# ESTIMATION OF VARIANCE COMPONENTS FOR THE U. S. CONSUMER PRICE INDEX VIA GIBBS SAMPLING

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

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

This is a report of the developmental step in a multiphase project to estimate certain components of variance for the United States Consumer Price Index (CPI). This report deals with estimating components for the housing part of the CPI. These variance components are estimated by a Hierarchical Bayes (HB) method employing Gibbs sampling and compared with the usual anova type estimators. It is seen that the HB method produces 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. For example, 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. In the 1987 revision the sample size for the housing part of CPI was based on a different type of allocation but for the upcoming revision there is an interest in performing the same type of optimization on the housing part of the CPI as was previously performed on C&S.

In this paper, the relative size of three components of variance associated with change in the housing index are estimated. The three components are related to the housing sample design which will be explained in 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 can be found in Dippo and Jacobs(1983). 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 housing part of the CPI the next step is to divide each PSU into block clusters which are based on Census block groups. Block clusters include

both Census enumeration districts and partial block groups. These are described in the *Handbook* p189. Segments are selected from block clusters and housing units (HU) are selected within segments.

The HUs are assigned to six panels. The six panels are collected on a rotating basis with rental units in panel one being collected in January and July, rental units in panel two collected in February and August, etc., but owner units are only collected once every two years.

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 shelter index is the index of interest in this paper and it is estimated at the PSU level although not all PSUs are published. Let $IX_{it,s}$ denote the index at time t, in pricing area 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 rents in PSU i for times t and s respectively.

## 2. The Model

The housing part of the CPI, as mentioned in the previous section, can be considered to have three components of variance corresponding to the three stages of housing 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 rent relative, the rent change from time s to time t, for each unit as

$$X_{ijkt,s} = \mu_{t,s} + a_{it,s} + b_{ijt,s} + e_{ijkt,s}$$

where $\mu_{t,s}$ is a fixed factor, $a_{it,s}$ is a random factor corresponding to PSU selection, $b_{ijt,s}$ is a random factor associated with segment selection, and $e_{ijkts,}$ is a random factor corresponding to unit selection. The assumptions on $\{a_{it,s}\}$, $\{b_{ijt,s}\}$, and $\{e_{ijkt,s}\}$ are that they are mutually independent with mean 0, the $a_{it,s}$ are identically distributed with variance $\sigma_a^2(t,s)$, the $b_{ijt,s}$ are identically distributed with variance $\sigma_b^2(t,s)$, and the $e_{ijkt,s}$ are identically distributed with variance $\sigma_e^2(t,s)$. No attempt will be made to model this as a time series so the dependence on the parameters t and s will be suppressed. Although the Rent Index depends directly on the rent levels and not directly on rent relatives, the Owners Index is though to be more directly related to rent relatives. The reason for using rent relatives is given in Lane and Sommers (1984).

Our current work is to estimate the three components of variance, $\sigma_a^2$, $\sigma_b^2$, and $\sigma_e^2$. 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 component of variance and the segment 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 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 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 similar but limited situations with BLS data. See Baskin (1992) where the PSU component of variance was estimated under a hierarchical Bayes model. In the previous work, since very few parameters were being estimated, a straight numerical approach was possible whereas in the current work there are a larger number of parameters being estimated which requires a different estimation technique. We investigate in the present work, a Bayes estimator of the components of variance derived under a hierarchical normal model similar to the estimator used in Baskin (1992). 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.

Consider the following hierarchical model. Let $X_{ijk}$ denote the observation from the $k^{th}$ unit in the $j^{th}$ segment in the $i^{th}$ PSU. Let $K_{ij}$ be the number of units in PSU i and segment j; let $J_i$ be the number of segments in PSU i; and let I denote the number of PSUs. Assume that $X_{ijk}=\zeta_{ij} + \varepsilon_{ijk}$, for i = 1,...,I; j = 1,...,$J_i$; k = 1,...,$K_{ij}$ where $\varepsilon_{ijk}$ given $\sigma_e^2$ are i.i.d. $N(0,\sigma_e^2)$, $\zeta_{ij}$ given $\theta_i$ and $\sigma_b^2$ are i.i.d. $N(\theta_i,\sigma_b^2)$ if i corresponds to a non-self-representing PSU $\zeta_{ij}$ given $\mu$ and $\sigma_b^2$ are i.i.d. $N(\mu,\sigma_b^2)$ if i corresponds to a self-representing PSU, and all are independent. Thus

$$X_{ijk} \mid \zeta_{ij}, \sigma_e^2 \sim N(\zeta_{ij},\sigma_e^2).$$

Now assume that $\theta_i$ given $\sigma_a^2$ are i.i.d. $N(\mu, \sigma_a^2)$, $\mu \sim N(\alpha,\tau)$, $\sigma_e^2 \sim IG[\alpha_1,\beta_1]$, $\sigma_b^2 \sim IG[\alpha_2,\beta_2]$, $\sigma_a^2 \sim IG[\alpha_3,\beta_3]$ 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. The posterior distribution of the vector $\zeta$ given the rest of the parameters and the observations is multivariate normal with entries of the mean vector given by $\frac{\sigma_b^2 X_{ij.} + \mu\sigma_e^2}{\sigma_b^2 K_{ij} + \sigma_e^2}$ if i corresponds to a self-representing PSU and $\frac{\sigma_b^2 X_{ij.} + \theta_i \sigma_e^2}{\sigma_b^2 K_{ij} + \sigma_e^2}$ if i corresponds to non-self-representing PSU. For the posterior means $X_{ij.}$ denotes the mean of the observations from the i,j segment. The variance of the distribution is a diagonal matrix with diagonal entry $\sqrt{\frac{\sigma_b^2\sigma_e^2}{\sigma_b^2 K_{ij} + \sigma_e^2}}$. The posterior distribution of $\theta$ given the rest of the parameters and the observations is multivariate normal with mean $\frac{\sigma_a^2 \zeta_{i+} + \mu\sigma_b^2}{\sigma_a^2 J_i + \sigma_b^2}$ where $\zeta_{i+}$ denotes the sum of the $\zeta_{ij}$ corresponding to PSU i. The variance of this distribution is a diagonal matrix with diagonal entry $\sqrt{\frac{\sigma_a^2\sigma_b^2}{\sigma_a^2 J_i + \sigma_b^2}}$. The posterior distribution of $\sigma_e^2$ given the rest of the parameters and the observations is inverse gamma,

$$f(\sigma_e^2 \mid rest) \sim IG[\alpha_1 + K_{++}/2; \beta_1 + \sum_i \sum_j \sum_k (X_{ijk} - \zeta_{ij})^2/2]$$

Also

$$f(\sigma_b^2 \mid rest) \sim IG[\alpha_2 + J_+/2; \beta_2 + \sum_i \sum_j (\zeta_{ij} - \theta_i)^2/2]$$

$$f(\sigma_a^2 \mid rest) \sim IG[\alpha_3 + N/2; \beta_3 + \sum_i (\theta_i - \mu)^2]$$

where $\theta_i$ is taken to be $\mu$ if i corresponds to a self-representing PSU; and N is the number of non-self-representing PSUs.

## 3. Methodology

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.

Let $Z_1,...,Z_d$ denote the parameters in some order. Initial values for the parameters are needed and will be denoted by $Z_1^0,...,Z_d^0$. The posterior conditional distributions from the previous section are employed in the sampling scheme. The systematic scan Gibbs sampler iterates the following loop:

1. Sample $Z_1^{i+1}$ from $f(Z_1 \mid Z_2^i,...,Z_d^i)$
2. Sample $Z_2^{i+1}$ from $f(Z_2 \mid Z_1^{i+1}, Z_3^i,...,Z_d^i)$

...

d. Sample $Z_d^{i+1}$ from $f(Z_d \mid Z_1^{i+1},...,Z_{d-1}^{i+1})$

This is the implementation of the Gibbs sampler employed in this paper, i.e., the order in which parameters are "visited". Other visiting schemes are also common. It is possible to update some parameters more often than others as long as each parameter is visited infinitely often.

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 question 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. One

solution to this problem is to take several independent sets of initial values and run independent streams of the Gibbs sampler. This is the approach in the present work. This gives the additional advantage of producing a distribution of independent estimates.

## 4. Findings

The current housing database has five consecutive reporting periods for each panel. This allows creation of ten price change variables, four corresponding to a six month change, three corresponding to a one year change, two corresponding to one and a half year change and a single two year change. Since this gives rise to sixty variables, ten variables for each of six panels, only results for six month price change will be reported.

In Table 1. the anova estimates for the variance components are presented. The PSU component is seen to be consistently the smallest while the segment component typically contains the negative estimates. The first panel also contains a negative estimate for PSU component in the first time period.

TABLE 1. ANOVA ESTIMATES

1-PSU denotes the PSU component, 1-Seg denotes the segment component, 1-Err denotes the error component from panel 1, etc.

| time period | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| panel | | | | |
| 1-PSU | -.0000 | .0002 | .0000 | .0008 |
| 1-Seg | .0010 | .0012 | .0015 | .0032 |
| 1-Err | .0083 | .0119 | .0105 | .0305 |
| 2-PSU | .0009 | .0001 | .0014 | .0004 |
| 2-Seg | .0056 | .0021 | -.0006 | .0018 |
| 2-Err | .0203 | .0084 | .0321 | .0180 |
| 3-PSU | .0004 | .0001 | .0005 | .0029 |
| 3-Seg | -.0054 | .0006 | .0003 | -.0127 |
| 3-Err | .0379 | .0096 | .0146 | .0763 |
| 4-PSU | .0006 | .0005 | .0006 | .0008 |
| 4-Seg | -.0006 | .0006 | -.0009 | .0002 |
| 4-Err | .0148 | .0152 | .0219 | .0212 |
| 5-PSU | .0009 | .0005 | .0006 | .0009 |
| 5-Seg | .0030 | -.0005 | .0013 | -.0056 |
| 5-Err | .0202 | .0119 | .0110 | .0257 |
| 6-PSU | -.0068 | .0010 | .0005 | .0004 |
| 6-Seg | .0479 | -.0025 | -.0005 | -.0008 |
| 6-Err | .0140 | .0180 | .0166 | .0274 |

In Table 2 the hierarchical Bayes estimates for the same panel and time periods are presented. The PSU and error components are typically comparable but the segment component in the cases where the anova estimates are negative can be quite different.

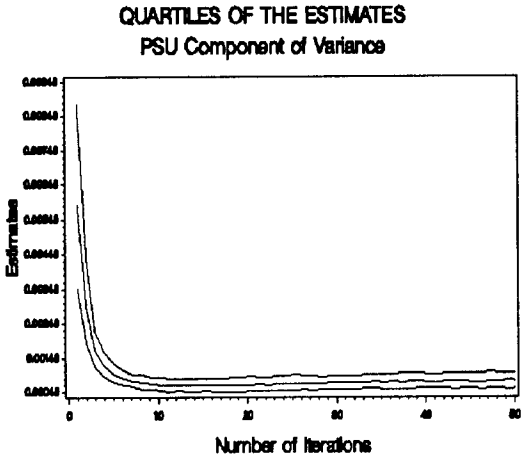
TABLE 2. HB ESTIMATES

1-PSU denotes the PSU component, 1-Seg denotes the segment component, 1-Err denotes the error component from panel 1, etc.

| time period | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Panel | | | | |
| 1-PSU | .0002 | .0001 | .0002 | .0008 |
| 1-Seg | .0013 | .0015 | .0015 | .0022 |
| 1-Err | .0085 | .0120 | .0109 | .0322 |
| 2-PSU | .0007 | .0002 | .0003 | .0001 |
| 2-Seg | .0099 | .0030 | .0114 | .0055 |
| 2-Err | .0198 | .0085 | .0038 | .0168 |
| 3-PSU | .0004 | .0001 | .0004 | .0003 |
| 3-Seg | .0063 | .0031 | .0029 | .0058 |
| 3-Err | .0298 | .0086 | .0119 | .0472 |
| 4-PSU | .0004 | .0003 | .0006 | .0001 |
| 4-Seg | .0052 | .0039 | .0066 | .0036 |
| 4-Err | .0249 | .0133 | .0373 | .0133 |
| 5-PSU | .0001 | .0001 | .0002 | .0002 |
| 5-Seg | .0165 | .0042 | .0072 | .0086 |
| 5-Err | .0164 | .0092 | .0098 | .0295 |
| 6-PSU | .0001 | .0001 | .0002 | .0002 |
| 6-Seg | .0165 | .0042 | .0072 | .0086 |
| 6-Err | .0164 | .0092 | .0098 | .0295 |

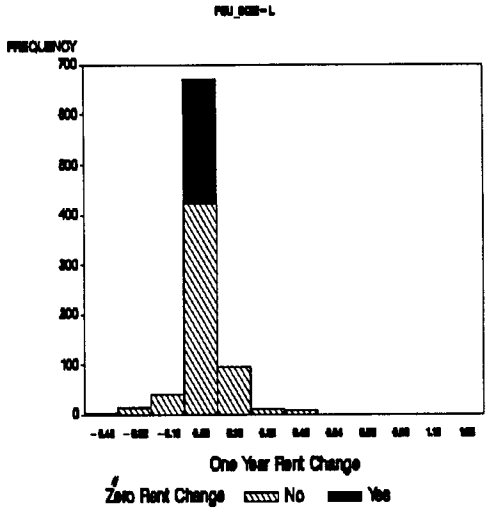
The CPI index data is generally considered to be accurate to five decimal places but the convergence of the estimates in the Gibbs sampling methodology had an interesting problem. Generally, the estimates would converge to four decimal places, usually after 20 iterations in each independent stream. However, the estimates never seemed to converge in the fifth decimal place. The process was tested up to 500 iterations and no convergence was ever achieved in the fifth decimal place. This is an indication that the data is actually accurate to four decimal places. Quartiles of the distributions of the independent streams was one

criterion used to judge convergence of the estimates. The quartiles typically show that the estimates are close to stable after as few as ten or twelve iterations with fairly obvious convergence after twenty or so iterations. The following example shows the quartiles for the PSU component of variance for one time period.




A second point of concern with these estimates is sensitivity to initial values. The test programs took fixed starting values for the initial values which derived from the best prior guess of the components of variance. A more advanced version of the Gibbs sampler used the initial values from the previous runs and added random numbers to each initial value to attempt to test sensitivity to the initial values. There was no difference in the final estimates between the two techniques.

Another concern is prior sensitivity. The histogram of rent change is known to be skewed to the right with some large values of rent change and several values of rent change of zero. While the observed data for each class of PSU was skewed it was not as skewed as expected. The following graph representates rent change with the units showing no rent change being represented in black.




This together with the fact that there are indications that HB methods are somewhat robust is an indication that prior sensitivity should not be a problem. Some initial tests of alternative specifications of the problem meet with little success because of computational difficulties.

The major drawback of this methodology is the computational inefficiency. Because of cost constraints this problem was not considered as appropriate for development on a mainframe. The only current alternative is the BLS desktop environment. BLS is currently operating 80386 machines under DOS as its 'work station' environment, but this environment is very slow for the current type of problem. Also an evaluation copy of OS/2 was used to compare two machines with exactly the same hardware configuration. To run the Gibbs sampler on one panel for five of the ten periods took over forty hours in a C program compiled under DOS. To run the same Gibbs sampler on the same data set in a C program compiled under OS/2 took 20 hours. A large amount of disk space was required to save the results of each program run in order to evaluate the convergence. This type of computing environment makes Gibbs estimation in our current setup an impractical option for production purposes. While the computations are currently infeasible on a production basis in the desktop environment this experiment was a valuable developmental tool and the test could be considered successful from a statistical point of view.

## 5. Conclusions

The estimates of the size of the components of variance of housing indicates that the previous sampling methodology may have been successful in controlling sources of variation for the variables considered here. The HB estimator performed well in the sense of producing estimates which have good properties. However, in terms of computational feasibility the Gibbs sampling methodology in the current environment proved to be problematic.

## 6. Acknowledgments

The author would like to thank Janet Williams, Rick Valliant, and Stuart Scott for their careful reading of this paper and for their helpful comments. The author would also like to thank Bill Miller, Shawn Jacobson, Jim Branscome, and Steve Henderson for insight into the BLS housing design. Thanks go to Ken Archer for help in accessing the housing database. The author would like to thank Janet Williams for support on this project. And special thanks go to Sylvia Leaver for help and support.

## 7. 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.

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.

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.

Lane, W. F., and Sommers, J. P., (1984) "Improved Measures of Shelter Costs," *Proceedings of the Business and Economics Statistics Section*, American Statistical Association, 49-55.

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