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

Robert M. Baskin and William H. Johnson
U.S. Bureau of Labor Statistics
2 Massachusetts Ave, N.E., Room 3655, Washington, D.C. 20212

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

KEY WORDS: Hierarchical Bayes, maximum likelihood, primary sampling unit, anova estimate.

This is a report of a project to estimate certain components of variance for the United States Consumer Price Index (CPI). This report deals with estimating components for the commodities and services part of the CPI. These variance components are estimated by Restricted Maximum Likelihood (REML) and the usual anova type estimators. It is seen that the REML methods produce nonnegative estimates of components of variance whereas the anova estimates will produce negative estimates of some components.

In section one the sampling design will be introduced. In section two the model for the components is built. The estimation methodology will be explained in section three and findings will be presented in section four.

1. Introduction and 1987 Design Description

The Bureau of Labor Statistics (BLS) is currently making preparations for the next 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. In the 1987 revision, the sample size for the commodities and services (C&S) part of the CPI was allocated using an optimization scheme in which components of variance were used as parameters, as reported in Leaver, et al (1987).

In this paper, the relative size of four components of variance associated with change in the commodities and services index are estimated. The four components are related to the sample design which will be explained in the following paragraphs.

For a full discussion of the CPI the reader is referred to Chapter 19 of the *BLS Handbook of Methods*, (1992). However, the following features of the CPI are important for the present discussion.

According to the *Handbook*, p 176, "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. 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 such as food, shelter, and transportation.

Pricing for the CPI is conducted in 88 PSUs in 85 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. The method of controlled selection is used and this complicates the actual randomization distribution. There are four classes of PSUs. The 32 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, 20 L PSUs and 24 M PSUs are in the current sample design. Urban areas not included in MSAs are classified as R PSUs. The current CPI contains 12 of these sampling units. The boundaries of these PSUs were defined by BLS. A description of the PSU selection for the 1987 revision can be found in Dippo and Jacobs (1983) and an update of the PSU selection for the 1998 revision is described in Williams et al (1993). The 32 A PSUs are referred to as certainty or self-representing PSUs. Thirty of these 32 PSUs are the largest metropolitan areas. For the remaining strata, the selected PSUs are referred to as non-self-representing PSUs.

The PSU stage of sample selection is common to both housing and C&S. In the C&S part of the CPI the next step is to independently sample outlets and items within each PSU replicate combination. The outlet sample is based on the Continuing Point of Purchase Survey (CPOPS),

conducted by the Bureau of the Census, and the item sample is based on the Consumer Expenditure Survey (CE), also conducted by the Bureau of the Census. Outlets are selected in eight POPS categories using a systematic pps sample. Items are selected in eight major groups using a stratified systematic pps sample. Selections from the independent samples are then matched by POPS expenditure category. For example, if Safeway at 1315 North Van Dorn street in PSU A315 is selected in the food and beverages category and the Entry Level Item, bananas, is selected in the major group food and beverages, then BLS will attempt to price bananas in this outlet. In the outlet the BLS field representative (FR) will follow a process called disaggregation to obtain a unique quote on bananas in this outlet. The entire process is described in the *Handbook* p185. The four stages of sampling lead to four components to be modeled.

The design is complicated by the fact that some items such as Food are priced monthly, while in all but the five largest PSUs other items such as Apparel are priced on an every other month basis. This makes comparing components across time difficult.

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. The index is estimated at the PSU level although not all PSUs are published. Let $IX_{it,s}$ denote the index at time t, in PSU i, relative to time period s. Then

$$IX_{it,s} = 100 * CW_{it} / CW_{is}$$

where CW_{it} and CW_{is} denote the aggregated weighted prices in PSU i for times t and s respectively.

2. The Model

The C&S part of 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. Thus we can write the price relative, the price change from time s to time t, for each unique item as

$$X_{ijkl,s} = \mu_{t,s} + \alpha_{it,s} + \beta_{ijt,s} + \gamma_{ikt,s} + \varepsilon_{ijkl,s}$$

where $\mu_{t,s}$ is a fixed factor, $\alpha_{it,s}$ is a random factor corresponding to PSU selection, $\beta_{ijt,s}$ is a random factor associated with outlet selection, $\gamma_{ikt,s}$ is a random factor associated with item selection, and $\varepsilon_{ijkl,s}$ is a random factor corresponding to selection of the unique quote. The assumptions on $\{\alpha_{it,s}\}$, $\{\beta_{ijt,s}\}$, $\{\gamma_{ikt,s}\}$, and $\{\varepsilon_{ijkl,s}\}$ are that they are mutually independent with mean 0, the $\alpha_{it,s}$ are identically distributed with variance $\sigma^2_\alpha(t,s)$, the $\beta_{ijt,s}$ are identically distributed with variance $\sigma^2_\beta(t,s)$, the $\gamma_{ikt,s}$ are identically distributed with variance $\sigma^2_\gamma(t,s)$, and the $\varepsilon_{ijkl,s}$ are identically distributed with variance $\sigma^2_\varepsilon(t,s)$. No attempt will be made to model this as a time series so the dependence on the parameters t and s will be suppressed. Our current work is to estimate the four components of variance, σ^2_α, σ^2_β, σ^2_γ, and σ^2_ε. Typically these estimates will be presented as proportions of the total variance. Note that because of the controlled selection of PSU's a true design-based estimate of the PSU component of variance is difficult, if not impossible to compute, leading us to use the model-based approach described here. Furthermore the form of the standard Anova estimators allows the estimate of the PSU, outlet and item components 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 type of problem can be found in Searle, Casella and McCulloch (1992). As can be seen from the estimates produced, this unfortunate phenomenon does actually occur so other methods of estimation are needed in this case. Among the limited options are taking the positive part of the anova estimator, restricted maximum likelihood, or using a Bayesian estimator. A Bayesian estimator under squared error loss is guaranteed to be

nonnegative and a Bayesian estimator was also considered because of certain successes in a similar situation with BLS data. See Baskin (1993) where three components of variance in the housing part of the CPI were estimated under a hierarchical Bayes (HB) model using Gibbs sampling. In the present work, a Bayes estimator of the components of variance is derived under a hierarchical normal model similar to the estimator used in Baskin (1993). This HB estimator has the desired property of being a smooth nonnegative estimator of the variance. Simulations in the balanced case have also shown that it performs satisfactorily for small and moderate sample sizes and for a variety of distributions including heavy tailed distributions.

For this estimator consider the following hierarchical model. Let X_{ijkl} denote the l^{th} unique quote from the k^{th} item selection and the j^{th} outlet selection in the i^{th} PSU. Let K_i be the number of items in PSU i, J_i be the number of outlets in PSU i, and L_{ijk} denote the number of unique quotes in PSU i, outlet j, and item k. Also let I denote the number of PSUs. Assume that $X_{ijkl} = \zeta_{ijkl,s} + \varepsilon_{ijkl,s}$, L_{ijk} where $\varepsilon_{ijkl,s}$ given σ_ε^2 are i.i.d. $N(0, \sigma_\varepsilon^2)$. Here, $\zeta_{ijkl,s}$ represent a sum of terms corresponding to PSU, outlet, and item selection. Thus

$$X_{ijkl} | \zeta_{ijkl,s}, \sigma_\varepsilon^2 \sim N(\zeta_{ijkl,s}, \sigma_\varepsilon^2).$$

Now assume that α_i given σ_α^2 are i.i.d. $N(0, \sigma_\alpha^2)$ if i corresponds to a non-self-representing PSU, α_i are 0 for self-representing PSUs, β_{ij} given σ_β^2 are i.i.d. $N(0, \sigma_\beta^2)$, γ_{ik} given σ_γ^2 are i.i.d. $N(0, \sigma_\gamma^2)$, μ_i are i.i.d. $N(\mu_0, \sigma_\tau^2)$,. σ_α^2~IG[a_1,b_1], σ_β^2~IG[a_2,b_2], σ_γ^2~IG[a_3,b_3], σ_ε^2~IG[a_4,b_4] and all are independent. (x~IG[a,b] means that x is inverse gamma with density $f(x)=b^a e^{-b/x}/\Gamma(a)x^{(a+1)}$ if (x>0)).

We are interested in finding the posterior distributions of the parameters given the observations and the other parameters. The posterior distribution of the vector ζ given the rest of the parameters and the observations is multivariate normal with entries of the mean vector given by $\frac{\sigma_\beta^2 X_{ij.} + \mu\sigma_\varepsilon^2}{\sigma_\beta^2 K_{ij} + \sigma_\varepsilon^2}$ if i corresponds to a self-representing PSU and $\frac{\sigma_\beta^2 X_{ij.} + \sigma_\varepsilon^2}{\sigma_\beta^2 K_{ij} + \sigma_\varepsilon^2}$ if i corresponds to non-self-representing PSU. Similarly for the μ, α, β, and γ the posterior means can be derived. For the parameters of interest, the sigmas, the posterior depending on the observations and the rest of the parameters are inverse gammas. The posterior distribution of σ_e^2 given the rest of the parameters and the observations is inverse gamma,

$$f(\sigma_\varepsilon^2 | \sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2, X, \zeta) \sim$$
$$IG[a_4 + L_{+++}/2; b_4 + \sum_i \sum_j \sum_k \sum_l (X_{ijkl} - \zeta_{ijk})^2/2]$$

Also

$$f(\sigma_\gamma^2 | rest) \sim IG[a_3 + K_{++}/2; b_3 + \sum_i \sum_k \gamma_{ik}^2/2]$$

$$f(\sigma_\beta^2 | rest) \sim IG[a_2 + J_{++}/2; b_2 + \sum_i \sum_j \beta_{ij}^2/2]$$

$$f(\sigma_\alpha^2 | rest) \sim IG[a_1 + N/2; b_1 + \sum_i \alpha_i^2]$$

where N is the number of non-self-representing PSUs.

3. Methodology

The maximum likelihood method is described in Searle, Casella and McCulloch (1992). This technique is based on the same model as the anova estimates but assumes that the data is normally distributed. The estimates are made by maximizing the log of the density function of the data. Using the notation from Searle, Casella and McCulloch (1992), let V denote the total variance matrix for the observations. The density to be maximized is

$$f(x) = \frac{e^{-\frac{1}{2}(x-\mu)'V^{-1}(x-\mu)}}{(2\pi)^{\frac{N}{2}}|V|^{\frac{1}{2}}}$$

where |V| denotes determinant of V and N denotes the full sample size. This maximization must be done by numerical techniques which are typically iterative in nature. The most common approach is some variant of the Newton-Raphson method although the EM algorithm is another popular technique.

The Gibbs sampling methodology which has been used in this work to estimate the components of variance is described in several recent papers but one of the standard references is Gelfand and Smith (1990) with several nice examples presented in Gelfand et. al. (1990). The Gibbs sampling methodology is both conceptually simple and easy to implement. The major drawback is the fact that it is computationally inefficient. In a problem such as the present problem with a large number of parameters computational efficiency is an issue. The issue of convergence is dealt with in Gelfand and Smith (1990) where it is shown that under relatively mild assumptions the rate of convergence is exponential.

Another typical problem which arises when using Gibbs sampling is sensitivity to initial values for the parameters. For estimates of variance, zero is an absorbing state for the Gibbs sampler so that values close to zero can "trap" estimates close to initial values. This situation seems to be the case for the estimation attempted and no viable solution was found. Further work may yield a solution but at this point the Gibbs sampling technique is not useable.

4. Findings

The current research uses price change for the periods from 9201 to 9411 (YYMM format) but December data was not available. Price change for different lengths of time was investigated. Two month, six month and one year price change were used. For one year price change, this allows creation of twenty two price change variables, since December data was not available. The intent is to calculate the components for each of the eight major groups within each of the four Census regions. It is clear from the results that there are very large differences by region so the regions cannot be combined.

In Table 1. the anova estimates and the restricted maximum likelihood estimates are presented for the first major group, apparel, and for the first time period, January 1993 to January 1994. For the variance component PSU, the order of the estimates appear to be similar. The outlet component shows the largest discrepancy. The anova estimates are negative which are usually interpreted to mean that the variable is not really significant in the model. The reml estimates indicate that the outlet component is the largest component. The anova estimate of item indicates that item is the most important. This discrepancy makes the results difficult to interpret.

TABLE 1.
ESTIMATES FOR ONE YEAR CHANGE
9301 to 9401

	PSU	OUT	ITEM	ERROR
NE region				
reml	0.0629	0.4440	0.1302	0.2769
anova	0.0411	-0.1658	0.7723	0.1637
MW region				
reml	0.0892	0.0913	0.0399	0.2085
anova	0.0888	-0.0554	0.1918	0.2113
SO region				
reml	0.0430	0.0913	0.0424	0.2364
anova	0.0220	0.0967	0.0583	0.2036
WE region				
reml	0.3133	0.3468	0.1299	0.3137
anova	0.2477	0.2818	0.1798	0.2871

Since estimates were made for several periods of one year change, the average of the estimates were calculated to see if there appeared to a consistency over time.

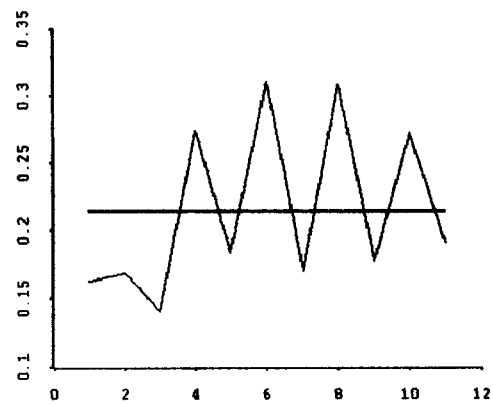
TABLE 2.
ESTIMATES FOR ONE YEAR CHANGE

	PSU	OUT	ITEM	ERROR
NE region				
mean	0.0297	0.1477	0.2737	0.2151
9401	0.0411	-0.1658	0.7723	0.1637
9402	0.0381	0.1262	0.1864	0.1694
9403	0.0760	0.0000	0.6429	0.1412
9404	0.0423	0.1804	0.1176	0.2743
9405	0.0158	0.1929	0.2211	0.1851
9406	0.0000	0.3379	0.1017	0.3109
9407	0.0257	0.1287	0.2240	0.1713

9408	0.0191	0.3367	0.1010	0.3088
9409	0.0101	0.0849	0.2355	0.1786
9410	0.0350	0.1301	0.1138	0.2721
9411	0.0234	0.1069	0.2946	0.1910
MW region				
mean	0.0385	0.0939	0.1154	0.1167
9401	0.0888	-0.0554	0.1918	0.2113
9402	0.0065	0.0788	0.0477	0.0530
9403	0.0846	0.1656	0.1534	0.1469
9404	0.0000	0.0746	0.0964	0.1029
9405	0.1102	0.1605	0.2192	0.1705
9406	0.0132	0.0945	0.0639	0.0688
9407	0.0798	0.0610	0.2665	0.1707
9408	0.0078	0.0392	0.0868	0.0877
9409	0.0267	0.1660	0.0577	0.1006
9410	0.0054	0.0793	0.0332	0.1117
9411	0.0009	0.1131	0.0523	0.0595
SO region				
mean	0.0560	0.1652	0.0934	0.1921
9401	0.0220	0.0967	0.0583	0.2036
9402	0.3288	0.7253	0.0959	0.1442
9403	0.0739	0.0922	0.1018	0.2178
9404	0.0000	0.1732	0.0075	0.1281
9405	0.0508	0.2523	0.1052	0.2036
9406	0.0063	0.2400	0.0452	0.1202
9407	0.0649	0.0000	0.3144	0.2345
9408	0.0345	0.1229	0.0495	0.2335
9409	0.0103	0.0577	0.0810	0.2274
9410	0.0180	0.0391	0.1082	0.2799
9411	0.0060	0.0182	0.0605	0.1198
WE region				
mean	0.1701	0.1644	0.2652	0.2803
9401	0.2477	0.2818	0.1798	0.2871
9402	0.1865	0.3492	0.2413	0.2647
9403	0.3717	0.0000	0.7634	0.4776
9404	0.4723	0.3789	0.3683	0.5053
9405	0.3092	0.1069	0.4538	0.4875
9406	0.1529	0.3131	0.1502	0.1571
9407	0.0592	0.3323	0.0947	0.2049
9408	0.0108	0.0000	0.3046	0.1168
9409	0.0213	0.0254	0.0970	0.1767
9410	0.0289	0.0192	0.1866	0.1811
9411	0.0111	0.0023	0.0776	0.2249

Table 2 presents the data for Apparel by region for the average of the collection periods and also for each of the one year changes for the collection periods. Since apparel has bimonthly collection the time period estimates can be seen to have a periodic movement around the averages. There is also some indication that early 1994 was less stable than the rest of 1994. Graph 1 presents the error component for Apparel in region 1 graphed over time with the mean displayed as a reference line.

Graph 1.



The estimates for the other components for Apparel display similar behavior. However the relative sizes across regions seem to be different. The means will be used at this point for the optimization purposes but further investigation into the behavior of the time period estimates in relation to the means is needed.

The hierarchical Bayes estimates for the same major groups and time periods are not presented. The HB estimates would not consistently converge. Due to the unstable nature of these estimates they are not further considered but different results for different prior values indicate that the Gibbs sampler is a flaming disaster.

5. Conclusions

The anova estimates of the size of the psu components of variance indicates that the item component is the smallest component while the error component is typically the largest. The reml estimator performed well in the sense of producing estimates which have some good properties. However, the order of the estimates appears to disagree with the more traditional anova estimates. In order to be sure of the usefulness of the estimates this discrepancy needs to be explained.

6. Future Work

The major task is to decide if the reml estimates are satisfactory for our purposes. This

includes more model checking and diagnostics. The current work also needs to be extended to investigate the change in the components over time.

7. Acknowledgments

The authors would like to thank Janet Williams and Rick Valliant for their careful reading of this paper and for their helpful comments. The author would also like to thank Jim Branscome, Sylvia Leaver, Frank Ptacek and Dave Swanson for insight into the C&S sample design. The author would like to thank Janet Williams for support on this project.

8. References

Bureau of Labor Statistics, *BLS Handbook of Methods* (1992), Washington. DC: U.S Government Printing Office, 176-235.

Baskin, R.M. (1992), "Hierarchical Bayes Estimation of Variance Components for the U.S. Consumer Price Index", *Proceedings of the Survey Research Methods Section*, American Statistical Association, 716-719.

Baskin, R.M. (1993), "Estimation of Variance Components for the U.S. Consumer Price Index via Gibbs Sampling", *Proceedings of the Survey Research Methods Section*, American Statistical Association (Vol. 2), 808-813.

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.

Gelfand, A.E. and Smith, A.F.M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association, 85*, 398-409.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling," *Journal of the American Statistical Association, 85*, 972-985.

Leaver, S. G., Weber, W.L., Cohen, M.P., and Archer, K.P. (1987), "Item-Outlet Sample Redesign for the 1987 U.S. Consumer Price Index Revision," International Statistical Institute, LII, Book 3 pp 173-185.

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

Williams, J.L., Brown, E.F., Zion, G.R. (1993), "The Challange of Redesigning the Consumer Price Index Area Sample"*Proceedings of the Survey Research Methods Section*, American Statistical Association (Vol. 1), 200-205.