# Chapter Thirteen

Regression and Correlation

**Summary**

• A scatter diagram (also called scatterplot) is a quick visual method used to display relationships between two interval-ratio variables.

• Equations for all straight lines have the same general form:

 = *a* + *b*(*X*)

* The best-fitting regression line is that line where the residual sum of squares, or ∑e2, is at a minimum. Such a line is called the least squares line, and the technique that produces this line is called the least squares method.

• The coefficient of determination (*r*2) and Pearson’s correlation coefficient (*r*) measure how well the regression model fits the data. Pearson’s r also measures the strength of the association between the two variables.

• The general form of the multiple regression equation involving two independent variables is

 = *a* + *b1*(*X1*) + *b2*(*X2*)

**Outline**

* The Scatter Diagram
  + A scatter diagram is a visual method used to display a relationship between two interval-ratio variables
  + A scatter diagram can suggest whether two variables are associated
* Linear Relations and Prediction Rules
  + A linear relationship allows us to approximate the observations displayed in a scatter diagram with a straight line
  + In a perfectly linear relationship, all the observations fall along a straight line
* Constructing Straight-Line Graphs
  + The best-fitting line is the one that generates the least amount of error, also referred to as the residual
  + The least squares line (best-fitting line) is a line where the residual sum of squares is at a minimum
* Statistics in Practice
  + Median household income and criminal behavior
* Methods for Assessing the Accuracy of Predictions
* Prediction errors
* The coefficient of determination (*r*2)
  + A PRE measure of association
  + It reflects the proportion of the total variation in the dependent variable, Y, explained by the independent variable, X
* Testing the Significance of *r2* Using ANOVA
  + Like other descriptive statistics, *r*2 is an estimate based on sample data
  + ANOVA and regression analysis can look very much the same
  + Regression sum of squares (SSR) reflects the improvement in the prediction error resulting from using the linear prediction equation, SST – SSE
  + Residual sum of squares (SSE) is the sum of squared differences between observed and predicted Y
  + Mean squares regression is an average computed by dividing the regression sum of squares (SSR) by its corresponding degrees of freedom
  + Mean squares residual is an average computed by dividing the residual sum of squares (SSE) by its corresponding degrees of freedom
  + In the social sciences, it is the square root of *r*2 or *r*—known as Pearson’s correlation coefficient—that is most often used as a measure of association between two interval-ratio variables
  + Pearson’s *r* is a measure of relationship or association for interval-ratio variables
  + Like gamma, it ranges from 0.0 to ±1.0, with 0.0 indicating no association between the two variables
* Statistics in Practice
  + Teen pregnancy and social inequality
  + Slope (b is the amount of change in a dependent variable per unit change in an independent variable
* Focus on Interpretation
  + The marriage penalty in earning
* Multiple Regression
  + Multiple regression is an extension of bivariate regression
  + It allows us to examine the effect of two or more independent variables on the dependent variable
  + The calculations involved in multiple regression are quite elaborate but are easily accomplished using SPSS or other statistical software
  + Multiple coefficient of determination (*R*2) is a measure that reflects the proportion of the total variation in the dependent variable that is explained jointly by two or more independent variables
  + Pearson’s multiple correlation coefficient (*R*) is a measure of the linear relationship between the independent variable and the combined effect of two or more independent variables
* ANOVA for Multiple Linear Regression
  + The ANOVA summary table for multiple regression is nearly identical to the one for simple linear regression, except that the degrees of freedom are adjusted to reflect the number of independent variables in the model
  + The existence of a correlation only denotes that the two variables are associated and not that they are causally related
    - Correlation is not causation
  + There are numerous examples in the research literature of spurious or confounded relationships
  + Uncovering spurious or confounded relations between an independent and a dependent variable can be accomplished by using multiple regression
  + Multiple regression helps us examine the effect of an independent variable on a dependent variable while holding constant one or more additional variables