ESTIMATION OF CORRELATION BETWEEN YEARS FOR THE AMERICAN HOUSING SURVEY

Jay J. Kim, Robert Fay III and George Train III
Jay J. Kim, Bureau of the Census, Suitland, MD 20233

I. Introduction

The Bureau of the Census conducts the American Housing Survey-National Sample (AHS-N) every other year. To investigate whether the change over time is significant or not for some survey variables of interest, the correlation over two different time points is required for the variable. Thus the correlation coefficients over two different time points have been calculated for some AHS-N survey variables. The survey panels involved in the correlation computations are 1985-1991 AHS-N. Since the survey was conducted biennially, the correlations we are interested in are for the survey variables that are two, four and six years apart. Pairs of AHS-N's which are separated by two years are 1985 and 1987 AHS-N, 1987 and 1989 AHS-N, finally 1989 and 1991 AHS-N. Pairs of AHS-N's which are separated by four years are 1985 and 1989 AHS-N, and 1987 and 1991 AHS-N. Only one pair of AHS-N's are separated by six years which are 1985 and 1991 AHS-N.

It should be noted that starting with the 1987 AHS-N and repeated every four years a rural supplement, which amounted to 50 percent of the usual rural sample cases was included. Table 1 shows that there are more common units between 1985 and 1989 AHS' and between 1987 and 1991 AHS' due to rural supplemental units. Its effect will show in the size of correlation later. To account for the additional units, base weights are adjusted for the rural sample housing units whenever applicable.

Table 1. Sample Size, 1985 - 1991

	Interview	Total
1985	42,500	44,300
1987	49,600	51,300
1989	44,800	46,700
1991	51,000	53,400

This paper shows the methodologies used for computing correlations and the results.

II. Calculation of Correlation Coefficients - AHS-N

The Bureau of the Census conducts AHS-N biennially. When new data comes in, the subject matter people sometimes get interested in whether the new value is significantly different from that of the previous years. In order to do the test for the change, the correlation over time is required for the variable. Thus the correlations were calculated using the balanced half-sample replication (BHR) approach. This means the variance and covariance were computed using BHR, since the correlation is composed of them.

The AHS-N sample is composed of housing units selected from the self-representative (SR) Primary Sampling Units (PSUs) and those from the nonself-representative (NSR) PSUs. For calculating the variance for the AHS, two NSR PSUs were collapsed into one stratum to mimic a collapsed stratum. Thus NSR PSUs have both a between- and within-PSU variance component. On the other hand SR PSUs have a within-PSU variance only. The within-PSU variance was computed by forming two sampling error computation units (SECUs) within a SR PSU or a combination of SR PSU's. These two SECUs constitute a (collapsed) stratum and a NSR PSU is a SECU.

Altogether, we used 48 replicates and the full sample for BHR estimation. Replicates were formed using Hadarmard Matrix. As an example 4x4 Hadarmard matrix is given in Table 2 below.

Table 2. Hadarmard Matrix (4x4)

		Stratum			
		1	2	3	4
Replicates	1	-	+	+	-
	2	+	+	-	-
	3	+	-	+	-
	4	-	-	-	-

Now a stratum has two SECUs. "+" in the table indicates the first SECU from the pair of SECUs is used to form a given replicate. Traditionally, for the pair of SECUs in a stratum, replicate factors of 2 and 0 are used depending on whether the SECU is in a replicate or not. However, Robert Fay proposed alternate scheme using 1+E and 1-E instead of 2 and 0. Fay's approach is adopted for this research. Replication factors were developed for each survey year, but it was decided to use 1991 factors for all years. The rationale is as follows. When replication factors were developed for each survey year, the possibility of calculating correlation over two survey years was not considered. Hence the replication factors were developed separately for each survey year and as a result they were not consistent over the pair of years. That is, for example, roughly speaking, the same replicate can receive a factor of .5 in one year, but 1.5 in the other year, or vice versa, rather than receiving .5 or 1.5 in both years.

Replicate estimates are calculated as follows;

Let

- G be the number of items defined;
- I be the number of replicates, i.e., 48
- S be the number of segments;
- R_{is} be the replicate factor assigned to segment s

for replicate i;

Y_{gs} be the weighted value for segment s and item g; and

X_{gi} be the replicate estimate for replicate I and item g.

Then

$$X_{gi} = \sum_{s}^{S} Y_{gi} R_{is} \qquad \text{for } g = 1,2,.....,G$$

$$i = 0,1,.....,I.$$

When i = 0, X_{gi}, i.e., X_{g0} is the full sample estimate for item g, since $R_{0s} = 1.0$ for s=1,2,...,S. Then the estimated variance for the g^{th} item can be computed as follows;

$$Var(X_g) = \frac{4}{I} \sum_{i=1}^{I} (X_{gi} - X_{g0})^2 \qquad .....(1)$$

Using (1.5, .5) instead of (2, 0) underestimates the variance by the factor of 4 which is from $\frac{1}{(1.5-2)^2}$ or $\frac{1}{(.5-0)^2}$. Thus the variance is multiplied by 4 in the above formula.

Let

X_{gt} t=a,b be the estimated total for the g^{th} item for time t;

X_{git}, t=a,b be the replicate estimate for replicate i and item g for time t;

and

X_{g0t}, t=a,b be the full sample estimate for item g for time t=a,b.

Then

$$Cov(X_{ga},X_{gb}) = \frac{4}{I} \sum_{i=1}^{I} (X_{gia} - X_{g0a})(X_{gib} - X_{g0b}) \qquad .....(2)$$

The estimated correlation is

$$Corr(X_{ga},X_{gb}) = \frac{\sum_{i=1}^{I} (X_{gia} - X_{g0a})(X_{gib} - X_{g0b})}{\sqrt{\sum_{i=1}^{I} (X_{gia} - X_{g0a})^2 \sum_{i=1}^{I} (X_{gib} - X_{g0b})^2}}.$$

VPLX was used for actual calculation of correlation.

IV. Computed Correlation Coefficients

In Table 3, three sets of correlations are given as well as the averages of the three sets except for "year built." When more than one set of correlations is available, averages were taken.

Fig 1. Corr for year built - 2 years apart

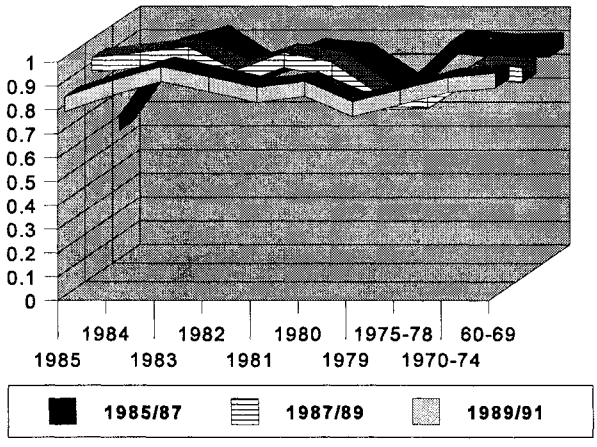
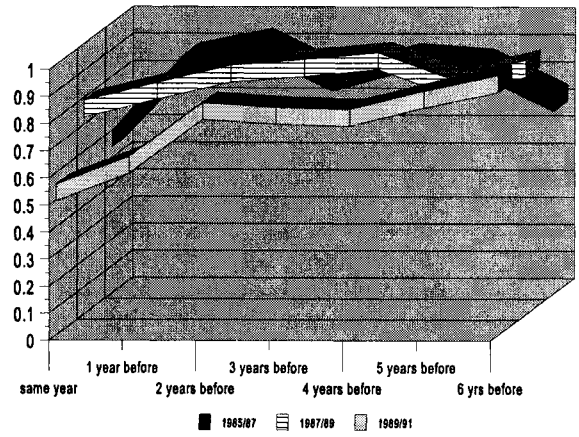


Fig 2 Correlation for year built - between two years



As for the "year built," there is a tendency that when one of

two years coincides with one (earlier year of the two) of two years in which the surveys were conducted, the correlation is much lower than the rest. For example, in Table 3 when the correlation is computed between 85 and 87 AHS-N for the year built which is 85, the correlation is .570. However, the same correlation for the built year which is 84, just one year earlier than 85, jumps to .854. Also when the earlier survey year is one year after the year built, the correlation tends to be slightly lower. Using the same pair of years in the same table, we can see the correlation for year built which is 84 is .854, but if the year built is 83, it goes up to .909. This phenomenon repeats for other pairs of years. This phenomenon is best illustrated by comparing Fig 1 and Fig 2. In Fig 1, correlations were charted in chronological order and in Fig 2, they were graphed based on the number of years after construction. The correlation is variable quite a bit in Fig 1, but stable in Fig 2. Thus an averaging technique different from other variables is taken for this variable, i.e., only three correlation coefficients were calculated, even if the number of years built is many (at least 15). One of the three is for the case when the year built (for earlier survey of the two) is the same as the year the survey was conducted, the second is for the case when the earlier survey year is one year after the year built, and the third for the remainder which is average of all the remaining years.

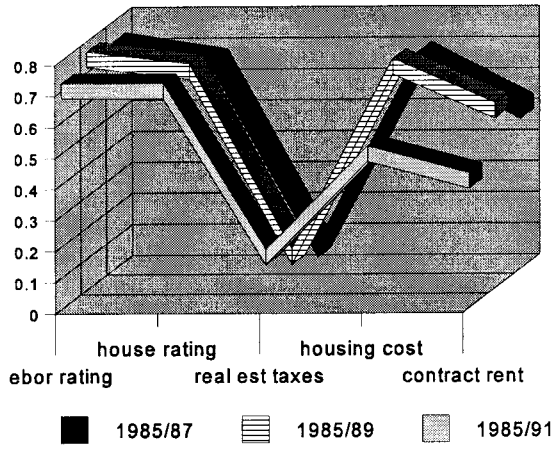
In Table 4, two sets of correlations as well as average correlations are given. One might assume that the correlations or average correlations between two years of AHS would diminish as the period between survey years increases. However, comparing Table 3 with Table 4 one can find that, 23 times out of 77, this was not the case. This might be explained as follows. Starting in 1987, and every four years thereafter, supplemental rural sample units were added. In these years, the existing rural sample units are assigned lower weights to assign weights to the supplemental units. Thus when correlation is calculated between two years of AHS which are separated by a period of two years as in Table 3, one of the two years always contains the

supplemental units. In this situation, we have an imbalance between the years. However, when correlation is calculated between two years of AHS which are separated by a period

of four years as in Table 3, both years have the supplemental units or neither have the supplemental units. Thus we have a perfect balance between the two years of AHS. Therefore, it is no surprise to see some higher correlations between the years of AHS which are separated by a longer period.

In Table 5, one set of correlations between two years of AHS six years apart is given. As seen in the table, some values are higher than those calculated for a shorter period. Most of them can be seen as a random variation as this set does not involve averaging. However, correlations for coal

Fig 3 Correlation - 2 to 6 years apart



and noncash renter in this table are higher than those in other tables disregarding the length of period, which can not be explained. We need more years of data to examine this type of behavior.

Fig 3 shows the correlations for five variables between years which are two, four and six years apart. Those for neighborhood rating, house rating and real estate taxes do not deteriorate as the length of period increases, but they decrease quite a bit for housing cost and contract rent.

V. References

Fay, R.E. (1995), VPLX Manual, internal Census Bureau document.

Judkins, D.R. (1990), "Fay's Method for Variance Estimation," Journal of Official Statistics, Vol. 6, No. 3, pp. 223-239.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," Journal of Official Statistics, Vol. 1, No. 4, pp. 381-397.

Rust, K. and Kalton, G. (1987), "Strategies for Collapsing Strata for Variance Estimation," Journal of Official Statistics, Vol. 3, No. 1, pp. 69-81.

Variance Work Group (1987), "Replicate Variance Estimation System," internal Census Bureau memorandum.

Table 3. Correlation in the Case of Survey Years Separated by Two Years

	85/87	87/89	89/91	Average
Year Built				
1989			.528	The year built is the
1988			.631	year the survey was
1987		.757	.828	conducted : .618
1986		.825	.812	
1985	.570	.889	.805	The survey year is
1984	.854	.911	.873	one year after the
1983	.909	.936	.931	year built : .770
1982	.779	.839	.889	
1981	.854	.911	.844	The survey year is
1980	.834	.872	.870	more than one year
1979	.705	.736	.785	after the year built
1975-78	.892	.742	.835	: .861
1970-74	.874	.864	.886	
1960-69	.879	.850	.904	
1950-59	.894	.793	.887	
1940-49	.905	.876	.942	
1930-39	.904	.883	.881	
1920-29	.884	.901	.907	
1919 or earlier	.803	.879	.868	
Neighborhood rating	.709	.793	.866	.789
House rating	.655	.802	.833	.763
Real Estate Taxes	.063	.359	.414	.279
Housing Cost	.676	.557	.528	.587
Contract Rent	.508	.498	.815	.607
Utility Cost				
Electricity	.607	.668	.713	.663
Gas	.573	.734	.616	.641
Oil	.719	.667	.629	.672
Coal	.679	.642	.581	.634
Tenure				
Owner	.677	.821	.731	.743
Cash Renter	.511	.622	.852	.662
Noncash Renter	.556	.493	.538	.529

Table 4. Correlation in the Case of Survey Years Separated by Four Years

	85/89	87/91	Average
Year Built			
1987		.828	The survey was conducted
1986		.812	in the year built : .636
1985	.444	.805	
1984	.838	.873	The survey year is one year
1983	.911	.931	after the year built: .825
1982	.837	.889	
1981	.820	.844	The survey year is more
1980	.790	.870	than one year after the
1979	.607	.785	year built : .845
1975-78	.740	.835	
1970-74	.892	.886	
1960-69	.832	.904	
1950-59	.778	.887	
1940-49	.816	.942	
1930-39	.834	.881	
1920-29	.860	.907	
1919 or earlier	.834	.868	
Neighborhood rating	.741	.866	.804
House rating	.702	.833	.768
Real Estate Taxes	.104	.414	.259
Housing Cost	.716	.528	.622
Contract Rent	.572	.815	.694
Utility Cost			
Electricity	.613	.713	.663
Gas	.607	.616	.612
Oil	.762	.629	.696
Coal	.678	.581	.630
Tenure			
Owner	.786	.731	.759
Cash Renter	.441	.852	.647
Noncash Renter	.414	.538	.476

Table 5. Correlation in the Case of Survey Years Separated by Six Years

	85/91	Suggested Value for yr built
Year Built		
1985	.425	The survey was conducted
1984	.821	in the year built : .425
1983	.881	
1982	.756	The survey year is one year
1981	.770	after the year built : .821
1980	.734	
1979	.503	The survey year is more
1975-78	.688	than one year after the
1970-74	.821	year built : .743
1960-69	.784	
1950-59	.721	.749 (incl 84)
1940-49	.814	
1930-39	.667	
1920-29	.790	
1919 or earlier	.731	
Neighborhood rating	.702	
House rating	.701	
Real Estate Taxes	.165	
Housing Cost	.497	
Contract Rent	.409	
Utility Cost		
Electricity	.597	
Gas	.558	
Oil	.650	
Coal	.729	
Tenure		
Owner	.578	
Cash Renter	.371	
Noncash Renter	.485	