A COMPARISON VIA REPLICATION ANALYSIS OF PREDICTIVE MODELS OF REAL ESTATE SELLING PRICES, MULTIPLE CORRELATION WEIGHTS VERSUS SIMPLE CORRELATION WEIGHTS

Steven W. Lamb, Indiana State University
Samuel C. Certo, Indiana State University

Abstract

The paper illustrates a procedure for developing a predictive model using simple correlation weights between the criterion variable and the set of predictor variables. Next, this predictive model is compared with a standard multiple regression model via replication analysis.

Multiple regression techniques periodically are subject to condemnation as well as periods of great praise. Cooley and Lohnes (1) cite a study that concludes that "conventional least-squares regression should be dropped from the applied statistician's repertiore in favor of prediction-criterion simple correlation weights for sample sizes less than 200." They also imply that if a researcher wants his multiple regression findings to be taken seriously it would be best to judge their validity using replication samples.

The purpose of this paper is to first illustrate the procedure for developing a predictive model using simple correlation weights between the criterion variable and the set of predictor variables. The next step is to compare this predictive model with a standard multiple regression model via replication analysis. A unique feature of this paper is the size of the norming sample, 1729 complete cases. This extremely large norming sample allows the author to present some interesting findings which may have broad application.

The Data Base

The norming sample was composed from 2,315 real estate transactions which took place during an eight year time span from 1969 to 1976 throughout a city of approximately 100,000 population. The test sample was composed from 105 real estate transactions which took place in 1976. The breakdown of the time series data by year and subdivision appear in Table 1.

TABLE I

BREAKDOWN OF DATA BASE,
NUMBER OF OBSERVATIONS BY YEAR

Year	Norming Sample	Test Sample
1969	30	---
1970	206	---
1971	272	---
1972	293	---
1973	373	---
1974	350	---
1975	369	---
1976	422	105
	2,315	105

The test sample was from the most recent time period since the purpose of developing a predictive model is, of course, to forecast the selling prices of homes in the future.

The following independent variables were considered as dummy variables in the initial norming equations; the existence of central air conditioning, of a built-in dishwasher, of a disposal, of a crawl space, of a slab, and of more than one story. Also considered as a dummy variable was the type of exterior construction, brick or wood. Independent variables considered other than the preceding dummy variables were as follows: square footage of lot size, age of house when sold, number of bathrooms, number of bedrooms, total number of rooms minus the number of bathrooms and bedrooms (to remove a possible source of multicolinearity), square footage of livable area minus 144 square feet for each bedroom (an estimate of the average size) and 40 square feet for each bathroom (again to remove a possible source of multicolinearity), the number of fireplaces, and the annual heating cost. A subjective variable indicating the quality of landscaping was included. Finally, time and time squared were included as independent variables. The variable time was the month the sale was made. The base month (the month the first sale was made in the data base) was numbered one. The dependent variable was the price at which the house was sold.

The decision was made to include those independent variables in the norming equations which did not exhibit obvious linear relationships with the other independent variables, and that were found significant, meaningful and interpretable in a multiple regression run.

The Multiple Regression Norming Equation

The norming equation developed from multiple regression procedures is as follows:

$$Y = 16507.41 + 6.028X_1 - 236.54X_2 + 5711.39X_3 + 18.909X_4 - 2807.18X_5 + 3083.35X_6 + 3158.84X_7 + 2923.66X_8 + 50.712X_9 \quad (1)$$

The number of complete cases was 1,709 from the 2,315 cases used in the norming sample. Note that the value of the coefficient of multiple correlation is equal to .84 as shown in Table II.

The next step was to place the values of the independent variables of the test sample into the norming equation deriving estimates for the dependent variable (the actual selling price for the house). Then a simple regression was run between the actual selling price and the estimated selling price. The results are shown in Table III.

TABLE II

REGRESSION OUTPUT ASSOCIATED WITH EQUATION ONE

Multiple R	.839
R Square	.704
Adjusted R Square	.703
Standard Error	7002.190

Variable		Beta Value	Standard Error of b value	F value
Sq. ft. living space	X_1	.295	.328	336.97
Age of the house	X_2	-.380	9.534	615.48
Central air conditioning	X_3	.216	400.459	203.41
Heating cost	X_4	.157	2.096	81.42
Quality of landscaping	X_5	-.158	260.420	116.20
Existence of fireplace	X_6	.119	373.996	67.97
More than one story	X_7	.104	455.318	48.13
Exterior construction	X_8	.095	442.538	43.65
Month house was sold	X_9	.091	9.878	26.36

TABLE III

CORRELATION OUTPUT ASSOCIATED WITH ESTIMATES GENERATED USING NORMING EQUATION ONE IN CONJUNTION WITH THE TEST DATA, CORRELATED WITH THE VALUES OF THE DEPENDENT VARIABLE OF THE TEST DATA

Multiple R	.883
R Square	.779
Adjusted R Square	.776
Standard Error	8076.623

Variable	b value	Beta value	Standard Error of b value	F value
Estimated selling price using norming equation 1	1.324	.883	.082	260.984
Constant - 6492.904				

It must be emphasized that this method of estimating the value of the coefficient of multiple correlation does not capitalize on chance. Since the regression coefficients used in conjunction with the set of independent variables of the test sample were derived from the norming sample, capitalization on chance will not be a problem when the derived estimates are correlated with the actual value.

The numeric value of the coefficient of multiple correlation is unusual in that it is greater than that of the norming samples. There is no reduction in correlation. The fact that the norming sample was extremely large increases the probability of this occurrence. Individuals using this equation for prediction purposes could reasonably expect the amount of shared variance to be equal to .78.

The Simple Correlation Values Used As Weights In The Norming Equation

The simple correlations found in the norming sample between the dependent variable and the set of independent variables used in the multiple regression norming equation were the weights used in the second norming equation.

$$Y_f = r_{y.1} Z_1 + r_{y.2} Z_2 + r_{y.3} Z_3 \text{ ------ } r_{y.9} Z_9$$

$$Y_f = .556Z_1 - .504Z_2 + .547Z_3 + .440Z_4 - .413Z_5 + .176Z_6 + .189Z_7 + .348Z_8 + .253Z_9 \quad (2)$$

The values in the test sample were standardized by using the means and standard deviations found in the test sample. Then estimates of Y_f were derived by placing the values of the independent variables from the test sample in equation (2).

The next step was to derive values that could be used as predicted values for the selling prices for the 1976 test sample observations. This was accomplished by first standardizing the value of Y_f. Then the standardized values of Y_f were altered so that they had the mean and standard deviation of the dependent variable of the norming sample. The mean and standard deviation of the dependent variable of the norming sample was used because the mean and standard deviation of the test sample would not be available in a real world situation. Finally, a simple correlation was run between the actual values of the dependent variable of the 1976 test observations and these predicted values. The multiple correlation coefficient was .872. The results are shown in Table IV.

TABLE IV

CORRELATION OUTPUT ASSOCIATED WITH ESTIMATES GENERATED USING NORMING EQUATION TWO IN CONJUNCTION WITH THE TEST DATA, CORRELATED WITH THE VALUES OF THE DEPENDENT VARIABLE OF THE TEST DATA

Multiple R	.872
R Square	.761
Adjusted R Square	.758
Standard Error	8401.664

Variable	b Value	Beta Value	Standard Error of b value	F Value
Estimated selling price using norming equation (2)	1.160	.872	.076	235.566
Constant - 491.305				

Conclusions

The amount of shared variance between the norming equation estimates and the actual values had a decrease of only 2 percentage points when the estimates were derived using simple correlation weights versus multiple correlation weight. Initially, it looks as if simple correlation weights fare rather well, when they are used for prediction purposes. But it must be realized that this particular set of independent variables did not suffer from serious multicolinearity problems. When the "independent" variables have no correlation among themselves then the standardized multiple regression equation weights become nothing more than the simple correlation values between the dependent variable and the specific independent variable. Thus, when the set of independent variables have little multicolinearity among themselves, one might as well use the simple correlation values as predictor weights; the multiple regression weights (which are certainly more difficult to interpret) will yield the same results. When the set of independent variables suffer from multicolinearity, one should distinguish between two cases, small sample size versus large sample size. When the sample size is small, the argument for simple correlation weights is strong due to their relative stability as compared to multiple regression weights which are influenced by the stability of the estimates of covariance among variables. However, when the sample size is large, the increase in the stability of the regression coefficients will generally yield more reliable estimates.

One, however, could always develop two predictive models; the first based upon simple correlation weights, the second based upon multiple regression weights. Then comparisons could be made using a split sample design to determine the more reliable model.

References

William W. Cooley and Paul R. Lohnes, Multivariate Data Analysis (New York: John Wiley & Sons, Inc., 1971, page 56)

Norman H. Nie, Statistical Package for the Social Sciences (New York: McGraw-Hill, 1975)