JOURNAL OF HOUSING ECONOMICS 1, 110–126 (1991)

Arithmetic Repeat Sales Price Estimators*

ROBERT J. SHILLER

Cowles Foundation, Yale University

Received December 30, 1990

Repeat sales price estimators are designed to infer price indexes of infrequently sold and unstandardized assets, such as houses, based only on changes in prices of those individual assets that are observed to be sold twice. Repeat sales price estimators that are *arithmetic,* and either *value-weighted* or *equally weighted* are proposed here. Moreover, variants are proposed that are *interval-weighted,* i.e., that correct for a form of heteroskedasticity, and that include additional regressors representing *changes in hedonic variables.* Some of these methods are applied to data on house prices in Atlanta, Chicago, Dallas, and San Francisco 1970–1986.

Until now, repeat sales price estimators have provided estimates that are essentially equally weighted geometric averages of individual prices.[1] It is well known that the geometric average of any set of positive numbers not all equal is less than the arithmetic average of them.[2] Portfolios of assets (let us say, houses) have values that are related to arithmetic, not geometric, averages of prices. Moreover, the geometric averages of house prices do not give more weight to the more valuable houses; they are equally weighted indexes rather than value-weighted indexes. The weighting may make a difference to the estimated index if price changes in more valuable houses are different from price changes in less valuable houses.

* Presented at AREUEA session, Allied Social Science Association Meetings, Washington, DC, December 30, 1990. The author is indebted to Karl Case, John Clapp, Ray Fair, William Goetzman, Henry Pollakowski, William Schauman, Christopher Sims, and Allan Weiss for helpful discussions. This research was supported by the National Science Foundation under Grant SES-8921257.

[1] There has recently been a great deal of interest in developing better price indexes for housing and other nonstandardized assets. For example, repeat sales price estimators have been studied on connection with housing prices by Abraham and Schauman (1990), Case and Quigley (1991), Case *et al.* (1990), Case and Shiller (1987, 1989), Mark and Goldberg (1984), Palmquist (1982), and Pollakowski and Wachter (1990) and with art prices by Goetzman (1990a).

[2] The arithmetic average of a set of n numbers is their sum divided by n; the geometric average is the nth root of their product; see Ito (1987, p. 807) for a discussion of inequalities.

1051-1377/91 $3.00

Those who wish to study the covariances between housing prices and prices of other assets for the purpose of constructing a well-diversified portfolio would prefer to use an arithmetic index that represents the value of a portfolio of housing, and may prefer to see a value-weighted index that provides an index of the total value of real estate.

Goetzman (1990b) has proposed an estimated index of housing prices that is produced by first forming a geometric repeat sales index and then correcting this index by multiplying by a factor that depends on the cross-sectional variance of asset prices. The correction factor may be motivated eight by truncating a Taylor series expansion or by assuming cross-sectional variation in log housing price changes is lognormal.

While Goetzman's method appears to be serviceable, I propose here arithmetic repeat sales estimators that are simpler and more direct than his, and that do not rely on approximations or lognormality assumptions. There are several variations on the arithmetic repeat sales estimators: the value-weighted arithmetic repeat sales estimator (VW-ARS), the equally weighted arithmetic repeat sales estimator (EW-ARS), and the interval-weighted and hedonic-variable-augmented variations on these.

The different variants may serve different purposes. For example, the value-weighted arithmetic repeat sales estimator gives an index of the price of the aggregate stock of housing, an index of the value of an investment in a portfolio of all real estate, whose value is more influenced by the appreciation of the more valuable houses in the portfolio. The equally weighted repeat sales estimator gives an index of the value of a portfolio that is more concentrated in smaller houses, holding equal dollar amounts of houses in each value category.

The differences among these different estimators may often be small, if the cross-sectional variation of prices is not too large, as we shall see in some examples below. However, the differences are not negligible in our examples, and in principle the differences between geometric and arithmetic indexes could be enormous. If there were ever an observation of a price equal to zero for one house, the geometric index, related as it is to products rather than sums of numbers, would be zero for that period, while the value of a portfolio of houses might hardly be affected by the zero. If a single house is sold for one dollar (as sometimes happens) this would, unless sample size is very large, have a devastating impact on a geometric index, but not on an arithmetic index. The alternative indexes proposed here are no more difficult to calculate than the geometric; in fact they are more natural analogues to familiar indexes, such as stock price indexes, than are the geometric indexes. It is thus worth getting the index calculations right; practitioners should in most applications use one of the methods proposed here rather than the geometric estimation methods.

I assume throughout that there are n observations of repeat sales of

individual assets (let us say, houses), $2n$ sales in total. Each observation consists of the first sale price, the time period of the first sale, the second sale price of house i, and the time period of the second sale. I suppose that the time period is sufficiently long (let us say, monthly) so that there is at least one sale in each time period, where there are $T + 1$ periods in the sample, numbered from $t = 0$ to $t = T$.

I. THE SIMPLE GEOMETRIC REPEAT SALES (GRS) PRICE ESTIMATOR

I begin with a review of the geometric repeat sales estimators in use today. The Bailey–Muth–Nourse procedure, here called the geometric repeat sales or GRS procedure, estimates an index of log prices by regressing log price changes on a matrix of dummy variables. The matrix of independent variables is the $n \times T$ matrix Z whose ijth element is -1 if the first sale of house i occurred in period j, is 1 if the second sale of house i occurred in period j, and is zero otherwise.[3] The first column of Z corresponds to $t = 1$; there is no column for $t = 0$ since the estimated (log) index will be zero at $t = 0$ (the base year) by construction, so that its antilog will be one at $t = 0$. The dependent variable vector y has ith element equal to the change in log price for the ith house, using $p_{ij} = \ln(P_{ij})$, where P_{ij} is the price of the ith house at time j. The model to be estimated asserts that $y = Z\gamma + e$, where the ith element of γ is the log price index for time t, and for the purpose of computing standard errors it is assumed that the elements of the vector of error terms e are independent of each other, reflecting the notion that individual house price variations unrelated to the city-wide variations are due to idiosyncratic value changes. Then the estimated log price index for time t is the tth element of the ordinary least-squares regression coefficient vector $\hat{\gamma} = (Z'Z)^{-1}Z'y$.

If the change in log price of a house is given by the change in a true city-wide price index γ plus a zero-mean error term that is uncorrelated with the error terms associated other houses, and if the variance of this error term is the same for all houses, then the standard error matrix of $\hat{\gamma}$ has the usual form $s^2(Z'Z)^{-1}$. The assumption that the error term has zero mean implies that the true γ to be estimated is a geometric, not arithmetic, index. Moreover, the Gauss–Markov theorem applies and the estimator $\hat{\gamma}$

[3] The same GRS estimator can be written in another way, so that the estimate is a vector $\hat{\delta}$ of estimated changes in the log price index; it is produced by regressing the same vector y on a matrix Z_D whose ijth element is 1 if house i was between sales at time j; i.e., time j was after the first sale but not after the second sale. The vector of estimated coefficients is $S\hat{\gamma}$ where S is a $T \times T$ lower triangular matrix with ones along the main diagonal and -1 along the first off-diagonal.

is best linear unbiased. In practice, it is likely that the variance of the error term depends on the interval between sales, implying that a more efficient estimator is a weighted regression, to be discussed below. But one might still desire to use the simple GRS estimator if one does not accept this model, of if one values simplicity and ease of understandability. Other published price indexes are also simple indexes that do not involve weighting of observations.

For an example of the estimator, let us consider, for simplicity, an extremely small data set consisting of only five houses, and only three time periods, i.e., two index values to estimate. Suppose houses 1 and 2 were each sold in periods 1 and 2, houses 3 and 5 were each sold in periods 0 and 1, and house 4 was sold in periods 0 and 2; then we have

$$Z = \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} p_{12} - p_{11} \\ p_{22} - p_{21} \\ p_{31} - p_{30} \\ p_{42} - p_{40} \\ p_{51} - p_{50} \end{bmatrix}. \tag{1}$$

The normal equations $Z'Z\hat{\gamma} = Z'y$ for the GRS estimator are easily interpreted: the ith equation gives the estimated log index for the ith period as the average log price of all houses sold in that period minus the average of their base-period log price inferred from their other sale price using the estimated index. In this example the GRS normal equations are

$$\hat{\gamma}_1 = \frac{p_{11} + p_{21} + p_{31} + p_{51}}{4} - \frac{(p_{12} - \hat{\gamma}_2) + (p_{22} - \hat{\gamma}_2) + p_{30} + p_{50}}{4} \tag{2}$$

$$\hat{\gamma}_2 = \frac{p_{12} + p_{22} + p_{42}}{3} - \frac{(p_{11} - \hat{\gamma}_1) + (p_{21} - \hat{\gamma}_1) + p_{40}}{3}. \tag{3}$$

The first normal equation, the equation for $\hat{\gamma}_1$, the index for period 1, is based on the four houses sold in that period, two of which (houses 3 and 5) had their other sales in the base period, and two of which (houses 1 and 2) had their other sales in period 2, which had to be corrected by subtracting $\hat{\gamma}_2$ to infer a base-period price. The second normal equation, the equation for $\hat{\gamma}_2$, is an average of the log prices of the three houses that were sold in period 2 minus the average inferred log price of these three houses in period 0.

Note that the estimated log price index is based on averages of log price changes of individual houses, so that if we take $\exp(\hat{\gamma})$ as an index of the level of housing prices, then this index is based on *geometric* averages of individual house price relatives.

II. THE VALUE-WEIGHTED ARITHMETIC REPEAT SALES (VW-ARS) ESTIMATOR

An arithmetic estimator that is entirely analogous to the GRS[4] can be obtained by defining a matrix of independent variables X by X_{ij} equals minus the price of the first sale of house i if the time of the first sale was j, equals the price of the second sale of house i if the time of the second sale was j, and zero otherwise. The vector Y of observations on the dependent variable is given by Y_i equals the price of the first sale of house i if the first sale was in period 0, and is zero otherwise. Moreover, let us define a vector β whose ith element is a *reciprocal* price index for time i, equal to the estimated price at time zero divided by the estimated price at time i. By estimating reciprocal price indexes, rather than price indexes themselves, we have that the elements of $X\beta$ are all based on prices expressed in base-year units.[5] Here, the base period, as with the GRS index, will again be period 0, but now the index at time 0 is 1, not zero. In the example here, one may write the X and Y matrices as

$$X = \begin{bmatrix} -P_{11} & P_{12} \\ -P_{21} & P_{22} \\ P_{31} & 0 \\ 0 & P_{42} \\ P_{51} & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 \\ 0 \\ P_{30} \\ P_{40} \\ P_{50} \end{bmatrix}. \tag{4}$$

Note that this X matrix has zeros in the same places as did Z, replaces -1 with minus a price, and replaces $+1$ with a price.

Let us call the error term u_i the price of a house on its second sale date j times a true city-wide reciprocal price index β_j on that date minus the price of that house on its first sale date times the reciprocal of price index β_i on that sale date. The vector of error terms is then $u = Y - X\beta$. We will suppose, as before, that these error terms are uncorrelated across houses, reflecting individual house price variation alone. Now, one could con-

[4] The simpler expedient of replacing logs of prices with their levels in the GRS estimator would not be desirable, since we expect that it is more likely that the percentage change in house prices, rather than the absolute change, may be constant across houses of different values. The resulting estimator would not effectively control for the change in mix of houses through time; if larger or better houses are sold at one time than at another, the estimator would show a larger price increase, even if all houses appreciated at the same rate. The different expedient of replacing log differences with percentage changes in the GRS estimator runs afoul of compounding problems; 10-year growth percentages are in effect treated as simple sums of two consecutive 5-year growth percentages.

[5] Using the price index itself rather than its reciprocal suggests an ARS estimator where the inferred missing prices in the normal equation are given nonunitary weights.

ceive of taking an estimate of the vector β as an ordinary regression coefficient vector $(X'X)^{-1}X'Y$, but since there are stochastic independent variables, there is an errors in variables problem. Let us take instead an instrumental variables estimator $\hat{\beta} = (Z'X)^{-1}Z'Y$, where Z is as defined in the preceding section, or Eq. (1).[6] Under the assumptions that $p\lim(Z'u/n) = 0$ and $p\lim(Z'X/n)$ is nonsingular, this VW-ARS $\hat{\beta}$ will be a consistent estimator of β. That $p\lim(Z'u/n) = 0$ is a representation of the assumption that β is an arithmetic, not geometric, index.

The normal equations $Z'X\hat{\beta} = Z'Y$ have an interpretation analogous to those of the GRS estimator except that the estimator is here based on arithmetic instead of geometric averages. The ith diagonal element of $Z'X$ equals the sum of all prices of homes sold in period i. The ijth element, $i \neq j$, of $Z'X$ is minus the sum of all j-period prices of houses sold both in period i and in period j. Thus, the ith normal equation gives the index for period i as the mean price of all houses that were sold in period i divided by their mean price in the base period, where base-period prices of those houses not actually sold in the base period are inferred from their other prices using the estimated index. In our example, these VW-ARS normal equations, closely analogous to the GRS equations (2) and (3), are

$$\hat{\beta}_1^{-1} = \text{Index}_1 = \frac{P_{11} + P_{21} + P_{31} + P_{51}}{\hat{\beta}_2 P_{12} + \hat{\beta}_2 P_{22} + P_{30} + P_{50}} \tag{5}$$

$$\hat{\beta}_2^{-1} = \text{Index}_2 = \frac{P_{12} + P_{22} + P_{42}}{\hat{\beta}_1 P_{11} + \hat{\beta}_1 P_{21} + P_{40}}. \tag{6}$$

It follows from these normal equations that should there by any time period i in which all houses sold in that period were also sold in period 0, the index in that period is the same as a value-weighted arithmetic price index: it is (dividing both numerator and denominator of (5) and (6) by their respective number of elements in the summation) the ratio of the average price of these houses in period i to the average price of these houses in period 0.[7]

[6] There would be no effect on the estimates if the Z_D of footnote 3 were used in place of Z.

[7] Note that if we next constructed the corresponding equation for the base year $t = 0$, the mean of prices of all houses sold in period 0 divided by the mean of their other sale prices each deflated by the estimated index for the period of this other sale, then this base year index value equals one by construction. This would not generally be true if we used median or mode instead of mean as a measure of central tendency in the numerators and denominators of (5) and (6); this discrepancy in the base year indicates the kinds of conceptual problems one faces if one replaces means with these other measures of central tendency to try to derive median- or mode-based repeat sales price indexes.

The estimated index also has an interpretation in terms of the value of a portfolio consisting of all houses. The index in period i is an estimated value of the portfolio of all houses in time period i divided by an estimated value of the portfolio of all houses in period 0. The denominator of the index, the estimated value of the portfolio of all houses in period 0, is made using the price in period 0 of all houses, or, when period 0 prices are not observed, an inferred price using the index and the price observed closest to period 0. The numerator of the index for period i, the estimated value of the portfolio of all houses in period i, is made using for each house the price observed in period i, or, failing that, a price inferred using the index from the price observed closest after period i, or, failing that, a price inferred using the index and the price used in the denominator. To see that the estimator has this interpretation, it is helpful to use a transformation of the normal equations $Z'X = Z'Y$ written[8] as follows. Let us also, for illustrative purposes, suppose that there was a sixth house that was sold only once, in period 1; we will include it in the portfolio although it will not affect the estimated index. The VW-ARS normal Eqs. (5) and (6) are thus rewritten in the form

$$\hat{\beta}_1^{-1} = \text{Index}_1 = \frac{P_{11} + P_{21} + P_{31} + \hat{\beta}_2 P_{42}/\hat{\beta}_1 + P_{51} + P_{61}}{\hat{\beta}_1 P_{11} + \hat{\beta}_1 P_{21} + P_{30} + P_{40} + P_{50} + \hat{\beta}_1 P_{61}} \tag{5'}$$

$$\hat{\beta}_2^{-1} = \text{Index}_2 = \frac{P_{12} + P_{22} + P_{30}/\hat{\beta}_2 + P_{42} + P_{50}/\hat{\beta}_2 + \hat{\beta}_1 P_{61}/\hat{\beta}_2}{\hat{\beta}_1 P_{11} + \hat{\beta}_1 P_{21} + P_{30} + P_{40} + P_{50} + \hat{\beta}_1 P_{61}} \tag{6'}$$

Standard errors of the estimator should take account of the heteroskedasticity of the errors $u = Y - X\beta$, a heteroskedasticity potentially related to the rows of Z. For example, repeat sales where the interval between sales is a long one may show a lot of error, due to drift in the value of the individual home. An asymptotic standard error of the estimate $\hat{\beta}$ that takes account of this is given by[9]

$$\text{var}(\hat{\beta} - \beta) = (Z'X)^{-1} V (X'Z)^{-1}, \tag{7}$$

where $V = \sum_{i=1}^{n} Z_i' \hat{u}_i \hat{u}_i' Z_i$ and where $\hat{u} = Y - X\hat{\beta}$. If, as we suppose, houses with a longer interval between sales tend to have errors u_i with a larger squared value, then this will tend to affect $\text{var}(\hat{\beta} - \beta)$; it may be very different than if this heteroskedasticity were not accounted for. Standard errors for the growth rate of prices from t to $t + k$ can be inferred by a

[8] This interpretation uses a linear combination of the normal equations, or more simply, the normal equations $Z_D X = Z_D Y$, where Z_D is as defined in footnote 3.

[9] White (1984, Theorem 4.26, pp. 69 and 136).

linearization, giving squared standard error equal to $S_{t,t}/\hat{\beta}_t^2 + S_{t+k,t+k}(\hat{\beta}_t^2/\hat{\beta}_{t+k}^4) - 2S_{t,t+k}(\hat{\beta}_t/\hat{\beta}_{t+k}^3)$, where $S = \text{var}(\hat{\beta} - \beta)$.

III. THE EQUALLY WEIGHTED ARITHMETIC REPEAT SALES ESTIMATOR

The error terms $Y - X\beta$ are likely to have variance that depends on the price. More valuable houses have larger price movements. To obtain an equally weighted arithmetic repeat sales estimator, we may divide each row of the matrix X and Y by the price of the first sale corresponding to that row, thereby converting the error term $Y_i - X_i\beta$ from a levels error to a proportional error, and weighting each asset, each house, the same. An advantage to the equally weighted index is that the estimated index may be more efficiently estimated since the estimation procedure in effect takes account of the greater variance in the error terms in homes that have a higher initial price.

If we assume that the error term from the value-weighted arithmetic repeat sales estimator is also independent of the first price, then the probability limit of the EW-ARS estimator is the same as that of the VW-ARS estimator. However, one may not wish to assume this. If, let us say, more valuable houses are appreciating more slowly than the less valuable houses, then a VW-ARS estimator may tend to show lower price growth through time than the EW-ARS estimator, and for good reason.

The equally weighted repeat sales index has a portfolio interpretation just as does the value-weighted index. The index in period i is an estimated value of a different portfolio of houses in time period i divided by an estimated value of this portfolio in period 0. In this case, the portfolio invests in a share in all houses such that each share is worth \$1 when the house is sold first. That the index has this form can be readily seen by dividing all terms in the numerator and denominator of (5′) and (6′) by the first price observed for that house. Note that if prices are generally rising, houses that were not sold until late in the sample are given less weight in the portfolio. Note also that one would not generally have had information in period 0 to invest in such a portfolio, since the amount to be invested in each house not actually sold in period 0 is not observed yet; in practice one might often approximate this portfolio by investing relatively heavily in smaller houses.

An alternative to this EW-ARS estimator is one that makes the index the estimated value of a portfolio of houses that had equal dollar-value investments in each house in the *base* period. The normal equations for this base period equally weighted arithmetic repeat sales price estimator are derived by dividing, for all j, the terms corresponding to house j in the

numerator and denominator of the normal equations (as exemplified here by (5) and (6) or (5′) and (6′)) by $\beta_i P_{ji}$, where i is the date of the first sale of this house. Unfortunately, one must generally use iterative methods to solve the resulting normal equations which are usually nonlinear in the parameters, and the equations do not have a simple instrumental variables interpretation. One might also, if data are available, use a different equally weighted estimator derived by dividing through the ith row of X and Y by some objective measure of the size (e.g., square footage) of the house corresponding to that row, thereby producing a physically equally weighted arithmetic repeat sales price index, which represents the value of a portfolio that invests in the same physical amount of each house.

IV. APPLICATION TO REPEAT SALES DATA FOR FOUR CITIES

The figures show estimated VW-ARS and EW-ARS estimators along with simple GRS estimators for four cities, Atlanta, Chicago, Dallas, and San Francisco, quarterly data 1970—1 to 1986—2 (1986—3 for San Francisco). These are the same data used by Case and Shiller (1987, 1989); there are 8945 repeat sales pairs in Atlanta over this sample; the corresponding number for Chicago was 15,530, for Dallas 6669, and for San Francisco 8066.

Despite the fact that ARS estimators are, under assumptions noted above, estimates of arithmetic averages, which are always greater than the geometric averages estimated by the GRS estimator, the ARS estimates are not always greater than the corresponding GRS estimates in the sample. The value-weighted ARS estimators in this sample are often less than the GRS estimator, and show no strong upward bias relative to the latter. It was noted above that since the VW-ARS estimator is value-weighted and the GRS estimator equally weighted, a slower growth rate for the VW-ARS estimator may reflect a slower growth path for houses in the high price range. The EW-ARS estimator is more consistently greater than the GRS estimate than is the VW-ARS estimator in these data. At the end of the sample, the EW-ARS estimator for Atlanta was 4.6% higher than the GRS estimator at the end of the sample; for Chicago the corresponding figure was 4.4%, for Dallas 7.3%, and for San Francisco 2.0%.

Since the VW-ARS estimator does not downweight observations corresponding to very expensive houses, there is some concern that it may be more influenced by an occasional sale of a very expensive house. This tendency is probably a disadvantage of the VW-ARS estimator, related to the fact that the estimator takes no account of heteroskedasticity. The importance of this disadvantage could be reduced by obtaining a larger sample, or by following a rule of tossing out all houses that are at the

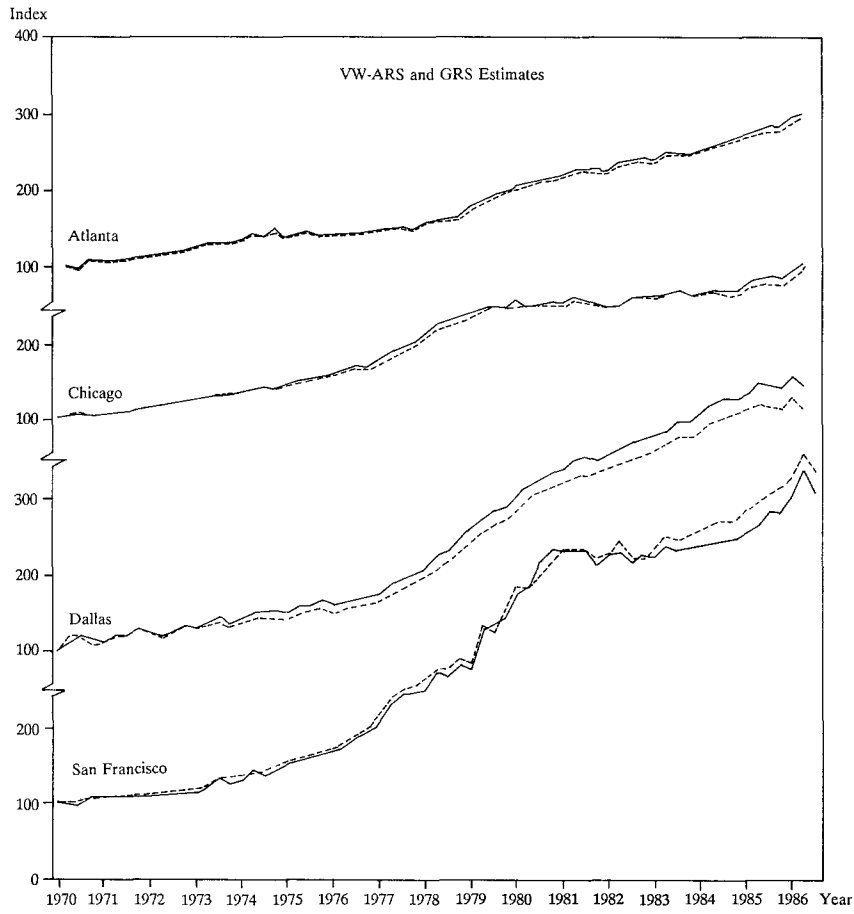

FIG. 1. Value-weighted repeat sales price index VW-ARS (solid line) and geometric repeat sales index GRS (dashed line). The VW-ARS index shown is $100/\beta_i$ plotted against i, where β_i is the ith element of coefficient vector estimated using the VW-ARS procedure. The geometric index shown is $100 \exp(\gamma_i)$ plotted against i, where γ_i is the ith element of coefficient vector estimated using GRS procedure. Data for Atlanta, Chicago, and Dallas are quarterly 1970—1 to 1986—2; data for San Francisco are quarterly for 1970—1 to 1986—3

extreme high limit of the range of house values. However, even without doing this, there is only a little suggestion in the figures presented here that the VW-ARS estimator is noisier than the EW-ARS estimator. This disadvantage of VW-ARS estimator does not appear to be very damaging here; one may wish to live with it in order to have a value-weighted estimator, which produces an index of the total value of all houses. The value-weighted index may be regarded as more representative of the direction of housing value; the equally weighted index could possibly be

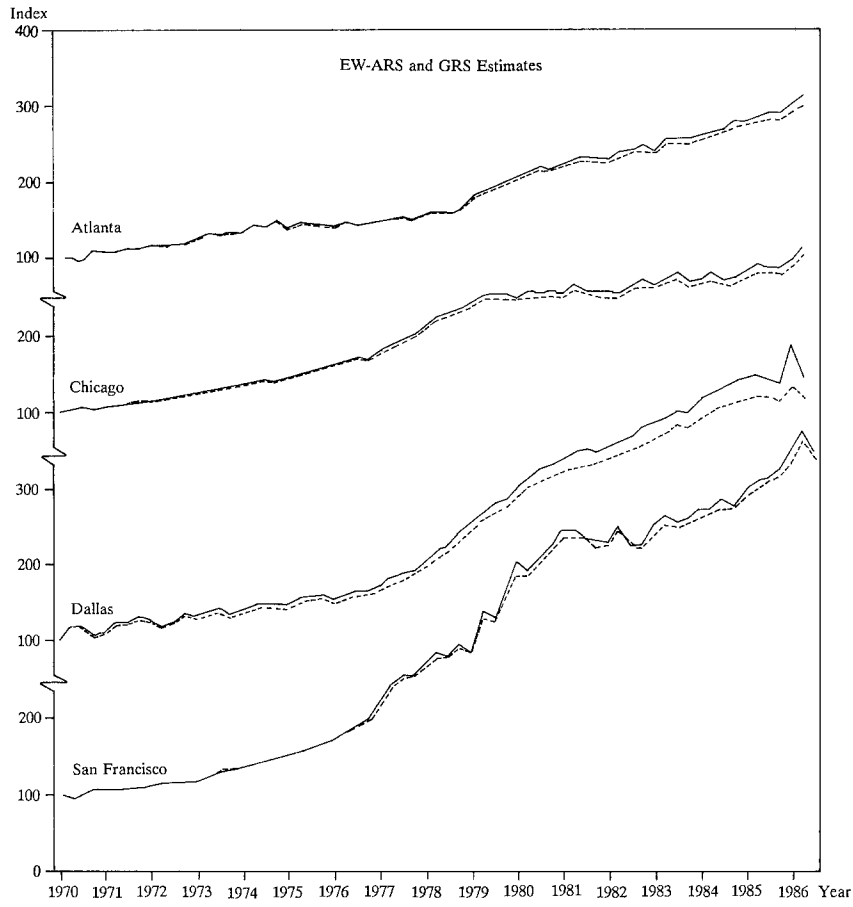

FIG. 2. Equally weighted repeat sales price index EW-ARS (solid line) and geometric repeat sales index GRS (dashed line). The EW-ARS index shown is $100/\beta_i$ plotted against i, where β_i is the ith element of coefficient vector estimated using the EW-ARS procedure. The geometric index shown is $100 \exp(\gamma_i)$ plotted against i, where γ_i is the ith element of coefficient vector estimated using GRS procedure. Data for Atlanta, Chicago, and Dallas are quarterly 1970—1 to 1986—2; data for San Francisco are quarterly for 1970—1 to 1986—3

unduly influenced by the effects of transactions of very small homes, homes that may contribute negligibly to overall housing value.

V. INTERVAL-WEIGHTED VERSIONS OF THE ABOVE ESTIMATORS

It was documented in Case and Shiller (1987) that when homes have a longer interval between sales, the squared error term in the GRS regression tends to be larger; we proposed a model of the error term in the GRS

regression in which the expected square error term is given by a constant plus a term proportional to the interval between sales. This then suggests a generalized least-squares estimator that should be more efficient than the ordinary least-squares estimator offered by Bailey, Muth, and Nourse. Analogous estimators were proposed by Webb (1988) and Goetzman (1990b). This same model can be used here to downweight the rows of X and Y in the GRS, the VW-ARS, and the EW-ARS that correspond to longer intervals between sales. These interval-weighed estimators will be referred to here as the I-GRS, IVW-ARS, and I-ARS estimators, respectively.[10]

Before defining these, it is important first to consider the issue of multiple sales of the same house; most earlier treatments of the repeat sales estimators assumed that no single house was observed sold more than twice.[11] The crudest way to handle the multiple sales problem is to represent the data on each such house as a number of pairs of sales, applying the above procedures and disregarding the fact that these are on the same house. When this is done, some arbitrariness is introduced into the estimator, since there is more than one way to decompose the multiple sales into pairs of sales. For example, a house sold in periods 1, 2, and 3 could be considered as repeat sales in 1 and 2 and 2 and 3, or as repeat sales in 1 and 2 and 1 and 3. The choice made affects the estimated coefficient vector. It *should* generally affect the estimated coefficient vector, since the assumption that regression errors are uncorrelated with each other cannot be compatible with both decompositions. The assumption that u_i is uncorrelated with u_j for all i and j is actually no longer plausible in either decomposition, since the same price appears in two different rows of the y vector. If multiple sales of the same house arise frequently, then a generalized least-squares estimator that takes account of the correlation across error terms should ideally be used.

As in Case and Shiller (1987), we assume that the natural log of the price of house i at time t is given by

$$\ln(P_{it}) = \text{constant}_i + C_t + H_{it} + N_{it}, \tag{8}$$

where constant_i is a house-specific constant term, reflecting such things as the size of the house, C_t is a city-wide price factor, H_{it} is a random walk (where ΔH_{it} has zero mean and variance σ_H^2) that is uncorrelated with C_T for all T, and N_{it} is noise term (which has zero mean and variance σ_N^2) and is uncorrelated with C_T and H_{jT} for all j and T and with N_{jT} unless $i = j$ and

[10] Case and Shiller (1987) referred to the I-GRS estimator as the weighted repeat sales (WRS) estimator. The new name for the estimator is introduced here to distinguish it more accurately from the others.

[11] But see Palmquist (1982).

$t = T$. Here, H_{it} represents the drift in house value through time (say, through changes in tastes or population distribution) and N_{it} represents noise at the time of sale (due, say, to random arrival of interested buyers or to errors in judgment). Then, the error term in the GRS estimator has variance equal to $2\sigma_N^2 + \sigma_H^2 \times$ (interval between sales). Case and Shiller (1987) estimated the parameters of this model for the four cities Atlanta, Chicago, Dallas, and San Francisco for 1970 to 1987, and the average (over the four cities) estimate of $2\sigma_N^2$ was 0.0084 and of σ_H^2 was 0.0011. This means that the standard deviation σ_N of the noise in price associated with the time of sale is about 6.5% of the value of the house, and the standard deviation σ_H of the quarter-to-quarter change in value of a house is 3.3% of the value of the house. A simple arithmetic average of price relatives (the price relative defined as the house price in the last quarter divided by house price in the first quarter of the sample, assuming these were observed) ought to be, assuming this model and lognormally distributed prices, at the end of a sample 66 quarters long, higher by a factor of $\exp(0.5\sigma_H^2 * 66)$ than the corresponding geometric average;[12] using the above estimate of σ_H^2 this suggests that the EW-ARS estimated index should be 3.7% higher than the GRS estimated index at the end of the 66-quarter sample; the estimated figures reported at the end of the preceding section are on average a little higher than that. The actual discrepancies between the EW-ARS and the GRS estimators may not follow this simple ratio rule for various reasons; for example, price relatives may not be lognormally distributed; actual price relatives may have "fat tails."

If we assume that multiple repeat sales are grouped together in our listing of repeat sales and arranged as consecutive pairs of repeat sales (so that there is no overlap in intervals between sales for a given house; a house sold in periods 1, 2, and 3 is considered as a repeat sale in periods 1 and 2 and a repeat sale in periods 2 and 3), then the covariance between consecutive repeat sales of the same house is $-\sigma_N^2$. The variance matrix Ω of the n-element vector of error terms is then block diagonal, with blocks corresponding to individual houses; each block is tridiagonal. Hence, since the size of the blocks is likely to be very small relative to the dimension of Ω, Ω is easily inverted. We can then use a generalized least-squares estimate of γ, called I-GRS, $\hat{\gamma} = (Z'\Omega^{-1}Z)^{-1}Z'\Omega^{-1}y$. [This estimator collapses to the WRS estimator of Case and Shiller (1987) if no house is sold more than twice.][13]

[12] Goetzman (1990) has, as noted above, proposed correcting the GRS estimator by multiplying by such a factor.

[13] Clapp and Giaccotto (1990) and Goetzman (1990b) find that in their applications the differences between this estimator and the GRS estimator $(Z'Z)^{-1}Z'y$ were small, and so the simpler GRS procedure (or, by extension, the simple VW-ARS or EW-ARS estimators) may

The estimators IVW-ARS and I-ARS are defined analogously. Using the same block-diagonal tridiagonal Ω, these estimators are $\hat{\beta} = (Z'\Omega^{-1}X)^{-1}Z'\Omega^{-1}Y$. Under general assumptions these GLS-like instrumental variables estimators are asymptotically efficient in the sense defined by White (1984, Theorem 4.57). In practice, there are two unknown elements of the matrix Ω that must be estimated: the dependence of the diagonal elements on the interval between sales, and the value of the off-diagonal element.

VI. COMBINING REPEAT SALES WITH HEDONIC ESTIMATORS

Methods of combining repeat sales estimators with hedonic estimators have been proposed by Case and Quigley (1989), Case *et al.* (1990), and Clapp and Giaccotto (1990). Some of these methods entailed using information about *changed* characteristics to improve the efficiency of repeat sales estimators. We may follow such methods in combination with arithmetic repeat sales estimators.

It is useful first to note that the GRS estimator can be derived as a sort of special case of a hedonic estimator where hedonic variables consist only of house dummy variables, one for each house; the ith element of the jth dummy variable is 1 if the ith observation is on the jth house, and is zero otherwise.[14] With the hedonic regression, all houses may be included, even those sold only once, although if there are house dummies in the regression those houses sold only once will have no effect on the estimated price index. The ith element of the dependent variable vector is the log price of the ith house sale; the matrix of independent variables consists of period dummy variables, one for each period in the sample, and the house dummy variables. There is, however, multicollinearity among the columns of this matrix of independent variables, so one must drop one column of the matrix of independent variables; let us drop the time dummy corresponding to the 0th time period. It might not be advisable to estimate the coefficient vector by ordinary least squares, since the error terms for any one house are likely to be correlated; the GRS procedure will turn out to be the same as a generalized least-squares estimate of this hedonic regression that takes account of this correlation. If we transform the vector of observations on the dependent variable and the matrix of observations on the independent variables by premultiplying both by a

suffice. If one takes this simpler route, there is no reason to give any special treatment to multiple sales of the same house, so long as sales pairs are chosen so that intervals between sales in the pairs for a given house do not overlap in time.

[14] See also Palmquist (1982).

nonsingular matrix S that replaces all but one of the rows for each house by consecutive differences of rows, and leaving one of the level observations for each house, then the ordinary least-squares estimate (which may be regarded as a generalized least-squares estimate of the model that takes into account the correlations structure of the errors) of the coefficient vector returns for us a coefficient vector consisting of the GRS estimator and the coefficients of the house dummies. Whenever there is a dummy variable in a regression which is zero except for one element, then the effect of including that dummy in the regression is the same, in terms of the coefficients other than the coefficient of that dummy, as dropping the corresponding observation from the regression. The transformed matrix of independent variables includes dummy variables that eliminate the effect on estimated coefficients of all the single-sale and level observations.

Consideration of this hedonic regression suggests the possibility of using changed characteristics in a repeat sales regressions, as suggested earlier in Palmquist (1982), Case and Quigley (1989), and Case et al., (1990). Suppose that we augmented the set of regressors for the original hedonic regression discussed above by some other hedonic variables, e.g., log number of rooms in the house. Individual houses must show some change in these hedonic variables between sales; otherwise the hedonic variables will show strict multicollinearity with the house dummy variables. Then, premultiplying the vector of independent variables and matrix of independent variables by the same matrix S would leave us with the GRS regression estimator augmented by some additional independent variables, the ith observation of each such additional independent variable being the *change* between the corresponding pair of repeat sales of the additional hedonic variable. Thus, for example, if one additional hedonic variable, the log number of rooms, were added to the original hedonic regression, then this would amount to adding to the GRS regression an additional regressor which is zero for repeat sales for which rooms did not change between sales, and equals the change in the log number of new rooms for repeat sales for which number of rooms did change.

One might then consider adding as additional regressors (and as additional instruments) to any of the estimators considered in this paper additional variables representing such changes in hedonic variables between repeat sales. This might be a useful alternative to dropping from the sample all repeat sales for which there is evidence of change in the house between sales. When there is a good deal of evidence about housing characteristics, we might find that most houses change between sales. We might not want to drop all such repeat sales observations. Including all repeat sales along with the additional regressors carrying information about the changes retains the desirable characteristic of repeat sales esti-

mators that there is no effect of the estimator of changes through time in the representation of individual houses, and at the same time takes account of some observed changes in houses.

It is worth noting, finally, that the above suggests another idea for ordinary hedonic regressions, without the house dummies that would convert them into repeat sales estimators. To obtain a value-weighted arithmetic hedonic regression estimator, we first form X and Y matrices with one observation per house. The vector Y of dependent variables has all of its observations equal to 1.00. The jth of the first $T + 1$ columns of the X matrix has ith element equal to zero unless the house corresponding to observation i was sold in period $j - 1$, in which case the element is the price of that house. Hedonic regressors can be appended as additional columns of the X matrix. The estimator is then $(Z'X)^{-1}Z'Y$, where the matrix Z of instruments is the X matrix where each price in the first $T + 1$ columns is replaced with the number 1. The first $T + 1$ elements of the vector of estimated coefficients are the reciprocal price index estimates. If there are no hedonic regressors, this estimator returns as a price index for each period just the average price of a house in that period, just as if one had simply regressed house price on period dummies. When hedonic regressors are added, however, the estimators change from the usual hedonic regression estimators. For example, if there is a single hedonic regressor equal to the number of square feet in the house, then an ordinary hedonic regression setup where the dependent variable is the price (not log price) in effect presumes that the number of square feet has the same linear effect on price in time periods when housing prices are low as in time periods when housing prices are high; the arithmetic hedonic price estimator proposed here would not. As above, we can also form an equally weighted hedonic regression estimator by dividing each row of X and Y by the price of the house corresponding to that row.

REFERENCES

ABRAHAM, J. M., AND SHAUMAN, W. S. (1990). "New Evidence on Home Prices from Freddie Mac Repeat Sales," reproduced, Federal Home Loan Mortgage Corp.

BAILEY, M. J., MUTH, R. F., AND NOURSE, H. O. (1963). "A Regression Method for Real Estate Price Index Construction," *J. Amer. Statist. Assoc.* **58,** 933–942.

CASE, B., POLLAKOWSKI, H. O., QUIGLEY, J. M., AND WACHTER, S. M. (1990). "On Choosing among House Price Index Methodologies," unpublished paper, Harvard University.

CASE, B., AND QUIGLEY, J. M. (1991). "The Dynamics of Real Estate Prices," *Rev. Econ. Statist.*, forthcoming.

CASE, K. E., AND SHILLER, R. J. (1987). "Prices of Single Family Homes Since 1970: New Indexes for Four Cities," *New Eng. Econ. Rev.* 45–56.

CASE, K. E., AND SHILLER, R. J. (1989). "The Efficiency of the Market for Single-Family Homes," *Amer. Econ. Rev.* **79,** 125–137.

CLAPP, J. M., AND GIACCOTTO, C. (1990). "Estimating Price Trends for Residential Property: A Comparison of Repeat Sales and Assessed Value Methods," unpublished paper, University of Connecticut.

GOETZMAN, W. N. (1990a). "Accounting for Taste: An Analysis of Art Returns Over Three Centuries," reproduced, Columbia University.

GOETZMAN, W. N. (1990b). "The Accuracy of Real Estate Indices: Repeat Sale Estimators," reproduced, Columbia University.

ITO, K. (1987). *Encyclopedic Dictionary of Mathematics,* Vol. II, MIT Press, Cambridge, MA.

MARK, J. H., AND GOLDBERG, M. A. (1984). "Alternative Housing Price Indices: An Evaluation," *AREUEA J.* **12,** 30–49.

PALMQUIST, R. B. (1982). "Measuring Environmental Effects on Property Values without Hedonic Regressions," *J. Urban Econ.* **11,** 333–347.

POLLAKOWSKI, H. O., AND WACHTER, S. M. (1990). "Effects of Land Use Constraints on Housing Prices," *Land Econ.* **66,** 315–324.

WEBB, C. (1988). "A Probabilistic Model for Price Levels in Discontinuous Markets," in *Measurement in Economics* (W. Eichhorn, Ed.). Physica-Verlag, Heidelberg.

WHITE, H. (1984). *Asymptotic Theory for Econometricians.* Academic Press, Orlando.