EVALUATION OF COMPOSITE ESTIMATION METHODS FOR COST WEIGHTS IN THE CPI

David C. Swanson, Sharon K. Hauge, Mary Lynn Schmidt, U.S. Bureau of Labor Statistics
David C. Swanson, 2 Mass. Ave., NE, Room 3655, Washington, DC 20212

Key Words: composite estimation, sample size, mean squared error

Any opinions expressed in this paper are those of the authors and do not constitute policy of the U.S. Bureau of Labor Statistics.

Introduction

The Consumer Price Index (CPI) uses data from the Consumer Expenditure Survey to weight individual area/item price indexes in order to aggregate them to higher-level price indexes. The accuracy of the weights of local area indexes is improved through a process called *composite estimation*, in which strength is borrowed from a larger geographic area. Larger geographic areas have similar expenditure patterns to those of local areas, but the patterns are more stable because they come from a larger sample. In this paper we describe the method of composite estimation used in the CPI's 1998 Revision, and then summarize research to improve the composite estimation process conducted over the past ten years at the U.S. Bureau of Labor Statistics.

Background

The CPI is one of the most important economic indicators available in the United States. It measures the change in prices of goods and services that are purchased by American consumers. The goods and services currently represented in the CPI are those purchased by urban Americans during the 3-year period 1993-95.

The samples and weights used to compute the CPI are updated periodically to reflect changing consumer expenditure patterns and changing demographics. In the past these updates occurred approximately every ten years, but in the future they will occur more often. The weights come from another survey conducted by the U.S. Bureau of Labor Statistics called the *Consumer Expenditure Survey*, and are updated every time the CPI is revised.

To produce the CPI, price data are collected every month from a random sample of stores in thirty-eight geographic areas across the United States. These geographic areas are called *index areas*. Examples of index areas are the Boston metropolitan area, the St. Louis metropolitan area, and the San Francisco metropolitan area. Within each of the 38 index areas price data are collected for 211 item categories called *item strata*. Together the 211 item strata cover all consumer purchases. Examples of item strata are Bananas, Women's Dresses, and Electricity.

Multiplying the number of index areas by the number of item strata gives us 8,018 ($= 38 \times 211$) different area/item combinations for which data need to be collected. Price indexes are calculated for each one of these 8,018 area/item combinations, and then the indexes are aggregated to form higher-level price indexes using expenditure estimates from the Consumer Expenditure Survey as their weights. Because of small sample sizes within individual index areas, composite estimation is used to increase the accuracy of the weights.

Current Method of Composite Estimation

In the CPI's 1998 revision composite estimation of Consumer Expenditure Survey data was performed in the following way:

First, expenditure estimates were obtained from the Consumer Expenditure Survey for the 3-year period 1993-95 for each of the 8,018 area/item combinations. Then each item stratum's *relative importance* was computed. Relative importance is the total expenditure made on a particular item stratum divided by the total expenditure made on *all* item strata. In other words, it is the proportion of total consumer expenditures made on a specific item stratum.

For example, according to the Consumer Expenditure Survey, consumers in the Boston metropolitan area spent $463 million on women's dresses during the 3-year period 1993-95, and $224 billion on all items combined. Dividing $463 million by $224 billion gives a relative importance of 0.0021, meaning that consumers in the Boston

metropolitan area spent 0.21 percent of their total expenditures on women's dresses.

Relative importances were computed by index area, and also by *major area*. For the purpose of composite estimation the United States was divided into 8 major areas, which are the 4 Census regions (Northeast, Midwest, South, West) cross-classified by the sets of self-representing and non-self-representing index areas.

After computing relative importances for each item stratum within each index area and major area, the composite estimation process simply involved replacing every index area relative importance by a weighted average of the index area and major area relative importances. The weighted average is expected to be a more accurate estimate of the index area's true relative importance because the index area and major area have similar expenditure patterns, but the major area has a larger sample size.

For example, in the Boston metropolitan area consumers spent 0.21 percent of their total expenditures on women's dresses. In all self-representing areas in the Northeast Region they spent 0.32 percent on women's dresses. In the composite estimation process Boston's estimate of 0.0021 is replaced by a weighted average of the two estimates: 0.0021 is replaced by $\alpha 0.0032 + (1-\alpha)0.0021$ where $0 \le \alpha \le 1$. Of course the entire Northeast Region has a larger sample size than the Boston metropolitan area alone, so the weighted average is expected to have less variability than Boston's original estimate.

Mathematically, let x be an item stratum's relative importance estimate for a major area, and y be its estimate for an individual index area within that major area. Then in composite estimation y is replaced by $y^* = \alpha x + (1-\alpha)y$, where $0 \le \alpha \le 1$. This is a weighted average of the index area and major area relative importance estimates. The weight α is chosen to be the number between 0 and 1 that minimizes the mean squared error (MSE) of y*.

Although MSEs are not exactly the same as variances, they are very similar. To give the basic idea of how the formula for α is derived, the following shows how to find the value of α that minimizes the variance of y*:

$$
\begin{aligned}
V(y^*) &= V(\alpha x + (1-\alpha)y) \\
&= \alpha^2 V(x) + (1-\alpha)^2 V(y) + 2\alpha(1-\alpha)Cov(x,y) \\
&= \alpha^2[V(x) + V(y) - 2Cov(x,y)] - \\
&\quad 2\alpha[V(y) - Cov(x,y)] + V(y) \\
&= \alpha^2 V(y-x) - 2\alpha[V(y) - Cov(x,y)] + V(y)
\end{aligned}
$$

Then to minimize the variance of y*, the derivative is taken with respect to α and set equal to 0:

$$\frac{dV(y^*)}{d\alpha} = 2\alpha V(y-x) - 2[V(y) - Cov(x,y)]$$

$$\frac{dV(y^*)}{d\alpha} = 0 \quad \text{implies} \quad \alpha = \frac{V(y) - Cov(x,y)}{V(y-x)}$$

This shows how to obtain the value of α that minimizes the *variance* of y*. When *MSE* is minimized the formula becomes the following:

$$\alpha = \frac{V(y) - Cov(x,y)}{E[(y-x)^2]}$$

When we look at all 8,018 area/item combinations, the median reduction in root mean squared error ($\sqrt{MSE}$, or RMSE) achieved by this method of composite estimation was 14 percent, so the method was fairly effective. It was also relatively easy to implement.

Research to Improve Composite Estimation

Although the current method of composite estimation worked fairly well, ten years ago the CPI program started to look for ways to improve the process even further. This led to a research program in which a number of alternative methods and models were developed. They are briefly described below.

First is a nearest neighbor method in which a weighted average of the relative importances from the index area and the index areas geographically closest to it is used. After that, a multivariate generalization of the current method is described, followed by several Bayesian models, and then finally a method in which the inputs to the current model are improved.

For more details on these methods, see the articles listed in the bibliography at the end of this paper.

Nearest Neighbor Method

This is the same as the current method, but with one slight change: instead of compositing an index area with its major area, the index area is composited with the index areas that are geographically closest

to it. For example, Baltimore is in the South region, which includes many index areas that are far from Baltimore, such as Miami, New Orleans, Houston, and Dallas. Baltimore's nearest neighbors, however, are mostly in the Northeast region, such as Philadelphia and New York. In the nearest neighbor method, instead of compositing Baltimore with the set of self-representing index areas in the South, Baltimore is composited with the set of index areas that are geographically closest to it. An index area's expenditure patterns are expected to be closer to its nearest neighbors than they are to those of distant index areas that just happen to be assigned to the same region of the country.

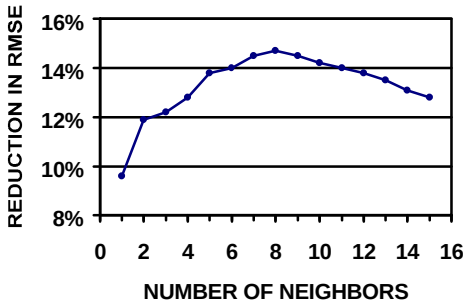
Here is an example. Baltimore's 2 nearest neighbors are Washington, DC and Philadelphia. The total expenditures on women's dresses and all items combined during the 3-year period 1993-95 are shown below. According to the Consumer Expenditure Survey, people in the Baltimore metropolitan area spent $230 million on women's dresses, and $80,395 million on all items. Thus women's dresses represented 0.29 percent of total expenditures in Baltimore. Similar numbers are shown for Washington, DC, Philadelphia, and all 3 areas combined.

Total Expenditures, 1993-95 ($ millions)

Index Area	Women's Dresses	All Items	Relative Importance
Baltimore	230	80,395	.0029
Washington, DC	557	180,951	.0031
Philadelphia	718	197,160	.0036
Total	**1,505**	**458,506**	**.0033**

The composite estimate of relative importance for Baltimore is then computed as $\alpha 0.0033 + (1-\alpha)0.0029$, where $0 \le \alpha \le 1$. The number α is computed the same way as in the current method – by taking the derivative of the MSE with respect to α, setting the derivative equal to 0, and then solving for α. In this example α turned out to be 0.59. When α=0.59 the composite estimate becomes 0.0031, which means that the new expenditure estimate for women's dresses in Baltimore is 0.0031 × $80,395 million = $249 million. Thus $249 million is the new (and presumably more accurate) estimate of total expenditures on women's dresses in the Baltimore metropolitan area for the 1993-95 period.

In the example above Baltimore was composited with its 2 nearest neighbors. Of course it could have been composited with its 3 nearest neighbors, its 4 nearest neighbors, and so on. A graph comparing the median reduction in RMSE ($\sqrt{\text{MSE}}$) for item stratum/index area relative importances against the number of neighbors used is shown below.



The graph shows that the optimal number of neighbors to use is 8. When 8 neighbors are used, the median reduction in RMSE is 15 percent, which is 1 percentage point better than the current method.

A Multivariate Generalization of the Current Method

When we look at the formula for α in the current method of composite estimation, it is clear that it takes into consideration covariances between index areas, but ignores covariances between item strata. In 1992 a multivariate procedure similar to the current method was developed in which covariances between different item strata were taken into consideration as well. This generalization of the current method naturally produced superior results. The median reduction in RMSE was 14 percent for the current method, 15 percent for the nearest neighbor method, and 18 percent for this multivariate method. Thus the multivariate method worked a little better than the current method.

The problem with the multivariate method was that it required calculating some matrix inverses, and whenever one or more of the item stratum/index area combinations had no reported expenditures in it, the matrices could not be inverted. In those situations the method broke down. As a result more research was needed.

Bayesian Models

Eight different Bayesian models were proposed to improve the direct survey estimates of expenditure. Two of them are described here.

In 1992 a linear empirical Bayes model was proposed to estimate the true expenditure. The basic idea of the model was that the true expenditure can be modeled as a linear combination of other variables, $\theta = \mathbf{x}^{\mathbf{T}}\beta + \varepsilon \quad = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$, where θ is the true expenditure, the x_i's are some exogenous variables, the β_i's are coefficients of those variables, and ε is a random error term with mean 0.

For example, the total expenditure on women's dresses might be modeled as a linear function of the number of women living in the index area ($\theta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where x_1 is the number of women aged 15-24, x_2 is the number of women aged 25-65, and x_3 is the number of women over 65); or the total expenditure on new cars might be modeled as a linear function of the number of families in various income categories ($\theta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, where x_1 is the number of families with incomes between \$0 and \$30,000, x_2 is the number of families with incomes between \$30,001 and \$75,000, and x_3 is the number of families with incomes over \$75,000).

Then under certain Bayesian assumptions the true expenditure was estimated to be a weighted average of the direct survey estimator and the linear model of it:

$$\hat{\theta} = \alpha y + (1-\alpha)\mathbf{x}^{\mathbf{T}}\beta, \text{ where } \alpha = \frac{V(\theta)}{V(\theta) + V(y \mid \theta)}$$

Of course this model assumes that we have some outside variables that produce good estimates of expenditure. However, the problem with the model is that finding such variables can be difficult, and using them to produce official expenditure estimates is rather difficult to justify at a survey organization such as the Bureau of Labor Statistics. As a result, we did not pursue this model any further.

In the same year a hierarchical Bayes model was proposed to estimate expenditures. Let θ_{kj} be the true total expenditure for item stratum k in index area j, let Y_{kj} be the direct survey estimator of the total expenditure, and let $\theta_j = [\theta_{1j}, \ldots, \theta_{Kj}]^T$ and $Y_j = [Y_{1j}, \ldots, Y_{Kj}]^T$ be vectors of those quantities. Also let us make the following assumptions:

(i) Conditional on the true expenditure θ_j, the Y_j's are independent with $Y_j \mid \theta_j \sim N(\theta_j, V_j)$

(ii) Conditional on $\mu \in R^K$ and $0 < r_0 < \infty$, the θ_j's are independent with $\theta_j \mid \mu, r_0 \sim N(\mu, r_0^{-1} I)$

(iii) The prior density function of μ and r_0 is given by $\pi(\mu, r_0) \propto r_0^{\frac{1}{2}f-1} \exp(-\frac{1}{2} e r_0)$

(iv) The values of V_j, f, and e are assumed to be known.

Under this model, the posterior distribution of $\theta = [\theta_1^{\mathbf{T}}, \ldots, \theta_J^{\mathbf{T}}]^{\mathbf{T}}$ given $Y = [Y_1^{\mathbf{T}}, \ldots, Y_J^{\mathbf{T}}]^{\mathbf{T}}$ and r_0 is a multivariate normal distribution with mean vector $E(\theta \mid Y, r_0) = [Q_1^T, \ldots, Q_J^T]^T$, where $Q_j = (I - V_j W_j) Y_j + V_j W_j \left(\sum_{j=1}^{m} W_j \right)^{-1} \sum_{j=1}^{m} W_j Y_j$, and where $W_j = (V_j + r_0^{-1} I)^{-1}$. The hierarchical Bayes estimator of θ was then obtained using the familiar iterated formula for expectations, $\hat{\theta} = E(\theta \mid Y) = \quad E[E(\theta \mid Y, r_0) \mid Y]$, where the expectation is computed using the posterior density of r_0 given Y, found to be:

$$g(r_0 \mid Y) \propto \left| \sum_{j=1}^{m} W_j \right|^{-\frac{1}{2}} \prod_{j=1}^{m} |W_j|^{\frac{1}{2}} \exp[-\tfrac{1}{2}\{e r_0 + \sum_{j=1}^{m} Y_j^T W_j Y_j - (\sum_{j=1}^{m} W_j Y_j)^T \left(\sum_{j=1}^{m} W_j \right)^{-1} \sum_{j=1}^{m} W_j Y_j\}]$$

Regardless of the model's effectiveness, it is clear that the model is rather hard to describe to non-statisticians, which makes it is hard to convince them of its value. It is also rather complicated to implement in a production environment. As a result, we did not pursue this model any further either.

In addition to the two models described above, six other Bayesian models were proposed. All of them were considered to be overly complex, both in terms of the ability to explain them to non-statisticians, and in terms of the ability to implement them in a production environment. As a result, none of them was pursued any further.

Current Research

As we mentioned before, the current method of composite estimation involves replacing an item stratum's relative importance for a particular index

area with a weighted average of the index area and major area relative importance estimates. That is, y is replaced by $y^* = \alpha x + (1-\alpha)y$, where y is the index area estimate, x is the major area estimate, and

$$\alpha = \frac{V(y) - Cov(x, y)}{E[(y-x)^2]}.$$

Recently a proposal was made to use the same model, but with an improved estimate of α. The method of improving the estimate of α is simply to improve the estimates of V(y), Cov(x,y), and $E[(y-x)^2]$ that go into its formula, which in turn improves the estimate of α. Specifically, the recommendation is to use *stratified* estimators of V(y), Cov(x,y), and $E[(y-x)^2]$ to improve their stability.

Currently we are evaluating this method. So far we have identified and tested over a dozen variables that were conjectured to be effective in stratifying the families in the Consumer Expenditure Survey. The variables consisted of several expenditure/income variables, several demographic variables, and several geographic variables.

After stratifying the families with these variables, we used the following formula to measure how much of the total variance in the Consumer Expenditure Survey was explained by each variable:

$$R^2 = 1 - \frac{\sum_i \sum_s \sum_{f \in s} (RI_{if} - RI_{is})^2}{\sum_i \sum_s \sum_{f \in s} (RI_{if} - RI_i)^2}$$

where RI_{if} is the relative importance of item category *i* for family *f*, RI_{is} is the relative importance of item category *i* for all families in stratum *s*, and RI_i is the relative importance of item category *i* for all families in the Consumer Expenditure Survey's sample. The numerator of this ratio is the stratified variance, and the denominator is the un-stratified variance. One minus the ratio is the proportion of total variance "explained" by the stratification variables, which is commonly called *R-squared*.

Stratifying Variable	R^2
Total family expenditures	9.4%
Total family pre-tax income	7.8%
Age	4.0%
Family Type	3.8%
Education	2.2%
Family size	2.0%
Race	0.7%
Tenure (owner/renter)	1.9%
Type of segment	1.2%
Degree urban	1.1%

The table above shows that expenditure/income variables explained the greatest amount of variance, followed by demographic variables, and then geographic variables. This means that stratifying by expenditure/income variables will probably improve the stability of α by the greatest amount, followed by demographic variables, and then geographic variables.

After measuring the stratification effect of individual variables, we also looked at whether stratifying by more than one variable would have a greater effect. For example, we looked at whether age and education together would produce better results than either variable by itself. We found that additional variables did not have a noticeable effect, so in the future we will be testing this method by stratifying with just one variable.

Summary

The current method of composite estimation works fairly well. It is easy to implement, and successfully reduces RMSE by a median amount of 14 percent. Some of the other methods reduce RMSE a little more, but they are also harder to implement, and sometimes break down. All eight Bayesian models proposed are somewhat complex, difficult to describe to non-statisticians, difficult to convince non-statisticians of their value, and difficult to implement in a production environment. Our current research focuses on returning to the current method of composite estimation and improving the stability of the weight α that goes into it. Further improvement of the current method looks promising.

References

Cohen, M. P. and Sommers, J. P. (1984). "Evaluation of Methods of composite Estimation of Cost Weights for the CPI", Proceedings, Section on Business and Economic Statistics, American Statistical Association, pp. 466-471.

Datta, G. S., and Lahiri, P. (1992). "Robust Hierarchical Bayes Estimation of Small Area Characteristics in Presence of Covariates," Technical

Report, University of Georgia, Department of Statistics.

Ghosh, M., and Sohn, S. Y. (1991). "An Empirical Bayes Approach towards Composite Estimation of Consumer Expenditure," Technical Report, University of Florida, Department of Statistics.

Lahiri, P. (1992). "Estimation of Consumer Expenditures for Small Areas – the Hierarchical Bayes Approach." Technical Report, University of Nebraska – Lincoln, Division of Statistics.

Lahiri, P., and Wang, W. (1992). "A Multivariate Procedure Towards Composite Estimation of Consumer Expenditure for the U.S. Consumer Price Index Numbers". Survey Methodology, Statistics Canada, vol. 18, no. 2, 279-292.

Lahiri, P. (1994). "Estimation of Consumer Expenditure: A Linear Empirical Bayes Approach." Technical Report, University of Nebraska – Lincoln, Division of Statistics.

U.S. Bureau of Labor Statistics, *BLS Handbook of Methods* (1997), Washington, DC, U.S. Government Printing Office.