Sample Truncation in FHA Data: Implications for Home Purchase Indexes*

JOHN S. GREENLEES
Bureau of Labor Statistics
Washington, D.C.

I. Introduction

The home purchase index is the single most important price series in the Consumer Price Index (CPI). In December 1980, for example, the relative importance of home purchase in the CPI for all urban consumers was 10.3 percent. Another component of homeownership cost, contracted mortgage interest, had a weight of 9.8 percent in the CPI. The mortgage cost index is computed as the product of the home purchase index and an index of mortgage interest rates. Thus, a ten percent increase in measured home purchase prices is sufficient to increase the all-items CPI by approximately two percent.

The importance of the home purchase index places great value on the accuracy of its measurement. However, the index is one of the most often criticized of all CPI series. The primary objections result from the decision by the Bureau of Labor Statistics (BLS) to base the index on sales data provided by the Federal Housing Administration (FHA). These FHA data have been employed in numerous econometric studies, including analyses of housing demand [14], place-to-place house price indexes [16], and the substitutability of land and capital in the production of housing services [18]. At the same time, the FHA samples are widely recognized to be unrepresentative of the universe of house sales, with possible distorting effects in some applications.[1] Also, in recent years the home purchase index has differed markedly from other national-level house price indexes published by the U.S. Bureau of the Census and the National Association of Realtors.[2]

Probably the most critical problem with the FHA data base results from the program's ceilings on the size of insurable mortgages. These ceilings historically have

* The author is grateful to his colleagues in the Division of Price and Index Number Research, and to an anonymous referee, for many valuable comments and suggestions. He also thanks Jerry Hausman for providing the computer program used in likelihood function maximization, and Albert Knopp for assistance in obtaining FHA Master File data. The views expressed are those of the author and do not represent an official position of the Bureau of Labor Statistics or the views of other staff members.

1. For example, de Leeuw [3] and Polinsky and Ellwood [14] argue that estimates of the income elasticity of demand for housing obtained from FHA data should be adjusted upward by 40 to 50 percent if they are to represent the United States population as a whole.

2. Comparisons of historical series have been presented by Mitchell [9] and Triplett [19].

confined FHA insurance to the low end of the house price spectrum. The home purchase index therefore relies on an implicit assumption that the price movements of smaller, older, less well-located, or otherwise less valuable homes can be taken as representative of the entire housing market. Furthermore, even if this assumption is valid, statistical theory suggests that the FHA sample will yield biased estimates of price change, since the mortgage ceilings cause the sample to be drawn indirectly on the basis of price. In any given period, the FHA sample mean price will lie below the mean of the population of house sales. During periods of inflation the population distribution of prices will rise relative to the mortgage ceilings, the FHA sample will become more truncated, and the ratio of the sample mean to the population mean will fall. Conversely, when FHA raises its dollar ceilings, the sample mean price will tend to rise independently of any movements in the population distribution. In general, therefore, month-to-month price change will be underestimated so long as the ceilings remain in effect, and overestimated during periods immediately following upward ceiling adjustments.

Tables I and II indicate the extent of FHA sample truncation and the importance of the mortgage ceiling levels. As shown in Table I, the median sales price in 1978 of new FHA-insured houses was more than $15,000 lower than the median for all single-family construction. Only four percent of FHA houses sold for more than $60,000, the level at which the mortgage ceiling was set in November 1977. Previously the ceiling had been raised in August 1974 from $33,000 to $45,000. Table II shows that these ceiling adjustments were associated with sharp increases in FHA's market share following periods of slow decline. The FHA percentage rose from 6 to 15 in the last two quarters of 1974, and from 6 to 11 between mid-1977 and early 1978. It is important to note that these patterns are not evident for the VA mortgage guaranty program, which shares FHA's interest rate policies but not its mortgage size restrictions.

The present CPI index computation method has two means of dealing with the problem of sample truncation. First, the house sample is stratified by age and living area, and price is measured in dollars per square foot, in order to correct for intertemporal changes in the average "quality" of FHA houses. It is readily seen that if age and square footage were the only determinants of house price, truncation of the sample would not bias the CPI inflation estimates. Second, whenever the FHA ceilings are raised, the BLS "links out" subsequent price changes in a manner similar to that used in other markets when products are replaced or experience marked quality change. Recently the BLS has

Table I. Percentage Distribution of New Single-Family House Sales, 1978, by Sales Price and Type of Financing

Price	All Homes[a]	FHA Homes[b]
Under $30,000	3	7
$30,000 to $39,999	14	41
$40,000 to $59,999	40	48
$60,000 or Over	42	4
Total	100	100
Median Price	$55,700	$40,400

[a]Source: [20, 9-14].
[b]Source: [21, 38].

undertaken an evaluation of the CPI home purchase index procedures at the local, regional, and national levels. In this paper I report on the first step in that research: a determination of whether the truncated nature of the FHA samples necessarily leads to a significant bias or loss of accuracy in any index based on FHA data. Using data from each of three metropolitan areas, I compare hedonic indexes estimated by ordinary least squares (OLS) to indexes generated by a regression technique which explicitly corrects for sample truncation.

The statistical model used for regression estimation in the truncated case is that presented by Hausman and Wise [6]. Its application to the case under consideration here is discussed in Section II below. I also discuss the use of a testing procedure (derived by

Table II. Percentage Distribution of New Single-Family House Sales, 1973-1978, by Type of Financing[a]

Quarter		FHA Insured	VA Guaranteed	Conventional and Other
1973	I	13	12	75
	II	10	12	77
	III	8	11	82
	IV	7	10	83
1974	I	5	13	82
	II	6	12	82
	III	10	15	75
	IV	15	14	71
1975	I	12	15	72
	II	11	13	76
	III	9	12	79
	IV	7	10	83
1976	I	9	13	78
	II	8	10	81
	III	8	11	80
	IV	9	13	78
1977	I	11	13	76
	II	8	11	81
	III	6	11	83
	IV	10	10	80
1978	I	11	11	79
	II	11	10	80
	III	12	9	79
	IV	12	10	78

[a]Source: [20, 13]

Hausman [5]) to determine whether the OLS and truncated regression estimation methods yield significantly different indexes. Section III of the paper describes the specification and estimation of the hedonic regression functions, and presents and compares the price indexes obtained for the Minneapolis-St. Paul, Chicago-Northwestern Indiana and San Francisco-Oakland SMSAs. Conclusions and implications of the analysis are the subject of Section IV.

II. Data and Methodology

The data bases for the analysis reported in Section III were drawn from the FHA's 1969–73 and 1974–78 Master Statistical Files. Together these files contain data on more than 1.6 million FHA-insured residential property sales, and should be approximately equal to a combination of all the monthly CPI tapes received by the BLS over a ten-year period. In addition to the price, age, square footage, and location data used in the CPI, each house record includes a wealth of financial and appraisal information describing the property and purchaser. Less complete information on neighborhood and site variables was available for years prior to 1972. I therefore restricted my attention to the 1972–78 period. I also deleted a number of house records as questionable or out of scope, using the same editing criteria found in the CPI processing algorithm.

The heterogeneity of the dwelling units found in the FHA sample and the detailed information on housing characteristics provided in the Master File records led naturally to my choice of hedonic regression as a means for constructing house price indexes. Griliches [4] summarizes this now well-known approach and reviews several applications. Others presenting theoretical discussions include Sherwin Rosen [17], Muellbauer [10], and Pollak [15]; for a recent application to housing prices, see Palmquist [12].

Hedonic indexes are based on the specification of a functional form relating product price to levels of measurable product characteristics. Regression estimation of this function for different time periods then yields an index of the cost of acquiring a given vector of characteristics. Pollak [15] demonstrates that, under certain conditions, this constant-characteristics index may be interpreted as an upper bound on the true cost-of-living index—i.e., of the cost of attaining the level of indifference implied by the base characteristics bundle.

The price-characteristics function is assumed to take the form

$$y_i = X_i\beta + \varepsilon_i \tag{1}$$

where y_i is the logarithm of the sale price of house i, X_i is a $1 \times k$ vector of house characteristics (such as age, square footage, the presence or absence of a garage, and neighborhood descriptors), β is a $k \times 1$ vector of parameter values, and ε_i is a random disturbance term with expectation zero and variance σ^2. This is a fairly standard hedonic specification;[3] however, its application to FHA sales data must be based on an adequate modelling of the implicit FHA sample selection process.

3. The popularity of the semi-logarithmic form exemplified by (1) has been noted by, for example, Griliches [4] and Muellbauer [10]. Most recently Palmquist [12] finds the semi-log to out-perform several alternative functional forms according to the criteria suggested by Box and Cox [2].

A number of distinctive program provisions are important in determining the relative attractiveness, to buyers and sellers, of FHA financing. As compared to typical conventional loan terms during our period of study, the liberal FHA loan-to-value ratios should have been attractive to buyers. Sellers, meanwhile, may have been discouraged by the requirement that they pay "points" to offset a lower-than-market interest rate. Variations in these relative terms cause the FHA market share and sample size to differ from area to area and month to month.[4] This creates no particular econometric problems in estimating equation (1). Neither are problems caused by the fact that the FHA sample is unrepresentative—i.e., the distribution of X differs from that in the population of house sales. Rather, the difficulties arise because the FHA mortgage ceiling creates an indirect negative relationship between the dependent variable in (1) and the probability of inclusion in the FHA sample. Under these conditions ordinary least squares becomes an unacceptable estimation method.

The FHA sample selection bias takes effect in the higher ranges of house prices, and can be illustrated by the provisions regarding house value and down payment requirements. For example, between 1974 and 1977 the mortgage ceiling was $45,000. The maximum FHA loan amount was 97 percent of the first $25,000 of appraised house value, 90 percent of the next $10,000, and 80 percent of any additional value.[5] A simple calculation yields $49,687.50 as that house price (actually, house value) which would have required the ceiling mortgage when the purchaser contributed the smallest allowable down payment. For houses priced below $49,687.50, therefore, the ceiling should have played no role in the choice between FHA and conventional financing. Above this price, the relative attractiveness of FHA terms deteriorated markedly. With the mortgage ceiling at $45,000, a house valued at $50,000 required a minimum FHA down payment of 10 percent. The minimum rose to 25 percent for a $60,000 home, 40 percent for a $75,000 home, and so on. Since conventional loan markets did not operate in such a fashion, we expect that in this value range the probability of FHA financing declined with price. In econometric terms, this means that house sales with high values of y_i and ε_i were less likely to be sampled. As a result, the expectation of ε_i is negative for all FHA observations. Further, ε_i and X_i cannot be assumed to be independent, since the expectation of ε_i is lower when $X_i\beta$ is high. The assumptions of the standard linear model are violated, and OLS can be expected to yield biased and inconsistent estimates of β.

In my initial empirical work for this paper, I experimented with a formal econometric model of stochastic sample truncation. The probability that buyer and seller would agree to FHA financing was specified as a probit function of sales price and other factors such as age of house, location, and the difference between FHA and conventional mortgage rates. The result, when combined with the hedonic function (1), was variant of other stochastic censoring and truncation models, such as those proposed by Nelson [11], Heckman [7], and Lee [8]. Unfortunately, this model appeared to be very sensitive to minor changes in specification, probably because of its critical and untestable distributional assumptions.

4. As Zerbst and Brueggeman [23] and others have pointed out, the differential between FHA, VA, and conventional mortgage interest rates also implies that the type of mortgage can have an independent effect on house price. That is, FHA houses will tend to sell for higher prices holding house quality constant, because the seller has to pay discount points to equalize yields to lenders. This should only affect the CPI to the extent that the interest rate differential shifts over time.

5. The requirements described here and in Section III are drawn from Title 25 of the Code of Federal Regulations as periodically amended.

Since it did not produce plausible index series, the model will not be discussed further here.

An alternative, and in some respects preferable, method of analyzing truncated samples within a regression framework has been developed and applied by Hausman and Wise [6]. This approach develops maximum likelihood estimates of β and σ^2 under the assumption that the disturbances ε_i are distributed normally in the population of sales, while the sample design is assumed to be such that cases are included when y_i is at or below some limit C_i, and excluded otherwise. As a result, all observations in the sample satisfy the condition

$$\varepsilon_i \leq C_i - X_i\beta. \tag{2}$$

The probability of such an event occurring is $\Phi((C_i - X_i\beta)/\sigma)$, where Φ represents the standardized normal c.d.f. This leads to the following expression for the likelihood of a sample observation:

$$L_i = (2\pi\sigma^2)^{-1/2} \exp(-(y_i - X_i\beta)^2/2\sigma^2)/\Phi((C_i - X_i\beta)/\sigma). \tag{3}$$

The log-likelihood function for a sample of T observations is given by

$$\log L = -(T/2)\log(2\pi) - T\log\sigma - (2\sigma^2)^{-1}\sum_{i=1}^{T}(y_i - X_i\beta)^2 - \sum_{i=1}^{T}\log\Phi((C_i - X_i\beta)/\sigma). \tag{4}$$

This differs from the sample log-likelihood in the standard linear model only in the inclusion of the final summation term. Hausman and Wise note that the parameter estimates $\hat{\beta}$ and $\hat{\sigma}$ resulting from maximization of (4) are consistent and asymptotically normal.

The above model can be adapted to our present purposes under the assumption that serious truncation of the FHA sample takes place only as a result of the mortgage ceiling, and that below some threshold value house price plays no role in determining the probability of FHA financing. As discussed earlier in this section, a natural choice for the threshold value C_i would be, for example, \$49,687.50 when the FHA mortgage ceiling was \$45,000. The FHA sample can then be edited to exclude all quotes above the threshold price, and consistent estimates of the parameters of equation (1) can be obtained by maximizing (4) using the edited sample. This approach has the possible weakness that it ignores any sample selection biases resulting from factors other than the mortgage ceiling. It also does not make use of all the sample observations. As compared to the above-mentioned stochastic truncation model, however, it has the advantage that the truncation is not required to follow any particular functional form (such as a probit or logit) and is allowed to be discontinuous at (or above) the threshold, as would be expected given the ceiling regulations.

In Section III, I employ both OLS and the Hausman-Wise technique (which I will refer to as truncated regression or TR) to estimate alternative hedonic indexes of house prices.[6] The OLS regressions are estimated using the full FHA samples, the TR regressions using edited samples as discussed above. If no sample truncation problem exists in the FHA data, OLS and TR should produce similar index series, since both methods yield

6. The generalized Gauss-Newton algorithm developed by Berndt, Hall, Hall, and Hausman [1] is used to maximize the likelihood function (4).

consistent estimators of the β vector. Conversely, if it is true that some kind of truncation rule is in effect at high price levels, the two parameter vectors should diverge. (Note that OLS using the *edited* sample would yield inconsistent estimates under any conditions.)

In this paper I follow the practice, common in hedonic studies, of constructing the indexes from a series of adjacent-period regressions. That is, the rate of price increase between period i and period $i + 1$ is estimated using a price regression which utilizes all sales during the two periods and which includes as one regressor a binary "time dummy" equalling unity in period $i + 1$. The coefficient on this variable is the estimate of price change, and an index series from periods 1 to N is obtained by "chaining" the price relatives from the N adjacent-period regressions.

The primary alternative to this adjacent-period or "pairwise aggregation" approach involves the estimation of N hedonic regressions each using data from only a single period. The index is then derived by evaluating each estimated regression at some representative set of explanatory variable values. A variety of operational rather than theoretical considerations motivated my selection of the adjacent-period method.[7] First, pairwise aggregation, by including an average of twice as many observations in each regression, is likely to result in parameter estimates which have smaller standard errors and are more stable from period to period. Second, the parameter restrictions inherent in the pairwise aggregation method are equivalent to a maintained assumption of this paper—namely, that the rate of inflation in price is unrelated to the X vector. If this assumption were invalid, construction of an accurate house price index would require a housing unit sample much more representative than the FHA data base. The final advantage of the pairwise aggregation approach is that the information on price change in each period is represented by a single coefficient estimate. This makes possible a convenient statistical test of the significance of FHA sample truncation, as described below.

Divergences between the OLS and TR indexes will be taken as evidence of a non-ignorable sample design problem in the FHA data base. It is therefore desirable to have a means of testing whether the two indexes are significantly different in a statistical sense. Since for any given period the two alternative index changes are estimated from different sized regression samples, and since the difference between the two specifications does not amount to a constraint on a parameter vector, the improvement in performance obtained by accounting for truncation cannot be measured by a standard likelihood ratio test. However, recently Hausman [5] has presented a test of model specification which can be adapted to the present problem.

In both the OLS and TR models, the estimate of period-to-period price change is given by the "time dummy" coefficient. As applied here, the Hausman specification test is based on a comparison in each period of the two alternative coefficient estimates. The null hypothesis is that the house price function takes the form (1), the disturbance terms ε_i are normally distributed with constant variance,[8] and there is no sample design effect.

7. Griliches [4] criticizes the use of pairwise aggregation indexes in the automobile price application, on the basis that persistent "model effects" may distort the parameter estimates. This is less likely to be a problem in the housing market, where the samples are larger, the quality spectrum is more nearly continuous, and the products are not usually classified by make or model.

8. Poirier [13] demonstrates the use of Box-Cox [2] transformations to compare alternative functional forms in the situation where the dependent variable is truncated. This procedure was not followed here, primarily because the Box-Cox class excludes certain plausible functional forms, such as a linear model with disturbance variance proportional to the expected value of the dependent variable. Zarembka [22] reports that the Box-Cox procedure is not robust to heteroscedasticity of the error term.

These assumptions imply that OLS estimation will yield a consistent estimator b_1 of the time dummy coefficient. Further, b_1 is asymptotically normal and efficient, attaining the Cramer-Rao bound.

Under the above null hypothesis the Hausman-Wise TR coefficient estimator b_2 is also consistent, but its asymptotic variance is higher than that of b_1. Hausman shows that the difference between b_2 and b_1 has zero asymptotic covariance with b_1, and therefore must have an asymptotic variance equal to the difference between the asymptotic variances of the two estimators. Let δ_1 and δ_2 be the estimated standard errors of b_1 and b_2, and let ${\sigma_1}^2$ and ${\sigma_2}^2$ be the OLS and TR estimators of σ^2. It can then be shown that the test statistic m, defined by

$$m = (b_2 - b_1)^2/[{\delta_2}^2 - {\delta_1}^2({\sigma_2}^2/{\sigma_1}^2)] \tag{5}$$

is asymptotically distributed as chi-square with one degree of freedom.[9] When m is sufficiently high, the null hypothesis is rejected, and I conclude that in that period the OLS estimate of price change is distorted by probabilistic truncation of the FHA sample at high price levels.

III. Simulated Home Purchase Indexes

In this section I present and compare alternative house price indexes for three metropolitan areas: Minneapolis-St. Paul, Chicago-Northwestern Indiana, and San Francisco-Oakland. Ordinary least squares and truncated regression indexes are defined and simulated on a quarterly basis for a six-year period beginning with the fourth quarter of 1972 and ending with the fourth quarter of 1978.

Two primary considerations led to the choice of these areas as the subjects of intensive examination. First, the sample sizes were large enough to permit more reliable statistical analysis than was possible for many other cities. Second, the three areas chosen provided a wide range of average house prices, so that some information could be gathered regarding interarea variation in the importance of sample truncation effects. For example, in the CPI home purchase sample for June, 1977, the average sale price was \$28,453 in Chicago, \$35,583 in Minneapolis, and \$40,014 in San Francisco.

In order to increase the regression sample sizes and smooth the period-to-period index movements, the indexes were simulated on a quarterly basis. Records from the FHA Master File were assigned to quarters according to endorsement (closing) date. The resulting quarterly sample sizes ranged widely, due to the cyclical natures of both the construction industry and FHA's market share. Average sample sizes were 493, 462, and 353 in Minneapolis, Chicago, and San Francisco, respectively.

The hedonic indexes were each derived from 24 adjacent-quarter regressions, as justified in Section II above. The regressions were semi-logarithmic in form, with the dependent variable being the logarithm of sale price. Quarterly price change was estimated by the coefficient on a dummy variable indicating the second quarter of the

9. Multiplying δ_1^2 by the parenthesized ratio in (5) yields an alternative consistent estimator of the asymptotic variance of b_1. This adjustment should increase the power of the test, as discussed by Hausman [5] in other contexts.

Table III. Means of Hedonic Regression Variables

Variable	Area					
	Minneapolis-St. Paul		Chicago-Northwestern Indiana		San Francisco-Oakland	
	1972 IV	1978 IV	1972 IV	1978 IV	1972 IV	1978 IV
Logarithm of Sale Price	9.926	10.726	9.884	10.406	10.167	10.842
Living Area (Sq. Feet)	1045.747	1055.063	1134.086	1108.892	1295.738	1149.789
Age (Years)	32.924	37.407	26.232	26.993	5.139	18.531
Number of Rooms	5.228	5.320	5.559	5.547	5.951	5.319
Lot Size (× 1000 Sq. Feet)	8.961	9.296	5.334	7.100	4.848	5.280
Number Full Bathrooms	1.038	1.040	1.125	1.151	1.605	1.484
Number Half Bathrooms	0.127	0.103	0.204	0.187	0.355	0.131
Dummy for Central Air Conditioning	0.114	0.201	0.151	0.324	0.102	0.131
Dummy for Garage	0.835	0.884	0.675	0.691	0.741	0.845
Dummy for Fireplace	0.152	0.196	0.069	0.101	0.481	0.549
Dummy for Suburban or Rural Neighborhood	0.544	0.563	0.303	0.496	0.843	0.549
Dummy for Ramsey County	0.203	0.185	—	—	—	—
Dummy for Outlying Counties	0.278	0.212	—	—	—	—
Dummy for Suburban Illinois Counties	—	—	0.099	0.396	—	—
Dummy for Suburban Indiana Counties	—	—	0.215	0.446	—	—
Dummy for Alameda County	—	—	—	—	0.667	0.493
Dummy for Contra Costa County	—	—	—	—	0.287	0.498
Dummy for Non-Detached or Semi-Detached	—	—	0.071	0.122	0.336	0.103
Sample Size	79	378	465	139	324	213

pooled sample.[10] Table III displays the sample means of each regression variable for the first and last quarters of the study period. The explanatory variable sets are identical in the OLS and TR models and in the three metropolitan areas, with a few exceptions. In Minneapolis, all but a few FHA dwelling units were detached, so I excluded the dummy for semi-detached or row houses. Distinct county location variables were also defined for each area.

10. Given the logarithmic nature of the dependent variable, the anti-log of the coefficient on the time dummy is the estimated price relative for the two pooled quarters.

Table IV. Index Series, Minneapolis-St. Paul

Quarter			OLS Index	Truncated Regression Index	*m* Statistic
1	1972	IV	100.0	100.0	—
2	1973	I	103.4	103.3	.003
3		II	103.6	103.5	.000
4		III	106.5	106.4	.000
5		IV	113.6	113.6	.018
6	1974	I	116.3	116.6	.057
7		II	115.7	116.5	.179
8		III	123.8	125.8	1.945
9		IV	125.7	124.7	16.047
10	1975	I	132.2	131.3	.073
11		II	136.8	136.2	.344
12		III	142.7	142.4	.184
13		IV	143.5	143.4	.283
14	1976	I	145.2	144.9	.211
15		II	150.6	150.0	.202
16		III	152.1	152.3	2.398
17		IV	157.1	158.1	2.418
18	1977	I	158.8	160.3	.649
19		II	167.0	169.1	.726
20		III	173.3	176.5	3.343
21		IV	182.1	186.5	1.777
22	1978	I	190.0	190.5	41.305
23		II	203.2	204.2	1.856
24		III	218.1	219.8	2.654
25		IV	219.8	221.7	.419

During the 25 quarters of the study period the FHA mortgage ceiling was changed twice: from $33,000 to $45,000 in August 1974, and to $60,000 in November 1977. Using the reasoning of Section II, the edited samples used in estimating the TR indexes excluded all sales above $36,812.50 during the first eight quarters, all sales above $49,687.50 in quarters 9 through 21, and sales above $67,812.50 thereafter.[11] Below these price levels truncation was assumed to cause no estimation problems. The editing process reduced the TR sample sizes by approximately two percent in Minneapolis, four percent in San Francisco, and 0.2 percent (22 sales) in Chicago.

The regression coefficient estimates are of little direct interest here, and will not be discussed in any detail. In general, signs of the coefficients were consistent with

11. The sample was also truncated from below at $1,000 as part of the editing process. Both limits were explicitly recognized in the TR estimation program.

Table V. Index Series, Chicago-Northwestern Indiana

Quarter			OLS Index	Truncated Regression Index	m Statistic
1	1972	IV	100.0	100.0	—
2	1973	I	98.2	98.1	.151
3		II	100.3	100.2	.004
4		III	101.8	101.7	.000
5		IV	101.6	101.6	.001
6	1974	I	106.3	106.2	.044
7		II	107.3	107.3	.318
8		III	108.4	108.1	1.126
9		IV	112.1	111.4	.860
10	1975	I	114.3	113.7	.053
11		II	114.5	113.9	.007
12		III	114.3	113.6	.032
13		IV	115.5	114.9	.046
14	1976	I	116.8	116.3	.261
15		II	117.1	116.6	.039
16		III	120.6	120.1	.018
17		IV	122.0	121.6	.009
18	1977	I	123.8	123.4	.009
19		II	130.2	130.2	.810
20		III	131.7	131.8	.005
21		IV	133.4	133.4	.014
22	1978	I	141.0	138.7	4.393
23		II	145.5	143.2	.009
24		III	152.9	150.8	.306
25		IV	158.2	156.5	.176

expectations. The R^2 statistics in the OLS regressions were almost always in the .55 to .75 range. As might be expected, the TR estimates of the disturbance variance tended to be slightly higher than the OLS estimates.

Table IV presents the two alternative index series for Minneapolis-St. Paul. Although the estimates of total price change over six years are fairly close, the effect of the mortgage ceiling is reflected in short-term patterns of divergence and convergence. The TR index exceeds the OLS index by 2.0 points in quarter 8, the quarter of the first ceiling change, following three quarters of divergence. This small difference is reversed in quarter 9. Again, between quarters 15 and 21 the indexes drift apart slowly, then converge in period 22. By the final quarter there is some evidence of a third divergence beginning to appear.

These trends are consistent with the hypothesis that the truncation inherent in the FHA sample causes the OLS index to be biased downward during quarters prior to changes in the mortgage ceiling. Following each ceiling adjustment, the OLS index

Table VI. Index Series, San Francisco-Oakland

Quarter			OLS Index	Truncated Regression Index	*m* Statistic
1	1972	IV	100.0	100.0	—
2	1973	I	101.2	101.2	.016
3		II	98.5	98.6	.003
4		III	101.4	101.6	.124
5		IV	103.8	104.1	.003
6	1974	I	107.7	108.3	.232
7		II	110.1	110.6	.039
8		III	112.2	110.6	1.823
9		IV	123.2	121.0	.050
10	1975	I	127.9	126.2	1.934
11		II	127.3	125.5	.004
12		III	130.6	128.9	.064
13		IV	133.7	132.1	.258
14	1976	I	136.1	134.7	.216
15		II	140.4	138.9	.000
16		III	142.8	141.7	.489
17		IV	147.6	146.6	.089
18	1977	I	152.6	152.1	.449
19		II	158.5	159.6	1.613
20		III	165.4	166.7	.005
21		IV	177.2	179.4	.071
22	1978	I	199.6	199.1	1.185
23		II	205.2	206.4	.842
24		III	209.2	209.6	.172
25		IV	218.0	221.8	1.635

overestimates price change as the downward bias in the index level is temporarily eliminated.

The last column of Table IV gives the Hausman *m* statistics measuring the significance of the differences between the OLS and TR estimates of price change. As noted earlier, these values are each distributed asymptotically as chi-square with one degree of freedom under the null hypothesis that the OLS specification (with normally distributed errors) is correct. The null hypothesis is strongly rejected in quarters 9 and 22, the quarters immediately following ceiling changes. The *m* statistic also exceeds the 90 percent point of the chi-square distribution in quarter 20. In general, the higher *m* values are clustered around the periods of the ceiling adjustments.

Results for the other two metropolitan areas are displayed in Tables V and VI. The relative index movements are similar in form to those observed in Minneapolis, but the divergences are not as wide. In Chicago the largest differential between the OLS and TR

Table VII. Index Movements in Selected Subperiods

Index	Average Quarterly Percentage Increase					
	1972 IV to 1974 III	1974 III to 1975 I	1975 I to 1977 IV	1977 IV to 1978 II	1978 II to 1978 IV	1972 IV to 1978 IV
Minneapolis-St. Paul						
OLS	3.1	3.3	3.0	5.6	4.0	3.3
Truncated Regression	3.3	2.2	3.2	4.6	4.2	3.4
Chicago-Northwestern Indiana						
OLS	1.2	2.7	1.4	4.4	4.3	1.9
Truncated Regression	1.1	2.6	1.5	3.6	4.5	1.9
San Francisco-Oakland						
OLS	1.7	6.8	3.0	7.6	3.1	3.3
Truncated Regression	1.4	6.8	3.2	7.3	3.7	3.4

series is 2.3 index points in quarters 22 and 23. The values of the m statistic do not approach the usual critical levels except in quarter 22, when the TR index rises by a smaller amount following the ceiling change.

I had expected to find the strongest evidence of truncation problems in the San Francisco data, because of the high average price levels observed there.[12] However, no such evidence is apparent in Table VI. The TR index exceeds the OLS index by only 2.2 points in the quarter of the 1977 ceiling change, and lies below the OLS index at the time of the increase to $45,000 in quarter 8. Application of the Hausman specification test to the two sets of time dummy coefficients produces no rejections of the OLS model.

Table VII summarizes the movements of the indexes in each city over five subperiods. The subperiods are chosen to highlight the patterns of divergence between the OLS and TR series. During the first, third, and fifth subperiods, which precede ceiling changes, the truncated regression indexes generally rise more quickly than the OLS series. Following the ceiling adjustments, in subperiods 2 and 4, the OLS increases are greater. Again, this supports the notion that the ceilings bias the OLS indexes downward. When the ceilings are relaxed, the OLS series move sharply upward to approximately the correct levels. The TR series, which are adjusted for ceiling effects, exhibit somewhat smoother upward trends in Table VII. As the last column of the table shows, however, over the long run the alternative methods produce similar estimates of price change.

IV. Conclusions

The measurement of the cost of shelter for homeowners is an issue which has generated a great deal of controversy, particularly in recent years as the homeownership index has

12. In the context of equation (4), the correction for truncation in the Hausman-Wise likelihood is more important when $C_i - X_i\beta$ is small—i.e., when the mean price level is near the ceiling.

risen much faster than other CPI components. Several alternative measurement techniques have been proposed or advocated. Some of these would estimate homeownership costs by means of an appropriately weighted average of residential rent levels; other "user cost" or "outlays" indexes would, like the current CPI, depend in part on an index of home purchase prices.[13] As of this writing the FHA sample is the only sales data base which is sufficiently large, and available in a sufficiently timely manner, to serve as the basis for national and local home purchase indexes. Therefore, the validity of price measurements obtained from FHA data is of critical importance in the choice among homeownership index methodologies.

The purpose of this paper has been to determine the extent and severity of sample design effects in the FHA data base during the period 1972–78. Conclusions were mixed in the three cities studied. Sample truncation was found to be quantitatively most important in Minneapolis-St. Paul, where, for example, house price inflation between the fourth quarters of 1976 and 1977 was estimated as 18.0 percent by truncated regression and 15.9 percent by OLS. In Chicago, where house prices were relatively low, divergence between the two series was less noticeable. More surprisingly, no large or statistically significant difference was found in the high-priced San Francisco area.

These truncation effects observed during the sample period are less severe than might have been expected on theoretical grounds. Apparently, even in some PSU's where the mean house price has approached the level of the mortgage ceiling, enough high-priced homes have been included in the FHA sample to avoid large short-term bias in an hedonic index which ignores truncation. Further, in none of the three cities did any long-term divergence appear between the regression indexes. These two results suggest that, historically, FHA has acted fairly promptly to adjust its mortgage ceilings in response to market pressure, and these ceiling increases have been large enough to temporarily eliminate the sample truncation bias.

There remain several important topics for additional study. One issue concerns the extent to which the behavior of the OLS indexes results from the detailed regression specification used. The degree of quality adjustment in the hedonic method should have assisted in damping fluctuations caused by sample truncation. Research is needed to develop new and operationally feasible methods of quality adjustment in the CPI, as well as to evaluate current BLS methods for handling FHA sample design problems.

It should also be emphasized that although the city-level results differed in degree, the three samples produced similar patterns of downward bias, and subsequent upward correction, in the OLS index series. More serious truncation biases could occur in the future, depending on the course of FHA policy with respect to the mortgage ceiling. Combined with the other known weaknesses of the FHA data—erratic sample sizes, uneven coverage of local areas, low average prices—the evidence presented here supports further work to identify alternative data bases for use in the CPI. Finally, the sample truncation could have a more significant impact in other applications, such as cross-sectional price comparison or the econometric analysis of individual purchase behavior. Research should therefore continue on econometric modelling of the FHA sample design, particularly through the analysis of data sets containing both FHA and non-FHA sales.

13. The BLS now publishes on an experimental basis five monthly indexes of homeownership costs in addition to the official CPI series.

References

1. Berndt, Ernst, Bronwyn Hall, Robert Hall, and Jerry Hausman, "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurment,* Fall 1974, 653–65.

2. Box, George and David Cox, "An Analysis of Transformations." *Journal of the Royal Statistical Society,* Series B, 1964, 211–52.

3. de Leeuw, Frank, "The Demand for Housing: A Review of Cross-Section Evidence." *Review of Economics and Statistics,* February 1971, 1–10.

4. Griliches, Zvi. "Introduction: Hedonic Price Indexes Revisited," in *Price Indexes and Quality Change,* edited by Zvi Griliches. Cambridge: Harvard University Press, 1971, 3–15.

5. Hausman, Jerry, "Specification Tests in Econometrics." *Econometrica,* November 1978, 1251–71.

6. ——— and David Wise, "Social Experimentation, Truncated Distributions, and Efficient Estimation." *Econometrica,* May 1977, 919–38.

7. Heckman, James, "Sample Selection Bias as a Specification Error." *Econometrica,* January 1979, 153–61.

8. Lee, Lung-Fei, "Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables." *Econometrica,* July 1979, 977–96.

9. Mitchell, Daniel, "Does the CPI Exaggerate or Understate Inflation?" *Monthly Labor Review,* May 1980, 31–3.

10. Muellbauer, John, "Household Production Theory, Quality, and the 'Hedonic Technique.' " *American Economic Review,* December 1974, 977–94.

11. Nelson, Forrest, "Censored Regression Models with Unobserved, Stochastic Censoring Thresholds." *Journal of Econometrics,* November 1977, 309–27.

12. Palmquist, Raymond, "Alternative Techniques for Developing Real Estate Price Indexes." *Review of Economics and Statistics,* August 1980, 442–8.

13. Poirier, Dale, "The Use of the Box-Cox Transformation in Limited Dependent Variable Models." *Journal of the American Statistical Association,* June 1978, 284–7.

14. Polinsky, A. M. and David Ellwood, "An Empirical Reconciliation of Micro and Grouped Estimates of the Demand for Housing." *Review of Economics and Statistics,* May 1979, 199–205.

15. Pollak, Robert, "The Treatment of 'Quality' in the Cost of Living Index." U.S. Bureau of Labor Statistics Working Paper No. 90, June 1979.

16. Rosen, Harvey, "Estimating Inter-city Differences in the Price of Housing Services." *Urban Studies,* October 1978, 351–5.

17. Rosen, Sherwin, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy,* January-February 1974, 34–55.

18. Sirmans, C. F. and Arnold Redman, "Capital-Land Substitution and the Price Elasticity of Demand for Urban Residential Land." *Land Economics,* May 1979, 167–76.

19. Triplett, Jack, "Does the CPI Exaggerate or Understate Inflation? Some Observations." *Monthly Labor Review,* May 1980, 33–5.

20. U.S. Bureau of the Census, "New One-Family Houses Sold and For Sale, December 1979." *Construction Reports* C25-79-12, February 1980.

21. U.S. Department of Housing and Urban Development, *Characteristics of FHA Single-Family Mortgages: Selected Sections of National Housing Act, Calendar Year 1978,* March 1979.

22. Zarembka, Paul, "Transformations of Variables in Econometrics." in *Frontiers in Econometrics,* edited by Paul Zarembka. New York: Academic Press, 1974, 81–104.

23. Zerbst, Robert and William Brueggeman, "Adjusting Comparable Sales for FHA and VA Financing." *Appraisal Journal,* July 1979, 374–80.