MODELING THE REALIZED OUTLET SAMPLE FOR THE COMMODITIES AND SERVICES COMPONENT OF THE U.S. CONSUMER PRICE INDEX.

William H. Johnson, Sylvia G. Leaver and Thomas S. Benson, U.S. Bureau of Labor Statistics
2 Massachusetts Avenue, NE, Room 3655, Washington, DC 20212

KEY WORDS: simulation, sample design, sample size

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

The U.S. Consumer Price Index (CPI) is made up of two large components, commodities and services (C&S) and housing. Each year approximately 25% of the sample for commodities and services is rotated in order to introduce new items into the sample for the CPI and to select a set of outlets which better reflects current expenditures among outlets. The variance and cost of the sample for commodities and services are modeled so that a sample design can be found which minimizes the variance of the estimator for the commodities and services component of the CPI. In order to do this, it is necessary to ascertain the number of distinct outlets that arise under repeated sampling of the outlet frames for distinct item categories.

This paper documents the method used to estimate the actual number of unique outlets as a function of the number of outlets specified to be selected and attempts to evaluate the effectiveness of this methodology.

1. The Design of the Commodities and Services sample

The U.S. Consumer Price Index measures the change in the cost of a fixed market basket of goods and services. The commodities and services component of the CPI accounts for 72.5% of the expenditure weight in the CPI. The sample for the commodities and services component is derived from two stages of samples.

The first stage of the design is the selection of primary sampling units (PSUs) which are used for both the housing and commodities and services components of the CPI. A PSU is a geographic area composed of counties, or minor civil divisions in the northeast United States. There are 87 PSUs of which 31 are metropolitan statistical areas selected with certainty and 56 are smaller non-self representing areas.

At the second stage of sampling for commodities and services there are two independent samples drawn within each PSU. A sample of outlets is drawn from frames constructed from the Telephone Point of Purchase Survey (TPOPS). Commodities and services is represented by 217 TPOPS categories and 16 non-POPS categories. Roughly 25% of the categories are rotated each year. In the Telephone Point of Purchase Survey, consumers are sampled using random digit dialing. The selected consumers are then asked from which outlets they bought items within a subset of the TPOPS categories and how much money was spent in each outlet for each TPOPS category within a given recall period ranging from one week to five years. The set of the reported outlets make up the frames for the TPOPS categories.

Separately a sample of item strata is selected. An item stratum is a fairly small set of items or services such as bananas or canned fruits and vegetables or laundry and dry cleaning services. The item strata are clustered into item groups and the sample design specifies the total number of item strata selections per item group. Each item stratum corresponds to one or more TPOPS categories. Each time an item stratum is selected, one of the corresponding TPOPS categories is selected with probability proportional to the relative importance of the TPOPS category within the item stratum. The relative importance for a TPOPS category is the sum of the relative importances of the entry level items (ELIs) which are mapped to the TPOPS category. The ELI level relative importances within an item stratum are determined from the Consumer Expenditure survey (CE). This produces a sample of item strata and associated TPOPS categories. Once the samples of outlets and item strata are selected they are merged by TPOPS category to give a sample of outlets and items within outlets.

2. The model of the Commodities and Services sample

The total variance of price change is a weighted sum of variances of index area – group of item strata price changes. An index area here is a self representing PSU or a collection of non-self-representing PSUs of the same size class and in the same Census region.

The variance of an index area - group of item strata price change is decomposed into three or four components; a component of variance attributable to variation across outlets, a component of variance attributable to variation across item strata, a residual component of variance and if there are multiple PSUs in the index area, a component of variance attributable to variance across PSUs within the index area. The variance due to variation in price change across outlets is modeled as $\dfrac{s^2_{unit,outlet}}{total\ number\ of\ unique\ outlets}$ where $s^2_{unit,outlet}$ is the unit component of variance attributable to variation across outlets for the index area – item group combination. Thus we need to know the number of unique outlets within the index area for each group of item strata in order to estimate these index area – item group level price change variances.

The cost model is broken apart into outlet related costs and quote related costs. The quote related costs depend on the number of quotes but not upon the number of unique outlets in which these quotes are found. However, the outlet related costs require a measure of the number of unique outlets across all groups of item strata. This is because an outlet associated cost such as traveling to an outlet or obtaining non quote related information from an outlet are independent of the number of groups of item strata from which quotes are to be priced or initiated. For details of the variance and cost models for the commodities and services sample see Leaver *et al (1996)*.

3. The particular choice of model

Estimates of the number of unique outlets within each index area and item group and of the number of unique outlets across item groups within each index area are achieved by modeling the number of unique outlets against the number of outlets selected. The model chosen is a quadratic polynomial with zero intercept $M_{unique} = a*M^2 + b*M$ where M is the designated outlet sample size. The intercept is taken as zero because if zero outlets are selected then zero unique outlets will be in the selected sample. This model has been used historically since the optimization of the commodities and services sample for the 1987 CPI revision. The quadratic coefficient is almost always small, but in a check of 286 cases, the quadratic coefficient was significantly different from zero at the 0.05 level in 260 of those instances.

In the future alternative models may be considered to determine if a more suitable model exists.

4. Difficulties in determining the coefficients of the model

The sample for the commodities and services is currently derived primarily from two separate sources. Through 1996, frames of outlets for rotation were derived from the Continuing Point of Purchase Survey (CPOPS). CPOPS collected data for all item categories within a subset of PSUs and each PSU was rotated about once every five years. Collection was done by the Census Bureau using household personal interviews. Currently the frames of outlets are derived from the Telephone Point of Purchase Survey (TPOPS). TPOPS surveys consumers about a subset of all item categories and is designed so that about 25% of item categories are rotated each year in all PSUs. Currently the outlet sample is a mixture of sample derived from frames from CPOPS sources and from frames from TPOPS sources, but outlets to be initiated are only from frames derived from TPOPS.

In case the allocation of items and outlets couldn't be revised between sample selections based on more recently available data it was decided to model the allocation as though one quarter of the categories within each group of item strata would be rotated each year in each PSU.

Outlets that are currently in the sample can be selected again from new TPOPS based frames. Thus we end up having to measure the number of unique outlets for a sample derived partly from CPOPS based frames and partly from TPOPS based frames. This will continue for several years until the commodities and services outlet sample has been completely replaced by sample from TPOPS based frames.

As noted earlier, samples of outlets are merged with samples of item strata by POPS category. The CPOPS categories were revised in 1996 so frames from old CPOPS categories had to be mapped into new frames for newer CPOPS categories.

PSUs are divided into 12 groups for the purposes of TPOPS. The same set of item categories are rotated for all of the PSUs within each of these groups. Currently only the first four of these groups of PSUs has had any frames derived from TPOPS. The TPOPS categories are divided into 16 groups. During this transition from a CPOPS based sample to a TPOPS based sample the TPOPS categories have been introduced on a staggered basis, e.g., PSU group 1 has TPOPS based frames for the first seven groups of TPOPS categories, PSU group 2 has TPOPS based frames for the first six groups of TPOPS categories, PSU group 3 has TPOPS based frames for the first five groups of TPOPS categories and PSU group 4 has TPOPS based frames for the first four groups of TPOPS categories. Thus, the mixture of CPOPS and TPOPS frames from which the totality of the sample for a PSU are drawn varies from PSU to PSU depending on which TPOPS group of PSUs it belongs to.

Finally, we usually had to work from frames that differed from the final frames used to select the actual sample. This is because the frames of outlets may contain duplicate outlets with different descriptive information or the outlets may now be known to be out of business. The process of sending out the frames to the individual PSUs to eliminate duplicate and out-of-business outlets occurs fairly late in the sample selection process and it was only in designing the sample for the August 1999 TPOPS rotation that we have had the final refined frames from which to do the simulations. In preparing to do the sample design for the February 1999 TPOPS rotation, we went through a subset of frames and refined them using information available on the internet and a street atlas program. Comparing the final regression coefficients obtained using the unrefined and refined frames for February and August 1999 showed almost no difference. Thus it appears that the effect of frame refinement itself is minimal.

5. The procedure for obtaining the coefficients

In order to see what the final realized sample would be, the three processes of drawing a sample of outlets, a sample of item strata and associated POPS categories and merging these into a sample of outlets and items within outlets were simulated.

To select outlet samples it was first determined for each PSU what was to be the source of frames for the totality of item categories. This mixture of sources for frames changes with each rotation as further sample based on TPOPS frames is introduced. Outlet samples of size two through twelve were drawn independently with outlets being selected with probability proportional to the expenditures within outlet for the POPS category. The range of sample sizes used was intended to reflect the range in which we expected the final specified number of outlets per POPS category to fall. There was some variation in the maximum sample size used ranging from six through twelve. Comparisons of the final model coefficients obtained based on differing maximum outlet sample sizes showed very little difference.

The above process was repeated from six to fifteen times in order to provide additional datapoints for determining the model coefficients. The number of times the process was repeated varies in part because of time restrictions for some of the times that the sample design was revised. However the primary reason for the variation is due to the unequal clustering of PSUs into groups of PSUs in order to simplify the optimization problem.

Next the item samples were selected. One input needed is the number of selections of each item stratum. For the purposes of the simulation we used the number of item stratum selections from the most recent sample selection. From having used multiple sets of item stratum selections and examining the resulting model coefficients it appears that the model coefficients are quite insensitive to this particular input.

Each item stratum is associated with one or more POPS categories. For instance items in item stratum AA01, which contains men's suits, outer wear, sports coats and tailored jackets, come from outlets in TPOPS categories A01 (Men's suits or sports coats) and A02 (Men's outerwear) for PSUs in TPOPS PSU group 1. Each POPS category has a relative importance within the item stratum and this relative importance varies depending on Census region. For instance, the relative importance of POPS category A01 within item stratum AA01 is 0.626 in Census region 1, the northeast United States, but the relative importance is 0.728 for Census region 3, the southern United States.

For each item stratum, a number of POPS categories were selected with probability proportional to relative importance with the number of POPS category selections being the specified number of selections for the item stratum. A POPS category is allowed to be selected multiple times and there is the possibility that a given POPS category won't be selected at all. Enough samples were selected in order to have a separate sample to match up with each of the outlet samples and separate samples were drawn for each region. The process was also repeated for each of the five combinations of CPOPS and TPOPS categories which occur in our current set of PSUs.

Now for each PSU, item samples drawn using the relative importances for the region where the PSU is located are matched up to the outlet samples which have been drawn for that PSU by POPS category which yields a sample of outlets and item strata to be collected in those outlets.

For each sample one now needs to determine two different measures of the unique number of outlets. For the variance model the number of unique outlets within a group of item strata is needed. This is done by unduplicating the list of outlets in each sample for each item group and counting the number of remaining outlet records.

For the cost model a different measure of uniqueness is needed. This is because the fixed costs of visiting an outlet do not depend on the items to be priced or initiated within the outlet. It is still necessary to apportion these fixed outlet related costs across the groups of item strata. This was done by treating the outlet as contributing $\frac{1}{n}$ outlets to each of the n groups of item stratum for which there is at least one item stratum within the outlet. Other possibilities exist such as apportioning the fixed outlet cost based on the percentage of item stratum selections within the outlet in each group of item strata. Thus the number of unique outlets associated with each PSU group j and item group i is

$$U_{i,j} = \sum_{k=1}^{n_{i,j}} \frac{1}{\#\ \ of\ item\ groups\ outlet\ ijk\ is\ in}$$

where $n_{i,j}$ is the number of unique outlets in item group i and index area group j.

With these two measures of unique outlets within a group of item strata one now has a collection of specified outlet sample sizes per POPS category and two numbers of unique outlets for each group of item categories. The unique numbers of outlets are then regressed against the specified number of outlet selections and the square of the specified number of outlet selections.

6. How good are the models?

There are some changes that are necessitated by the conversion to TPOPS based frames. Previously the number of unique outlets per CPOPS category was modeled as a function of the number of outlet selections per CPOPS category. The number of unique outlets predicted was then multiplied by the number of CPOPS categories to get an estimate of the number of unique outlets per group of item strata from the number of outlet selections specified per CPOPS category.

With the current sample being a mixture of sample derived from TPOPS and CPOPS based frames, there is no longer a well defined number of POPS categories per group of item strata and there won't be a well defined number until the sample has converted entirely to being based on TPOPS frames. This is because the categories were changed considerably when switching to TPOPS and are usually more refined than comparable CPOPS frames. Thus, it was necessary to absorb the number of POPS categories into the model itself and predict the number of unique outlets per groups of item strata as a function of the specified number of outlet selections per POPS category.

It was desired to conduct as pure a test as possible. As selected outlets have their outlet code renamed it is impossible to determine overlap between outlets selected from old CPOPS based frames and the outlets selected in simulations from TPOPS frames. Thus it was decided to perform a test of the methodology using the actual selected but not yet initiated outlets for TPOPS categories which are part of the February 1999 TPOPS rotation. Six A-sized PSUs were selected for test based on their being contained entirely in the first TPOPS PSU group and thus having a full five quarters of TPOPS frames to work with.

The procedure described above was applied to just the TPOPS frames for these six PSUs through the first quarter of 1998 and the regression coefficients were computed at the individual PSU level.

First is a summary of the comparison of the predicted number of unique outlets within a group of item strata to the actual realized sample size within the same group of item stratum.

Group of item strata	Mean realized outlets	Mean predicted outlets	Mean of ratios
Overall	20.4	20.5	0.996
Food staples	18.8	18.6	0.992
Fruits and vegetables	9.0	8.8	0.964
Other food	21.2	20.6	0.976
Household	39.2	41.4	1.057
Apparel	21.1	20.8	0.978
Transportation	4.8	4.9	1.002
Education and communication	12.5	12.4	0.998
Entertainment and other	33.9	34.0	1.001

Next is a summary of the comparison of the predicted number of unique outlets across all groups of item strata to the actual realized number of unique outlets across all groups of item strata as distributed among groups of item strata.

Group of item strata	Mean realized outlets	Mean predicted outlets	Mean of ratios
Overall	18.8	18.7	0.958
Food staples	16.3	15.9	0.972
Fruits and vegetables	7.4	6.7	0.770
Other food	17.9	17.7	0.973
Household	37.4	39.1	1.050
Apparel	19.4	18.8	0.921
Transportation	4.8	4.9	1.002
Education and communication	12.5	12.2	0.977
Entertainment and other	32.3	32.4	1.009

Thus it appears that the regression formulas generated by these frames do a good job of predicting the unique number of outlets within a group of item strata and they do only slightly worse for predicting the number of unique outlets across groups of item strata with the likelihood that the estimates produced are slightly too low.

7. The effect of switching from a CPOPS based sample to a TPOPS based sample.

It is of interest to determine how much effect the introduction of frames for new TPOPS categories has on the unique outlet predicting function. This was done by comparing regression coefficients for all A-sized cities in TPOPS PSU group one using frames from the first quarter of 1997 through the third quarter of 1998 to ones using frames from the first quarter of 1997 through the second quarter of 1998 as well as the ones for the A-sized PSUs in TPOPS PSU group 2 using all of the frames through the third quarter of 1998. This allows examining the effect of the introduction of one quarter of TPOPS frames for PSUs in TPOPS PSU group one, the PSUs with the most TPOPS based frames available, and it allows one to compare two sets of PSUs that have the same total set of frames. Of interest are the groups of item stratum in which there were new associated TPOPS categories in the third quarter 1998 frames for TPOPS PSU group one.

	TPOPS group 1 Q983	
Group of item strata	Linear coefficient	Quadratic coefficient
Food staples	7.68	-0.24
Other food	15.49	-0.45
Apparel	18.65	-0.44
Entertainment and other	26.50	-0.68

	TPOPS group 1 Q982	
Group of item strata	Linear coefficient	Quadratic coefficient
Food staples	8.65	-0.23
Other food	15.49	-0.42
Apparel	19.71	-0.47
Entertainment and other	29.94	-0.75

	TPOPS group 2 Q983	
Group of item strata	Linear coefficient	Quadratic coefficient
Food staples	3.95	-0.03
Other food	11.94	-0.26
Apparel	15.96	-0.47
Entertainment and other	23.77	-0.76

As shown in this comparison, while there is some change in the model coefficients due to the introduction of new TPOPS categories, this amount of change is smaller than the difference in model coefficients using the same set of TPOPS categories but different PSUs.

8. Conclusions

It appears that the quadratic models used to estimate the number of unique outlets do a reasonable job of approximating the final realized outlet sample sizes. There is some effect on the coefficients of the quadratic models due to the introduction of frames for TPOPS categories over time, but it appears that there is more variation in the coefficients across PSUs. The refinement of frames seems to have little effect on the coefficients of the quadratic model. Due to the changes produced by the introduction of new TPOPS based frames with each rotation it is advisable to re-estimate the coefficients for each rotation if possible.

9. Acknowledgments

The authors would like to thank Janet Williams for her careful reading of this paper and her helpful comments. In addition, we would like to thank Owen Shoemaker for his thoughts on this particular problem.

10. References

Leaver, S., Johnson, W. H., Baskin, R., Scarlett, S., and Morse, R. (1996) "Commodities and Services Sample Redesign for the 1998 Consumer Price Index Revision," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Vol. I, pp. 239-244**.**