Pr(i \in A_1) = \pi_{1i}$. Traditionally, an inexpensive or easy to observe auxiliary variable vector, $\boldsymbol{x}$, is observed in A_1. A second-phase sample A_2 is selected from A_1 with a design $p_{2|1}(\bullet)$ with conditional inclusion probabilities $Pr(i \in A_2 | i \in A_1) = \pi_{2i|1i}$. Often, the variables of interest $\boldsymbol{y}$ are observed in A_2. The information from $\boldsymbol{x}$ is used in $p_{2|1}(\bullet)$, in an estimator for the mean or total of $\boldsymbol{y}$, or in both. A common estimator is the two-phase regression estimator (Särndal et al 1992).

Longitudinal samples with a fixed number of panels can be viewed as two-phase samples with many second-phase samples. The first-phase sample is composed of all units that will be observed at some time. The second-phase samples, A_{2p} for $p = 1, 2, \ldots, P$, are disjoint panels of units. All units in A_{2p} are observed at times determined by a longitudinal observation scheme. The second-phase sample design may be thought of as a procedure to partition the first-phase sample into disjoint panels. The selection of A_1 can be done implicitly by selecting panels directly from the population.

Let

$$\overline{y}_{2pt} = \left[\sum_{i \in A_{2p}} \pi_{1i}^{-1} \pi_{2i|1i}^{-1} \right]^{-1} \sum_{i \in A_{2p}} \pi_{1i}^{-1} \pi_{2i|1i}^{-1} y_{it} \quad (1)$$

be the second-phase mean from panel p of characteristic y observed at time t. When more than one panel is observed at time t, we have more than one estimator of the population mean at time t. Panel means from the same panel observed at different times are correlated since they contain the same sampled units. We propose an estimator for population time means that incorporates the correlation structure of panel means and give a central limit theorem for the estimator. We close with suggestions on how to utilize the estimator in creating an analysis dataset.

Estimated Generalized Least Squares Estimator

We write the cell-mean model for panel means as

$$\overline{\boldsymbol{y}}_{\boldsymbol{2}} = \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{e}, \quad (2)$$

where $\overline{\boldsymbol{y}}_{\boldsymbol{2}}$ is the vector of second-phase panel means $\overline{y}_{2pt}$, $\boldsymbol{X}$ has elements

$$x_{pt,j} = \begin{cases} 1 & \text{if } t = j \\ 0 & otherwise \end{cases} \quad (3)$$

that link $\overline{y}_{2pt}$ to the corresponding population mean μ_t, $\boldsymbol{\mu}$ is the vector of population means , and $\boldsymbol{e}$ is a vector of errors due to sampling and measurement. Assume $E(\boldsymbol{e}) = \boldsymbol{0}$ and $Var(\boldsymbol{e}|\mathcal{F}_N) = \boldsymbol{V}$,

Table 1: Toy Example

Panel	Time		
	1	2	3
1	X	X	X
2		X	
3			X

where $\{\mathcal{F}_N\}$ is a sequence of finite populations. Assume uncorrelated panels, $Corr(e_{pt}, e_{qu}|\mathcal{F}_N) = 0$ for $p \neq q$, and assume $Corr(e_{pt}, e_{p,t+j}|\mathcal{F}_N) = \rho(j)\ \forall t$. Assume $\boldsymbol{V}$ is a function of a fixed number of parameters and of stratum sample sizes. Assume

$$\overline{y}_{pt} - \mu_t|\mathcal{F}_N = O_p(n^{-1/2}). \tag{4}$$

Consider the toy example of an observation scheme of three periods depicted in Table 1. The components of model (2) for the representation of Table 1 are

$$\overline{\boldsymbol{y}}_{\boldsymbol{2}} = \begin{pmatrix} \overline{y}_{2,11} \\ \overline{y}_{2,12} \\ \overline{y}_{2,13} \\ \overline{y}_{2,22} \\ \overline{y}_{2,33} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{5}$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \tag{6}$$

and assuming equal panel sizes and equal stratum compositions,

$$\boldsymbol{V} \propto \begin{pmatrix} 1 & \rho(1) & \rho(2) & 0 & 0 \\ \rho(1) & 1 & \rho(1) & 0 & 0 \\ \rho(2) & \rho(1) & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{7}$$

where $\rho(h)$ is the correlation of panel means at lag h.

The generalized least squares estimator (GLSE) for $\boldsymbol{\mu}$,

$$\breve{\boldsymbol{\mu}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}\overline{\boldsymbol{y}}_{\boldsymbol{2}} \tag{8}$$

can be computed when $\boldsymbol{V}$ is known. Skinner and Holmes (2003) describe uses of generalized least squares in longitudinal survey analysis.

In many applications, the parameters in $\boldsymbol{V}$ will need to be estimated. Let $\hat{\boldsymbol{V}}$ be an estimator of $\boldsymbol{V}$ with error that is $O_p(n^{-1/2})$. The estimated generalized least squares estimator (EGLSE) for $\boldsymbol{\mu}$ is

$$\hat{\boldsymbol{\mu}} = \left(\boldsymbol{X}'\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}\hat{\boldsymbol{V}}^{-1}\overline{\boldsymbol{y}}_{\boldsymbol{2}} \tag{9}$$

Many methods exist for estimating the covariance matrix of $\overline{\boldsymbol{y}}_{\boldsymbol{2}}$. In our applications, we fit a model to the empirical correlations. We use nonlinear least squares to compute the estimated correlation parameters. Fuller (1987) provides theory for using nonlinear least squares to estimate variance parameters used in an EGLSE. By our assumptions,

$$\hat{\boldsymbol{\mu}} - \breve{\boldsymbol{\mu}}|\mathcal{F}_N = O_p(n^{-1}) \tag{10}$$

and

$$\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|\mathcal{F}_N = O_p(n^{-1/2}). \tag{11}$$

The GLSE is superior to direct time mean estimators of the form

$$\overline{y}_{P_t^\star} = \left[\sum_{i\in P_t^\star} \pi_{1i}^{-1}\pi_{2i|1i}^{-1}\right]^{-1} \sum_{i\in P_t^\star} \pi_{1i}^{-1}\pi_{2i|1i}^{-1} y_{it}, \tag{12}$$

where $P_t^\star$ is the collection of panels observed at time t. For the GLSE, $Var(\breve{\boldsymbol{\mu}}) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}$. The EGLSE variance can be estimated by replacing $\boldsymbol{V}$ with the consistent estimator $\hat{\boldsymbol{V}}$. Given a consistent first-phase replication variance for time means, a consistent replication variance estimator for the EGLSE has been proposed (Legg, Fuller, and Nusser 2005).

Central Limit Theorem

To facilitate the construction of confidence intervals for time means using the EGLSE, we provide conditions on the population and sample designs that give asymptotic normality of the EGLSE. Central limit theorems for first-phase means have been given for Poisson sampling, simple random sampling, and stratified random sampling under mild assumptions. Conditional on $(A_1, \mathcal{F}_N)$ the same theorems may be applied to second-phase means for conditional asymptotic normality. The following lemma adapted from Schenker and Welsh (1988) provides conditions for combining first and second-phase results.

Lemma. Let $\{\mathcal{F}_N\}$ be a sequence of finite populations and let $\boldsymbol{\theta}_{\boldsymbol{N}}$ be a function in $\mathcal{R}^k$ of the elements in $\mathcal{F}_N$ such that

$$N^{1/2}(\boldsymbol{\theta}_{\boldsymbol{N}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_k(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{11}}). \tag{13}$$

Let a design, an estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{N}}$, and a sequence of conditional variance matrices $\boldsymbol{V}_{\boldsymbol{22,N}}$ be such that

$$N^{1/2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{N}} - \boldsymbol{\theta}_{\boldsymbol{N}})|\mathcal{F}_N \xrightarrow{\mathcal{L}} N_k(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{22}})\ a.s. \tag{14}$$

and

$$\lim_{N\to\infty} \boldsymbol{V_{22,N}} = \boldsymbol{V_{22}} \ a.s., \tag{15}$$

where $\boldsymbol{V_{11}} + \boldsymbol{V_{22,N}}$ is positive definite for all N. Then

$$N^{1/2}(\boldsymbol{V_{11}} + \boldsymbol{V_{22,N}})^{-1/2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{N}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N_k(\boldsymbol{0}, \boldsymbol{I_k}), \tag{16}$$

where $\boldsymbol{I_k}$ is the $k \times k$ identity matrix.

We apply the lemma for fixed-rate second phase sampling.

Theorem. Let $\{(\boldsymbol{y_i}, \boldsymbol{x_i})\}$ be a sequence of independent and identically distributed random vectors, where $\boldsymbol{y_i}$ is a vector of response variables of length T with fifth moments and $\boldsymbol{x_i}$ is a vector of second-phase stratum indicators of length $T \times G$. Let $\{\mathcal{F}_{N_k}, A_{1k}\}_{k=1}^{\infty}$ be a sequence of populations and first-phase samples, where A_{1k} is a sample of size n_{1k} from $\mathcal{F}_{N_k}$. Assume $\mathcal{F}_{N_k} \subset \mathcal{F}_{N_{k+1}}$ and $A_{1k} \subset A_{1,k+1}$, where $\mathcal{F}_{N_k}$ contains the first N_k elements of $\{(\boldsymbol{y_i}, \boldsymbol{x_i})\}$. Assume that each finite population $\mathcal{F}_{N_k}$ is divided into G strata. Let N_{gk} be the size of stratum g in $\mathcal{F}_{N_k}$. Let $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}$ be the EGLSE defined in (9) for the k^{th} sample. Assume $\{\mathcal{F}_{N_k}, A_{1k}\}$ is such that the first-phase mean vector satisfies,

$$V(\overline{\boldsymbol{y}}_{\boldsymbol{1k}}|\mathcal{F}_{N_k})^{-1/2}(\overline{\boldsymbol{y}}_{\boldsymbol{1k}} - \boldsymbol{\mu}_{\boldsymbol{N_k}})|\mathcal{F}_{N_k} \xrightarrow{\mathcal{L}} N(\boldsymbol{0}, \boldsymbol{I_T}) \ a.s., \tag{17}$$

where $\overline{\boldsymbol{y}}_{\boldsymbol{1k}}$ is the vector of first-phase weighted means with elements $\overline{y}_{1kt}$ for $t = 1, 2, \ldots, T$, $\boldsymbol{\mu}_{\boldsymbol{N_k}}$ is the vector of finite population means, and

$$n_{1k} V(\overline{\boldsymbol{y}}_{\boldsymbol{1k}}|\mathcal{F}_{N_k})\boldsymbol{\Sigma}_{\boldsymbol{1}}^{-1} \longrightarrow \boldsymbol{I_T} \ a.s. \tag{18}$$

for some positive definite matrix $\boldsymbol{\Sigma_1}$. Assume the first-phase design is such that

$$\begin{array}{l} \lim_{k\to\infty} N_k^{-1} \sum_{i\in A_{1k}} \pi_{1i}^{-1}(1, \boldsymbol{x_i'}, \boldsymbol{y_i'}, vech(\boldsymbol{y_i}\boldsymbol{y_i'})')' \\ (1, \boldsymbol{x_i'}, \boldsymbol{y_i'}, vech(\boldsymbol{y_i}\boldsymbol{y_i'})') = \boldsymbol{B} \ a.s., \end{array}, \tag{19}$$

where $\boldsymbol{B}$ is a matrix of constants. The notation $vech(\boldsymbol{C})$, where $\boldsymbol{C} = \{c_{ij}\}$ is a $T \times T$ matrix, denotes the vector $(c_{11}, \ldots, c_{1T}, c_{22}, \ldots, c_{2T}, \ldots, c_{TT})'$. Assume a sequence of first-phase inclusion probabilities $\pi_{11}, \pi_{12}, \ldots$, satisfying

$$K_L < n_{1k}^{-1} N_k \pi_{1i} < K_M \tag{20}$$

for positive K_L and K_M. Let the second-phase samples be A_{2pk} for $p = 1, 2, \ldots, P$ where the A_{2pk} are mutually exclusive, and let the sample size be n_{2pgk} for stratum g in panel p. Let the second-phase sampling rates, $n_{2pgk} n_{1gk}^{-1}$, be fixed and constant for each p, g pair, where n_{1gk} is the sample size in stratum g of A_{1k}. Then

$$\boldsymbol{\Sigma}_{\boldsymbol{ck}}^{-1/2}(\hat{\boldsymbol{\mu}}_{\boldsymbol{k}} - \boldsymbol{\mu}_{\boldsymbol{N_k}}|\mathcal{F}_{N_k}) \xrightarrow{\mathcal{L}} N(\boldsymbol{0}, \boldsymbol{I_T}) \tag{21}$$

where

$$\boldsymbol{\Sigma_{ck}} = \boldsymbol{\Lambda}(\boldsymbol{X}[V(\overline{\boldsymbol{y}}_{\boldsymbol{1k}}|\mathcal{F}_{N_k})]\boldsymbol{X'} + \boldsymbol{\Sigma_{2|1,k}})\boldsymbol{\Lambda'}, \tag{22}$$

$\boldsymbol{X}$ is the model matrix from the cell-mean model, and $\boldsymbol{\Lambda}$ is the probability limit of

$$\hat{\boldsymbol{\Lambda}} = \left(\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}. \tag{23}$$

Denote the second-phase panel p mean estimator of μ_{N_kt} by $\overline{y}_{2pkt}$. The components of $\boldsymbol{\Sigma_{2|1,k}}$ are

$$\begin{array}{l} Cov(\overline{y}_{2qkt_1} - \overline{y}_{1kt_2}, \overline{y}_{2ukt_2} - \overline{y}_{1kt_2}|\mathcal{F}_N) \\ = E\left[\sum_{g=1}^{G} n_{1gk}^2 (n_{2pgk}^{-1} - n_{1gk}^{-1}) S_{1kg,t_1,t_2}|\mathcal{F}_{N_k}\right] \end{array} \tag{24}$$

for $q = u$ and

$$\begin{array}{l} Cov(\overline{y}_{2qkt_1} - \overline{y}_{1kt_2}, \overline{y}_{2ukt_2} - \overline{y}_{1kt_2}|\mathcal{F}_N) \\ = E\left[\sum_{g=1}^{G} -n_{1gk}^2 N_{gk}^{-1} S_{1kgt_1,t_2}|\mathcal{F}_{N_k}\right] \end{array} \tag{25}$$

for $q \neq u$, where

$$\begin{array}{l} S_{1kgt_1,t_2} = (n_{1gk} - 1)^{-1} \\ \times \sum_{i\in A_{1gk}} (w_{1ki} y_{it_1} - \tilde{y}_{1k\pi,gt_1})(w_{1ki} y_{it_2} - \tilde{y}_{1k\pi,gt_2}), \end{array} \tag{26}$$

$$w_{1ki} = \left(\sum_{j\in A_{1k}} \pi_{1j}^{-1}\right)^{-1} \pi_{1i}^{-1}, \tag{27}$$

$$\tilde{y}_{1k\pi,gt} = n_{1gk}^{-1} \sum_{i\in A_{1gk}} w_{1ki} y_{it}, \tag{28}$$

and A_{1gk} is the set of indices in stratum g in A_{1k}.

Note that in the theorem, means from different panels are correlated. In the cell-mean model, we assume panel means are uncorrelated. For second-phase stratified sampling, the correlation between panel means arises due to the finite population correction factors. The cross-panel correlations will be small when the first-phase sampling fractions in second-phase strata are small. The theorem is a direct application of the lemma. Conditional almost sure convergence is obtained by (19). The sequence of populations and first-phase samples considered is not a standard sequence due to the nesting of the samples. However, sequences of poisson samples, simple random samples, and stratified random samples can be given such a representation.

Using the EGLSE

The EGLSE may be used to provide estimates of totals or means. However, for many large-scale surveys, the data are used in estimation problems that are not specified to the survey statisticians. A typical output of an initial estimation process is a dataset with a set of weights. Practitioners decide what estimators are of interest. Often the parameters being investigated can be written as smooth functions of totals.

The EGLSE returns a vector of weights, $\hat{\mathbf{\Lambda}}$, for the vector of panel means. Different weights are applied to the panel mean vector depending on the target time point. The EGLSE weights are also particular to the variable being analyzed. Therefore, supplying the EGLSE weights for all possible variables would be computationally burdensome. We propose three methods for applying the EGLSE when the output of the survey is a dataset.

From a dataset creation standpoint, the simplest way to include the EGLSE in output would be to supply a set of $\hat{\mathbf{\Lambda}}$ matrices. Practitioners can compute panel means for their variables of interest, then apply appropriate EGLSE coefficient matrices for their variables. The coefficient matrices differ only by the components in $\hat{\boldsymbol{V}}$. Therefore, the EGLSE coefficient matrix for a variable with a variance matrix similar to the variable of interest may be used to construct an approximation to the EGLSE. The responsibility for choosing the coefficient matrices falls on the practitioner, making estimation more cumbersome than the use of direct weighted sums.

The consistency of the replication variance estimator does not require that the variance estimator $\hat{\boldsymbol{V}}$ be consistent for $\boldsymbol{V}$. The requirement is that $\hat{\boldsymbol{V}}$ converges to a positive definite matrix. Therefore, the replication variance estimator may be used when the EGLSE coefficient matrix from another variable is used. Similarly, the central limit theorem only assumes the EGLSE coefficient matrix is consistent for a matrix. The EGLSE constructed with the $\hat{\mathbf{\Lambda}}$ matrix from another variable does not necessarily have the asymptotic efficiency of the GLSE. The efficiency loss will depend on the degree to which the variance matrices for the variable of interest and the EGLSE variable agree.

The EGLSE may be used as the control total in a regression estimator. For specified variables, the regression estimators will match the EGLSEs. For the remaining variables, the direct weighted estimators may be improved by the adjustment for EGLSEs. Because the sum of the weights must be an estimate of the total number of units, the regression weights can only be applied to a set of elements with data for all time points.

Let A be the portion of the sample from a longitudinal two-phase survey that is always observed. Let $\hat{\boldsymbol{\mu}}_q$ be the EGLSE for characteristic q. The first step is to construct the EGLSE for Q different variables of interest. Let $w_{i,old}$ be the analysis weight for segment i, ratio adjusted so that $\sum_{i\in A} w_{i,old} = \hat{N}$, where $\hat{N}$ an either the known population size or an estimator of the population size. We shall consider $w_{i,old}$ that are proportional to $\pi_i^{\star -1} = \pi_{1i}^{-1}\pi_{2i|1i}^{-1}$, the two-phase probability for including segment i in A. A set of regression weights $w_{i,new}$ are those that minimize

$$\sum_{i\in A} \alpha_i^{-1}(w_{i,new} - w_{i,old})^2 \tag{29}$$

subject to

$$\sum_{i\in A} w_{i,new}\boldsymbol{y}_{qi} = N\hat{\boldsymbol{\mu}}_q \tag{30}$$

for $q = 1, 2, \ldots, Q$, where $\boldsymbol{y}_{qi}$ is the observation vector of unit i for characteristic q and the α_i's are constants chosen by the practitioner (Deville and Särndal 1992). The multiplier N in (30) may be replaced by $\hat{N}$. The regression weights are

$$\begin{aligned} w_{i,new} &= w_{i,old} + \alpha_i \boldsymbol{y_i'} \left[\textstyle\sum_{i\in A} \alpha_i \boldsymbol{y_i} \boldsymbol{y_i'}\right]^{-1} \\ &\quad \times \left(N\hat{\boldsymbol{\mu}} - \textstyle\sum_{i\in A} w_{i,old}\boldsymbol{y_i}\right). \end{aligned} \tag{31}$$

When we apply the weights in (31) to a variable z_i, the estimator for the total of z_i is

$$\begin{aligned} &\textstyle\sum_{i\in A} w_{i,new} z_i \\ &= \textstyle\sum_{i\in A} w_{i,old} z_i + \left(N\hat{\boldsymbol{\mu}} - \sum_{i\in A} w_{i,old}\boldsymbol{y_i}\right)' \hat{\boldsymbol{\beta}}_{\boldsymbol{zy}}, \end{aligned} \tag{32}$$

where

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{zy}} = \left[\sum_{i\in A} \alpha_i \boldsymbol{y_i} \boldsymbol{y_i'}\right]^{-1} \sum_{i\in A} \alpha_i \boldsymbol{y_i} z_i. \tag{33}$$

The mean square error improvement for the the regression estimator for z_i depends on the strength and shape of the relationship between z and $\boldsymbol{y}$. The improvement will be large for a highly correlated linear relationship between z and $\boldsymbol{y}$.

To establish the limiting properties of the regression estimator, write the regression estimator of the mean

as

$$\overline{z}_{reg,A} = \left[\sum_{i\in A} w_{i,new}\right]^{-1} \sum_{i\in A} w_{i,new} z_i. \quad (34)$$

Assume

$$\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} = O_p(n_1^{-1/2}), \quad (35)$$

$$\overline{z}_A - \overline{z}_N = O_p(n_1^{-1/2}), \quad (36)$$

$$\overline{\boldsymbol{y}}_{\boldsymbol{A}} - \boldsymbol{\mu} = O_p(n_1^{-1/2}), \quad (37)$$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{zy}} - \boldsymbol{\beta}_{\boldsymbol{zy}} = O_p(n_1^{-1/2}), \quad (38)$$

and

$$\hat{\boldsymbol{\mu}} - \breve{\boldsymbol{\mu}} = O_p(n_1^{-1}), \quad (39)$$

where

$$\boldsymbol{\beta}_{\boldsymbol{zy}} = \left[\sum_{i\in 1}^{N} \alpha_i w_{i,old}^{-1} \boldsymbol{y_i} \boldsymbol{y_i'}\right]^{-1} \sum_{i\in 1}^{N} \alpha_i w_{i,old}^{-1} \boldsymbol{y_i} z_i, \quad (40)$$

$\hat{\boldsymbol{\mu}}$ is the EGLSE of the vector of time means $\boldsymbol{\mu}$, $\overline{z}_N$ is the population mean of z, and n_1 is the first-phase sample size. The error in $\overline{z}_{reg,A}$ is

$$\begin{aligned} \overline{z}_{reg,A} - \overline{z}_N &= \overline{z}_A - \overline{z}_N + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}}_{\boldsymbol{zy}} - \\ &\quad (\overline{\boldsymbol{y}}_{\boldsymbol{A}} - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}}_{\boldsymbol{zy}} \\ &= \overline{z}_A - \overline{z}_N + [(\breve{\boldsymbol{\mu}} - \boldsymbol{\mu}) + (\hat{\boldsymbol{\mu}} - \breve{\boldsymbol{\mu}})]' \\ &\quad \times \hat{\boldsymbol{\beta}}_{\boldsymbol{zy}} - (\overline{\boldsymbol{y}}_{\boldsymbol{A}} - \boldsymbol{\mu})'\hat{\boldsymbol{\beta}}_{\boldsymbol{zy}} \\ &= \overline{e}_A + (\breve{\boldsymbol{\mu}} - \boldsymbol{\mu})'\boldsymbol{\beta}_{zy} + O_p(n_1^{-1}) \\ &=: d_{reg} + O_p(n_1^{-1}), \end{aligned} \quad (41)$$

where

$$d_{reg} = \overline{e}_A + (\breve{\boldsymbol{\mu}} - \boldsymbol{\mu})'\boldsymbol{\beta}_{zy} \quad (42)$$

and

$$\overline{e}_A = \left[\sum_{i\in A} w_{i,old}\right]^{-1} \sum_{i\in A} w_{i,old}(z_i - \overline{z}_N - (\boldsymbol{y_i} - \boldsymbol{\mu})'\boldsymbol{\beta}_{\boldsymbol{zy}}). \quad (43)$$

The variance of the approximating variable is

$$V(d_{reg}) = V(\overline{e}_A) + \boldsymbol{\beta}'_{\boldsymbol{zy}} \boldsymbol{V}(\breve{\boldsymbol{\mu}}) \boldsymbol{\beta}_{\boldsymbol{zy}} + 2\boldsymbol{\beta}'_{\boldsymbol{zy}} C(\overline{e}_A, \breve{\boldsymbol{\mu}}). \quad (44)$$

From the uncorrelated panel assumption, $C(\overline{e}_A, \breve{\boldsymbol{\mu}})$ depends only on A. By our assumptions, the variance of d_{reg} is $O_p(n_1^{-1})$.

We now extend the replication variance estimator to the regression estimator. Suppose we have a consistent replication variance estimator for $V(\hat{\boldsymbol{\mu}})$ as defined in Legg, Fuller, and Nusser (2005). Since $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{zy}}$ are functions of the sample containing the $\boldsymbol{y_q}$ and z values, the same replicate sample may be used to construct a replicate for

$$\hat{t}_{z,reg} = \sum_{i\in A} w_{i,new} z_i. \quad (45)$$

Let $A^{(k)}$ be the k^{th} replicate sample. A replicate for $\hat{t}_{z,reg}$ is

$$\hat{t}^{(k)}_{z,reg} = \hat{t}^{(k)}_z + (N\hat{\boldsymbol{\mu}}^{(\boldsymbol{k})} - \hat{\boldsymbol{t}}^{(\boldsymbol{k})}_{\boldsymbol{y}})'\hat{\boldsymbol{\beta}}^{(\boldsymbol{k})}_{\boldsymbol{zy}}, \quad (46)$$

where

$$\hat{t}^{(k)}_z = N\left[\sum_{i\in A^{(k)}} w_{i,old}\right]^{-1} \sum_{i\in A^{(k)}} w_{i,old} z_i, \quad (47)$$

$$\hat{\boldsymbol{\beta}}^{(\boldsymbol{k})}_{\boldsymbol{zy}} = \left[\sum_{i\in A^{(k)}} \alpha_i \boldsymbol{y_i} \boldsymbol{y_i'}\right]^{-1} \sum_{i\in A^{(k)}} \alpha_i z_i \boldsymbol{y_i}, \quad (48)$$

and

$$\hat{\boldsymbol{\mu}}^{(\boldsymbol{k})} = (\boldsymbol{X'}\hat{\boldsymbol{V}}_{\boldsymbol{N}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'}\hat{\boldsymbol{V}}_{\boldsymbol{N}}^{-1}\overline{\boldsymbol{y}}^{(\boldsymbol{k})}_{\boldsymbol{2\pi,N}}, \quad (49)$$

where

$$\overline{y}^{(k)}_{2\pi,pt,N} = \left[\sum_{i\in A^{(k)}_{2p}} \pi^{-1}_{2i|1i}\pi^{-1}_{1i}\right]^{-1} \sum_{i\in A^{(k)}_{2p}} \pi^{-1}_{2i|1i}\pi^{-1}_{1i} y_i, \quad (50)$$

where $A^{(k)}_{2p}$ is the set of indices in $A_{2p} \bigcap A^{(k)}$ and the components of $\hat{\boldsymbol{t}}^{(\boldsymbol{k})}_{\boldsymbol{y}}$ are calculated as in (47).

The regression estimator requires a set of data with a completely observed time series. One option is to use imputation to create a complete dataset. If the regression estimator is computed for the imputed dataset, the regression weights will depend on both the real and imputed values. An alternative is to select imputed values so that the EGLSE is attained for weighted sums of selected variables.

Let A_t be the portion of the sample observed at time t. Let

$$\overline{z}_{A_t} = \left[\sum_{i\in A_t} w_{i,old}\right]^{-1} \sum_{i\in A_t} w_{i,old} z_{it}, \quad (51)$$

$$\overline{y}_{q,A_t} = \left[\sum_{i\in A_t} w_{i,old}\right]^{-1} \sum_{i\in A_t} w_{i,old} y_{qit}, \quad (52)$$

and

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{zy,A_t}} = \left[\hat{\boldsymbol{V}}(\overline{\boldsymbol{y}})\right]^{-1} \hat{\boldsymbol{C}}(\overline{\boldsymbol{y}}, \overline{z}_{A_t}) \quad (53)$$

where $\hat{\boldsymbol{V}}(\overline{\boldsymbol{y}})$ and $\hat{\boldsymbol{C}}(\overline{\boldsymbol{y}}, \overline{z}_{A_t})$ are pooled consistent estimators of $\boldsymbol{V}(\overline{\boldsymbol{y}})$ and $\boldsymbol{C}(\overline{\boldsymbol{y}}, \overline{z}_{A_t})$, respectively, and $\overline{\boldsymbol{y}}$ is the vector of $\overline{y}_{q,A_t}$ for $q = 1, \ldots, Q$. Let

$$\overline{z}_{reg,A_t} = \overline{z}_{A_t} + (\hat{\boldsymbol{\mu}} - \overline{\boldsymbol{y}})'\hat{\boldsymbol{\beta}}_{\boldsymbol{zy,A_t}} \quad (54)$$

be the regression estimator for the mean at time t.

Assume an imputation procedure that imputes a value for each unobserved z_{it} and y_{qit} and denote the full first-phase sample after imputation by A_I. Further, assume that an imputed value is a function of observations from the same panel. Let $w_{i,full}$ be the weight associated with the first-phase sample. Let

$$\overline{z}_{I,t} = \left[\sum_{i\in A_I} w_{i,full}\right]^{-1} \sum_{i\in A_I} w_{i,full} z_{I,it}, \qquad (55)$$

where $z_{I,it}$ is the observed z_{it} for panels observed at time t and is the imputed z_{it} for unobserved panels. It is assumed that the imputed z_{it} are such that $\overline{z}_{I,t} = \overline{z}_{reg,A_t}$. Assume that

$$\overline{y}_{qtp,I} - \mu_{qt} = O_p(n_1^{-1/2}) \qquad (56)$$

and

$$\overline{z}_{p,I} - N^{-1} t_z = O_p(n_1^{-1/2}), \qquad (57)$$

where $\overline{y}_{qtp,I}$ and $\overline{z}_{p,I}$ are means from imputed panel p data at time t. Then, the order of the error in $\overline{z}_{I,t}$ is the same as for $\overline{z}_{reg,A_t}$.

Selecting imputed values to satisfy a set of restrictions generated by the regressions can be difficult. Since the imputed values will form a time series, the series for each unit should be internally consistent. One procedure is to first impute values and then use the EGLSE as a diagnostic for imputation. If weighted total estimators after imputation differ from the EGLSEs for selected variables by more than a specified tolerance, the imputation procedure requires adjustment. Otherwise, the dataset after imputation is accepted.

Discussion

Estimated generalized least squares solutions are commonly used in linear modeling applications. For longitudinal samples, combining information across observation time periods using an EGLSE requires assumptions for the sequence of populations, first-phase sample designs, and second-phase sample designs. In our work, we assume standard properties hold for the first-phase sample and prove results for fixed rate stratified random sampling for the second-phase.

The theory developed in this paper and in Legg, Fuller and Nusser (2005) rebuild the standard tools for analyzing large survey data in the framework of longitudinal two-phase sampling. The EGLSE automates constructing composite estimators using the information from different panels and years.

Note on Lemma 1

After completion of this work, J. N. K. Rao sent a copy of a manuscript on limiting normality for two-phase sampling designs (Chen and Rao 2006). In Theorem 2 of Chen and Rao the almost sure conditional convergence assumption of our lemma is replaced by a strong conditional weak convergence statement.

Acknowledgements

This work was supported by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

References

Chen, J. and Rao, J. N. K. (2006). Asymptotic normality under two-phase sampling designs. To appear in *Statistica Sinica*.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Fuller, W. A. (1987). *Measurement Error Models*. John Wiley and Sons, Inc.

Legg, J. C., Fuller, W. A., and Nusser, S. M. (2005). Estimation for longitudinal surveys with repeated panels of observations. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag New York, Inc.

Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics*, 16(4):1550–1566.

Skinner, C. J. and Holmes, D. J. (2003). Random Effects Models for Longitudinal Survey Data. In R.L. Chambers and C.J. Skinner (eds.), *Analysis of Survey Data*. John Wiley and Sons, Inc., pp. 205-219.