

# **R**REAL ESTATE STATISTICS WITHOUT FEAR

*STATISTICAL ANALYSIS IN REAL ESTATE APPRAISAL*

---

*Mark R. Linné MAI, SRA, CRE, CAE, ASA, FRICS*



# **I**ntroduction

## *What You Need to Know*

---

### **Is It For You?**

What becomes clear through the process of anticipating the future of the profession, is that appraisers will be utilizing statistics far more frequently than is the currently the case. The need to accurately analyze and interpret the available database becomes compelling, as the data becomes more readily available. Appraisers who fail to adapt will not be providing the best analysis to their clients; appraisers who meld their appraiser judgment to the needs of their clients will gain the edge in providing better appraisal analysis. As a by-product, the appraisal process should become easier and more reliable, relying on judgment and analysis of sales which have been vigorously analyzed via the appropriate statistical mechanism.

Appraisers who are able to integrate the best elements of judgment and technology, combining a statistically validated computer technology with actual appraiser inspection and review will become the model of the profession into the next century. Such an integration ensures that human judgment is augmented, not eliminated, by statistical use of computer technology.

The future belongs always to those who adapt; those who are willing to make the changes that will ensure their success in the future.

We are in the nexus of a transformational series of events: the significant synergism between the availability of significant computing power, which has advanced significantly over the last

decade, and the application of this computer power against the available database, utilized within the framework of statistical techniques and evaluative tools.

By approaching this process with an open mind, and with a desire to enhance the appraisal process, the appraiser who chooses to adapt will be the appraiser who survives.

# Chapter One

## *Fundamentals of Appraisal Analysis*

---

### **Introduction to Statistics and Analysis**

To many people, the word “statistics” is synonymous with “mathematics,” and, of course, there are many synonyms for mathematics. Most are fairly negative, created during periods when math students in school were either bored or terrified by what was being presented by their instructor. This experience has tainted many adults and has affected the way they look at math (and statistics) in their professions. Indeed, companies often hire persons who majored in math or a related field simply because such persons do not fear math. Statistics, as it was (and is) taught traditionally in school, was often included with all the other math stuff that students discarded after final exams, their experience teaching them that math-related skills were either not useful for everyday life or something to be avoided at all costs. Unfortunately, this practice has also affected the appraisal profession, where the mere mention of statistics (let alone things like regression analysis) causes many in the audience to draw up in an intellectual fetal position.

The notion that statistical analysis can actually prove practical in a real world setting is a surprising revelation to many. The fact that its concepts can be straightforward and readily understood can be amazing. The purpose of this handbook is of this ilk. Appraisers can and should utilize statistical analysis in every appraisal. The concepts need to be understood by the user and definitely not feared. This chapter will focus on general concepts surrounding statistical analysis. Chapter 5 will actually go into a detailed description of how and when to apply one of the most powerful statistical tools out there, regression analysis.

## **The Scope of Analysis**

The use of statistical analysis gives one the ability to **identify, measure, and interpret** “things” in nature. These “things,” which academicians call phenomena, can be thought of as any measurable activity, such as human action or naturally occurring events. The growth of a particular crop, the relationship between high blood pressure and strokes, the buying patterns of bald men in Minnesota, all of these “things” represent examples of particular phenomenon that can be statistically categorized and analyzed.

What constitutes phenomenon that cannot be analyzed? Obvious examples include things that cannot be identified or measured, like UFOs and psychic phenomenon. Perhaps such things do exist, but we do not have the ability to measure them as of now. For example, could scientists in the 1800’s analyze the spread of a disease like cholera before they knew what to study (i.e. germs)? Yet, cholera did in fact exist, as the morticians could readily attest to back in old London.

In the real property appraising arena, the phenomenon we are attempting to analyze concerns a property’s value or factors that affect value; the actual variables analyzed depends on which appraisal method is employed (either cost, income or direct sales comparison techniques).

The cost approach attempts to look at the construction costs of a given improvement on a given piece of land, with the land value determined by some kind of separate analysis (usually the direct sales comparison approach). The cost approach sometimes references cost indices, provided by vendors such as Marshall and Swift. It is important to know that these indices are themselves some form of statistical measurement. That is, they utilize the usual or typical costs of construction, which is a form of averaging, which in turn is termed by statisticians “a measure of central tendency.” Other modifying factors such as time or location factors can be used to modify the cost index values. All of these are results of statistical analysis and are

themselves what we term “statistics.” The appraiser, therefore, already utilizes statistical analysis when employing this valuation method for a given appraisal problem.

The income approach also utilizes statistics and analysis, in that comparable rental properties are arrayed (i.e. lined up) and the appraiser employs some form of analysis to select the “best” properties to determine a market rent rate. The process of selecting, adjusting and reconciling these arrayed properties is a form a statistical analysis. It can be straight forward or complex; obviously, discounted cash flow analysis can certainly involve the use of statistical analysis. The determination of an appropriate capitalization rate or gross rent multiplier also can involve statistical analysis to determine the appropriate rate or factor. Many commercial appraisers are already using statistical analysis in their appraisal reports.

It is in the third method, the direct sales comparison approach, where much of the traditional statistical analysis is employed. The D.S.C. approach also gives us a good example of the three applications of statistical analysis described previously. In the direct sales comparison approach, the “thing” we are measuring is usually the sale price of the property. It is often referred to as the “outcome variable.” This appraisal valuation method simply relates the sale prices of similar properties to the subject property either by a grid mechanism or some other form of factor comparison, where selected comparable sale properties are used as the basis for deriving a value for the subject property. Statistical analysis helps the appraiser select comparable properties (identifying), measure differences in property characteristics (measuring) and apply adjustment amounts (interpreting) to arrive at an estimate of value.

The following will examine these three elements of statistic analysis in appraising, using the direct sales comparison approach to analyze residential data:

## **Identification**

Identification becomes the first important step in our analysis. The crucial question of identifying what particular phenomenon is under scrutiny depends on the goal of your analysis. For example, is your goal to model the actual market transaction that determines property value for a given area? Or perhaps the goal is to create a broad valuation model that can be used by the public. The answer is important, in that it often determines what type of modeling strategy should be employed. For example, it makes no sense to create a 20-plus variable regression model to value residential properties if the market transaction that sets those values in the real world is limited to a smaller number of variables. Indeed, after location, size, style and perhaps two or three more factors, what else is there that determines residential property values on a consistent basis? Factors such as swimming pools may theoretically increase property value, but that increase can be affected by location (is the property located in Alaska or Arizona?) or whether it is a typical upgrade feature. If there are insufficient number of sale properties with swimming pools, then the appraiser cannot adequately analyze its impact on property values. If there are too many properties with swimming pools, like areas with warmer climates, then it may be impossible to determine the impact that factor has on value. Identification, therefore, becomes important not only in correctly analyzing the overall valuation problem, but also in identifying the correct variables to be ultimately analyzed. This topic will be covered in a later chapter.

## **Measurement**

Measuring the variables that affect property value is the next important step of statistical analysis. Once the data is identified, measuring it allows the appraiser to quantify the phenomenon and draw conclusions from it. Correctly measuring the phenomenon links the identification and interpretive steps. Correctly measuring data means the appraiser must

understand the basic concepts behind data; basic questions about what type of data there are, their limitations, and the actual source of the data.

What are the sources of data? In the case of real estate valuation, that data can often be found in county assessor files available to the public. Sale information can also be found in local MLS data downloads, the information source used by real estate brokers. Integrating data from sources such as these has its own inherent challenges, such as the interpretation of variables. For example, what is the definition of Gross Living Area being used by your data source? If the county assessor defines living area as all above-grade heated square footage, while M.L.S. source considers living area as all above grade square footage minus any heated porches or attics, then a decision has to be made as to what definition to use. One of the sources of data, therefore, must then be altered systematically to “fit” with the other source. It is critical that whatever process is employed “filters” data to fit agreed upon definitions consistently.

Other questions can also arise when looking at linked variables, such as house style and living area. For example, if MLS data considers living area with split-levels to include the garden level floor, while assessor data groups the garden level area with finished basement square footage, then a decision needs to be made as to a consistent definition for split-levels (this particular problem with bi-levels can be even more problematic!).

The aforementioned example was one of inter-definition inconsistency, where one data source does not “agree” with another source in terms of a variable definition. There is also the problem of INTRA-definition inconsistency; i.e. when definitions are not always consistent within the same data source. MLS, for example, has made great strides to maintain consistency in its data. Problems remain, however, when one real estate agent’s finished basement is another agent’s garden level component of total living area.

Errors that create data problems are generally of two types. The first, systematic error, occurs when the data is consistently inaccurate or unreliable. The above are examples of such errors,



and they need to be addressed before any valuation analysis can occur. The second problem is random error, where the error occurs infrequently and uniquely. If a particular living area is measured incorrectly by the county appraiser, the impact on the analysis is usually less important than in the case of a systematic error. Usually if the random error is extreme, the analyst can readily identify it based on the relative sale performance of the other sale properties; the next step can be the removal of the property in question, or the exchange of the “bad” value with another value (sometimes referred to as the “proxy” value). This book will illustrate strategies to get around both types of these “bad data” problems in later chapters.

Once the data is edited (i.e. cleaned up), the actual measurement of variables is the next step in the measurement process. It is often the easiest step in statistical analysis, especially with the advent of personal computers. This step can often take only a few seconds to complete and uses software such as spreadsheets (Excel, Lotus, Quattro) or databases (Access, Foxpro, Dbase).

The question arises as to how to measure the edited data. This can depend on the type of data being analyzed. Appraisers need to understand the types of data out there and their limitations and strengths. Data such as sale price totals are very robust, in that they contain all the basic features of numeric information. We can tell, for example, that a \$200,000 home is twice as expensive as a \$100,000 home. Other data, such as house style or quality levels, need further analysis before they can be used by the appraiser.

Statisticians have developed several methods of classifying what types of data there are. These categories are based on several factors, all of which determine what type of measurements are possible. This determination results in the statistician determining what types of analysis are then appropriate to use. Appraisers need at least a basic understanding of this data analysis to adequately evaluate statistical analysis they perform.

The classic three-level statistical data scheme breaks down all data into these groupings:

**Nominal-** Nominal variables represent an identification with a particular group, with no numeric difference implied between the groups. A generic example is Fireplace, where properties possessing a fireplace equal the value of “1” and those without a fireplace are represented by “0”. There is no inherent numerical difference between these values; “0” and “1” are simply markers (or names, hence the term “nominal”) to distinguish between these two groups. Another example would be a location number, such as a subdivision number used by a county assessor office. Each subdivision number is in effect a label that identifies that subdivision area uniquely. A real world, non-appraising example would be a person’s social security number.

**Ordinal-** This variable type are also called rank-order variables, since they refer to some kind of ordering of data. A real world example would be a school report card, where a grade of “A” is above “B,” which is above “C,” and so on. Although there is additional information provided to the appraiser as compared with nominal data, in that we now have a scale, there is no information referring to any numeric differences between these ordered numbers. Whether the school grading system is based on 90-100% for “A”, 80-90% for “B”, or 90-100% for “A” and 60-90% for “B” is not known. What is known is that “A” is greater than “B,” but not by what amount. The spacing between values, therefore, is unknown.

An appraising example of this would be a qualitative scale, such as property physical condition. If the assessor data contains a four level variable called Condition, with 4 = Excellent, 3 = Good, 2 = Average, and 1 = Poor, one still does not know the real differences between these values. For example, how much better is a good property from an average property? Is the spacing between the rank categories consistent?

Assume that Construction Quality is included in a valuation model, and that it equals \$5000 based on differences between the average selling prices of homes. Properties with a rating of “Excellent” would receive \$20,000 ( $\$5,000 \times 4$ ) while “Average” properties would receive only

\$5000 ( $\$5,000 \times 1$ ). The actual valuation data may show something entirely different, where “Excellent” properties were \$25,000 above “Good” properties, which in turn were \$3000 above “Average” properties. This problem with spacing occurs whenever there is more than two categories within the variable. Sometimes appraisers can utilize methods such as paired sales analysis to create separate values for each quality level and there are ways to treat this problem using more advanced procedures such as regression analysis.

The important point for the appraiser is to remember that ordered data can be utilized, as long as distance between variable values is not taken as is.

**Interval-** Nominal and ordinal data are also known as “categorical” data, since the primary information derived pertains to group membership (for example, is the property part of Subdivision A or Subdivision B?). The remaining data category is termed **Interval** data, where the data value itself gives intrinsic information. This type of data category is also referred to as continuous data, in that the intervals between the numerals actually mean something.

An example of continuous data would be a variable such as sale price. A home that sells for \$200,000 has a market value worth twice that of a home that sells in the same area for \$100,000 during the same time period. On the other hand, a home with a quality of “Excellent” is not twice as valuable as a home with a quality rating of “Average.” This is true even if excellent quality homes are given a rating of 2 and average homes are given a rating of 1.

The bottom-line difference between nominal/ordinal data and interval data is that the latter yields more information. The types of statistical tests, therefore, can yield more information. We not only know that a 2000 square foot house differs in size from a 1000 square foot home (nominal level information) or that it is bigger (ordinal level information), but that it is twice the size of the home (interval level information).

## **Interpreting**

Once the data is captured, choosing the type of analyses becomes the final logical step. Some procedures simply indicate the spread of data, either through descriptive means (frequency) or through index numbering (scores and rankings). In real estate, there are many examples of these types of measurements, such as average sale price or the top ten selling neighborhoods. Here nominal and ordinal data can be used alongside interval data, such as neighborhood and sale price. It is when the appraiser wants to quantify differences between groups that the data needs to be carefully scrutinized.

Procedures such as regression analysis can compare many separate characteristics at once with a group of sale properties, and estimate their contributory weight to the value of the subject parcel. It is this ability to simultaneously compare many homes in a particular area that gives the modern appraiser an analytical leg up over the traditional appraiser, so long as the area in question contains enough sale properties and the data is properly prepared. These are considerable qualifiers. There are areas, such as custom home neighborhoods, mansion-dominated markets, areas crisscrossed with outside influences such as highways and boulevards, mountain resorts, and other locations, where traditional appraising methods are superior to statistical methods that rely on a certain level of similarity between properties (termed “homogeneity” by statistical gurus). In these cases, the statistical information garnered may still yield valuable adjustment data for the appraiser, although it should be applied within traditional fee appraisal methodologies.

There are statistical “tricks” one can employ to make certain nominal or ordinal data behave as interval/ratio data, but it is out of the scope of this book to venture too far in this direction. Some of this will be discussed briefly in Chapter 5. For now, knowing that different levels of measurements exist is sufficient; knowing the question one wants answered is more important than trying to remember twenty separate statistical measurement procedures.

Interpreting output from statistical analysis does not require the interpreter to be a graduate level statistician, as long as the above steps have been followed. Fortunately, the appraiser analyst has a large body of appraisal theory to judge the “soundness” of statistic output. For example, if a regression-based valuation model yields an incremental value for fireplaces at \$20,000 in a given neighborhood, the appraiser can compare that with other statistical information about this same area. If the average sale price is \$350,000, then the appraiser may decide that the \$20,000 per fireplace is a sound amount. On the other hand, if the average sale price is \$80,000, then appraisal judgment may force the appraiser to reject this variable value as too extreme. In that instance, the variable fireplace may be masking another valuation factor not represented in the model. This masking can involve a related variable (such as a first floor den) or another variable not related at all. For example, if in our fireplace example, there were 5 homes with one fireplace and 5 homes with two fireplaces. For some reason, let’s say the one-fireplace homes all sold early in our sale year, while all of the two-fireplace homes sold in the latter part of our sale year. It is possible that all of some of the fireplace value could capture any sale price appreciation present in the market. One way to test this would be to look at the market adjustment variable in the model. If it was absent or had an unsound value (even negative), then it could be affected by the fireplace variable. Statisticians call this “multicollinearity,” which means simply that two variables are interacting in the analysis. It is important that appraisers understand completely what variables are present in any statistical analysis presented to them (such as from a vendor-supplied A.V.M. product).

To correctly interpret data, the appraiser needs to understand the basics behind statistical analysis. As previously explained, appraisers already use statistical analysis in their everyday work. Statistical analysis encompasses the three steps or elements of basic analysis (identify, measure, interpret). The goal is to provide a framework to utilize data for the purposes of appraisal work.

## **Descriptive versus Inferential Statistics**

The two major areas of statistical analysis employed by appraisers involve the use of descriptive and inferential statistics. The difference is obvious when looking at a real world example. Suppose an appraiser wishes to include a section on sale price appreciation in his appraisal report. The first step would be to describe the actual appreciation factor, such as the quarterly trend in the average sale price for residential homes in a given area. The next step would apply this trend in some manner to the appraisal valuation process in the report as it relates to the subject property. This may be as simple as applying a percentage factor to the subject property's concluded value.

This two-step process illustrates in a simplified manner the difference between descriptive and inferential statistics. Descriptive statistics describe data in some way. The distribution of sale price or type of home style are examples. Inferential statistical analysis, on the other hand, would attempt to infer some described phenomenon to the subject property. In our above example, once the sale price appreciation was described in step one, step two "connected" it to the subject property valuation.

It is usually in the second step that appraisers find the most difficulty. Applying a described event to the subject property implies risk, in that the appraiser is placing their judgment and analytical skill on the line (and in writing). Is that not, however, the entire focus of the appraisal report? Is not the entire valuation process one of inferring value on a given property, based on described related phenomenon? As this book prescribes, most of the "new" knowledge or techniques described herein are already being used by appraisers.

## **Descriptive Statistics**

With the introduction of computers, analysts of all types (including appraisers) are faced with a myriad of choices. Some of these computer users, including the authors, think the choices are overwhelming. As with the case of learning to operate a personal computer for the first time, a

successful strategy to overcome this information overload is to start at the right place. In this instance, it would be for the appraiser to ask questions from the “appraising side” when approaching data. In other words, ask pertinent appraising questions and seek information from your data on an “as need” basis. Questions such as the size of the property, or the style, or the number of comparable sales (and any adjustments necessary) are the places to begin your analysis. Where not to begin your analysis is buried in the index or table of contents of a statistics textbook. Even worse, in the table of contents of a statistical programmers text book! Rather, ask of the data and then use the appropriate statistical procedure to answer those questions.

The first descriptive statistic everyone learns is the central measure of tendency, also known as the average. There are actually three measures of central tendency taught in most beginning statistical courses: the mean, median, and mode, all collectively known as the average. In most cases, the mean is synonymous with the term “average,” although the appraiser should always be specific as to the exact measurement behind this term. Many inferential statistical tools, such as regression analysis, utilize the mean in their calculation process. The mean average is basically calculated by listing every value of a variable (such as sale price) and dividing the sum of these values by the number of observations. If the sale file contains ten sales, then the mean average would be calculated by summing all of the sale prices and dividing that number by the number “10.” Obviously, if the sale file contains a sale that is significantly larger or smaller than the other sales, then the mean average can be affected significantly. If the sale file contains 100 sales, the effect of this same “outlier” sale would be less. But how can the appraiser know for sure?

Utilizing the median average to verify the mean average is often a good method of checking for unwanted affects of extreme outlier cases. The median average is simply the middle value, or the one that occurs at the 50<sup>th</sup> percentile. If the median and mean averages are similar, then the appraiser can assume that the mean is not affected significantly by any outliers. It does not mean, of course, that outlier cases are not present in the sale file. If an outlier case “pulls” the

mean average from the middle value, then the median and mean averages will not be close to one another. On the other hand, if there are outlier cases at both ends of the sale distribution (i.e. sale prices significantly above and below the mean average), then the appraiser might still want to restrict the analysis to sales that are within a certain distance from the middle value.

The modal value is the value that occurs most frequently in an array of data (an array is simply a column of data, such as the above sale price example). Comparing the mode with the other two measures of central tendency can also support the mean value as an accurate representation of the true average, and can help describe how the data is distributed across all values. If the mode, median and mean all agree, then that tells the appraiser something about the distribution of the data. In our sale data example, having these three averages agree can tell us that the data is not skewed, meaning it does not bulge in either direction. The spread is fairly uniform, therefore. Often the distribution of the data is typified by the bell curve, where most of the values occur in the center (where the mean, median, and modal averages would lie in this case). Another type of data distribution is the uniform distribution, where the data is evenly spread out; an example of this would be a sale file, where the sale dates are evenly spread out over the sale period.

There are other descriptive statistical tools available to the user. Two very easy and simple descriptive statistical tools to help describe what your data “looks like” are the frequency distribution and crosstab table (also known as the contingency table). The frequency distribution tells the appraiser how many times a value occurs for a particular variable. For example, if an appraiser has 10 homes to analyze, it would be helpful to see the distribution of the living areas of these homes. It may also be helpful to see what style of homes are represented in this 10 home sample. Frequency analysis is an easy and elegant way to view this data; with it one can determine if the homes are similar in size and style between one another and with the subject property.



What if the appraiser wishes to compare both the size and the style of the home at the same time? For example, if the subject home is a large ranch home, but the 10 comparable sale properties contain only much smaller ranch homes, then using a crosstab table could alert the appraiser that further property characteristic adjustments could be warranted. Or that another sample of more comparable properties needs to be gathered.

Time is also another variable that a frequency or crosstab table can readily display. The same questions regarding comparability can be answered. In our 10 home comparable sample, the appraiser may want to know when these 10 sales occurred. In areas of sale price appreciation, it may be important to know that our large ranch home is better represented by earlier sales, which were more predominantly made up of this type of home. A frequency table can easily tell the appraiser if the sale dates are evenly distributed across the sale period (i.e. a uniform distribution).

The important point is that the appraisal analyst needs to know the limitations and possible pitfalls of his or her data file. Running adequate descriptive frequency and cross tab tables is the first critical step taken in developing a valuation model that makes appraising sense. Always assume that your analysis will potentially have to undergo the same appraisal scrutiny as that of typical fee appraisal quantitative methods. Creating the qualitative checks is vital for the entire analysis process. Without it, the appraising analyst can leave himself or herself open to criticism.

Other descriptive tools include the standard deviation, boxplots, stem and leaf plots, and other statistical methods whose goal is to describe the data. These descriptive statistical tools all describe the distribution (or spread) of data. Knowing how your data is distributed can help the appraiser determine way that the data might behave, which is the subject of the next section on inferential statistical tools.

## **Inferential Statistics**

Inferential statistical tools differ from descriptive tools in that they help the user define the association between a set of independent variables and dependent variables. One of the primary inferential statistical tools, termed regression analysis

Independent variables are those that are given in the analysis. In the cause and effect language of the next section, the independent variables “cause” the dependent variables to behave in a particular way. Some inferential tools simply describe the strength of the relationship between the independent variable and the dependent variable. Here strength is measured as the amount of change in the independent variable and the resulting change in the dependent variable. Any unexplained variation in the dependent variable is treated as an unknown or error term.

More powerful inferential statistical tools attempt to model, or explain in numeric detail, the association between independent and dependent variables in a systematic form. For example, regression analysis attempts to build a “linear” (meaning straight line) relationship between two sets of variables (generally, the dependent variable set has one variable and the independent set has one or more variables). The idea behind regression analysis is that if you change an independent variable by one unit amount, the regression model then changes the dependent variable by some amount. The regression analysis can also tell the user the amount of variation the model actually explains. These facets of regression analysis will be covered in greater detail in Chapter 5.

Inferential tools can be used to describe a set relationship, such as the association between the size of homes and the sale price in a given neighborhood. Here the analysis involves analyzing the data in the file, with the purpose of explaining the relationship between the independent variables and the dependent variable and is termed “closed set analysis,” since it involves only

the data in the inferential analysis; in other words, all cases under study have both a dependent and independent variable set.

Another useful application involves predicting the value of the dependent variable based on the value of the independent variables, termed “open set analysis.” The term open set is used because the results of the closed set analysis are then applied to cases where there is no dependent variable value. For example, an analysis of the relationship between the sale price and the size and age of single-family homes in a given area can be readily modeled by regression analysis. This can be limited to a simple study of the effects of housing characteristics on market value (close set analysis), or the appraiser can use this same information to predict the value of unsold homes in the same area (open set analysis). This second application can be used independently or with traditional direct sales comparison grid analysis. For example, a multiple regression model can help develop the adjustment amounts used in a sales adjustment grid.

Although at first these two applications appear one in the same, the risk varies greatly. To be successful, both close set and open set analyses need the independent variable set to be comprehensive enough to adequately explain the relationship between the independent and dependent variable sets. Open set analysis also requires that the sale sample adequately represents the population of all homes in the area of study. For example, if the regression sale sample includes only ranch style homes, it may not be a good model for 2-story homes in the same area.

Inferential analysis begs the question of causation. The next section discusses the pitfalls of assuming causation in economic relationships.

## **The Case for Causation**

It is impossible in mathematics to prove causation. Simply relating two sets of phenomenon does not by itself “prove” that one causes the other. Yet, this leap of analytic faith is what drives the entire basis of inferential analysis. A few rules of thumb should help guide the appraiser as to when to assume causation and when to avoid it at all costs.

The following eight rules are necessary, but not sufficient individually, to “prove” causation. Appraisers need to be careful not to overstate certain assumed economic relationships, even those that are supported by the following rules.

### **1. Analysis and Bias**

The statistical analysis employed to support a causal relationship needs to be as bias-free as possible. For example, suppose the appraiser makes a statement in a report that rising income “causes” an increase in the demand for greater retail space in a given area. Supporting analysis in the appraisal that is sponsored by the local chamber of commerce may be biased and therefore less credible than a more objective source (such as a local university study on retail demand). The appraiser should always question the source of data and analysis, especially when relying on such second hand support in their own appraisals.

Analytical bias can appear in many forms, such as sampling error or a flaw in the measurement process. An example of the first would be a sale data file that purports to represent all single family homes in an area, when in fact it represents only those homes that actually sold. The question arises whether that sale sample represents all homes; suppose there were two distinctive home builders in a neighborhood with significant differences in building quality. The higher quality homes may sell at a greater rate, and therefore be over-represented in the sales sample. The appraiser, in this case, would have to identify the home builder in his analysis and treat it as a variable to prevent sampling bias from undermining the conclusion of the analysis.

The other method of dealing with this problem would be for the appraiser to state that the valuation model pertains only to the higher quality homes.

The second major source of bias arises from measurement error. For example, if an appraiser purchases a single-family home sale file from the local assessor office, that data may contain errors due to measurement mistakes. This can arise from poor building measurements, inconsistent definitions of building characteristics, and other instances where the data reported differs from reality (this was covered earlier in this chapter). This is a common problem in some MLS data files also.

## **2. Strength of Association**

The stronger the measurement of association, the better the appraiser can support the assertion of causality. If your support is simply relative, demonstrated by the statement that “since income is expected to increase over the next five years, we can assume that retail demand will also increase during that same time,” it can be difficult to defend the causal link and can even create doubt with the reader. This is true even for associated phenomenon that “we all know are related.” Assuming that the reader shares the same beliefs (or biases) that the appraiser possesses can be risky.

A more defensible assertion would be an actual measurable statement, such as “a 5% increase in aggregate income results in a 3% increase in aggregate demand for housing.” This causal link assertion would then be supported by some stated mathematical relationship.

## **3. Consistency**

Can the causal association be proven in differing locales and from differing sources? If the appraiser can cite several sources from different location, it can help support the case for causation.

#### **4. Correct Temporal relationship**

Your data and analysis to support causation must have the correct time relationship. For example, if one is purporting that an increase in personal income “causes” an increase in retail demand, then the income increase should precede the increase in demand.

#### **5. Dose-response**

In some causal relationships, the more of “X causes more of Y”; in our example, the greater the increase in personal income can cause greater increases in retail demand. Realize that some relationships are not linear, meaning an increase in X may cause an increase in Y only after a certain threshold is achieved. For example, an increase in personal income may not affect the demand for luxury homes until a certain income level is reached.

#### **6. Plausibility**

The stated causal relationship should make appraisal and economic sense. Even if your analysis supports the relationship that “X causes Y,” if X represents a decrease in personal income and Y represents an increase in retail demand, the reader will question the entire construct. This is not an attempt to state that only obvious relationships can be proven causal, but it becomes more difficult when the stated relationship ultimately does not make appraising sense.

#### **7. Specificity**

Generally speaking, this occurs when a single effect causes another. Rather, many economic relationships are banded together and cause several outcomes. Linking one cause with one effect can often say more about the limited scope of analysis rather than make a case for causation.

## **8. Analogy**

If another similar relationship can be cited, this can support your contention of causation. These other relationships can take place in differing locations or involve similar economic variables. For example, the link between personal income and retail demand can be similar to the relationship between disposable income and demand for a particular good.

Any one of these rules can support the case for causation. Typically, utilizing several of these rules makes your contention stronger. Satisfying all eight rules, while a noble goal, would probably overwhelm the appraisal report; if the causal link is critical in some way to the appraisal conclusion, however, utilizing as many of these rules as possible would probably be prudent. The main point is to always question your assumptions regarding causation and state clearly your reasoning behind any causal construct presented in your appraisal report. And to realize that ultimately, the case for causation is based on the weight of evidence and not absolutes.

### **Summary**

In this chapter, we discussed statistical analysis, focusing on several topics. The purpose of statistical analysis is to identify, measure, and interpret things or events (also called phenomenon). To be studied, events and things must be known and understood; it is this fact that gives the appraiser the upper hand in dealing with statistical analysis. Appraisal theory, not statistical theory, drives the machinery. The appraiser needs to be in control to correctly interpret whatever output is placed before them, even if that output comes from a sophisticated software package or from an available statistician on duty.

There are three basic steps used in statistical analysis; identification, measurement and interpretation of phenomenon. Statistics allow the appraiser to perform these three steps in

analysis; often the appraiser is already doing this function using traditional appraisal methodologies.

Asking the correct question becomes as important as employing the correct statistical procedure. Indeed, it determines what correct statistical procedure is needed. The nature of inferential analysis was compared with descriptive analysis, with both similarities and differences highlighted. Statistical tools needed to correctly perform both types of statistical analysis were briefly identified, with examples that will be further explored in an upcoming chapter.

The final section on causation was a very brief primer on a concept that is often taken for granted by analysts. Causation per se cannot be proven, but can be inferred using eight rules as support.

Regression analysis is a powerful statistical tool that appraisers can use to create valuation models. Used primarily in residential appraisal setting, it can measure the influences of several variables on a property's value. The array of variables utilized must simulate the real estate market and be realistic in the total number of variables in the model; this will be discussed further in Chapter 4.



# Chapter Two

## *Review of Statistical Analysis*

---

Statistical analysis gives one the ability to identify, measure, and interpret events in nature. These events, or phenomena, can be thought of as any measurable activity; in the case of real estate valuation, this activity involves human action, such as buying, selling, renting, or developing real property. Such phenomena are measured by monetary transactions or other quantitative indices, allowing the appraiser to categorize, gauge, and compare the activity.

What phenomena cannot be analyzed? Obvious examples include real estate activity of a confidential nature, where the data are concealed. In such cases, market- or industry-derived factors can be used, as long as the appraiser makes it clear in the appraisal that specific property data are unavailable. In a case where standard factors or comparable properties are not available, an appraiser may not be able to perform the appraisal assignment at all.

### **Components of Statistical Analysis**

The following sections will examine these three components of statistic analysis from an appraisal perspective:

- 1. Identification**
- 2. Quantification**
- 3. Interpretation**

For example, when performing any direct sales comparison analysis, the market-relevant variables must be identified, as well as the unit of comparison. Data is then quantified through the adjustment process, with the results are interpreted and applied to the subject. While both traditional valuation and appraisal valuation modeling employ all three components, the difference lies in the scope of the

analysis. Only appraisal valuation modeling can effectively use all of the available market data. The results from the latter are superior because of the broader and deeper scope of market analysis, as well as the quantification methods employed.

### **Identification**

The question of identifying what particular real estate phenomenon (or variable) is of interest depends on the purpose of the appraisal analysis. The goal could be one of the following:

- **To specify important elements of comparison that control overall value**
- **To value a group of properties in a given area**
- **To create a broad valuation model**
- **To “mark to market” (value) a portfolio of real property, financial securities, or derivatives**
- **To estimate risk**

The goal of the analysis determines what type of valuation strategy should be employed. For example, an appraiser should not create a regression model with numerous variables to value residential properties if the market that determines those values uses a smaller set of variables. The presence of a swimming pool may theoretically increase property value, but that increase can be affected by other factors, such as location (e.g., if the property is in Alaska or Arizona) or whether a pool is a typical upgrade feature. If there are no sale properties with swimming pools, then the appraiser cannot adequately analyze its impact on property values anyway. If most properties have swimming pools, like in wealthy areas in warmer climates, then it may be impossible to separate the impact that factor has because it is associated with all large, quality homes.

**Identification**, therefore, is important not only in analyzing the overall valuation problem correctly, but also in choosing the specific variables that will be analyzed. It sets the stage for the entire appraisal valuation analysis. It is not limited to the set of variables chosen; identification includes all preliminary steps in the valuation process.

## **Quantification**

Once the scope of the appraisal and all relevant variables are identified, measuring them and calculating their impact allows the appraiser to quantify their effect. Calculating the influence of a phenomenon links the identification and interpretive steps in analysis.

The actual calculation of variable influence can be the easiest step in statistical analysis, depending on the appraiser's "toolbox." If the appraiser is limited to a pad of paper and a financial calculator, the analysis can take hours or even days. It can even limit the scope of analysis. When appraisers use software such as electronic spreadsheets (Microsoft Excel, Lotus, Corel Quattro Pro) or databases (Access, dBase), the quantification step can take much less time. Specialized analytical software, such as SPSS and MiniTab, offer the best package for data analysis and are highly recommended.

To correctly measure data the appraiser must understand:

- **Some basic concepts behind data**
- **The types of data**
- **Limitations of the data**
- **Some considerations about the source of the data**

This analytical step produces output. It is the interpretation of this output where the appraiser will now apply appraisal theory to solve the appraisal problem at hand.

## **Interpretation**

This step essentially concludes the analytical process by evaluating all output from an appraisal perspective. The appraiser must apply appraisal valuation theory to correctly interpret this output. This theory is not part of a new branch of appraisal theory, but rather a restatement of existing and accepted appraisal standards of practice.

The primary objective of this final step is to insure that the output from your analysis makes appraisal “sense,” but if the process includes statistical applications, then as the appraiser needs to understand the statistical analysis that drove the measurement process as well. This requires base competence (the goal of this book) and must be presented in a manner that the reader can understand. It does not require the appraiser to be at the graduate level of statistics, but it does require basic competence concerning the correct interpretation of any analytical output.

Now that the basic three step model of analysis has been briefly explained, the next section provides for a brief review of common statistical terms. Some of these will be used in following applications and need to be understood.

### SOME IMPORTANT TERMS AND CONCEPTS

In statistics, the term **population** refers to all items that are being studied and the term **sample** refers to a specific subset of the entire population. For example, if you were retained to make an appraisal of the market value of a specific neighborhood shopping center (subject property), the population would be all shopping centers in the same geographic area. A sample could be all recent sales of centers that are comparable to the subject property in similar markets.

A **variable** is the term used for a property attribute that may take on different values across different properties. For example, the parking ratio of a shopping center is a variable, since different centers have different parking ratios. One center may have a parking ratio of 4.5 spaces-per-thousand square feet of gross leasable area, while another center has a ratio of 5.2 spaces-per-thousand. The sale price is another variable, since different properties sell for different prices.

**VALUE** is a variable that is typically not part of the appraisal process, although it often the goal of the entire appraisal. Market value and sale prices of comparable properties are used interchangeably, indicating the importance of correctly defining value in the appraisal.

An important variable concept when performing any valuation analysis involves independent and dependent variables, diagramed as follows:

**DEPENDENT  
VARIABLES**

**SALE PRICE**  
OR  
**RENTAL RATE**  
OR  
**CAP RATE**  
OR  
**COST FACTOR**

**INDEPENDENT  
VARIABLES**

**IMPROVED AREA**  
**AGE**  
**LAND AREA**  
**CONDITION**  
**QUALITY**  
**PARKING**  
**FIREPLACES**  
**OFFICE FINISH**  
**STORIES**

Generally, the appraiser is interested in deriving an estimate of the dependent variable value from the independent variables. This is true whether the appraisal is a traditional appraisal or one that employs a statistical valuation model. Using this approach is critical in any statistical analysis of the data as well. In both instances, the relationship between the dependent variable and the independent variables is the same. Although causation between variables is technically not provable, the association between these variable sets must make appraisal sense, and must be explained in those terms. For example, a single family residence with a greater total living area (an independent variable) would be expected to result in greater sale price (the dependent variable) than a smaller home; the appraiser needs to explain such a relationship whether performing a traditional appraisal with a manual adjustment grid, or when using a statistically-based appraisal valuation model.

**IN** a traditional direct sales comparison approach, the dependent variable would be the sale price per square foot (for example), while the independent variable would be those adjustment factors such as market trending, location, and any appropriate physical or economic element of comparison.

## Categories of Data

There are four different categories of data. Each is relevant in the description and analysis of data. Each of these categories represents measurement, and each represents a different level of measurement. These levels are useful in determining which statistical procedures are appropriate to utilize in understanding the data. The classic statistical data classification of data is as follows:

1. **Nominal (qualitative)**
2. **Ordinal (qualitative)**
3. **Interval (quantitative)**
4. **Ratio (quantitative)**

### Nominal Data

Nominal data uses numbers to categorize and classify information. While these types of data are useful, there are limitations beyond their use in classifying data. This type of data contains too little information to be used with many statistical techniques, but is useful in categorizing data. Examples can include neighborhood, construction type, roof material, or other variables that identify a property, but does not quantify it.

While each of these data points describes the data, they do not lend themselves to mathematical equations. It is, for example, impossible to add two of these data together. Even though nominal data cannot be utilized in their raw form in mathematical operations, they can be utilized to define the amount of number of observations that fall into a given category. For instance, in descriptive statistics such as frequency distributions or percentages, it would be useful to know the number of occurrences within a neighborhood defined as a nominal variable.

Nominal variables signify membership in a particular group, with no quantifiable difference implied between the groups. In other words, a nominal variable simply names a phenomenon (nominal derives from the Latin root *nomen*, which means name). A generic example is a variable for location based on two neighborhoods, labeled **1** and **2**, respectively. There is no inherent numerical difference between these values; they simply distinguish whether a property is located in one neighborhood or the other. The appraiser could have just as easily used the letters “A” and “B” to label these neighborhoods. .

Some additional examples of nominal variables are listed below:

- ◆ **Type of Shopping Center (e.g. Neighborhood Center, Community Center, Regional Mall, Power Center, Specialty Center, Lifestyle Center, Entertainment Retail Center, or Discount Center) and**
- ◆ **Type of Industrial Property (e.g. Warehouse, Flex, R&D).**
- ◆ **Type of Single Family Residence (Ranch, 2-Story, Bi-Level)**

Nominal variables allow a qualitative classification of property attributes. Nominal variables can be measured only as to belonging to a specific named classification, or not. They serve only to distinguish one item from another, but there is no order ranking placed on these identified differences. You can think of this type of variable as having only the power to **identify differences**.

#### **Ordinal Data**

Think of “order” when you see this classification of data. Data described as ordinal includes data that is organized as rankings, e.g. first, second, third, etc. Ordinal data can be very useful in situations where it is difficult to obtain more specific information about data. Ordinal data contains more information than nominal data, but is still limited from widespread usage in many statistical techniques. Ordinal data represents a measure of order within a hierarchy. Ordinal data measures a rising order of inequality within a category. Examples would include the comparison of any higher scale point to a lower scale point. It is important to note that the distance between the points is unspecified. The most frequent use of this type of data in real estate would include categories such as the quality of a given house. This ranking is demonstrated as follows:

1. **Excellent**
2. **Very Good**
3. **Good**
4. **Average**
5. **Fair**

## 6. Poor

The appraiser would still have some questions about the differences between the above scores. For example, how much better is a good property from an average property? Is the spacing between the rank categories consistent? The important concept to understand is that ordinal data incorporates the property of rankings, though the distance between scale values is undefined. Each ranking should be in order sequence, i.e. lesser to greater.

### ARRAYS

use ordinal, interval and ratio level data. You cannot array a nominal variable.

Assume that construction quality is included in a valuation model that allocates \$5,000 to each increment of that characteristic based on differences between the average selling prices of homes. For example, properties with a rating of excellent would receive \$20,000 ( $\$5,000 * 4$ ) while average properties would receive only \$10,000 ( $\$5,000 * 2$ ). The actual sales data may show something entirely different. For example, where excellent properties were \$25,000 above good properties, good properties are only \$3,000 above average properties. The scale simply does not reveal the difference in value between each condition level. This problem with spacing may be evident whenever there are more than two categories within the variable. Sometimes methods can be used such as paired sales analysis to create separate values for each quality level (but there are better ways to treat this problem using more advanced procedures such as regression analysis). Ordered data can be used as long as distance between variable values is not taken "as is."

Nominal and ordinal data are also known as **categorical** data because the primary information derived pertains to group membership (for example, if the property is part of Subdivision A or Subdivision B); these categories are either unranked (nominal) or ranked (ordinal). Additional examples that are germane to real estate would also include:

- ◆ **Socio-economic income classes in a community (e.g. poverty, low income, middle income, upper income, and affluent)**
- ◆ **Class of Office Buildings (e.g. Class A, Class B, and Class C)**

#### Interval and Ratio Data




Interval variables have quantifiable differences between each value. An example is the year of construction, where the number of years between two values represents the numeric difference between each year. A property with a year of construction of 1964 is 30 years older than a property with a year of construction of 1994. These variables have a meaningful scale yet, no absolute zero point on the scale. A non-real estate example of an interval variable is temperature measurements. On the Fahrenheit scale, we know that 60 degrees is 30 degrees warmer than 30 degrees, but it isn't twice as warm, since zero degrees is simply a benchmark; it does not represent the complete absence of heat.

In terms of relevance to appraisers, the year of construction variable is the most wide spread interval level variable. In terms of valuation modeling, it is recommended that the appraiser change the variable to an age-equivalent variable. Using an age variable, which is a ratio level variable, allows for adjustments to be more easily interpreted. The following valuation equation was derived from a simple regression model for industrial properties, using the year of construction:

$$\text{Sale Price} = -\$17,000,000 + \$24.99*(\text{Bldg Sf}) + \$313,884*(\text{Land Acres}) + \$8132*(\text{YOC})$$

While certainly "correct" statistically, using the YOC variable results in a large negative constant value and a large coefficient value for the YOC adjustment. When this YOC is replaced by the age of the improvements, the resulting valuation equation results:

$$\text{Sale Price} = \$368,054 + \$24.99*(\text{Bldg Sf}) + \$313,884*(\text{Land Acres}) + \$8132*(\text{Age})$$


Mathematically, both equations yield the same value for the subject. The statistics measuring the accuracy of the model are also the same, but the coefficient value for **AGE** is now easily interpreted, and could be used directly in any valuation adjustment. The constant, (represented

above by the red arrow) which measures the “base” value of the valuation equation prior to adjustments, is also smaller and more in line with the coefficient values.

Ratio variables are like interval variables, but the scale has an identifiable absolute zero point for the attribute. If you transform the **Year of Construction** to **Age**, then you have transformed an interval level variable to a ratio level data, since **Age** does have a “zero” point (i.e. **Age** = 0). For most descriptive statistical calculations we need to have interval or ratio type variables.

Both interval and ratio data are similar. Both involve measurable distance between numbers. Thus comparisons can be made between one distinct data point and the next. Ratio data differs from Interval level data because its ratios are meaningful. You divide a ratio level variable by another and have a meaningful ratio statistic; you cannot do this with interval level data.

Both interval and ratio data are often called **parametric data**. Parametric data allows for the use of statistical techniques that based on measures of central tendency such as the mean average.

**IF** you divide 1964 by 1994, you get 0.98. If you transform these interval variables into AGE (a ratio variable) using 2003 as the current year, you get 39 / 9, or 4.3. The first ratio is meaningless; the second ratio indicates that a building with an age of 39 years is 4.3 times as old as one that is 9 years old.

can

are

Ratio data incorporates both the concept of zero and the nature of interval data, providing a ranking of data that can be utilized numerically in the data analysis process. Ratio measurement is based on an ordered series of number rankings beginning with zero. The relationship between the number rankings indicates the absolute value of the data, such that a sales price of a \$1,000,000 retail building is twice as much as a retail building with a sales price of \$500,000.

For quantitative data (both interval and ratio data), the data value itself provides explicit information, in that equal differences have equal meaning. There is always a unit of measure involved, such as square feet, acres, roof pitch, or number of units. Most data of this type is continuous, in that it can always be measured more and more precisely, like square feet of area (e.g., 423.6 square feet). Or it can be discrete, like numbers of apartment bedrooms (1, 2, or 3) or number of baths (1, 1½, 2, 2½, etc.). With

all quantitative data, either continuous or discrete, the intervals between the values are quantitatively meaningful.

As noted, ratio data does allow multiplication and division (as well as adding and subtracting). For example, a home that has 2,200 square feet of GLA is 83% larger than a home that has 1,200 square feet. The difference between the size of the two homes, 1,000 square feet, has real meaning that can be measured and interpreted as a percentage difference (i.e., the difference in size divided by the size of the smaller house). On the other hand, a home with a quality rating of excellent and an ordinal value rating of 2 is not necessarily twice as valuable as a home with a quality rating of average that corresponds to a rating of 1.

The bottom-line difference between qualitative data and quantitative data is that the latter yields more **information**. Statistical analysis that uses interval or ratio data therefore yields more information than analyses using just nominal or ordinal level data.

**Limitations of Data:**

As has been detailed in the preceding descriptions, the type of data chosen limits the type of statistical techniques that can be utilized in analysis. Care must always be exercised to ensure that the proper data type is utilized in the decision-making process.

## Examples of Four Types of Data

### Qualitative

#### Nominal data

- Census tract
- Type of building
- Type of zoning

#### Ordinal data

- Quality
- Condition
- Utility

### Quantitative

#### Interval data

- Time (years, months)

#### Ratio data

- Sale price
- Size of building
- Age of building
- Size of parcel
- Number of stories

You have now completed the most theoretically oriented chapter in this book. The remaining chapters will now present applications using the analytical framework presented in this chapter; limited theoretical presentations will be provided as well, to better orient the reader.

# Chapter Three

## *Advanced Statistical Analysis*

---

The use of statistical analysis, both descriptive and inferential, must be understood by the appraiser, whether he or she has actually employed statistical analysis or if the appraiser is relying on a statistical analysis performed by another. It is particularly important that appraisers understand the types of analysis and the goals of the output when evaluating the analysis; the evaluation should be focused on the appraising veracity (i.e. reasonableness) of the analysis.

This chapter puts to task the analytic tools described in previous chapters, with the goal of providing a statistical guide for appraisers. This guide will hopefully help appraisers navigate the myriad of statistical tools and methods that are available in books or on computers. As stated before, much of the methodology is familiar territory, albeit with different names. Appraisers already perform much of the methods to be described in this chapter. The goal of this chapter is to place that knowledge in a systematic framework, with the ultimate goal of developing a useful method for readers to evaluate, build and utilize statistical valuation models.

Although the authors have had experience with valuation models on a macro-level (i.e., modeling entire metropolitan areas), the focus of this chapter will be on individual applications that fee appraisers can utilize in their everyday work environment. The authors do not want to discourage any appraiser who wishes to model entire neighborhoods or even multiple neighborhoods, but most applications will probably be on a more localized level, where an

appraiser is applying these modeling skills to value an individual property. The good news is that the principles employed either at the macro or micro level are the same.

### **Market Value Estimation**

Market Value Estimation (MVE) differs from the more traditional Automated Valuation Model approach (AVM). With AVM's, the appraisal process itself is generally automated. With MVE the modeling design concerns the actual real estate market mechanism present in a defined neighborhood. In general, the MVE model is much simpler, with fewer predictor variables. The reason for this simplicity has to do with the market transaction process. When one is purchasing a home, what are the common factors that influence all sales in the same neighborhood? Location, of course. Next, the size of the home, expressed as total living area square footage, number of bedrooms, bathrooms, or a combination of these. Other variables, such as year of construction, house style, subdivision, number of car spaces, lot size, and basement finished square footage can also play a significant role. All of these are considered Level One variables.

Other, perhaps less important variables (termed Level Two) such as fireplaces, garage type, and air conditioning may also be significant and be included. Level Three variables are those that tend to be subjective: construction quality, physical condition, and functional utility are examples.

Out of these 20 or so possible variables, the model may arrive at five or six variables that are actually used in the modeling process. The question to answer with MVE modeling is this: during the market transaction process, what common variables actually influence market value for typical purchasers? In a neighborhood where all homes fall between 1200 and 1350 square feet in living area, one would expect that living area square feet would probably not come into the valuation equation. Other variables, less related to dwelling size, would be expected to play a greater role.

Automated Valuation Models in most cases properly refer to Automated Appraising Valuation Models. The difference between this and Market Valuation Estimation is profound. The correct application of modeling real estate market transactions is just that: the modeling of the real estate market transaction, not the traditional appraisal process as we know it. This important application distinction also applied in general to other automation schemes, such as artificial intelligence and iteration-based models (where valuation is part of a step-by-step automated process). Most of these valuation schemes are also attempts to replicate the appraiser process, and hence the appraiser, with computers.

By changing the analytical paradigm away from appraisers and more toward the actual market process, the MVE modeling process can become an aid for appraisers, rather than be a threat.

Assessment analysts and mass appraisers for years have grappled with large, unwieldy statistical models to value residential properties in their jurisdiction. Some assessment-based models contain as many as 30 or more variables to predict the market value of residential properties. These variables can include add-on items such as hot tubs, front and rear porches, finished attics, wood decks, patios, and even built-in barbecues. Common sense forces appraisers to ask this questions when presented with these big, unwieldy approaches to valuation: is it reasonable to build a model with that many variables, when the typical real estate transaction may involve five or six common factors?

One of the rules of analysis is that any phenomenon under scrutiny that one wishes to predict or explain must be typical. "Typical" here means that one would expect that persons behave in a pattern that is repeatable and quantifiable. For example, if 100 persons purchasing 100 different homes in the same area purchased those homes for reasons that were all unique to their particular transaction, then it would become impossible to quantify the major factors contributing to real estate value. Fortunately, in most areas people sell and purchase real estate for common and quantifiable reasons.

The central theory behind MVE modeling concerns this market modeling approach. Central to this valuation approach is the controversial proposition that the true market value of a property is determined best when 100 persons bid on the same property, rather than the more traditional appraising axiom that states that the actual sale transaction of a property determines the truest market value. Under laboratory conditions, the true market value of a property could be derived if a large number of persons were allowed to bid on the same property. This bidding process would factor out conditions of a sale that affect its “arm’s length” nature; the value could be displayed as a distribution, with a mean sale price and standard deviation. One could then compare this sale price bid with other similar properties.

The problem with this scenario is that it is not real world. Home purchasers do not generally gang up to bid on single properties in most markets. The real estate purchase mechanism is most often pointed toward single offers for single properties. What mass appraising can offer, however, is the ability to approximate this multi-bid process. By valuing properties based on multiple sales in a given area, with general property homogeneity, the mass appraising estimate of value can often be the “next best option” to the multiple bid scenario.

Traditional mass appraisal methods have originated with the county assessors. The county assessor, however, is confined by the realities and goals of the ad valorem process. The county assessor usually builds mass appraising models with wide nets; that is, models with many variables. Part of this is due to the fact that the goal of the assessor valuation process is an equitable distribution of the tax burden across all properties. The major emphasis is with equitable treatment of properties vis-à-vis property values. For example, if the assessor consistently undervalues properties in a jurisdiction by 10%, the net impact is zero in terms of the effective tax burden. The property tax load is distributed across properties by the same effective tax rate (the mill levy would be higher for all properties by the same factor). The problem with assessor-based modeling is when properties are not valued equitably. For example, if one neighborhood is under valued as compared to a similar neighborhood in the



same jurisdiction, then homes in the former area would pay less than homes in the latter. Assessor models, therefore, focus on equity, rather than valuation veracity.

MVE models, on the other hand, focus on the valuation accuracy on an individual property basis, much as traditional fee appraising. This focus means that only variables that truly contribute to value are included, variables that approximate the market mechanisms determining real estate value in a particular area. An MVE model, adjusting for small differences in property characteristics, results in perhaps the truest market valuation approximation possible.

It is beyond the scope of this book to discuss the merits of the valuation debate concerning whether the truest “gold standard” of value is with a single market transaction or with an M.V.E. model. It is sufficient to note that M.V.E. modeling can offer appraisers a concise modeling approach with a reliable market value estimate in many instances. And by approximating the market transaction, it allows the appraiser to evaluate the model, using appraising theory.

### **Regression Explained**

Once the analytical scope is defined using an MVE approach, then the next step is to use a mathematical modeling program to create valuation coefficients based on a unique set of property characteristics. This set may vary between modeling areas, even with areas that appear to be similar in overall characteristics. The reason for this variation concerns the fact that the variables used in the model are selected by the market. The appraisers will enter the same amount of variables into each model, but the variables actually used by the models will be based on market performance. If fireplaces don't add to value in a particular neighborhood, then it will not be included in the valuation process.

Fortunately, there are software packages on the market today that can make the mechanics of this step fairly easy (although the interpretation of the output is not always easy). The most common modeling processes utilize regression analysis, which was briefly discussed in Chapter 3.

Regression is a powerful tool in inferential analysis and has received much attention over the past 40 years. Part of it has to do with the advent of personal computers and part of it has to do with its own elegance as a way to relate and interpret the relationships between two or more variables (such as sale price and living area square footage). The statistical method is termed “regression analysis,” and the engine driving it is termed “least squares” (in regression’s most common form).

### **Least Squares Approach**

The “least squares” process was developed hundreds of years ago as a way to explain the relationship between a set of two or more variables; in this case, these variables include items such as living area square feet, basement square feet, number of bedrooms and bathrooms, and the property’s value (or sale price). Regression modeling tests each variable entered into the model and performs two steps simultaneously. First, it selects which variables contribute significantly to value (based on statistical probability theory). Second, it creates a valuation formula that has the least amount of “distance” between the estimated sale price and the actual sale price of every property in the sale area. This distance, termed error, is based on all sales used; obviously, if one sale differs from all the others by a sizable amount, it can affect the relationship with these other sales. It is important, therefore, to first scrutinize all the sales in the area and determine that they represent typical homes in the sales area. Caution needs to be exercised, since the analyst doesn’t want to tailor the data to fit into a neat pattern. If enough sales fall out of the expected range, then the sale area may be too unstable to model with regression.

Simply put, the least square method is an analytical procedure that minimizes this distance between data along a given straight line. Imagine, for example, building a straight road across the United States that touches both the Atlantic Ocean and the Pacific Ocean at each end. The person building this road wishes to place it as close as possible to the 50 largest population centers. What steps could that person employ?

The first step would be to identify what centers of population exist in the United States, which, for this example, would involve identifying the 50 largest cities in the country. The second step would be to plan to build the road with shortest distances from every city. The method of least squares measures those distances from each city to the road and shifts the road location until the total distance of every city to the road is minimized. Least squares builds this road by first placing its midpoint in the center of the United States in terms of the 50 largest cities (say somewhere in rural Iowa). Obviously the eastern portion of the country will influence the placement of the road more than the western U.S., since more of the large cities are located in that region of the country.

The least squares method builds that line by pivoting it by its center point. Imagine the roadway moving like a seesaw, until the total distance as measured from the line to every city is minimized. Regression uses this method to relate variables.

Assume an example where we have 10 single-family sales in the same neighborhood. The variable of interest, the sale price, is called the "dependent" variable, in that its value depends on the values of the other variables in the regression analysis, say living area and basement area. These variables are termed "independent", since they are already determined in our analysis. They are givens, in that housing characteristics are provided by the county assessor, multi-list service, or other data source. The sale prices of our properties are also given, but in this analysis, we are assuming that the sale price is influenced by these independent variables in some way. Appraisal theory and economic theory, for example, would tell the analyst that larger homes would probably sell for a higher sale price than a smaller home, if all other factors

were basically the same. Note that nowhere have we invoked any statistical theory known only to statisticians. What is driving our analysis at all times is real property appraising theory.

Regression analysis uses the least squares method to build a model, which is a fancy name for a defined quantitative relationship between two or more variables. A related example of this would be the adjustment grid used on U.R.A.R. report forms for residential appraising. A regression model, in a slightly different form, is similar to an adjustment grid, in that it relates changes in one or more variables with a variable of interest (in our example the estimated sale price). The general form of the regression equation is this:

$$\text{Dependent Var.} = \text{Constant} + \text{Coefficient} * (\text{Indep Var \#1}) + \text{Coefficient} * (\text{Indep Var \#2})$$

Or, more precisely

$$\text{Est. Sale Price} = \text{Base Neighborhood Value} + (\$50 * \text{Square Feet}) + (\$25 * \text{Basement SF})$$

The values for living area square feet and basement square feet (termed coefficients) are constructed by the regression analysis. These values are then combined with (i.e. multiplied by) the values from our data source. The estimated sale price is then computed for each of our ten sale properties. Note two important facets with regression analysis: first, there is no place where the appraiser can enter what the owner or lender thinks that the property is worth. In other words, the process is “blind” in terms of the expectations of concerned parties. Second, there is an actual sale price lying out there. Our next step will be to compare our estimate of the sale price with the actual sale price for each of the sale properties.

It is this comparison of the expected sale price to the actual sale price that drives such statistics as the R-square, which many persons remember from their statistics courses (it may be the only thing they remember). Another important statistic is the Coefficient of Variation, which in

effect is the average difference between the actual sale price and the estimated sale price, usually expressed as a percentage (like 5%).

Once the model is evaluated for accuracy (which will be discussed shortly), the next step is to apply the model's values (called parameters) to the subject property itself. This is much like the process in the traditional direct sales comparison approach, where the subject property is compared with the sale properties. What is different is that here, the regression model provides the appraiser with the actual values for each significant valuation factor, such as living area square feet and number of bathrooms.

The rules and guidelines governing what constitutes a good array of comparable data is the same as with the direct sales comparison approach. If the subject property is similar to the sale sample, then the model will provide an estimate of value that can be used with some level of confidence. That degree of confidence is dependent on several factors, such as the homogeneity of the neighborhood (that is, how similar are the properties?), the degree of accuracy of the model, the number of sales, and the "appraisal veracity" of the coefficient values. For example, if the model has a living area square foot value of \$500 per square foot, and the neighborhood is a typical suburban tract subdivision, then the model is obviously poor. This is the case even if the statistical accuracy is good. The M.V.E. method also has an advantage over the direct sales grid, in that statistical measurements that relate to the comparability of the sale properties are also provided.

# Chapter Four

## *Output Statistics-Evaluating the Model*

---

### **Output Statistics**

How does the appraiser gauge the accuracy of the regression model? This is where the term “statistic” is often used to refer to numbers generated by the model that tell whether the “fit” of the model is within some range of acceptability. And what is the definition of an appropriate fit? This depends on completely separate guidelines away from the field of statistic and mathematics. This is where the particular discipline determines what ranges are acceptable. For example, hog farmers in eastern Colorado may have much less stringent accuracy requirements for predicting hog futures than commercial aircraft engineers designing metal fatigue tolerances of new metal alloys. One question worth asking when determining statistical tolerances is the cost of being wrong. The cost to the farmer for overestimating demand for pork is probably less than the cost of underestimating the tensile strength of a commercial jet. The need for accuracy for appraisers probably falls somewhere in between these two applications.

Many output statistics are available from statistical software programs. The R-square statistic is useful, in that it ranges from 0.00 to 1.00, with 1.00 being the desired value. It can be interpreted in our example as the overall fit of the arrayed sale properties. In terms of everyday practical use, the R-square value is often expressed in terms of a percentage, so that an R-square of 0.50 becomes 50%. That 50% can be thought of as the amount of variation explained by the regression model. The reason 1.00, or 100%, is desired is that it describes a perfect model; in other words, an R-square of 100% means that the regression model completely describes the sale price relationship between living area square footage and

basement square footage. In our 10-sale example, the estimated sale prices would match the actual sale prices exactly. This rarely occurs in the real world. It is, in fact, so rare that its occurrence is cause for concern!

Having a statistic such as the R-square might seem like the ultimate measure of the accuracy of a model. The general rule of thumb is that any regression model used in real estate valuation with an R-square greater than 60% is considered fairly accurate; obviously, the closer this statistic approaches 100%, the better. The astute appraiser, however, recognizes that this statistic describes the general accuracy of the valuation model. It does not provide a useful measure of how accurate the model is on an individual property basis. In other words, the model's average error is usually more useful to those concerned with the predictive power of a regression-based valuation model. In our example, an R-Square of 75% might be interesting, but a statistic that describes the average error (or in a positive light, the average accuracy of the model) is a better tool for evaluation. One such statistic that measures the average error is the C.O.V., or Coefficient of Variation, for models with sale in excess of approximately 25.

This statistic has been used for years by county assessor appraisers in their mass appraisal modeling. It is a good measure of how well the model can predict the value of homes in the modeling area. Obviously, it pertains to typical homes that do not have significant differences. The appraiser needs to know the physical characteristics of the subject home and how it relates to the sale properties, much as the direct sales comparison approach. Any special adjustments to account for differences between the sale property and the arrayed sales would have to be dealt with.

For regression models using fewer than 25-30 sales, the alternative form of this statistic, the Coefficient of Dispersion, is the preferred measurement. The difference between the two is that the C.O.V. divides the standard deviation by the mean sale price, while the C.O.D. divides the absolute deviation by the median sale price. The term "deviation" is synonymous with difference, in this case the difference between the actual sale price and the estimated sale

price. Both the C.O.V. and C.O.D. statistics derive an accuracy percentage based on dividing the average difference between the estimated and actual sale price by the average sale price. It's the way they calculate the average difference that makes them differ; the C.O.V. calculates the variation by squaring each deviation (the differences between the actual and estimated sale price) and then adding these squared deviations together, dividing by the number of sales and taking the square root. In other words, it takes the standard deviation of the sale price and divides it by the mean sale price.

The C.O.D., on the other hand, takes the absolute value of each deviation and sums these values, and then takes the average. This average absolute deviation is then divided by the median sales price.

Why all the squaring and absolute valuing? It's because if you simply add the deviations, you would get a sum of zero, since some of the deviations are negative and the others are positive.

The following illustrates the differences between each method with a small sales sample:

<b>Sale Number</b>	<b>Est Sale Price</b>	<b>Difference from Average</b>	<b>Squared Difference</b>	<b>Absolute Difference</b>
1	\$100,000	-\$5,000	\$25,000,000	\$5,000
2	\$105,000	\$0	\$0	\$0
3	\$110,000	+\$5,000	\$25,000,000	\$5,000
4	\$100,000	-\$5,000	\$25,000,000	\$5,000
5	<u>\$110,000</u>	+\$5,000	\$25,000,000	\$5,000
<b>Average</b>	<b>\$105,000</b>	<b>\$0</b>	<b>\$100,000,000</b>	<b>\$20,000</b>

If one simply added the differences in our example, the amount would be zero. Instead, the COV method squares the differences and the COD takes the differences without reference to



whether it is positive or negative. The COV difference results in a total squared differences of 100,000,000; this amount is then divided by 5 (the number of sales) and then the square root is taken. The result is 4.76%. The COD method takes the \$20,000 total deviation, calculates the average (\$4,000) and divides this by the median average \$105,000, resulting in a ratio of 3.8%. Looking at the arrayed data illustrates that the COD is probably more accurate with its 4% average error; the COV works better with larger sale amounts. Since appraisers are often faced with sale samples smaller than 25 sales, the COD is probably the preferred statistic.

Both the C.O.D. and C.O.V. take their respective deviation totals and divides them by either the mean sale price (C.O.V.) or the median sale price (C.O.D.). What happens if you divide by the wrong average sale price? It depends; generally, when you calculate the C.O.V. or the C.O.D. with large sale totals (greater than 25), there is not much difference, since the mean and median averages tend to be very close. In small samples, however, the mean and median can differ significantly, so caution is recommended whenever using these statistics when the sale total is less than 30. One way of remembering is that the COD uses the meDian as its divisor.

Most computer programs that calculate regression models provide the R-Square statistic as part of their standard output. Getting the C.O.V. or C.O.D., on the other hand, is more problematic. One way to approximate it with standard output is to take the regression standard error, which is always supplied, and dividing that number by the average sale price. If this method is used, it provides a good C.O.V. clone. Since it better approximates the C.O.V. (and not the C.O.D.), caution should be exercised with sale totals under 30.

Another cautionary note concerns any sales that are especially atypical. For example, a sale analysis uses ten sales, nine of these ranging between \$100,000 and \$120,000, with an average error of 6%. The tenth sale has a sale price of \$200,000 and an error of 10%. The C.O.V. statistic will be greater than the C.O.D. statistic, based on the influence of the \$200,000 sale on the mean sale price. On the average, the C.O.V. will tend to overstate variation and therefore understate the accuracy of the model if it is used in small sale samples. Of course, it is probably better to have a statistic that tends to overstate the average error of a model than to

understate it, so this approximation method can still be useful when getting the C.O.D. statistic is difficult.

### **What is an Accurate Model?**

What is an acceptable accuracy range for a regression-based appraisal model? Based on the experience of the authors, a range of less than 7% average error is considered acceptable in most cases. A model under 5% is considered very good, but in any case, appraising principles become the guiding protocol. Even if model accuracy is acceptable, the same principles used in traditional appraisal applications still apply. The subject property generally needs to be similar enough to the comparable sales for the valuation estimate to be reasonable.

For example, assume that the subject property in question is in a deteriorated condition, while all of the sale properties in the model are in average condition. The appraiser must then adjust the model estimate of value for the subject property by some factor accounting for the subject property's deteriorated condition. The appraiser may even decide that the market value estimate is not appropriate at all. In a subsequent chapter, the use of global adjustment factors is discussed, which can assist the appraiser in making adjustments for variables not directly used in the local model.

### **Selecting Variables**

Selecting which variables to include in the model is an important step, one where the appraiser needs to use appraising theory, not statistical theory, and some common sense. There are basically three levels of variables that confront the appraiser when modeling. First level variables are those that are considered fundamental to the valuation process, such as living area, bathroom counts, age, and basements. Second level variables are property amenities such as air conditioning, garage type, fireplaces, and swimming pools. Whether a variable is considered a first or second level variable can depend on local building characteristics and

market demand. Air conditioning, for example, is a second level variable in locales such as Denver, whereas in Phoenix, it is probably a first level variable. The difference between each level concerns data availability. The appraiser must decide at the outset whether the data set being used is sufficient in terms of available variables. For example, if living area is not consistently or accurately available, then the appraiser may decide that the data is insufficient for analysis. If, on the other hand, a second level variable is unavailable, such as fireplace, then it may still be possible to perform an adequate appraisal analysis. Again, it is important to note that appraisal theory is driving such decisions, rather than statistical theory.

Once run, the regression equation presents the valuation model in the following format:

$$\begin{aligned} \text{Estimated Sale Price} = & \text{Constant} & + & (\text{Characteristic\#1 X Value}) \\ & & + & (\text{Characteristic\#2 X Value}) \\ & & + & (\text{Characteristic\#3 X Value}) \\ & & + & \text{etc.} \end{aligned}$$

The value for characteristics reside in the sale file; again, these are variables such as basement square feet and number of bedrooms. The values to be multiplied against these characteristics are derived from the regression equation. It is these values that control the model and bear appraisal scrutiny. Replacing the above theoretical equation with a real world equation illustrates how the appraising analyst can bring his or her experience into play:

$$\text{Estimated Sale Price} = \$100,000 + (\# \text{ of Bedrooms X } \$5000) + (\text{Basement SF X } \$25)$$

**House A: 4 bedrooms, 450 SF basement, 2 fireplaces.**

**House B: 3 bedrooms, 300 SF bedrooms, 1 fireplace.**

**Estimated values:**

**House A:  $\$100,000 + (4 \times \$5000) + (450 \times 25) = \$131,250$**

**House B:  $\$100,000 + (3 \times \$5000) + (300 \times 25) = \$122,500$**

The values created by the equations are the estimates of market value by the model. The model consists of the edited data and the equation(s) derived to estimate the value of the property. The differences between the estimated value and the sale price of the arrayed sale properties allow the appraiser to evaluate how accurate the model is, based on our COV and COD description above.

**Time-** One of the Level One variables that needs scrutiny involves changing market conditions. Under the heading of “time” or “sales trend”, this factor needs to be accounted for, especially if the appraiser is required to look back further than one year from the effective date of the appraisal. Typically, the appraiser adjusts the comparable sale prices before physical, location, and economic factors are accounted for. Since M.V.E. modeling typically analyzes physical and location factors, the appraiser can test for any sale price appreciation while controlling for these other valuation factors.

Other changes over the sale period can also happen. These are changes that must be accounted for before any analysis is performed to estimate appreciation in sale price. These changes in the definitions of model variables can significantly affect the M.V.E. analysis. If the appraiser cannot adjust for these changes, the analysis may have to be terminated.

First and foremost, the valuation variables such as living area, basement area, subdivision number and other must remain consistent over time. Apples must remain apples. If, for example, the appraiser utilizes county assessor data in its database, and the assessor changes its definition of gross living area during the sales period, then an adjustment to the data will have to be made. Even if these differences can be controlled for, caution must be exercised when interpreting them. For example, assume your county data source includes finished

basement square footage in gross living area at the beginning of the sale period, but excludes it beginning with the second half of the period. Even if one can “pull out” the finished basement area from the living area totals of the first half, the data may still be difficult to compare. Are the finished basement totals reliable enough? Were they collected and verified in the same manner for both periods? It may be that the variable should be excluded from the model entirely. It may also be that the entire data base would need to be re-verified.

Another issue with time is clustering of data. The sales should occur fairly even throughout the time period in question. If the sale period covers two years, and nine out of ten of the sales occur during the first two months, then the derived appreciation (or depreciation) trend will be over-influenced by what occurs during the first few months. Building a trend based on that distribution is risky. Remember that the ultimate goal of any time analysis is to project the trend into the future, based on accurate and consistent historical data.

### **Adjustment Factors**

The use of the direct sales comparison approach requires that appraisers account for adjustments to comparable sales in a particular order. The issues of property rights conveyed, terms of sale, conditions of sale, and expenditures immediately after the sale are all pertinent to the MVE process. As stated before, the use of regression analysis applies only to adjustments made for physical and market (time) factors. Location adjustments with regression-based modeling are difficult and should not be usually attempted. Adjustments to sale prices to account for factors impacting the “arm-length” nature of the comparable sales should be determined and made prior to their use in the MVE model. This procedure should be followed for every computer-assisted appraisal.

Whether the MVE process is to create a comprehensive valuation model or to assist in deriving individual adjustment amounts for a traditional appraisal does not negate the need to follow

the approved methods of the direct sales comparison approach. Appraisers need to be ready to adjust the sales data file with MVE modeling as with any appraisal.

### **Strategies When Dealing with Time**

One approach to account for the changes in market price is simply to treat it as any other continuous variable, grouped with living area, basement area, garage type, etc.. As a continuous variable, time can be entered into the statistical analysis and controlled for. If time is deemed a significant variable, then it can be used by the appraiser analyst to adjust the dependent variable, in this instance the sale price of the sample properties.

For example, assume that 20 sale properties are used to calculate the value of homes in a particular subdivision. Next, assume that these 20 properties sold over a one-year period. If the statistical analysis determines that properties have been appreciating by the rate of 1% per month, then these properties would be adjusted by that amount, depending on the month the property sold. A property that sold during the first month of the sale period would be adjusted downward by 12% (or 1% times 12 months). A property sold six months into the sales period would be adjusted downward by 6%, and so forth. Assumptions necessary for model accuracy would include that the appreciation was constant (no significant seasonal fluctuations) and that other significant differences were accounted for. These other differences would include the usual array of property characteristics, such as living area, style, and age. In effect, a time trend amount that adjusts the sale price takes time completely out of the picture, as if every sale occurred at the end of the time period. Next, these adjusted sale prices would be used and the statistical analysis would be rerun to determine the estimate of value equation with all significant variables except for time (since it is already in the sale price).

An alternative method would be to simply leave in time as a variable and create the valuation equation with time in the model explicitly. In this instance, the estimate of value is as of the sale date, not at the end of the sale period. The overall effect is generally the same as the

previous method, but the appraiser must understand the subtle differences between both approaches.

The following example illustrates how time would be coded in our analysis:

Sale	Sale Date	Sale Amount	Coded Time Value
1	1/97	\$100,000	12
2	12/97	\$110,000	1
3	06/97	\$108,000	6
4	09/97	\$107,000	4
5	04/97	\$103,000	9

The beginning month of the time period is coded as the last month in the time period, while the last month of the period is coded as the number “1”; if it is a two year span, then sales occurring in the first month would be coded as “24.” The reason for this reverse order is that this allows for the time value derived from the model to be positive. Other time periods, such as quarters and semi-annual periods, could also be used.

If the model determines that time is a significant factor, then the appraiser can adjust the sale price much as adjustments made for time are performed on sale adjustment grids. Or the appraiser can simply run the model with time and the other variables included.

### **Building an MVE Model**

The following section will take the reader through a step-by-step process of creating a Market Value Estimation model. This can apply to either a comprehensive model where property valuations are the goal of the MVE model, or a limited set of variable estimates to be used in a traditional sales grid (this can also be used to make market-derived adjustments for external obsolescence). The purpose of this section is to define the steps that need to take place to actually create appraising models.

Having a clear understanding of the data set is important, in that the appraiser needs to ask questions of data sources to collect and edit the data before any modeling can commence. The first several steps detailed concern the preparation of the sale file. The following steps actually deal with the modeling process. These modeling steps will be presented in a generic format that can apply to any of the popular spreadsheet and statistical software packages.

These steps will guide the appraiser through the set-up process, and allow for a better understanding of what will be necessary to analyze the data. The alternative, where the appraiser simply collects whatever data is available and proceeds to “wring-out” whatever is there has been attempted by many before (including the authors). Based on our own experience and those of others appraiser analysts, we recommend the first process wholeheartedly!

### **Step 1-Designation of Modeling Areas**

The first step involves the appraiser understanding how the property data will be grouped. Many county assessor databases contain neighborhood identification numbers to distinguish between areas. These groupings are generally superior to those found in M.L.S. sources, in that M.L.S. data is often grouped by the collective opinions of brokers, rather than a systematic grouping analysis of neighborhood areas. This is not always the case; the authors recommend that the appraiser first check with both sources to determine which data source has the best grouping construct.

Grouped data is usually provided by the county in a comprehensive manner. Often the assessor’s subdivision or neighborhood numbers can be provided to the appraiser at a nominal charge; there are even Internet-based data sources that will download sales based on subdivision numbers. It is critical that the appraiser understands how the neighborhoods are identified. For example, a neighborhood may be known in the community by a single name,



while the county assessor has the neighborhood broken down into several sub-neighborhoods, based on subdivision filing number. These groupings can be combined by the appraiser and the entire neighborhood can then be modeled. If the sub-neighborhood represents significantly different houses in the same overall neighborhood, then it might be preferable to model the sub-neighborhoods separately. House style, age of the home, lot size, and other factors can clue the appraiser as to the veracity of combining sub-neighborhoods. The best approach, however, may be to simply speak with your colleague at the county assessor's office as to the nature of the neighborhood. Often these mass appraisers have already developed a grouping strategy and are more than willing to discuss this with fee appraisers. Again, the appraiser needs to make this determination, not the computer.

Based on the experience of the authors, the ideal range of sale properties usually range between 18 and 32 sale properties. This range is significantly dependent on the characteristics of the residential area. It may be possible to combine separate neighborhoods, treating each area with a unique neighborhood variable in the model. The method and result of this approach will be discussed below.

If the appraiser is faced with many small areas that are too small to be modeled separately, then a process called cluster analysis can be employed. The authors utilized this type of analysis in areas where the assessor had neighborhoods broken down into very small units. The neighborhoods were grouped based on the mean age, mean size and mean county actual value of properties in each neighborhood. The reason for using the assessor's estimate of market value is that this allows a shortcut to grouping the data. For example, if the average age of two neighborhoods was 10 years, and the average size was 2000 square feet of living area, then the county assessor average market price could alert the appraiser to other factors that are significantly different. In our example, if Neighborhood A had an average assessor value of \$100,000, while Neighborhood B had an average assessor value of \$500,000, then the appraiser would probably opt to not combine these two areas.

Cluster analysis assumes that the average age, size, and assessor value is a good representation of each neighborhood. If any of these three variables are affected by outlier properties, then the cluster analysis should not be used. It is beyond the scope of this hand book to detail the theory behind cluster analysis. The best approach appraisers can use is to monitor closely the grouping of neighborhoods. As with the adjustment process, appraisal principles guide the process, not the quantitative method used by the appraiser.

Once the sale properties are correctly grouped, the appraiser can then proceed to the next step, which involves coding variables such as house style, age, garage size, etc.

### **Step 2 – Creating the Sale File**

Once the modeling area is determined, the next step the appraiser needs to take is to create a sale file. This sale file will then be used in step three, which is when the appraiser will utilize statistical analysis to create a model.

A sale file can contain any number of sales in it. When an appraiser uses the traditional direct sales comparison approach, the sales collected and analyzed can be considered a manual sales file. All of the required data editing that appraisers employ with the direct sales comparison approach needs to be undertaken with MVE modeling. Some of the adjustment considerations, particularly those performed before adjustments for market and physical factors, will need to be accomplished before the sales file can be analyzed. Conditions of sale, property rights transferred, any significant expenditures undertaken by the grantee after sale, and any unusual financing terms are factors that could affect the market-based sale price. As with the direct sales comparison approach, any sale that requires too much adjusting may need to be excluded from the MVE analysis.

Once the data are verified internally for sale factors, the next step involves verifying and coding certain variables that will be utilized by the regression model. These variables include Level 1

factors, such as living area, basement area, number of bathrooms, house style, time trend and age of the improvements. Level 2 factors should also be evaluated; these include items such as fireplaces, garage type and size, air conditioning, evaporative cooler units, and other variables that may contribute to value. The major difference between Level 1 and Level 2 concerns what variables are necessary for basic residential dwellings. This can vary by region and local building demand, as stated before. Things like Living Area, Bathrooms and House Style are obvious Level 1 factors; all homes need these items. It is not critical that the appraiser gets the exact definition with a Level 1 or Level 2 variable; often, the market in a particular area may determine that a Level 1 factor is not significant, while a Level 2 factor is. An example of this would be a condominium project, where the unit size consists of two very similarly sized units. In this example, the model might determine that gross living area is not a significant valuation factor, while the presence or lack of a fireplace is. Another example would be a neighborhood where housing style consists of one type.

The important point that the appraiser needs to remember here is that the overall variable array needs to make appraisal sense. If the appraiser knows that a particular neighborhood borders a golf course, and that location feature significantly affects value, then the appraiser needs to make sure that variable is included in any MVE model he or she creates.

There is a third level set of variables that can be utilized by the MVE process. Level 3 variables are subjective factors such as condition, construction quality, and functional utility. These factors, while often important, can also indicate that a particular neighborhood or area contains a wide array of residential property types: perhaps too wide an array. Again, the appraiser must determine if the neighborhood's inventory of homes are homogenous enough (i.e. similar enough) to use MVE modeling.

Once data is collected and sale conditions are accounted for, the appraiser can create a sales file with valuation variables. The following is an example of a simple sales file in a spreadsheet:

### SAMPLE SALES FILE

Sale	Sale Price	Sale Date	Living Area	Basement	YOC	House Style
1	\$100,000	Jan-98	2,000	0	1976	Ranch
2	\$105,000	May-98	1,800	600	1978	2-Story
3	\$110,000	Jun-97	2,300	1,000	1978	2-Story
4	\$105,000	Sep-98	1,450	800	1976	Tri-Level
5	\$112,000	Nov-97	2,100	750	1977	2-Story

In this example, some of the variables are Level 1 and some are Level 2. The appraiser can determine several things about the arrayed sales immediately. First, the year of construction of the properties in this neighborhood are so similar (ranging from 1976 to 1978) that this Level 1 factor will probably not be significant. If the model does determine that it is significant, the appraiser needs to make sure that the resulting coefficient values make appraisal sense, and that they are not replacing or taking the place (acting as a proxy) for another variable. Age, for example, can sometimes behave like building style or quality. A good way to check for this is to evaluate the coefficient value. If the age value is not negative (i.e. older homes are more valuable), then there may be older, Victorian-styled homes in your neighborhood that have undergone renovations and remodeling. The appraiser can either model them separately or simply exclude homes beyond a certain year of construction. Regardless, the appraiser needs to know the data well enough that 100 year-old Victorian homes are not included with 25 year old tract homes.

Regression analysis needs the variables in the model to be continuous. Recall that in Chapter 3, there were three basic levels of data: nominal, ordinal, and interval. Also, interval level data was also termed continuous, in that the distance between values could be taken as is. The

other two types of data were used to identify (nominal) or order (ordinal). Variables such as living area, basement area, lot size, age, market (time), and number of bathrooms are already interval level (and therefore continuous). Other variables that may be important in the MVE process need to be recoded before they can be included.

For example, house style can be significant in determining property value. Living area square footage rates are often different for ranch style homes, as opposed to two-story homes or tri-level homes. But how can the appraiser include these housing styles as continuous variables? One approach would be to code house style as follows:

**Ranch = 1**  
**2-story = 2**  
**Tri-level = 3**

While this coding scheme is a convenient method from an identification perspective, recall that it is a nominal data array. The numbers used are meant for identification purposes only. One could very well reverse the coding scheme (with Ranch = 3 and Tri-levels = 1) with the same results. If the appraiser used this coding in the regression model, the computer would take the numbers literally; based on the coefficient value, Tri-levels would receive three times the value as ranch homes. These values would be affected more by the coding scheme (what the appraiser decided arbitrarily in terms of coding) than the market data itself.

The correct method to code nominal level data is to create separate variables for every house style less one. If there were four basic house styles in the neighborhood, then there should be three house style variables. The reason for using one less than the number of housing styles has to do with the internal mathematics of regression analysis: something appraisers don't need to worry too much about. What in effect happens is that the house style not selected becomes the "base" variable. The following example illustrates this coding process and results:

STYLE	VAR 1 (RANCH)	VAR 2 (2-STORY)	VAR 3 (BI-LEVEL)
RANCH	1	0	0
2-STORY	0	1	0
BI-LEVEL	0	0	1
TRI-LEVEL	0	0	0

RESULTING COEFFICIENT VALUES:

$$\begin{aligned} \text{Var 1 (Ranch)} &= \$5,000 \\ \text{Var 2 (2-Story)} &= \$2,500 \\ \text{Var 3 (Bi-Level)} &= \$3,000 \end{aligned}$$

RESULTING VALUATION EQUATIONS:

	Value	Var 1	Var 2	Var 3	Other Vars
Sale #1 (Ranch)	\$105,000	= (\$5,000*1)	+ (\$2,500*0)	+ (\$3,000*0)	+ 100000
Sale #2 (2-Story)	\$102,500	= (\$5,000*0)	+ (\$2,500*1)	+ (\$3,000*0)	+ 100000
Sale #3 (Bi-Level)	\$103,000	= (\$5,000*0)	+ (\$2,500*0)	+ (\$3,000*1)	+ 100000
Sale #4 (Tri-Level)	\$100,000	= (\$5,000*0)	+ (\$2,500*0)	+ (\$3,000*0)	+ 100000

The above equations illustrate how the mechanics work utilizing dichotomous variables; the reason the term “dichotomous” arises out of the nature of the 0,1 coding scheme. The fourth type of house, tri-levels, become the base home in the valuation equation. For simplicity, the authors have included all of the other valuation factors under the last variable category. Mechanically, the equations include all of the house style variables, but each sale property receives value for a house style only with one of the variables (or not at all with the case of tri-level houses).

This scenario works with other categorical variables, such as fireplaces, garage style, basement style, swimming pools, location factors such as golf course proximity. The appraiser needs to do this recoding before any modeling occurs.

Suppose that there are twenty sale properties in the sale file, and only two have fireplaces. The question arises as to how many “hits” must occur with a variable for it to be included in the modeling process. Based on empirical experience, the authors recommend that at least 5-6 sales must possess the property characteristic for it to become part of the modeling variables. Furthermore, if a variable is excluded from the modeling process because it has too few hits, then the appraiser should carefully evaluate such sales after the modeling process has created sale estimates, to determine that their exclusion does not bias (i.e. influence) the model.

For example, the appraiser is modeling a neighborhood that borders a golf course. The sale file contains 25 sales, with only three sales bordering the golf course. The appraiser models the entire neighborhood, but excludes any variable that identifies the three properties bordering the golf course because there are too few sales with that property characteristic (remember, such a variable would be used, it would be coded (0,1), with 1 representing properties that bordered the golf course). Now after the modeling process is complete, the appraiser now must evaluate the output. One of the necessary steps would be to look at the value estimates of those three golf course properties. If the sale price is less than the actual selling price for all three properties, then the appraiser should analyze the data further to determine if golf course proximity adds value (it probably does). The appraiser, for example, might exclude those three properties from the analysis and rerun the model; the reason for this that these three sales might “contaminate” the other, non-golf course sales. The results from this new model would represent home in the neighborhood without the golf course influence. The appraiser could then determine the added value of golf course proximity via traditional appraisal methods, such a paired sale analysis.

Coding and editing data is a significant step in the overall modeling process. It requires that the appraiser understand how data is processed by the regression analysis, but more importantly, it requires that the appraiser understand the valuation impact of each variable. Fortunately, this is exactly the same consideration that must be employed with every appraisal already produced.

A final step in creating the sales file variable set is easy. It involves looking at the arrayed variables that will be used by the M.V.E. model as the independent (i.e. valuation) variables. The appraiser needs to make sure that duplicative variables are not included together in the data set. For example, assume that Living Area SF, Bedroom Count Total, and Total Rooms are all included in the sales file. The appraiser needs to select one of these variables as the improvement size variable. Including all three would result in the model either partitioning the valuation effect of size across two or three variables, or developing coefficient values that do not make sense:

	Case A	Case B
<b>Living Area Square Feet Coefficient Value</b>	<b>\$20.00</b>	<b>\$75.00</b>
<b>Bedroom Total Coefficient Value</b>	<b>\$10,000</b>	<b>-\$5,000</b>
<b>Total Rooms Coefficient Value</b>	<b>\$5,000</b>	<b>-\$1,500</b>
<b>Real value of Property Size = \$50/sf OR \$20,000/bedroom OR \$10,000/room</b>		

Interpretation of above table: if the model used either living area or bedroom total or total room count in the model, the value of the coefficient would be correct, in that the model would be assigning a coefficient value to only one variable related to property improved area. By including all three, the real contribution of property improved area is either diluted (Case A) or confusing (Case B). In each case, the model does not make appraising sense. In this particular case, the authors have generally used living area square footage as the improvement size variable. This preference, however, is based on the real estate practices of their local markets (Colorado and Arizona).



Other linked variables that appraisers need to be aware of include the following:

**Condition vs. Quality**

**Bathroom # vs. Plumbing Fixture #**

**Car Space # vs. Garage Size (sf)**

The general rule of thumb is that if two variables in effect measure the same property characteristic, then one of them should be excluded.

# Chapter Five

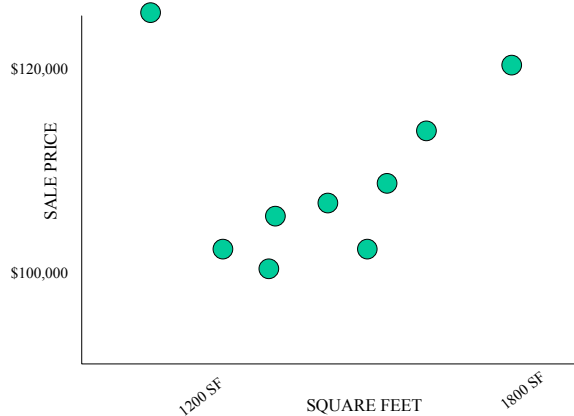
## *The Modeling Process*

---

### **Preliminary Tools for Evaluating Data**

The key to surviving the modeling process and to not get overwhelmed by the statistical output generated by today's software is to always keep focused on the appraisal aspect of the process. One method that the authors developed when using M.V.E. modeling concerned the use of visual output. Often a picture of the data can be more readily understood than a table of confusing numbers. Two basic "picture" methods are available with most software packages. These methods are easy to understand and, more importantly, they allow the appraiser to evaluate the data.

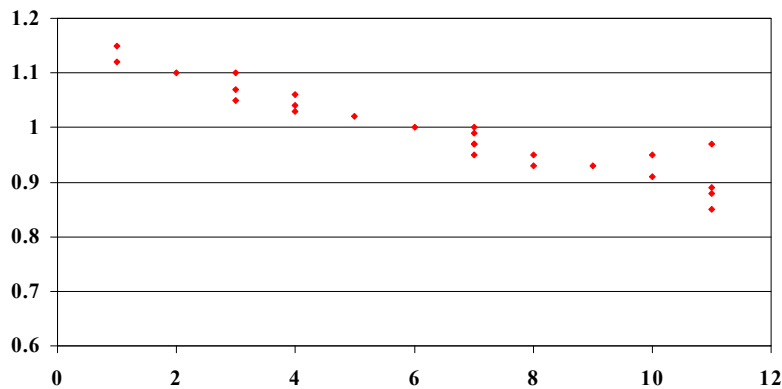
The first visual output is termed a "scatter plot." Scatter plots are exactly what they sound like; a plot of data that can tell an appraiser whether a sale property "fits" with other data in the sale sample. The following example illustrates a simple scatter plot of sale price and living area square feet:



Examining this scatter plot can tell the appraiser several things immediately. First, the sale prices range between \$100,000 and approximately \$130,000, with a corresponding living area size range of approximately 1,000 square feet and 1,800 square feet. What if the subject property is 3,000 square feet? What if the subject property's value estimate is \$175,000? Generally, caution must be exercised when the subject property lies outside of the range of critical Level 1 variables, such as living area and sales price. While this type of analysis can be gleaned from tables and arrays of numbers, graphs can present the data in a readily understood format. Another feature is that the graphical presentation such as the scatter plot can often identify outlier cases: sales that lie outside of the range of the other sales. For example, the above scatter plot has one sale that lies outside the sale price and living area range of the other sales. The appraiser can run other scatter plots or examine the sale file to determine possible reasons for the aberrant nature of the outlier. Perhaps it is the only sale property that has a basement. Or perhaps it is the only sale property that sold in 1998 (while the other sales occurred in 1997). What is important is that the appraiser can use this type of graphical analysis to determine if data needs to be either excluded from the analysis, or more importantly, if there are other factors that must be accounted for in the analysis.

The second graphical tool available to the appraiser concerns the M.V.E. output after the modeling process takes place. It involves graphing the ratio of estimated value to sale price:

## SALE RATIO USING MONTH OF SALE



$$\text{SALE RATIO} = (\text{SALE PRICE ESTIMATE}) // (\text{ACTUAL SALE PRICE})$$

The closer this ratio is to 1.0, the more accurate the estimate. Obviously, the average ratio is of interest to the appraiser in terms of evaluating the accuracy of a particular sale. Its real utility as an evaluative tool is when the appraiser graphs all sales using the sales ratio. With it, patterns can be readily checked, using a similar methodology as with scatter plots. Generally, the sale ratios are plotted with the sale ratio along the “Y”, with the independent variable, such as living area, month of sale, age or other valuation variable one wished to verify. The sale ratio should be randomly scattered along the 1.0 latitude of the sale ratio axis. Below are several examples:

Consistent patterns may point to interactive effects of one or more variables affecting the sale price estimate. For example, if a sale ratio graph is performed using month of sale, the appraiser can verify that any sale appreciation has been accounted for. If there is a pattern, then the model needs to be re-examined. For example, suppose the following sales ratio graph is created using month of sale:

The sale ratio clearly indicates that there is a consistent over-estimating of sale price occurring early in the sale year, while there is a consistent under estimation occurring in the latter part of the year. Faced with this result, the appraiser can return to the modeling coefficients (this will

be covered later in this section) and determine if first, is there a coefficient for market trends, and second, if there is, what is the value. Other tests can be undertaken to check to make sure that the model variable makes appraising sense. Obviously, if the appraiser suspects that the market is appreciating during the sale period and the model indicates that the market trend is either missing or depreciating, then further analysis must be undertaken.

Based on the experience of the authors, every model must include sale ratio checks on the following independent variables:

- 1. Living Area**
- 2. Age of property**
- 3. Subdivision (if more than one subdivision is analyzed)**
- 4. House style (if there is more than one)**
- 5. Sale Month (or whatever sale period unit is employed)**
- 6. Any other variable that is used by the model**

Generally, any problems with variables that do not possess random sale ratios can be dealt with by remixing the variables in the model; in other words, the problem often arises with variables interacting with one another that adversely affect one or both of them. Other times, variables are excluded because they do not meet the statistical requirements of the model. The appraiser can force such variables into the equation to determine if that solves the problem. The above variable checks should be undertaken even if the variable is not in the modeling process. For example, if a model does not include age, the appraiser still must check this factor using the above sale ratio analysis.

The above graphical analyses are easily available through spread sheet graphics. These methods provide quick, easy checks on model output. There are, of course, numeric checks that also must be undertaken, such as verifying the appraisal soundness of every coefficient value in

the model. Before this process is explained, it would be wise to first detail the actual modeling process itself.

### **Regression Modeling**

Once the sale file data has been edited for errors and outlier sales, the next step is the actual modeling process. Since it is outside the scope of this book to teach statistical software code, the step to create and run an M.V.E. model will be presented generically. The reader can then take these steps and apply them to whatever software package they have.

Regression analysis software usually requires that the dependent variable is specified separately from the independent variables. The appraiser needs to then list all of the independent variables that will be analyzed. This includes all recoded variables.

Once the data set is selected, the regression model usually requires that you indicate the method of variable selection. Variables can be forced into the model (termed “forward” or “backward” selection) or they can be entered one at a time (termed “stepwise”). Generally, the authors recommend that the appraiser specify stepwise selection. The reason for this is that by using stepwise selection, the variables are entered one at a time, with the most significant variable selected first (this is usually the living area variable or bedroom total variable). The appraiser can then gauge the impact of adding each variable to the model. How does the regression model know when to stop? This is usually determined by the software, which looks at the contributory value of each added variable. Remember, each added variable lessens the amount of unexplained variation in the relationship between the dependent variable (sale price) and the independent variables (living area, basement area, age, time, etc.). The model theoretically would add new valuation variables to the model until all of the variation is explained; in this case, the estimates of each sale price would be exactly the same as the actual sale prices (and the R-square statistic would be 1.00, or 100%). As previously stated, this almost never happens. Most good M.V.E. models will explain 60% to 70% of the

economic relationship between sale price and property characteristics such as living area, basement size, age, market trend, etc. (the R-square statistic), and have 3% to 7% average errors per sale property (the COD/COV statistics). The regression model will usually stop adding variables after the contributory effect falls below a certain threshold.

The regression model adds variables in the stepwise process in the following manner. First, it adds the “best” independent variable, based on the explanatory power of the variable as it relates to sale price. In most cases, this is the living area variable. The regression model then reruns the regression equation with the independent variable, and selects the next best independent variable. It then adds this variable to the model and reruns. This cycle is repeated until all of the best variables are included. What happens if the model finds no independent variables to include? Then the modeling process is terminated.

The appraiser can then do three things: not continue with the modeling process, select a new set of variables, or lower the threshold that allows variables to enter the model. This threshold is usually specified by a “pin” or “pout” value. This value is usually set by the software at .05 or .10, which can be interpreted as the probability that the variable to be added actually contributes to the model (if this explanation sounds fuzzy, do not worry). The important thing to remember is that most software packages allow the user to change the criteria used by the regression analysis to include or exclude variables. Adjusting these upward will sometimes allow marginal variables to enter the equation. It is beyond the scope of this book to discuss the pitfalls of allowing in variables by raising the inclusion threshold. The major pitfall is that a variable that does not really contribute to the value of a property will be allowed into the model, with the result that the model will contain a bogus coefficient value for that variable. Although this scenario is possible, the appraiser has one factor in their favor; that is, the appraisal experience possessed by the appraiser. The appraiser knows the market better than the M.V.E. model. The fact that the coefficients can be evaluated based on their appraisal veracity means that the appraiser can simply look at the coefficient value and make a determination as to the “reasonableness” of the model. For example, if the model returns a

coefficient value of \$50,000 for a fireplace, the appraiser already knows there is a problem with the model. The chance that a variable value slips through the modeling process without actually contributing to value in the real world is far greater in applications outside of real property appraising. The appraising profession already has years of real world theory and experience that can be utilized to evaluate the output from the regression process. It is simply a matter of setting up the modeling process correctly to allow the appraiser to adequately evaluate the model process from both an input and output perspective.

#### **Step 4 – Output (Verifying)**

This book has already discussed output statistics such as the R-square and the COD/COV. The other important output step, other than evaluating the coefficient values, involve using scatter plots and sales ratio. If all of these factors make appraising sense, then the model can be used as either a supportive tool to the traditional appraising process, or as a stand-alone valuation model.

The important point in verifying the output of ANY appraisal model is that the values for ALL of the valuation variables must make appraising sense. Whether they make statistical sense can be verified with the above tools. Any valuation variable that does not make sense renders the model and its output unusable from an appraising perspective. The appraiser needs to exclude the “bad” variable and run the model program again, or the sale file needs to be examined for outlier sales that do not belong with the other sale properties. The third possibility is that the neighborhood cannot be modeled using regression analysis. Hopefully, in the preparation phase of the modeling effort, this fact would have been discovered by examining the descriptive data about the neighborhood.



## **Summary**

This chapter focused on the actual “nuts and bolts” of regression modeling utilizing residential real property. The purpose of this chapter was to get the appraiser familiar with mechanically constructing an M.V.E. model. Important points to remember include understanding that the output of the model must be such that the appraiser can verify the appraising veracity of the model. There are methods of editing and coding data that allows for non-continuous data to be included in the model.

Evaluating the model for appraisal veracity is a vital step in the modeling process. The appraiser can and must be the final judge in determining that the output from such models are true indicators of value and not simply interesting mathematical relationships.

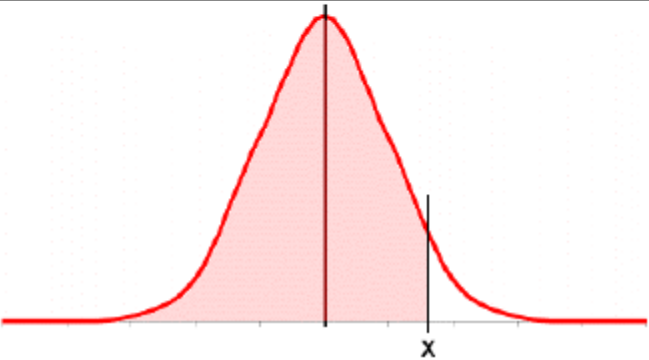
# CHAPTER SIX

## *PUTTING EVERYTHING IN CONTEXT*

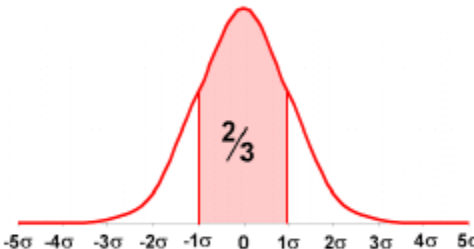
To help you remember the relationship to all of the various statistics and concepts presented in the Real Estate Statistics Without Fear, here are some helpful definitions and relationships:

MEASURE	DESCRIPTION	FORMULA
<b>Mean/ Average</b>	The sum of the sample values divided by the sample size. The sample statistic $\bar{x}$ is an unbiased estimate of the population mean $\mu$ .	$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n}$
<b>Median</b>	The middle value after the sample values have been sorted into order by magnitude. If there are an even number of values in the sample, the average of the two middle values.	
<b>Mode</b>	The most common value in the sample.	
<b>Range</b>	The difference between the largest and smallest values in the sample.	
<b>Variance</b>	An estimate of the variation or dispersion of the process from which the sample was drawn. The sample statistic 's <sup>2</sup> ' is an unbiased estimator of the population parameter $\sigma^2$ .	$s^2 = \frac{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}{n - 1}$
<b>Standard Deviation</b>	The square root of the variance. Often preferred as a measure of process variation. The sample statistic 's' is an estimator of the population parameter ' $\sigma$ '.  This method of calculating the standard deviation is known as the Root Mean Square Error (RMSE) method.	$s = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}{n - 1}}$

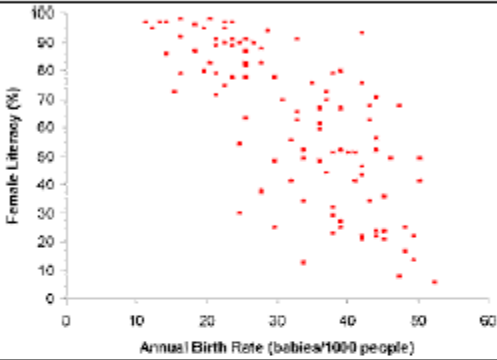
## NORMAL DISTRIBUTION

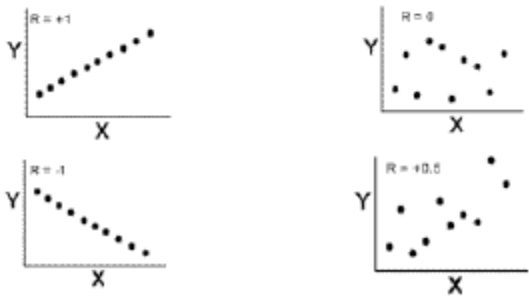
DESCRIPTION	<p>The output of many types of processes can be represented by the normal distribution. The normal distribution is often used to estimate the proportion of process output that will lie within a specific range of values (for example, the proportion of the process output that will be within specification).</p> <p>The normal distribution shape results from the 'common cause' variation in a process. It can be used to reveal the presence of 'special causes'.</p>
DISTRIBUTION	

## USEFUL VALUES FROM THE NORMAL DISTRIBUTION

DISTRIBUTION	<p>About two thirds of the process output lie within one standard deviation either side of the process mean. Other values to remember:</p> 														
VALUES	<table border="1" style="width: 100%; text-align: center;"> <thead> <tr> <th>Standard Deviations either side of the mean</th> <th>Approximate amount</th> <th>Exact amount</th> </tr> </thead> <tbody> <tr> <td>±1</td> <td>two-thirds</td> <td>68.27%</td> </tr> <tr> <td>±2</td> <td>95%</td> <td>95.45%</td> </tr> <tr> <td>±3</td> <td>99.7%</td> <td>99.73%</td> </tr> </tbody> </table>	Standard Deviations either side of the mean	Approximate amount	Exact amount	±1	two-thirds	68.27%	±2	95%	95.45%	±3	99.7%	99.73%		
Standard Deviations either side of the mean	Approximate amount	Exact amount													
±1	two-thirds	68.27%													
±2	95%	95.45%													
±3	99.7%	99.73%													

## SCATTER GRAPHS & CORRELATION

CHART TYPE	Scatter Graphs
DESCRIPTION	Scatter graph are used to explore associations between two variables. They are normally used when the data form natural pairs, and where there are many pairs. Standard X-Y graphs are normally used to explore a causal relationship between X and Y; typically X is varied systematically and the effect on Y measured.
EXAMPLE	
COMMENTS	Female Literacy and Birth Rate are associated, but there is not necessarily a causal relationship.

Correlation	
DESCRIPTION	Correlation is a measure of the strength of the relationship between the input and the output of a process. Correlation is measured by the 'Pearson Product Moment Correlation', known as 'R'. The value of 'R' varies from +1 to -1.
EXAMPLES	
COMMENTS	An R value of + 1 is perfect correlation. Values between -0.5 and + 0.5 show weak relationships.

# Levels of Measurement and Measurement Scales



Differences between measurements, true zero exists

**Ratio Data**

Highest Level

Differences between measurements but no true zero

**Interval Data**

Strongest forms of measurement

Ordered Categories (rankings, order, or scaling)

**Ordinal Data**

Higher Level

Categories (no ordering or direction)

**Nominal Data**

Lowest Level

Weakest form of measurement

# **C** HAPTER SEVEN

## ***STEPS TO COMPETENCE***

---

In this book, we have identified some basics of statistical analysis. The purpose of statistical analysis is to identify, calculate, and interpret phenomena. In appraisal practice these events are often market transactions that help determine value in real property. To be studied, these transactions (and the variables involved) must be clearly defined and understood. The importance of appraisal knowledge and experience to the proper incorporation of statistical analysis in the appraisal process is clear—appraisal theory, not statistical theory, always drives the process.

Classifying, arranging, and just looking at data are key parts of statistical analysis. Data can “name,” “order,” or measure phenomena. Data can also be continuous, discrete (choice), or binomial (yes/no). Often certain nominal or ordinal data can be made to behave as interval/ratio data. Similarly, continuous data and discrete data can sometimes be “converted.” How data are structured is a part of the analytical process.

Appraisers must be able to interpret whatever output is placed before them correctly, whether that output comes from a statistical software package, from an available statistician on duty, or from a canned AVM. Understanding the structure of statistical analysis is fundamental to analysts and users of appraisals. USPAP requires that the appraiser understand the data and analysis used in the appraisal process and to report that information in a manner that is credible and not misleading. For many readers, a report containing statistical output with confusing, complicated, or “black box” algorithms and unqualified conclusions can be misleading or just plain wrong.

### **Where do you start?**

- 1. Understanding of Appraisal**
- 2. Augmentation with Basic Statistics**
- 3. You must work with data in a practical environment**

This handbook is not meant to replace a standard statistics textbook, which should be a fundamental reference. Nor will it replace an understanding of basic appraisal principles and practices. Rather, it should be seen as a supplement to such knowledge and a guide through specific issues and problems.

### **Dangers:**

Statistical programs are so powerful that output can be generated that can make a poor outcome appear to be a good result...."a little knowledge is a dangerous thing"

### **Introduction:**

Appraisers need to consider both classes and texts to augment their knowledge in the profession. While there are numerous books on statistics that a student may consult, it is critical that the text be accessible and appropriate for the late career student. The following texts are available from a variety of sources, and each is useful and appropriate for those who wish to examine statistical techniques on their own. The authors have carefully considered the manner in which the narrative material is presented, and selected from more than 40 books, the following texts. A narrative on each text and its suitability is provided as follows:

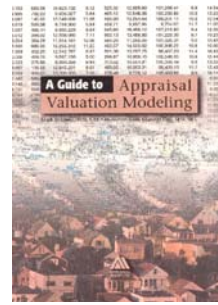
**A Guide to Appraisal Valuation Modeling**

**Mark R. Linne, MAI, CAE**

**M. Steven Kane**

**George Dell, MAI, SRA**

**2000: The Appraisal Institute**

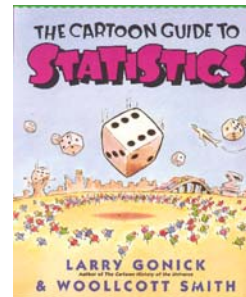


This text, written by two of the authors of this text, provides an overview of the entire Appraisal Valuation Modeling process. The book focuses on introducing readers to the mathematical modeling of market behavior. The handbook provides historical perspectives, statistical fundamentals, support for assertions of causality in appraisal reports, and the basics of regression analysis and model construction. The topics are brought together in a case study on the valuation of lots in an actual residential subdivision. Throughout the text, the authors highlight the interplay of evolving statistical theory, traditional appraisal standards and practices, and simple common sense.

**The Cartoon Guide To Statistics**

**Larry Gonick and Woolcott Smith**

**1993: Harper Perennial**



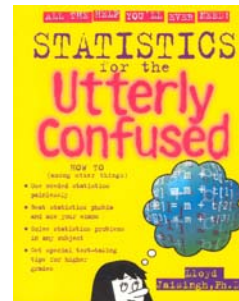
This book is perhaps the most reader friendly of the texts available on the subject of statistics. The text is engaging and provides excellent insight and background for the reader. The text describes itself as covering all of the central ideas of modern statistics, including the summary and display of data, probability in gambling and medicine, random variables, Bernoulli Trials, the Central Limit Theorem, hypothesis testing, confidence interval estimation, and much more. The text is easy to read and the illustrations provide both humor and clarity of understanding.



**Statistics For The Utterly Confused**

**Lloyd R. Jaisingh, Ph.D.**

**2000: McGraw-Hill**

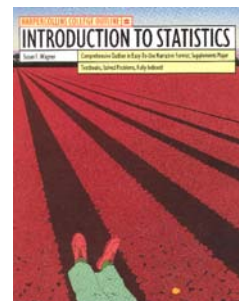


Similar to the “Cartoon Guide To Statistics”, this book provides a reader-friendly introduction to statistics. It is especially written for students in introductory non-calculus-based courses, and is targeted towards professionally who use statistics in the workplace. The author is a teacher, and the text is written to enhance understanding. This text also includes multiple-choice questions at the end of each chapter that is helpful in gauging how well the reader has grasped the essentials of the statistical material.

**Introduction to Statistics**

**Susan F. Wagner, Ph.D.**

**1992: HarperCollins Publisher Inc.**

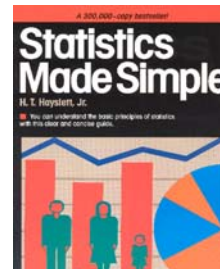


This text is touted as a college outline text, and while it is not entirely as engaging as the previous texts presented, it is nonetheless a fairly easy read, providing a comprehensive outline of statistics in a narrative format, that can be used to supplement other texts. The book is very well indexed and includes solved problems for the reader.

### **Statistics Made Simple**

**H. T. Hayslett, Jr., M.S.**

**2001: Broadway Books**



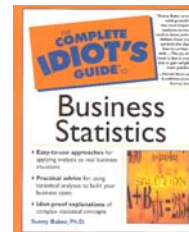
This guide provides a concise and straightforward introduction to statistics, and provides a platform for understanding basic principals with only a basic algebra background. The text contains exercises in each chapter that help the reader in reviewing the knowledge presented. In addition, the text contains step-by-step directions for applying statistical techniques.

### **The Complete Idiot's Guide To Business**

#### **Statistics**

**Sunny Baker, Ph.D.**

**2002: Pearson Education Company**

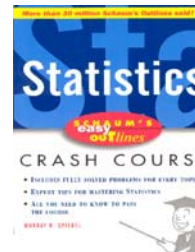


Similar to both of the texts already presented, this text uses both comic's graphics and easily understood statistical presentations to painlessly present the major topics in statistics to the reader. This book is somewhat more focused towards business statistics than the earlier texts, and this focus provides a real world framework that is helpful to appraisers.

**Statistics: Crash Course**

**Murray R. Spiegel; David P. Lindstrom**

**2000: McGraw-Hill**



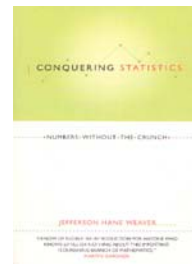
This text is part of Schaum’s Outline Series, and provides a condensed and abridged version of a larger text: Outline of Theory and Problems of Statistics. The text is built for quick, effective study, and though abridge, covers all of the statistics that are necessary for the real estate appraiser. The text includes

**Conquering Statistics:**

**Numbers Without The Crunch**

**Jefferson Hane Weaver**

**2000: Perseus Publishing**

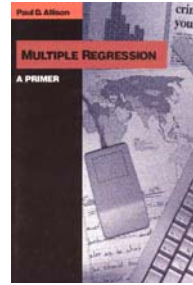


This book is an excellent introduction to statistics. It is somewhat different than the other books in this survey, in that it relies primarily on narrative, with no equations, and virtually no mathematical formulae. The text tells the story behind the statistics, and provides a fascinating backdrop to the history and understanding of the reason these numeric is important. The book is written for the statistically scared, and uses simple mathematical principles to breakdown the often-confusing concepts behind probabilities, means, and samples.

### **Multiple Regression: A Primer**

**Paul D. Allison**

**1999: Pine Forge Press**



This is an introductory text for those who want to begin understanding the uses for which multiple regression can provide insight. The book is constructed in a question and answer format, and uses modules to focus instruction to specific areas of statistics. Though the book focuses on methods for the social sciences, the presentation is appropriate for a broad-based introduction to basic statistical methods.

### **A Mathematician Reads The Newspaper**

**John Allen Paulos**

**1995: Anchor Press**



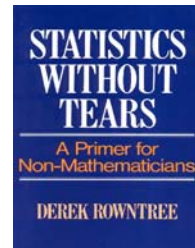
Less a book on statistics, and more an analysis on how numbers can be mis-understood, this text is nonetheless an interesting examination of how a lack of statistical knowledge can hinder our understanding of them.

**Statistics Without Tears:**

**A Primer for Non-Mathematicians**

**Derek Rowntree**

**1981: Charles Scribner's Sons**



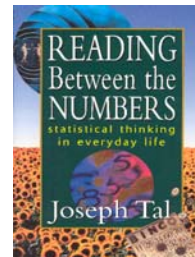
This book is a fine mix of statistics grouped with an understanding of the role statistics plays in our everyday lives. The text operates on the principal that statistics can be learned without having to perform calculations. The book provides an introduction to the basic concepts and terminology of statistics, providing the reader with the ideas of the subject before getting involved in the associated calculations.

**Reading Between The Numbers:**

**Statistical Thinking in Everyday Life**

**Joseph Tal, Ph.D.**

**2001: McGraw Hill**

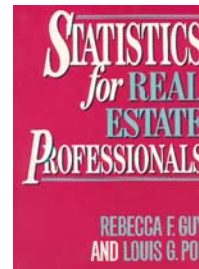


The goal of this text is to bring statistics down to earth for the average reader. The book focuses on the story and psychology behind the numbers that we see in our everyday lives. The book provides an understanding of the manner in which statistics are used in our decision-making.

**Statistics For Real Estate Professionals**

**Rebecca F. Guy and Louis G. Pol**

**1989: Quorum Books**



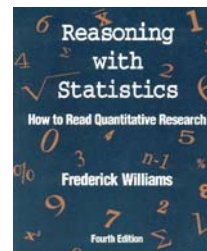
This excellent text is the only one of the books in our survey that specifically examines statistics as it relates to real estate and valuation. The text focuses on the role that statistics plays in measuring neighborhood change, economic value, market effects, and the role of statistics in measuring fluctuations and variances in the value of real estate. The book provides a strong introduction to statistics, and then covers descriptive statistics, index construction, probability, the normal distribution, sampling and inferential statistics. This is a superb book, and should be one of the first for those who want to integrate statistics and real estate. It should be noted that this book is out of print from time to time.

**Reasoning With Statistics:**

**How To Read Quantitative Research**

**Frederick Williams**

**1991: Holt, Rinehart and Winston, Inc.**

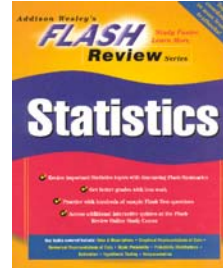


This text is intended for students interested in understanding quantitative methods with a focus on the social sciences. Given this focus, the book provides good explanation of statistics without resorting to complex formulae.

**Statistics: Addison Wesley Review Series**

**Julie Sawyer**

**2002: Addison Wesley**



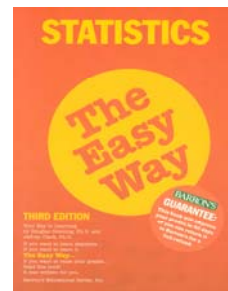
This book is a review of basic statistics, and provides a good outline with questions at the end of each chapter. The book should be utilized in tandem with one of the other texts noted above in order to gain the greatest understanding of statistical methods.

**Statistics The Easy Way**

**Douglas Downing, Ph.D. &**

**Jeffrey Clark, Ph.D.**

**1997: Barron's Educational Series, Inc.**



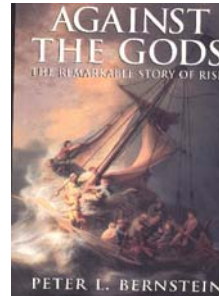
This text provides the fundamentals of statistical theory and applications. The book is especially relevant for those who have limited knowledge of mathematics. The book is structured with exercises and questions with answers at the end of each chapter. This is a good book that covers all of the elements that are needed to understand and apply statistical techniques.

**Against the Gods: A Remarkable Story of**

**Risk**

**Peter L. Bernstein**

**1996: John Wiley & Sons, Inc.**



This remarkable book provides an overview of the origins of statistics and provides strong biographical information on those who pioneered the development of the statistical techniques that we use today. A very interesting treatment that underscores the human element in the statistical process.

**Forgotten Statistics**

**Douglas Downing, Ph. D.**

**Jeffrey Clark, Ph. D.**

**1996: Barron's Educational Series, Inc.**



# CHAPTER SEVEN

## *GLOSSARY*

---

### **Absolute deviation**

The difference between an observation and the measure of central tendency (such as the arithmetic mean) for that array, without regard as to the sign (positive or negative) of the difference. For example, if the observed value is 6 and the mean of these observations is 10, then the difference (or deviation) would be  $6 - 10$  or  $-4$ ; dropping the negative sign would yield the absolute deviation, which in this case would be 4.

### **ANOVA (analysis of variance)**

Generally, just a two-variable regression. This is a well-developed body of techniques for analyzing and comparing data without the use of multiple regression. It is particularly useful for testing and comparing group mean scores and variances and for integrating categorical and continuous variables in an analysis. Considered outdated by some, it is well-suited for many appraisal applications building on the underlying economic and appraisal theory, and it is a good pedagogical tool. It can be especially useful in estimating transaction adjustments, sale conditions (motivation), market conditions (time), and space (location). The concepts underlie visual graphic analysis.

### **Array**

A set of data identifying some phenomenon. A set of sale properties would include an array of sale prices as well as other arrays pertaining to property attributes.

### **Attribute**

A characteristic of a person, thing, or event. In appraising, an element of comparison.

### **Average deviation**

The arithmetic mean of the absolute deviations of a set of numbers from a measure of central tendency, such as the median. Used to calculate the coefficient of dispersion.

### **AVM**

Appraisal valuation model or automated valuation model. Both terms refer to electronic algorithm-assisted valuation decision making.

### **Base-home approach**

A method of appraising single-family residential properties whereby each residence to be appraised is compared with one having common or typical characteristics and of known value, called the base home. The differences between the two in terms of attributes such as size, functional utility, condition, and other

factors are weighted by the appraiser to derive the subject property's valuation conclusion. This approach forms the basic construct for multiple regression valuation methodology used in appraisal valuation modeling.

### **Bias**

A statistic is biased if it systematically differs from the population parameter being estimated, such as the average sale price. An example in appraisal valuation modeling would be if the model results consistently overvalued properties that sold above the average sale price and undervalued properties that sold below the average sale price.

### **CAMA (computer-assisted mass appraisal)**

A system of appraising property, usually only certain types of real property, that incorporates computer-supported statistical analyses such as multiple regression analysis and adaptive estimation procedures to assist the appraiser in estimating value.

### **Categorical variable**

A variable that assigns data based on group identification. The group identification can be ranked or simply be an identifying group label. The power of statistical comparison is limited because the information provided by the categorization is limited. The data values are termed discrete, in that there are no values between each data category.

### **Central tendency**

The result of data clustering about a value or values. In a normal distribution, the central tendency is constant when measured by the mean, median, or mode. These three statistics are termed measures of central tendency.

### **COD (coefficient of dispersion)**

The average deviation of a group of numbers from the median expressed as a percentage of the median. In ratio studies, the average percentage deviation from the median ratio. This statistic can be effective as a measure of accuracy in regression modeling small data sets (less than 25 sales).

### **Coefficient (in a regression)**

The multiplier applied to the variable. (Analogous to an adjustment factor, such as \$54 per square foot. The \$54 is the coefficient; the square foot is the independent variable.)

### **Coefficient of determination**

The square of the correlation coefficient. Also, a term identifying the  $R^2$  statistic, which is used to evaluate regression models. This statistic measures the amount of total variation explained by the regression model; its utility when estimating point values, such as with appraisal valuation models, is perhaps less useful than statistics that measure the average error, such as the COV or COD. More recent statistical references are dropping this moniker in favor of using the term  $R^2$  directly to identify this statistic. It also avoids confusion with other statistical tools that use the term coefficient, such as the coefficient of dispersion or the coefficient of variation.

**Coefficient of variation (COV)**

1. A measure of relative variation. It is given by the ratio of the standard deviation of a set of data to the mean of that set of data. This is a way of normalizing the statistic so that it can be compared across all data sets.
2. A standard statistical measure of the relative dispersion of the sample data about the mean of the data; the standard deviation expressed as a percentage of the mean. In regression modeling, this statistic can be estimated by dividing the standard error of the estimate by the average sale price of the sales data set.

**Collinearity**

Correlation between two independent variables in a regression. (Same as multicollinearity, if more than two variables are involved.)

**Continuous variable**

A variable that can never be measured exactly. Examples include living area, improved area, and age. Examples of non-continuous data include the number of fireplaces, bedrooms, and bathrooms. It is sometimes useful to group continuous data into discrete groupings, such as grouping single-family residences into living area groups of 1,000-1,500 square feet, 1,501-2,000 square feet, and over 2,000 square feet.

**Correlation**

1. Association between two variables or a measure of such association.
2. A statistical phenomenon where an identified phenomenon implies a corresponding result in another phenomenon. One example would be the size of a property affecting the sale price. In most instances (and controlling for other important factors), the larger a property would garner a greater sale price. This type of correlation between property size and sale price can be measured and evaluated. For example, it may involve a partial correlation, where the relationship is less clear; this often occurs due to other factors affecting the dependent variable.

**Data**

Groups of observations; either qualitative or quantitative.

**Data set**

Refers to a sample of data used in a statistical analysis such as regression modeling. An appraisal example would be the sales file used in an appraisal valuation model.

**Dependent variable**

The variable that may be or is believed to be predicted by or caused by independent variables; response variable; explained variable.

**Descriptive statistic**

A statistic (number) used to describe a data set. (See inferential statistic.)

**Deterministic**

Certain, as opposed to probabilistic.

**Deviation**

The amount an individual observation differs (i.e., deviates) from the mean.

**Discrete data**

Data that are categorized into two or more groups. Examples include number of bedrooms, fireplaces, garage spaces, and bathrooms. Discrete data that belongs to only two groups, such as swimming pools (0  no, 1  yes), are termed binary data. The type of data determines what type of statistical tests can be performed on the data set. Discrete data can sometimes be treated as continuous data for mathematical analysis.

**Discrete variable**

Nominal and ordinal variables are discrete, in that the numeric identification for each is a discrete total and does not include any inherent information about the magnitude of differences between each group.

**Dispersion**

A generic word for “spread.” See standard deviation, standard error, and interquartile range.

**Disturbance**

The “error” or stochastic part of the analysis. It is the way of facing the reality that economic data are not deterministic and will not provide exact answers.

**Econometrics**

The application of mathematical and statistical techniques to economic situations. It can be considered the “scientific” approach to valuation. Its primary importance in the appraisal context is that it accommodates the analysis of electronically delivered market data from a comprehensive database. Traditional methods can be considered more the “art” of appraising, well-suited to difficult-to-obtain, difficult-to-verify market information.

**Empirical**

Using scientific evidence from the real world (i.e., using mathematics and statistics instead of language-based arguments for research and analysis).

**Error**

The unexplained or stochastic part of the analysis. A measurement of how “wrong” the value might be. It is the way of facing the reality that economic data are not deterministic and will not provide exact answers.

**Factor**

A variable used to identify a property attribute or the underlying characteristic of a property that may indirectly affect value or the reciprocal of a rate (multiplying net income by the appropriate factor will yield the same result as dividing the same net income by the direct capitalization rate).

**Function**

A mathematical relationship.

**Functional form**

The mathematical function used to model an economic relationship (linear, logarithmic, exponential, piecewise, inverse, among many others).

**Fuzzy systems**

A system that accommodates uncertainty or probability.

**Graphical analysis**

Using graphs, charts, and tables to “see” data and thereby uncover relationships, analyze effects, and assist in interpretation.

**Graphical presentation**

Using graphs, charts, and tables in reports to help the reader see the data and to support, explain, and justify interpretations of that data.

**Hedonic regression**

A multivariate analysis that analyzes a compound good, such as a house, that provides several functions in varying proportions.

**Heuristic**

Instructive or pedagogical, exploratory, trial-and-error, rule-of-thumb methodology. More akin to the “art” of appraising.

**IAAO**

International Association of Assessing Officers, which is the overseeing authority for the nation’s assessors. It functions much the same way as the Appraisal Institute.

**Inferential statistic**

A statistic used to infer results based on a relatively small set of data onto a larger set. This type of analysis goes one step beyond simply describing phenomena by attempting to predict outcomes or relationships.

**Information**

Any data that throw light on the estimated parameter, measured by the “Fisher Information number.” As the number increases, the variance of the estimate decreases (reliability increases).

**Information technology**

The broad discipline of data gathering, analysis, and interpretation, generally in the electronic, computerized software domain (process technology); utilizing product technologies (hardware).

**Instrumental variable**

A variable that can function as a proxy for something unmeasured or even unmeasurable. Sometimes called a surrogate variable.

**Intercept**

In regression analysis, this is the value of the dependent variable when the independent variables are set to zero. In property valuation models, it is often tempting to define the intercept as the inherent land value of a property without any improvements, but this is not the case because the linear relationship may not hold between the data set and the Y-axis. The appraiser should consider the intercept an artifact of the regression equation, rather than a land attribute in the appraisal model.

**Interquartile range**

A measure of spread. It measures the “distance” between the values one-quarter and three-quarters of the way along a data set.

**Interval variable**

A variable based on the actual numeric value of the data itself. For example, the variable Sale Price contains all of the attributes of nominal and ordinal data, but it also possesses meaning in terms of the differences between values. A property that sells for \$100,000 has twice the market value of a property that sells for \$50,000.

**Mass appraisal**

The process of valuing a set of properties, using a single effective date, a retrospective sales period, common data elements and definitions, and statistical analysis to evaluate the outcome.

**Mean (greek letter  $\bar{x}$ , pronounced mü, “mew”)**

The sum of n numbers divided by n.

**Median**

The middle-most value in a vector of values—e.g., 7, out of {1, 4, 6, 6, 7, 9, 11, 22, 41}.

**Mode**

The value that occurs most often—e.g., 8, out of {2, 4, 5, 6, 6, 7, 7, 8, 8, 8, 8, 10, 12}.

model

1. A representation that attempts to explain in as great detail as possible the relationship between a dependent variable, such as sale price, and independent variables that reflect factors of supply and demand.
2. A copy or representation that describes the underlying logical structure, mathematical relationships, and the behavior of the agents in the system (buyers, sellers, agents, lenders, and

regulators). It can be a single-equation model with one or more independent variables, or it can be a system of equations. It deals with the purely logical aspects of valuation theory (as differentiated from the application of real data, empirical testing of the appraisal theory, and interpretation of the results).

### **Multicollinearity**

Correlation between three or more independent variables in a regression (same as collinearity, if two variables are involved).

### **Nominal variable**

A variable used for group identification, such as house style or neighborhood. This type of data variable is often used in descriptive analysis, although it also has limited applications in inferential analysis.

### **Nonprobability sample**

A sample not produced by a scientific random process; for example, it may be a sample based upon an appraiser's judgment about which cases to select. It is well-suited for some poorly developed data sets.

### **Observation**

The words or numbers that represent an attribute for a particular case.

### **Ordinal variable**

A variable based on a ranked order of data, such as a measurement of quality of construction based on a scale of 1 through 5. This type of variable provides more information than nominal data, although it is more limited than interval variables, since the rankings themselves do not provide any information concerning the distance between each ranking. In other words, a ranking variable for quality does not imply that a score of 4 is "twice as good" as a score of 2; the correct interpretation would be that the score of 4 is two levels above the score of 2.

### **Outlier**

An observation that is extreme, in that it is out of the typical range. It is usually a suspect of error or economic inappropriateness and demands individual attention. Outliers can unduly influence the result of the regression model. The further an observation is from the measures of central tendency, the greater the influence may be.

### **Precision**

The degree of refinement in the performance of a model. Precision relates to the quality of the model, as opposed to the term validity, which relates to the result of the analysis. The precision of an appraisal model would include the types and quality of the variables and data used. The accuracy of the same model would refer to the overall architecture of the model; that is, does the model measure what it purports to measure?

### **Qualitative data**

Data that are based on subjective measures, where the data tend to fall into nominal or ordinal categories; usually represented in the form of words (see quantitative data). An amenity such as View may indeed affect market value but is nevertheless difficult to measure and quantify. Quantitative data, on the other

hand, are more objective, in that they are based on interval data that can be measured and compared with much more precision. Qualitative data are still valuable as a source of information, and when correctly ranked or systematically treated, they can significantly improve the appraisal modeling process.

**Quantitative data**

Data in the form of numbers. (Note: qualitative data can often be quantified in the econometric context.)

**R<sup>2</sup>**

See coefficient of determination.

**Random**

A number or value that can take on one of several specific possible values. It is not indeterminate but is usually connected with a probability of taking on that value.

**Rational expectations**

The assumption of consistent rational behavior.

**Regression analysis**

A method for determining the association between two or more variables.

**Reliability**

Freedom from random error. Statistically, this relates to the consistency, unbiasedness, and efficiency of a model. Economically, in terms of valuation practice, this may include a rating of the truthfulness and bias of a data source, measurement and transmission error or bias, and verification breadth, depth, and detail. In the future, the analysis of reliability will be as important as the value estimate and will be inseparable from scientific analysis.

**Risk**

The odds or probability of an unfavorable outcome. (See reliability.)

**Robust**

The quality of a statistic that is useful in spite of the violations of one of its basic assumptions.

**Sales ratio**

The ratio of the estimated sale price divided by the actual sale price. This is a powerful tool that can be graphically displayed to determine the accuracy of a model in terms of relating the dependent variable (sale price) by a set of independent variables (such as property attributes). It can also be used to evaluate whether there are systematic biases present in the model output.

**Sample**

Any data taken to represent a larger population of data. There are judgment samples, random samples, probability samples, self-selecting samples, etc. A random sample is more akin to the science of appraising, while a judgment sample is more akin to the art of appraising. Each has its place.



**Skewed**

A distribution that is not symmetrical.

**Standard deviation**

The square root of the variance; the square root of the sum of the deviations squared.

**Standard error (S.E.)**

Exactly the same as standard deviation, except that it applies to a sample rather than to the population.

**Stochastic**

Probabilistic (as opposed to deterministic).

**Variable**

A logical collection of attributes. For example, each possible size of a house is an attribute; the collection of all such attributes is the variable Square Feet. (See random.)

