Chapter Seventeen

Bivariate Correlation And Regression

LEARNING OBJECTIVES

1. Learn the bivariate analysis of association.
2. Understand bivariate regression analysis.
3. Define the correlation analysis.

Bivariate Analysis of Association

Bivariate Techniques:

- Statistical methods of analyzing the relationship between two variables.

Independent Variable:

- Variable believed to affect the value of the dependent variable.

Bivariate Analysis of Association

Dependent Variable:

- Variable expected to be explained or caused by the independent variable.

Bivariate Regression Analysis:

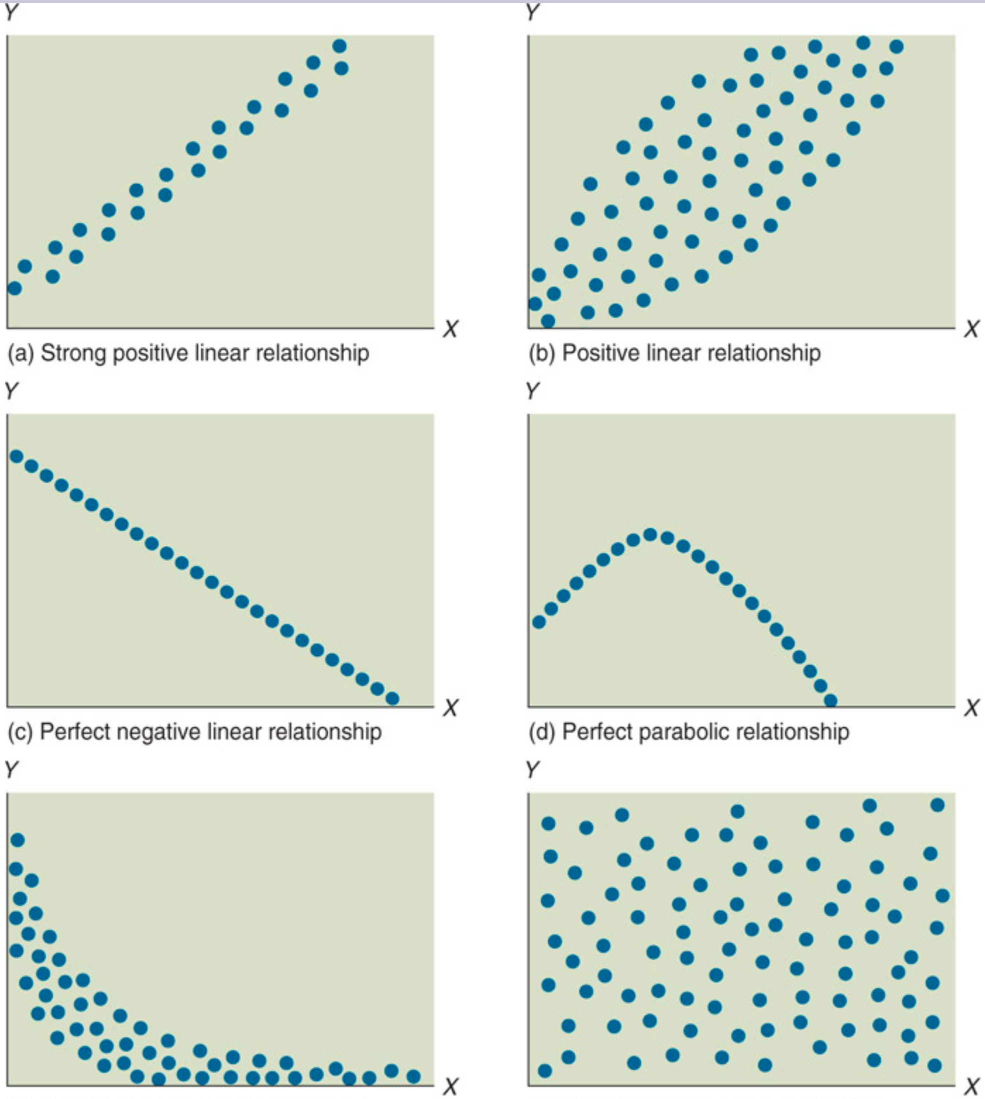
- The analysis of the strength of the linear relationship between variables when one is considered the independent variable and the other is the dependent variable.

Scatter Diagram

Scatter Diagram:

Graphic plot of the data with dependent variable on the Y (vertical) axis and the independent variable on the X (horizontal) axis. Shows the nature of the relationship between the two variables, linear or nonlinear.

Types of Relationships Found in Scatter Diagrams



(a) Strong positive linear relationship

(b) Positive linear relationship

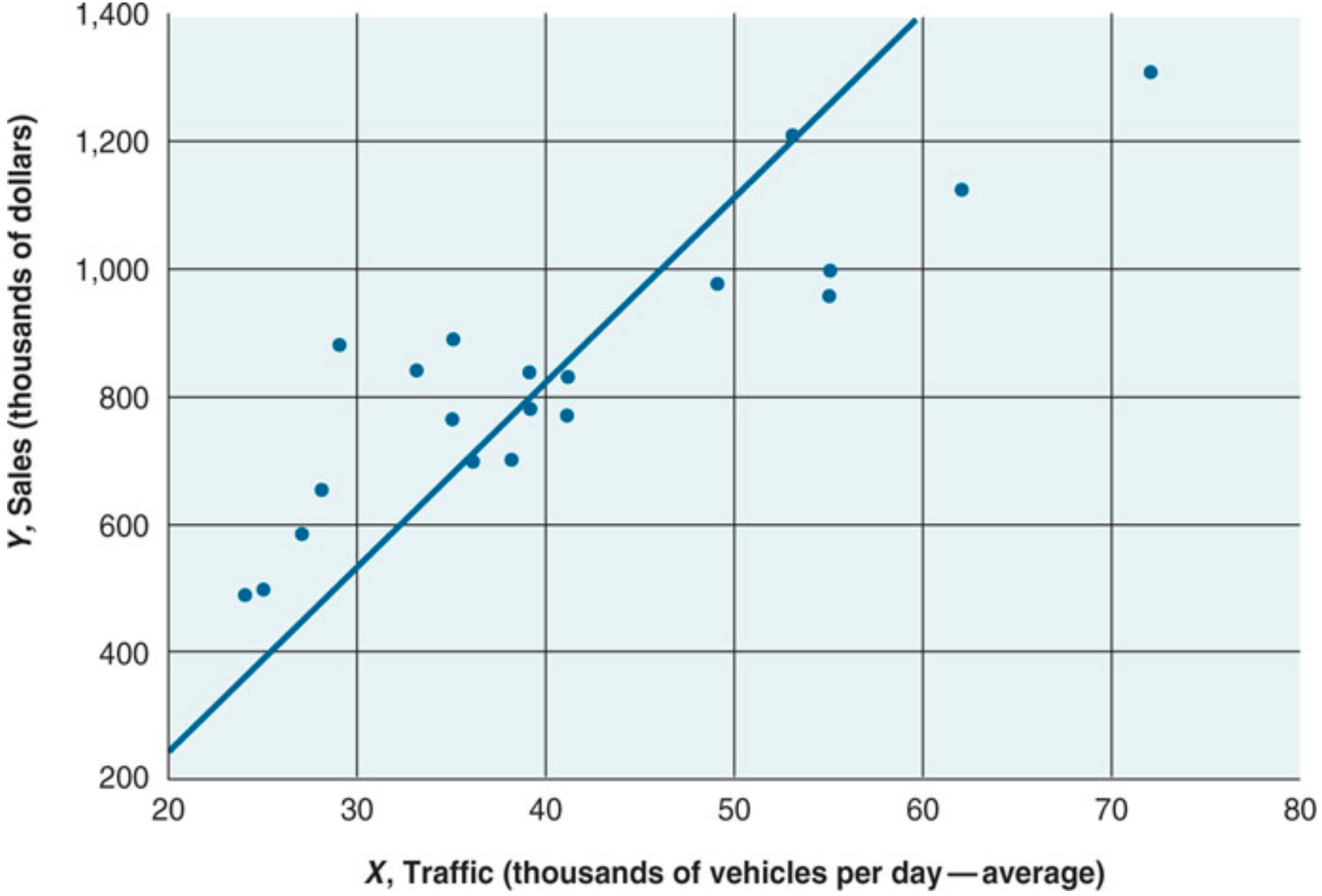
(c) Perfect negative linear relationship

(d) Perfect parabolic relationship

(e) Negative curvilinear relationship

(f) No relationship between *X* and *Y*

Types of Relationships Found in Scatter Diagrams



Least-Squares Estimation Procedure

Estimating the best line of fit:

$$Y = \hat{a} + \hat{b}X + e$$

Where:

Y = dependent variable

$\hat{a}$ = estimated Y intercept

$\hat{b}$ = estimated slope of the regression line

X = independent variable

e = error

Values for “a” and “b” can be calculated as follows:

$$\hat{b} = \frac{\Sigma X_i Y_i \; - \; n\bar{X}\bar{Y}}{\Sigma X^2_i \; - \; n(\bar{X})^2}$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

Where:

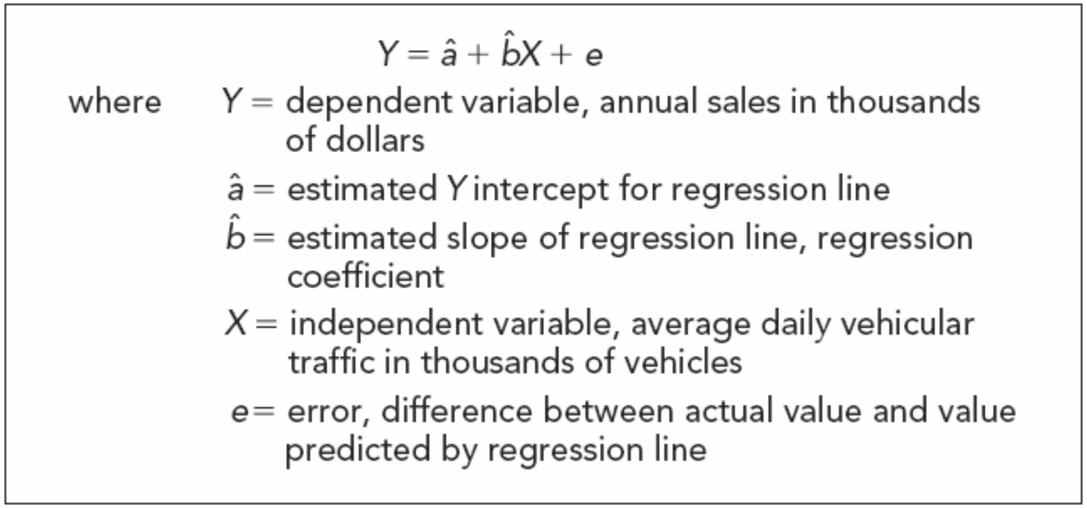
$\bar{\mathbf{X}}$ = mean of value X

$\bar{\mathbf{Y}}$ = mean of value y

n = sample size

Least-Squares Estimation Procedure

The least-squares procedure is a fairly simple mathematical technique that can be used to fit data for X and Y to a line that best represents the relationship between the two variables.



$$Y = \hat{a} + \hat{b}X + e$$

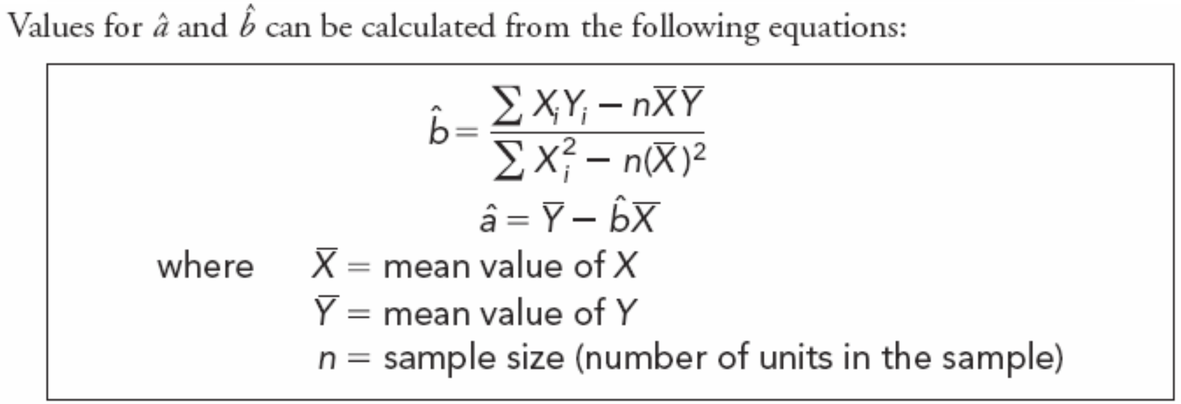
where Y = dependent variable, annual sales in thousands of dollars

$\hat{a}$ = estimated Y intercept for regression line

$\hat{b}$ = estimated slope of regression line, regression coefficient

X = independent variable, average daily vehicular traffic in thousands of vehicles

e = error, difference between actual value and value predicted by regression line



Values for $\hat{a}$ and $\hat{b}$ can be calculated from the following equations:

$$\hat{b} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n(\bar{X})^2}$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

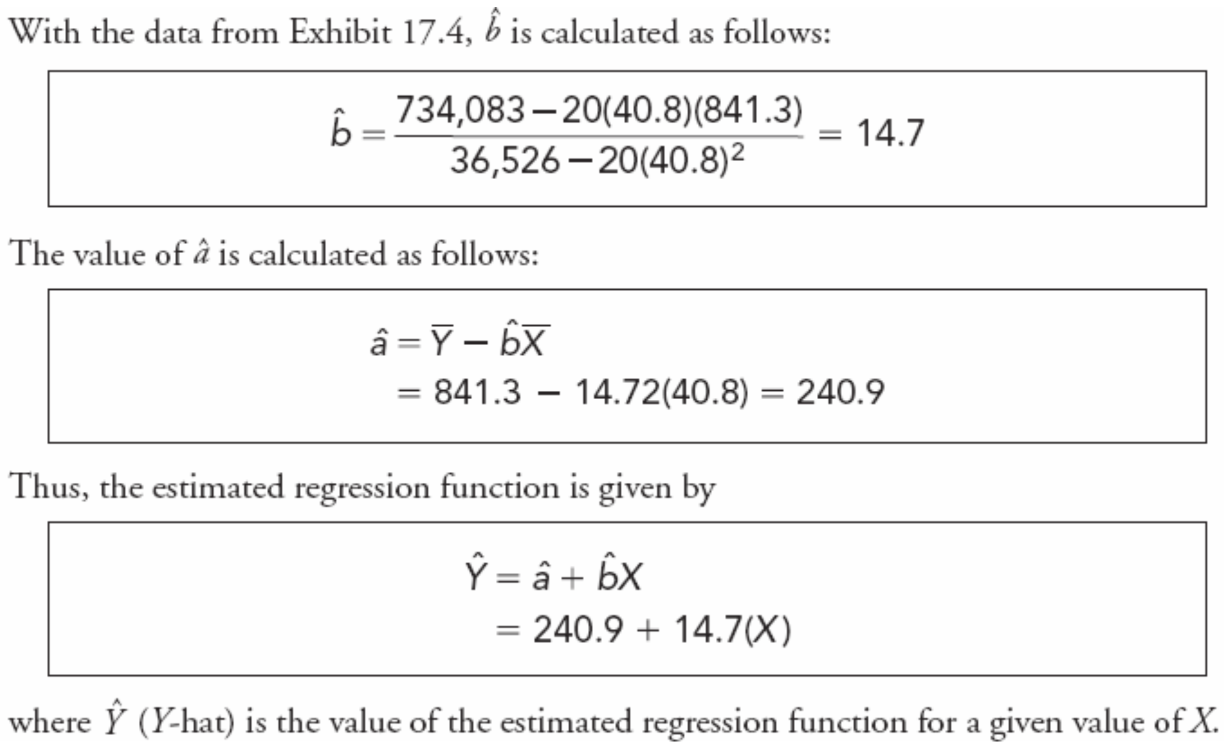
where $\bar{X}$ = mean value of X

$\bar{Y}$ = mean value of Y

n = sample size (number of units in the sample)

Least-Squares Estimation Procedure

The least-squares procedure is a fairly simple mathematical technique that can be used to fit data for X and Y to a line that best represents the relationship between the two variables.



With the data from Exhibit 17.4, $\hat{b}$ is calculated as follows:

$$\boxed{\hat{b} = \frac{734{,}083 - 20(40.8)(841.3)}{36{,}526 - 20(40.8)^2} = 14.7}$$

The value of $\hat{a}$ is calculated as follows:

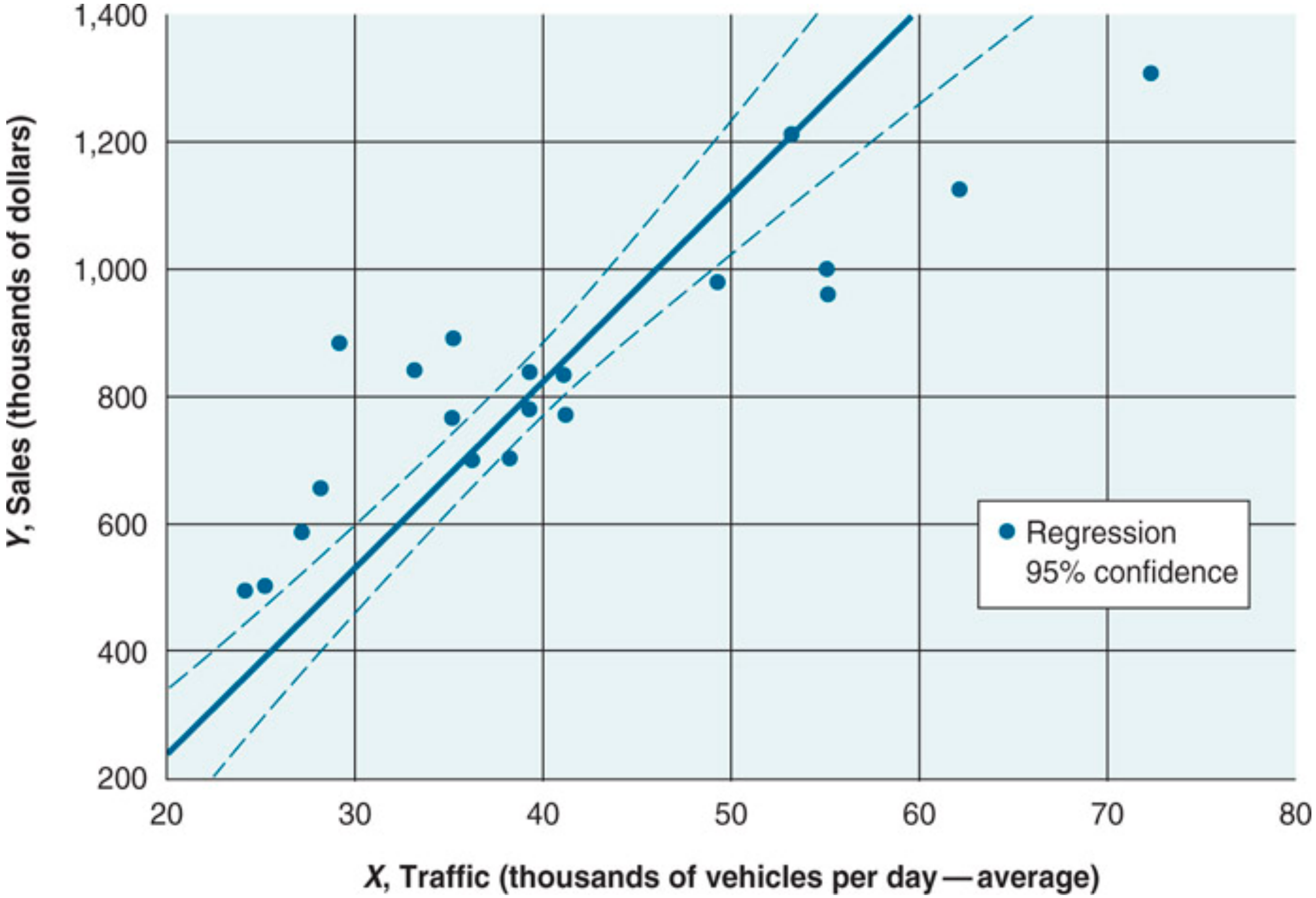
$$\boxed{\begin{aligned} \hat{a} &= \overline{Y} - \hat{b}\overline{X} \\ &= 841.3 - 14.72(40.8) = 240.9 \end{aligned}}$$

Thus, the estimated regression function is given by

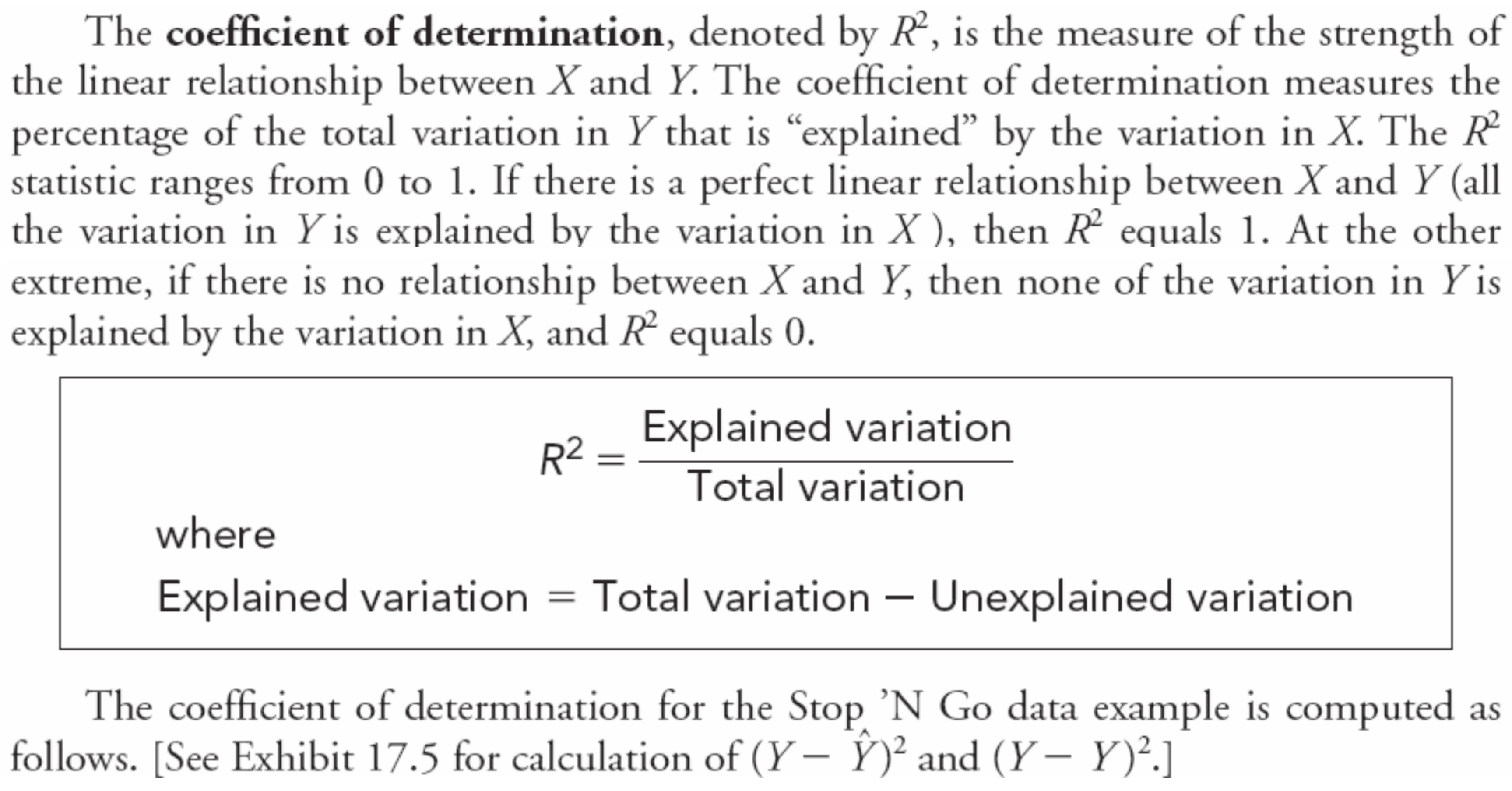
$$\boxed{\begin{aligned} \hat{Y} &= \hat{a} + \hat{b}X \\ &= 240.9 + 14.7(X) \end{aligned}}$$

where $\hat{Y}$ (Y-hat) is the value of the estimated regression function for a given value of X.

Coefficient of Determination



Coefficient of Determination



The **coefficient of determination**, denoted by R^2, is the measure of the strength of the linear relationship between X and Y. The coefficient of determination measures the percentage of the total variation in Y that is "explained" by the variation in X. The R^2 statistic ranges from 0 to 1. If there is a perfect linear relationship between X and Y (all the variation in Y is explained by the variation in X), then R^2 equals 1. At the other extreme, if there is no relationship between X and Y, then none of the variation in Y is explained by the variation in X, and R^2 equals 0.

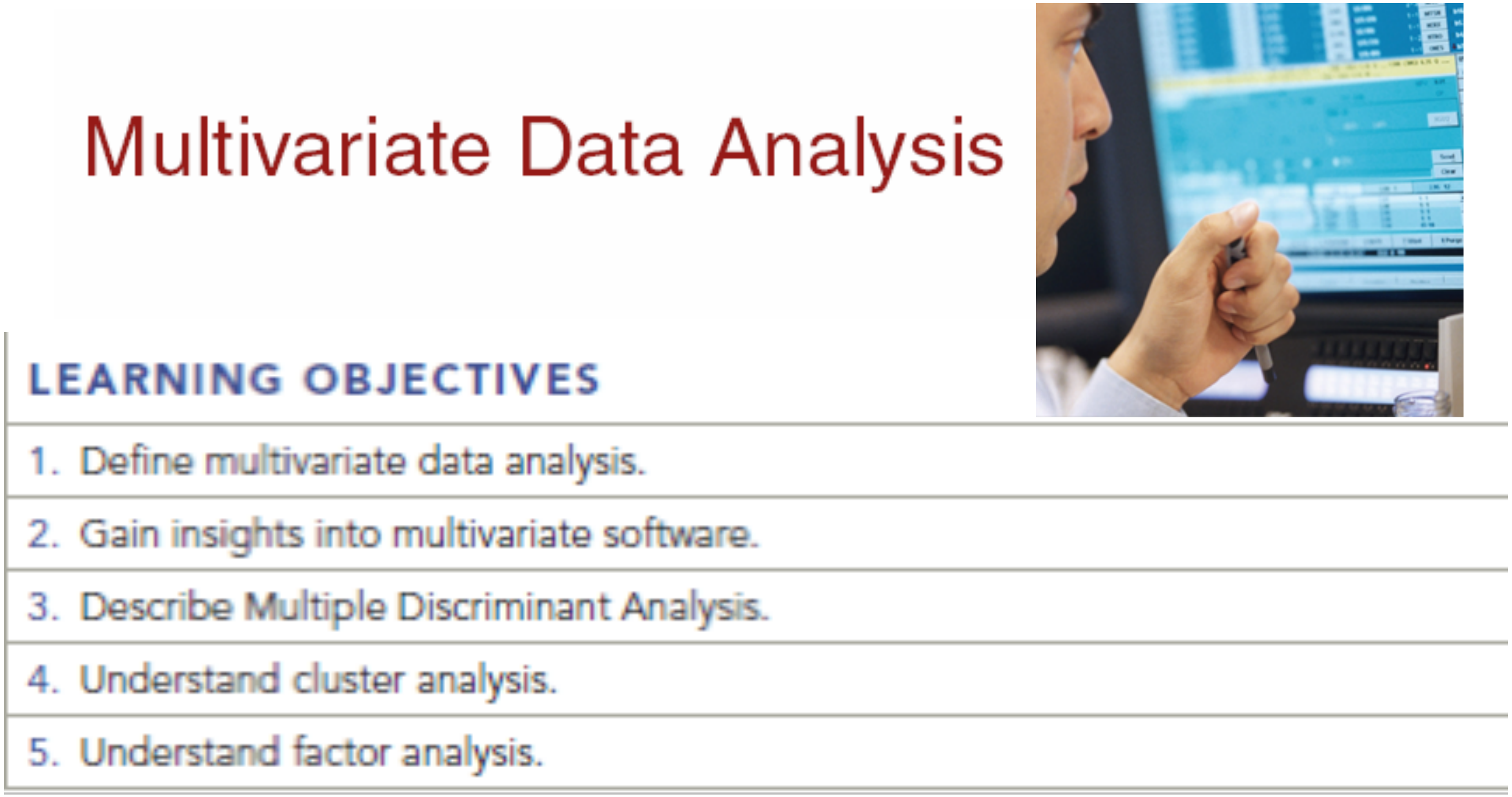
$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

where

Explained variation = Total variation − Unexplained variation

The coefficient of determination for the Stop 'N Go data example is computed as follows. [See Exhibit 17.5 for calculation of $(Y - \hat{Y})^2$ and $(Y - Y)^2$.]

Chapter Eighteen



Multivariate Data Analysis

LEARNING OBJECTIVES

1. Define multivariate data analysis.
2. Gain insights into multivariate software.
3. Describe Multiple Discriminant Analysis.
4. Understand cluster analysis.
5. Understand factor analysis.

Multivariate Analysis

A general term for statistical procedures that simultaneously analyze multiple measurements on each individual or object under study.

Multivariate Regression Analysis

Multiple Discriminant Analysis

Cluster Analysis

Factor Analysis

Conjoint Analysis

Multivariate Analysis

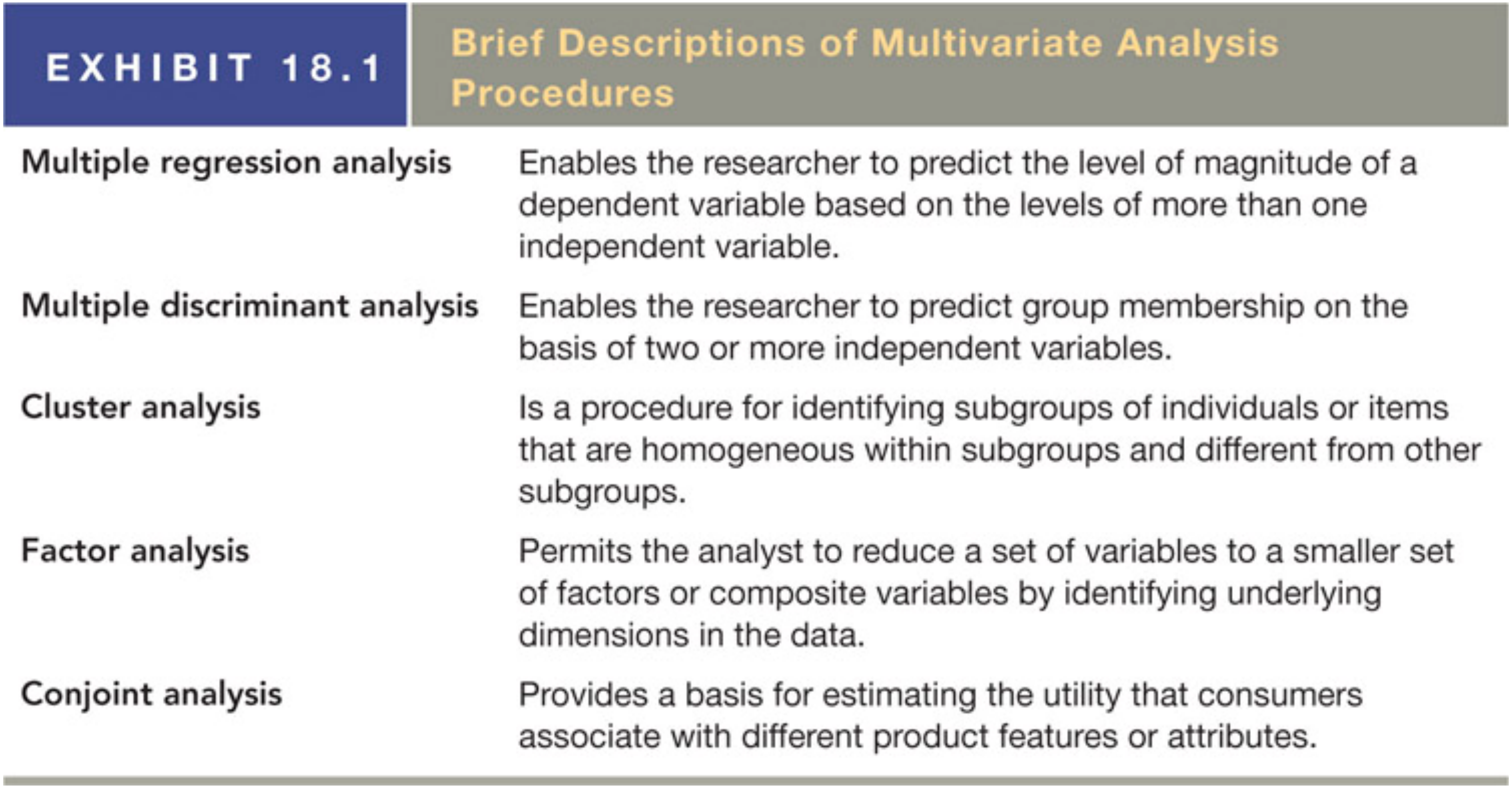


EXHIBIT 18.1 Brief Descriptions of Multivariate Analysis Procedures

Multiple regression analysis	Enables the researcher to predict the level of magnitude of a dependent variable based on the levels of more than one independent variable.
Multiple discriminant analysis	Enables the researcher to predict group membership on the basis of two or more independent variables.
Cluster analysis	Is a procedure for identifying subgroups of individuals or items that are homogeneous within subgroups and different from other subgroups.
Factor analysis	Permits the analyst to reduce a set of variables to a smaller set of factors or composite variables by identifying underlying dimensions in the data.
Conjoint analysis	Provides a basis for estimating the utility that consumers associate with different product features or attributes.

Multiple Regression

Key Concepts

A procedure for predicting the level or magnitude of a (metric) dependent variable based on the levels of multiple independent variables.

Coefficient of Determination:

- Measured changes in the dependent and independent variables.

Regression Coefficients:

- Effect of the independent variable on the dependent variable.

Dummy Variables:

•They're nominally scaled variables included in regression analysis.

Multiple Regression

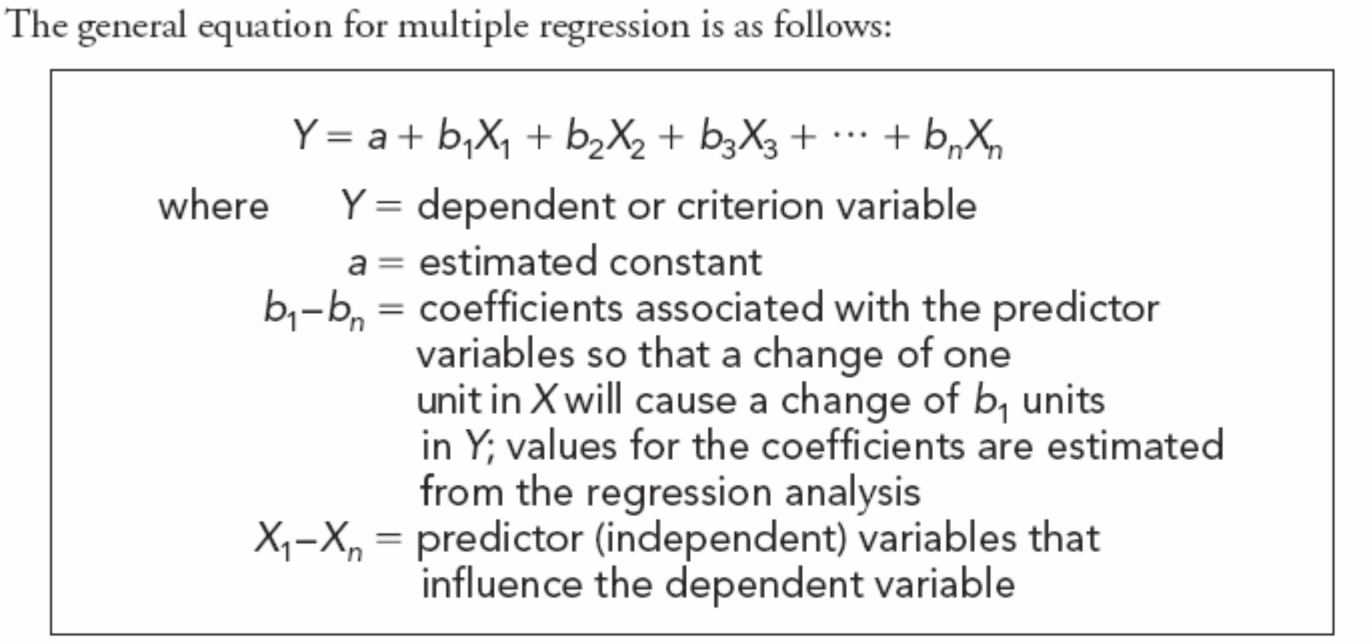
Key Concepts

A procedure for predicting the level or magnitude of a (metric) dependent variable based on the levels of multiple independent variables.

Dummy Variables:

- They're nominally scaled variables included in regression analysis.
- An example of a dummy variable for a measure of location of birth would be
- 0 = born in the United States, 1 = born outside the United States

Multiple Regression Analysis



The general equation for multiple regression is as follows:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_nX_n$$

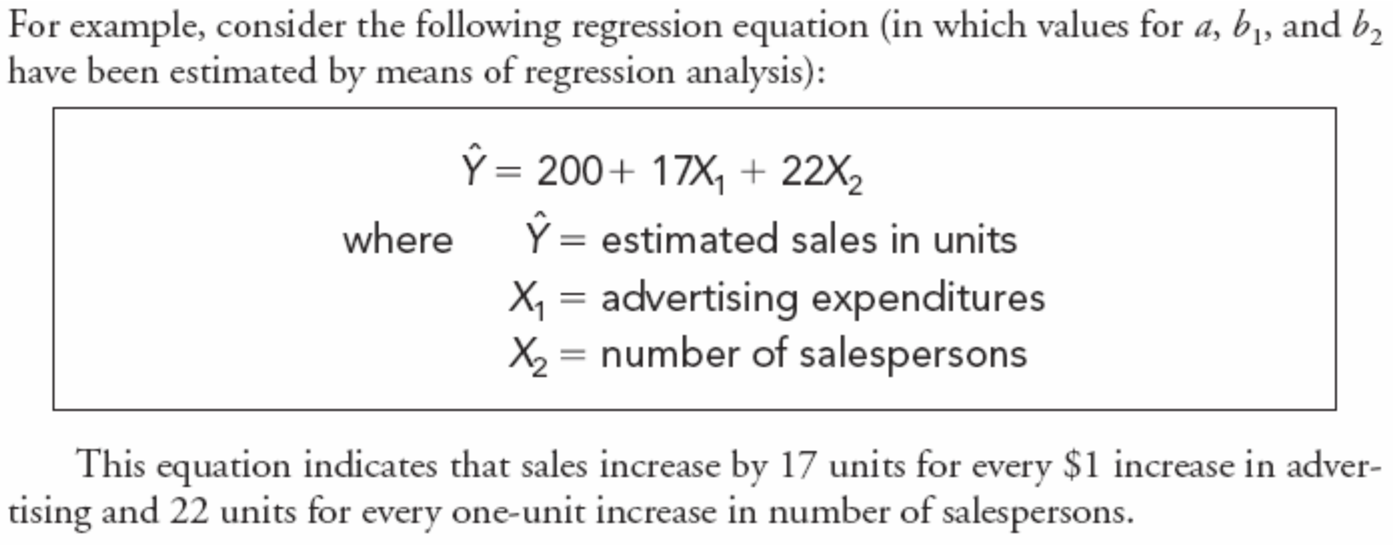
where Y = dependent or criterion variable

a = estimated constant

b_1–b_n = coefficients associated with the predictor variables so that a change of one unit in X will cause a change of b_1 units in Y; values for the coefficients are estimated from the regression analysis

X_1–X_n = predictor (independent) variables that influence the dependent variable

Multiple Regression Analysis



For example, consider the following regression equation (in which values for *a*, b_1, and b_2 have been estimated by means of regression analysis):

$$\hat{Y} = 200 + 17X_1 + 22X_2$$

where $\hat{Y}$ = estimated sales in units
X_1 = advertising expenditures
X_2 = number of salespersons

This equation indicates that sales increase by 17 units for every $1 increase in advertising and 22 units for every one-unit increase in number of salespersons.

Examples

Suppose a marketer was trying to predict the sales (Y) of a product for a given level of advertising (X_1) and sales person performance (X_2). The resulting output was as follows:

$$Y = 2{,}300 + 34(X_1) + 15.5(X_2) \text{ with a } R^2 = .12$$

Given the above results, what would you suggest to the marketer?

Examples

the sales (Y) of a product for a given level of advertising (X_1) and sales person performance (X_2).

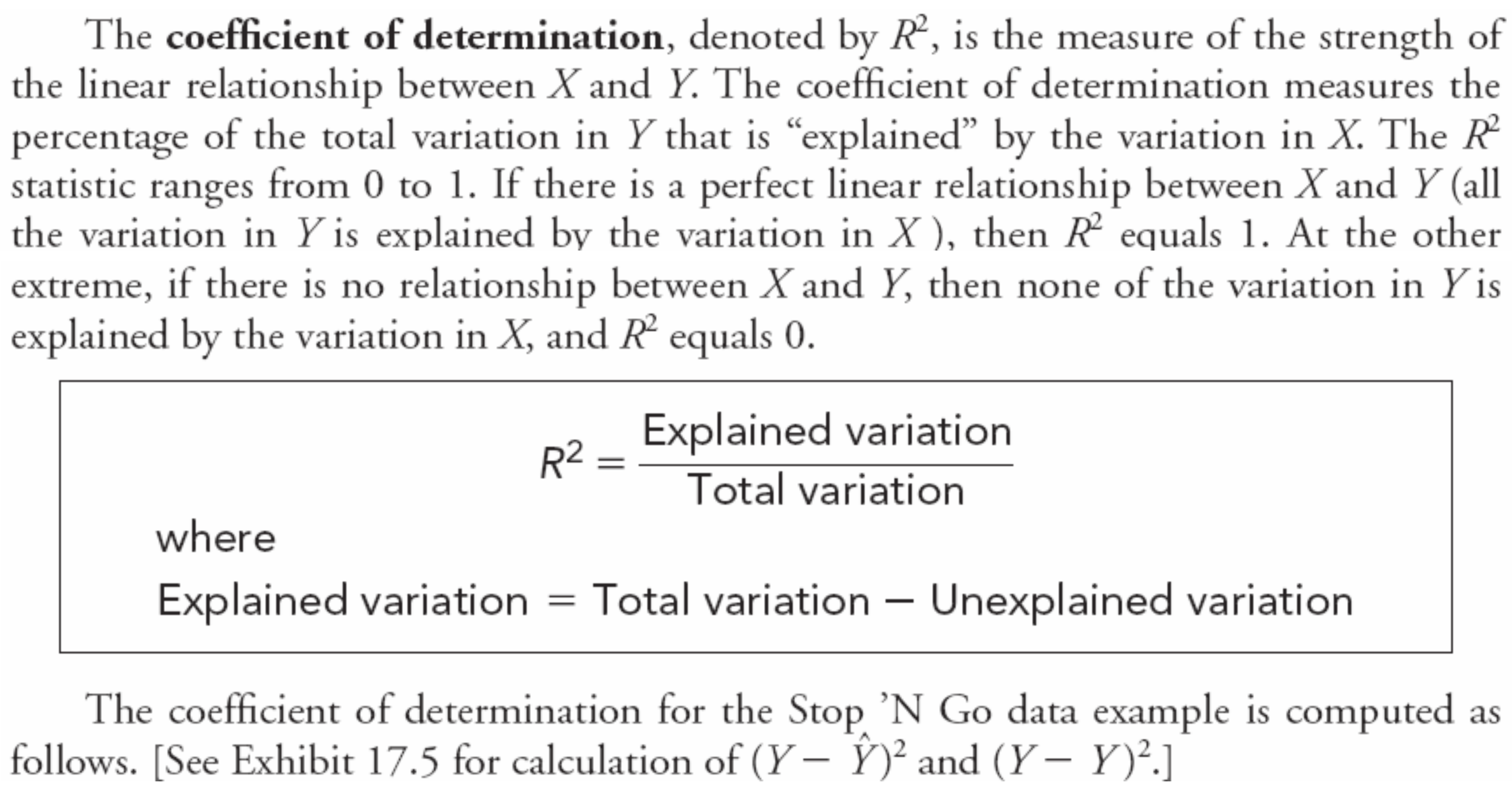
$$Y = 2{,}300 + 34(X_1) + 15.5(X_2) \text{ with a } R^2 = .12$$

Answer: The resulting regression equation depicts a positive relationship between Sales and both Advertising and Salesperson Performance, which means greater inputs of advertising and higher salesperson performance result in higher sales.

However, the R^2 is only .12, meaning that only 12% of the variation in the dependent variable Sales (Y) is explained jointly by Advertising (X_1) and Salesperson Performance (X_2).

There must be other performance variables that would explain more of the variation in Sales. The consultant should suggest additional analysis of the company database to find additional variables that would explain more of the variation in Sales.

Coefficient of Determination



The **coefficient of determination**, denoted by R^2, is the measure of the strength of the linear relationship between X and Y. The coefficient of determination measures the percentage of the total variation in Y that is "explained" by the variation in X. The R^2 statistic ranges from 0 to 1. If there is a perfect linear relationship between X and Y (all the variation in Y is explained by the variation in X), then R^2 equals 1. At the other extreme, if there is no relationship between X and Y, then none of the variation in Y is explained by the variation in X, and R^2 equals 0.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

where

Explained variation $=$ Total variation $-$ Unexplained variation

The coefficient of determination for the Stop 'N Go data example is computed as follows. [See Exhibit 17.5 for calculation of $(Y - \hat{Y})^2$ and $(Y - Y)^2$.]

Examples

- An advertising agency has been doing work for a client selling widgets.
- The three-month campaign has produced a low correlation between advertising expenditures and sales for its client.
  - Hence, the client is considering firing the ad agency.
- The ad agency counters that consumer sales: not a fair assessment of the effectiveness of the ad campaign after only three months.
- An analysis of advertising expenditures in relation to
  - number of requests for information about the widgets;
  - number of distributors stocking widgets;
  - number of retailers requesting shipments of widgets.

Examples

- What kind of analysis would best assist the ad agency in making their case for the effectiveness of their ad campaign?

- A multiple regression analysis.
  - requests for information about widgets,
  - number of distributors stocking widgets,
  - number of retailers requesting shipments of widgets.

  - AD Exp. = $B_0 + B_1(X_1) + B_2(X_2)+B_3\ (X_3)$

Multiple Discriminant Analysis

Multiple Discriminant Analysis:

Procedure for predicting group membership for a (**nominal or categorical**) dependent variable on the basis of two or more independent variables.

Metric Scale:

A type of quantitative that provides the most precise measurement.

Nominal or Categorical:

A type of non-metric qualitative data scale that only uses numbers to indicate membership in a group (e.g., 1=male, 2=female). Most mathematical and statistical procedures cannot be applied to nominal data.

Multiple Discriminant Analysis

Discriminant Score:

Score that is the basis for predicting to which group a particular object or individual belongs; also called Z score.

Discriminant Coefficient:

Estimate of the discriminatory power of a particular independent variable; also called discriminant weight.

Classification Matrix:

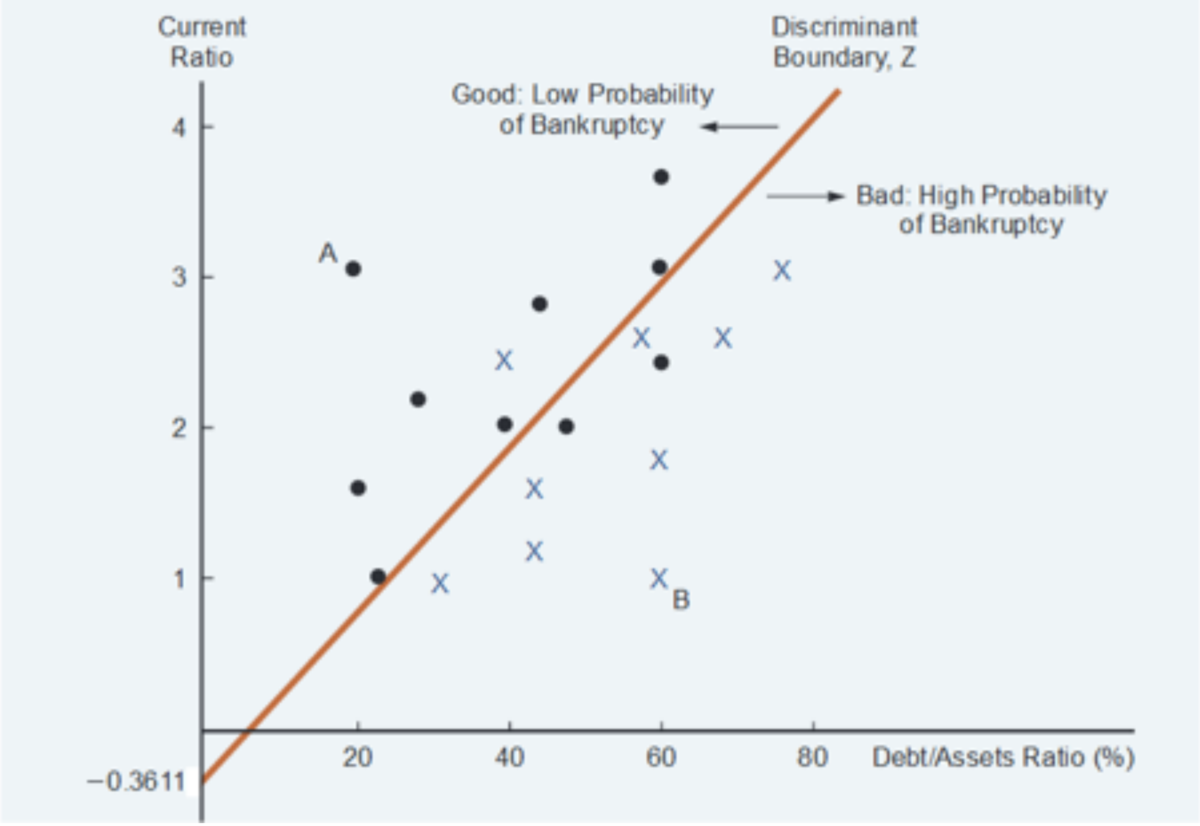
A matrix or table that shows the percentages of people or things correctly and incorrectly classified by the discriminant model.

Multiple Discriminant Analysis

- Z=a+b1* (current ratio)+b2(Debt ratio)
- Fit a discrimination function using historic data.
- Plot the equation in the figure.
- Companies lie to the left of the line(Z<0) are unlikely to go to bankrupt,
- Companies lie to the right (Z>0) are likely to go bankrupt.

Multiple Discriminant Analysis

FIGURE 25B-1 Discriminant Boundary between Bankrupt and Solvent Firms



Key difference

- In the case of multiple regression analysis, the dependent variable must be **metric**;
- In multiple discriminant analysis, the dependent variable is **nominal or categorical** in nature.

Example

A client has a data set that would be appropriate for a linear model. He wants the resulting model to predict whether or not a person would buy a particular product.

The client has been advised that a multiple regression analysis would be the best approach. What would you suggest?

Example

A client has a data set that would be appropriate for a linear model. He wants the resulting model to predict whether or not a person would buy a particular product. The client has been advised that a multiple regression analysis would be the best approach. What would you suggest?

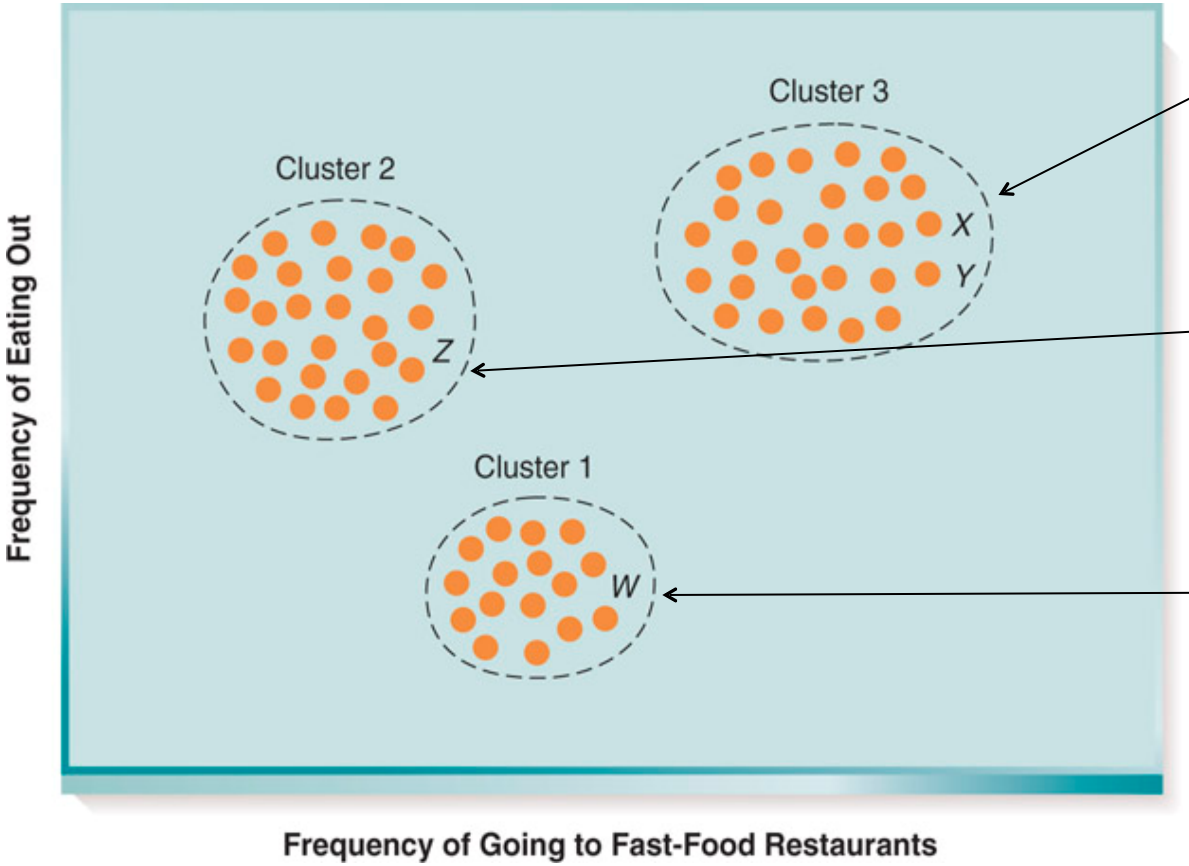
Ans: Multiple regression analysis would be appropriate if the goal is to predict what sales would be given a set of predictor variables.
However, if the goal was to simply predict whether or not a particular type of purchase behavior would occur, then the proper method would be multiple discriminant analysis.

Cluster Analysis

- Cluster analysis
  - A multivariate approach for grouping observations based on similarity among measured variables.
    - Cluster analysis is an important tool **for identifying market segments**.
    - Cluster analysis classifies individuals or objects into a small number of mutually exclusive and exhaustive groups.
    - Objects or individuals are assigned to groups so that there is great similarity within groups and much less similarity between groups.
    - The cluster should have high internal (within-cluster) homogeneity and external (between-cluster) heterogeneity.

Cluster Analysis

The general term for statistical procedures that classify objects, or people, into some number of mutually exclusive and exhaustive groups on the basis of two or more classification variables.



People who frequently eat out and also frequently eat at fast-food restaurants.

Consumers who frequently eat out but seldom eat at fast-food restaurants.

People who do not frequently eat out or frequently eat at fast-food restaurants.

Examples

- An analyst is trying to group together persons who attend theatre performances and who prefer a dinner to accompany the theatre performance. Hence, the two pieces of information are 1) frequency of attending theatre performances and 2) preference for a dinner-theatre format.

- What type of analysis should be undertaken to determine if those persons preferring the dinner-theatre format are those who attend theatre performances more frequently?

Examples

- An analyst is trying to group together persons who attend theatre performances and who prefer a dinner to accompany the theatre performance.
- Hence, the two pieces of information are
  - 1) frequency of attending theatre performances
  - 2) preference for a dinner-theatre format.
- What type of analysis should be undertaken to determine if those persons preferring the dinner-theatre format are those who attend theatre performances more frequently?

Examples

- This problem would be best solved using cluster analysis, which would simultaneously group persons preferring the dinner-theatre format with how frequently they attend theatre performances

Techniques of Cluster Analysis

- Which of the following is not an example of a cluster technique?

K-means

Nearest neighbor

Decision trees

BIRCH

All of these are examples

Factor Analysis

Factor: A linear combination of variables that are correlated with each other.

A procedure for simplifying data by reducing a large set of variables to a smaller set of factors of composite variables by identifying dimensions of the data.

Factor Analysis

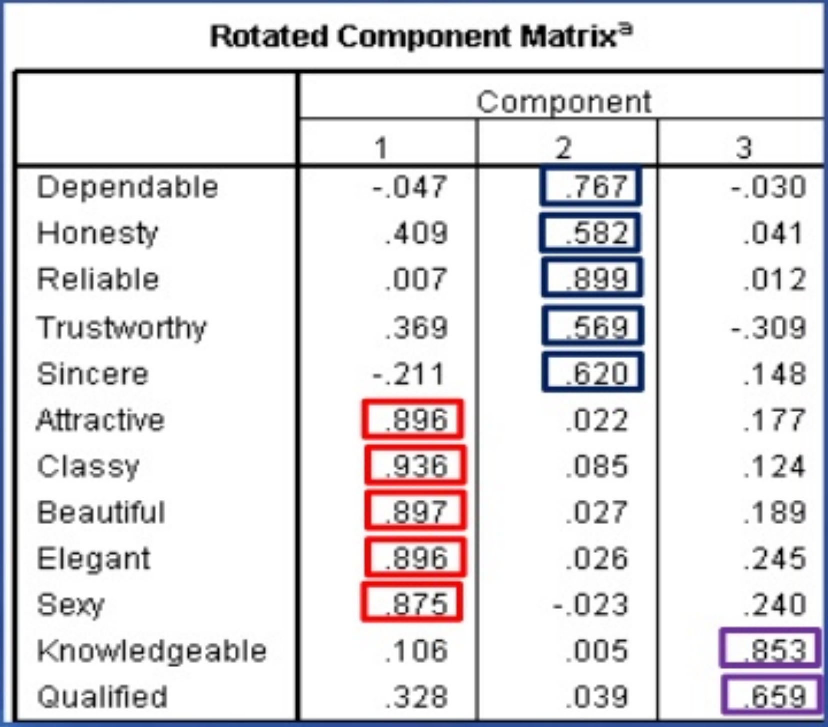
- Data Reduction Technique
  - Overlapping information among p variables
- For a set of p variables we want to extract a smaller set of n factors that adequately describes the key measures of differentiation
  - Uncover the salient underlying dimensions
- The correlation matrix
- If x1 and x2 are highly correlated the two variables may reflect the same latent construct

Questionnaire to obtain responses regarding the celebrity for a Jewelry Brand

- How do you rate the celebrity “ Katrina Kaif “ on the following factors:

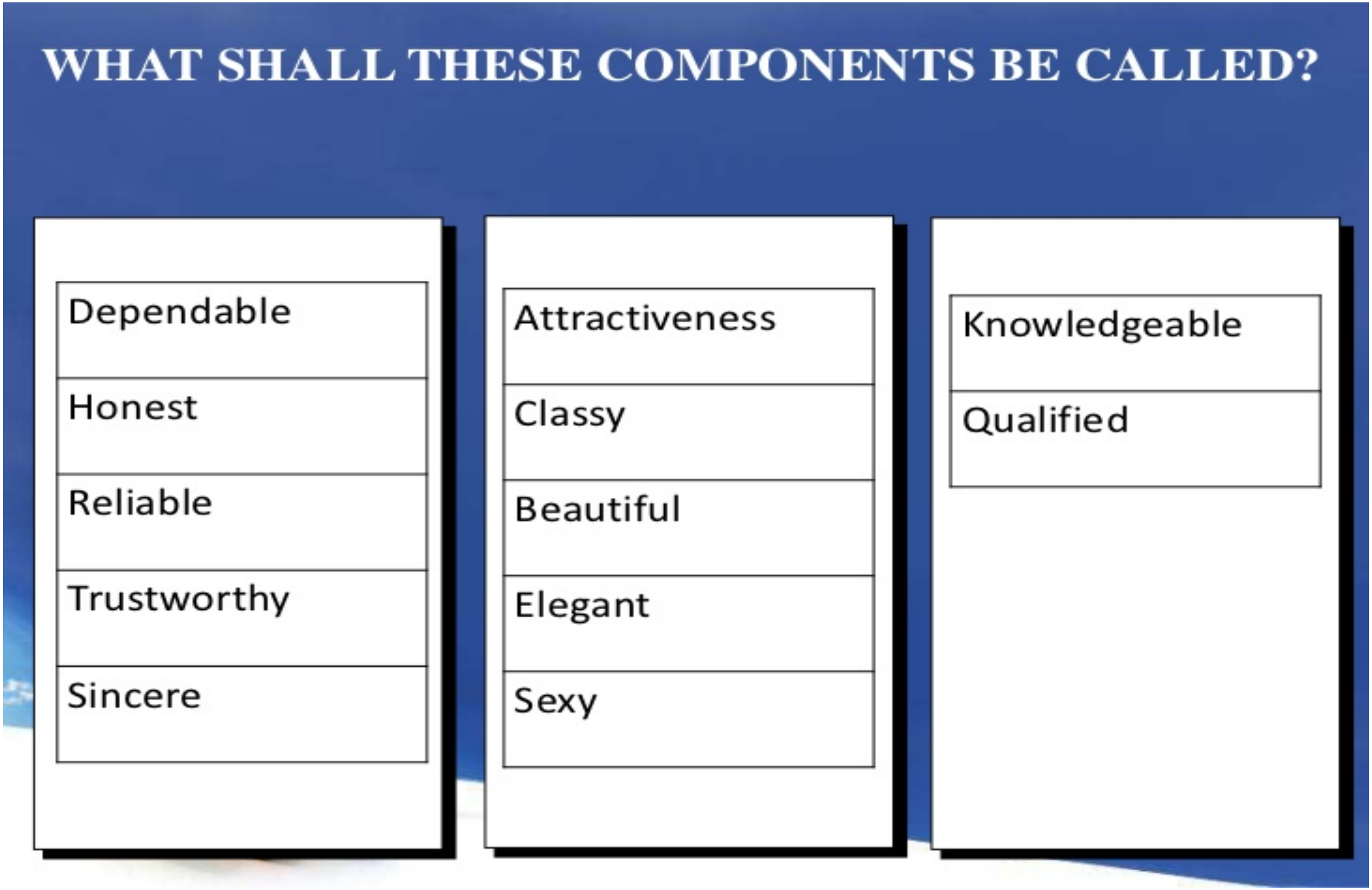
Dependable 1 2 3 4 5
Classy 1 2 3 4 5
Beautiful 1 2 3 4 5
Elegant 1 2 3 4 5
Attractive 1 2 3 4 5
……. 1 2 3 4 5

COMPONENT MATRIX OBTAINED FROM SPSS



Rotated Component Matrix[a]

	Component		
	1	2	3
Dependable	-.047	.767	-.030
Honesty	.409	.582	.041
Reliable	.007	.899	.012
Trustworthy	.369	.569	-.309
Sincere	-.211	.620	.148
Attractive	.896	.022	.177
Classy	.936	.085	.124
Beautiful	.897	.027	.189
Elegant	.896	.026	.245
Sexy	.875	-.023	.240
Knowledgeable	.106	.005	.853
Qualified	.328	.039	.659

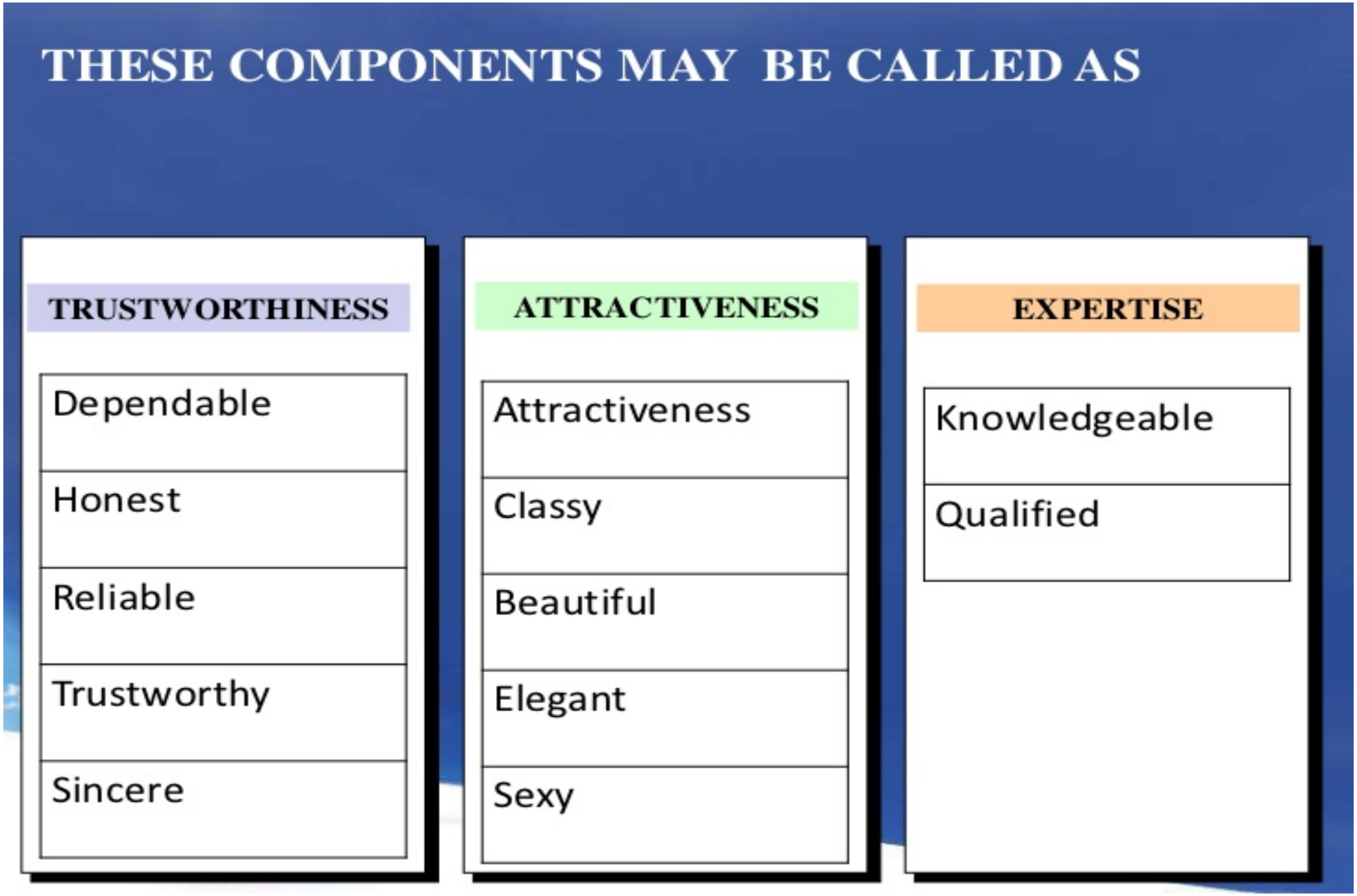


WHAT SHALL THESE COMPONENTS BE CALLED?

- Dependable
- Honest
- Reliable
- Trustworthy
- Sincere

- Attractiveness
- Classy
- Beautiful
- Elegant
- Sexy

- Knowledgeable
- Qualified



THESE COMPONENTS MAY BE CALLED AS

TRUSTWORTHINESS

- Dependable
- Honest
- Reliable
- Trustworthy
- Sincere

ATTRACTIVENESS

- Attractiveness
- Classy
- Beautiful
- Elegant
- Sexy

EXPERTISE

- Knowledgeable
- Qualified

Examples

- A client wanted to know which preventive health care information (PHC) sources were related to each other. The following table has factor loadings for 10 types of preventive health care information using a Varimax rotation. Based on the table of results below, what would you tell your client?

Sources of PHC information	**Factor 1**	**Factor 2**	**Factor 3**
Web MD	.88	.15	.23
Nutritional labels	.11	.67	.14
Health fairs	.25	.32	.88
Health magazines	.22	.74	.30
Health books	.43	.64	.31
Hospital seminars	.03	.21	.85
Television	.70	.31	.34
Wellness center newsletters	.27	.70	.20
Medical Encyclopedias	.33	.76	.31
Hospital websites	.72	.42	.19

Examples

a) Analysis of the factor loadings coupled with the number of factors developed during the principal components step, there are 3 distinct factors or groups of PHC information that co-vary together.

b)The next step for the analyst would be to examine the factor loadings and assign the individual factors (sources of PHC info) to the composite factor group they are related to.

c)
Factor 1: Electronic and video media
Factor 2: Print Media
Factor 3: Institutional related health events

Question

- Independent variables are important components in multiple regression analysis and ___________________.

  a. multiple discriminant analysis

  b. factor analysis

  c. conjoint analysis

  d. cluster analysis

Question

Cluster analysis is particularly valuable for what type of marketing strategy?

a. product differentiation

b. positioning

c. segmentation

d. cost leadership