UNIT III - DESCRIBING RELATIONSHIPS

Correlation –Scatter plots –correlation coefficient for quantitative data –computational formula for correlation coefficient – Regression –regression line –least squares regression line – Standard error of estimate – interpretation of r2 –multiple regression equations –regression towards the mean

3.1. Introduction to Correlation

3.1.1. Correlation

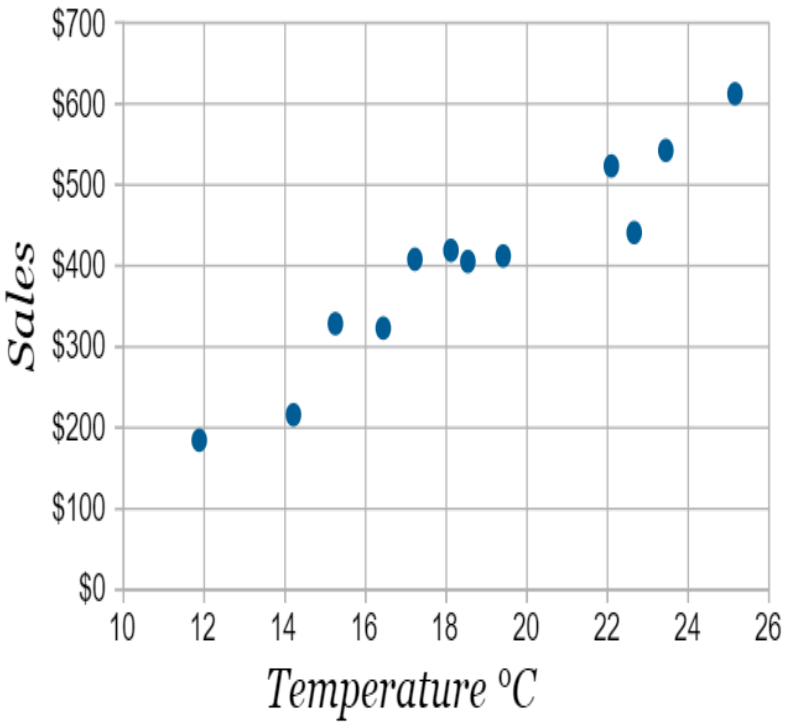
- Correlation refers to a process for establishing the relationships between two variables.
- Correlation is used to test relationships between quantitative variables or categorical variables.
- In other words, it's a measure of how things are related.
- The study of how variables are correlated is called correlation analysis.

3.1.2. Scatter Plots

- A graph containing a cluster of dots that represents all pairs of scores
- A scatter diagram is a diagram that shows the values of two variables X and Y, along with the way in which these two variables relate to each other. The values of variable X are given along the horizontal axis, with the values of the variable Y given on the vertical axis.

Example1: A dataset contains the sales of the ice-cream based on temperature

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



We can easily see that warmer weather and higher sales go together. The relationship is good but not perfect.

Example2: A dataset for greeting card exchange

Table 6.1
GREETING CARDS SENT AND RECEIVED BY FIVE FRIENDS

	NUMBER OF CARDS	
FRIEND	SENT	RECEIVED
Andrea	5	10
Mike	7	12
Doris	13	14
Steve	9	18
John	1	6

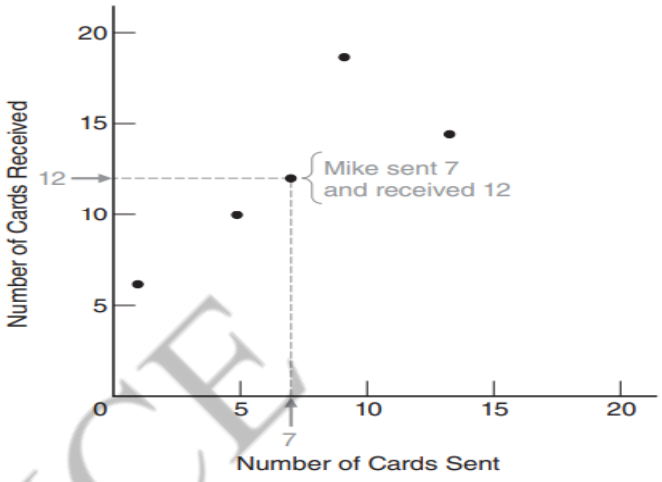


FIGURE 6.1
Scatterplot for greeting card exchange.

3.2.1. Types of Correlation

- The scatter plot explains the correlation between the two attributes or variables.
- It represents how closely the two variables are connected.
- There can be three such situations to see the relation between the two variables

1. **Positive Correlation**
    - When the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable
2. **Negative Correlation**
    - When the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.

3. **No Correlation**

   - When there is no linear dependence or no relation between the two variables.

Positive, Negative, or Little or No Relationship?

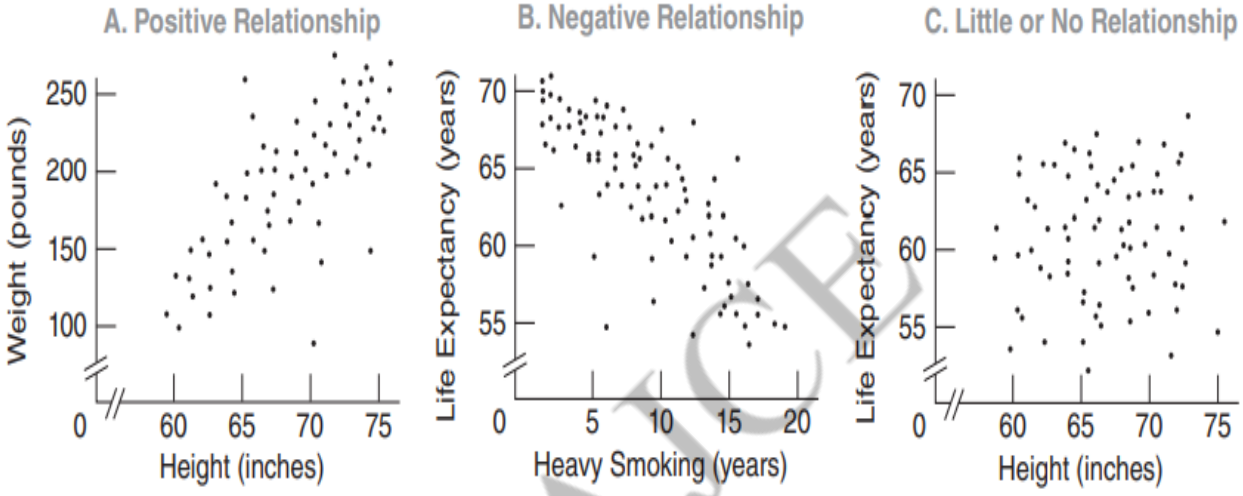
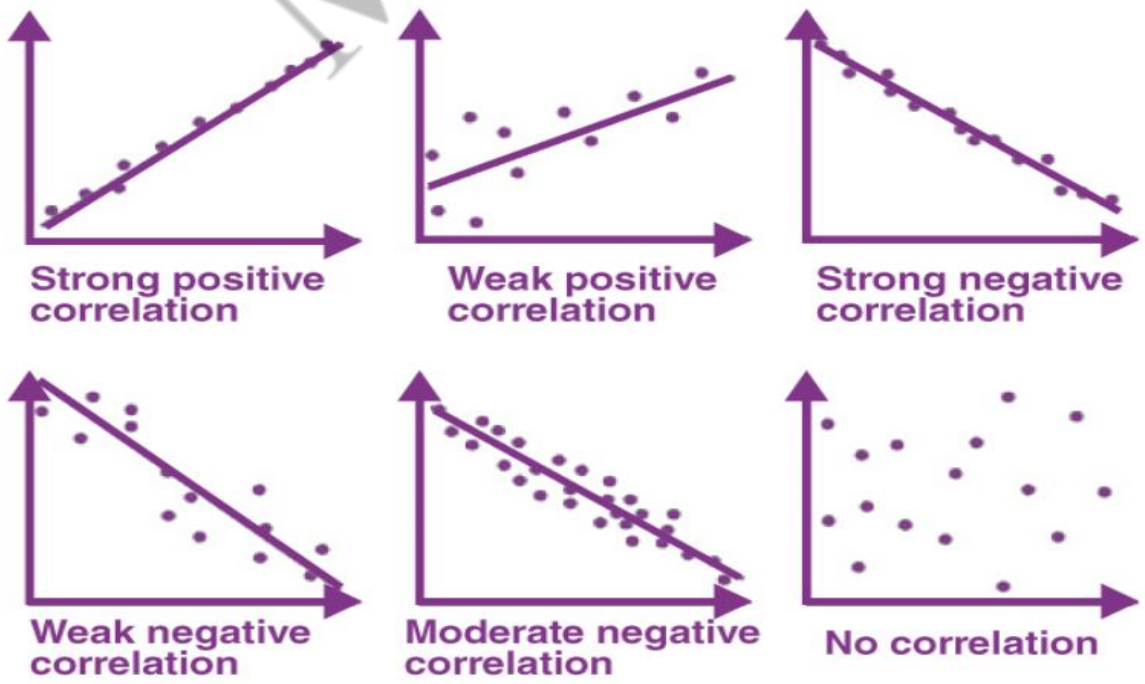


FIGURE 6.2

Three types of relationships.



Linear Relationship

- A relationship that can be described best with a straight line

Curvilinear Relationship

- A relationship that can be described best with a curved line

Example:

Age vs. Physical Strength

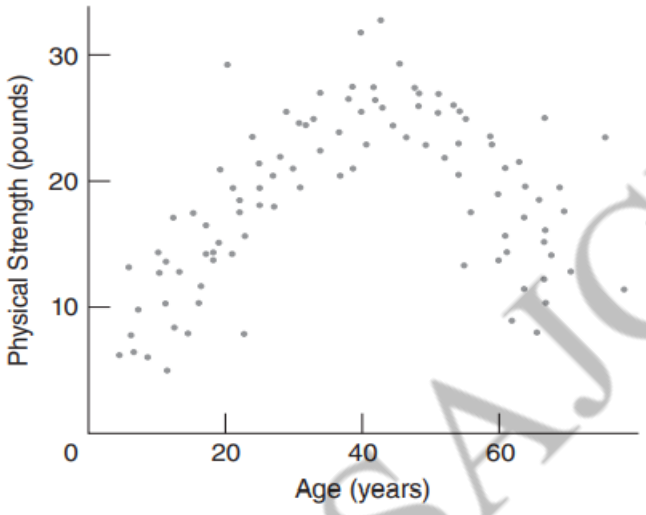


FIGURE 6.4
Curvilinear relationship.

Exercises

Progress Check *6.1 Indicate whether the following statements suggest a positive or negative relationship:

(a) More densely populated areas have higher crime rates.

(b) Schoolchildren who often watch TV perform more poorly on academic achievement tests.

(c) Heavier automobiles yield poorer gas mileage.

(d) Better-educated people have higher incomes.

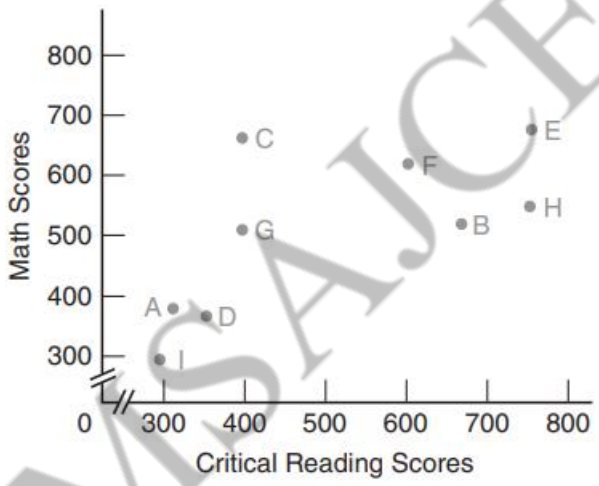
(e) More anxious people voluntarily spend more time performing a simple repetitive task.

Answer

(a) Positive. The crime rate is higher, square mile by square mile, in densely populated cities than in sparsely populated rural areas.
(b) Negative. As TV viewing increases, performance on academic achievement tests tends to decline.
(c) Negative. Increases in car weight are accompanied by decreases in miles per gallon.

(d) Positive. Increases in educational level—grade school, high school, college—tend to be associated with increases in income.
(e) Positive. Highly anxious people willingly spend more time performing a simple repetitive task than do less anxious people.

Progress Check *6.2 Critical reading and math scores on the SAT test for students A, B, C, D, E, F, G, and H are shown in the following scatterplot:



(a) Which student(s) scored about the same on both tests?

(b) Which student(s) scored higher on the critical reading test than on the math test?

(c) Which student(s) will be eligible for an honors program that requires minimum scores of 700 in critical reading and 500 in math?

(d) Is there a negative relationship between the critical reading and math scores?

Answer

6.2 **(a)** I, D, F **(c)** E, H
(b) B, H, E **(d)** No. The relationship is positive.

3.1.3. Correlation Coefficient for Quantitative Data: r

Correlation coefficient (r)

- Correlation Coefficient A number between –1 and 1 that describes the relationship between pairs of variables
- A correlation coefficient is a way to put a value to the relationship.
- Correlation coefficients have a value of between -1 and 1. A "0" means there is **no relationship** between the variables at all, while -1 or 1 means that there is a **perfect negative or positive correlation**
- A correlation coefficient quite close to 0, but either positive or negative implies little or no relationship between the two variables.
- A correlation coefficient close to plus 1 means a positive relationship between the two variables, with increases in one of the variables being associated with increases in the other variable.
- A correlation coefficient close to -1 indicates a negative relationship between two variables, with an increase in one of the variables being associated with a decrease in the other variable.

Pearson Correlation Coefficient (r)

- A number between –1.00 and +1.00 that describes the linear relationship between pairs of quantitative variables.

Key Properties of r

Named in honor of the British scientist Karl Pearson, the Pearson correlation coefficient, r, can equal any value between –1.00 and +1.00.

Furthermore, the following two properties apply:

1. The sign of r indicates the type of linear relationship, whether positive or negative.

2. The numerical value of r, without regard to sign, indicates the strength of the linear relationship.

Sign of r

- A number with a plus sign (or no sign) indicates a positive relationship, and a number with a minus sign indicates a negative relationship

Numerical Value of r

- The more closely a value of r approaches either –1.00 or +1.00, the stronger (more regular) the relationship.
- Conversely, the more closely the value of r approaches 0, the weaker (less regular) the relationship.
- For example, an r of –.90 indicates a stronger relationship than does an r of –.70, and an r of –.70 indicates a stronger relationship than does an r of .50.

Interpretation of r

Located along a scale from –1.00 to +1.00, the value of r supplies information about the direction of a linear relationship whether positive or negative and, generally, information about the relative strength of a linear relationship whether relatively weak because r is in the area of 0, or relatively strong because r deviates from 0 in the direction of either +1.00 or –1.00.

r is Independent of Units of Measurement

- The value of r is independent of the original units of measurement.
- In fact, the same value of r describes the correlation between height and weight for a group of adults, regardless of whether height is measured in inches or centimeters or whether weight is measured in pounds or grams.

Verbal Descriptions

- When interpreting an r value, you'll find it helpful to translate the numerical value of r into a verbal description of the relationship.
- An r of .70 for the height and weight of college students could be translated into "Taller students tend to weigh more" (or some other equally valid statement, such as "Lighter students tend to be shorter");
- An r of –.42 for time spent taking an exam and score could be translated into "Students who take less time tend to make higher scores"; and an r in the neighborhood of 0 for shoe size and IQ could be translated into "Little, if any, relationship exists between shoe size and IQ."

Progress Check *6.3 Supply a verbal description for each of the following correlations. (If necessary, visualize a rough scatterplot for *r*, using the scatterplots in Figure 6.3 as a frame of reference.)

(a) an *r* of –.84 between total mileage and automobile resale value

(b) an *r* of –.35 between the number of days absent from school and performance on a math achievement test

(c) an *r* of .03 between anxiety level and college GPA

(d) an *r* of .56 between age of schoolchildren and reading comprehension

Answer

6.3 **(a)** Cars with more total miles tend to have lower resale values.
(b) Students with more absences from school tend to score lower on math achievement tests.
(c) Little or no relationship between anxiety level and college GPA.
(d) Older schoolchildren tend to have better reading comprehension.

Correlation Not Necessarily Cause-Effect

- A correlation coefficient, regardless of size, never provides information about whether an observed relationship reflects a simple cause-effect relationship or some more complex state of affairs.

Example:

Speculate on whether the following correlations reflect simple cause-effect relationships or more complex states of affairs.

(a) Caloric intake and body weight

(b) Height and weight

(c) SAT math score and score on a calculus test

(d) Poverty and crime

Answers:

(a) Simple cause-effect

(b) Complex

(c) Complex

(d) Complex

3.1.4. Computation Formula for r

Calculate a value for *r* by using the following computation formula:

CORRELATION COEFFICIENT (COMPUTATION FORMULA)

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} \tag{6.1}$$

where the two sum of squares terms in the denominator are defined as

$$SS_x = \sum \left(X - \bar{X}\right)^2 = \sum X^2 - \frac{\left(\sum X\right)^2}{n}$$

$$SS_y = \sum \left(Y - \bar{Y}\right)^2 = \sum Y^2 - \frac{\left(\sum Y\right)^2}{n}$$

and the sum of the products term in the numerator, SP_{xy}, is defined in Formula 6.2.

SUM OF PRODUCTS (DEFINITION AND COMPUTATION FORMULAS)

$$SP_{xy} = \sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right) = \sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n} \tag{6.2}$$

Where n = Quantity of Information

Σx = Total of the First Variable Value

Σy = Total of the Second Variable Value

Σxy = Sum of the Product of first & Second Value

Σx^2 = Sum of the Squares of the First Value

Σy^2 = Sum of the Squares of the Second Value

Table 6.3
CALCULATION OF *r*: COMPUTATION FORMULA

A. COMPUTATIONAL SEQUENCE

Assign a value to n (1), representing the number of pairs of scores.
Sum all scores for X (2) and for Y (3).
Find the product of each pair of X and Y scores (4), one at a time, then add all of these products (5).
Square each X score (6), one at a time, then add all squared X scores (7).
Square each Y score (8), one at a time, then add all squared Y scores (9).
Substitute numbers into formulas (10) and solve for SP_{xy}, SS_x, and SS_y.
Substitute into formula (11) and solve for r.

B. DATA AND COMPUTATIONS

	CARDS		4	6	8
FRIEND	**SENT, *X***	**RECEIVED, *Y***	***XY***	***X*²**	***Y*²**
Doris	13	14	182	169	196
Steve	9	18	162	81	324
Mike	7	12	84	49	144
Andrea	5	10	50	25	100
John	1	6	6	1	36

1 $n = 5$ 2 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

10 $$SP_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

11 $$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$$

Here r = 0.80, so there is a positive relation between x and y

Exercises:

1. Estimate whether the following pairs of scores for X and Y reflect a positive relationship, a negative relationship, or no relationship.

X	Y
64	66
40	79
30	98
71	65
55	76
31	83
61	68
42	80
57	72

(a) Construct a scatterplot for X and Y. Verify that the scatterplot does not describe a pronounced curvilinear trend.

(b) Calculate r using the computation formula

2. Calculate the value of r using the computational formula for the following data

X	Y
2	8
4	6
5	2
3	3
1	4
7	1
2	4

3. On the basis of an extensive survey, the California Department of Education reported an r of –.32 for the relationship between the amounts of time spent watching TV and the achievement test scores of schoolchildren. Each of the following statements represents a possible interpretation of this finding. Indicate whether each is true or false

(a) Every child who watches a lot of TV will perform poorly on the achievement tests.

(b) Extensive TV viewing causes a decline in test scores.

(c) Children who watch little TV will tend to perform well on the tests.

(d) Children who perform well on the tests will tend to watch little TV.

(e) TV viewing could not possibly cause a decline in test scores.

Answer:

(a) False. This statement would be true only if a perfect negative relationship (−1.00) described the relationship between TV viewing time and test scores.

(b) False. Correlation does not necessarily signify cause-effect.

(c) True

(d) True

(e) False. Although correlation does not necessarily signify cause-effect, it opens the possibility of cause-effect.

3.2. Regression

- Regression is a statistical method used to determine the strength and character of the relationship between one dependent variable and a series of independent variables
- It is also called simple linear regression or Ordinary Least Squares (OLS)
- Linear regression establishes the linear relationship between two variables based on a line of best fit
- Regression captures the correlation between variables observed in a data set, and computes whether those correlations are statistically significant or not.
- The two basic types of regression
    - Simple linear regression
    - Multiple linear regression
- Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable while multiple linear regression uses two or more independent variables to predict the outcome.

Understanding Regression

- If two variables are correlated, description can lead to prediction.
- For example, if computer skills and GPAs are related, level of computer skills can be used to predict GPAs.
- Regression can also help predict sales for a company based on weather, previous sales, GDP growth, or other types of conditions.
- Regression can help finance and investment professionals as well as professionals in other businesses

Example Scenario for Regression Analysis

A correlation analysis of the exchange of greeting cards by five friends for the most recent holiday season suggests a strong positive relationship between cards sent and cards received. When informed of these results, another friend, Emma, who enjoys receiving greeting cards, asks you to predict how many cards she will receive during the next holiday season, assuming that she plans to send 11 cards

Two Rough Predictions

Predict "Relatively Large Number"

Predict "between 14 and 18 Cards"

Table 6.1
GREETING CARDS SENT AND RECEIVED BY FIVE FRIENDS

	NUMBER OF CARDS	
FRIEND	SENT	RECEIVED
Andrea	5	10
Mike	7	12
Doris	13	14
Steve	9	18
John	1	6

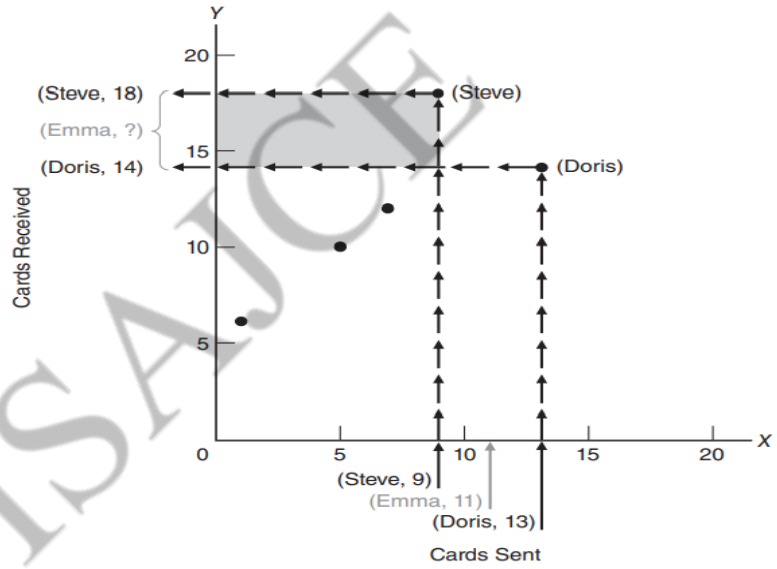


FIGURE 7.1
A rough prediction for Emma (using dots for Steve and Doris).

- To obtain a slightly more precise prediction for Emma, refer to the scatter plot for the original five friends.
- Notice that Emma's plan to send 11 cards locates her along the X axis between the 9 cards sent by Steve and the 13 sent by Doris.
- Using the dots for Steve and Doris as guides, construct two strings of arrows, one beginning at 9 and ending at 18 for Steve and the other beginning at 13 and ending at 14 for Doris.
- Focusing on the interval along the Y axis between the two strings of arrows, you could predict that Emma's return should be between 14 and 18 cards, the numbers received by Doris and Steve.

3.2.1. Regression line

- A regression line indicates a linear relationship between the dependent variables on the y-axis and the independent variables on the x-axis.
- The regression line is plotted closest to the data points in a regression graph.
- This statistical tool helps analyze the behavior of a dependent variable y when there is a change in the independent variable x by substituting different values of x in the regression equation.

- The regression line is a straight line rather than a curved line because of the linear relationship between cards sent and cards received.
- All five dots contribute to the more precise prediction, illustrated in Figure 7.2, that Emma will receive 15.20 cards.
- Look more closely at the solid line designated as the regression line in Figure 7.2, which guides the string of arrows, beginning at 11, toward the predicted value of 15.20
- Regardless of whether Emma decides to send 5, 15, or 25 cards, it will guide a new string of arrows, beginning at 5 or 15 or 25, toward a new predicted value along the Y axis.

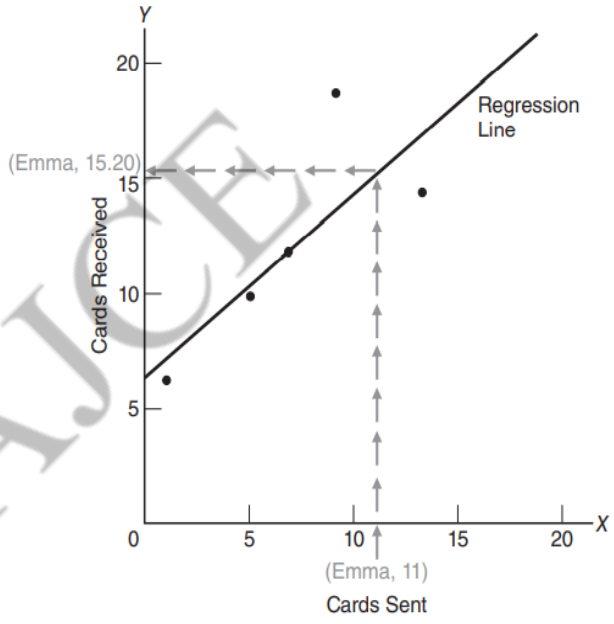


FIGURE 7.2
Prediction of 15.20 for Emma (using the regression line).

Placement of Line

- If all five dots had defined a single straight line, placement of the regression line would have been simple; simply let it pass through all dots.
- When the dots fail to define a single straight line, as in the scatterplot for the five friends, placement of the regression line represents a compromise.
- It passes through the main cluster, possibly touching some dots but missing others.

Predictive Errors

- Figure 7.3 illustrates the predictive errors that would have occurred if the regression line had been used to predict the number of cards received by the five friends.
- Solid dots reflect the actual number of cards received, and open dots, always located along the regression line, reflect the predicted number of cards received.
- The largest predictive error, shown as a broken vertical line, occurs for Steve, who sent 9 cards. Although he actually received 18 cards, he should have received slightly fewer than 14 cards, according to the regression line.

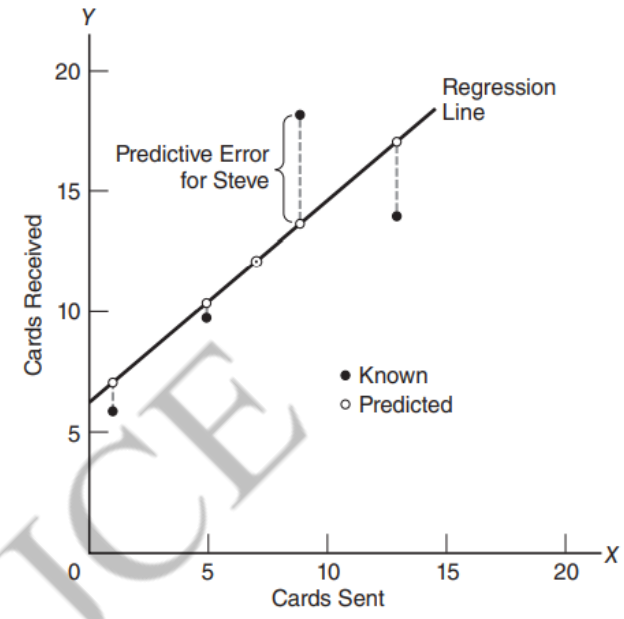
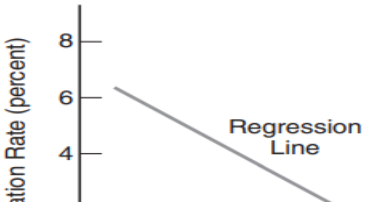


FIGURE 7.3
Predictive errors.

- The smallest predictive error—none whatsoever—occurs for Mike, who sent 7 cards. He actually received the 12 cards that he should have received, according to the regression line.

Progress Check *7.1 To check your understanding of the first part of this chapter, make predictions using the following graph.



(a) Predict the approximate rate of inflation, given an unemployment rate of 5 percent.

(b) Predict the approximate rate of inflation, given an unemployment rate of 15 percent.

Answer

(a) Approximately 5–6 percent

(b) Approximately 2–3 percent

3.2.2. Least Squares Regression Line

- The Least Squares Regression Line is the line that makes the vertical distance from the data points to the regression line as small as possible. It's called a "least squares" because the best line of fit is one that minimizes the variance
- The equation that minimizes the total of all squared prediction errors for known Y scores in the original correlation analysis
- To avoid the arithmetic standoff of zero always produced by adding positive and negative predictive errors (associated with errors above and below the regression line, respectively), the placement of the regression line minimizes not the total predictive error but the total squared predictive error, that is, the total for all squared predictive errors.
- When located in this fashion, the regression line is often referred to as the least squares regression line.

Least Squares Regression Equation

LEAST SQUARES REGRESSION EQUATION

$$Y' = bX + a \qquad (7.1)$$

- Y´ represents the predicted value (the predicted number of cards that will be received by any new friend, such as Emma);
- X represents the known value (the known number of cards sent by any new friend);
- b and a represent numbers calculated from the original correlation analysis

Finding Values of b and a

To obtain a working regression equation, solve each of the following expressions, first for b and then for a, using data from the original correlation analysis.

The expression for b reads:

SOLVING FOR *b*

$$b = r\sqrt{\frac{SS_y}{SS_x}} \qquad (7.2)$$

- Where r represents the correlation between X and Y (cards sent and received by the five friends);
- SSy represents the sum of squares for all Y scores (the cards received by the five friends);
- SSx represents the sum of squares for all X scores (the cards sent by the five friends).

The expression for a reads:

SOLVING FOR *a*

$$a = \bar{Y} - b\bar{X} \qquad (7.3)$$

- Where Y and X refer to the sample means for all Y and X scores, respectively, and b is defined by the preceding expression.
- The values of all terms in the expressions for b and a can be obtained from the original correlation analysis either directly, as with the value of r, or indirectly, as with the values of the remaining terms: SSy ′ SSx ′ Y, and X.

Table 6.3
CALCULATION OF *r*: COMPUTATION FORMULA

A. COMPUTATIONAL SEQUENCE

Assign a value to *n* (1), representing the number of pairs of scores.
Sum all scores for *X* (2) and for *Y* (3).
Find the product of each pair of *X* and *Y* scores (4), one at a time, then add all of these products (5).
Square each *X* score (6), one at a time, then add all squared *X* scores (7).
Square each *Y* score (8), one at a time, then add all squared *Y* scores (9).
Substitute numbers into formulas (10) and solve for SP_{xy}, SS_x, and SS_y.
Substitute into formula (11) and solve for *r*.

B. DATA AND COMPUTATIONS

	CARDS		4	6	8
FRIEND	**SENT, *X***	**RECEIVED, *Y***	***XY***	***X*²**	***Y*²**
Doris	13	14	182	169	196
Steve	9	18	162	81	324
Mike	7	12	84	49	144
Andrea	5	10	50	25	100
John	1	6	6	1	36

1 $n = 5$ 2 $\Sigma X = 35$ 3 $\Sigma Y = 60$ 5 $\Sigma XY = 484$ 7 $\Sigma X^2 = 325$ 9 $\Sigma Y^2 = 800$

10 $$SP_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 484 - \frac{(35)(60)}{5} = 484 - 420 = 64$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 325 - \frac{(35)^2}{5} = 325 - 245 = 80$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 800 - \frac{(60)^2}{5} = 800 - 720 = 80$$

11 $$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{64}{\sqrt{(80)(80)}} = \frac{64}{80} = .80$$

Table 7.1
DETERMINING THE LEAST SQUARES REGRESSION EQUATION

A. COMPUTATIONAL SEQUENCE

Determine values of SS_x, SS_y, and r (1) by referring to the original correlation analysis in Table 6.3.
Substitute numbers into the formula (2) and solve for b.
Assign values to $\overline{X}$ and $\overline{Y}$ (3) by referring to the original correlation analysis in Table 6.3.
Substitute numbers into the formula (4) and solve for a.
Substitute numbers for b and a in the least squares regression equation (5).

B. COMPUTATIONS

1 $SS_x = 80^*$
$SS_y = 80^*$
$r = .80$

2 $$b = r\sqrt{\frac{SS_Y}{SS_X}} = .80\sqrt{\frac{80}{80}} = .80$$

3 $\overline{X} = 7^{**}$
$\overline{Y} = 12^{**}$

4 $$a = \overline{Y} - (b)(\overline{X}) = 12 - (.80)(7) = 12 - 5.60 = 6.40$$

5 $$Y' = (b)(X) + a$$
$$= (.80)(X) + 6.40$$

- Once numbers have been assigned to b and a, as just described, the least squares regression equation emerges as a working equation with a most desirable property: It automatically minimizes the total of all squared predictive errors for known Y scores in the original correlation analysis
- Solving for Y′ In its present form, the regression equation can be used to predict the number of cards that Emma will receive, assuming that she plans to send 11 cards.

  Simply substitute 11 for X and solve for the value of Y′ as follows:

$$\begin{aligned} Y' &= .80(11) + 6.40 \\ &= 8.80 + 6.40 \\ &= 15.20 \end{aligned}$$

X sent cards	Y received cards	Y' (Predicted cards)
13	14	16.8
9	18	13.6
7	12	12
5	10	10.4
1	6	7.2

Calculate Y'

Where X= 13

Y' = .80(13) + 6.40

Y' = 16.8

Where X = 9

Y' = .80(9) + 6.40

Y' = 13.6

Table 7.2 lists the predicted card returns for a number of different card investments.

Table 7.2
PREDICTED CARD RETURNS (Y') FOR DIFFERENT CARD INVESTMENTS (X)

X	Y'
0	6.40
4	9.60
8	12.80
10	14.40
12	16.00
20	22.40
30	30.40

Progress Check *7.2 Assume that an r of .30 describes the relationship between educational level (highest grade completed) and estimated number of hours spent reading each week. More specifically:

EDUCATIONAL LEVEL (X)	WEEKLY READING TIME (Y)
$\overline{X} = 13$	$\overline{Y} = 8$
$SS_x = 25$	$SS_y = 50$
$r = .30$	

(a) Determine the least squares equation for predicting weekly reading time from educational level.

(b) Faith's education level is 15. What is her predicted reading time?

(c) Keegan's educational level is 11. What is his predicted reading time?

7.2 **(a)** $b = \sqrt{\frac{50}{25}}(.30) = .42;\ a = 8 - (.42)(13) = 2.54$

(b) $Y' = (.42)(15) + 2.54 = 8.84$

(c) $Y' = (.42)(11) + 2.54 = 7.16$

3.2.3. Standard Error of Estimate, s y | x

- The standard error of the estimate is a measure of the accuracy of predictions.
- Simply, it is used to check the accuracy of predictions made with the regression line.
- A rough measure of the average amount of predictive error
- A standard deviation which measures the variation in the set of data from its mean, the standard error of estimate also measures the variation in the actual values of Y from the computed values of Y (predicted) on the regression line. It is computed as a standard deviation, and here the deviations are the vertical distance of every dot from the line of average relationship.
- Although designed to minimize predictive error, the least squares equation does not eliminate it. Therefore, our next task is to estimate the amount of error associated with our predictions. The smaller the estimated error is, the better the prognosis will be for our predictions

STANDARD ERROR OF ESTIMATE (COMPUTATION FORMULA)

$$s_{y|x} = \sqrt{\frac{SS_y\left(1-r^2\right)}{n-2}} \qquad (7.5)$$

The standard error of estimate represents a special kind of standard deviation that reflects the magnitude of predictive error

The value of 3.10 for sy|x , as calculated in Table 7.3, represents the standard deviation for the discrepancies between known and predicted card returns originally shown in Figure 7.3. In its role as an estimate of predictive error, the value of sy|x can be attached to any new prediction.

Table 7.3
CALCULATION OF THE STANDARD ERROR OF ESTIMATE, $s_{y|x}$

A. COMPUTATIONAL SEQUENCE

Assign values to SS_y and r (1) by referring to previous work with the least squares regression equation in Table 7.1.
Substitute numbers into the formula (2) and solve for $s_{y|x}$.

B. COMPUTATIONS

1 $SS_y = 80$

$r = .80$

2 $s_{y|x} = \sqrt{\frac{SS_y(1-r^2)}{n-2}} = \sqrt{\frac{80\left(1-[.80]^2\right)}{5-2}} = \sqrt{\frac{80(.36)}{3}} = \sqrt{\frac{28.80}{3}} = \sqrt{9.60}$

$= 3.10$

Thus, a concise prediction statement may read: "The predicted card return for Emma equals 15.20 ± 3.10," in which the latter term serves as a rough estimate of the average amount of predictive error, that is, the average amount by which 15.20 will either overestimate or underestimate Emma's true card return.

3.2.4. Interpretation of r^2

- R-Squared (R^2 or the coefficient of determination)
- R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable.
- In other words, r-squared shows how well the data fit the regression model
- R-squared can take any values between 0 to 1.

- The most common interpretation of r-squared is how well the regression model explains observed data.
- For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model. Generally, a higher r-squared indicates more variability is explained by the model.

Expressing the equation for r^2 in symbols, we have:

r^2 INTERPRETATION

$$r^2 = \frac{SS_{Y'}}{SS_Y} = \frac{SS_Y - SS_{Y|X}}{SS_Y} \quad (7.6)$$

SS_y = 80 (refer the previous calculation)

$$SS_{y|x} = \sum(Y - Y')^2$$

X sent cards	Y received cards	Y' (Predicted cards)
13	14	16.8
9	18	13.6
7	12	12
5	10	10.4
1	6	7.2

$= [(14-16.8)^2 + (18-13.6)^2 + (12-12)^2 + (10-10.4)^2 + (6-7.2)^2]$

$= (-2.8)^2 + (4.4)^2 + (0)^2 + (-0.4)^2 + (-1.2)^2$

$=7.84 + 19.36 + 0 + 0.16 + 1.44$

$= 28.8$

$$r^2 = \frac{80 - 28.8}{80} = 0.64$$

An r-squared of 64% reveals that 64% of the variability observed in the target variable is explained by the regression model.

Exercises 1:

X(observations)	10	20	30	40	50
Y(responses)	12	8	14	11	7

a. **Construct a scatter plot**
b. **Find the correlation coefficient**
c. **Determine the suitable regression line and find least square equation**

b. Find correlation coefficient (r)

X	Y	XY	X^2	Y^2
10	12	120	100	144
20	8	160	400	64
30	14	420	900	196
40	11	440	1600	121
50	7	350	2500	49
$\Sigma X =$ 150	$\Sigma Y =$ 52	$\Sigma XY =$ 1490	$\Sigma X^2 =$ 5500	$\Sigma Y^2 =$ 574

n=5

$$SP_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n} = 1490 - \frac{(150)(52)}{5} = 1490 - 1560 = -70$$

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 5500 - \frac{(150)^2}{5} = 5500 - 4500 = 1000$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 574 - \frac{(52)^2}{5} = 574 - 540.8 = 33.2$$

$$r = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{-70}{\sqrt{(1000)(33.2)}} = \frac{-70}{182.20} = -0.38$$

r = - 0.38 so there is a strong or moderate negative relation between variables X and Y

c. Find least square regression equation

LEAST SQUARES REGRESSION EQUATION

$$Y' = bX + a \qquad (7.1)$$

SOLVING FOR *b*

$$b = r\sqrt{\frac{SS_y}{SS_x}} \qquad (7.2)$$

SOLVING FOR *a*

$$a = \bar{Y} - b\bar{X} \qquad (7.3)$$

Find b

$$b = -0.38\sqrt{\frac{33.2}{1000}} = (-0.38)\ (0.182) = -0.07$$

Find a

$$\bar{X} = \frac{10+20+30+40+50}{5} = 30 \qquad \bar{Y} = \frac{12+8+14+11+7}{5} = 10.4$$

a = 10.4 - (-0.07) (30) = 12.5

LEAST SQUARES REGRESSION EQUATION

$$Y' = bX + a \qquad (7.1)$$

a = 12.5

b = -0.07

Y' = -0.07(X) + 12.5

Here

X= 12 then Y' = 11.66

X=27 then Y' = 10.61

X(Independent Variable)	Y(Dependent Variable) Actual Value	Y' Predicted Value
10	12	11.8
20	8	11.1
30	14	10.4
40	11	9.7
50	7	9

Find Standard error of estimate

STANDARD ERROR OF ESTIMATE (COMPUTATION FORMULA)

$$s_{y|x} = \sqrt{\frac{SS_y(1-r^2)}{n-2}}$$

Here

SS_y = 33.2

r= -0.38

n=5

$$S_{y/x} = \sqrt{\frac{33.2(1-0.38^2)}{5-2}} = 3.07$$

The value of 3.07 for sy|x , as calculated represents the standard deviation for the discrepancies between actual and predicted value for Y'

The predicted Y' value may be Y'± 3.07

Interpretation of r^2

Expressing the equation for r^2 in symbols, we have:

r^2 INTERPRETATION

$$r^2 = \frac{SS_{Y'}}{SS_Y} = \frac{SS_Y - SS_{Y|X}}{SS_Y} \quad (7.6)$$

SS_y = 33.2

$$SS_{y|x} = \sum(Y - Y')^2$$

X(Independent Variable)	Y(Dependent Variable) Actual Value	Y' Predicted Value
10	12	11.8
20	8	11.1
30	14	10.4
40	11	9.7
50	7	9

$= [(12-11.8)^2 + (8-11.1)^2 + (14-10.4)^2 + (11-9.7)^2 + (7-9)^2]$

$= (0.2)^2 + (-3.1)^2 + (3.6)^2 + (1.3)^2 + (-2)^2$

$= 0.04 + 9.61 + 12.96 + 1.69 + 4$

$=28.3$

$$r^2 = \frac{33.2 - 28.3}{33.2} = 0.15$$

An r-squared of 15% reveals that 15% of the variability observed in the target variable is explained by the regression model

Exercises 2:

Assume that an r of $-.80$ describes the strong negative relationship between years of heavy smoking (X) and life expectancy (Y). Assume, furthermore, that the distributions of heavy smoking and life expectancy each have the following means and sums of squares:

$$\bar{X} = 5 \qquad \bar{Y} = 60$$
$$SS_x = 35 \qquad SS_y = 70$$

(a) Determine the least squares regression equation for predicting life expectancy from years of heavy smoking.

(b) Determine the standard error of estimate, $s_{y|x}$, assuming that the correlation of $-.80$ was based on $n = 50$ pairs of observations.

(c) Supply a rough interpretation of $s_{y|x}$.

(d) Predict the life expectancy for John , who has smoked heavily for 8 years.

(e) Predict the life expectancy for Katie, who has never smoked heavily.

Exercises 3:

Each of the following pairs represents the number of licensed drivers (X) and the number of cars (Y) for seven houses in my neighborhood:

DRIVERS (X)	CARS (Y)
5	4
5	3
2	2
2	2
3	2
1	1
2	2

(a) Construct a scatterplot to verify a lack of pronounced curvilinearity.

(b) Determine the least squares equation for these data. (Remember, you will first have to calculate r, SS_y and SS_x)

(c) Determine the standard error of estimate, $s_{y|x}$, given that $n = 7$.

(d) Predict the number of cars for each of two new families with two and five drivers.

Exercises 4:

At a large bank, length of service is the best single predictor of employees' salaries. Can we conclude, therefore, that there is a cause-effect relationship between length of service and salary?

3.2.5. Multiple Regression Equations

Multiple regressions is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables.

The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

$$Y = a + b_1X_1 + b_2X_3 + \ldots + b_nX_n$$

Here Y is the dependent variable, and $X_1,\ldots,X_n$ are the *n* independent variables. In calculating the weights, a, $b_1,\ldots,b_n$, regression analysis ensures maximal prediction of the dependent variable from the set of independent variables. This is usually done by least squares estimation.