Appendix D

Fundamentals of Mass Appraisal

Robert Gloudemans
Richard Almy

International Association of Assessing Officers
Kansas City, Missouri

Chapter 7
Multiple Regression Analysis

Multiple regression analysis (MRA) is a statistical technique for estimating unknown data on the basis of known and available data. MRA is the workhorse of mass appraisal. It can be used to help determine the relationship between two variables, for example, between sale price per unit and time of sale or between percent good and effective age, as illustrated in Chapter 4, "Key Issues in Mass Appraisal." At a more sophisticated level, MRA can be used to estimate market values or income parameters (rent per unit, expense ratios, gross income multipliers, and capitalization rates) from an analysis of many variables. This chapter provides a primer on MRA in mass appraisal and provides the bridge between the mass appraisal concepts and methods described previously and the specification and calibration of MRA models for various property types taken up in Chapters 8, "Land and Residential Models," and 9, "Commercial Models."

MRA models can be additive, multiplicative, or hybrid. Additive models are the least flexible but the simplest and most common. This chapter illustrates MRA using additive model structures and then discusses multiplicative and hybrid models.

The general structure of an additive MRA model in which sale price is the dependent variable is

$$S = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p, \qquad (1)$$

where

S = sale price (dependent variable)

$X_1, X_2, \ldots, X_p$ = the independent variables

$b_1, b_2, \ldots, b_p$ = coefficients or prices assigned by the algorithm to the independent variables

b_0 = a constant determined by the algorithm.

This general model structure can be used to estimate any dependent variable, usually abbreviated Y in statistical textbooks. S is used throughout this discussion because it is the dependent variable of interest in sales comparison models.

As a simplified illustration, consider the equation,

$$S = 45{,}600 + 124.20 \times X_1 - 1{,}400 \times X_2$$

where
X_1 = square feet of living area
X_2 = effective age.

In this case, b_0 is 45,600, b_1 is 124.20, and b_2 is −1,400. For a house with 2,000 square feet and an effective age of 15 years, the predicted value is

$$S = 45{,}600 + (124.20 \times 2{,}000) - (1{,}400 \times 15)$$

$$S = 45{,}600 + 248{,}400 - 21{,}000 = 273{,}000.$$

The coefficients calculated for the variables are derived from sales analysis and reflect their respective contributions to the estimation of sale price. A more realistic example would contain additional independent variables.

As with any valuation technique, accurate MRA models require reliable market and property characteristics data. MRA tends to work well when sales are sufficient and property characteristics are coded consistently. Predicted values are particularly accurate for parcels with typical characteristics. Predicted values for parcels with atypical characteristics can have high margins of error and should be reviewed.

Theory and Method

The objective of MRA applications of the sales comparison approach is to model the relationship between property characteristics and value, so that unknown property values can be estimated from known property characteristics. Using the 35 sales in Table 7-1, Figure 7-1 graphs the relationship between living area and sale price and fits a trend line to the data. The sale price of an unsold property can be estimated by noting its size and reading the corresponding estimated sale price from the trend line. For example, to estimate the value of an unsold house with 2,000 square feet of living area, a vertical reference line is drawn at 2,000 square feet. Then a horizontal reference line is drawn through the point at which the vertical line intersects the trend line. This process is illustrated by the dashed lines in Figure 7-2. The estimated value of the house is approximately $170,000.

Regression analysis fits the trend line to the data using the principle that a straight line can be determined by one point on the line and its slope. In fact, the regression equation to estimate sale price based on only one independent variable is

$$S = b_0 + b_1X_1, \tag{2}$$

Table 7-1.
Regression Data

Sale Number	Square Feet	Sale Price
1	750	109,500
2	778	75,900
3	860	132,000
4	924	105,000
5	1,020	160,000
6	1,116	100,00
7	1,190	129,90
8	1,292	945,0
9	1,350	140,00
10	1,487	128,30
11	1,500	169,5
12	1,555	130,0
13	1,650	119,5
14	1,724	219,0
15	1,750	159,
16	1,800	185,
17	1,842	110,
18	1,912	146,
19	2,110	185
20	2,297	145
21	2,450	22
22	2,504	16
23	2,524	18
24	2,590	21
25	2,638	14
26	2,773	2
27	2,876	1
28	2,940	2
29	3,042	2
30	3,110	[illegible]
31	3,288	
32	3,430	
33	3,498	
34	3,740	
35	3,910	

is large. One means of minimizing $\sum e_i^2$ is to add additional variables. In Figures 7-1 and 7-2, some points lie below the regression line because they represent properties with negative features, such as minimal construction quality or poor physical condition. Other points lie above the line because they represent properties with positive attributes, such as above-average construction quality or good physical condition.

The model might be respecified as

$$\sqrt{I} = b_0 + b_1X_1 + b_2X_2 + b_3X_3, \tag{4}$$

where

X_2 = construction quality

X_3 = physical condition.

Again, MRA would calculate the regression coefficients b_0, b_1, b_2, and b_3 to minimize $\sum e_i^2$, where, in this case, the predicted values are a function of living area, construction quality, and physical condition. Note that the importance of any one variable in the regression equation is directly related to its contribution in reducing $\sum e_i^2$.

Evaluation of Regression Results

Users of MRA should be familiar with key statistics that help evaluate the accuracy and reliability of models. Some of these are measures of *goodness of fit* and relate to evaluation of the predictive accuracy of the equation. The most important are the coefficient of determination (R^2), the standard error of the estimate (*see*), the coefficient of variation (*COV*), and the average percentage error. In different ways, each indicates how well the equation succeeds in minimizing $\sum e_i^2$ and predicting the dependent variable. Other regression statistics relate to the importance and reliability of individual variables in the model. They include the coefficient of correlation (r), t-statistic, F-statistic, and beta coefficient.

Coefficient of Determination

The coefficient of determination, R^2, is the percentage of the variation in the dependent variable explained by the regression model. Assuming that no records are kept of the physical description, site amenities, or other characteristics of properties, other than sale prices, how would the market value of any given property be estimated? One obvious answer is the average sale price. For properties that have sold, the sum of the squared errors, SSE, associated with this estimate is

$$SSE = \sum(S_i - \overline{S})^2 \tag{5}$$

where $\overline{S}$ = the average sale price.

t-Statistic

The t-statistic is a measure of the significance or importance of a regression variable in explaining differences in the dependent variable (sale price). It is calculated as the ratio of the regression coefficient, b_j, to its standard error, s_j (not to be confused with *see*):

$$t_j = b_j \div s_j. \qquad (16)$$

The standard error of b_j, s_j, is akin to a standard deviation; it measures the error associated with using b_j as an estimator of the true but unknown relationship between X_j and the dependent variable in the model.

When t_j is large, we can be confident that X_j is a significant predictor. Conversely, when t_j is small, we cannot reject the null hypothesis that $b_j = 0$ and thus we cannot conclude that X_j is an important predictor. However, this does not mean that X_j is not correlated with the dependent variable. The t-value measures the *marginal contribution* of an independent variable in predicting the dependent variable when all other variables included in the model are held constant. Because some variables duplicate information provided by others, they may be highly correlated with sale price, but are insignificant predictors as indicated by their t-values. Conversely, other variables possess the peculiarity of predicting sale prices in combination, although individually none may be highly correlated with sale prices.

The significance of t-statistics can be evaluated by reference to a t table (see Table A-2 in Appendix A), where degrees of freedom $= n - p - 1$ and p is the number of independent variables in the model. In general, provided that sample size is large (at least 50), a t-value in excess of ± 2.00 indicates that we can be 95 percent confident that $b_j = 0$ and therefore that X_j is a significant predictor variable. Similarly, a t-value in excess of ± 2.58 indicates that we can be 99 percent confident that X_j is a significant predictor. Most statistical software reports a probability statistic that indicates the significance of the t-value, sparing the need to reference a t-table.

For the 35 sales in Table 7-1, the regression coefficient for square feet of living area is 61.884 and the standard error is 6.799 (see Table 7-2). Thus the t-value is

$$t = 61.884 \div 6.799 = 9.102.$$

The associated significance value, namely 0.000, is the probability that $b_j = 0$, that is, that square feet of living area is *not* a significant predictor of sale price. Thus, in this case, we can be virtually 100 percent confident that square feet of living area is a significant predictor of sale prices.

F-Value

The F-value is direct
individual regressior
F-values are based o

F = variance ex

Additional variance
unexplained variar
variance, the more
can be of the varia
case with the t-va
importance of an
all other variables
equation).

In MRA, the

$$F = t^2.$$

That is, the F-va
F-values of app
the 95 percent
confidence leve

Some regre
both measure t
one or the oth
changed to F-

Beta Coe

Beta coefficie
importance o
variable. Bet
independent
of one. For e
by its standa
or leverage
coefficients
dependent

F-Value

The *F*-value is directly related to the *t*-value and is also used to test whether or not individual regression variables are significant predictors of the dependent variable. *F*-values are based on the ratio

$$F = \text{variance explained by } X_j \div \text{unexplained variance.} \tag{17}$$

Additional variance refers to the amount by which the inclusion of X_j reduces unexplained variance. Obviously, the larger this amount relative to unexplained variance, the more important is X_j in reducing $\sum e_i^2$ and the more confident we can be of the variable's significance in predicting the dependent variable. As is the case with the *t*-value, however, the *F*-value provides a measure of the marginal importance of an individual variable in explaining the dependent variable when all other variables are also taken into account (by including them in the regression equation).

In MRA, the *F*- and *t*-values are mathematically related:

$$F = t^2. \tag{18}$$

That is, the *F*-value is the square of the *t*-value. Provided that sample size is large, *F*-values of approximately 4.0 or larger indicate that a variable is significant at the 95 percent confidence level. Again, the probability statistic indicates the exact confidence level.

Some regression programs report *t*-values, and others report *F*-values. However, both measure the same thing, and some software gives the user the ability to select one or the other (SPSS software reports *t*-values by default, although this can be changed to *F*-values in syntax mode).

Beta Coefficients

Beta coefficients are *standardized* regression coefficients that measure the relative importance of the independent variables in explaining or predicting the dependent variable. Beta coefficients are obtained by transforming the dependent and independent variables so that they all have a mean of zero and standard deviation of one. For each variable, this is accomplished by subtracting its mean and dividing by its standard deviation. A beta coefficient, B_j, thus measures the relative influence or leverage that an independent variable exerts on the dependent variable. Beta coefficients can be loosely thought of as representing the percentage change in the dependent variable associated with a percentage change in the independent variable with all other variables held constant. Beta coefficients are related to regression coefficients by the formula

$$B_j = b_j \times (s_j \div s_y)\,, \quad (19)$$

where

s_j = the standard deviation of X_j

s_y = the standard deviation of the dependent variable in the model.

(In a one-variable linear regression, B_j equals the correlation coefficient between the independent and dependent variables.)

Beta coefficients are useful in evaluating the relative importance of independent variables in the model. Assume the data shown in Table 7-4 for the variables SFLA (square feet of living area), QUAL (construction quality), and EFFAGE (effective age). Because all three variables are measured in different units, their regression coefficients cannot be meaningfully compared. However, based on their beta values, we can conclude that SFLA is the dominant variable in the model, followed by EFFAGE, and then QUAL.

Table 7-4. Beta Coefficients for Three Variables

Variable	Mean	Coefficient	Beta
SFLA	1,534	84.68	.585
QUAL	3.48	15,459	.183
EFFAGE	33.10	−2,785	−.266

Stepwise and Backward Regression

Model builders have several options in applying MRA. The default method in most software is automatic inclusion of all candidate variables. Two useful alternatives that filter redundant or otherwise insignificant variables are *stepwise regression* and *backward regression.*

In stepwise regression, variables are entered one at a time until all significant predictors have been included. The variable entered first, say, X_1, is that variable most highly correlated with the dependent variable. A least-squares regression is performed, the residuals (errors) are saved internally, and a search is made to determine the remaining variable most highly correlated with and thus able to reduce the errors from the first model. Suppose that this variable is X_4. A second regression is performed with X_1 and X_4 as independent variables. The remaining variables are searched to determine which has the highest correlation with the residuals from the second regression. That variable, say, X_6, is then included in a third regression. The process continues until all variables have been included or the remaining variables fail to meet some predetermined significance level for inclusion (0.05, which implies 95 percent confidence, is the most common default level). At each step the algorithm may either add a new variable or delete a variable that falls below a minimum significance level for retention in the model. (The significance level for entry must be set stricter than the significance level for retention to prevent the repeated entry and removal of the same variable.) The procedure eliminates insignificant variables and helps prevent the model from being more complex than necessary.

In backward elimination, the algorithm begins with all variables and iteratively eliminates those that are not significant while adding back any that were previously eliminated but subsequently achieve significance. Aside from the possibility of a variable being deleted and then added back, backward regression leaves the variables in the same order that the modeler listed or entered them. Stepwise regression output, on the other hand, lists variables in the sequence in which they entered the final model, so that the most important or significant variables are listed first and the least significant are listed last.

For illustrative purposes, consider the statistics for 667 residential sales shown in Table 7-5. The mean sale price is $255,898, and there are 16 potential predictor variables. Some of these, such as SFLA and FINBSMT, are quantitative variables. QUAL is a discrete (categorical) qualitative variable. The air conditioning, pool, masonry exterior, and neighborhood variables are binaries, coded 1 if the feature or condition is present and 0 if not. For these variables, the mean represents the percentage of cases with the feature. For example, 29.7 percent of the homes have air conditioning, 9.8 percent have swimming pools, and 10.6 percent have masonry exterior walls. In this example, neighborhood 403 represents the *base*, or *reference*, neighborhood. It is a neighborhood with typical sale prices and good market activity. Regression coefficients determined for the other neighborhoods thus reflect market preferences relative to neighborhood 404.

Table 7-6 displays the correlation matrix for the variables. It shows the correlation coefficients between the independent variables and the dependent variable, PRICE, as well as with each other. The correlations of PRICE with SFLA and QUAL are particularly strong. The matrix also reveals potential information overlaps or interrelationships among the independent variables. For example, QUAL is rather highly correlated with SFLA, UNFBSMT, GARSIZE, and NBHD_406.

Property Appraisal and Assessment Administration

General editor
Joseph K. Eckert, Ph.D.

Senior technical editors
Robert J. Gloudemans
Richard R. Almy

The International Association of Assessing Officers

the tax rate is expressed as lar. For example, a 2 percent per $100, $20 per $1000, or dollar.

. The right to extract ore, pe- her minerals from a property.

etric. Any of a family of of measuring distance. **Eu- nce**, a member of this fam- traight-line distances (as the quared. In mass appraisal g, Minkowski metric usually sum of absolute differences in each dimension, and axicab" or city block pattern. tives are possible, including as calculated only for the greatest difference, but the tance is most common.

ue most often assumed by extension for grouped **data**, hich a plurality of the ob-

esentation of how some- or purposes of appraisal, a (in words or an equation) the relationship between ted sale price and **variables** **pply and demand factors**.

The interaction of buyers ort-term credit instruments.

term used in land survey- to mean a permanent ob- e ground marking a point, which is known, on the rth. *See also* **geodetic** con-

Mortgage coefficient. A component of the basic rate in the Ellwood variant of **mortgage-equity analysis.**

Mortgage constant. Annual **debt service** expressed as a percentage of the initial principal amount of the loan.

Mortgage-equity analysis. A technique used to estimate the value of a property from a knowledge of the equity **yield rate,** typical mortgage terms (including the interest rate, the loan-to-value ratio, the term of the loan, and the amortization provisions), the **holding period**, and the percentage by which the property will **appreciate** or **depreciate** over the holding period.

Moving average. A **statistic** used to smooth the values of a **variable** when those values are erratic over distance or time, as in the case of land values and mortgage commitments. For example, a five-block simple moving average of land values along a major street would assign to block 16 the average of the values for blocks 14–18; it would assign to block 17 the average of the values for blocks 15–19, and so on.

Multicollinearity. The phenomenon of two or more **variables** being correlated. If the two correlated variables are both **independent variables** (note that if they are correlated they are not truly *independent* in the relationship sense) used to predict the value of some other, dependent variable, then modeling problems will arise. If the multicollinearity is perfect, the **multiple regression** algorithms simply will not work; if the multicollinearity is serio imperfect, the coefficients generated b **algorithm** will be individually mear less (although the **model** as a whole still be useful).

Multiple regression, multiple regressi analysis (MRA). A particular statisti technique, similar to **correlation**, used analyze **data** in order to predict the val of one **variable** (the **dependent variable** such as **market value**, from the know values of other variables (called **indepen- dent variables**), such as lot size, numbe of rooms, and so on. If only one independent variable is used, the procedure is called simple regression analysis and differs from correlation analysis only in that correlation measures the strength of relationship, whereas regression predicts the value of one variable from the value of the other. When two or more variables are used, the procedure is called multiple regression analysis. *See* **linear regression.**

Multiplicative model. A model in which the **coefficients** of **independent variables** serve as powers (**exponents**) to which the independent variables are raised or in which independent variables themselves serve as exponents; the results are then multiplied to estimate the value of the **dependent variable**.

Multiplicative transformation. A **transformation** of a set of **variables** accomplished by multiplying a variable by one or more other variables. For example, room area is a multiplicative transformation of length and width.

Linear regression. A kind of statistical analysis used to investigate whether a **dependent variable** and a set of one or more **independent variables** share a linear **correlation** and, if they do, to predict the value of the dependent variable on the basis of the values of the other variables. Regression analysis of one dependent variable and only one independent variable is called simple linear regression, but it is the word *simple* (not *linear*) that distinguishes it from **multiple regression analysis** with its multiple independent variables.

Listing. The process by which the assessor ensures that records for the taxable property identified during **discovery** are preserved with **integrity**, available for use in **valuation** activities, and ultimately reflected in the **assessment roll**.

Locational obsolescence. A component of **economic obsolescence**; loss in value due to suboptimal siting of an **improvement**.

Location variable. A **variable**, such as the distance to the nearest commercial district or the traffic count on an adjoining street, that seeks to measure the **contribution** of locational factors to the total property value.

Logarithm; log. The number that, when used as an **exponent** for another number (called the base), results in a third number of some practical interest (called the antilogarithm). There are two bases that are used with any frequency; the base 10 produces what are called common logarithms, and the base 2.71828 (e) produces what are called natural logarithms. For example, $\log_{10}100 = 2$; $10^2 = 100$. Logarithms were originally used to simplify complex calculations involving multiplications inasmuch as two numbers can be multiplied by adding their logarithms and taking the antilog of the result. Logarithms are also used as means of transforming variables in regression analysis.

Log-linear relationship. A **correlation** between two **variables** such that if the **value** of one variable changes by a certain percentage, the value of the other changes by a certain amount. (Recall that **logarithms** permit multiplication to be done by means of adding logs.) For example, there is a log-linear relationship between x and y in the following sequence:

x	5	6	7	8
y	20	30	45	67.5

Long run. A planning period long enough for a firm to be able to vary quantities of all resources it uses.

Macroeconomics. The economics of the economy as a whole—the forces causing recession, depression, and inflation together with the forces resulting in economic growth.

Mann-Whitney test. A test in **inferential statistics**, similar to the **Kruskal-Wallis test**, that seeks to determine whether the differences in values between two sets of observations from any population are statistically significant.

Map book and page system. A system for parcel identification in which a code (usually numeric) is used to identify each parcel, each code containing four elements: the volume or book of maps in which the parcel is to be found, the page on which it is to be found, the block, and the individual parcel on the block.