Liner Regression and Correlation

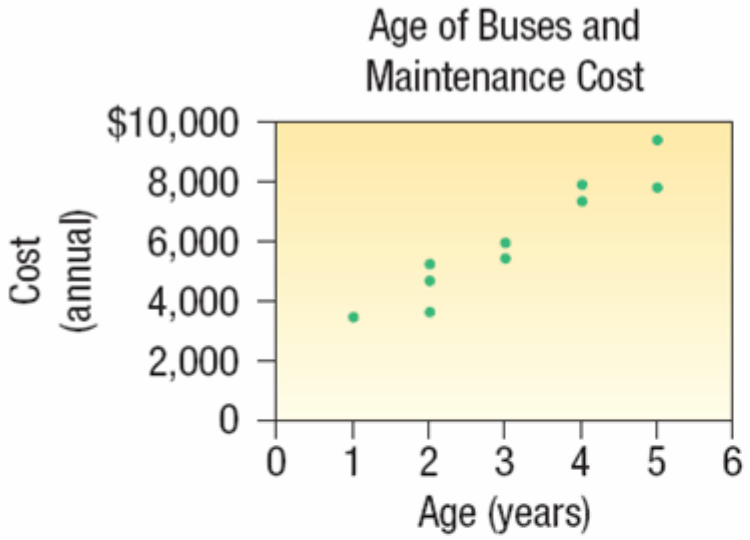
Chapter 13

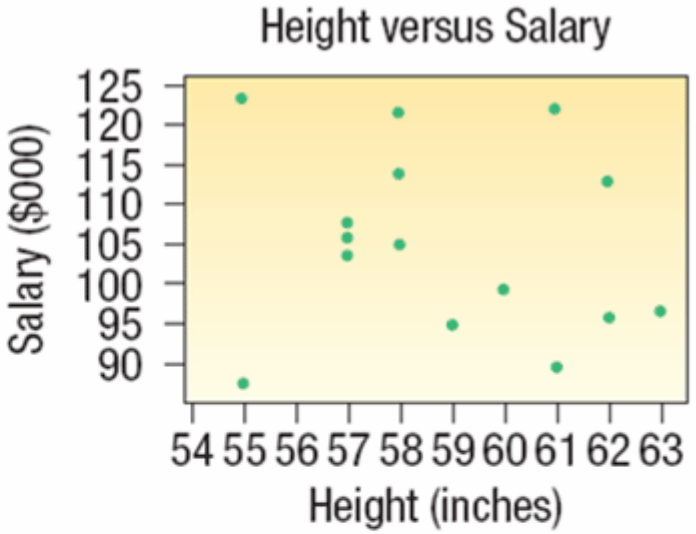
Learning Objectives

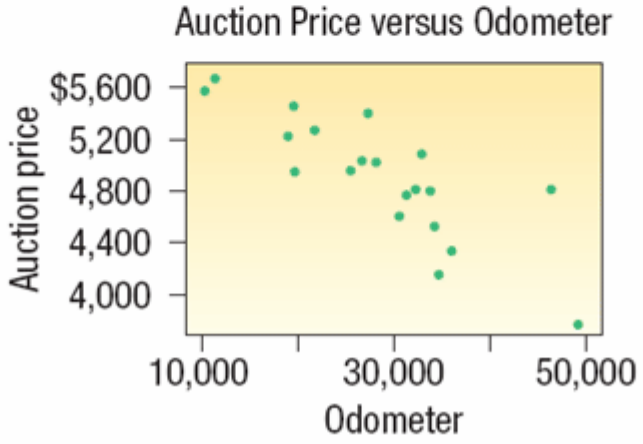
- Understand and interpret the terms dependent and independent variable.
- Calculate and interpret the coefficient of correlation and the coefficient of determination.
- Calculate the least squares regression line.

Introduction

- Recall in Chapter 4 the idea of showing the relationship between two variables with a scatter diagram was introduced.
- In that case we showed that, as the age of the bus increased the maintenance cost for the bus also increased.
- In this chapter we carry this idea further. Numerical measures to express the strength of the relationship between two variables are developed.
- In addition, an equation is used to express the relationship between variables, allowing us to estimate one variable on the basis of another.


Age of Buses and
Maintenance Cost
Cost
(annual)
$10,000
8,000
6,000
4,000
2,000
0
0 1 2 3 4 5 6
Age (years)


Height versus Salary
Salary ($000)
125
120
115
110
105
100
95
90
54 55 56 57 58 59 60 61 62 63
Height (inches)


Auction Price versus Odometer
Auction price
$5,600
5,200
4,800
4,400
4,000
10,000
30,000
50,000
Odometer

Uses

Some examples:

- Is there a relationship between the miles per gallon achieved by large pickup trucks and the size of the engine?
- Is there a relationship between the number of hours that students studied for an exam and the score earned?

Correlation Analysis

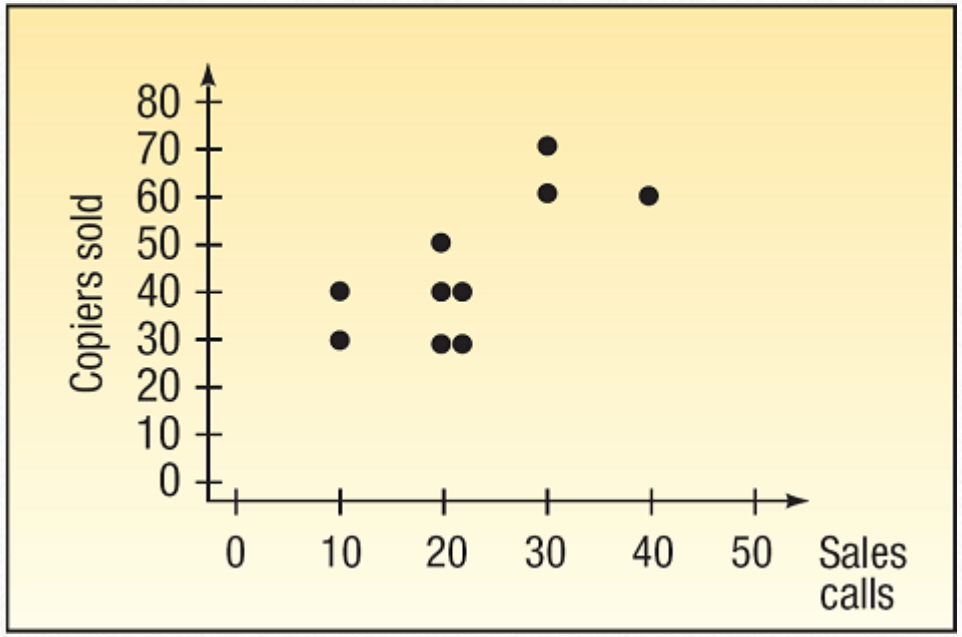
- Correlation Analysis is the study of the relationship between variables. It is also defined as group of techniques to measure the association between two variables.
- A Scatter Diagram is a chart that portrays the relationship between the two variables. It is the usual first step in correlations analysis.
- The Dependent Variable is the variable being predicted or estimated.
- The Independent Variable provides the basis for estimation. It is the predictor variable.

Example

- The sales manager of Copier Sales of America, which has a large sales force throughout the United States and Canada, wants to determine whether there is a relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 10 representatives and determines the number of sales calls each representative made last month and the number of copiers sold.

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Scatter Diagram

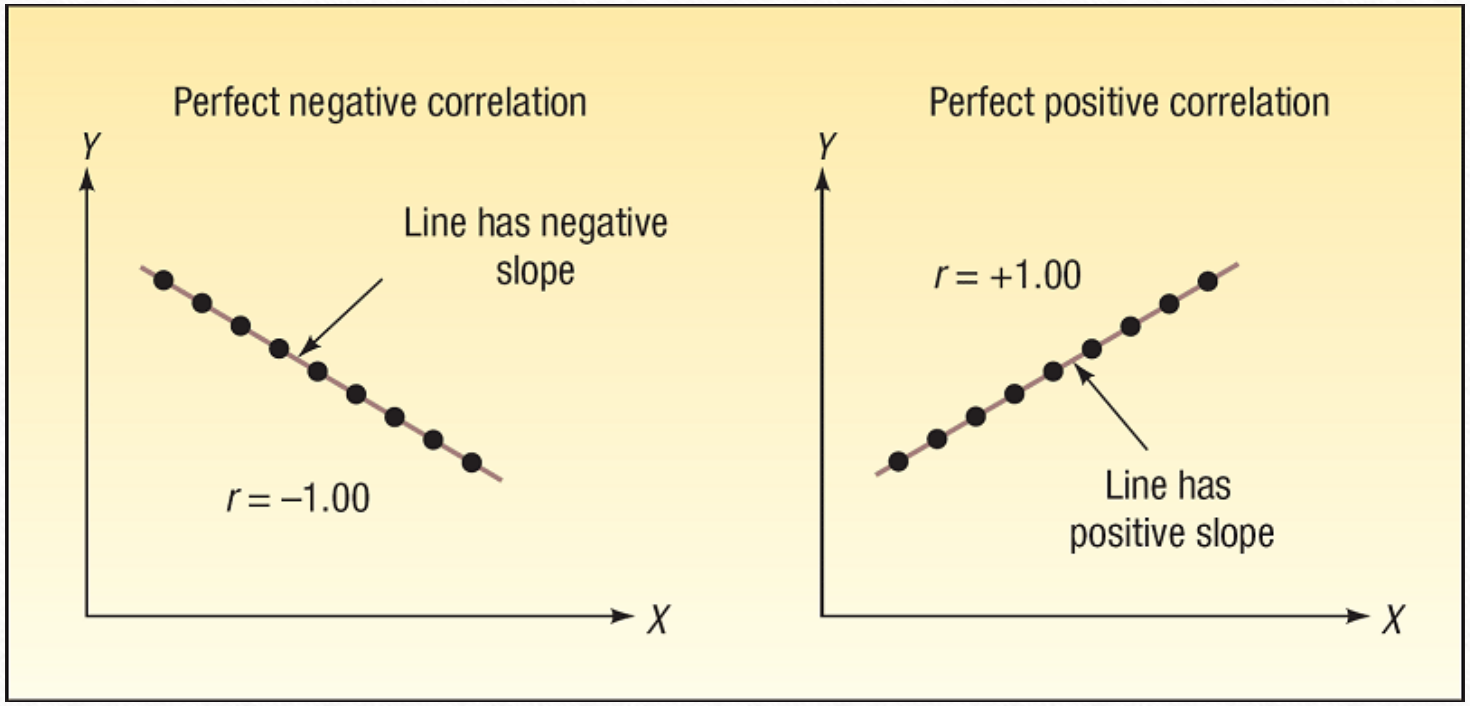


The Coefficient of Correlation, *r*

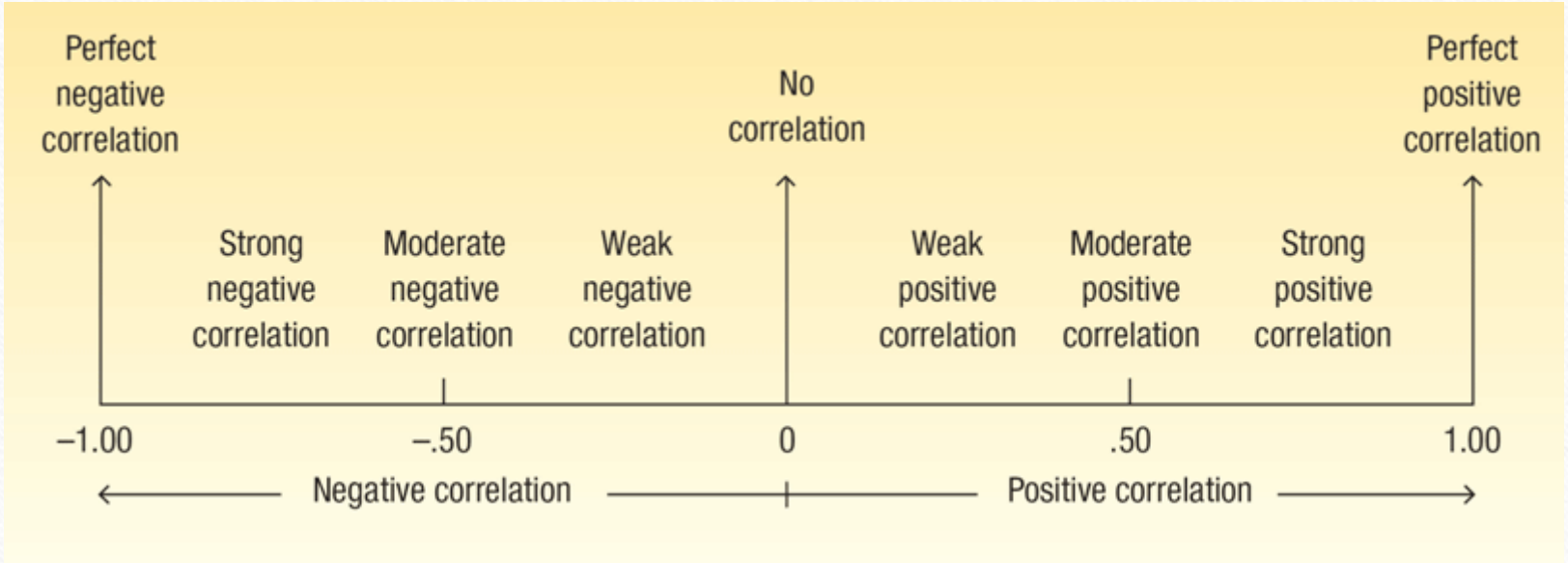
The Coefficient of Correlation (*r*) is a measure of the strength of the relationship between two variables. It requires interval or ratio-scaled data.

- It can range from -1.00 to 1.00.
- Values of -1.00 or 1.00 indicate perfect and strong correlation.
- Values close to 0.0 indicate weak correlation.
- Negative values indicate an inverse relationship and positive values indicate a direct relationship.

Perfect Correlation



Correlation Coefficient - Interpretation



Correlation Coefficient - Formulas

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n-1)S_X S_Y}$$

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

Example

- Using the Copier Sales of America data which a scatterplot was developed earlier, compute the correlation coefficient.

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Example

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{(n-1)s_x s_y} = \frac{900}{(10-1)(9.189)(14.337)} = 0.759$$

- How do we interpret a correlation of 0.759?
- First, it is positive, so we see there is a direct relationship between the number of sales calls and the number of copiers sold. The value of 0.759 is fairly close to 1.00, so we conclude that the association is strong.

- However, does this mean that more sales calls ***cause*** more sales?
- No, we have not demonstrated cause and effect here, only that the two variables—sales calls and copiers sold—are related.

Coefficient of Determination

The coefficient of determination (r^2) is the proportion of the total variation in the dependent variable (Y) that is explained or accounted for by the variation in the independent variable (X). It is the square of the coefficient of correlation.

- It ranges from 0 to 1.
- It does not give any information on the direction of the relationship between the variables.

Example

- Using the Copier Sales of America data which a scatterplot was developed earlier, compute the coefficient of determination.

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Example

- The coefficient of determination, r^2 ,is 0.576, found by $(0.759)^2$
- This is a proportion or a percent; we can say that 57.6 percent of the variation in the number of copiers sold is explained, or accounted for, by the variation in the number of sales calls.

Regression Analysis

In regression analysis we use the independent variable (*X*) to estimate the dependent variable (*Y*).

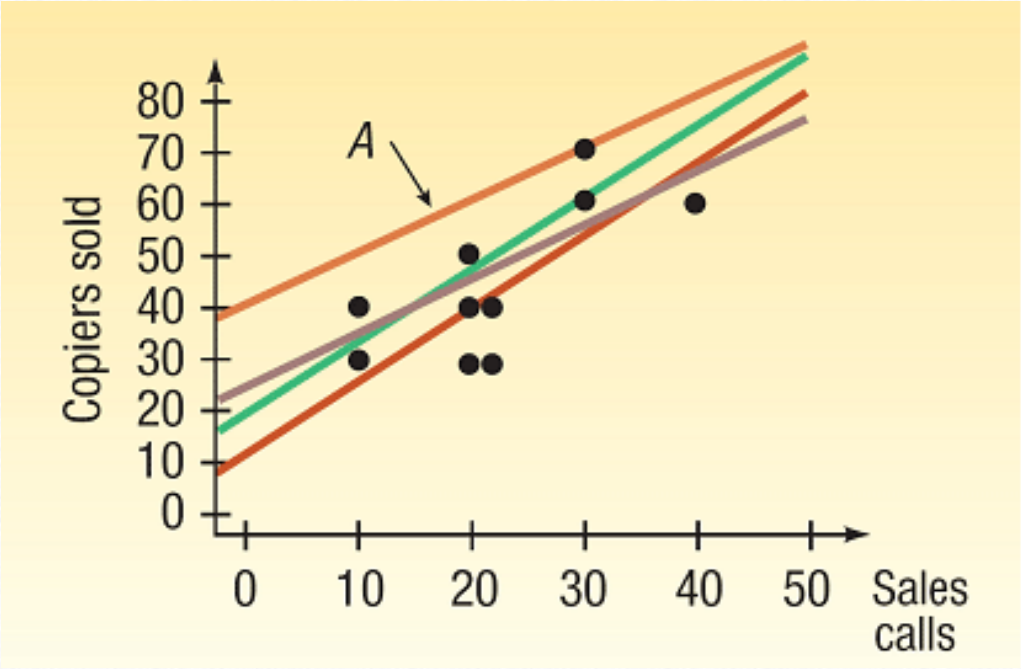
- The relationship between the variables is linear.
- Both variables must be at least interval scale.
- The least squares criterion is used to determine the equation.

Regression Analysis

REGRESSION EQUATION An equation that expresses the linear relationship between two variables.

LEAST SQUARES PRINCIPLE Determining a regression equation by minimizing the sum of the squares of the vertical distances between the actual *Y* values and the predicted values of *Y*.

Illustration of the Least Squares Regression Principle



Simple Linear Regression Model

GENERAL FORM OF LINEAR REGRESSION EQUATION $\hat{Y} = a + bX$

where

$\hat{Y}$ read Y hat, is the estimated value of the Y variable for a selected X value.

a is the Y-intercept. It is the estimated value of Y when $X = 0$. Another way to put it is: a is the estimated value of Y where the regression line crosses the Y-axis when X is zero.

b is the slope of the line, or the average change in $\hat{Y}$ for each change of one unit (either increase or decrease) in the independent variable X.

X is any value of the independent variable that is selected.

Computing the Slope of the Line and The Y- intercept

$$b = r\frac{S_y}{S_x}$$

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

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

Example

Recall the example involving Copier Sales of America. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 10 sales representatives. Use the least squares method to determine a linear equation to express the relationship between the two variables.

What is the expected number of copiers sold by a representative who made 20 calls?

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

Additional Example

- Bradford Electric Illuminating Company is studying the relationship between kilowatt-hours (thousands) used and the number of rooms in a private single-family residence. A random sample of 10 homes yielded the given table.

number of rooms	kilowatt-hours (thousands)
12	9
9	7
14	10
6	5
10	8
8	6
10	8
10	10
5	4
7	7