12 Bivariate Data Analysis: Regression and Correlation Methods

12.1 Introduction (P.187-191)

Many scientific investigations often involve two continuous variables and researchers are interested to know whether there is a (linear) relationship between the two variables. For example, a researcher wishes to investigate whether there is a relationship between the age and the blood pressure of people 50 years or older.

A Motivational Example: Suppose that a medical researcher wishes to investigate whether different dosages of a new drug affect the duration of relief from a particular allergic symptoms.

To study this, an experiment is conducted using a random sample of ten patients and the following observations are recorded:

Dosage (x)	3	3	4	5	6	6	7	8	8	9
Duration of relief (y)	9	5	12	9	14	16	22	18	24	22

Note: There are two variables in the problem and they are labelled by x and y for our convenience. The common sense suggests that the variable y depends on x and x can be independently selected and controlled by the researcher. However, the variable y cannot be controlled and is dependent on x.

Further Examples

Variable 1 (x)	Variable 2 (y)
Income	Expenditure
Temperature	Reaction time of a chemical
Alcohol consumption	Cholesterol level
Age of cancer patients	Length of survival

In these cases the variable y may depend on x. Therefore, x can be considered as an *independent* variable and the variable y as *dependent* (on x) variable. Statisticians are often interested to see whether there is a *linear* relationship (or linear association) between the two variables x and y. Such observations as collected as pairs on x and y (or (x, y)) are called bivariate data.

In the previous bivariate example, '$x_1 = 3$' corresponds to '$y_1 = 9$' and there are $n = 10$ pairs.

A Notation

In general, there are n pairs of such bivariate data given by

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

12.2 Graphical Representation of Bivariate Data

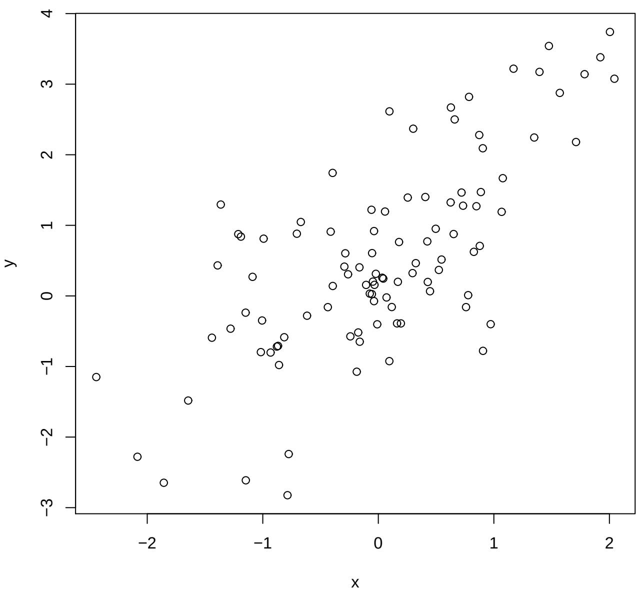
A standard plot on a grid paper of y (y-axis) against x (x-axis) gives a very good indication of the behaviour of data. This coordinate plot of the points (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ on a grid paper is called a *scatter plot.*

Note: The first step of the analysis of bivariate data is to plot the observed pairs, (x, y) and obtain a scatter plot. This plot gives a clear picture of a possible relationship between x and y.

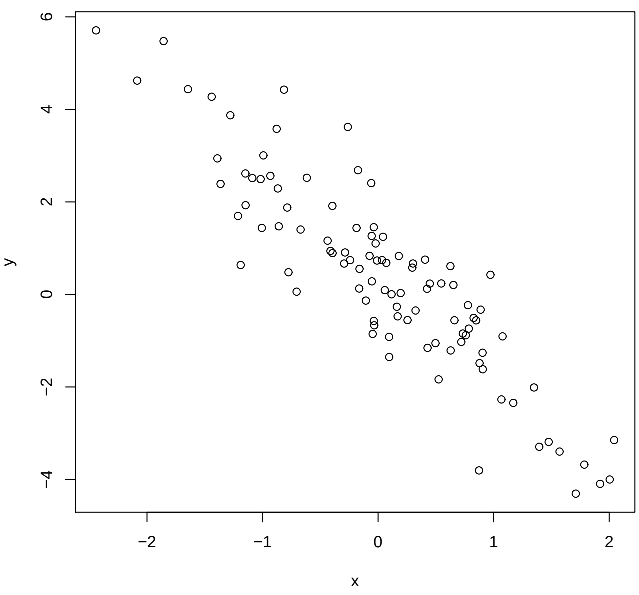
Now we look at a number of other possible scatter plots we may observe in data analysis.

Some typical scatter plots

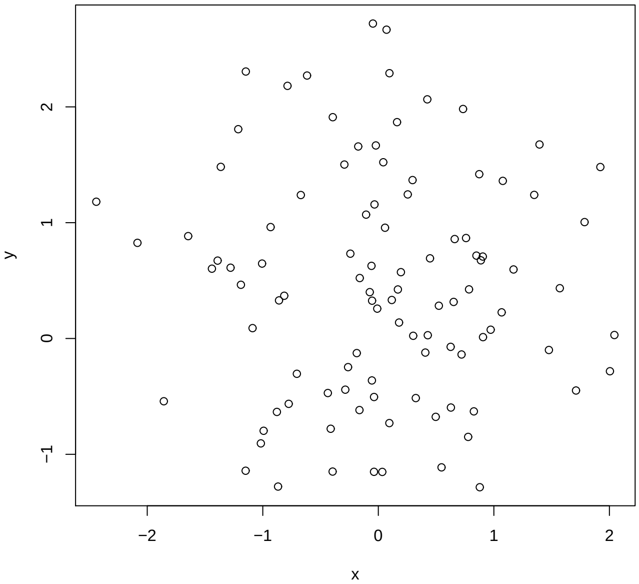
(a) Positive slope (or upward trend)



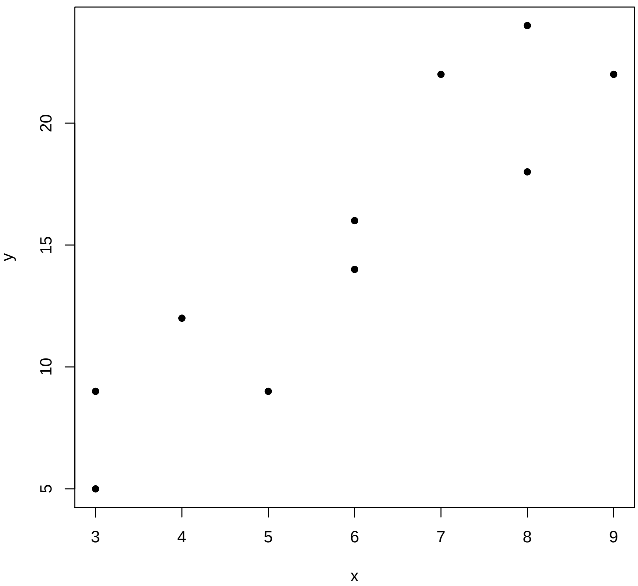
(b) Negative slope (or downward trend)



(c) Random scatter (or no apparant pattern)



Example: A scatter plot of the above bivariate data is:



This scatter plot shows that the points seem to cluster around a straight line. It tells us that there is a possible (approximate) liner relationship between x and y.

Then we want to further investigate:

Is the linear relationship between the variables x and y clear and significant ?

We need to find a measure to investigate the strength of a possible linear relationship between two variables x and y. This measure is known as the *correlation coefficient* (or Pearson's correlation coefficient) between x and y.

12.3 The correlation coefficient (P.215-218)

Recall the following:

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right),$$

$$L_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

and

$$L_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

(L_{xx} is used in the calculation of s^2). Now the correlation coefficient between x and y is denoted by r and is given by the ratio

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}.$$

Important Property: It can be shown that the values of r lie between -1 and 1, or

$$-1 \leq r \leq 1$$

for all such calculations. When r is large and close to +1 or -1 (ideally $0.5 < r < 1$ or $-1 < r < -0.5$), we conclude that there is a linear relationship between x and y.

Note: All these formulae are available on the formulae sheet.

Read: P.222-223.

Example: Find the correlation coefficient, r, between x and y for the data in the dosage example.

Dosage (x)	3	3	4	5	6	6	7	8	8	9
Duration of relief (y)	9	5	12	9	14	16	22	18	24	22

Solution:

It is easy to obtain that:

$$\sum_{i=1}^{10} x_i = 59, \ \sum_{i=1}^{10} x_i^2 = 389,$$

$$\sum_{i=1}^{10} y_i = 151, \ \sum_{i=1}^{10} y_i^2 = 2651, \sum_{i=1}^{10} x_i y_i = 1003.$$

Now,

$$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \underline{\hspace{4cm}},$$

$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = \underline{\hspace{4cm}}$$

and

$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = \underline{\hspace{5cm}}.$$

Hence

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \underline{\hspace{4.5cm}}.$$

This correlation coefficient is close to one. From the scatter plot, we've noticed that the points are very close to a straight line with a positive slope.

In general:

- It can be seen that when the points lie perfectly on a straight a line, then r is either $+1$ (when the slope is positive) or -1 (when the slope is negative).
- If the points are close to a straight line (not perfectly), then r is close to either 1 or -1 (but not exactly equal).

Illustrative Examples:

Find the correlation coefficient between x and y for the following two data sets:

1.

x	0	1	2	3
y	3	8	13	18

Note that here $y = 5x + 3$.

2.

x	0	1	2	3
y	8	6	4	2

Note that here $y = -2x + 8$.

Solution: We have

1. $\sum_i x_i = 6, \sum_i x_i^2 = 14, \sum_i y_i = 42, \sum_i y_i^2 = 566, \sum_i x_i y_i = 88$

$$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \underline{\hspace{4cm}};$$

$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = \underline{\hspace{4cm}};$$

$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = \underline{\hspace{4cm}}$$

and

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \underline{\hspace{3cm}}.$$

2. $\sum_i x_i = 6, \sum_i x_i^2 = 14, \sum_i y_i = 20, \sum_i y_i^2 = 120, \sum_i x_i y_i = 20$

$$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \underline{14 - 6^2/4 = 5};$$

$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = \underline{120 - 20^2/4 = 20};$$

$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = \underline{20 - 6(20)/4 = -10}$$

and

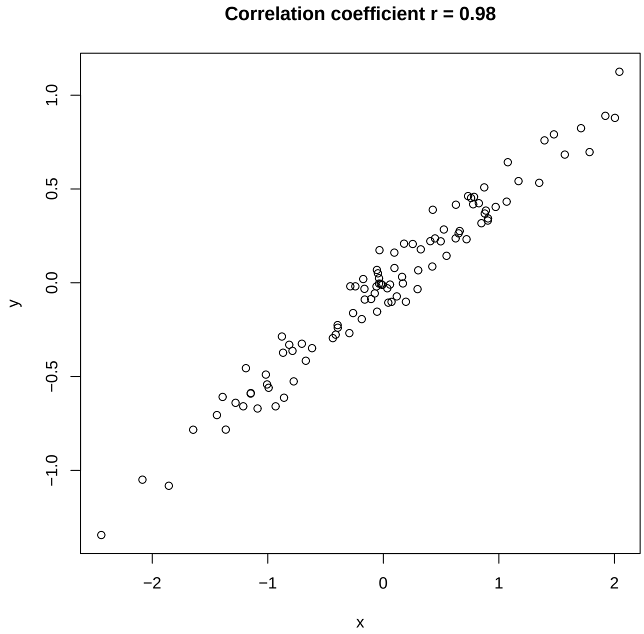
$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \underline{\hspace{3cm}}.$$

Exercise: Draw a scatter plot for each of the examples.

Notes:

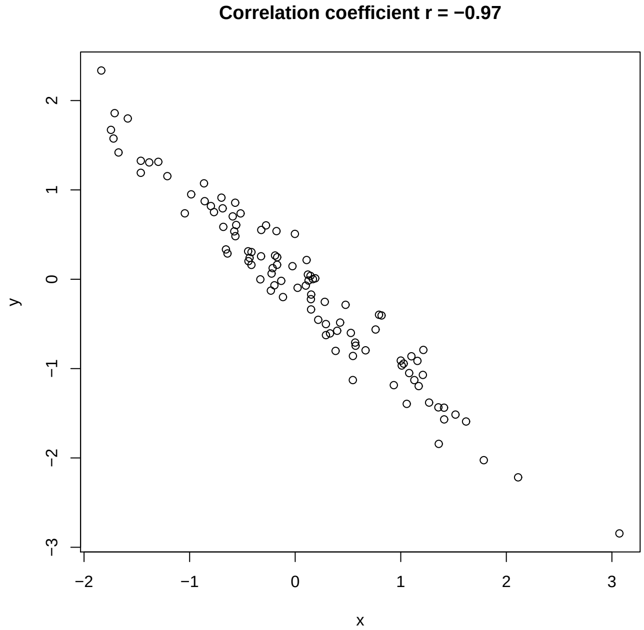
1. Using a similar argument, it can be seen that when r is close to +1 or -1, the points are very close to a straight line as given below:

Case I: $r \approx +1$**:**



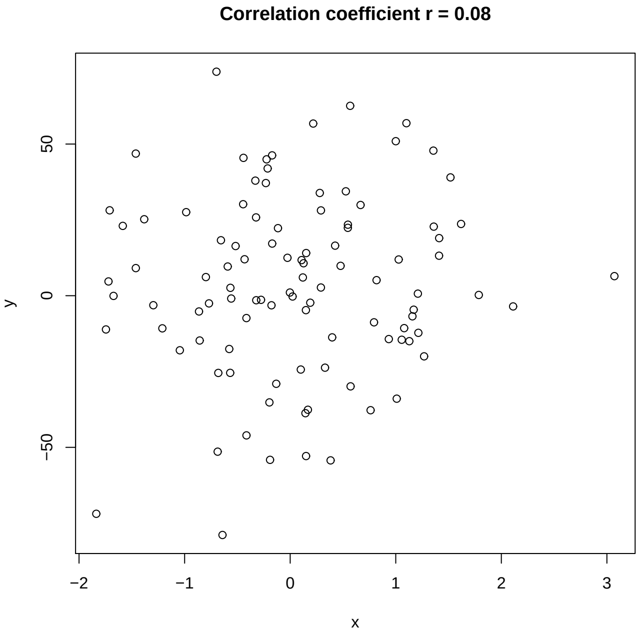
In this case r is close to $+1$ and we say that there is a strong positive linear relationship between x and y.

Case II: $r \approx -1$**:**



In this case r is close to -1 and we say that there is a strong negative linear relationship between x and y.

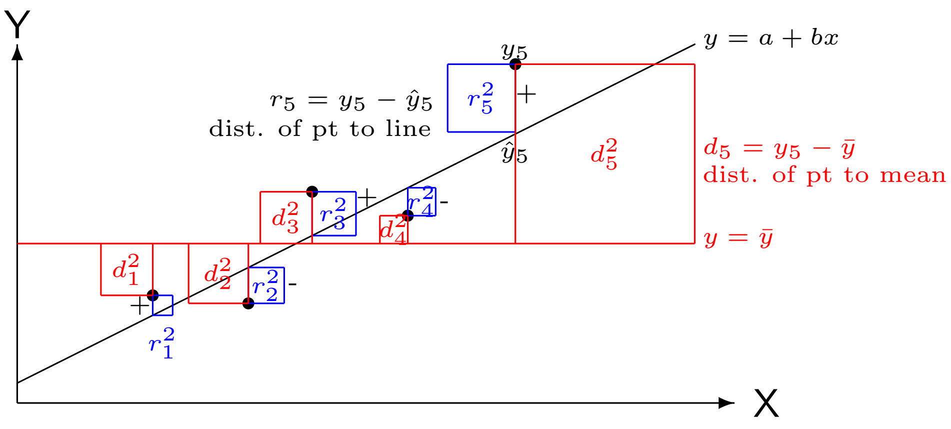
2. When r is close to zero (0), the points are far away from a straight line as shown below:



In this case we say that there is a very weak (or no) linear relationship between x and y since the points are randomly scattered.

A Linear Relationship between x and y

When r is close to +1 or -1, there is a possible *linear relationship* between x and y as described by $y = a + bx$. In such cases, it can be shown that r^2 gives the *proportion of variability explained by a linear relationship between x and y.*



Residuals

In other words, it is 'one minus the proportion of variation not explained by the model' as given below:

$$r^2 = 1 - \frac{\sum_i r_i^2}{\sum_i d_i^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Hence it measures an overall fit of the regression model.

Note: Linear relationship between x and y should be investigated only when the variables are meaningful. For example, you may expect a large positive correlation between

x = number of TV sets per capita and
y = average life expectancy.

However, these variables are not directly related to each other and the correlation coefficient is meaningless.

Read: examples 11.22 (P.215); 11.23 and 11.24 (P.216) and 11.25 (P.217).

Example: Soil temperature (x_i, in ^{0}C) and germination interval (y_i, in days) were observed for winter wheat at 10 localities:

x_i	12.5	5.0	3.0	5.0	6.5	6.0	4.0	7.0	5.5	4.0
y_i	10	26	41	29	27	19	18	20	28	33

Find the correlation coefficient between x and y.

Solution: We have $n = 10$. It is easy to obtain:

$$\sum_{i=1}^{n} x_i = 58.5, \sum_{i=1}^{n} y_i = 251, \sum_{i=1}^{n} x_i^2 = 404.75,$$

$$\sum_{i=1}^{n} y_i^2 = 6985, \sum_{i=1}^{n} x_i y_i = 1310.5$$

$$L_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = 404.75 - \frac{58.5^2}{10} = 62.525$$

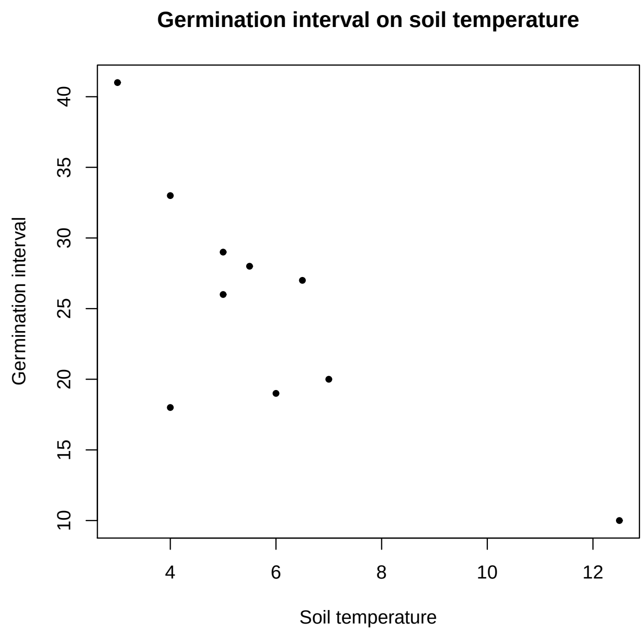
$$L_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = 6985 - \frac{251^2}{10} = 684.9$$

$$L_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = 1310.5 - \frac{(58.5)(251)}{10} = -157.85$$

Now the sample correlation coefficient is

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \underline{\hspace{6cm}}.$$

The scatter plot is:



Notes:

1. The two variables are negatively correlated and r is fairly close to -1.

2. $r^2 =$ ________________.

Therefore, a linear relationship between x (independent variable) and y (dependent variable) will explain about 58% of the variability in y.

Now we consider the fitting of a straight line for a given set of such data with large $|r|$ values. This problem is known as the *Simple Linear Regression.*

12.4 Simple Linear Regression (P.191-198)

Suppose that we have n pairs of observations,

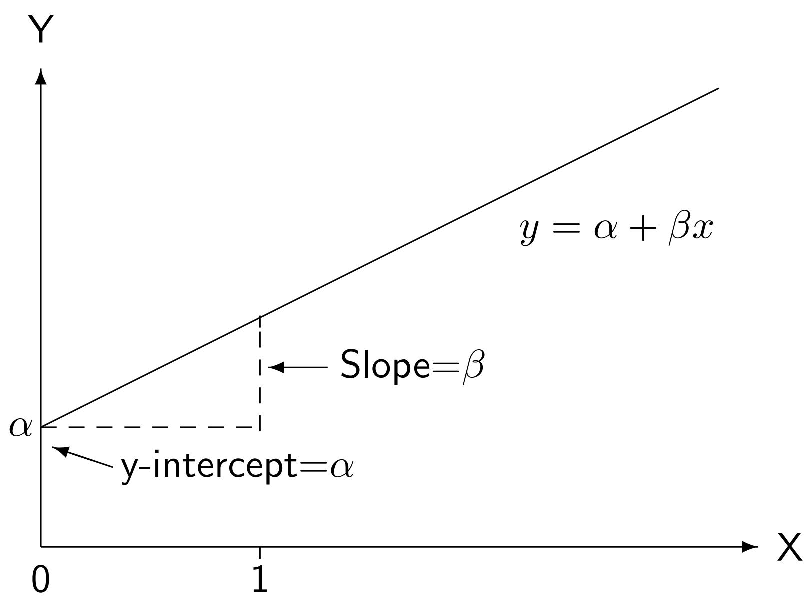
$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

and the scatter plot indicates that there is a possible straight line fit for these data. Now we wish to find (or fit) a straight line to these points.

To find a straight line we need the slope (or the gradient) and the y-intercept. Recall the following from your year 8-12 work:

- Equation of a straight line is $y = \alpha + \beta x$.
- β is the slope and α is the y intercept.

Diagram



The regression line

12.4.1 Estimates of α and β

Let a and b be the estimates of the true parameters α (y-intercept) and β (slope). It is known that

$$b = \frac{L_{xy}}{L_{xx}}$$

is an estimator of β and

$$a = \bar{y} - b\bar{x}$$

is an estimator of α.

Regression line: The estimated regression (or fitted) line is given by

$$\hat{y} = a + bx$$

Interpretation of the regression slope:

For every additional unit increase in x, the average change in y is b units.

Note:

1. a and b are called the *least squares* estimators, because they minimise the sum of squared distances between observed y_i's and estimated $\hat{y}_i$'s.

2. These formulae are also available on the formulae sheet.

Read: Examples 11.8 to 11.11 on P.194-195.

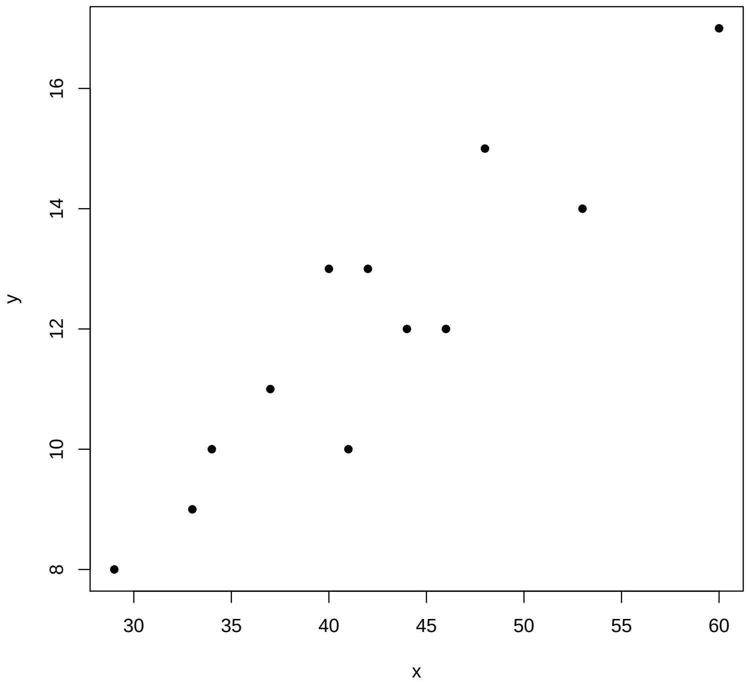
Example: The dose x (in gms) and concentration in urine, y (in mg/gm) of a fluid after intravenous administration are measured for 12 people:

x:	46	53	37	42	34	29	60	44	41	48	33	40
y:	12	14	11	13	10	8	17	12	10	15	9	13

(i) Prepare a scatter plot for these data. What do you notice?
(ii) Find the correlation coefficient between x and y. What proportion of variability is explained by a simple linear regression model?
(iii) Fit the least squares regression for these data.
(iv) Estimate the urine concentration when the dose is 50g.

Solution:

(i)



The scatter plot shows a positive relationship between x and y.

(ii) We have $n = 12$ and

$$\sum_i x_i = 507 \Rightarrow \overline{x} = \frac{507}{12} = 42.25$$

$$\sum_i y_i = 144 \Rightarrow \overline{y} = \frac{144}{12} = 12$$

$$\sum_i x_i^2 = 22265, \sum_i y_i^2 = 1802, \sum_i x_i y_i = 6314$$

$$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = \underline{22265 - \frac{507^2}{12} = 844.25}$$

$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = \underline{1802 - \frac{144^2}{12} = 74}$$

$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = \underline{6314 - \frac{507(144)}{12} = 230}$$

Correlation coefficient:

$$r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} = \underline{\qquad\qquad\qquad\qquad}.$$

$$r^2 = \underline{\qquad\qquad\qquad}.$$

and therefore a simple linear regression model explains about 85% of the variability in y.

(iii) Regression line:

$$b = \frac{L_{xy}}{L_{xx}} = \underline{\qquad\qquad\qquad},$$

$$a = \bar{y} - b\bar{x} = \rule{7cm}{0.4pt}.$$

Fitted line: $\rule{9cm}{0.4pt}$

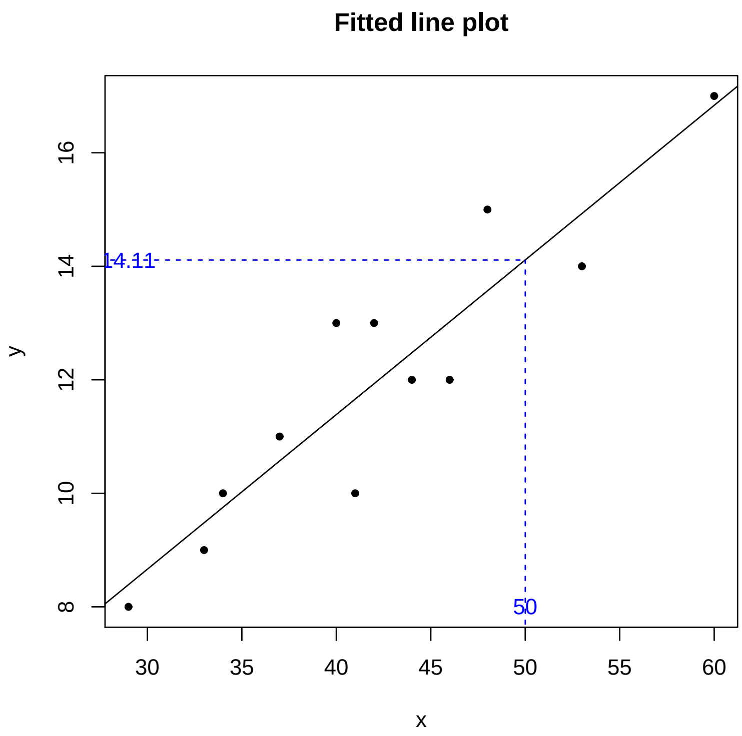
or $\rule{4cm}{0.4pt}$.

Hence for each additionally administered gram of fluid the concentration in urine increases by 0.272 mg/gm.

(iv) When the dose is 50 g, put $x_0 = 50$. We have

$$\hat{y}_0 = a + bx_0 = \rule{5.5cm}{0.4pt}.$$

The following graph gives the fitted line plot with fitted value $\hat{y}$ for this example.



Note that $\hat{y}_0$ provides a good prediction only when x_0 lies within the range of x, i.e. 33 to 60 in this example, or at least close to the range of x.

12.5 Assessing the Goodness of Fit of a Regression Line

The value of r^2 is often reported as a measure of the overall goodness of fit of the regression model. In other words, the closer the value of r^2 is to 1, the better is the model fit. Notice that perfect straight lines with $r = -1$ or 1, result in perfect $r^2 = 1$.

12.5.1 Residuals

The regression line is a mathematical model for the overall pattern of a linear relationship between x and y. The deviations from this overall pattern are called *residuals*.

Definition: A residual is the difference between an observed value and its corresponding predicted value:

$$\text{Residual} = y - \hat{y}$$

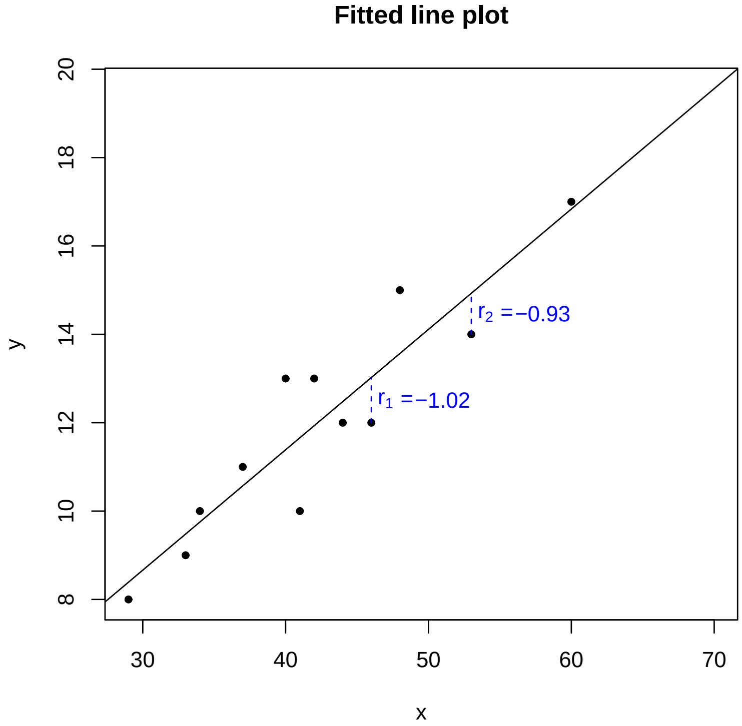
Example: Compute the residuals from the dose and concentration data discussed before.

Solution: The fitted line for this data is $\hat{y} = 0.490 + 0.272\ x$. We first find $\hat{y}$ for each value of x. For example, at $x_1 = 46$,

$$\hat{y}_1 = a + bx_1 = \underline{\hspace{6cm}}.$$

Therefore the residual (or error) at $x_1 = 46$ is
$r_1 = y_1 - \hat{y}_1 = \underline{\hspace{5cm}}$.

Similarly, at $x_2 = 53$, the residual is
$r_2 = y_2 - \hat{y}_2 = \underline{\hspace{5cm}}$



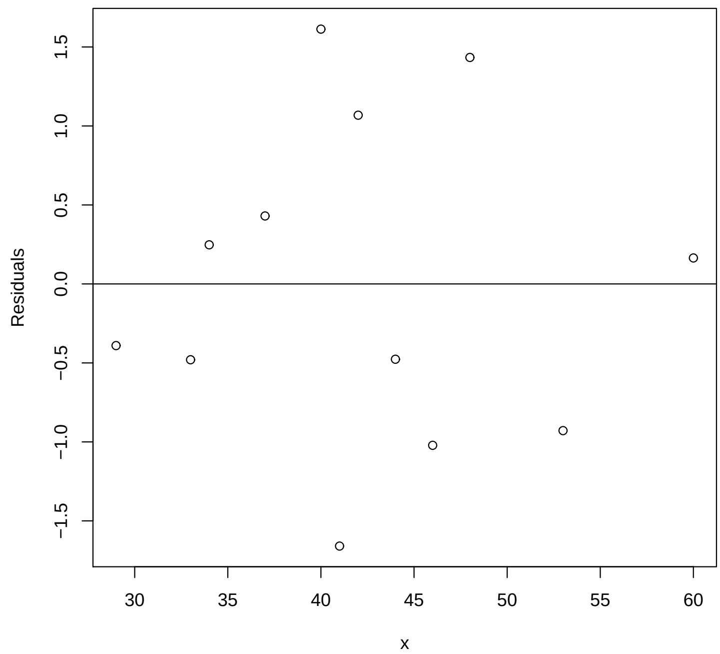
Exercise: Find the remaining 10 residuals for the above regression.

Ans: 0.446, 1.086, 0.269, -0.378, 0.190, -0.458, -1.642, 1.454, -0.466, 1.630

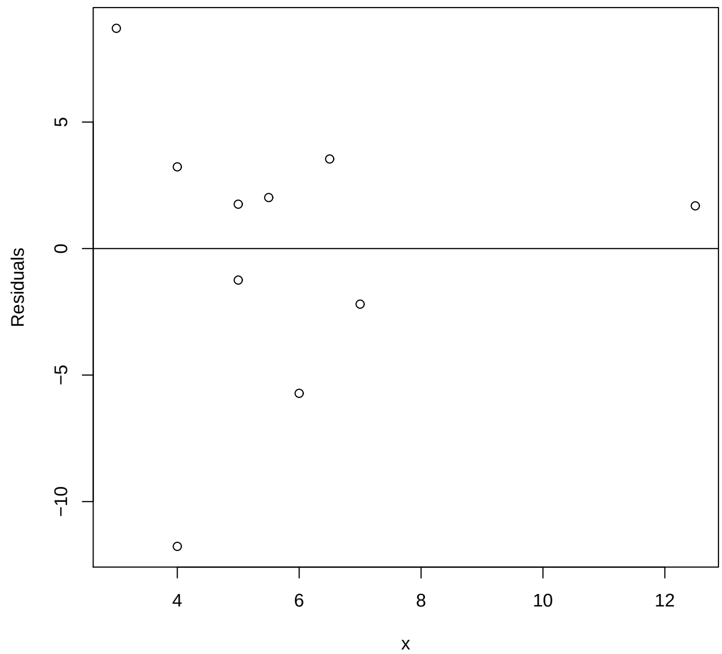
12.5.2 Residual plots

Scatter plots of the residuals e vs. predicted $\hat{y}$ or vs. the x variable can be used to assess the goodness of fit of the regression line.

Example 1: Dose vs. concentration



Example 2: Temperature vs. germination time



Note: The fit is good if the dots are a random scatter around zero.