REGRESSION AND CORRELATION

APPLIED STATISTICS (STAT200)

INTRODUCTION

INTRODUCTION TO CHAPTER 12. (DEVORE & BERK 2012)

THE OBJECTIVE OF REGRESSION

Regression analysis attempts to determine the relationship between several variables so that we can gain information about one of them (response) through knowing values of the others (predictors).

When we analyze only two variables, we denote them with x and y. Two cases exist:

- x and y are deterministically related: once we are told the value of x, the value of y is fully specified
- x and y are non-deterministically related: the value of y cannot be determined just from knowledge of x, but patterns can be seen in the data

Regression analysis investigates the latter relationship.

DETERMINISTIC EXAMPLES

Example 1: We rent a van for one day and the rental cost is $25.00 plus $0.30 per mile driven. If we let $x =$ the number of miles driven and $y =$ the rental change, then $y = 25 + 0.3x$. Thus, driving the van 100 miles results in a fee of $y = 25\$ + 0.3\$ \cdot 100 = 55\$$.

Example 2: If the starting velocity of a particle is v_o and it undergoes constant acceleration a, then the distance traveled $y = v_0 x + \frac{1}{2} a x^2$ where x corresponds to time.

NON-DETERMINISTIC EXAMPLES

Example 3: Consider the variables $x =$ high school grade point average (GPA) and $y =$ college GPA. Y cannot be determined just by knowing x, but there is a tendency: those students who have high school GPAs usually also have high college GPAs.

Example 4: $x =$ age of child, $y =$ size of child's vocabulary.

Example 5: $x =$ size of an engine in cubic centimeters, $y =$ fuel efficiency for an automobile equipped with that engine.

PROGRAM OF THE CHAPTER

In this chapter we will see:

- Simple linear regression: a linear probabilistic model for relating two variables
- Study procedures for making inferences based on the data. This will lead us to the correlation coefficient, a quantitative measure of the relationship of two variables
- Assess the adequacy of any particular regression model
- Multiple regression analysis: it relates the variable y to two or more variables. Example: relating fuel efficiency (y) to the four variables weight, engine size, number of cylinders, and transmission type

SIMPLE LINEAR REGRESSION MODELS

SECTION 12.1 (DEVORE & BERK 2012)

SIMPLE LINEAR REGRESSION

Let y be the dependent / response variable, and x be the independent / predictor variable.

Once we know the value of x, there is still uncertainty in the value of y in probabilistic models. A typical approach is given by the model equation

$$\begin{aligned} y &= \text{some particular deterministic function } x + \text{a random deviation} \\ &= f(x) + \varepsilon \end{aligned}$$

ε is called random deviation or random error, and is assumed to have mean value 0. Its role is to allow for a non-deterministic relationship.

THE ROLE OF ε

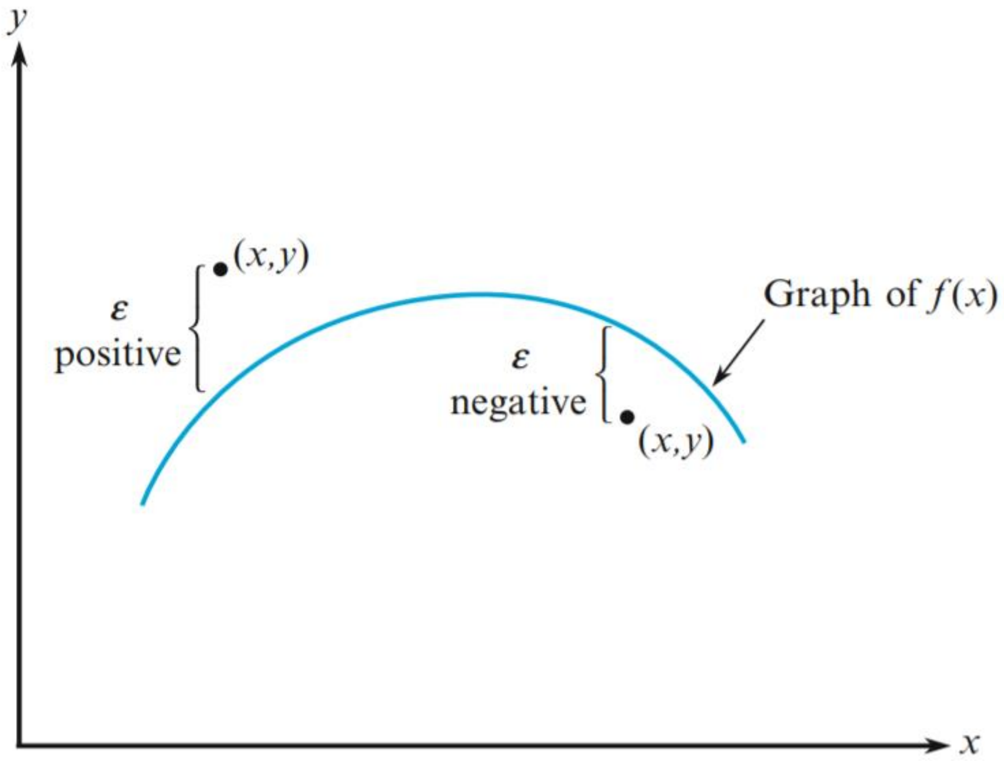


Figure 12.1 Observations resulting from the model equation (12.1)

- If $\varepsilon > 0$: then (x, y) falls above the graph of $f(x)$
- If $\varepsilon < 0$: then (x, y) falls below the graph of $f(x)$.

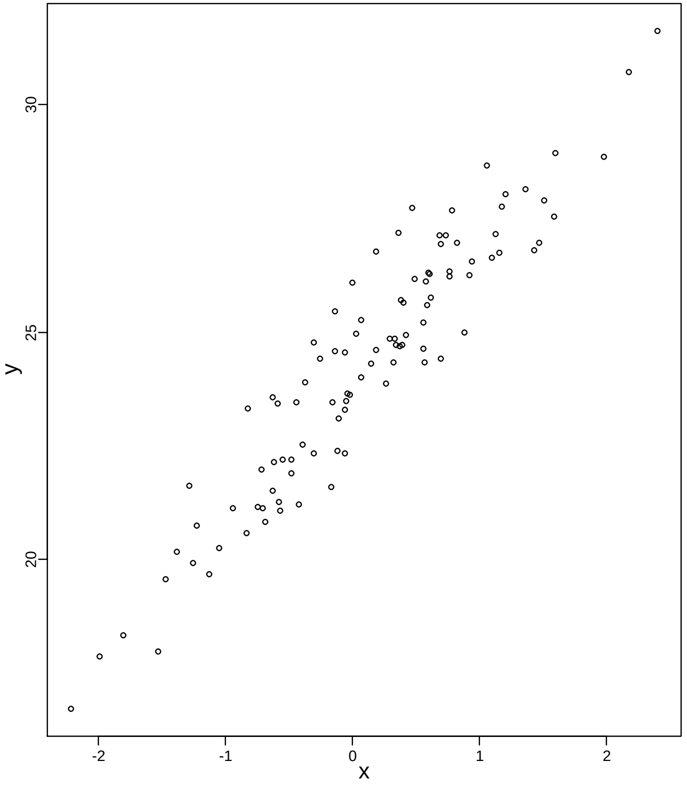
Since the mean of ε is 0, we could expect (x, y) to lay on the graph - but this actually (almost) never happens.

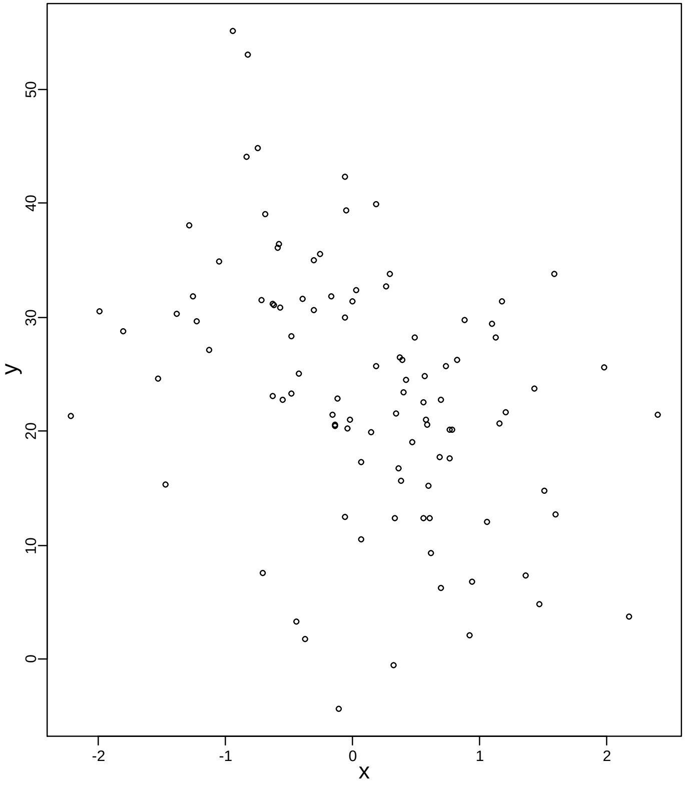
SCATTER PLOTS

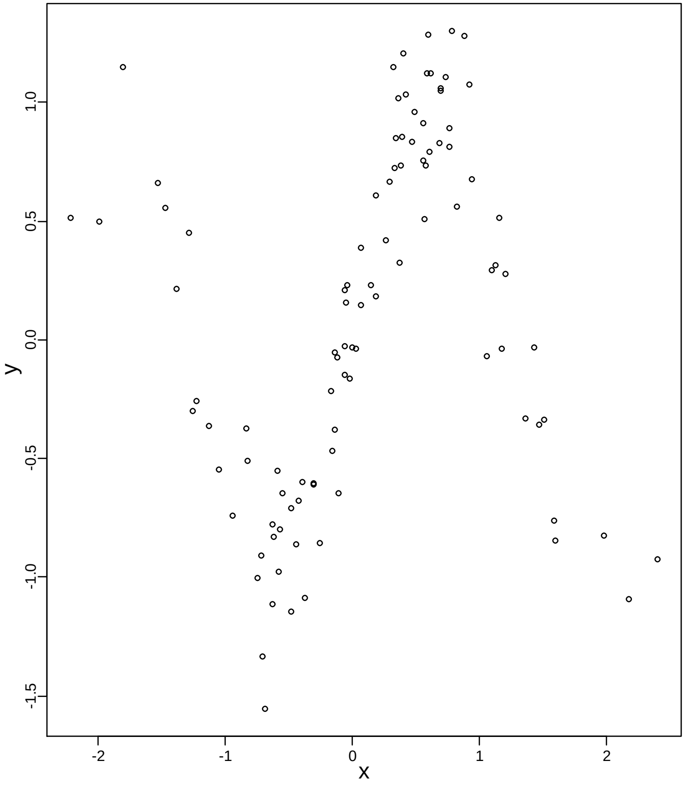
How do we choose the correct $f(x)$?

- The sample data we analyze are in the form of n pairs (x, y). We create a picture of the observations $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$ called scatter plot
- The pattern of points in the plot should suggest an appropriate $f(x)$

Examples:





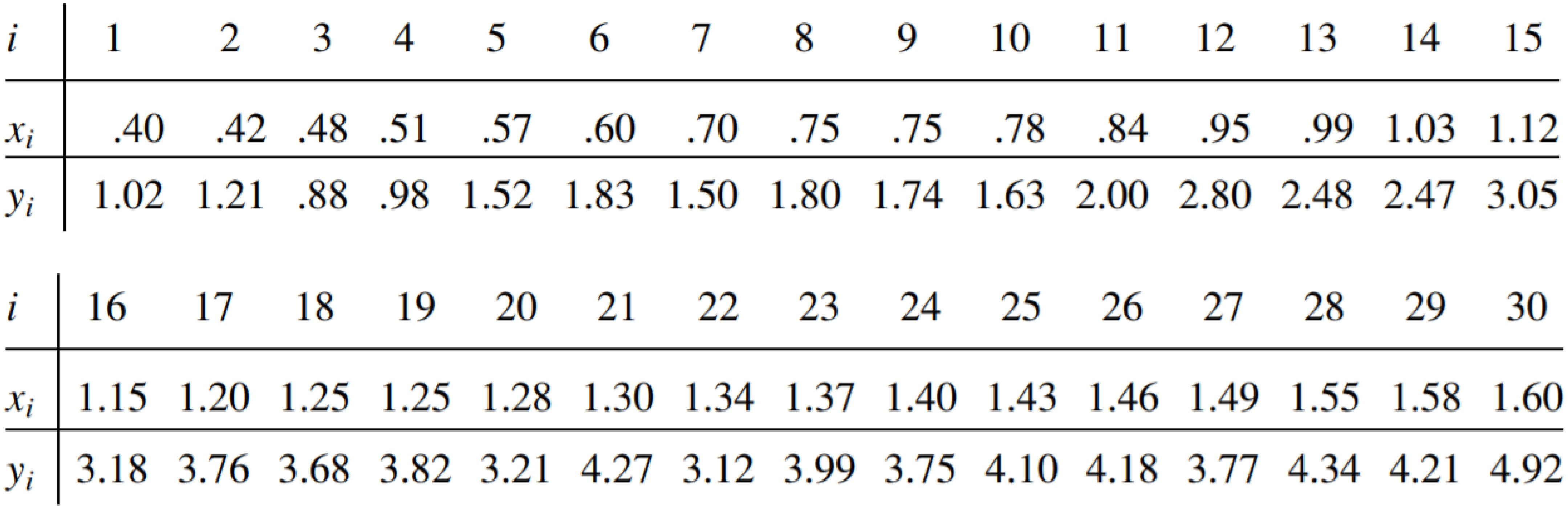


EXAMPLE 12.1

We possess thirty observations (x_i, y_i) where

- $x =$ horizontal width of the eye opening, in cm, and
- $y =$ ocular surface area (OSA).

For example, we observe $(x_1, y_1) = (.40, 1.02), (x_2, y_2) = (.42, 1.21)$ etc

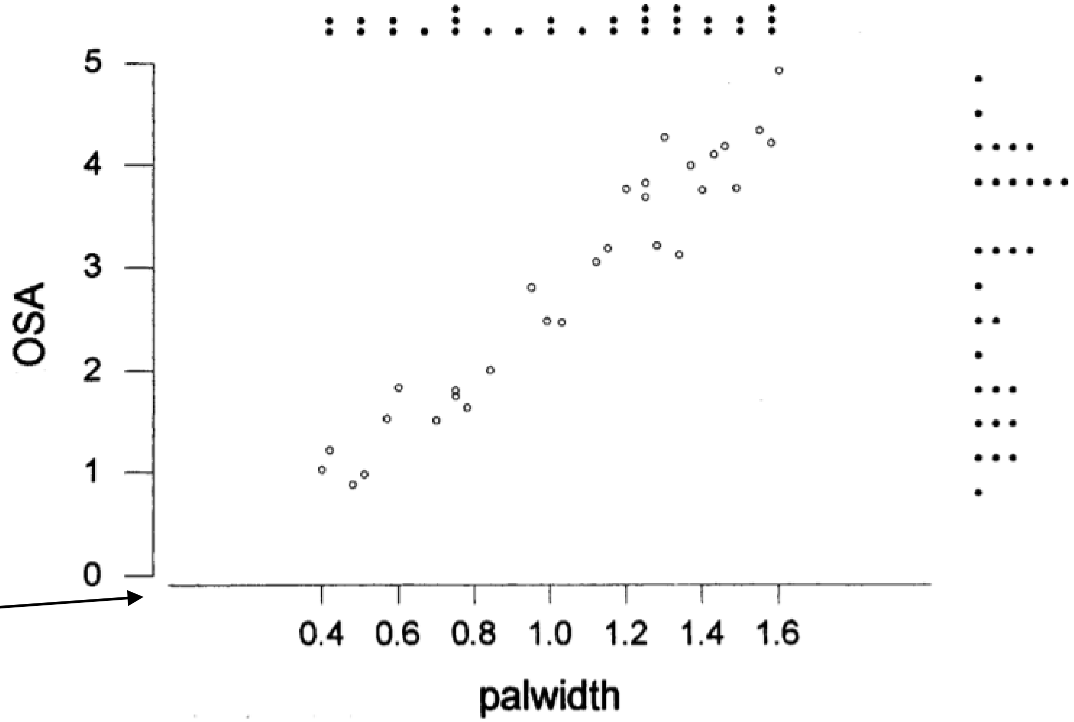


i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	.40	.42	.48	.51	.57	.60	.70	.75	.75	.78	.84	.95	.99	1.03	1.12
y_i	1.02	1.21	.88	.98	1.52	1.83	1.50	1.80	1.74	1.63	2.00	2.80	2.48	2.47	3.05

i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y_i	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99	3.75	4.10	4.18	3.77	4.34	4.21	4.92

EXAMPLE 12.1

- Some observations have the same x values, but different y values (for example $x_8 = x_9 = .75$, while $y_8 = 1.80, y_9 = 1.74$)
- There is a strong tendency for y to increase as x increases.
- It appears that the value of y could be predicted from x by finding a straight line that is reasonably close to the points in the plot.



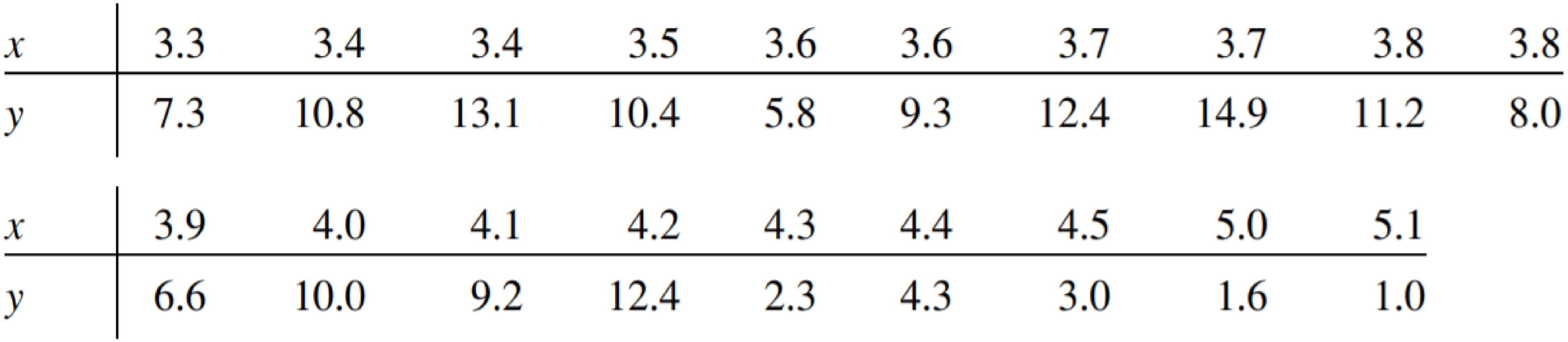
Scatter plot for the data in the previous table

The horizontal and vertical axes intersect at the point $(0,0)$

EXAMPLE 12.2

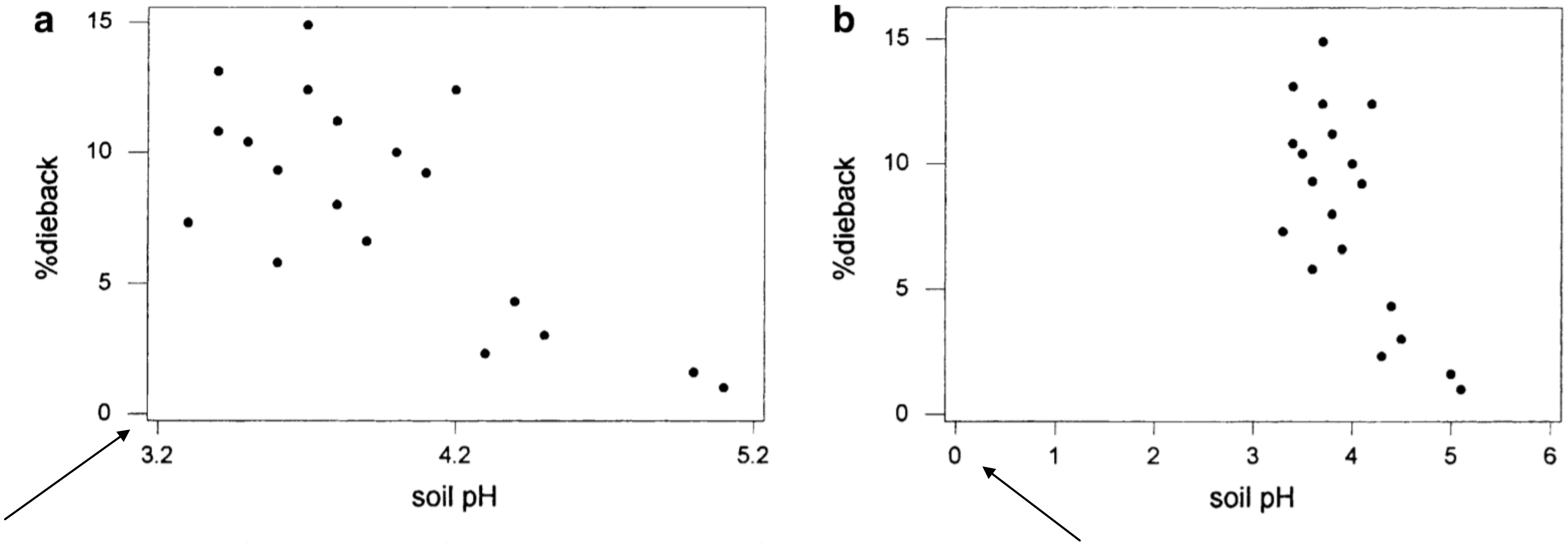
Another table shows observations (x, y) where

- $x =$ soil pH and
- $y =$ mean crown dieback (%) (indicator or growth retardation)



x	3.3	3.4	3.4	3.5	3.6	3.6	3.7	3.7	3.8	3.8	3.9	4.0	4.1	4.2	4.3	4.4	4.5	5.0	5.1
y	7.3	10.8	13.1	10.4	5.8	9.3	12.4	14.9	11.2	8.0	6.6	10.0	9.2	12.4	2.3	4.3	3.0	1.6	1.0

EXAMPLE 12.2



Automatic selection from the software MINITAB of the scale of the axes.

We chose manually the scales for the axes so that the intersection was at $(0,0)$.

- Large values of % dieback tend to be associated with low soil pH
- The two variables appear to be at least approximately linearly related

A LINEAR DETERMINISTIC MODEL

For a deterministic linear relationship $y = \beta_0 + \beta_1 x$. That is

- the slope coefficient β_1 guarantees change of magnitude β_1 in y when x increases / decreases by one unit
- the intercept coefficient β_0 corresponds to value of y when $x = 0$
- the graph of $y = \beta_0 + \beta_1 x$ is a straight line

Example: the line $y = 100 - 5x$ specifies an increase of -5 (i.e. a decrease of 5) for each one-unit increase in x, and the vertical intercept of the line is 100.

A LINEAR PROBABILISTIC MODEL

Definition - Simple linear Regression Model: There are parameters β_0, β_1 and σ^2 such that for any fixed value of the independent variable x, the dependent variable is related to x through the model equation

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where the random deviation ε is assumed to be normally distributed with mean value 0 and variance σ^2. That is, homoscedasticity means that variance remains constant, regardless of the fixed x value (and so does the mean value).

The n observed pairs $(x_1, y_1), \ldots, (x_n, y_n)$ are regarded as having been generated independently of each other from the model equation.

SIMPLE LINEAR REGRESSION MODEL

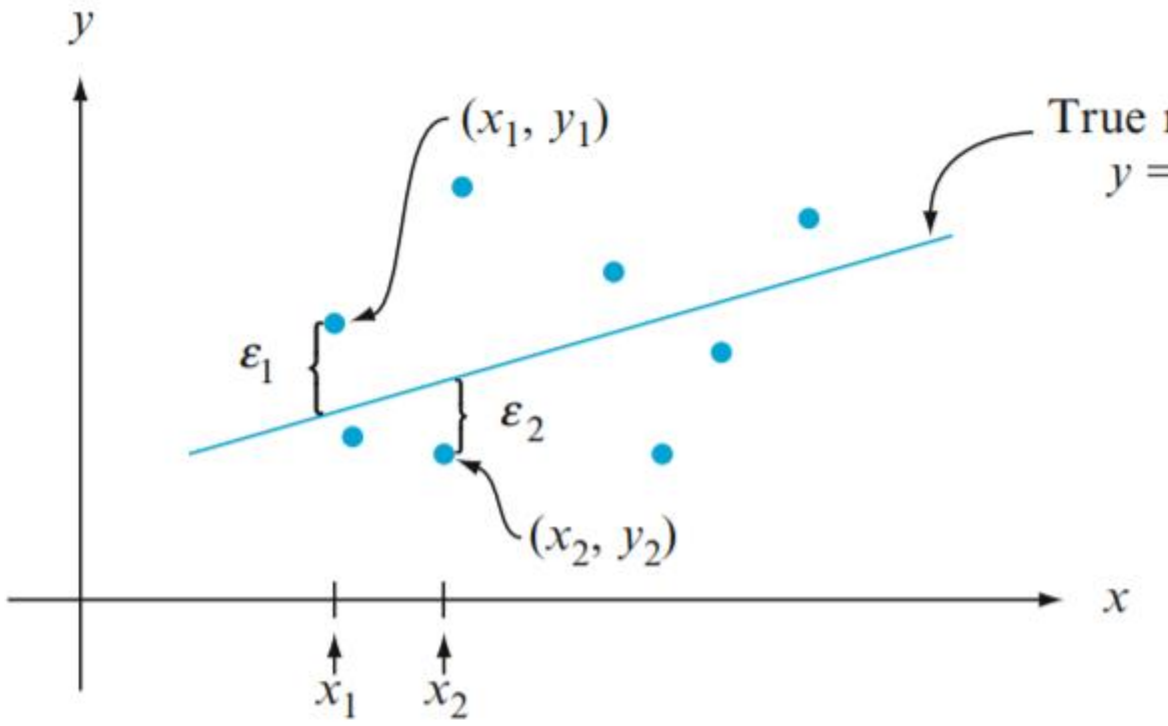


Figure 12.4 Points corresponding to observations from the simple linear regression model

INTERPRETATION OF THE COEFFICIENTS

- The parameters β_0 and β_1 are the coefficients of the population or true regression line $\beta_0 + \beta_1 x$.
- The slope parameter β_1 is now interpreted as the expected or true average increase in y associated with a one-unit increase in x.
- The variance parameter σ^2 controls the amount of variability in the data. If σ^2 is very close to 0, virtually all the pairs (x_i, y_i) in the sample should correspond to points quite close to the population regression line. But if σ^2greatly exceeds 0, a number of points in the scatter plot should fall far from the line.

NOTATION

Let x^* denote a particular value of the independent variable x, and

$$\mu_{y\cdot x^*} = E(y|x^*) = \text{the mean value of } y \text{ when } x = x^*$$

$$\sigma^2_{y\cdot x^*} = V(y|x^*) = \text{the variance of } y \text{ when } x = x^*$$

Example: let $x =$ applied stress (kg / mm^2)and $y =$ time to fracture (h). Then $\mu_{y\cdot 20}$ denotes the expected time to fracture when applied stress is 20 kg / mm^2. If we conceptualize an entire population of (x, y) pairs resulting from applying stress to specimens, then $\mu_{y\cdot 20}$ is the average of all values of the dependent variable for which $x = 20$.

$\sigma^2_{y\cdot 20}$describes the spread in the distribution of all y values for which applied stress is 20.

SOME CONSIDERATIONS

Consider replacing x in the model equation by the fixed value x^*. Then the only randomness results from ε. Hence, we can write

$$\begin{aligned} E(y|x^*) = \mu_{y\cdot x^*} &= E(\beta_0 + \beta_1 x^* + \varepsilon) \\ &= \beta_0 + \beta_1 x^* + E(\varepsilon) = \beta_0 + \beta_1 x^* \end{aligned}$$

$$V(y|x^*) = \sigma^2_{y\cdot x^*} = V(\beta_0 + \beta_1 x^* + \varepsilon)$$

$$= V(\beta_0 + \beta_1 x^*) + V(\varepsilon) = 0 + \sigma^2 = \sigma^2$$

- The first sequence of equalities says that the population regression line is the line of mean y values – the mean y value is a linear function of the predictor.
- The second sequence of equalities says the amount of variability in the distribution of y is the same at any x value. The constant variance property implies that points should spread out about the population regression line to the same extent throughout the range of x values

FIGURE 12.5

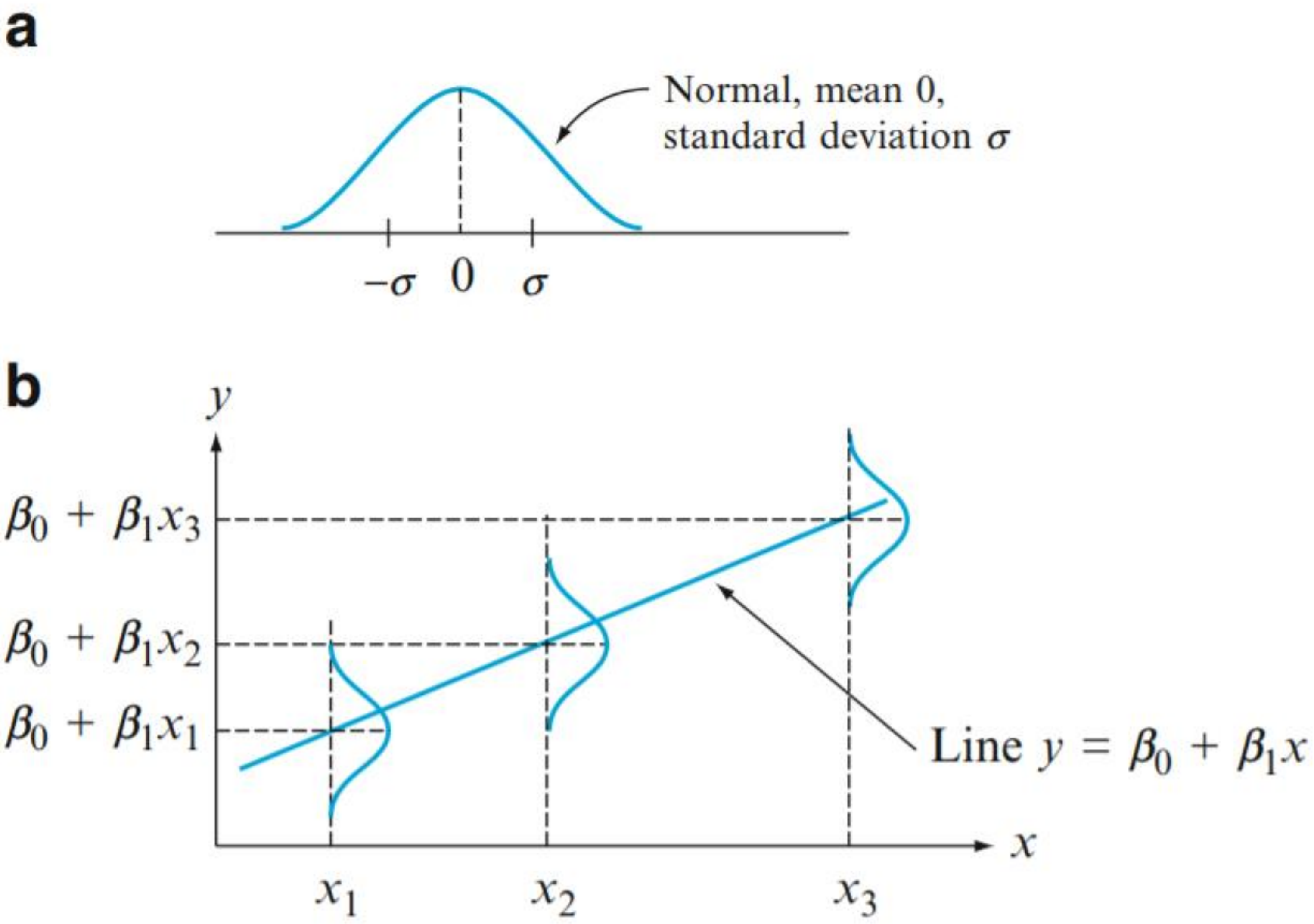


Figure 12.5 (a) Distribution of ε, (b) distribution of Y for different values of x

EXAMPLE 12.3

Suppose the relationship between applied stress x and time-to-failure y is described by the simple linear regression model with true regression line $y = 65 - 1.2x$ and $\sigma = 8$.

For $x = 20, y$ has mean value $\mu_{y \cdot 20} = 65 - 1.2(20) = 41.$ Hence

$$P(y > 50 \text{ when } x = 20) = P\left(Z > \frac{50 - 41}{8}\right) = 1 - \Phi(1.13) = .1292$$

When the applied stress is $25,$ then $\mu_{y \cdot 25} = 35.$ Hence

$$P(y > 50 \text{ when } x = 25) = P\left(Z > \frac{50 - 35}{8}\right) = 1 - \Phi(1.88) = .0301$$

FIGURE 12.6

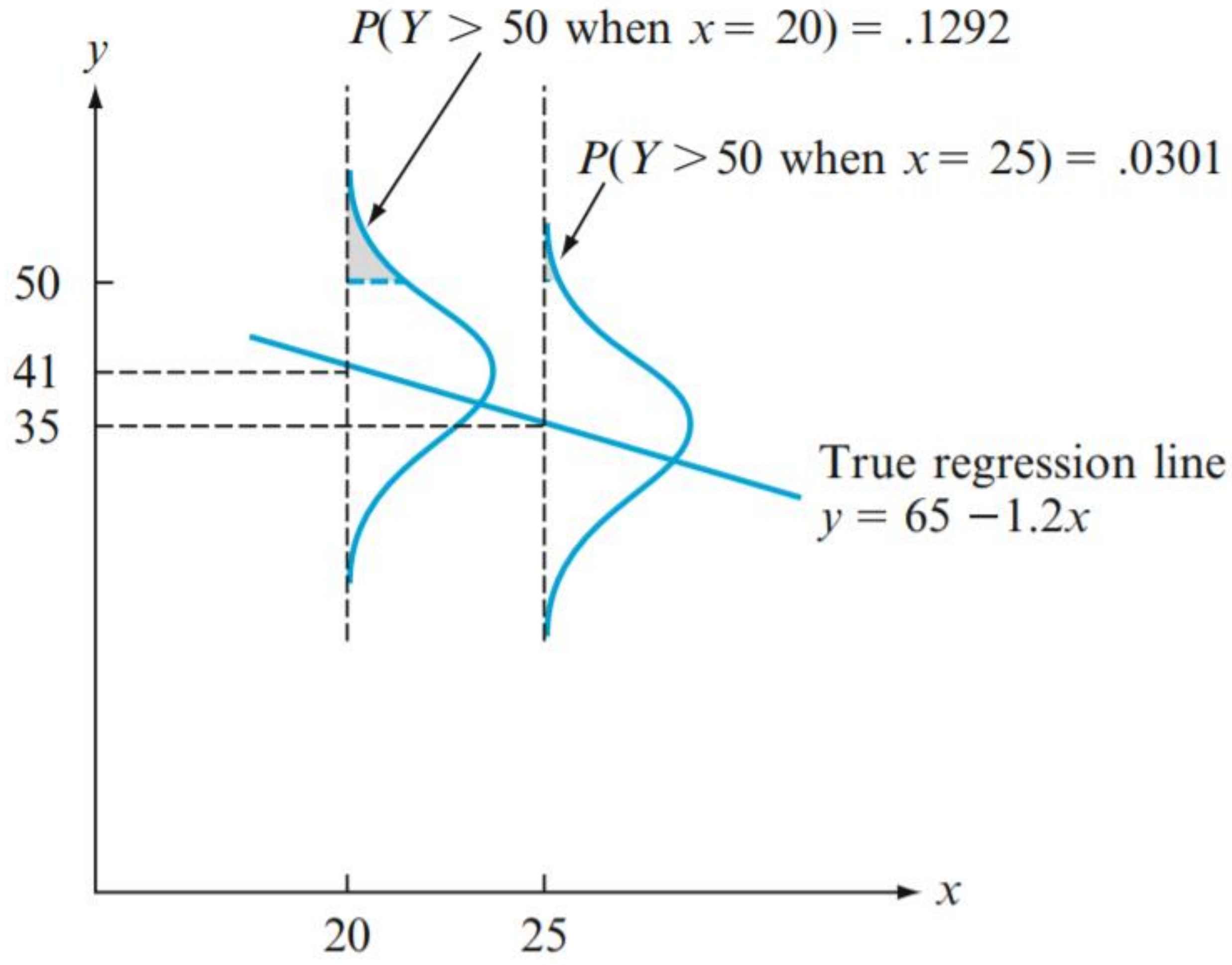

$P(Y > 50 \text{ when } x = 20) = .1292$
$P(Y > 50 \text{ when } x = 25) = .0301$
y
50
41
35
True regression line
$y = 65 - 1.2x$
20
25
x

Figure 12.6 Probabilities based on the simple linear regression model

EXAMPLE 12.3 - CONTINUED

Suppose that y_1 denotes an observation on time-to-failure made with $x = 25$ and y_2 denotes an independent observation made with $x = 24$. Then

- the difference $y_1 - y_2$ is normally distributed with mean value $E(y_1 - y_2) = \beta_1 = -1.2$, variance $V(y_1 - y_2) = \sigma^2 + \sigma^2 = 128$, and standard deviation $\sqrt{128} = 11.314$
- $\Rightarrow$ the probability that $y_1 > y_2$ is

$$P(y_1 - y_2 > 0) = P\left(Z > \frac{0 - (-1.2)}{11.314}\right) = P(Z > .11) = .4562$$

That it, even though we expected y to decrease when x increases by one unit, the probability is fairly high (but lower than .5) that the observed y at $x + 1$ will be larger than the observed y at x.

ESTIMATING MODEL PARAMETERS

SECTION 12.2 (DEVORE & BERK 2012)

ESTIMATION OF PARAMETERS

In this section we assume that x and y are related according to a simple linear regression model.

- The values of β_0, β_1 and σ^2 will almost never be known; what we can do is estimate them, using our set of n observations $(x_1, y_1), \dots, (x_n, y_n)$.
- y_i is the observed value of a random variable Y_i, where $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
- ε_i's are independent $\Rightarrow$ Y_i's are independent $\Rightarrow$ y_i's are independent.

FIGURE 12.8

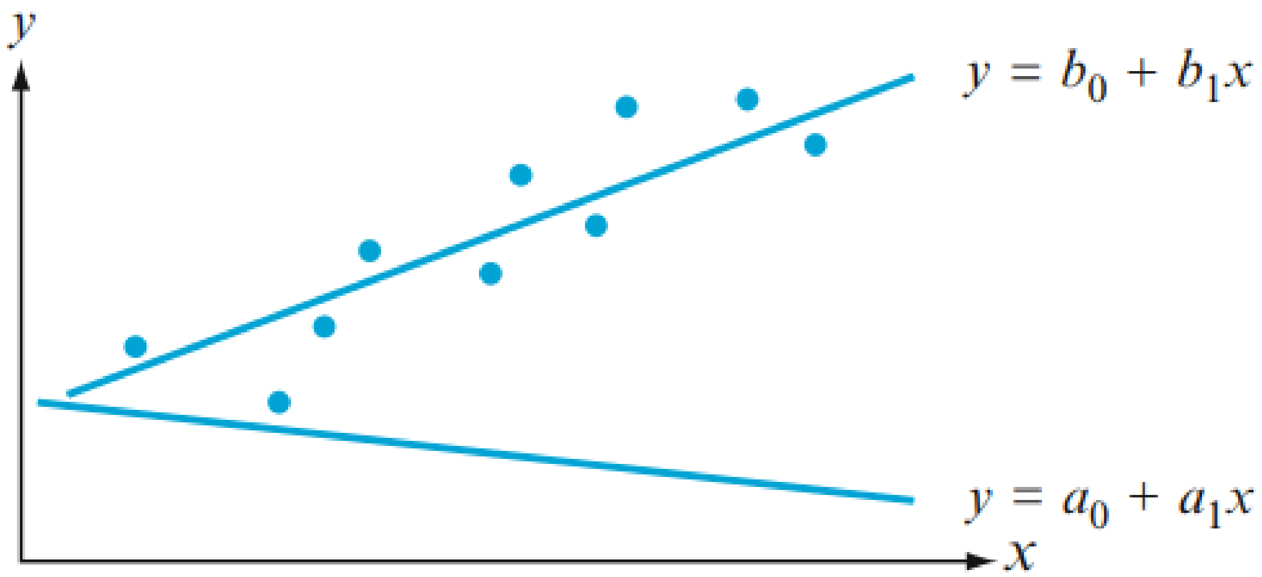


Figure 12.8 Two different estimates of the true regression line

Two candidates: $y = a_0 + a_1x$ and $y = b_0 + b_1x$. Intuitively,

- the first line is not a reasonable estimate of the true line $y = \beta_0 + \beta_1 x$, because the points are too far away.
- The second candidate is a better choice, because the points are scattered rather closely about this line.

PRINCIPLE OF LEAST SQUARES

An estimate of $y = \beta_0 + \beta_1 x$ should be a line that provides in some sense a best fit to the observed data points.

Principle of least squares (Gauss and Legendre, around 1800): a line provides a good fit to the data if the vertical distances (deviations) from the observed points to the line are small. The best fit-line is then the one having the smallest possible sum of squared deviations.

FIGURE 12.9

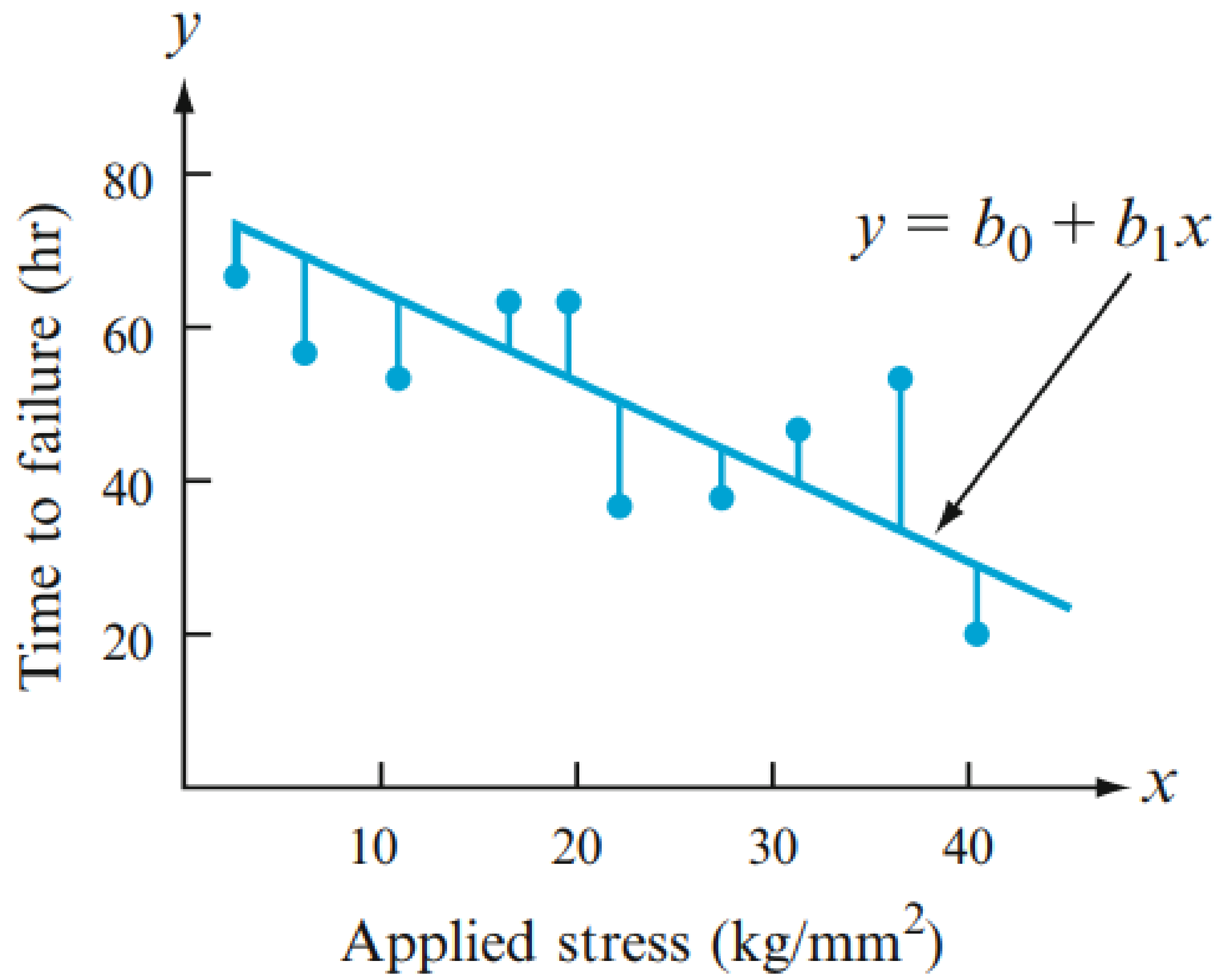


Figure 12.9 Deviations of observed data from line $y = b_0 + b_1x$

PRINCIPAL OF LEAST SQUARES – CONTINUED

- The vertical deviation of the point (x_i, y_i) from the $y = b_0 + b_1 x$ is
$$\text{height of point } - \text{height of line} = y_i - (b_0 + b_1 x_i)$$
- The sum of squared vertical deviations from the points $(x_1, y_1), \dots, (x_n, y_n)$ to the line then is
$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$
- The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the least squares estimates, are those values that minimize $f(b_0, b_1)$.
That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfy $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ for any b_0 and b_1. The estimated regression line or least squares line is then the line with equation
$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

MINIMIZING THE VALUES OF b_0 AND b_1

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero, and solving the equations

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

MINIMIZING THE VALUES OF b_0 AND b_1

We can simplify and rearrange the equations, gaining the system of equations, called the normal equations:

$$nb_0 + \left(\sum x_i\right) b_1 = \sum y_i$$

$$\left(\sum x_i\right) b_0 + \left(\sum x_i^2\right) b_1 = \sum x_i y_i$$

- The normal equations are linear in the two unknowns b_0 and b_1.
- Provided that at least two of the $x_i's$ are different, the least squares estimates are the unique solution to this system.

ESTIMATION OF b_0 AND b_1

- The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- Computing formulas for the numerator and denominator of b_1 are

$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \qquad \mathrm{S}_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

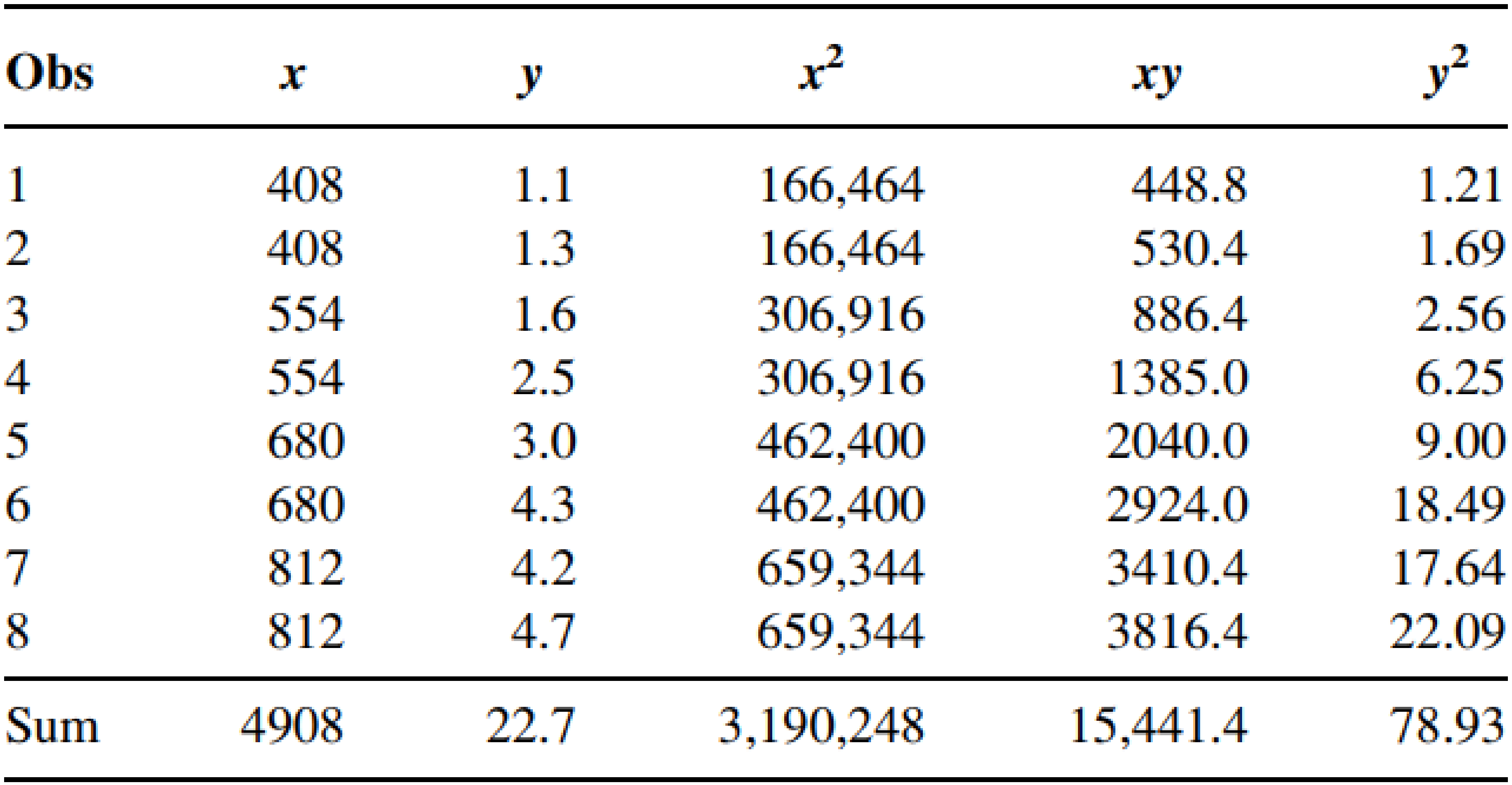
- The least squares estimate of the intercept b_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

Remark: $\hat{\beta}_0$ and $\hat{\beta}_1$ are also the maximum likelihood estimates because of the normality assumption.

EXAMPLE 12.5

x = atmospheric concentration of CO_2 in microliters per liter, and
y = mass in kilograms



Obs	x	y	x^2	xy	y^2
1	408	1.1	166,464	448.8	1.21
2	408	1.3	166,464	530.4	1.69
3	554	1.6	306,916	886.4	2.56
4	554	2.5	306,916	1385.0	6.25
5	680	3.0	462,400	2040.0	9.00
6	680	4.3	462,400	2924.0	18.49
7	812	4.2	659,344	3410.4	17.64
8	812	4.7	659,344	3816.4	22.09
Sum	4908	22.7	3,190,248	15,441.4	78.93

EXAMPLE 12.5

Hence

- $\bar{x} = \frac{4908}{8} = 613.5,$
- $\bar{y} = \frac{22.7}{8} = 2.838,$
- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{15{,}441.4 - (4908)(22.7)/8}{3{,}190.248 - (4908)^2/8} = \frac{1514.95}{179{,}190} \approx .00845$
- $\hat{\beta}_0 = 2.838 - (.00845)(613.5) = -2.349$

The equation of the estimated regression line (least squares line) then is

$$y = -2.35 + .00845x$$

EXAMPLE 12.5

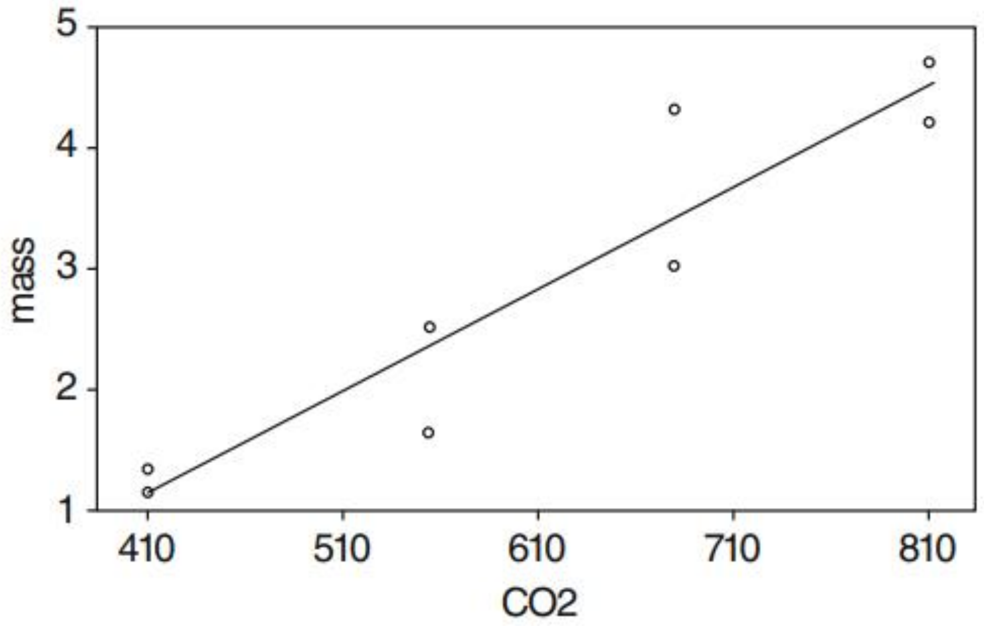


Figure 12.10 A scatter plot of the data in Example 12.5 with the least squares line superimposed, from R

Remark: the least squares line should not be used to make a prediction for an x value much beyond of the range data, such as $x = 250$ or $x = 1000$ in Example 12.5.

The danger of extrapolation is that the fitted relationship may not be valid for such values. Sometimes, this is obvious: for $x = 250, \hat{y} = -.235$ which is an impossible value for the mass. Other times it's not obvious.

ESTIMATING σ^2 AND σ

The parameter σ^2 determines the amount of variability inherent in the regression model.

- If σ^2 is large, (x_i, y_i)'s are quite spread out
- If σ^2 is small, (x_i, y_i)'s will tend to fall very close to the true line

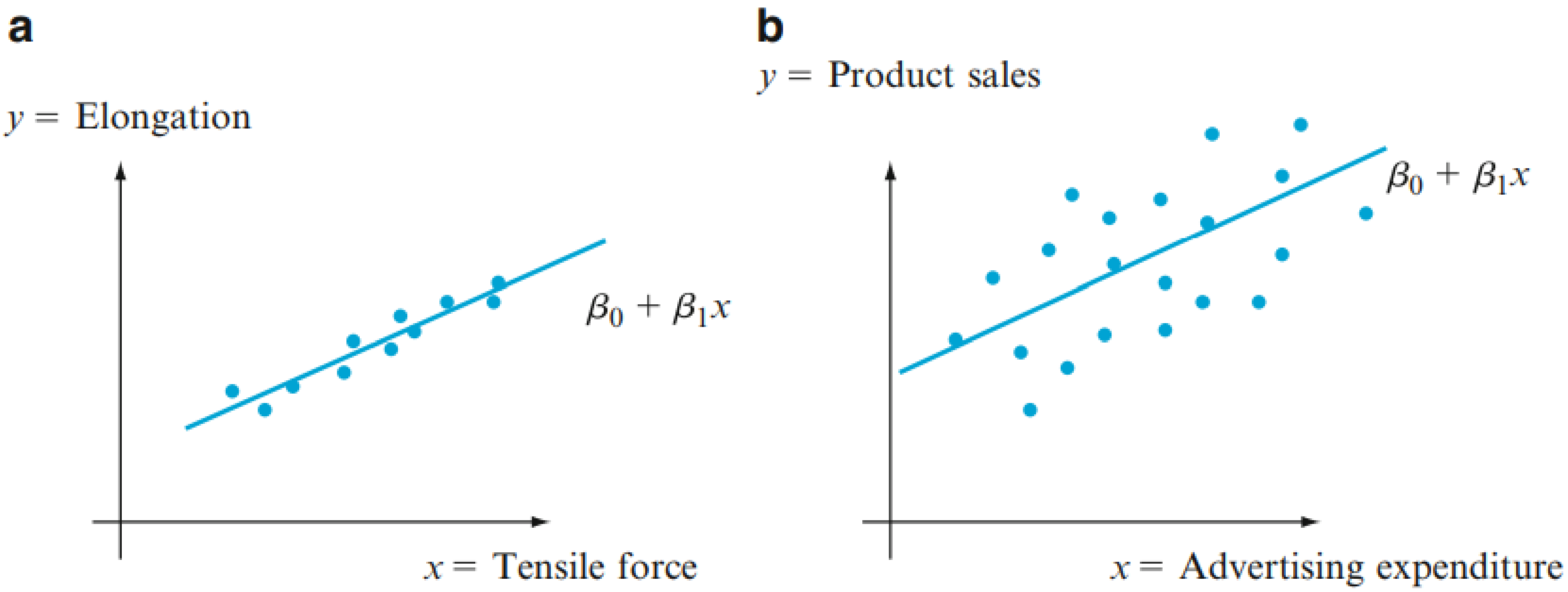


Figure 12.11 Typical sample for σ^2: (a) small; (b) large

SOME DEFINITIONS

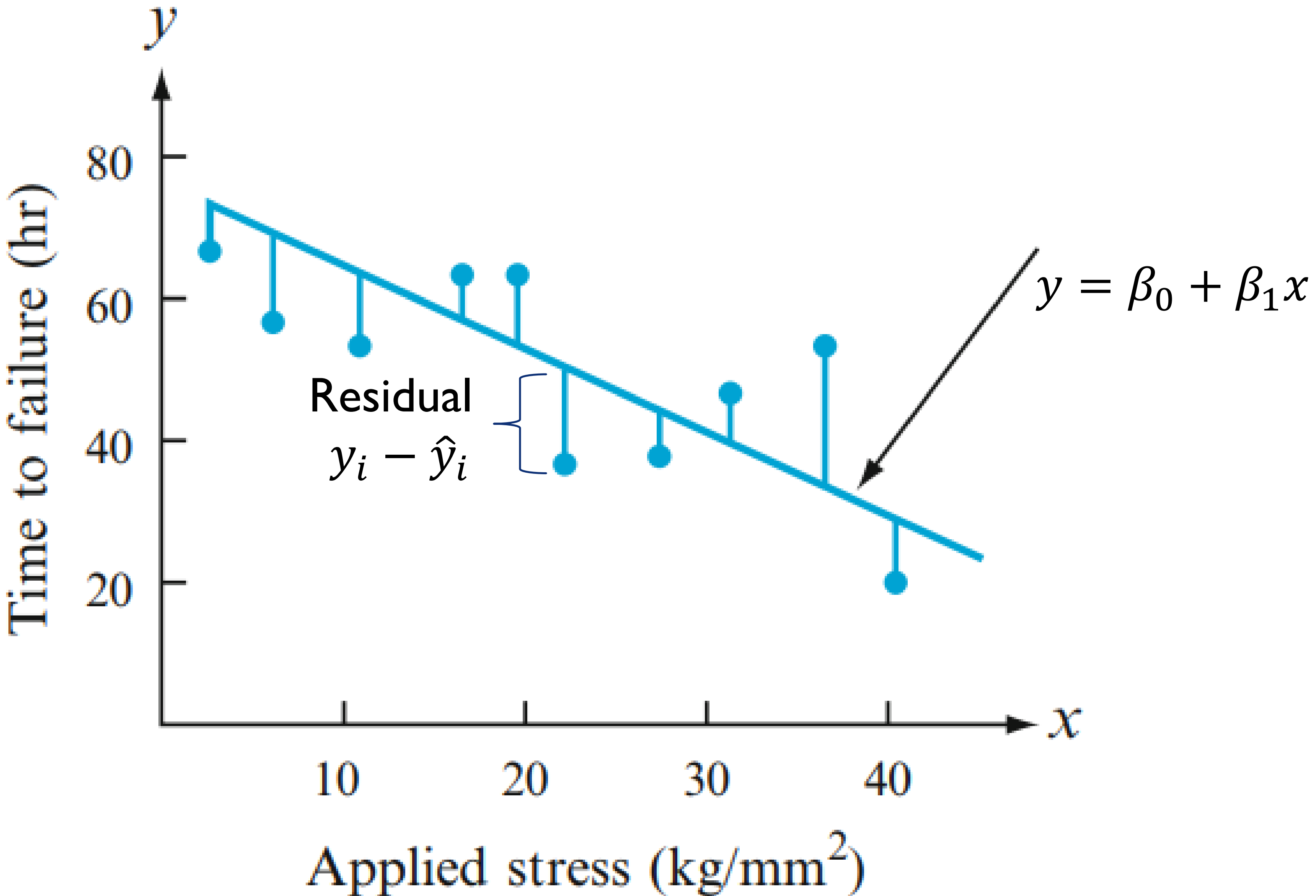
To estimate σ^2, we need to introduce some quantities.

Definitions: The fitted (or predicted) values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are obtained by successively substituting the x values $x_1, x_2, \ldots, x_n$ into the equation of the estimated regression line: $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2, \ldots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$. The residuals are the vertical deviations $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$ from the estimated line.

- If the residuals are small, then much of the variability in observed y values appears to be due to the linear relationship between x and y
- If the residuals are large, it suggests some inherent variability in y relative to the amount due to the linear relation.

Remark: When we estimate the regression line through the principle of least squares, the sum of the residuals should be zero.

RESIDUALS


y
80
60
40
20
Time to failure (hr)
Residual
$y_i - \hat{y}_i$
$y = \beta_0 + \beta_1 x$
x
10
20
30
40
Applied stress (kg/mm^2)

ERROR SUM OF SQUARES

The error sum of squares (or residual sum of squares), denoted by SSE, is

$$\mathrm{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

And the least square estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{\mathrm{SSE}}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

- We will continue to use the symbol s^2 for the estimated variance, but don't confuse it with our previous s^2!
- The divisor $n-2$ in s^2 is the number of degree of freedom (df) associated with the estimate. This is because, to obtain s^2, the parameters β_0 and β_1 must first be estimated, which results in a loss of 2 df.
- We can rewrite the formula for SSE as

$$\mathrm{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

SOME PLOTS

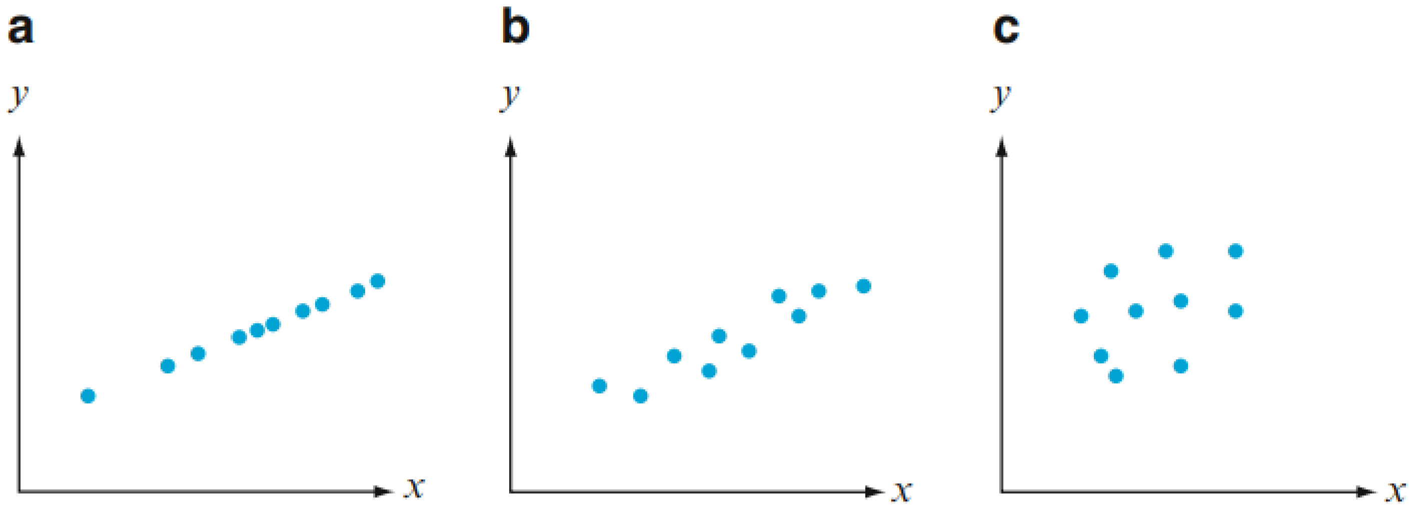


Figure 12.12 Explaining *y* variation: (a) all variation explained; (b) most variation explained; (c) little variation explained

- A.: all points fall exactly on a straight line ⇒ all sample variation in y can be attributed to the fact that x and y are linearly related
- B.: Not on a straight line, but the deviations from the least squares line are small
- C.: There are substantial variation about the least squares line relative to overall y variation.

SOME OBSERVATIONS

The error sum of squares SSE can be interpreted as a measure of how much variation in y is left unexplained by the model – that is, how much cannot be attributed to a linear relationship.

- In A., $\mathrm{SSE} = 0$, and there's no unexplained variation
- In B., The unexplained variation is small for the data
- In C., The unexplained variation is much larger than in B.

A quantitative measure of the total amount of variation in observed y values is given by the total sum of squares

$$\mathrm{SST} = \mathrm{S}_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2 / n$$

TOTAL SUM OF SQUARES

- Just as SSE is the sum of squared deviations about the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$, SST is the sum of squared deviations about the horizontal line at height $\bar{y}$.
- $\mathrm{SSE} < \mathrm{SST}$ unless the horizontal line is the least squares line; in fact, the sum of squared deviations about the least squared line is smaller than the sum of squared deviations about any other line, by definition!
- The ratio $\mathrm{SSE}/\mathrm{SST}$ is the proportion of total variation that cannot be explained by the simple linear regression model, and $1 - \mathrm{SSE}/\mathrm{SST} \in [0,1]$ is the proportion of observed y variation explained by the model.

COEFFICIENT OF DETERMINATION

The coefficient of determination, denoted by r^2, is given by

$$r^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \in [0,1]$$

- r^2 is the proportion by which the error sum of squares is reduced by the regression line compared to the horizontal line.
- The higher the value of r^2, the more successful is the simple linear regression model in explaining y variation
- If r^2 is small, an analyst may want to search for an alternative model.

ANALYSIS OF VARIANCE IDENTITY FOR REGRESSION

$$\text{SST} = \sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Of the two sums on the right, the first is $\text{SSE} = \sum (y_i - \hat{y}_i)^2$. The second is something new that we call the regression sum of squares, $\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$.

The analysis of variance identity for regression is $\text{SST} = \text{SSE} + \text{SSR}$.

The coefficient of determination can now be written as follows:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}.$$

LINEAR REGRESSION IN R

- Use the function `lm()`
- The sintax is the following:

```
response ~ predictor
```

or, if you have more than one predictor,

```
response ~ pred1 + pred2 + pred3 + ...
```

In the case of polynomial regression, define a new predictor `predsq=pred1`^{2}.

Then, write

```
response ~ pred1 + predsq
```

Use the function `summary()` to obtain p-values and standard errors

```
R script on screen
```