# ONE WAY TO BUILD AN ESTIMATOR WITH APPLICATIONS TO SAMPLING THEORY

Steve Woodruff, Bureau of Labor Statistics
Room 4985, 2 Massachusetts Ave. N.W. , Washington D.C. 20212

Key Words: Nonresponse, Response Error, Multivariate Normal

## 1) INTRODUCTION

In survey sampling problems data is usually being collected on many study variables, some of which (variables of primary interest) are positively correlated with the sample design variable(s) and other variables which are not (peripheral variates). A sample design which is nearly optimal for estimating (via Horvitz-Thompson, HT) the means of those variates which are positively correlated with the design variable may be extremely inefficient for estimating (via HT) means of peripheral variates. This problem can be addressed by post stratifying on one or more other variates which are correlated with the peripheral variates being gathered from the sample units. This paper describes a way of building an estimator, which makes use of these other sources of data together with the relationships between these data items to reduce the error of estimates of peripheral variate means.

The class of estimators to be built are generally multivariate and utilize data dependencies that can be captured in a covariance matrix. These estimators minimize MSE in the presence of nonresponse, response bias, and weak relationships with the sample design variables. Since they directly use the type of data dependencies that are exploited by composite, Bayes, ratio, and regression estimation, they can achieve, by default, the same reductions in mean square error that these techniques also provide. The same is true about nonresponse adjustment, the information that is used to perform this operation is also included in the estimation process.

A vector of finite population means (target means) of the study variables (target variables) is to be estimated under the following setup.

1) A sampling frame which identifies members of the finite population and contains quantitative data on a fixed set of characteristics (auxiliary variables) for each population unit.

2) From this sampling frame, a sample is selected. This sample may be selected by a known randomization procedure, possibly a function of the auxiliary data. A stratified, clustered, multistage sample is the general rule.

3) For each member of the sample, some subset of the target variables is observed and this subset may vary from sample member to sample member (item nonresponse). Let $a_i$ be the row vector of auxiliary variables attached to the $i^{th}$ member of the population and let $t_i$ be the row vector of target variables attached to the $i^{th}$ member of the population. Let $y_i = (a_i , t_i)$, then $a_i$ is known for all $i\varepsilon U$ (the population), $t_i$ is unknown for all i not in the sample (s), and for each $i\varepsilon s$ some subset of the components of $t_i$ is known and this subset varies from sample member to sample member (item nonresponse).

4) Suppose it is appropriate to describe every $y_i$ (whether its components are observed or not) in the population as the outcome of a vector valued random variable ,$Y=(A,T)$, with mean $\mu = (\mu_A, \mu_T)$ and variance/covariance matrix $\Sigma$ . Thus the $\{y_i\}$ are the outcomes of independent and identically distributed random variables $\{Y_i\}$ , each distributed as $Y \sim (\mu, \Sigma)$.

Using only the data outlined in 1) through 3) and data relationships described in 4), an estimator of the target means ($\mu_T$) that minimizes expected squared distance between itself and $\mu_T$ is constructed. Expectation is over the distribution generated by the distribution of Y (which can include the sampling distribution and, if known, the response mechanism).

The solution described here is a direct application of the Gauss Markov theorem. The estimator is linear in the available data and unbiased in the appropriate space. With these constraints, knowledge of second moments of Y ($\Sigma$) will be enough to construct a best (minimum variance) solution to the problem of estimating $\mu_T$. This solution is also Maximum Likelihood given the missing sample data (see Little & Rubin 1987) and a fixed point of the EM-algorithm.

In repeated surveys, $\Sigma$ may be estimated from historical data and relatively stable over time as the expected values of the targets and auxiliaries change. In such cases, it may be appropriate to use this estimate of $\Sigma$ as the actual quantity. $\Sigma$ may be derived from superpopulation models, sampling distributions, or combinations of both. The sample indicator function and possibly the response indicator functions may be treated as auxiliary (poststratification) variables. Using sample and response indicator variables in this way provides the sampler with a way to adjust for nonignorable non response (& response bias) in cases where the nonresponse mechanism can be given an accurate stochastic description.

Result 1 in section 3 quantifies the effect on MSE of deleting some of the target variables from the data structure, described above for Y. Result 2 in section 3 quantifies the effect on MSE of deleting auxiliary variables from Y. The theory presented here is applied in the Johnson, Woodruff (1990) paper.

## 2) BUILDING AN ESTIMATOR

**a) An Example-** Suppose, the problem is to estimate the number of production workers and women workers in a small industry. A simple random sample of three of the 26 firms in this industry was selected but the firms in this sample were less than cooperative. Only one firm provided both the numbers of its production workers (P) and its women workers (W), one gave data on only women workers and one gave only production workers. In addition, the total employment (E) in each firm in this industry is known as is the matrix of variances and covariances for and between the random variables describing total employment (E), women workers (W) and production workers (P). This sample data (outcomes of (E,W,P)) are:

|  | E | W | P |
|---|---|---|---|
| 1 | 92 | 35 | 71 |
| 2 | 90 | 44 | -- |
| 3 | 85 | -- | 75 |

The average employment in this industry is 76 and the variance/covariance matrix of (E,W,P) is:

$$\begin{pmatrix} 205 & 131 & 170 \\ 131 & 101 & 120 \\ 170 & 120 & 190 \end{pmatrix}$$

Note that all three employment variables are positively correlated. Employment (E) in this example is the auxiliary variable and the pair, (Women Workers, Production Workers) = (W,P) are the target variables.

Suppose (E,W,P) are approximately normally distributed. Since the average employment in the industry is 76, the data on the variables of interest from all three sample members are positively biased and everything needed to compute the bias (conditional on the known Employment data) in the observed values of W and P for each sample member is at hand. These biases (under normality) for a sample member are the sample member's employment minus 76 times .64 for W and times .83 for P (for example, the bias in W for firm one is 10.24 and the bias in P is 13.28). The coefficients, .64 and .83, come from the expression for conditional expectation of the targets given the auxiliary variable under multivariate normality. The vector (.64, .83) results from multiplying the reciprocal of the (1,1) component of the covariance matrix , 1/205 , by the row, (131,170).

Subtract these biases from the observed measurements on the targets to get bias adjusted observations.

| Establishment | W | P |
|---|---|---|
| 1 | 24.76 | 57.72 |
| 2 | 35.04 | — |
| 3 | ---- | 67.53 |

These bias adjusted variables are now more appropriately thought of as outcomes of iid random vectors with the target mean as their common expected value and with a common variance/covariance matrix that measures a smaller dispersion around this target mean. This new (conditional) variance/covariance matrix is $\begin{pmatrix} 17.29 & 11.37 \\ 11.37 & 49.02 \end{pmatrix}$ as opposed to $\begin{pmatrix} 101 & 120 \\ 120 & 190 \end{pmatrix}$ prior to conditioning on E.

The responding item means of the adjusted data would be unbiased estimates of the target means. These target means may also be estimated by imputing for the missing data items and taking the column means of both adjusted and imputed data. This second estimate of the target means would have a smaller variance than the first. Finally, by applying the EM-algorithm under normality to reimpute and reestimate until convergence a third estimate is obtained with a smaller variance than either of the two above. In this example, this estimate under the EM-algorithm converges to (30.288, 64.188). When the variance/covariance matrix is known (as in this example) and must not be reestimated with each iteration of the EM-algorithm, there is a direct closed form solution that yields exactly the same result as the EM-algorithm. This is done via the Gauss theorem on minimum variance estimation by using the following linear relation between the bias adjusted data and the target mean, $\mu_t$.

$$\begin{pmatrix} 24.78 \\ 57.72 \\ 35.04 \\ 67.53 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \mu_T + \varepsilon$$

where $\mathrm{Var}(\varepsilon) = \begin{pmatrix} 17.29 & 11.37 & 0 & 0 \\ 11.37 & 49.02 & 0 & 0 \\ 0 & 0 & 17.29 & 0 \\ 0 & 0 & 0 & 49.02 \end{pmatrix}$

and $E(\varepsilon)=0$.

Rewriting this as $,Y = X\mu_T + \varepsilon$ where $\mathrm{Var}(\varepsilon) = \Sigma$, we get: $\hat{\mu}_T = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Y =$ (30.288, 64.188)' , exactly as with the EM-algorithm.

The observed data maximum likelihood estimator (MLE) as described in Chapter 7 of Little & Rubin is, under normality, the GLS estimator, $\hat{\mu}_T = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} Y$. This estimate is also a fixed point of the EM-algorithm by the theorems in the theory section of this chapter. Thus the Normal EM-algorithm with $\Sigma$ known will converge to the GLS given above.

**b) Generalization of the Example in a)-**

There are two phases to estimation once the set of appropriate target variables and auxiliary variables has been determined. Phase one is poststratification on the auxiliary variables (using the regression adjustments under conditionality). Phase two is equivalent to imputation and estimation without explicitly going through the computational rigors of the EM-algorithm.

The available data on the population of interest is represented as the partial outcome of an $N x(k_a+k_t)$ matrix W of random variables. N= the number of units in the population, $k_a$ = the number of auxiliary variables, $k_t$ = the number of target variables, and n= the sample size. The i$^{th}$ row of W contains the random variables (auxiliaries and targets) associated with the i$^{th}$ population unit.

By definition, the auxiliary outcomes are known for all population units and the target variable outcomes are observed only for sample units and, due to nonresponse, we may observe only a subset of the $k_t$, targets for each sample unit and this subset varies from unit to unit..

Let $Y_i = (A_i , T_i)$, the i$^{th}$ row vector of W containing the two random vectors $A_i$ , the $1xk_a$ vector of auxiliaries (outcomes known), and $T_i$, the $1xk_t$ random vector of target variables, the outcomes of which may be unknown (if i is not a sample member) or partially known if i is a sample member. From this point on, redefine i slightly to denote only sample members. i runs from 1 to n (the sample size) since only sample members will be considered in what follows.

The observed targets and all the auxiliary variables for the i$^{th}$ sample unit are the components of the realization of the random vector, $Y_i^o = (A_i, T_i\chi_i)$, where $\chi_i$ is the response indicator matrix for the row vector of target variables $T_i$ attached to the i$^{th}$ sample unit. $\chi_i$ is constructed from the identity matrix of order $k_t$ (the number of target variables) by deleting each column j of this identity matrix for which the j$^{th}$ target variable is a nonresponse. Thus $Y_i^{\circ}$ contains the i$^{th}$ units' auxiliary variables and only those target variables observed for the i$^{th}$ unit. Letting I be the identity matrix of order $k_a$ (the number of auxiliary variables) and $X_i = \begin{pmatrix} I & 0 \\ 0 & \chi_i \end{pmatrix}$, $Y_i^o = (A_i, T_i\chi_i)$, $= (A_i,\ T_i)\begin{pmatrix} I & 0 \\ 0 & \chi_i \end{pmatrix} = (A_i,\ T_i)X_i.$

Let $\varepsilon_i$ model the difference between realized and expected values of the components of $Y_i^{\circ}$ , then:

$$Y_i^o = (\mu_A,\ \mu_T)X_i + \varepsilon_i \cdot$$

All available information (population data, sample data, and the stochastic relationships between them) is summarized in a linear model:

$$\left(Y_1^o,\ Y_2^o,\ Y_3^o, \ldots\ .\ , Y_n^o\right) =$$
$$(\mu_A,\ \mu_T)(X_1,\ X_2,\ X_3, \ldots\ .\ , X_n)$$
$$+ (\varepsilon_1,\ \varepsilon_2,\ \varepsilon_3, \ldots\ .\ , \varepsilon_n)$$

where $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots\ldots \varepsilon_n) - N(0, \Sigma_{Y^0})$,

$\Sigma_{Y^0}$ is the block diagonal matrix of the $\{X_i{}' \Sigma X_i\}$, and $\Sigma$ is the covariance matrix of the vector , $(A_i,T_i)$.

Writing this in compact form : $Y^o = \mu X + \varepsilon$ where $\varepsilon$ has mean zero and covariance matrix $\Sigma_{Y^0}$.

Now auxiliary variables are used to adjust for the peculiarities of the realized sample (and possibly response bias) This is done by post stratification on the auxiliary variables, using regression adjustments. Then by a procedure that is similar to imputation or the EM-algorithm, the observed data items are used to compensate for missing data.

Let $\Sigma = \begin{pmatrix} \Sigma_A & \Sigma_{AT} \\ \Sigma_{TA} & \Sigma_T \end{pmatrix}$, where $\Sigma_A$ is the variance/covariance matrix of an $A_i$ , $\Sigma_T$ is the variance/covariance matrix of a $T_i$ , and $\Sigma_{AT}$ is the matrix of covariances between $A_i$ and $T_i$ . The expected value of $Y_i = (A_i , T_i)$ is $\mu = (\mu_A ; \mu_T)$, and by definition $\mu_A$ is known. When Y is multivariate normal the conditional properties of the multivariate normal distribution are exploited to make inferences about $\mu_T$ based on the conditional distribution of the target variables given the auxiliary variables. $E(T_i | A_i = a_i) = \mu_T + (a_i - \mu_A)\Sigma_A^{-1}\Sigma_{AT}$. The expected value of $T_i$ is influenced by it's associated auxiliary outcomes $A_i = a_i$. This suggests that when $a_i$, $\mu_A$, and $\Sigma$ are all known, the target data should be translated as follows:

Let $Z_i = [T_i - (A_i - \mu_A)\Sigma_A^{-1}\Sigma_{AT}]$ . Then the conditional expected value of $Z_i$ given $A_i = a_i$ is $\mu_T$ , the quantity to be estimated. The $\{Z_i\}$ are centered on $\mu_T$ and the dispersion of the $\{Z_i\}$ about $\mu_T$ is tighter by $\Sigma_{TA}\Sigma_A^{-1}\Sigma_{AT}$; that is, the conditional variance/covariance matrix of $(Z_i|A_i=a_i)$ is $\Sigma_\delta = \Sigma_T - \Sigma_{TA}\Sigma_A^{-1}\Sigma_{AT}$.

The observed components of $Z_i$ (responses) are separated from the unobserved components (non responses) by post multiplying $Z_i$ by $\chi_i$ and in order to avoid introducing another variable name, now let

$$Z_i = [T_i - (A_i - \mu_A)\Sigma_A^{-1}\Sigma_{AT}]\chi_i, \qquad (2.0)$$

then $E(Z_i|A_i) = \mu_T\chi_i$ and $V(Z_i|A_i) = \chi_i'(\Sigma_T - \Sigma_{TA}\Sigma_A^{-1}\Sigma_{AT})\chi_i = \chi_i'\Sigma_\delta\chi_i$. Letting $\delta_i$ model the difference between the realized and expected value of $(Z_i|A_i)$:

$Z_i = \mu_t\chi_i + \delta_i$ and summarizing over all sample units:

$$(Z_1, Z_2, \ldots . Z_n) = \mu_t(\chi_1, \chi_2, \ldots . \chi_n) + (\delta_1, \delta_2, \ldots . \delta_n)$$

or

$$Z = \mu_T\chi + \delta, \qquad (2.1)$$

where $\delta$ has mean zero and covariance matrix $\Sigma_Z$, and $\Sigma_Z$ is the block diagonal matrix of the $\{\chi_i\Sigma_\delta\chi_i'\}$,

$$\Sigma_Z = \begin{pmatrix} \chi_1\Sigma_\delta\chi_1' & 0 & . & . & 0 \\ 0 & \chi_2\Sigma_\delta\chi_2' & 0 & . & 0 \\ . & 0 & . & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & \chi_n\Sigma_\delta\chi_n' \end{pmatrix} \qquad (2.2)$$

From (2.1), the generalized least squares (GLS) estimator for $\mu_T$ is: $\hat{\mu}_{T|A} = Z\Sigma_z^{-1}\chi'(\chi\Sigma_z^{-1}\chi')^{-1}$. The variance/covariance matrix of $\hat{\mu}_{T|A}$ is $(\chi\Sigma_z^{-1}\chi')^{-1}$.

This estimator seems to ignore the sampling distribution. As explained in the next section, the sampling distribution is often implicitly included in the estimation process. Section 3 contains two results (or theorems) which quantify the effect on bias and variance of omitting target and auxiliary variables from the construction of the estimator outlined above.

## 3) STRUCTURE THEOREMS

Let $A=(A^1,A^2)$, and $T = (T^1,T^2)$ be partitions of A and T into subvectors. In this section, two results are proved which, quantify the effect on MSE of deleting the data on $T^2$ from W and the construction of the GLS for $\mu_{T^1} = E(T^1)$ and the effect on MSE of deleting the data on $A^2$ from the construction of the GLS for $\mu_T = E(T)$. Superscripts denote subvectors.

*Bias and variance of the GLS estimators under these two forms of reduced data structure are evaluated with respect to the probability space that these GLSs inherit from the stochastic structure on (A,T) and all the data on these variates.*

***Definition 1.*** $A^2$ is redundant for estimating $\mu_T$ if and only if the conditional random vector, T given $A^1$, is independent of , $A^2$ given $A^1$, $[(T|A^1)\perp(A^2|A^1)]$.

***Definition 2.*** $T^2$ is redundant for estimating $\mu_{T^1}$ if and only if $T^2$ given A, is independent of $T^1$ given A $[(T^1|A)\perp(T^2|A)]$.

This last definition is reciprocal, it also implies that $T^1$ is redundant for estimating $\mu_{T^2}$.

A target or auxiliary variable is redundant for estimating a particular target mean if neither its inclusion in nor omission from the construction of the GLS has any effect on the estimator of that target mean.

**Definition 3.** A variate (or vector of variates) that is not redundant is called pertinent.

The next two results are the structure theorems. Their proofs in the case of multivariate normality are by direct computation and the Gauss (-Markov) theorem on minimum variance estimation. Bias and variance of these "deleted" GLS estimators are with respect to the full stochastic structure on (A,T), all the data on these variates, and conditioned on the known sample outcomes for A. The proofs below depend on (and possibly clarify) this last statement.

Result 1 - Deleting pertinent target variables, $T^2$ from W, will increase the variance of the components of $\hat{\mu}_{T^1|A}$, the GLS for $\mu_{T^1}$ compared to the variance of the corresponding components of $\hat{\mu}_{T|A}$ and will have no effect on their bias.

Proof of 1

a) By rearranging the components of Z, (2.1) can be rewritten with block diagonal X-matrix as:

$$Z=(Z^1,Z^2)=(\mu_{T^1},\mu_{T^2})\begin{pmatrix} \chi_{11} & 0 \\ 0 & \chi_{22} \end{pmatrix}+(\delta^1,\delta^2) \quad (3.1)$$

where $Z^1$ contains only members of the first set of target variables, $T^1$, $Z^2$ contains only members of the second set of targets, $T^2$, and the covariance matrix of $(\delta^1,\delta^2)$ is $\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, a rearrangement of $\Sigma_Z$.

b) Note that by ignoring $T^2$ in the construction of (3.1) we would have:

$$Z^1 = \mu_{T^1}\chi_{11} + \delta^1 \qquad (3.2)$$

where the variance/covariance matrix of $\delta^1$ is $\Sigma_{11}$.

c) The GLS for $\mu_{T^1}$ under (3.2) is $\hat{\mu}_{T^1|A}$. It is linear and unbiased under (3.1). By the Gauss theorem, the first $T^1$ components of $\hat{\mu}_{T|A}$ are minimum variance linear unbiased for $\mu_{T^1}$ under (3.1). Therefore, $V(\hat{\mu}_{T|A}Q) \le V(\hat{\mu}_{T^1|A})$ where Q is the matrix of zeros and ones that picks out the $T^1$ components of $\hat{\mu}_{T|A}$ , V denotes variance of the enclosed vector, and ≤ is component by component comparison.

Finally, since $E(\hat{\mu}_{T^1|A}) = E(\hat{\mu}_{T|A}Q)$ under (3.1) , deleting targets has no effect on bias. **END PF1.**

Corollary -

$\hat{\mu}_{T|A}Q = \hat{\mu}_{T^1|A} \Leftrightarrow Cov(T^1, T^2 | A) = 0.$ This corollary shows that the definition of redundant target variable makes sense; when $T^2$ is redundant, the estimate of $\mu_{T^1}$ can borrow no strength from data on $T^2$.

Result 2. - Let $\hat{\mu}_{T|A^1}$ be the GLS for the mean of T when the auxiliary variables in $A^2$ are omitted from W and the estimator construction described in the previous section. Conditional on $A=(A^1,A^2)$, $\mathrm{Var}(\hat{\mu}_{T|A^1}) \ge \mathrm{Var}(\hat{\mu}_{T|A})$ and $\hat{\mu}_{T|A^1}$ is a biased estimate for the mean of T, $\mu_T$. "≥" is component by component comparison.

Proof of 2

Let the variance/covariance matrix of $(A,T)=(A^1,A^2,T)$ be:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{1T} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{2T} \\ \Sigma_{T1} & \Sigma_{T2} & \Sigma_{TT} \end{pmatrix}.$$

Then ignoring $A^2$ and conditioning on $A^1$ alone, the analog to $Z_i$ is $Z_i^* = T_i - (a_i^1 - \mu_{a^1})\Sigma_{11}^{-1}\Sigma_{1T}$ , where $a_i^1$ is the $i^{th}$ unit's realized value for $A^1$ and $\mu_{a^1}$ is the expected value of $A^1$. The variance/covariance matrix of $Z_i^*$ is $\Sigma_{TT} - \Sigma_{T1}\Sigma_{11}^{-1}\Sigma_{1T}$. The variance/covariance matrix, $\Sigma_Z$, given in (2.2) for the vector $Z=(Z_1,Z_2,Z_3,\ldots\ldots\ldots,Z_n)$ has an analog for $Z^* = (Z_1^*, Z_2^*, Z_3^*, \ldots\ldots\ldots, Z_n^*)$, that is denoted ,$\Sigma_{z^*}$, and is given by (2.2) with $\Sigma_\delta = \Sigma_{\delta^*} = \Sigma_{TT} - \Sigma_{T1}\Sigma_{11}^{-1}\Sigma_{1T}$. Conditioning on both subvectors of A, $T_i$ transforms to $Z_i$ by subtracting the following expression from $T_i$.

$$[(a_i^1 - \mu_{a^1}),(a_i^2 - \mu_{a^2})]\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{1T} \\ \Sigma_{2T} \end{bmatrix}$$

$$= (a_i^1 - \mu_{a^1})\Sigma_{11}^{-1}\Sigma_{1T} + (a_i^1 - \mu_{a^1})GFG'\Sigma_{1T}$$

$$-(a_i^1 - \mu_{a^1})GF\Sigma_{2T} - (a_i^2 - \mu_{a^2})FG'\Sigma_{1T}$$

$$+(a_i^2 - \mu_{a^2})F\Sigma_{2T}$$

where $G = \Sigma_{11}^{-1}\Sigma_{12}, F = (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}$ , (') denotes transpose, and $a_i^2$ and $\mu_{a^2}$ are defined analogously to $a_i^1$ and $\mu_{a^1}$. Thus the difference in bias adjustment between conditioning on A and conditioning on $A^1$ alone is given by:

$$\mathrm{Diff}_i = (a_i^1 - \mu_{a^1})GFG'\Sigma_{1T} - (a_i^1 - \mu_{a^1})GF\Sigma_{2T} - (a_i^2 - \mu_{a^2})FG'\Sigma_{1T} + (a_i^2 - \mu_{a^2})F\Sigma_{2T} =$$

$$Z_i^* - Z_i,$$

Conditioning on both $A^1$ and $A^2$, $\mathrm{Diff}_i$ is the bias in $Z_i^*$ when $A^2$ is ignored. Let $D = (\mathrm{Diff}_1\chi_1, \mathrm{Diff}_2\chi_2, \ldots\ldots\ldots, \mathrm{Diff}_n\chi_n)$, then, after deleting the nonresponses from $Z_i$ and $Z_i^*$ , (see 2.0) and forming Z and $Z^*$ ,

$Z^*=Z + D$. The GLS given both $A^1$ and $A^2$ is:

$\hat{\mu}_{T|A} = Z\Sigma_Z^{-1}\chi'(\chi\Sigma_Z^{-1}\chi')^{-1}$ , and the GLS ignoring $A_2$ is:

$$\hat{\mu}_{T|A^1} = (Z+D)\Sigma_{Z^*}^{-1}\chi'(\chi\Sigma_{Z^*}^{-1}\chi')^{-1} =$$

$$Z\Sigma_{Z^*}^{-1}\chi'(\chi\Sigma_{Z^*}^{-1}\chi')^{-1} + D\Sigma_{Z^*}^{-1}\chi'(\chi\Sigma_{Z^*}^{-1}\chi')^{-1} =$$

$$Z\Sigma_{Z^*}^{-1}\chi'(\chi\Sigma_{Z^*}^{-1}\chi')^{-1} + \mathrm{BIAS},$$

where conditional on both $A^1$ and $A^2$ , BIAS is the constant $D\Sigma_{Z^*}^{-1}\chi'(\chi\Sigma_{Z^*}^{-1}\chi')^{-1}$. The variance/covariance matrix of $\hat{\mu}_{T|A^1}$ given A is the variance/covariance of $Z\Sigma_{Z^*}^{-1}\chi'(\chi\Sigma_{Z^*}^{-1}\chi')^{-1}$ conditional on $A=(A^1,A^2)$, and this matrix is

$(\chi\Sigma_{Z^*}^{-1}\chi')^{-1}\chi\Sigma_{Z^*}^{-1}\Sigma_Z\Sigma_{Z^*}^{-1}\chi'(\chi\Sigma_{Z^*}^{-1}\chi')^{-1}$ . Note that term-by-term, the diagonal of $(\chi\Sigma_{Z^*}^{-1}\chi')^{-1}\chi\Sigma_{Z^*}^{-1}\Sigma_Z\Sigma_{Z^*}^{-1}\chi'(\chi\Sigma_{Z^*}^{-1}\chi')^{-1}$ is greater than the diagonal of $(\chi\Sigma_Z^{-1}\chi')^{-1}$ , the variance/covariance matrix of $\hat{\mu}_{T|A}$ . This inequality follows because $\hat{\mu}_{T|A}$ is the GLS conditional on A and

$Z\Sigma_{Z^{\bullet}}^{-1}\chi'(\chi\Sigma_{Z^{\bullet}}^{-1}\chi')^{-1}$ is linear and unbiased.

Thus $\text{MSE}(\hat{\mu}_{T|A^1}) =$

$$(\chi\Sigma_{Z^{\bullet}}^{-1}\chi')^{-1}\chi\Sigma_{Z^{\bullet}}^{-1}\Sigma_Z\Sigma_{Z^{\bullet}}^{-1}\chi'(\chi\Sigma_{Z^{\bullet}}^{-1}\chi')^{-1}+ \text{(BIAS)'(BIAS)}$$

and the diagonal elements of the first term in this MSE are greater than the corresponding diagonal elements of $\text{Var}(\hat{\mu}_{T|A})$.

**END Pf 2**

$Z\Sigma_{Z^{\bullet}}^{-1}\chi'(\chi\Sigma_{Z^{\bullet}}^{-1}\chi')^{-1}$ in the above proof is an unbiased estimate of the mean of T but, conditional on $A=(A^1,A^2)$, this estimator has less than optimal weight vectors, $\Sigma_{Z^{\bullet}}^{-1}\chi'(\chi\Sigma_{Z^{\bullet}}^{-1}\chi')^{-1}$, and thus a larger variance than $\hat{\mu}_{T|A^1}$. GLS estimators are generally robust against misspecification of the optimal weight vectors, $\Sigma_Z^{-1}\chi'(\chi\Sigma_Z^{-1}\chi')^{-1}$ . In these cases, the increase in MSE due to omitting pertinent auxiliary variables is dominated by bias and the additional variance due to using the wrong weight vectors is relatively negligible.

**Corollary -** $\text{Diff}_i = 0 \Leftrightarrow \text{Cov}[(A^2,T)|A^1]=0$. Thus the definition of redundant ( not pertinent) auxiliary variable at the beginning of this section makes sense.

In summary, target variables control variance and auxiliary variables control mostly bias. There are three ways to reduce variance: increase the sample size, increase the number of pertinent target variables in T, and increase the number of pertinent auxiliary variables in A. Increasing the number of pertinent auxiliary variables can reduce bias.

If one includes target or auxiliary variables in the data structure (data matrix, W) that are redundant for a particular target mean, then the GLS estimator of that target mean is algebraically identical to the GLS estimator derived from the data structure, W, that excludes them but in every other way is the same. This means that there is no penalty (except possibly computer time) for including unnecessary (redundant) variables in the data matrix, W. In particular, the sampling distribution may be included in the estimation process by using the sample indicator function as an auxiliary variable, but in many common situations it will prove to be redundant. Many estimators which seem to ignore the sampling distribution implicitly use it through other auxiliary variables in the data matrix that already contain all the pertinent information that the sample indicator provides.

The observed data maximum likelihood estimator (MLE) as described in Chapter 7 of Little and Rubin is, under normality, the GLS estimator, $\hat{\mu}_{t|A} = Z\Sigma_z^{-1}\chi'(\chi\Sigma_z^{-1}\chi')^{-1}$. This estimator is also a fixed point of the EM-algorithm (see theorems in the theory section of that chapter). Thus the Normal EM-algorithm with Σ given, will converge to this GLS. The GLS estimator contains it's own nonresponse adjustment by default.

## 5) CONCLUSIONS

Although the sampling distribution is a necessary part of inference from sample survey data, it is rarely sufficient because of many features of applied sampling. These features include nonresponse, response bias, and data relationships that make applied sampling a multivariate discipline where univariate methods generally fail to produce optimal inferences. In spite of this, sample survey inference has remained largely univariate with an encyclopedia of corrective techniques to handle these negative features of sample data.

This paper discussed theory and applications of multivariate methods for estimating finite population mean vectors assuming data deficiencies like nonresponse (both item and total, ignorable and otherwise) and response bias, but exploiting data dependencies modeled by the covariance matrix of survey variables (both design and target variables). These data dependencies are used to minimize mean square error in the presence of the data deficiencies. The estimator so derived automatically handles many missing data problems that practitioners face by fully exploiting known data dependencies. Its use is indicated in repeated surveys where nonresponse is a problem and strong data dependencies are present.

The theory presented here is at least two centuries old, Gauss (1809). Sampling theory is much newer; its standard methodologies work and can be applied with primitive computational aids (say 1950s technology). The material presented here would have been totally impractical in 1950 but in the 1990s the computer revolution has made Gauss' methods quite practical. Present day computers allow instant availability of huge supplementary data bases and the computational power to make short work of complex estimators. The estimation process described in this paper is applied in Johnson and Woodruff (1990).

## REFERENCES

Gauss K.F. (1809), *Werke* 4, 1-93, Gottingen.

Johnson C. and Woodruff S. (1990). An Application of Regression Superpopulation Models in the Current Employment Statistics Survey, Proceedings of the American Statistical Association, Survey Research Methods.

Little J.A. and Rubin D. B. (1987). *Statistical Analysis with Missing Data*, Wiley.