

WEIGHTING SAMPLE DATA SUBJECT TO INDEPENDENT CONTROLS

Cary T. Isaki¹, Julie H. Tsay² and Wayne A. Fuller³

ABSTRACT

In the U.S. Census of Population and Housing a fraction, on the order of one-in-six, of the households receives a longer version of the census questionnaire called the long form. All others receive a version called the short form. In the past, raking and selected control totals from the short form have been used to create two sets of weights for long form estimation; one for individuals and one for households. We describe a weight construction method based on quadratic programming that produces household weights such that the weighted sum for individual characteristics and household characteristics match closely selected short form totals.

KEY WORDS: Raking; Regression; Quadratic programming; Weighting area.

RÉSUMÉ

Dans le recensement de la population et des ménages aux États-Unis, une fraction, de l'ordre d'un sur six, reçoit une version plus longue du questionnaire de recensement appelé le formulaire long. Tous les autres reçoivent une version appelée le formulaire court. Dans le passé, le ratissage et les totaux de contrôle choisis du formulaire court ont été utilisés pour créer deux ensembles de pondérations pour les estimateurs du formulaire long, un pour les individus et un pour les ménages. Nous décrivons une méthode de construction de la pondération basée sur la programmation quadratique qui produit des poids pour les ménages tels que la somme pondérée pour les caractéristiques individuelles et pour les caractéristiques de ménages s'apparentent étroitement aux totaux choisis du formulaire court.

MOTS-CLÉS : Ratissage; régression; programmation quadratique; région de pondération.

1.1. Introduction

Given the availability of known characteristic totals, it is common among survey practitioners to use such information in estimators of the post stratified, ratio and regression type. The known characteristic totals are sometimes called independent controls because they are derived outside of the survey situation. Use of such independent controls tends to reduce the variance of most estimates. Independent controls also often partially compensate for coverage problems in surveys.

National periodic surveys generally use independent controls. For example, in the U.S., the official monthly labor force survey uses person characteristic totals based on administrative records to construct sample weights. An annual survey of manufacturers utilizes previous census data in a difference estimator to estimate current annual totals.

The U.S. decennial census utilizes a sampling and estimation procedure for the measurement of some characteristics. The questionnaire for detailed characteristics is called the long form. The long form sample consists of a random sample of addresses. The long form questionnaire requests information that is asked of all individuals (called short form information) as well as information on additional characteristics. In previous Censuses weights were derived by raking initially weighted sample responses to various independent controls based on short form information. Two sets of sample weights were derived, one for person characteristics and one for housing unit characteristics.

One set of categories of short form information that was used for person weighting was a classification of individuals by race, Hispanic origin, age and sex; another person classification was by family type and household size. For households, categories include the

¹ Cary T. Isaki, Statistical Research Division, Washington, D.C. 20233, Cary T. Isaki@ccmail.census.gov.

² Julie H. Tsay, Statistical Research Division, Washington, D.C. 20233, Julie H. Tsay@ccmail.census.gov

³ Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA 50010, waf@iastate.edu.

cross classification, race by Hispanic origin of householder by tenure. In the 1990 Census long form weighting process, initially weighted sample counts of persons and of housing units were each classified by four sets of classifications. Raking in four dimensions were used to construct weights. When raking was completed, long form sample weights were converted to integers. For more details, see Schindler, et. al. (1992).

Our objective is to present a method of obtaining a single set of household sample weights for the long form sample that maintains both person and household controls.

1.2 Some Long Form Weighting Methods

1.2 A. U.S. 2000 Census

Two possibilities exist for the U.S. 2000 Census. One possibility is that the 2000 Census situation will be essentially like the 1990 Census situation. That is, the independent controls are tabulated from the short form for Census respondents.

The other possibility for the 2000 Census is to provide two sets of numbers. The two sets of controls would come from the short form Census totals and from the estimates from the post enumeration survey, called the Accuracy and Coverage Evaluation (ACE) survey. The set of numbers from the Census will be used to determine the number of representatives in each of the 50 states. The ACE results will be used to produce a second set of numbers to be used for all other purposes. Such purposes include re-districting of political boundaries, revenue sharing, and public use data.

Under the ACE situation, the current plans for the 2000 Census call for long form weights being developed in two stages. The first stage is the raking method using short form data as controls, except that no integer rounding of weights is attempted. The second stage is to use the ACE survey to develop adjustment factors (ratios of ACE totals to enumeration totals) for various person and housing unit categories. The products of the raked weights and the adjustment factors are integer rounded to complete the operation. Separate weights for persons and for housing units are constructed. The ACE survey is designed to estimate person characteristics only. A separate operation is required to estimate household totals. In the following, we consider using controls based on Census short form information only.

1.2 B. Statistics Canada

Long form weighting in the 100% information situation is a part of the Canadian Census of population and housing which is carried out every five years. Canada adjusts the long form to the short form Census data. Unlike the U.S., the Canadians seek a single set of household weights subject to constraints (independent controls). Their method is essentially regression estimation with the additional aim of providing good estimates of smaller geographic areas contained within the area of interest. See Bankier, et. al. (1997).

Regression weights are not always positive unless special efforts are made. See Huang and Fuller (1978). If the regression weights in the Canadian procedure exceed prescribed bounds, collapsing of cells defining explanatory variables is also carried out. Near linear dependencies among the explanatory variables may require use of methods to identify the number of linear dependencies, the variables involved and their elimination. See Bankier, et. al. (1992). Deville and Sarndal (1992) tie the regression and raking approaches together under the heading of calibration estimators.

1.2 C. Our Proposed Method - Quadratic Programming

Our proposed long form weighting method is regression based and, like the Statistics Canada approach, provides a single set of household weights that maintain given independent controls. We solve for household weights using quadratic programming with the restrictions that the weights fall within an acceptable range and maintain control totals. The household weights maintain the area control totals used in the U.S. Census. In the following, we call our method the quadratic programming method or QP. Apart from the specified ranges for acceptable weights, the QP and regression methods provide the same results.

1.3 The Quadratic Programming Method (QP)

1.3 A. An Earlier Application

An earlier application of quadratic programming (QP) in a Census environment can be found in Isaki et. al. (1999) where household weights for 100% Census households were obtained using person totals as controls. In that application, the person totals were obtained via a post enumeration survey called the Integrated Coverage

Measurement (ICM) survey. The household weights were interpretable as adjustment factors to adjust for the undercount / overcount of the persons in the Census. That application involved the creation of a Census data file, called a transparent file. Rather than integer household weights, the objective was to provide integer estimates of households of the required type.

Finding household weights in the 100% Census situation is roughly six times as large of a problem as constructing weights for the long form sample. Consequently, in the 100% Census situation households were grouped by category and the weights obtained by QP were the same for each household in the same household category.

1.3 B. Long Form Sample Weights - QP

We first introduce the mathematical setting for the QP application and then discuss the details.

Let

- i) $\{W_i\}$ denote the set of housing unit weights where i denotes the i^{th} long form sample household
- ii) $\{W_i^{(2)}\}$ denote the set of initial housing unit weights
- iii) X_j denote the j^{th} person control
- iv) Y_j denote the j^{th} household control
- v) K denote a positive integer to be specified by the user
- vi) A denote the long form sample

The quadratic programming method seeks

$W_i, i = 1, 2, \dots$ that minimize the function

$$f = \sum_{i \in A} (W_i - [W_i^{(2)}])^2 [W_i^{(2)}]^{-1}$$

subject to

- i) $\sum_{i \in A} W_i X_{ij} = X_j$, for $j = 1, 2, \dots$
- ii) $\sum_{i \in A} W_i Y_{ij} = Y_j$, for $j = 1, 2, \dots$
- iii) $1 \leq W_i \leq K$,

where the summations are over housing units in the long form sample.

1.3 C. Details of the QP Method

1.3 C. 1 Controls Used in the QP Method in the 100% Census Situation

The person control (X_j) categories include a cross classification of age and sex and race / ethnicity. Other characteristics such as number of renters, number of males, number of persons by broad age groups were used as additional controls.

The majority of the household control (Y_j) categories are defined by a cross classification of household type (eg., family with children under 18) and household size (number of persons in the family). The Y_j also include race / ethnicity of the householder cross classified by tenure.

1.3 C. 2 Integer Rounding of W_i

The last step in long form weighting is to round the W_i to integers. For our work, we used the cumulate and round method by grouping sample housing units by race / ethnicity of the householder and tenure. Then within each such group, we sorted the sample by family type by household size and conducted integer rounding. In this manner, 100% Census housing unit counts by race by tenure were maintained.

2. Numerical Results for the 100% Census Situation

We used the Bureau's 1990 Census data, located in a file called CenSAS, to illustrate the application of the QP method on actual data. The CenSAS provides household data and data for persons in households together with long form weights, as developed for the 1990 Census. The data are in SAS format. Hence, the CenSAS data file provided a data source appropriate for numerically comparing the performance of the Bureau's 1990 long form weighting method with the QP method.

Long form sample weighting is done by weighting area, where the weighting areas contain, on the average, about two to three thousand housing units. Hence, much work had already been completed by the time the raking process was started. There were about 56,000 weighting areas in 1990. With so many weighting areas, it is desirable to use a weighting method that is fast and automatic in execution. The execution time for QP ranged from one to three minutes on a Sun Unix machine, depending on the sample size.

For our numerical work we chose three weighting areas of various sizes of occupied housing units from the Houston, TX metropolitan area all of which exhibited similar results. We exhibit results for WA 1788 which contained 8034 occupied housing units (25,145 persons).

In the QP method, only occupied housing units are considered. For vacant housing units ratio estimation to the Census vacant totals or to estimated vacant totals via ACE is suggested. Housing units in WA 1788 were either occupied or vacant.

The table below provides the 100% Census counts in the first column. Estimates using the Bureau's long form weights are given in the column headed CenSAS. In CenSAS, the person based weights are used for person characteristics and the housing unit based weights are used for housing unit characteristics. The QP estimates are given in the third column and the estimated standard errors for the QP estimates are given in the last column. The estimates are presented as a percent of the 100% Census counts.

Table Comparison of 100% Census, CenSAS and QP Estimates for Housing Units and Persons in Long Form Weighting for WA 1788

<u>Characteristic</u>	<u>100% Census</u>	<u>Percent CenSAS</u> 100% Census	<u>Percent QP</u> 100% Census	<u>Estimated Standard Error of Percent QP/100% Census</u>
<u>1. Number of Housing Units</u>				
with own children	4349	100	100	.1
not with own children	3685	100	100	.2
with 1 to 4 persons	6785	100	100	.08
with 5+ persons	1249	100	100	.4
rented	2559	100	100	.08
owned	5475	100	100	.04
<u>2. Number of Persons</u>				
0 - 4 years	2493	98	99	2.4
5 - 17 years	6339	101	101	.8
18 - 44 years	12,711	100	100	.2
45 - 64 years	3028	102	100	1.9
65+ years	574	94	100	4.9
males	12,473	100	100	.4
females	12,672	100	100	.3
Hispanic	2385	103	100	1.2
Non Hispanic	22,760	100	100	.12
Black	1285	102	101	2.5
White	23,372	100	100	0.2
Asian	257	81	106	9.9
Remainder	1231	104	97	2.2
Renter	7978	95	100	.2
Owner	17,167	102	100	.08

For housing unit characteristics in WA 1788, the two sets of weights are equivalent, differing from the Census controls by rounding error.

The person categories 18-44, 45-64, males and renters were used as controls in the QP procedure. Hence any differences for those categories, should they occur, are due to rounding. In other categories the QP estimates are generally closer to the Census 100% values. The QP estimates based on a single set of weights are as close or closer to the 100% counts as the CenSAS household estimates based on household weights, and as close or closer to the 100% counts as the CenSAS person controls based on person weights. Here, we judge the QP procedure to be the preferred procedure.

A jackknife variance estimation procedure was used to estimate the variance of the QP based estimator. We used 16 replicates.

3. Conclusions

The QP method has been shown to work well on actual long form data sets. A single set of weights gave favorable results in comparison with the Bureau method used in previous censuses. The QP estimation module can replace the raking estimation module in the Census operational setting. Our experience has been that the program operates smoothly. The QP method is quite flexible. It can be used to provide long form sample weights for units other than occupied housing units such as group quarters (e.g., jails). It can also produce long form sample weights in an ACE situation with a limited number of controls.

4. References

1. Bankier, M.D., Rathwell, S. and Majkowski, M (1992), "Two Step Generalized Least Squares Estimation in the 1991 Canadian Census", Working Paper-Methodology Branch, Census Operations Section, Social Survey Methods Division, Statistics Canada, 24 pages.
2. Bankier, M., Houle, A.M., and Luc, M. (1997), "Calibration Estimation in the 1991 and 1996 Canadian Censuses", Statistics Canada, (draft), 8 pages.
3. Deville, J., and Sarndal, C. (1992), "Calibration Estimators in Survey Sampling", *JASA*, vol. 87, no. 418, pages 376-382.
4. Fuller, W.A. (1999), "Variance Estimation of Regression Estimator with ICM Controls", Iowa State University, (draft), 5 pages.
5. Huang, E.T. and Fuller, W. A. (1978), "Nonnegative Regression Estimation for Sample Survey Data", Proceedings of the Social Statistics Section, American Statistical Association, pages 300-305.
6. Isaki, C.T., Tsay, J.H., and Fuller, W.A. (1998) "Estimation of Census Adjustment Factors", submitted to Survey Methodology, 34 pages.
7. Isaki, C.T., Ikeda, M.M., Tsay, J.H., and Fuller, W.A. (1999), "An Estimation File that Incorporates Auxiliary Information", submitted to *Journal of Official Statistics*, 28 pages.
8. Isaki, C.T., and Ikeda, M.M. (1996), "Some Estimates of Housing Unit Category Totals in the 95 ICM Via Dual System Estimation and Census Plus". Unpublished manuscript dated 11/13/96, 30 pages. U.S. Bureau of the Census.
9. Schindler, E., Griffin, R., and Swan, C. (1992), "Weighting the 1990 Census Sample", Proceedings of the American Statistical Association, pp 664-669.