Updating the Housing Age-Bias Regression Model in the Consumer Price Index

This paper examines the U.S. Bureau of Labor Statistics (BLS) past design and results of the age-bias regression model, as well as its current design. This narrative describes what has been done and presents current research on areas where improvement is warranted. The fundamental question is, "Does this model really capture depreciation for U.S. housing in the CPI?"

Louise Leonard Campbell

Louise Leonard Campbell is an economist in the Office of Consumer Prices and Price Indexes, U.S. Bureau of Labor Statistics. Email: Campbell.Louise@bls.gov

Background of Housing Aging Bias Estimation

The measurement goal of the Consumer Price Index (CPI), produced by the U.S. Bureau of Labor Statistics (BLS or the Bureau), is to accurately estimate the price change of a set of constant-quality consumer goods and services. The components of the CPI Housing Survey are Rent of Primary Residence (Rent) and Owners' Equivalent Rent of Primary Residence (OER). Together, these two components comprise more than 29 percent of the national CPI; that is, more than 29 percent of all consumption spending is attributed to renter-occupied and owner-occupied residential dwelling space. The indexes for the housing components measure changes in the cost of shelter services for renters and homeowners. Rent estimates are based on contract rents—what tenants pay their landlords for the flow-of-housing services as provided in the lease. This may include items, such as utilities or furnishings that are in addition to shelter. OER also uses a flow-of-services concept; estimates that owners of housing units would have to pay to rent equivalent shelter—without additions such as utilities.[1]

The CPI Housing Survey is the data source for calculating Rent and OER estimates. The CPI began using the current housing sample, beginning with the index for January 1999.[2] (This sample replaced the one used since 1987.) The current sample is based on the 1990 Census and is composed entirely of rental units selected to represent units in both largely renter and largely owner-occupied neighborhoods. OER estimates are based on changes in rents of renter units in the largely owner-occupied neighborhoods.[3]

With each passing year, the dwellings in the CPI housing sample deteriorate, losing some value (depreciate), and deliver less shelter service to their occupants. If this were not taken into account, the CPI would have a downward bias. To offset this quality loss, the aging bias due to depreciation, staff at BLS developed an age-bias adjustment.[4] This tool, which BLS began using with the CPI for January 1988, makes a monthly incremental adjustment to each housing unit in the sample to maintain a constant quality index over time.

The theoretical basis for the hedonic method used in estimating age bias is found in Randolph (1988). Randolph estimated that the age bias due to depreciation was 0.3 to 0.4 percent annually.[5] His 1988 regression model, adopted by BLS for the sample selected from 1980 Census data, accounted for the influence of structure type and

the rent control status of units. The age-bias regression model includes structural characteristics of housing units. Structural changes examined are changes in the number of bedrooms, number of bathrooms, number of other rooms (not previously mentioned), and additions or deletions of central air conditioning units. Subsequently, Bureau researchers realized that the model missed other dimensions of quality change, although parameter estimates for most housing structural components had reasonable algebraic signs and magnitudes. Since the age-bias regression provided, as a by-product, estimates of the value of structural changes, the CPI began using these estimates in 1989, to quality adjust for structural changes when changes occurred to sample units.

The CPI used the (old) 1980-Census-based housing sample from January 1987 through December 1998. In 1999, BLS replaced that sample with a new one, based on the 1990 Census. Along with the new sample, BLS specified a new regression model to adjust for age bias (depreciation).[6] Staff were forced to change the model, because some of the variables used in the pre-1999 model were no longer available. Additionally, this revision was the opportunity to align age-bias processing with the new sample.

This paper focuses on the change between the pre-1999 hedonic regression model and the model used since then in producing depreciation estimates for the housing stock.

The Pre-1999 Sample

The hedonic regression model used from 1988 through 1999 follows:

Log(rent) = f [**13 structural characteristics variables**
(detached, bedrooms, other rooms, complete kitchen, dishwasher, washer/dryer, oil heat, electric heat, central air conditioning, extra bathroom, rent control),
various location and survey variables, 13 neighborhood characteristics variables
(renters, race white, large buildings, two or more autos, without complete plumbing, air-conditioned, children age 6 to 18, college students, families below poverty level, elderly over 65, mobile homes, unemployment, with college education),
7 dummy variables for services provided with rent
(gas, electric, parking, furnishings, swimming pool, other recreation),
6 depreciation variables
(age, age x squared, age x rent control, age x old, age x detached, age x rooms)]
+ a random error term.

Both models rely on the CPI's geographic structure. The CPI geographic sample consists of selected metropolitan areas and urban places (primary sampling units or PSUs) where the CPI observes rents (and other prices) over time. BLS generates age-bias estimates for each PSU. In January 1999, the CPI housing component moved to the revised structure that the rest of the CPI had adopted one year earlier. Until this revision of the CPI, there were 88 PSUs–14 in the Northeast, 22 in the Midwest, 33 in the South, and 19 in the West. The model had dummy variables for these PSUs, and there were additional dummy variables for structural characteristics for each PSU within each Census region. Preparing data for the regressions was quite cumbersome. We weighted rent values by their relative cost on a unit-by-unit basis, according to the number of renters they represented within their area and the estimated owner and renter expenditures of the areas, and matched the CPI-collected data to 1980 Census data. From 1988 through 1999, we used the specification structure developed by Randolph. This structure generated depreciation, or age-bias factors, with rent levels of the most recently collected CPI data. For the purpose of this analysis, data for the year 1997 was chosen, because it was the most logical data to compare with the most recent annual regression estimates. This regression model was run for each Census region individually. The model specificity was unique to that region, and the regression estimates were particular to the geography of that Census region. Following are the variance analyses for each region:

Age-Bias
January 1997 Regression Model— **NORTHEAST** (CP 199612)
Dependent variable: LOGRENT
Analysis of Variance

Source	Degrees of freedom	Mean square	F-Value	Prob>F	R-Square	Adjusted R-square
Model	110	5.10886	75.239	0.0001	0.6945	0.6853
Error	3640	0.06790				
Total	3750					

Age-Bias
January 1997 Regression Model— **MIDWEST** (CP 199612)
Dependent variable: LOGRENT
Analysis of Variance

Source	Degrees of freedom	Mean square	F-Value	Prob>F	R-Square	Adj R-sq
Model	125	3.74195	66.206	0.0001	0.6647	0.6546
Error	4175	0.05652				
Total	4300					

Age-Bias
January 1997 Regression Model— **SOUTH** (CP 199612)
Dependent variable: LOGRENT
Analysis of Variance

Source	Degrees of freedom	Mean square	F-Value	Prob>F	R-Square	Adj R-sq
Model	168	5.57731	85.781	0.0001	0.7253	0.7169
Error	5457	0.06502				
Total	5625					

Age-Bias
January 1997 Regression Model— **WEST** (CP 199612)
Dependent variable: LOGRENT
Analysis of Variance

Source	Degrees of freedom	Mean square	F-Value	Prob>F	R-Square	Adj R-sq
Model	110	5.58910	110.900	0.0001	0.7453	0.7386
Error	4169	0.05040				
Total	4279					

Each regression model used at least 110 variables to generate the estimates. Although it is not shown here, the number of units used in the regressions decreased each year over the life of the 1987 sample. Because the sample used in age-bias processing was so closely tied by location to the 1980 Census data, units constructed after 1980 could not be added. (These units were added to the sample but could not be used in age-bias regression.) The result was that each year, fewer units were used in age-bias processing. Dwellings dropped out for one of four reasons: They were no longer rental units, they were condemned, they were converted to businesses, or they were destroyed. Although the regression model was well specified, the sample used to estimate it through 1999 became increasingly unrepresentative of the current pool of housing units.

Focusing on two other descriptive statistics from the tables above, the adjusted R-square indicates the amount of variation that the model explains with a correction for the degrees of freedom. The average adjusted R-squared for the old model structure is quite high at 0.7. Roughly speaking, this implies that 70 percent of the variation in log rent is explained by variation in the explanatory variables. The F Value is an indication of confidence in the independent variables in the model, showing the degree to which this set of variables has statistically significant impact on the dependent variable. The largest F value generated in this model structure is a bit over 100, at 110.9 for the West region. The fact that it is not larger may indicate that the models were over-specified.

The model used a complex array manipulation that simulated a maximum likelihood function. Each PSU, and several of the other independent variables, were arrayed in tandem with the weighted rent for each housing unit, to generate an age-bias factor for that specific PSU. The statistical program (SAS) consistently generated messages indicating co-linearity and warnings that the results were likely biased. Furthermore, staff were not comfortable with the number of zero values that the model created.

The Current Sample: The New Model

When the housing sample drawn from the 1990 Census, the model was revised. One major decision was to use ZIP Codes

to match Census data to CPI data in the set-up phase of the process. Doing this allowed new construction (built after the 1990 Census) sample units to become part of the analysis. The other decision was to use unweighted rent data, eliminating the need for maximum likelihood estimation. The new regression model is:[7]

Log (rent) = *f*[**10 structural characteristics variables, various location and survey variables** (detached, bedrooms, bedrooms squared, other rooms, other rooms squared, oil heat, electric heat, central air conditioning, window air conditioning, bathrooms), **2 survey variables** (A-size, B size), **10 neighborhood characteristics variables** (race white, large buildings, two or more autos, air-conditioned, children age 6 to 18, some college, families below poverty level, elderly 65 & over, mobile homes, unemployment), **3 dummy variables for services provided with rent** (gas, electric, parking), **5 depreciation variables, and a random error term** (age, age squared, age x old, age x detached, age x allrooms)] **+ a random error term.**

The CPI Housing Survey, based on the 1990 Census, includes only rental units and asks significantly fewer questions of respondents about their dwellings than does its predecessor. The number of PSUs was reduced from 88 to 87, and the various PSUs were not included as dummy variables in the age-bias model specifications. As a result, the regression specification became significantly simpler. For the years since its implementation in 1999, the annual estimate of age bias has been in the 0.2 to 0.3 range.[8] Results of the current model design are specified by the analysis of variance as follows:

The current age-bias model has significantly fewer regression variables, significantly more observations, and allows community variables from Census data to be matched by ZIP Code to new construction units, as long as the ZIP Codes existed in 1990. The F value of 854.4 is significantly higher than the one generated through the previous regression model. However, much of that increase can be explained by the adjustments in the degrees of freedom for the model and the error. Just as before, the probability of exceeding the F value is exceedingly small, so it can be stated with confidence that the fewer number of independent variables, the bigger the impact on log rent. The independent variables that were deleted did not contribute that much in explaining the variation in the dependent variable. That is why it was previously stated that the old model may have been over-specified. Generating the age-bias factor no longer requires the complex maximum likelihood function algorithm.

Naturally, omitted variable bias cannot be ruled out. The combination of changes may have resulted in the lower R-squared statistics and the lower values of the annually generated age-bias factor estimates. (See appendix I.) Appendix 1 shows the effect of age-bias values for the Nation annually since their integration in the production process. The revised sample, implemented in 1999, consistently yields values less than 0.3 percent annually.

Data Source Considerations

The most important independent variable in the current model is age, determined from the *year built* question. The year the unit was built is asked of respondents during initiation of the survey instrument, although we do not need to know the year built of age of a sample unit to use it to estimate rent change. We do need it, however, to use the unit in the age-bias regression. All units in the housing survey are rental units, and respondents often do not know the year the dwelling was built. At initiation, the majority of the units in the current sample, 55 percent, were missing year built. Subsequently, both national office staff in Washington, D.C., and data collectors in the field made a strenuous effort to learn the year built from other sources and BLS has been able to collect year-built data for many sample units. From 2003 through 2005, the number of units with year-built data increased to 89 percent.

ZIP Code data for units is important so Census data that contain neighborhood characteristics can be matched to new construction built in the 1990s and later. (The database is updated periodically with appropriate ZIP Codes.)

Future Research

BLS is planning to replace the current housing sample over a 6-year period, beginning in 2009. Up to now, Census informa-

Age-Bias
January 2006 Regression Model— **ALL UNITED STATES** (CP 200511)
Dependent variable: LOGRENT
Analysis of Variance

Source	Degrees of freedom	Mean square	F-Value	Prob>F	R-Square	Adj R-sq
Model	30	89.89777	854.37	0.0001	0.5202	0.5196
Error	23642	0.10522				
Total	23672					

tion has been the basis for the housing survey's sample frame. The new sample will be drawn from the 2000 Census. In the 2000 Census, however, detailed neighborhood questions moved to the *long form*, so fewer respondents were asked them. The Census Bureau is implementing a new survey, the American Community Survey (ACS) to replace much of the housing data they collect in the Decennial Census. If the ACS grows as planned, it will be a major data source for the aging bias model for the Housing CPI.

Current research may eventually lead to changes in the way the factors are regressed and applied. Age-bias factors might be processed by regression procedures at index area levels and applied at the unit level. Preliminary research found that the weight of units designated as *old*, units built before 1900, tends to result in smaller age-bias factors.

Other data sources that BLS researchers believe will enhance regression results have been identified. For example, researchers will use average income by ZIP Code as one of the independent variables in the functional form. (This data has only recently become available from the Internal Revenue Service.[9])

A Small Global Survey

The Ottawa Group is an international organization of individuals responsible for price programs worldwide. In 2003, BLS staff asked members of the Ottawa Group if their price programs use hedonics of rental units to estimate depreciation for the housing stock. Over 90 percent of those members surveyed do not estimate depreciation. Just as the current Bureau model is not truly comparable to the previous one, this lack of comparability is even greater for other countries and their structure for estimating price change. The Bureau's regression model is unique; our method of estimating shelter is also rare.

Conclusion

The Bureau's two (former and current) age-bias regression models are not completely comparable. Clearly, each produced, or produces, estimates used in the production of the Rent of Primary Residence (Rent) and Owners' Equivalent Rent of Primary Residence (OER) Indexes. The R-square value for the current model structure is significantly lower than it had been previously. However, this is to be expected, since there are fewer independent variables—and combining all four regions into one—almost certainly increases the variation. There are still the potential problems of the data source considerations mentioned above and a now-higher likelihood of omitted variable bias. The CPI program treats the hedonic regression effort as an area of constant improvement, because the quality of the index is affected by its being applied correctly.

The first graph in appendix II shows the effect of age-bias estimates on the Rent and OER Indexes, since 1988. The second graph shows the effect of age-bias adjustments on the published All Items Consumer Price Index for the same period. Over time, the average increase in the CPI due to age-bias estimates is approximately 3.0 percent.[10] Unit-level estimation may yield better results than the practice of generating yearly estimates at the PSU level, and then applying them to individual units on a monthly basis.

Notes

Special thanks go to Walter Lane, Frank Ptacek, Randal Verbrugge, and Ronald Johnson for their helpful and insightful input.

[1] Details regarding the estimation of price change for shelter can be found in the *Handbook of Methods*, Chapter 17. The Consumer Price Index, pp. 23-25.

[2] See "Revision of the CPI housing sample and estimators," *Monthly Labor Review* by Robert Baskin and Frank Ptacek, pp. 31-39, December 1996, **http://stats.bls.gov/opub/mlr/1996/12/art5full.pdf**.

[3] Current statistical methods and economic theory regarding how the BLS handles owner-occupied rent can be viewed on line. Also, see *Treatment of Owner-Occupied Housing in the CPI*, Robert Poole, Frank Ptacek and Randal Verbrugge. This paper was presented before the Federal Economic Statistics Advisory Committee (FESAC), December 9, 2005, **http://stats.bls.gov/bls/fesacp1120905.pdf .**

[4] See "Adjusting the CPI shelter index to compensate for the effect of depreciation," Walter F. Lane, William C. Randolph and Stephen A. Berenson, *Monthly Labor Review*, pp. 34-37, Technical Notes, October 1988— **http://stats.bls.gov/opub/mlr/1988/10/rpt1full.pdf**.

[5] See Housing *"Depreciation and Aging Bias in the Consumer Price Index*," William C. Randolph, *Journal of Business & Economic Statistics*, July 1988, Vol. 6, No. 3, p. 365.

[6] William Thompson, a supervisory economist in the Division of Consumer Prices and Price Indexes, designed this model in 1999.

[7] Ibid.

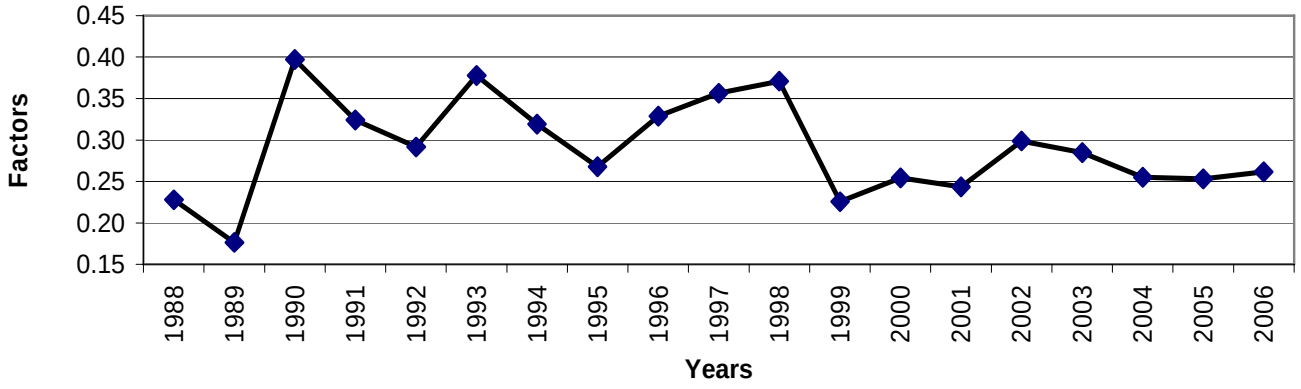
[8] Current statistical methods and economic theory regarding how the BLS handles owner-occupied rent can be viewed on line. Also, See *Treatment of Owner-Occupied Housing in the CPI*, Robert Poole, Frank Ptacek and Randal Verbrugge. This paper was presented before the Federal Economic Statistics Advisory Committee (FESAC), December 9, 2005, **http://stats.bls.gov/bls/fesacp1120905.pdf**, p 29.

[9] Regressions have been performed using 2002 IRS average gross income (AGI) data by ZIP Codes. Researchers used AGI and log AGI as independent variables, yielding small improvement in the variance analysis statistics.

[10] See "What has happened to price measurement since the Boskin Report? The U.S. Experience," David S. Johnson, Stephen B. Reed and Kenneth J. Stewart, U.S. Bureau of Labor Statistics; "OECD Conference: Inflation Measures: Too High – Too Low – Internationally Comparable?" Paris 21-22 June 2005, pp. 10-11.

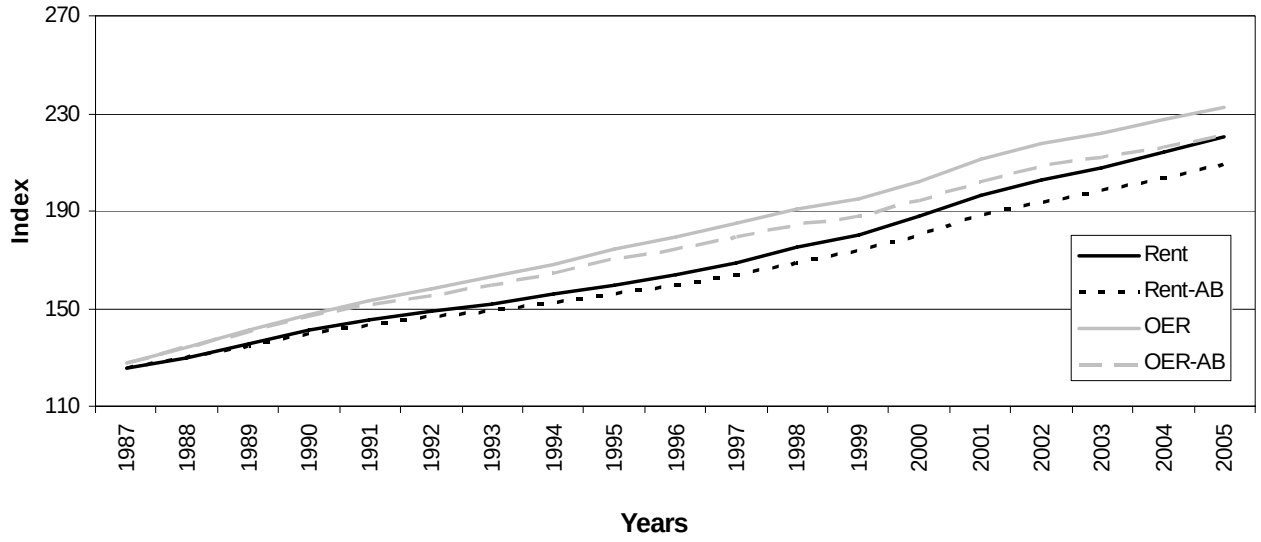
Appendix I

Annual CPI housing age-bias (AB) factors



Appendix II

Graph 1. **CPI Rent and OER Indexes with and without housing AB factors**



Graph 2. **Published CPI & CPI without housing AB factors**