Regression Analysis and Statistical Applications

B

The most commonly used statistical application in the appraisal of real property, tax assessment, automated valuation modeling, and other forms of real estate analysis is undoubtedly regression analysis. As discussed in Chapter 14, regression analysis allows comparison of a dependent variable, usually price or rent, and either a single independent variable (in simple linear regression) or many independent variables (in multiple regression). This appendix supplements the discussion of essential statistical topics in Chapter 14 with more detailed discussion of the application of simple linear regression and multiple regression. Other topics include model specification, model validation, underlying regression model assumptions, and the potential misuse of statistical methods.

Simple Linear Regression

In its simplest form, linear regression captures a relationship between a single dependent variable and a single independent, or predictor, variable. This relationship is usually written as follows:

$$Y_i = \alpha + \beta x_i + \varepsilon$$

which reflects an underlying deterministic relationship of the linear form $Y = \alpha + \beta x$ plus the stochastic (i.e., random) component ε. As shown on a graph, the slope of the regression line is b, and the intercept is a. The effect of any variables, other than the single independent variable, that may influence the value of the dependent variable is not included in a simple linear regression model.

In an appraisal application of simple linear regression, for example, the Y variable in the model could represent market rent and the x variable could be apartment living area. The random component would reflect sampling error plus the imperfections of real estate markets, which include the influence of factors such as informational advantages, the negotiating strengths of the parties to a sale or lease transac-

tion, and any other variables not included in the model. The simple linear regression model yields an estimate of the equation

$$\hat{Y}_i = a + bx_i + e$$

where

a is an estimator of α,
b is an estimator of β,
and *e* is an estimator of ε.

The outcome variable $\hat{Y}_i$ is the expected market price (for example, the model's estimate of market rent) of property *i*, given the value of the independent variable *x*.

The presence of the random error term is an indication that regression models are inferential (or "stochastic"), rather than deterministic. Regression models provide estimates of the outcome variables that should be accompanied by a statement about the degree of uncertainty associated with the estimate. In addition, they provide estimates of the coefficient on the independent variable, *b* in this context, which also incorporate a degree of uncertainty.

In Table B.1, the apartment rent data set that was introduced in Chapter 14 is augmented by adding living area to demonstrate a simple linear regression model. Note that the range in rent per square foot is $0.35 ($1.20 – $0.85), an indication that living area probably is not the sole factor determining rent. Otherwise, rent per square foot would exhibit minimal variation.

A simple linear regression model will uncover the extent to which rent is explained by the living area variable. The model can be run on a number of statistical software packages. Figure B.1 shows the output that was derived using Excel.

This output illustrates that the best-fitting linear relationship between living area and rent is a line with intercept $336.17 and a slope of $0.57359 per square foot of floor area:

$$Price = \$336.17 + \$0.57359 \times Floor\ Area$$

The model *F*-statistic, 42.85908, is highly significant, meaning that the model predicts rent better than merely relying on mean unit rent. The *t*-statistic on floor area, 6.546685, is also highly significant, meaning that living area is an important factor for rent estimation. The coefficient of determination, R^2, can vary from 0 to 1, with 0 indicating no explanatory power whatsoever and 1 indicating perfect explanatory power (i.e., a deterministic model). The R^2 of .557632 indicates that 55.8% of the variation in rent is accounted for by variation in floor area. Adjusted R^2 is useful for comparing multiple competing models with differing sets of independent variables because the measure accounts for the number of explanatory variables in relation to sample size. The model having the highest adjusted R^2 is usually the preferred model. In this instance, with only one independent variable under consideration, there is no competing model.

Obtaining an understanding of the intercept and slope is referred to as structural modeling because the model uncovers the structure of the relationship between the dependent variable and the independent variable. A simple linear model facilitates development of a "best fit" line in two-dimensional space, which can be overlaid on a scatter plot of the data to demonstrate unexplained variation in the dependent variable, as shown in Figure B.2.

The scatter plot shows that rent generally rises linearly with floor area. The regression line shown on the chart is the best-fitting straight line, which minimizes

Table B.1 Living Area and Monthly Rent

	Rent	Living Area (Sq. Ft.)	Rent per Sq. Ft.
	$600	650	$0.92
	650	670	0.97
	695	655	1.06
	710	755	0.94
	715	695	1.03
	730	770	0.95
	735	840	0.88
	735	820	0.90
	760	865	0.88
	760	760	1.00
	785	740	1.06
	800	740	1.08
	800	730	1.10
	805	890	0.90
	815	850	0.96
	820	850	0.96
	820	740	1.11
	825	970	0.85
	825	970	0.85
	825	770	1.07
	825	690	1.20
	850	850	1.00
	850	970	0.88
	850	970	0.88
	850	970	0.88
	850	805	1.06
	850	850	1.00
	860	830	1.04
	860	790	1.09
	890	860	1.03
	890	850	1.05
	920	970	0.95
	920	1,030	0.89
	930	890	1.04
	970	1,050	0.92
	995	1,000	1.00
Median	$825.00	845	$0.985
Mean	$815.83	836	$0.983
S	$84.71	110	$0.087
Minimum	$600.00	650	$0.85
Maximum	$995.00	1,050	$1.20

Figure B.1 Excel Summary Output of Simple Linear Regression

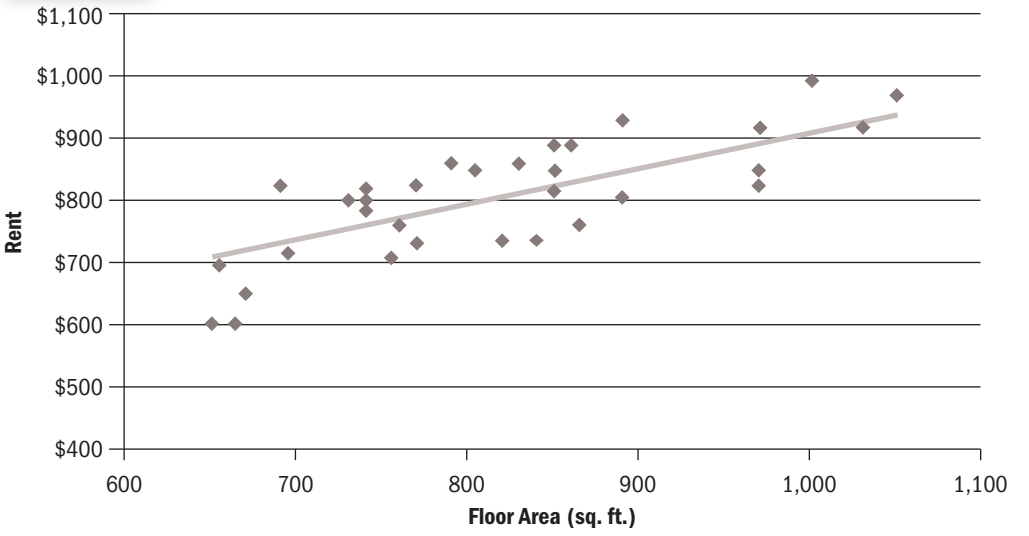
SUMMARY OUTPUT	
Regression Statistics	
Multiple *R*	0.746748
R Square	0.557632
Adjusted *R* Square	0.544621
Standard Error	57.16637
Observations	36

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	140063.2	140063.2	42.85908	1.69E-07
Residual	34	111111.8	3267.994		
Total	35	251175			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	336.1697	73.88506	4.549901	6.53E-05
X Variable 1	0.573589	0.087615	6.546685	1.69E-07

Figure B.2 Rent and Floor Area



the squares of the errors between the data and the line's fit to the data. Differences between actual prices and the regression line can be attributed to one of two causes: (1) randomness in pricing (i.e., the stochastic element of price) or (2) other unaccounted-for variables that are also important in determining rent. Those elements might include unit characteristics such as bedroom counts, bathroom counts, and tenant amenities such as a pool, spa, and exercise facility. Simple linear regression becomes multiple linear regression when more than one independent variable is included in a model to account for additional elements of comparison.

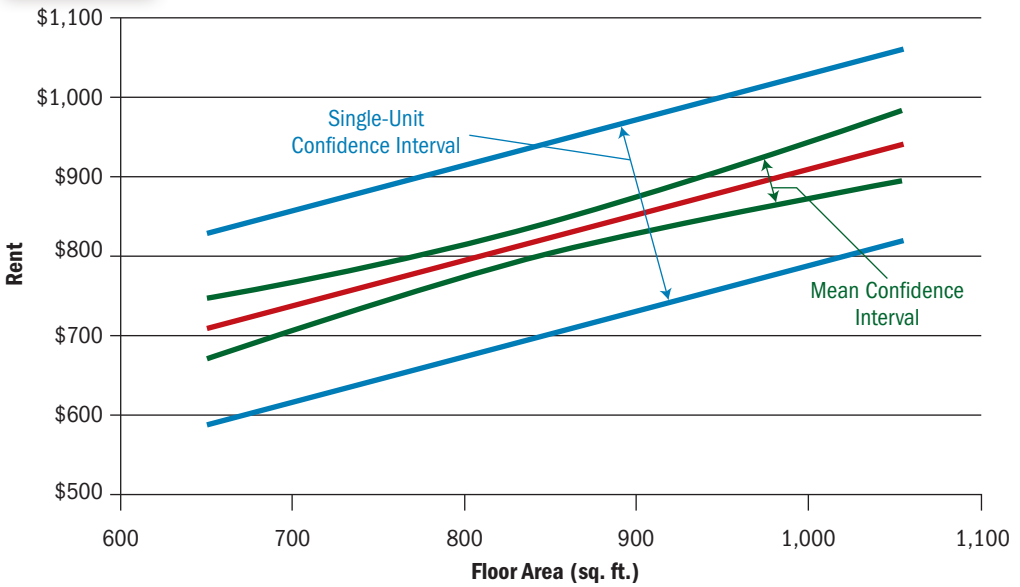
Regression models can either be predictive or structural (i.e., constructed for the purpose of understanding the structure of the relationship among variables). Predictive models are predominant in most valuation settings. Two forms of predictive models are generally employed. One form is used to estimate the mean outcome, and the other form estimates a single, specific outcome. The primary difference is that the confidence interval for an estimation of the mean outcome is narrower than the confidence interval for estimation of a single, specific outcome. Furthermore, regression models are not usually employed to estimate outcomes using inputs that are outside the ranges of the independent variables.

For example, assume the appraiser wants to predict rent for an 810-sq.-ft. apartment using the sample data. The predicted mean rent for units of this size and the predicted rent for a single, specific 810-sq.-ft. unit are the same at $800.78. However, the confidence interval widths vary considerably, as follows:

95% confidence interval on mean rent of 810-sq.-ft. units:	$780.86 to $820.70
95% confidence interval on rent of a single 810-sq.-ft. unit:	$682.91 to $918.65

SPSS and Minitab are capable of calculating and reporting confidence intervals for the mean and for a single outcome. The confidence intervals for the data are illustrated in Figure B.3 along with the regression line for unit rent. Note that the prediction confidence intervals are narrowest near the mean unit size and grow wider for single units. For this reason, the confidence intervals must be calculated separately for any given value of the independent variable (or values for the independent variables in multiple linear regression). This is a time-consuming process, which is best accomplished electronically in SPSS or Minitab. Note also that the limits of the known data are shown at the ends of each plotted line. Beyond the limits of known data, any

Figure B.3 Regression Line with Confidence Intervals for Mean and Single-unit Rent Estimates



conclusions drawn by the appraiser will constitute forecasts or predictions and that usable confidence is further reduced or eliminated statistically.

The equations for calculating prediction confidence intervals for simple linear regression are as follows:

Prediction Confidence Interval for the Mean *Y* Outcome

$$\text{Confidence Interval} = \hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Prediction Confidence Interval for an Individual *Y* Outcome

$$\text{Confidence Interval} = \hat{Y}_i \pm t_{n-2} S_{YX} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

The confidence intervals widen as they depart from the mean because the numerator $(x_i - \bar{x})^2$ gets larger as the distance of the independent variable from the mean increases. For further clarification, the value symbolized as S_{YX} in these equations is reported as the "Standard Error of the Estimate" in SPSS, "*S*" in Minitab, and "Standard Error" in Excel. The summation $\Sigma (x_i - \bar{x})^2$ is often referred to as *SSX* (sum of squares for the *x* variable) and is calculated as $S_{YX} \div S_b$. S_b is reported as "standard error" for the independent *x* variable coefficient in both SPSS and Excel and as "standard deviation" for the independent *x* variable coefficient in Minitab. Given this information, it is possible to calculate confidence intervals by hand for a simple linear regression if the need arises. The confidence interval calculations become more complex with multiple linear regression and are best calculated using statistical software.

Multiple Linear Regression

As stated earlier, additional independent variables can be included in a regression model to account for more than one element of comparison. In real estate appraisal, multiple linear regression is often a more realistic representation of the interplay of the variety of transactional and property characteristics that can affect the value of a predictor variable like *price* or *rent* than simple linear regression can be.

As a demonstration of a multiple linear regression model, suppose that further investigation of the rent data reveals variation in bedroom counts, bath counts, and common amenities. Characteristics such as these can be modeled by use of numerical variables and by the creation of indicator variables (also known as *dummy variables*) to convert categorical data such as common amenities into numerical variables. (Obviously, other elements of comparison may be important such as differences in age and condition, location, access, neighboring land use, and other characteristics. This example is simplified for demonstration purposes.) To create a common amenity variable that indicates the presence of a pool, spa, and exercise facility, units in apartment complexes that have the feature are coded 1 and units in apartment complexes that do not have a pool, spa, and exercise facility are coded 0. Bedroom and bath counts are entered as discrete numerical data. The revised sample data set is shown in Table B.2.

Table B.2 Rent, Living Area, Room Counts, and Amenities

Rent	Living Area (Sq. Ft.)	Bedrooms	Baths	Pool/Spa/Exercise
$600	650	1	1	0
650	670	1	1	0
695	655	1	2	1
710	755	1	1	0
715	695	1	2	1
730	770	2	1	0
735	840	2	1	0
735	820	2	1	0
760	865	2	1	0
760	760	1	2	0
785	740	1	1.5	1
800	740	1	2	1
800	730	1	2	1
805	890	2	2	0
815	850	2	2	0
820	850	2	2	0
820	740	1	2	1
825	970	2	2	0
825	970	2	2	0
825	770	1	2	1
825	690	1	2	1
850	850	2	1	1
850	970	2	2	0
850	970	2	2	0
850	970	2	2	0
850	805	2	1	1
850	850	2	2	0
860	830	2	2	0
860	790	2	1	1
890	860	2	2	0
890	850	2	2	1
920	970	2	2	0
920	1,030	2	2	0
930	890	2	2	1
970	1,050	2	2.5	0
995	1,000	2	2.5	0

A multiple regression model using Minitab yields the following price equation:

Unit Rent = $209.06 + $0.4703 × Living Area (sq. ft.) + $50.10 × Bedrooms + $58.27 × Bathrooms + $79.77 × Pool/Spa/Exercise		
t-statistics:	Living Area (sq. ft.)	3.83 (p = .001)
	Bedrooms	2.06 (p = .048)
	Baths	3.45 (p = .002)
	Pool/Spa/Exercise	5.22 (p = .000)
Model *F*-statistic = 37.80 (p = .000)		
R^2 = .830		
Adjusted R^2 = .808		

This result indicates that living area, bedroom count, bath count, and an amenity consisting of a pool, spa, and exercise facility are all significant in the determination of unit rent. The *t*-statistics are all significant at $\alpha \leq .05$. The p values stated after the *t*- and *F*-statistics are the probabilities of the model result occurring by chance. When the p value is less than .05, then the variable (or model in the case of the *F*-statistic) is said to be significant at the 5% level (i.e., $\alpha \leq .05$). Here, most of the results are significant at the 1% level. The model's *F*-statistic is also highly significant. This model is preferred to the simple linear regression model because adjusted R^2 has gone up from .545 to .808, despite the loss in degrees of freedom resulting from adding more variables while keeping sample size constant. The expanded multiple linear regression model accounts for 83% of the variation in unit rent, which is a vast improvement over the 55.8% coefficient of determination for the simple linear regression model.

To predict mean rent and a specific unit rent for an 810-sq.-ft. apartment unit having 2 bedrooms, 1½ baths, and use of an on-site pool, spa, and exercise facility, the calculation would be

$$\text{Unit Rent} = \$209.06 + \$0.4703 \times 810 + \$50.10 \times 2 + \$58.27 \times 1.5 + \$79.77 \times 1 = \$857.38$$

Note that the Minitab estimate is $857.40, which is unaffected by rounding.

The associated 95% confidence intervals derived in Minitab are

95% confidence interval on mean rent:	$827.55 to $887.24
95% confidence interval on a single-unit rent:	$776.00 to $938.79

One benefit of the expanded multiple regression model's higher explanatory power is more predictive precision in comparison to the simple linear regression model, as indicated by the tighter confidence intervals for the predicted mean and for a single-unit rent prediction.

Another way to develop such a model would be to create indicator (or "dummy") variables for discrete numerical variables such as bedroom counts and bath counts. This allows the rent contributions of these features to vary instead of being constrained to a single linear coefficient. Often the creation of indicator variables to describe discrete numerical variables will improve the model fit. For example, adding dummy variables to reflect 1, 1½, 2, and 2½ bath categories to this model increases R^2 to .854 and adjusted R^2 to .824.

Model Specification

Model specification issues fall into two broad categories for valuation purposes: (1) the functional form of the relationship between the dependent variable and the independent variables and (2) the choice of variables to include in the model.

Functional Form

Functional form issues arise because of a regression model's presumed linear relationship between dependent and independent variables, even though many of these relationships are likely to be curvilinear. (Curvilinear relationships are characterized by curved lines instead of straight lines. Examples include logarithmic curves, exponential curves, inverse curves, and polynomial curves.) Many characteristics of real property are thought to be subject to increasing or diminishing marginal utility. Consider bathroom counts. Keeping floor area and bedroom count constant, adding bathrooms could initially result in increasing marginal utility. However, as more bathrooms are added above some optimum level, the contribution to value begins to diminish. Consider a three-bedroom home with six baths and the contribution to value added by the fourth, fifth, and sixth baths. Other independent variables that may have a curvilinear relationship to price or rent include property age, floor area, lot area, garage stall count, bedroom count, and proximity (i.e., distance) measures. Furthermore, the nature of the functional relationship between these variables and price or rent can vary by market area whether defined geographically (e.g., region of the country) or economically (e.g., market norms).

Because the underlying functional form of the relationship between an independent variable set and a price or rent outcome variable is unknown, regression model builders must search for the functional form that best fits the data being analyzed. This involves variable transformations such as logarithms, exponents, polynomials, reciprocals, and square roots. In some cases, a transformation applies to an entire equation. In others, transformations apply only to certain variables.

Examples of transformations of entire equations include a hypothesized multiplicative model and a hypothesized exponential model. Transformations are done in these cases to convert the underlying relationships from a nonlinear form to a linear form that is more amenable to regression analysis. These transformations are illustrated in Figure B.4.

In the transformed multiplicative model, the logs of the independent and dependent variables have a linear relationship, and the exponents of these variables are

Figure B.4 Multiplicative and Exponential Model Transformations

Underlying Multiplicative Price (*P*) Model	Underlying Exponential Price (*P*) Model
$P = \alpha x_1^{\beta_1} x_2^{\beta_2} \varepsilon$	$P = e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon)}$
Log Transformation to Linear Form	**Log Transformation to Linear Form**
$\ln (P) = \ln \alpha + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \ln \varepsilon$	$\ln (P) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

transformed into the linear regression coefficient estimates. The estimated coefficients can either be placed into the underlying model to directly estimate price (or value), or the linear model can be used to estimate the log of price, which can then be converted to price. This sort of multiplicative model accommodates a variety of variable relationship shapes, depending on the value of the exponents (the βs). Models of this type are used extensively in mass appraisal for property tax assessment. Transformations of exponential models into the log-linear form and the prior log-log transformation are often useful for controlling heteroscedasticity, a concept that is explained later in this appendix.

It is also possible, and often appropriate, to include other variable transformations. For example, one variable may be curvilinear while others are linear in relation to the dependent variable. The curvilinear variable could be modeled in quadratic form (e.g., floor area) while the other variables are modeled in linear form. An estimation model of this sort would be similar to the following:

$$P = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

In this case x_2 is entered in a quadratic form. If x_2 represented floor area, a positive coefficient for x_2 along with a negative coefficient for x_2^2 could indicate price increasing with floor area but at a decreasing rate as the negative x_2^2 variable diminishes the positive contribution of the x_2 variable. The decision to include a quadratic term should be based on whether its inclusion is theoretically supported and it significantly improves the model, which would be shown by a significant *t*-statistic for the coefficient of the squared variable, improvement in adjusted R^2, or both.

Indicator variables are another form of variable transformation—e.g., the dummy variable used in the apartment illustration to indicate the presence of a pool, spa, and exercise facility in the apartment complex. Indicator variables transform categorical variables into numerical variables so that their effects can be included in a regression model. Dummy variables are the simplest single-category form of indicator variables, coded 1 if the observation is included in the category and 0 if it is not. Sometimes more than one category is required to completely exhaust categorical variable possibilities. For example, suppose a data set spans four years (2005 to 2008), and the year of sale is being entered as a set of indicator variables. Dummy variables would be created for the years 2005, 2006, and 2007, each variable coded 1 or 0 depending on the year of sale for each observation, assuming the valuation date is 2008. The year 2008 is accounted for in the model when the variables $2005 = 0$, $2006 = 0$, and $2007 = 0$. As a result, no variable is created for 2008. The coefficients of the variables 2005, 2006, and 2007 indicate the adjustments required to account for these earlier transactions. The general rules are

1. Create one less dummy variable than the number of categories.
2. All of the dummy variables from an indicator variable set must be included in the model even though some of them may not be significant. That is, the decision to include or exclude a categorical variable implies that all of the dummy variables related to the categorical variable set must either be included or excluded.[1]

1. See Terry Dielman, *Applied Regression Analysis for Business and Economics*, 3rd ed. (Pacific Grove, Calif.: Duxbury, 2001), 406. "[I]ndicator variables are designed to have a particular meaning as a group. They are either all retained in the equation or all dropped from the equation as a group. Dropping individual indicators changes the meaning of the remaining indicators."

Variable Inclusion

Decisions to include or exclude variables determine whether or not a model is underspecified or over-specified. Two problems arise relating to variable inclusion. First, if relevant variables are excluded from a model, the ability of the model to account for change in the dependent variable is diminished. Second, misspecification leads to biased estimates of population parameters (i.e., the independent variable coefficients) because correlation among independent variables causes the model to adjust coefficient estimates when the model is underspecified or overspecified. Coefficients of variables that are included are altered in the regression model to account for their correlations with relevant variables that are excluded. Conversely, coefficients of relevant variables that are included are altered to account for correlations with irrelevant variables that are included.

The apartment unit rent data illustration demonstrates the effect of underspecification. The model was initially underspecified because it included only one independent variable—*living area*. However, three other variables were found to be significant—*bedroom count*, *bath count*, and *on-site pool/spa/exercise facility*. These additional variables are correlated with *living area*. A correlation matrix (Table B.3) quantifies these relationships.

Table B.3 Rent Data Variable Correlations

	Living Area	Bedrooms	Baths	Pool/Spa/Exercise
Living Area	1			
Bedrooms	.780	1		
Baths	.419	.041	1	
Pool/Spa/Exercise	-.493	-.450	-.008	1

Note: Correlation, symbolized as *r*, can range from -1 to +1. Perfect negative correlation is -1, whereas perfect positive correlation is +1. When $r = 0$, two variables are uncorrelated (i.e., independent or orthogonal).

All three of the additional variables are significantly correlated with living area, indicating that omission of these variables from the model would distort the coefficient of the living area variable. This, in fact, occurred. The coefficient of living area was $0.574 per square foot in the simple linear regression model but was reduced to $0.47 in the multiple regression model. The $0.574 coefficient value was distorted by omitting variables that should have been included in the model. The multiple regression model provides a better estimate of unit rent and a less-distorted estimate of the effect of the amount of living area on rent.

In addition, the newly included variables are correlated with each other, sharing some explanatory power. For example, the variables *living area* and *bedrooms* are correlated with *pool/spa/exercise facility*. It appears as though these amenities are more prevalent when the amount of living area is smaller and bedroom counts are lower. As a result of this correlation, the coefficient of the *pool/spa/exercise facility* variable would be distorted if the *living area* and *bedroom count* variables were inadvertently omitted from the model. The multiple regression model provides a better estimate of unit rent and a less-distorted estimate of the contributory value of additional living area, unclouded by the simple regression model's attempt to account for the number of bedrooms and baths and the presence of amenities. Because all four of the variables are significant, all of them should be included in the multiple regression model.

Model Validation

Reference books on statistics offer several suggestions for regression model validation, including

- Collecting new data to assess the model's predictive ability on the new data
- Comparing results with theory and with previously published empirical studies
- Data splitting

Collecting new data is often not a practical option in applied valuation settings. Nevertheless, it is possible and recommended that analysts assess the signs of the variables in the regression equation and compare them with theoretical and intuitive expectations. Staying current on relevant published studies is an obvious priority and needs little discussion. The third option, data splitting, provides the most practical sample-specific and model-specific means of model validation and is worthy of further examination.

Data splitting, which is also known as *cross-validation*, requires that the data set be divided into two subsets: (1) a model-building set and (2) a validation set, usually referred to as a *holdout sample*. The holdout sample, which should be randomly chosen from the full data set, can be a small proportion of the full data set (e.g., 10% to 20%).

Two possible validation routines are recommended. The first routine is to compare the coefficients and significance levels derived from the model-building set with the coefficients and significance levels derived from a regression model using all of the data. The results should be consistent, otherwise a small number of influential observations may be affecting the model disproportionately. The second routine is to use the regression model derived from the model-building set to predict the dependent variable values for the holdout sample. One measure of how well the model predicts is to compute the correlation between the actual values in the holdout sample and the predicted values. The correlation should be high when the model is valid.

If the data set is too small to accommodate data splitting into a model-building sample and a holdout sample, then an alternative, but time-consuming, data-splitting procedure may be employed. The alternative procedure is to (1) remove one observation from the data set, (2) run the regression model with the remaining $n - 1$ observations, (3) use the model to predict the value for the omitted observation, and (4) repeat the procedure by sequentially omitting each observation in turn and reestimating the model and predicting the value for each omitted observation. This procedure will generate n holdout samples of $size = 1$. The predicted value for each holdout observation should correlate highly with the actual observed values. A subroutine in SAS statistical software can automate this procedure. Unfortunately, the procedure cannot be automated in SPSS, Minitab, or Excel.

If the results from these two validation routines are satisfactory, the model is likely to be valid. A final regression model employing all of the data would therefore be appropriate for valuation purposes.[2]

Data Sufficiency

The thought process involved in making a decision regarding how many data observations are necessary for application of a regression model differs from the calcula-

2. See John Neter, William Wasserman, and Michael H. Kutner, *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*, 3rd ed. (Homewood, Ill.: Irwin, 1990), 465-470, for a more complete discussion of model validation.

tion of sample size for inferences about a mean, which was presented in Chapter 14. In regard to a regression model, the measure of data sufficiency is based on degrees of freedom—i.e., the relationship between the number of observations (n) and the number of independent variables in the model (k). When the ratio of n to k is too low, the model is considered "overfitted" and the regression outcome is in danger of being data-specific, not representative of the underlying population.

For example, consider a ratio of n to k of 2:1. It is always possible to connect two points with a straight line. In this case, the coefficient of determination, R^2, would always be equal to 1 in a simple linear regression model. However, the model may not actually explain anything. Since R^2 and the ability to generalize from a sample to a population are affected by the ratio of n to k, many researchers suggest that the minimum ratio should be in the range of 10 to 15 observations per independent variable,[3] with a ratio of 4:1 to 6:1 as an absolute minimum.[4] One indication of an overfit model due to a ratio of n to k that is too low is an increase in adjusted R^2 as the least-significant variables are removed from the model.

The multiple regression model example using the apartment rent data includes 36 observations (n) and four independent variables (k). The ratio of n to k is 9:1, which is less than optimal but more than the absolute minimum. If additional variables such as *apartment age*, *location*, *condition*, *parking ratio*, and the like were to be added to the regression model, then more data would be required to accommodate the expansion of the model.

Underlying Regression Model Assumptions

In addition to the linearity of the relationship of variables, regression modeling has several other important theoretical underpinnings, generally referred to as the *assumptions of regression*.[5] The additional assumptions are that

- Errors are normally distributed.
- Variance is homoscedastic.
- Errors are independent.
- The explanatory variables are not highly interrelated.

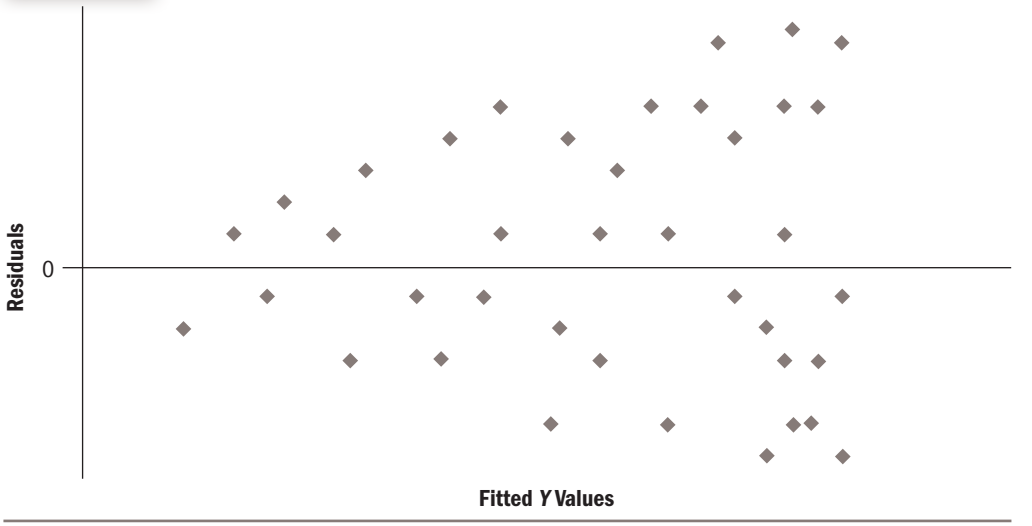
The normality assumption means that the errors around the regression line are normally distributed for each independent variable value. Regression models are fairly resistant to violations of the normality assumption as long as error distributions are not dramatically different from normal.[6] This assumption is important because it is the basis for the validity of the *F*-tests and *t*-tests of model and variable significance, and it provides the mathematical basis for the calculation of confidence intervals. The detrimental effects of non-normality are diminished as sample size increases.

---

3. Joseph F. Hair, Rolph E. Anderson, Ronald L. Tatham, and William C. Black, *Multivariate Data Analysis with Readings*, 3rd ed. (New York: Macmillan, 1992), 46.
4. Hair, et al., caution readers that a ratio of 4:1 is an absolute minimum, whereas Neter, et al., refer to a ratio of 6:1 to 10:1 as a minimum.
5. An easy-to-read and understandable text dealing solely with regression modeling is Terry Dielman, *Applied Regression Analysis for Business and Economics*, 3rd ed. (Pacific Grove, Calif.: Duxbury, 2001). The book devotes an entire chapter to identification of and correction for violations of underlying regression model assumptions.
6. David M. Levine, Timothy C. Krehbiel, and Mark L. Berenson, *Business Statistics: A First Course*, 3rd ed. (Upper Saddle River, N.J.: Prentice Hall, 2003), 436.

Homoscedasticity refers to variation around the regression line that is equal for all values of the independent variable. When this assumption is violated (i.e., when the data is *heteroscedastic*), significant variable coefficients are apt to appear to be insignificant and confidence intervals will be skewed due to systematic variation in error variance. A simple way to check for violation of the homoscedasticity assumption is by examining a plot of residuals against the independent variables or the fitted values of the dependent variable. The data set is homoscedastic, not heteroscedastic, when the distribution of residuals is similar across the range of each independent variable or the fitted values of the dependent variable. A plot showing systematic narrowing or widening of the range of residual values as the values of an independent variable or fitted values of the dependent variable change is an indication of a model that violates the assumption of homoscedasticity. Figure B.5 shows the residuals (e) plotted against the fitted values of the dependent variable ($\hat{Y}$). Note that the residuals are more tightly packed when the fitted values of the dependent variable are small and less tightly packed when the fitted values are large. The data appear to be heteroscedastic, and error variance is directly related to the value of the dependent variable.

Figure B.5 Scatter Plot Illustrating Nonconstant Error Variance



Suggested corrections for violation of this assumption include

- Replacing the values of the dependent variable with the natural logarithm of the dependent variable (i.e., a log transformation)
- Replacing the values of the dependent variable with the square root of the dependent variable (i.e., a square root transformation)

These two transformations replace the dependent variable with less-variable functional forms. However, the replacement variables are undefined for negative numbers. If the size of the residual is correlated with the values of one of the independent variables, then the values of the correlated independent variable can assist in stabiliz-

ing the variance by dividing the regression equation by the correlated independent variable (known as *weighted least squares*). For example, consider a simple regression equation $price = \alpha + \beta\ (size)$, where size is measured in square feet. If a plot shows residual variance increasing as size increases, then division of the model by size should correct the heteroscedasticity problem. The resulting regression model would be

$$\frac{Price}{Size} = \alpha \frac{1}{(Size)} + \beta$$

In the new size-weighted equation, α becomes the regression coefficient on the reciprocal of size and β becomes the constant term. The resulting regression model would estimate price per square foot as a function of the reciprocal of size, which can be easily transformed into a price estimate. More precise corrections can often be obtained by raising the independent variable divisor (the *size* variable in this example) to an exponential power. For instance, SPSS includes a weighted least squares procedure that tries numerous exponents and identifies the one that works best.

Violation of the assumption of error independence most often occurs with time-series data. Residuals in sequential time periods may be correlated as a result of occurrences in a prior time period influencing subsequent time periods. This phenomenon is referred to as *serial correlation* or *autocorrelation*. Variable coefficient estimates remain unbiased under conditions of autocorrelation. However, the standard errors of the coefficients are biased, which affects the validity of *t*-statistics produced by a regression model. The Durbin-Watson test is one well-known means of testing for first-order autocorrelation (i.e., correlation between a residual and the next residual in a time sequence).

High interrelation among independent variables is referred to as *multicollinearity*. When this occurs, the independent variables share explanatory power and consequently the coefficients on the correlated independent variables are biased. Multicollinearity is often difficult to correct. When possible, gathering more data (i.e., increasing *n*) may help. Also, data reduction methods such as factor analysis and the use of proxy variables can be employed to gather correlated variables together into a single representative construct. Ridge regression has also historically been suggested as a means of dealing with multicollinearity.[7]

It is important to note that multicollinearity has no effect on a model's predictive ability, assuming that the model is well specified. Multicollinearity does seriously affect structural interpretation of a model's coefficients. If multicollinearity results in inclusion of superfluous variables that would otherwise be excluded, then the loss in degrees of freedom due to their inclusion will lead to a loss of some predictive power. Investigation of the existence of multicollinearity includes analysis of a matrix of independent variable correlation and an examination of regression model multicollinearity diagnostics including variance inflation factors (VIFs). Most statistical packages will generate VIFs, but they are not available in Excel. The general rule of thumb is that no VIF should be greater than 10 and the mean VIF should not be considerably larger than 1.[8] Note that a VIF of 10 equates to multiple correlation of 0.95, which

7. See also Graeme J. Newell, "The Application of Ridge Regression to Real Estate Appraisal," *The Appraisal Journal* (January 1982): 116-119; Alan K. Reichert, James S. Moore, and Chien-Ching Cho, "Stabilizing Statistical Appraisal Models Using Ridge Regression," *The Real Estate Appraiser and Analyst* (Fall 1985): 17-22; Doug Sweetland, "Ridge Regression: A Word of Caution," *The Appraisal Journal* (April 1986): 294-300; and Jonathon Mark, "Multiple Regression Analysis: A Review of the Issues," *The Appraisal Journal* (January 1988): 89-109.

8. Neter, et al., 409-410.

may be excessive in many instances. Some analysts suggest a maximum VIF of 5 as a criterion for multicollinearity, which implies multiple correlation below 0.90.[9] In the multiple regression example using the garden apartment observations, variance inflation factors are 4.7 (*living area*), 3.4 (*bedrooms*), 1.7 (*baths*), and 1.4 (*pool/spa/exercise*).

Misuse of Statistical Methods

Statistical methods are powerful tools for summarizing and describing data. They are also useful for making inferences about population parameters and the construction of predictive models. Unfortunately, they are also easily and frequently misused. Abuse usually falls into one or both of two categories:

- Overt attempts to mislead
- Ignorance

Manipulating the scale of charts, providing insufficient categories in frequency distributions and related histograms, and intentionally omitting variables in regression models are examples of attempts to mislead. Other practices, such as unknowingly violating the underlying assumptions of regression, using too low a ratio of n to k, and failing to recognize the limitations on sample representativeness, could be the result of simple ignorance.

One rarely discussed problem in appraisal applications of statistical analysis is how well a statistical sample represents the larger population. This problem stems from the fact that real property sales are generally not randomly selected from the population that they are purported to represent. In some instances, sales are representative even though they have not been randomly selected, and inferences are appropriate. However, in other instances, some underlying cause may have had a temporary or location-specific influence on the decision to offer certain properties for sale, and data affected by that influence may not be representative of the market as a whole. In these situations, inferences derived from sales data may not provide a true picture of the overall market.

It is incumbent upon professional analysts to provide charts, tables, and graphs that accurately reflect the data being presented. In addition, those who employ inferential statistical methods should be competent—i.e., educated in inferential methods and experienced with the software being used. The burden of proof of competence and lack of bias ultimately lies with the analyst.

Frequently encountered problems of statistical misuse include

- Failure to fully understand the ramifications of violating the assumptions underlying regression models
- Failure to test and assess the validity of a regression model and its underlying assumptions
- Failure to correct regression models when necessary to adequately comply with the underlying assumptions

Three particularly problematic areas explained earlier are multicollinearity, heteroscedasticity, and autocorrelation. Multicollinearity often results in variable signs that

9. Hair, et al., 48.

are theoretically or intuitively incorrect and the apparent insignificance of variables that share explanatory power. Heteroscedasticity masks the significance of otherwise significant explanatory variables. Autocorrelation fails to account for historical influence on a time-series variable.

Other common problems result from the misspecification of regression models including "overfitting" where the ratio of n to k is too low, inclusion of irrelevant variables, and omission of important variables. Note that inclusion of any variable, relevant or not, will result in an increase in the coefficient of determination, R^2. Adjusted R^2 provides a test of whether inclusion of an additional variable adds sufficient explanatory power. When adjusted R^2 does not increase with the addition of another variable, the additional variable is most likely irrelevant. (The additional variable should probably be included, however, if there is strong theoretical support for its importance to the relationship being studied.)

In conclusion, credible regression modeling includes an assessment of data sufficiency, a residual analysis, an assessment of which variables should be included in a model, and model validation. Regarding data sufficiency, due to the required ratio of n to k, an analyst often has too few observations to facilitate inclusion of all of the variables known or thought to be important. To ensure credibility the analyst must assess the need for and availability of additional data or explore means of variable reduction such as factor analysis or proxy variables. In addition, the appraiser's workfile should include an analysis of residuals regarding the assumptions underlying any model employed and an assessment of functional form and support for the variables included.

As a final caution, be aware that although modern statistical software is easy to use, its use can contribute to the production of a less-than-credible work product when the steps required to ensure credible model building are overlooked.