# WEIGHTING SCHEMES FOR HOUSEHOLD PANEL SURVEY

## FINAL REPORT

**Prepared By:**

**Graham Kalton and J. Michael Brick**

**Westat, Inc.,**
**1650 Research Blvd.,**
**Rockville, MD 20850**

## TABLE OF CONTENTS

# 1. INTRODUCTION

National panel surveys of household economics have been mounted in many countries in recent years. The U.S. Panel Study of Income Dynamics (PSID), conducted by the Survey Research Center of the University of Michigan, began in 1968 and has been collecting data on an annual basis since that time (see Hill, 1992, for a description of PSID). Household panel surveys similar to PSID are in progress or are being planned in most European countries, including Belgium, Britain, France, Germany, Hungary, Ireland, Luxembourg, the Netherlands, Spain and Sweden. In 1983, the U.S. Bureau of the Census started to conduct the Survey of Income and Program Participation (SIPP) (Nelson *et al.*, 1985; Kasprzyk, 1988; Jabine *et al.*, 1990; Citro and Kalton, 1993). In 1994, Statistics Canada introduced the Survey of Labour and Income Dynamics (SLID) (Lavallée *et al.*, 1993).

A common feature to most of these household panel surveys is that they start with a national sample of households, and then follow all the members of those households for the life of the panel. Over the course of time, household compositions change in a variety of ways. Some members of original sampled households leave those households to set up on their own or to join other households. For example, a daughter may leave her parental household to get married. New members may join original sampled households, as, for example, when an elderly parent moves in with the family of a child. In order to be able to describe the economic circumstances of sample individuals at different points of time, household panel surveys usually collect data not only for the sample individuals but also for the individuals living with the sample individuals at the particular point of time. These latter individuals are termed associated persons in this paper.

As the panel duration increases, the proportion of associated persons in the sample at a wave rises. Table 1 illustrates this point from the 1984 SIPP panel. In interpreting the results in Table 1 it needs to be borne in mind that the 1984 SIPP panel had two short waves (waves 2 and 8) in which only three-quarters of the sample was interviewed, and also experienced sample reductions in waves 5 and 6 for budgetary reasons. Nevertheless, the results in Table 1 provide a general indication of the increase in the proportions of associated persons over time. After a year associated persons comprised about 8.6 percent of the sample and after

two years they comprised about 12.6 percent of the sample. With a long-term household panel survey, the proportion of associated persons becomes substantial after several years.

**Table 1: Percentage of Persons (Adults and Children) by Wave of Entry to the 1984 SIPP Panel for Each Wave**

| Wave of entry | Wave | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2* | 3 | 4 | 5† | 6† | 7 | 8* | 9 |
| 1 | 100.0 | 96.3 | 93.8 | 91.4 | 89.8 | 88.6 | 87.4 | 86.6 | 86.0 |
| 2 | | 3.7 | 2.7 | 1.6 | 1.4 | 1.3 | 1.2 | 1.0 | 1.1 |
| 3 | | | 3.5 | 3.4 | 2.4 | 2.0 | 1.8 | 1.5 | 1.5 |
| 4 | | | | 3.6 | 3.4 | 2.5 | 2.1 | 1.8 | 1.7 |
| 5 | | | | | 2.9 | 3.0 | 2.1 | 1.8 | 1.6 |
| 6 | | | | | | 2.6 | 2.6 | 1.9 | 1.7 |
| 7 | | | | | | | 3.0 | 3.1 | 2.6 |
| 8 | | | | | | | | 2.3 | 1.8 |
| 9 | | | | | | | | | 2.0 |
| Total | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| Numbers | 53726 | 41443 | 56197 | 55993 | 51075 | 45007 | 44500 | 32334 | 41807 |

*Only 3 of 4 rotation groups were interviewed in waves 2 and 8.
†Sample reductions were made in waves 5 and 6.

Source: Kasprzyk and McMillen (1987).

The purpose of this paper is to consider how weights can be developed for the data collected from both original and associated sample persons to be used in producing unbiased (or approximately unbiased) estimates of population parameters. In considering the analysis of a household panel survey, three different types of analysis may usefully be distinguished:

- Cross-sectional analyses of households at a particular point in time;
- Cross-sectional analyses of individuals at a particular point in time;
- Longitudinal analyses of individuals over a period of time.

These three types of analysis are discussed below. Longitudinal analyses of households over a period of time are not treated here because of the problematic nature of this type of analysis caused by changes in household composition (see, for example, Duncan and Hill, 1985).

In order to prepare for the discussion of weighting schemes, the next section elaborates on the household changes that can occur over time, and the types of individual involved. The following sections then discuss weighting schemes that may be used for the different forms of analysis. The discussion of weighting schemes relies heavily on previous work by Ernst (1989), Gailly and Lavallée (1993), Huang (1984), Judkins *et al.* (1984), Lavallée and Hunter (1992), and Little (1989).

## 2. CHANGES IN POPULATION AND HOUSEHOLD COMPOSITION OVER TIME

In analyzing a panel survey, it needs to be recognized that survey populations change over time. With household panel surveys two types of change need to be distinguished:

- Changes in population composition;
- Changes in household composition.

The composition of a survey population changes over time because some individuals leave the population, some enter the population, and some may leave and join the population more than once. Individuals leave the population through death, emigration, or entering an institution (for surveys of the noninstitutional population). Individuals enter the population through birth (or reaching the specified minimum age), immigration, and leaving an institution.

Households change composition over time for many different reasons, including deaths, births and marriages. Figure 1 illustrates some of the changes that can arise between two points of time. Household 1 at time 1 contains 4 individuals, who end up in three different households (households 1′, 2′ and 3′) at time 2. One of these individuals (D) joins two household members (E and F) from another household at time 1. Household 3 contains a single individual at time 1 who joins a household with two individuals from household 2 at time 2. Households 4, 5 and 6 all combine into a single household at time 2. Household 7 splits into three separate households at time 2. The lower case letters in households 8 and 9 denote persons who were present in the population at one time but not both times. Person $s$ in household 8 left the population between the two time points and person $v$ in household 10′ entered the population in that period.

Consider now a simple sample design of selecting 3 households from the 9 households at time 1 in Figure 1 by systematic sampling. The three possible samples are (1,4,7), (2,5,8) or (3,6,9). Using the enumeration rule that all original sample persons and all persons living with them at time 2 are to be interviewed, it can be seen that the population members at time 2 have unequal selection probabilities. For instance, individuals A, B and C have inclusion probabilities of 1/3 at time 2 since they are included only if household 1 was

sampled at time 1; individuals D, E and F have inclusion probabilities of 2/3 at time 2 since they are included if either household 1 or household 2 was sampled at time 1; and individuals J, K, L, M and N have inclusion probabilities of 1 at time 2 since they are included if either household 4, or household 5, or household 6 was sampled at time 1. Similarly, households 1′ and 2′ have inclusion probabilities of 1/3 at time 2 since they are included only if household 1 was sampled at time 1; household 3′ has an inclusion probability of 2/3 at time 2 since it is included if household 1 or household 2 was sampled at time 1; and household 5′ has an inclusion probability of 1 since it is included if either household 4, household 5, or household 6 was sampled at time 1. Weighting schemes that compensate for these unequal selection probabilities are needed for the analysis of the resultant data.

As noted, households 8 and 9 at time 1 are subject to changes in population composition at time 2. Person $s$ in household 8 leaves the population before time 2, and hence is not represented in the sample at that time. This does not affect cross-sectional estimates at time 2. Person $v$ in household 10′ is a new entrant to the population, and is included in the sample if household 9 was sampled at time 1. In this way the household panel survey enumeration rule incorporates some new entrants into the population for cross-sectional estimates for later time points. However, it does so only for new entrants who join households that contain individuals who were eligible for the initial sample. New entrants who set up their own households are not represented in later waves of the panel. Equally, households composed of only new entrants are not represented in household analyses at later waves.

The failure of household panel surveys to cover households composed of only new entrants presents a problem for cross-sectional analyses of later waves of the panel. If these households and their members constitute a negligible proportion of the population, the solution to the problem may be to simply ignore it. However, if the proportion is appreciable, as can occur in later waves of a long-term panel, alternative solutions may be called for. One possibility is to add a sample of new entrants (e.g., immigrants) to the panel. This solution is, however, often impracticable. Another solution is to limit the population of inference to persons who were members of the population at the start of the panel. With this solution, new entrants found living with sample members are also excluded from the sample. This solution has the virtue of providing the clearest definition of the population of inference. Whether the solution is appropriate depends on whether that definition can adequately satisfy the survey objectives.

Changes in population composition pose problems for longitudinal analyses of individuals. For many purposes, the population of inference is restricted to those who were present in the population throughout the time period of observation specified for the analysis. The inclusion of associated persons in longitudinal analysis also creates problems. If the time period for the longitudinal analysis starts at the beginning of the panel, the analysis can be restricted straightforwardly to original sample members. If the time period starts later, it is tempting to include both original sample members and associated persons joining the panel before the start of the analytic time period. However, the usual enumeration rules for household panel surveys specify that data are collected for associated persons only while they continue to live with original sample members, that is, they are not followed if they cease to live with such persons. Unless the time period is short enough that the number of associated persons who cease to live with sample persons in that period is negligible, this enumeration rule makes it problematic to include associated persons in longitudinal analyses. This point is discussed further below.

## 3. CROSS-SECTIONAL ESTIMATES FOR HOUSEHOLDS

This section considers weighting schemes that may be used to produce cross-sectional estimates for households for any wave of a household panel survey after the first. At the first wave a sample of households is selected and all the individuals in the sampled households become panel members to be followed throughout the life of the panel or until they leave the survey population. At a subsequent wave, wave $t$, the household sample comprises all the households in which panel members reside. Households that consist of new entrants only are not represented in the sample at later waves. Such households are assumed to be a negligible proportion of all households and are ignored here. Initially, complications of nonresponse are also ignored.

Consider the estimation of the total $Y$ for all $H$ households in the population at time $t$:

$$Y = \sum_{i=1}^{H} Y_i . \tag{3.1}$$

A general estimator for this total can be expressed as

$$\hat{Y} = \sum_{i=1}^{H} w_i Y_i$$

where $w_i$ is a random variable that takes the value $w_i = 0$ if household $i$ is not in the sample. The expectation of $\hat{Y}$ is

$$E(\hat{Y}) = \sum_{i=1}^{H} E(w_i) Y_i . \tag{3.2}$$

By comparing equations (3.1) and (3.2), it can be seen that $\hat{Y}$ is unbiased for $Y$ for any weighting scheme for which $E(w_i) = 1$.

There are many ways to satisfy the condition $E(w_i) = 1$. Three will be treated here. First, consider a standard *inverse selection probability weighting scheme*. As Figure 1 illustrates, some households at time $t$ have several paths leading to their selection from the households at time 1. The probability of a household being in the sample at time $t$ is the probability of one or more of the households at time 1 from which it has drawn members being selected for the original sample. The probability of household $H_i$ being in the sample at time $t$ is then

$$\begin{aligned} P(H_i) &= P(h_j \cup h_k \cup h_\ell \cup ...) \\ &= \Sigma P(h_j) - \Sigma\Sigma P(h_j \cap h_k) + \Sigma\Sigma\Sigma P(h_j \cap h_k \cap h_\ell) ... \end{aligned} \tag{3.3}$$

where $P(h_j \cup h_k \cup h_\ell \cup ...)$ is the selection probability of the union of original households $h_j$, $h_k$, $h_\ell$, etc., for the original sample, $P(h_j)$ is the selection probability of original household $h_j$ for the original sample, $P(h_j \cap h_k)$ is the joint selection probability of original households $h_j$ and $h_k$ for the original sample, etc., and where households $h_j$, $h_k$, $h_\ell$, etc., each contain at least one member who is currently in household $H_i$. The weight for each sampled household is then $w_i = 1/P(H_i)$. With this weighting scheme,

$$E(w_i) = P(H_i)[1/P(H_i)] + [1-P(H_i)]0 = 1,$$

satisfying the condition for an unbiased estimator of a population total.

In practice, the computation of $P(H_i)$ will generally not be as complex as equation (4) might suggest because the number of original households represented in household $H_i$ is usually small. With, say, two original households involved, $P(H_i)$ reduces to

$$P(H_i) = P(h_1 \cup h_2) = P(h_1) + P(h_2) - P(h_1 \cap h_2). \tag{3.4}$$

A problem with the application of the inverse selection probability approach is that $P(h_j)$ may be known only for households selected for the original sample, and not for other households. Also the joint probability may not be known. Even when the original sample was an equal probability one, so that all the $P(h_j)$ are the same, the joint probability may depend on the sample design (for instance, whether the two households were in the same segment or not). The difficulty of obtaining $P(H_i)$ is a major drawback with this form of inverse selection probability approach.

An alternative strategy for developing the weights for time $t$ is to base them only on the selection probabilities of households selected for the original sample, thus avoiding the difficulty in obtaining $P(H_i)$ noted above. One approach is to identify the set of households $h_j$ at time 1 that would result in household $H_i$ being in the sample at time $t$, and compute the weight for household $H_i$ as

$$w_i = \sum_j \alpha_{ij} w'_{ij} \tag{3.5}$$

where $w'_{ij} = 1/p_j$ if household $h_j$, which has at least one member in household $H_i$, was selected for the original sample and $w'_{ij} = 0$ if not, where $p_j = P(h_j)$, and where $\alpha_{ij}$ are any set of constants satisfying $\sum_j \alpha_{ij} = 1$.

With this approach,

$$E(w'_{ij}) = p_j(1/p_j) + (1-p_j)0 = 1,$$

and hence

$$E(w_i) = \sum_j \alpha_{ij} = 1.$$

Thus, the use of weights $w_i$ will yield unbiased estimators of totals for the household population for any choice of constants $\alpha_{ij}$, provided that $\Sigma\alpha_{ij} = 1$. As indicated above, the principal advantage of this type of scheme is that it requires information only on the initial selection probabilities of the original households that were sampled at time 1, which are known. It does not require information on the initial selection probabilities of the other original households that have members in the current household, which are often not known.

A natural choice of $\alpha_{ij}$ is to make them equal for all the original households that lead to the selection of household $H_i$ at time $t$. Huang (1984) terms this scheme a multiplicity approach. Here the scheme will be called an *equal household weighting scheme*. With the equal household weighting scheme

$$w_i = \Sigma w'_{ij}/C_i, \tag{3.6}$$

where $C_i$ is the number of original households represented in household $i$ at time $t$.

An alternative version of the above approach is one based on original sample persons rather than households. In this case, let $I_{ijk}$ denote individual $k$ from original household $j$ in household $i$. Then

$$w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk}$$

where $w'_{ijk} = 1/p_j$ if individual $k$ in household $h_j$ was in the original sample and $w'_{ijk} = 0$ if not, and where the $\alpha_{ijk}$ are any set of constants satisfying $\sum_j \sum_k \alpha_{ijk} = 1$. Since the probability of an individual being selected for the original sample is the same as that of that individual's household,

$$E(w'_{ijk}) = p_j(1/p_j) + (1-p_j)0 = 1.$$

In this case, the natural choice of the constants $\alpha_{ijk}$ is to make them equal for all members of the current household who were eligible for selection for the original sample. This produces what has been termed the fair share weighting scheme (Huang, 1984; Ernst, 1989). This scheme is termed here an *equal person weighting scheme*. With this scheme

$$w_i = \frac{1}{M_i}\Sigma M_{ij} w_{ij}'$$

where $w_{ij}' = w_{ijk}'$ is constant for all individuals in household $H_i$ emanating from the same sampled household at time 1, $M_{ij}$ is the number of individuals in household $H_i$ coming from household $h_j$, and $M_i = \sum_j M_{ij}$ is the number of individuals in household $H_i$ who were eligible for the sample at time 1. The equal person weighting scheme is applied in the SIPP and is proposed for use in the SLID.

Although developed here in terms of persons rather than households, it is readily apparent that the equal person weighting scheme could equally have been generated in terms of households. As shown above, the household weight $w_i = \sum_j \alpha_{ij} w_{ij}'$ satisfies the condition $E(w_i) = 1$ for any set of constants $\alpha_{ij}$ such that $\sum_j \alpha_{ij} = 1$. The equal household weighting scheme chooses $\alpha_{ij} = 1/C_i$, with $\Sigma\alpha_{ij} = 1$. The choice $\alpha_{ij} = M_{ij}/M_i$, with $\sum_j \alpha_{ij} = 1$, leads to the equal person weighting scheme.

It is instructive to compare the inverse selection probability weighting scheme with the equal household and equal person weighting schemes in a simple case. Following Little (1989), consider household $H_i$ selected at time $t$ with household members coming from two original households. Let $p_1$ and $p_2$ denote the selection probabilities for the original households, and let $p_{12}$ denote their joint selection probability. Under the inverse selection probability approach, the household weight is

$$w_i^* = \frac{1}{p_1 + p_2 - p_{12}},$$

as indicated above.

Under the equal person weighting scheme the weight for household $H_i$ depends on which household or households were selected for the original sample:

$w_i = P_1/p_1$ if only household $h_1$ was selected;

$w_i = P_2/p_2$ if only household $h_2$ was selected;

$w_i = (P_1/p_1) + (P_2/p_2)$ if both $h_1$ and $h_2$ were selected;

where $P_1$ and $P_2$ are the proportions of members of household $H_i$ who came from households $h_1$ and $h_2$, respectively (excluding any new entrants to the population). The probability of only household $h_1$ being selected is $(p_1 - p_{12})$, of only household $h_2$ being selected is $(p_2 - p_{12})$, and of both households being selected is $p_{12}$. The expected value of the weight conditional on household $H_i$ being in the sample is thus

$$E(w_i|H_i \text{ in sample}) = \frac{(p_1-p_{12})(P_1/p_1)+(p_2-p_{12})(P_2/p_2)+p_{12}[(P_1/p_1)+(P_2/p_2)]}{p_1+p_2-p_{12}}$$

i.e.,

$$E(w_i|H_i \text{ in sample}) = \frac{1}{p_1+p_2-p_{12}} = w_i^*.$$

As this result demonstrates, the weight for household $H_i$ varies depending on which original households were selected, but in expectation the weight is the same as that obtained from the inverse selection probability approach.

Results for the expectation of the weight of household $H_i$ under the equal household weighting scheme can be readily obtained as a special case of the above derivation

in which $P_1 = P_2 = ½$. In expectation, the weight is the same as that for the inverse selection probability approach.

Given that the weight $w_i = \Sigma\alpha_{ij}w_{ij}'$ satisfies the condition $E(w_i) = 1$ for any set of $\alpha_{ij}$ such that $\Sigma\alpha_{ij} = 1$, the question arises as to the optimal choice of the $\alpha_{ij}$. As illustrated above, the weight for a household in the sample at time $t$ ($H_i$) depends on which of the households at time 1 ($h_j$) leading to the inclusion of $H_i$ at time $t$ were selected. In expectation, the weight is equal to the standard weight from the inverse selection probability approach, $w_i^*$. It varies around $w_i^*$ according to which of the $h_j$ were selected. One approach is to choose the $\alpha_{ij}$ to minimize the variance of the estimated total $\hat{Y}$.

The variance of $\hat{Y}$ may be expressed as

$$V(\hat{Y}) = VE(\hat{Y}|s) + EV(\hat{Y}|s) \tag{3.7}$$

where $s$ denotes the set of households in the sample at time $t$. Now

$$E(\hat{Y}|s) = E\left(\sum_{i=1}^{H} w_i Y_i | s\right)$$

$$= \sum^{s} E(w_i|H_i)Y_i = \sum w_i^* Y_i = \hat{Y}^*,$$

where $\hat{Y}^*$ is the standard inverse selection probability estimator. Thus

$$VE(\hat{Y}|s) = V(\hat{Y}^*).$$

The first term in equation (3.7) is thus the variance of the standard inverse selection probability estimator, and the second term is the additional variance resulting from the use of weighting schemes from the class (3.5), $w_i = \Sigma\alpha_{ij}w_{ij}'$. The $\alpha_{ij}$ may then be chosen to minimize $EV(\hat{Y}|s)$.

Consider

$$V(\hat{Y}|s) = V(\sum^{s} w_i Y_i|s)$$

$$= \sum^{s} Y_i^2 V(w_i|H_i) + \sum_{i\neq i'}\sum Y_i Y_{i'} Cov(w_i, w_{i'}|H_i, H_{i'}).$$

Assuming $Cov(w_i, w_{i'}|H_i, H_{i'}) = 0$,

$$V(\hat{Y}|s) = \sum Y_i^2 V(w_i|H_i)$$

$$= \sum Y_i^2 \left[E(w_i^2|H_i) - w_i^{*2}\right]$$

since, as noted above, $E(w_i|H_i) = w_i^*$. Thus, assuming $Cov(w_i, w_{i'}|H_i, H_{i'}) = 0$, $V(\hat{Y}|s)$ is minimized when $E(w_i^2|H_i)$ is minimized.

Consider the application of this approach to the simple case discussed above in which $H_i$ is composed of members from two original households and let $w_i = \alpha_i w_{i1}' + (1-\alpha_i) w_{i2}'$. Then

$$E(w_i^2|H_i) = \frac{(p_1 - p_{12})\dfrac{\alpha_i^2}{p_1^2} + (p_2 - p_{12})\dfrac{(1-\alpha_i)^2}{p_2^2} + p_{12}\left(\dfrac{\alpha_i}{p_1} + \dfrac{1-\alpha_i}{p_2}\right)^2}{p_1 + p_2 - p_{12}}.$$

Minimizing $E(w_i^2|H_i)$ is equivalent to minimizing

$$\Delta = (p_1 - p_{12}) p_2^2 \alpha_i^2 + (p_2 - p_{12}) p_1^2 (1-\alpha_i)^2 + p_{12}\left[(p_2 - p_1)\alpha_i + p_1\right]^2.$$

Then

$$\frac{\partial \Delta}{\partial \alpha_i} = 2(p_1 - p_{12})p_2^2\alpha_i - 2(p_2 - p_{12})p_1^2(1-\alpha_i) + 2p_{12}(p_2 - p_1)[(p_2 - p_1)\alpha_i + p_1].$$

Solving $\partial\Delta/\partial\alpha_i = 0$ for $\alpha_i$ gives the optimum $\alpha_i$ as

$$\alpha_{oi} = \left(1 + \frac{p_2 - p_{12}}{p_1 - p_{12}}\right)^{-1}. \tag{3.8}$$

If the original households are selected independently, i.e. $p_{12} = p_1 p_2$,

$$\alpha_{oi} = \left[1 + \frac{p_2(1-p_1)}{p_1(1-p_2)}\right]^{-1} = \left[1 + \frac{\psi_2}{\psi_1}\right]^{-1}, \tag{3.9}$$

where $\psi_j = p_j/(1-p_j)$ is the odds of original household $h_j$ being selected.

Irrespective of whether the households are sampled independently, in the special case of an equal probability (epsem) sample of households initially, with $p_1 = p_2$,

$$\alpha_{oi} = \tfrac{1}{2}.$$

Thus, in the two-household case, the equal household weighting scheme minimizes the variance of the household weights around the inverse selection probability weight when the initial sample is an epsem one.

The optimal choice of $\alpha_{oi}$ given by (3.8) requires knowledge of $p_1$, $p_2$ and $p_{12}$, and that given by (3.9) requires independence and knowledge of $p_1$ and $p_2$. If these probabilities were known, then the standard inverse selection probability weight could be employed and would be preferable. In the case of an approximately epsem sample, the equal

household weighting scheme should be close to the optimal, at least for the two-household case. This would apply, for instance, in the case of an epsem initial sample, with perhaps a few departures from epsem, and with nonresponse adjustments that are not too variable. With the equal household weighting scheme, when only one of the $C_i$ original households, $h_j$, in $H_i$ was selected for the original sample (as will generally be the case), then the weight for $H_i$ is simply $1/C_i p_j$.

In the case of a non-epsem initial sample, the choice of the $\alpha_{ij}$ would ideally depend on the original household selection probabilities. However, since these probabilities are unknown, that approach cannot be applied. By default, the equal household or equal person weighting schemes may therefore be employed in this case. It should be noted that the use of these schemes (or any scheme with constant $\alpha_{ij}$'s satisfying $\sum_j \alpha_{ij} = 1$) with a non-epsem initial sample still results in an unbiased estimate $\hat{Y}$. The drawback to these schemes in such a case is only that the $\alpha_{ij}$ are suboptimal in terms of minimizing the variance of $\hat{Y}$.

It should be noted that the equal household weighting scheme requires information on the number of original households $h_j$ contributing members to household $H_i$ at time $t$. That number may be difficult to determine in some cases. Consider, for example, a household at time $t$ that contains two associated persons. It may sometimes be difficult to determine whether these two persons were in a single household or in two separate households at the time of the initial sample selection. The equal person weighting scheme has the attractive feature of avoiding the need for Wave 1 household information. This feature provides an important reason for preferring the equal person to the equal household weighting scheme.

The discussion thus far has assumed that there is no nonresponse. In practice, however, there is household nonresponse at the initial wave and further nonresponse at wave $t$ among households responding at the initial wave. Weighting adjustments for household nonresponse at the initial wave can be readily handled by treating response as another stage of selection, assumed at random within weighting classes. Then the quantity $p_j$ in the above discussion is redefined to be the product of the selection probability of original household $h_j$ for the first wave and the weighted response rate for the weighting class in which $h_j$ falls. The initial weight $w_{ij}'$ is then the inverse of this redefined $p_j$. With this change, the results given above generalize to the situation where weighting adjustments are applied for household nonresponse at the initial wave.

The same approach can also be extended to cover weighting adjustments for households responding at the initial wave that lead to no responding households at wave $t$. In this case, the responding households at the initial wave can be divided into weighting classes based on responses given at the initial wave, and the weights of households leading to one or more responding households at wave $t$ can be further adjusted to compensate for those leading to no responding households at wave $t$. The revised $w_{ij}^{/}$ can then be employed in equation (3.5) and subsequently.

Both the above nonresponse adjustments are applied in relation to the original households. A further type of household nonresponse cannot be handled in this way. This type of nonresponse involves the situation where an original household splits into two or more separate households at wave $t$, and where some but not all of those households respond at wave $t$. In this case the adjustment for the nonresponding households needs to be made in relation to the wave $t$ households, $H_i$, rather than the original households, $h_j$. If the number of original households having members in each wave $t$ nonresponding household of this type were known, the weights $w_i$ for these households could be computed using the approach described above. Then weighting adjustments could be readily applied within weighting classes of the wave $t$ households to compensate for the nonresponding households. In practice, however, the number of original households having members in a nonresponding household at wave $t$ may often be unknown. One approach for handling this situation is to estimate this number by the average number for responding households at wave $t$ that have similar characteristics to (e.g., they are also splits from original households), and are in the same weighting class as, the nonresponding household. Using such estimated numbers where necessary, the weights $w_i$ can be determined for all nonresponding households of the type being discussed. Standard weighting adjustments can then be applied to the responding households at wave $t$ to compensate for these nonresponding households.

# 4. CROSS-SECTIONAL ESTIMATES FOR INDIVIDUALS

This section considers weighting schemes that may be used to produce cross-sectional estimates for individuals for any wave of a household panel survey after the first. At a subsequent wave, wave $t$, the survey population has changed: some members of the original population will leave and some new entrants will join the survey population in the period from wave 1 to wave $t$. Leavers result in a reduction of the panel sample size from wave 1 to wave $t$, but do not cause any bias in the survey estimates for wave $t$. New entrants who join households that contain one or more members of the original population can be represented in cross-sectional estimates for later waves of a household panel survey, but new entrants living in households that do not contain any members of the original population are not covered (unless a special sample of them can be taken). Given this situation, a decision needs to be made as to whether to restrict the population of inference for cross-sectional analyses at wave $t$ to individuals who were in the survey population at time 1 and who are in the population at time $t$, thus excluding all new entrants, or whether to define the population of inference at wave $t$ as all those present at wave $t$, accepting the noncoverage that will occur with respect to new entrants who live in households containing no members of the original population. The latter approach is mainly adopted here.

Let there be $N$ individuals in the population at time $t$, with $N_i$ individuals in household $H_i$ ($i = 1, 2, \ldots H$) and $\Sigma N_i = N$ (new entrants living in households containing no members of the original population are not included here, and are ignored throughout the rest of this section). The members of household $H_i$ come from households $h_j$, $h_k$, $h_\ell$, etc., at time 1. Let $M_{ij}$ denote the number of members of household $H_i$ at time $t$ who were in household $h_j$ at the start of the panel. The sum $M = \Sigma\Sigma M_{ij}$ is less than the population size at time 1 because of leavers from the population in the period from time 1 to time $t$, and $M < N$ because of new entrants to the population who are in households containing members from the original population.

Consider now the estimation of a total for the population of individuals at time $t$:

$$Y = \sum_{i=1}^{H} \sum_{k=1}^{N_i} Y_{ik}. \qquad (4.1)$$

where $Y_{ik}$ is the value for individual $k$ in household $H_i$. As in the household case discussed in the previous section, a general estimator for this total can be expressed as

$$\hat{Y} = \sum_{i=1}^{H} \sum_{k=1}^{N_i} w_{ik} Y_{ik} \tag{4.2}$$

where $w_{ik}$ is a random variable that takes the value $w_{ik} = 0$ if individual $k$ in household $H_i$ is not in the sample. The estimator $\hat{Y}$ is unbiased for $Y$ provided that $E(w_{ik}) = 1$.

There are many ways to satisfy the condition $E(w_{ik}) = 1$. It is instructive to consider three of them. First, let $w_{ik} = 0$ for all individuals not in the original sample. In this case, the estimator $\hat{Y}$ discards the associated persons who are included in the sample because they are living with original sample persons at time $t$. Let $p_{ik}$ denote the probability of a member of the original population, individual $k$ residing in household $H_i$ at time $t$, being selected for the initial sample, and let $w_{ik} = 1/p_{ik}$. Then, for such an individual

$$E(w_{ik}) = p_{ik}(1/p_{ik}) + (1-p_{ik})0 = 1.$$

With this scheme, all new entrants to the population have $w_{ik} = 0$ with certainty. Thus $\hat{Y}$ in (4.2) provides an unbiased estimator of the total for the original population that is still present at time $t$, but does not include a component for the new entrants.

Modifications to the above procedure can be made to cover certain types of new entrants. For instance, births to sampled mothers can be included by assigning them the weight of their mothers, or if, as in the SIPP, the survey population is taken to be adults aged 16 and over, those under 16 at the start of the panel can be treated as sampled persons with assigned probabilities, and they can be included in the analyses of later waves after they have attained the age of 16. Such modifications do not, however, handle all types of new entrants. Provided that the proportion of other types of new entrants is small, this deficiency may not be a serious concern.

The weighting scheme that restricts the analysis to original sample persons, plus certain specified new entrants, is employed with the PSID. Its limitation is that it fails to make direct use of data collected for associated persons. Such data may be used to provide information on the situation of sample persons, but the associated persons are excluded from the sample for the analysis.

In order to include associated persons in cross-sectional analyses for time $t$, they need to be assigned positive weights. Following this approach, the only individuals given weights of zero for cross-sectional analysis at wave $t$ are those for whom no data are collected at that time. Noting that the probability of an individual being selected for the sample is the same as that of his or her household, weighting schemes for cross-sectional analyses of individuals at wave $t$ can be obtained directly from those for households given in Section 3. Here we will develop the general strategy of producing weights for cross-sectional analysis at time $t$ based only on the selection probabilities of members of the original sample, thus avoiding the problems with the inverse selection probability approach noted in Section 3.

Let $I_{ijk}$ denote individual $k$ from original household $h_j$ who is now in household $H_i$. Let $w_i$ denote the weight for every member of household $H_i$ for cross-sectional analyses at time $t$, and let

$$w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk}$$

where $w'_{ijk} = 1/p_j$ if household $h_j$ was in the original sample and $w'_{ijk} = 0$ if not. Then, as before, $E(w'_{ijk}) = 1$ for members of the original population. New entrants, for whom $p_j = 0$, may be handled by setting $\alpha_{ijk} = 0$. Then

$$E(w_i) = \sum_j^{N_i} \sum_k \alpha_{ijk} E(w'_{ijk}) = \sum_j^{M_i} \sum_k \alpha_{ijk} = 1$$

provided that $\Sigma\Sigma\alpha_{ijk} = 1$. Under this condition $\hat{Y}$ is unbiased for $Y$.

A natural choice of $\alpha_{ijk}$ is to set $\alpha_{ijk} = 1/M_i$ for all members of the original population. This is the equal person weighting scheme in which every member of household $H_i$ at time $t$ (including new entrants) receives the weight

$$w_i = \sum_j \sum_k w'_{ijk}/M_i.$$

Another choice of the $\alpha_{ijk}$ is that used for the equal household weighting scheme. Let $C_i$ denote the number of original households that have members in household $H_i$ at time $t$. Then $\sum_j \sum_k \alpha_{ijk} = 1$ can be divided equally between households, with each member of original household $h_j$ being assigned a value of $\alpha_{ijk} = 1/C_i M_{ij}$. Then for original household $h_j$

$$\sum_k^{M_{ij}} \alpha_{ijk} = 1/C_i.$$

The derivation of the $\alpha_{ijk}$ to minimize the variance of the estimated total $\hat{Y}$ for the population of individuals follows directly from the corresponding derivation for the population of households given in Section 3. The estimated total for the population of individuals is

$$\hat{Y} = \sum_i^s \sum_k^{N_i} w_{ik} Y_{ik} = \sum_i^s \sum_k^{N_i} w_i Y_{ik}$$

since the weights for every individual in sampled household $H_i$ are the same. This estimated total can be expressed as

$$\hat{Y} = \sum_i^s w_i Y_i,$$

where $Y_i = \Sigma Y_{ik}$ is the household total for $H_i$. Thus $\hat{Y}$ can be expressed as a household total, and the results of Section 3 can be applied directly.

Consider the example from Section 3 in which $H_i$ is composed of members from only two original households, perhaps together with one or more new entrants. In this case the person-level weight $w_i = \Sigma\Sigma\alpha_{ijk}w'_{ijk}$ reduces to

$$w_i = \left(\sum_k \alpha_{i1k}\right)w'_{i1} + \left(\sum_k \alpha_{i2k}\right)w'_{i2}$$
$$= \alpha_i w'_{i1} + (1-\alpha_i)w'_{i2}$$

where $\alpha_i = \Sigma\alpha_{i1k}$. As shown in equation (3.8), the optimum value of $\alpha_i$ is

$$\alpha_{oi} = \left(1 + \frac{p_2 - p_{12}}{p_1 - p_{12}}\right)^{-1}.$$

The individual values $\alpha_{ijk}$ are not needed for computing the $w_i$; only the original household totals $\sum_k \alpha_{ijk}$ are required. If individual values are needed for the $\alpha_{ijk}$, they may be simply assigned as $\sum_k \alpha_{ijk}/M_{ij}$.

As in the household case, the optimum weighting $\alpha_{oi}$ requires knowledge of $p_1$, $p_2$ and $p_{12}$. If these probabilities are known, the standard inverse selection probability weight $w_i^*$ can be computed, and would be preferred. In the case of an approximately epsem sample, the equal household weighting scheme should fare well.

In cross-sectional analyses of individuals in a household panel survey, weighting adjustments for nonresponse and noncoverage can be considered at either the household or the person level. For the weighting scheme that restricts cross-sectional analyses to original sample persons, person- and household-level weighting adjustments may be readily applied. At the first wave, the weight based on each sample respondent's original selection probability can be modified to compensate for nonresponse at that wave, employing adjustment cells that take

account of person-level and household-level characteristics. Nonresponse at subsequent waves can also readily be based on person- and household-level characteristics. It should be noted that person-level nonresponse adjustments lead to a variability of the weights for the original sample persons within households, even when no change in household composition takes place.

Household-level weighting adjustments can be applied straightforwardly in cross-sectional analyses of individuals that include associated persons. All that is needed is to adjust the original weight $w'_{ijk} = w'_{ij}$ to compensate for household nonresponse at the first wave, and then to make a further adjustment at wave $t$, as described in Section 3.

Person-level weighting adjustments for cross-sectional analyses of individuals including associated persons introduce a further complexity. If person-level weighting adjustments are made at the first wave, the values of the nonresponse adjusted $w'_{ijk}$ will be different for different members of the original household $h_j$. As before, let the weight for household $H_i$ at wave $t$ be

$$w_i = \sum_j \sum_k \alpha_{ijk} w'_{ijk}$$

where $w'_{ijk}$ is the nonresponse-adjusted weight for sample person $k$ if household $h_j$ was in the original sample and $w'_{ijk} = 0$ if household $h_j$ was not sampled. Then, treating nonresponse as a further stage of selection within weighting classes, $E(w'_{ijk}) = 1$, as before. Thus $w_i$ satisfies the condition that $E(w_i) = 1$ provided that $\Sigma\Sigma\alpha_{ijk} = 1$.

With unequal values of $w'_{ijk}$ for different members of the same original household, the question arises as to how the $\alpha_{ijk}$ should be chosen. With the equal person and equal household weighting schemes, $\alpha_{ijk} = \alpha_{ij}$, a constant for all members of the original household. These schemes thus in effect use the average $w'_{ijk}$ instead of the constant $w'_{ijk}$ that applied before. A range of other choices of $\alpha_{ijk}$ is also possible. For example, one possibility would be to set $\alpha_{ijk} = 1/C_i$ for the member of original household $h_j$ with the smallest weight and $\alpha_{ijk} = 0$ for all the other members of that household. In the SIPP, the complexities of person-level nonresponse adjustments are avoided by restricting the adjustments to household-level characteristics.

As has been noted, the basic person-level weighting schemes described above assign the same weight to every member of household $H_i$ at time $t$, and that weight is the same as the household weight. Moreover, this property is retained when nonresponse adjustments are made only at the household level. Thus, in this case, there is no need to record a separate household weight, since it can be readily obtained by taking the weight of any one of the household members.

This equivalence of the household weight and the weight of each household member is lost, however, if person-level nonresponse adjustments are applied. It is also lost when only household-level nonresponse adjustments are used but when poststratification adjustments are made at the person level, as for example is commonly done to make the weighted sample distribution by age, sex and race conform to the known population distribution for these variables. These adjustments cause household members to have different weights. An approach for determining a household weight in this situation is to define one household member as the "principal" person in the household, and use that person's weight as the household weight. This is the approach to household weighting adopted in the SIPP.

By taking the household weight as that of the principal person the need to compute a separate household weight is avoided. Although this approach should be generally satisfactory for a survey where the main analytic focus is on person-level statistics, it has its disadvantages. First, there are no poststratification controls to ensure that the weighted sample number of households conforms to the known number of households in the population or that the weighted sample household distributions for key variables conform to known household population distributions for those variables. Secondly, the use of principal person weights can potentially lead to some distortions in household distributions. Consider, for instance, the effect of a poststratification adjustment at the person level that assigns high adjustments to young black males. If a young black male is the defined principal person in a responding household, that household will receive the high adjustment, and this may not be appropriate. Thus, if household-level statistics are an important product of the survey, the development of separate household weights seems advisable.

# 5. LONGITUDINAL ANALYSES OF INDIVIDUALS

A key analytic advantage of a panel survey is the ability to conduct longitudinal analyses relating variables for the same sampled units measured at different time points. Since all persons in original sampled households are followed throughout the life of the panel or until they leave the survey population, the data they provide may be readily analyzed longitudinally for any time period within the panel's time span (although nonresponse adjustments may be needed for panel attrition). Thus, for example, in a ten-year panel, data for original sampled persons may be analyzed from year 1 to year 10, from year 5 to year 9, or for any other period. New entrants (e.g., births) may be included in the analysis for periods beginning after the start of the panel provided that they are treated as panel members who are followed throughout the panel even when they leave the households of original sampled persons.

Given the weighting schemes described in the previous section, associated persons can be included in cross-sectional analyses of later waves. These weighting schemes provide a cross-sectional representation of the population at any wave of the panel (apart from new entrants). It is then possible to consider all the sample of original sample members and associated persons at time $t$ as the initial sample of a new panel that may be used for longitudinal analyses from time $t$ to $(t+k)$. This procedure is, for instance, used in the SIPP, where all original sample members and associated persons present at the start of the second year of the panel are included in analyses relating to that year.

The limitation to the inclusion of associated persons in longitudinal analysis is that the following rules used in most household panel surveys specify that associated persons are dropped from the panel if they cease living with original sample persons. Thus, associated persons who live with original sample members at the start of the analysis period but who cease to live with them before the end of the analysis period effectively become nonrespondents. If the analysis period is relatively short, the number of such nonrespondents may be small and the risk of serious nonresponse bias may be negligible. If the analysis period is a long one, however, the number of not-followed associated persons may be appreciable, causing concerns about potential bias. The issue here is one of a trade-off between the reduced variance due to the increase in sample size from including associated persons in the analysis versus the increased

bias resulting from the additional nonresponse caused by failing to follow associated persons leaving the households of original sample persons.

It may be noted that it is possible to avoid the additional nonresponse bias by changing the following rules to specify that associated persons are to be followed from the time they join the panel for the rest of the life of the panel, or until they leave the survey population, irrespective of whether they continue to live with original sample members. This change, however, leads to an expanding panel, and the need for additional resources. Not only do data need to be collected for associated persons at waves after they cease to live with sample persons, but data also need to be collected for any persons with whom the associated persons live at later waves.

# 6. SUMMARY AND CONCLUDING REMARKS

This paper has described weighting schemes that enable all households for which, and all individuals for whom, data are collected in the later waves of a household panel survey to be included in cross-sectional analyses of those waves. These weighting schemes can accommodate new entrants to the population who move in to live with members of the original population, but not other new entrants.

The usual inverse selection probability weighting scheme requires information on the household selection probabilities of all members of the households sampled at a later wave, as well as the joint selection probabilities of the original households that contribute members to the later wave households. The inverse selection probability weighting scheme can often not be applied because these probabilities are unknown. To deal with this problem, an alternative approach that requires information on only the selection probabilities of sampled original households is described.

This alternative approach produces a class of weighting schemes including the equal person (fair share) scheme used in SIPP and the equal household weighting scheme. All the schemes in this class produce weights that are in expectation equal to those produced by the usual inverse selection probability scheme. The variance in the weights around the inverse selection probability weights gives rise to an increase in the variance of the survey estimates. When the original households are selected with approximately equal probability, the equal household weighting scheme is near optimal for both household and individual level analyses to control this increase in variance.

The alternative class of weighting schemes produces unbiased estimates of population totals for any choice of constant $\alpha_{ij}$ that satisfies the condition $\sum_j \alpha_{ij} = 1$ and for any initial sample design. The equal household and equal person weighting schemes are, however, suboptimal for non-epsem initial samples. One of them may nevertheless be the appropriate scheme for such designs, because the optimal choice of the $\alpha_{ij}$ depends on the unknown initial selection probabilities, and hence cannot be determined. The equal household and equal person weighting schemes have different data requirements, in that the former requires information on

Wave 1 household composition whereas the latter does not. The fact that this information may not always be readily obtainable thus argues in favor of the equal person weighting scheme.

The cross-sectional individual weights for a particular wave can be used as the starting weights for a longitudinal analysis that begins at that wave. This procedure includes associated persons present at that wave in the longitudinal analysis. However, if associated persons are not followed when they cease to live with sampled persons, those who leave sample persons before the end of the period of the longitudinal analysis become nonrespondents. Before associated persons are included in a longitudinal analysis, a check should therefore be made to ensure that their inclusion will not give rise to risks of serious nonresponse bias.

The class of weighting schemes described has a broader range of application than that indicated here. It can in fact be usefully applied in any situation where an inverse selection probability weighting scheme would be appropriate, but where not all the inclusion probabilities and joint inclusion probabilities are known. Consider, for instance, the modified version of the Mitofsky-Waksberg random digit dialing sampling procedure for telephone surveys described by Brick and Waksberg (1991). A sample of telephone numbers (primes) is selected at the first stage of this two-stage sample design. If a prime number is found to be a working residential number, that household is selected and a fixed number of additional telephone numbers in the same 100-bank is selected. The households found at these numbers are then all included in the sample. If a prime number is not a working number, the sampling process stops. With this procedure, the probability of a working residential number being selected depends on the number of working residential numbers in its 100-bank, and hence differs across 100-banks. This probability can be estimated from the sample of telephone numbers in the 100-bank. A complication arises, however, when a sampled household has two or more telephone numbers. In this case, the selection probability of the sampled telephone number can be estimated, but those of the nonsampled numbers cannot. Thus, the standard inverse selection probability weighting scheme cannot be used. However, the alternative weighting scheme described here can be employed.

## 7. REFERENCES


Brick, J.M. and Waksberg, J. (1991). Avoiding sequential sampling with random digit dialing. *Survey Methodology*, **17**, 27-41.

Citro, C.F. and Kalton, G. (1993). *The Future of the Survey of Income and Program Participation*. Washington D.C.: National Academy Press.

Duncan, G.J. and Hill, M.S. (1985). Conceptions of Longitudinal Households: Fertile or Futile? *Journal of Economic and Social Measurement*, **13**, 361-375.

Ernst, L.R. (1989). Weighting Issues for Longitudinal Household and Family Estimates. In *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh, New York: John Wiley, pp. 139-159.

Gailly, B. and Lavallée, P. (1993). *Inserer des Nouveaux Membres dans un Panel Longitudinal de Menages et D'Individus: Simulations*. Walferdange, Luxembourg: CEPS/Instead.

Hill, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.

Huang, H. (1984). Obtaining Cross-Sectional Estimates from a Longitudinal Survey: Experiences of the Income Survey Development Program. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 670-675.

Jabine, T.B., King, K.E. and Petroni, R.J. (1990). *Survey of Income and Program Participation: Quality Profile*. Bureau of the Census, Washington D.C.: U.S. Department of Commerce.

Judkins, D., Hubble, D., Dorsch, J., McMillen, D., and Ernst, L. (1984). Weighting of Persons for SIPP Longitudinal Tabulations. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 676-687.

Kasprzyk, D. (1988). *The Survey of Income and Program Participation: An Overview and Discussion of Research Issues*. SIPP Working Paper No. 8830. Washington D.C.: U.S. Bureau of the Census.

Kasprzyk, D. and McMillen, D.B. (1987). SIPP: Characteristics of the 1984 Panel. *Proceedings of the Social Statistics Section, American Statistical Association*, 181-186.

Lavallée, P. and Hunter, L. (1992). Weighting for the Survey of Labour and Income Dynamics. *Proceedings of Statistics Canada Symposium 92: Design and Analysis of Longitudinal Surveys*, 65-75.

Lavallée, P., Michaud, S. and Webber, M. (1993). The Survey of Labour and Income Dynamics, Design Issues for a New Longitudinal Survey in Canada. *Bulletin of the International Statistical Institute*, 49th Session, Contributed Papers, Book 2, 99-100.

Little, R.J.A. (1989). Sampling Weights in the PSID: Issues and Comments. Panel Study of Income Dynamics Working Paper, Ann Arbor: University of Michigan.

Nelson, D., McMillen, D. and Kasprzyk, D. (1985). *An Overview of the SIPP, Update 1.* SIPP Working Paper No. 8401. Washington D.C.: U.S. Bureau of the Census.