

ADJUSTMENTS FOR NONRESPONSE IN LONGITUDINAL SURVEYS

W. A. Fuller and A. B. An¹

ABSTRACT

Methods of using characteristics of nonrespondents and auxiliary data to make adjustments for nonresponse are investigated. Properties of ordinary regression estimators, and of regression estimators with initial weights based on estimated response probabilities are obtained. The methods are applied to data from the United States Survey of Income and Program Participation (SIPP).

KEY WORDS: Sample surveys; Regression Estimation; Missing Data.

RÉSUMÉ

Sont étudiées ici, les méthodes utilisant les caractéristiques des non-répondants et les données auxiliaires afin de faire l'ajustement pour la non-réponse. Les propriétés des estimateurs par régression simple, et des estimateurs par régression avec les poids initiaux basées sur les probabilités estimées de la réponse sont établies. Ces méthodes sont appliquées à des données de l'enquête nommée Survey of Income and Program Participation (SIPP), conduite aux États-Unis.

MOTS CLÉS: Sondages; estimation par régression; données manquantes.

1. INTRODUCTION

Longitudinal surveys, also called panel surveys, are characterized by observations made at different time points on some of the elements of the sample. A pure longitudinal sample (a pure panel) of length k is one in which every unit in the sample is observed at each of k time points included in the study. A rotating panel is one in which a unit is observed for a partial set of time points and is not observed for the remaining set of time points in the study. There are many ways in which the observation pattern can be specified. The Canadian Labor Force Survey and the U.S. Current Population Survey are examples of surveys designed to run continuously in which units rotate into the sample for a fixed period (or periods) and then permanently rotate out of the observation set.

There exist an array of designs combining individuals observed at some time points and individuals observed at all time points of the study set of time points. The simplest such design is a two-phase sample in which the observations at the second of two time points is a subsample of those observed at time one. The book edited by Kasprzyk *et al.* (1989) contains an excellent discussion of various aspects of

panel surveys. Duncan and Kalton (1987) discuss different types of repeated surveys and the objectives of such surveys.

The largest fraction of the survey literature has been devoted to rotating surveys. An early study describing the use of least squares to incorporate information from a previous occasion into the estimate of the current occasion is that of Jessen (1942). Patterson (1950) investigated estimation for rotating samples. This work was followed by a number of authors, including Eckler (1955), Rao and Graham (1964), Gurney and Daly (1965), Raj (1965), Wolter (1979), Huang and Ernst (1981), and Kumar and Lee (1983). These authors treated the unknown quantities at each occasion as fixed parameters. Blight and Scott (1973), Scott and Smith (1974), Scott, Smith and Jones (1977), Smith (1978), and Jones (1979) considered estimation under the assumption that the underlying true values are the realization of a time series.

Essentially, every survey conducted with human respondents has some nonresponse. Therefore, there is an extensive literature on methods of adjusting for nonresponse. References devoted to the general area include Madow *et al.* (1983), Kalton (1983), Little and Rubin (1987), and Lessler and Kalsbeek (1992, Ch. 7).

¹ Wayne A. Fuller, Statistical Laboratory, Iowa State University, Ames, IA 50010; and Anthony B. An, SAS Institute, SAS Campus Drive, Cary, NC, 27513.

These references describe a number of different procedures. We shall be interested in procedures that combine information from several sources to construct estimators. As described, surveys of a longitudinal nature may be designed to observe individuals only some of the time. In addition, nonresponse leads to an undesigned pattern of incomplete observations on individuals.

There is a close relationship between samples with nonresponse and two-phase samples. See Särndal and Swensson (1987), Kott (1994), and An and Fuller (1995). We shall exploit this relationship and the theory of linear estimation in developing estimators for longitudinal studies.

2. REGRESSION ESTIMATION

We shall use linear estimation procedures heavily in our development of estimators. Regression estimation, or a modification of regression estimation, is used for nonresponse adjustment as well as an estimation procedure for repeated surveys with a rotation design. The following material is taken from Fuller, Loughin, and Baker (1994). Assume that a sample containing n units has been selected and that the probability of selecting unit i is π_i . Assume that a k -dimensional vector of population means, denoted by $\bar{\mathbf{X}}=(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)$ is known, that the vector $(y_i, x_{i1}, x_{i2}, \dots, x_{ik})$ is observed for every unit in the sample and that an estimator of the mean of y is desired. We assume that the first element of \mathbf{x}_i is one for all i . Hence, the first element of $\bar{\mathbf{X}}$ is also one. The vector $\mathbf{x}_i=(x_{i1}, x_{i2}, \dots, x_{ik})$ is sometimes called the vector of control variables. A regression estimator of the mean of y is

$$\bar{y}_r = \bar{\mathbf{X}} \hat{\boldsymbol{\beta}}, \quad (2.1)$$

$$\text{where } \hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \mathbf{x}'_i \pi_i^{-1} \mathbf{x}_i \right)^{-1} \sum_{i=1}^n \mathbf{x}'_i \pi_i^{-1} y_i, \quad (2.2)$$

and we have assumed $\sum \mathbf{x}'_i \pi_i^{-1} \mathbf{x}_i$ to be nonsingular. The estimator (2.1) can also be written as

$$\bar{y}_r = \sum_{i=1}^n w_i y_i \quad (2.3)$$

where

$$w_i = \bar{\mathbf{X}} \left(\sum_{i=1}^n \mathbf{x}'_i \pi_i^{-1} \mathbf{x}_i \right)^{-1} \mathbf{x}'_i \pi_i^{-1}, \quad (2.4)$$

and the weights have the property,

$$\sum_{i=1}^n w_i \mathbf{x}_i = \bar{\mathbf{X}}. \quad (2.5)$$

If the vector \mathbf{x}_j is replaced by the vector

$$(1, \mathbf{z}_j) = (1, x_{j2} - \bar{X}_2, x_{j3} - \bar{X}_3, \dots, x_{jk} - \bar{X}_k), \quad (2.6)$$

the estimator can be written in the form

$$\bar{y}_r = \bar{y}_\pi + (\bar{\mathbf{Z}} - \bar{\mathbf{z}}_\pi) \hat{\boldsymbol{\beta}}_z = \bar{y}_\pi - \bar{\mathbf{z}}_\pi \hat{\boldsymbol{\beta}}, \quad (2.7)$$

where $\bar{\mathbf{Z}} = \mathbf{0}$ is the population mean of \mathbf{z}_j , $\bar{\mathbf{z}}_\pi = \bar{\mathbf{x}}_\pi - \bar{\mathbf{X}}$

$$(\bar{y}_\pi, \bar{\mathbf{z}}_\pi) = \left(\sum_{i=1}^n \pi_i^{-1} \right)^{-1} \sum_{i=1}^n \pi_i^{-1} (y_i, \mathbf{z}_i)$$

and

$$\hat{\boldsymbol{\beta}}_z = \left[\sum_{j=1}^n (\mathbf{z}_j - \bar{\mathbf{z}}_\pi)' \pi_i^{-1} (\mathbf{z}_j - \bar{\mathbf{z}}_\pi) \right]^{-1} \sum_{j=1}^n (\mathbf{z}_j - \bar{\mathbf{z}}_\pi)' \pi_i^{-1} y_j.$$

Fuller (1975) gave theory that covers the regression estimator of the mean and of the total. To construct an estimator of the variance of the regression estimator for a stratified two-stage sample, replace our single subscript i with the triple ℓjt . Then, omitting the finite correction term, a variance estimator is

$$\hat{\mathbf{V}} \{ \bar{y}_r \} = (n-k)^{-1} \sum_{\ell=1}^L (n_\ell - 1)^{-1} \sum_{j=1}^{n_\ell} (d_{\ell j} - d_{\ell.})^2,$$

where

$$d_{\ell j} = \sum_{t=1}^{m_{\ell j}} w_{\ell jt} (y_{\ell jt} - \mathbf{x}_{\ell jt} \hat{\boldsymbol{\beta}}),$$

$$d_{\ell.} = n_\ell^{-1} \sum_{j=1}^{n_\ell} d_{\ell j},$$

n_ℓ is the number of sample primary sampling units in stratum ℓ , $m_{\ell j}$ is the number of sample elements in primary sampling unit j of stratum ℓ , $\hat{\boldsymbol{\beta}}$ is the vector of coefficients defined in (2.2), n is the total number of elements in the sample, and $w_{\ell jt}$ is the weight for element t in primary sampling unit j of stratum ℓ .

The theoretical development for regression estimation in surveys typically assumes the sample to be a probability sample from the population. However, given auxiliary information, regression estimation provides a method of reducing nonresponse bias. The

degree to which the bias is reduced depends upon the relationship between the control variables, the variables of interest, and the response probabilities.

Let p_i be the conditional probability of observing the unit given that the unit is selected. Then

$$E\left\{\sum_{i=1}^n \mathbf{x}'_i \pi_i^{-1} \mathbf{x}_i - \sum_{i=1}^N \mathbf{x}'_i p_i \mathbf{x}_i\right\} = \mathbf{0} \quad (2.9)$$

and

$$E\left\{\sum_{i=1}^n \mathbf{x}'_i \pi_i^{-1} y_i - \sum_{i=1}^N \mathbf{x}'_i p_i y_i\right\} = \mathbf{0}. \quad (2.10)$$

Therefore, under conditions such as those used by Fuller (1975),

$$p \lim_{n \rightarrow \infty} (\hat{\beta} - \gamma) = \mathbf{0}, \quad (2.11)$$

where $\hat{\beta}$ is defined in (2.2) and

$$\gamma = \left(\sum_{i=1}^N \mathbf{x}'_i p_i \mathbf{x}_i \right)^{-1} \sum_{i=1}^N \mathbf{x}'_i p_i y_i. \quad (2.12)$$

Then

$$\bar{Y} = \bar{\mathbf{X}} \gamma + \bar{A}, \quad (2.13)$$

where $\bar{A} = N^{-1} \sum_{i=1}^N a_i$ and $a_i = y_i - \mathbf{x}_i \gamma$. Thus, the regression estimator (2.1) will be a consistent estimator of \bar{Y} if $p \lim_{N \rightarrow \infty} \bar{A} = 0$. The probability limit of \bar{A} will be zero if the finite population is a random sample from an infinite population in which the linear model

$$y_i = \mathbf{x}_i \beta + e_i, \quad E\{e_i\} = 0$$

holds for all i .

The mean \bar{A} is zero when p_i is a constant for all i and an element \mathbf{x}_i is one for all i because then

$$\gamma = \beta = \left(\sum_{i=1}^N \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \sum_{i=1}^N \mathbf{x}'_i y_i \quad (2.14)$$

and $\sum_{i=1}^N (y_i - \mathbf{x}_i \beta) = 0$. A sufficient condition for \bar{A} to be zero is the existence of a row vector \mathbf{c} such that

$$\mathbf{c} \mathbf{x}'_i = p_i^{-1} \quad (2.15)$$

for $i = 1, 2, \dots, N$ (see Zyskind, 1967). Thus, if the response probability is a linear function of the control variables, the regression estimator is a consistent estimator of the mean of y . One way in which (2.15) can be satisfied is for the elements of \mathbf{x}_i to be dummy variables that define subgroups and for the response probabilities to be constant in each subgroup. This

situation is sometimes described by saying that elements are missing at random in each subgroup.

The error in $\hat{\beta}$ as an estimator of γ can be approximated by

$$\hat{\beta} - \gamma = \mathbf{G}^{-1} T^{-1} \sum_{i=1}^n \mathbf{x}'_i \pi_i^{-1} a_i,$$

where a_i is defined in (2.13),

$$T = \sum_{i=1}^N p_i \text{ and } \mathbf{G} = T^{-1} \sum_{i=1}^N \mathbf{x}'_i p_i \mathbf{x}_i.$$

Under reasonable assumptions,

$$(\hat{T}, \hat{\mathbf{G}}) = \left(\sum_{i=1}^n \pi_i^{-1}, \hat{T}^{-1} \sum_{i=1}^n \mathbf{x}'_i \pi_i^{-1} \mathbf{x}_i \right)$$

are consistent estimators of T and \mathbf{G} . Thus, the variance of the regression estimator can be estimated by estimating the variance of $\sum_{i=1}^n \mathbf{x}'_i \pi_i^{-1} a_i$. If we assume that the conditional probabilities of response in one primary sampling unit are independent of those in all other primary sampling units and that at least one observation unit is observed in each selected primary sampling unit, then (2.8) remains an appropriate estimator of the variance of the regression estimated mean of y .

3. RESPONSE PROBABILITY AND REGRESSION ESTIMATION

The procedure of making a first adjustment to the selection probabilities and corresponding adjustment to the sample weights is very common in survey sampling. The most frequently used procedure is to form adjustment cells and to ratio adjust the weights in the cell so that the sum of the weights is equal to the estimated (or known) total for the cell (see, for example, Little and Rubin, 1987, p. 250). Frequently, this adjustment is followed by another estimation scheme such as ratio estimation or regression estimation. This two-step procedure is currently used by the U.S. Census Bureau in SIPP (see Waite, 1990). A modification of the procedure using an estimated response probability function is discussed by Folsom and Witt (1994). We consider the theory for such procedures in this section.

We assume that the inverse of the response probability for individual i is given by

$$p_i^{-1} = g(\mathbf{z}_i; \theta^0), \quad (3.1)$$

where θ^0 is the true value of θ , $g(\mathbf{z}_i; \theta)$ is

continuous in θ with continuous first and second derivatives in an open set containing θ^0 for all \mathbf{z}_i . We also assume that p_i is bounded below by a positive number. We assume a finite population of size N that is a sample from an infinite superpopulation. Let a sample of size n be selected from the finite population. We begin the discussion under the assumption of a simple random nonreplacement sample. The vector $(\mathbf{x}_i, \mathbf{z}_i)$ is observed on each of the n elements of the original sample. There may be, and usually will be, elements that appear in both \mathbf{x} and \mathbf{z} . The response mechanism with probabilities p_i produces a smaller sample of size m on which the vector $(Y, \mathbf{x}_i, \mathbf{z}_i)$ is observed. Let δ_i be an indicator variable with $\delta_i=1$ if a response is obtained and $\delta_i=0$ if a response is not obtained. Using the vector $(\delta, \mathbf{x}_i, \mathbf{z}_i)$, the response probability function is estimated. Assume that $\hat{\theta}-\theta=O_p(n^{-1/2})$, where $\hat{\theta}$ is the estimator of θ .

Let

$$\gamma = \left(\sum_{i=1}^N \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i=1}^N \mathbf{x}_i' y_i \quad (3.2)$$

denote the finite population regression vector and let $a_i = y_i - \mathbf{x}_i' \gamma$. We assume

$$\sum_{i=1}^N a_i = 0. \quad (3.3)$$

The sum of the a_i is zero by construction when the vector \mathbf{X}_i contains an element that is identically equal to one.

Let

$$\hat{\gamma} = \left(\sum_{i=1}^m \mathbf{x}_i' \mathbf{x}_i \pi_i^{-1} p_i^{-1} \right)^{-1} \sum_{i=1}^m \mathbf{x}_i' y_i \pi_i^{-1} p_i^{-1} \quad (3.4)$$

$$\tilde{\gamma} = \left(\sum_{i=1}^m \mathbf{x}_i' \mathbf{x}_i \pi_i^{-1} \hat{p}_i^{-1} \right)^{-1} \sum_{i=1}^m \mathbf{x}_i' y_i \pi_i^{-1} \hat{p}_i^{-1}, \quad (3.5)$$

where π_i 's are the selection probabilities used to select the sample of n from N and $\hat{p}_i^{-1} = g(\mathbf{z}_i; \hat{\theta})$. We assume a sequence of finite populations and samples such that

$$\hat{\gamma} - \gamma = O_p(n^{-1/2}) \quad (3.6)$$

$$\mathbf{A}_n^{-1/2} (\hat{\gamma} - \gamma) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}), \quad (3.7)$$

where

$$\mathbf{A}_n = \mathbf{B}_{xx}^{-1} \hat{\mathbf{V}} \left\{ \sum_{i=1}^m \mathbf{x}_i' a_i \pi_i^{-1} p_i^{-1} \right\} \mathbf{B}_{xx}, \quad (3.8)$$

$$\mathbf{B}_{xx} = \sum_{i=1}^m \mathbf{x}_i' \mathbf{x}_i \pi_i^{-1} p_i^{-1}$$

and $\hat{\mathbf{V}} \left\{ \sum_{i=1}^m \mathbf{x}_i' a_i \pi_i^{-1} p_i^{-1} \right\}$ is the estimated variance computed by the appropriate sampling formula. Sufficient conditions for (3.8) and (3.9) are given in Fuller (1975).

Now,

$$\begin{aligned} \hat{p}_i^{-1} - p_i^{-1} &= \frac{\partial g(\mathbf{z}_i; \theta^*)}{\partial \theta'} (\hat{\theta} - \theta^0) \\ &= \frac{\partial g(\mathbf{z}_i; \theta^0)}{\partial \theta'} (\hat{\theta} - \theta^0) \\ &\quad + 0.5 (\hat{\theta} - \theta^0)' \frac{\partial^2 g(\mathbf{z}_i; \theta)}{\partial \theta \partial \theta'} (\hat{\theta} - \theta^0), \end{aligned} \quad (3.9)$$

where θ^* and $\check{\theta}$ are on the line segment joining $\hat{\theta}$ and θ^0 . Let \mathbf{h}_i denote the row vector of first derivatives of $g(\mathbf{z}_i; \theta)$ evaluated at $\theta = \theta^0$ and let $\mathbf{B}(\mathbf{z}_i; \check{\theta})$ denote the matrix of second derivatives of $g(\mathbf{z}_i; \theta)$ evaluated at $\theta = \check{\theta}$. Then we can write

$$\begin{aligned} \hat{p}_i^{-1} - p_i^{-1} &= \mathbf{h}_i (\hat{\theta} - \theta^0) \\ &\quad + 0.5 (\hat{\theta} - \theta^0)' \mathbf{B}(\mathbf{z}_i; \check{\theta}) (\hat{\theta} - \theta^0). \end{aligned} \quad (3.10)$$

We require the sample design to be such that sample moments converge to population moments. Therefore, we assume

$$\begin{aligned} N^{-1} \sum_{i=1}^m \pi_i^{-1} p_i^{-1} \mathbf{W}_i' \mathbf{W}_i \\ - N^{-1} \sum_{i=1}^N \mathbf{W}_i' \mathbf{W}_i = O_p(n^{-1/2}), \end{aligned} \quad (3.11)$$

where we assume one is an element of \mathbf{W}_i and

$$\begin{aligned} \mathbf{W}_i = & \left(y_i, \mathbf{x}_i, \mathbf{z}_i, \mathbf{h}_i, \right. \\ & \left. (\text{vec } \mathbf{x}_i' a_i p_i \mathbf{h}_i)', (\text{vec } \mathbf{x}_i' p_i \mathbf{h}_i)' \right). \end{aligned}$$

We note that some of the moment assumptions can be weakened. We also assume that the sample mean squares of the vector of second derivatives, $\text{vec } \mathbf{B}(\mathbf{z}_i; \theta^0)$, converge to the corresponding

population moments. We have $\tilde{\gamma} - \gamma =$

$$\mathbf{B}_{xx}^{-1} \sum_{i=1}^m \pi_i^{-1} p_i^{-1} \mathbf{x}'_i a_i \left[1 + p_i \mathbf{h}_i (\hat{\theta} - \theta^0) \right] + O_p(n^{-1}). \quad (3.12)$$

Under assumption (3.11), the matrix

$$N^{-1} \sum_{i=1}^m \pi_i \mathbf{x}'_i a_i \mathbf{h}_i - N^{-1} \sum_{i=1}^N p_i \mathbf{x}'_i \mathbf{h}_i a_i = O_p(n^{-1/2}). \quad (3.13)$$

Therefore, $\tilde{\gamma} - \gamma = O_p(n^{-1/2})$ and

$$N^{-1} \sum_{i=1}^N Y_i - \hat{\mu}_y = O_p(n^{-1/2}). \quad (3.14)$$

If we assume

$$\mathbf{C}_h = N^{-1} \sum_{i=1}^N p_i \mathbf{x}'_i \mathbf{h}_i a_i = o_p(1), \quad (3.15)$$

then

$$\tilde{\gamma} - \gamma = \left(\sum_{i=1}^m \mathbf{x}'_i \mathbf{x}_i \pi_i^{-1} p_i^{-1} \right)^{-1} \sum_{i=1}^m \mathbf{x}'_i a_i \pi_i^{-1} p_i^{-1} + o_p(n^{-1/2}). \quad (3.16)$$

The matrix \mathbf{C}_h is $O_p(N^{-1/2})$ if the finite population is a sample from a superpopulation in which a is independent of $(\mathbf{x}_i, \mathbf{z}_i, \pi_i, p_i)$.

4. APPLICATION TO SIPP

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

The SIPP is a multistage stratified (72 strata) cluster systematic sample of the noninstitutionalized resident population of the United States, where the cluster is a household. 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, composed of seven waves, contains the SIPP-interviewed people from February 1987 through May 1989. The survey produces two kinds of estimates: cross-sectional and longitudinal. In order to be a part of the longitudinal sample, the respondent must provide data at each of seven interview periods. About 79% of

those that responded at the first interview (Wave One) of Panel 1987 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 (CPS). We compare alternative estimators that use the information in different ways.

We treat the Panel 1987 SIPP data as a three-phase sample, where the phase I sample is the Current Population Survey. In the analysis, we assume zero error in the estimates of the phase I sample. The phase II sample is the 1987 Wave One data. The phase II included all the people who were eligible and participated in the survey during Wave One. The phase III sample is defined as a subsample from the 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.

The current estimation scheme makes an initial adjustment in the panel weights, equivalent to an initial estimate of response probabilities, based on 80 adjustment cells (see Waite, 1990). The 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 observed at the first interview. The second stage adjustment is a raking procedure based on 97 variables with estimated means taken from the Current Population Survey. These 97 variables are based on gender, age, race, family type and household type.

To extend the estimator of the response probabilities beyond that based on adjustment cells, let \hat{r}_{ijk} denote the estimated response probabilities based on the adjustment cells for individual k in cluster j of stratum i . The \hat{r}_{ijk} are constant in a cell and are the ratio of weighted respondents to the total sample in a cell where the weights are the inverses of the selection probabilities. Let δ_{ijk} denote an indicator variable equal to one if the individual responds on all seven periods and zero otherwise. To reduce the number of variables to be included in the nonlinear estimation of a logistic model, we first computed the linear regression of δ_{ijk} on $(\mathbf{x}_{ijk}, \mathbf{y}_{ijk})$, where \mathbf{x}_{ijk} is the vector of second stage adjustment variables and \mathbf{y}_{ijk} is the vector of indicator variables for the adjustment cells. Let $\hat{\delta}_{ijk}$ be the predicted values from this

regression and let \bar{r} be the mean response rate. Then

$$\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \text{ of the logistic model}$$

$$(1 - p_{ijk})^{-1} = 1 + \exp\{\theta_0 + \theta_1 \log[\hat{r}_{ijk}(1 - \hat{r}_{ijk})^{-1}]\}$$

$$+ \theta_2(\hat{\delta}_{ijk} - \hat{r}_{ijk}) + \theta_3(\hat{\delta}_{ijk} - \hat{r}_{ijk})^2$$

$$+ \theta_4(\hat{r}_{ijk} - \bar{r})(\hat{\delta}_{ijk} - \hat{r}_{ijk})\}$$

was estimated. The estimated vector is

$$\hat{\theta} = (0.035, 1.032, 6.158, -5.280, 6.576).$$

$$(0.052)(0.036)(0.316) (1.794)(2.506)$$

If no other variables are included in the model, $\hat{\theta}_1 = 1.000$. Recall that \hat{r}_{ijk} are based on 80 cells and $\hat{\delta}_{ijk} - \hat{r}_{ijk}$ reflects the effect of 97 variables. Even after adjusting for the degrees of freedom hidden in the variables, the fit indicates that the adjustment cells model for response probability can be improved. However, by the results of Section II, if p_{ijk}^{-1} is well approximated by a linear function of \mathbf{X} , then the regression estimator is approximately unbiased.

Table 1 contains estimated standard errors of three three-phase procedures expressed relative to the estimated standard error of the Census procedure. The

procedure called "3-phase element" is a regression estimation procedure that uses the 97 phase I (CPS) variables and the 79 phase II variables in a three-phase estimator. The regressions at each stage are computed using sums of squares and products based on individuals. The weights used in the regression computations are the initial sampling weights. The "3-phase cluster" procedure is the same estimator with regression coefficients based on cluster totals. The procedure "3-phase cluster, \hat{p}_{ijk} " uses the estimated weights from the logistic function in the calculation of the regression coefficients. There are modest differences in the standard errors. It is not surprising that some estimated standard errors are larger for the procedure that uses estimated response probabilities. If the response probabilities vary, this produces a wider range of weights which can increase the variance.

It is felt that the primary effect of using estimated probabilities will be on the bias of the estimators. In the case of SIPP, few significant differences in the final estimators were observed. One exception was labor force. The estimator using response probabilities produced estimates of the fraction of individuals in the labor force that were larger than those using the initial sampling weights.

Table 1: Estimated standard errors for alternative estimation procedures divided by estimated standard error of Census Procedures

Variable	Procedure		
	3-Phase Element ¹	3-Phase Cluster ¹	3-Phase Cluster ² \hat{p}_{ijk}
Jan 87 Personal Income	0.99	0.96	0.95
Jan 89 Personal Income	0.99	0.99	0.98
Jan 87 Household Income (10's)	1.01	0.97	0.97
Jan 89 Household Income (10's)	1.01	0.98	0.98
Jan 87 Labor Force (%)	0.98	0.94	0.96
Jan 89 Labor Force (%)	0.99	0.98	1.01

¹ Regressions computed using selection probability weights

² Regressions computed using estimated response probability weights

ACKNOWLEDGMENT

This research was partly supported by Cooperative Agreement 43-3AEU-3-80088 with the National Agricultural Statistics Service and the U.S. Bureau of the Census.

REFERENCES

An, A. B., Breidt, F. J., and Fuller, W. A. (1994). "Regression weighting methods for SIPP data", *Proceedings of the Survey Research Methodology Section, American Statistical Association*, 434-439.

- An, A. B., and Fuller, W. A. (1995). "Regression adjustment for nonresponse", *Proceedings of the Survey Research Methodology Section, American Statistical Association*.
- Blight, B. J. N., and Scott, A. J. (1973). "A stochastic model for repeated surveys", *Journal of the Royal Statistical Society, Series B*, 35, 61-66.
- Breidt, F. J., and Fuller, W. A. (1993). "Regression weighting for multiphase samples", *Sankhyā, Series B*, 55, 297-309.
- Cochran, W. G. (1977), *Sampling Techniques, Third Edition*, New York: Wiley.
- Duncan, G. J. and Kalton, G., (1987). "Issues of design and analysis of surveys across time", *International Statistical Review*, 55, 97-117.
- Eckler, A. R. (1955). "Rotation sampling", *Annals of Mathematical Statistics*, 26, 664-685.
- Folsom, R. E., and Witt, M. B. (1994). "Testing a new attrition nonresponse adjustment method for SIPP", Technical report. Research Triangle Institute, Research Triangle Park, North Carolina.
- Fuller, W. A. (1975). "Regression analysis for sample survey", *Sankhyā, Series C*, 37, 117-132.
- Fuller, W. A. (1990). "Analysis of repeated surveys", *Survey Methodology*, 16, 167-180.
- Fuller, W. A., (1996). "Replication Variance Estimation for Two Phase Samples". Unpublished manuscript, Iowa State University, Ames, Iowa.
- Fuller, W. A., and Isaki, C. T. (1981). "Survey design under superpopulation models", in *Current Topics in Survey Sampling*, New York: Academic Press.
- Fuller, W. A., Kennedy, W., Schnell, D., Sullivan, G., and Park, H. J. (1986). *PC CARP*, Statistical Laboratory, Iowa State University, Ames, Iowa.
- Fuller, W. A., Loughin, M. M., and Baker, H. D. (1994). "Regression weighting for the 1987-88 National Food Consumption Survey", *Survey Methodology*, 20, 75-85.
- Gurney, M., and Daly, J. F. (1965). "A multivariate approach to estimation in periodic sample surveys", *Proceedings of the American Statistical Association, Section on Social Statistics*, 242-257.
- Huang, L. R., and Ernst, L. R. (1981). "Comparison of an alternative estimator to the current composite estimator in the Current Population Survey", *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 303-308.
- Jessen, R. J. (1942). "Statistical investigation of a sample survey for obtaining farm facts", *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.
- Jones, R. G. (1979). "The efficiency of time series estimators for repeated surveys", *Australian Journal of Statistics*, 21, 45-56.
- Kalton, G. (1983). *Compensating for Missing Survey Data*, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan.
- Kasprzyk, D., Duncan, G., Kalton, G., and Singh, M. P. (1989). *Panel Surveys*, New York: Wiley.
- Kott, P. S. (1994). "A note on handling nonresponse in sample surveys", *Journal of the American Statistical Association*, 89, 693-696.
- Kumar, S. and Lee, H. (1983). "Evaluation of composite estimation for the Canadian Labor Force Survey", *Survey Methodology*, 9, 1-24.
- Lessler, J. T., and Kalsbeek, W. D. (1992). *Nonsampling Error in Surveys*, New York: Wiley.
- Little, R. J. A., and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, New York: Wiley.
- Madow, W. G., Nisselson, N., and Olkin, I. (eds.) (1983). *Incomplete Data in Sampling Surveys*, New York: Academic Press.
- Patterson, H. D. (1950). "Sampling on successive occasions with partial replacement of units", *The Journal of the Royal Statistical Society, Series B*, 12, 241-255.

- Petroni, R. J., Singh, R. P., and Kasprzyk, D. (1992). "Longitudinal weighting issues and associated research for the SIPP", *Proceedings of the Survey Research Methodology Section, American Statistical Association*, 548-553.
- Raj, D. (1965). "On sampling over two occasions with probability proportionate to size", *Annals of Mathematical Statistics*, 36, 327-330.
- Rao, J. N. K. (1994). "Estimating totals and distribution functions using auxiliary information at the estimation stage", *Journal of Official Statistics*, 10, 153-165.
- Rao, J. N. K., and Graham, J. E. (1964). "Rotation designs for sampling on repeated occasions", *Journal of the American Statistical Association*, 59, 492-509.
- Särndal, C.-E., and Swensson, B. (1987). "A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse", *International Statistical Review*, 55, 279-294.
- 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.
- Scott, A. J., Smith, T. M. F., and Jones, R. G. (1977). "The application of time series methods to the analysis of repeated surveys", *International Statistical Review*, 45, 13-28.
- Smith, T. M. F., (1978). "Principles and problems in the analysis of repeated surveys", pages 201-216, in N. Krishnan Namboodiri, ed. *Survey Sampling and Measurement*, New York: Academic Press.
- 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.
- Wolter, K. (1979). "Composite estimation in finite populations", *Journal of the American Statistical Association*, 74, 604-613.
- Zyskind, G. (1967). "On canonical forms, nonnegative covariance matrices and best and simple least squares linear estimators in linear models", *Annals of Mathematical Statistics*, 38, 1092-1109.