THE SELECTION OF VARIABLE COMBINATIONS FOR PREDICTING HOUSING CHARACTERISTICS

William M. Shenkel and Cathie Bennett*

Housing characteristics and their prediction have assumed increased importance among urban economists. Interest has centered on slum and blight and on neighborhood changes associated with the demographic environment. These developments have encouraged multiple regression analysis as a means of measuring housing price changes. Predictive models are directed to other aspects of housing.

Of major importance are special purpose models that treat housing values as the dependent variable relative to a series of housing characteristics. The latter examples have been developed for local property tax purposes, mortgage analysis and urban planning. In these instances, the recommended procedure rests on the prediction of housing values by multiple regression analysis of selected housing characteristics. The general objective of these studies is to formulate an equation to predict housing values from a set of independent variables.

Errors in predictive models arise mainly from one or more of four deficiencies: (1) Selection of the wrong combination of independent variables, (2) invalid coding procedures, (3) poor fitting of curvilinear relationships and (4) non-homogeneous samples. It is the latter point that assumes the greatest importance relative to the other statistical deficiences and the point to which this paper is directed.

To expand on this issue consider the problem of predicting housing values from coded characteristics of houses recently sold. The evidence suggests that independent variables associated with lower priced housing are poor predictors to explain the value of higher priced housing. It would appear that independent variables associated with lower priced housing carry different weights and fall in different combinations compared to independent variables associated with higher priced dwellings.[1] The evidence indicates further, that predictive errors are reduced by selecting independent variables unique to discrete housing subsamples. While appropriate independent variables may be readily identified statistically, the optimum stratification of housing samples is much more difficult.

The difficulty is partly resolved by variance analysis. The housing data in hand have shown marked improvement in prediction after housing samples have been partitioned by variance analysis. For instance, locational variables may assume different degrees of importance for discrete housing submarkets. Distance to employment may be highly correlated with dwelling values of modestly priced dwellings but insignificant relative to values of upper income dwellings. In short, it is difficult to explain values without stratifying dwelling samples into homogeneous groups that share a common set of significant, independent variables. Apparently manipulation of independent variables, by coding variations or by transgeneration, will not correct for non-homogeneous samples.

*Chairman and graduate student, Department of Real Estate and Urban Development, College of Business Administration, University of Georgia, Athens.

To illustrate, assume that dwelling prices are treated as the dependent variable. To predict housing values within the same submarket, a combination of housing characteristics must be coded and treated as independent variables associated with housing prices. If the predictive equation proves valid, the value of all houses drawn from the same universe and meeting similar sampling parameters may be predicted within limits of the error term.

SAMPLE STRATIFICATION BY VARIANCE ANALYSIS

Suppose that prices of single family dwellings are treated as repeated observations associated with relevant property characteristics. To improve predictability, the sample of housing sales must be stratified into subsamples before formulation of the multiple regression formula. At the outset, it is presumed that a randomly selected sample of housing prices will vary such that the predictability of independent variables and other coefficients will be relatively low.

To explain further, first identify that set of subgroups which gives the maximum reduction in errors in predicting dwelling values from a series of independent variables (housing characteristics). Starting with the parent sample, data are grouped by the binary division which provides the largest reduction in the unexplained sum of squares. Binary division is continued to reduce further the unexplained sum of squares. This process is repeated for successive subgroups provided a sufficient number of cases justifies multiple regression analysis and provided that multiple regression analysis, applied to each subgroup, continues to reduce the error in predicting the dependent variables.[2]

Begin with the proposition that for the group as a whole, the sum of squares explained by the mean is equal to

$$N\bar{X}^2 = \frac{(\Sigma X)^2}{N}$$

Then the total sum of squares (TSS) unexplained by the mean is equal to

$$\Sigma (X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

If the original group is now divided into two groups, the explained sum of squares is found by

$$\text{Explained sum of squares} = N_1\bar{X}^2{}_1 + N_2\bar{X}^2{}_2$$

Thus the stratification that increases this expression most over $N\bar{X}^2$ represents the optimum division for predictive purposes. Hence let the latter term represent the between group sum of squares (BSS). The optimum solution is given by the maximum value of the between group sum of squares divided by the total sum of squares (BSS/TSS).[3] To illustrate, two housing samples were subjected to this procedure: the first in Montgomery County, Pennsylvania, and the second in Bibb County, Georgia.

PREDICTION OF HOUSING VALUES IN MONTGOMERY COUNTY, PENNSYLVANIA

Dwelling sales completed between January 1967 and December 1970 were subjected to multiple regression analysis. This initial study produced a formula that predicts dwelling values given the relevant housing characteristics. The average dwelling price for some 481 houses in the Montgomery County sales sample was $30,406. These houses had an average square foot floor area of 2,221 square feet and dealt with houses constructed between 1972 and 1970, though only ten dwellings were constructed before 1900. Regression analysis on the entire sample using 29 independent variables gave a standard error of estimate of $3,757, a multiple coefficient of determination of .77, and an average residual of $2,940.

Each sample stratification, to be acceptable, should improve predictive results as shown by the standard error of the estimate, the multiple coefficient of determination and the average residual. The standard error of the estimate and the average residual values are also calculated as relatives of the mean sale price. Relative measures of prediction correct for differences in the means of each subsample. For the original sample these relative measures were 12 percent for the average residual. It should also be noted that predictive accuracy was improved by converting certain independent variables to dichotomous form. After stratification of the sample, prediction may be improved by changing coding techniques and by transgenerating coded variables to show curvilinear relationships.

To apply variance analysis to stratify the sample, each independent variable was grouped into uniform frequency intervals. For example, to show variations in price according to the date of sale, each month beginning with January, 1967, was coded from 1 to 48. The year of construction was divided into intervals beginning with interval 1 for buildings completed before 1900. Houses completed in 1900 or later were grouped by intervals of five years. For certain variables with a limited range, e.g., the number of rooms or the number of fireplaces, data were entered in their original format. Such grouping produced valid results and markedly reduced the number of calculations.

The BSS value was then calculated for each independent variable for the original sample of 481 observations. Sample stratification was based on the variable showing the largest BSS value. In the first step, BSS values indicated that the sample should be stratified into two groups: dwellings with a land area of less than 20,000 square feet and dwellings with 20,000 square feet or more. This division produced the largest BSS value which, expressed as a relative of the total sum of squares, was 34.3 percent.

The first division stratified the sample into two groups of 306 and 175 cases. Before dividing this sample further, stepwise multiple regression analysis was applied to both groups. Multiple regression analysis for these two samples indicated considerable improvement in predictability relative to the formula derived from the 481 cases.

Since multiple regression analysis showed an improvement in predictability as the sample was stratified by BSS values, the two subgroups were subjected to a second binary division. In turn each succeeding subgroup was stratified into a series of binary groups provided (1) each successive group included an adequate number of observations for multiple regression analysis and (2) multiple regression analysis applied to each successive subgroup improved predictability.

The original group of 481 cases was first divided into two groups according to the number of square feet of lot area. These two groups were divided in turn by the number of bathrooms and the date of sale. For the 162 dwellings sold from September 1968 to April 1970, the model showed an additional improvement in prediction by subdividing according to the number of square feet of floor area.

It will be recalled that one of the conditions for guiding binary stratification turns on the improvement in multiple regression results. For this purpose it is relevant to compare changes in (1) the mean value of the dependent variable, (2) the standard error of the estimate, (3) the standard error of the estimate relative to the mean value of the dependent variable, (4) the multiple coefficient of determination, and (5) residual values for each sample group. Table 1 summarizes these figures for the nine groups derived by the stratification process.

For group A the average residual value is $2, 940. Item (7) of Table 1 shows this value reduced to a relative of the mean--9. 6.percent of $30, 406. This calculation reveals a reduction in the error term as a relative value for successive groups, though the absolute value may increase (See group A_1). In item (6) each residual value is expressed as a percent of the dependent variable and averaged for the sample group. The figure for item (7) shows the average residual value in absolute terms reduced to a relative of the dependent variable mean.

The initial sample, group A, shows a relatively high standard error of estimate, representing 12 percent of the dependent variable mean. The residual value in absolute terms, and as a relative of the dependent variable mean, is also comparatively high--9. 6 percent.

As each group is stratified, multiple regression analysis continues to reduce predictive errors. For instance, the first stratification of groups A_1 and A_2 reduces the standard error of estimate (as a relative) to 11 percent and lowers residual values, as a percent of the dependent variable mean, to 8. 6 percent and 7. 7 percent. A review of Table 1 indicates, for group A_{122} and group A_{21}, that residual values have been reduced to a nominal error of 2. 2 percent and 2. 9 percent. Therefore binary stratification according to BSS values tends to improve predictability.

Certain tentative generalizations may be drawn from these data. First, it will be observed that the significance of independent variables varies widely between the five final groups. Secondly, the constant term and the importance of variables significant in more than one group varies considerably. For example, the attached garage shows a numerical importance of 934 for group A and 6, 280 for group A_{21}. Similar relationships were observed for other variables. Thirdly, the number of significant variables ranged from 33 for group A_2, to 11 items for group A_{122}.

Part of the explanation for the wide range of data coefficients, negative terms and highly variable constant terms lies in the degree of autocorrelation between independent variables. Since the objective is to predict housing values, there is little point in refining these terms to show functional relationships. In the present case, each beta value shows the association between the independent variable and the dependent variable. In fact the variation in beta values suggests a high degree of intercorrelation which is quite irrelevant for predictive success.

Support for sample stratification according to BSS values lies partly in the differences in significance observed between variables of sample subgroups. These relationships are most outstanding in groups A_{21} and A_{122}. The first group, A_{21}, shows a low end mean price of $18, 506, while the second group, A_{122}, has a high end mean price of $43, 796. Of the twelve independent variables proving significant for group A_{21} only five of these are recorded for group A_{122}, the high value group. Also note that the most significant item for group A_{21}, attached garage, is seventh on the list for the group of more valuable houses.

Hence, in applying multiple regression analysis for predictive purposes, housing data show substantial variation in significant variables, their beta coefficients change widely, and the relative importance of common variables varies among subsamples. For these reasons, predictive errors appear to be magnified if dwelling samples are not stratified into fairly homogeneous groups.

PREDICTION OF HOUSING CHARACTERISTICS IN BIBB COUNTY, GEORGIA

To further test the feasibility of stratifying samples according to BSS values, some 321 dwellings sold in Bibb County, Georgia, between January 1969 and June 30, 1971 were subjected to the same procedure. Initially the sample revealed a mean sale price of $25,211. The multiple regression formula predicted housing values to a standard error of estimate of $2,198 (8.7 percent of the mean price) and an average residual of $1,599 (6.3 percent of the mean sale price). Stated differently, the mean deviation of $1,599 showed an average 6.5 percent deviation from sales prices. See Table 2.

Variance analysis divided the 321 cases into two groups of 198 and 123 observations. The first binary stratification divided cases between dwellings with floor areas equal to or greater than 1,500 square feet and dwellings with less than 1,500 square feet. The average residual as a percent of the mean sales price reduced to 5.5 percent for both groups--a reduction of 1.8 points from data applying to 321 cases. The standard error of the estimate as a relative of the mean price was also lower for the two subgroups.

Those houses with 1,500 square feet or more were further stratified into dwelling groups with carports and without carports. In both instances, the standard error of the estimate, as a relative of the mean value of the dependent variable, and the average residual value were reduced. Therefore, the data of Bibb County, Georgia, exhibited results similar to the data shown for Montgomery County, Pennsylvania.

Beta coefficients for the original sample and the samples resulting from the stratifications again showed different values. Further, the number of variables varied between the four final groups. In this regard, data of Bibb County, Georgia, again showed statistical results comparable to data of Montgomery County, Pennsylvania.

CONCLUSIONS

Variance analysis has substantially reduced errors in predicting dwelling values. The data show that a series of binary divisions based on between group sum of squares values tends to improve predictability of formulas derived from multiple regression analysis. In the present instance, data from two dwelling samples reveal a general improvement in R^2 values, a reduction in the standard error of the estimate (relative to mean values of the dependent variable) and a lowering of residuals. These findings are consistent with other studies.

While this procedure has been tested for data that apply to predicting values, prediction of other housing characteristics, it is believed, may be subjected to this same analysis. Housing surveys are underway to record the condition of housing, neighborhood qualities and related data. In these efforts variance analysis, in the light of the present study, should precede refinement of multiple regression techniques. There seems little merit in revising data coding or transgenerating to account for curvilinear relationships if the sample does not meet the optimum tests of variance analysis.

One further point deserving emphasis relates to data refinement. Experience has shown that predictive success for relatively small samples (less than 500 observations) may be improved by omitting extreme values of independent variables. There seems little advantage in retaining a variable greatly removed from the measure of central tendency observed for that variable. To retain independent variables showing extreme values forces a beta coefficient that may not apply to typical non-sample data. Thus the reduction in predictive errors by combining variance analysis with multiple regression analysis depends on a refined formula applied only to data conforming closely to multiple regression samples.

In short there seems adequate support for sample stratification by between group sum of squares techniques before applying multiple regression analysis. Predictive results have been improved for each instance in which this procedure was used. Though the cases at hand are confined to relatively small samples, it is strongly suspected that even greater improvements would result if larger samples were available so that a greater number of binary divisions would prove feasible.

Table 1. A Summary of Multiple Regression Results Showing the Effects of Sample Stratification

Statistical Measures	Sample Group Designation								
	A	A_1	$*A_{11}$	A_{12}	$*A_{121}$	A_{122}	$*A_2$	$*A_{21}$	$*A_{22}$
(1) $\bar{Y}$	\$30,406	\$33,797	\$30,319	\$36,889	\$35,255	\$43,796	\$24,476	\$18,506	\$25,812
(2) SEE	3,757	3,823	2,756	2,999	2,902	1,796	2,794	957	2,703
(3) SEE/$\bar{y}$	.12	.11	.09	.08	.08	.04	.11	.05	.10
(4) R^2	.77	.69	.74	.81	.72	.96	.79	.98	.74
(5) $[\Sigma(Y - Y_c)]/N$	2,940	2,927	1,903	2,045	1,952	974	1,893	538	1,761
(6) $\Sigma[(Y - Y_c)/Y]/N$	10.04	8.87	6.33	5.56	5.49	2.30	8.24	3.16	7.17
(7) $\frac{[\Sigma(Y - Y_c)]/N}{\bar{Y}}$	9.6	8.6	6.3	5.5	5.5	2.2	7.7	2.9	6.8
(8) N	481	306	144	162	131	31	175	32	143
(9) Number of variables	23	28	25	23	14	11	33	12	28

*Final groups used to predict the dependent variable.

Table 2. A Summary of Multiple Regression Results Showing the Effects of Sample Stratification: Bibb County, Georgia

Statistical Measures	B	B_1	$*B_{11}$	$*B_{12}$	$*B_2$
			Sample Group Designation		
(1) $\bar{Y}$	$25,211	$20,723	$20,046	$21,114	$32,441
(2) SEE	2,198	1,663	1,664	1,496	2,585
(3) SEE/$\bar{Y}$	8.7	8.0	8.2	7.9	7.9
(4) R^2	.94	.94	.96	.96	.89
(5) $[\Sigma(Y-Y_c)]/N$	1,599	1,243	1,090	1,051	1,786
(6) $\Sigma[(Y-Y_c)/Y]/N$	6.5	6.1	5.8	5.0	5.5
(7) $\frac{[\Sigma(Y-Y_c)]/N}{\bar{Y}}$	6.3	5.5	5.4	4.9	5.5
(8) N	321	198	73	125	123
(9) Number of variables	22	23	19	17	17

*Final groups used to predict the dependent variable.

FOOTNOTES

[1] For this study it is assumed that data conform to the assumptions necessary to multiple regression analysis. Consult Donald J. Bogue and Dorothy L. Harris. Comparative Population and Urban Research Via Multiple Regression and Covariance Analysis. Published Jointly by the Scripps Foundation Research and Training Center at the University of Chicago. Oxford and Chicago, 1954, pp. 11-12.

[2] For a more complete explanation see John A. Sonquist, Multivariate Model Building (Ann Arbor, Michigan: Survey Research Center, Institute for Social Research, The University of Michigan, 1970), 244 pp.; and John A. Sonquist and James N. Morgan, The Detection of Interaction Effects, Monograph No. 35 (Ann Arbor, Michigan: Survey Research Center, Institute for Social Research, The University of Michigan, 1970), 296 pp.

[3] See also James N. Morgan and John A. Sonquist, "Problems in the Analysis of Survey Data, and a Proposal," American Statistical Association Journal (June, 1963), pp. 415-434 and Bruce L. Gensemer, Jane A. Lean, and William B. Neenan, "Awareness of Marginal Income Tax Rates among High-Income Taxpayers," National Tax Journal (September, 1965), pp. 258-267.