THE SURVEY OF INCOME AND
PROGRAM PARTICIPATION

REGRESSION WEIGHTING METHODS FOR SIPP DATA

No. 202

A. B. An, F. J. Breidt & W. A. Fuller
Iowa State University

U. S. Department of Commerce BUREAU OF THE CENSUS

REGRESSION WEIGHTING METHODS FOR SIPP DATA

Anthony B. An, F. Jay Breidt, and Wayne A. Fuller, Iowa State University
Anthony B. An, Statistical Laboratory, Iowa State University, Ames, Iowa 50011

Key Words: Nonresponse, Two-phase Estimation, Multi-phase Estimation

I. INTRODUCTION

The Census Bureau designed the Survey of Income and Program Participation (SIPP) to provide improved information on income and participation in government programs. Characteristics associated with persons and households which may have impact on income and program participation are collected in the SIPP surveys.

The SIPP is a multistage stratified systematic sample of the noninstitutionalized resident population of the United States. The sample is the sum of four equal sized rotation groups. Each month one rotation group was interviewed. One cycle of four interviews for the four groups is called a wave. Several waves which cover a period of time are called a panel. For example, Panel 1987, which contains seven waves, is the sample of the SIPP-interviewed people from February 1987 through May 1989. The survey produces two kinds of estimates: cross-sectional and longitudinal. We consider estimation for the panel 1987 longitudinal sample. In order to be a part of the longitudinal sample, the respondent must provide data at each of seven interview periods. About 80% of those that responded at the first interview (Wave One) also responded at the remaining six interviews. A total of 30,766 people interviewed in Wave One were eligible for the 1987 panel longitudinal sample. A total of 24,429 individuals completed all seven interviews. Estimation for the longitudinal sample uses information from all Wave One respondents and also uses control information from the Current Population Survey. We compare alternative estimators that use the information in different ways.

Longitudinal estimators are derived from the weights assigned to the people in the longitudinal sample. Many weighting procedures have been investigated for the longitudinal sample. The current weighting scheme at the U.S. Census Bureau is described by Waite (1990). The procedure makes two adjustments to the base weights, where the base weights are the reciprocals of the probabilities of selection. The adjustments attempt to compensate for nonresponse and undercoverage, using variables thought to be highly correlated with SIPP variables of interest. The first stage adjustment is of the post stratification type. The cells are defined by characteristics of people who were eligible in the Wave One sample. The second stage adjustment is a raking procedure performed after the first adjustment using data from the Current Population Survey as controls.

We treat the Panel 1987 SIPP data as a three-phase sample. We consider the phase I sample to be the Current Population survey. In the analysis, we assume zero error in these estimates. The phase II sample is the 1987 wave one data. Phase II included all the people who were eligible and participated in the survey during Wave One. The phase III sample is defined as the subsample from phase II which includes all people who participated in the survey from Wave One through Wave Seven unless they died or moved to an ineligible address. The phase III sample is also called the longitudinal sample of panel 1987. We will treat the SIPP sample as a stratified (72 strata), cluster probability sample, where the cluster is a household.

II. NOTATION AND SIMPLE ESTIMATORS

Throughout this paper, the subscript will indicate the individual. For example, x_{ijk} is the vector of observations on the x-variables for the k-th individual in the j-th cluster of stratum i, where

$$x_{ijk} = \left(x_{ijk1}, x_{ijk2}, \ldots, x_{ijkp}\right),$$

$i = 1, 2, \ldots, L$ is the stratum identification, $j = 1, \ldots, n_i$ is the cluster identification, $k = 1, 2, \ldots, m_{ij}$ is the element-within-cluster identification, and x_{ijkl} is the ijk-th observation for the l-th variable, where $l = 1, 2, \ldots, p$. Characteristics in different samples are identified by I, II, or III according to the phase. In sample τ, we define the data matrices

$$\left[X^{(\tau)}, Y^{(\tau)}, Z^{(\tau)}\right] = \left[\left(x_{ijk}\right), \left(y_{ijk}\right), \left(z_{ijk}\right)\right]$$

which is an $n^{(\tau)} \times (p+q+r)$ matrix, and

$$G^{(\tau)} = \left[1, X^{(\tau)}\right], \quad E^{(\tau)} = \left[1, Y^{(\tau)}\right],$$

where $\tau = II, III$, $n^{(II)}$ and $n^{(III)}$ are the total number of elements in Sample II and Sample III, respectively. If no confusion will result, the sample marks will be omitted and we will simply write, for example, X.
The x-variables are control variables for phase I, the Y-variables are control variables for phase II, and the Z-variables are the variables of interest. The initial

weights matrices are denoted by $W^{(II)} = \left(W_{ijk}^{(0,II)} \right)$ and $W^{(III)} = \left(W_{ijk}^{(0,III)} \right)$, respectively.

We assume that in the phase I sample, only X-variables are observed and that the vector of population totals of the X-variables, denoted by X_I, is available. In the phase II sample, we observe Y and X, and in the phase III sample, we observe X, Y, and Z.

We will consider regression estimation and the regression coefficient matrices are identified by the sample phase where the regression is applied. For example, $\hat{\beta}_{Y \cdot X}^{(II)}$ is the least squares estimate of the $p \times q$ regression coefficient matrix $\beta_{Y \cdot X}$ obtained by regressing Y on X in Sample II. Therefore, we have

$$\begin{pmatrix} \hat{\beta}_0^{(\tau)} \\ \hat{\beta}_{Y \cdot X}^{(\tau)} \end{pmatrix} = \left(G^{(\tau)'} W^{(\tau)} G^{(\tau)} \right)^{-1} G^{(\tau)'} W^{(\tau)} Y^{(\tau)}, \qquad (1)$$

where $(\tau = II, III)$. The total number of elements in the population is denoted by N and the population means of the variables are denoted by μ.

A subscript indicating the phase of the sample is applied to estimated totals. For example,

$$\hat{X}_{II} = \sum_{i=1}^{L} \sum_{j=1}^{n_i^{(II)}} \sum_{k=1}^{m_{ij}^{(II)}} w_{ijk}^{(0,II)} x_{ijk} \qquad (2)$$

is the estimated total for X computed from Sample II using the initial weights. Let f_i be the sampling rate for the i-th stratum, where $f_i = N_i^{-1} n_i$ and let m_{ij} be the number of elements in the ij-th cluster. Then the estimated covariance matrix for $\hat{X}_{II}$ is

$$\hat{v}\left(\hat{X}_{II}\right) = \sum_{i=1}^{L} (n_i - 1)^{-1} n_i (1 - f_i) \sum_{j=1}^{n_i} \left(x_{ij\cdot} - \bar{x}_{i\cdot} \right)' \left(x_{ij\cdot} - \bar{x}_{i\cdot} \right), \quad (3)$$

where

$$x_{ij\cdot} = \sum_{k=1}^{m_{ij}} w_{ijk}^{(0,II)} x_{ijk}, \quad \bar{x}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij\cdot}.$$

Similarly, if we have a variable Y, with the ijk-th observation y_{ijk}, then the estimated covariance matrix between $\hat{X}$ and $\hat{Y}$ is

$$\hat{\mathrm{Cov}}(\hat{X}_{II}, \hat{Y}_{II}) = \sum_{i=1}^{L} (n_i - 1)^{-1} n_i (1 - f_i) \sum_{j=1}^{n_i} \left(x_{ij\cdot} - \bar{x}_{i\cdot} \right)' \left(y_{ij\cdot} - \bar{y}_{i\cdot} \right) \quad (4)$$

These are the basic estimates for totals based on weights associated with the sampling design.

III. ESTIMATORS

We will compare the approximate variances of three estimation procedures for the SIPP data. The three procedures use the auxiliary information in different ways.

A. Three-Phase Estimator (Estimation Scheme One)

We give the steps for constructing the three-phase regression estimator.

Step 1. In Sample II, construct weights by regressing Y on X. Let the regression weights be

$$w_{ijk}^{(1,II)} = w_{ijk}^{(0,II)} + \left[0, X_I - \hat{X}_{II} \right] \left(G^{(II)'} W^{(II)} G^{(II)} \right)^{-1} \begin{bmatrix} 1 \\ x_{ijk}' \end{bmatrix} w_{ijk}^{(0,II)} \qquad (5)$$

$i = 1, \ldots, L; \quad j = 1, \ldots, n_i^{(II)}; \quad k = 1, 2, \ldots, m_{ij}^{(II)}$.

The weights are such that $\sum_{ijk} w_{ijk}^{(1,II)} \left[1, x_{ijk} \right] = \left[N, X_I \right]$.

Step 2. In Sample II, estimate the mean of Y, μ_Y, using the weights in (5)

$$\hat{\mu}_Y^{(1)} = \frac{1}{N} \sum_{ijk} w_{ijk}^{(1,II)} y_{ijk} = \hat{\bar{Y}}_{II} + \left(\bar{X}_I - \hat{\bar{X}}_{II} \right) \hat{\beta}_{Y \cdot X}^{(II)}, \qquad (6)$$

where $\hat{\beta}_{Y \cdot X}^{(II)} = \left(G^{(II)'} W^{(II)} G^{(II)} \right)^{-1} G^{(II)'} W^{(II)} Y^{(II)}$.

Step 3. In Sample III, using (6) as the controls, regress Z on X and Y to construct the regression weights

$$w_{ijk}^{(1,III)} = w_{ijk}^{(0,III)} \left\{ 1 + N \left[0, \bar{X}_I - \hat{\bar{X}}_{III}, \hat{\mu}_Y^{(1)} - \hat{\bar{Y}}_{III} \right] \left(F^{(III)'} W^{(III)} F^{(III)} \right)^{-1} \left[1, x_{ijk}, y_{ijk} \right] \right\} \qquad (7)$$

where $\left(\hat{\bar{X}}_{III}, \hat{\bar{Y}}_{III} \right) = \left[\sum_{ijk} w_{ijk}^{(0,III)} \right]^{-1} \sum_{ijk} w_{ijk}^{(0,III)} \left(x_{ijk}, y_{ijk} \right)$,

and $F^{(III)} = \left[1, X^{(III)}, Y^{(III)} \right]$.

Step 4. In Sample III, estimate μ_Z based on the weights in (7):

$$\hat{\mu}_Z^{(1)} = \left[\sum_{ijk} w_{ijk}^{(1,III)} \right]^{-1} \sum_{ijk} w_{ijk}^{(1,III)} z_{ijk} = \hat{\bar{Z}}_{III} + \left[\bar{X}_I - \hat{\bar{X}}_{III}, \hat{\mu}_Y^{(1)} - \hat{\bar{Y}}_{III} \right] \hat{\beta}_{Z \cdot XY}, \qquad (8)$$

Table 1. Estimated Means 1987 SIPP Panel Data

Characteristic	Estimate #1 ($)	s.e. #1	s.e. #2 / s.e. #1	s.e. #3 / s.e. #1	Mean s.e. / 3-Ph. s.e.	Reg. *X* s.e. / 3-Ph. s.e.	2-Ph. *Y* s.e. / 3-Ph. s.e.
Jan 87 Personal Income	982.4	7.71	1.003	1.012	1.240	1.044	1.174
Jan 89 Personal Income	1038.2	7.64	1.003	1.007	1.235	1.036	1.172
Jan 87 Personal Earnings	755.4	7.09	1.003	1.011	1.247	1.044	1.173
Jan 89 Personal Earnings	791.9	6.73	1.003	1.007	1.272	1.036	1.201
Jan 87 Family Income	2744.6	24.00	1.001	0.992	1.138	1.046	1.084
Jan 89 Family Income	2849.7	23.92	1.000	0.993	1.125	1.034	1.081
Jan 87 Family Earnings	2247.4	23.44	1.000	0.989	1.160	1.041	1.106
Jan 89 Family Earnings	2313.7	21.74	0.999	0.991	1.173	1.034	1.125
Jan 87 Family Property Income	150.6	5.39	1.000	1.000	1.053	1.016	1.035
Jan 89 Family Property Income	153.5	5.26	1.000	1.000	1.048	1.011	1.033
Jan 87 Family MTT	31.2	1.70	1.000	0.993	1.073	1.016	1.051
Jan 89 Family MTT	29.3	1.67	1.003	1.000	1.055	1.019	1.031
Jan 87 Family Other Income	315.3	5.72	1.000	0.998	1.186	1.006	1.166
Jan 89 Family Other Income	353.3	8.79	1.000	0.999	1.085	1.002	1.076
Jan 87 HH Income	2819.9	24.32	1.001	0.993	1.128	1.046	1.074
Jan 89 HH Income	2923.5	23.96	1.000	0.993	1.120	1.035	1.076
Jan 87 HH Earnings	2311.7	23.75	1.000	0.990	1.154	1.042	1.099
Jan 89 HH Earnings	2364.9	21.90	0.999	0.991	1.170	1.035	1.120
Jan 87 HH Property Income	152.4	5.40	1.000	1.000	1.053	1.016	1.035
Jan 89 HH Property Income	155.0	5.28	1.000	1.000	1.048	1.012	1.033
Jan 87 HH MTT	32.9	1.84	1.000	0.994	1.066	1.015	1.045
Jan 89 HH MTT	30.3	1.70	1.003	1.000	1.054	1.019	1.031
Jan 87 HH Other Income	322.8	5.91	1.000	0.998	1.179	1.006	1.160
Jan 89 HH Other Income	360.6	8.86	1.000	0.999	1.085	1.002	1.076
Jan 87 Labor Force	45.9	0.22	1.021	1.044	1.532	1.070	1.407
Jan 87 Labor Force	47.0	0.24	1.008	1.014	1.444	1.028	1.360

*HH = Household, **MTT = Means Tested Transfers

phase estimator, and the ratio of the standard error of the two-phase estimator using only *Y*-variables to the standard error of the three-phase estimator. As expected, each of these three procedures is uniformly inferior to the three-phase estimator.

Also, the regression procedure using *X*-variables is uniformly superior to the two-phase estimator using only *Y*-variables. The gains from using the *Y*-variables in addition to the *X*-variables ranges from 0.2% for Jan. '89 "Other Income" to 7.0% for Jan. '87 Labor Force Status.

REFERENCES

Breidt, F. J. and Fuller, W. A. (1993). Regression weighting for multiphase samples. *Sankhyā Series B*, 55, 297-309.

Folsom, R. E. and Witt, M. B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. Technical report. Research Triangle Institute.

Mosbacher, R. A., Darby, M. R. and Bryant, B. E. (1991). Survey of income and program participation user's guide. 2nd ed. Technical documentation. U.S. Dept. of Commerce, Washington, D.C.

Petroni, R. J., Singh, R. P. and Kasprzyk, D. (1992). Longitudinal weighting issues and associated research for the SIPP. ASA Proc. of Survey Rsch. Methods Sect., 548-553.

Waite, P. J. (1990). Sipp 1987: specifications for panel file longitudinal weighting of persons. Internal Census Bureau memorandum from Waite to Courtland. June 1, 1990.

IV. APPLICATION TO THE SIPP DATA

We compare regression weighting methods for the Panel 1987 data from SIPP. Sample I is the Current Population Survey. We assume that there is zero error for estimated means from Sample I. Sample II is the Panel 1987 Wave One sample. The sample size of Sample II is 30,766 individuals. Initial weights for Sample II were weights constructed using the Census Bureau control variables.

Sample III is the Panel 1987 longitudinal sample. The sample size of Sample III is 24,429 individuals. The initial weight for Sample III is the weight for Sample II multiplied by the ratio of sample sizes.

Equation (11) defines the weights used in calculating the covariances between means in different samples. The weights used for Sample III in these calculations depend on the way in which Sample III is selected from Sample II. In the SIPP situation the sample is self selecting so that it is necessary to use a model for the selection procedure. The model used in our variance comparison is that Sample III is a simple random sample from Sample II. Under this assumption, the weight for an element appearing in Sample III is a simple multiple of the Sample II weight.

The regression variables are based on the non-interview adjustment cells and on the Current Population Survey variables used by the Census Bureau to construct weights for the Panel 1987 longitudinal sample. The X-variables are the variables associated with the second-stage adjustment used by the Census Bureau. The second-stage adjustment variables are based on gender, age, race, family type, and household type. There are 97 X variables in our analysis.

The Y variables are indicator variables for the non-interview adjustment cells in the first stage adjustment procedure described in Waite (1990). The non-interview adjustment cells are formed using variables such as level of income, race, education, type of income, type of assets, labor force status, and employment status. There are 79 Y variables for the 80 cells used in our analysis. The Z variables used in our analysis are Personal Income, Personal Earnings, Family Income, Family Earnings, Family Property Income, Family Means Tested Transfers, Family Other Income, Household Earnings, Household Property Income, Household Means Tested Transfers, and Household Other Income. All variables are recorded for January 1987 and for January 1989. For example, Personal Income for January 1987 was the total income of the person in January of 1987. Family income for January 1989 is the total income of the family with which the interviewed person lived when the survey was conducted. Similarly, Household Earnings for January 1987 is the total earnings of the household in which the interviewed person lived. The Census Bureau defined family and household differently. The household is the sample unit for the SIPP. A household may have more than one family. The terms income, earnings, property income, means-tested income transfers and "other income" are different sources of income for individuals and households.

Estimated standard errors for the three schemes are compared in Table 1. The estimated means of the Z variables are listed in the column of "Estimate".

These estimates were calculated using the three-phase estimator. Estimates of the means computed by other schemes are omitted to simplify the table. The estimated standard errors for the means from scheme #1 are listed under the column "s.e. #1"

The ratios of estimated standard errors from other schemes to the one from scheme #1 are also listed in the table. The difference among the standard errors from the three schemes are small. Because the phase III sample is about 80% of the phase II sample, there must be very large differences in the regression correlations to produce noticeable differences between the standard errors.

If the regression coefficients were computed using cluster totals, the three-phase estimator would always dominate the other two estimators to the degree of accuracy employed in the Taylor approximations. Because the regression coefficients are computed using individuals as observations, it is possible for the estimated standard deviations for schemes two and three to be less than the estimated standard deviation for three-phase estimation.

The results are mildly surprising in that procedure two, the approximation to the current Census Bureau procedure, performs marginally better than the other two procedures. It must be realized that these are estimated variances and, in particular, that the ratio of the variance of the phase II sample to the variance of the phase III sample is estimated. There may be a hidden bias in that the variables used in the analysis are those selected by the Census Bureau.

The variance approximations assume random sampling is used to select the phase III sample from phase II. Of course, this is not true and a primary objective of the first adjustment is to reduce the nonresponse bias. Unfortunately, outside information would be required in order to compare the bias properties of the three estimators.

Table 1 also contains the ratio of the standard error of the mean of Z in the small sample to the standard error of the three-phase estimator, the ratio of the standard error of the regression estimator using only X-variables to the standard error of the three-

$$u_{ijk}^{(2,III)} = w_{ijk}^{(0,III)}\left\{1+N\left[0, \hat{\mu}_Y^{(2)} - \hat{\bar{Y}}_{III}\right]\right.$$

$$\left.\left(E^{(III)\prime} W^{(III)} E^{(III)}\right)^{-1}\left[1, y_{ijk}\right]'\right\}. \qquad (15)$$

These weights satisfy $\sum_{ijk} u_{ijk}^{(2,III)}\left[1, y_{ijk}\right] = N\left[1, \hat{\mu}_Y^{(2)}\right]$.

Step 3. Estimate μ_Z and μ_X using weights (15):

$$\dot{\mu}_Z^{(2)} = \sum_{ijk} u_{ijk}^{(2,III)} z_{ijk} = \hat{\bar{Z}}_{III} + \left(\hat{\bar{Y}}_{II} - \hat{\bar{Y}}_{III}\right)\hat{\beta}_{Z\cdot Y}^{(III)}.$$

(16)

$$\dot{\mu}_X^{(2)} = \sum_{ijk} u_{ijk}^{(2,III)} x_{ijk} = \hat{\bar{X}}_{III} + \left(\hat{\bar{Y}}_{II} - \hat{\bar{Y}}_{III}\right)\hat{\beta}_{X\cdot Y}^{(III)}.$$

Step 4. Construct weights using the regression of Z on X, and using $\bar{X}_I$ as the control:

$$w_{ijk}^{(2,III)} = u_{ijk}^{(2,III)}\left\{1+N\left[0, \bar{X}_I - \dot{\mu}_X^{(2)}\right]\right.$$

$$\left.\left(G^{(III)\prime} U_3^{(III)} G^{(III)}\right)^{-1}\left[1, x_{ijk}\right]'\right\} \qquad (17)$$

where $U_3^{(III)} = \mathrm{diag}\left(u_{ijk}^{(2,III)}\right)$.

Step 5. Estimate μ_Z using weights (17):

$$\tilde{\mu}_Z^{(2)} = \sum_{ijk} w_{ijk}^{(2,III)} z_{ijk} = \dot{\mu}_Z^{(2)} + \left(\bar{X}_I - \dot{\mu}_X^{(2)}\right)\hat{\beta}_{Z\cdot X}^{(III)}. \qquad (18)$$

The estimate of the covariance matrix of $\tilde{\mu}_Z^{(2)}$ is based on the Taylor expansion

$$\tilde{\mu}_Z^{(2)} = \hat{\bar{Z}}_{III} + K\gamma + O_p\left(n^{(III)-1}\right),$$

where $K = \left[\bar{X}_I - \hat{\bar{X}}_{III}, \hat{\bar{Y}}_{II} - \hat{\bar{Y}}_{III}\right]$ and $\gamma = \left[\beta'_{Z\cdot X}, \left(\beta_{ZY} - \beta_{X\cdot Y}\beta_{Z\cdot X}\right)'\right]'$. Using the same procedure as used for three-phase estimation, we can estimate $\mathrm{Cov}\left(\tilde{\mu}_Z^{(2)}\right)$.

C. Estimation Scheme Three

Estimation scheme three differs from scheme two only in that the totals for the first nonresponse adjustment are regression estimated totals using the Current Population Survey data as control variables. We outline the steps in the estimation.

Step 1. As in Steps 1 - 2 of three-phase estimation, define regression weights for Sample II and estimate the mean of Y:

$$w_{ijk}^{(3,II)} = \left\{1+\left[0, X_I - \hat{X}_{II}\right]\left(G^{(II)\prime} W^{(II)} G^{(II)}\right)^{-1}\left[1, x_{ijk}\right]'\right\}$$

$$\hat{\mu}_Y^{(3)} = \sum_{i,j,k} w_{ijk}^{(3,II)} y_{ijk} = \hat{\bar{Y}}_{II} + \left(\bar{X}_I - \hat{\bar{X}}_{II}\right)\hat{\beta}_{Y\cdot X}^{(II)}. \qquad (20)$$

Step 2. In Sample III, regress Z on Y, using the $\hat{\mu}_Y^{(3)}$ in (20) as the control for Y, to create weights

$$u_{ijk}^{(3,III)} = w_{ijk}^{(0,III)}\left\{1+N\left[0, \hat{\mu}_Y^{(3)} - \hat{\bar{Y}}_{III}\right]\left(E^{(III)\prime} W^{(III)} E^{(III)}\right)^{-1}\left[1, y_{ijk}\right]'\right\}. \qquad (21)$$

These weights satisfy

$$\sum_{ijk} u_{ijk}^{(3,III)} y_{ijk} = N\left[1, \hat{\mu}_Y^{(3)}\right].$$

Step 3. In Sample III, use the weights in (21) to estimate the mean of X and Z:

$$\dot{\mu}_Z^{(3)} = \frac{1}{N}\sum u_{ijk}^{(3,III)} z_{ijk} = \hat{\bar{Z}}_{III} + \left(\hat{\mu}_Y^{(3)} - \hat{\bar{Y}}_{III}\right)\hat{\beta}_{Z\cdot Y}^{III},$$

(22)

$$\dot{\mu}_X^{(3)} = \frac{1}{N}\sum u_{ijk}^{(3,III)} x_{ijk} = \hat{\bar{X}}_{III} + \left(\hat{\mu}_Y^{(3)} - \hat{\bar{Y}}_{III}\right)\hat{\beta}_{X\cdot Y}^{(III)}.$$

Step 4. In Sample III, construct the regression weights, using the regression of Z on X and $\dot{\mu}_Z^{(3)}$ and $\dot{\mu}_X^{(3)}$ as the controls, to create

$$w_{ijk}^{(3,III)} = u_{ijk}^{(3,III)}\left\{1+\left[0, X_I - \hat{X}_{II}\right]\left(G^{(III)\prime} U_2^{(III)} G^{(III)}\right)^{-1}\left[1, x_{ijk}\right]'\right\} \qquad (23)$$

where $U_2^{(III)} = \mathrm{diag}\left(u_{ijk}^{(3,III)}\right)$. These weights satisfy

$$\sum_{ijk} w_{ijk}^{(3,III)}\left[1, x_{ijk}\right] = N\left[1, \dot{\mu}_X^{(3)}\right].$$

Step 5. Estimate μ_Z using weights (23),

$$\tilde{\mu}_Z^{(3)} = \sum_{ijk} w_{ijk}^{(3,III)} z_{ijk} = \dot{\mu}_Z^{(3)} + \left(\bar{X}_I - \dot{\mu}_X^{(3)}\right)\hat{\beta}_{Z\cdot X}^{(III)} \qquad (24)$$

The covariance matrix, $\mathrm{Cov}\left(\tilde{\mu}_Z^{(3)}\right)$, can be estimated as described for three-phase estimation.

where

$$\left[\hat{\beta}_0^{(III)}, \hat{\beta}'_{Z\cdot XY}\right] = \left(F^{(III)'} W^{(III)} F^{(III)}\right)^{-1} \left(F^{(III)'} W^{(III)} Z^{(III)}\right).$$

To estimate the covariance matrix of $\tilde{\mu}_Z^{(1)}$, we use the Taylor expansion,

$$\tilde{\mu}_Z^{(1)} = \hat{\bar{Z}}_{III} + H\delta + O_p\left(n^{(III)-1}\right) \qquad (9)$$

where

$$H = \left[\bar{X}_I - \hat{\bar{X}}_{III}, \bar{X}_I - \bar{X}_{II}, \hat{\bar{Y}}_{II} - \hat{\bar{Y}}_{III}\right] \text{ and } \delta = \left[\beta'_{Z\cdot XY}, (\beta_{Y\cdot X}\beta_{Z\cdot XY})', \beta'_{Z\cdot XY}\right].$$ Thus

$$V\left(\tilde{\mu}_Z^{(1)}\right) \approx V\left(\hat{\bar{Z}}_{III}\right) + \operatorname{cov}\left(\hat{\bar{Z}}_{III}, H\right)\delta + \delta' \operatorname{cov}\left(H, \hat{\bar{Z}}_{III}\right) + \delta' V(H)\delta. \qquad (10)$$

Covariance matrices between two mean estimators from different samples are estimated using the larger sample by assigning the observations not in the small sample zero weights, and noting that estimated means are ratios. For example, the weights for Sample II that can be used to construct Sample III estimates are

$$\ddot{w}_{ijk}^{(0,II)} = \begin{cases} w_{ijk}^{(0,III)} & \text{if } (i,j,k) \in III \\ 0 & \text{otherwise} \end{cases} \qquad (11)$$

for $(i,j,k) \in II$. Then

$$\hat{X}_{III} = \sum_{III} w_{ijk}^{(0,III)} x_{ijk} = \sum_{II} \ddot{w}_{ijk}^{(0,II)} x_{ijk}.$$

If we assume the finite population correction is negligible, some estimated covariance matrices are

$$\hat{\operatorname{Cov}}\left(\hat{\bar{Z}}_{III}, \hat{\bar{Y}}_{II}\right) = \sum_{i=1}^{L} \left(n_i^{(II)} - 1\right)^{-1} n_i^{(II)} \sum_{j=1}^{n_i^{(II)}} \left(a_{ij\cdot} - \bar{a}_{i\cdot}\right)' \left(b_{ij\cdot} - \bar{b}_{i\cdot}\right) \qquad (12)$$

$$\hat{V}\left\{\hat{\bar{Z}}_{III}\right\} = \sum_{i=1}^{L} \left(n_i^{(II)} - 1\right)^{-1} n_i^{(II)} \sum_{j=1}^{n_i^{(II)}} \left(a_{ij\cdot} - \bar{a}_{i\cdot}\right)' \left(a_{ij\cdot} - \bar{a}_{i\cdot}\right),$$

$$\hat{V}\left\{\hat{\bar{Y}}_{III}\right\} = \sum_{i=1}^{L} \left(n_i^{(II)} - 1\right)^{-1} n_i^{(II)} \sum_{i=1}^{n_i^{(II)}} \left(c_{ij\cdot} - c_{i\cdot}\right)' \left(c_{ij\cdot} - \bar{c}_{i\cdot}\right),$$

and

$$\hat{V}\left\{\hat{\bar{Y}}_{II}\right\} = \sum_{i=1}^{L} \left(n_i^{(II)} - 1\right)^{-1} n_i^{(II)} \sum_{i=1}^{n_i^{(II)}} \left(b_{ij\cdot} - \bar{b}_{i\cdot}\right)' \left(b_{ij\cdot} - \bar{b}_{i\cdot}\right), \qquad (13)$$

where

$$\left(a_{ij\cdot}, b_{ij\cdot}, c_{ij\cdot}\right) = \hat{N}^{-1} \sum_{k=1}^{m_{ij}^{(II)}} \ddot{w}_{ijk}^{(0,II)} \left[\left(z_{ijk} - \hat{\bar{Z}}_{III}\right), \left(y_{ijk} - \hat{\bar{Y}}_{II}\right), \left(y_{ijk} - \hat{\bar{Y}}_{III}\right)\right].$$

$\left(\bar{a}_{i\cdot}, \bar{b}_{i\cdot}, \bar{c}_{i\cdot}\right) = n_i^{-1} \sum_{j=1}^{n_i} \left(a_{ij\cdot}, b_{ij\cdot}, c_{ij\cdot}\right), i = 1, 2, \ldots, L$, and the weights are such that

$$\hat{N} = \sum_{ijk \in II} w_{ijk}^{(0,II)} = \sum_{ijk \in III} w_{ijk}^{(0,III)} = \sum_{ijk \in II} \ddot{w}_{ijk}^{(0,II)}.$$

To be totally correct, the multiplier in $\hat{V}\left\{\hat{\bar{Z}}_{III}\right\}$ should be $\left(n_i^{(III)} - 1\right)^{-1} n_i^{(III)}$, where $n_i^{(III)}$ is the number of primary sampling units in Sample III. We use the multiplier $\left(n_i^{(II)} - 1\right)^{-1} n_i^{(II)}$ for simplicity for both sample sizes, because, with about 100 primary sampling units per stratum, the multiplier has little effect. We estimate δ in (10) by the least squares procedure of Step 3. The estimated covariance matrices and regression coefficients are used to estimate the covariance matrix, $\operatorname{Cov}\left(\tilde{\mu}_z^{(1)}\right)$.

B. Estimation Scheme Two

Estimation scheme two is an approximation to the procedure currently used by the Census Bureau to construct weights for the SIPP panel. In this procedure, the information from the respondents of wave one is used to construct weights for the panel adjusted for nonresponse. Then the population information from the Current Population Survey is used to create final weights. We give the steps required to construct the estimator.

Step 1. In Sample II, estimate the controls for Y using initial sampling weights:

$$\hat{Y}_{II} = \sum w_{ijk}^{(0,II)} y_{ijk}, \qquad \hat{\mu}_Y^{(2)} = \hat{\bar{Y}}_{II} = n^{-1} \hat{Y}_{II}. \qquad (14)$$

Step 2. In Sample III, construct weights using $\hat{\mu}_Y^{(2)}$ as the population control: