SAMPLING WEIGHTS FOR THE HOUSING SAMPLE OF THE CONSUMER PRICE INDEX

Eugene F. Brown
Bureau of Labor Statistics, 2 Mass Ave NE #3655, Washington, DC 20212

KEY WORDS: Sample design, regression, rent

INTRODUCTION

The Consumer Price Index (CPI) is a measure of the average change in the prices paid by urban consumers for a market basket of goods and services across the United States. Currently, the housing sample is used in pricing both residential rent and owners' equivalent rent which together make up the largest component of the CPI, 5.8% and 19.3%, respectively. The rent index is a measure of rent change during a certain specific time period. Rather than a direct measure of homeowners cost, the Bureau of Labor Statistics uses the amount of rental income that a homeowner foregoes by living in the home instead of renting it out. This proxy is called owners' implicit rent and the index of price change in implicit rent is known as the rental equivalence (REQ) index.

The first step in selecting the housing sample for the 1999 Consumer Price Index was to choose the areas in which the sample is to be conducted. Any metropolitan statistical area with more than 1.5 million in population was automatically included as a self representing area in the sample. Other areas were selected with probability proportional to population. After the 1998 revision, the CPI will consist of 87 different PSU's, of which 77 are metropolitan areas defined by the Office of Management and Budget, and 10 are non-metropolitan areas. (See Williams, *et. al.* (1993)) Segments inside sampled areas were formed from Census blocks. The segments had to contain at least 50 owners plus renters in A and B size (metropolitan) areas, and at least 30 owners plus renters in C size (non-metropolitan) areas. Blocks with less than the prescribed number of owners plus renters were combined using geographical proximity as the criterion for combination into segments. After selecting segments and listing all housing units in them, sample housing units will be selected. This paper describes the segment selection methodology.

Stratification of segments was done using a six stratum design described in Brown and Johnson(1994). This process used latitude and longitude data to group the segments within primary sampling units (PSUs) into geographical strata. Two of the strata were found in the region of densest expenditure weight as defined below and are called the center city, while the rest of the strata form the suburbs. The center city was split into either East-West regions or North-South regions, depending upon rent levels in these regions. The suburbs were split into four quadrants, with approximately equal weight in each quadrant.

For the 1998 revision, a new system of assigning measures of size for segment sample selection will be used. Expenditure weights based on average rental values and imputed owners' average rental values at the Census block level will be used as measures of size for each segment.

WEIGHTING OF SEGMENTS FOR SELECTION

In past revisions, segments have been weighted by the number of housing units in the segment. This is not a good measure of housing expenditure, since higher value housing units were treated the same as lower value housing units. It also can cause other problems in the sample. Since dangerous neighborhoods often contain lower value housing units, this representation can cause us to sample more heavily in these neighborhoods than would a size more directly related to expenditures. The difficulty of collecting data in those areas causes more segments to be eliminated from the sample, and therefore less yield from the sample is obtained. Housing segment weights proportional to expenditure also should cause less public housing and other problem housing to be chosen within each PSU. Public housing is ineligible for the survey but the counts of such units cannot be eliminated from the segment housing counts. Since the amount of increase in price is also related to expenditure, it was felt this new method would lower the variance of the housing index. It is therefore desirable to find some method for weighting units based on expenditure rather than number of housing units per segment.

Weighting by expenditure also has some problems associated with it. The biggest of these is what expenditure should be used? We do have Census information on several different levels, with variables that include average rent, average owner value, number of renters, number of owners, number of housing units, and several others. If we were just to use average rent value, what about areas where there are only a few renters and many owners? Would the average rent value be indicative of those owners or not? One solution is to include an imputed average rent value for owners in the calculation of segment weight. The final weight for the segment would then be defined as (number of renters)*(average rent)+(number of owners)*(owners imputed rent). This gives an estimate of expenditure for the segment, and is exactly what we were looking for in a weight. The first three quantities in the equation are available from the 1990 Census. The last, owners imputed rent, was derived from regression analyses described in the next section.

REGRESSION ANALYSIS

Block group data from the 1990 Census was used in all of the analyses. In particular, average rent values and average owner home values were used in all regressions. There were some difficulties in using this data. First of all, average rent level had a ceiling of $1250. This meant that the owners' implicit rent also had a ceiling of $1250. This ceiling is not very high, especially in areas of the country like New York City, Los Angeles, and Washington D.C. Also, in any block group in which there were less than six renters, the value for average rent was imputed by the Census Bureau. Therefore, any block group with less than six owners or renters was deleted from the regressions. These small block groups were later assigned weights using the fitted regression models.

The only imputation that is needed is owners' imputed rent, since all of the other variables come from the Census data. Regression analysis was used to impute owners' rent from average rent of the segment.

Nonlinear regression was used to determine imputed average rent values from average owner values. Two different nonlinear functions were compared. These were the exponential function and the square root function. The exponential function is of the form $\boldsymbol{y = \beta_0 * (1 - \exp(-\beta_1 * x))}$, where y is the average rent and x is the average owner value of the block group. SAS procedure NLIN was used to fit $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$. β_0 is the maximum value that the average rent can attain, while β_1 is a steepness coefficient. The square root function is of the form $\boldsymbol{y = b_0 * \sqrt{x}}$, where y is the average rent for the block group and x is the average owner value for the block group. Figure 3.1 shows a plot of the exponential function for A207(Chicago), while figure 3.2 is a plot of the square root function. Figure 3.3 is a plot of the residuals of the exponential function for A207, and figure 3.4 is a plot of residuals for the square root function.

Generally, the square root function did fairly well in predicting average rent from average owner value, although not as well as the exponential function. This is to be expected, however, since only one coefficient is found using the square root function, while two are found using the exponential function. The following table lists R^2 values for all self-representing PSU's in the sample for both the exponential and square root functions.

Table 3.1 R^2 values for self-representing PSU's

PSU	City	R^2 for exponential regression	R^2 for square root regression
A102	Philadelphia	50.7	53.6
A103	Boston	37.7	34.6
A104	Pittsburgh	46.4	48.0
A109	New York City	24.1	15.8
A110	NY-Conn. Suburbs	33.5	29.8
A111	NJ-PA Suburbs	42.6	41.2
A207	Chicago	51.5	47.3
A208	Detroit	50.4	56.0
A209	St. Louis	47.1	48.2
A210	Cleveland	58.7	58.6
A211	Minneapolis	33.1	26.2
A212	Milwaukee	49.0	47.9
A213	Cincinnati	37.3	36.8
A214	Kansas City	50.6	49.5
A312	Washington, DC	52.0	48.3
A313	Baltimore	43.9	41.9
A316	Dallas	48.4	45.8
A318	Houston	53.5	51.5
A319	Atlanta	48.5	44.8
A320	Miami	40.6	37.4
A321	Tampa	42.1	41.7
A419	Los Angeles County	44.0	43.2
A420	Los Angeles Suburbs	56.8	52.2
A422	San Francisco	40.7	37.4
A423	Seattle	42.3	42.4
A424	San Diego	36.5	32.9
A425	Portland	30.0	26.4

A426	Honolulu	19.8	20.7
A427	Anchorage	28.9	30.8
A429	Phoenix	46.7	45.3
A433	Denver	37.2	34.7

The above table shows that R^2 values for the negative exponential function generally are somewhat better than those for the square root function. The performance of the two models is very similar, and so the choice between the two models was difficult. The plot of residuals for the exponential function showed much more of a scattering effect than the plot of residuals for the square root function for A207. The square root function's residuals seemed to show a downward trend as average owner value went up. This may be because of the truncation of rental values at $1250. Since the exponential function is basically constant above owner values of $300,000, this same effect is not seen in the residual plot for the A207 exponential curve. A similar effect can be observed in plots of residuals for other PSU's.

The exponential function showed a little better performance, and also did not weight the high owner value housing units as much as they might have been weighted using the square root function. Therefore, we decided to go with the negative exponential function.

After choosing the exponential function, we looked at the resulting effect in expected numbers of segments chosen for different categories of composite rent. The following table summarizes the findings for the Chicago metropolitan statistical area for a total sample of 288 segments.

Composite rent category	Number of expected segments using expenditure weighting	Number of expected segments using current method
rent<$336.54	31.7	57.6
$336.54<=rent<$421.67	46.9	57.6
$421.67<=rent<$502.97	57.0	57.6
$502.97<=rent<$596.50	67.7	57.6
rent>=$596.50	84.7	57.6

As expected, expenditure weighting causes higher composite rent segments to be chosen more often than the current method.

CONCLUSION

The weights for the sample after 1999 should be much improved over the current weighting system. Using average rent values and imputed average rent values in the weighting scheme should help to choose fewer segments in dangerous areas, and should give better predicted expenditure weights. This is a much needed improvement over our current system.

REFERENCES

Brown, E.F., Johnson, W.H., (1994) "Stratification Designs for the Housing Sample of the Consumer Price Index", *Proceedings of the Government Statistics Section*, American Statistical Association.

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