Journal of Official Statistics
Vol. 3. No. 2, 1987, pp. 141–153

Linear Weighting of Sample Survey Data

Jelke G. Bethlehem and Wouter J. Keller[1]

Abstract: To improve the quality of estimates in sample surveys some kind of weighting is often carried out. Post-stratification is a popular weighting method. Two major problems can make applications of post-stratification difficult: empty strata and lack of adequate population information. This paper presents a general method for weighting, in which weights are obtained from a linear model which relates the target variables of the survey to auxiliary variables. Post-stratification is a special case of this method. Because of the generality of the method, different weighting schemes can be applied which take advantage of the available population as much as possible and at the same time avoid the mentioned problems. The theory is illustrated with an example.

Key words: Weighting; post-stratification; regression estimator.

1. Introduction

A sample survey is an instrument for making inferences about a finite population using observations on only some of the elements in the population. If sufficient population information is available, estimates of population parameters can be improved by assigning weights to the observed elements. Estimates are obtained by simple summation of the weighted observations. Bailar et al. (1978) describe weighting as a frequently used adjustment method to correct for potential bias caused by nonresponse. Platek and Gray (1980) and Lindström et al. (1979) present weighting as an important method to correct for this bias. Little (1982) discusses the use of models including weighting for correcting for nonresponse. Even in the case of complete response it may still be worthwhile to perform some kind of weighting. Post-stratification (see e.g. Holt and Smith (1979)) is a well-known and much used weighting method. It increases precision when strata are homogeneous with respect to the target variable, and reduces bias when strata are homogeneous with respect to target variables or response probabilities.

This paper presents a general framework for weighting based on estimators constructed from linear models. It will be shown that classical weighting can be derived from the linear model theory.

In the application of post-stratification two important problems may arise. The first is the problem of empty strata. If the stratification is obtained by crossing a number of

[1] Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.
Acknowledgement: The authors want to thank the unknown referee for his/her thorough reading and valuable comments.

variables, the resulting number of strata may be so large that one or more strata have no observations. In the empty strata no estimates can be computed and proper estimation of population parameters is not possible. This problem can be solved by collapsing strata. This can be a tedious and time consuming process, especially for large sample surveys. Another way to solve the problem is not to use all available variables for the construction of the strata. Since each variable may play an important role in the reduction of bias and precision, this solution is also not very satisfactory.

The second stratification problem concerns the availability of population information; applying post-stratification requires that the stratum sizes are known. Although it may be desirable to use a large number of variables for the construction of strata, stratification cannot be carried out if the population sizes of the strata are not available, even if the stratification will not result in empty strata.

With the proposed method, one can avoid the above mentioned post-stratification problems, and still use all the variables that are important for weighting purposes. The idea is best illustrated by an example. Suppose that in a sample survey five variables are available for weighting: sex (two categories), age (ten categories), marital status (four categories), region (eleven categories), and degree of urbanization (six categories). An ordinary post-stratification using these variables would result in $2\times10\times4\times11\times6 = 5\ 280$ strata. The number of strata may be reduced by collapsing strata or by leaving out one or more variables. Alternatively, one might use our method to carry out a number of post-stratifications simultaneously. It is, e.g., possible to post-stratify by sex, age and marital status (80 strata) and at the same time by region and degree of urbanization (66 strata). Due to the decreased number of strata in each post-stratification, the problem of empty strata will appear less frequently. Of course the amount of population information used is less than that used in full post-stratification. But information on all the variables is still taken into account.

The same type of solution is possible when the population sizes of all 5 280 strata are not available. If, e.g., the distribution by sex, age and marital status is available at only the regional level (and not by degree of urbanization) and the age distribution is available by region and degree of urbanization, it is possible to carry out both post-stratification by sex, age, marital status and region, and by age, region and degree of urbanization.

Our method shows some resemblance with raking ratio estimation. In raking ratio estimation the resulting weights are the product of the weight coefficients. Our method, however, produces weights which are sums of weight coefficients. Computation of raking ratio weights uses an iterative process, whereas our weights are obtained by least squares techniques which require matrix inversion only. Furthermore, the theory behind our method is so straightforward that simple approximations of the variance of estimators can be obtained.

In this paper, inference is based on probability sampling where randomization is introduced through the sampling design: the randomization or design-based approach. Hansen et al. (1983) emphasize that for large enough samples the validity of design-based inference does not depend on assumptions concerning the distribution of characteristics in the population. Still, information about the population can be used to improve the efficiency of the estimates. The decrease in the variance is determined by the extent to which the auxiliary information is associated with the target variable.

In Sections 2 and 3 we introduce the basic notation. Section 4 presents the regression

estimator. In Section 5 we apply the theory to simple random sampling. Section 6 shows that post-stratification is a special case of the theory. Section 7 offers a solution for the post-stratification problems. An application of the theory is given in Section 8. Section 9 gives some suggestions for models for weighting which are not included in the theory. More details on the theory can be found in Bethlehem and Keller (1982, 1983).

2. Population and Sample

Let the target population U consist of N identifiable elements, which may be labeled 1,2, ,..., k, ..., N. Associated with each element k is an (unknown) q-vector y_k of values of q quantitative target variables and a p-vector x_k of values of p auxiliary variables. Let $Y=(y_1, y_2, \ldots, y_N)'$ denote the $N\times q$-matrix of values of the target variables for all elements and let $X=(x_1, x_2,\ldots,x_N)'$ be the $N\times p$-matrix of values of the auxiliary variables for all elements.

We assume the objective of the sample survey to be estimation of the q-vector of population means

$$\bar{y} = Y' \iota/N, \tag{2.1}$$

where ι is the N-vector consisting of ones. The p-vector of population means for the p auxiliary variables is denoted by

$$\bar{x} = X' \iota/N. \tag{2.2}$$

We restrict ourselves to sampling without replacement. A sample from the finite population U can be denoted an $N\times N$-diagonal matrix T. The k-th diagonal element of T assumes the value 1 if the corresponding element k is in the sample, and it assumes the value 0 if this is not the case. The sample size, i.e. the sum of the diagonal elements of T, is denoted by n. The expected value of T is equal to

$$E(T) = \Pi, \tag{2.3}$$

where Π is the $N\times N$-diagonal matrix of the first-order inclusion probabilities $\pi_1, \pi_2,\ldots, \pi_N$. Observe that in this notation the Horvitz-Thompson (1952) estimator for the vector $\bar{y}$ of population means can be written as

$$\hat{\bar{y}}_{HT} = Y'\Pi^{-1}T\iota/N. \tag{2.4}$$

The variance is equal to

$$V(\hat{\bar{y}}_{HT}) = Y' \Delta Y/N^2, \tag{2.5}$$

where element Δ_{ij} of $N\times N$-matrix equal to

$$\Delta_{ij} = (\pi_{ij} - \pi_i\pi_j)/\pi_i\pi_j, \tag{2.6}$$

and π_{ij} is the second-order inclusion probability of elements i and j, with $\pi_{ii}=\pi_i$. This is, in matrix notation, the well-known expression for the variance of the Horvitz-Thompson estimator, see e.g. Raj (1968, p. 54).

3. The Regression Model

If the auxiliary variables are correlated with the target variables, an estimator that is more precise than the Horvitz-Thompson estimator can be constructed. This relationship implies that for a suitably chosen $p\times q$-matrix B of regression coefficients, the elements in the $N\times q$-matrix of residuals

$$E = Y - XB \tag{3.1}$$

vary less than the values of target variables themselves. Observe that all quantities in (3.1) are fixed numbers. There are no random variables. Applying the ordinary least squares method results in

$$B = (X'X)^{-1}X'Y. \tag{3.2}$$

An estimator for B, based on sample data, is defined as

$$\hat{B} = (X'\Pi^{-1}TX)^{-1}X'\Pi^{-1}TY. \tag{3.3}$$

The estimator $\hat{B}$ is not unbiased, but it can be shown that the bias is of the order $n^{-1/2}$, where n is the sample size. So $\hat{B}$ is approximately unbiased for large samples. We will call $\hat{B}$ an asymptotically design unbiased (ADU) estimator. ADU-estimators are discussed by Särndal (1980) and Wright (1983). For the estimator in (3.3), the matrix of cross-products is weighted by the inverse first order inclusion probabilities. Särndal (1980) discusses in great detail weighting of cross-product matrices.

4. The Regression Estimator

It is not our first objective to estimate B. What we need is an estimator for the population mean $\bar{y}$. We define the regression estimator of $\bar{y}$ by

$$\hat{\bar{y}}_R = \hat{B}'X'\iota/N = \hat{B}'\bar{x}. \tag{4.1}$$

Since $\hat{B}$ is an ADU-estimator of B, $X\hat{B}$ is an ADU-estimator of XB. But (4.1) is an ADU-estimator of $\bar{y}$ only if $B'\bar{x}=\bar{y}$, and that is the case only if there exists a p-vector c of fixed numbers such that $Xc=\iota$, see e.g. Bethlehem (1985a). The condition $Xc=\iota$ can also be found in Särndal (1980) and Isaki and Fuller (1982). Wright (1983) gives more general conditions for asymptotic design unbiasedness.

Under the condition $Xc=\iota$ the regression estimator can be written in the somewhat different form

$$\hat{\bar{y}}_R = \hat{\bar{y}}_{HT} + \hat{B}'(\bar{x} - \hat{\bar{x}}_{HT}), \tag{4.2}$$

where $\hat{\bar{x}}_{HT}$ is the p-vector of Horvitz-Thompson estimates for x. We will call estimator (4.2), or its equivalent (4.1) the regression estimator. This estimator is also given by a number of other authors, see e.g. Robinson and Särndal (1980) and Isaki and Fuller (1982). However, those authors study the distributional properties of the estimator under superpopulation models (the model-based approach), in contrast to the design-based approach in this paper. Under superpopulation assumptions, validity of inference depends on the correctness of the specified model. Särndal (1982) discusses regression estimation of linear functions under a general design-based approach.

In the case of simple random sampling and use of only one auxiliary variable, the estimator in (4.2) reduces to the simple regression estimator as, e.g., given in Cochran (1977). The estimator presented in (4.2) can be considered a generalized version of this simple regression estimator. The estimator is generalized in two ways: more than one auxiliary variable can be used and any design using sampling without replacement may be applied.

The regression estimator is not an unbiased estimator. However, it can be shown that its bias is of the order $1/n$. The variance-covariance of the estimator can be approximated by

$$V(\hat{\bar{y}}_R) \doteq E'\Delta E/N^2. \tag{4.3}$$

This result is obtained by writing the regression estimator as

$$\hat{\bar{y}}_R = \bar{y} + \hat{\bar{e}}_{HT} + d, \tag{4.4}$$

where d is a random variable of order n^{-1}, and where

$$\hat{\bar{e}}_{HT} = E'\Pi^{-1}T\iota/N, \tag{4.5}$$

is the Horvitz-Thompson estimator of the vector of residual means $\bar{e}$. Since $\bar{e}$ equals zero, it will be clear that (4.4) is an ADU-estimator of $\bar{y}$. The variance of (4.4) is approximately equal to the variance of the ordinary Horvitz-Thompson estimator (4.5), and thus (4.3) is obtained by substituting E for Y in (2.5). Expressions for variances of residuals are also given by Särndal (1982, 1985), and Särndal and Wright (1984).

In fact, application of regression estimation comes down to making inferences about the residuals E instead of making inferences about Y. It is now clear that choosing auxiliary variables that result in small residuals improves the estimates of population parameters. So, standard errors are based on the residuals. This is also the case in simple unweighted or weighted regression analysis. There are differences, however. Simple unweighted regression assumes that the observations are independent with identical variances. Neither condition is satisfied when complex sampling designs are used. In weighted analysis, coefficients are computed after weighting the observations with the square root of the covariance matrix of the disturbances. In our approach, first-order inclusion probabilities are used. In terms of design-based inference the weighted least squares estimator will not be consistent.

Weights are a part of the proposed regression estimator. Introducing the N-vector of weights

$$w = \Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}\bar{x}, \tag{4.6}$$

and recalling that the regression estimator can be written as

$$\hat{\bar{y}}_R = Y'\Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}\bar{x}. \tag{4.7}$$

It is obvious that

$$\hat{\bar{y}}_R = Y'w. \tag{4.8}$$

Notice that the weights do not depend on the target variable, but only on the auxiliary variables. So, for the computation of weights only auxiliary variables are required. However, the quality of the resulting estimator is determined by the strenght of the relationship between the target variables and the auxiliary variables. If we use the weights to estimate the means of the auxiliary variables we get

$$X'w = X'\Pi^{-1}TX(X'\Pi^{-1}TX)^{-1}\bar{x} = \bar{x}. \tag{4.9}$$

The weights balance the sample so that the sample distribution of the auxiliary variables agrees with the population distribution of these variables.

It is theoretically possible that negative weights can be obtained. Such weights may produce negative estimates of population means that are already known to be positive. In practice this will hardly ever happen as long as the sampling design is specified correctly, the linear model has sufficient explanatory power, and the sample is large enough to allow parameter estimation.

There are useful applications for the variance-covariance matrix (4.3). Since the matrix takes into account both the (possibly complex) sampling design and the weighting scheme, it provides the appropriate information for further analysis of the vector of population estimates, e.g., by multivariate methods. See, for example, Landis and Lepkowski (1983), where qualitative data from complex samples are analyzed using loglinear models. The variance-covariance matrix used there, however, takes into account only the sampling design and not the applied weighting scheme.

5. Simple Random Sampling

The first illustration of the theory developed in Section 4 uses simple random sampling

and only one target variable. Introducing y_s as the n-vector of sampled values of the target variable and X_s as the $n \times p$-matrix of auxiliary variables corresponding to sampled elements, the regression estimator reduces to

$$\hat{\bar{y}}_R = \bar{y}_s + \hat{\beta}' (\bar{x}-\bar{x}). \quad (5.1)$$

$\bar{y}_s$ is the mean of the elements in y_s, x_s is the p-vector of sample means of the auxiliary variables, and

$$\hat{\beta} = (X_s' X_s)^{-1} X_s' y_s. \quad (5.2)$$

In (5.1) we again recognize the simple regression estimator, where $\bar{x}$, $\bar{x}_s$ and $\hat{\beta}$ are vectors instead of scalars. Working out the variance (4.3) in this case gives

$$V(\hat{\bar{y}}_R) \doteq \frac{1-f}{n} (y-X\beta)' (y-X\beta) / (N-1), \quad (5.3)$$

where $f=n/N$ is the sampling fraction. This result confirms the approximation given by Cochran (1977), in the case of the simple regression estimator.

6. One-Way Stratification

The use of the regression estimator is not restricted to quantitative auxiliary variables. In Sections 6, 7, and 8 we will explore the case of qualitative auxiliary variables. In this section we will consider the use of one such auxiliary variable ($p=1$), and one target variable ($q=1$).

When a qualitative variable is included in a linear model, the variable β is replaced by as many dummy variables as it has categories. Suppose our auxiliary variable has L categories. Thus the population is divided into L non-overlapping sub-populations (strata). For each category there is a dummy variable which assumes the value 1 if the particular element belongs to that stratum, otherwise it assumes the value 0. For every element, only one dummy variable assumes the value 1; all other values are 0. Consequently the matrix X consists of N rows, each row contains exactly one 1. The columns of X sum up to the sub-population totals N_1, $N_2,\ldots,N_L$, where $N_1+N_2+\ldots+N_L=N$.

From this population a simple random sample without replacement of size n is selected. We can retain the notation used in Section 5. The columns of X_s will sum up to the (random) sample totals n_1, $n_2,\ldots,n_L$ in the strata, where $n_1+n_2+\ldots+n_L=n$. The vector of population means of the auxiliary variables is equal to $\bar{x}=(N_1, N_2,\ldots,N_L)'/N$ and the corresponding vector of sample means is equal to $\bar{x}_s=(n_1, n_2,\ldots,n_L)'/n$.

Due to the special structure of the matrix X the matrix $X_s' X_s$ is a diagonal matrix with diagonal elements equal to $n_1, n_2,\ldots,n_L$. Substituting the diagonal matrix into (5.2) results in

$$\hat{\beta} = (\bar{y}_s^{(1)}, \bar{y}_s^{(2)},\ldots,\bar{y}_s^{(L)})', \quad (6.1)$$

where $\bar{y}_s$ is the sample mean of the target variable in stratum h ($h=1,2,\ldots,L$). Substituting (6.1) into (5.1) gives as the regression estimator, in this case,

$$\hat{\bar{y}}_R = \sum_{h=1}^{L} N_h \bar{y}_s^{(h)} / N = \hat{\bar{y}}_{PS}, \quad (6.2)$$

where the subscript PS denotes the traditional post-stratification estimator. So post-stratification is a special case of the regression estimator. Since only one qualitative variable is used, we will call this case one-way stratification. Section 7 will deal with multi-way stratification.

The post-stratification estimator (6.2) is only defined if there is at least one observation available in every stratum. The same applies for the regression estimator. If there are no observations in one or more strata, then some of the diagonal elements of $X_s' X_s$ are zero, in which case $X_s' X_s$ is singular. We

can make $X_s'X_s$ non-singular by collapsing strata, or we can apply the technique of incomplete multi-way stratification which is treated in the next section.

Application of (5.3) gives an approximated variance equal to

$$V(\hat{\bar{y}}_{PS}) \doteq \frac{1-f}{n} \sum_{h=1}^{L} \frac{N_h-1}{N-1} S_h^2, \qquad (6.3)$$

in which S_h^2 is the variance (with denominator N_h-1) in sub-population h. This is a somewhat different expression than the approximation given by e.g., Cochran (1977, p. 135). The difference is caused mainly by omitting terms of the order n^{-2} and N^{-1} in (6.3).

Note that we have computed the unconditional variance, i.e., the average over all possible samples. In the literature on post-stratification, a main issue is whether the conditional or unconditional variance should be used. Holt and Smith (1979) argue that the unconditional variance should be used when comparing sampling strategies before the sample is drawn, and for inference after the sample is drawn, the conditional variance is appropriate. We chose the unconditional variance because this variance emerges in a natural way from the theory of general regression estimation. Only for the case of ordinary post-stratification is there a simple interpretation of the conditional variance, i.e., the variance conditional on the realized sample sizes in the strata. If quantitative auxiliary variables are used in the regression model, the meaning of the conditional variance is not clear. Recently Hidiroglou and Särndal (1986) have developed conditional variances for an estimator similar to the one in (4.2) in the case of simple random sampling without replacement. They found that confidence intervals based on conditional variances are smaller than confidence intervals based on unconditional variances. So, in spite of the general form of the unconditional variance (4.3), it might be better to use the conditional variance if the design lends itself to simple formulation and interpretation.

7. Multi-Way Stratification

Post-stratification is not restricted to using one qualitative auxiliary variable. The theory is equally applicable for a number of qualitative auxiliary variables. Suppose we have m qualitative variables with p_1, p_2,...,p_m categories. Now every combination of values of the auxiliary variables forms a stratum, the total number of strata being equal to $p=p_1\times p_2\times\ldots\times p_m$. If the m qualitative auxiliary variables are replaced by p dummy variables, then the theory given in Section 6 can be applied.

If the theory of linear models is applied to qualitative independent variables, it is usually called the analysis of variance. For this reason we use a terminology that has its roots in the analysis of variance. The auxiliary variables correspond to factors and the strata to cells. Stratification in which strata are constructed by crossing all the auxiliary variables corresponds to an analysis of variance in which the model contains the highest-order interaction. For this reason we call this type of post-stratification complete multi-way stratification. Complete multi-way stratification is not always practicable because of the problems of empty strata and lack of sufficient population information. However, regression estimation permits new weighting methods that are not a part of ordinary post-stratification.

In incomplete multi-way stratification a number of subsets of auxiliary variables are selected. For every subset, a complete multi-way stratification is constructed (based on less variables than the original complete multi-way stratification). By proper specification of the design matrix X, these complete multi-way sub-stratifications can be carried out simultaneously. If the highest-order

interactions are removed from the model and replaced by lower-order interactions, then in many cases the problems mentioned disappear.

To describe an incomplete multi-way stratification it is convenient to use a simple notational language. Crossing auxiliary variables is denoted by the operator "×." So, complete multi-way stratification by sex, age, marital status, region and degree of urbanization is denoted by

SEX × AGE × MARITAL STATUS × REGION × DEGREE OF URBANIZATION.

Combining several stratifications into one incomplete stratification is denoted by the operator "+." Thus, an incomplete multi-way stratification which uses the population distribution by sex, age and marital status on the one hand and the distribution by region and degree of urbanization on the other hand is denoted by

(SEX × AGE × MARITAL STATUS) + (REGION × DEGREE OF URBANIZATION).

The rows of the matrix X will in this last example not contain one 1, but two 1s. One set of dummy variables indicates the combination of sex, age and marital status, and another set of dummy variables denotes the combination of region by degree of urbanization. A consequence of these stratification designs is that the matrix $X_s'X_s$ is singular. However, this singularity can easily be removed by deleting redundant columns from X.

Stratification comes down to estimation of the parameter vector β. The number of parameters to be estimated is smaller in an incomplete stratification than in a complete stratification. So, incomplete stratification decreases estimation problems. For instance, if sex has two categories, age ten categories, marital status four categories, region eleven categories and degree of urbanization six categories, then

SEX × AGE × MARITAL STATUS × REGION × DEGREE OF URBANIZATION

requires the estimation of 5 280 parameters, whereas

(SEX × AGE × MARITAL STATUS) + (REGION × DEGREE OF URBANIZATION)

requires the estimation of, at most, 146 parameters. On the other hand the incomplete stratification model might not fit as well as the complete stratification model.

However, we believe that in practice, the incomplete stratification model is based on enough parameters that it serves almost as well as the complete stratification model.

8. An Example

In an example we illustrate the possible outcomes of the application of different weighting schemes. Using data from the Dutch Housing Demand Survey, 1977/'78, the average household income in a large town is estimated. Three auxiliary variables are available for weighting purposes: SEX (sex in two categories), AGE (age in six categories) and MAR (marital status in two categories). In the case of traditional post-stratification seven weighting schemes are possible: SEX, AGE, MAR, SEX × AGE, SEX × MAR, AGE × MAR and SEX × AGE × MAR. Using the theory of linear models eleven more weighting schemes are possible. Table 1 gives for all possible weighting schemes, which may make use of at most these three variables, the estimates of the average income, the estimate of its standard error and the approximate 95% confidence interval.

Table 1. Estimation of the Average Household Income in the Dutch Housing Demand Survey 1977/1978

Weighting scheme	Number of parameters	Estimate	Standard error	95% confidence interval	
				Lower bound	Upper bound
None	1	23494	182	23137	23852
SEX	2	23613	179	23263	23963
AGE	6	23990	170	23657	24323
MAR	2	23624	161	23308	23940
SEX×AGE	12	24012	167	23684	24340
SEX+AGE	7	24065	168	23736	24349
SEX×MAR	4	23809	160	23496	24123
SEX+MAR	3	23675	160	23361	23990
AGE×MAR	12	23987	153	23687	24287
AGE+MAR	7	24071	154	23769	24374
SEX+AGE+MAR	8	24104	154	23802	24405
(SEX×MAR)+AGE	9	24172	153	23871	24472
(SEX×AGE)+MAR	13	24078	153	23777	24379
(AGE×MAR)+SEX	13	24004	152	23705	24302
(SEX×AGE)+(SEX×MAR)	14	24149	153	23849	24449
(AGE×MAR)+(SEX×MAR)	14	24076	152	23778	24374
(SEX×AGE)+(AGE×MAR)	18	23985	152	23687	24283
(SEX×AGE)+(AGE×MAR) +(SEX×MAR)	19	24054	152	23757	24352
SEX×AGE×MAR	24	24048	152	23751	24345

Two important trends can be observed in the table: (1) using more auxiliary information increases the precision of the estimates (the standard error reduces from 182 to 152), and (2) the estimate tends to shift as more information is used (the average income increases). Further analysis showed that the sample contained too few unmarried young people and this group is known to have relatively low incomes. However, this is more than compensated for by an even greater over-representation of old unmarried people (who have very low income). Weighting by AGE and MAR gives a higher income estimate. The confidence intervals of the two extreme cases (no weighting and complete multi-way stratification) are nearly non-overlapping. If the strata are homogeneous with respect to the target variable or the response probabilities then this can indicate that weighting is efficient; it reduces bias and increases precision.

Furthermore it is clear from the table that omitting the highest order interaction has little impact. Differences are small in the group of weighting schemes that uses all

three variables. The model SEX+AGE+MAR (eight parameters) performs almost as well as the model SEX×AGE×MAR (24 parameters). We can draw the rough conclusion that any weighting scheme will do, as long as it contains the variables AGE and MAR. This particular weighting scheme requires knowledge of only the distribution by age and the distribution by marital status.

From the example it is clear that our method of weighting, based on linear models, offers a better alternative when a preferred weighting scheme cannot be carried out in a conventional way. Incomplete multi-way stratification is therefore expected to produce more accurate estimates than the ordinary post-stratifications.

9. Other Models

The optimal value of the coefficients in formula (3.3) is determined by ordinary least squares. If the residuals depend on the values of the auxiliary variables, the estimator can be improved by minimizing the sum of squares of residuals in

$$E^* = V^{1/2}E = V^{1/2}Y - V^{1/2}XB, \qquad (10.1)$$

where V is the user supplied $N \times N$-matrix specifying the relationship between residuals and auxiliary variables. The justification for the use of V has been discussed by Särndal (1980). Application of ordinary least squares to (10.1) gives

$$B = (X'V^{-1}X)^{-1}X'V^{-1}Y. \qquad (10.2)$$

The theory of the regression estimator can be developed in the same way as the theory in the previous sections, where now

$$\hat{B} = (X'V^{-1}\Pi^{-1}TX)^{-1}X'V^{-1}\Pi^{-1}TY. \qquad (10.3)$$

The ratio estimator is a special case of the regression estimator corresponding to (10.3). We get this ratio estimator by using one target variable and one auxiliary variable. We assume that the order of magnitude of the residuals is proportional to the square root of the value of the auxiliary variable. The approximated variance and estimated variance agree with the results given in Cochran (1977) and other textbooks.

Another approach to weighting based on linear models is given by Bethlehem and Keller (1982). If practical problems affect the values of the inclusion probabilities, it might be better to estimate the true inclusion probabilities on the basis of the sample. To do this we estimate a model that relates the inclusion probabilities to the available auxiliary information. An estimator of the population mean is now obtained by replacing the true, but unknown, inclusion probabilities in the usual Horvitz-Thompson estimator with the estimated inclusion probabilities.

The approaches discussed in this paper are based on the use of linear models. There is, however, no reason to restrict weighting to linear models. In fact, since weights should be non-negative, a multiplicative model might be more appropriate. Examples of the use of such models can be found in, e.g., Chapman (1976) and Bailar et al. (1978). In the literature on sampling theory this method is usually called raking estimation, but in other statistical literature the method based on the multiplicative model is also called RAS-technique and iterative proportional fitting. Just as the concept of post-stratification can be extended to linear models for weighting, the concept of raking can be extended to loglinear models for weighting.

A disadvantage of the raking estimator is that no simple formula for its variance is available. This is mainly due to the iterative way the estimator is constructed. Until recently there were only some partial results by Brackstone and Rao (1979) and Konijn (1973, 1981). They gave approximations to

the variance in the case of incomplete multi-way weighting using two auxiliary variables under simple random sampling, and stopping the iteration process after two steps. Recently, Bankier (1986) presented a new method for producing estimators in multiple frame surveys, and applied his method to the raking estimator. He suggests avoiding the discouraging complex variance formula by using a technique of repeated numerical linearization of the quantities appearing in the variance.

10. Summary and Conclusions

In this paper we proposed a general method for the computation of weighting schemes. The method is based on applications of the theory of linear models to describe the relationship between the target and the auxiliary variables. The approach is design-based and not model-dependent, i.e., we do not assume the existence of a superpopulation. The crucial hypothesis in our model is that there exists a relationship between target variables and auxiliary variables such that the variance of the target variables in the population is substantially larger than the conditional variance given the values of the auxiliary variables. If this is the case, we can improve the efficiency of our estimates by using the auxiliary information.

We show that the use of a linear model to estimate the target variables amounts to assigning weights to the observed values of the target variables. The magnitude of these weights depends only on the sample values of the auxiliary variables and their population distributions. Therefore, knowledge of the appropriate information on the auxiliary variables enables us to assign these weights to the observed elements, independent of the choice of the target variables. However, the choice of the auxiliary variables is guided by their relationship to the target variables.

The auxiliary variables could be quantitative or qualitative. In the case of one quantitative auxiliary variable, our procedure results in the simple regression estimator. In the case of one qualitative auxiliary variable, our weighting method is equivalent to the traditional post-stratification procedure. By including more qualitative auxiliary variables, new methods of stratification become possible. The most interesting one is, in our opinion, the incomplete multi-way stratification. This corresponds to a model where, in ANOVA terms, higher order interactions between the auxiliary variables are removed. In other words, the weights are obtained by fitting the sample distribution to only certain marginals of the population distribution of the auxiliary variables. This offers possibilities for solving the problem of empty strata in a traditional post-stratification by using only lower-order marginal distributions instead of all interactions. Our experience suggests that leaving out higher-order interactions seldom changes the values of the estimates or of the variances substantially. Our procedure also provides ways to deal with situations where the population distribution of certain interactions is unknown. By allowing the selection of the interaction terms to be carried out automatically, given the available auxiliary information and a lower bound for strata filling, the usual practice of collapsing strata by hand can be avoided.

Since our weighting method is based on the theory of linear models, the weights can be expressed by an explicit formula that involves some simple matrix manipulations. Additionally, we can provide an explicit formula for the (approximate) variance-covariance matrix of the estimators of the means of the target variables, taking into account both the (possibly complex) sample design and the (complete or incomplete) stratification used. Computation of the weights and the variance-covariance matrix for a given set of tar-

get variables can be done automatically, saving time and cost in the processing of sample survey data.

The Netherlands Central Bureau of Statistics has implemented the theory of weighting based on linear models in the computer program LINWEIGHT. A version which runs on an IBM-PC/XT or compatible is available, see Bethlehem (1985b).

11. References

Bailar, B.A., Bailey, L. and Corby, C. (1978): A Comparison of Some Adjustment and Weighting Procedures for Survey Data. In N. Krishnan Namboodiri, (Ed.), Survey Sampling and Measurement. Academic Press, Cambridge.

Bankier, M.D. (1986): Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys. Journal of the American Statistical Association, 81, pp. 1074–1079.

Bethlehem, J.G. (1985a): The Non-response Bias of Some Estimators. Internal CBS-report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G. (1985b): LINWEIGHT User Manual. Internal CBS-report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G. and Keller, W.J. (1982): Linear Models for Weighting of Sample Survey Data. Internal CBS-report, Netherlands Central Bureau of Statistics, Voorburg.

Bethlehem, J.G. and Keller, W.J. (1983): Weighting Sample Survey Data Using Linear Models. Internal CBS-report, Netherlands Central Bureau of Statistics, Voorburg.

Brackstone, G.J. and Rao, J.N.K. (1979): An Investigation of Raking Ratio Estimators. Sankhyā, Ser. C, 41, pp. 97–114.

Chapman, D.W. (1976): A Survey of Nonresponse Imputation Procedures. Proceedings of the American Statistical Association, Social Statistics Section, pp. 245–251.

Cochran, W.G. (1977): Sampling Techniques. Third edition, Wiley, New York.

Hansen, M.H., Madow, W.H., and Tepping, B.J. (1983): An Evaluation of Model-dependent and Probability-sampling Inferences in Sample Surveys. Journal of the American Statistical Association, 78, pp. 776–793.

Hidiroglou, M.A. and Särndal, C.E. (1986): Conditional Inference for Small Area Estimation. Invited paper presented at the Survey Research Section of the Annual meeting of the American Statistical Association held in Chicago.

Holt, D. and Smith, T.M.F. (1979): Post-stratification. Journal of the Royal Statistical Society, A, 142, pp. 33–36.

Horvitz, D.G. and Thompson, D.J. (1952): A Generalization of Sampling Without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, pp. 663–685.

Isaki, C.T. and Fuller, W.A. (1982): Survey Design under the Regression Superpopulation Model. Journal of the American Statistical Association, 77, pp. 89–96.

Konijn, H.S. (1973): Statistical Theory of Sample Survey Design and Analysis. North-Holland, Amsterdam.

Konijn, H.S. (1981): Biases, Variances, and Covariances of Raking Ratio Estimators for Marginal and Cell Totals and Averages of Observed Characteristics. Metrika, 28, pp. 109–121.

Landis, J.R. and Lepowski, J.M. (1983): The Analysis of Categorical Data from Complex Sample Surveys. Presented at the 143rd annual meeting of the American Statistical Association, Toronto.

Lindström, H., Wretman, J., Forsman, G., and Cassel, C. (1979): Standard Methods

for Non-response Treatment in Statistical Estimation. (National Central Bureau of Statistics, Stockholm) Statistics Sweden.

Little, R.J.A. (1982): Models for Non-response in Sample Surveys. Journal of the American Statistical Association, 77, pp. 237–250.

Platek, R. and Gray, G.B. (1980): Imputation Methodology. Total Survey Error. Statistics Canada, Ottawa.

Raj, D. (1968): Sampling Theory. McGraw-Hill, New York.

Robinson, P.M. and Särndal, C.E. (to appear): Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling. Sankhyā, B.

Särndal, C.E. (1980): On Π-inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling. Biometrika, 67, pp. 639–650.

Särndal, C.E. (1982): Implications for the Survey Design for Generalized Regression Estimation of Linear Functions. Journal of Statistical Planning and Inference, 7, pp. 155–170.

Särndal, C.E. (1985): How Survey Methodologists Communicate. Journal of Official Statistics, 1, pp. 49–63.

Särndal, C.E. and Wright, R.L. (1984): Cosmetic Form of Estimators in Survey Sampling. Scandinavian Journal of Statistics, 11, pp. 146–156.

Wright, R.L. (1983): Finite Population Sampling with Multivariate Auxiliary Information. Journal of the American Statistical Association, 78, pp. 879–884.

Received August 1986
Revised May 1987