AN APPLICATION OF REGRESSION AND CALIBRATION ESTIMATION TO POST-STRATIFICATION IN A HOUSEHOLD SURVEY

Bodhini Jayasuriya, Richard Valliant, U.S. Bureau of Labor Statistics
Richard Valliant, 2 Massachusetts Av., N.E., Rm 4915, Washington, D.C. 20212

Key Words: General regression estimator, Principal person method, Replication variance

ABSTRACT

This paper empirically compares three estimation nethods—regression, calibration, and principal person—ised in a household survey for post-stratification. Post-tratification is important in many household surveys to djust for nonresponse and the population undercount that esults from frame deficiencies. The correction for opulation undercoverage is usually achieved by adjusting stimated people counts in each post-stratum to equal the orresponding population control counts typically available rom an external source such as a census. We will compare stimated means from the three methods and their stimated standard errors for a number of expenditures rom the Consumer Expenditure Survey sponsored by the Bureau of Labor Statistics in an attempt at understanding iow each estimation method accomplishes this step in ost-stratification.

1. INTRODUCTION

In large household surveys, post-stratification is a neans of reducing mean square errors by adjusting for lifferential response rates among population subgroups and rame deficiencies that often result in undercoverage of the arget population. In general, the population is subdivided nto groups (post-strata) at the estimation stage based on nformation that affect the response variables. The stimator is constructed in such a way that the estimated otal number of individuals falling into each post-stratum is qual to the true population count. Post-stratum opulation counts are typically available from an external ensus for numbers of persons but not for numbers of iouseholds. If household estimates are needed, a single veight must be assigned to each household while using the erson counts for post-stratification. Regression stimators of totals or means accomplish this by using erson counts in each household's auxiliary data. Calibration estimation, with a least-squares distance unction, is closely related to regression estimation but possibility that each person in a household ma different weight. The weight associated with the person is then assigned to the household. Thi method is difficult to analyze theoretically. The r estimator discussed in this paper, while easily adj the population under count, automatically pr household weight that is not based on any particu its members. Lemaître and Dufour (1987) Statistics Canada's use of the regression estimat regard.

There are a growing number of precedents fo of regression estimators in surveys both in the t literature and in actual survey practice. Statistic has incorporated the general regression estimato generalized estimation system (GES) software th used in many of its surveys. Fuller, Loughin a (1993) discuss an application to the USDA N Food Consumption Survey. One of the attra regression estimation is that many of the techniques in surveys including the post-stra estimator mentioned above are special cases of r estimators. It also more flexibly incorporates auxi than other more common methods. Other works regression estimation and post-stratification Bethlehem and Keller (1987), Casady and Vallia Deville and Särndal (1992), Deville, Särndal, and (1993), and Zieschang (1990).

In this study we compare the regression estim the PP estimator currently in use at the Bureau Statistics (BLS). The ordinary least-squares r estimator has the disadvantage that it can nonpositive weights. A number of ways are sug the literature on how to overcome this problem. the most flexible is the calibration method intro Deville and Särndal (1992) which can remove any weights as well as control extreme weight calibration estimators produced by these new we also compared to the original regression estimato PP estimator.

In Section 2, the three different estim presented. Section 3 is an application of these p to the Consumer Expenditure (CE) Survey at B same setting as in Zieschang (1990). We con

2. REGRESSION, CALIBRATION, AND PRINCIPAL PERSON ESTIMATION

First, we give a brief introduction to the regression stimator. A sample s of size n is selected from a finite opulation U of size N. Let the probability of selection of he i^{th} unit be π_i. The sample could be two-stage and the nit could be either the primary sampling unit or the econdary sampling unit. There is no need here to omplicate the notation with explicit subscripts for the lifferent stages of sampling. Let the variable of interest be lenoted by y and suppose that its value at the i^{th} unit, y_i, s observed for each $i \in s$. Assume the existence of K uxiliary variables $x_1, x_2, \ldots, x_K$ whose values at each $i \in s$ re available. Define $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{iK})'$, for each $i \in U$, vhere x_{ik} denotes the value of the variable x_k at unit i. et $\mathbf{X} = (X_1, \ldots, X_K)'$ denote the K-dimensional vector of nown population totals of the variables $x_1, x_2, \ldots, x_K$. The egression estimator is then motivated by the working nodel ξ:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i \qquad (2.1)$$

or $i = 1, \ldots, N$. Here, $\beta_1, \ldots, \beta_K$ are unknown model arameters. The ε_i are random errors with $E_\xi(\varepsilon_i) = 0$ and $\text{var}_\xi(\varepsilon_i) = \sigma_i^2$ for $i = 1, \ldots, N$. The term working model" is used to emphasize the fact that the nodel is likely to be wrong to some degree. In the CE, y_i night be the total food expenditures by the consumer unit CU) and the x_{ik}'s might be various CU characteristics like umbers of people of different ages, or CU income, that ave an effect on the CU's expenditure on food. The ariance of expenditures might be dependent on CU size so hat having σ_i^2 proportional to the number of persons in he CU might be reasonable. Then, a linear regression stimator of the population total of y is defined to be

$$\hat{y}_R = \hat{y}_\pi + \left(\mathbf{X} - \hat{\mathbf{x}}_\pi\right)' \hat{\beta} \qquad (2.2)$$

vhere $\hat{y}_\pi$ denotes the π-estimator (or Horvitz-Thompson stimator) of the population total of y, i.e., $\hat{y}_\pi = \sum_{i \in s} a_i y_i$, vhere $a_i = 1/\pi_i$. Also, $\hat{\mathbf{x}}_\pi = \left(\hat{x}_{1\pi}, \ldots, \hat{x}_{K\pi}\right)'$ is the vector f π-estimators of the population totals of the variables $x_1, x_2, \ldots, x_K$ and

$$\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_K\right)' = \left[\sum_{i \in s} \frac{a_i \mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2}\right]^{-1} \sum_{i \in s} \frac{a_i \mathbf{x}_i y_i}{\sigma_i^2}. \qquad (2.3)$$

Even if model (2.1) fails to some degree, $\hat{y}_R$ will still have easonable design-based properties because, even though

The regression estimator $\hat{y}_R$ can also be expres weighted sum of the sample y_i's with i th weight,

$$w_i = a_i \left[1 + \left(\mathbf{X} - \hat{\mathbf{x}}_\pi\right)' \left(\sum_{i \in s} \frac{a_i \mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2}\right)^{-1} \frac{\mathbf{x}_i}{\sigma_i^2}\right].$$

From (2.4) it is easily seen that the known populat are exactly reproduced for the auxiliary variables.

The estimator of $\boldsymbol{\beta}$ in (2.3) does not accour correlation among the errors in model (2.1). In populations, units that are geographically near ea e.g., CU's in the same neighborhood, may be c Using a full covariance matrix $\mathbf{V}$ may be mo optimal (e.g., see Casady and Valliant 1993 1992). Though use of a full covariance matri: lower the variance of $\hat{\beta}$, the elements of $\mathbf{V}$ will d the particular y being studied, and estimation generally a nuisance. Consequently, it is intere practical to consider the simple case of $\mathbf{V} = \text{diag}($ leads to (2.2). Note that when the desigr $\text{var}_p(\hat{y}_R)$ is estimated, it will be necessary to use that properly reflects clustering and othe complexities.

The regression estimator has the disadvantage weights can be unreasonably large, small or even The calibration estimators of Deville and Särnda introduced next, add constraints to restrict the si weights. Calibration estimators are formed by min given distance, F, between some initial weight final weight, subject to constraints. The constr involve the available auxiliary variables thus inco them into the estimator. The regression presented above is a special case of the c estimator in which F is defined to be the general squares (GLS) distance $F(w_i, a_i) = a_i c_i (w_i / a_i - 1)^2 / 2$ for $i = 1, \ldots, n$, known, positive weight (e.g., $c_i = \sigma_i^2$ or $c_i = 1$) a with unit i, and w_i, the final weight. The tota distance $\sum_{i \in s} F(w_i, a_i)$ is minimized subject constraints, $\sum_{i \in s} w_i \mathbf{x}_i = \mathbf{X}$. In this form, the weig regression estimator of the population total of y (2.4) can be written as,

$$w_i = a_i g(c_i^{-1} \boldsymbol{\lambda}' \mathbf{x}_i)$$

for $i = 1, \ldots, n$ where

$$g(u) = 1 + u,$$

for $u \in \Re$ and $\boldsymbol{\lambda}$ is a Lagrange multiplier evalua minimization process. The calibration weights c hosen in such a way as to reflect the desired restrictions n the weights. Choosing $L>0$ ensures that the weights re positive, and U is picked to be appropriately small to rohibit large weights. The calibration weights must be olved for iteratively; one easily programmed algorithm is iven in Stukel and Boyer (1992).

In most household surveys, post-stratification serves rimarily as an adjustment for undercoverage of the target opulation by the frame and the sample. In the U.S., there re no reliable population counts of households to use in ost-stratification. Consequently, population counts of ersons are used for the post-strata control totals. This isagreement in the unit of analysis (the household) and the nit of post-stratification (the person) when a household haracteristic is of interest led to the development of the P method that is used in the CE and Current Population urveys.

In the PP method described in Alexander (1987), a ousehold begins the weighting process with a single base veight, a_i, that is then adjusted for nonresponse. The djusted weight is assigned to each person in the household nd the person weights are then further adjusted to force hem to sum to known population controls of persons by ge, race, and sex. This last adjustment can result in ersons having different weights within the same ousehold. The household is then assigned the weight of he person designated as the "principal person" in the ousehold. This method has an element of arbitrariness nd is difficult to analyze mathematically. The regression nd calibration estimators can be formulated in such a way hat population person controls are satisfied, all persons in household retain the same weight, and no arbitrary choice mong person weights is needed to assign a household veight.

3. AN APPLICATION

We compare the three estimators (i.e., regression, estricted calibration (with $L=.5$, $U=4$), and principal erson) by an application to the estimated means and their stimated standard errors for a number of expenditures rom the CE Survey sponsored by the Bureau of Labor tatistics.

The CE Survey gathers information on the spending atterns and living costs of the American consumers. here are two parts to the survey, a quarterly interview and weekly diary survey. The Interview Survey collects etailed data on the types of expenditures which espondents can be expected to recall for a period of three

$n=5156$ CU's were used. The CE Survey's pri of analysis is the consumer unit, an economic fam a household. A consumer unit (CU) consists of i in the household who share expenditures. Thus, t be more than one CU in a household.

Five different sets of auxiliary variables were They were chosen by testing the adequacy of m for the selected expenditures with different combi the available auxiliary variables. The 56 post-str on age/race/sex currently in use in the CE were The combinations of auxiliaries used to form the weights are given in Table 1. The number of variables in each model is given within parenthese on this information, weights (2.5) were compute given in (2.6)—regwts—and (2.7)—calwts. For regression and restricted calibration weights, w equal to the adjusted base weight, i.e., $1/\pi_i$ nonresponse adjustment.

Table 1. Weights and their corresponding auxiliar variables. Number of cells are in parantheses.

Weight	Auxiliary Variables
regwts0	age/race/sex (56)
regwts1	inter., age/race/sex, region, urban×region (18)
regwts2	intercept, age/race/sex, region, urban×region, age of reference person, housing tenure, family income before taxes (24)
calwts0	age/race/sex (45)
calwts1	inter., age/race/sex, region, urban×region (18)
calwts2	intercept, age/race/sex, region, urban×region, age of reference person, housing tenure, family income before taxes (24)
calwts3	intercept, age/race/sex, region, urban×region, family income before taxes (truncated at \$500,000) (19)
calwts4	intercept, age/race/sex, region, urban×region, age of reference person, housing tenure (23)
PP	age/race/sex (56)

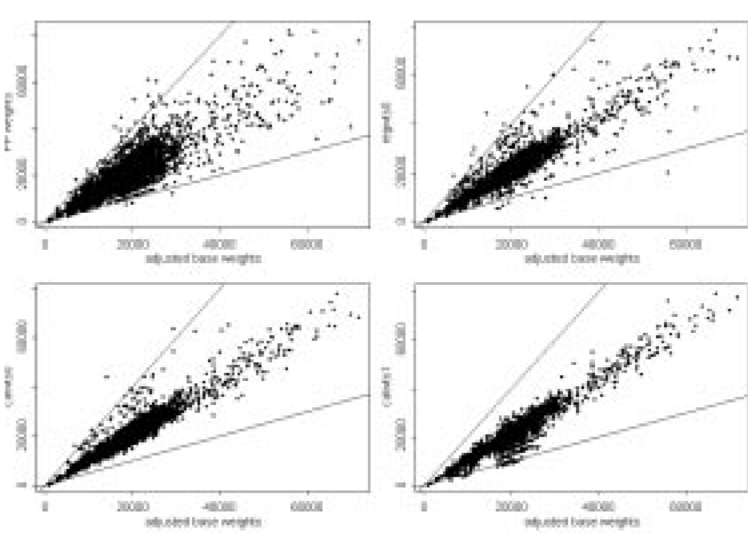
For this application, the population totals nec evaluate $\mathbf{X}=(X_1,\ldots,X_K)'$ were obtained mostly 1990 Census figures projected to 1992 and the Population Reports published by the U.S. Bure Census.

3.1 Comparisons of Weights and Estimated C Counts

eplicate weights, nearly half the sets for each of regwts0, egwts1 and regwts2 had some negative weights though he maximum number of negative weights for any replicate vas 3. The negatives are a potential cause of inflated tandard errors, since the negative weights will be offset by arge positive weights in order for the fixed population ontrol totals to be met in every replicate. Calwts, which estrict the deviation from the base weights by choosing and U appropriately, (in this instance, L=0.5 > 0) aturally did not produce any negative weights.

On examining scatter plots (not shown here) comparing ome of the different weights to each other, the PP and egwts0, while being substantially different from each ther, exhibited final weights that can be considerably lifferent from the adjusted base weights. The adjustments an be either up or down. A less variable set of djustments was apparent in regwts1, calwts0, and calwts1. alwts1 and calwts4 were quite similar and both were lose to regwts1. The two sets of weights that involve the uantitative variable family income before taxes, calwts2 nd calwts3, were closely related. Some CU's had calwts2 alues larger than 60,000 but had calwts0, calwts1, calwts4 30,000. These CU's all had family incomes before taxes f a quarter of a million dollars or more. Thus, the nclusion of that variable in the calibrations did have a ubstantial impact on some units. We did use a control nly on the grand total income; having controls by income lasses might have changed the weights on some of these ases.

igure 1. Four sets of weights plotted against adjusted ase weights. Reference lines correspond to L=.5 and J=2.



indicate that the PP weights and regwts0 do not to the restriction $a_i / 2 \le w_i \le 2a_i$.

Previous studies at BLS regarding ge regression estimation in the CE had concluded number of single person CU's was under estimate compared to the estimate produced by the PP met found minimal evidence of that phenomenon here indicated by the ratios shown in Table 2. It co ratio of the estimated number of CU's under the a procedures to that of the PP estimation procedur of CU.

Table 2. Estimated counts in thousands of CU's b size for PP weights and ratios of other estimated c the PP weights estimates. Ratios greater than 1.02 than 0.98 are highlighted.

Weights	CU Size				
	1	2	3	4	5+
PP	28,784	30,680	15,409	15,068	9,993
regwts0	0.96	0.99	1.01	0.99	1.02
regwts1	1.00	1.00	1.02	0.98	0.99
regwts2	1.00	1.01	1.01	1.01	0.97
calwts0	0.96	0.99	1.00	1.00	1.02
calwts1	1.00	1.00	1.02	0.98	0.99
calwts2	0.99	1.01	1.01	1.00	0.97
calwts3	0.98	1.02	1.01	1.01	0.96
calwts4	1.01	1.00	1.01	0.99	0.99

A similar table constructed by Composition of CU that while regwts0 and calwts0 estimato substantially different from PP for the category Or 1+ children, for single person CUs they were not.

3.2 Precision of Estimates from the Different M

Although comparison of weights is instruc methods must ultimately be judged based on the estimated CU means and their precision. The errors of these estimators were computed via th of balanced half sampling (BHS) using 44 rep currently implemented in the CE for the PP estim BHS estimator is constructed to reflect the stra and the clustering that is used in the CE. expenditure estimates from the CE Survey are for various domains of interest, we computed t and the standard errors for a few chosen domain For each of these, the coefficient of variation computed and then its ratio to the cv of the PP was calculated.

xpenditures, and for each of the following domains: Age f Reference Person, Region, Size of CU, Composition of Iousehold, Household Tenure, and Race of Reference Person.

Table 3. Ratios to CE cv to cv's for the different weighting methods. The minimum ratio is highlighted in each row.

Expenditure	regwts			calwts		
	0	1	2	0	1	2
All Exp.	0.98	0.90	0.79	0.98	0.90	0.78
Shelter	0.93	0.85	0.75	0.93	0.85	0.74
Utilities	1.08	1.03	0.94	1.07	1.03	0.88
Furniture	1.08	1.21	3.52	1.06	1.21	2.58
Maj. ap.	1.08	1.06	1.04	1.06	1.08	1.09
All vehi.	0.90	0.89	0.98	0.91	0.89	0.98
Cars, (n)	0.95	0.91	1.01	0.96	0.91	1.02
Cars, (u)	0.98	0.94	0.96	0.97	0.94	0.97
Gasol.,	1.17	1.11	1.03	1.12	1.10	0.99
Health	1.05	0.97	0.86	1.07	0.97	0.85
Educat.	0.92	0.93	1.04	0.91	0.93	1.06
Contrib.	1.01	1.02	1.28	1.01	1.02	1.30
Pers. ins.	1.00	0.97	1.64	1.01	0.98	1.24
Life, Ins.	1.08	1.02	1.53	1.08	0.98	1.38
Pensions	1.00	0.99	1.75	1.01	0.99	1.34

n addition, ratios for all CU's, i.e., the total across the lomains, were computed for each expenditure and those or regwts 0, 1, 2 and calwts 0, 1, 2 are shown in Table 3. For All Expenditures, regwts2 and calwts2 with ratios of 79 and .78 provide substantial reduction in cv compared to PP. For less aggregated expenditures, regwts1 or calwts1 provide reasonably consistent improvements over PP without the losses incurred by some of the other weights or expenditures like Furniture, Personal insurance and pensions, and its subcategory Pensions and social security.

A trellis plot (Cleveland 1993) of the cv and mean atios for calwts0 and calwts1 by age of reference person is given in Figure 2. Calwts0 is pictured because it is the nearest calibration equivalent to the current method of post-stratification. Calwts1 appears to be the best of the lternatives we have examined in the sense of improving he All Expenditures estimates while providing consistent performance for individual expenditure groups. In each panel of the plot a vertical reference line is drawn at 1, the point of equality between the calibration results and those or the PP method. The lower tier in the plot presents

ratios tend to be less than 1, for most dom expenditures, and calwts1 is somewhat better than

Calwts2 and calwts3, which used family incor taxes as one of the auxiliaries, had somewh performance for domains, sometimes makin improvements over PP but occasionally showin losses. This is connected to the nature of t income variable which had a substantial number with negative and zero values. These CU's v usefulness of this variable in predicting expenditur

Taking all of the above into consideration, calwts1 and calwts4 can be deemed a clear imp over the PP estimator. Calwts1 has the advantag negative weights over regwts1. Since calwts4 re auxiliary variables as opposed to calwts1's recommend calwts1 over all the other types of w have considered.

4. CONCLUSION AND FUTURE RESEA

The objective of this study was to i alternatives to the principal person method for household weights that did not depend on the one single member of the household. Different weights based on the regression estimation procec presented and their relative merits evaluated. R estimation incorporates the current surve stratification methods in which the weighted su persons in each post-stratum is forced to be eq independent census count of that number. accomplished via auxiliary variables that are inc into the regression model. It also automatically for each sample household a weight that does n on any single one of its members. In order to elir undesirable negative weights that can result from least-squares regression estimation, calibration e were adapted to the present problem. The c estimation procedure has the flexibility to re possible deviation of each final weight from its ba while adhering to the properties discussed above particular allows the constraint of positive weig calibration weights are easily computed via matrix software like S-Plus™.

Overall, the ordinary regression estimator calibration estimator both appeared to be an imp over the Principal Person estimator in term coefficient of variation. For the future, the c estimators can be further refined by using the pro regression estimation to choose the auxiliary vari

REFERENCES

ALEXANDER, C. H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.

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

CASADY, R J., and VALLIANT, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.

CLEVELAND, W.S. (1993). *Visualizing Data*. Summit, New Jersey: Hobart Press.

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

DEVILLE, J.C., SÄRNDAL, C.E., AND SAUTORY, O. (1993). Generalized Raking Procedures In Survey Sampling. *Journal of the American Statistical Association*, **88**, 1013-1020.

FULLER, W. A., LOUGHIN, M. M. and BAKER, H. D. (1993). Regression weighting for the 1987-1988 Nationwide Food Consumption Survey. Un report submitted to the United States Depa Agriculture.

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

RAO, J.N.K. (1992). Estimating totals and di functions using auxiliary information at the e stage. Presented at the Workshop on Uses of Information in Surveys, Statistics Sweden.

SÄRNDAL, C. E., SWENSSON, B. and WRETM (1992). *Model assisted survey sampling*. New Springer - Verlag.

STUKEL, D. M. and BOYER, R. (1992). Calibra estimation: an application to the Canadian labo survey. Working Paper, Statistics Canada : SS 009E, 1992

ZIESCHANG, K. D. (1990). Sample weighting and estimation methods in the Consumer Ex Survey. *Journal of the American S Association*, 85, 986-1001.

Figure 2. Ratios to CE of cv's and means for two weighting methods by age of reference person.