APPLICATIONS OF SUPERPOPULATION MODELS TO ESTIMATION IN REPEATED SURVEYS

Stephen M. Woodruff, Bureau of Labor Statistics
441 G St. N.W., Room 2128, Washington D.C. 20212

1. INTRODUCTION

A best linear unbiased estimator (BLUE) is derived using least-squares. This BLUE makes use of two types of supplementary information. The first type consists of sample survey results from prior time periods and the second type is a superpopulation model which links these prior survey variables to the survey variables of current interest. This BLUE can also be used as an index estimator and evolves naturally as an alternative to superpopulation prediction theory (SPT) in cases where the auxiliary variable is not known for all members of the sampling frame.

When a superpopulation model can be confidently specified, it can be inefficient to base inferences on the sampling distribution. In such cases it is normally better to condition on the sample which was selected and apply SPT. There is a growing body of empirical evidence to support this approach to survey inference, for example, Royall and Cumberland (1981), Royall and Herson (1973), Bardsley, P. , and Chambers, R.C. (1984), and Ericson, W.A. (1969).

The superpopulation method suggested in this paper is not prediction but rather estimation of parameters in a superpopulation model. For moderate to large size universes this method is essentially solving the same problem as predicting a finite population mean.

This paper applies a multivariate analog of the regression superpopulation model which stochastically links a random variable Y_i to a known auxiliary variable X_i and is given by:

$$Y_i = \beta X_i + \varepsilon_i \quad \text{for} \quad i=1,2,\ldots\ldots,N$$

Where the $\{\varepsilon_i: 1\le i\le N\}$ are pairwise uncorrelated, $\mathcal{E}(\varepsilon_i)=0$ for all i, $\mathcal{V}(\varepsilon_i)=\sigma^2 G(X_i)$, the function G and the set $\{X_i: 1\le i\le N\}$ are known, β and σ^2 are unknown. Y_i is observed for each member of a sample, s_y, of n units and $\bar{Y}$ is to be estimated where:

$$\bar{Y} = (1/N)\sum_{i=1}^{N} Y_i$$

This model is discussed and referenced by Cassel, Särndal, and Wretman (1977). The best linear unbiased estimator (BLUE) for $\bar{Y}$ (the universe mean) under this model reduces to the ratio estimator when G(X)=X. When $G(X)=X^2$ and the sampling fraction is small the BLUE under this model is very nearly the Horvitz-Thompson estimator under PPX sampling: Cassel, Särndal, and Wretman (1977) page 120. Apparently this superpopulation model can be used to justify several common sampling design based estimators.

We will be interested in applications of this regression model where only a subset of the set of auxiliary variables, $\{X_i: 1\le i\le N\}$ is known. This situation may arise when this auxiliary data becomes stale with passing time and only a portion of these X-values are sufficiently recent to be reliable. It may also arise when X_i is the same datum as Y_i but at some prior time period and therefore is known only for the units which were sampled at that prior time.

Let s_x denote the subset of known (or reliable) X-values. The naive superpopulation model will be used to relate the values of the X variable in s_x to these values in the entire universe. (i.e. a superpopulation model which is often implicitly assumed in cases of complete ignorance about the X-values outside s_x as well as how s_x was selected). This naive superpopulation model essentially assumes that the set $\{X_i: 1\le i\le N\}$ are outcomes of N iid random variables from a distribution function, F, with finite second moment. This superpopulation model expresses the assumptions underlying the use of the sample mean to estimate the population mean when nothing is known about how the sample was selected or any other relation between the sample and the universe. This particular superpopulation model can be inefficient if not disastrous for inferential purposes when more information about the process which generates the data is known and this information belies this naive model. Under the naive model the set $\{X_i: 1\le i\le N\}$ consists of exchangeable random variables and thus as Ericson (1969) suggests, simple random sampling is an appropriate sampling strategy for making inferences about the universe mean $\bar{X}$ or universe total, X. Exchangeability also implies that (for inferential purposes) s_x is essentially a simple random sample regardless of its method of selection. Let the first two moments of F be μ and σ_x^2 then:

$$X_i = \mu + \delta_i \quad \text{for } i=1,2,3,..,N$$

where the set $\{\delta_i: 1\le i\le N\}$ consists of iid random variables with $\mathcal{E}(\delta_i)=0$ for all i and $\mathcal{V}(\delta_i)=\sigma_x^2$ for all i and the δs are independent of the εs, μ is unknown, and σ_x^2 is unknown.

Then unconditionally $\mathcal{E}(Y_i)=\beta\mu$, and its variance is:

$$\sigma_y^2 = \sigma^2\mathcal{E}(G(X_i)) + \beta^2\sigma_x^2,$$

$$\text{and } \mathrm{Cov}(Y_i,X_i) = \beta\sigma_x^2.$$

Let: $\eta=\mathcal{E}(Y_i)$. In matrix notation, the random vectors $(Y_i,X_i)'$ for $i=1,2,\ldots\ldots,N$ are iid with mean and covariance matrix given as:

$$\theta = \begin{bmatrix}\eta\\ \mu\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}\sigma_y^2 & \beta\sigma_x^2\\ \beta\sigma_x^2 & \sigma_x^2\end{bmatrix} \quad \text{respectively.}$$

If this covariance matrix is known then the least squares estimator of θ together with the variance of this least squares estimator are available once the size of s_x, size of s_y, and size of the overlap between s_x and s_y are known. When this covariance matrix must be estimated then its form under this bivariate superpopulation model will suggest a more precise covariance estimate ($\hat{\beta}\hat{\sigma}_x^2$) than is available from the usual sample covariance estimator based on the sample units common to both s_x and s_y (which is often denoted s_{xy}). Since least squares BLUEs as well as composite estimators are usually dependent upon good variance and covariance estimates the use of a superpopulation model to derive these second moment estimates should be expected to give (and indeed does give) superior BLUE estimators.

The rest of this paper will derive and test the k-dimensional version of these BLUEs. One particularly attractive feature of these BLUEs is that under the multivariate regression model given in the following sections, the covariance matrices are directly algebraically invertible and these inverses have a relatively simple form (tri-diagonal). This simplicity of form can contribute to computational simplicity which will further enhance precision.

The multivariate superpopulation model to be discussed in this paper does seem to capture the essential structure of many common data sources at the Bureau of Labor Statistics. Simulations show that its application to estimation can yield results that are significantly superior to composite estimation and non superpopulation based least squares BLUEs.

2. A MULTIVARIATE SUPERPOPULATION MODEL

The multivariate superpopulation model to be considered here is given as:

$$Z_i = \theta + \Delta_i \quad \text{for } i=1,2,3,\ldots\ldots\ldots,N.$$

Where Z_i, θ, and Δ_i are k-dimensional vectors, the set of random vectors $\{\Delta_i\}$ are mutually independent each with expectation the zero vector and with covariance matrix, Σ_z.

This paper addresses the problem of estimating the components of θ or linear functions of these components given sample data on units where some subset of the k components are observed for each sample unit. Some units may have observations on all the k components of Z_i, and some sample units will have data on only a proper subset of these k components.

Let u be a subset of $\{1,2,3,\ldots\ldots,k\} = I_k$. Let s(u) denote the set of units in the sample each of which has data on exactly those components of the Z-vector which are contained in u. Note that the universe is the disjoint union of all these 2^k subsets. Let n(u) denote the size of (the number of units in) s(u). Let c(u) be the number of elements in u (an integer between 0 and k). Let z(u) be the n(u)·c(u)×1 column vector consisting of n(u), c(u)×1, column subvectors for each of the n(u) sample units in s(u). Each of these n(u) subvectors contains the observed data for the components of the Z-vector corresponding to the elements in u. The ordering within the above vectors and subvectors will be from smallest component or unit to largest. That is, the first subvector of z(u) is the smallest element of s(u) (smallest i $1\le i\le N$ such that $i\varepsilon s(u)$) and the first component of that subvector the smallest element of u and so on.

For the remainder of this paper, the vector Z_i, $1\le i\le N$, will represent the data for unit i for each of the k time periods. For a core sample s_c of size m (=$s(I_k)$ in the above notation), data for each of the k time periods is collected. In addition, a supplemental sample of previously unsampled units is collected at each time t ($1\le t\le k$) and data for time t is observed. This supplementary sample is of size n-m and is denoted s_t (=$s(\{t\})$ in the above notation). Using the sample data from the k time periods and the superpopulation model we will be interested in estimating the k^{th} component of θ (the mean for the current time period).

This sample data can be written as the column vector, Y. The transpose of Y is: $Y' = (z(\{1\})',z(\{2\})',\ldots\ldots\ldots\ldots,z(\{k\})',z(I_k)')$ where $(\cdot)'$ denotes transpose. Then:

$$Y = X\theta + \varepsilon \qquad (2.1)$$

where Y is kn×1, X is kn×k, and ε is kn×1. The transpose of the design matrix is:

$X' = (X_1',X_2',X_3',\ldots\ldots\ldots,X_k',M')$ where X_j is an (n-m)×k matrix of zeros everywhere except in the j^{th} column which contains all ones. M is km×k, consisting of m, k×k identity matrices stacked vertically. Defining the components of Δ_i by the equation: $\Delta_i' = (\delta_{i1},\delta_{i2},\ldots\ldots,\delta_{ik})$, we can write the transpose of ε as:

$\varepsilon' = (\delta(\{1\})',\delta(\{2\})',\ldots\ldots\ldots,\delta(\{k\})',\delta(I_k)')$

where $\delta(u)$ is defined analogously to z(u).

The dispersion matrix of ε, Σ, is a block diagonal matrix function of Σ_z and the sampling scheme. It has Σ_z in the last (lower right) m, k×k blocks and the appropriate variance (diagonal element of Σ_z) in the upper left diagonal. If we let $\Sigma_z = (a_{ij})$ then these upper left block diagonal elements of the dispersion matrix of ε consist of $a_{11}I,a_{22}I,\ldots,a_{kk}I$ where I is the (n-m)×(n-m) identity matrix.

To apply least-squares to the sample data to get the BLUE of θ it only remains to specify Σ_z. In the next section a superpopulation model which yields a specific form of Σ_z will be considered.

3. THE MULTIVARIATE REGRESSION MODEL

Letting $\Sigma_z = (a_{ij})$, the BLUE and its variance will be derived for Σ_z given by:

$$a_{ij} = \begin{cases} \sigma_i^2 \prod_{\ell=i+1}^{j} \beta_\ell & \text{for } j>i \\ \sigma_i^2 & \text{for } i=j \\ \sigma_j^2 \prod_{\ell=j+1}^{i} \beta_\ell & \text{for } j<i \end{cases} \qquad (3.1)$$

where σ_i^2, i=1,2,.....,k and β_i, i=2,3,.....,k are positive real numbers. This form for the covariance matrix is the result of the following superpopulation model. Let

$Z_i' = (z_{i1}, z_{i2}, \ldots\ldots, z_{ik})$ i=1,2,....,N

where $z_{i1} = \beta_1 + \delta_{i1}$

and $z_{ij} = \beta_j \cdot z_{ij-1} + \delta_{ij}$ for j=2,3,...,k (3.2)

and the $\Delta_i = (\delta_{i1}, \delta_{i2}, \ldots\ldots, \delta_{ik})$ are iid random vectors with mean that is the zero vector. Then the covariance matrix of Δ_i, Σ_z, is necessarily of the form in 3.1 above.

If the components of Z_i are measurements on the same datum at different time periods then this type of linear relationship will often provide an adequate description for the underlying process which generates the data at business establishment level in many of the Bureau of Labor Statistics surveys. Two special cases of 3.1 will now be analyzed in detail.

Case 1: $\sigma_i^2 = \sigma^2$ for all i, $1 \le i \le k$ and $\beta_i = \rho$ for all i, $1 \le i \le k$, $0 < \rho < 1$.

Case 2: a_{ij} as specified in 3.1. Letting $\theta' = (\mu_1, \mu_2, \ldots\ldots, \mu_k)$, the problem of finding the BLUE for the current mean, μ_k, given the superpopulation model and the sampling plan described in the last section will be solved for these two cases.

Case 1 arises when $a_{ij} = \sigma^2 \cdot \rho^{|i-j|}$ for all i and j. This case does not arise naturally from the superpopulation model, 3.2, unless extreme limitations are placed on the superpopulation parameters in this model. This case is still useful because the estimator it suggests appears to be extremely robust. That is, it works well in many commonly found instances where case 2 more accurately models the data. The BLUE for μ_k is the last component of the vector:

$\hat{\theta} = (X'\Sigma^{-1}X)^{-1}(X'\Sigma^{-1}Y)$ where X, Σ, and Y are defined in the last few paragraphs of section two ($\hat{\theta}$ is the generalized least squares estimate of θ in 2.1). For case 1 this last component can be written as:

$$\hat{\mu}_k = \left\{(1-\rho^2)/nD_k\right\} \cdot \sum_{i=1}^{k} D'_{i-1}\alpha^{k-i}\rho^{k-i}H_i$$

where $H_i = \left[z_s(i) + (1/\{1-\rho^2\})T_i\right]$ and $T_i =$

$$\left[-\rho z_c(i-1) + (1+q_i\rho^2)z_c(i) - \rho z_c(i+1)\right]$$

and where:

$\alpha = m/n$

$z_s(i) = \sum_{j \varepsilon s_i} z_{ji}$ $\qquad$ $z_c(i) = \sum_{j \varepsilon s_c} z_{ji}$

for i=1,2,3,.....,k.

$z_c(0) = z_c(k+1) = 0$ and $q_i = \begin{cases} 0 \text{ if } i=1 \text{ or } k \\ 1 \text{ otherwise} \end{cases}$

Define $D_0=1$. D_j and D'_j for $1 \le j \le k$ are given as:

$$D_j = \left[(A_2-1)^2A_1^{j-1} - (A_1-1)^2A_2^{j-1}\right]/(A_1-A_2)$$

$$D'_j = \left[(A_1-1)A_2^j - (A_2-1)A_1^j\right]/(A_1-A_2)$$

with $A_1 = (1/2)\left[1 + \rho^2 + [(1+\rho^2)^2 - 4\rho^2\alpha^2]^{\frac{1}{2}}\right]$

and $A_2 = (1/2)\left[1 + \rho^2 - [(1+\rho^2)^2 - 4\rho^2\alpha^2]^{\frac{1}{2}}\right]$. The variance of $\hat{\mu}_k$ is given by:

$$V(\hat{\mu}_k) = (D'_{k-1}/D_k)\cdot(\sigma^2(1-\rho^2)/n).$$

Everything in $\hat{\mu}_k$ is a function of known sample design parameters and two unknown superpopulation parameters which will normally have to be estimated from the data, ρ and σ^2.

The estimation of these two parameters may be relatively easy as well as precise and this may explain why this estimator of μ_k is robust. In case 2 the quantity to be estimated is $\mu_k = \beta_1\cdot\beta_2\cdot\beta_3\cdot\ldots\ldots\cdot\beta_k$ under 3.2 and the vector θ of unconditional means is:

$\theta' = (\mu_1, \mu_2, \mu_3, \ldots\ldots, \mu_k) =$

$(\beta_1, \beta_1\cdot\beta_2, \beta_1\cdot\beta_2\cdot\beta_3, \ldots\ldots, \beta_1\cdot\beta_2\cdot\ldots\cdot\beta_k)$.

The BLUE for β_i (i=2,3,4, ...,k) under (3.2) is approximately

$$\hat{\beta}_i = z_c(i)/z_c(i-1) \text{ where } z_c(i) = \sum_{j \varepsilon s_c} z_{ji}.$$

$\hat{\beta}_i$ is exactly the BLUE when the conditional variance of δ_{ji} given z_{ji-1} is proportional to z_{ji-1}. 3.2 implies that $\hat{\beta}_i$ is both unbiased and consistant for β_i under very mild conditions.

An unbiased estimator for σ_i^2 is:

$$\hat{\sigma}_i^2 = (1/(n-1)) \sum_{j \varepsilon s_i \cup s_c} (z_{ji} - \bar{z}_i)^2$$

where $\bar{z}_i = (z_c(i)+z_s(i))/n$ and $z_s(i) = \sum_{j \varepsilon s_i} z_{ji}$.

Note that 3.2 implies that the covariance matrix Σ_z has the form given by 3.1 which is a function of only β_i i=2,3,.....,k and σ_i^2 i=1,2,3,.....,k. Replacing these superpopulation parameters with their respective estimates in Σ^{-1} gives an estimate of Σ^{-1} which is noticeably more precise than the estimator for the inverse of a covariance matrix given by:

$$\left[(1/(m-1)) \sum_{j \varepsilon s_c} (Z_j - \bar{Z})(Z_j - \bar{Z})'\right]^{-1} \quad (3.3)$$

The covariance matrix given in 3.1 can be inverted algebraically and this inverse is tridiagonal. This computationally pleasing

property means that the k^{th} component of $(X'\Sigma^{-1}X)^{-1}(X'\Sigma^{-1}Y)$ can then be written in its explicit algebraic form which can then be estimated by substituting $\hat{\beta}_i$ and $\hat{\sigma}_i$ for β_i and σ_i $i=1,2,\ldots\ldots,k$.

The BLUE for μ_k and the variance of this estimator require some notation.

Let $C_i = \sigma_i^2 - \sigma_{i-1}^2\beta_i^2$ $i=2,3,\ldots\ldots,k$

and $C_1 = \sigma_1^2$.

Let $G_i = -m\beta_i/C_i$ $i=2,3,\ldots\ldots,k$.

Let $F_i = (n-m)(1/\sigma_i^2) + m((1/C_i) + (\beta_{i+1}^2/C_{i+1}))$ $i=1,2,3,\ldots\ldots,k-1$

and $F_k = (n-m)(1/\sigma_k^2) + m/C_k$.

Let $B_0 = 1$, $B_1 = F_1$, and

and $B_i = F_iB_{i-1} - G_i^2B_{i-2}$ for $i=2,3,4,\ldots\ldots,k$.

Let $x_i = (B_{i-1}/B_k)\prod_{j=i+1}^{k}(-G_j)$ $i=1,2,3,\ldots\ldots,k$.

Let $w_1 = (1/\sigma_1^2)z_s(1) + ((1/C_1) + (\beta_2^2/C_2))z_c(1) - (\beta_2/C_2)z_c(2)$

and $w_i = (1/\sigma_i^2)z_s(i) - (\beta_i/C_i)z_c(i-1) + ((1/C_i) + (\beta_{i+1}^2/C_{i+1}))z_c(i) - (\beta_{i+1}/C_{i+1})z_c(i+1)$ $i=2,3,4,\ldots\ldots,k$

and $w_k = (1/\sigma_k^2)z_s(k) - (\beta_k/C_k)z_c(k-1) + (1/C_k)z_c(k)$.

The BLUE for μ_k can finally be given as:

$$\hat{\mu}_k = \sum_{i=1}^{k} w_ix_i$$

and its variance is $V(\hat{\mu}_k) = B_{k-1}/B_k$.

A natural estimate for its variance is $V(\hat{\mu}_k)$ with $\hat{\beta}_i$ and $\hat{\sigma}_i^2$ substituted for σ_i^2 and β_i in its formula.

It should be noted that all of the derivations in this section which are dependent upon the sampling scheme described in section two. can be easily if tediously repeated for quite general sampling plans which involve far more complicated patterns of overlap and may also incorporate nonresponse as will. In spite of the total sample size, the matrix, $X'\Sigma^{-1}X$, can be algebraically derived and it is k×k independent of this total sample size. The same is true of the k-vector $X'\Sigma^{-1}Y$ and thus potential computer related difficulties encountered with the inversion of large matrices may be avoided.

4. SOME EMPIRICAL RESULTS

In this section the estimators which were derived in the last section are compared empirically using replications of simulated data which are generated according to model 3.2 and some perturbations from this model. The tables in this section compare four estimators. These four estimators are defined and denoted as follows:

a) M_1 is $\hat{\mu}_k$ under case 1.

b) M_2 is $\hat{\mu}_k$ under case 2.

c) M_s is the sample mean of the data for time k from the units in $s_k \cup s_c$.

d) M_e is $\hat{\mu}_k$ under case 2 where the exact value of Σ is used in place of its estimated value.

The entries in the following tables are estimated mean squared errors from 100 replications of the process which generates the universe data, the sample selection, and the estimators. Thus for a particular estimator, M_j, j=1,2,s, or e, the quantity tabulated is:

$$(1/100)\sum_{r=1}^{100}(M_{jr} - U_r)^2$$

where r is the replicate number and where U_r is the universe mean in the r^{th} replication.

The estimator of Σ using (3.1) is roughly twice as precise (half the average squared error) as the "often used" estimator of a covariance matrix given in (3.3). It is also apparent that M_1 is pleasantly robust even in cases of relatively severe deviations from the assumptions on which it is based ($\beta_i = \beta < 1$ for all i & $\sigma_i^2 = \sigma^2$ for all i).

Table 1.

$\vec{\beta} = (1.0,1.01,0.99,1.00,1.02,1.01)$

k	M_s	M_e	M_1	M_2
3	.62	.36	.39	.42
	.64		.41	.44
4	.62	.29	.36	.34
	.67		.37	.34
5	.96	.44	.53	.49
	1.10		.61	.55
6	.80	.46	.56	.49
	1.02		.75	3.19
7	.86	.44	.65	.72
	1.12		.89	.94

Table 2.

$\vec{\beta} = (.98,1.00,0.99,0.97,1.00,0.99)$

k	M_s	M_e	M_1	M_2
3	.59	.34	.37	.40
	.60		.38	.42
4	.60	.28	.35	.32
	.61		.45	.33
5	.87	.40	.49	.44
	.91		.52	.48
6	.71	.41	.49	.44
	.77		.56	390.47
7	.74	.39	.57	.60
	.74		.60	2.05

Table 3.

$\vec{\beta}$ = (1.13,1.21,1.09,1.12,1.05,1.23)

k	M_s	M_e	M_1	M_2
3	1.05	.56	.70	.68
	1.24		.88	.79
4	1.20	.51	.61	.71
	2.18		1.38	.96
5	2.23	.90	1.22	1.11
	7.25		5.67	4.08
6	1.88	.96	1.33	1.34
	17.15		22.52	10.89
7	2.75	1.25	1.97	4.24
	95.34		141.34	66.74

Table 4.

$\vec{\beta}$ = (1.13,1.08,.98,1.12,1.06,1.07)

k	M_s	M_e	M_1	M_2
3	.85	.47	.53	.56
	1.02		.68	.66
4	.81	.38	.43	.45
	1.31		.69	.58
5	1.56	.65	.84	.77
	3.26		2.03	1.61
6	1.34	.72	.94	8.62
	4.93		4.98	2.83
7	1.57	.76	1.09	1.61
	11.55		13.07	7.39

Table 5.

$\vec{\beta}$ = (1.03,1.02,.98,1.02,1.01,1.07)

k	M_s	M_e	M_1	M_2
3	.66	.38	.41	.44
	.71		.46	.48
4	.65	.31	.37	.35
	.76		.41	.35
5	1.04	.47	.57	.52
	1.33		.73	.64
6	.85	.48	.58	.52
	1.28		.95	1.38
7	1.01	.50	.75	.86
	1.64		1.35	1.23

For each value of k and each estimator (except M_e) there are two entries in the tables. The upper entry is the estimated mean square error when the hypothesized model (3.2) holds and the lower entry is estimated mean square error when bias is introduced into the data generation. The reason M_e has but one mean square error estimate is the difficulty of finding the exact value of Σ when the bias is introduced into the data generation.

For the forgoing tables the population parameters are as follows:

N = 2000, n = 60, m = 20, β_1 = 50, σ_1^2 = 25,

The vector $(\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7)$ is given in each Table.

$z_{i1} = \beta_1 + \delta_{i1}$ where $\delta_{i1} \sim N(0,\sigma_1^2)$

$z_{ij} = \beta_j z_{ij-1} + \delta_{ij}$ where $\delta_{ij} \sim N(0,(1/9)z_{ij-1})$

for i=1,2,3,.....,N=2000 and j=2,3,.....,k.

The bias perturbation is accomplished by adding $(50 - z_{ij-1})^2(.007)$ to z_{ij} in the above expressions for z_{ij}.

All of the above δ_{ij} are stochastically independent.

For M_1, the parameters β and σ^2 were estimated as:

$\hat{\beta}$ = .99 and

$$\hat{\sigma}^2 = (1/k)\sum_{i=1}^{k} \hat{\sigma}_i^2$$

It is apparent from these tables that in spite of mild to severe model failure the case 1 estimator, M_1, is pleasantly robust. The actual BLUE, M_e, (the BLUE given the actual Σ_z) is generally only slightly better than M_1 and M_2. Thus relatively little is lost in estimating this covariance matrix using the information given by the superpopulation models for the two cases.

The sample mean at time k has roughly twice the mean square error of the three versions of the least squares estimator. This ratio of mean square errors does not seem to vary greatly with k and this suggests that a k of three or four will suffice to give a good current estimate at least for the data of these simulations. Apparently sample data more than three or four time periods in the past carries little information about the present.

The correlation between data from adjacent time periods in these simulations is roughly .9. Composite estimation under this correlation structure can be expected to give a variance ratio of about .8 (the ratio of the composite estimator to M_s). Note that this ratio for M_1 and M_2 generally runs between .5 and .7 for k≤5.

5. CONCLUSIONS

Both the simulation results and the relative tractability of Σ_z under the regression superpopulation model support this superpopulation approach to estimation in repeated surveys where the model given by 3.2 captures the essential behavior of the sampling universe. There are several repeated surveys at the Bureau of Labor Statistics (BLS) which can be adequately modelled by 3.2 and more simulations will be run to test M_1 and M_2 on some of these data series. In particular, M_1 and M_2 should be compared to the standard forms of composite estimators and design based BLUEs (see page 156, Raj, for the design based BLUE when k=2). The simulations summarized in section four show that M_1 and M_2 can be strong (if not overwhelming) competitors to composite estimation and design based BLUEs. This strength is not surprising given the additional stochastic structure (superpopulation model) used to derive M_1 and M_2.

The reduction in mean square error of M_1 and M_2 comes in part from improved estimates for second moments (covariance matrix). The

precision of the covariance matrix estimate that is needed for application of generalized least squares can often be the fatal weakness of generalized least squares estimators. The multivariate regression superpopulation model gives a covariance matrix as a function of a reduced number of superpopulation parameters all of which can be easily and accurately estimated. The result is an estimated covariance matrix with less than half the mean square error of the estimate given by 3.3.

In addition to being easier to estimate this covariance matrix, Σ_z, is directly algebraically invertible and its inverse is tridiagonal. This feature simplifies the computational problems sometimes associated with generalized least squares.

Two types of data linkage are used in the derivation of M_1 and M_2. The first type is the data linkage given by the sample overlap between time periods. This overlap permits the estimation of superpopulation parameters that require sample units with data for several time periods. The second form of linkage is the superpopulation model that stochastically relates the data for a given unit at different times and thus models the information contained in the historical data that is relevant to estimation for the current time period.

The regression superpopulation model together with the sample overlap carry the essential information which is pertinent to the least squares estimation setup. This information takes the form of the linear relationship given by 2.1. From this linear relationship the least squares BLUE of θ is derived.

Applications of this superpopulation model to index estimation or estimation of linear combinations of time period means follow directly from the Gauss-Markov theorem. This theorem also gives a variance expression for estimators of a given index or linear combination of time period means.

Some topics that were not addressed by this paper but need consideration are as follows. The variance estimators suggested for M_1 and M_2 need to be tested. BLUEs under different patterns of sample overlap can be derived. Simulation studies on a greater variety of data sets need to be done. These will be addressed in a future paper.

I hope that this paper has suggested some useful approaches to improved estimation in repeated surveys.

REFERENCES

1) Bardsley, P. and Chambers, R.C. (1984), "Multipurpose Estimation from Unbalanced Samples," Applied Statistics, 33, 290-299.

2) Cassel,C. , Särndal, C. and Wretman, J.H. (1977), Foundations of Inference in Survey Sampling, John Wiley & Sons.

3) Ericson, W.A. (1969), "Subjective Bayesian Models in Sampling Finite Populations I," Journal of the Royal Statistical Society, Ser. B., 31, 195-234.

4) Godambe V.P., and Thompson M.E. (1986), "Parameters of Superpopulation and Survey Population: Their Relationships and Estimation," International Statistical Review, 54, 127-138.

5) Raj D. (1968), Sampling Theory, McGraw-Hill.

6) Royall, R.M., and Cumberland. W.G. (1981a), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," Journal of the American Statistical Association, 76, 66-77.

7) Royall, R.M., and Cumberland, W. G. (1981b), "The Finite-Population Linear Regression Estimator and Estimators of its Variance _ An Empirical Study," Journal of the American Statistical Association, 76, 924-930.

8) Royall, R. M., and Herson, J. H. (1973), "Robust Estimation in Finite Populations I," Journal of the American Statistical Association, 68, 880-889.

9) Scott A.J., and Smith T.M.F. (1974), "Analysis of Repeated Surveys Using Time Series Methods," Journal of the American Statistical Association, 69, 674-678.

10) Tenenbein, A. (1970), "A Double Sampling Scheme for Estimating from Misclassified Binomial Data," Journal of the American Statistical Association, 65, 1350-1361.