Multiple Regression Analysis for Regulators

There is an old saying in real estate that value is derived from three factors: location, location, location. There is a similar saying in the field of valuation modeling, but the term used is data, data, data. Regression analysis involves the use of known data to estimate the value of the unknown, the value of a given subject property. In order to be effective this process has to involve a skilled appraiser using good data.

I came to my current position in IAAO after working for 27 years in local assessment offices in various roles, always dealing with mass appraisal. In my last jurisdiction, we used 30 different valuation models to value over 150,000 residential parcels every year. During the time I worked there the jurisdiction was never out of compliance with state requirements to appraise property within 10% of its market value. In addition to that practical application, I have helped write and teach classes on mass appraisal and mass appraisal modeling in the United States and abroad.

My experience has taught me several things about regression analysis: (1) you don't have to be a statistician to develop sound models; (2) you must never lose sight of basic appraisal principles and (3) you should spend 95% of your time on the data. Developing accurate models that reflect sound appraisal principles is easy when the data is good, but nearly impossible when the data is bad.

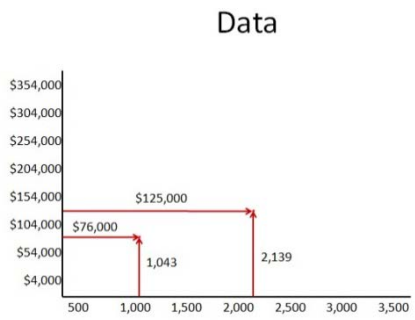
Soundness of the data is measured in two ways: quantity and quality. We will look at the latter first.

The "known data" mentioned earlier are sales, and the quality of those sales is measured by their accuracy and consistency. An investigation into accuracy begins with the selling price. The appraiser should verify the accuracy of the sale amount of every sale used in the modeling process just as he or she would for a single property appraisal. Any third party responsible for reviewing a valuation model and/or the appraisal resulting from that model has a right to be assured that all the sales used to specify and calibrate that model were individually verified for accuracy.

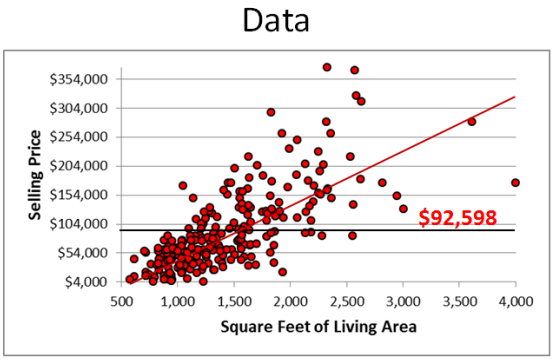
As important as the selling price is the data it represents. Knowing a home sold for $250,000 is only useful when coupled with accurate characteristics of the property. In this case accuracy relates as much to consistency as it does to objective measurements. For example, a large portion of the value of a residential property is attributable to the quality of its construction and design. Quality is a subjective judgement made by the appraiser or whomever collected that data. Even characteristics that seem objective, such as the number of bathrooms, may become problematic when one data collector refers to three fixtures as a full bath and another a three-quarter bath. Unless the collector of the data takes great care to enforce consistency in description and coding of characteristics, the appraiser may be left with property characteristics that cannot support valuation modeling.

Earlier I mentioned the term “quantity”. A major consideration in valuation modeling is the number of sales (quantity) available to the appraiser. One of the first questions students ask me when I teach a workshop on multiple regression analysis is how much data is enough and the answer is always the same – it depends. The purpose of a valuation model is to replicate market activity. Specifically, in the area of regression analysis we are trying to answer the question “what causes the sale price to change from one property to another?” That question may be answered with just a few sales when the pattern is clear and unchanging or it may take several hundred sales when prices are “all over the board”.

This can be illustrated using actual sales plotted in a scatter diagram. Each of the diagrams used in this presentation were plotted using Excel and they represent the intersection of selling price on the vertical or ‘Y’ axis and the square feet of living area on the horizontal or ‘X’ axis.



Plotting all of our sales in this particular sample produced the diagram shown below, which begins to show a pattern but is not very useful for finding market value without adding something to it. That something is an estimation line that allows the user to locate the square feet of living area on the horizontal axis, draw a line perpendicular to that point until it intersects the estimation line and then draw a second line from that point, perpendicular to the vertical axis, to a point on that axis representing the value estimate.



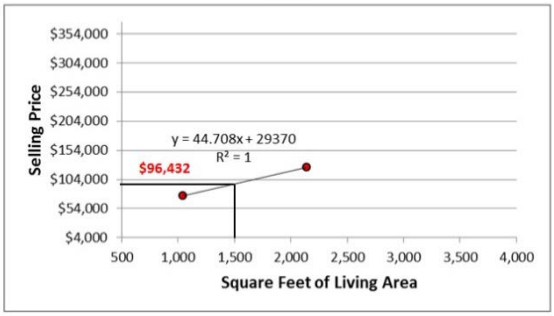
One of the obvious choices is the average value, which for this group of sales is $92,598. If selling price was the only item of data provided, average value would have to be used. However, it is easy to see how poor a job it does in replicating the values within this sample.

On the other hand, regression analysis fits a line to all the data points in such a way as to minimize the difference between any point on that line and the actual selling prices. The result is an estimation line that more closely replicates market activity. For those interested in the numbers, the average ratio of estimated value to actual selling price is 2.07 for the average price line and 1.46 for the regression line while the standard deviations are 2.8 and 1.4 respectively.

To illustrate how changes in the quantity and type of data used to build the regression or estimation line can affect the final value conclusion, we will examine very small sales samples beginning with just two sales.

The diagram to the right uses the two sales shown in our first diagram. Notice the intersection points are connected by a diagonal line which is the regression line. If the subject property contains 1,500 square feet of living area, this regression line will produce a value estimate of $96,432. This is derived from the formula y=44.708x + 29370 by substituting 1,500 for "x" such that it becomes 44.708(1,500) + 29370.

Data

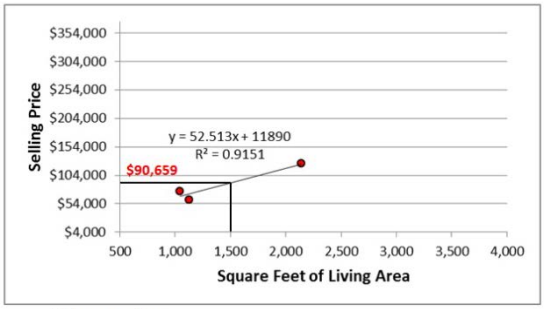


Two sales
Subject property at 1,500 square feet

Adding one additional sale changes the regression line, because it alters the underlying formula. Instead of $96,432, the new formula is estimating the subject value at

52.513(1,500) + 11890 = $90,659

Data

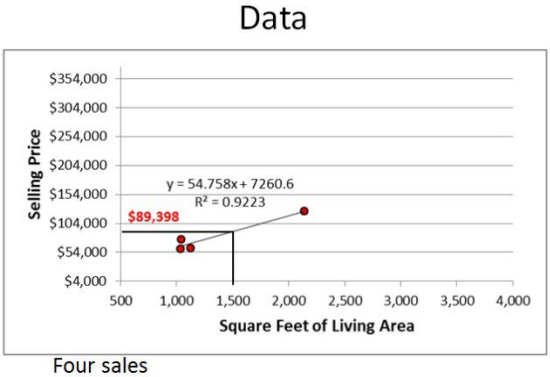


Three sales
Subject property at 1,500 square feet

The addition of one sale altered the estimation by 6%.

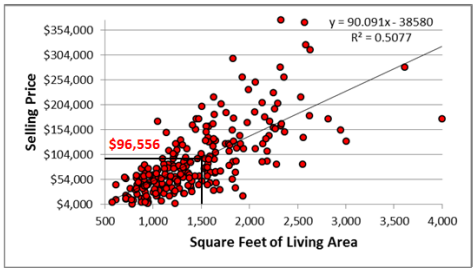
I did not attempt to order the sales by price or size, only by neighborhood. That allowed the changes produced to be more random in nature. The fourth sale, for example, did not alter the previous estimation significantly. It only changed by $1,261. However, this should illustrate the importance of the underlying data. Imagine, for example, the result of including incorrect sales or square foot data.

Data



Four sales
Subject property at 1,500 square feet

Data



(90.091 * 1,500) – 38,580 = 96,556
Two hundred thirty-two sales
Subject property at 1,500 square feet

Also consider the difference between adding a sale at a time and throwing all the sales from that area into the model at once.

The illustrations thus far have used two variables, a dependent variable (selling price) and one independent variable (square footage). For that reason, this is called simple regression. It is easier to illustrate in two dimensions and the basic principles are the same for multiple regression analysis. The objective is to calculate a regression line that best fits all the data points and can then be used to estimate values.

The number of sales needed to support a model depends on the number of variables needed to replicate the market. Every variable in the model must be supported with enough sales to ensure its accuracy. A rule of thumb often cited in mass appraisal is 15 sales for every independent variable. The issue is not whether a model can be built with fewer sales; it is whether the model output can be relied upon.

Let's talk about some terms. A regression model is composed of a **dependent** variable and one or more **independent** variables and their coefficients. It will look something like the following output, which was generated using Excel.

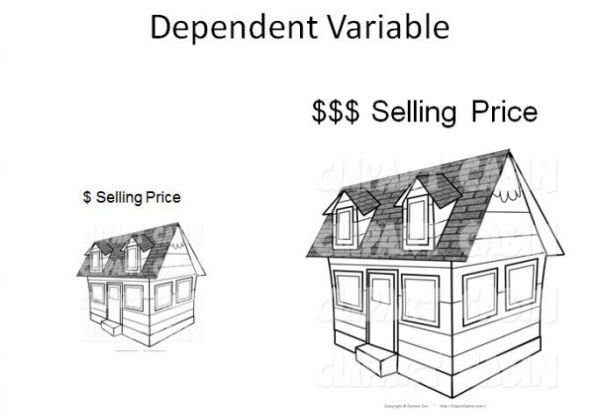
SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.9669
R Square	0.9349
Adjusted R Square	0.9341
Standard Error	33,025.1339
Observations	861

ANOVA

	df	*SS*	*MS*
Regression	10	13,305,646,471,636	1,330,564,647,164
Residual	850	927,060,548,124	1,090,659,468
Total	860	14,232,707,019,760	

	Coefficients	*Standard Error*	*t Stat*
Intercept	25,288.9214	10,218.0979	2.47
SFGF	80.6051	1.2922	62.38
CDUEDX	142,318.2551	6,821.8708	20.86
CDUVG	78,177.4253	4,201.8137	18.61
CDUGD	39,196.7333	2,967.9862	13.21
AGE	-4,638.4940	280.8507	-16.52
PLMBTFX	3,941.1210	564.6623	6.98
CDUFR	-32,703.5625	3,391.8770	-9.64
CDUPR	-72,628.7910	19,152.1854	-3.79
TOTLSF	1.2091	0.3354	3.61
MONTHS	60.3232	304.6368	0.20



A variable is said to be **dependent** when any change in its value depends on changes in the value of other variables. In appraisal, selling price is typically used as the dependent variable in modeling. Selling prices depend on other variables for their changes. For example, we expect a larger house to sell for a higher price than a smaller one. Therefore, the selling price in that case depends on the size of the house, which is usually represented by square feet of living area. Likewise, we expect selling price to reflect the relative quality of construction, making the quality rating an independent variable on which selling price depends.

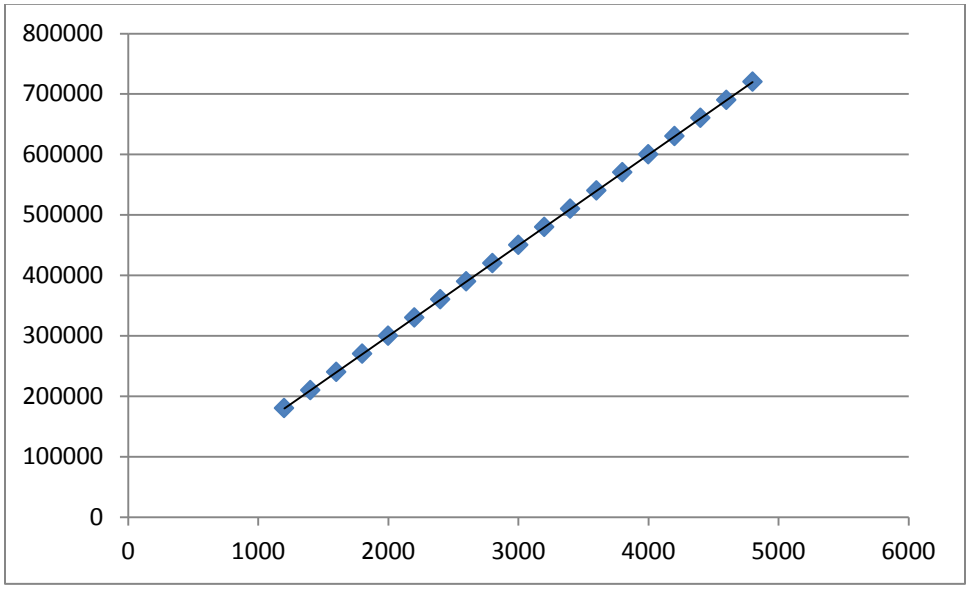
The dependent variable, selling price, is not shown in the model output above. The **independent** variables are listed in the bottom, left most column beginning with SFGF. This represents a combination of the square feet of living area with a numeric representation of a quality rating called grade factor. Whether independent variables are used just as they appear on the property record or are combined, as in this case, is a function of the skill of the modeler and the makeup of the data.

The modeler is said to be **specifying** the model as determinations are made of which variables will be included. Again, there is no objective way of establishing the precise number of variables to include in a model. The model building process is iterative in nature. The model is built, **specified**, and run. The output is represented by several factors and statistics that the modeler uses to determine the quality of performance and therefore, the need to include or exclude variables and try the process again. Part of the output takes the form of coefficients for each of the independent variables. Developing these coefficientws is called **calibration**. In an additive model, which is the most common form, each independent variable is multiplied by its coefficient and added to the result of that same calculation for all other variables. That is finally added to a constant amount, called the “intercept” in the above example, and compared to the dependent variable, selling price. The closer these estimates are to the actual selling prices, the more accurate the model is said to be.

The question for the model builder then becomes where to start, and the answer to that question requires a combination of appraisal and model building skills. An appraiser experienced in the local market should be aware of the factors that influence value. A tool that is helpful to provide a starting place or support a prior decision is the **correlation coefficient**. Correlation

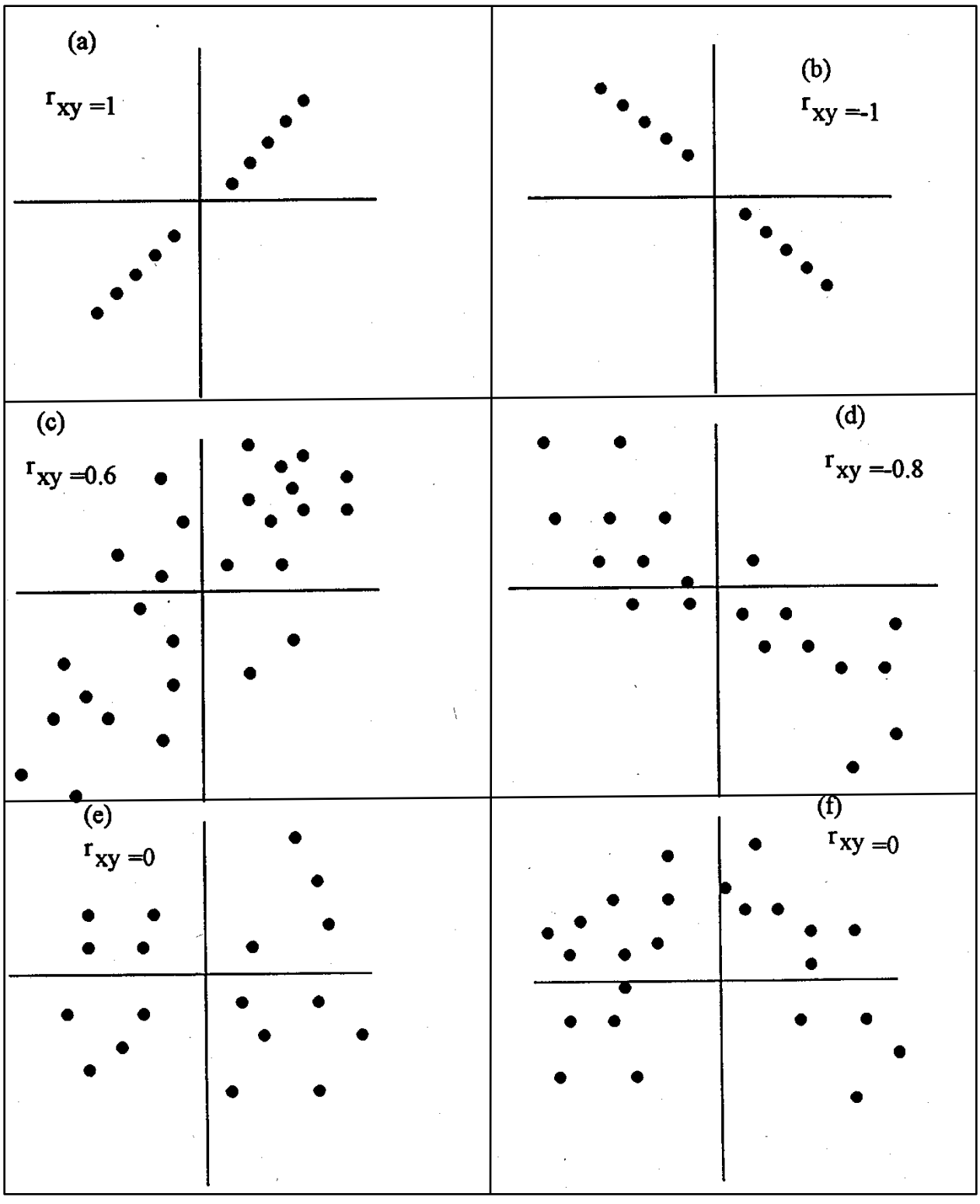
refers to any of a broad class of statistical relationships involving dependence. The Pearson correlation coefficient is one of the most common measures of correlation. It measures the linear relationship between two variables, which is especially helpful in basic regression modeling.

Consider, for example, the relationship between selling price and square feet of living area. It is generally assumed that larger homes will command higher prices than smaller ones when everything else is held constant. If the relationship between these two variables is in direct proportion and we plot selling prices on a vertical axis against square feet of living area on the horizontal axis, the resulting graph will look like the following.

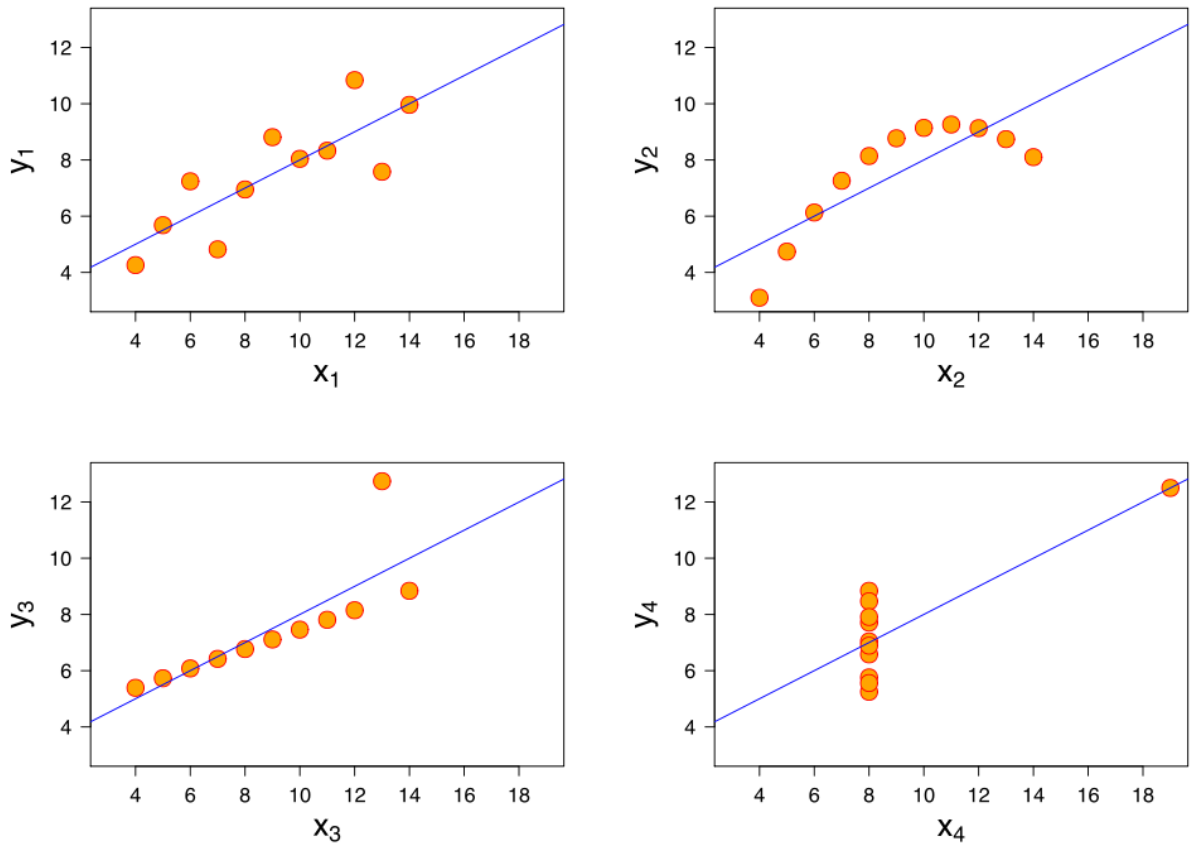


Placing a square at the intersection of the selling price and size of each sale property produces a straight line on a diagonal. The Pearson correlation coefficient in this case would be 1.0 on a scale from 0 to 1.0, which represents perfect, positive, linear correlation. Any deviation from that perfectly straight line would result in a coefficient of less than 1. That would mean there is a weaker linear correlation between the two variables.

The diagrams on the following page illustrate various correlations. Diagram (a) illustrates a positive correlation of 1.0, while diagram (b) shows a negative correlation of 1.0. Diagrams (c) and (d) shower weaker correlations and diagram (e) shows no correlation at all between the two variables. Diagram (f) is interesting because there is a pattern formed by the intersection of each x and y, but it is not linear. So while the two variables are related, that relationship is not linear. Therefore, the particular measure we are using shows a zero correlation.


(a)
r_{xy} =1
(b)
r_{xy} =-1
(c)
r_{xy} =0.6
(d)
r_{xy} =-0.8
(e)
r_{xy} =0
(f)
r_{xy} =0

The next set of diagrams is referred to as “Anscombe’s quartet” after their creator, Francis Anscombe. They were constructed to illustrate the importance of graphing data before drawing a final conclusion regarding any calculated summary statistics. Each dataset consists of 11 data points. All four of these datasets have the same average (mean) values for both x and y. They also share the exact same correlation coefficients. The obvious lesson to take from this is that those building or evaluating valuation models should not rely on summary statistics alone.



Before we leave the topic of the coefficient of correlation, it is important to look at a tool the model builder has to review the correlation between and among several variables. The tool is called the correlation matrix and an example is shown below.

	SPR	*SFGF*	*CDUEDX*	*CDUVG*	*CDUGD*	*AGE*	*PLMBTFX*	*CDUFR*	*CDUPR*	*TOTLSF*	*MONTHS*
SPR	1.0000										
SFGF	0.9170	1.0000									
CDUEDX	0.3054	0.1627	1.0000								
CDUVG	0.2658	0.1424	-0.0553	1.0000							
CDUGD	0.0784	-0.0024	-0.0898	-0.1594	1.0000						
AGE	-0.2536	-0.1929	0.0411	0.1045	0.1089	1.0000					
PLMBTFX	0.5333	0.5352	0.0692	0.1328	-0.0297	0.0176	1.0000				
CDUFR	-0.1691	-0.0275	-0.0724	-0.1285	-0.2087	-0.0657	-0.0591	1.0000			
CDUPR	-0.0122	0.0314	-0.0104	-0.0185	-0.0301	-0.0224	-0.0040	-0.0243	1.0000		
TOTLSF	0.5285	0.5288	0.1489	0.1852	0.0302	0.0870	0.3189	-0.0988	0.0121	1.0000	
MONTHS	-0.0099	0.0030	-0.0166	-0.0130	0.0237	-0.0154	-0.0134	0.1361	-0.0145	-0.0305	1.0000

Each variable used in this analysis is displayed both in the first column and in the first row. That is why the intersection of the SPR column and SPR row displays a coefficient of 1.0; each variable is perfectly correlated with itself. For this reason, some statistical packages will show only one half of the matrix, or, as in this case, leave one half of the matrix blank. Each half is the mirror image of the other.

The reason for displaying the variables in this manner is to allow the model builder to quickly see which variables are most strongly related to the dependent variable selling price (SPR) and whether they are correlated with each other. Notice the variable SFGF (square footage times grade factor) is highly correlated with SPR. It is also correlated with PLMBTFX (total plumbing fixtures), possibly because they both reflect the relative size of the dwelling. Placing two independent variables that are highly correlated with each other in the same model may cause the model to behave erratically and produce poor results.

The model builder selects those variables that are most highly correlated with selling price and not with each other. Individually, we look for a correlation coefficient approaching either a positive or a negative 1.0. There is a summary statistic that measures the overall strength of the model and it is called the **coefficient of determination** or R^2. It measures the percent of change in the dependent variable that is explained by the independent variables, or how much of the change in selling price is being explained by the current model. The goal is 100%. Generally speaking, the closer R^2 is to 100, the better the model is performing.

Three things must be kept in mind when evaluating the coefficient of determination: (1) the R^2 will continue to increase as long as variables added to the model are correlated in any way with the selling price; (2) a model in a neighborhood of very similar homes may have a low R^2 simply because selling prices do not vary much, leaving nothing to explain; and (3) models relate to the specific selling prices and the characteristics of the sales used. The first issue can be dealt with by watching the relationship between the R Square and the Adjusted R Square. As variables are entered relative to their correlation with selling price, this relationship will begin to diverge at some point and that is when the model builder should stop introducing new variables.

Keep in mind the purpose of the model is to determine what is causing selling prices to change from one property to another. If all properties in a given area are selling for the same or

nearly the same price the R^2 may fall toward zero. Because the R^2 measures the amount of variation in the selling price being explained by the model, when there is no variation, there is nothing to explain. A model with a low coefficient of determination and few variables may be very good at estimating values that are very similar to each other.

The job of the model builder is to replicate the market. The group of sales used to build the model should, therefore, proportionately match the market being appraised in terms of their particular characteristics and current market conditions. A model built using sales of tract homes should not be used to value unique estate style homes and a model built eighteen months ago in a fast changing market should not be used to estimate current values.

In the same manner, the modeler should evaluate each variable and its coefficient in relation to all other variables in the model. This process culminates in an estimate of value with each variable contributing some portion of the whole. As new variables are introduced into the model those portions change. Notice the difference in the coefficients assigned to SFGF on the first and the last modeling run.

Model 1			*Model 10*	
	Coefficients			*Coefficients*
Intercept	3,894.0816		Intercept	25,288.9214
SFGF	95.9139		SFGF	80.6051
			CDUEDX	142,318.2551
			CDUVG	78,177.4253
			CDUGD	39,196.7333
			AGE	-4,638.4940
			PLMBTFX	3,941.1210
			CDUFR	-32,703.5625
			CDUPR	-72,628.7910
			TOTLSF	1.2091
			MONTHS	60.3232

Using the 1,500 square foot home used previously, model 1 would yield a value estimate of:

95.9139 x 1500 = 143,869.5 + 3,894.0816 = 147,763.58

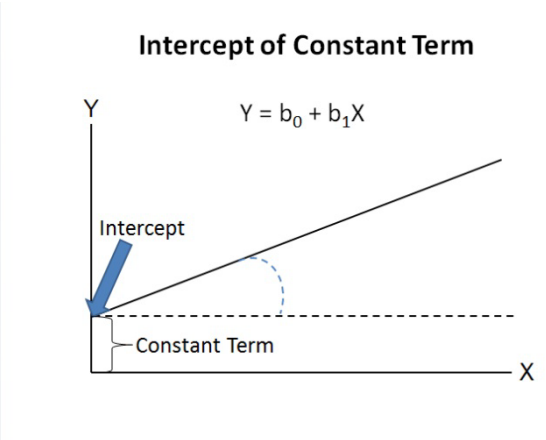
The contribution of the SFGF variable to the total value estimate was $143,869.50. When the other variables are included in the model that contribution declines to:

80.6051 x 1500 = 120,907.65

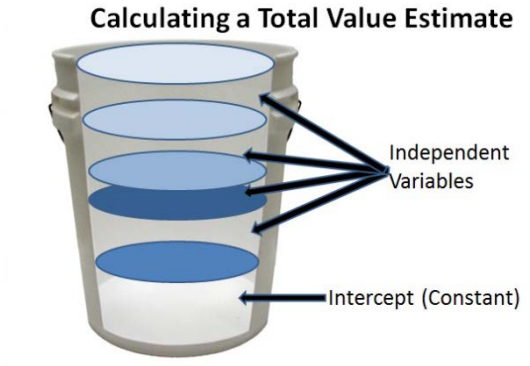
Coefficients should never be considered apart from their participation in a given model. They only contribute to the total value in relation to the other variables in the model. There is a tendency on the part of appraisers new to modeling to treat coefficients the same as rates such

that the coefficient in the final model becomes $80.61/square foot. That can be true only if the modeler assumes the model constructed is perfect, lacking nothing. The best that can be said is that a particular coefficient accurately contributes to value in relation to the remaining variables in the model until the market changes.

Another complicating factor in the use of coefficients to develop contributory values lies in the **intercept** or **constant** term. It gets its name from the fact that in simple regression it is represented as that point where the regression line intersects the Y or vertical axis. Remember from high school math that a straight line can be defined by two points or by one point and the slope of the line. The simple regression lines we showed earlier adopt the second approach with the intercept forming one point and the coefficient for X forming the slope.



Values on the Y axis represent the value estimates found by adding the result of multiplying a value of X by a coefficient and then adding it to the amount at the intercept. It is clear the value at the intercept will be the same for each value of X; it remains constant for each value of X.



Another way of looking at the constant term is to picture it as the starting point for constructing the value of any property. From that point, values for each of the independent variables are added until the total value is calculated. The complicating factor for appraisers is not knowing what is included in the constant term. A portion of the value of each of the independent variables may be included in that constant term such that their coefficients reflect an added value above and beyond the constant. If one of the variables in a model is square feet of living area the appraiser cannot say with any degree of certainty that the coefficient represents the total contributory value of square feet of living area.

Valuation models are time sensitive, just as a single property appraisal. Local assessors use those models to estimate value as of a statutory lien date. Anyone using multiple regression should specify the effective data of the value conclusions.

By virtue of the data used to build a model, that model has a range of effectiveness. The dominant characteristics of the model dataset establish that range. For example, a model that depends on sales of homes in a range of sizes from 1,000 to 2,000 square feet will not be a reliable indicator of the value of a 3,000 square foot house. Likewise, if all the sale homes were built after 1980, the model should not be used to value a home built in 1955.

Models are also geograhically defined. Location is always important in valuing real estate and requires some way to identify and quantify differenes in location. Theoretically a model can be built with a very broad geographic coverage as long as the variation in location influences can be accurately captured in a usuable form. That is rarely the case.

Questions for Regulators to Ask

How were sales selected and verified?

What is the appraisal date and how current are sales to that date?

How does the range of selling prices compare to the value assigned to the subject?

How were the model variables selected?

Did you use an additive or multiplicative model?

What training and experience do you have in model building?