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

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 methods—regression, calibration, and principal person—used in a household survey for post-stratification. Post-stratification is important in many household surveys to adjust for nonresponse and the population undercount that results from frame deficiencies. The correction for population undercoverage is usually achieved by adjusting estimated people counts in each post-stratum to equal the corresponding population control counts typically available from an external source such as a census. 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 Bureau of Labor Statistics in an attempt at understanding how each estimation method accomplishes this step in post-stratification.

1. INTRODUCTION

In large household surveys, post-stratification is a means of reducing mean square errors by adjusting for differential response rates among population subgroups and frame deficiencies that often result in undercoverage of the target population. In general, the population is subdivided into groups (post-strata) at the estimation stage 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 for numbers of households. If household estimates are needed, a single weight must be assigned to each household while using person counts for post-stratification. Regression estimators of totals or means accomplish this by using person counts in each household’s auxiliary data. Calibration estimation, with a least-squares distance function, is closely related to regression estimation but has the possibility that each person in a household may have different weight. The weight associated with the person is then assigned to the household. This method is difficult to analyze theoretically. The principal person estimator discussed in this paper, while easily adjusting the population undercount, automatically provides household weight that is not based on any particular members. Lemaitre and Dufour (1987) Statistics Canada’s use of the regression estimator is relevant.

There are a growing number of precedents for regression estimators in surveys both in the statistical literature and in actual survey practice. Statistics Canada has incorporated the general regression estimator into its generalized estimation system (GES) software that is used in many of its surveys. Fuller, Loughin and others (1993) discuss an application to the USDA National Food Consumption Survey. One of the attractive features of regression estimation is that many of the techniques in surveys including the post-strata estimator mentioned above are special cases of regression estimators. It also more flexibly incorporates auxiliary information than other more common methods. Other works on regression estimation and post-stratification can be found in Bethlehem and Keller (1987), Casady and Valliant (1992), Deville and Särndal (1992), and Zieschang (1990).

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

In Section 2, the three different estimators are presented. Section 3 is an application of these estimators to the Consumer Expenditure (CE) Survey at the same setting as in Zieschang (1990). We con...
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 \( \pi_i \). 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 \( x_1, x_2, \ldots, x_K \) whose values at each \( i \in s \) are available. Define \( x_i = (x_{i1}, x_{i2}, \ldots, x_{ik})' \), for each \( i \in U \), where \( x_{ik} \) denotes the value of the variable \( x_k \) at unit \( i \).

Let \( X = (x_1, \ldots, x_K)' \) denote the \( K \)-dimensional vector of known population totals of the variables \( x_1, x_2, \ldots, x_K \). The regression estimator is then motivated by the working model:

\[
y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{ik} + \varepsilon_i \quad (2.1)
\]

or \( i = 1, \ldots, N \). Here, \( \beta_1, \ldots, \beta_K \) are unknown model parameters. The \( \varepsilon_i \) are random errors with \( \mathbb{E}(\varepsilon_i) = 0 \) and \( \text{var}(\varepsilon_i) = \sigma_i^2 \) for \( i = 1, \ldots, N \). The term "working model" is used to emphasize the fact that the model is likely to be wrong to some degree. In the CE, \( y_i \) might be the total food expenditures by the consumer unit (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 \( \sigma_i^2 \) proportional to the number of persons in the CU might be reasonable. Then, a linear regression estimator of the population total of \( y \) is defined to be

\[
\hat{y}_R = \bar{y}_n + (\hat{x} - \bar{x})' \hat{\beta}
\]

(2.2)

where \( \hat{y}_n \) denotes the \( \pi \)-estimator (or Horvitz-Thompson estimator) of the population total of \( y \), i.e., \( \hat{y}_n = \sum_{i \in s} a_i y_i \), where \( a_i = 1/\pi_i \). Also, \( \hat{x}_\pi = (\hat{x}_{i1}, \ldots, \hat{x}_{iK})' \) is the vector of \( \pi \)-estimators of the population totals of the variables \( x_1, x_2, \ldots, x_K \) and

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

(2.3)

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

The regression estimator \( \hat{y}_R \) can also be expressed as a weighted sum of the sample \( y_i \)'s with \( i \)th weight,

\[
w_i = a_i \left[ 1 + (X - \bar{x})' \left( \sum_{i \in s} \frac{a_i x_i x_i'}{\sigma_i^2} \right)^{-1} X \right]^{-1}.
\]

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

The estimator of \( \beta \) in (2.3) does not account for correlation among the errors in model (2.1). In 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 optimal (e.g., see Casady and Valliant 1993). Though use of a full covariance matrix lowers the variance of \( \hat{\beta} \), the elements of \( V \) will depend on the particular \( y \) being studied, and estimation generally a nuisance. Consequently, it is interpractical to consider the simple case of \( V = \text{diag}(\sigma_i^2) \) which lead to (2.2). Note that when the design variance \( \varphi(y_n) \) is estimated, it will be necessary to use weights that properly reflect clustering and other complexities.

The regression estimator has the disadvantage that the variances of the weights can be unreasonably large, small or even negative. The calibration estimators of Deville and Särndal introduced next, add constraints to restrict the scale of the weights. Calibration estimators are formed by minimizing the total distance, subject to constraints. The constrained version involves the available auxiliary variables thus improving the estimator. The regression presented above is a special case of the calibration estimator in which \( \hat{F} \) is defined to be the general squares (GLS) distance

\[
F(w_i, a_i) = a_i c_i (w_i / a_i - 1)^2 / 2 \quad \text{for} \quad i = 1, \ldots, n,
\]

with \( a_i = 1/\pi_i \), and \( w_i \) the final weight. The total distance \( \sum_{i \in s} F(w_i, a_i) \) is minimized subject to constraints, \( \sum_{i \in s} w_i x_i = X \). In this form, the weighted regression estimator of the population total of \( y \) (2.4) can be written as

\[
w_i = a_i g(c_i^2 \lambda, x_i)
\]

for \( i = 1, \ldots, n \) where

\[
g(u) = 1 + u,
\]

for \( u \in \mathbb{R} \) and \( \lambda \) is a Lagrange multiplier evaluated in the minimization process. The calibration weights can be made to reflect certain complex sampling design.
hosen in such a way as to reflect the desired restrictions on the weights. Choosing $L > 0$ ensures that the weights are positive, and $U$ is picked to be appropriately small to prohibit large weights. The calibration weights must be solved for iteratively; one easily programmed algorithm is given in Stukel and Boyer (1992).

In most household surveys, post-stratification serves primarily as an adjustment for undercoverage of the target population by the frame and the sample. In the U.S., there are no reliable population counts of households to use in post-stratification. Consequently, population counts of persons are 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 nonresponse. 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 regression and calibration 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.

3. AN APPLICATION

We compare the three estimators (i.e., regression, restricted calibration (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

$n = 5156$ CU’s were used. The CE Survey’s pri

of analysis is the consumer unit, an economic family household. A consumer unit (CU) consists of it in the household who share expenditures. Thus, there be more than one CU in a household.

Five different sets of auxiliary variables were chosen by testing the adequacy of model for the selected expenditures with different combinations of the available auxiliary variables. The 56 post-strata 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 parentheses on this information, weights (2.5) were computed given in (2.6)—regwts—and (2.7)—calwts. For regression and restricted calibration weights, $w$ is equal to the adjusted base weight, i.e., $1 / \pi$, nonresponse adjustment.

Table 1. Weights and their corresponding auxiliary variables. Number of cells are in parentheses.

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

For this application, the population totals needed to evaluate $X = (X_1, \ldots, X_p)'$ were obtained mostly from 1990 Census figures projected to 1992 and the Population Reports published by the U.S. Bureau of the Census.

3.1 Comparisons of Weights and Estimated CU Counts
Replicate weights, nearly half the sets for each of regwts0, egwts1 and regwts2 had some negative weights though the maximum number of negative weights for any replicate was 3. The negatives are a potential cause of inflated standard errors, since the negative weights will be offset by large positive weights in order for the fixed population control totals to be met in every replicate. Calwts, which restrict the deviation from the base weights by choosing $L$ and $U$ appropriately, (in this instance, $L=0.5 > 0$) naturally did not produce any negative weights.

On examining scatter plots (not shown here) comparing some of the different weights to each other, the PP and egwts0, while being substantially different from each other, exhibited final weights that can be considerably different from the adjusted base weights. The adjustments can be either up or down. A less variable set of adjustments was apparent in regwts1, calwts0, and calwts1. Calwts1 and calwts4 were quite similar and both were close to regwts1. The two sets of weights that involve the quantitative variable family income before taxes, calwts2 and calwts3, were closely related. Some CU’s had calwts2 values larger than 60,000 but had calwts0, calwts1, calwts4 > 30,000. These CU’s all had family incomes before taxes of a quarter of a million dollars or more. Thus, the inclusion of that variable in the calibrations did have a substantial impact on some units. We did use a control only on the grand total income; having controls by income lasses might have changed the weights on some of these cases.

Figure 1. Four sets of weights plotted against adjusted base weights. Reference lines correspond to $L=.5$ and $U=2$.

Previous studies at BLS regarding regression estimation in the CE had concluded number of single person CU’s was under estimated 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 correlation ratio of the estimated number of CU’s under the a procedures to that of the PP estimation proced 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.

<table>
<thead>
<tr>
<th>Weights</th>
<th>CU Size</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5+</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td></td>
<td>28,784</td>
<td>30,680</td>
<td>15,409</td>
<td>15,068</td>
<td>9,993</td>
</tr>
<tr>
<td>regwts0</td>
<td>0.96</td>
<td>0.99</td>
<td>1.01</td>
<td>0.99</td>
<td>1.02</td>
<td></td>
</tr>
<tr>
<td>regwts1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.02</td>
<td>0.98</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>regwts2</td>
<td>1.00</td>
<td>1.01</td>
<td>1.01</td>
<td>1.01</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>regwts3</td>
<td>0.98</td>
<td>1.02</td>
<td>1.01</td>
<td>1.01</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>regwts4</td>
<td>1.01</td>
<td>1.00</td>
<td>1.01</td>
<td>0.99</td>
<td>0.99</td>
<td></td>
</tr>
</tbody>
</table>

A similar table constructed by Composition of CU that while regwts0 and calwts0 estimator 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 N

Although comparison of weights is instructive methods must ultimately be judged based on the estimated CU means and their precision. The errors of these estimators were computed via the of balanced half sampling (BHS) using 44 repl currently implemented in the CE for the PP estimator is constructed to reflect the strata and the clustering that is used in the CE. expenditure estimates from the CE Survey are for various domains of interest, we computed the 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.
expenditures, and for each of the following domains: Age of Reference Person, Region, Size of CU, Composition of Household, 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.

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

In addition, ratios for all CU’s, i.e., the total across the domains, 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 ratios for calwts0 and calwts1 by age of reference person is given in Figure 2. Calwts1 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 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 domains expenditures, and calwts1 is somewhat better than Calwts2 and calwts3, which used family income as one of the auxiliaries, had somewhat performance for domains, sometimes making improvements over PP but occasionally showning losses. This is connected to the nature of the income variable which had a substantial number with negative and zero values. These CU’s usefulness of this variable in predicting expenditure.

Taking all of the above into consideration, calwts1 and calwts4 can be deemed a clear improvement over the PP estimator. Calwts1 has the advantage of negative weights over regwts1. Since calwts4 relies on auxiliary variables as opposed to calwts1’s recommend calwts1 over all the other types of weights we have considered.

4. CONCLUSION AND FUTURE RESEARCH

The objective of this study was to identify 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 process presented and their relative merits evaluated. The estimation incorporates the current survey stratification methods in which the weighted sum 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 adjusts for each sample household a weight that does not depend on any single one of its members. In order to eliminate undesirable negative weights that can result from least-squares regression estimation, calibration weights were adapted to the present problem. The estimation procedure has the flexibility to restrict the possible deviation of each final weight from its base while adhering to the properties discussed above.

Particular allows the constraint of positive weights, and calibration weights are easily computed via matrix software like S-Plus™.

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


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