

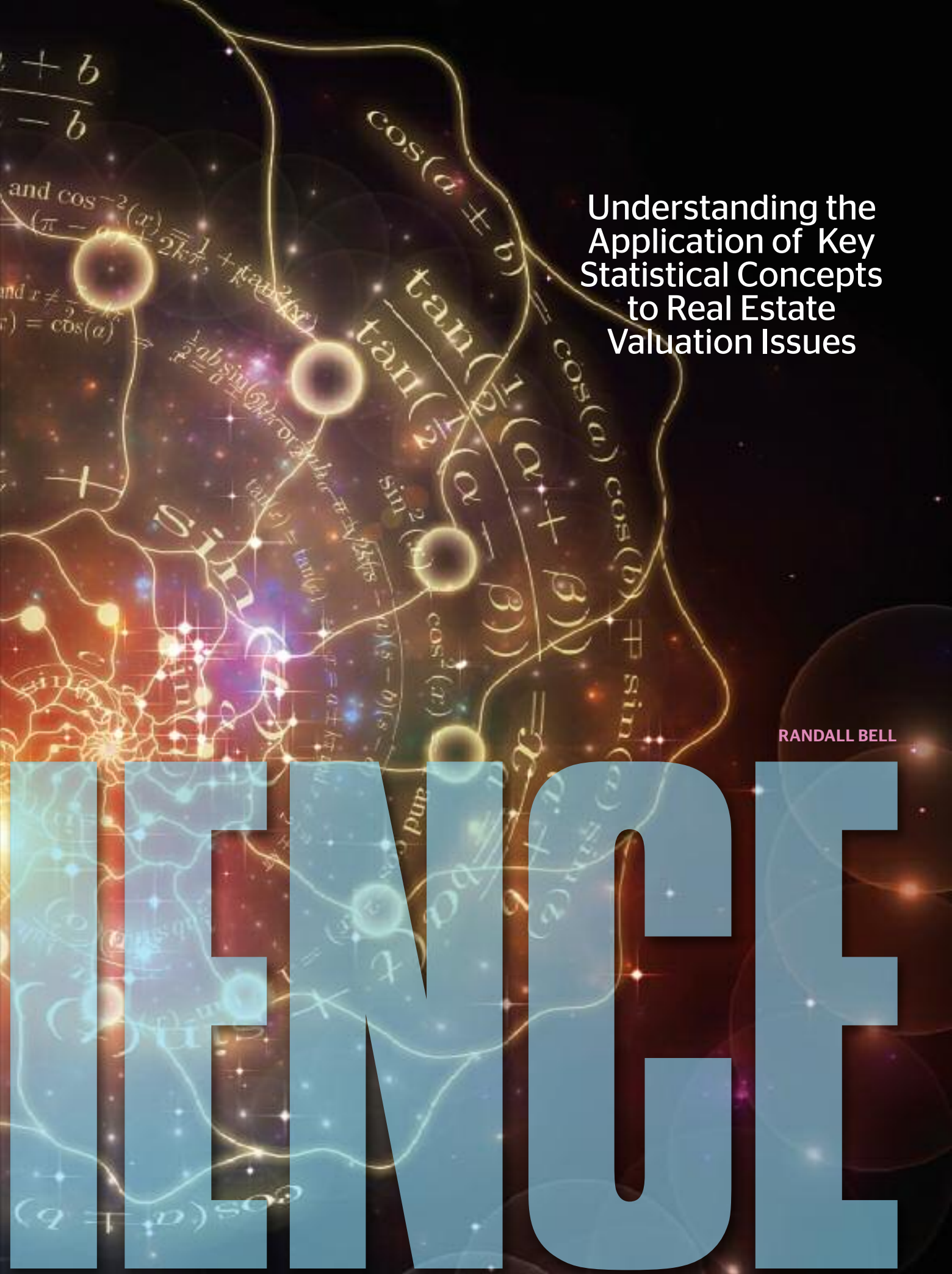
WHERE

ARIT

MEETS

SSO

WITH PRAGMATIC USE OF DESCRIPTIVE STATISTICS  
WHEN ANALYZING TRANSACTIONAL DATA,  
APPRAISERS CAN PLAY AN IMPORTANT ROLE  
IN BRIDGING THE GAPS BETWEEN  
PURELY MATHEMATICAL VALUATION MODELS,  
LOCAL MARKET CONDITIONS, AND  
THE PHYSICAL REAL ESTATE BEING EXAMINED.



Understanding the  
Application of Key  
Statistical Concepts  
to Real Estate  
Valuation Issues

RANDALL BELL

# THE MODEL

# STATISTICS

has historically been a particularly difficult topic of study for many people.<sup>1</sup> However, due to the advances of standard computer software, the use of statistics has become far more practical. Today, using algebraic algorithms in a professional context is no more necessary than referring to electrical schematics in order to use a calculator. What is more important—in the context of real estate economics—is to understand the key concepts in statistics and their direct application to valuation issues. This article addresses statistics in the context of real estate economics and valuation issues, focusing primarily on descriptive statistics which analyze transactional market data. Algebraic formulas are set aside in favor of pragmatic Excel analyses, with statistics outlined in progressive one, two, and multiple variable formats.

## Overview

With the advances of technology, statistics has evolved from their use by a few academics and practitioners into a practical tool that can be useful for many. In a valuation context, statistical studies of market data are considered a refined version of the sales comparison approach, which is often used for larger data sets. Tables of data sometimes have limited utility in visualizing numerical relationships.<sup>2</sup> Statistical charts and graphs, on

the other hand, provide a pictorial view of the data and can be especially useful for presentation purposes.

**Outdated Approaches.** One problem with the study of statistics is that some outdated courses and texts fail to address it in an intelligent and pragmatic way. Inordinate amounts of time may be spent in doing algebraic calculations with a handheld calculator, when a laptop computer can do these calculations virtually instantly. Some academics unconventionally replace standard sales adjustment grids with statistical modeling. And some use specialized statistical programs, even though conventional software, (e.g., Microsoft Excel, SPSS, or MiniTab) is well-suited for real estate valuation issues.<sup>3</sup> Even more problematic is that the cloak of complexity may be used to mask a biased analysis, or to pass off junk science as a legitimate study.

Practitioners now have the means to conduct an intelligent, user-friendly, and competent statistical analysis. Indeed, according to *The Appraisal of Real Estate*:

The traditional real property appraiser plays an important role in this changing world by bridging the gaps between purely mathematical valuation models, local market conditions, and the physical real estate being analyzed.<sup>4</sup>

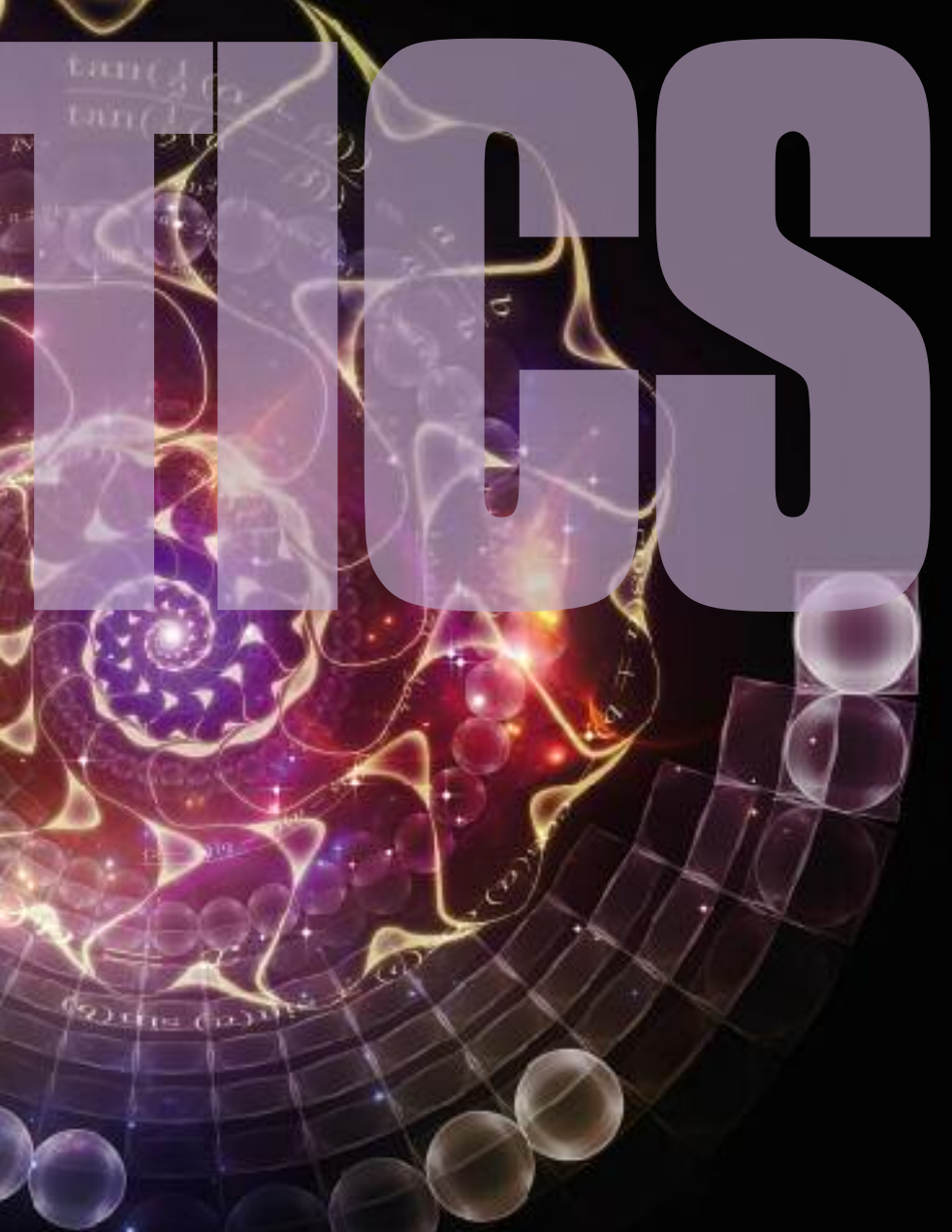
**Need to Focus on Relevant Aspects.** A key step to gaining a working knowledge of statistics is to appreciate that it is a very broad field, with corresponding nuances and lingo. Just as no single medical doctor will know all aspects of medicine, also no one person knows all aspects of statistics. What

is important is to focus on those aspects of statistics that are relevant to real estate generally, or damage economics specifically.

## Inferential Statistics

Inferential statistics draws inferences or conclusions about a population based on observations from a smaller sample. In *The Dictionary of Real Estate Appraisal*, inferential statistics is defined as, “The process of drawing conclusions about population characteristics through analysis of sample data.”<sup>5</sup> Historically, much of statistics dealt with inferential statistics as, prior to the computer age, it was difficult to collect complete data sets. Accordingly, it was necessary to collect samples of a population from which conclusions could be inferred. Today, complete data sets are often

RANDALL BELL, Ph.D, MAI, specializes in real estate damage economics. He is the chief executive officer of Landmark Research, LLC, based in Laguna Beach, California. Mr. Bell has consulted on property damage assignments around the world, and is the author of the text *Real Estate Damages*, published by the Appraisal Institute. He may be contacted at [bell@realestatedamages.com](mailto:bell@realestatedamages.com).



available to real estate professionals, so inferential statistics are generally of less importance; however inferential statistics may still be used, for example, when market surveys are employed.

Statistics involves the study of issues or “variables.” There are several types of numerical variables that are most relevant to real estate economics. Not all of the literature uses the same names and descriptions, but more modern statistical references cite variables as coming under one of four categories:

1. Nominal.
2. Ordinal.
3. Interval.
4. Ratio.

These four modern categories of variables are described below, along with other traditional variables relevant to the real estate professions.

#### **Categorical (Qualitative).**

- *Nominal* variables, meaning that they have a name, e.g., the exterior of a residence may be stucco, wood, or metal siding.<sup>6</sup>
- *Ordinal* variables, which mean that they are ranked, such as “no view, limited view, or outstanding view,” or “fair, average, good, or excellent condition.”<sup>7</sup> Traditionally, there are also subsets of ordinal variables—(1) discrete variables, where a feature is counted, such as the number of bedrooms, number of bathrooms, or number of loading docks;<sup>8</sup> (2) binary or “dummy” variables, that are simple “yes” or “no” questions, or “have” or “not have” features. For example, a house has or does not have a pool, and a warehouse either has or does not have dock-high loading doors. “Dummy variables are used in sta-

istics to convert categorical information into numerical data.”<sup>9</sup>

#### **Measured (Quantitative).**

- *Interval* or continuous variables are features that are measured, such as in inches, feet, or square feet.<sup>10</sup>
- *Ratio* variables have a real zero point, such as percentage of yard area or percentage of building to land ratios.

Not all variables can be used for the same mathematical functions. For example, a zip code (a nominal variable) cannot be added or multiplied to produce a meaningful number. The chart in Exhibit 1 sets forth what mathematical computations may be made for the four primary types of variables.<sup>11</sup>

All of these variables have a role to play in real estate analyses; however, discrete and continuous variables stand out, as they are the ones found in the many two-dimensional charts used in the real estate professions. Furthermore, variables can generally be placed into two categories, dependent variables and independent variables. Dependent variables “depend” on independent variables. With real estate statistical analyses, there is one dependent variable, usually price per unit or price per square foot, where there can be one or more independent variables. A dependent variable is defined as “the variable being estimated.”<sup>12</sup> For example, the value of a house (a dependent variable) can depend upon the date of sale, the square footage, the number of bedrooms, the number of bathrooms and the quality, which are all potential independent variables.

**Data Size.** A study of statistics may include either an entire population of variables or a sampling of the population. The law of large numbers essentially states that the more data used the better.<sup>13</sup> In probability theory, the central limit theorem essentially states that, with sufficient sample size, the sampling distribution of the mean tends to be normally distributed, regardless of the distribution of the underlying population. When sampling from a symmetrical population distribution, 15 samples are required.<sup>14</sup> For non-symmetrical distributions, a sample size of 30 is required to assure approximate normality of the sampling distribution of the mean.

A *normal distribution curve*, nicknamed a bell curve<sup>15</sup> sets forth the most probable mean or average (the high point on

the curve) as well as the *variance* (the degree of dispersion of a variable's value) from the mean, and also the standard deviations from the mean.<sup>16</sup> A standard deviation (the square root of the variance) is the normal probabilities where one standard deviation will account for 68% of the data, two standard deviations account for 95% of the data, and three standard deviations will account for 99% of the data.<sup>17</sup>

For example, if one wanted to determine typical real estate closing costs, it may be impractical to interview the entire population of brokers or agents. Accordingly, 30 brokers or agents could be surveyed out of a larger population. The data would be inputted and then observed to see if it falls within a standard normal distribution curve or not. In cases where data is normally distributed, it would look like the illustration in Exhibit 2.

## Descriptive Statistics

Descriptive statistics, as the name implies, describe or summarize data. Descriptive statistics is defined as, "A branch of statistics concerned only with characterizing, or describing, a set of numbers."<sup>18</sup> In the context of descriptive statistics and real estate valuation, there are three fundamental categories of descriptive statistical studies.

1. Single variable studies, such as pie charts, histograms, or bar charts.
2. Simple regression analyses, which includes two variable graphs or scatter diagrams, typically on an "x" axis and "y" axis that demonstrate the relationship between two variables.
3. Multiple regressions, which involves three or more variables and use more advanced studies.

Graphing three or more variables requires multiple dimensions; it is impossible to visualize these analyses on a table or a chart and requires the use of more sophisticated, yet less presentation-friendly, statistical techniques.

## Single-Variable Statistics

Single-variable statistics are relatively straightforward and are of considerable benefit in the real estate professions. The general public is well familiar with single, independent variable statistics, as newspapers and magazines often use these

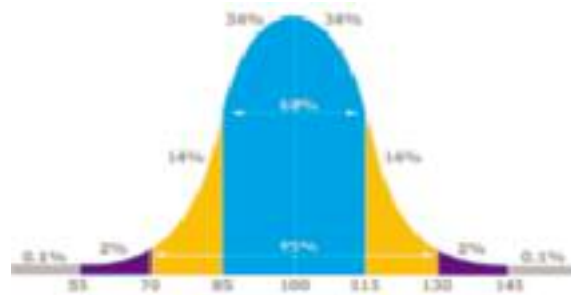
### EXHIBIT 1

#### Potential Computations for Four Primary Variable Types

OK to Compute....	Nominal	Ordinal	Interval	Ratio
Frequency distribution.	Yes	Yes	Yes	Yes
Median and percentiles.	No	Yes	Yes	Yes
Add or subtract.	No	No	Yes	Yes
Mean, standard deviation, standard error of the mean.	No	No	Yes	Yes
Ratio, or coefficient of variation.	No	No	No	Yes

### EXHIBIT 2

#### Normal Distribution Curve



### EXHIBIT 3

#### Pie Chart



statistical models to present numerical information, which often take form in pie charts, bar charts, or tables. In chart form, bar and pie charts provide a picture summarizing information from a table.<sup>19</sup>

For example, a pie chart can be used to set forth the age distribution in a community (a typical chart is illustrated in Exhibit 3).

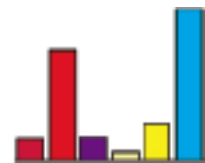
Furthermore, a bar chart, or histogram, illustrated in Exhibit 4, could be used in a variety of ways. A histogram is a type of bar chart showing the distribution of a dataset, where the total area of the bars equals either the number of observations or 100%, depending on how it is constructed.

## Two-Variable (Bivariate) Statistics

Simple regression analyses are relatively straightforward statistical models that

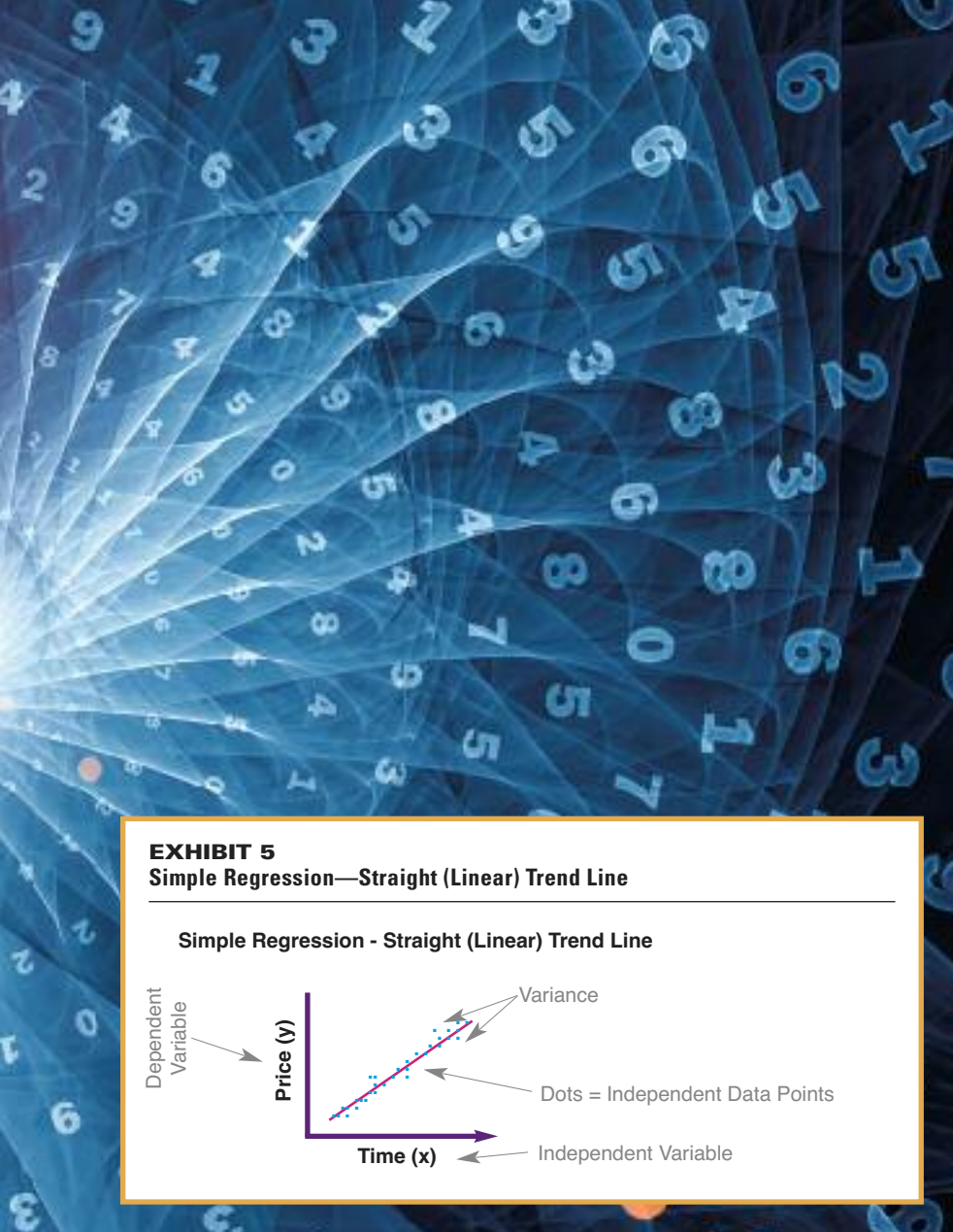
### EXHIBIT 4

#### Side-By-Side Bar Chart



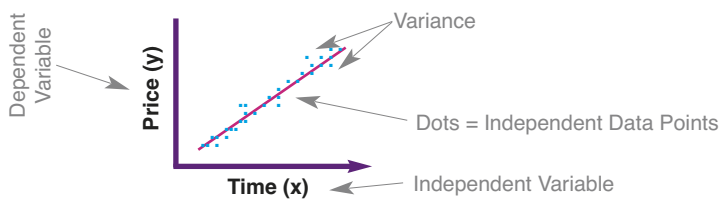
can be highly effective in setting forth the relationship of two variables, such as price and time, price and size, or other relationships. Because they use two variables, they are commonly graphed on an "x" axis and a "y" axis that show the relationship between these two variables. The dependent variable, usually price, typically goes on the "y" axis, whereas the independent variable, often time or size, typically goes on the "x" axis. Correlation is defined as "A general term for interdependence between pairs of variables."<sup>20</sup> When two variables move together, then they are said to be "correlated."<sup>21</sup> It is important to be careful about using the term "value" when describing data. Price is an observation, value is not. A sample "x and y" graph is illustrated in Exhibit 5.

Analysis of continuous or discrete data, usually shown graphically in a scat-



**EXHIBIT 5**  
**Simple Regression—Straight (Linear) Trend Line**

**Simple Regression - Straight (Linear) Trend Line**



ter diagram, can be greatly enhanced by using a trend line, which is a simple regression. Standard software, such as Excel, will facilitate a simple regression analysis.<sup>22</sup> There are multiple options for selecting trend lines, including linear (straight-line) or polynomial (curve linear) trend lines. In fact, there are a variety of polynomial trend options which effectively allows for the number of curves in the line. There are also logarithmic, power, and exponential trend lines. Ultimately, trend lines should be selected that best generally fit the market data. While a linear form is simplest, a polynomial form may best describes data over a significant time frame, particularly when the direction or magnitude of a market trend changes at some point.

**Coefficient of Determination.** The *coefficient of determination* ( $r^2$ ) describes the “goodness of fit” and can provide valuable insight into the appropriate form for the trend line.<sup>23</sup> In theory, the coefficient of determination ranges from 0% (no correlation) to 100% (perfect correlation); although in practice  $r^2$  will typically fall somewhere between these two extremes.

Simple regression analyses can be used in a variety of ways. A time-series study charts time on the “x axis” of a graph and price on the “y axis”. Such studies are invaluable for the purposes of evaluating market trends.

**Detrimental Conditions.** Analysis of detrimental conditions using large data sets often involves the selection of a *study* or *test area*, which is essentially a group of properties associated with some distinct attribute (e.g., a neighborhood affected by contamination, airport noise, or construction defects) that is then compared to one or more otherwise similar *control areas* for the purpose of determining the impact, if any, of the particular attribute on market pricing.<sup>24</sup> Such studies compare the affected test area to unaffected but otherwise similar control areas to determine whether a certain attribute has any impact on value or marketability. In a time-series analysis, data can be analyzed for discrete time periods (e.g., months, quarters, years) or continuously over a specified time frame. Discrete time periods often use the mean or median price per period.<sup>25</sup>

Using time and the price per square foot analyses of homogeneous neigh-

1 There is an old joke that says if someone has only one day to live, it should be spent in a statistics class, because it seems to go on forever.  
 2 Goddard, “Graphics Improve the Analysis of Income Data,” 68 *The Appraisal Journal* 388 (October 2000).  
 3 Wolverton, *An Introduction to Statistics for Appraisers* (Appraisal Institute, 2009), page 5.  
 4 Appraisal Institute, *The Appraisal of Real Estate, 13th ed.* (Appraisal Institute, 2008), page 597.  
 5 Appraisal Institute, *The Dictionary of Real Estate Appraisal, 5th ed.* (Appraisal Institute, 2010), page 314.  
 6 Wolverton, note 3, *supra*, pages 38-39.  
 7 Wolverton, *id.*  
 8 *Id.*  
 9 *The Dictionary of Real Estate Appraisal*, note 5, *supra*, page 313.  
 10 Wolverton, note 3, *supra*.  
 11 <http://www.graphpad.com/faq/viewfaq.cfm?faq=1089> (retrieved 2/27/2012).  
 12 *The Dictionary of Real Estate Appraisal*, note 5, *supra*, page 312.  
 13 *Id.*, page 310.

14 *The Dictionary of Real Estate Appraisal*, note 5, *supra*, page 310.  
 15 No relationship to the author!  
 16 Upton and Cook, *Oxford Dictionary of Statistics, 2nd ed.* (Oxford University Press, 2008), page 32; *The Dictionary of Real Estate Appraisal*, note 5, *supra*, page 321.  
 17 *The Dictionary of Real Estate Appraisal*, note 5, *supra*, page 320.  
 18 *Id.*, page 312.  
 19 Wolverton, note 3, *supra*, page 57.  
 20 Everitt and Skrondal, *The Cambridge Dictionary of Statistics, 4th ed.* (Cambridge University Press, 2010), page 107.  
 21 Wolverton, note 3, *supra*, page 52.  
 22 *Id.*, page 5.  
 23 Bell, *Real Estate Damages, 2nd ed.* (Appraisal Institute, 2008), page 35.  
 24 *Id.*, pages 34-35.  
 25 *Id.*  
 26 Wolverton, note 3, *supra*, page 292.  
 27 *Id.*, page 291.  
 28 *The Dictionary of Real Estate Appraisal*, note 5, *supra*, page 310.  
 29 *The Appraisal of Real Estate*, note 4, *supra*, page 625.

borhoods is a particularly powerful tool, as these models are relatively straight-forward yet they inherently take into account multiple neighborhoods, the market trends, the prices of the properties, the size of the properties, and general locational factors to the extent that the neighborhoods are homogeneous. Furthermore, these studies provide a clear and straightforward graph for presentation purposes. As the data can be visually inspected by the analyst, along with colleagues, clients, judges, and juries, they have an inherent transparency that adds to their creditability.

### Time-Value Model Example

The example presented in Exhibit 6 shows a time-value model where a neighborhood sits over an area of contaminated groundwater called a plume. Here the test areas, which have contamination but no human exposure issues, perform no differently relative to the control areas after the discovery of the contamination. In other words, in this case there is no indicated diminution in value to the test properties resulting from the non-exposure contamination issues.

However, another study reflected a situation where contaminated materials were disbursed in air emissions from a manufacturing facility and settled on properties surrounding the plant. Testing showed that the residents did have direct exposure to the contaminants that settled on the top of their homes and the surface of their yards. In this case, the time-series model showed that the test market performed normally prior to the discovery of the contamination, but that values fell notably after the discovery of the direct exposure to the contamination.

### Multiple-Variable Statistics

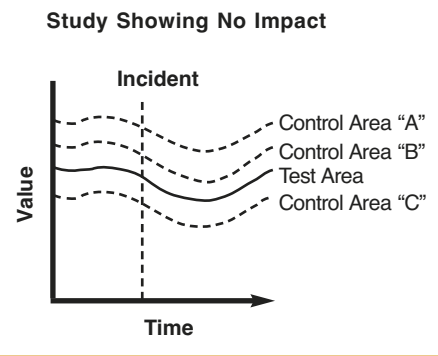
In contrast to simple regression, a multiple regression is considered a multivariate model, which analyzes the relationship among several independent variables at the same time.<sup>26</sup> Multiple regression analysis is a statistical model that expresses the value of a dependent variable, usually price in a real estate context, as a function of two or more independent variables, such as location,

site area, age, building size, views, and pools.<sup>27</sup> Mathematically, the objective of a multiple regression is to minimize the sum of the squared residuals relative to the data inputted into the study. Yet stated simply, the objective of a multiple regression is to find the relationship, if indeed it exists, between the depending variable and the independent variables, and if it does exist, to what extent they are correlated. In other words, in the context of real estate damage economics, the objective is to determine if the price (the dependent variable) is affected by some factor within the marketplace. While regression modeling can be used to estimate value of individual properties, its use in a detrimental condition analysis tends to focus on the coefficients relating to specific independent variables associated with some detrimental condition, such as airport noise, landslide zones, and environmental contamination. A coefficient is defined as “A statistic used as a measure of change, association, or dispersion.”<sup>28</sup> A sample chart reflecting multiple valuation trends is presented in Exhibit 7.

**Drawbacks.** Performed correctly, multiple regression analyses can be a powerful tool with statistical validity. Statistical modeling acknowledges the potential for error, but is able to quantify that error potential. The approach may be applicable in the study of detrimental conditions with a sufficient quantity of data; however, it also has drawbacks. As multiple regression analysis inherently involves multiple variables, once combined the data cannot be visually graphed or charted in the same way as single variable or simple regression models can. In contrast, the data is input into a computer and the results are set forth in “an analysis of variance” (ANOVA) printout, which can be generated by a standard version of Excel software. These printouts are not generally straightforward to the layman, as they include numerous intermittent calculations that are less important than others. In other words, not all statistical indicators are of equal importance; some are simply intermittent calculations and thus have limited importance.

An additional drawback is that, unlike a simple regression model graph, multiple regression analyses are not visual, thus some consider them to lack

**EXHIBIT 6**  
Trend Study—No Impact



transparency or to be a “black box” valuation approach where the complexities are not easily visualized or understood. Accordingly, considerable trust is put into the quality of the market data, the applicability of the analysis for the issue at hand, and the analyst’s competence. Furthermore, subjective choices are made regarding the selection and weight given to the multiple variables, and ongoing reiterations are often made that are dependent on the skills and judgment of the analyst. An appraiser or analyst will typically conduct numerous statistical reiterations or “runs” to eliminate extraneous variables and outliers. Along the way, many subjective decisions are made regarding inclusion of independent variables and the elimination of outlying data, and for these reasons some courts may not allow multiple

## EXHIBIT 7 Multiple-Variable Trend Charts

### Multiple-Variables (Multiple Regression):

#### 1. SUMMARY INPUT TABLE

This study measures the impact, if any, of being in a known landslide zone.

(Study Area: 0=No, 1=Yes)

The market data is laid out initially in a table with each variable in a column. Note this is a sample of 1,006 observations.

No.	Street Address	Price	Sale Date	Months from Landslide	Lot SqFt	Age	House SqFt	Study Area
1	423 S. Paseo Serena	\$247,000	05/29/96	41	11,000	26	2,117	0
2	483 S. Paseo Serena	\$245,000	10/21/97	58	6,500	26	2,117	0
3	482 S. Paseo Real	\$190,000	11/20/96	47	6,300	26	1,607	0
4	432 S. Paseo Real	\$285,000	09/24/98	69	11,750	26	1,607	0
5	411 S. Paseo Real	\$227,500	05/15/97	53	11,700	26	2,386	0
Data Summarized								
1005	6709 E. Leafwood Dr	\$325,000	06/18/99	78	8,400	23	2,196	1
1006	1068 S. Burlwood Dr	\$375,000	06/04/99	78	9,100	22	2,582	1

#### 2. DESCRIPTIVE

##### STATISTICS TABLE

This output calculates and summarizes the key statistics of the data.

	Price	Sale Date	Months from Landslide	Lot SqFt	Age	House SqFt	Study Area
Mean	\$289,312	11/10/96	46.8	10,590	22.7	2,413	0.1
Median	\$280,000	03/07/97	51.0	9,035	22.0	2,436	0.0
Mode	\$275,000	08/31/98	78.0	7,000	22.0	1,932	0.0
Standard Deviation	\$61,549	N/A	22.8	5,793	3.6	464	0.2
Range	\$422,000	N/A	80.0	87,156	20.0	2,252	1.0
Minimum	\$135,500	01/11/93	1.0	4,200	12.0	1,332	0.0
Maximum	\$557,500	09/30/99	81.1	91,356	32.0	3,584	1.0
Count	1,006	1,006	1,006	1,006	1,006	1,006	1,006

#### 3. RESIDUAL TABLE

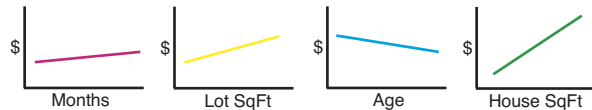
The multiple regression creates a mathematical model for predicting value. The actual sales prices can be compared to the model's predicted values to (1) assess the accuracy of the model itself, and (2) identify and discard "outlier" data with high residuals. The regression can then be re-run.

Note that this shows only a sample of 4 out of 1,006 items of market data.

Observation	Predicted Sold Price	Residuals
34	281,147.81	5031.8873
61	345765.092	12989.0765
512	256789.998	87453.8977
602	412966.5255	42565.6789

#### 4. DESCRIPTIVE STATISTICS "ANOVA" TABLE

ANOVA combines multiple variables into one analysis. Variables are all computed on a linear (straight line) basis, even if the data is actually exponential (ever growing) or polynomial (curved).



$r^2$  or the coefficient of multiple determination is a primary indicator of model validity on a scale of 0% to 100%.

(Adjusted  $r^2$  adjusts for the number of variables. (Multiple  $r^2$ )<sup>2</sup> =  $r^2$ )

Regression Statistics		ANOVA					Coefficients							
Multiple $r^2$	0.7144	Regression	5	1.94306E+11	3.88612E+11	208.461604	0.00000000	Intercept	90,421.85	14,552.62	6.213441	0.00000000	61,864.67	118,979.04
$r^2$	0.5104	Residual	1,000	1.86419E+12	1.864188203	0.00000000	0.00000000	Months from Landslide	926.45	59.93	15.459222	0.00000000	808.85	1,044.05
Adjusted $r^2$	0.5079	Total	1,005	3.80725E+12		0.00000000	0.00000000	Lot Size (SqFt)	1.56	0.24	6.493514	0.00000000	10.9	2.03
Standard error	43176.25							Age of Improvements	-1,623.82	410.01	-3.960463	0.00008011	-2,428.39	-819.24
Observations	1,006							Gross Living Area (SqFt)	73.78	3.19	23.1604	0.00000000	667.53	80.03
								Landslide Study Area	-34657.19	5,709.57	-6.070019	0.00000000	-45,861.31	-23,453.08

The "Answer" means that homes in the landslide area typically are worth \$34,657 less than a similar home outside of the landslide area.

Intercept is a hypothetical home value with all the variables set to "zero."

P-Value is the "Student's T" distribution 2-tailed. Possibility (if divided by 2) that the variable is meaningless. Ideal < 0.10.

Sig F is the possibility the analysis is meaningless.

95% chance the true value that coefficients lie in this range

t-stat is coefficient/standard error. Higher the t-stat the more relevant the variable. Check for sign and reasonableness. Ideally >2 or <-2.

Observations is the number of items of market data.

**Standard Error:** Standard deviation of a regression coefficient (Residual MS)<sup>0.5</sup>

#### Regression Calculations

**df = Degrees of Freedom**

(Number of independent variables)

**SS = Sum of Squares** (Total SS - Residual SS)

**MS = Mean of Squares** (Regression SS / Regression df)

**F = F-Stat**, Overall significance, higher = better (Regression MS / Residual MS)

#### Residual Calculations

**df = Degrees of Freedom** (Total df - Regression df)

**SS = Sum of Squares** (Predicted y - actual y, where "y" is the dependent variable. "0" if perfect.

**MS = Mean of Squares** (Residual SS / Residual df)

#### Total Calculations

**df = Degrees of Freedom** (n-1, where "n" is number of observations)

**SS = y - mean of y = (n-1)(standard deviation of y)<sup>2</sup>**  
(y = model's predicted value)



regression analyses into evidence. On the other hand, single-variable and simple regression models are generally not problematic in this regard.

If one is to employ a multiple regression analysis, there are two primary cautions. The first is that there may be significant correlation between two or more independent variables, resulting in *multicollinearity*, which often produces unreliable estimates.<sup>29</sup> For example, the number of bedrooms may be closely correlated with square footage, so it may be preferable to pick one or the other as a variable, but not both. While multicollinearity does not necessarily affect the ability of a model to predict the dependent variable, the validity of conclusions about collinear independent variables is a concern. As a result, multicollinearity should generally be avoided.

The other caution is in the area of heteroskedasticity, or non-constant regression error.<sup>30</sup> Both simple and multiple regression modeling are based on the premise that there is a constant variance of the data, which is termed homoskedasticity. If there is non-constant variance, called heteroskedasticity, this can be problematic. Exhibit 8 illustrates constant and non-constant regression errors.

## Airport Case Study

An effective means of illustrating the concepts of a multiple regression is to provide an example of its use and application. The following overview shows an application of regression modeling in a situation involving homes located in noise contours adjacent to an airport.

The independent variables considered in this case are:

1. Sales data.
2. Living area.
3. Lot area.
4. Pool.
5. Location inside or outside the 65 DNL noise contour area.
6. Location inside or outside the 70 DNL noise contour area.
7. Year built.
8. Distress sale (short sale or foreclosure).
9. Gated community.
10. Central air-conditioning.
11. Dock.

12. Waterfront with open ocean access.
13. Waterfront with six-bridge ocean access.
14. Lakefront.
15. Impact windows.

## Study Area

The general parameters for the study area included the homes located in the 65+ DNL noise contours surrounding the Ft. Lauderdale International Airport, extending outward to include areas that would be impacted as a result of the runway expansion, as well as areas that would not be effected. Specifically, the area of study is approximately 21 square miles surrounding the airport. The northern boundary is approximately S 24 Street, in the greater Fort Lauderdale area. The western boundary is the turnpike near the west end of Dania Beach. The south side is Sterling Road, also in the Dania Beach area. The east side of the study area is the Intracoastal Waterway.

These 21 square miles are also known as range 41, township 50, sections 24, 25, and 36. Included are range 41, township 50, sections 19 through 36. In the middle of this area is the airport, with most of the commercial and industrial land use towards the east. The west and center of the areas are generally single-family residential properties. All sales that were identified since 2000 were included in the study area, which ultimately totaled a population of 530 sales transactions, which were all screened by the local appraiser for accuracy.

**Specific Neighborhood Features.** A variety of externalities that may negatively effect portions of the subject neighborhood study area were considered. One is the Florida Power and Light power transmission lines, that run from Port Everglades through part of the area. A second externality considered was the Waste Station on the west end of the airport known as South Broward County Resource Recovery. Third, a gas transmission line which crosses just east of I-95 at 1487 NW 10th Street was noted. These neighborhood externalities were identified in the market data set. In addition, there are at least four fixed bridges in the study area, the clearance of which restricts open ocean access to boats with heights of less than nine feet. These factors were also considered in the study.

## Research Results

**Simple Regression.** The simple regression study measures a single independent variable, price per square foot over time, the dependent variable. The study used 145 single family residential sales within the 65 and 70 DNL contours against 2,843 sales outside of the noise contours. All the sales occurred between 9/1/1992 and the current date. The data sets were outputs of MLS and Broward County Assessor public records. Only properties which are coded as single family residential and arm's length transactions are included. The data sets were then plotted in a graph and polynomial (curved) trend lines to determine how the two data sets varied over time. The resulting graph, illustrated in Exhibit 9, shows that since 1992, properties inside the contours lines consistently sold for a significantly lower value than those outside of the contours. While a simple regression does not incorporate all of the independent variables, it generally indicates a diminution in value of over 15%.

**Multiple Regression.** The multiple regression addresses the question of whether residential properties located in the 65 DNL noise area sell for more or less as compared to similar properties located outside the contour area. In all, 530 sales were analyzed over a period since 2000. "Home price" is the dependent variable. The independent variables considered in this case included sales date, living area, lot area, pool, inside or outside the 65 DNL noise contour area, inside or outside the 70 DNL noise contour area, year built, distress sale (short sale or foreclosure), gated community, central air conditioning, dock, waterfront with open ocean access, waterfront with fix-bridge ocean access, lakefront, and impact windows.

Further analysis indicated that some of these independent variables were *not statistically significant* in a mass appraisal context. These included lot area, inside or outside the 70 DNL noise contour area, central air-conditioning, dock, and impact windows. Lot area is normally statistically significant; however, as

<sup>30</sup> Everitt and Skrondal, note 20, *supra*, page 204.

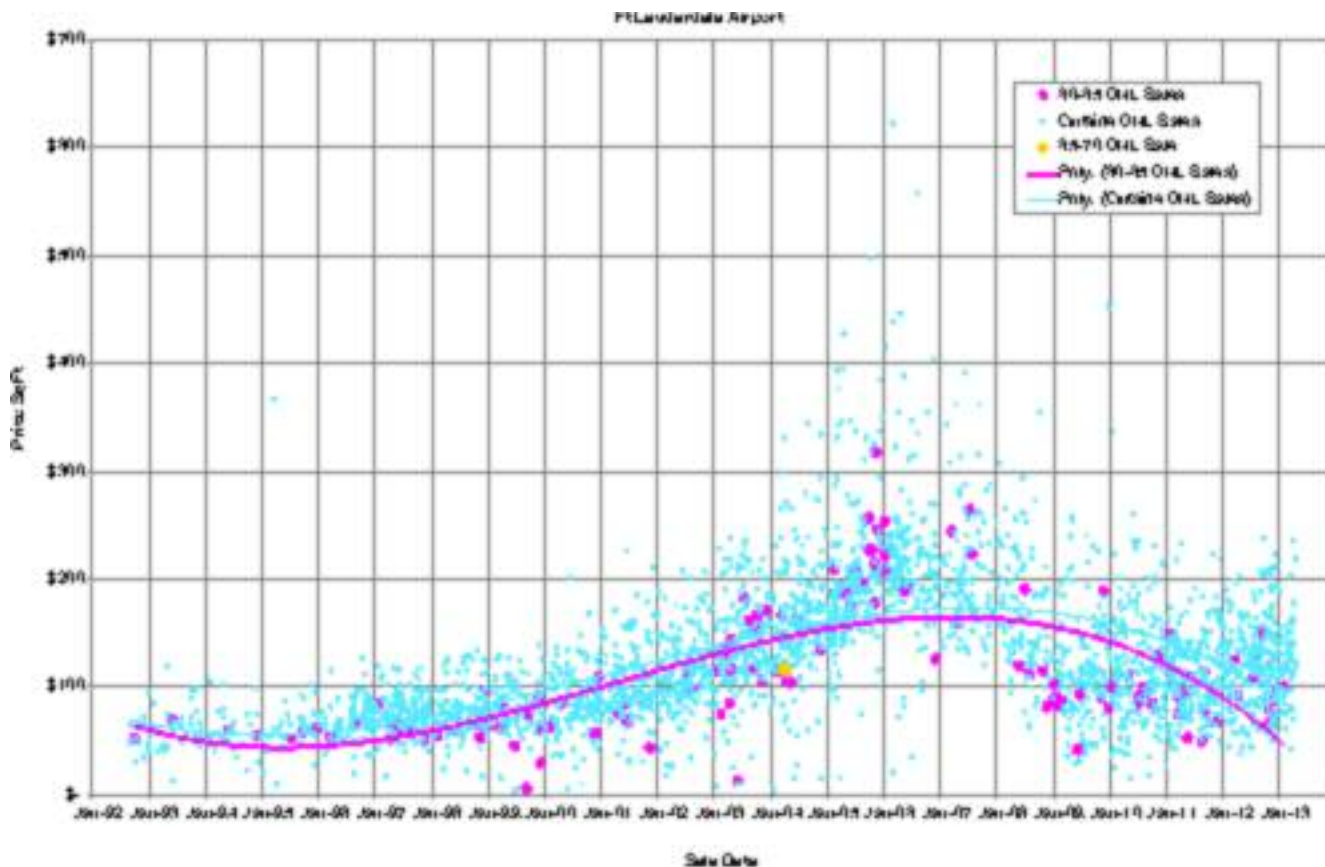
<sup>31</sup> Wolverton, *An Introduction to Statistics for Appraisers*, note 3, *supra*, page 70.

<sup>32</sup> Appraisal Institute, *The Dictionary of Real Estate Appraisal*, 5th ed., note 5, *supra*, page 320.

**EXHIBIT 8**  
Constant and Non-Constant Regression Errors



**EXHIBIT 9**  
Price Per Square Foot



many of the homes sit on small, ocean-front lots, the price per SqFt did not statistically correlate to larger lots off the water. There was only one sale identified within the 70 DNL noise contour, which is not a valid population size. Central air was skewed due to the fact that over 90% of the properties already have central air and those that do not are disproportionately older homes with a mean value of approximately \$169,000, relative to the overall mean home value of approximately \$281,000. The “dock” independent variable was considered to pose a multicollinearity (redundancy) issue with the other water-front features already being studied. The impact windows also indicat-

ed a low statistical reliability from a mass-appraisal perspective

**Summary Input Table**

The summary input table sets forth all of the market data. In this study, the variables studied are the sales price (the dependent variable) and the independent variables. The table illustrated in Exhibit 10 sets forth a sample of the market data used.

The first three indicators are mean, median, and mode, which are different ways of describing “central tendency.” Central tendency asks the question “Is there a single number that best represents the variable in question?”<sup>31</sup>

These are often called the *typical* or *average* number. The mean is the most commonly used measurement where the variables are totaled and divided by the number of the population. The mean is the same as the arithmetic average of the data. In this case, the mean selling price for the transactions included in the data set is approximately \$281,185. The median is the midpoint between the lowest and highest variable. The mode is the most frequently occurring variable.

The standard deviation is a measurement of dispersion, which essentially states how far the data sits away from the typical or the mean.<sup>32</sup>

Mathematically it is the positive square root of (Continued on page 45)

## Real Estate Statistics

(Continued from page 13) the variance, but unlike the variance which is a mathematical factor, it is expressed in terms of the units defined, or in this case, it is simply expressed as dollars. The standard deviation is a measure of variability from the mean. Measures of variance generally describe the dispersion of the data, while the standard deviation, though complicated to calculate manually, is often preferred in situations requiring more sophisticated analyses.

If data is normally distributed (represented by a bell-shaped curve), there is a 68% probability that a predicted value will be within one standard deviation of the mean, a 95% probability that a predicated value will be within two standard deviations of the mean, and a probability in excess of 99% that a predicated value will be within three standard deviations of the mean. This information is often used to construct confidence intervals around point estimates. For a normal distribution, approximately 95% of the data would fall within  $\pm 2$  standard deviations, or in this case, approximately \$150,618.

The range simply expresses the highest and lowest variable.<sup>33</sup>

The range indicator here is the difference between the maximum and the minimum price. The *count* is simply the number of transactions analyzed in this analysis. The data set includes 530 total single family residential sales, of which 65 are within the 65 and 70 DNL contours

### ANOVA Summary Output Table

Excel produces an “Analysis of Variance” (ANOVA) summary output table using a mathematical model that essentially does two things:

1. Provides the “answer” to the question.
2. Provides statistical indicators that relate to the reliability of the “answer.”



Of the four tables typically used in a multiple regression analysis, this is the most involved.

This dataset includes 530 property sales occurring between January 1, 2000 and May, 2013. All the sales are verified MLS transactions, and many were field inspected. Additionally, all the sales had a verified Broward County public record. Only single-family residential, arm’s-length transactions are included in the final dataset. The output table is illustrated in Exhibit 11.

**Primary Indicators.** Indicators that fall within the category of primary indicators include:

- Analysis Conclusion: The foremost indicator, which also reflects the “answer” to the study, is the “Coefficients” of -61,471. In this case it reflects that the “study area” (those properties within the 65 DNL noise contour airport area) typically sell for \$61,471 less than an otherwise similar property outside of the study area.
- Coefficients: Other coefficients are calculated for each independent vari-

able, reflecting a marginal rate of change.<sup>34</sup> Coefficients are normally reviewed for expected magnitude and signage. In this case, the coefficient for age is negative, indicating a value change of approximately -\$761 per year; the coefficient for gross living area is positive, indicating that each additional square foot is worth approximately \$130. Values also increase, as expected, for water frontage, gated communities and pools. Each of these figures makes sense and easily reconcile with the experience and observations of real estate professionals.

- R-Squared: The *coefficient of determination* ( $r$ ) ranges from 0% to 100%, reflecting the amount of price variation explained by the model. In this case it reflects 0.073, or 73%. An alternative calculation is *Adjusted  $r^2$*  of 0.73, or 73% which adjusts the coefficient of determination for the number of independent variables, assisting in the identification of extraneous or redundant independent variables. In multiple regression analysis, a low  $r^2$  will normally indicate a significant amount of unexplained variation, although this does not necessarily invalidate the model. However, the  $r$ -squared with this airport study indicates a strong, statistically valid study.

<sup>33</sup> *Id.*, page 318.

<sup>34</sup> *The Dictionary of Real Estate Appraisal*, note 5, *supra*, page 310.

<sup>35</sup> Bell, *Real Estate Damages*, 2nd ed. (Appraisal Institute, 2008), page 39.

<sup>36</sup> *The Dictionary of Real Estate Appraisal*, note 5, *supra*, page 321.

<sup>37</sup> *Id.*

<sup>38</sup> *Id.*, page 320.

<sup>39</sup> *Id.*, page 317.

<sup>40</sup> *Id.*, page 315.

<sup>41</sup> *Id.*, page 312.

<sup>42</sup> Upton and Cook, note 16, *supra*, page 9.

<sup>48</sup> Wolverton, note 3, *supra*, page 61.

**EXHIBIT 10**  
Summary Input Table

No.	City	Sale Price	Sale Date	Living Area (Sq Ft)	Pool	GG DNL	Year Built	Distress	Gated	Water Oper	Water Fibred	Lake Front	Central AC	Impact Windows
1	Dania Beach	\$ 124,000	02/18/00	1,090	1	1	1972	0	0	0	0	0	1	0
2	Dania Beach	\$ 84,300	04/27/00	1,007	0	1	1955	0	0	0	0	0	1	0
3	Dania Beach	\$ 83,100	12/04/00	1,464	0	1	1960	1	0	0	0	0	0	0
4	Ft Lauderdale	\$ 129,000	12/29/00	1,176	0	1	1969	0	0	0	0	0	1	0
5	Ft Lauderdale	\$ 122,500	06/04/01	1,608	0	1	1998	0	0	0	0	0	1	0
6	Ft Lauderdale	\$ 118,000	06/07/01	1,394	0	1	1961	0	0	0	0	0	1	1
7	Dania Beach	\$ 119,000	07/03/01	1,747	1	1	1976	0	0	0	0	0	1	0
8	Hollywood	\$ 175,000	09/05/01	3,198	1	0	1998	0	1	0	0	0	1	0
9	Ft Lauderdale	\$ 119,000	10/16/01	1,394	1	1	1971	0	0	0	0	0	0	0
10	Ft Lauderdale	\$ 90,000	11/21/01	1,147	0	1	1954	0	0	0	0	0	0	0
11	Ft Lauderdale	\$ 154,900	06/02/02	1,408	0	1	1958	1	0	0	0	0	1	0
12	Dania Beach	\$ 92,000	06/27/02	922	0	1	1955	0	0	0	0	0	1	0
13	Ft Lauderdale	\$ 120,000	02/07/03	1,043	0	1	1955	0	0	0	0	0	1	0
14	Dania Beach	\$ 115,800	04/05/03	1,479	0	0	2004	0	0	0	0	0	1	0
15	Dania Beach	\$ 150,000	04/25/03	1,761	0	1	1957	0	0	0	0	0	1	0
16	Dania Beach	\$ 180,000	06/02/03	1,566	0	1	1973	0	0	0	0	0	1	0
17	Hollywood	\$ 275,000	06/05/03	2,212	0	0	1997	0	1	0	0	1	1	0
18	Hollywood	\$ 305,000	06/05/03	2,237	0	0	2000	0	1	0	0	1	1	0
19	Hollywood	\$ 260,000	06/05/03	2,137	0	0	2000	0	1	0	0	0	1	0
20	Hollywood	\$ 310,000	06/05/03	2,493	0	0	2000	0	1	0	0	0	1	0
21	Hollywood	\$ 301,500	07/05/03	2,237	0	0	2000	0	1	0	0	0	1	0
22	Dania Beach	\$ 150,000	07/28/03	1,924	0	1	2002	0	0	0	1	0	1	0
23	Hollywood	\$ 163,000	09/05/03	3,669	0	0	2003	0	1	0	0	0	1	0
24	Dania Beach	\$ 250,000	09/30/03	1,678	1	1	1978	0	0	0	1	0	1	0
25	Dania Beach	\$ 178,000	11/24/03	1,680	0	1	2003	0	0	0	0	0	1	0
26	Ft Lauderdale	\$ 209,000	12/10/03	1,219	1	1	1961	0	0	0	0	0	1	1
27	Hollywood	\$ 285,000	01/05/04	2,185	0	0	2000	0	1	0	0	0	1	0
28	Dania Beach	\$ 125,000	03/31/04	1,942	1	1	1973	0	0	0	1	0	0	1
29	Ft Lauderdale	\$ 75,000	04/14/04	644	0	0	1941	1	0	0	0	0	0	0
30	Ft Lauderdale	\$ 302,000	04/21/04	2,819	1	1	1972	0	0	0	0	0	1	0
31	Ft Lauderdale	\$ 103,000	06/05/04	1,139	0	0	1952	1	0	0	0	0	1	0
32	Hollywood	\$ 310,000	06/05/04	2,231	0	0	1998	0	1	0	0	0	1	0
33	Dania Beach	\$ 105,000	06/24/04	1,014	0	1	1956	0	0	0	0	0	1	0
34	Ft Lauderdale	\$ 189,000	07/05/04	1,517	0	0	1955	0	0	1	0	0	1	0
35	Ft Lauderdale	\$ 395,000	07/05/04	1,432	0	0	1956	1	0	1	0	0	1	1
36	Dania Beach	\$ 215,000	07/30/04	1,532	0	1	1971	0	0	0	0	0	1	0
37	Ft Lauderdale	\$ 215,000	06/02/04	1,483	0	1	1972	0	0	0	0	0	1	0
38	Ft Lauderdale	\$ 420,000	06/05/04	2,702	0	0	2000	0	1	0	0	0	1	0
39	Ft Lauderdale	\$ 525,000	09/05/04	1,720	0	0	1963	1	0	1	0	0	1	0
40	Dania Beach	\$ 157,000	09/05/04	1,666	1	0	1978	0	0	0	0	0	1	0
41	Ft Lauderdale	\$ 300,000	10/05/04	1,586	0	0	1956	0	0	1	0	0	0	0
42	Hollywood	\$ 348,500	11/05/04	2,237	0	0	2001	0	1	0	0	0	1	0
43	Ft Lauderdale	\$ 470,000	01/05/05	1,380	0	0	1956	0	0	1	0	0	1	0
44	Dania Beach	\$ 121,500	01/05/05	961	0	0	1940	0	0	0	0	0	1	0
45	Ft Lauderdale	\$ 280,000	02/28/05	1,337	0	1	1960	0	0	0	0	0	0	0
46	Hollywood	\$ 180,000	03/05/05	2,212	0	0	1997	0	1	0	0	0	1	0
47	Ft Lauderdale	\$ 315,000	04/05/05	1,670	0	0	1953	0	0	1	0	0	1	0
48	Hollywood	\$ 590,000	04/05/05	3,690	0	0	1998	0	1	0	0	0	1	0
49	Ft Lauderdale	\$ 258,000	04/22/05	1,390	0	1	1960	0	0	0	0	0	1	0
50	Hollywood	\$ 610,000	06/05/05	3,198	1	0	1997	0	1	0	0	1	1	0
51	Hollywood	\$ 525,000	06/05/05	2,882	1	0	2000	0	1	0	0	1	1	0
52	Hollywood	\$ 491,500	06/05/05	2,148	1	0	1997	0	1	0	0	1	1	0
53	Ft Lauderdale	\$ 287,000	06/27/05	1,296	1	0	1956	1	0	0	0	0	1	0
54	Ft Lauderdale	\$ 315,000	07/05/05	1,704	0	0	1955	0	0	1	0	0	1	0

**EXHIBIT 11**  
**Summary Output Table**

SUMMARY OUTPUT 7

Regression Statistics	
Multiple R	88%
R Square	73%
Adjusted R Square	73%
Standard Error	78,303
Observations	530

ANOVA						
	df	SS	MS	F	Significance F	
Regression	10	8,818,558,151,891	881,855,815,189	143.83	0.00	
Residual	519	3,182,144,829,053	6,131,300,249			
Total	529	12,000,812,980,944				

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	\$ 3,197,430	593,045	5.39	0.00	2,032,355	4,352,494	2,032,355	4,352,494
Sale Date	\$ (42.15)	3.54	(11.92)	0.00	(49.10)	(35.21)	(49.10)	(35.21)
Living Area (Sq Ft)	\$ 130.57	7.85	16.63	0.00	115.14	145.00	115.14	145.00
Pool	\$ 24,572	8,352	2.95	0.00	8,253.51	41,079.44	8,253.51	41,079.44
65 DNL	\$ (51,471)	11,455	(5.37)	0.00	(83,975)	(38,955)	(83,975)	(38,955)
Year Built	\$ (751)	294	(2.59)	0.01	(1,339)	(184)	(1,339)	(184)
Distress	\$ (51,585)	7,842	(7.85)	0.00	(75,991)	(45,180)	(75,991)	(45,180)
Gated	\$ 35,905	13,411	2.75	0.01	10,559	53,251	10,559	53,251
Water-Open	\$ 141,915	9,822	14.45	0.00	122,520	151,212	122,520	151,212
Water-Fixed	\$ 107,351	13,143	8.17	0.00	81,541	133,181	81,541	133,181
Lakefront	\$ 42,758	15,251	2.80	0.01	12,787	72,750	12,787	72,750

Mean: \$281,175

- Significance F: *Significance F* measures the probability that the results were obtained purely by chance.<sup>35</sup> In this case, there is a near zero chance that these results were obtained purely by chance.
- T-Statistic: *T-statistics (t-stats)* are also calculated for each independent variable and reflect the difference between a sample mean and a hypothesized value of the population mean divided by the standard error of the mean.<sup>36</sup> The larger the *t-stat*, the more one can be confident that the coefficient is meaningful. All of the T-Statistics indicated in this study are valid.
- P-Value: The *P-value* measures the probability that the true value of the coefficient is zero (the significance level).<sup>37</sup> An acceptable significance level for coefficients in

most applications is 10%, or a *P-value* of 0.10 or less.

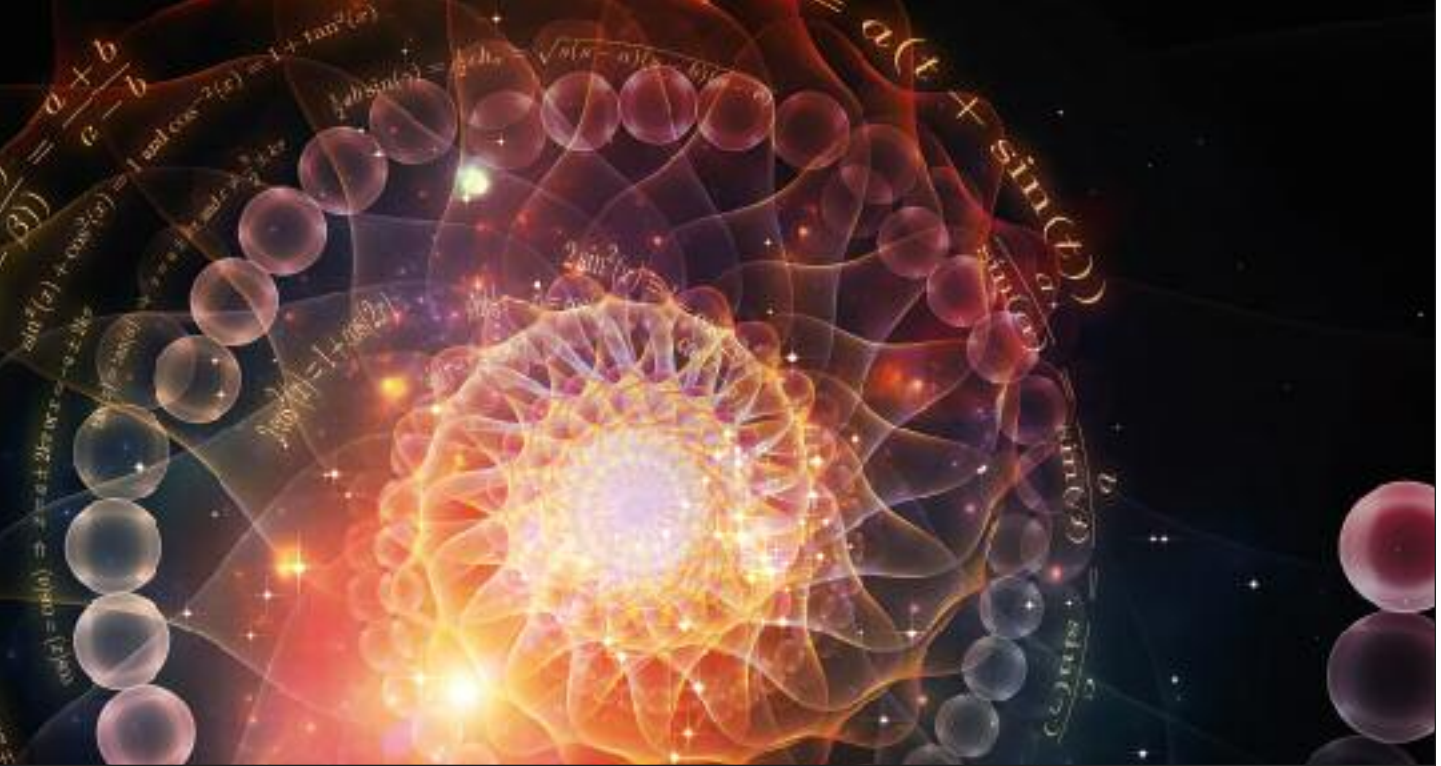
- Standard Error: The standard error is the standard deviation of the regression coefficient estimate.<sup>38</sup> In this case it is 78,303, which essentially means that the home prices are typically \$281,175 (the mean discussed above) plus or minus \$78,303. As this is one standard deviation, this would only account for about 68% of the variability in *y*.
- Observations: There are 530 items of market data, or observations in this analysis.<sup>39</sup>

In this study, the coefficient calculated for the airport study area is - \$61,471. In other words, the conclusions of the analysis show that properties located within the airport area do incur a diminution in value of typically of

\$61,471 or approximately 21.9% of the mean value.

**Secondary Indicators.** An ANOVA printout also has a number of indicators that are intermittent calculations or are secondary in importance. (For example a *t-stat*, which is a primary indicator, is simply the *coefficient* divided by the *standard error*.)

- The *Intercept*, represents an estimate of the value of the dependent variable when the values of the independent variables are set to zero.<sup>40</sup>
- The *Degrees of Freedom (df)* is the number of values in a calculation that are allowed to vary.<sup>41</sup>
- The *total df* is calculated as  $n-1=df$ , where *n* represents the sample size.
- The *regression df* reflects the number of independent variables, which are five in this example, which is used in calculating the mean of squares.



- In statistics, sometimes numbers are squared in order to consistently produce positive numbers for comparison. The *Sum of Squares* (SS) is the total variability in a set of data, and the *Mean of Squares* (MS) is the Sum of Squares divided by the regression degrees of freedom.<sup>42</sup>
- The *Coefficients Standard Error* is the respective standard error, or the standard deviation of a regression coefficient estimate, for each of the variables.
- The *Lower 95%* and the *Upper 95%* reflect the range for each respective coefficient, with a 95% degree of certainty.

## Conclusion

Historically, statistics was a highly specialized and complex topic which only a few academics and fewer practitioners were proficient in. The barriers of entry were high, as statistical calculations were made using exhaustive algebraic formulas and handheld calculators, mainframe computers, or Fortran programming on punch cards.

Today, virtually every desktop and laptop computer has the ability to efficiently perform statistical calculations in seconds that once took hours. This article has set aside the longhand algebraic formulas and instead focused on practical applications for real estate professionals using Excel software.

The application of inferential statistics is nearly limitless, but often draws inferences or conclusions about a population based on observations from a smaller sample. In other words, this is the sampling of a large population, such as for surveys. It could be argued that using regression modeling to estimate value or even the impact of some condition is an application of inferential statistics, since the analyst is effectively making an inference based on a sample of sales from a larger population of properties. This is where the *normal distribution curve*, or *bell curve* is used. On the other hand, descriptive statistics describe or summarize data. This is the area of statistics that is most relevant to real estate professionals.

Variables are those features in the real estate market that are often of interest to professionals, such as square footage, number of bedrooms, and location in or outside of some particular neighborhood. In descriptive statistics, this can be studied with one-variable models (pie-charts, bar charts), two-variable models (graphs, simple regressions), and multiple-variant modeling (Excel ANOVA) studies. All of these analyses can provide creditable insights, depending on the circumstances of the study. Both one-variable and two-variable studies have a high degree of transparency, where other professionals and laymen alike can examine the data and the resulting analyses in a straightforward manner. On the other hand, while

multiple-variable studies are more complex, they may lack transparency. Multiple regression analyses can be problematic in that:

1. It can be a subjective process of selecting numerous independent variables.
2. The criteria for eliminating outliers can be criticized.
3. There is an opportunity to manipulate data.

As multiple regressions are premised on all of the variables being linear in nature, this is obviously problematic when the data is curved, such as prices that rise and fall over the time period being studied.

Certainly, statistics cannot be oversimplified, and many issues clearly go beyond the scope of this article. Real estate professionals should use the core appraisal literature, including those sections of *The Appraisal of Real Estate* and *The Dictionary of Real Estate Appraisal* that are devoted to statistics, as well as books specifically devoted to the topic such as *An Introduction to Statistics for Appraisers*.

Regardless of the statistical methodology employed, it is critical that real estate professionals present their studies in a manner that makes clear sense to other real estate professionals. A competent professional will provide their clients and colleagues with studies that avoid distortion, enhance the users' comprehension of the topic, include relevant data, and serve a clear purpose.<sup>48</sup> ■