Survey Methodology, June 1994
Vol. 20, No. 1, pp. 75–85
Statistics Canada

Regression Weighting in the Presence of Nonresponse with Application to the 1987-1988 Nationwide Food Consumption Survey

WAYNE A. FULLER, MARIE M. LOUGHIN and HAROLD D. BAKER[1]

ABSTRACT

A regression weight generation procedure is applied to the 1987-1988 Nationwide Food Consumption Survey of the U.S. Department of Agriculture. Regression estimation was used because of the large nonresponse in the survey. The regression weights are generalized least squares weights modified so that all weights are positive and so that large weights are smaller than the least squares weights. It is demonstrated that the regression estimator has the potential for large reductions in mean square error relative to the simple direct estimator in the presence of nonresponse.

KEY WORDS: Non-negative weights; Consistency.

1. INTRODUCTION

In many sampling situations, the population means of auxiliary variables are known, but the particular values of the variables for individual elements are not used in the sample selection. Although the information is not used in the sampling design, it may be highly desirable to incorporate the information about population means into the estimation procedure. Common estimation procedures utilizing auxiliary information are ratio estimation, post-stratification, regression estimation, and raking. Regression estimation is the most general procedure in that the regression method can handle multiple auxiliary variables, continuous auxiliary variables, and discrete auxiliary variables. Post-stratification can be considered a special case of regression estimation in which the regression variables are indicator variables for the post strata. The raking procedure, also known as iterative proportional fitting, is restricted to auxiliary information in the form of discrete categories. Deming and Stephan (1940), Stephan (1942), El-Badry and Stephan (1955), Ireland and Kulblack (1968), Darroch and Ratcliff (1972), Brackstone and Rao (1979), and Oh and Scheuren (1987) are references on raking.

Early applications of regression estimation are Watson (1937), Cochran (1942) and Jessen (1942). Cochran (1977, Ch. 7) contains the basic theory. Regression estimation for survey samples has been discussed by numerous authors, including Mickey (1959), Fuller (1975), Royall and Cumberland (1981), Isaki and Fuller (1982), Wright (1983), Luery (1986), Alexander (1987), Bethlehem and Keller (1987), Copeland, Pritzmeier, and Hoy (1987), Lemaître and Dufour (1987), Särndal, Swensson and Wretman (1989), Deville and Särndal (1992), Zieschang (1990), and Rao (1992).

In much of the cited literature, regression estimation is described as a procedure for reducing variance in probability samples. In practice, one of the motivations for regression estimation is the potential for reducing bias associated with selective nonresponse. See, for example, Little and Rubin (1987, p. 55) for the special case of adjustment cells, and Bethlehem (1988) for the generalized regression estimator.

Nonresponse prompted the use of regression estimation in our application and we discuss regression estimation in the response adjustment context in Section 3. The standard regression estimator and the modified procedure that produces positive weights are introduced in Section 2. Application of the regression weighting procedure to the Nationwide Food Consumption Survey is described in Section 4.

2. REGRESSION ESTIMATOR

To introduce the regression estimator used in our study, assume that a sample containing n units has been selected and that the probability of selecting unit i is π_i. For our purposes, it is sufficient for π_i to be proportional to the selection probabilities. The sample might be a two-stage stratified sample, and the unit can be either the primary sampling unit or the observation unit. In our application, the unit is the observation unit. Assume that a k-dimensional vector of population means, denoted by $\bar{X} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ is known, that the vector $(y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$ is observed for every unit in the sample and that an estimator of the mean of y is desired. We assume that the first element of x_i is one for all i. Hence, the first element of $\bar{X}$ is also one. The vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ is sometimes

[1] Wayne A. Fuller, Marie M. Loughin and Harold D. Baker, Iowa State University.

called the vector of control variables. A regression estimator of the mean of y is

$$\bar{y}_r = \bar{X}\hat{\beta}, \tag{2.1}$$

where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \right)^{-1} \sum_{i=1}^{n} x_i' \pi_i^{-1} y_i, \tag{2.2}$$

and we have assumed $\sum x_i' \pi_i^{-1} x_i$ to be nonsingular. This definition of the regression estimator follows Mickey (1959) who suggested restricting the term regression estimator to estimators that are location and scale invariant. The estimator (2.1) can also be written as

$$\bar{y}_r = \sum_{i=1}^{n} w_i y_i, \tag{2.3}$$

where

$$w_i = \bar{X}\left(\sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \right)^{-1} x_i' \pi_i^{-1}, \tag{2.4}$$

and the weights have the property,

$$\sum_{i=1}^{n} w_i x_i = \bar{X}. \tag{2.5}$$

The weights of expression (2.4) are relatively easy to compute, and once computed, can be used for the estimation of any y-variable. If the vector x_j is replaced by the vector

$$(1, z_j) = (1, x_{j2} - \bar{X}_2, x_{j3} - \bar{X}_3, \ldots, x_{jk} - \bar{X}_k), \tag{2.6}$$

the estimator can be written in the form

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

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

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

and

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

In the form (2.7), $\bar{y}_r$ is the intercept in the regression of y on z. Thus, the theory given by Fuller (1975) for regression coefficients is applicable to the regression estimator of the mean. If the population total of units is known and denoted by N, the estimated population total is $N\bar{y}_r$.

By defining a sequence of populations and samples, it is possible to show that the estimator (2.1) is a consistent estimator of the mean of y. See, for example, Fuller (1975). The estimator of the variance of the regression estimator is a function of the joint probabilities. Consider a stratified two-stage sample and replace our single subscript i with the triple ℓjt. Then, omitting the finite correction term, a variance estimator is

$$\hat{V}\{\bar{y}_r\} = (n - k)^{-1} n \sum_{\ell=1}^{L} (n_\ell - 1)^{-1} n_\ell \sum_{j=1}^{n_\ell} (d_{\ell j.} - d_{\ell ..})^2, \tag{2.8}$$

where

$$d_{\ell j.} = \sum_{t=1}^{m_{\ell j}} w_{\ell jt} (y_{\ell jt} - x_{\ell jt} \hat{\beta}),$$

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

n_ℓ is the number of sample primary sampling units in stratum ℓ, $m_{\ell j}$ is the number of sample elements in primary sampling unit j of stratum ℓ, $\hat{\beta}$ is the vector of coefficients defined in (2.2), n is the total number of elements in the sample, and $w_{\ell jt}$ is the weight for element t in primary sampling unit j of stratum ℓ. The factor $n - k$ is used by analogy to the divisor for the unbiased estimator of the error variance in ordinary regression. When the vector of control variables is coded as in (2.6), the estimator (2.8) is the estimated variance of the first element of $\hat{\beta}$, the estimated intercept. The estimator (2.8) was suggested in Hidiroglou, Fuller and Hickman (1976) and the consistency of the estimator was established by Fuller (1975). Also see Särndal, Swensson and Wretman (1989).

The estimators, constructed with weights (2.4), have good large sample properties. However, they may have undesirable behavior in small samples. Because the weights are linear functions of x_i, it is possible for some of the weights to be negative. Negative weights make it possible for estimates of positive parameters to be negative. Early research on methods of constructing nonnegative regression weights was conducted by Husain (1969). Huang (1978) designed a computer program to produce nonnegative regression weights. Huang and Fuller (1978) described the weight generation procedure and showed

that the large sample distribution of the modified estimator is the same as that of the ordinary regression estimator. Also see Goebel (1976).

The computer algorithm of Huang (1978) is an iterative procedure based upon the ideas of generalized least squares. The goal of the Huang algorithm is a set of weights w_i, $i = 1, 2, \ldots, n$, satisfying (2.5) that do not differ greatly from the initial weights, where difference is a function of the initial weight. The Huang algorithm attempts to compute weights w_i satisfying

$$(1 - M) \max_{1 \le i \le n} w_i \pi_i^{-1} \le (1 + M) \min_{1 \le i \le n} w_i \pi_i^{-1},$$

where the parameter M, $0 < M \le 1$, is specified by the user and is generally chosen in the interval $[0.8, 1.0]$. If the first round regression weights defined by (2.4) do not satisfy the requirements, a second round of regression weights is computed. The second round weights are weighted regression weights in which a control factor is assigned to each observation. Small control factors are assigned to observations with large or small first round weights. Relatively large control factors are assigned to observations with first round weights close to π_i^{-1}. The second round regression weights are checked and if they fail to satisfy the criteria, the control factors are modified, and so on. The algorithm is given in the Appendix.

The control weighting used in the Huang algorithm has much in common with bounded-influence and robust regression methods. That is, in the final estimator, the contribution to the estimation of the slope vector is reduced for observations that are far from the mean. See Hampel (1978), Krasker (1980), and Mallows (1983). Recent research on this type of estimator for survey samples is that of Deville and Särndal (1992), Akkerboom, Sikkel, and van Herk (1991), Hulliger (1993) and Singh (1993).

It is not always possible to construct weights meeting the criteria and also satisfying (2.5). For example, if all of the observations on x_{t2} exceed the mean, there is no set of positive weights summing to one that also satisfy $\sum_{t=1}^{n} x_{t2} w_t = \bar{X}_2$. Therefore, the weight generation program will terminate if weights meeting the specified criteria cannot be constructed after a specified number of iterations.

In some situations it is desirable to restrict the weights to the nonnegative integers. This is true when estimates of totals are being constructed and the population contains well defined units, such as people. Nonnegative integer weights then provide more comfortable estimates, in that the estimates are physically attainable. Integer weights can be constructed so that no rounding is necessary when building tables. With such integer weights, all multiple way tables will automatically be internally consistent.

The Huang program contains an option to round the real weights to integer weights in a manner that maintains the sum of the weights. After rounding, the equalities (2.5) will generally no longer hold exactly. We have found that by iterating the Huang algorithm using the first-round integer weights as initial weights, integer weights can be constructed such that there is a very modest deviation from equality for expression (2.5). Cox (1987), Cox and Ernst (1982), and Fagan, Greenberg and Hemmig (1988) discuss rounding.

3. REGRESSION ESTIMATION WITH NONRESPONSE

The early theoretical developments for regression estimation assumed the sample to be a probability sample from the population. However, it has long been recognized that regression estimation can be used to reduce the bias that arises from imperfections in the data collection procedure. The most obvious of these imperfections is nonresponse. In all large samples of human subjects, some of the subjects fail to provide information. If the nonrespondents differ from the respondents, direct estimates constructed from the respondents will be biased. Given auxiliary information, regression estimation provides a method or reducing the bias. The degree to which the bias is reduced depends upon the relationship between the control variables, the variables of interest, and the response probabilities. See Little and Rubin (1987) for a general discussion of these issues.

Let π_i^* denote the inclusion probability equal to the product of π_i and the conditional probability of observing the unit given that the unit is selected. Then

$$E\left\{ \sum_{i=1}^{n} x_i' \pi_i^{-1} x_i \mid \xi_N \right\} = \sum_{i=1}^{N} x_i' \pi_i^{-1} \pi_i^* x_i \qquad (3.1)$$

and

$$E\left\{ \sum_{i=1}^{n} x_i' \pi_i^{-1} y_i \mid \xi_N \right\} = \sum_{i=1}^{N} x_i' \pi_i^{-1} \pi_i^* y_i, \qquad (3.2)$$

where the expectations are conditional on the given finite population ξ_N, and n is the realized sample size. In the case of nonresponse, the ratio $p_i = \pi_i^* \pi_i^{-1}$ is the response probability for individual i. Therefore, under conditions such as those used by Fuller (1975),

$$\operatorname*{plim}_{\substack{n \to \infty \\ N \to \infty}} (\hat{\beta} - \gamma) = 0, \qquad (3.3)$$

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

$$\gamma = \left(\sum_{i=1}^{N} x_i' \pi_i^{-1} \pi_i^* x_i \right)^{-1} \sum_{i=1}^{N} x_i' \pi_i^{-1} \pi_i^* y_i. \qquad (3.4)$$

Then

$$\bar{Y} = \bar{X}\gamma + \bar{A}, \tag{3.5}$$

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

$$y_i = x_i\beta + e_i, \quad E\{e_i\} = 0$$

holds for all i.

The mean $\bar{A}$ is zero when $\overset{*}{\pi}_i = \pi_i$ for all i and an element of x_i is one for all i because then

$$\gamma = \beta = \left(\sum_{i=1}^{N} x_i'x_i\right)^{-1}\sum_{i=1}^{N} x_i'y_i \tag{3.6}$$

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

$$cx_i' = \overset{*}{\pi}_i^{-1}\pi_i = p_i^{-1}, \tag{3.7}$$

for $i = 1, 2, \ldots, N$. Thus, if the ratio of nominal probabilities to true probabilities is a linear function of the control variables, the regression estimator is a consistent estimator of the mean of y, where the limit is for sequences as defined in Fuller (1975). One way in which (3.7) can be satisfied is for the elements of x_i to be dummy variables that define subgroups and for the response probabilities to be constant in each subgroup. This situation is sometimes described by saying that elements are missing at random in each subgroup. We take the assumption that $\bar{A} = 0$ as our working assumption in the empirical analysis.

In any regression problem, it is impossible to use the sample to verify some of the assumptions. For example, in ordinary least squares regression, the residuals $\hat{e}_i = y_i - x_i\hat{\beta}$ are uncorrelated with x_i and, hence, the residuals cannot be used to check the assumption that the true errors are uncorrelated with x. Thus, in a survey with nonresponse, one searches for control variables that are correlated with y and (or) that one believes are correlated with the response probabilities. But one cannot guarantee that all bias has been removed by regression estimation based on a particular set of control variables.

In practice, one can often identify x-variables that are correlated with the probability of response and (or) correlated with the y-variables. For example, in the 1987-1988 Nationwide Food Consumption Survey, the response rate was low among high-income households. Therefore, use of variables for household income in a regression estimator is expected to reduce the bias in the estimated mean for characteristics that are correlated with income.

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

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

where a_i is defined in (3.5),

$$T = \sum_{i=1}^{N}\pi_i^{-1}\overset{*}{\pi}_i$$

and

$$G = T^{-1}\sum_{i=1}^{N} x_i'\pi_i^{-1}\overset{*}{\pi}_i x_i.$$

Under reasonable assumptions

$$\hat{T} = \sum_{i=1}^{n}\pi_i^{-1}$$

and

$$\hat{G} = \hat{T}^{-1}\sum_{i=1}^{n} x_i'\pi_i^{-1}x_i$$

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

The estimator of variance (2.8) also remains appropriate if the regression weights are constructed by a procedure other than (2.4). For example, let the weights be defined by

$$w_{gi} = \bar{X}\left[\sum_{i=1}^{n} x_i'\pi_i^{-1}g_i x_i\right]^{-1} x_i'\pi_i^{-1}g_i,$$

where the g_i are functions of the x_i. Assume

$$\text{plim}\,\hat{\beta}_g = \gamma_g,$$

where

$$\hat{\beta}_g = \left[\sum_{i=1}^{n} x_i'\pi_i g_i x_i\right]^{-1}\sum_{i=1}^{n} x_i'\pi_i^{-1}g_i y_i.$$

Also assume

$$\underset{N\to\infty}{\text{plim}}\, N^{-1}\sum_{i=1}^{N}(y_i - x_i\gamma_g) = 0.$$

Then expression (2.8) with w_{gi} replacing w_{tjt} is a consistent estimator of the variance of the estimator. The estimator (2.8) will be used in our empirical analyses.

Formula (2.8) identifies the two effects of regression estimation on the variance of an estimated mean. The correlation effect reduces the variance of the estimated mean while the increase in the sum of squares of the weights increases the variance of the estimated mean. To understand these effects, consider a simple random sample. If the y variable is correlated with x, the correlation tends to reduce the variance of the regression estimator relative to that of the simple estimator because

$$E\{(y_i - x_i\beta)^2\} \leq E\{[y_i - E(y_i)]^2\}.$$

If the sample means of the control variables differ from the population means, then

$$\sum_{i=1}^{n} w_i^2 > n^{-1},$$

where n^{-1} is the sum of squares of the simple weights for a simple random sample.

When comparing the variance of the sample mean with the variance of the regression estimator, one should not forget that one of the reasons for using regression estimation in samples with nonresponse is to produce an estimator with less bias than that of the direct estimator. Thus, in the next section we compare an estimator of the mean square error of the simple estimator to an estimator of the variance of the regression estimator.

4. APPLICATION TO THE NATIONWIDE FOOD CONSUMPTION SURVEY

The 1987-1988 Nationwide Food Consumption Survey was conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture. The original sample was a self-weighting stratified sample of area primary sampling units within the 48 conterminous states. Primary sampling units were divided into secondary units called area segments. Households within the sample segments were contacted by personal interview. The field operation was conducted during the period April 1987 through August 1988 by a contractor under contract to the Human Nutrition Information Service.

Approximately 37% of the housing units identified as occupied provided complete household food use information. The realized household sample contains 4,495 households. Because of the low response rate, the Human Nutrition Information Service decided to use regression weighting in the estimation. Population totals for all characteristics except urbanization were estimated by the Human Nutrition Information Service from the March 1987 Current Population Survey. See Bureau of the Census (1987). The population totals for urbanization classes were furnished by the contractor. In our analysis, we treat the estimated population totals as if they were known population totals.

Table 1

Sample and population characteristics of households

Characteristic	Category	Household Sample Frequency	Household Sample Percent	Population Percent
Season	Spring	1,828	40.7	25.0
	Summer	678	15.1	25.0
	Fall	717	16.0	25.0
	Winter	1,272	28.3	25.0
Region	Northeast	905	20.1	21.2
	Midwest	1,172	26.1	24.7
	South	1,567	34.9	34.4
	West	851	18.9	19.6
Urbanization	Central Cities	1,064	23.7	31.2
	Suburban	2,122	47.2	46.0
	Nonmetro	1,309	29.1	22.9
Household Income as % of Poverty	< 131%	1,041	23.2	20.0
	131-300%	1,564	34.8	32.2
	301-500%	1,108	24.6	25.9
	> 500%	782	17.4	21.8
Household Food Stamps	Yes	314	7.0	7.4
	No	4,181	93.0	92.6
Ownership of Domicile	Yes	2,998	66.7	64.1
	No	1,497	33.3	35.9
Race of Household Head	Black	519	11.5	11.1
	Nonblack	3,976	88.5	88.9
Age of Household Head	< 25	338	7.5	7.9
	25-39	1,588	35.3	36.1
	40-59	1,369	30.5	30.5
	60-69	660	14.7	13.0
	70+	540	12.0	12.6
Household Head Status	Both Male and Female	3,057	68.0	60.8
	Female Only	1,044	23.2	26.0
	Male Only	394	8.8	13.2
Female Head Worked	Yes	1,792	39.9	41.5
	No	2,703	60.1	58.5
Exactly One Adult in Household	Yes	1,211	26.9	29.7
	No	3,284	73.1	70.3
Exactly Two Adults in Household	Yes	2,616	58.2	54.2
	No	1,879	41.8	45.8
Presence of Child < 7 Years Old	Yes	1,009	22.4	20.1
	No	3,486	77.6	79.9
Presence of Child 7-17 Years Old	Yes	1,309	29.1	26.5
	No	3,186	70.9	73.5
Household Size	(Mean)		2.731	2.642
Household Size, Squared	(Mean)		9.546	9.125

Characteristics of the population and of the household sample are given in Table 1. The original sample was unbalanced with respect to time of interview with nearly 41% of the interviews in the spring quarter and about 16% of the interviews in each of the summer and fall quarters. Interviews for the spring and summer quarters were done in both 1987 and 1988.

The sample was also unbalanced with respect to urbanization. There was a lower fraction of central city households in the sample than in the population (24% versus 31%), and a higher fraction of nonmetropolitan households in the sample than in the population (29% versus 23%).

The fraction of high income households was smaller in the sample than in the population. The sample contained a higher fraction of households with both a male and female head than the population (68% versus 61%). A higher fraction of the sample than of the population consisted of households with children. The sample was only mildly unbalanced with respect to several other socio-demographic characteristics.

The characteristics listed in Table 1 are believed by the staff of the Human Nutrition Information Service to be related to food consumption behavior. Therefore, variables based on those characteristics were used in the regression weighting procedure. To implement the weight generation program, each of the categorical variables of Table 1 was converted to a set of indicator variables. For example, three variables were created for the characteristic, household income as a percent of poverty. These were

$Z_{t1} = 1$ if income $< 131\%$ for t-th household
$= 0$ otherwise,

$Z_{t2} = 1$ if income is 131-300% for t-th household
$= 0$ otherwise,

$Z_{t3} = 1$ if income is 301-500% for t-th household
$= 0$ otherwise.

Using this procedure, 25 indicator variables were created. In addition, household size and the square of household size were used as continuous variables.

The twenty-seven variables were used to generate regression weights using Huang's program. The parameter M of the weight generation program was set equal to 0.9 in the computation. The weights were rounded to integers, where each integer weight is a weight in thousands. The sum of the weights is 88,942, which is the number of households in the population in thousands. The average weight is 19.787, the smallest weight is 6, and the largest weight is 47. Thus, the largest weight is 2.38 times the average weight. The sum of squares of the weights is 2,317,930. The average weight squared and multiplied by the sample size is 1,759,884. Thus, if a variable has zero multiple correlation with the 27 variables, the variance of an estimate computed with the weights will be about 1.32 times the variance of the simple unweighted estimator.

Figure 1 shows the regression weights computed by the Huang algorithm plotted against the ordinary least squares weights. Because there are 4,495 households, many points are hidden. Both weights are standardized by dividing by the average weight. Thus, the average for weights coded in this manner is one. Because there are 27 control variables used in the construction, the Huang weights tend to form a swarm of points about an S-shaped function of the original weights. If there were only one control variable, the points would fall on an S-shaped curve. The original weights for observations to the left of zero were negative.

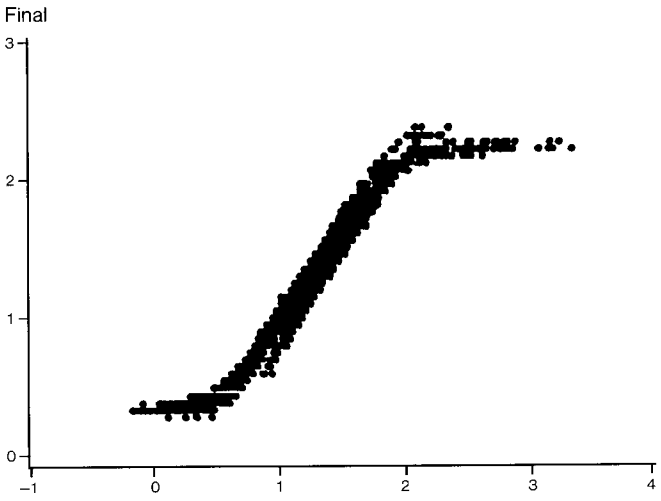

Figure 1. Plot of final weights against the ordinary least squares weights, both expressed relative to the average weight.

To compare estimates constructed with weights to unweighted estimates, we use the variables

Y_1 = adjusted total number of meals away from home (meals away),

Y_2 = total money value of food used at home (home food),

Y_3 = household size in 21-meal-equivalent persons (meal persons),

Y_4 = indicator to identify housekeeping households (housekeeping).

The household size in 21-meal-equivalent persons is the total adjusted meals eaten from household food supplies in the past 7 days divided by 21. "Meal persons" is the sum of two terms. The first term is the sum of the proportions of meals eaten at home in the interview week by each household member. The second term is the number of meals served to guests, boarders, and employees during the interview week, divided by 21. In other words:

$$\text{meal persons for } j\text{-th household} = \sum_i (h_{ij} + a_{ij})^{-1} h_{ij} + (21)^{-1} b_j,$$

where h_{ij} = meals eaten at home by the i-th individual in the j-th household during the interview week, a_{ij} = meals eaten away from home by the i-th individual in the j-th household during the interview week, b_j = number of meals eaten by nonhousehold members in the j-th household during the interview week.

The adjusted total number of meals bought and eaten away from home is the sum of the proportions of meals eaten away from home in the interview week by household members, multiplied by 21. In the notation used to define meal persons,

$$\text{meals away for } j\text{-th household} = 21 \sum_i (h_{ij} + a_{ij})^{-1} a_{ij}.$$

The total value of food used at home is the expenditures for purchased food plus the money value of home-produced food and food received free-of-cost that was used during the survey week. Expenditures for purchased food were based on prices reported as paid regardless of the time of purchase. Sales tax was excluded. Purchased food with unreported prices, food produced at home, food received as a gift, and food received instead of pay were valued at the average price per pound paid for comparable food by survey households in the same region and season.

A housekeeping household is a household with at least one person having ten or more adjusted meals from the household food supply during the seven days before the interview. Household food-use analyses generally consider only housekeeping households.

Table 2
Properties of alternative estimators

Variable	Unweighted Mean	Weighted Mean	Difference	Relative Efficiency of Regression
Meals away				
Housekeeping	7.75 (0.22)	7.93 (0.17)	−0.18 (0.09)	2.52
Nonhousekeeping	18.31 (1.14)	18.12 (1.19)	0.19 (0.68)	0.92
All	8.27 (0.22)	8.57 (0.22)	−0.30 (0.12)	2.56
Home food				
Housekeeping	61.10 (1.14)	59.56 (0.98)	1.54 (0.41)	3.65
Nonhousekeeping	25.99 (1.25)	26.39 (1.46)	−0.40 (1.00)	0.73
All	59.37 (1.12)	57.49 (0.91)	1.88 (0.39)	5.60
Meal persons				
Housekeeping	2.42 (0.03)	2.33 (0.01)	0.09 (0.01)	89.00
Nonhousekeeping	0.51 (0.03)	0.49 (0.03)	0.02 (0.02)	1.00
All	2.33 (0.03)	2.22 (0.01)	0.11 (0.01)	129.00
Housekeeping (%)	95.06 (0.40)	93.77 (0.58)	1.29 (0.10)	5.30

The means of the variables computed using unweighted data are given in Table 2 in the column headed, "Unweighted mean". Three means are given for meals away, home food, and meal persons. Two means are computed for the two subpopulations defined by the housekeeping variables. The third mean, designated by "all" in the table, is the estimated mean for the entire population. The standard errors of the estimates are given in parentheses below the estimates. The estimates and standard errors for the unweighted estimates were computed in PC CARP. See Fuller *et al.* (1986). The computations accounted for the fact that the sample is an area stratified cluster sample.

Because the sample is a two-stage sample, the estimated variances are larger than the variance of a simple random sample containing the same number of households. The ratio of the variance for a sample estimate to the variance of a simple random sample containing the same number of individuals is called the design effect. The estimated design effect is about 2.5 for meals away and meal persons, is about 4.1 for home food, and is about 1.5 for housekeeping for the "all" means for the unweighted sample.

The column headed "Weighted mean" contains the estimates computed with the regression weights. The standard errors were computed in PC CARP using formula (2.8) with the regression weights replacing the π_i^{-1}. The variance calculation requires computing a regression for every y-variable. The estimated means for the subpopulations are ratios of regression estimators. The variances for the subpopulation means were computed by calculating the variances of the Taylor deviates for the ratio using formula (2.8). The standard errors for unweighted and weighted estimates are similar for meals away and home food. However, the standard errors for the regression estimate of the population mean of meal persons is about one third of the standard error of the unweighted estimate. The standard error of the regression estimator is smaller because meal persons is highly correlated with the household size variables used as controls in the regression procedure.

The estimated squared multiple correlation between the variables of the table and the 27 control variables is 0.29, 0.44, 0.82, and 0.12 for meals away, home food, meal persons, and housekeeping, respectively. If the sample means of the control variables were nearly equal to the population means, the standard error of the regression estimate of meals away would be about $(1 - 0.29)^{1/2} = 0.84$ times the standard error of the unweighted estimate. In fact, the estimated standard error of the regression is about 0.97 times the standard error of the unweighted estimate. The difference is due to the fact that $\sum_{i=1}^{n} w_i^2$ is considerably bigger than n^{-1} because the sample is unbalanced on a number of items. Note that

$$0.97 \doteq [(0.71)(1.32)]^{1/2},$$

where $0.71 = (1 - 0.29)$ is one minus the squared correlation and $1.32 = n \sum_{i=1}^{n} w_i^2$. The situation for housekeeping is more extreme. The correlation is not large, and, apparently, the large deviations from the regression line are associated with large weights. The estimated variance for the regression estimator is about twice the estimated variance of the unweighted estimator.

Table 2 also contains the estimated differences between the unweighted and weighted estimators. The difference between the unweighted and the weighted estimated total is

$$\sum_{t=1}^{n} Nn^{-1}y_t - \sum_{t=1}^{n} w_t y_t = \sum_{t=1}^{n} (n^{-1}N - w_t)y_t.$$

The difference between the estimated means is the difference between the totals divided by the population size. To compute the variance of the difference between the means, we note that the hypothesis of a zero difference is equivalent to the hypothesis that the correlation between w_t and y_t is zero. Therefore, we computed the unweighted regression of y_t on w_t and computed the variance of the regression coefficient under the design using PC CARP. The standard errors for the difference in Table 2 are such that the "t-statistic" for the hypothesis of zero difference is equal to the "t-statistic" for the coefficient of w_t in the regression of y_t on w_t.

For all four characteristics, the difference between the weighted and unweighted estimators of the population mean is significant at traditional levels. Thus, under the assumption that the regression estimators are unbiased, there are significant biases in the unweighted estimators.

The bias picture is mixed for the estimates of the subpopulation means. The difference between the two estimators is significant for the three means for the housekeeping subpopulation, which is the population of interest. The difference is nonsignificant for the three means for the nonhousekeeping subpopulation. The sample contains only 222 nonhousekeeping households. Therefore, the variance of the difference between the weighted and unweighted estimates is much larger for the nonhousekeeping households than for the housekeeping households.

The differences between the two estimates of the population means are a function of the differences between the two estimates of the subpopulation means and the two estimates of the fraction of households in the two categories. This explains why the difference for "all" can be larger than both the "housekeeping" and "nonhousekeeping" differences.

The last column of Table 2 contains the ratio of the estimated mean square error of the unweighted estimator to the variance of the regression estimator. The estimated mean square errors for the unweighted estimators were computed as

$$\widehat{\text{MSE}}_u = \hat{V} + \max\{0, (\text{Diff})^2 - (\text{s.e. diff})^2\},$$

where $\hat{V}$ is the estimated variance of the unweighted estimate, Diff is the difference between the two estimates from Table 2, and s.e. diff is the standard error of the difference from Table 2. The estimated variance $\hat{V}$ for the unweighted estimator is variance formula (2.8) with constant w_{ijt}, and with $x_{ijt}\hat{\beta}$ replaced by $\bar{y}_{t..}$. The second term of the estimated mean square error is the estimated squared bias. Under the assumption that the regression estimator is unbiased, the expected value of $(\text{Diff})^2$ is the squared bias plus the variance of the difference. Hence, under the assumption that the regression estimator is unbiased, the estimated mean square error of the unweighted estimator is a consistent estimator. The estimated mean square errors of the weighted estimators are the variances of the weighted estimators computed as the squares of the standard errors of Table 2.

Of the four characteristics for which the population mean was estimated, the estimated relative efficiency of the regression estimator to the simple mean ranges from 2.5 to 129. The regression estimator for meals away has the smallest estimated efficiency. The variances of the two estimators are similar, but because of the estimated bias, the regression estimate for meals away is estimated to have a mean square error that is about 40% of that of the unweighted estimate. The mean square error of the regression estimate for home food is less than 20% of that of the unweighted estimate, that for meal persons is about 1% of that of the unweighted estimate, and that for housekeeping is about 20% that of the unweighted estimator. In all cases, the squared bias is a very important component of the estimated mean square error.

Because the unweighted subpopulation estimates for the nonhousekeeping households showed little bias, the unweighted estimates are estimated to be somewhat more efficient than the regression estimates. The nonhousekeeping subpopulation is only about 6% of the population and the deviations from the subpopulation mean show little correlation with the control variables. On the other hand, the regression estimates for the housekeeping subpopulation are estimated to be much more efficient than the unweighted estimates. The relative efficiencies for the housekeeping subpopulation are close to those of the total population estimates.

Even after allowing for the fact that the population totals from the Current Population Survey are not known population totals, it is clear that large gains are associated with regression estimation for the population means. Although the regression estimator for the means of the small subpopulation is estimated to be less efficient than the unweighted estimators, the loss in efficiency is small relative to the large gains in efficiency estimated for the other variables.

ACKNOWLEDGEMENTS

This research was partly supported by Research Support Agreement 58-3198-9-032 with the Human Nutrition Information Service, U.S. Department of Agriculture. We thank Phil Kott, Patricia Guenther, and the referees for useful comments.

APPENDIX
WEIGHT GENERATION PROGRAM

In this appendix, we present the regression weight generation procedure of Huang and Fuller (1978). The procedure we describe contains the option of specifying maximum and minimum weights. This option was not part of the original program. For a discussion of related weight generation procedures, see Singh (1993).

Suppose that the population means $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)$ of the k auxiliary variables $(X_1, X_2, \ldots, X_k)$ are known. Let a sample of n observations be available and let

$$X = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix}, \tag{A.1}$$

where X_{ij} is the observation on variable j for individual i.

In addition to the array of sample observations and the populations means, two initial factors v_i and $g_i^{(0)}$, $i = 1, 2, \ldots, n$, are required to initiate the computations. The v_i are typically inversely proportional to the probabilities of selection. The default values for $g_i^{(0)}$ are $g_i^{(0)} = 1$. For stratified samples or data with unequal variances, the user may choose other values for $g_i^{(0)}$. (See Huang 1978 or Goebel 1976.) The program input includes the sample size n, the population size N, the parameter M, the maximum number of iterations LI, the upper bound of the ratios of weights to the average weight U_B, and the lower bound of the ratios of weights to the average weight L_B. It is required that $0 \leq L_B < 1 < U_B$. In our description, we assume $\sum_{i=1}^{n} v_i = n$. The program normalizes the v_i so that the sum is n.

The program can be used to construct weights to estimate means or to estimate totals. The weights for totals are the weights for the means multiplied by N. For means, the program attempts to construct weights w_i such that

$$\sum_{i=1}^{n} w_i(1, X_i) = (1, \bar{X}), \tag{A.2}$$

$$L_B < nw_i < U_B, \tag{A.3}$$

$$(1 - M) \max_{1 \leq i \leq n} w_i v_i \leq (1 + M) \min_{1 \leq i \leq n} w_i v_i, \tag{A.4}$$

for $i = 1, 2, \ldots, n$.

The program is iterative, where an iteration consists of computing the generalized least squares weights, where a control factor h_i is applied to each observation. The h_i is a product of v_i and g_i, where g_i for iterations after the first is a "bell" shaped function of the distance (in a suitable metric) that the observation is from the population mean. At each iteration, the weights satisfy (A.2) but may fail (A.3) or (A.4).

It will not always be possible to construct weights satisfying the specified restrictions in the specified number of iterations. If the sample is such that the restriction cannot be met, the program outputs the weights of the last iteration. In the single x case, when the criterion cannot be satisfied, there will be two weights, one for those greater than the population mean, and one for those less than the population mean.

To describe the algorithm, let

$$Z_{ij} = X_{ij} - \bar{X}_j,$$

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \ldots & Z_{1p} \\ \vdots & \vdots & & \vdots \\ Z_{n1} & Z_{n2} & \ldots & Z_{np} \end{pmatrix},$$

$$V = \text{diag}(v_1, v_2, \ldots, v_n),$$

$$J_n = (1, 1, \ldots, 1)',$$

$$A^{(0)} = Z'H^{(0)}Z,$$

$$G^{(0)} = \text{diag}(g_1^{(0)}, \ldots, g_n^{(0)})$$

and

$$H^{(0)} = VG^{(0)}.$$

The algorithm consists of iterating three steps.

1. The initial calculation is for $\alpha = 0$. At iteration α, the vector of regression weights, denoted by $w^{(\alpha)}$, is

$$w^{(\alpha)} = [1 + n\bar{u}_v^{(\alpha)}]^{-1} V(n^{-1}J_n + u^{(\alpha)}) = (w_1^{(\alpha)}, \ldots, w_n^{(\alpha)})', \tag{A.5}$$

where

$$u^{(\alpha)} = G^{(\alpha)}Z(A^{(\alpha)})^{\dagger}(\bar{X} - \bar{x}_v) = (u_1^{(\alpha)}, \ldots, u_n^{(\alpha)})',$$

$$\bar{x}_v = \left(\sum_{i=1}^{n} v_i \right)^{-1} \sum_{i=1}^{n} v_i x_i,$$

$(A^{(\alpha)})^{\dagger}$ is a symmetric generalized inverse of $A^{(0)}$,

$$n\bar{u}_v^{(\alpha)} = \max\{J_n'Vu^{(\alpha)}, n^{-1} - 1\}, \tag{A.6}$$

and

$$A^{(\alpha)} = Z'H^{(\alpha)}Z.$$

2. The weights of Step 1 are checked to see if they satisfy the criteria.

   (a) Is $| nu_i^{(\alpha)} | \leq M$ for all i?

   (b) Is

   $$L_B \leq nw_i^{(\alpha)} \leq U_B$$

   for all i?

   If either (a) or (b) fails for any i and LI iterations have not been completed, go to Step 3. If (a) and (b) are satisfied, or if LI iterations have been completed, the weights are output.

3. The control factors $h_i^{(\alpha)}$, $i = 1, 2, \ldots, n$, are modified. Set

   $$H^{(\alpha)} = H^{(\alpha-1)}G^{(\alpha)},$$

   where

   $$G^{(\alpha)} = \text{diag}(g_1^{(\alpha)}, g_2^{(\alpha)}, \ldots, g_n^{(\alpha)}),$$

   $$g_i^{(\alpha)} = 1 \qquad 0 \leq d_i^{(\alpha)} < 0.5$$

   $$= 1 - 0.8(d_i^{(\alpha)} - 0.5)^2 \qquad 0.5 \leq d_i^{(\alpha)} \leq 1$$

   $$= 0.8(d_i^{(\alpha)})^{-1} \qquad d_i^{(\alpha)} > 1,$$

   $$d_i^{(\alpha)} = 1.33[D_i^{(\alpha-1)}]^{-1} n \mid u_i^{(\alpha-1)} \mid,$$

   $$D_i^{(\alpha-1)} = \min\{M, C_{\text{Li}}^{(\alpha-1)}\} \qquad \text{if } w_i^{(\alpha-1)} < v_i$$

   $$= \min\{M, C_{\text{Bi}}^{(\alpha-1)}\} \qquad \text{if } w_i^{(\alpha-1)} \geq v_i,$$

   $$C_{\text{Li}}^{(\alpha-1)} = \max\{| v_i^{-1}(1 + n\bar{u}_v^{(\alpha-1)})L_B - 1 |, 0.1\, M\},$$

   $$C_{\text{Bi}}^{(\alpha-1)} = \max\{| v_i^{-1}(1 + n\bar{u}_v^{(\alpha-1)})U_B - 1 |, 0.1\, M\}.$$

   Go to Step 1 to compute new regression weights.

The constant 1.33 in the definition of $d_i^{(\alpha)}$ and the constant of 0.8 in the definition of $g_i^{(\alpha)}$ were chosen to speed convergence. The control factors $g_i^{(\alpha)}$ are chosen to downweight observations on the basis of a distance from the population mean.

The definition of $w^{(\alpha)}$ in (A.5) is an alternative way of writing the vector of generalized least squares weights of (2.4) when $\pi_i^{-1} = h_i^{(\alpha)}$.

REFERENCES

AKKERBOOM, J.C., SIKKEL, D., and van HERK, H. (1991). Robust weighting of financial survey data. Contributed paper presented at meeting of the International Statistical Institute, Cairo, Egypt.

ALEXANDER, C.H. (1987). A model based justification for survey weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 183-188.

BETHLEHEM, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

BETHLEHEM, J.G., and KELLER, W.A. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.

BRACKSTONE, G.J., and RAO, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyā*, Series C, 97-114.

BUREAU OF THE CENSUS (1987). Current Population Survey, March 1987: Technical Documentation. Washington, D.C.

COCHRAN, W.G. (1942). Sampling theory when the units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.

COCHRAN, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley.

COPELAND, K.R., PEITZMEIER, F.K., and HOY, C.E. (1987). An alternative method of controlling current population survey estimates of population counts. *Survey Methodology*, 13, 173-182.

COX, L.H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, 82, 520-524.

COX, L.H., and ERNST, L.R. (1982). Controlled rounding. *INFOR*, 20, 423-432.

DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

DEMING, W.E., and STEPHAN, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

DARROCH, J.N., and RATCLIFF, D. (1972). Generalized iterative sealing for log-linear models. *Annals of Mathematical Statistics*, 43, 1470-1480.

EL-BADRY, M.A., and STEPHAN, F.F. (1985). On adjusting sample tabulations to census counts. *Journal of the American Statistical Association*, 50, 738-762.

FAGAN, J.T., GREENBERG, B.V., and HEMMIG, B. (1988). Controlled rounding of three dimensional tables. Statistical Research Division Report Census/SRD/RR-88/02. U.S. Bureau of the Census, Washington, D.C.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhyā C*, 37, 117-132.

FULLER, W.A., KENNEDY, W., SCHNELL, D., SULLIVAN, G., and PARK, H.J. (1986). PC CARP. Statistical Laboratory, Iowa State University, Ames Iowa.

GOEBEL, J.J. (1976). Application of an iterative regression technique to a national potential cropland survey. *Proceedings of the Social Statistics Section, American Statistical Association*, 350-353.

HAMPEL, F.R. (1978). Optimally bounding the gross-error-sensitivity and the influence of position in factor space. *Proceedings of the Statistical Computing Section, American Statistical Association*, 59-64.

HIDIROGLOU, M.A. (1974). Estimation of regression parameters for finite populations. Unpublished Ph.D. thesis, Iowa State University, Ames, Iowa.

HIDIROGLOU, M.A., FULLER, W.A., and HICKMAN, R.D. (1976). SUPER CARP, Statistical Laboratory, Iowa State University, Ames, Iowa.

HUANG, E.T. (1978). Nonnegative regression estimation for sample survey data. Unpublished Ph. D. thesis. Iowa State University, Ames, Iowa.

HUANG, E.T., and FULLER, W.A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Social Statistics Section, American Statistical Association 1978*. 300-303.

HULLIGER, B. (1993). Robustification of the Horvitz-Thompson estimator. Contributed paper 49th Session of the International Statistical Institute. Book 1, 583-584.

HUSAIN, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.

IRELAND, C.T., and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 169-188.

JESSEN, R.J. (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Experiment Station Research Bulletin, 304.

KRASKER, W.A. (1980). Estimation in linear regression models with disparate data points. *Econometrica*, 48, 1333-1346.

LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-207.

LITTLE, R.J.A., and RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

LUERY, D. (1986). Weighting survey data under linear constraints on the weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 325-330.

MALLOWS, C.L. (1983). Discussion of Huber: Mimimax aspects of bounded-influence regression. *Journal of the American Statistical Association*, 78, 77.

MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.

OH, H.L, and SCHEUREN, F. (1987). Modified raking ratio estimation. *Survey Methodology*, 13, 209-219.

RAO, J.N.K. (1992). Estimating totals and distribution functions using auxiliary information at the estimation stage. Paper presented at the Workshop on Users of Auxiliary Information in Surveys, Örebo, Sweden, October, 1992.

ROYALL, R.M., and CUMBERLAND, W.G. (1981). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76, 924-930.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

SAS INSTITUTE INC. (1989). SAS/STAT User's Guide, Version 6 Fourth Edition, Volume 1. Cary, NC: SAS Institute Inc.

SINGH, A.C. (1993). On weight adjustment in survey sampling. Unpublished manuscript. Statistics Canada, Ottawa, Canada.

STEPHAN, F.F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13, 166-178.

WATSON, D.J. (1937). The estimation of leaf areas. *Journal of Agricultural Science*, 27, 474-483.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

ZIESCHANG, K.D. (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.