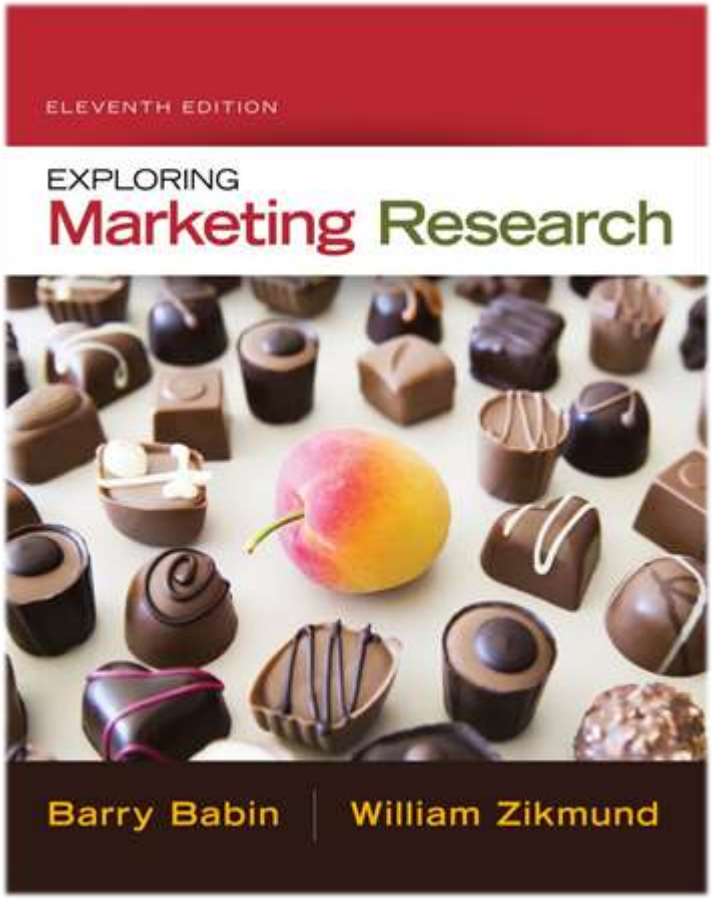


Chapter 18
Advanced Topics in Linear Analytics

LEARNING OUTCOMES

After studying this chapter, you should

1. Understand the meaning of covariance and correlation theoretically
2. Compute a covariance and correlation matrix
3. Separate causal relationships from other types of relationships to build explanations
4. Use multiple regression for predictive purposes
5. Appreciate the arithmetic of OLS

Understanding Covariance and Correlation

- Measures of association capture how much one variable changes as another variable(s) changes
  - Covariance is the absolute amount of association between two variables determined by how a change in one variable corresponds systematically to a change in another
    - The formula is:

$$S_{xy} = \mathbf{Cov}_{(x,y)} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})\ (Y_i - \overline{Y})}{n - 1}$$

    - The population covariance:

$$\sigma_{xy} = \mathbf{Cov}_{(x,y)} = \frac{\sum_{i=1}^{n} (X_i - \mu_x)(Y_i - \mu_y)}{n}$$

Correlation

- A correlation coefficient is a statistical measure of association between two variables expressed on a range of -1 to +1
  - Standardized representation of covariance
    - If the value of *r* equals +1, a perfect positive relationship exists
    - If the value of *r* equals -1, a perfect negative relationship exists
    - If the value of *r* equals 0, no relationship exists
    - A correlation coefficient indicates both the magnitude and the direction of a relationship

Correlation (cont'd.)

- Formula:

$$r_{xy} = r_{yx} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

  - Or

$$r_{xy} = r_{yx} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

  - where

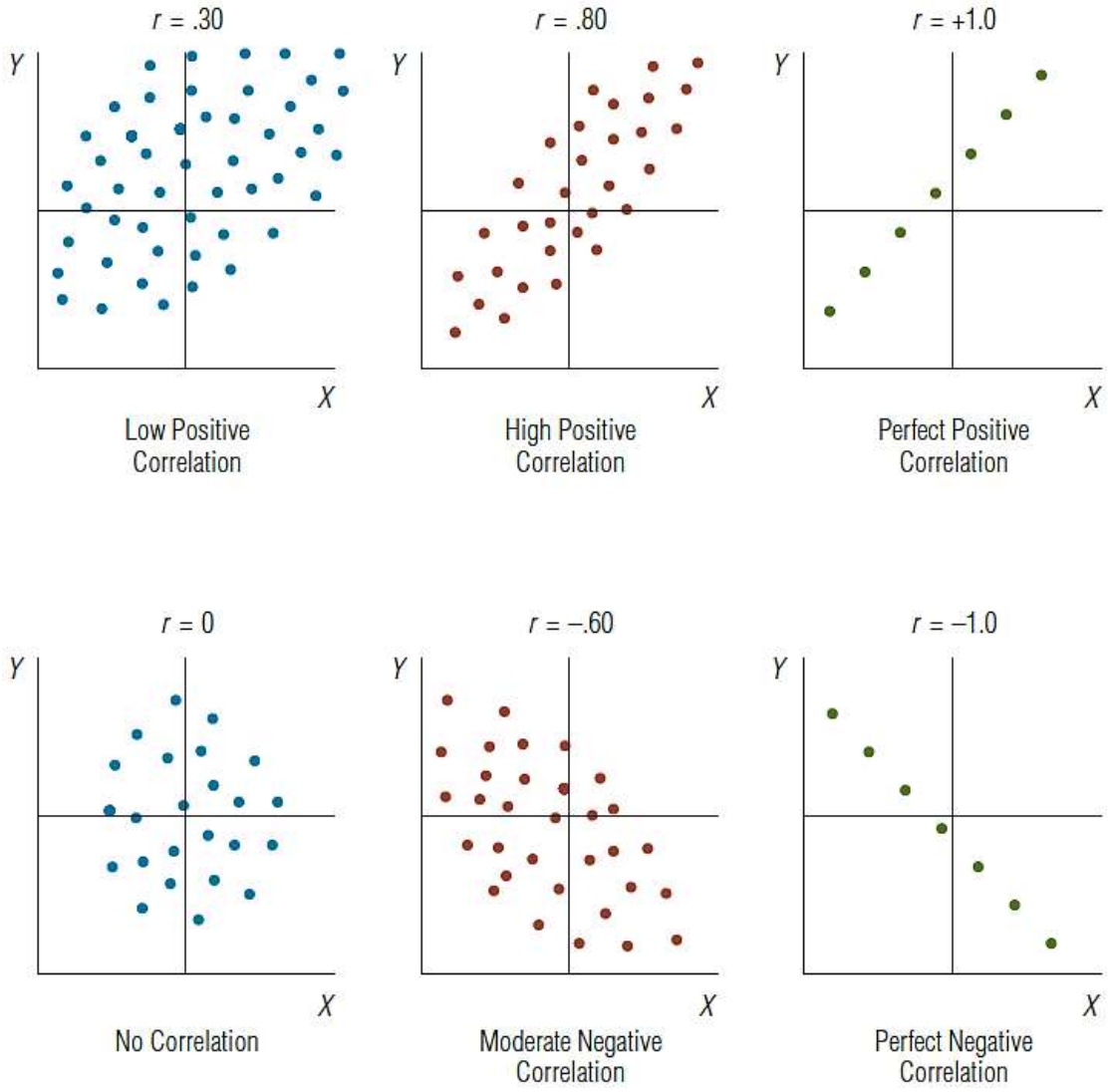
σ_x^2 = overall variance of X

σ_y^2 = overall variance of Y

σ_{xy} = overall covariance of X and Y

  - *X* and *Y* values are plotted on one another in a Cartesian plane is called a scatter plot

EXHIBIT 18.2 Scatter Diagram to Illustrate Correlation Patterns



Correlation Calculation Illustrated

- Consider an investigation made to determine whether the average number of hours worked in manufacturing industries relates to unemployment
  - A correlation analysis of the data is carried out in Exhibit 18.4
    - Notice that the exhibit breaks down the different components needed to compute variances as well as covariance and correlation
      - The correlation between the two variables (hours worked and unemployment rate) is -0.635, indicating a negative (inverse) relationship
      - When the number of hours goes up, unemployment comes down

EXHIBIT 18.4 Correlation Analysis of Number of Hours Worked in Manufacturing Industries with Unemployment Rate

Unemployment Rate (X_i)	Number of Hours Worked (Y_i)	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
5.5	39.6	.51	.2601	−.71	.5041	−.3621
4.4	40.7	−.59	.3481	.39	.1521	−.2301
4.1	40.4	−.89	.7921	.09	.0081	−.0801
4.3	39.8	−.69	.4761	−.51	.2601	.3519
6.8	39.2	1.81	3.2761	−1.11	1.2321	−2.0091
5.5	40.3	.51	.2601	−.01	.0001	−.0051
5.5	39.7	.51	.2601	−.61	.3721	−.3111
6.7	39.8	1.71	2.9241	−.51	.2601	−.8721
5.5	40.4	.51	.2601	.09	.0081	.0459
5.7	40.5	.71	.5041	.19	.0361	.1349
5.2	40.7	.21	.0441	.39	.1521	.0819
4.5	41.2	−.49	.2401	.89	.7921	−.4361
3.8	41.3	−1.19	1.4161	.99	.9801	−1.1781
3.8	40.6	−1.19	1.4161	.29	.0841	−.3451
3.6	40.7	−1.39	1.9321	.39	.1521	−.5421
3.5	40.6	−1.49	2.2201	.29	.0841	−.4321
4.9	39.8	−.09	.0081	−.51	.2601	.0459
5.9	39.9	.91	.8281	−.41	.1681	−.3731
5.6	40.6	.61	.3721	.29	.0841	.1769

Coefficient of Determination

- If we wish to know the proportion of variance in *Y* that overlaps *X* (or vice versa), we can calculate the coefficient of determination by squaring the correlation coefficient:

$$R_2 = r_{xy}^2 = \left(\frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} \right)^2$$

- The coefficient of determination represents a measure of association between a dependent variable and the independent variable(s) employed to predict it

Covariance and Correlation Matrix

- A covariance matrix contains the covariance for every pair of variables among a set of metric variables with the off-diagonal elements displaying covariances and the diagonal elements displaying variance
  - The covariance matrix, often abbreviated with a bold capital **S**, is a square matrix summarizing all the relationships among variables that retains the absolute scale values
  - **S** captures both the common scatter (covariance and variance) and the level (higher values result from larger numbers) of the set of variables

Covariance and Correlation Matrix (cont'd.)

- A correlation matrix is a standardized covariance matrix
  - While it retains all the information about relationships in the off-diagonal elements (the actual correlations), the information about the level of variables is removed as all the variances become standardized to 1
  - Every diagonal element is 1
  - One way to produce the correlation matrix is to first standardize all variables, yielding *Z*-scores that express variable values in standard deviations away from the mean, and then compute all possible covariances
  - Researchers often compute correlation matrices early on as they provide a quick summary of how variables are corresponding with each other

Causality and Explanation

- Three conditions necessary for causal conclusions:
  - Temporal sequence
  - Concomitant variance
  - Non-spurious association
- In regression analysis, we provide evidence of concomitant variance in the form of significant parameter coefficients
  - The issue of temporal sequence must be sorted out logically

Control Variables

- Variables that are not involved in any causal assertion or hypothesis, but are measured and included in an effort to understand the true effect of hypothesized variables on dependent variables
  - The ovals shown in Exhibit 18.5 represent the variance of the independent variable X, the dependent variable *Y*, and a control variable C
  - Exhibit 18.5 also illustrates the notion of multicollinearity, which occurs when multiple predictor variables included in a regression analysis correlate with each other

Control Variables (cont'd.)

- How much correlation, or shared variance, among independent variables is large?
  - The **variance inflation factor (VIF)** represents an overall estimate of variance overlap among the independent variables
  - How much overlap is too much?
    - In typical marketing research, a VIF of or 5 of more is a clear indication of problems with multicollinearity
    - When VIFs are between 2 and 5, the analyst becomes cautious and looks for signs that indicate the results are suspicious
    - The adjusted R^2 takes into account multiple independent variables overlapping, as shown in Exhibit 18.5

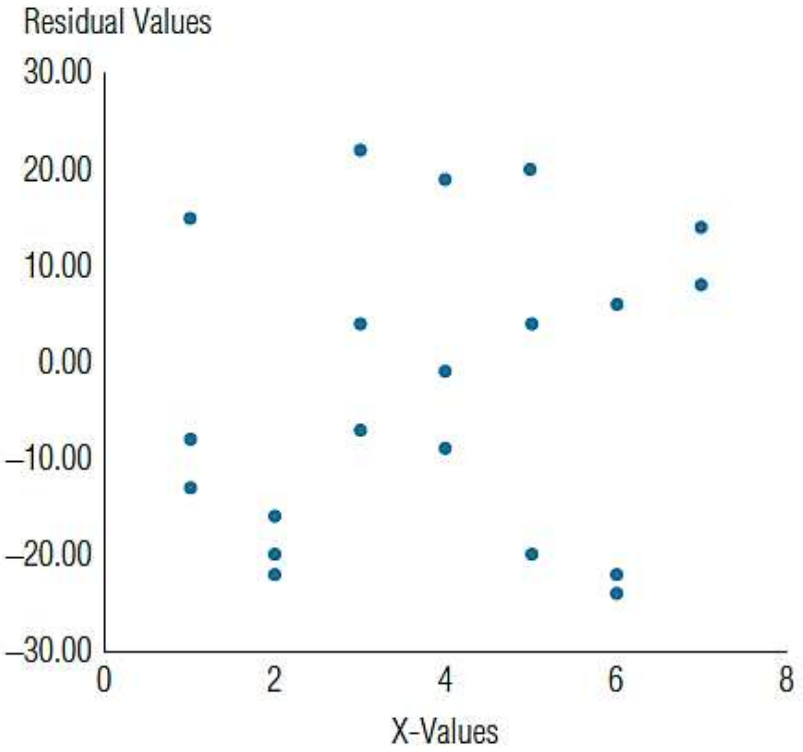
Residuals

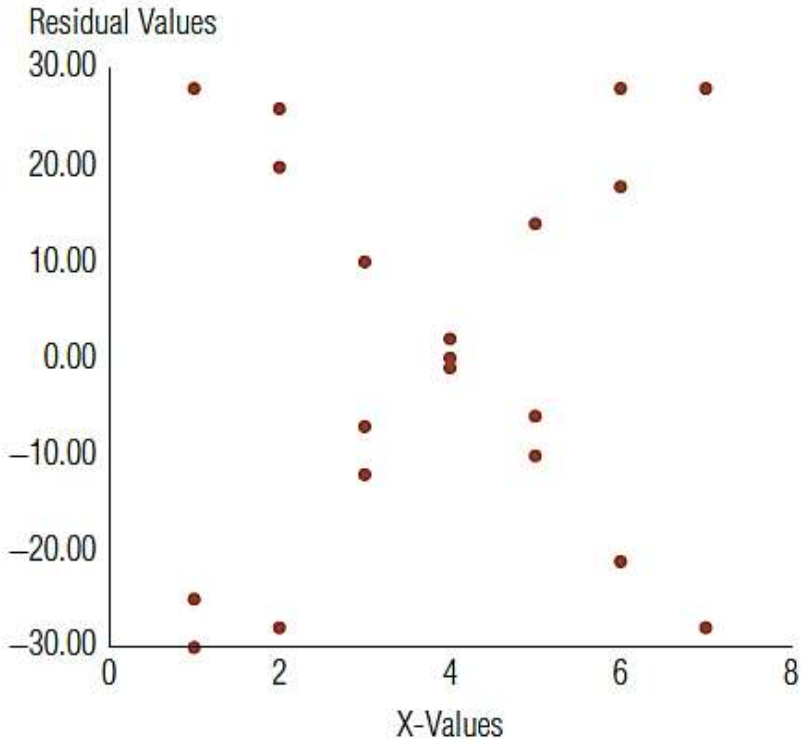
- A residual means the portion left over (unexplained) after the predictor variables are included in a model intended to represent *Y*
  - A major assumption of any GLM, including linear regression, is that the residual terms are independent
    - Most regression programs provide an easy mechanism to plot residuals
    - When residual plots reveal a random distribution of dots, the researcher can conclude that the error is white noise
    - Underspecification means a model does not include a complete set of variables that would be necessary to offer a complete explanation of the dependent variable

Residuals (cont'd.)

- Exhibit 18.7 contrasts plots of systematic versus random residuals
  - In the frame on the left, the plot of residuals against the independent variable, *X*, displays a random pattern with no easily detectable difference in the size of residuals for any value of *X*
    - The right frame depicts larger values (in absolute value) of residuals for smaller values or larger values of *X* and smaller values of residuals for *X*-values close to the scale midpoint of 4
    - The left frame of no pattern depicts a state known as **homoscedasticity**, while the right frame displays a condition of **heteroscedasticity** where the residuals change with the values of some variable

EXHIBIT 18.7 Contrasting Homo- and Heteroscedasticity





Steps in Regression Aimed At Explanation

1. Select the appropriate dependent variable and independent variables based on the research questions or the hypotheses derived from them
2. Select any relevant control variables available in the data set and include as additional independent variables
3. Perform multiple regression with appropriate software
4. Interpret overall Model F and *R squared*. If the model is statistically and practically significant, proceed to step 5

Steps in Regression Aimed At Explanation (cont'd.)

5. Examine VIFs for potential multicollinearity. If no concerns emerge, proceed.
6. Interpret individual parameter estimates. Consider the statistical significance of each variable, the valence of the relationship (positive or negative), and the relative size of the relationship as best captured with the standardized regression coefficients.
7. Examine residual plots. If no patterns emerge, and homoscedasticity is evident, the researcher can confidently interpret the resulting parameter coefficients for explanatory meaning.

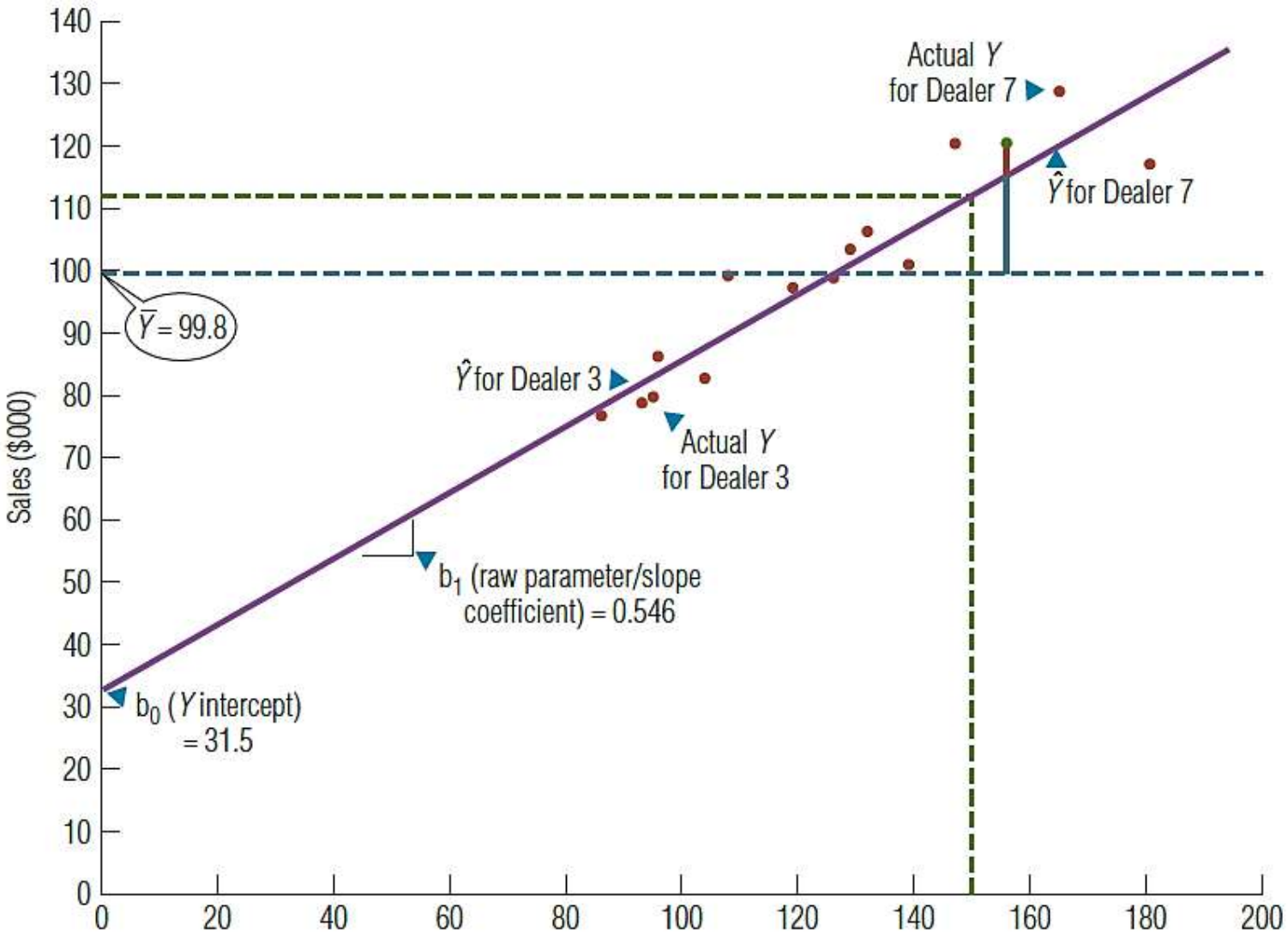
Regression For Prediction

- Not all regressions are intended to explain some dependent variable
  - At times, the researcher's concern is merely prediction
  - In other words, the question involves forecasting what *Y* might be more than on why *Y* comes to take on that value
  - In these cases, the researcher charges the analysts with generating an accurate prediction of *Y* for any values of *X* as a priority over the explanatory interpretation of β
    - Predicted values of *Y* are represented as *Y-hat*

Visual Estimation Of A Simple Regression Model

- One way to determine the relationship between *X* and *Y* is to simply visually draw the best-fit straight line through the points in the figure
  - That is, try to draw a line that goes through the center of the plot of points
  - For any given value of the independent variable, a prediction can be made by selecting the dependent variable that goes along with that value

EXHIBIT 18.10 The Best-Fit Line or Knocking Out the Pins



Errors in Prediction

- We would like an estimation technique that will place our line so that the total sum of all errors over all observations is minimized
  - In other words, one in which no line fits better
- Although with good guess work, visual estimation may prove somewhat accurate, perhaps there is a more certain method

Time-Series Analysis

- Generally, time-series models place an emphasis on prediction over explanation
  - The goal is to forecast some dependent variable for a future time period
    - In terms of explanatory power, a problem with time-series analysis is auto-correlation
    - Auto-correlation refers to the fact that observations from different time periods are almost always correlated with each other
    - Thus, the residuals are unlikely to be independent of each other without corrective measures
    - One simple corrective measure involves including a variable that accounts for the time period

Ordinary Least-Squares Illustrated

- OLS (ordinary least-squares) is a relatively straightforward mathematical technique that guarantees that the resulting straight line will produce the least possible total error in using *X* to predict *Y*

Using Squared Deviations

- The OLS criterion minimizes the total squared error of prediction:

$$\text{Sum of squared errors} = SSE = \sum_{i=1}^{n} e_i^2$$

  - where

$e_i = Y_i - \hat{Y}_i$ (the residual = the difference between the actual observed value and the estimated value of the dependent variable for any of i observations)

Y_i = actual observed value of the dependent variable

$\hat{Y}_i$ = estimated or predicted value of the dependent variable (pronounced Y-hat)

n = number of observations

i = number of the particular observation

Regression Equation & Parameter Estimates

- Regression equation:

$$Y_i = b_0 + b_1 X_1 + e_i$$

- Parameter estimates

$$b_1 = \frac{n(\sum X_i Y_i) - (\sum X_i)(\sum Y_i)}{n(\sum X_i^2) - (\sum X_i)^2} \qquad b_0 = \overline{Y} - b_1 \overline{X}$$

➢ where

Y_i = ith observed value of the dependent variable
X_i = ith observed value of the independent variable
$\overline{Y}$ = mean of the dependent variable
X = independent variable
$\overline{X}$ = mean of the independent variable
n = number of observations
b_0 = intercept estimate
b_1 = slope estimate (regression weight)

Statistical Significance of Regression Model

- An ***F*-test (regression),** or analysis of variance, can be applied to a regression to test the relative magnitudes of the SSR (sums of squares—regression) and SEE (sums of squared errors) with their appropriate degrees of freedom

The Equation for the *F*-test (Regression)

$$F_{(k-1)(n-k)} = \frac{\text{SSR}/(k-1)}{\text{SSE}/(n-k)} = \frac{\text{MSR}}{\text{MSE}}$$

- where

MSR is an abbreviation for mean squared regression

MSE is an abbreviation for mean squared error

k is the number of independent variables (always 1 for simple regression)

n is the sample size

R-Squared

- The coefficient of determination, *R*-squared, reflects the proportion of variance explained by the regression line
- R^2 can be found with this formula:

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

- Good and bad values for the coefficient of determination depend on so many factors that no precise guideline is appropriate
- Thus, the focus should be on the *F*-test