APPROXIMATING THE VARIANCE OF THE SURVEY REGRESSION ESTIMATOR USING POSTSTRATIFICATION

Kelli A. Leonard, Providian Bankcorp,
Anthony B. An, Sarah M. Nusser, and F. Jay Breidt, Iowa State University
Sarah M. Nusser, Department of Statistics, Iowa State University, Ames, Iowa 50011

Key Words: Variance Approximations, Regression Estimation, Poststratification

INTRODUCTION

Regression estimation is used to include auxiliary information from sample units in the estimation of population parameters from survey data. Known population totals or means for auxiliary variables are incorporated into the estimation of weights to insure that the weighted data properly reflect population characteristics.

Although regression estimation allows the use of large numbers of continuous and discrete control variables, variance estimation can be cumbersome because it involves two estimation passes for each variable. First, the regression coefficients are estimated; second, the variance of the estimated mean is calculated from the residuals. Computational effort can be excessive when analyses are required for several variables and subpopulations because separate multiple regressions must be run for each analysis variable/subpopulation combination.

Variance calculations can be simplified by using an approximation to the regression variance estimator. One approach is to use poststratification, which provides a piecewise constant approximation to the regression function used to define the weights. The variance estimator for the poststratification estimator of the mean requires only one pass through the data to calculate the estimated variance for all analysis variables and subpopulations, and variances for several dependent variables can be estimated in a single run. A related idea is developed in Relles (1981).

An efficient method of calculating variances for regression estimates is especially useful for large surveys containing many variables and several subpopulations, such as the 1989-1991 Continuing Surveys of Food Intake of Individuals (CSFII) conducted by the USDA Human Nutrition Information Service. The survey is used to develop policies relating to food production and marketing, food safety, food assistance, and nutrition education. The 1989-1991 CSFII database contains approximately 40,000 observations at the person-day level. The report generated from these data includes tables of estimated means, standard errors, and selected percentiles for 15-20 dietary components and 30 sex/age groups. Calculating regression standard errors for these tables is extremely time consuming due to the large number of dependent variables and subpopulations. The poststratified variance estimator is considered as a potential approach to simplifying calculations for the estimated standard errors included in these tables.

The performance of three different poststratification schemes was investigated using the 1989-1991 CSFII data. The standard errors of each of the three poststratification estimators were compared to that of the regression estimator to judge the performance of the approximations for eight sex/age groups and six dietary components. This paper describes the three poststratification methods and summarizes the results of analyses indicating their relative performance.

1989-1991 CSFII DATA

The 1989-1991 CSFII was conducted by the USDA's Human Nutrition Information Service. Separate studies were conducted during each of the three years. While the emphasis of the study objectives shifted somewhat from year to year, data collection methods remained constant. Each year, the CSFII consisted of two independent samples, a basic (all income) sample and a low income sample. A stratified sample of area primary sampling units (PSUs) was selected from the 48 conterminous states. Sixty strata were defined based on geographic, urbanization, and population density considerations. Census area segments were systematically sampled within each PSU, and households were systematically selected from the segments. Data were collected for the household and for each individual in the selected household, including three consecutive 24-hour dietary intake records.

In this paper, we analyze a subset of the 1989-1991 CSFII data containing the 11,912 individuals who provided three complete days of dietary intake data. To compare the variance estimators, we focus on four age groups for each sex and six dietary components. The age groups are less than one year

old, 30-39 years old, 80 years and older, 20 years and older. The dietary components are energy, iron, vitamin A, cholesterol, fiber, and zinc. These categories were selected to provide a broad range of sample sizes and intake behaviors.

REGRESSION ESTIMATION FOR THE CSFII

Regression methods were used to estimate individual weights for the combined 1989-1991 three-day intake data (An, McVey, and Fuller, 1994). Weights were calculated separately for the three sex/age groups of males 20 and older, females 20 and older, and individuals younger than 20. The control variables used in the weight procedure were indicators for geographic regions, urbanization, income as percent of poverty level, presence of a child six years old or younger in the household, presence of a child 7 to 17 years old in the household, exactly one adult in the household, exactly two adults in the household, household received food stamps in last twelve months, ownership of domicile, race of the individual (Black, Non-Black), ethnicity of the individual (Hispanic, Non-Hispanic), age class of the individual, employment during previous week (female head employed last week for persons under 20), female head is younger than 40 and has no child less than 18 (for individuals 20 or older), day of week on which food intake was recorded, and quarter of interview (Jan. - Mar., Apr. - Jun., Jul. - Sep., Oct. - Dec; 12 quarters in the three-year period).

There were a total of 43 indicator variables for males 20 and older and females 20 and older, and a total of 40 indicator variables for individuals younger than 20. A modification of Huang and Fuller's (1978) procedure was applied to produce non-negative integer weights for each of the three sex/age groups such that

$$\sum_{i=1}^{n} w_i x_{ij} = \tau_{x_j}$$

for each of the control variables, where w_i is the weight for the i-th observation, x_{ij} is the value of the j-th control variable for observation i, n is the number of observations in the group, and τ_{x_j} is the population total for the j-th control variable for the group.

For stratified single-stage cluster sampling, the variance of the regression mean is

$$\hat{V}(\bar{y}_r) = \frac{n}{n-k}\sum_{i=1}^{L}\frac{n_i(1-\frac{n_i}{N_i})}{(n_i-1)}\sum_{j=1}^{n_i}(d_{ij.}-\bar{d}_{i.})^2, \quad (1)$$

where

$$d_{ij.} = \sum_{k=1}^{m_{ij}} w_{ijk}(y_{ijk} - \mathbf{x}'_{ijk}\hat{\boldsymbol{\beta}});$$

$$\bar{d}_{i.} = \frac{1}{n}\sum_{j=1}^{n_i} d_{ij.};$$

w_{ijk} is the regression weight, y_{ijk} is the analysis variable, and $\mathbf{x}_{ijk}$ is the vector of control variables for element k in cluster j of stratum i; $\hat{\boldsymbol{\beta}}$ is the estimated vector of regression coefficients; L is the number of strata; n is the total number of clusters in the sample; n_i and N_i are the number of clusters in the sample and the population, respectively, in stratum i; and m_{ij} is the number of elements in cluster j in stratum i (Fuller, Loughin, and Baker, 1991). See Cochran (1977) and Fuller (1975) for a more complete discussion of regression estimation.

POSTSTRATIFICATION

Poststratification involves grouping the sampling units into strata after the sample has been drawn, and is most commonly used to stratify the sample based on information that is not available until after the sample is collected. However, in this study, the poststrata are used to form a piecewise constant approximation to the weight function. The approximation is expected to perform well if the variables used to create the poststrata are correlated with the variables used in the regression weighting.

Three poststratification schemes are proposed. All three methods split the original data set into three sex/age groups: males 20 and older (3,381 individuals), females 20 and older (4800 individuals), and individuals younger than 20 (3,781 individuals). Subsequent stratification of the three sex/age groups vary with each scheme. The first poststratification method further splits the individuals by income categories, and then by the magnitude of the individual's weight. The second method uses a cluster algorithm to define substrata of the sex/age groups. A principal components

scheme is used for the third poststratification. The methods are described in detail below.

The first method of poststratifying is referred to as the *weight-based* method. Each of the three sex/age groups was split into four groups based on the household income as percent of poverty. The four income classes were those whose household income was less than 75% of poverty level, between 76% and 130% of poverty level, between 131% and 300% of poverty level, and greater than 300% of poverty level. The eight sex/age groups of males 20 and older and females 20 and older were further divided based on whether there was a child less than 17 in the household. Each of the resulting 20 groups was divided in two based on whether the regression weight was above or below the median of the individual regression weights. Forty poststrata were defined using this method.

In the second poststratification method, referred to as the *cluster* method, observations were clustered using the final regression weight and all regression control variables except those that were highly correlated with the regression weight (geographic division, quarter of interview, and day of week). The 26 clustering variables were standardized to zero mean and unit variance to give each variable equal influence in the clustering algorithm. Ward's clustering method was used to produce 16 clusters for each sex/age group, resulting in 48 poststrata. Ward's clustering method was selected because it produces clusters of approximately equal size (Johnson and Wichern, 1982).

The third poststratification scheme used principal component analysis (Johnson and Wichern, 1982) and is referred to as the *principal components* method. For each of the three sex/age groups, first principal component scores were calculated for each individual based on the same variables used in the cluster method. Then each sex/age group was split based on whether the individual's first principal component score was above or below the median score. The procedure of calculating principal component scores and splitting each new group into two smaller groups based on the median of the scores was repeated three times, producing 48 poststrata.

Using the notation in (1), the poststratified estimator of the mean of y is

$$\bar{y}_{post} = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w^{*}_{ijk} y_{ijk} \ , \qquad (2)$$

where w^{*}_{ijk} are the poststratified weights for the k-th element of cluster j of stratum i. The poststratified weights are defined by

$$w^{*}_{ijk} = \frac{U_{\cdots r}}{\hat{U}_{\cdots r}} w_{ijk} \ ,$$

where $U...r$ is the known population total number of elements in poststratum r, w_{ijk} is the original weight for element ijk, $\hat{U}_{\cdots r} = \sum_{i=1}^{L}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} w_{ijk} U_{ijkr}$, and $U_{ijkr} = 1$, if element ijk is in poststratum r, 0 otherwise. For our poststratification schemes, the $U_{\cdots r}$ are unknown. Because in this application control variables were used to define poststrata, $U_{\cdots r} \doteq \hat{U}_{\cdots r}$. Thus, the poststratified weights were set equal to the original weights $\left(w^{*}_{ijk} = w_{ijk}\right)$.

The poststratification variance estimator is similar to the regression variance estimator in (1), except that the mean of the poststratum containing element ijk replaces $\boldsymbol{x}'_{ijk}\hat{\boldsymbol{\beta}}$. See Fuller et al. (1986) for a more complete explanation.

RESULTS AND DISCUSSION

None of the three methods produced poststrata that had an excessively small or large number of individuals. The weight-based method and cluster method had close to the same amount of variation in poststratum counts, with the weight-based method poststratum counts ranging from 85 individuals to 558 individuals and the cluster method poststratum counts ranging from 89 individuals to 573 individuals. The range for the principal components method poststratum counts was 211 individuals to 300 individuals.

The relationship between the poststrata and the control variables used in the poststratification methods was investigated by regressing each control variable on indicator variables for the poststrata. The regression weight was also regressed on the poststrata indicator variables. The R^2 values from these regressions are presented in Table 1. The degree of correlation between the regression weight and poststratum indicators varies as expected. The weight-based procedure produces the strongest relationship, and clustering, which involves less restrictive splitting algorithms than principal

Table 1. R^2 values obtained from regressing each control variable and the regression weight on poststratum indicator variables for each of the three poststratification methods.

Weight regression variable	Weight-based Method	Cluster Method	PCA Method
Age:			
Under 5	0.23	0.47	0.78
5-9	0.20	0.44	0.53
10-14	0.18	0.49	0.68
15-19	0.16	0.65	0.61
20-24	0.06	0.69	0.18
25-39	0.31	0.54	0.65
40-59	0.15	0.48	0.50
60-69	0.13	0.86	0.44
Ethnicity (Hispanic, Non-Hispanic)	0.05	0.59	0.08
Female head < 40 and no child < 18 in household	0.14	1.00	0.39
Household received food stamps in last 12 months	0.38	0.37	0.53
Income as percent of poverty level:			
0-75%	1.00	0.78	0.65
76-100%	0.44	0.77	0.22
101-130%	0.46	0.72	0.17
131-300%	1.00	0.62	0.36
301-500%	0.58	0.66	0.30
Ownership of domicile	0.22	0.26	0.41
Person employed last week	0.20	0.20	0.34
Exactly 1 adult in household	0.17	0.47	0.58
Exactly 2 adults in household	0.10	0.26	0.50
Presence of child ≤ 6 in household	0.38	0.35	0.50
Presence of child 7-17 in household	0.52	0.32	0.44
Race (Black, Non-Black)	0.12	0.35	0.20
Urbanization:			
Central cities	0.05	0.09	0.24
Suburban areas	0.07	0.08	0.25
Average R^2(excluding final regression weight)	0.29	0.50	0.42
Standard deviation	0.26	0.24	0.18
Regression weight	0.64	0.40	0.29

Table 2. Estimated mean, standard error for the regression estimated mean, and ratios of poststratified standard error estimates to the regression standard error estimate for each dietary component and sex/age category.

Sex	Age	n	Estimated Mean	Regression Standard Error	Weight-based Ratio	Cluster Ratio	PCA Ratio
Energy (kcal)							
F	<1	95	519.76	79.38	1.01	1.01	0.99
F	30-39	1051	1581.97	24.54	1.02	1.00	1.02
F	80+	271	1366.33	49.38	1.01	1.04	1.01
F	20+	4800	1502.85	15.40	1.04	1.05	1.02
M	<1	107	670.68	43.60	0.97	0.98	0.98
M	30-39	784	2206.65	57.00	1.01	0.99	1.00
M	80+	102	1770.65	77.87	1.01	1.00	1.00
M	20+	3381	2142.39	30.75	1.04	0.95	1.03
Iron (mg)							
F	<1	95	10.72	1.67	1.03	1.04	1.04
F	30-39	1051	11.95	0.34	1.00	1.00	1.00
F	80+	271	11.51	0.55	0.98	1.00	0.98
F	20+	4800	11.66	0.15	1.07	1.07	1.07
M	<1	107	13.02	1.69	0.98	0.96	0.99
M	30-39	784	16.48	0.62	1.00	0.97	0.97
M	80+	102	14.29	0.95	1.02	1.01	1.01
M	20+	3381	16.09	0.26	1.04	0.96	1.00

components, does better than the principal components approach.

The weight-based method produces poststrata that correlate well with the income categories, in part because they are part of the splitting procedure, but the poststrata are not highly correlated with other control variables. The cluster poststrata are most strongly correlated with individual control variables, particularly income and age categories. Principal components poststrata are moderately well associated with age categories and household composition, rather than with income.

Ratios of the estimated standard errors for the three poststratified estimators relative to the regression estimator are presented in Table 2 for energy and iron, which are representative of patterns exhibited by the other four dietary components. The ratio of the poststratified estimator to the regression estimator is very close to one in all cases, ranging from 0.95 to 1.07. There were no apparent systematic patterns across poststratification methods, sex/age groups, or dietary components, although there appears to be a slight increase in bias as the subgroup size increases for the weight-based method.

More variation among poststratification methods was expected since the correlation between the regression weight or control variables and the poststratum indicators varied considerably across methods. It is possible that regression estimation is not providing much improvement in estimation for these subgroups and these dietary components. This can occur if the regressions that define the weights do not adequately reflect relationships for the smaller sex/age groups in this study, or if intakes for the dietary components used in this study are not highly correlated with the regression variables. A comparison of results using a simple variance estimator with the regression estimator would indicate whether regression estimation is providing much gain in precision. Also, variance estimation for variables that are more highly correlated with control variables will provide a better comparison among the variance estimation approaches.

Because of the large size of the database, it is useful to consider computational difficulties. The principal component analysis method required 16 passes through the principal component procedure for each of the three sex/age groups, which was time consuming and required considerable computer memory and disk space to store the intermediate data sets. The cluster method required only one pass through the clustering procedure for each of three sex/age groups. However, cluster algorithms are computationally intensive and the total time for

creating the poststrata using the clustering method far exceeded the time required to create poststrata using the principal components method. Assigning cluster identification numbers after completing the cluster algorithm required another pass through the TREE procedure in SAS. The weight-based method required only data manipulation in SAS. Although many intermediate data sets were created, this method was the most straightforward to implement and required the least amount of time to create the poststrata.

ACKNOWLEDGEMENT

We thank Dan McCaffrey of the Rand Corporation for bringing the Relles (1981) reference to our attention.

REFERENCES

An, A. B., McVey, A. M., and Fuller, W. A. (1994), Construction of regression weights for the 1989-1991 CSFII data. Progress report for the Human Nutrition Information Service, U.S. Department of Agriculture. Department of Statistics, Iowa State University, Ames, Iowa.

Cochran, W. G. (1977), *Sampling Techniques*, 3rd ed. Wiley, New York.

Fuller, W. A. (1975), Regression analysis for sample survey. *Sankhyá 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.

Fuller, W. A., Loughin, M. M., Baker, H. D. (1991), Regression weighting for the 1987-1988 Nationwide Food Consumption Survey. Report to the Human Nutrition Service, U.S. Department of Agriculture. 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 of the American Statistical Association 1978*. Washington, D.C. 300-303.

Johnson, R. A. and Wichern, D. W. (1982), *Applied Multivariate Statistical Analyses*, Prentice-Hall, New Jersey.

Relles, D. A. (1981), Using weights to estimate population parameters from survey records. N-1136-HUD. The Rand Corporation, Santa Monica, California.