Survey Methodology, December 1996
Vol. 22, No. 2, pp. 127-137
Statistics Canada

An Application of Restricted Regression Estimation in a Household Survey

BODHINI R. JAYASURIYA and RICHARD VALLIANT[1]

ABSTRACT

This paper empirically compares three estimation methods – regression, restricted regression, and principal person – used in a household survey of consumer expenditures. The three methods are applied to post-stratification which is important in many household surveys to adjust for under-coverage of the target population. Post-stratum population counts are typically available from an external census for numbers of persons but not for numbers of households. If household estimates are needed, a single weight must be assigned to each household while using the person counts for post-stratification. This is easily accomplished with regression estimators of totals or means by using person counts in each household's auxiliary data. Restricted regression estimation refines the weights by controlling extremes and can produce estimators with lower variance than Horvitz-Thompson estimators while still adhering to the population controls. The regression methods also allow controls to be used for both person-level and household-level counts and quantitative auxiliaries. With the principal person method, persons are classified into post-strata and person weights are ratio adjusted to achieve population control totals. This leads to each person in a household potentially having a different weight. The weight associated with the "principal person" is then selected as the household weight. We will compare estimated means from the three methods and their estimated standard errors for a number of expenditures from the Consumer Expenditure survey sponsored by the U.S. Bureau of Labor Statistics.

KEY WORDS: Calibration; Principal person method; Replication variance; Restricted regression.

1. INTRODUCTION

A signal problem in large household surveys is under-coverage of the target population often arising from differential response rates among population subgroups and frame deficiencies. Post-stratification is one method used at the estimation stage to reduce mean square errors based on information that affect the response variables. The estimator is constructed in such a way that the estimated total number of individuals falling into each post-stratum is equal to the true population count. Post-stratum population counts are typically available from an external census for numbers of persons but not always for numbers of households. If household estimates are needed, a single weight must be assigned to each household while using the person counts for post-stratification. Regression estimators of totals or means accomplish this by using person counts in each household's auxiliary data. Restricted regression estimation controls extreme weights and can produce estimators with lower variance than the Horvitz-Thompson estimator while still adhering to the population controls. An alternative used by some surveys is the Principal Person (PP) method (Alexander 1987) in which the household weight is based on the individual designated as the"principal person" in each household. Persons are classified into post-strata and person weights are ratio adjusted to achieve population control totals, leading to the possibility that each person in a household may have a different weight. The weight associated with the principal person is then assigned to the household. This *ad hoc* method is difficult to analyze theoretically. The regression estimators discussed in this paper, while easily adjusting for the population under-count, automatically provide a household weight that is not based on any particular one of its members. Lemaître and Dufour (1987) address Statistics Canada's use of the regression estimator in this regard.

There are a growing number of precedents for the use of regression estimators in surveys both in the theoretical literature and in actual survey practice. Statistics Canada has incorporated the general regression estimator into its generalized estimation system (GES) software that is now used in many of its surveys (Estevao, Hidiroglou and Särndal 1995). Fuller, Loughin and Baker (1993) discuss an application to the USDA Nationwide Food Consumption Survey. One of the attractions of regression estimation is that many of the standard techniques in surveys including the post-stratification estimator mentioned above are special cases of regression estimators. The regression estimator also more flexibly incorporates auxiliary data than other more common methods. In a household survey, for example, both person-level and household-level auxiliaries that can be qualitative or quantitative are easily accommodated. Other works related to regression estimation and post-stratification include Bethlehem and Keller (1987), Casady and Valliant (1993), Deville and Särndal (1992), Deville, Särndal and Sautory (1993) and Zieschang (1990).

In this study we compare the regression estimator with the PP estimator currently in use at the Bureau of Labor Statistics (BLS). Each estimator can be written in the form of a weighted sum of the sample values of the response variable. Then each weight is traditionally interpreted as the number of

[1] Bodhini R. Jayasuriya and Richard Valliant, U.S. Bureau of Labor Statistics, 2 Massachusetts Avenue, N.E., Room 4915, Washington, DC 20212, U.S.A.

individuals in the population who would have the corresponding value of the response variable. This interpretation requires that each weight be greater than or equal to one. The ordinary least-squares regression estimator has the disadvantage that it can produce non-positive weights. A number of ways are suggested in the literature on how to overcome this problem. Possibly the easiest is the method introduced by Deville and Särndal (1992) which can remove any negative weights as well as control extreme weights. The restricted regression estimators produced by these new weights are also compared to the original regression estimator and the PP estimator.

In Section 2, the three different estimators are presented. Section 3 is an application of these procedures to the Consumer Expenditure (CE) Survey at BLS – the same setting as in Zieschang (1990). We compare the coefficients of variation for a number of the survey target variables for the full population and for a number of domains. Section 4 provides a summary of our conclusions.

2. REGRESSION, CALIBRATION AND PRINCIPAL PERSON ESTIMATION

First, we give a brief introduction to the regression estimator. A sample s of size n is selected from a finite population U of size N. Let the probability of selection of the i-th unit be π_i. The sample could be two-stage and the unit could be either the primary sampling unit or the secondary sampling unit. There is no need here to complicate the notation with explicit subscripts for the different stages of sampling. Let the variable of interest be denoted by y and suppose that its value at the i-th unit, y_i, is observed for each $i \in s$. Assume the existence of K auxiliary variables $x_1, x_2, ..., x_K$ whose values at each $i \in s$ are available. Define $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iK})'$, for each $i \in U$, where x_{ik} denotes the value of the variable x_k at unit i. Let $\boldsymbol{X} = (X_1, ..., X_K)'$ denote the K-dimensional vector of known population totals of the variables $x_1, x_2, ..., x_K$. The regression estimator is then motivated by the working model ξ:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_K x_{iK} + \varepsilon_i \tag{2.1}$$

for $i = 1, ..., N$. Here, $\beta_1, ..., \beta_K$ are unknown model parameters. The ε_i are random errors with $E_\xi(\varepsilon_i) = 0$ and $\mathrm{var}_\xi(\varepsilon_i) = \sigma_i^2$ for $i = 1, ..., N$. The term "working model" is used to emphasize the fact that the model is likely to be wrong to some degree. In the CE, the unit of analysis, indexed by i, is a consumer unit (CU), which is similar to a household and defined in more detail in Section 3. The value y_i might be the total food expenditures by the CU and the x_{ik}'s might be various CU characteristics like numbers of people of different ages, or CU income, that have an effect on the CU's expenditure on food. The variance of expenditures might be dependent on CU size so that having σ_i^2 proportional to the number of persons in the CU might be reasonable. We include an intercept in some of our models by setting the first auxiliary variable, x_1, equal to 1.

A linear regression estimator of the population total of y is defined to be

$$\hat{y}_R = \hat{y}_\pi + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)'\hat{\beta} \tag{2.2}$$

where $\hat{y}_\pi$ denotes the π-estimator (or Horvitz-Thompson estimator) of the population total of y, *i.e.*,

$$\hat{y}_\pi = \sum_{i \in s} a_i y_i \tag{2.3}$$

with $a_i = 1/\pi_i$. Also, $\hat{\boldsymbol{x}}_\pi = (\hat{x}_{1\pi}, ..., \hat{x}_{K\pi})'$ is the vector of π-estimators of the population totals of the variables $x_1, x_2, ..., x_K$ and

$$\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_K)' = \left[\sum_{i \in s} \frac{a_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_i^2}\right]^{-1} \sum_{i \in s} \frac{a_i \boldsymbol{x}_i y_i}{\sigma_i^2}. \tag{2.4}$$

We assume that $\sum_{i \in s} a_i \boldsymbol{x}_i \boldsymbol{x}_i' / \sigma_i^2$ is nonsingular. Even if model (2.1) fails to some degree, $\hat{y}_R/N$ is a design consistent estimator of the population mean $\bar{Y}$ irrespective of whether the assumed model is true or false. This is clear from (2.2). If $\hat{y}_\pi/N$ and $\hat{\boldsymbol{x}}_\pi/N$ are design consistent estimators of $\bar{Y}$ and of $\bar{\boldsymbol{X}}$, the vector of population means of the auxiliaries, then the second term in $\hat{y}_R/N$ converges to zero while the first converges to $\bar{Y}$. For more details, see Särndal, Swensson and Wretman (1992).

The regression estimator $\hat{y}_R$ can also be expressed as a weighted sum of the sample y_i's, which is a desirable feature for survey operations. It is easily seen that (2.2) can be re-written as $\hat{y}_R = \sum_{i \in s} w_i y_i$ with

$$w_i = a_i \left[1 + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)' \boldsymbol{A}^{-1} \frac{\boldsymbol{x}_i}{\sigma_i^2}\right] \tag{2.5}$$

where $\boldsymbol{A} = \sum_{i \in s} a_i \boldsymbol{x}_i \boldsymbol{x}_i' / \sigma_i^2$. The weights do depend on the sample through the x_i's that are in the sample, but this is also true of many survey estimators, including the post-stratification estimator. However, these weights do not depend on the particular y variable being studied, implying that one set of w_i weights can be used for all estimates.

A mean per unit is estimated in the obvious way: $\hat{\bar{y}}_R = \hat{y}_R/\hat{N}$ where $\hat{N} = \sum_{i \in s} w_i$. If we estimate the totals of the auxiliaries $\boldsymbol{x}_i$, then

$$\begin{aligned} \sum_{i \in s} w_i \boldsymbol{x}_i' &= \sum_{i \in s} \left[a_i \boldsymbol{x}_i' + (\boldsymbol{X} - \hat{\boldsymbol{x}}_\pi)' \boldsymbol{A}^{-1} \frac{a_i \boldsymbol{x}_i \boldsymbol{x}_i'}{\sigma_i^2}\right] \\ &= \boldsymbol{X}', \end{aligned} \tag{2.6}$$

i.e., we reproduce the known population totals. This is also a characteristic of the post-stratification estimator.

The estimator of β in (2.4) does not account for any correlation among the errors in model (2.1). In clustered populations, units that are geographically near each other, *e.g.*, CU's in the same neighborhood, may be correlated. Using a full covariance matrix V may be more nearly optimal (*e.g.*, see Casady and Valliant 1993 and Rao 1994). Though use of a full covariance matrix V may lower the variance of $\hat{\beta}$, the elements of V will depend on the particular y being studied, and estimation of V is generally a nuisance. Consequently, it is interesting and practical to consider the simple case of $V = \text{diag}(\sigma_i^2)$ that leads to (2.2). Note that when the design-variance $\text{var}_p(\hat{y}_R)$ is estimated, it will be necessary to use a method that properly reflects clustering and other design complexities.

The regression estimator has the disadvantage that the weights can be unreasonably large, small or, even negative. The restricted calibration estimators of Deville and Särndal (1992), introduced next, add constraints to control the size of the weights. Calibration estimators are formed by minimizing a given distance, F, between some initial weight and the final weight, subject to constraints. The constraints can involve the available auxiliary variables thus incorporating them into the estimator. The regression estimator presented above is a special case of the calibration estimator in which F is defined to be the generalized least squares (GLS) distance function,

$$F(w_i, a_i) = \frac{a_i c_i}{2}\left(\frac{w_i}{a_i} - 1\right)^2$$

for $i = 1, ..., n$, with c_i a known, positive weight (*e.g.*, $c_i = \sigma_i^2$ or $c_i = 1$) associated with unit i, and w_i, the final weight. The total sample distance $\sum_{i \in s} F(w_i, a_i)$ is minimized subject to the constraints,

$$\sum_{i \in s} w_i \boldsymbol{x}_i = \boldsymbol{X}. \tag{2.7}$$

In this form, the weights of the regression estimator of the population total of y given in (2.5) can be written as,

$$w_i = a_i g(c_i^{-1} \boldsymbol{\lambda}' \boldsymbol{x}_i) \tag{2.8}$$

for $i = 1, ..., n$ where

$$g(u) = 1 + u, \tag{2.9}$$

for $u \in \Re$ and λ is a Lagrange multiplier evaluated in the minimization process. The particular form of w_i with $c_i = \sigma_i^2$ for the regression estimator was given in (2.5). To eliminate extremes, the weights can be refined by restricting g so that

$$g(u) = \begin{cases} L & \text{if } u < L - 1 \\ 1 + u & \text{if } L - 1 \leq u \leq U - 1 \\ U & \text{if } u > U - 1. \end{cases} \tag{2.10}$$

With this definition of g, the weights w_i satisfy

$$L < w_i / a_i < U \tag{2.11}$$

for $i = 1, ..., n$ so that L and U can be chosen in such a way as to reflect the desired deviation from the initial weights a_i. Choosing $L > 0$ ensures that the weights are positive, and U is picked to be appropriately small to prohibit large weights. The restricted regression weights must be solved for iteratively; one easily programmed algorithm is given in Stukel and Boyer (1992). Another method of restricting weights is ridge regression as used by Bardsley and Chambers (1984).

In most household surveys, post-stratification serves primarily as an adjustment for under-coverage of the target population by the frame and the sample. In the U.S., there are few reliable population counts of households to use in post-stratification. Consequently, population counts of persons are usually used for the post-strata control totals. This disagreement in the unit of analysis (the household) and the unit of post-stratification (the person) when a household characteristic is of interest led to the development of the PP method that is used in the CE and Current Population Surveys.

In the PP method described in Alexander (1987), a household begins the weighting process with a single base weight, a_i, that is then adjusted for non-response. The adjusted weight is assigned to each person in the household and the person weights are then further adjusted to force them to sum to known population controls of persons by age, race, and sex. This last adjustment can result in persons having different weights within the same household. The household is then assigned the weight of the person designated as the "principal person" in the household. This method has an element of arbitrariness and is difficult to analyze mathematically. The intent of this research was not to see if the PP method could be improved upon, but rather to use the current implementation of PP as a convenient baseline for measuring the performance of other estimators.

The regression and restricted regression estimators can be formulated in such a way that population person controls are satisfied, all persons in a household retain the same weight, and no arbitrary choice among person weights is needed to assign a household weight. This is accomplished by defining the auxiliary variables at the household level. For example, if there were three age post-strata and household i has 1, 0, and 2 persons in these post-strata, the auxiliary data vector would be $\boldsymbol{x}_i = (1,0,2)'$. Note that this formulation is different from Lemaître and Dufour (1987) who defined the auxiliary variables at the person level and assigned the average of the household data – (1/3, 0, 2/3) in the example – to each person. Those authors used this "average" method because they were interested in estimates both for persons, *e.g.*, number employed, and for households, *e.g.*, economic families. We, on the other hand, need only a household weight since our target variables (*i.e.*, y) like shelter or utility expenditures are collected at the household level.

3. AN APPLICATION

We compare the three estimators (*i.e.*, regression, restricted regression (with $L = .5$, $U = 4$), and principal person) by an application to the estimated means and their estimated standard errors for a number of expenditures from the CE Survey sponsored by the Bureau of Labor Statistics.

The CE Survey gathers information on the spending patterns and living costs of the American consumers. There are two parts to the survey, a quarterly interview and a weekly diary survey. The Interview Survey collects detailed data on the types of expenditures which respondents can be expected to recall for a period of three months or longer (*e.g.*, property, automobiles, major appliances) – an estimated sixty to seventy percent of total household expenditures. The Diary Survey is completed at home by the respondent family for two consecutive 1-week periods and collects data on all the expenses of the family in that time period. The sample is selected in two stages with geographic primary sampling units at the first stage and households at the second.

We evaluated the estimators described above for a number of expenditures from the Interview Survey. Data collected during the second quarter of 1992 consisting of $n = 5156$ CU's were used. The CE Survey's primary unit of analysis is the consumer unit, an economic family within a household. A consumer unit (CU) consists of individuals in the household who share expenditures. Thus, there may be more than one CU in a household.

Five different sets of auxiliary variables ($\boldsymbol{x}_i$'s in the notation of Section 2) were studied. They were chosen by testing the adequacy of model (2.1) for the selected expenditures with different combinations of the available auxiliary variables. Combinations of auxiliaries were identified in which each estimated regression coefficient was significant in an ordinary least squares regression at the 5% level. A key step that substantially improved the fit of the models was simply including an intercept. Factored into the selection of auxiliaries was also the knowledge that the survey has more under-coverage of Blacks than non-Blacks and that this needed to be accounted for by post-stratification. We viewed this method of variable selection as exploratory and, consequently, a number of combinations were studied to determine which set produced the best estimators of mean expenditures. The 56 post-strata based on age/race/sex currently in use in the CE were included. (The 56 are routinely collapsed in actual CE operations because of small sample sizes in some cells.) Other variables that were statistically significant in various combinations were region (NE, MW, S, W), urbanicity (urban/rural) by region, age of reference person of the CU (< 25, 25-34, 35-44, 45-64, 65+), household tenure (owner/renter), income before taxes of the CU, and the 56 post-strata collapsed by sex and some of the age categories to form 10 age/race categories. Based on this information, weights (2.8) were computed using g given in (2.9) – regwts – and (2.10) – calwts. For both the regression and restricted regression weights, we set a_i equal to the adjusted base weight, *i.e.*, $1/\pi_i$ times a non-response adjustment. In order for the matrix $\boldsymbol{A}$ in Section 2 to be nonsingular, one of the categories in some auxiliaries, like region, was omitted from each $\boldsymbol{x}_i$. For this application, the population totals necessary to evaluate $X = (X_1,...,X_K)'$ were obtained mostly from the Statistical Abstract of the United States (1993) whose sources are the 1990 Census figures and the Current Population Reports published by the U.S. Bureau of the Census. When an intercept is used, the appropriate control total for that variable is the number of CU's in the population for which we used the PP estimate as a surrogate. The combinations of auxiliaries used to form the different weights are given in Table 1. RegwtsO, with 56 age/race/sex post-strata uses the largest number of post-strata. The 56 are the starting point for the PP method but are usually collapsed to 30-40 because of small cell sizes. When computing calwtsO, those 56 post-strata were collapsed to 45 since the constraints imposed by the L and U bounds could cause singularity in the matrix based algorithm.

Table 1

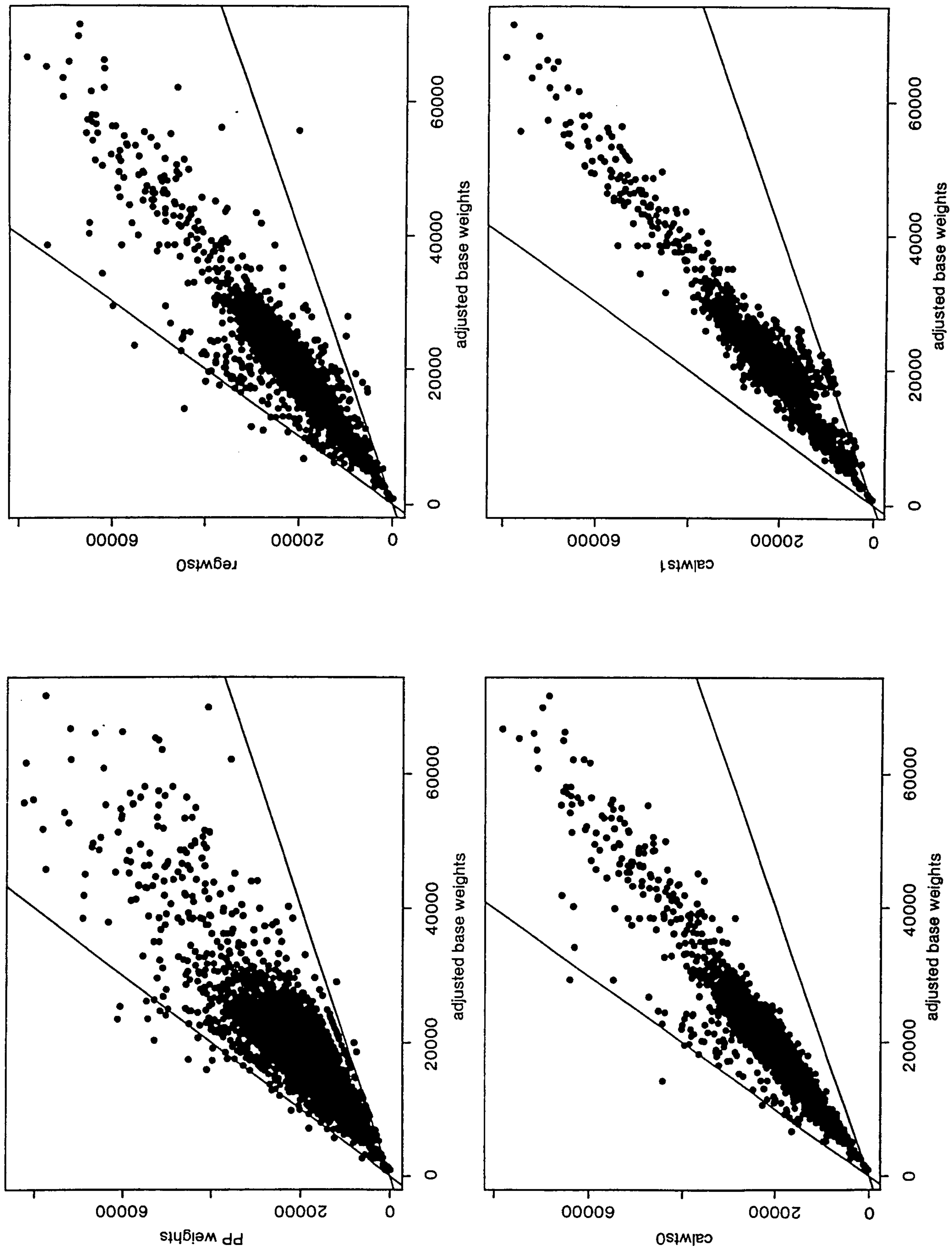
Weights and Their Corresponding Auxiliary Variables

Weights	Auxiliary Variables	K
regwts0	Age/race/sex	56
regwts1	Intercept, 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	Intercept, 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[1]

[1] The initial set of 56 is usually collapsed to 30-40 because of small sample sizes in some cells.

3.1 Comparisons of Weights

A variety of comparisons of weights produced by the different methods were made, only a few of which can be mentioned here. Figure 1 shows plots of the PP weights, regwtsO, calwtsO, and calwts1 versus the adjusted base weights. For PP and regwtsO, the adjustments to go from a_i to w_i are much more variable than for calwtsO and calwts1, which employ the $L = 0.5$ and $U = 4$ restrictions. High variability among the w_i can lead to expenditure estimates with high variance and to poor confidence interval coverage since large sample normality may not hold. Even though (2.11) implies that $a_i/2 < w_i < 4a_i$ for each i for the calwts, the lower right panel in Figure 1 shows that the calwts1 satisfy $a_i/2 < w_i \leq 2a_i$, for each i. Thus, setting $U = 2$ or 3 would have little effect on calwts1. CalwtsO would have been slightly affected by setting $U = 2$ since a few points were outside the upper reference line. The upper two panels indicate that the PP weights and regwtsO do not conform to the restriction $a_i/2 < w_i < 2a_i$.



(Reference lines correspond to L = .5 and U = 2)

Figure 1. Four sets of weights plotted versus adjusted base weights

The concern about negative regression weights was minor in the application. In the full sample, only one CU had a negative weight for regwts1 and regwts2 while regwts0 had no negative weights. However, in the replicates used for variance estimation, described in Section 3.2, 2 or 3 CU's did have negative weights in many replicates so that using the L restriction was more important there.

3.2 Precision of Estimates from the Different Methods

Although comparison of weights is instructive, the methods must ultimately be judged based on the level of estimated CU means and their precision. The standard errors of these estimators were computed via the method of balanced half sampling (BHS) using 44 replicates as currently implemented in the CE for the PP estimator. The BHS estimator is constructed to reflect the stratification and the clustering that is used in the CE. A half sample is constructed in a prescribed way (McCarthy 1969) to contain one half of the first-stage sample units in a survey. Defining the mean per CU based on CU's in half-sample α to be $\hat{\bar{y}}_{R(\alpha)}$ and that for the full sample to be $\hat{\bar{y}}_R$, the BHS estimate of variance is $V_{\text{BHS}}(\hat{\bar{y}}_{R(\alpha)}) = \sum_{\alpha=1}^{44}(\hat{\bar{y}}_{R(\alpha)} - \hat{\bar{y}}_R)^2/44$. To compute each $\hat{\bar{y}}_{R(\alpha)}$, the same estimation steps used for the full sample are repeated for the CU's in the half-sample. As the expenditure estimates from the CE Survey are published for various inter domains of interest, we computed the means and the standard errors for a few chosen domains as well. For each of these, the coefficient of variation (cv) was computed and then its ratio to the cv of the PP weight estimate was calculated.

For each type of weight, if the ratio of each expenditure cv to that of the PP weights is less than one, an improvement over the PP estimate is indicated since, for all the weights, the expenditure mean estimates were very close to those of the PP estimates. We computed the ratios of cv's and the ratios of means for each of the sets of weights described in Table 1, for each of the chosen expenditures, and for each of the following domains:

(1) Age of Reference Person: < 25, 25-34, 35-44, 45-54, 55-64, 65+
(2) Region: NE, MW, S, W
(3) Size of CU: 1, 2, 3, 4, 5+
(4) Composition of Household: Husband and wife only, Husband and wife + children, Other Husband and wife, One parent + at least one child < 18, Single person and other CU's
(5) Household Tenure: Owner, Renter
(6) Race of Reference Person: Black, Non-Black.

Table 2

Ratios to PP cv of cv's for the Different Weighting Methods
The Minimum Ratio is Highlighted in Each Row

Expenditure	regwts			calwts				
	0	1	2	0	1	2	3	4
All expenditures	0.98	0.90	0.79	0.98	0.90	0.78	0.75	0.87
Shelter	0.93	0.85	0.75	0.93	0.85	0.74	0.72	0.84
Utilities	1.08	1.03	0.94	1.07	1.03	0.88	0.91	0.92
Furniture	1.08	1.21	3.52	1.06	1.21	2.58	2.57	1.17
Major appliances	1.08	1.06	1.04	1.06	1.08	1.09	1.00	1.03
All vehicles	0.90	0.89	0.98	0.91	0.89	0.98	0.97	0.90
New cars, trucks	0.95	0.91	1.01	0.96	0.91	1.02	1.02	0.91
Used cars, trucks	0.98	0.94	0.96	0.97	0.94	0.97	0.96	0.95
Gasoline, motor oil	1.17	1.11	1.03	1.12	1.10	0.99	0.94	1.10
Health care	1.05	0.97	0.86	1.07	0.97	0.85	0.87	0.94
Education	0.92	0.93	1.04	0.91	0.93	1.06	1.07	0.88
Cash contributions	1.01	1.02	1.28	1.01	1.02	1.30	1.29	1.03
Personal insurance, pensions	1.00	0.97	1.64	1.01	0.98	1.24	0.98	0.95
Life, other personal insurance	1.08	1.02	1.53	1.08	0.98	1.38	1.33	1.01
Pensions, social security	1.00	0.99	1.75	1.01	0.99	1.34	1.06	0.97

We will discuss only domains (1) – (3) here. In addition, ratios for all CU's, *i.e.*, the total across the domains, were computed for each expenditure and are shown in Table 2. For All Expenditures, regwts2, calwts2, and calwts3, with ratios of .79, .78, and .75, 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 for expenditures like Furniture, Personal insurance and pensions, and its sub-category Pensions and social security.

Trellis plots (Cleveland 1993) of the cv and mean ratios for calwts0 and calwts1 are given in Figures 2-4. 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 alternatives we have examined in the sense of improving the All Expenditures estimates while providing consistent performance for individual expenditure groups. In each panel of the plots a vertical reference line is drawn at 1, the point of equality between the calibration results and those for the PP method. The lower row in each plot presents ratios of means from calwts0 and calwts1 to the PP means and illustrates that with a few exceptions the levels of the means from the two restricted regression choices are about the same as from PP.

The two calibration choices, in the main, improve cv's compared to PP, *i.e.*, cv ratios tend to be less than 1, for most domains and expenditures, and calwts1 is somewhat better than calwts0. For the age-of-reference-person domains < 25 and 65+, for example, 12 of the 15 expenditures have calwts1 ratios of less than 1. For CU sizes 1-4 the numbers of cv ratios less than or equal to 1 are 12, 9, 9, and 11. There are exceptions, of course. For the South region only 6 of 15 expenditures have calwts1 cv ratios less than or equal to 1.

Calwts2 and calwts3, which used family income before taxes as one of the auxiliaries, had somewhat erratic performance for domains, sometimes making major improvements over PP but occasionally showing serious losses. This is connected to the nature of the family income variable itself. For the entire data set of 5156 CU's, income before taxes was positive for 4698 CU's, zero for 450 CU's and negative for 8 CU's. The zeroes are incomplete income reporters while the negatives are for families that had business losses added to other income. In either case, these CU's vitiate the usefulness

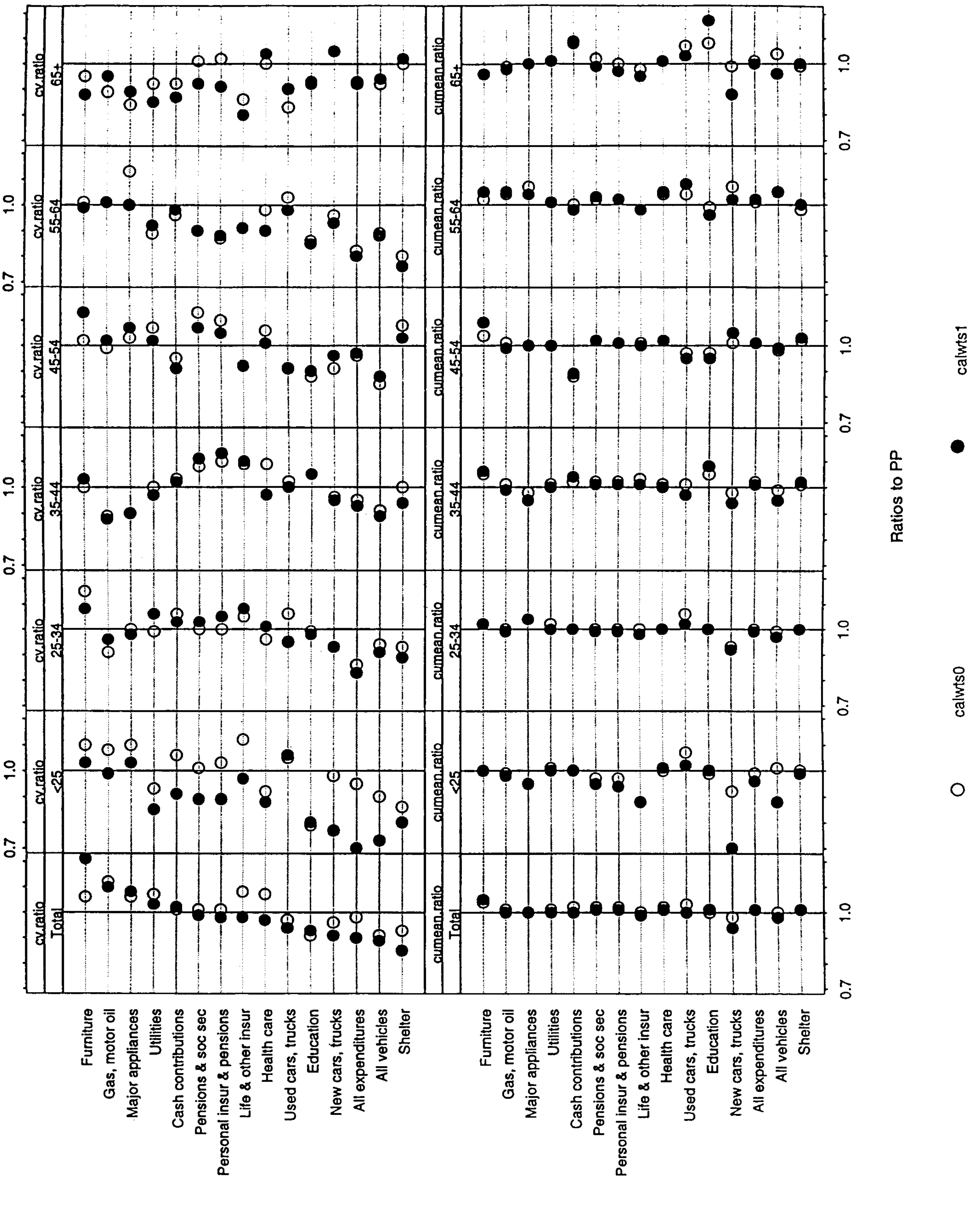


Figure 2. Ratios to PP of cv's and means for two weighting methods by age of reference person

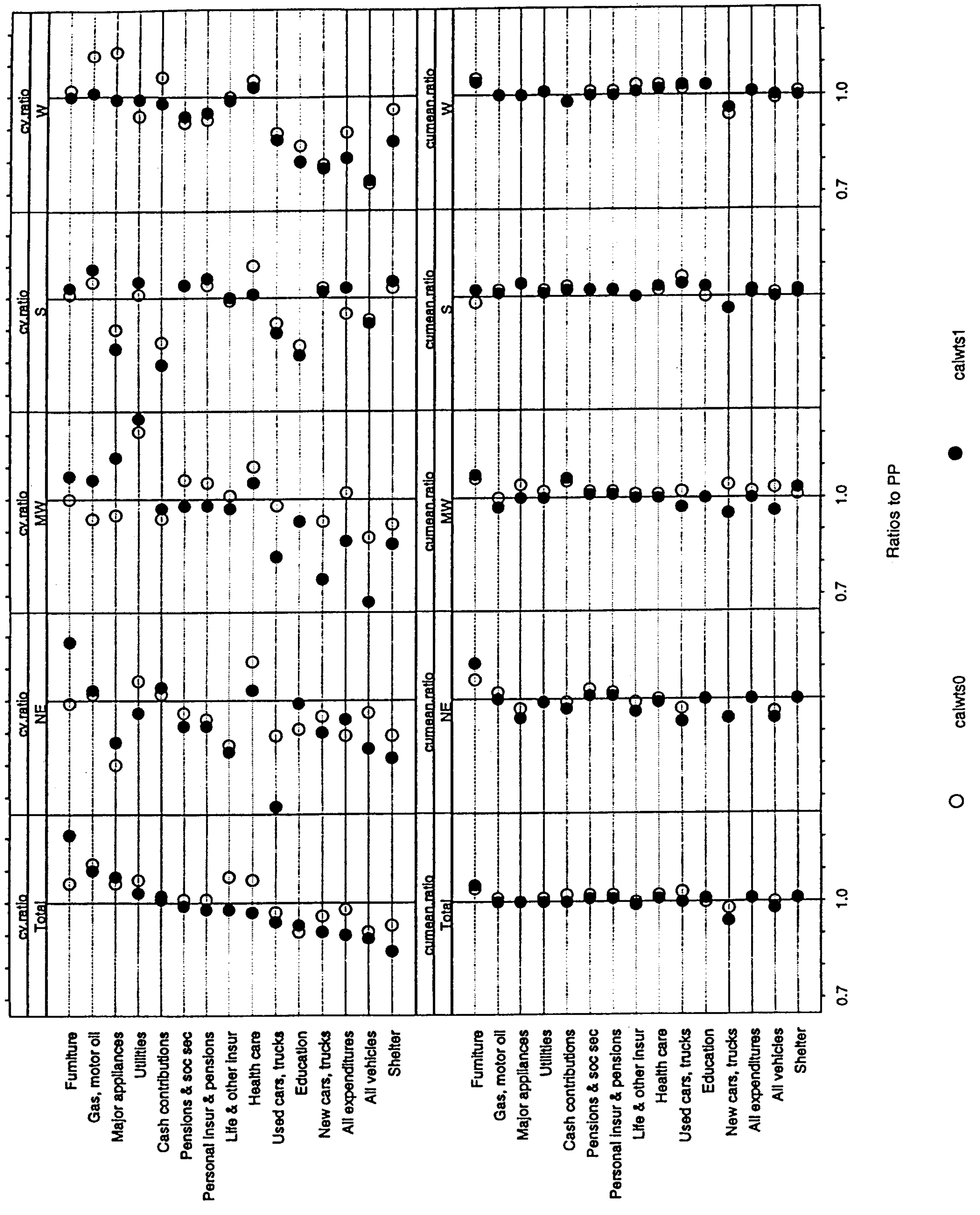


Figure 3. Ratios to PP of cv's and means for two weighting methods by region

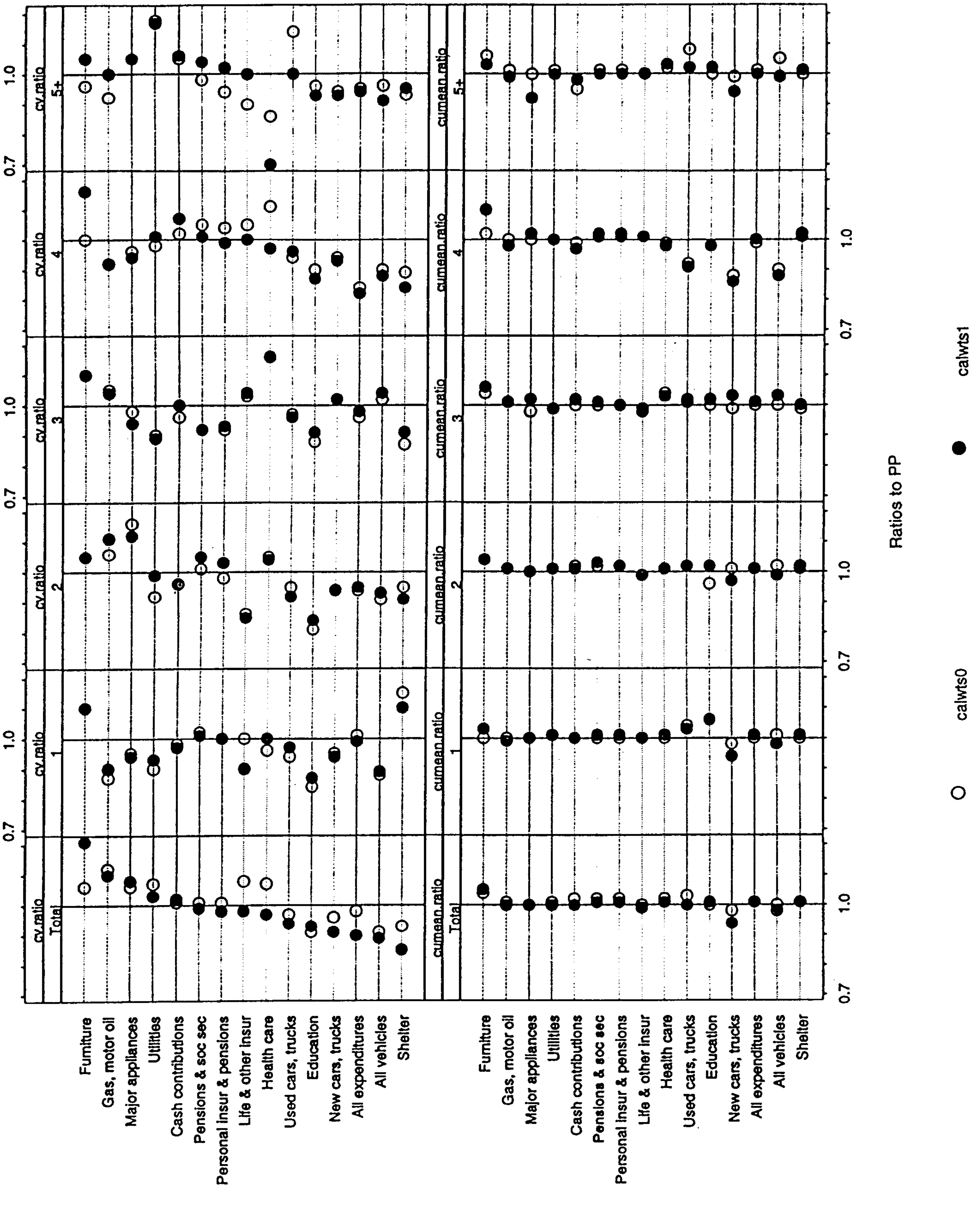


Figure 4. Ratios to PP of cv's and means for two weighting methods by size of CU

of this variable in predicting expenditures. Perhaps, use of another measure of income combined with item imputations for missing incomes would improve calwts2 and calwts3 for domain estimation.

Taking all of the above into consideration, regwts1, calwts1 and calwts4 are efficient choices in this application. Calwts1 has the advantage of non-negative weights over regwts1. Since calwts4 requires 23 auxiliary variables as opposed to calwts1's 18, calwts1 is the more parsimonious choice. Subsequent to the analysis discussed here, we performed a similar study using a full year's data for both the Interview and Diary Surveys for 1990. Results were similar to those reported here and a final set of 24 auxiliaries was adopted based on number of persons by age, race, sex, region, urban × region, and number of CU's by tenure, and an intercept. The conversion of CE estimation to restricted regression is now underway.

4. CONCLUSION

The objective of this study was to investigate methods for deriving household weights that did not depend on the weight of one single member of the household. Different types of weights based on the regression estimation procedure were presented and their relative merits evaluated. Regression estimation incorporates the current survey post-stratification methods in which the weighted sum of the persons in each post-stratum is forced to be equal to an independent census count of that number. This is accomplished via auxiliary variables that are incorporated into the regression model. It also automatically produces for each sample household a weight that does not depend on any single one of its members.

We studied eight types of weights that came from five different regression models. In order to eliminate the undesirable negative weights that can result from ordinary least-squares regression estimation, restricted regression estimators were adapted to the present problem. Restricted regression has the flexibility to restrict the possible deviation of each final weight from its base weight while adhering to the properties discussed above. This, in particular, allows the constraint of positive weights. The restricted regression weights are easily computed via matrix-oriented software like S-Plus™ or SAS/IML™.

Restricted regression, and more generally, restricted calibration have a number of attractive features for household surveys, like the one studied here, but also for surveys of other types of units like hospitals, schools, or business establishments where a variety of auxiliary data may be available. Given past data on target variables, standard model building procedures can be used for the selection of auxiliary variables. The properties of regression estimation can be used to choose the predictors optimally in order to reduce the redundancy of information that gets incorporated into the survey estimation procedure. This is one of the greatest advantages of using an estimator that has a vast and tested literature behind it. Good predictors may include qualitative variables, *e.g.*, age, race, type of hospital (general medical, psychiatric, *etc.*), type of business (manufacturing, retail trade, *etc.*) that might be often used in stratification or post-stratification. The predictors can also be quantitative variables like family income, annual sales, number of students at different levels, or the number of inpatient days to name but a few. In our application, including an intercept also led to noticeably smaller standard errors of survey estimates. The regression approach also allows data at different levels to be easily incorporated in estimation. In the household survey studied here, auxiliaries on both persons and households were included.

The immense flexibility of regression gives practitioners options they might not otherwise have. If new, pertinent predictor variables become available, software for regression estimation can accommodate them simply by changing the matrix of auxiliaries and vector of population controls. Software that is rigidly written to perform only post-stratification or ratio estimation with a single auxiliary, for example, might have to undergo a major overhaul to change the estimator. Of course, if the estimator is one of the less general post-stratification or the ratio types, regression software will often handle it as a special case. In the United States, an extremely large continuing household survey is being contemplated (Love, Alexander and Dalzell 1995) that will provide very precise estimates of many characteristics that may be used as control totals in smaller surveys. The restricted regression approach positions the CE Survey to smoothly incorporate such new data in estimation should it become available.

ACKNOWLEDGMENTS

Any opinions expressed are those of the authors and do not represent policy of the Bureau of Labor Statistics. We are grateful to the referees and editor for their useful comments.

REFERENCES

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

BARDSLEY, P., and CHAMBERS, R.L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290-299.

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.

ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimating system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

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

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

LOVE, S., ALEXANDER, C.H., and DALZELL, D. (1995). Constructing a major survey – operational plans and issues of continuous measurement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. To appear.

McCARTHY, P.J. (1969). Pseudo-replication: half-samples. *Review of the International Statistical Institute*, 37, 239-264.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

STUKEL, D.M., and BOYER, R. (1992). Calibration estimation: an application to the Canadian Labor Force Survey. Working Paper, SSMD-92-009E, Ottawa: Statistics Canada.

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