Partial Overlap Samples and Combining Information

Wayne A. Fuller

Statistical Laboratory, Iowa State University, USA

ABSTRACT: In many surveys, different quantities of information are obtained from different units. Common examples are multiple phase samples, partially overlapping samples, and partial response. Estimators of the regression type are exhibited for such situations, with particular application to longitudinal surveys. An example applied to the National Resources Inventory is described.

1. Introduction

The study of the dynamics of populations is currently receiving considerable attention. Such study requires observations made at multiple time points on units in the population. Surveys designed specifically for such study include the Survey of Income and Program Participation (SIPP) conducted by the U.S. Census Bureau and the National Resources Inventory (NRI) conducted by the Natural Resources Conservation Service of the U.S. Department of Agriculture. The NRI is nearly a pure panel survey of certain land areas. We define a pure panel survey to be a survey in which the same units are observed at each time point of a survey conducted at more than one time point. A longitudinal survey is a survey conducted at more than one time point with some units observed at more than one time point. The term longitudinal survey is generally used for surveys conducted at more than two points in time with multiple observations on some units planned as part of the survey design.

A rotation survey is one in which a unit is observed for a partial set of time points and is not observed for the remaining set of time points in the study. There are many ways in which the observation pattern can be specified. The Canadian Labor Force Survey and the U.S. Current Population Survey are examples of surveys designed to run continuously in which units rotate into the sample for a fixed period (or periods) and then permanently rotate out of the observation set.

There exist an array of designs combining individuals observed at some time points and individuals observed at all time points of the study set of time points. The simplest such design is a two-phase sample in which the observations at the second of two time points is a subsample of those observed at time one. The book edited by Kasprzyk et al. [1989] contains an excellent discussion of various aspects of panel surveys. Duncan and Kalton [1987] and Schreuder, Gregorie, and Wood [1993] discuss different types of repeated surveys and the objectives of such surveys.

The largest fraction of the survey literature on repeated surveys has been devoted to rotating surveys. An early study describing the use of least squares to incorporate information from a previous occasion into the estimate of the current occasion is that of Jessen [1942]; also see Cochran [1942]. Patterson [1950] investigated estimation for rotating samples. Patterson's work was followed by a number of authors, including Eckler [1955], Rao and Graham [1964], Gurney and Daly [1965], Raj [1965], Smith [1978], Wolter [1979], Jones [1980], Huang and Ernst [1981], Breau and Ernst [1983], and Kumar and Lee [1983].

We shall compare some designs for studies directed toward longitudinal dynamics. Because such studies are multiple purpose, we cannot hope to develop a design that is optimal for all objectives. A part of our investigation will be to identify the trade-offs.

2. Supplemented Panel Designs

Consider a simple three period survey in which one-fourth of the units are observed over all three periods and each of three sets of one-fourth of the units is observed over exactly one of the three periods. Thus, if the total sample size is n, then 0.5 n of the units are observed at each time point. Let (Y_1, Y_2, Y_3,) denote the value of a characteristic observed at times one, two, and three, respectively. Assume that the correlation between observations at time i and time j on the same element is that of a first order autoregressive process with parameter ρ. Assume simple random sampling for the selection of all samples.

Let $(\bar{y}_{11}, \bar{y}_{21}, \bar{y}_{31})'$ denote the estimated mean at time one, two and three, of the sample elements that are observed all three periods. Let $(\bar{y}_{12}, \bar{y}_{23}, \bar{y}_{34})'$ denote the sample means for the three periods for the sample elements that are observed once. Let $\mu = (\mu_1, \mu_2, \mu_3)'$ denote the population means for the three periods. Then we can write

$$\bar{y} = X\mu + e \tag{1}$$

where $\bar{y}' = (\bar{y}_{11}, \bar{y}_{21}, \bar{y}_{31}, \bar{y}_{12}, \bar{y}_{23}, \bar{y}_{34})$, $X' = (I_3, I_3)$ and the covariance matrix of e is

$$\textstyle\sum_{ee} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & 0 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 0 & 0 & 0 \\ \rho_2 & \rho_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \sigma^2 \tag{2}$$

where σ^2 is the variance of a mean of $n/4$ observations, $\rho_1 = \rho$, and $\rho_2 = \rho^2$. It follows that the best linear unbiased estimator of μ using this amount of information is

$$\hat{\mu}_g = (X' \textstyle\sum_{ee}^{-1} X)^{-1} X' \textstyle\sum_{ee}^{-1} \bar{y} \tag{3}$$

and

$$V\{\hat{\mu}_g\} = (X' \textstyle\sum_{ee}^{-1} X)^{-1} . \tag{4}$$

We compare the covariance matrix (4) with the covariance matrix of a pure panel survey in which the same 0.5 n units are observed on all three periods. The covariance matrix for the pure panel design is

$$V\{\hat{\mu}_{panel}\} = \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \frac{\sigma^2}{2}.$$

Table 1 contains variances for several functions of means for ρ ranging from 0 to 0.99, relative to the corresponding variance for the pure panel design. If $\rho = 0$, the variance of the best linear unbiased estimator of the three means is the variance of the simple mean at each period. If the total number of elements is n, the variance at each time period is $2n^{-1}\sigma^2$. In the remaining discussion we assume $\sigma^2 = 1$. If observations on the same element are correlated ($\rho \neq 0$), the variance of the estimated means for the supplemented panel design is less than $2n^{-1}$. The limit of the correlation is one. This can occur for characteristics such as the age of an individual at a fixed point. The lower bound for the variance of the supplemented panel design at $\rho = 1$ is n^{-1} because this is the number of different individuals in the study. Correspondingly, the limit of the relative efficiency for period means of the supplemented panel, relative to the pure panel, is 2.0. The variances of the estimated means for the supplemented panel design for the first and last period are the same. The variance of the middle period is somewhat smaller because the middle observation has one period correlation with the first and third observations on the same element.

Table 1. Variances of Functions of the Estimators for Three Period Design with 50% New Observations at Each Time Relative to Variances of Pure Panel Design

ρ	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_1 - \bar{y}_2$	$\bar{y}_1 - \bar{y}_3$
-0.70	0.838	0.755	0.674	1.253
-0.50	0.929	0.875	0.768	1.143
0.00	1.000	1.000	1.000	1.000
0.50	0.929	0.875	1.304	1.143
0.70	0.838	0.755	1.488	1.253
0.90	0.660	0.595	1.773	1.681
0.99	0.520	0.510	1.973	1.951

If the correlation is positive, the variance of the difference of two means is smaller for the pure panel survey than for the supplemented panel design. As ρ approaches one, the variance of the difference of two means approaches zero for both designs. Thus, for example, the variance of period-to-period change for the pure panel is only two percent of the variance of the mean if $\rho = 0.99$. See Table 2. The pure panel design has an efficiency approaching twice that of the supplemented panel design as ρ approaches one.

Table 2. Variances of Functions of Means for Pure Panel Study

ρ	$nV\{\bar{y}_1\}$	$nV\{\bar{y}_2\}$	$nV\{\bar{y}_1 - \bar{y}_2\}$	$nV\{\bar{y}_1 - \bar{y}_3\}$
-0.70	2.00	2.00	6.80	2.04
-0.50	2.00	2.00	6.00	3.00
0.00	2.00	2.00	4.00	4.00
0.50	2.00	2.00	2.00	3.00
0.70	2.00	2.00	1.20	2.04
0.90	2.00	2.00	0.40	0.76
0.99	2.00	2.00	0.04	0.08

The roots of the covariance matrices for the two designs and various values of ρ are given in Table 3. Except for $\rho = 0$, the supplemented panel design has a smaller largest root and a larger smallest root.

If one uses the trace criterion to choose a design, then one would always use the design in which some new elements are observed at each time period. For example, if $\rho = 0.5$, the ratio of the trace of the covariance matrix of the pure panel design to that of the supplemented panel design is 1.10. If $\rho = 0.7$, the ratio is 1.23.

Table 3. Roots of Covariance Matrix for Three Period Study

Type	ρ	r_1	r_2	r_3	Sum
Pure Panel	-0.7	4.53	1.02	0.45	6.00
Supplemented		2.77	1.35	0.74	4.86
Pure Panel	-0.5	3.69	1.50	0.81	6.00
Supplemented		2.59	1.71	1.16	5.46
Pure Panel	0.0	2.00	2.00	2.00	6.00
Supplemented		2.00	2.00	2.00	6.00
Pure Panel	0.5	3.69	1.50	0.81	6.00
Supplemented		2.59	1.71	1.16	5.46
Pure Panel	0.7	4.53	1.02	0.45	6.00
Supplemented		2.77	1.35	0.74	4.86
Pure Panel	0.9	5.48	0.38	0.14	6.00
Supplemented		2.93	0.64	0.26	3.83
Pure Panel	0.99	5.95	0.04	0.01	6.00
Supplemented		2.99	0.08	0.03	3.34

Table 4 contains the variance of some functions of means for the supplemented panel design relative to the pure panel design for a five period study. The variances of the first period means are similar to those of the three period design. However, the variances of the second and third period means are considerably smaller than the second period mean of the three period design. The limit of the relative efficiency for the means of the supplemented panel relative to the pure panel as ρ approaches one is 3.0. On the other hand, the limit of the relative efficiency for the difference of two means is 0.5, as it was for the three period design.

Table 4. Variances of Estimated Means for Five Period Supplemented Design Relative to Variances of Pure Panel Design

ρ	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_2 - \bar{y}_1$	$\bar{y}_3 - \bar{y}_1$	$\bar{y}_4 - \bar{y}_1$
-0.70	0.833	0.734	0.721	0.661	1.250	0.627
-0.50	0.928	0.871	0.867	0.765	1.108	0.815
0.00	1.000	1.000	1.000	1.000	1.000	1.000
0.50	0.928	0.871	0.867	1.301	1.108	1.007
0.70	0.833	0.734	0.721	1.478	1.250	1.104
0.90	0.615	0.517	0.492	1.750	1.577	1.450
0.99	0.375	0.357	0.351	1.966	1.937	1.914

The general structure of the roots of the covariance matrix of the five means for the five period design is similar to that of the three period design. The largest root of the pure panel design is always larger than the largest root of the supplemented panel design. The supplemented panel design always has a larger smallest root than the pure panel design. The dominance of the supplemented panel relative to the pure panel design with respect to the trace criterion is greater for the five period design than for the three period design. This is a further reflection of the fact that estimates of period means are being improved more than estimates of differences are being degraded, as the number of periods is increased. The trace for the five period pure panel is 5.00 and the traces for the supplemented panel design are 3.86, 4.46, 5.00, 4.46, 3.86, 2.76 and 1.82 for ρ =-0.70, -0.50, 0.00, 0.50, 0.70, 0.90 and 0.99, respectively.

As a second illustration of the variance of a panel with a supplemented component, consider a 2×2×2 table constructed for a binomial variable observed at three points in time. We assume that the probability of being in state one is 0.5 and the probability of being in state i at time t, given state i at time t -1 is 0.8 for both states zero and one. Let

X_j = 1 if state one at time j, for j = 1, 2, 3,
= 0 otherwise;

X_{2+j} = 1 if state one at time 1 and state one at time j, for j = 2, 3,
= 0 otherwise;

X_6 = 1 if state one at time 2 and state one at time 3,
= 0 otherwise; and

X_7 = 1 if state one at times 1, 2 and 3,
= 0 otherwise.

The covariance matrices for estimators constructed with the pure panel design and the supplemented panel design are given in Table 5. The entries are for the variance of a sample of size n at each time period multiplied by n. The supplemented panel design has smaller variances for the three time period marginals and for the marginals of the adjacent 2 x 2 tables (period 1 and 2, and period 2 and 3). The pure panel design has smaller variances for the one-three second order interaction (X_5) and the third order interaction (X_7). For those estimators in the supplemented panel design with smaller variances, the covariances are also smaller than those of the pure panel design. As a result, for example, the estimated change from period one to period two is smaller for the pure panel design than for the supplemented panel design.

Table 5. Covariance Matrices for Alternative Designs for 2×2×2 Table

Covariance Matrix of Pure Panel Design						
0.2500	0.1500	0.0900	0.2000	0.1700	0.1200	0.1600
	0.2500	0.1500	0.2000	0.1500	0.2000	0.1600
		0.2500	0.1200	0.1700	0.2000	0.1600
			0.2400	0.1840	0.1600	0.1920
				0.2244	0.1840	0.2112
					0.2400	0.1920
						0.2177
Covariance Matrix of Supplemented Panel Design						
0.2226	0.0750	0.0274	0.1488	0.1250	0.0512	0.1120
	0.2050	0.0750	0.1400	0.0750	0.1400	0.0928
		0.2226	0.0512	0.1250	0.1488	0.1120
			0.2244	0.1480	0.0956	0.1664
				0.2338	0.1480	0.2144
					0.2244	0.1664
						0.2324

The characteristic roots of the covariance matrix for the pure panel design are

$$(1.249,\ 0.209,\ 0.113,\ 0.056,\ 0.031,\ 0.010,\ 0.003),$$

and the characteristic roots for the supplemented panel design are

$$(0.942,\ 0.265,\ 0.176,\ 0.096,\ 0.059,\ 0.021,\ 0.006).$$

The sum of the roots for the pure panel design is 1.671 and the sum for the supplemented panel design is 1.565. The orthogonal linear combination with largest variance has a variance about 30 percent smaller for the supplemented panel design. This linear combination is essentially a sum of the seven estimates. The remaining six linear combinations have much smaller variances for both designs, and the variances are smaller for the pure panel design than for the supplemented panel design.

3. Estimation for Additional Characteristics

In the previous section, we outlined an estimation scheme for a vector of time means for a y-characteristic. While we considered a scalar y, there is no conceptual difficulty in extending the procedure to a vector of y-characteristics.

In the 2×2×2 table, the variables for change in classification, $X_4\ X_5\ X_6\ X_7$, are observed only in the panel portion of the sample. We constructed estimates for these variables using the generalized least squares approach. This is an inefficient computational procedure if one is dealing with many characteristics. Also, if the generalized least squares procedure is applied to different sets of variables with some overlap, then the estimates are not internally consistent. For applications, we suggest the procedure of Fuller [1990]. In that procedure, a set of variables from each time period is chosen as the control variables. The generalized least squares procedure is then used to estimate the means (or totals) for the vector of control variables. Given the estimated control means, a set of regression weights is constructed for the panel portion of the sample such that

$$\sum\nolimits_{i\in s_1} w_i\, y_i \;=\; \hat{\mu}_y ,$$

where s_1 is the panel portion of the sample, w_i is the regression weight for the ith element in the panel portion, y_i is the vector of control variables for the ith element, and $\hat{\mu}_y$ is the estimated generalized least squares estimator of the mean of y. The estimator of the mean of a characteristic z is

$$\hat{\mu}_z = \sum_{i \in s_1} w_i \, z_i .$$

This estimator is a type of two-phase estimator. It differs from the classical two phase estimator in that portions of the control vector are observed in different parts of the large (first phase) sample. The estimator can also be written

$$\hat{\mu}_z = \bar{z}_{(1)} + (\hat{\mu}_y - \bar{y}_{(1)}) \, \hat{\beta}_{z \cdot y} ,$$

where $\hat{\beta}_{z \cdot y}$ is the regression coefficient for the regression of z on y computed from s_1 and $(\bar{z}_{(1)}, \bar{y}_{(1)})$ is the vector of estimated means based upon the panel portion of the sample. The variance of the approximate distribution of $\hat{\mu}_z$ is

$$V\{\hat{\mu}_z\} = V\{\bar{e}_{(1)}\} + 2C\{\bar{e}_{(1)}, \, \hat{\mu}_y \, \beta_{z \cdot y}\} + V\{\hat{\mu}_y \, \beta_{z \cdot y}\} ,$$

where

$$\bar{e}_{(1)} = \bar{z}_{(1)} - \bar{y}_{(1)} \, \beta_{z \cdot y}$$

and $\beta_{z \cdot y}$ is the population analog of $\hat{\beta}_{z \cdot y}$. If the estimator $\hat{\beta}_{z \cdot y}$ is root n consistent for $\beta_{z \cdot y}$ and if $\beta_{z \cdot y}$ is such that

$$C\{\bar{e}_{(1)}, \, \hat{\mu}_y \} = 0 ,$$

the variance estimation procedure suggested by Fuller (1997) can be applied to estimate the variance of $\hat{\mu}_z$. If one is unwilling to assume zero covariances, then a replication scheme in which replicates are drawn from the entire sampling procedure can be used. See Rao and Sitter [1995] and Sitter [1997].

4. The National Resources Inventory

The National Resources Inventory is a survey of the nonfederal land area of the United States conducted by the Natural Resources Conservation Service of the U.S. Department of Agriculture. It is a large survey of about 300,000 primary sampling units. A primary sampling unit is a segment of land. In the Midwest, the segment is 160 acres, but the primary sampling units vary across the country. In the western part of the U.S., there are some that are as big as 640 acres, and, in the east, some segments are on the order of 100 acres. Within the primary sampling unit, points are designated for observations. There are either two or three points per primary sampling unit. The basic survey has been conducted in each of the years 1982, 1987, 1992, and is currently underway for 1997.

In 1995, a sample of 3,000 segments was selected from the 300,000 for a study that was called the Erosion Update Study. This study was a subsample of the large sample, but the primary sampling units were different. In 26 states, counties were sampling units and for 22 of the states, states were primary sampling units. The 1992 basic NRI sample of 300,000 segments and the 1995 subsample of 3,000 segments form a classical two-phase sample. In the first-phase sample of 300,000 sampling units, vector X is observed. In the subsample of 3,000 units, an extended vector (X, Y) is observed.

A second special study was conducted in 1996. The original 1995 sample of 3,000 was augmented by another 1,000 segments to obtain a total of 4,000 segments. A special study is underway in 1997 in which the 1996 sample has been augmented by another 2,000 segments. Thus, we can divide the original 300,000 segments into four subgroups: 292,000 are observed only in 1992; 3,000 are observed in 1992, 1995, 1996, and 1997; 1,000 are observed in 1992, 1996, and 1997; and 2,000 are observed in 1992 and 1997.

The regression model for a characteristic y for these data is

$$\begin{bmatrix} \bar{y}_{92,292} \\ \bar{y}_{92,3} \\ \bar{y}_{95,3} \\ \bar{y}_{96,3} \\ \bar{y}_{97,3} \\ \bar{y}_{92,1} \\ \bar{y}_{96,1} \\ \bar{y}_{97,1} \\ \bar{y}_{92,2} \\ \bar{y}_{97,2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{92} \\ \mu_{95} \\ \mu_{96} \\ \mu_{97} \end{bmatrix} + \begin{bmatrix} e_{92,292} \\ e_{92,3} \\ e_{95,3} \\ e_{96,3} \\ e_{97,3} \\ e_{92,1} \\ e_{96,1} \\ e_{97,1} \\ e_{92,2} \\ e_{97,2} \end{bmatrix}$$

where $\bar{y}_{t,j}$ is the mean for the characteristic at time t on a group of j(000) segments, and μ_t is the mean of y at time t. If we assume that the four subgroups are independent and that the correlation is of the autoregressive form with $\rho = 0.9$, the efficiency of the generalized least squares estimator relative to the simple mean is 1.00, 2.94, 2.27 and 1.52 for μ_{92}, μ_{95}, μ_{96} and μ_{97}, respectively.

The estimation method used in the actual study was closely related to the generalized least squares procedure with a selected subset of control variables as outlined in Section 3.

References

Breau, P. and Ernst, L.R. (1983), "Alternative Estimators to the Current Composite Estimator," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 397-402.

Cochran, W.G. (1942), "Sampling Theory when the Sampling Units are of Unequal Sizes," *Journal of the American Statistical Association,* 37, pp. 199-212.

Duncan, G.J. and Kalton, G. (1987), "Issues of Design and Analysis of Surveys Across Time," *International Statistical Review*, 55, pp. 97-117.

Eckler, A.R. (1987), "Rotation Sampling," *Annals of Mathematical Statistics*, 26, pp. 664-685.

Fuller, W.A. (1990), "Analysis of Repeated Surveys," *Survey Methodology*, 16, pp. 167-180.

Fuller, W.A. (1997), "Replication Variance Estimation for Two Phase Samples," unpublished manuscript, Iowa State University, Ames, Iowa.

Gurney, M. and Daly, J.F. (1965), "A Multivariate Approach to Estimation in Periodic Sample Surveys," *Proceedings of the American Statistical Association, Section on Social Statistics*, pp. 242-257.

Huang, L.R. and Ernst, L.R. (1981), "Comparison of an Alternative Estimator to the Current Composite Estimator in the Current Population Survey," *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 303-308.

Jessen, R.J. (1942), "Statistical Investigation of a Sample Survey for Obtaining Farm Facts," *Iowa Agricultural Experiment Station Research Bulletin,* 304, pp. 54-59.

Jones, R.G. (1980), "Best Linear Unbiased Estimators for Repeated Surveys," *Journal of the Royal Statistical Society, Series B,* 42, pp. 221-226.

Kasprzyk, D., Duncan, G.J., Kalton, G. and Singh, M.P. (1989), *Panel Surveys*, New York: John Wiley & Sons.

Kumar, S. and Lee, H. (1983), "Evaluation of Composite Estimation for the Canadian Labor Force Survey," *Survey Methodology,* 9, pp. 1-24.

Patterson, H.D. (1950), "Sampling on Successive Occasions with Partial Replacement of Units," *The Journal of the Royal Statistical Society, Series B,* 12, pp. 241-255.

Raj, D. (1965), "On Sampling over Two Occasions with Probability Proportionate to Size," *Annals of Mathematical Statistics,* 36, pp. 327-330.

Rao, J.N.K. and Graham, J.E. (1964), Rotation Designs for Sampling on Repeated Occasions," *Journal of the American Statistical Association,* 59, pp. 492-509.

Rao, J.N.K. and Sitter, R.R. (1995), "Variance Estimation under Two-phase Sampling with Application to Imputation for Missing Data," *Biometrika*, 82, pp. 453-460.

Schreuder, H.T., Gregorie, T.G. and Wood, G.B. (1993), *Sampling Methods for Multi-Resource Forest Inventory*, New York: John Wiley &Sons.

Sitter, R.R. (1997), "Variance Estimation for the Regression Estimator in Two-phase Sampling," *Journal of the American Statistical Association*, 92, pp. 780-787.

Smith, T.M.F. (1978), "Principles and Problems in the Analysis of Repeated Surveys," in *Survey Sampling and Measurement*, ed. N. Krishnan Namboodiri, New York: Academic Press, pp 201-216.

Wolter, K. (1979), "Composite Estimation in Finite Populations," *Journal of the American Statistical Association,* 74, pp. 604-613.