Composite Estimation Techniques Used for the CPIR Weights

Harry Marks, Bureau of Labor Statistics

The consumer price index (CPI) is a Layspeyres index which is used to measure the change of prices for a set of items whose quality and quantity are fixed over a period of time. An item stratum is defined as a specified set of items, goods or services that can be purchased in the retail market at time t=0 by a specified set of consumer units. Let Q_i be the quantity of all items in the i^{th} item stratum purchased at time t=0 and let P_{it} be the average price of all items in the item stratum at time t. Then the index, I_T, at time t=T can be written

$$(1) \quad I_T = \frac{\sum_{i \in I} Q_i P_{iT}}{\sum_{i \in I} Q_i P_{io}}$$

where the sum extends over the set of item strata denoted by I. This can be rewritten as:

$$(2) \quad I_T = \frac{\sum_{i \in I} (Q_i P_{io}) P_{iT}/P_{io}}{\sum_{i \in I} (Q_i P_{io})}$$

where the term $(Q_i P_{io})$ represents the total expenditure at time t=0 for item stratum i (for a given population of consumers) and the term P_{it}/P_{io} is the price "relative" for item stratum i between time t=0 and t=T.

This paper presents the methodology used by BLS to estimate the total expenditures $Q_i P_{io} = C_i$, for each item stratum, hereafter referred to as "cost weights".

I. Structure of the Cost Weights

The country was divided into "4" regions corresponding to the Census definition of North East, South, North Central, and West. Each region was further divided into 9 or 11 mutually exclusive pricing areas so that there are 40 index or local publication areas. An index is computed for two definitions of consumer units, hereafter referred to as populations. One population consisted of all urban consumer units, representing approximately 80 percent of the total U.S. population. The other population consisted of urban wage and clerical worker consumer units, representing between 35-40 percent of the total U.S. population. The set of item strata was partitioned and the elements of the partition are referred to as expenditure classes (EC's). Cost weights were needed for each item stratum and for each expenditure class within each index publication area for each population.

II. Estimation of the Cost Weights

A household survey called the 1972-73 consumer expenditure survey was conducted by the Census Bureau for the Bureau of Labor Statistics. This survey provided the data from which average annual expenditures were computed for the two populations of consumer units for each item stratum in the 40 index areas. For the i^{th} item stratum and the p^{th} index area within the r^{th} region, denote this quantity (average annual expenditures) by $\bar{X}_{ipr}$. If W_{pr} denotes the proportion of consumer units in the region for the p^{th} index area, then

$$(3) \quad \bar{X}_{ir} = \sum_p W_{pr} \bar{X}_{ipr}$$

is the average annual expenditure for the i^{th} item stratum in the r^{th} region.

Generally, a decrease in the mean square error (MSE) of the index could be achieved by taking a weighted average of the regional and index area average expenditures for the item stratum or EC. Using this weighted average times the number of consumer units in the index area, the expenditure (i.e., cost weight) for the p^{th} index area was estimated. A factor influencing the decision to use a weighted average was that the CV's (coefficient of variation) of the item stratum cost weights at the index area level were estimated to be between .1 to .9. This method of taking a weighted average is called "composite estimation". More precisely, the "composite estimator" of the average expenditure for the i^{th} item stratum for the p^{th} index area was determined by estimating the value of b_{ipr} that minimizes the MSE of $\bar{X}'_{ipr}$, where $\bar{X}'_{ipr}$ is defined as:

$$\bar{X}'_{ipr} = b_{ipr}\bar{X}_{ir} + (1-b_{ipr})\bar{X}_{ipr}$$

Let σ^2_{ipr} be the variance of $\bar{X}_{ipr}$, and let σ^2_{ir} be the variance of $\bar{X}_{ir}$. Further let $B^2_{ipr} = \{E(\bar{X}_{ipr}) - E(\bar{X}_{ir})\}^2$. Then the mean square error of $\bar{X}'_{ipr}$ is:

$$(5) \quad MSE(\bar{X}'_{ipr}) = b^2_{ipr}(B^2_{ipr} + \sigma^2_{ipr} + \sigma^2_{ir} - 2W_{pr}\sigma^2_{ipr}) - 2b(\sigma^2_{ipr}(1 - W_{pr})) + \sigma^2_{ipr}$$

where $cov(\bar{X}_{ipr}, \bar{X}_{ir}) = W_{pr}\sigma^2_{ipr}$.

The value of b_{ipr} that minimizes MSE $(\bar{X}'_{ipr})$ is seen to be:

$$(6) \quad b_{ipr} = \frac{\sigma^2_{ipr}(1 - W_{pr})}{B^2_{ipr} + \sigma^2_{ipr}(1-2W_{pr}) + \sigma^2_{ir}}$$

In order to estimate b_{ipr} it was necessary to evaluate σ^2_{ipr} and B^2_{ipr}. Variances were estimated from the consumer expenditure survey not for the item stratum, but in general for classes of items that were subsets of the item stratum. The exception was for the expenditure class, for which estimate of variances were made. Thus, it was necessary to "generalize" the variances, using regression methods so that estimates of variance could be made for the item strata.

Specifically, rel-variances were computed by dividing the variances by the square of the mean expenditure. For each region, for each population and for each EC, the generalization was made on the estimated "unit" rel-variance of the respondents with positive expenditure. That is, if we let V^2_g be the measured rel-variance for item g from the CEX survey, q_g the percentage of responses with positive expenditure for the appropriate population, and m the sample size of the survey, then

$$(7) \quad V^2_{ug} = m\, q_g\, V^2_g - (1 - q_g).$$

which is the "unit" rel-variance of set of units with positive response, assuming a random sample from a universe with q_g percentage of positive expenditure. This was computed and regressed against the average expenditure of those respondents with positive expenditures.

The regressions were done independently for self-representing SR primary sampling units (PSU) and non-self-representing NSR PSU's. PSU are groups of contiguous counties, selected as the sample areas for the consumer expenditure survey. For the SR PSU's, the unit rel-variances for consumer units with positive expenditures were computed using equation (7), i.e.,

$$(8) \quad V^2_{uipr} = m_{pr}\, q_{ipr}\, V^2_{ipr} - (1 - q_{ipr})$$

These values were regressed using the average expenditure of those units with positive expenditure as the independent variable, and then the generalized rel-variance was computed by:

$$(9) \quad \hat{V}^2_{ipr} = (\hat{V}^2_{uipr} + 1 - q_{ipr})/m_p\, q_{ipr},$$

where $\hat{V}^2_{uipr}$ is the regressed value.

For the NSR PSU's, or more precisely, for those publication areas with NSR PSU's, the between PSU rel-variance was estimated by generalizing the unit total rel-variance, $\hat{V}^2_{tipr}$ and the within unit PSU rel-variance in the same manner as described above. Then, by subtraction the unit between PSU rel-variance was computed. The intra-PSU correlation was computed and restricted to be a non-negative number which is represented by d in equation (10). Finally the estimated rel-variance was computed as:

$$(10) \quad \hat{V}^2_{ipr} = \frac{\hat{V}^2_{tipr}\,(1 + d(f_p\, \bar{n}_p - 1))}{m_{pr} q_{ipr}}$$

where $\bar{n}_p$ was the average sample size in the NSR-CEX PSU's in the publication area, and f_p was the ratio of the number of CEX PSU divided by the number of NSR-CES PSU's in the publication area. The generalization was made like this because f_p was not always equal to 1 and the SR PSU's in the publication areas with NSR PSU's were different in size from the SR PSU's that represented large metropolitan areas. In characteristics of importance they were more like the NSR PSU's.

In general, very small multiple regression coefficients were observed, so that for the generalized value of the unit rel-variance of the units with positive expenditure, the average value was taken. The exception was for those EC's that contained food items. Here the correlations were larger, and the prediction at the EC level was much closer to the actual computed value using the independent variable.

The estimated variance was computed as: $\hat{\sigma}^2_{ipr} = \bar{X}^2_{ipr} \hat{V}^2_{ipr}$ and these estimates of σ^2_{ir}' were computed as:

$$(11) \quad \hat{\sigma}^2_{ir} = \sum_p W^2_{pr} \hat{\sigma}^2_{ipr}$$

Next an estimate of B^2_{ir} was made. The number B^2_{ir} can be thought of as a between area variance and the resulting estimate can be interpreted as an estimate of an a-priori variance on the parameter $E(\bar{X}_{ipr})$. Thus, the resulting composite estimate could be viewed as an empirical Bayesian estimator with the adjustment of $(1 - W_{pr})^{(1)}$. A weighted analysis of variance with weights W_{pr} was done in order to measure the average bias. This was accomplished by first estimating the "intra-market basket" correlation, and then applying a factor based on this correlation to the estimated within market basket unit variance. The steps involved in this procedure are described below:

1) The "between place sum of squares adjusted by the mean" was computed within the region i.e.,

$$(12) \quad S^2_{ir} = \sum_p W_{pr} (\bar{X}_{ipr} - \bar{X}_{ir})^2$$

from this was subtracted

$$(13) \quad Z^2_{ir} = \sum_p W_{pr} (1 - W_{pr})\, \hat{\sigma}^2_{ipr}$$

to get the true between sum of squares (unbiased),

$$(14) \quad B^2_{ir} = S^2_{ir} - Z^2_{ir}.$$

where $\hat{\sigma}^2_{ipr}$ was the estimated variance for the i^{th} item, p^{th} area, r^{th} region.

(2) The weighted average of the within unit variance was computed using:

$$(15) \quad U^2_{ir} = \sum_p W_{pr}\, \hat{\sigma}^2_{iupr},$$

where $\hat{\sigma}^2_{iupr} = m_p\, \hat{\sigma}^2_{ipr}$, for the SR PSU's, and $\hat{\sigma}_{iupr} = \bar{X}^2_{ipr} \hat{V}^2_{tipr}$

where $\hat{V}^2_{tipr}$ is defined in equation (10), for publication areas with NSR CEX PSU's. The sum here could be thought of as a "within unit variance" for the r^{th} region and depends on the survey design. The "intra" area correlation was computed using:

$$(16) \quad d_{ir} = \frac{B^2_{ir}}{B^2_{ir} + U^2_{ir}}$$

3) These d_{ir}'s were averaged over a set of item strata and across regions. The set of items chosen are items in the same "major group", e.g., all food items, all clothes items, all transportation items, etc. Since the statistics d_{ir} is

an unstable statistic , and skewed to the left, it was felt by averaging the d_{ir}'s over a set of items that have similar economic characteristics the mean square error of intra-area correlation for a particular item would be reduced. The average value was labeled d'. The intra index area correlation was then computed as:

$$(17) \qquad d = \max (0, \ d')$$

since the intra-area correlation was assumed to be a positive number. The estimate of B^2_{ir} was:

$$(18) \qquad \hat{B}^2_{ir} = \frac{d}{1-d} U^2_{ir}$$

From equation (6), using estimates of $\hat{\sigma}^2_{ipr}, \hat{\sigma}^2_{ir}$ and B^2_{ir} we have

$$(19) \quad b_{ipr} = \min \left\{ 1, \ \frac{\hat{\sigma}^2_{ipr}(1 - W_{pr})}{B^2_{ipr} + \hat{\sigma}^2_{ipr}(1-2W_{pr}) + \hat{\sigma}^2_{ir}} \right\}$$

and using this value of $\hat{b}_{ipr}$, we computed the "initial" composite estimate:

$$(20) \quad \bar{X}'_{ipr} = \hat{b}_{ipr} \bar{X}_{ir} + (1 - \hat{b}_{ipr}) \bar{X}_{ipr} .$$

In order to estimate the mean square error of the composite estimate, it must be taken into account that $\hat{b}_{ipr}$ is an estimate and not a known number. Let $\hat{b}_{ipr} = \gamma_{ipr} b_{ipr}$, where γ is an unknown factor. If we compute the MSE of $\bar{X}'_{ipr}$, using the value of b_{ipr}, we have

$$(21) \quad MSE(\bar{X}'_{ipr}, \hat{b}_{ipr}) = \sigma^2_{ipr}(1 - b_{ipr}(1-W_{pr})(2\gamma - \gamma^2))$$

The above equation shows that

$$MSE(\bar{X}'_{ipr}, \hat{b}_{ipr}) > \sigma^2_{ipr} \text{ iff}$$

$$\gamma > 2, \text{ i.e., } \hat{b}_{ipr} > 2b_{ipr} .$$

Thus, overestimates of b are more "dangerous" than underestimates of b. The estimates were adjusted so as not to be outside a "confidence" interval of $\bar{X}_{ipr}$, i.e. The final estimate took the form:

$$(22) \quad \bar{X}''_{ipr} = \bar{X}'_{ipr} \text{ if } (\bar{X}'_{ipr} - \bar{X}_{ipr}) < k_o \hat{\sigma}_{ipr}$$

$$\bar{X}_{ipr} - \text{sgn}(\bar{X}_{ipr} - \bar{X}'_{ipr}), \text{ otherwise}$$

where

$$\text{sgn}(u) = \begin{cases} +1 & \text{if } u > o \\ o & u = o \\ -1 & \text{if } u < o \end{cases}$$

and k_o was chosen after the following computations and analysis were completed:

1) First, k_{ipr} was computed for each item stratum and publication area where:

$$(23) \quad k_{ipr} = \frac{|\bar{X}_{ipr} - \bar{X}'_{ipr}|}{\hat{\sigma}^2_{ipr}}$$

We next let $P_b(k_o)$ be the probability that $k<k_o$. This was estimated by estimating the c.d.f. of k_{ipr} over all items defined in a major group (for each population separately) as a function of b, i.e., the c.d.f. was computed over sets of items for $\hat{b}$ ranging from (0, 0.25), (0.25, 0.5), (0.5, 0.75)and (0.75, 0.95)and (0.95, 1) . An estimate of the MSE $(\bar{X}'_{ipr})$ was made by

$$(24) MSE_1(\bar{X}'_{ipr}) = p_{\hat{b}}(k_o) MSE(\bar{X}'_{ipr}, \hat{b}) + (1 - p_{\hat{b}}(k_o)) (1 + k_o^2) \hat{\sigma}^2_{ipr}$$

where MSE $(\bar{X}_{ipr}, b)$ is defined in equation (21).

To estimate γ , item strata were pooled in the same major group, and for these items, the value of b_{ipr} was taken to be the average of the b_{ipr} for item stratum within the major group; b.

$$(25) \qquad \gamma_{ipr} = b_{ipr}/\bar{b}$$

The resulting estimate of the MSE_1 $(\bar{X}'_{ipr})$ was visualized as an overestimate of the true value of MSE $(\bar{X}''_{ipr})$ due to the method of computing γ_{ipr}, so (24) was recomputed using a value of $\gamma_{ipr} = 1$, which was defined as $MSE_2(\bar{X}'_{ipr})$. This MSE_2 is an underestimate of the true value of the MSE $(\bar{X}'_{ipr})$.

Next, the average value of $MSE_j(X'_{ipr})$ over all item stratum within a major group was computed, defined as $\overline{MSE}_j$, j = 1, 2 and graphs were made of $\overline{MSE}_j$ as a function of k_o. These graphs identified an initial value of k where the change in the MSE_j $(\bar{X}''_{ipr})$, j = 1, 2 became "small".

Next, the average percent difference and the total number of percent differences greater than 0.5 was computed to identify potential areas where a "large" number of large percent differences were occurring. Percent differences were computed by:

$$(26) \qquad W_{ipr} = \frac{|\bar{X}'_{ipr} - \bar{X}_{ipr}|}{\min (\bar{X}'_{ipr}, \bar{X}_{ipr})}$$

With the above information and discussions with the economist of BLS as to the reasonableness of the final results the values of k_o were determined. Values of k_o were between 0.5 and 2.0, and the values of MSE_j (b, k_o)/MSE_j (b=0), were 0.75 for the food items, i.e. a 25 percent reduction in the average mean square error, and for the commodity and service items the ratios ranged from 0.85 - 0.90, i.e. a 10-15% reduction in mean square error. A method of simulation will be used to further evaluate the properties of the composite estimator.

After the k_o were chosen, the final composite estimate made for the area was multiplied by the number of consumer units in the area as measured from the CEX survey, so that an estimate of total expenditure was made. A final procedure called "raking" was performed, which involved ensuring that the sum of the total expenditures as estimated using the composite estimate for the

areas in a region equal to the total expenditure in the region using the preliminary estimate, at the item stratum level; and that the item stratum expenditures using composite estimation equaled the EC expenditure using composite estimation within a publication area. This procedure involves an iterative process, which did not stop until the boundry conditions with respect to the marginals were satisfied to within 1×10^{-4}.

III. Method of Evaluation

The method to be used for measuring the reduction in the mean square error, and also to determine the distribution of some of the variables used in the estimation procedure will be that of replication. Up to 36 replicates will be formed. For the i^{th} replicate the index or other statistics will be computed, say, X_i, and the variance of the statistic x, will be estimated by

$$(27) \quad \text{Var}(X) = \frac{1}{N} \sum_{i=1}^{N} (X_i - X)^2$$

where N = the number of replicates.

An estimate of the bias that is introduced by using the composite estimate will be made by computing the statistic with the composite estimate say X_{Li} and the statistic without the composite estimate, say X_{Wi}, for the i^{th} replicate. Let X_L, X_W be the statistics computed using the whole sample. An estimate of bias squared, B^2, will be computed as:

$$(28) \quad (X_L - X_W)^2 - \text{Var}(X_L - X_W)$$

where $\text{Var}(X_L - X_W)$ is computed by formula (26), where $X = X_L - X_W$. If we take X_W to be an unbiased estimate, the difference in MSE of X_L and X_W will then be computed.

These measurements will allow us to estimate the distribution of the various parameters used for the composite estimator. With this we will measure the robustness of the composite estimator with respect to these variables, and this in turn will help us identify areas where improvements would be most beneficial. For example, what effect does the "generalized variances" have on the final estimate, or what effect does the intra-Index area correlation have on the estimator; with respect to the criteria of decreasing the MSE.

IV. Further Research

Questions have been asked concerning the information that is being used to help improve the estimate of cost weights. Specifically, does there exist other demographic or socio-economic variables associated with the Index area, or even the PSU that could be used in improving the estimation of the cost weights? For example, instead of taking a weighted average between the regional mean expenditure and the place mean expenditure, one could take a weighted average between a regressed mean expenditure for the place and the place mean expenditure.

Following Hansen and Madow,[2] the problem is to find a constant, b, such that the average MSE of the X'_{ipr} within the region is minimized, where

$$(29) \quad X'_{ipr} = b\,\bar{X}_{ipr} + (1 - b)\hat{X}_{ipr}$$

and where $\hat{X}_{ipr}$ is the regressed value of $\bar{X}_{ipr}$, when regressing $\bar{X}_{ipr}$ onto a set of known independent variables within the region. For mathematical simplicity, and since the gains made by adding an extra variable with respect to the average of mean square error are correlated with the gains with respect to the average decrease in mean square error for the particular places, we are using the criteria of minimizing the average of mean square errors. Further, for simplicity of notation, let us assume that the value of the variance of X_{ipr} is one (which can be achieved by dividing X_{ipr} by σ^2_{ipr}).

Let $\bar{X}_{ir}$ be the NX1 vector with components $\bar{X}_{ipr}$ and let $\hat{X}_{ir}$ be the NX1 vector of regressed values where N is the number of places in the region, and let $X'_{ir} = b\bar{X}_{ir} + (1-b)\hat{X}_{ir}$. Let Z denote the design matrix of known independent variables (measured without error);, then

$$(30) \quad \hat{X}_{ir} = Z\,\hat{\beta}_{q+1,ir}$$

where $\hat{\beta}_{q+1}$ are the regression coefficients

$$(31) \quad \hat{\beta}_{q+1} = (Z^t Z)^{-1} Z^t \bar{X}_{ir} \quad Z$$

and that $(Z^t Z)$ is of full rank $= q + 1$. Further, let $\mu_{ir} = E(\bar{X}_{ir})$ and let $\hat{\mu}_{ir} = E(\hat{X}_{ir})$. Then, we are looking for a value of b which minimizes

$$(32) \quad E(X'_{ir} - \mu_{ir})^t (X'_{ir} - \mu_{ir})$$
$$= b^2 E(\hat{X}_{ir} - \bar{X}_{ir})^t (\hat{X}_{ir} - \bar{X}_{ir}) + 2bE(\hat{X}_{ir} - \bar{X}_{ir})(\bar{X}_{ir} - \mu_{ir}) + N$$

The value of b which minimizes (28) is

$$(33) \quad b_{opt} = \frac{E(\bar{X}_{ir} - \hat{X}_{ir})^t (\bar{X}_{ir} - \mu_{ir})}{E(\hat{X}_{ir} - \bar{X}_{ir})^t (\hat{X}_{ir} - \bar{X}_{ir})}$$

and the reduction is mean square error is

$$(34) \quad R = E(\bar{X}_{ir} - \mu)^t (X_{pr} - \mu) - E(X'_{pr} - \mu)^t (X'_{pr} - \mu)$$
$$= \frac{[E(\bar{X}_{ir} - \hat{X}_{ir})^t (X_{ir} - \mu_{ir}]^2}{E(\hat{X}_{ir} - \bar{X}_{ir})^t (\hat{X}_{ir} - \bar{X}_{ir})}$$

The numerator of b_{opt} can be written,

$$(35) \quad E(\bar{X}_{ir} - \hat{X}_{ir})^t (\bar{X} - \mu_{ir})$$
$$= E[(\bar{X}_{ir} - \mu_{ir}) - (\hat{X}_{ir} - \mu)]^t (\bar{X}_{ir} - \mu_{ir})$$
$$= N - E[(\hat{X}_{ir} - \hat{\mu}_{ir})^t (\bar{X}_{ir} - \mu_{ir})$$
$$= N - E[(\bar{X}_{ir} - \bar{\mu}_{ir})^t Z (Z^t Z)^{-1} Z^t (\bar{X}_{ir} - \mu_{ir})]$$
$$= N - q - 1$$

since $Z(Z^t Z)^{-1} Z^t$ is of rank $q + 1$ and is idem-

potent. The denominator of b_{opt} can be written

$$(36)\quad E((\hat{X}_{ir}-\bar{X}_{ir})^t(\hat{X}_{ir}-\bar{X}_{ir}) = N - (q+1) + N(1 - R^2)\Sigma(\mu_{ipr} - \mu_{ir})^2/N$$

where $R_q 2$ is the multiple regression coefficient when regressing the expected value of $\bar{X}_{ipr}$ onto z. The term $\Sigma_p(\mu_{ipr} - \mu_{ir})^2/N$ was estimated, by equation (18) in the general case, and for the assumptions we have made, namely that $\sigma^2_{ipr} = 1$, we have that

$$(37)\quad \Sigma(\mu_{ipr} - \mu_{ir})^2 = \frac{\delta}{1-\delta} \cdot n$$

where n is the sample size in each place (assumed equal here for sake of simplicity) and δ is the true intra-Index area correlation. Thus the optimal value of b is:

$$(38)\quad b_{opt} = \frac{N - q - 1}{N - q - 1 + \frac{N\ n\ \delta}{1-\delta}(1 - R_q 2)}$$

and the reduction in mean square error is:

$$(39)\quad T_q = \frac{(N - q - 1)^2}{(N - q - 1) + \frac{N\ n\ \delta}{1-\delta}(1 - R^2)}$$

where T_q denotes the reduction, where there are q independent variables (excluding the constant).

The ratio of T_q to T_o is:

$$(40)\quad T_q/T_o = \left(\frac{(N-q-1)}{N-1}\right)^2 \frac{N - 1 + \frac{N\,n\,\delta}{1-\delta}}{N - q - 1 + \frac{N\,n\delta}{1-\delta}(1 - R_q 2)}$$

In order to benefit by using the additional q independent variables, we want

$T_q/T_o \geq 1$, or that

$$(41)\quad R_q 2 \geq 1 + \frac{N - q - 1}{N - 1}\left(\frac{(N-1)(q - \theta) + q\theta}{(N-1)\theta}\right)$$

where $\theta = \frac{Nn\theta}{1-\delta}$

If we put N = 10, q = 1, n equal to a typical value of 500, and $\delta = 1 \times 10^{-3}$ so that $\theta = 5$, then the right side of equation (41) is .387, i.e., in order to gain using one independent variable the true multiple regression coefficient must be $\geq \sqrt{.387} = .623$.

Measurement of multiple regression coefficient for various variables are being planned to be done:

In conclusion questions that need to be answered are

1) What is the effect of the MSE of cost weights on the MSE of the index, and the effect of composite estimation on the MSE of the index? Some work has begun in this area, and will be a subject of a future paper.

2) What improvement can be made to the estimates of the parameters of the composite estimator, and what effect does this improvement have on the MSE of the cost weights?

3) What variables can aid in decreasing the MSE of the cost weights?

References

1) Efron, Bradley and Morris, Carl, 1975, "Data Analysis Using Stein Estimator and Its Generalizations", Journal of the American Statistical Association pp. 311-319

2) Madow, William G. and Hansen, Morris, "On Statistical Models and Estimation in Sample Surveys", Unpublished paper