Gabriele Brondino[†]

Mathematical/statistical interpretation of direct valuation methods

Keywords: **direct valuation methods, market value, transaction price, multiple linear regression, sales comparison approach, nearest neighbors appraisal technique.**

Abstract Starting from the conditions dictated by the valuation principles of cost forecasting, normality, and comparison, the links between estimated Market Value and Transaction Price can be explained from a mathematical point of view.

The Transaction Price Model and a significant estimate sample are crucial to the development of any one of the direct methods for estimating Market Value.

The goal of this study is to describe the mathematical and statistical relationships - and the difficulties of application - at the base of direct valuation methods, utilizing pseudo-real case studies. The article will take into consideration a particularly large sample to provide an estimate of Market Value of the asset to be assessed through the statistical methodology known as multiple linear regression. The so-called Sales Comparison Approach (SCA) will then be applied to estimate the Market Value in the case of a relatively small sample. Finally, the practical utility and the conditions for the application of a valuation method known as Nearest Neighbors Appraisal Technique (NNAT) will be explored.

Let us consider an apartment of 80 square meters (sm) located in a low-income neighborhood in Turin, built in the 1950s, partially renovated, with two exposures, a cellar, etc. Let us suppose that the apartment must be appraised. The appraiser should first verify whether a sample can be made in order to make a direct estimate of the Market Value of that apartment. The sample will be made up of a number of apartments similar to the one to be appraised, recently sold and in the same geographical area. The expert will proceed with the appraisal according to the cost forecasting, normality (principio di ordinarietà in Italian in which "ordinary" is meant to be the most frequently occurring value in a normal distribution coincides with the average – translator's note) and comparison principles of valuation.

PRICE AND MARKET VALUE

The *Market Value* of a property being appraised is defined as the most likely transaction price of that property. The purchase price, however, is the quantity of money exchanged between buyer and seller in a deed of sale.

From the above definition, it appears that the estimate of the Market Value is somehow related to Transaction Price. In the following article, we will try to clarify this relationship from a mathematical point of view. For the moment, it is sufficient to point out that in the appraisal practice, Market Value can be estimated directly only if the real Transaction Price is known for all the units in the sample survey.

The direct estimate of Market Value must respect the normality principle. Thus, the value obtained will not coincide exactly with the effective Price that will be exchanged between the parties, but

[†] Researcher in Estimate at the Second School of Architecture Politecnico of Torino

rather with the most probable Price! Or with the price that would be formed in a perfect market and which the majority of appraisers would estimate. On the other hand, the estimate of Market Value must satisfy the principle of cost forecasting. Therefore the estimated Market Value will not deviate much from the real Transaction Price established after the appraisal. It is evident that the two assumptions seem apparently contradictory. We will see how this obstacle can be overcome from the statistical point of view.

All methods referring to the direct valuation of the Market Value of an asset (Sales Comparison Approach, Regression, Direct Monoparametric Estimation, etc..) can be defined through mathematical or statistical models. In this way similarities, differences, advantages and limitations of individual methods can be understood.

THE MARKET VALUE MODEL

Let us return to the 80 square meter (sm) apartment. The task of the appraiser is to provide an estimate of its Market Value. This requires that the appraiser ask him/herself which characteristics (or variables) define the Market Value of the apartment subject to appraisal. Far from claiming to make a comprehensive list, we sought to identify some of the characteristics that might influence the formulation of the apartment's Market Value:

- geographic location
- neighborhood safety
- services in the vicinity of the apartment
- total floor area
- year of construction
- state/conditions of the apartment
- state/conditions of the building
- number of exposures
- presence of cellar
- real estate market dynamics
- mortgage interest rates
- indices regarding the future of the economy
- ...

The characteristics listed (and those not listed) can be grouped in different ways. For example, we could agree with the fact that some refer to location (location, neighborhood, ...), others to the technical characteristics of the dwelling unit (floor area, year of construction, ...) while others focus on global and local economic and financial trends (real estate market dynamics, interest rates, indices, ...).

Some characteristics are easily identifiable (for example, number of exposures) while others are more difficult to specify (for example, degree of neighborhood safety). There are characteristics that can be specified without error (for example, number of exposures) but there are also characteristics that can give rise to measurement errors (for example, year of construction). The characteristics can also be divided into quantitative ones (for example, floor area) because the information collected is numeric, and qualitative ones (for example, condition of the dwelling) because the data are categories and not numbers.

Let us assume that we can identify the values of all the characteristics that determine Market Value (which we will indicate as VM_0) of the property to be appraised. For example, let us suppose that the coordinates of the apartment are x_{01}=7.1403 (latitude) e x_{02}=45.0456 (longitude). Similarly, we can assume that the degree of neighborhood safety is described as x_{03}= "Good" and so on for the other characteristics. From a mathematical point of view, it is easy to assume that VM_0 is a function (as complex as you want) of the characteristics that determine value, so that

$$VM_0 = h_0(x_{01}, x_{02},...) \quad (1)$$

If, in addition to the values of the characteristics x_{0j} (j=1,2,...), we also knew the value of the function h_0 there would be no major problems, and the appraisal would be a very simple matter. Unfortunately, reality is much more complicated.

THE TRANSACTION PRICE MODEL

That the Transaction Price of a property is somehow linked to Market Value is a commonly accepted fact. It is no coincidence that the cost forecast principle requires that the appraiser provide a value judgment that can approach the real Transaction Price (forecast). On the other hand, the Market Value resulting from an expert appraisal almost never coincides exactly with the real Transaction Price of the property. How can we explain this (sometimes significant) difference?

We must first ask ourselves where this difference originates. If the housing market were perfect, in fact, the difference would tend to annul itself. In a perfect market, the buyers and sellers would be numerous, driven by rationality, fully informed about market trends, etc.. It is not difficult to conclude that a property transaction rarely meets these conditions. For example, sellers often do not make their decisions to sell based on rational criteria (consider for example the case in which a person must sell a property for money reasons). Similarly, buyers may be less informed regarding the market than an institutional seller such as a builder/developer. In the deed of sale of a real estate property, it is not difficult to find many market imperfections. In other words, it is common to find characteristics that influence the formation of Price but not the formation of Market Value (for example, the seller's rush to sell a property). For this reason, it is reasonable to assume that the following relationship exists between Price and Market Value:

$$P_0 = VM_0 + \delta_0 = h_0(x_{01},x_{02},...) + \delta_0 \quad (2)$$

with P_0 being the purchase price of the property and δ_0 the difference due to the variables that do not influence the determination of Market Value but only the formulation of Price (i.e. the seller's economic situation).

Before continuing further, the term δ_0 should be explored in greater depth. Mathematically it should be considered a random variable. This statement can be justified by the definition that the appraisal discipline gives to Market Value: the most likely Transaction price. In fact, if δ_0 is a random variable, then the definition of Price translates into imposing that the expected value of that variable is null, or that:

$$E(\delta_0) = 0 \quad (3)$$

The link between Price and Market Value described in (2) allows us to explain (very loosely), on the one hand, the formation of the Transaction price, but it also allows us to find a way to estimate Market Value (the goal of any direct appraisal procedure).

To accomplish this mathematically, let us suppose that the 80 sm apartment has a market value of €180000. The economic needs of the seller could lead him/her to sell at a price of just € 150000. The difference of € 30000 would be one of the possible values of the random variable δ_0.

SIMPLIFICATIONS OF THE MARKET VALUE MODEL

To determine Market Value directly, the appraisal discipline requires that a sample survey of properties similar to the one being appraised be made. For each of the *n* properties in the sample, the estimated Transaction Price and the value of each of the characteristics that influence the formation of VM_0 will be identified by

$$P_i, x_{i1}, x_{i2}, \ldots. \forall i \in \{1, 2, \ldots, n\} \subseteq I_0 \tag{4}$$

where I_0 is the set of properties similar to the one being appraised.

Direct valuation methods use data collected in the sample survey to determine the estimate of Market Value. Some of these methods hypothesize the form of the function h_0 that tie the characteristics influencing VM_0 together. Other methods determine the estimate of the function through the use of sample data. However in both methods, it is necessary to introduce some simplifications to the Market Value model defined in equation (1).

First Simplification: Comparable assets

Who can help us define the function h_0 that can then allow us to determine VM_0 and how can this be done? Unfortunately, no one can define such a function precisely. To try to overcome this obstacle, the underlying principle of direct valuation methods, that of *comparison*, should be used. The mathematical translation of this postulate is the following:

$$VM_i = h(x_{i1}, x_{i2}, \ldots) \text{ per } \forall i \in \{0, 1, 2, \ldots, n\} \tag{5}$$

In practice, it is assumed that the Market Value of assets (here indicated by i) comparable to the one the being appraised is a function h of their characteristics. For comparable properties, it is assumed that the formation of their Market Value follows the same law that generates VM_0 (that is function h). Generally speaking, comparable assets are properties whose characteristics have values that are only partially different from the corresponding values recorded for the property being appraised. In other words, at least one characteristic j exists for which $x_{0j}=x_{ij}$ for any asset i, and they tend to have values that are mostly the same as those of the property being assessed. A more rigorous mathematical definition of similarity can be provided using the so-called similarity indices which will be discussed in the following paragraph.

In standard appraisal practice, the *geographic location* of the apartment is managed by using the concept of similarity. That is, only apartments located near the property being assessed and which have been sold recently are taken into consideration. Obviously this strategy implies that the concept of proximity be further clarified. In general, this clarification is based on the appraiser's knowledge of the marketplace in which the properties are being appraised. In even simpler terms, it is the appraiser who determines that the sample survey can include only those apartments that are located at a distance of less than 500 m from the property being appraised. There are, of course, more formal methods to define the concept of proximity. They refer to statistical methods that allow us to automatically obtain the distance that will help define the sample survey.

The recent development of technologies that enable georeferencing of property is encouraging the adoption of mathematical models that can explicitly manage a property's geographical characteristics. These models are alternatives to solutions that require taking into account only the apartments in the sample survey near the property being appraised. However, spatial models require a large amount of georeferenced data, appropriate statistical software and good knowledge of statistics - all of which are elements that do not favor the adoption of these methodologies in appraisal practice. Despite the promising results expected from the spatial models, in the following pages the geographical component of real estate will be managed by utilizing using the concept of proximity previously described.

Starting with the comparison principle, it is therefore possible to assume that there exists a single function h that can determine the Market Value of a group of similar assets, so you can reasonably redefine VM_0 as

$$VM_0 \equiv h(x_{01}, x_{02}, \ldots) \quad (6)$$

The introduction of this simplification allows us to reformulate the first question that we started with who can help us define the function *h* (and no longer h_0) that can allow us to determine VM_0 and how can they do this? This small modification will help us over come the obstacle. But to know how to do this we must wait a moment. First, in fact, we must understand the need for further simplification.

Second simplification: Unknown and unmeasurable characteristics

The mathematical correctness of the model introduced comes up against the difficult realities of appraisal practice. It is not at all taken for granted that the characteristics influencing Market Value are detectable or that is simple to identify all the characteristics that can affect VM_0. For example, it might happen that a characteristic such as "the distance from a school complex" is not considered influential in determining Market Value and therefore it is not identified. But this does not mean that this is true.

More often, it happens that the characteristic is considered influential on the Market Value but that is impossible (or too expensive) to gather the relevant data. For example, the "neighborhood safety" characteristic can be difficult to measure (unlike in the U.S.).

There might also be variables that are easy to identify (for example, the presence of a cellar) but information for some of the elements in the sample is not available. In this case, if the characteristic is not considered crucial in the formation of Market Value, not considering it can be contemplated.

Regardless of whether one or more characteristics are unknown, not detectable or not detected, from a mathematical point we need to adjust (5) and (6) which define the Market Value both of the property being appraised as well as that of comparable assets

$$VM_i = h_R\,(x_{i1}, x_{i2}, \ldots, x_{iJ}) + \eta_i \quad \forall\, i \in \{0, 1, 2, \ldots, n\} \quad (7)$$

In this new definition, Market Value is the sum of the value assumed by the restricted function h (h_R), which is assessed based only on the identified characteristics, and an error term (η_i).

Formula (7) expresses a very simple concept in mathematical form: if, in the definition of Market Value, we either forget, do not detect or identify some characteristics, then an error is introduced, generically indicated by the term η_i.

This error is negligible if the characteristics excluded, namely those with an index greater than J, are such that they do not significantly affect Market Value. In this case then

$$VM_i \approx h_R\,(x_{i1}, x_{i2}, \ldots, x_{iJ}) \quad \forall\, i \in \{0, 1, 2, \ldots, n\} \quad (8)$$

Sometimes, however, this approximation may not be reasonable; or the errors η_i may not be negligible. Unfortunately, in this case, mathematical tools cannot do very much. At best, we can try to replace the missing characteristics with others that are positively correlated to them. If this cannot be done, then the mathematical appraisal process should be terminated. But appraisal practice requires, in any case, the estimation of a Value and the error η_0 is compensated through non-mathematical methods.

Before moving ahead, let us sum up what we have obtained thus far. Assuming that we take into account a set of assets comparable to the one being appraised and select the J characteristics that influence the determination of VM_0 then Market Value is defined by (8) as a function of the identified values $x_{01}, x_{02}, \ldots, x_{0J}$. At this point, after having introduced two simplifications to the model defining Market Value and having gathered the data on the J characteristics of the assets being valued, we would think that the estimate of VM_0 is now terminated. Unfortunately it is not; on the contrary, up to now, we have simply highlighted the approximations and errors introduced into the estimation of Market Value. The h_R function, in fact, is not yet known (and alas ... it never will be!).

Third simplification: The form of the h_R function

In the introduction to this section, it was stated how direct valuation methods fall into two groups: those that hypothesize the form of the function h_R (referring, at the time, to h_0) and those that estimate h_R starting from the collected data samples. In both cases, the risk of error is extremely high. Assuming, for example, that the Market Value for the 80 sm apartment considered previously, and for the comparable apartments, depends on the Area characteristic in a linear way, meaning that

$$VM_i = ... + \beta_j \, Superficie_{ij} + \eta_i \quad \forall \, i \in \{0, 1, 2, ..., n\}$$

But this does not mean that that this assumption is wrong and that instead the true function of Market Value is the most complex

$$VM_i = ... + \beta_j \, Superficie_{ij} + \beta_{j+1} \, I(Superficie_{i(j+1)} > 75) + \eta_i \quad \forall \, i \in \{0, 1, 2, ..., n\}$$

We do not wish to go into the detail of the aforementioned expressions. They represent just one example of how, in the definition of the function of Market Value, there can be errors of form when assumptions are made.

An analogous observation can be made regarding those methods, such as regression, for which there are procedures (model selection) that can independently identify the form of the function that best fits the data samples collected.

Therefore, from the above, it is essential to make further corrections to (7) of Market Value

$$VM_i = f(x_{i1}, x_{i2}, ..., x_{iJ}) + \nu_i + \eta_i \quad \forall \, i \in \{0, 1, 2, ..., n\} \tag{9}$$

where ν_i is the error due to the incorrect definition of the form of the h_R function. Fortunately, (9) does not need to be simplified any further. It represents the mathematical model of Market Value from which to begin to estimate VM_0.

Before continuing, however, it is important to ask what effect the errors η_i and ν_i can have on the estimate of VM_0.

In this regard, from the mathematical point of view there can be three cases:

(a) the errors are null or neglibible: $\nu_i \approx 0$, $\eta_i \approx 0$;

(b) the errors are not negligible but they have a random pattern; in other words, they are random *Gaussian* variables with expected null value: $\nu_i \sim N(0, \sigma_\nu)$, $\eta_i \sim N(0, \sigma_\eta)$;

(c) the errors are not negligible and do not have a random pattern; in other words they are not distributed like *Gaussians.*

When the errors are type (a) or (b), the effects on the estimation of VM_0 are not generally negligible but can still be handled by several valuation methods. When errors are type (c), the situation is different; in this case the estimation procedure should be interrupted or at the very least, adjustments should be made to the results obtained. But in order to make adjustments, the magnitude of these errors should be established through the analysis of large samples (to assess the error of form specification) performed by surveying a large number of characteristics (to assess the error due to non-detected and/or non-detectable characteristics).

Unfortunately, the reality of appraisal practice is characterized by the small size of the samples and a certain difficulty in obtaining data relating to certain characteristics. Large studies are infrequent and, therefore, it is not very easy to ascertain whether η_i and ν_i errors are type (c).

Before proceeding, we must state very clearly that in order for the estimated Market Value of a property being valued (VM_0) to be "good," the errors due to the limited number of characteristics considered (η_i) and those dictated by the error of form (ν_i) must be type (a) or (b). If not, then all the appraisal procedures and associated errors would be affected by inaccuracies and the estimates obtained would not have any practical significance. Just as a great chef cannot cook a an important dish if the raw materials are of poor quality, so a theoretically accurate statistical or valuation method as cannot provide a good estimate in the presence of type (c) errors.

MODIFICATIONS TO THE TRANSACTION PRICE MODEL

It has already been said that to estimate the Market Value of a property, the Transaction Prices found in a sample survey should be used. To accomplish this task, regardless of the valuation method used, it is necessary to:

- extend the Transaction Price model (2) to all elements in the sample (this is absolutely legitimate and reasonable);
- replace in the extended model (2) the definition of Market Value (9) introduced as a result of the aforementioned simplifications;
- assume that the η_i and ν_i errors are type (a) or (b), in other words, assume that they are negligible or random.

The performance of these operations can construct the following algebraic expression:

$$P_i = VM_i + \delta_i = f(x_{i1}, x_{i2}, \ldots, x_{iJ}) + \nu_i + \eta_i + \delta_i \quad \forall\, i \in \{0, 1, 2, \ldots, n\} \tag{10}$$

With the assumption that the η_i and ν_i errors are type (a) or (b), then (10) can be further simplified in what we can define as the model of formation of the Transaction Price:

$$P_i = f(x_{i1}, x_{i2}, \ldots, x_{iJ}) + \varepsilon_i \quad \forall\, i \in \{0, 1, 2, \ldots, n\} \tag{11}$$

ensuring that the random variables ε_i (with expected value 0) are defined as the sum of errors η_i, ν_i and δ_i for any i between 0 and n. Next, we will describe the main methods of direct valuation of Market Value, or in other words, the algorithms that allow us to obtain a numerical value which we will indicate as

$$\hat{VM}_0 = \hat{f}(x_{01}, x_{02}, \ldots, x_{0J}) \tag{12}$$

with $\hat{f}$ being the estimate of the function f of Market Value.

At this point, it should be clarified that in valuation jargon, ***Market Value*** *means the appraisal opinion provided by the expert appraiser at the end of his/her consultancy/assessment*. Therefore, logically, we should have defined Market Value as the monetary amount indicated generally in (12). However, from the mathematical point of view, the adoption of this convention would have led to much confusion. *So it is more correct (mathematically speaking) to consider Market Value VM_0* as an unknown value and unknowable in any way.

And to define the monetary amounts resulting from the direct valuation $\hat{VM}_0$ procedure as the **Estimated Market Value**.

DIRECT VALUATION METHODS OF MARKET VALUE

The Transaction Price model (11) and a survey are necessary tools for any one of the direct methods for estimating Market Value. Below, we will consider some of these methods and attempt to describe their mathematical details and the difficulties in their application taking our cue from pseudo-real case studies. First we will consider an especially large survey sample and provide an estimate of the Market Value of the property through the statistical method known as Regression. Later, we will apply the so-called Sales Comparison Approach (SCA) to estimate Market Value in the case of a small survey sample.

Finally, we will assess the practical utility of a valuation method known as the Nearest Neighbors Appraisal Technique (NNAT) and the conditions under which it can be applied. The mathematical description and comparison of the Regression and SCA methods will follow the scheme and language introduced by Isakson (2002) in an article entitled “The Linear Algebra of the Sales Comparison Approach.” It stresses how the regression is actually a class of methods in continuous expansion and revision. The availability of statistical software and the construction of ever wider and more comprehensive databases of transactions suggest that in the coming years, this class of methods will be increasingly used in the valuation field.

To get a clearer and more complete picture of Regression methods, see the text by Hastie et. al. (2001) entitled “The Element of Statistical Learning. Data Mining, Inference and Prediction.” In the same text, it is also possible to find the theoretical foundations of the method referred to as NNTA.

MULTIPLE LINEAR REGRESSIONS

In the valuation field, an important role (especially outside Italy) is played by so-called mass appraisal. These appraisals are used primarily in the fiscal domain, allowing an assessment of the market value of a large number of goods.

There are many mathematical and statistical methods that can be used in mass appraisal. Among these methods, multiple linear regression is one of the best known. Although this method is not necessarily the most suitable in many practical cases, it is certainly the easiest to understand and apply. In particular, regression is a well-known statistical method that is now available in a wide variety of software. In the following paragraphs, a case study is presented and, through it, we will describe the multiple linear regression models.

The Survey Sample

In mass appraisal, the survey sample is made up of a large number of recently sold goods for which a certain number of characteristics will be surveyed. The multiple linear regression method can be used if the sample consists of assets located in the same vicinity.

The sample analyzed below consists of a database of 52 recent transactions for which we have information relating to the following characteristics: Price, Area, Exposures, Balcony, Cellar, Floor, Elevator, Year of Construction and Conditions of the housing unit.

The transactions took place in a specific geographical area characterized by low-cost housing construction. It is assumed that the characteristics analyzed are those that affect the determination of the market value of properties bought and sold in the area.

This hypothesis is certainly difficult to accept. It is more reasonable and correct to think that other characteristics also influence the formation of Market Value. Therefore, for each property considered the error component η_i (7) cannot be considered null. However, it appears equally reasonable to assume that the η_i errors are type (b), or in other words not negligible but random $\eta_i \sim N(0,\sigma_\eta)$.

The data collected for the 52 elements in the survey sample are organized in the following table (as an excerpt)

Table 1 Excerpt pf the data base of the 52 transcations that make up the sample survey

Price	Area	N° Exposu-res	Balconies	Cellar	Floor	Elevator	Year	Conditions
79800	45	2	1	present	0	absent	1930	Not renovated
65100	45	2	1	absent	0	absent	1950	Not renovated
••••	••	•	•	••••	•	••••	••••	••••
76800	55	2	1	present	3	absent	1930	Renovated-new

Table 1 shows how certain characteristics (Price, Area, Exposures, Balcony, Piano and Year of Construction) are quantitative. Others (Cellar, Elevator and Conditions), however, are qualitative characteristics. For the latter, a numerical codification must be performed before applying the regression formula (as well as the Sales Comparison Approach). The Cellar and Elevator characteristics will be coded using the number 0 (1) when it is Absent (Present). Similarly, the Conditions characteristic will be coded by substituting the number 0 (1) if the unit is Not Renovated (Renovated-New).

We will now present the equations at the heart of multiple linear regression. At the same time we will try to facilitate their understanding by inputting the values contained in the database excerpt (Table 1).

Hypothesis of model linearity

Taking the Transaction Price model in the version described in (11) and introducing the hypothesis of linearity of the characteristics' parameters, we obtain the following

$$P_i = x_{i1}\beta_1 + x_{i2}\,\beta_2 + \ldots + x_{iJ}\,\beta_J + \varepsilon_i \quad \forall\, i \in \{1, 2, \ldots, n\} \tag{13}$$

This model is known in statistics as a Model of Multiple Linear Regression. Price is therefore a function of a sum of terms. The first J terms are due to the characteristics considered to be potentially influential in price formation. The last term represents the error due to the components described in (11). The system of n equations (13) is linear and therefore can be synthetically represented by the following matrix

$$P = X\beta + \varepsilon \tag{14}$$

where P, β and ε are respectively the price vectors, parameters to be estimated and errors

$$P = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix} \in M_{(nx1)} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_J \end{bmatrix} \in M_{(Jx1)} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in M_{(nx1)} \tag{15}$$

while X is the matrix of the characteristics

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1J} \\ x_{21} & x_{22} & \cdots & x_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nJ} \end{bmatrix} \in M_{(nxJ)} \qquad (16)$$

Substituting the values collected in the sample (Table 1) in the price vector and in the characteristic matrix we obtain

$$P = \begin{bmatrix} 79800 \\ 65100 \\ \vdots \\ 76800 \end{bmatrix} \in M_{(52x1)} \qquad X = \begin{bmatrix} 1 & 45 & \cdots & 0 \\ 1 & 45 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 55 & \cdots & 1 \end{bmatrix} \in M_{(52x9)}$$

The first column in matrix X contains a vector of unit values, or 1s; this ensures that the so-called "intercept" is estimated (we recall that the intercept is the intersection of the straight line with the y-axis). The second column refers to the Area characteristic values for individual elements in the sample. The last column in the matrix is obtained by transforming the Conditions characteristic according to the codification previously introduced.

Returning to the matrix in (14), the problem lies in estimating the vector of the β parameters. This vector can be calculated using several statistical methods.

However, the simplest and most well-known is called the Least Squares method. In this method, in fact, the parameters are estimated by trying to minimize the squared residuals, or trying to minimize the quantity $\varepsilon' \varepsilon$.

Least squares

The goal of this paper is not to describe the least squares method for calculating model parameters (14). Please refer to Hastie et. al. (2001) or other statistical texts for an understanding of how the following is derived:

$$\hat{\beta}=(X'X)^{-1}X'P \qquad (17)$$

(17) yields the estimate of the parameter vector using only the data in our possession, the characteristics matrix and the Price vector.

It is important to point out that $\hat{\beta}$ is an estimate of the parameter vector β so that it generates an error that is managed explicitly through the use of confidence intervals (which will be discussed below).

Substituting in (17) the values collected in the survey, the vector containing the unknown parameters of the model (13) can be obtained

$$\hat{\beta} = \begin{bmatrix} -863649.3 \\ 1071.8 \\ -6775.1 \\ 6403.4 \\ -3078.4 \\ -642.6 \\ 6906.0 \\ 459.1 \\ 18008.0 \end{bmatrix} \in M_{(9x1)}$$

The vector of the estimates of the Parameters obtained contains nine numerical values to which an appraisal significance must be assigned; omitting the first term (the intercept) which mainly has mathematical significance. Let us focus now on the second value of 1071.8; this refers to the Area characteristic and represents the increase in Market Value of an asset for each additional square meter. Please note this is not Unit Price, which, in fact, is obtained by dividing Price by Area. The significance of the Area parameter is different. Mathematically, it represents a first derivative, while from the valuation point of view it is the difference in value between two completely identical properties except for the fact that one of them has a larger area expressed in square meters.

Continuing with the analysis of the values in the vector of the parameter estimates, we find -6775.1. This result has mathematical significance, but certainly cannot have significance for appraisal. In fact, it represents the increase in Market Value of a property resulting in an additional exposure. It would be like saying that if there are two apartments, one with only one exposure and the other with two, then the second is worth less because it has two exposures! This is clearly a contradiction. But we should not be too surprised at this apparent contradiction. The reason will be clarified in the next section when we highlight how the practical application of regression analysis requires a great deal of experience and the adoption of methods called “model selection”.

Continuing to look through the values contained in the estimate vector we find that an additional Balcony causes Market Value to increase by € 6403.40, while the presence of a cellar leads it to decrease in value of € 3078.40 (in this case, this is obviously a result that is has no significance in terms of appraisal). Interestingly, the numbers obtained show that the apartments on upper floors tend to have a lower market value, since the parameter estimate is negative: € -642.60. Though apparently contrary to common sense, this can have its foundation in the fact that in some cases the buildings in which the sample units are located have no elevators. The presence of the elevator contributes to an increase in value (€ +6906.00) but perhaps it would be more appropriate to study the combined effects (interaction) of the Floor and Elevator characteristics.

Towards the bottom of the estimate vector is the value 459.10, referring to the Year of Construction characteristic. So if there are two units, the first built in 1940 and the second in 1950, the difference in Market Value due to the year of construction is equal to 459.10 * (1950-1940) = € 4591. Finally, the last element in the estimate vector indicates that the difference in value between a renovated and a non-renovated apartment is equal to € 18008.00.

Despite the fact that the mathematical solution (17) to the problem (14) is (almost) always achievable, it may happen that the result has no value in terms of appraisal. In our case study, we have already seen how the estimates obtained for the parameters of the Exposure and Cellar characteristics were found to be unreasonable. This might suggest that regression is not a suitable method for analyzing data from a sample survey. In reality, however, equation (17) obtained through the least squares method is only the first step in a long and often complex process. While reiterating that it is not the aim of this paper to explore the topic of regression in detail, the following sections will present the results obtained by performing a methodical analysis of the data in Table 1.

Model Selection

Thus far, it has been assumed that all the characteristics collected in the survey relate to the estimated Market Value. This is as reasonable as assuming that there are other characteristics that influence Market Value. However we must remember that the collected data are collected from a sample and not the entire population of transactions. It follows that sample data can only have limited explanatory power; or in other words, that there are difficulties in correctly identifying all the characteristics that are significantly influential and in correctly assessing the different levels of individual influences. In other words again, due to the limited number of elements in the sample and the variability of Prices, a regression can identify and assess the influence of only a subset of characteristics that truly influence the formation of Market Value.

In the statistics literature, there are different approaches to so-called model selection. Among the most widely used is the so-called "stepwise regression". This procedure requires the construction of an initial model containing the characteristics that are assumed to influence Market Value and for which data is available. In successive steps, the procedure gradually eliminates the characteristics that data analysis shows to be not statistically significant in the formation of Market Value. At the end of the procedure, we obtain a final model consisting of a subset of characteristics and the resulting parameter estimates. In the case study considered, the procedure was applied starting from an initial model made up of the eight characteristics in Table 1 plus the interaction between the Floor and Elevetor characteristics. The "winning" model from the selection process is specified in Table 2.

Table 2 Winning model for the stepwise regression procedure papplied to the sample

Characteristic	Estimate	Std.Error	t-value	Pr(>|t|)
(Intercept)	-829802.12	223256.58	-3.717	0.000556 ***
Area	1135.86	254.53	4.463	5.37e-05 ***
Floor	-1289.04	1031.02	-1.250	0.217671
Elevator	70.34	6085.52	0.012	0.990829
Conditions	18022.50	3792.45	4.752	2.09e-05 ***
Year	438.22	116.79	3.752	0.000500 ***
Floor: Elevator	2637.16	1738.06	1.517	0.136186

To quickly read the results in Table 2, one should look at the last column, denominated Pr (> |t |). This column contains the so-called p-values, or numbers that indicate the presence of statistical significance when their values are less than 0.05. In this case there are four rows where the p-values are below 0.05. They can be easily identified by looking for the symbol ***. So it turns out that intercept, Area, Year of construction and Conditions are the characteristics that are statistically significant in the formation of Market Value. This statement must be understood correctly. It does not mean that only these characteristics influence the formation of Market Value. It means rather that the regression analysis of the data shows that these characteristics affect the value with a high confidence level. Obviously there is a possibility that the other characteristics considered in the initial model could affect the Market Value. The data, however, does not provide sufficient indications to consider that these characteristics (eg Exposure, Cellar, Balcony, ...) will be influential.

After having identified the characteristics that statistically influence the formation of Market Value, it is necessary to assess their influence. From the observation of the values in Table 2, we see that the increase in value due to an additional sm is equal to €1135.86. Similarly, when two units are very similar except for the fact that one is not renovated and the other is renovated/new, the difference in their estimated values is € 18022.50. The so-called "marginal price" for the Year of construction characteristic is equal to € 438.22 per year. For non-influential characteristics (Floor, Elevator and their interaction) the estimated values obtained are not significant and therefore should not be taken into consideration.

Before continuing the regression analysis, it should be noted that the parameter estimates in Table 2 were obtained through the analysis of a sample. This means that estimates are affected by an element of uncertainty due to the sampling procedure. If a different sample were chosen, the same estimates would not have been obtained. Values close to those reported in Table 2 would have been obtained but they would not have been exactly the same. To manage this component of uncertainty which is due to the sampling procedure, and cannot be eliminated, we must define the "confidence intervals" for the parameter estimates.

The parameter estimate for Area is € 1135.86 / sm. An indication of the uncertainty of the estimation is shown in the Standard Error column: € 254.53 / sm. Through these two values, it is possible to construct a confidence interval (95% confidence) by subtracting and adding twice the estimate of Standard Error: $1135.86 \pm 2 \cdot 254.53 = (626.80; 1644.92)$.

This formula, while not exact (standard error should be multiplied by a coefficient dependent on Student-t distribution), allows us to state that: with a confidence level of 95%, the true value of the Area parameter lies between a minimum of approximately € 627 and a maximum of approximately € 1645. From a practical standpoint, this is a very wide range that can only be reduced by increasing the sample size.

In any case, the indication of a confidence interval instead of a single point value shows, with immediacy, the uncertainty inherent in the appraisal due to the variability of prices and the type of sampling.

The point estimates (and intervals) of the parameters of the significant characteristics of the winning model can be very useful for the Sales Comparison Approach.

Testing the model

The appraisal procedure based on the least squares method and the construction of confidence intervals are based on assumptions about the vector of errors ε.

These procedures require, in fact, that the components of the vector be independent, have an expected value of zero, constant variance and be normally distributed.

If these assumptions are not met, then all results would be questionable.

But how to verify the reasonableness of such assumptions if we cannot observe the errors directly? In fact, vector ε is unknowable! If we look at (14), we see that in order to determine ε, we should know the value of β.

In reality, however, we can only obtain an estimate of the vector of the parameters, that is $\hat{\beta}$.

It follows that to verify the validity of the assumptions regarding the random vector ε, it is necessary to analyze the behavior of its realization called the vector of residuals

$$\hat{\varepsilon} = P - X\hat{\beta} \tag{18}$$

Once the vector of residuals has been found, it is necessary to ask which tool is best suited for testing the assumptions made on errors. The most frequent answer to this question in the valuation field is the so-called coefficient of multiple determination, known as R^2.

This coefficient can have values in the interval [0,1] and represents the proportion of variability explained by the model (14) compared to the natural variability present in the price trend P.

The closer the value of R^2 is to 1, the greater the proportion of the variability of prices. In the case study under consideration, the value of the linear coefficient of determination is equal to 0.755, a value tending towards unity but not too close to it.

Can we think of testing the validity of the hypotheses regarding the residuals by means of this numerical value?

The answer is definitely not! However, it is not unusual to see texts in the valuation field that reduce the analysis of the goodness-of-fit of the model only to the observation of the value of R^2.

Regarding this tendency, it should be stressed that the tools for testing the assumptions about errors are mostly graphic; and that the value of R^2 can only assess the degree of variability explained by the model. The graphs for the study of validity of the hypothesis of the model are varied and not always easy to read. Yet they are indispensable tools for maintaining that a regression analysis is valid.
The residuals obtained for the data in the case study in question are represented in

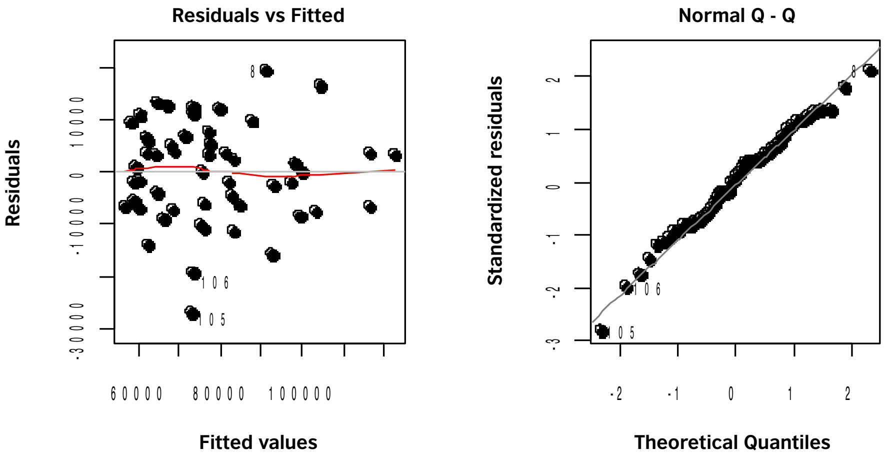

Illustration 1 Residual plots.
Left distribution of the residuals compared to the values predicted by the model.
Right graph of normal probability**.**

The pattern of the residuals in relation to the expected values using the parameter estimates (plot on the left in Illustration 1) is characterized by a random distribution of the data (without the presence of periodicity), by an average trend (red line) close to zero and by constant variability of the data compared to expected values. These considerations allow us to state that there is no reason to reject the hypothesis that the errors are independent, with expected value null and constant variance. The normal probability plot (on the right in Illustration 1) is the representation of the cumulative distribution of residuals (circles) and theoretical cumulative distribution of a Gaussian random variable (line segment). The normality assumption of errors (at the base of the construction of confidence intervals) cannot be refuted if the circles overlap with the line segment. Noting the pattern obtained for the case study, the normality assumptions can also be accepted with a certain degree of reasonableness.

From the aforementioned considerations, it is clear that the winning model resulting from the selection procedure satisfies the assumptions regarding errors. This is therefore a statistically correct model. But is it also a useful model? This can be answered by observing how R^2 takes on a definitely high value (0.755), but not too high. In fact about 25% (1-0.755) of the price variability is unexplained by the model. This percentage is due to all the causes and simplifications listed in the previous pages. In appraisal terms, the model obtained may prove useful to have an estimate of so-called “marginal prices” (the elements of the vector of the estimates). It can be considered partially useful for making general appraisals, although it must be said that the model could be greatly improved by increasing the sample size and the number of detected characteristics (in particular those relating to the spatial component). Clearly, the estimated model is not a suitable tool to perform the appraisal of a single property. To achieve this goal, the SCA must be used.

ESTIMATED MARKET VALUE

In the valuation world, Multiple Linear Regression is considered a method for so-called mass appraisals; or in other words, for those appraisals that require the formulation of value judgments on a large number of assets simultaneously. The mathematical operation to be performed, after obtaining the estimation of the vector parameter, is in fact very simple,

$$\hat{VM}_0 = \hat{f}\,(x_{01}, x_{02}, \dots, x_{0J}) = X_0\hat{\beta} \tag{19}$$

where

$$X_0 = [x_{01}\ x_{02}, \dots, x_{0J}] \in M_{(1\times J)} \tag{20}$$

In practical terms, the estimated market value of an property is obtained by multiplying the vector X_0 by the vector of the estimates of the parameters.

Vector X_0 is made up of the values of the characteristics observed for the property being appraised. In the case of mass appraisals, vector X_0 becomes a matrix with a number of rows equal to the number of properties to be assessed, so that we do not obtain a single estimate of Market Value but a vector.

Recalling the case study, if we want to appraise a non-renovated property with an area of 50 square meters in a building built in 1950, it would result in

$$\hat{VM}_0 = X_0\hat{\beta} = [1\ 50\ 0\ 1950] \begin{bmatrix} -829802.12 \\ 1135.86 \\ 18922.50 \\ 438.22 \end{bmatrix} = 81512.3$$

As previously noted, the description of the multiple linear regression method given here is not intended to be exhaustive. This presentation has two basic goals.

The first is to point out the close mathematical similarity between multiple linear regression and SCA. And the second is to highlight how the results obtained from the regression procedure can be a starting point for the practical application of SCA.

SALES COMPARISON APPROACH

The questions that require the formulation of an opinion regarding the estimated Market Value of a single asset are the norm in appraisal practice.

These estimates are used in different contexts and often serve to resolve legal disputes.

Among the many estimating methods, both direct and indirect, a central role is played by the Sales Comparison Approach. In the following paragraphs, we will focus on the mathematical justification of the method and the connections between SCA and multiple linear regression.

All of this starting from a case study.

The survey sample

An estimation problem requires the formulation of an opinion on the Market Value of a property of 45 square meters, with two exposures, a balcony, cellar, on the ground floor, not renovated, located in a 1950s building without an elevator (Table 3).

Table 3. Surveyed characteristics of subject property and the sample survey

Price	Area	Exp.	Balconies	Cellar	Floor	Elevator	Year	Conditions
?????	45	2	1	present	0	absent	1950	Not renovated
68300	38	2	2	present	1	absent	1940	Not renovated
91300	60	2	1	present	5	present	1950	Not renovated
78400	40	2	1	absent	1	present	1950	Not renovated
53900	45	2	1	present	2	absent	1950	Not renovated

To provide a direct estimate through SCA, it is necessary to select a sample consisting of four units comparable to the property to be assessed (Table 3) located in the same geographical area. The property and the elements in the sample are in the same area as the units considered in Table 1, so that the use of the results obtained from the regression can be justified.

As easily evidenced by Table 3, the property being appraised and the sample elements have in common number of exposures and conditions. Therefore, these two characteristics can be eliminated because they will not be able to explain any change in value. Then proceeding with the the encoding of qualitative variables and recalling the definitions introduced thus far,

$$P = \begin{bmatrix} 68300 \\ 91300 \\ 78400 \\ 53900 \end{bmatrix} \in M_{(4\times1)} \qquad X = \begin{bmatrix} 38 & 2 & 1 & 1 & 0 & 1940 \\ 60 & 1 & 1 & 5 & 5 & 1950 \\ 40 & 1 & 0 & 1 & 1 & 1950 \\ 45 & 1 & 1 & 2 & 0 & 1950 \end{bmatrix} \in M_{(4\times6)}$$

$$X_0 = [45\ 1\ 1\ 0\ 0\ 1950] \in M_{(1x6)}$$

Note that the penultimate column in X is obtained by multiplying the number of Floors by the column of the coded variable Elevator; this is to emphasize the fact that the presence of the elevator affects the value based mainly on the floor on which the unit is located (interaction between the two characteristics).

Starting from the general equation (11) of the Transaction Price model, the SCA develops a strategy to correct the price vector P starting from the existing differences between elements in the sample and the subject property $X\text{-}IX_0$.

Hypothesis of model linearity

The mathematical basis of SCA is the famous Taylor series, a topic well known in Mathematical Analysis. A continuous real function can be well approximated by a linear model around a specific point. So if we want to approximate a linear function of Market Value f in the neighbourhood of the point represented by vector X_0, we obtain

$$f(X_i) = f(X_0) + (X_i - X_0)\, f'(X_0) + \rho_i \quad \forall i \in \{1, 2, \ldots, n\} \tag{21}$$

The Market Value corresponding to property i in the survey sample is a sum of three terms: the Market Value (unknown) of the subject property, the matrix product between the vector of the variations in the characteristics and the first derivative f at point X_0, a term containing the approximation error that occurs if higher order terms are not considered.

To transform (21) into matrix form, it is sufficient to introduce the column vector I of dimension n

$$f(X) = I\, f(X_0) + (X - I\, X_0)\, \beta + \rho \tag{22}$$

Note that in this case, the vector of the parameters has the same significance as the one introduced in the regression (14). It is therefore a vector that contains the price variations corresponding to a unit change in characteristics.

The mathematical limit of the mathematic approximation lies in the fact that it applies only in the neighborhood of point X_0. Outside this neighbourhood, the approximation error ρ may not be negligible. It is all about understanding what the neighbourhood is for which the error is negligible. And this is precisely the weak point of the SCA procedure from a practical point of view. There are no tools to verify if the elements in the sample survey belong to a neighborhood in which the approximation error can be considered negligible.

Valuation model

Through some algebraic operations and recalling the relationship (11) between the vector of Prices and the function of Market Value, we can reach the mathematical equation of the SCA (result obtained by adjusting the equations contained in Isakson's article 2002)

$$\begin{aligned} I\, f(X_0) &= f(X) - (X - I\, X_0)\, \beta - \rho \\ &= P - \varepsilon - (X - I\, X_0)\, \beta - \rho \\ &= P + (I\, X_0 - X)\, \beta - (\rho + \varepsilon) \end{aligned} \tag{23}$$

with $I = [1, 1, .., 1]' \in M_{(nx1)}$. The linear system (23) consists of n linear equations, each having as a result the Market Value of the property being valued. In fact, $If(X_0)$ is a column vector of dimension n whose elements are all equal to Market Value $f(X_0)$.

Equation (23) allows us to factor the Market Value into two components: the so-called vector of adjusted prices and the vector of errors.

The adjusted prices S are therefore defined by

$$S = P + (I\, X_0 - X)\tilde{\beta} \tag{24}$$

In other words, they are the prices recorded for the elements in the sample to which an adjustment is made due to differences in the values of the characteristics. Recalling the values of the characteristics found in the sample survey and those of the subject property we have

$$(I\,X_0 - X) = \begin{bmatrix} 7 & -1 & 0 & -1 & 0 & 10 \\ -15 & 0 & 0 & -5 & -5 & 0 \\ 5 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 \end{bmatrix}$$

But the meaning of the newly introduced vector $\tilde{\beta}$ is still not clear. Why did we not use vector β indicated in equation (23)?

Vector β has as elements the first derivative of function f evaluated at point X_0 (known as marginal prices). The problem is that the function f is unknown along with its partial derivatives with respect to the characteristics. But then how to define the components of β? In practice, these values are defined by the appraiser based on his/her experience and knowledge of the local real estate market. However, the values of the components of β can be obtained through data analysis, for example by means of regression analysis or so-called “Paired Analysis”. Whichever road is travelled, expert opinion or analytical methods, the result is that vector β can only be approximated by a vector that has been named $\tilde{\beta}$.

In this case study, the results of the regression performed on the 52 transactions considered previously can be used. Thus, the marginal prices for the Area, Conditions and Year of construction characteristics can be defined respectively: € 1135.86, € 18922.50, and € 438.22. Of these values, the one for Conditions is useless, since this characteristic is the same for all elements in the sample and for subject property. We presume that the marginal prices of the other characteristics are the result of the experience of the appraiser so that

$$\tilde{\beta} = \begin{bmatrix} 1135.86 \\ 3180 \\ 640 \\ -2240 \\ 3760 \\ 438.22 \end{bmatrix} \longrightarrow S = \begin{bmatrix} 79693 \\ 66662 \\ 83199 \\ 58380 \end{bmatrix}$$

Before continuing the study of the mathematical foundations of the SCA and moving into the so-called phase of data reconciliation, we need to refer to (23) substituting in it the definition of a vector of adjusted prices (24).

The result is the appearance of a new component of error due to the approximation introduced replacing the unknown value of β with $\tilde{\beta}$, defining $\tilde{\beta} = \beta + \theta$, with $\theta \in M_{(Jx1)}$, then

$$I\, f(X_0) = S - ((I\,X_0 - X)\,\theta+\rho+\varepsilon) \;\mapsto\; S - I\, f\,(X_0) = (I\,X_0 - X)\,\theta+\rho+\varepsilon \qquad (25)$$

In summary, (25) states that the difference between the adjusted prices obtained through each element in the sample and VM_0 is due to three components of error:

- due to the use of marginal prices estimated or chosen by experts (θ);
- due to linear approximation in the neighborhood of X_0, and which therefore takes into account any non-negligible effect of nonlinearity (ρ);
- due to the variability of prices compared to market values (ε).

Estimating Market Value

Once the vector of the so-called adjusted prices has been obtained from a mathematical point of view, the so-called reconciliation is simply a matrix multiplication

$$\hat{VM}_0 = W'S \tag{26}$$

between the vector S and the column vector of weights $W = [w_1, w_2, .., w_n]' \in M_{(nx1)}$, that is a vector whose sum of elements is equal to 1 or, equivalently, I'W=1. In this way, an estimate of Market Value of the subject property analyzed thus far in the case study can be obtained by using a uniform weight vector (that is by assigning the same weight to all elements in the sample)

$$\hat{VM}_0 = W'\,S = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 79693 \\ 66662 \\ 83199 \\ 58380 \end{bmatrix} = 71983.65$$

Therefore, the application of the SCA to the case study produces a Market Value of the property of approximately € 71980.

Once the numerical result of the estimate has been obtained, the question arises: is this really meaningful? Certainly, an expert in the local real estate market might give a sensible answer to this question in a few moments based on his/her experience and knowledge.

However observing the question from a mathematical-statistical point of view, the question can be explored further by addressing the following issues:

- is there a way to evaluate the goodness-of–fit of the estimate obtained (as with multiple linear regression) or even to measure of its level of uncertainty with it? Note in this regard that in the valuation field, SCA is a deterministic method ... but uncertainty remains in any case as evidenced by the components of error in (25).
- is it possible to study the effects generated by the error θ due to the use of the vector $\tilde{\beta}$?
- is there a way to try to identify the neighborhood of point X_0 in order to limit the non-negligible effects that the non-linearity components (ρ) may have on the estimate result?
- is there a method to determine the vector of weights W so that the estimate is as correct as possible, or, in other words, as close to the unknown true value of Market Value?

We will seek answers to these questions through the use of graphs, distance measurement and simulations. By contrast, current SCA practice states that the question regarding the significance of the results is answered based on indices that seek to monitor the variability present in the vector S.

For example, if we look at the components of the vector of adjusted prices, the difference between the maximum and the minimum is

$$\max(S) - \min(S) = 83199 - 58380 = 24819$$

This figure was 34.5% of the Estimated Market Value. Such a great variation immediately calls for the rejection of the result. Therefore from an appraisal point of view, the problem is solved by looking for a different sample survey, or reviewing the vector of weights, etc.

The fact that the SCA procedure does not always allow to obtain an estimate of Market Value brings up another mathematical question: under what conditions and how often does the SCA not provide acceptable results?

Paired Analysis

Before seeking answers to the questions raised in the previous section, a few words should be spent to understand so-called *Paired Analysis* (PA) from a mathematical point of view. It has already been mentioned that this method is used to provide an evaluation for the vector elements $\tilde{\beta}$. It was not mentioned however, that PA is just a specific case of the SCA and, more precisely of (24).

PA can be used if there are two elements in the sample that have identical values for all characteristics except one. Unfortunately, in the case study in question, the sample does not have two elements that satisfy this condition (see Table 3). Therefore, we assume that we have two units, similar in all respects (for example, because they are located in the same building) except that one is on the second floor and the other on the third. The price for the first was € 84000, and for the second € 88000. One wonders what change in value is due to the interaction between the Elevator and Floor characteristics. Without thinking too much, one would say € 4000, because that amount is the difference in price between the third floor unit and the one on the second floor. Well, this is exactly the reasoning behind PA, which finds its justification in (24) where the role of the property being valued is played by the unit called *r*

$$p_r = p_s + (x_{rj} - x_{sj})\,\tilde{\beta}_j \quad \mapsto \quad \tilde{\beta}_j = \frac{p_r - p_s}{x_{rj} - x_{sj}} \tag{27}$$

Proposals for testing the goodness-of-fit of the SCA

The SCA is a deterministic appraisal method which is based on the linear approximation of the Market Value function in the neighborhood of point X_0. It is therefore necessary to verify that:

- the linear approximation is reasonable;
- the values assigned to the vector $\tilde{\beta}$ components are sensible;
- the elements of the sample survey are in the neighborhood of X_0.

To try to verify the validity of these assumptions, we first propose the use of **scatter charts**.

Plotting on a Cartesian plane the Transaction prices of the properties in the survey on the y axis and the corresponding values of the characteristics on the x axis, we obtain as many scatter diagrams as there are characteristics used in the SCA (Illustration 2).

Two curves are added to the diagrams thus obtained; the first (dashed) is obtained by nonparametric methods of approximation. The second (red segment) is a straight line whose slope is the marginal price of the characteristic which passes through the estimated Market Value (red square). By comparing the patterns of the two curves, it is possible to make some considerations regarding the assumption that the linear approximation is reasonable. The more the two curves overlap, the more the assumption is acceptable. Conversely, if the two curves follow different patterns then the linear approximation can be considered weak and, consequently, the error θ may not be as negligible as was believed. Observation of the scatter charts shows that in the case study the linear approximation introduced for the Floor characteristic is unacceptable. In this case, from the sample data it seems that Price increases on a quadratic trend as the Floor grows higher. However, the straight line plotted using the marginal price of the Floor characteristic indicates that the price should decrease linearly. The assumption of linearity is therefore not reasonable in this case. The selected value of the marginal price is also questionable.

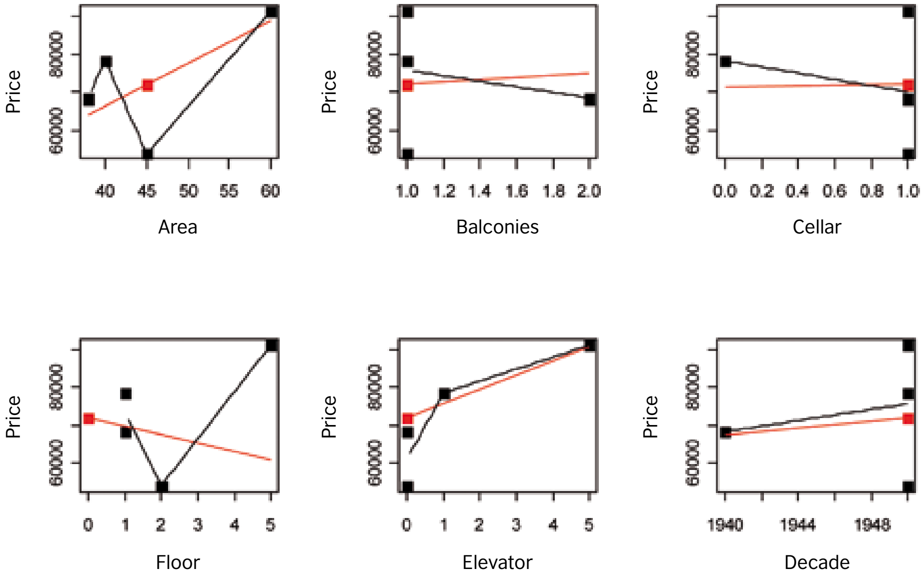

Figure 2 Scatter Charts of Price in relation to the characteristics used in the SCA. The black squares are the values in the sample survey. The dashed curve is the nonparametric estimate of the price trend in relation to the characteristic considered. The red square represents the estimated Market Value of the property being valued. The red curve is obtained by using the marginal price of the characteristic analyzed

The scatter charts described above are very useful to test the goodness-of-fit of the linear approximation and the sensibleness of the components of $\tilde{\beta}$. However, there are not very helpful in verifying that the elements in the sample are in the neighborhood of point X_0. To try to do this, the construction of a matrix of distances between the property being valued and the elements in the sample survey is proposed. This matrix must contain numbers that can express the distance between one property and another in relation to the differences due to different values of the characteristics taken into consideration. To construct this matrix, we must begin, first of all, from the following definition

$$H=\begin{bmatrix} 0_{1xJ} \\ I\,X_0 - X \end{bmatrix} \in M_{(n+1)xJ} \qquad (28)$$

The matrix H consists of *n* +1 rows representing the differences between the values of the characteristics of the property being valued, and those of the elements in the sample survey. The first row consists of J null values because it refers to the very property being valued. In the case study we analyzed we obtain

$$H=\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 7 & -1 & 0 & -1 & 0 & 10 \\ -15 & 0 & 0 & -5 & -5 & 0 \\ 5 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 \end{bmatrix}$$

The matrix H is formed by J columns that refer to the respective characteristics used in the estimate of Market Value. The numerical values contained in the different columns are expressed in different units of measurement. For example, the first column contains the differences in existing sm between the property to be appraised and the elements in the sample, while the second expresses a difference in the number of balconies. From the above, it is therefore impossible to calculate a distance matrix between the rows in H without prior normalization or some other adjustment. Another observation should be made regarding the weights of each characteristic in the estimation of VM_0. The fact that elements of the vector of marginal prices are not all the same shows how the columns in the matrix I X_0-X influence the estimate of Market Value in different ways. Based on this consideration it is suggested to calculate the distance matrix starting from

$$H \text{ diag } (\tilde{\beta}) \in M_{(n+1)\times J} \tag{29}$$

This new matrix is composed of columns, all having the same unit of measurement: the Euro! The values of this matrix are the corrections (or adjustments) due to the different properties and different characteristics. In fact, diag ($\tilde{\beta}$) is a square matrix with all elements null except those on the main diagonal which are equal to the marginal price

$$\text{diag } (\tilde{\beta}) = \begin{bmatrix} 1135.86 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3180 & 0 & 0 & 0 & 0 \\ 0 & 0 & 640 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2240 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3370 & 0 \\ 0 & 0 & 0 & 0 & 0 & 438.22 \end{bmatrix}$$

then

$$H\ \mathrm{diag}\ (\tilde{\beta}) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 7951 & -3180 & 0 & 2240 & 0 & 4382 \\ -17038 & 0 & 0 & 11200 & -18800 & 0 \\ 5679 & 0 & 640 & 2240 & -3760 & 0 \\ 0 & 0 & 0 & 4480 & 0 & 0 \end{bmatrix}$$

The matrix thus obtained allows us to state that, because of the difference in Area between the property being valued and the first element in the sample (second line), the price adjustment amounted to € 7951. While the fact that the first element in the sample survey has two balconies, one more than the property being valued, implies that its selling price has to be devalued by an amount equal to € 3180. The matrix thus obtained is composed of elements measured with the same unit of measurement (€) and the columns are properly weighed. It follows that this is an ideal candidate with which to create the distance matrix

$$D = \mathrm{dist}\ (H\ \mathrm{diag}\ (\tilde{\beta})) \in M_{(n+1)x(n+1)} \tag{30}$$

The matrix D is obtained by applying the Euclidean distance between the pairs of row vectors of H diag ($\tilde{\beta}$). The elements in this matrix measure the distance between the property being appraised and the elements in the sample in €. So if we want to have an indication of the distance between the property being appraised and the elements in the sample, it is sufficient to observe the values of the first row (or column).

$$D = \begin{bmatrix} 0 & 9877 & 27734 & 7199 & 4480 \\ 9877 & 0 & 32977 & 7002 & 9877 \\ 27734 & 32977 & 0 & 28687 & 26247 \\ 7199 & 7002 & 28687 & 0 & 7199 \\ 4480 & 9877 & 26247 & 7199 & 0 \end{bmatrix}$$

The distances between the property being appraised and the elements in the sample are estimated respectively at € 9877, € 27734, € 7199 and € 4480. These values allow us to identify the elements in the sample survey that are closest to the property being appraised. And, consequently, they can be very useful in defining a neighborhood for point X_0. In this regard, it is suggested to apply the statistical methodology known as "Multidimensional Scaling" (Hastie et al., 2001) to obtain a graphic representation of the distance matrix D.

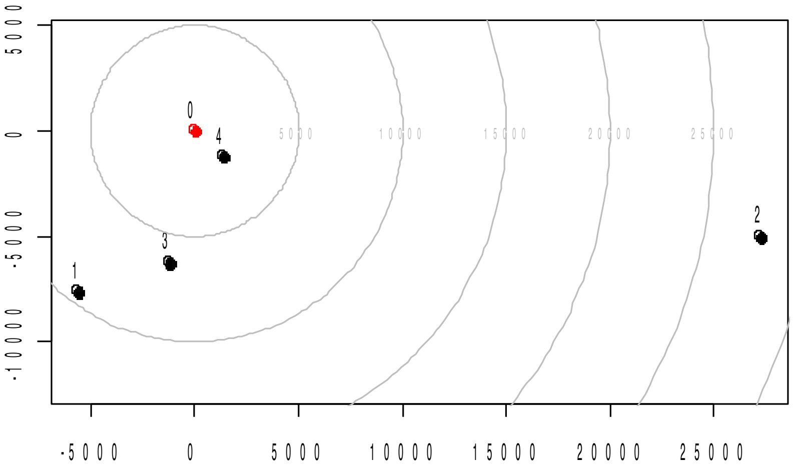

Illustration 3
Graphic representation
of the orthographic projection
of the distance matrix
(Multidimensional Scaling)

The chart shown in illustration 3 was obtained by using the metric Multidimensional Scaling method on the distance matrix D. The graph is easy to read. The red dot represents the property to be appraised while the other points are the four elements in the sample survey. Without great difficulty, one can see that the second element in the sample has a distance from the property being appraised greater than € 25000 and occupies a marginal position compared to the other three elements in the sample.
To attempt to limit the errors introduced by the linear approximation of the function of Market Value, we could therefore decide to delete the second element from the sample survey. However, this decision would further decrease the already small sample.
Then, always starting from the simple observation of the two-dimensional chart, we can choose to use a weight vector W, which depends on the distances of the elements from the property being appraised.

Proposals for the definition of the weight vector W

By algebraically developing the formulation of the estimated Market Value (26), we can reach the definition of the difference between the estimated value of the function VM_0 and the Market Value at point X_0.

$$|V\hat{M}_0 - f(X_0)| = |W'S - W'I\, f(X_0)| = W'((I\, X_0 - X)\, \theta + \rho + \varepsilon) \tag{31}$$

(31) shows that to try to minimize the estimation error, it is necessary to minimize the following terms $W'(IX_0 - X)\,\theta$ and $W'\rho$. The first of the two terms of error is a function of the differences in values observed between the property being appraised and the elements in the sample $I\,X_0$-X. It follows that the properties in the sample more distant from the property being appraised generate the larger errors. Similarly, the second error term assumes greater values when considering the sample properties more *distant* from the property being appraised. In fact, in the presence of functions of Market Value that are not too irregular, the linear approximation error becomes larger as the distance from the property being appraised progressively increases. To try to limit the negative

effects generated by the sample properties that are more distant from the property being appraised, a weight vector W defined by the first row (or column) of the matrix D can be used. For example

$$W' = \begin{bmatrix} \dfrac{d_{12}^{-1}}{\Sigma d_{1j}^{-1}} & \dfrac{d_{13}^{-1}}{\Sigma d_{1j}^{-1}} & \cdots & \dfrac{d_{1(n+1)}^{-1}}{\Sigma d_{1j}^{-1}} \end{bmatrix} \qquad (32)$$

This vector assigns a weight to the elements in the sample which is inversely proportional to their distance from the property being appraised. Obviously the application of such a weight vector generates an estimate of Market Value that is different from the one obtained previously (by adopting a uniform weight vector), in fact

$$\hat{VM}_0 = W'S = [0.202 \;\; 0.072 \;\; 0.278 \;\; 0.447] \begin{bmatrix} 79693 \\ 66662 \\ 83199 \\ 58380 \end{bmatrix} = 70202.04$$

The estimated Market Value thus obtained is to be preferred to that obtained by using a uniform weight vector. In fact, the negative effects due to the second element in the sample survey (see Illustration 3) are muted by the weight value used: 0.072 against the previous ¼.

It must be stressed that this vector (32) represents only one of many viable alternatives in the definition of the weight vector.

Proposals for the study of uncertainty of the estimation

We have already mentioned that the vector β of the so-called marginal prices is, in reality, an unknown. It is estimated as $\tilde{\beta}$ by appraisers giving rise, however, to an error component θ.

Previously, it was asked if it was possible to study the effects generated by θ on the estimation. In light of (31), the answer is affirmative. In fact, assuming a probability distribution for the vector θ, the distributional pattern of the difference between the estimated Market Value and the value of the function X_0 can be derived. This solution is certainly very interesting from a theoretical point of view because it allows us to evaluate the accuracy of the estimator of market value obtained with the SCA. However, from the practical point of view, another very interesting question arises: is it possible to attach a measure of uncertainty to an appraisal opinion?

The SCA is considered to be a deterministic method and the estimated Market Value is not accompanied by any measure of uncertainty. The so-called interval estimates, in fact, are currently the prerogative only of statistical procedures like multiple linear regression. But there are also components of uncertainty in the estimates obtained by the SCA. In particular, the components of vector $\tilde{\beta}$ are defined by experts in conditions that can rarely be defined as *certain*. For example, consider the marginal price of the Balcony variable in the case study analyzed: € 3180.This value can be obtained by consulting the industry bulletins, or by interviewing a number of experts in the local housing market. In this second case, the question arises whether it is possible that all the experts interviewed indicated exactly € 3180, the increase in value due to the presence of an extra balcony. More reasonably, the experts contacted would have reported values lying within a range from € 3100 to € 3300 for example. The marginal price used in all probability is simply the average of the values obtained by interviewing a sample of experts. But then the question arises whether it is possible to study the effects on the estimated Market Value generated by the uncertainty in the definition of the marginal price of a given characteristic.

By using Monte Carlo simulation procedures, it is possible to study the uncertainty in estimating market value that is generated starting from uncertain values of the components of vector $\tilde{\beta}$. Suppose, for example, that all components of the vector of marginal prices are affected by uncertainty. In particular, we assume that the exact values of vector $\tilde{\beta}$ are replaced by random variables as shown in Table 4.

Table 4 Distributions and indices of Marginal Prices of the characteristics

Characteristic	Distrib	Min	Max	Average	St.Dev
Areas	Normal			1135.86	254.53
Balconies	Uniform	3100	3300		
Cellar	Uniform	600	650		
Floor	Uniform	-2250	-2230		
Elevator	Uniform	3700	3800		
Year	Normal			438.22	116.79

The marginal prices of the Area and Year of construction characteristics were obtained by regression analysis. For this reason, they have been assigned a Gaussian (normal) distribution having as an average value the one previously used as a point value and as standard deviation the standard error value in Table 2. The uncertainty in the marginal price of the remaining characteristics was assumed to be described by means of uniform random variables. The extremes of the uniform random variables were chosen at random. The goal of this discussion is, in fact, to show the ease with which one can explicitly handle the uncertainty present in an SCA process. It is not the goal of this discussion to go into the most appropriate method for representing uncertainty in the attribution of marginal prices (we hope, however, for a deeper exploration of this topic).

Once the components of the vector $\tilde{\beta}$ have been transformed into random variables, the Monte Carlo simulation consists in repeating the procedure for estimating the Market Value for a desired number of times (in our case 10000). In each of these iterations, the marginal prices of the distributions (Table 4) are randomly extracted. The result obtained from the Monte Carlo simulation is a set of 10000 Market Value estimates. Illustration 4 contains a graphic representation of the empirical distribution of this set of realizations. The calculated average for the 10000 estimates obtained with Monte Carlo amounted to € 70160. This value is close to the point value previously obtained, € 70202.04. But it is far more interesting to note that, thanks to the simulation, it is also possible to associate an indication of uncertainty to the point estimate due to the estimation procedure for the attribution of marginal prices. For example, it is possible to indicate that the 10000 simulations generated estimates of Market Value between a minimum and a maximum of € 68740 and € 71050. Note that this is not a confidence interval.

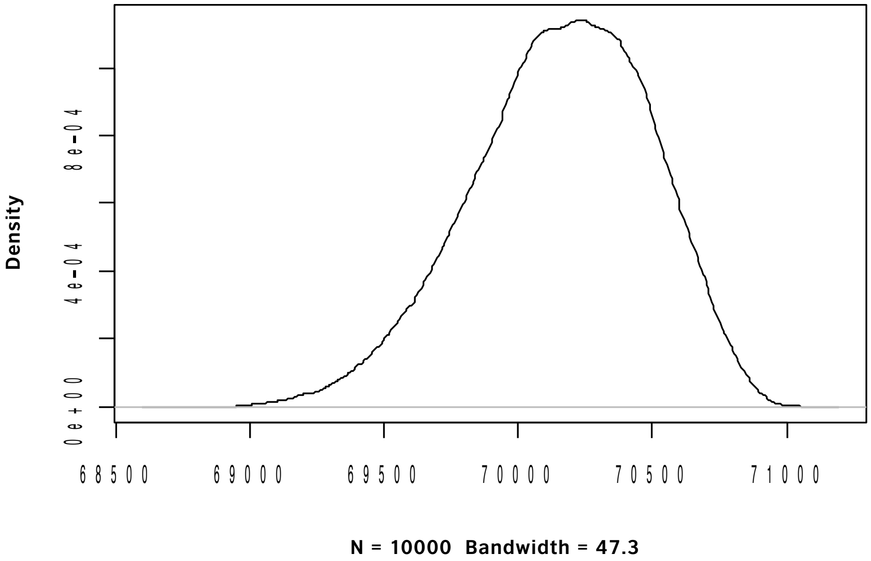

Illustration 4 Empirical density function of the estimated Market Value obtained through the 10000 iterations of the Monte Carlo simulation. Statistical indices: minimum 68740, first quartile 69940, median 70180, average 70160, third quartile 70400, maximum 71050

References

Hastie T., Tibshirani R., Friedman J. (2001). *The Element of Statistical Learning. Data Mining, Inference and Prediction*. Springer.

Isakson H. R. (2002). *The Linear Algebra of the Sales Comparison Approach*, Journal of Real Estate Research, 24(2), 117-128.