Simple Linear Regression

- If X is the *independent variable* and Y the *dependent variable*, a statistical relation is of the form:

$$Y = f(X) + \epsilon.$$

  In such cases, we call X an *explanatory variable* and Y a *response variable.*

- In a *simple linear regression* model, the response variable Y is linearly related to one *explanatory* variable X. That is,

$$Y_i = (a + bx_i) + \epsilon_i. \qquad i = 1, 2, \ldots, n.$$

  Assumptions:

  1. The mean of ϵ_i is 0 and the constant variance of ϵ_i is σ^2.

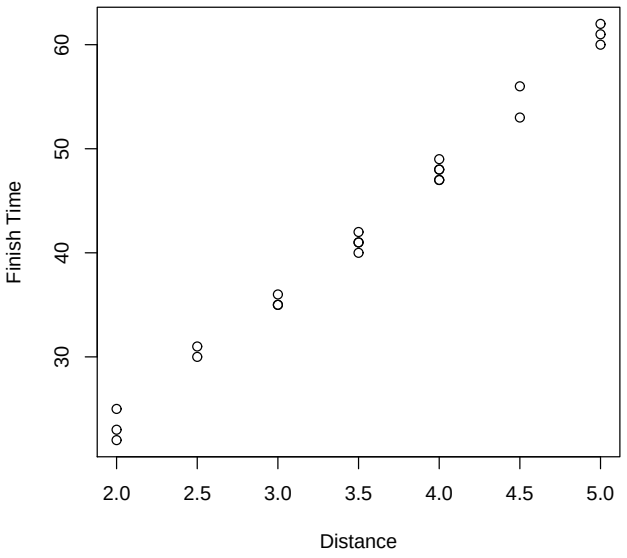
  2. The random errors ϵ_i are uncorrelated.

  3. a and b are parameters.

  4. x_i is a known constant.

- For example, let x denote the distance that a person plans to jog and y the time that it will take this person to finish it. Consider his 22 jogging distances and times from last month shown in the table below. If this person plans to jog for 5.5 miles tomorrow, predict how long it will take him to finish the run.

	1	2	3	4	5	6	7	8	9	10	11
Distance (x)	2	2	3	3	2	2.5	2.5	3	3.5	3.5	4
Time (y)	25	22	35	36	23	30	31	35	41	40	49
	12	13	14	15	16	17	18	19	20	21	22
Distance (x)	4	4	4	4.5	4.5	5	5	5	3.5	3.5	4
Time (y)	47	48	48	56	53	62	60	61	42	41	47



- **Equation of the Least-Squares Regression Line .** Suppose 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

$$\hat{y} = \hat{a} + \hat{b}x$$

with *slope*

$$\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{(\Sigma xy) - \frac{1}{n}(\Sigma x)(\Sigma y)}{(\Sigma x^2) - \frac{1}{n}(\Sigma x)^2} = r\frac{s_y}{s_x} \tag{1}$$

and *intercept*

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{2}$$

and *intercept*

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{3}$$

- **Practice.**

  1. The table below displays data on age (in years) and price (in $100) for a sample of 11 cars.

Age (x)	5	4	6	6	5	5	6	6	2	7	7
Price (y)	85	102	70	80	89	98	66	90	169	68	50

  a. Determine the values of SS_{xx}, SS_{yy}, and SS_{xy}.

  b. Determine the correlation coefficient r.

  c. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

  d. Determine the regression line.

  e. Estimate the expected value of a car that is 3 years old.

2. Tree Circumference and Height. Listed below are the circumferences (in feet) and the heights (in feet) of trees in Marshall, Minnesota (based on data from "Tree Measurements" by Stanley Rice, *American Biology Teacher*.

x (circ)	1.8	1.9	1.8	2.4	5.1	3.1	5.5
y (height)	21.0	33.5	24.6	40.7	73.2	24.9	40.4
x (circ)	5.1	8.3	13.7	5.3	4.9	3.7	3.8
y (height)	45.3	53.5	93.8	64.0	62.7	47.2	44.3

a. Determine the correlation coefficient r.

b. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

c. Determine the regression line.

d. Estimate the expected height of a tree that has a circumference of 10 feet.

3. A criminologist studying the relationship between population density and robbery rate in medium-sized U.S. cities collected the following data for a random sample of 16 cities; X is the population density of the city (number of people per unit area), and Y is the robbery rate last year (number of robberies per 100,000 people). Assume that the simple linear regression model is appropriate.

i	1	2	3	4	5	6	7	8
X_i	59	49	75	54	78	56	60	82
Y_i	209	180	195	192	215	197	208	189
i	9	10	11	12	13	14	15	16
X_i	69	83	88	94	47	65	89	70
Y_i	213	201	214	212	205	186	200	204

a. Determine the correlation coefficient r.

b. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

c. Determine the regression line.

d. Estimate the expected robbery rate (no. of robberies per 100,000 people) of a city with population density of $x = 90$.

4. To study the relationship between age x (in years) and body fat y, 18 adults (with ages from 33 to 48) were randomly selected. A summary of the data obtained is given below:

$$SS_{xx} = 2970,\ SS_{xy} = 1998,\ SS_{yy} = 1683,\ \sum x_i = 834 \text{ and } \sum y_i = 499$$

a. Determine the correlation coefficient.

b. Determine the regression line.

c. Using the regression line that you obtained in part (b), estimate the body fat of a 50-year-old person.

d. Based on your result in part (b), what can you expect will happen to someone's body fat as he/she ages by one year?

e. Based on your result in part (a), what can you say about the degree of linear relationship of age and body fat? Explain your answer.

f. Do you think it was appropriate to use linear regression to predict the body fat of a 50-year-old person? Explain your answer.

- **Recommended problems:**
  Section 12.2: (pp. 487-490) # 17, 19, 23, 25, 26.