THE EQUIVALENCE OF GENERALIZED DOUBLE SAMPLING REGRESSION ESTIMATORS, WEIGHT ADJUSTMENTS, AND RANDOMIZED HOT DECK IMPUTATIONS

Ralph Folsom, Jr., Research Triangle Institute

1. Introduction

The results reported in this paper relating to double sampling regression estimation for complex sample designs were motivated by the need to provide efficient sample design and estimation methods for a household based national medical care expenditure survey (NMCES) that incorporated a subsample based medical provider record check. To reduce the computation burden associated with regression estimation methods, simplified double sampling ratio and poststratified mean estimators are derived as special cases of the regression estimator. Attempting to combine the efficiency of regression based methods with the computational ease of the simple Horvitz-Thompson weighted sum, the equivalence of regression estimators, subsample reweighting strategies, and randomized hot deck imputation of medical provider data to the complement of the record check subsample is explored.

2. Double Sampling Regression Estimators

The NMCES double sampling design is characterized by a three stage area household cluster sample expected to yield $n = rst\bar{u}$ household members where r denotes the number of primary sample counties selected, s depicts the number of second stage area segments drawn per PSU, $\bar{t}$ represents the average number of responding third stage dwelling units selected per area segment, and $\bar{u}$ is the average number of persons per household in the universe. The first phase household data collection yields a vector $\underset{\sim}{X}(i,j,k,\ell)$ of annual health care expenditure and utilization variables for participating person ℓ from household k in segment j of PSU i. A subsample of m household sample participants is then selected for a medical provider record check survey (MPS). All medical doctors, doctors of osteopathy, and all health care facilities administered by MD/DOs that were reported as health care providers or usual sources of care by the MPS sample members are then queried by mail regarding the details of visits recorded in their records including associated expenditures, sources of payment, and diagnosed conditions.

If the MPS survey members are indexed by a single label ℓ and $Y(\ell)$ represents a medical provider reported variable, then the universe total Y(+) of the medical record based variable can be estimated by a generalized double sampling regression statistic of the form

$$\hat{Y}_+(DSR) = \hat{Y}_+(M) - \sum_{q=1}^{p-1} [\hat{X}_+(q|M) - \hat{X}_+(q|H)]\ \hat{\beta}_q(M)$$

where

$$\hat{X}_+(q|H) = \sum_{\ell \varepsilon HHS} X_q(\ell)/\pi(\ell)$$

is the household sample total for the q-th element of $\underset{\sim}{X}(\ell)$. The $\pi(\ell)$ denote HHS sample inclusion probabilities. The MPS sample totals are of the form

$$\hat{Y}_+(M) = \sum_{\ell \varepsilon MPS} m(\ell)\ Y(\ell)\ /\ E\{m(\ell)\}\ \pi(\ell)$$

where $m(\ell)$ depicts the frequency of selection for household participant ℓ in the MPS sample, allowing for multiple selections. The quantity $E\{m(\ell)\}$ denotes the expected number of MPS selections for household survey participant ℓ. The class of second phase designs considered specify $E\{m(\ell)\}$ as follows

$$E\{m(\ell)\} = m\{V(\ell)/\pi(\ell)\ \hat{V}_+(HHS)\} = mv(\ell)$$

where $V(\ell)$ is an element of the HHS data vector $\underset{\sim}{X}(\ell)$. The statistics

$$\hat{V}_+(HHS) = \sum_{\ell \varepsilon HHS} V(\ell)/\pi(\ell)$$

imbeded in $E\{m(\ell)\}$ is the HHS based Horvitz-Thompson estimator for the V variate universe total. The specific form of $V(\ell)$ envisioned, namely

$$V(\ell) = [X(\ell)]^{g/2} \text{ for } g = 0,\ 1, \text{ or } 2,$$

is motivated by optimality considerations under a superpopulation variance model ξ of the form

$$\underset{\xi}{\mathrm{Var}}[Y(\ell)] = \sigma^2 [X(\ell)]^g$$

where $X(\ell)$ might represent the HHS reported total annual health care expenditure. The MPS based X variate total estimators

$$\hat{X}_+(q|M) = \sum_{\ell \varepsilon MPS} m(\ell)\ X_q(\ell)/\pi(\ell)E\{m(\ell)\}\ ,$$

are defined similarly.

The regression coefficient estimator proposed for $\hat{Y}_+(DSR)$ is

$$\hat{\underset{\sim}{\beta}}(M) = [X'(V)^{-2}X]^{-1}_M\ [X'(V)^{-2}Y]_M$$

with

$$[X'(V)^{-2}X]_M = [\sum_{\Pi \varepsilon MPS} m(\ell)\underset{\sim}{X}'(\ell)\underset{\sim}{X}(\ell)/V^2(\ell)]$$

and

$$[X'(V)^{-2}Y]_M = [\sum_{\ell \varepsilon MPS} m(\ell)\underset{\sim}{X}'(\ell)\underset{\sim}{Y}(\ell)/V^2(\ell)]\quad .$$

Recalling the variance function superpopulation model that motivated the choice of $V(\ell)$; namely

$$\underset{\xi}{\mathrm{Var}}[Y(\ell)] = \sigma^2\ V^2(\ell)$$

one recognizes $\hat{\underset{\sim}{\beta}}(M)$ as the model based best linear unbiased estimator (BLUE) for β. With this specification of $\underset{\sim}{\beta}(M)$ the Taylor Series variance approximation for the DSR estimator is derived in the expanded version of this paper, Folsom (1981). By conditioning on the observed HHS person level sample size n, the combined effects of clustering, stratification, and without replacement selections at the PSU, area segment, and housing unit selection stages are captured by a single HHS covariance component $\rho(H)$. This variance-covariance component representation was derived for general PPS without replacement designs by Gray (1975) and subsequently extended by Folsom (1980) to a wider class of PPS designs including with replacement selections and Chromy's (1979) probability minimum replacement method. Combining this general variance-covariance component representation with an explicit form for the Taylor Series linearization of $\underset{\sim}{\beta}(M)$ derived independently by Folsom (1974) and Fuller (1974), the following variance approximation for the double sampling regression estimator $Y_+(DSR)$ was obtained

$$\mathrm{Var}[\hat{Y}_+(DSR)] = \sigma^2(H)\ [1+(n-1)\rho(H)]/n$$
$$+\ \sigma_e^2(M)\ [1+(m-1)\rho_e(M)]/m$$

where

$$\sigma^2(H) = \sum_{\ell \varepsilon U} P(\ell)D_Y^2(\ell)$$

with

$$D_Y(\ell) = \{\frac{Y(\ell)}{P(\ell)} - Y(+)\}$$

is analogous to the single stage PPS with replacement variance component based on single draw probabilities $P(\ell) = \pi(\ell)/n$ depicting the probability that frame unit ℓ is included in the HHS sample and is randomly assigned sample person label i. The HHS covariance component

$$\sigma^2(H)\rho(H) = \sum_{\ell \varepsilon U}\ \sum_{\ell' \varepsilon U} P(\ell\ell')\ D_Y(\ell)D_Y(\ell')$$

is defined similarly in terms of the double draw probability

$$P(\ell\ell') = \pi(\ell\ell')/n(n-1)$$

that frame units ℓ and ℓ' belong to the HHS sample and are randomly assigned sample person labels i and j respectively. Similarly, one can show that

$$\sigma_e^2(M)=[\frac{(n-1)}{n}] \sum_{\ell \varepsilon U}\ \sum_{\ell' \varepsilon U} \{\frac{P(\ell\ell')}{P(\ell)P(\ell')}\}V(\ell)V(\ell')\Delta_e^2(\ell\ell')/2$$

with

$$\Delta_e(\ell\ell') = \{\frac{e(\ell)}{V(\ell)} - \frac{e(\ell')}{V(\ell')}\}$$

and

$$e(\ell) = Y(\ell) - \underset{\sim}{X}(\ell)\underset{\sim}{\beta}$$

denoting the regression residual or prediction bias based on the universe level coefficient vector

$$\underset{\sim}{\beta} = [X'(V)^{-1}X]_U^{-1}\ [X'(V)^{-1}Y]_U \quad .$$

The MPS covariance component is

$$\sigma_e^2(M)\rho_e(M) = \underset{H}{E} \sum_{\ell \varepsilon H}\ \sum_{\ell' H} [v(\ell)v(\ell')-v(\ell\ell')]\delta_e^2(\ell\ell')/2$$

with $\underset{H}{E}$ indicating the HHS sampling expectation operator,

$$\delta_e(\ell\ell') = \{\frac{e(\ell)}{\pi(\ell)v(\ell)} - \frac{e(\ell')}{\pi(\ell)V(\ell')}\} \quad ,$$

and $v(\ell\ell')$ denoting a second phase double draw probability. The corresponding Taylor Series HHS component estimators are

$$\hat{\sigma}^2(H) = \sum_{\ell \varepsilon M}\ \sum_{\ell \varepsilon M'} \{\frac{m(\ell)m(\ell')}{m(m-1)}\}\{\frac{WH(\ell\ell')}{v(\ell\ell')}\}\{\frac{\delta_Y^2(\ell\ell')}{2n(n-1)}\}$$

and

$$\hat{\sigma}^2(H)\hat{\rho}(H)= \sum_{\ell \varepsilon M}\ \sum_{\ell' \varepsilon M} \{\frac{m(\ell)m(\ell')}{m(m-1)}\}\{\frac{WH(\ell\ell')-1}{v(\ell\ell')}\}\{\frac{\delta_Y^2(\ell\ell')}{2n(n-1)}\}$$

with

$$WH(\ell\ell') = P(\ell)P(\ell')/P(\ell\ell')$$

and

$$\delta_Y(\ell\ell') = \{\frac{Y(\ell)}{P(\ell)} - \frac{Y(\ell')}{P(\ell')}\}.$$

The second phase MPS variance component is estimated by

$$\hat{\sigma}_e^2(M) = \sum_{\ell \varepsilon M}\ \sum_{\ell' \varepsilon M} \{\frac{m(\ell)m(\ell')}{m(m-1)}\}\ WM(\ell\ell')\ d_e^2(\ell\ell')/2$$

where

$$WM(\ell\ell') = V(\ell)V(\ell')/V(\ell\ell')$$

and

$$d_e(\ell\ell') = \{\frac{e_M(\ell)}{\pi(\ell)v(\ell)} - \frac{e_M(\ell')}{\pi(\ell)v(\ell')}\}$$

with

$$e_M(\ell) = Y(\ell) - \underset{\sim}{X}(\ell)\hat{\underset{\sim}{\beta}}(M)$$

denoting the observed residual for MPS participant ℓ. Similarly

$$\hat{\sigma}_e^2(M)\hat{\rho}_e(M) = \sum_{\ell \varepsilon M}\ \sum_{\ell' \varepsilon M} \{\frac{m(\ell)m(\ell')}{m(m-1)}\}[WM(\ell\ell')-1]d_e^2(\ell\ell')/2$$
$$= \hat{\rho}_e^2(M) - \sum_{\ell \varepsilon M} m(\ell)e_M^2(\ell)/\pi^2(\ell)v^2(\ell)(m-1).$$

An interesting special case of $\mathrm{Var}[\hat{Y}_+(DSR)]$ is obtained when single stage PPS with replacement samples are drawn at both phases of selection. In this case

$$\mathrm{Var}[\hat{Y}_+(DSR)] = \sigma^2(H)/n + \sigma_e^2(M)/m$$

with

$$\sigma_e^2\ (M) = [(n-1)/n]\ V_+(U) \sum_{\ell \varepsilon U} e^2(\ell)/V(\ell)$$

and $V_+(u)$ denoting the universe total for $V(\ell)$. Considering the statistical properties of the proposed DSR estimator, Folsom (1981) notes that $Y_+(DSR)$ is a double sampling multivariable extension of Cassel, Sárndal, and Wretman's (1977) generalized regression estimator. The double sampling version has the analogous property that under the regression super population model with β known it is the probability unbiased (p-unbiased) estimator with smallest expected (ε-model expectation) sampling variance, among all linear probability and model ($p\varepsilon$) unbiased estimators. Since β must be estimated from the sample, $Y_+(DSR)$ is not probability (sampling expectation E) unbiased, but is approximately unbiased, to orders $O(1/m)$, in large second phase samples.

An alternative variance approximation strategy that avoids the burdensome calculation of double drawn probabilities $P(\ell\ell')$ and $v(\ell\ell')$ is suggested in section 5. In the following section, a built-in record check pilot survey is proposed that permits one to empirically fit the $\varepsilon[Y(\ell)]$ model and the variance function $\underset{\xi}{\mathrm{Var}}[Y(\ell)] = \sigma^2[x(\ell)]^g$ for g = 0, 1, or 2.

3. Adaptive Double Sampling Designs

In this section, an adaptive double sampling design strategy is proposed. For multi-wave household panel surveys like NMCES and NMCUES one could initiate a prospective provider record check during the first wave of household interviews. For a prespecified epsem subsample of the cooperating round 1 households, permission would be obtained to contact all providers for reported medical care visits as well as the household members usual sources of care (USOC) if no visits to the USOC providers is otherwise reported in round 1. A mail survey of the reported round 1 providers would be initiated on a flow basis as soon as computer generated summaries of the reported round 1 visits, visit charges, and sources of payment can be produced. These summaries would be sent to the providers along with a questionnaire designed to validate the reported visit dates to correct or supply missing visit charge and source of payment information, to diagnose the medical conditions associated with each visit, and to supplement the record with unmatched visits that appear in the providers records but not in the patients self report. This approach would solicite the provider's assistance in matching the patient's self-reported visits to the medical records. With data collection waves lasting 10 to 12 weeks, the preliminary prospective record check could be carried through two rounds of household and provider interviews. With the first round of household interviews beginning around the middle of February with a retrospective reporting period extending back to January 1, two rounds of interviews would record from four to six months of household self-reports and matching provider record data. Using the matched health care utilization and expenditure data from the initial record check subsample for the first two interview rounds would permit one to emperically fit the household to provider prediction equation and the associated variance function. A class of prediction equations that would seem ideally suited to this provider data modeling task is the class of cubic polynomial spline models frequently used in economic prediction equations.

In terms of the model based expectation ε and variance V operators, a cubic spline model for predicting provider variables $Y(h\ell)$ for person ℓ from the h-th poststratum is

$$\varepsilon[Y(h\ell)] = \sum_{q=0}^{3} [X(h\ell)]^{q/2}\,\hat{\beta}_h(q)$$

and

$$V[Y(h\ell)] = \sigma^2(h)\ [X(h\ell)]^g \text{ for } g = 0,\ 1, \text{ or } 2.$$

The poststrata h are defined in terms of intervals on the X axis. The prediction equation is specified as a cubic polynomial in the square root of X so that when scaled by $V(\ell) = X(h\ell)^{g/2}$ the resulting equation will have an intercept on the $X(\ell)/V(\ell)$ scale. The separate poststratum h polynomials are caused to connect smoothly at join points XBD(h) at the boundary between poststratum h and h+1 by imposing side conditions

$$\sum_{q=0}^{3} [XBD(h)]^{q/2}\hat{\beta}_h(q) = \sum_{q=0}^{3} [XBD(h)]^{q/2}\hat{\beta}_{h+1}(q)$$

that force the connection. To guarantee a smooth transition, the first derivative functions are also required to join on the boundaries. The corresponding side conditions are imposed for all second phase stratum boundaries h = 1,2,...,H-1. In the expanded version of this paper, the use of a robust regression procedure called weighted slice regression originally proposed by W. A. Larsen and I. J. Terpenning (1971) and subsequently evaluated by L. Denby and W. A. Larsen (1977) is illustrated as a method for generating robust squared residuals

$$r^2(h\ell) = [Y(h\ell) - \sum_{q=0}^{3} [X(h\ell)]^{q/2}\ \hat{\beta}_h(q)]^2$$

which are in turn fit to the model

$$\varepsilon[r^2(h\ell)] = \sigma^2_g(h)[X(h\ell)]^{g(h)}$$

to determine the optimum value of g for poststratum h. Having obtained the best fitting value of g(h) among the alternative values 0, 1, or 2, the second phase selections from poststratum h would be made with probabilities proportional to

$$V(h\ell) = [X(h\ell)]^{g(h)/2}$$

Having fit the cubic spline models and the variance functions to the built-in pilot study data one can proceed to determine the optimum second phase sample allocation to strata. Extending results of Cassel, Sárndal, and Wretman, one can estimate the model expectation of the second phase variance contribution in terms of the weighted residual mean square

$$\underset{\xi}{\hat{\sigma}^2}(h) = \sum_{\ell=1}^{m'(h)} \{Y(h\ell) - \underset{\sim}{X}(h\ell)\ \hat{\underset{\sim}{\beta}}_h\ \}^2/v^2(h\ell)[m'(h)-4]$$

where

$$v(h\ell) = \{X(h\ell)^{g(h)/2}\ /\ \sum_{\ell\varepsilon H(h)} X(h\ell)^{g(h)/2}\}.$$

With these definitions the super population model yields

$$\varepsilon\{\underset{HHS}{E}\ \underset{MPS}{Var}\ [\hat{Y}_h\ (DSR)]\}\ \hat{=}\ (1-f_hA_h)\ \hat{\sigma}^2_\xi\ (h)/m(h)$$

with

$$f_h = m(h)/n(h)$$

depicting the second phase subsampling rate from poststratum h, and

$$A_h = n(h) \sum_{\ell=1}^{n(h)} v^2(h\ell)$$

The optimum allocation of the second phase sample to the second phase strata so as to minimize cost subject to a total variance constraint of the form

$$Var\ [\hat{Y}\ (DSR)] \le V(0)$$

translates into the problem of minimizing

$$\sum_{h=1}^{H} \$(h)m(h)$$

subject to

$$\sum_{h=1}^{H} \hat{\sigma}_\xi{}^2(h)/m(h) \le V^*(0)$$

where

$$V^*(0)=V(0)-var\{\underset{HHS}{E}\ \underset{MPS}{E}\ [\hat{Y}_+(DSR)]\} - \sum_{h=1}^{H} A_h\ \hat{\sigma}_\xi{}^2(h)/n(h).$$

The household survey variance contribution

$$var\{\underset{HHS}{E}\ \underset{MPS}{E}\ [\hat{Y}(DSR)\}$$

in $V^*(0)$ could be approximated for this purpose by combining a simple random sampling HHS variance approximation for the provider survey Y variate derived from the built-in epsem pilot study sample with an HHS survey based design effect DEFF(X) estimated for the matching household reported X variate total adjusting the SRS variance approximation for the combined effects of HHS stratification, clustering, and disproportionate sampling (or unequal weighting).

Having developed the generalized double sampling regression estimator and a strategy for jointly optimizing the second phase sample design and estimation procedures, the next section explores some important special cases of the generalize regression estimator, namely the double sampling ratio estimator and the double sampling for stratification estimator. A simple reweighting strategy for producing the double sampling for stratification estimator is presented.

4. Ratio and Poststratification Estimators

The double sampling regression estimator presented in section 2 can be shown to include as a special case most of the familiar double sampling estimators. Two of these special cases will be displayed in this section. The first special case of interest is related to a super-population model of the form

$$\varepsilon[Y(\ell)] = \sum_{c=1}^{C} \delta_c(\ell)\ X(\ell)\ \beta(c)$$

with variance function

$$V[Y(\ell)] = \sigma_o\ X(\ell)$$

where

$$\delta_c(\ell) = \begin{cases} 1 & \text{if person } \ell \text{ belongs to poststratum c} \\ 0 & \text{otherwise} \end{cases}$$

The poststrata indexed by c = 1,...,C above could include population subgroups identified in terms of X variate intervals crossed with various person classification variables. Letting $\underset{\sim}{X}(\ell) = X(\ell)\ \underset{\sim}{\delta}(\ell)$ where $\underset{\sim}{\delta}(\ell)$ is the vector of C one-zero poststratum indicators for person ℓ, then the MPS selections would be with probabilities $v(\ell)$ proportional to $\sqrt{X(\ell)/\pi(\ell)}$ where $\pi(\ell)$ denotes person ℓ's first phase inclusion probability. The estimator for $\beta(c)$ in this instance is

$$\hat{\beta}(c) = y_+(c|MPS)\ /\ x_+(c|MPS)$$

with

$$x_+(c|MPS) = \sum_{\ell\varepsilon MPS} \delta_c(\ell)\ X(\ell)$$

$$y_+(c|MPS) = \sum_{\ell\varepsilon MPS} \delta_c(\ell)\ Y(\ell)$$

The associated double sampling estimator is then

$$\hat{Y}_+(DSR) = \sum_{\ell\varepsilon HHS} X(\ell)\ \underset{\sim}{\delta}(\ell)\ \underset{\sim}{\hat{\beta}}\ (MPS)/\pi(\ell)$$

$$= \sum_{c=1}^{C} \hat{X}_+(c|HHS)\ \hat{\beta}(c)$$

$$= \sum_{c=1}^{C} \hat{X}_+(c|HHS)\ [y_+(c|MPS)/x_+(c|MPS)]$$

Since $\underset{\sim}{Z}(\ell) = X(\ell)\underset{\sim}{\delta}(\ell)/V(\ell)$ has no intercept term, the associated estimator is not necessarily consistent. Note that the $\underset{\sim}{\beta}(c)$ ratios are not weighted. The familiar consistent estimator with

$$\hat{\beta}^*(c) = \frac{\sum_{\ell\varepsilon MPS} m(\ell)\delta_c(\ell)Y(\ell)/\pi(\ell)Em(\ell)}{\sum_{\ell\varepsilon MPS} m(\ell)\delta_c(\ell)X(\ell)/\pi(\ell)Em(\ell)} = \frac{\hat{Y}_+(c|MPS)}{\hat{X}_+(c|MPS)}$$

is produced by weighted least squares using the weight matrix

$$W = \text{Diag}[X(\ell)^{-3/2}].$$

The double sampling for stratification estimator derives from a model with

$$\varepsilon[Y(\ell)] = \sum_{c=1}^{c} \delta_c(\ell)\mu(c)$$

$$V[Y(\ell)] = \sigma^2.$$

In terms of this model, the BLUE for $\hat{\underline{\mu}}(c)$ is the unweighted Y mean for cell c, namely,

$$\bar{y}(c|MPS) = \sum_{\ell\varepsilon MPS} m(\ell)\underset{c}{\delta}(\ell)Y(\ell) / \sum_{\ell\varepsilon MPS} m(\ell)\underset{c}{\delta}(\ell)$$

$$= \sum_{\ell\varepsilon MPS} \delta_c(\ell)Y(\ell)/m(c).$$

The associated DSR is

$$\hat{Y}_+(DSR) = \sum_{\ell\varepsilon HHS} \underset{\sim}{\delta}(\ell)\hat{\underset{\sim}{\mu}}/\pi(\ell)$$

$$= \sum_{c=1}^{C} \hat{N}_+(c|HHS)\bar{y}(c|MPS).$$

Notice that since

$$Em(\ell) = m[1/\pi(\ell)] / \sum_{\ell\varepsilon HHS} [1/\pi(\ell)]$$

$$= m / \pi(\ell)\hat{N}$$

the properly weighted consistent estimator for $\mu(c)$ is the unweighted cell mean $\bar{y}(c|MPS)$. Forming the ratio adjusted MPS sample weights

$$aw(\ell) = \sum_{c=1}^{C} \delta_c(\ell)[\hat{N}_+(c|HHS)/\hat{N}_+(c|MPS)]$$

it is clear that one can form $\hat{Y}_+(DSR)$ as the simple weighted total

$$\hat{Y}_+(DSR) = \sum_{\ell\varepsilon MPS} aw(\ell)\, Y(\ell) \quad .$$

In the following section, an imputation strategy is proposed as an alternative to using the general double sampling regression estimator presented in section 2. While admittedly adding some extraneous variation to the full sample estimates, the imputation strategy makes the regression estimation transparent to analyst processing 'public use' tapes.

5. Model Based Predictions and Randomized Imputations

An alternative to using the general double sampling regression estimator developed in section 2 is to use the prediction equation

$$\hat{Y}(\ell) = \hat{\beta}_o(M) + \sum_{q=1}^{P-1} X_q(\ell)\, \hat{\beta}_q(M)$$

to provide medical provider data for household survey participants that were not selected for the medical provider survey. Utilizing these predicted values along with the household survey inclusion probabilities $\pi(\ell)$, the simple weighted total of the predicted values is equivalent to the double sampling regression estimator.

Another robust prediction estimator that has some intuitive appeal makes use of the optimum medical provider survey weights so that the observed Y values for the MPS participants are utilized instead of their model based predictions. This estimator has the form

$$\hat{Y}_+(DSR) = \sum_{\ell\varepsilon H} m(\ell)\hat{Y}(\ell)/2\pi(\ell)Em(\ell)$$

$$+ \sum_{\ell\varepsilon H} [1-f(\ell)]\hat{Y}(\ell)/2\pi(\ell)[1-F(\ell)\}$$

where $F(\ell)$ is the fractional part of $Em(\ell)$ when the expected number of selections for unit ℓ exceeds one; otherwise $F(\ell)=Em(\ell)$. Similarly, if $I(\ell)$ is the integer part of $Em(\ell)$, then $f(\ell)=m(\ell)-I(\ell)$is the zero-one sample indicator for the event that $m(\ell)=I(\ell)+1$. If we restrict our attention to minimum replacement designs, then $E\{f(\ell)\}=F(\ell)$. The first term in $Y_+(DSR)$ sums the predicted values for the MPS participants with $m(\ell)>0$. The second term sums the predicted values for the compliment of the MPS sample where $m(\ell)=f(\ell)=0$ plus those units with $Em(\ell)>1$ and $f(\ell)=0$, say $\bar{M}_+$. For the regression model treated in section 2 with $\underset{\sim}{X}(\ell)$ including $V(\ell)$, one can show that the weighted residuals

$$m(\ell)[Y(\ell)-\underset{\sim}{X}(\ell)\hat{\underset{\sim}{\beta}}(M)]/\pi(\ell)Em(\ell)$$

sum to zero over the entire MPS sample. Therefore, one can recast Y^*_+ (DSR) as

$$\hat{Y}^*_+(DSR) = \sum_{\ell\varepsilon M} m(\ell)Y(\ell)/2m\sigma(\ell)$$

$$+ \sum_{\ell\varepsilon \bar{M}_+} \hat{Y}(\ell)/2\pi(\ell)[1-F(\ell)]$$

where

$$\sigma(\ell) = [V(\ell)/ \sum_{\ell\varepsilon HHS} V(\ell)/\pi(\ell 0].$$

The estimator $\hat{Y}^*_+$ (DSR) is produced simply as a weighted sum using weights

$$w(\ell) = \begin{cases} m(\ell)/2m\sigma(\ell) \text{ for } \ell\varepsilon M \\ 1/2\pi(\ell)[1-F(\ell) \text{ for } \ell\varepsilon\bar{M}_+ \end{cases}$$

and imputing values $\hat{Y}(\ell) = \underset{\sim}{X}(\ell)\hat{\underset{\sim}{\beta}}(M)$ for the NonMPS members in $\bar{M}_+$. For the certainty units with $m(\ell)>1$ and $f(\ell)=0$ that belong to both M and $\bar{M}_+$, a weighted average of the observed and predicted values, say $\tilde{Y}(\ell)$ could be used along with the aggregate weight

$$\tilde{w}(\ell) = \{m(\ell)/m\sigma(\ell) + 1/\pi(\ell)[1-F(\ell)]\}/2.$$

For the cell mean model, this strategy would call for imputing the Y mean Y(c|MPS) for members of poststratum c that were not selected for the medical provider record check. While this simple prediction model imputation strategy is appropriate for estimating the universe total $Y_+(U)$, the failure to account for substantial amounts of unexplained Y variation when using the model predictions may seriously bias the estimation of universe level Y variable distributions. A weighted sequential hot deck imputation algorithm developed by Brenda Cox (1980) is tailor made for this purpose having the property that over repeated randomized imputations the weighted mean of the imputed residuals equals the weighted mean of the MPS observed residuals within imputation classes. If, for example, one defined imputation classes c based on X variate intervals and person characteristics, then Cox's weighted sequential hot deck selects a donnor member from the MPS sample of cell c with an X value close to that of the Non-MPS household member ℓ such that over repeated randomized imputations

$$E\{ \sum_{\ell \varepsilon \bar{M}_+} \delta_c(\ell)\hat{r}(\ell)/\pi(\ell)[1-F(\ell)]\}/ \sum_{\ell \varepsilon \bar{M}_+} \delta_c(\ell)/\pi(\ell)[1-F(\ell)]$$

$$= \{ \sum_{\ell \varepsilon M} m(\ell)\delta_c(\ell)r(\ell)/m\sigma(\ell)\}/\{ \sum_{\ell \varepsilon M} m(\ell)\delta_c(\ell)/m\sigma(\ell)\}$$

This property has the effect of causing the nonMPS weighted distribution of imputed residuals to reproduce the MPS weighted distribution of observed residuals. To the extent that c represents a second phase stratum or a subgroup classification variable in $\underset{\sim}{X}$ that is interacted with the element of $\underset{\sim}{X}$ that equals $V(\ell)$, then the MPS weighted mean of the observed residuals will be zero and the associated weighted Y total for cell c is equivalent over repeated impuations to an efficient, consistent, double sampling regression estimator. When the cell mean model is appropriate, it is not difficult to see that the two pronged model prediction and subsequent hot deck residual imputation is equivalent to a direct hot deck imputation of Y values to Non MPS members using the weighted sequential method within the poststratification cells c identified for the cell mean model.

A simplified method for approximating the variance of the DSR prediction plus residual imputation estimator $Y_+^r(DSR)$ is proposed in the expanded version of this paper. The method recommended calls for selecting the second phase sample independently within the two PSU level half-sample replicates formed for each of ten or twelve collapsed primary superstrata. With this structure, one can form partially balance half sample replicates across the superstrata and perform second phase regression predictions and residual imputations independently for each partially balanced half-sample and its complement. The corresponding half-sample replicate estimator $Y_\alpha(DSR)$ and its complementary half sample statistic $Y_\alpha^c(DSR)$ can then be contrasted to form a reasonably unbiased variance estimator for $Y_+^r(DSR)$; namely

$$var\{Y_+^r(DSR)\} = \Sigma \quad [Y_\alpha(DSR)-Y_\alpha^c(DSR)]^2/4R.$$

REFERENCES

Casel, Särndal, C. E. and Wretman, J. H. (1977), Foundations of Inference in Survey Sampling, John Wiley and Sons, New York.

Chromy, James R. (1979) Sequential Sample Selection Methods. American Statistical Assocation 1979 Proceedings of the Survey Research Methods Section.

Cox, Brenda G. (1980). The weighted sequential hot deck imputation procedure. American Statistical Assocation 1980 Proceedings of the Survey Research Methods Section, 721 to 725

Folsom, Ralph E., Jr. (1974). National Assessment Approach to Sampling Error Estimation, Sampling Error Monograph, Prepared for National Assessment of Educational Progress (First Draft).

Folsom, Ralph E., Jr. (1980). U-statistics estimation of variance components for unequal probability samples with nonadditive interviewer and respondent errors. American Statistical Association 1980 Proceedings of the Survey Research Methods Section, 137 to 142.

Folsom, Ralph E., Jr. (1981). The Equivalence of Generalized Double Sampling Regression Estimators, Weight Adjustment, and Randomized Hot Deck Imputation. Paper prepared for presentation at the American Statistical Assocation 1981 Meetings Section on Survey Research Methods.

Fuller, W. (1974). "Regression Analysis for Sample Survey," report prepared for the U.S. Bureau of the Census on work conducted under the Joint Statistical Agreement, Iowa State University, Ames, Iowa.

Gray G. B. (1975). Components of variance model in multi-stage stratified sample. Survey Methodology, 1: 27 to 43.

Larsen, W. A. and Terpenning, I. J. (1971). Talk presented at the American Statistical Assocation meeting. Fort Collins, Colorado.

Denby, L. and W. A. Larsen (1977). Robust regression estimators compared via monte carlo. Communications in Statistics. Theory and Methods, A6(4), 335-362.