REGRESSION WEIGHTING FOR THE 1987—1988 NATIONWIDE FOOD CONSUMPTION SURVEY

Marie M. Loughin, Wayne A. Fuller, and Harold D. Baker
Marie M. Loughin, Iowa State University, Ames, IA 50011

Key Words: Nonresponse, auxiliary information

1. Introduction

The 1987—88 Nationwide Food Consumption Survey was conducted by the Human Nutrition Information Service of the U.S. Department of Agriculture. The purpose of the household portion of the NFCS was to estimate the kinds and amounts of foods used by households in the United States. 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. Within sample segments, personal interviewing was used to collect household data. 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 4495 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 were furnished by the contractor. In our analysis, we treat the estimated population totals as if they were known population totals.

2. Regression Weighting

Regression estimation for survey samples was introduced by Cochran (1942) and Jessen (1942). Cochran (1977, Ch. 7) contains the basic theory. When a vector of population means (or totals) is known, the regression estimator of the mean for a simple random sample of n observations is

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

where

$$w_i = \bar{\mathbf{X}} \left[\sum_{j=1}^{n} \mathbf{x}_j' \mathbf{x}_j \right]^{-1} \mathbf{x}_i' , \qquad (2)$$

$\mathbf{x}_j$ is the k—dimensional vector of control variables, $\bar{\mathbf{X}}$ is the row vector of population means of the control variables, the first element of $\mathbf{x}_j$ is always one, and the first element of $\bar{\mathbf{X}}$ is one. The weights have the property

$$\sum_{i=1}^{n} w_i x_{ij} = \bar{X}_j , \quad j = 1, 2, ..., k , \qquad (3)$$

where $\bar{X}_j$ is the population mean of the j—th control variable. An estimator of the variance of $\bar{y}_r$ is

$$\hat{V}_r\{\bar{y}_r\} = n(n-k)^{-1} \sum_{i=1}^{n} w_i^2 (y_i - \mathbf{x}_i \hat{\beta})^2 , \qquad (4)$$

where the finite correction term is omitted. The estimator (4) was suggested in Hidiroglou, Fuller and Hickman (1976) and the consistency of the estimator was established by Fuller (1975). Also see Royall (1981), Wright (1983), and Särndal, Swensson and Wretman (1989).

The construction of the weights is easily extended to samples for which initial unequal weights provide unbiased estimators. Let $\pi_1, \pi_2, ... \pi_n$ be proportional to the selection probabilities for a sample of size n. A regression weight for mean estimation is

$$w_i = \bar{\mathbf{X}} \left[\sum_{t=1}^{n} \pi_t^{-1} \mathbf{x}_t' \mathbf{x}_t \right]^{-1} \pi_i^{-1} \mathbf{x}_i' , \qquad (5)$$

where we assume that the matrix $\sum_{t=1}^{n} \pi_t^{-1} \mathbf{x}_t' \mathbf{x}_t$ is nonsingular. The weights of (5) minimize

$\sum_{i=1}^{n} \pi_i w_i^2$ subject to the restrictions (3). Thus, the weights w_i are as "close" to the initial weights as is possible under the restrictions. The weights of expression (5) are relatively easy to compute and, once computed, can be used for estimation of any quantity. Several authors have discussed the construction of regression weights. Recent discussions include Bethlehem and Keller (1987) and Lemaitre and Dufour (1987), Copeland, Peitzmeier and Hoy (1987), and Deville and Särndal (1990).

The weights of (5) provide estimators with desirable large sample behavior. However, they may have undesirable properties in small samples. Because the weights are linear functions of the control variables, 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) and Huang (1988).

The computer algorithm of Huang (1978) is an iterative procedure based upon the ideas of generalized least squares. If the first—round weights fall outside the desired range, then a second round of calculation is completed in which relatively small control weights are assigned to observations that are far from the mean of $\mathbf{x}$ and relatively large control weights are assigned to observations that are close to the mean of $\mathbf{x}$. This type of control weighting has much in common with procedures that are now known as 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 in the area for survey samples is that of Deville and Särndal (1990), and Akkerboom, Sikkel, and van Herk (1991).

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 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 (3) will generally no longer hold exactly. To construct integer weights satisfying (3) exactly would require solving an integer programming problem. 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 (3).

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 of reducing the bias. The degree to which the bias is reduced depends upon the relationship between the control variables and the variable of interest. See Little and Rubin (1987) for a general discussion of this issue.

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—88 Nationwide Food Consumption Survey, the response rate was lower than expected among high—income households. Therefore, use of this variable in a regression estimator is expected to reduce the bias in estimated characteristics that are correlated with income. However, one cannot guarantee that all bias has been removed by regression estimation.

If the sample is a probability sample, the use of regression estimation will generally reduce the variance of the estimated mean of y, and this is the reason for using the estimator. Furthermore, if the item of interest is not correlated with the control variables, the possible increase in variance is of small order.

In samples that are unbalanced because of nonresponse, it is possible for the variance of the regression estimator to be larger than the variance of the simple estimator by a term that is order n^{-1}.

Formula (4) identifies the two effects of regression estimation on the variance of an estimated mean for an original simple random sample. If the y variable is correlated with $\mathbf{x}$, the correlation tends to reduce the variance of the regression estimator relative to the simple estimator because

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

The second effect is through the weights. 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. 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.

3. Application to Nationwide Food Consumption Survey

Fifteen characteristics were selected by the Human Nutrition Information Service for use in generating regression weights. These characteristics were season of interview, region, urbanization, household income as a percent of poverty, household receives food stamps, ownership of domicile, race of household head, age of household head, household head status, female head of household worked, exactly one adult in household, exactly two adults in household, presence of child < 7 years old, presence of child 7—17 years old, and household size. These characteristics were chosen because the information was gathered by the questionnaire, the population estimates are available, and they are considered to be related to eating habits.

Population and sample percents differed significantly for several of these characteristics. Although an attempt was made to distribute interviews evenly over the year, the original sample was unbalanced with respect to season 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 than the population (24% versus 31%), and a higher fraction of nonmetropolitan households than the population (29% versus 23%).

The fraction of high income households was smaller in the sample than in the population. The sample percent with income > 500% of poverty was 17.4, compared to 21.8% of the population. The sample contained a higher fraction of households with both a male and female head than the population (68% versus 61%). The sample was mildly unbalanced with respect to several other socio—demographic characteristics.

Because the characteristics above are believed to be related to food consumption behavior, the regression weighting procedure was used to bring the sample into balance. To implement the weight generation program, each of the categorical variables with k classes was converted to a set of $k - 1$ indicator variables. For example, three variables were created for the four—category characteristic, household income as a percent of poverty. If the household is in category i, then the value of the corresponding indicator variable is one. Otherwise, the indicator variable is set to zero. 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 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. The initial weights, analogous to π_i^{-1}, used in the computation of the regression weights are all equal to the constant 19.787. The ratio of the largest weight to the smallest weight is less than $(0.1)^{-1}(1.9) = 19$.

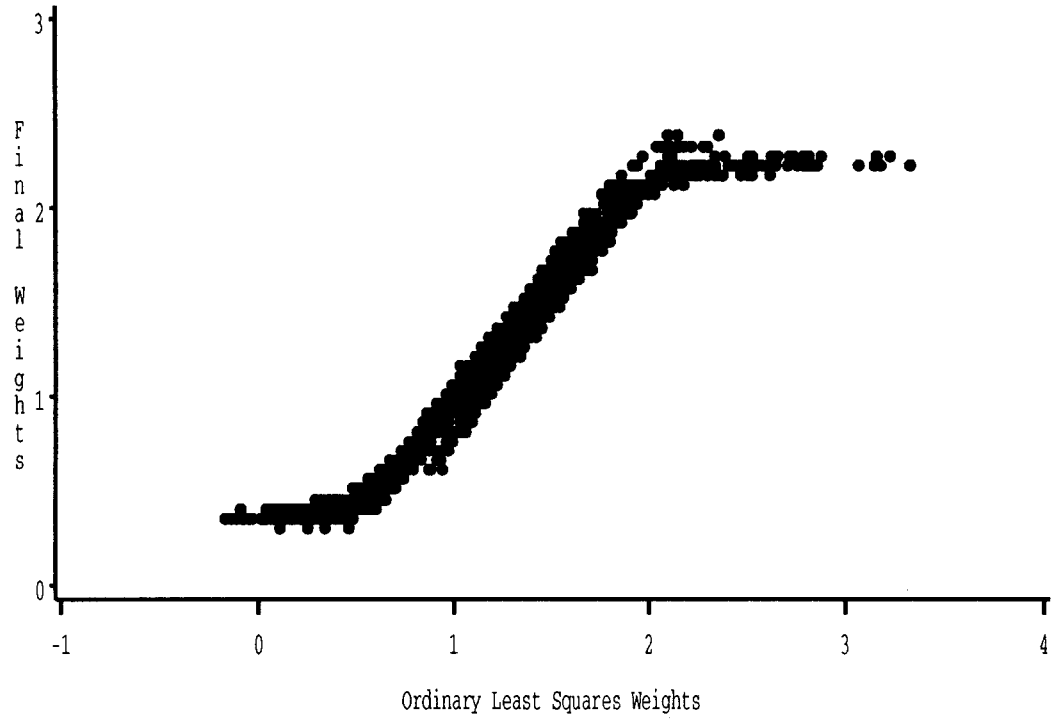

Figure 1. Plot of final weights against the ordinary least squares weights, both expressed relative to the average weight.

Figure 1 shows the regression weights computed by the Huang algorithm plotted against the ordinary least squares weights. 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.

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),

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.

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 unit price paid for comparable food by survey households in the same region and season.

The means of the variables, meals away and home food, computed using unweighted data, are given in Table 1 in the row headed, "Unweighted means." 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.

The row headed "Weighted mean" contains the estimates computed with the regression weights. The standard errors were computed in PC CARP using formula (4). The variance calculation requires computing a regression for every variable. The standard errors for unweighted and weighted

Table 1. Properties of alternative estimators.

Variable	Meals away	Home food
Unweighted mean	8.27 (0.22)	59.37 (1.12)
Weighted mean	8.57 (0.22)	57.49 (0.91)
Difference	—0.30 (0.12)	1.88 (0.39)
Relative efficiency of regression	2.56	5.60

estimates are similar for meals away and home food.

The estimated multiple correlations between the variables of the table and the 27 control variables are 0.29 and 0.44 for meals away and home food, 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 estimate is about 0.97 times the standard error of the unweighted estimate. The difference is due to the fact that $\Sigma_{i=1}^{n} w_i^2$ 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 $(1 - 0.29) = 0.71$ is the squared correlation and $1.32 = n\Sigma_{i=1}^{n} w_i^2$.

Table 1 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 and y is zero. Therefore, we computed the unweighted regression of y on w and computed the variance of the regression coefficient under the design using PC CARP. The standard errors for the difference in Table 1 are such that the "t—statistic" for the hypothesis of zero difference is equal to the "t—statistic" for the coefficient of w in the regression of y on w.

For both 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. We do not know that the regression estimator is unbiased, but it seems reasonable to assume that the regression adjustment reduces the bias in the estimators of the population mean.

The last row of Table 1 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

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

where $\hat{V}$ is the estimated variance of the unweighted estimate, Diff is the difference between the two estimates from Table 1, and s.e. diff is the standard error of the difference from Table 1. The second term of the estimated mean square error is the estimated squared bias. 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 1. Under the assumption that the regression estimator is unbiased, the expression for the estimated mean square error of the unweighted estimator is a consistent estimator.

The estimated relative efficiency of the regression estimator to the simple mean was 2.56 for meals away and 5.60 for home food. The regression estimator for meals away has the smaller 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. In both cases, the squared bias is a very important component of the estimated mean square error.

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.

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.

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.

Bethlehem, J. C. and Keller, W. A. (1987), Linear weighting of sample survey data. Journal of Official Statistics **3**, 141—153.

Bureau of the Census (1987), Current Population Survey, March 1987: Technical Documentation. Washington, DC.

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. Wiley, New York.

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.

Deville, J. C. and Särndal, C. E. (1990). Calibration estimators in survey sampling. Journal of the American Statistical Association **87**, 376—382.

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, ASA, Part 1:350—353.

Hampel, F. R. (1978), Optimally bounding the gross—error—sensitivity and the influence of position in factor space. Proceedings of the ASA Statistical Computing Section, American Statistical Association, Washington, D.C., 59—64.

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. (1988), A regression method to adjust census count. Unpublished manuscript. U.S. Bureau of the Census, Washington, D.C.

Huang, E. T. and Fuller, W. A. (1978), Nonnegative regression estimation for survey data. Proceedings of the Social Statistics Section of the American Statistical Association 1978. Washington, D.C. 300—303.

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

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.

Lemaitre, 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. Wiley, New York.

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

Royall, R. M. (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.

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