HIERARCHICAL BAYES ESTIMATION OF VARIANCE COMPONENTS FOR THE U. S. CONSUMER PRICE INDEX

Robert M. Baskin, U.S. Bureau of Labor Statistics
2 Massachusetts Ave, N.W., Room 3655, Washington, D.C. 20212

KEY WORDS: Laspeyres index, jackknife, anova estimate.

This is a report of the first step in a multiphase project to estimate the components of variance of the Consumer Price Index (CPI) for the period January 1987 to December 1991. This covers the period since the last major revision of the CPI in December 1986. The estimates of components of variance are conditional on December 1986 expenditure weights derived from the 1982 - 1984 Consumer Expenditure Surveys. These variance components are estimated by a Hierarchical Bayes (HB) method as well as the usual anova type estimators.

1. Introduction

The Bureau of Labor Statistics (BLS) is currently making preparations for the 1997 revision of the CPI. Decisions must be made on methodology and allocation of resources for the upcoming revision and relative sizes of the components of variance will be a factor in this process. For example, selection of Primary Sampling Units (PSUs), which will be defined in section two, for the 1997 revision has been scheduled for the summer of 1992. The relative size of the PSU component of variance gives information about the importance of such activities as selection of variables for stratification of the PSUs. Similarly, as other selections arise, the relative size of the corresponding component of variance should give an indication as to the relative importance of each activity.

In this paper, the relative size of the PSU component of variance is estimated for the indices from January 1987 to December 1991. This estimation is made first by a Hierarchical Bayes (HB) method and then compared to the usual anova type of estimator. The relative size of the component of variance due to PSU is seen to be consistently small. Furthermore, in order to give an idea of the accuracy of the estimates, these estimates are jackknifed. This method produces an estimate of the variance of the aforementioned estimators. The HB estimators are seen to have a similar estimate of error to the usual type estimators. Also simulations are presented which support the effectiveness of the HB estimator.

2. The Consumer Price Index

For a full discussion of the CPI the reader is referred to Chapter 19 of the *BLS Handbook of Methods*, (1988). However, the following features of the CPI are important for the present discussion.

According to the *Handbook*, p 154, "The CPI is a measure of the average change in the prices paid by urban consumers for a fixed market basket of goods and services." It is calculated monthly for the population of all urban families and also for the population of wage earners and clerical workers. This paper calculates estimates only for the all-urban index. The CPI is estimated for the total US urban population for all consumer items, but it is also estimated at other levels defined by geographic area and groups of items. Pricing for the CPI is conducted in 94 PSUs in 91 geographic areas (New York city consists of 3 PSUs and Los Angeles consists of 2 PSUs). In the CPI area design there is random selection of PSUs according to a stratified design in which one PSU is selected from each stratum. There are four classes of PSUs. The 34 A PSUs are metropolitan statistical areas (MSAs) which because of size or unique characteristics are selected with certainty. Other MSAs are classified as either large (L) PSUs or medium (M) PSUs. Of these MSAs, 22 L PSUs and 24 M PSUs were selected for the CPI during the 1987 revision. Urban areas not included in MSAs are classified as R PSUs. The 1987 CPI contains 14 of these sampling units. The boundaries of these PSUs were defined by BLS. A description of the PSU selection can be found in Dippo and Jacobs(1983). The 34 A PSUs are referred to as certainty or self-representing PSUs. These 34 PSUs are the largest metropolitan areas. For the remaining strata, the selected PSUs are referred to as non-self-representing PSUs. The next step is to divide

the universe of goods and services into item strata. Then, within each PSU, a selection of Entry Level Items (ELIs) is made from the item strata. A single ELI is chosen from each item stratum. Examples of item strata are 1) fruit juices and frozen fruits, 2) boys apparel and 3) eyeglasses and eyecare. Examples of ELIs in the item stratum fruit juices and frozen fruits are 1) frozen orange juice, 2) other frozen fruits and fruit juices, and 3) fresh canned/or bottled juices. Note that some item strata, such as the item stratum white bread, contain only one ELI, in this case again white bread, so that these ELIs are certainties. Further sampling occurs to determine outlets in which to price the selected ELIs. Most of the sample frames used in outlet selection are derived from the Current Point of Purchase Survey (CPOPS). For each PSU a selection of outlets is made corresponding to CPOPS category. Finally, there is sampling, by field representatives, within an outlet and within a previously selected ELI in order to determine the particular product to be priced. These four stages of sampling lead to the idea that the total variance of the CPI can be decomposed into four components of variance corresponding to these four stages.

The CPI is a modified Laspeyres index, which is a ratio of the costs of purchasing a set of items of fixed quality and quantity in two different time periods. Let IX(i,m,t,0) denote the index at time t, in pricing area m, for item stratum i, relative to base period 0. Then
IX(i,m,t,0)= 100*CW(i,m,t)/CW(i,m,0)
where CW(i,m,t) and CW(i,m,0) denote cost weights, which are estimates of expenditures in index area m on item stratum i for times t and base period 0 respectively. Cost weights for item strata are updated on a monthly or bimonthly basis depending on the particular item and index area. They are summed to estimate cost weights for higher level item aggregates (HLIAs) such as all items and major groups, e.g., food, medical or transportation.

3. The Model

The CPI, as mentioned in the previous section, can be considered to have four components of variance corresponding to the four stages of sampling. In order to model variance components it is typical to write the random variable of interest as a sum of fixed components and random components with a random component corresponding to each component of variance. However, since the present work is attempting to estimate only the PSU component of variance, which comes only from the non-self-representing PSUs, relative to the rest of the variance, it will suffice to consider the index as a sum of two random components, one corresponding to PSU variance and the other corresponding to the rest of the random components. Thus we can write

$$IX(i,m,t,0) = mean(t) + a(m,t) + e(i,m,t)$$

where the mean is a fixed factor, a(m,t) is a random factor corresponding to PSU selection and e(i,m,t) is a random factor corresponding to the rest of the randomness. The assumptions on {a(m,t)} and {e(i,m,t)} are that they are mutually independent with mean 0, the a(m,t) are identically distributed with variance $\sigma_a^2(t)$ and the e(i,m,t) are identically distributed with variance $\sigma_e^2(t)$. No attempt will be made to model this as a time series so the dependence of the variance components on the parameter t will be suppressed.

Our current work is to estimate the size of the PSU component of variance, σ_a^2, relative to the rest of the variance, σ_e^2. This ratio, which will be denoted by $\Delta = \sigma_a^2 / \sigma_e^2$, is convenient to calculate. It should be pointed out that the ratio of the PSU component of variance to the total variance is a 1-1 function of Δ so that if Δ is known, then σ_a^2/σ^2 can be calculated.

The following notation will be useful in what follows. Let IX(+,m,t,0) denote the sum of the indices IX(i,m,t,0) over the item strata and let IX(.,m,t,0) denote the mean of the indices over the item strata. Similarly, let IX(.,.,t,0) denote the mean of IX(.,m,t,0) over the PSUs. The usual anova estimators of σ_a^2 and σ_e^2 are based on what are referred to as the between sum of squares (BSS)

$$BSS = \sum_{m \in M} (IX(.,m.t.0) - IX(.,.,t,0))^2$$

where M is the set of index areas in the HLGA, and the within sum of squares (WSS)

$$WSS = \sum_{m \in M} \sum_{i \in I} (IX(i,m,t,0) - IX(.,m,t,0))^2$$

where I is the set of item strata in the HLIA and M is the set of index areas in the HLGA.

The only level at which the sampling for the CPI can be considered balanced is at the item strata level, because the number of item

strata in each nonself-representing PSU is the same. Thus the estimate of σ_e^2 is

$$\hat{\sigma}_e^2 = WSS / |M| \times (|I| - 1)$$

and the estimate of σ_a^2 is

$$\hat{\sigma}_a^2 = (BSS/(|M|-1)) - WSS/(|I| \times |M| \times (|I|-1))$$

where |M| denotes the number of elements in M and |I| denotes the number of elements in I.

The form of the last estimator allows the estimate of the PSU component of variance to be negative, although the probability of this happening is guaranteed to converge to zero as the sample size increases. A discussion of this can be found in Searle, Casella and McCulloch (1992). As can be seen from the estimates actually produced, this unfortunate phenomenon does actually occur so other methods of estimation are needed in this case. First of all, if the anova estimates are negative in a balanced model, maximum likelihood and restricted maximum likelihood do not help. Among the limited options are taking the positive part of the anova estimator or using a Bayesian estimator. A Bayesian estimator under squared error loss (or any quadratic loss) is guaranteed to be nonnegative. We investigate in the present work, a Bayes estimator of the variance ratio derived under a hierarchical normal model for balanced data proposed in Datta and Ghosh (1989). This HB estimator has the desired property of being a smooth nonnegative estimator of the variance. Simulations have also shown that it performs satisfactorily for small and moderate sample sizes and for a variety of distributions including heavy tailed distributions.

Consider the following hierarchical model as presented in Datta and Ghosh (1989). $Y_1,...,Y_k$ will denote the indices for PSUs 1,...,k and S will denote the within sum of squares WSS. These random variables will depend on the unknown parameters τ, b, r, λ.

I. Conditional on $T=\tau$, $B=b$, $R=r$, $\Lambda=\lambda$

$Y_1,...,Y_k$ and S are mutually independent with $\mathbf{Y} = (Y_1,...,Y_k)^T$ distributed as $N(\tau,(nr)^{-1}I)$ and S distributed as $r^{-1}\chi^2_{k(n-1)}$.

II. Conditional on $B=b$, $R=r$, $\Lambda=\lambda$,

T is distributed as $N(Xb,(\lambda r)^{-1}I)$ where X is a known kxp matrix of rank $p < k$.

III. B, R, and $Z=\Lambda R$ are marginally mutually independent, with B distributed as uniform(R^p), Z has pdf $f(z) \propto z^{-2}$ and R has pdf $f(r) \propto r^{(g-2)/2}$. Thus B, R, and Z have improper pdf's.

Stages I and II of the above hierarchical model can be identified as a balanced mixed effects model. To see this let

$$Y_{ij} = \mathbf{x}_i^T b + a_i + e_{ij}$$

for i=1,...,k and j=1,...,n. In the above $x_1,...,x_k$ are known vectors, **b** is the vector of regression coefficients, a_i's and e_{ij}'s are mutually independent with a_i's i.i.d. $N(0,\sigma_a^2)$ and e_{ij}'s i.i.d. $N(0,\sigma_e^2)$ where $\sigma_a^2 = (\lambda r)^{-1}$, $\sigma_e^2 = r^{-1}$ and $\sigma_a^2 / \sigma_e^2 = \lambda^{-1}$. The minimal sufficient statistic for this problem is $(Y_1,...,Y_k,S)^T$ where $Y_i = \Sigma Y_{ij}/n$ is the mean of the Y_{ij} and $S = \Sigma\Sigma (Y_{ij}-Y_i)^2$. Then $(Y_1,...,Y_k, S)^T$ has a distribution specified in I. and II.

We are interested in finding the posterior distribution of the variance ratio Λ^{-1}, and more particularly, the mean of the posterior distribution of Λ^{-1} given Y=y and S=s. From Datta and Ghosh (1989) we see that for $U = \Lambda/(n+\Lambda)$ the posterior distribution of U given Y=y and S=s is

$$f(u|y,s) \propto u^{(k-p-4)/2}(1 + uZ)^{-\phi/2}$$

where $\phi = nk - p - 2 + g$, P_X denotes the projection onto the column space of **X**, and $Z = nY^T(I - P_X)Y / S$ is a multiple of a usual F statistic. In our case, where **X** is a column of all 1s, P_X is a k by k matrix with every element equal to 1/n so Z=nBSS/WSS. Then the HB estimator of the ratio is

$$e_{HB} = E(\Lambda^{-1}|y,s) = \{ E(U^{-1}|y,s) - 1\}/n$$

where E(U|y,s)=

$$\frac{\int u^{(k-p-2)/2}(1 + uZ)^{-\phi/2}\,du}{\int u^{(k-p-4)/2}(1 + uZ)^{-\phi/2}\,du}$$

This estimator, while it has many nice properties, must be evaluated by numerical integration. It can be shown by tedious algebraic manipulations that the HB estimator is equal to the usual estimator plus a nonnegative term which converges to zero rapidly.

4. Findings

The HB estimate of the ratio of the PSU component of variance to the rest of the variance is seen to be consistently small for all major groups and all items. The anova estimator

clearly has problems because it produces negative estimates of variance for the major groups apparel, medical, entertainment and other. The variance of the HB estimator is also seen to be quite good in comparison to the anova estimator since they agree to three decimal places. One positive point for the anova estimator is that even in situations where the estimator was producing negative estimates the jackknifed version would sometimes correct this deficit.

For the month of January 1987, which was in the first year of the revision, the anova estimates produce nonnegative estimates for the major groups of apparel, medical, entertainment and other. Possibly the most important estimate is for all items, in which case the estimate is not negative. This pattern continued for all years. The HB estimates are all small and it appears that the variances of the HB estimator indicate a well behaved estimator. These values are presented in Table 2.

Considering how the variances change over time, Table 2 seems to indicate that the variances are showing a slight increase over time. There are previously reported results from Leaver (1990) in which the total variance is seen to increase over time. It appears that all items has the largest relative size for PSU component but in all cases the relative size of the PSU component is very small. The width of the confidence intervals are seen to generally decrease over time with the notable exception of medical which has an increase in recent months.

The most important thing to observe is that the confidence intervals always contain zero so that zero can never be rejected as a value for the PSU component of variance. In any case it appears that the actual estimate of the PSU component of variance for all items is always very small with a maximum of 2.6% of the total variance in January of 1987. Among the major groups food always has the largest relative PSU component of variance with a maximum of 7.5% of the total variance in January of 1987. Transportation has the next largest estimate with 4% of the total variance in January of 1987. All other estimates for major groups and all collection periods are at most 1% of the total variance. Plots of the confidence intervals are presented in Figure 1.

5. Conclusions

The HB estimate of the relative size of the PSU component of variance appears to be the best estimate of this component. It seems to give fairly stable estimates over time. It produces nonnegative estimates and the variance of the estimator seems to be very good. Using the results of the HB estimator we see that the relative size of the PSU component of variance is very small in most cases, especially for the groups in the index which are more highly weighted, such as food and housing.

6. Acknowledgements

The author would like to thank Sylvia Leaver and Rick Valliant for helpful discussions relating to these issues. The author also wishes to thank David Swanson, Janet Williams and Rick Valliant for their careful reading of this paper and for their helpful comments.

8. References

Bureau of Labor Statistics, *BLS Handbook of Methods* (1988), Washington. DC: U.S Government Printing Office, 166-167 and 174-176.

Corbeil, R.R. and Searle, S.R. (1976)," A Comparison of Variance Components Estimators," *Biometrics, 32*, 779-791.

Datta, G.S. and Ghosh, M. (1989), "Asymptotic Optimality of Hierarchical Bayes Estimators and Predictors," *Technical Report No 349. Dept. of Statistics, University of Florida.*

Dippo, C. S., and Jacobs, C. A. (1983), "Area Sample Redesign for the Consumer Price Index," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 118-123.

Leaver, S.G. (1990), "Estimating Variances For the U.S. Consumer Price Index for 1978-1986," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 290-295.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992), *Variance Components*, New York: John Wiley.