2.3 Correlation

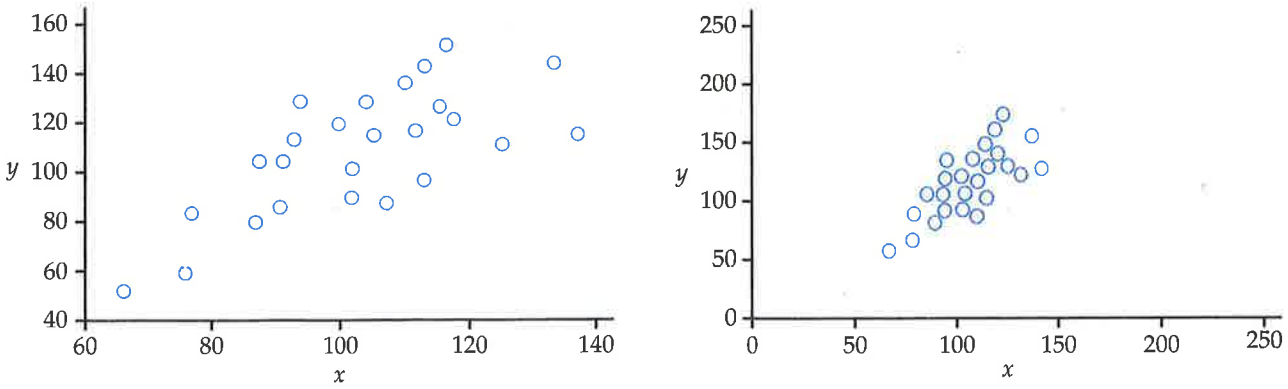


FIGURE 2.14 Two scatterplots of the same data. The linear pattern in the plot on the right appears stronger because of the surrounding space.

CORRELATION

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r.

Suppose that we have data on variables x and y for n individuals. The means and standard deviations of the two variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation r between x and y is

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

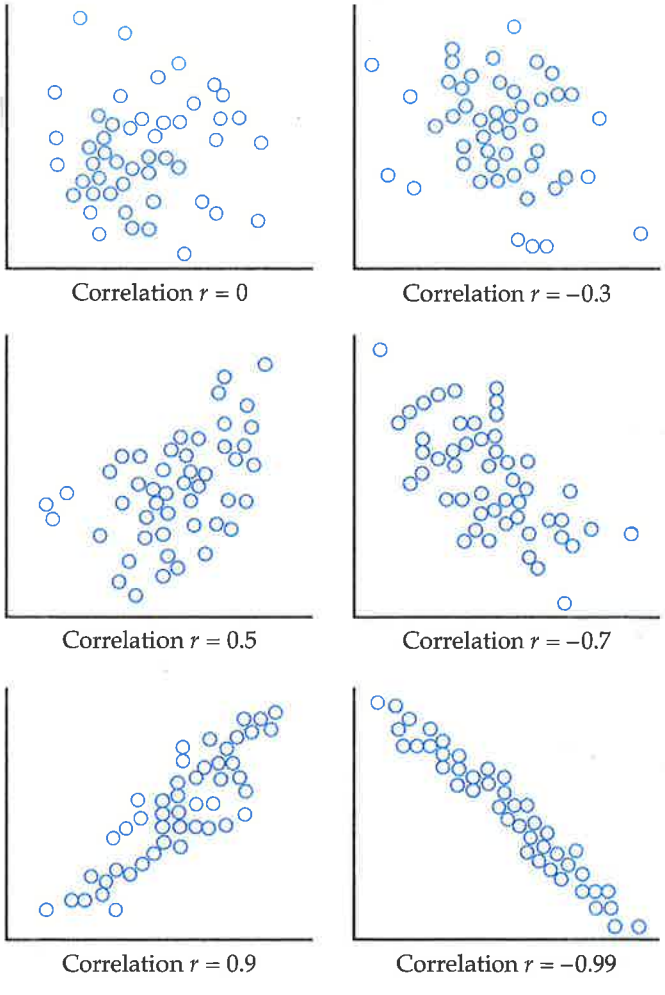


FIGURE 2.15 How the correlation r measures the direction and strength of a linear association.

RESPONSE VARIABLE, EXPLANATORY VARIABLE

A **response variable** measures an outcome of a study. An **explanatory variable** explains or causes changes in the response variable.

REGRESSION LINE

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to **predict** the value of y for a given value of x. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

FIGURE 2.17 A regression line fitted to the nonexercise activity data and used to predict fat gain for an NEA increase of 400 calories for Examples 2.19 and 2.20.

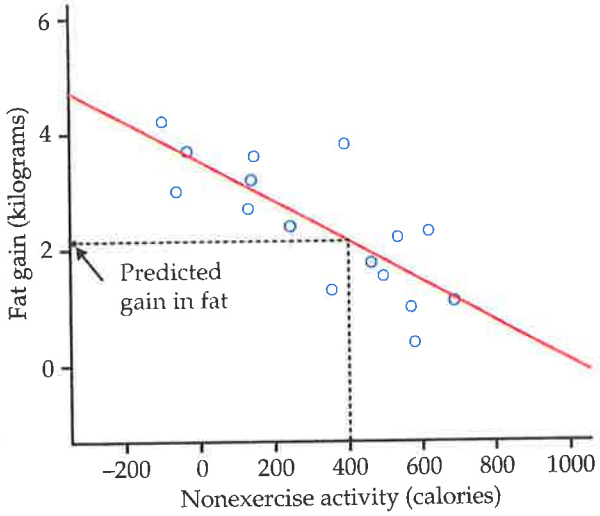
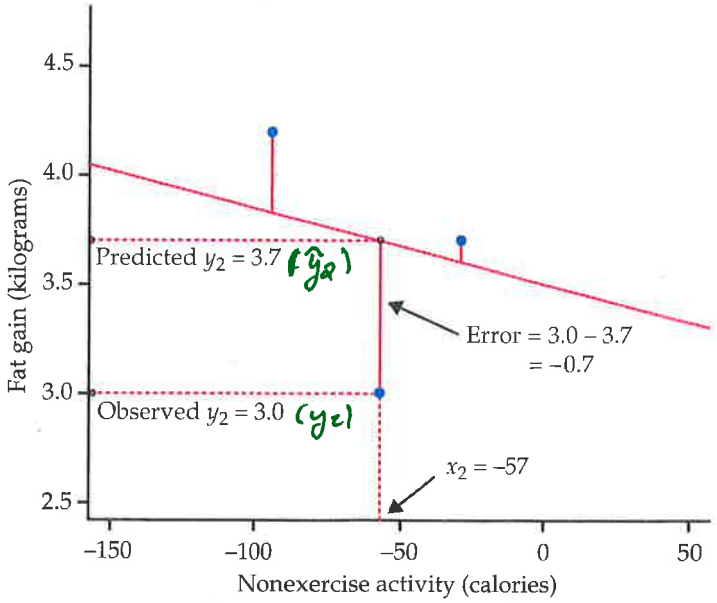


FIGURE 2.18 The least-squares idea: make the errors in predicting y as small as possible by minimizing the sum of their squares.



LEAST-SQUARES REGRESSION LINE

The **least-squares regression line of y on x** is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

STRAIGHT LINES

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form

$$y = b_0 + b_1 x$$

In this equation, b_1 is the **slope,** the amount by which y changes when x increases by one unit. The number b_0 is the **intercept,** the value of y when $x = 0$.

EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r. The **equation of the least-squares regression line** of y on x is

$$\hat{y} = b_0 + b_1 x$$

with **slope**

$$b_1 = r\frac{s_y}{s_x}$$

and **intercept**

$$b_0 = \bar{y} - b_1\bar{x}$$

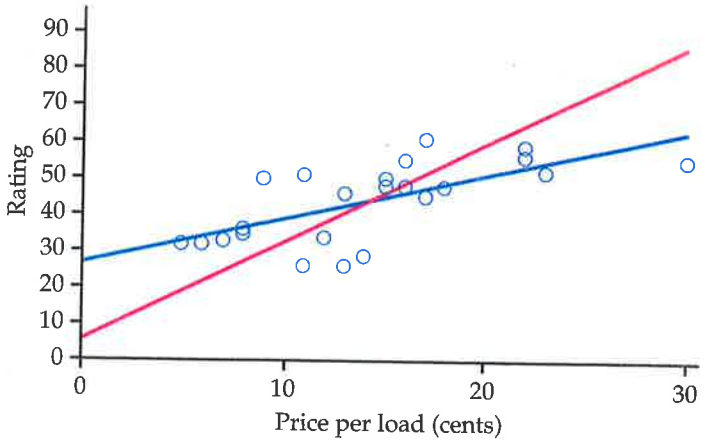


FIGURE 2.20 Scatterplot of price per load versus rating for 24 laundry detergents, from Example 2.8. The two lines are the two least-squares regression lines: using price per load to predict rating (red) and using rating to predict price per load (blue), for Example 2.24.

r^2 IN REGRESSION

The **square of the correlation, r^2,** is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.

$$r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y}$$

EXTRAPOLATION

Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x used to obtain the line. Such predictions are often not accurate and should be avoided.