BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION

SRD Research Report Number: Census/SRD/RR-88/25

CONSTRUCTION OF A HOUSING INDEX

by

Beverley Causey
Statistical Research Division
Bureau of the Census
Room 3134, F.O.B. #4
Washington, D.C. 20233 U.S.A.

Recommended: Larry Ernst

Report completed: November 21, 1988

Report issued: November 21, 1988

Construction of a Housing Index

This paper describes methodology for a study done specifically for HUD. The principles, however, can be considered in other settings.

0. Introduction. We want to construct a "constant quality" housing index. Suppose that we have a (sample from a) set of housing units at time A. We have a measure of housing quality: for us, separately, rent per month for rental units and price for condos; we do two separate analyses for rents and condos. In addition to quality we observe a set of characteristics for each unit: floor space, availability of air conditioning, etc. We look at the relationship between quality and characteristics.

Suppose we consider (a sample from) the same units at a later point in time, B. Once again we consider the relationship between quality and characteristics. We compare (a) the (actual) overall quality of the time-A units, against (b) the overall quality for the time-A units that would be obtained if we had (1) the characteristics of the time-A units combined with (2) the relationship between quality and characteristics that has been formulated for the time-B units. The ratio of (b) to (a) is the so-called constant-quality index. (I am not sure this terminology is appropriate; but it is just terminology.) As in reports of the Dept. of Commerce on the "price index of new one-family houses sold," we want to compute this ratio for a base year "A" and a succession of future years "B."

We will compute the ratio for the entire U.S.A. and also, with limitations because of scantiness of data, separately for each of 4 regions of the U.S.

Sections are as follows:

(1) A description of the data to be used.

(2) Methodology for a single region and a single year. For this relatively simple situation we explain the needed principles.

(3) (Results based on) pooling of regions and years.

(4) A conclusion as to what we recommend and suggest: a use of pooled results, in conjunction with the log of our measure of quality (that is, rent or condo price).

(5) Some additional results associated with our multiple-regression fit.

1. Data comes from the Survey of Market Absorption, supplemented with inputs from the Survey of Construction. For rental units the base year A is 1977, for condos 1983. The future years, B, are each of 1978-86 and 1984-86. Details of our characteristics, and of how we choose our predictor variables, appear in Section 4.

Originally, a single data "record" is a "building," i.e., a single development of units. There are 92,159 such buildings in all to start with in our data base, but we have excluded a few of these because of incomplete information. Within each building, units may be grouped according to number of bedrooms/bathrooms. For us the "group" within the building is the observation of interest. All of our characteristics (availablity of air conditioning, whether there is an elevator, etc.) are the same for all units within a group and in fact, except for number of bedrooms/bathrooms and area, the same for all units within a building. Likewise rent and price are the same within a group.

Each building has an associated weight which corresponds to the reciprocal of probability of inclusion in sample. We allocate this weight to each of our groups in proportion to the number of units in the group.

2. A single region. We start with the fit of a relationship between quality and characteristics at a particular point in time, for a single region.

Subsections are as follows:

(2.1) Definition of independent variables corresponding to our characteristics.

(2.2) The use of group weights, as described in the introduction, in weighted least squares.

(2.3) Our fit of nonnegative coefficients.

(2.4) Formation of the constant-quality ratio of the introduction.

(2.5) Use of logarithms of quality measures.

(2.6) A minor annoyance: the case where an independent variable's value is always zero.

2.1. Independent Variables. We need the following notation. Let j be the index for our observations: groups in the sense of the introduction. Let y_j be the measure of quality (rent or price) and w_j be the weight associated with group j. Let i be the index for our characteristics, and let x_{ij} be the value of characteristic i for unit j.

We now define, along with dependent variable Y, a set of independent variables X_i: that is, the numerical values for x_{ij}.

(1-2) Square Feet: the area of a unit (the same for all units in a group) expressed in square feet. Preliminary investigation of a linear relationship

$$\log Y = A + B \log Z$$

with Y=rent or condo price and Z=area in square feet, yielded B=.40 and .46 for rent and condo respectively. On this basis we surmised that Y itself might be more linearly related to $Z^{.5}$, the square root, than to Z itself. As independent (predictor) variables we thus include the characteristics $X_1=Z^{.5}$ and $X_2=Z$ itself, for both rents and condos.

(3-4) Number of bathrooms. Units have, as reported, 1, 1 1/2 or 2 or more bathrooms. We let the dummy variable X_3 be equal to -1 if (each unit in) a group has 1 bathroom, and equal to 0 otherwise. We let the dummy variable X_4 be equal to +1 if there are 2 or more bathrooms, equal to 0 otherwise. Thus 1 1/2 bedrooms is the "excluded middle." Note that we have 2 separate dummy variables here, rather than a single one which takes on values -1, 0, +1. These and our other dummy variables are deliberately chosen so as to be (apparently) positively correlated with y, that is, so that β would be positive if each of them were considered singly.

Choice of dummy variables is based also on (a) our computer frequency of the various categories (e.g., if there were essentially no units with exactly 1 1/2 bathrooms, we would just lump those very few units with either "1" or "2 or more" and create a single dummy variable) and (b) differences in "y" averages (e.g., if rents were essentially the same for units with 1 1/2 and for units with 2 or more, then there would be no point in making a distinction between these two types of units.)

(5-6) Number of bedrooms. Here we work as for bathrooms with the 3-way distinction now between 1, 2, and 3-or-more bedrooms.

(7) Heat. We lump "no answer" and "electric" (with or without heat pump) to correspond to $X_7=0$. A value $X_7=1$ corresponds to other types of heat including gas, oil, and solar.

(8) Air conditioning. A clear pattern is not evident in the data, but we lump "no answer" and "none" as ($X_8=$)0, and let 1 correspond to "yes."

(9) Elevator in building. We let 0 denote no answer or none, and 1 denotes yes.

(10-11) We let X_{10} equal -1 if the group of units is outside a metropolitan area, 0 otherwise. We let $X_{11}=+1$ if the units are in a central city (inside a metropolitan area) and =0 otherwise.

(12-13) Number of floors in building. We work as for bedrooms with the distinction again 1, 2, and 3-or-more.

(14-23) All the above characteristics are for both rents and condos. For rents only we have 10 additional items, for each of which we have lumped "not available" and "no answer" (value 1). On all 10 items the data suggests this lumping, except for A/C where as in (8) the pattern is fuzzy. These items are: (14) A/C, (15) swimming pool, (16) electricity, (17) gas, (18) heat, (19) hot water, (20) range for cooking, (21) refrigerator, (22) dishwasher, (23) parking place.

2.2 Weights. We do a multiple-regression fit of form

$$\hat{Y}=\alpha+\sum\beta_i X_i$$

with α and the quantities β chosen so as to minumize the sum

$$\sum_j w_j(y_j-\alpha-\sum_i\beta_i x_{ij})^2.$$

That is, we perform a customary weighted-least-square fit.

2.3 Nonnegativity. We constrain the coefficients β_i to be nonnegative,

because this makes sense for our characteristics as defined in the above. Thus we do a constrained least squares fit. As a result, much of the theory for residual variances, confidence intervals, etc., is, strictly speaking, not 100% properly applicable. We do address such issues in Section 5; but we will first obtain the quantities α and β_i as above.

Typically, many β's will be 0, with predictor variables thus disappearing from the fitted equation: Our characteristics viewed singly will as a rule yield positive β's but when put together will no longer all do so because of positive correlations among them. In at least one instance, a negative β_i may be obtained even for a characteristic viewed singly. Within the West, a single region, rents have in other studies been found higher for units without air conditioning than with it. It is believed that subregional differences in rent levels, rather than A/C as such, lend to this result. That is, an unexpectedly negative relationship arises, in effect, because of a correlation with an explanatory variable that is not included in our set of predictor variables. In our own study we do not attempt to incorporate this possibility.

2.4. The Constant-Quality Ratio. Suppose that, for two different years A and B, we do the above fit. An easily interpreted formulation of the constant-quality index in the introduction is "R" given by

$$R=[\bar{y}(B)+\sum\beta_i(B)(\bar{x}_i(A)-\bar{x}_i(B))]/\bar{y}(A).$$

Here $\bar{y}$ and $\bar{x}_i$ are weighted means, and A and B in parentheses giving the year of computation. That is, $\bar{y}$ is $(\sum w_j y_j)/(\sum w_j)$ and $\bar{x}_i$ is likewise. Year A is the base year, and year B is a subsequent year.

Some computational short cuts may be taken using (a) as in Section 2.1 the dummy-variable dichotomous structure of most of our x's, and (b) the decomposition of needed cross products and sums of squares into within-region and among-region components.

2.5 Logarithms. We also do an analysis which uses the variables of Section 2.1 except that we (a) replace Y, the quality measure, by log Y, and (b) re-define X_1 equal to log Z. (Somewhat superfluously, we still let X_2 be equal to Z itself.) Thus the fitted coefficients β_i, corresponding to our characteristics, now represent additive effects for log Y, and proportionate effects for Y itself. Formerly they represented additive effects for Y itself. In the next section this distinction will become important.

2.6 Zeros. Sometimes, for a given region and year, the values x_{ij} for given i will be the same for all (groups) j. In such instance we omit the variable X_i from our calculations; this omission does not change our results.

3. Pooling. In Section 2 we did a least-squares fit for each region-year separately. But we are dealing with relatively sparse amounts of data, and there is some volatility in the results. For example, for region 1 and rents the ratios R for each of the subsequent years 1978-86 (with base year 1977) are:

1.237, 1.323, 0.886, -3.280, 1.537, 1.815, 1.539, 5.085, 1.610.

We will resolve this issue by fitting a composite single set of coefficients β_i (although we still will fit α separately for each region-year).

Section 3.1 shows how we do this composite fit. Section 3.2 shows what we obtain when we use the logs of Section 2.5. For these computations we may obtain ratios R for each region considered separately. Yet we want, also, to form ratios for the entire U.S.A., i.e., for all regions combined. Section 3.3 considers the entire U.S.A.

3.1. The Full Sum of Squares. In this section we explicitly fit the composite set of coefficients. Along with the subscript i for variable and j for group, we have a subscript r for region and t (time) for year. We choose the coefficients β_i and α_{rt} (for each r-t combination) so as to minimize the sum of squares

$$\sum_{rtj}\sum\sum w_{rtj}(y_{rtj}-\alpha_{rt}-\sum_i \beta_i x_{rtij})^2$$

with, as in Section 2.3, the coefficients β_i constrained to be nonnegative.

3.2. Use of Logs. Once the coefficients α_{rt} and β_i are obtained from the last section, we can use them to compute the ratios R, and achieve much more stability than is displayed above from region 1. But if we do the pooling of the last section, the use of logs, as in Section 2.5, becomes particularly advantageous, for the following reason. As we discussed in Section 2.5, when we use logs, the fitted coefficients β_i represent proportionate rather than additive effects. Thus, for example, having an eleavator in the building might represent, on average, a 5% increase in rent per month if we use logs, but it might represent a $20 increase if we do not use logs. Over an extended period of years inflation will change the level of rents and prices and, similarly, the dollar value which an elevator (for the building) adds to the quality of an apartment. Hence we would work with proportionate rather than dollar values.

For the index R let us return to the constant-quality ratio of Section 2.4, for a single region with base year A and subsequent year B. Working with logs, we replace y_j by the quantity $u_j = \log y_j$, and form the weighted mean $\bar{u}=(\sum w_j u_j)/\sum w_j)$. We now obtain R given by

$$\log R = \bar{u}(B) - \bar{u}(A) - \sum\beta_i(\bar{x}_i(B)-\bar{x}_i(A)).$$

(The regression fits for year A and B differ only in the constant quantities α. This difference is equal to log R.)

We have obtained a composite set of values β_i; these are instructive in themselves. For rents and condos we give these values in Table 1. For the rest of this Subsection 3.2 we digress from our principal purpose, the construction of R, to discuss these values. The reader will need to refer to Section 2.1, where the variables X_i are defined.

Referring to Table 1 and rents, we have β_1=.061. With X_1 = log Z and Z=square feet, this suggests a relationship

log Y = constant + .061 x log Z

Table 1
Values of Beta (rounded)

Rents

(1)	.061	(2)	0	(3)	.073	(4)	.074	(5)	.105	(6)	.185	(7)	.045
(8)	0	(9)	.195	(10)	.119	(11)	.017	(12)	.038	(13)	.067	(14)	0
(15)	.115	(16)	0	(17)	0	(18)	.025	(19)	0	(20)	.066	(21)	.088
(22)	.099	(23)	.088										

Condos

(1)	.227	(2)	0	(3)	0	(4)	.102	(5)	.047	(6)	.250	(7)	.080
(8)	0	(9)	.283	(10)	.037	(11)	0	(12)	.093	(13)	.145		

whereas Section 2.1 suggested that .061 should instead be (B=).40. In Section 2.1 we considered log Z singly; here we consider it in conjunction with 22 other predictor variables. In the manner indicated in Section 2.3, the inclusion of these other predictors apparently crowds out much of the predictive value of X_1.

All the other β's correspond to dummy variables. To illustrate their interpretation, consider X_3 and X_4. With β_4=.074, the additive increase in log Y that results from having 2 or more bathrooms. as opposed to 1½ bathrooms, is .074. That is, Y itself gets multiplied by the factor exp(.074), about 1.077. With β_3=.073, the decrease in log Y that results from having only 1 bathroom is .073. Thus Y itself is divided by the factor exp(.073), about 1.076, to reflect the fact that there is only 1 bathroom.

3.3. Combining Regions. Up to now we have considered R for only a single region. To obtain R for the aggregate of 4 regions, again for the two years "A" and "B," we proceed as follows. For Section 2.5 we use for each year, weighted means (that is, $\bar{y}$ and $\bar{x}_i$ of Section 2.4) based on summing over all regions instead of over a single region. We continue to use the composite coefficients β_i of Section 3.1. For the logs of Section 3.2, with coefficients β_i again as in Section 3.1, we again use

weighted means based on all regions. In effect we form for each year a single α for the 4 combined regions.

From Section 3.2 we obtain for rents, with base year 1977, values for R as in Table 2. We have also included the quantities Q given by

$$\log Q = \bar{u}(B) - \bar{u}(A)$$

which we would obtain as R for β_i all 0, i.e., without taking into account the characteristics X_i. That is, Q is just a simple comparison of rent levels at two points in time. For example, for region 1 in 1978 we have Q=1.143 and R=1.147. The coefficients β_i do not have a strong inpact here, although they do for other region-years. For example, for all regions combined which we designate as "region 0," in 1986 we get Q =1.966 and R=1.469. (What do we say about interpretation of Q and R?)

Table 2
Constant-Quality Ratios
"Region 0" is 4 regions combined.

Rents

Reg.	1Q	1R	2Q	2R	3Q	3R	4Q	4R	0Q	0R
1978	1.143	1.147	1.081	1.084	1.140	1.123	1.122	1.089	1.105	1.089
1979	1.373	1.353	1.206	1.209	1.265	1.252	1.285	1.257	1.244	1.227
1980	1.195	0.989	1.312	1.004	1.382	1.063	1.363	1.000	1.323	1.006
1981	1.557	1.246	1.502	1.190	1.562	1.192	1.514	1.114	1.490	1.139
1982	1.685	1.319	1.645	1.280	1.850	1.423	1.714	1.294	1.722	1.314
1983	1.648	1.404	1.633	1.276	1.750	1.362	1.703	1.277	1.650	1.270
1984	1.551	1.323	1.543	1.174	1.817	1.413	1.742	1.252	1.706	1.290
1985	1.580	1.250	1.841	1.396	1.942	1.480	1.891	1.337	1.883	1.389
1986	2.099	1.585	1.807	1.451	1.984	1.531	2.037	1.434	1.966	1.469

Condos

Reg	1Q	1R	2Q	2R	3Q	3R	4Q	4R	0Q	0R
1984	0.902	0.961	1.072	1.073	1.019	1.021	1.067	1.076	1.029	1.038
1985	1.189	1.148	1.116	1.102	1.073	1.029	1.174	1.145	1.145	1.111
1986	1.404	1.333	1.172	1.240	1.152	1.099	1.434	1.370	1.365	1.320

4. Conclusions. With relatively small amounts of data for many region-years we have adopted the pooling to compute the quantities β_i and the use of logs, for reasons discussed above. In future computations we might proceed likewise. For rents, if we want R for an additional year, 1987, we might recompute β_i based on the pooled years 1977-87. That is, we make use of all the available data.

5. Regression Results. We now describe some further computations for the multiple-regression fit that leads to the results in Tables 1 and 2. We (a) consider an analysis of variance, that is, a breakdown of sources of variation, and (b) give 95% confidence intervals for the pooled-region constant-quality ratios in column OR of Table 2. As indicated in Section 2.3, we did a constrained least squares fit: all fitted coefficients β_i are nonnegative. Here we act as though our full set of predictor variables is the reduced set of those for which $\beta_i > 0$. If in fact we had started with this reduced set of predictor variables, we would have gotten the same set of fitted β_i's, with or without the nonnegativity constraint.

(a) Accounting for region-time differences is straightforward. Thus we view our overall sum of squares (SSQ) as that which measures differences among logs of rents (likewise of prices) within region-years. We break this overall SSQ into 3 components:
(1) That which is accounted for by the single explanatory variable X_1: log of square feet.
(2) That which is accounted for by the remaining explanatory variables beyond what is accounted for by X_1.
(3) That which is not accounted for by X_1 and the other explanatory variables.

As percents of the overall SSQ we have for rents (1)23.46, (2)35.55, and (3)40.99. For condos we have (1)17.15, (2)23.21, and (3)59.64.

Along with the above SSQ measuring differences within region-year, one may be interested in the SSQ which measures differences among regions but within years. The sum of these is the SSQ which measures, simply, differences within years. Accordingly, we have computed the percent of the differences within years that is accounted for by regional differences. For rents this percent is 10.22, for condos it is 14.08. In other words, most of the within-year variation is not accounted for by regional differences.
(b) Using conventional multiple-regression methodology, we form 95% confidence intervals for the ratios in column OR of Table 2; note that

our model works with logs, and that we thus use an "exp" transformation to get intervals for the ratios themselves. The ratios (as in Table 2) and usefully narrow confidence intervals appear in Table 3.

Table 3
Confidence Intervals

Rents

	Ratio	Interval
1978	1.089	1.084 to 1.094
1979	1.227	1.221 to 1.233
1980	1.006	0.993 to 1.019
1981	1.139	1.124 to 1.154
1982	1.314	1.296 to 1.332
1983	1.270	1.253 to 1.287
1984	1.290	1.274 to 1.307
1985	1.389	1.371 to 1.407
1986	1.469	1.451 to 1.488

Condos

1984	1.038	1.026 to 1.052
1985	1.111	1.097 to 1.125
1986	1.320	1.302 to 1.337