CHAPTER 13

SIMPLE LINEAR REGRESSION

Prem Mann, *Introductory Statistics, 8/E*

13.1 Simple Linear Regression

- Simple Regression
- Linear Regression

Definition

A regression model is a mathematical equation that describes the relationship between two or more variables. A ***simple regression*** model includes only two variables: one independent and one dependent. The dependent variable is the one being explained, and the independent variable is the one used to explain the variation in the dependent variable.

Linear Regression

Definition

A (simple) regression model that gives a straight-line relationship between two variables is called a ***linear regression*** model.

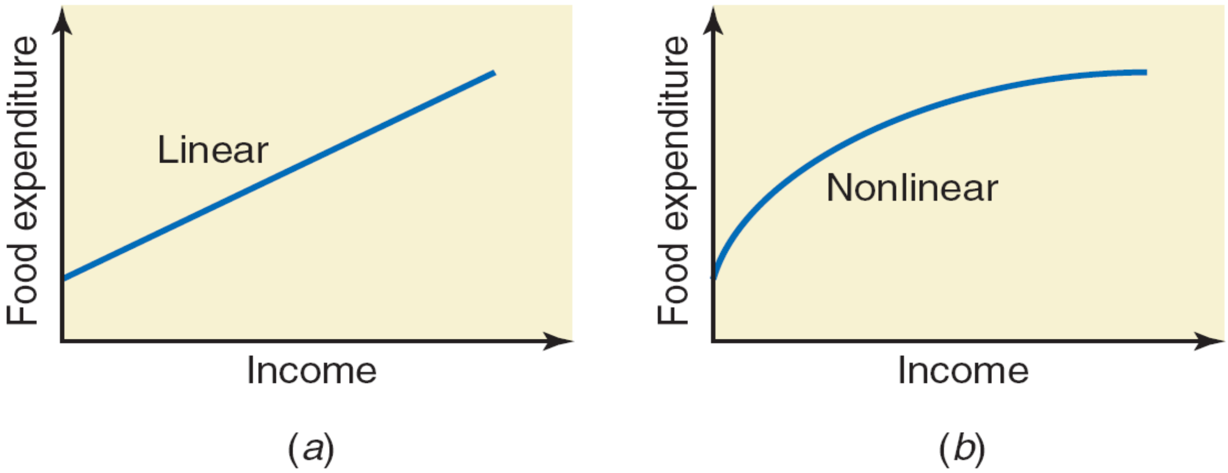


Figure 13.2 Plotting a linear equation.

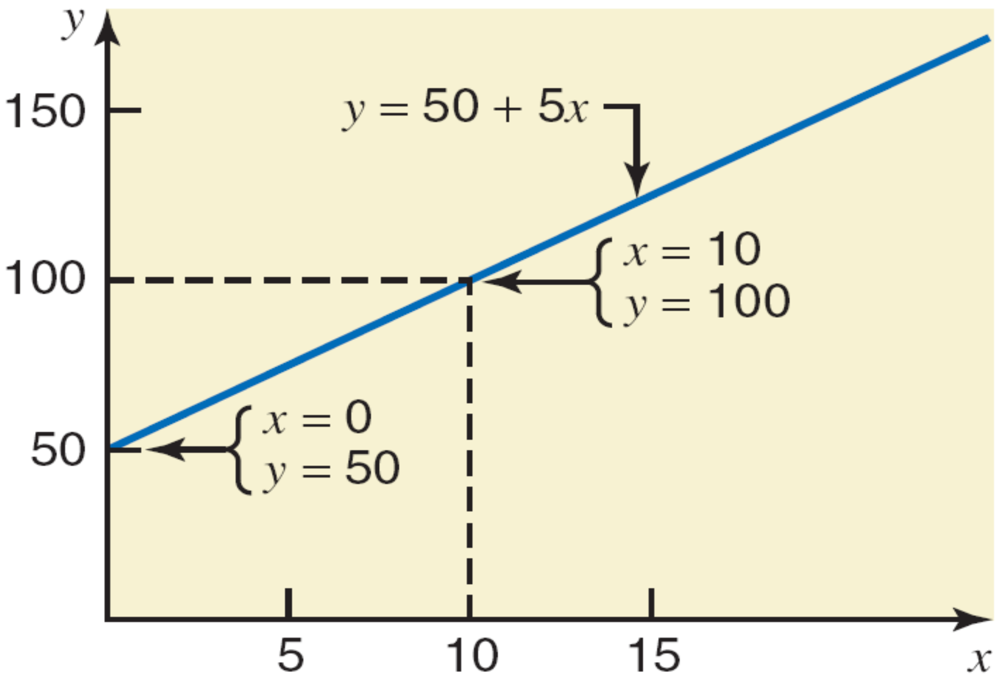
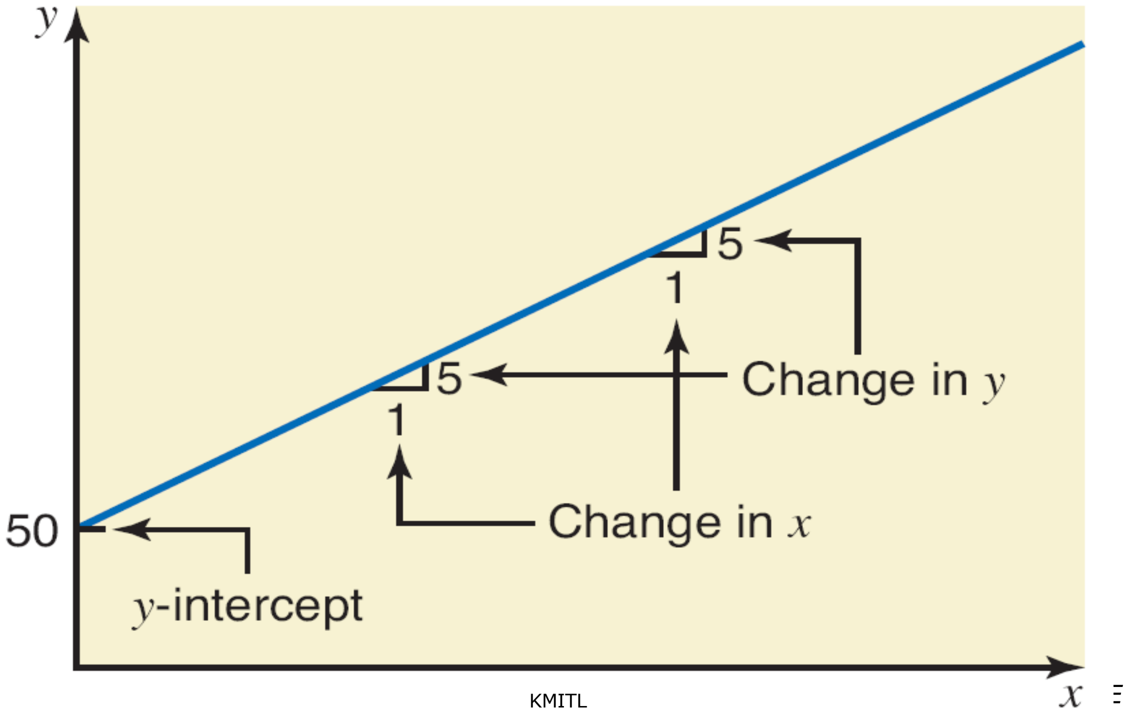


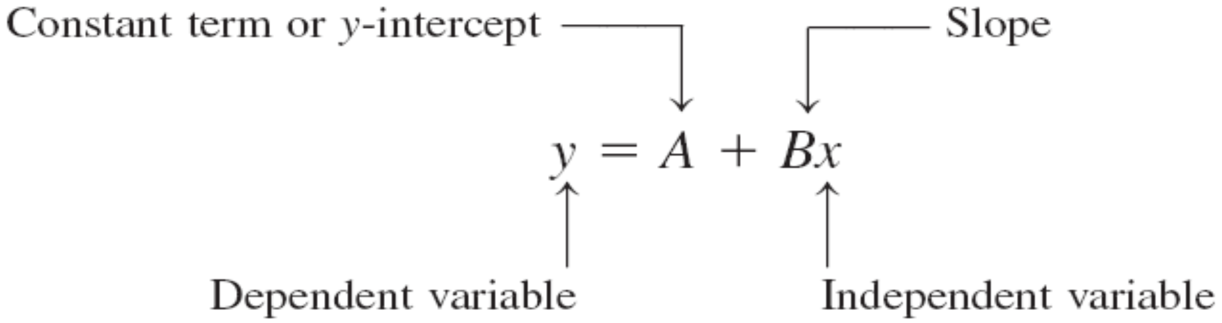
Figure 13.3 y-intercept and slope of a line.



SIMPLE LINEAR REGRESSION ANALYSIS

Definition

In the ***regression model*** $y = A + Bx + \varepsilon$, A is called the y-intercept or constant term, B is the slope, and ε is the random error term. The dependent and independent variables are y and x, respectively.



Constant term or y-intercept

Slope

$$y = A + Bx$$

Dependent variable

Independent variable

SIMPLE LINEAR REGRESSION ANALYSIS

Definition

In the model $\hat{y} = a + bx$, *a* and *b*, which are calculated using sample data, are called the ***estimates of A and B***, respectively.

Table 13.1 Incomes (in hundreds of dollars) and Food Expenditures of Seven Households

Income	Food Expenditure
55	14
83	24
38	13
61	16
33	9
49	15
67	17

Scatter Diagram

Definition

A plot of paired observations is called a ***scatter diagram***.

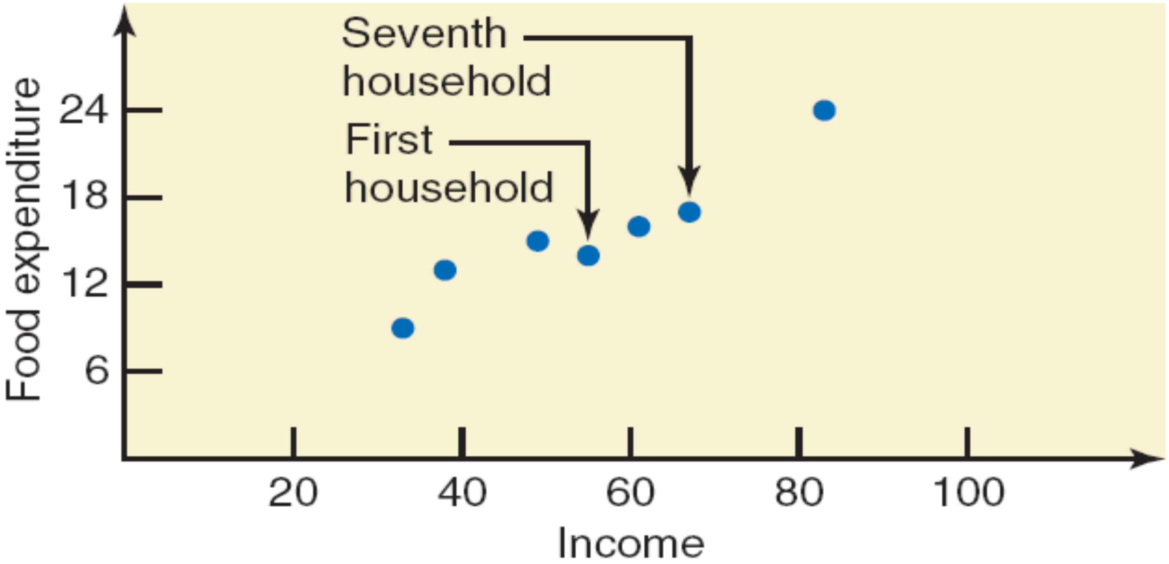


Figure 13.5 Scatter diagram and straight lines.

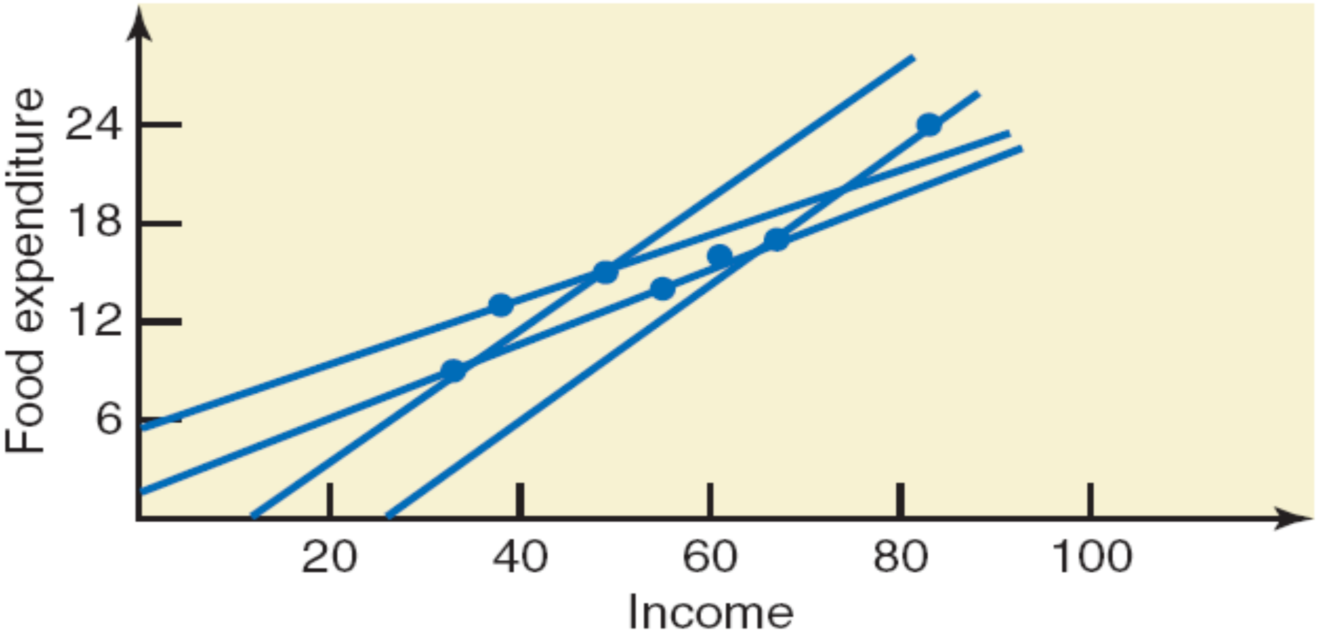
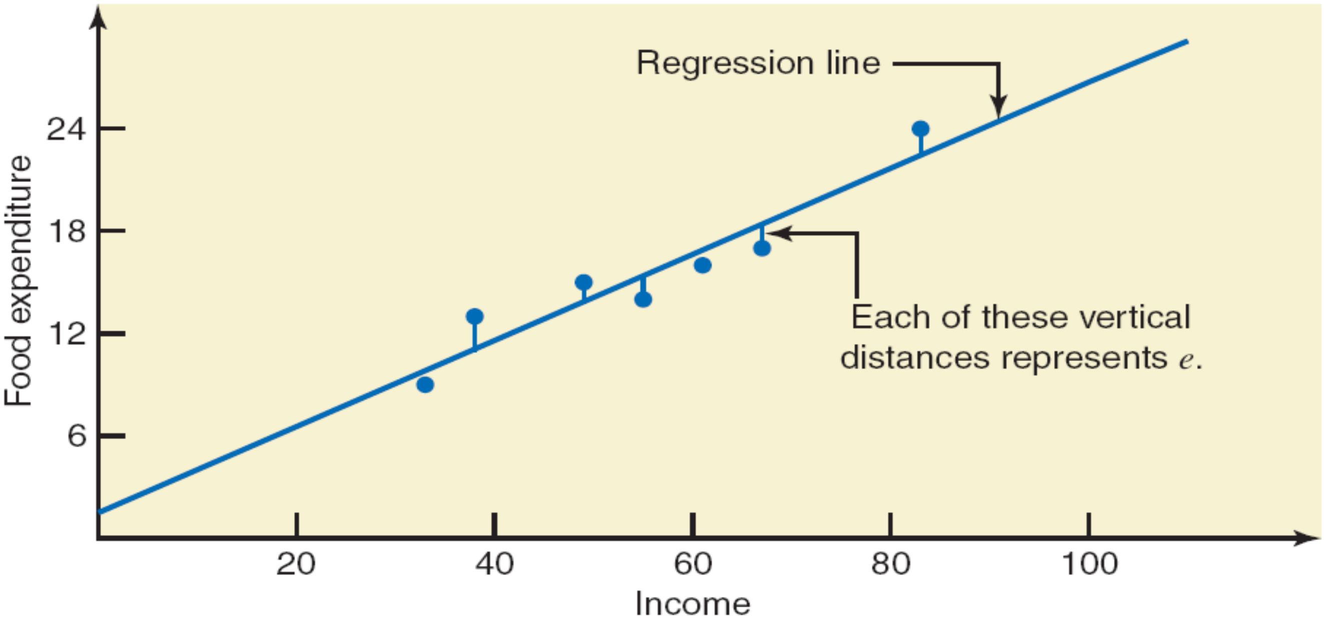


Figure 13.6 Regression Line and random errors.



Error Sum of Squares (SSE)

The ***error sum of squares***, denoted SSE, is

$$\mathbf{SSE} = \sum \mathbf{e}^2 = \sum (\boldsymbol{y} - \boldsymbol{\hat{y}})^2$$

The values of *a* and *b* that give the minimum SSE are called the ***least square estimates*** of *A* and *B*, and the regression line obtained with these estimates is called the ***least squares line***.

The Least Squares Line

For the least squares regression line $\hat{y} = a + bx$,

$$b = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

where

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} \quad \text{and} \quad SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

and SS stands for "sum of squares." The least squares regression line $\hat{y} = a + bx$ is also called the regression of y on x.

Example 13-1

Find the least squares regression line for the data on incomes and food expenditure on the seven households given in the Table 13.1. Use income as an independent variable and food expenditure as a dependent variable.

Table 13.2

Income x	Food Expenditure y	xy	x^2
55	14	770	3025
83	24	1992	6889
38	13	494	1444
61	16	976	3721
33	9	297	1089
49	15	735	2401
67	17	1139	4489
$\Sigma x = 386$	$\Sigma y = 108$	$\Sigma xy = 6403$	$\Sigma x^2 = 23{,}058$

Example 13-1: Solution

$$\sum x = 386 \qquad \sum y = 108$$

$$\bar{x} = \sum x / n = 386/7 = 55.1429$$

$$\bar{y} = \sum y / n = 108/7 = 15.4286$$

$$SS_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n} = 6403 - \frac{(386)(108)}{7} = 447.5714$$

$$SS_{xx} = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = 23{,}058 - \frac{(386)^2}{7} = 1772.8571$$

Example 13-1: Solution

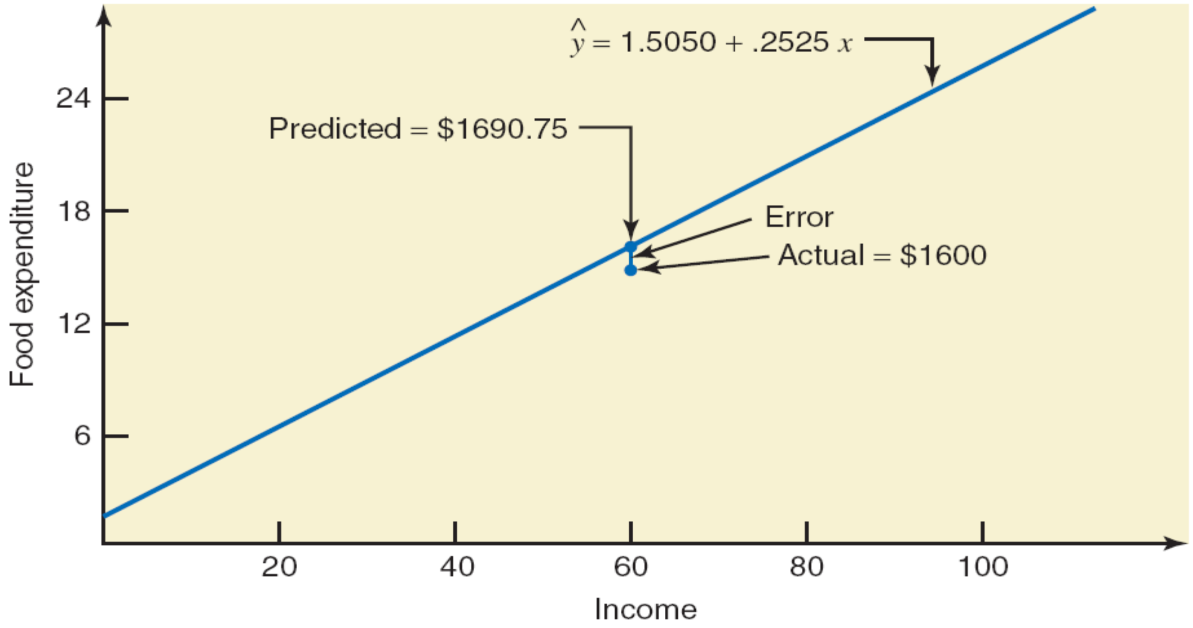
$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{447.5714}{1772.8571} = .2525$$

$$a = \bar{y} - b\bar{x} = 15.4286 - (.2525)(55.1429) = 1.5050$$

Thus, our estimated regression model is

$$\hat{y} = 1.5050 + .2525\,x$$

Figure 13.7 Error of prediction.



Interpretation of *a* and *b*

Interpretation of *a*

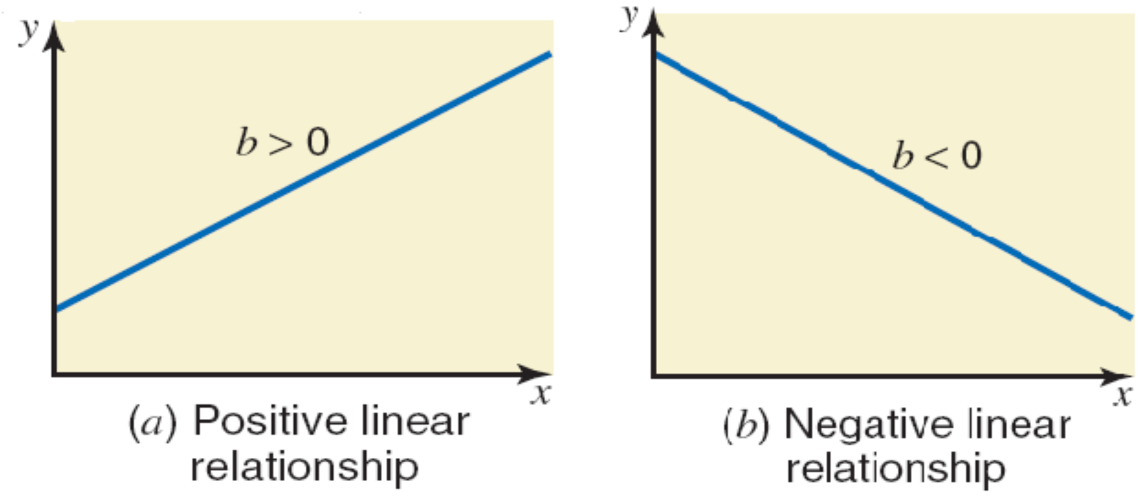
- Consider a household with zero income. Using the estimated regression line obtained in Example 13-1,
  - $\hat{y}$ = 1.5050 + .2525(0) = $1.5050 hundred.
- Thus, we can state that a household with no income is expected to spend $150.50 per month on food.
- The regression line is valid only for the values of x between 33 and 83.

Interpretation of *a* and *b*

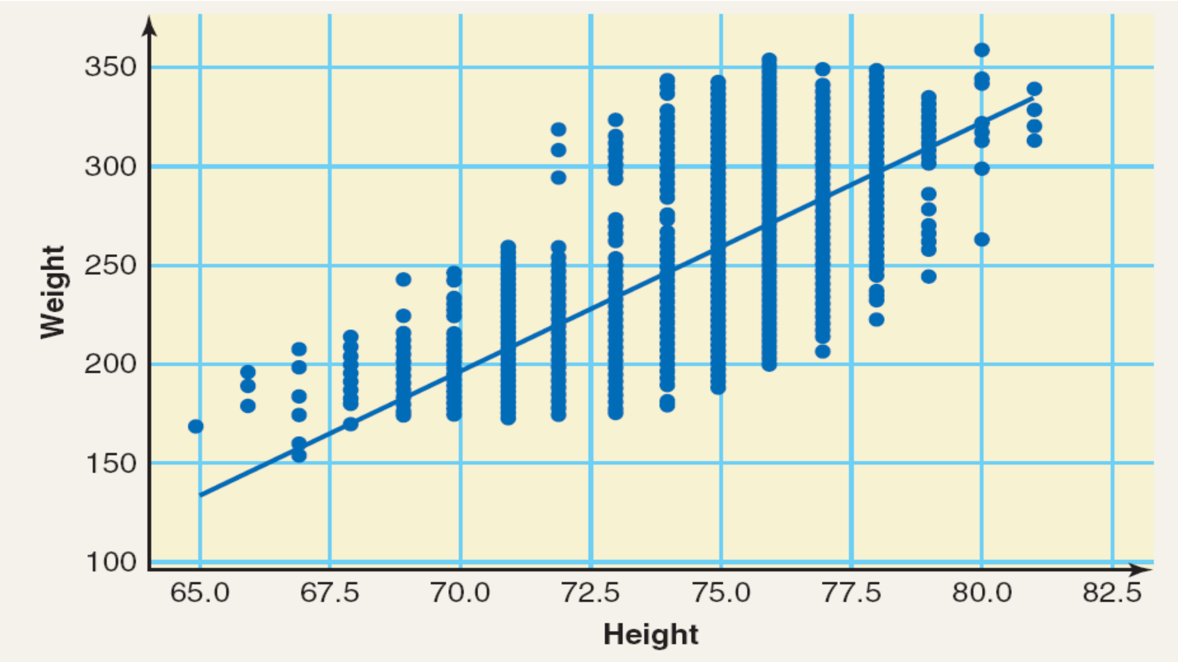
Interpretation of *b*

- The value of b in the regression model gives the change in y (dependent variable) due to a change of one unit in x (independent variable).
- We can state that, on average, a $100 (or $1) increase in income of a household will increase the food expenditure by $25.25 (or $.2525).

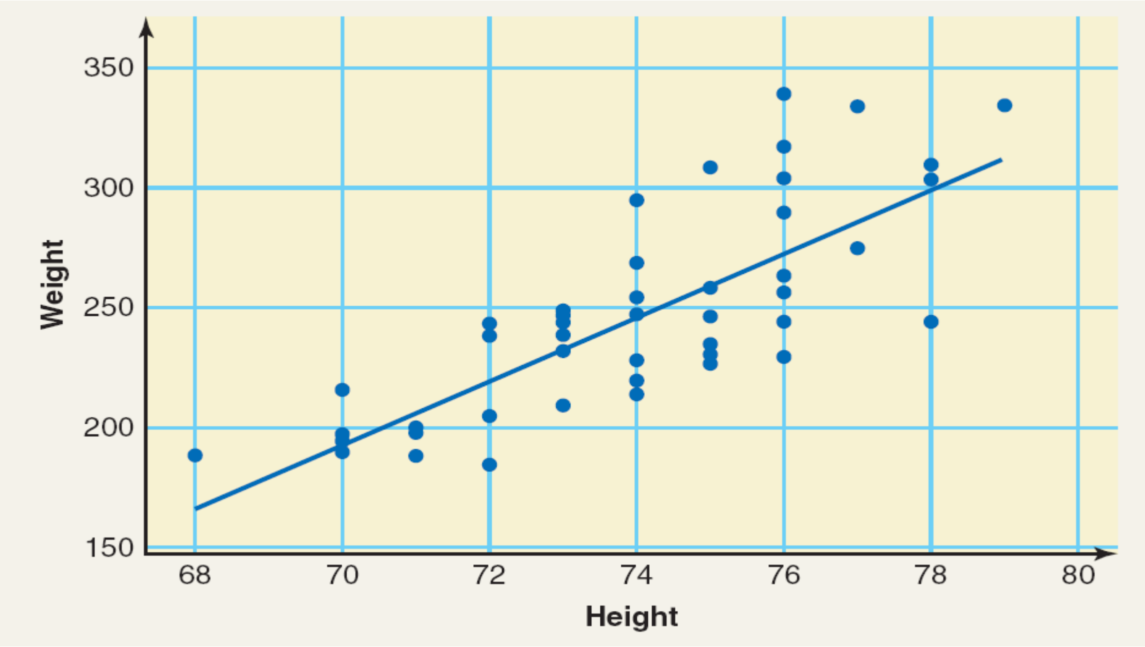
Figure 13.8 Positive and negative linear relationships between x and y.


y
$b > 0$
x
(a) Positive linear relationship
y
$b < 0$
x
(b) Negative linear relationship

Case Study 13-1 Regression of Weights on Heights for NFL Players



Case Study 13-1 Regression of Weights on Heights for NFL Players



Assumptions of the Regression Model

Assumption 1: The random error term Є has a mean equal to zero for each x

Assumption 2: The errors associated with different observations are independent

Assumption 3: For any given x, the distribution of errors is normal

Assumption 4: The distribution of population errors for each x has the same (constant) standard deviation, which is denoted σ_{ϵ}

Figure 13.11 (a) Errors for households with an income of $4000 per month.

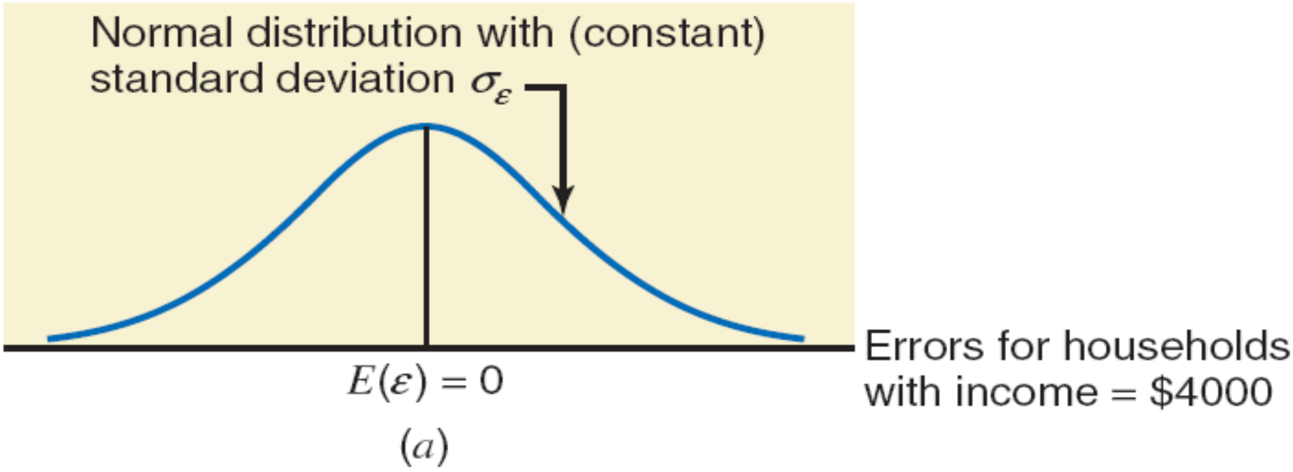


Figure 13.11 (b) Errors for households with an income of $ 7500 per month.

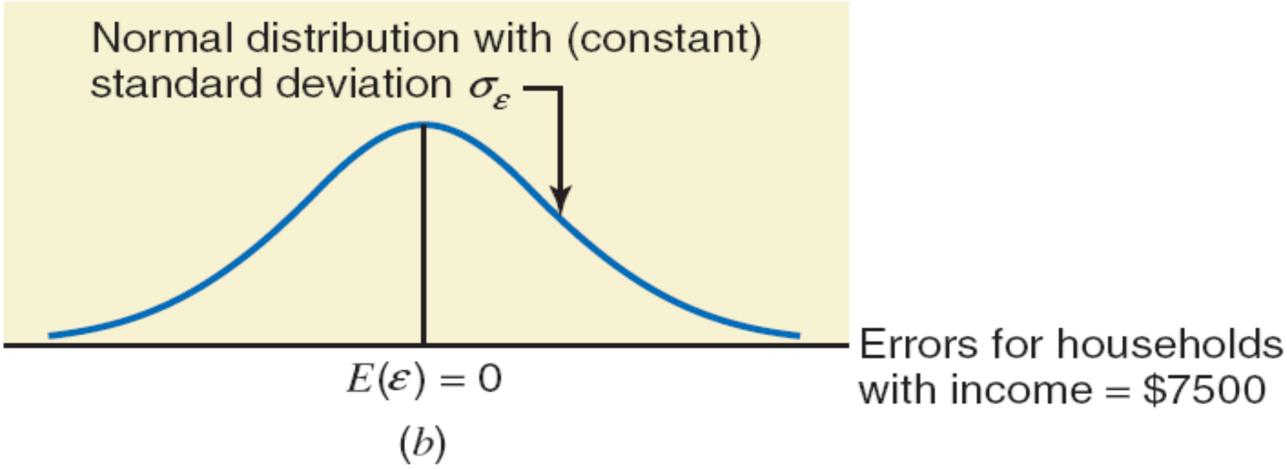


Figure 13.12 Distribution of errors around the population regression line.

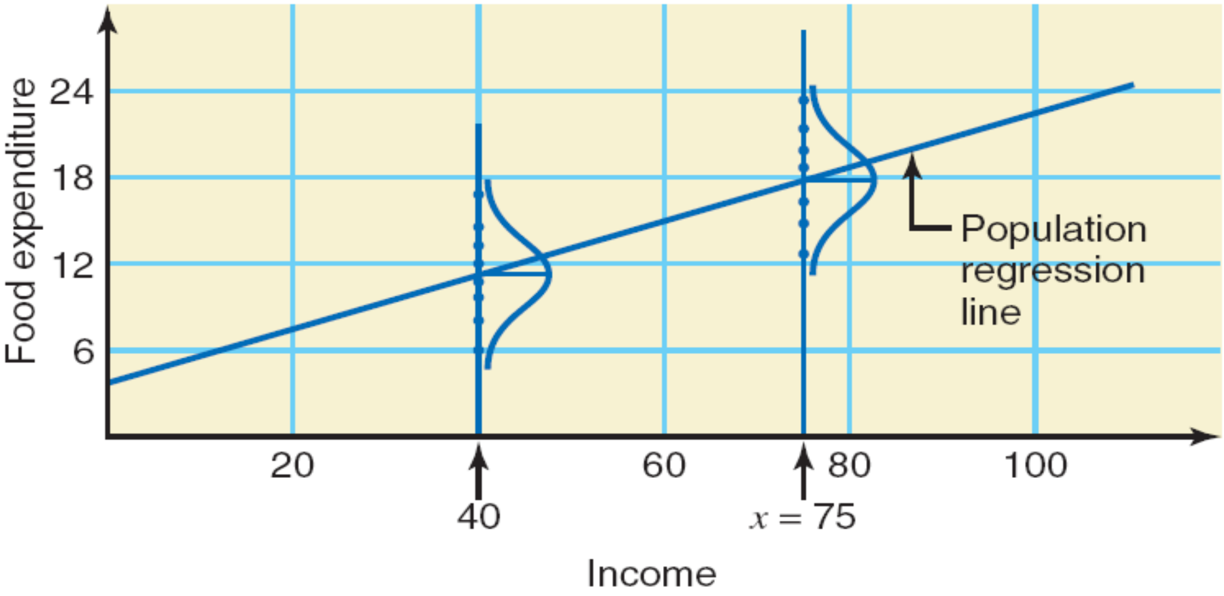
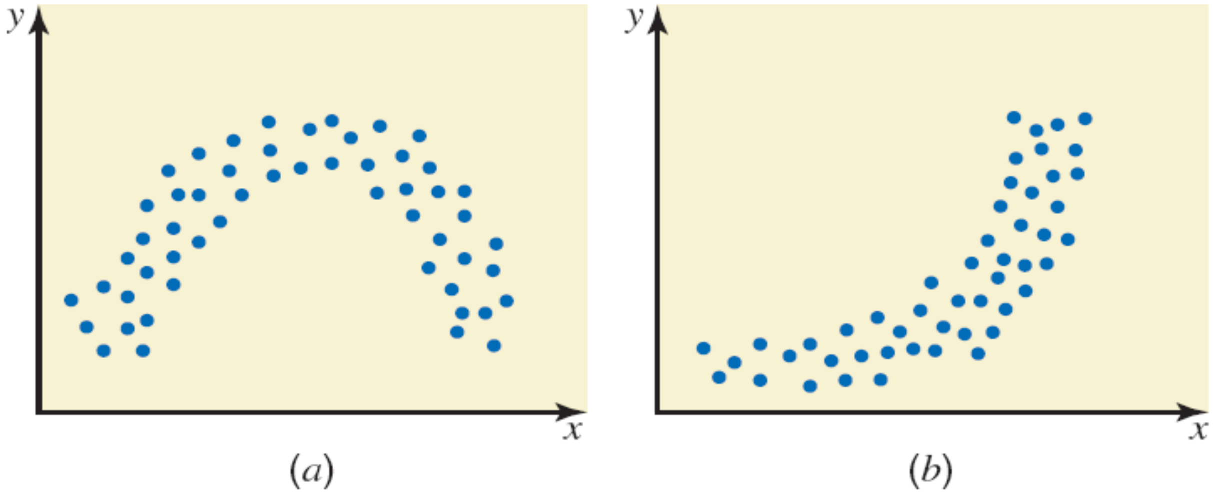
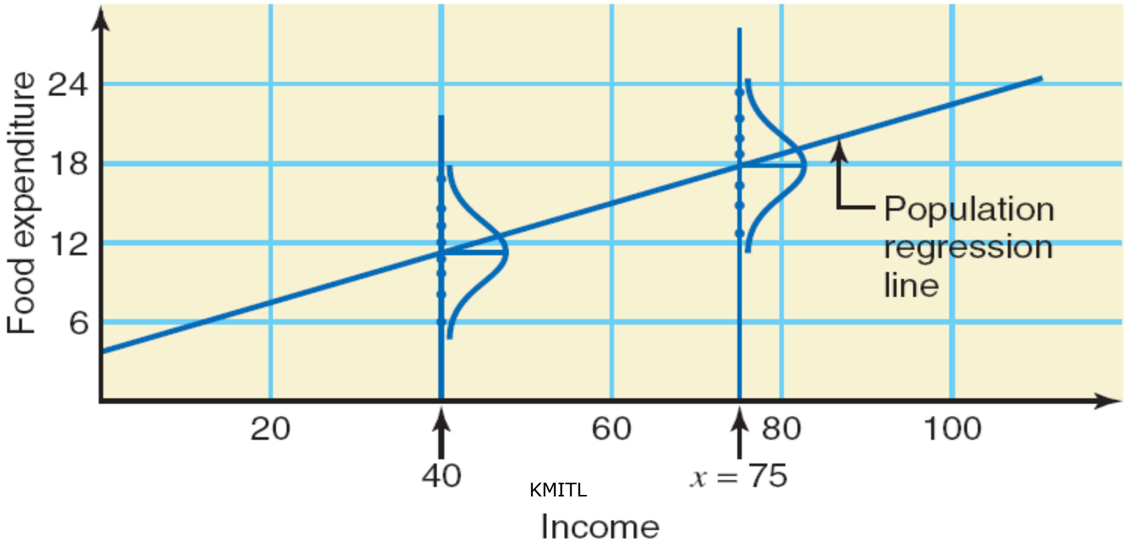


Figure 13.13 Nonlinear relations between *x* and *y*.



13.2 Standard Deviation of Errors and Coefficient of Determination

Degrees of Freedom for a Simple Linear Regression Model The ***degrees of freedom*** for a simple linear regression model are $df = n - 2$



STANDARD DEVIATION OF ERRORS AND COEFFICIENT OF DETERMINATION

The ***standard deviation of errors*** is calculated as

$$s_e = \sqrt{\frac{SS_{yy} - bSS_{xy}}{n-2}}$$

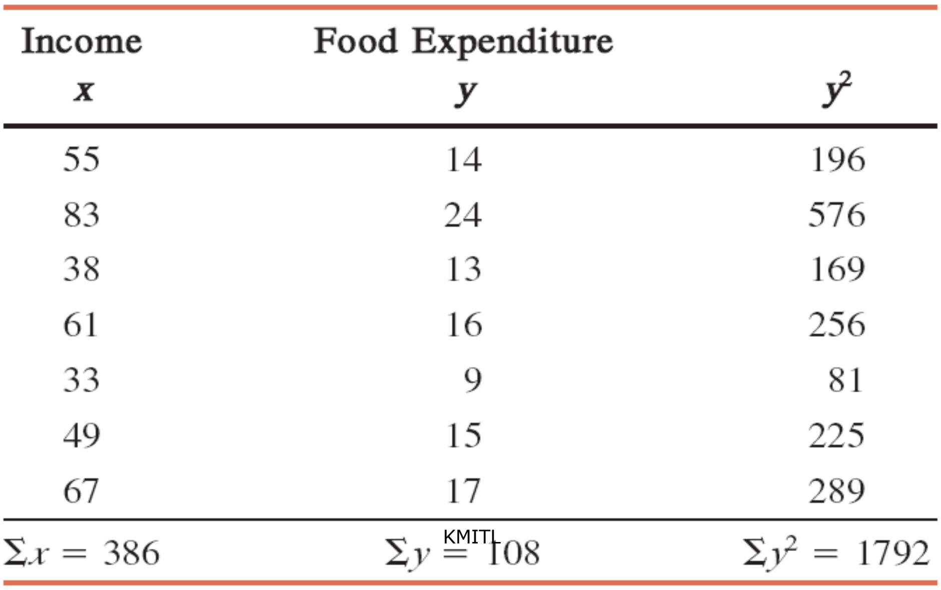
where

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

Prem Mann, *Introductory Statistics*, 8/E

Example 13-2

Compute the standard deviation of errors s_e for the data on monthly incomes and food expenditures of the seven households given in Table 13.1.



Income x	Food Expenditure y	y^2
55	14	196
83	24	576
38	13	169
61	16	256
33	9	81
49	15	225
67	17	289
$\Sigma x = 386$	$\Sigma y = 108$	$\Sigma y^2 = 1792$

Example 13-2: Solution

$$SS_{yy} = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = 1792 - \frac{(108)^2}{7} = 125.7143$$

$$s_e = \sqrt{\frac{SS_{yy} - bSS_{xy}}{n-2}} \sqrt{\frac{125.7143 - .2525(447.5714)}{7-2}} = 1.5939$$

COEFFICIENT OF DETERMINATION

Total Sum of Squares (SST)

The ***total sum of squares***, denoted by **SST**, is calculated as

$$SST = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

Note that this is the same formula that we used to calculate SS_{yy}.

Figure 13.15 Total errors.

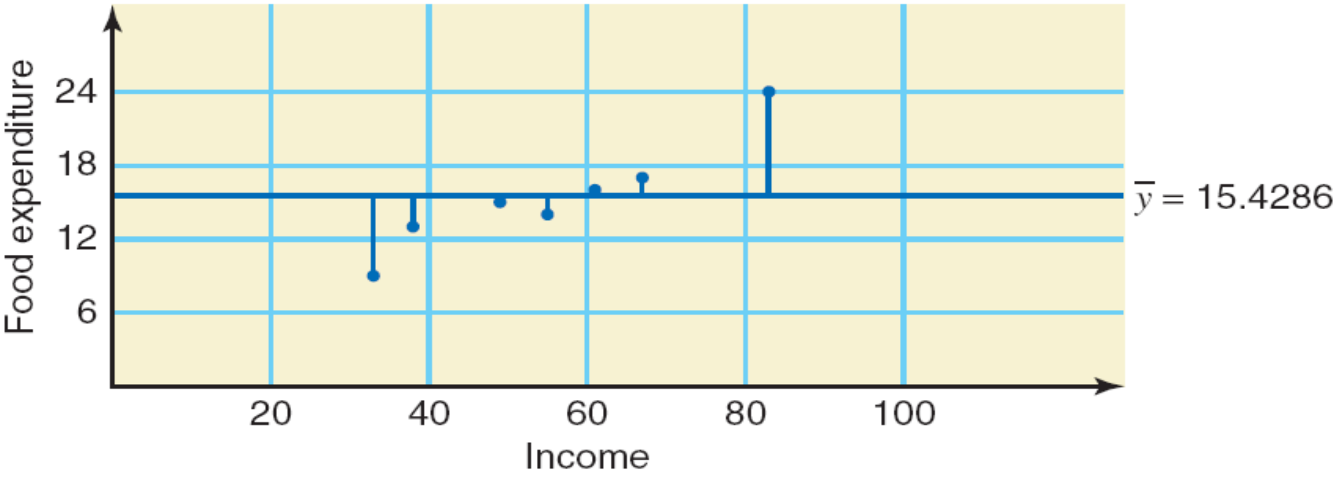
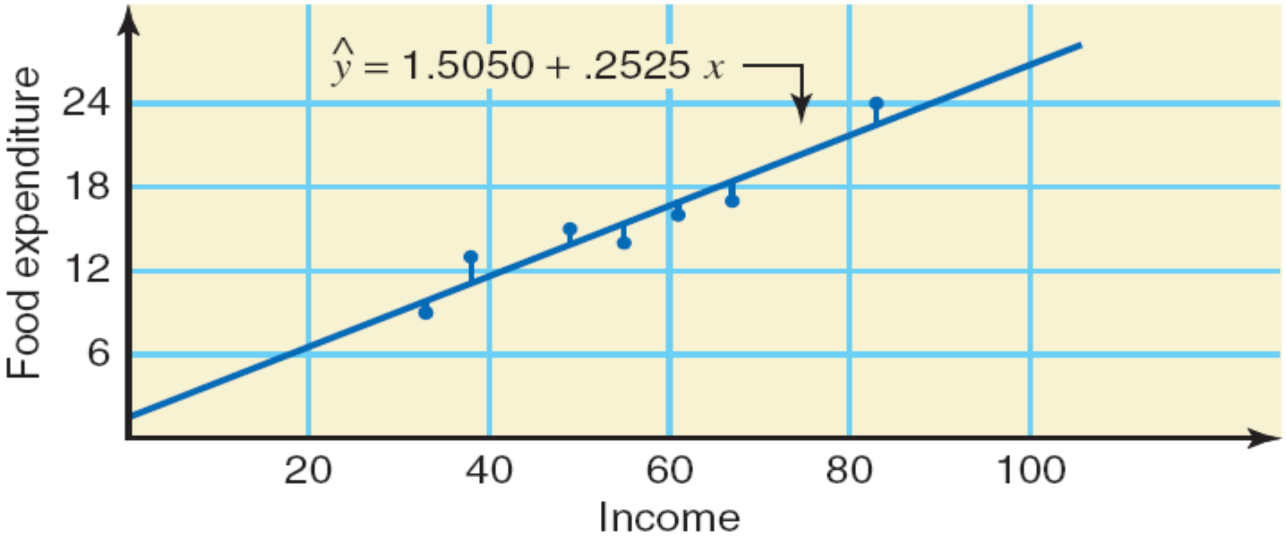


Table 13.4

x	y	$\hat{y} = 1.5050 + .2525x$	$e = y - \hat{y}$	$e^2 = (y - \hat{y})^2$
55	14	15.3925	−1.3925	1.9391
83	24	22.4625	1.5375	2.3639
38	13	11.1000	1.9000	3.6100
61	16	16.9075	−.9075	.8236
33	9	9.8375	−.8375	.7014
49	15	13.8775	1.1225	1.2600
67	17	18.4225	−1.4225	2.0235
				$\Sigma e^2 = \Sigma(y - \hat{y})^2 = 12.7215$

Figure 13.16 Errors of prediction when regression model is used.



COEFFICIENT OF DETERMINATION

Regression Sum of Squares (SSR)

The ***regression sum of squares*** , denoted by **SSR**, is

$$SSR = SST - SSE$$

COEFFICIENT OF DETERMINATION

Coefficient of Determination

The ***coefficient of determination***, denoted by r^2, represents the proportion of SST that is explained by the use of the regression model. The computational formula for r^2 is

$$r^2 = \frac{b\ SS_{xy}}{SS_{yy}}$$

and $0 \le r^2 \le 1$

Example 13-3

For the data of Table 13.1 on monthly incomes and food expenditures of seven households, calculate the coefficient of determination.

- From earlier calculations made in Examples 13-1 and 13-2,
- ***b*** **= .2525,** ***SS***$_{xx}$ **= 447.5714,** ***SS***$_{yy}$ **= 125.7143**

$$r^2 = \frac{b\ SS_{xy}}{SS_{yy}} = \frac{(.2525)(447.5714)}{125.7143} = .90$$

13.4 Linear Correlation

- Linear Correlation Coefficient
- Hypothesis Testing About the Linear Correlation Coefficient

Value of the Correlation Coefficient

The ***value of the correlation coefficient*** always lies in the range of −1 to 1; that is,

$$-1 \leq \rho \leq 1 \quad \text{and} \quad -1 \leq r \leq 1$$

Figure 13.18 Linear correlation between two variables.

(a) Perfect positive linear correlation, r = 1

(b) (b) Perfect negative linear correlation, r = -1

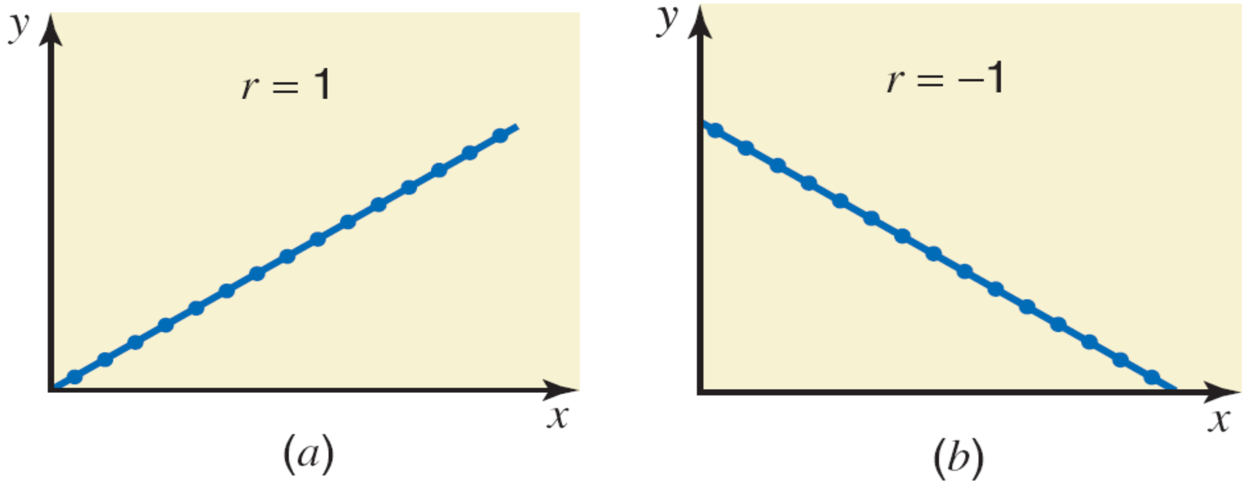


Figure 13.18 Linear correlation between two variables.

(c) No linear correlation, , $r \approx 0$

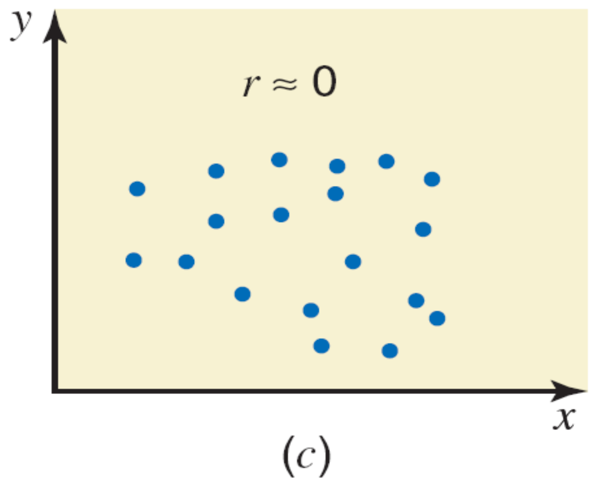
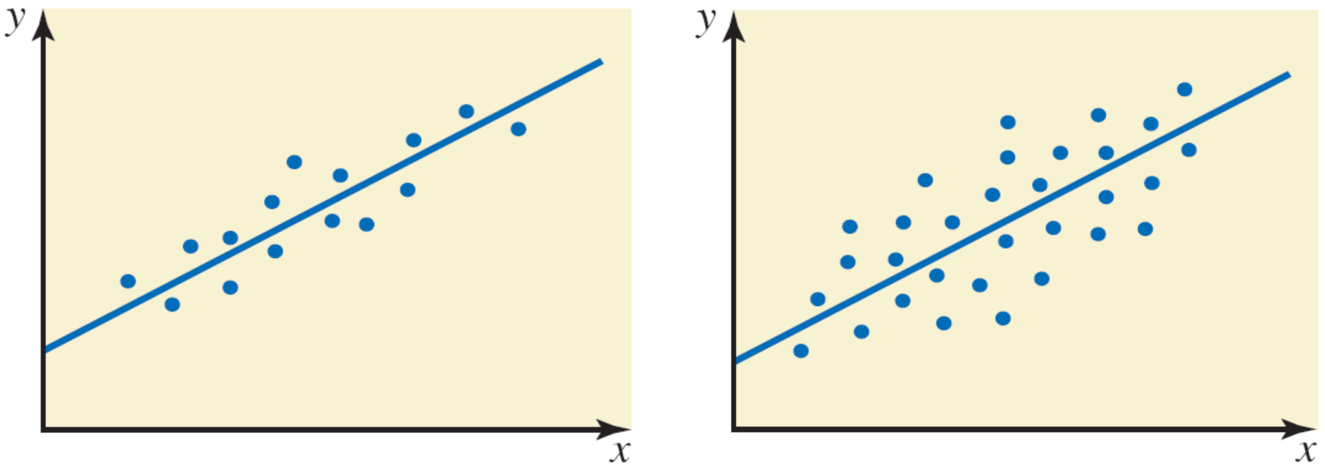


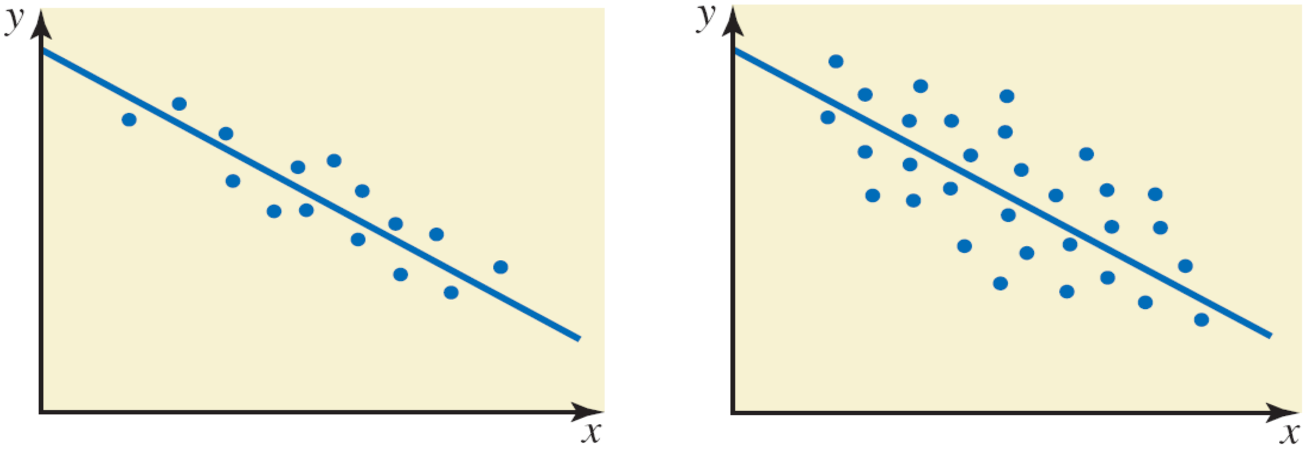
Figure 13.19 Linear correlation between variables.



(*a*) Strong positive linear correlatio (*r* is close to 1)

(*b*) Weak positive linear correlation (*r* is positive but close to zero)

Figure 13.19 Linear correlation between variables.



(*c*) Strong negative linear correlation (*r* is close to –1)

(*d*) Weak negative linear correlation (*r* is negative and close to zero)

Linear Correlation Coefficient

Linear Correlation Coefficient

The ***simple linear correlation coefficient***, denoted by *r*, measures the strength of the linear relationship between two variables for a sample and is calculated as

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Example 13-6

Calculate the correlation coefficient for the example on incomes and food expenditures of seven households.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

$$= \frac{447.5714}{\sqrt{(1772.8571)(125.7143)}} = .95$$

Hypothesis Testing About the Linear Correlation Coefficient

Test Statistic for *r*

If both variables are normally distributed and the null hypothesis is H_0: $\rho = 0$, then the value of the test statistic t is calculated as

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Here $n - 2$ are the degrees of freedom.

Example 13-7

Using the 1% level of significance and the data from Example 13-1, test whether the linear correlation coefficient between incomes and food expenditures is positive. Assume that the populations of both variables are normally distributed.

Example 13-7: Solution

- **Step 1:**

  H_0: ρ = 0 (The linear correlation coefficient is zero)
  H_1: ρ > 0 (The linear correlation coefficient is positive)

- **Step 2:**

  The population distributions for both variables are normally distributed. Hence, we can use the *t* distribution to perform this test about the linear correlation coefficient.

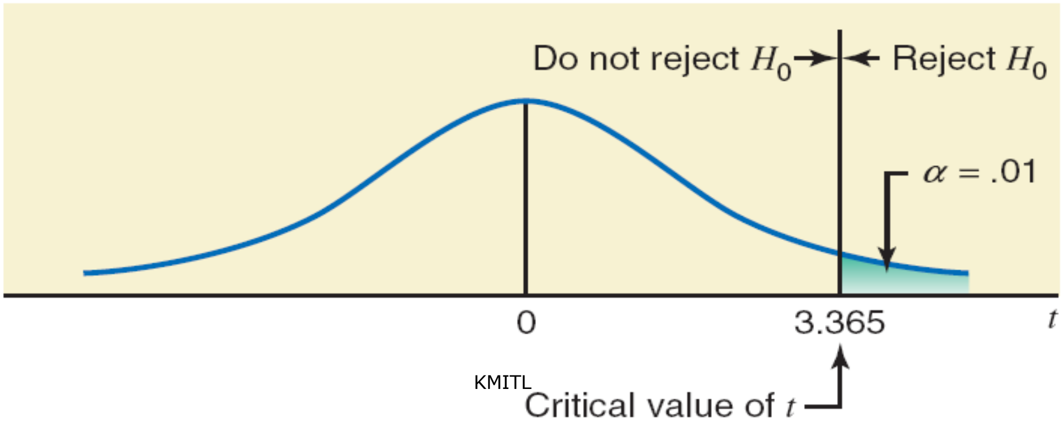
Example 13-7: Solution

- **Step 3:**

  Area in the right tail = .01

  $df = n - 2 = 7 - 2 = 5$

  The critical value of t = 3.365



Example 13-7: Solution

- ❑ **Step 4:**

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$t = 0.9481\sqrt{\frac{7-2}{1-(0.9481)^2}}$$

$$= 6.667$$

Example 13-7: Solution

- **Step 5:**

  The value of the test statistic t = 6.667

  - It is greater than the critical value of $t=3.365$
  - It falls in the rejection region

  Hence, we reject the null hypothesis.

  We conclude that there is a positive relationship between incomes and food expenditures.