Scatterplots and Correlation

Al Nosedal
University of Toronto

Summer 2019

My momma always said: "Life was like a box of chocolates. You never know what you're gonna get."

Forrest Gump.

Definitions

- A *response variable* measures an outcome of a study.
- An *explanatory variable* may explain or influence changes in a response variable.

Coral reefs

How sensitive to changes in water temperature are coral reefs? To find out, scientists examined data on sea surface temperatures and coral growth per year at locations in the Red Sea. What are the explanatory and response variables? Are they categorical or quantitative?

Solution

Sea-surface temperature is the explanatory variable; coral growth is the response variable. Both are quantitative.

Scatterplot

A *scatterplot* shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.
Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

Do heavier people burn more energy?

Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. We have data on the lean body mass and resting metabolic rate for 12 women who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours.

Data

Mass	Rate	Mass	Rate
36.1	995	40.3	1189
54.6	1425	33.1	913
48.5	1396	42.4	1124
42.0	1418	34.5	1052
50.6	1502	51.1	1347
42.0	1256	41.2	1204

The researchers believe that lean body mass is an important influence on metabolic rate. Make a scatterplot to examine this belief.

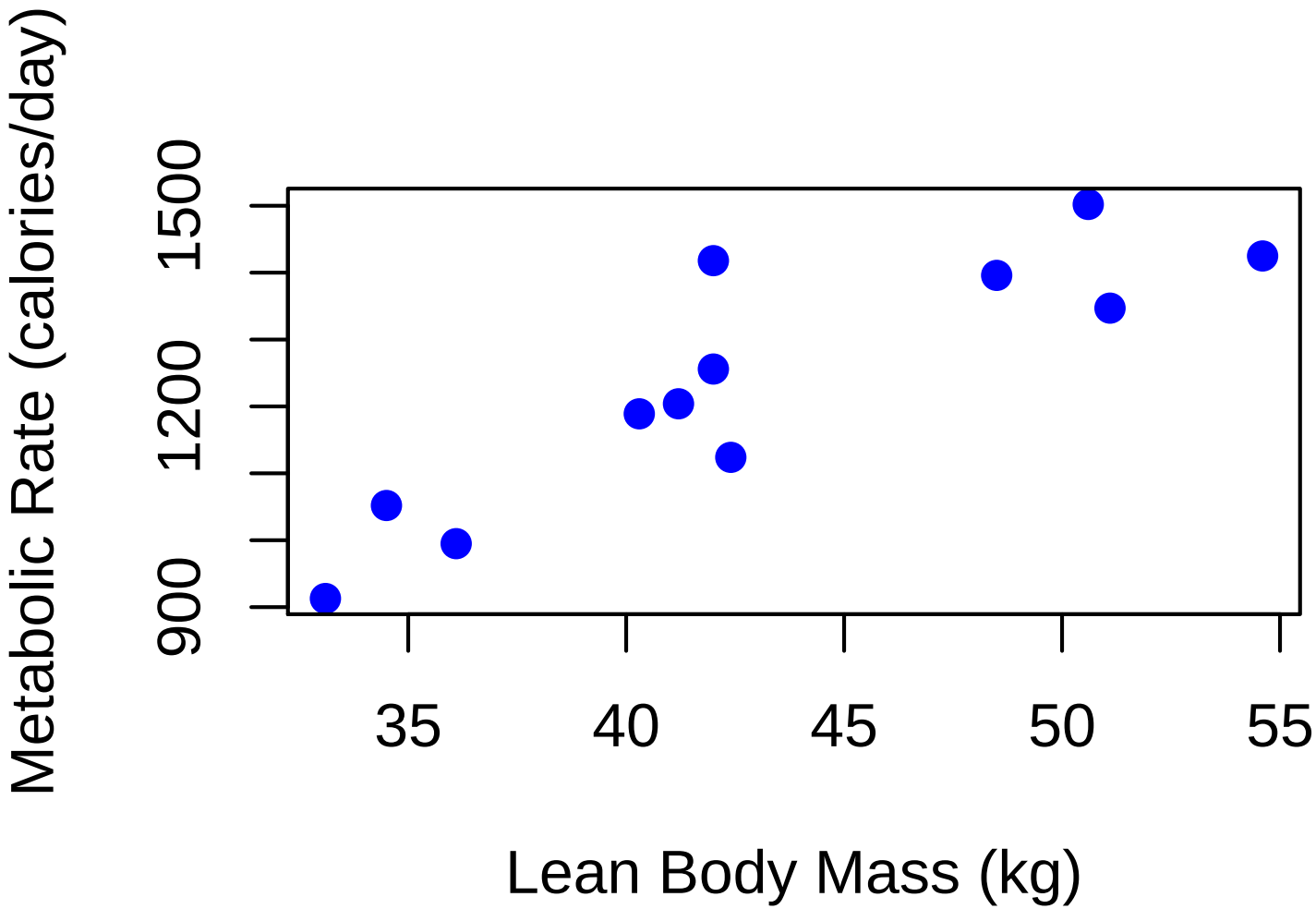
R code

```
# Step 1. Entering data;

mass=c(36.1, 54.6, 48.5, 42.0, 50.6, 42.0,
40.3, 33.1, 42.4,  34.5, 51.1, 41.2);

rate=c(995, 1425, 1396, 1418, 1502, 1256,
1189, 913, 1124, 1052, 1347, 1204);
```

R code

```
# Step 2. Making scatterplot;

plot(mass, rate ,pch=19,col="blue",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)");
```

Scatterplot



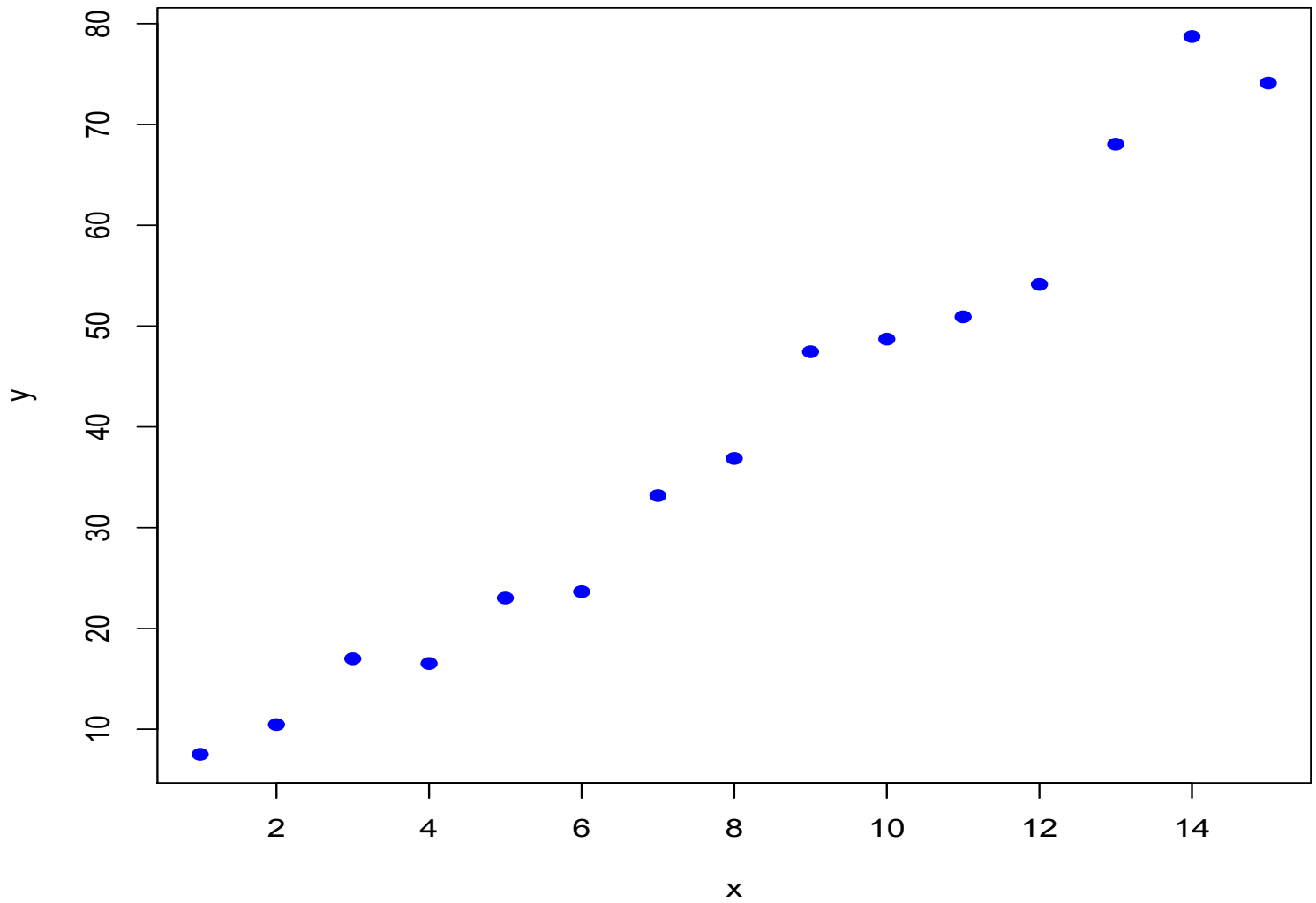
Examining a Scatterplot

In any graph of data, look for the *overall pattern* and for striking *deviations* from the pattern.
You can describe the overall pattern of a scatterplot by the *direction*, *form*, and *strength* of the relationship.
An important kind of deviation is an *outlier*, an individual value that falls outside the overall pattern of the relationship.

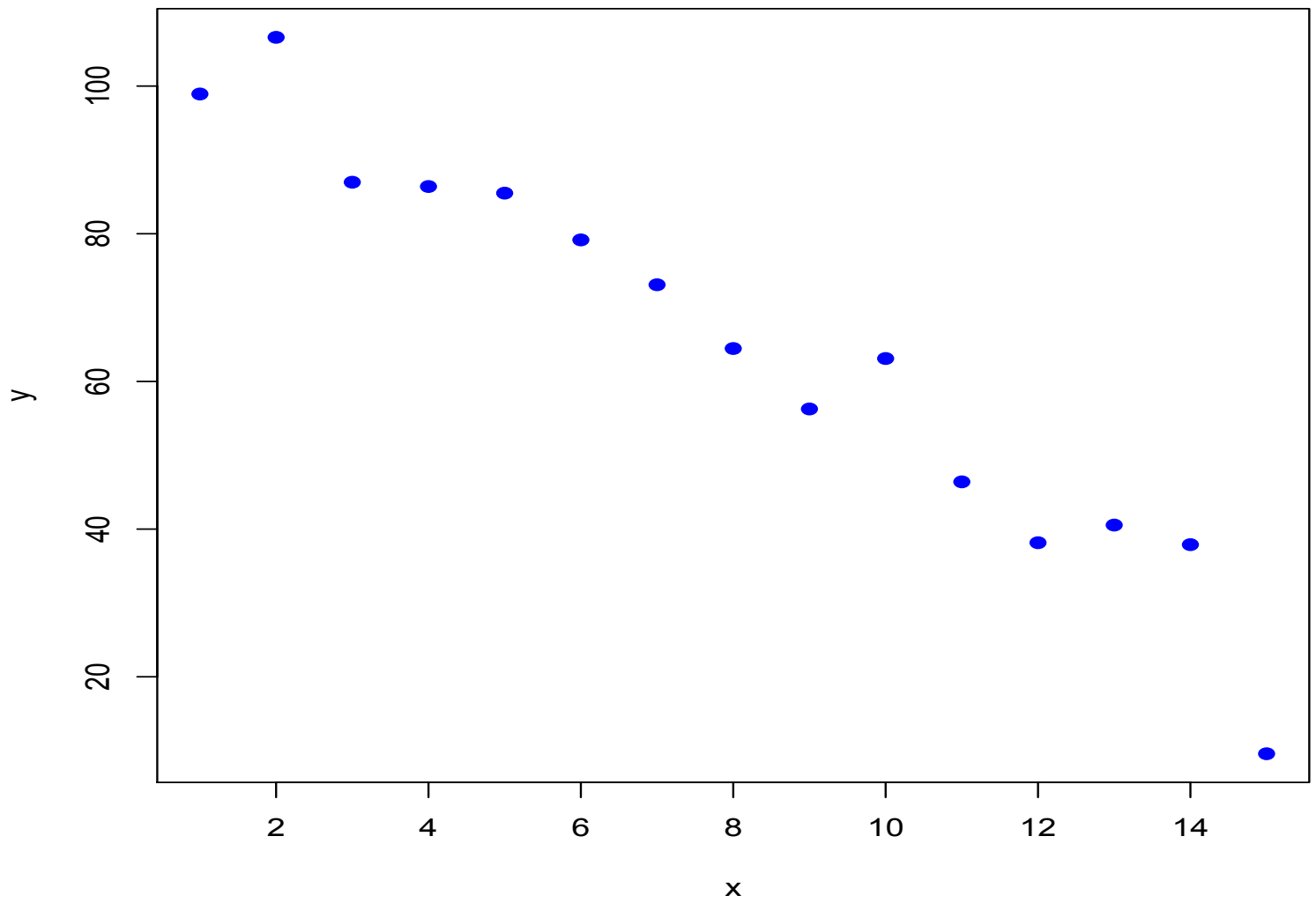
Positive Association, Negative Association

Two variables are *positively associated* when above-average values of one tend to accompany above-average values of the other, and below-average values also tend to occur together.
Two variables are *negatively associated* when above-average values of one tend to accompany below-average values of the other, and vice versa.
The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form.

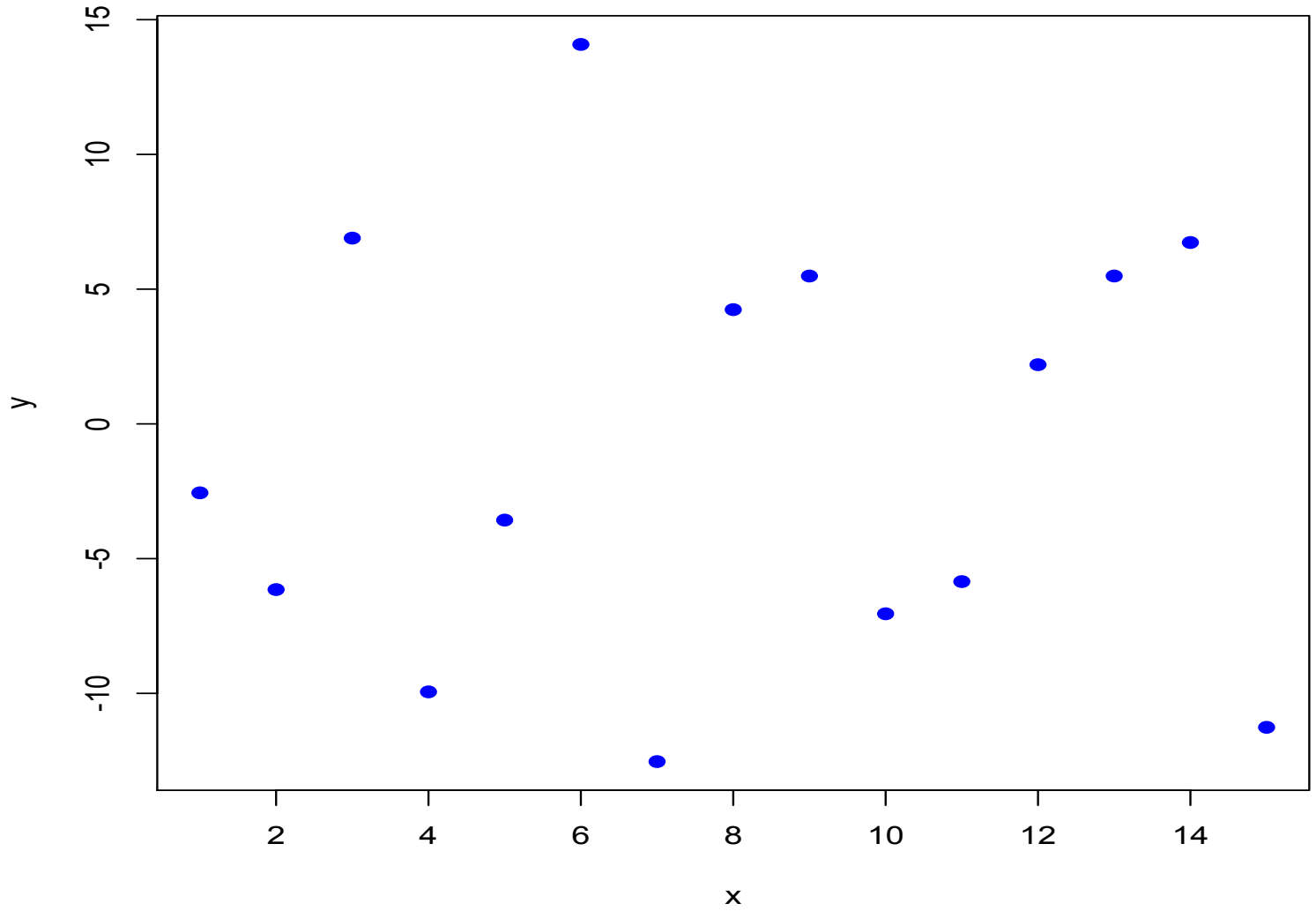
Positive Association (scatterplot)



Negative Association (scatterplot)



NO Association (scatterplot)



Do heavier people burn more energy? (Again)

Describe the direction, form, and strength of the relationship between lean body mass and metabolic rate, as displayed in your plot.

Solution

The scatterplot shows a positive direction, linear form, and moderately strong association.

Do heavier people burn more energy? (Again)

The study of dieting described earlier collected data on the lean body mass (in kilograms) and metabolic rate (in calories) for both female and male subjects.

Mass	Rate	Sex	Mass	Rate	Sex
36.1	995	F	40.3	1189	F
54.6	1425	F	33.1	913	F
48.5	1396	F	42.4	1124	F
42.0	1418	F	34.5	1052	F
50.6	1502	F	51.1	1347	F
42.0	1256	F	41.2	1204	F

More data

Mass	Rate	Sex	Mass	Rate	Sex
51.9	1867	M	47.4	1322	M
46.9	1439	M	48.7	1614	M
62.0	1792	M	51.9	1460	M
62.9	1666	M			

a) Make a scatterplot of metabolic rate versus lean body mass for all 19 subjects. Use separate symbols to distinguish women and men. (This is a common method to compare two groups of individuals in a scatterplot)
b) Does the same overall pattern hold for both women and men? What is the most important difference between women and men?

Reading our data

```
# Step 1. Entering data;

# url of metabolic rate data;

meta_url=
"https://mcs.utm.utoronto.ca/~nosedal/data/metabolic2.txt"

# import data in R;

data = read.table(meta_url, header = TRUE);
```

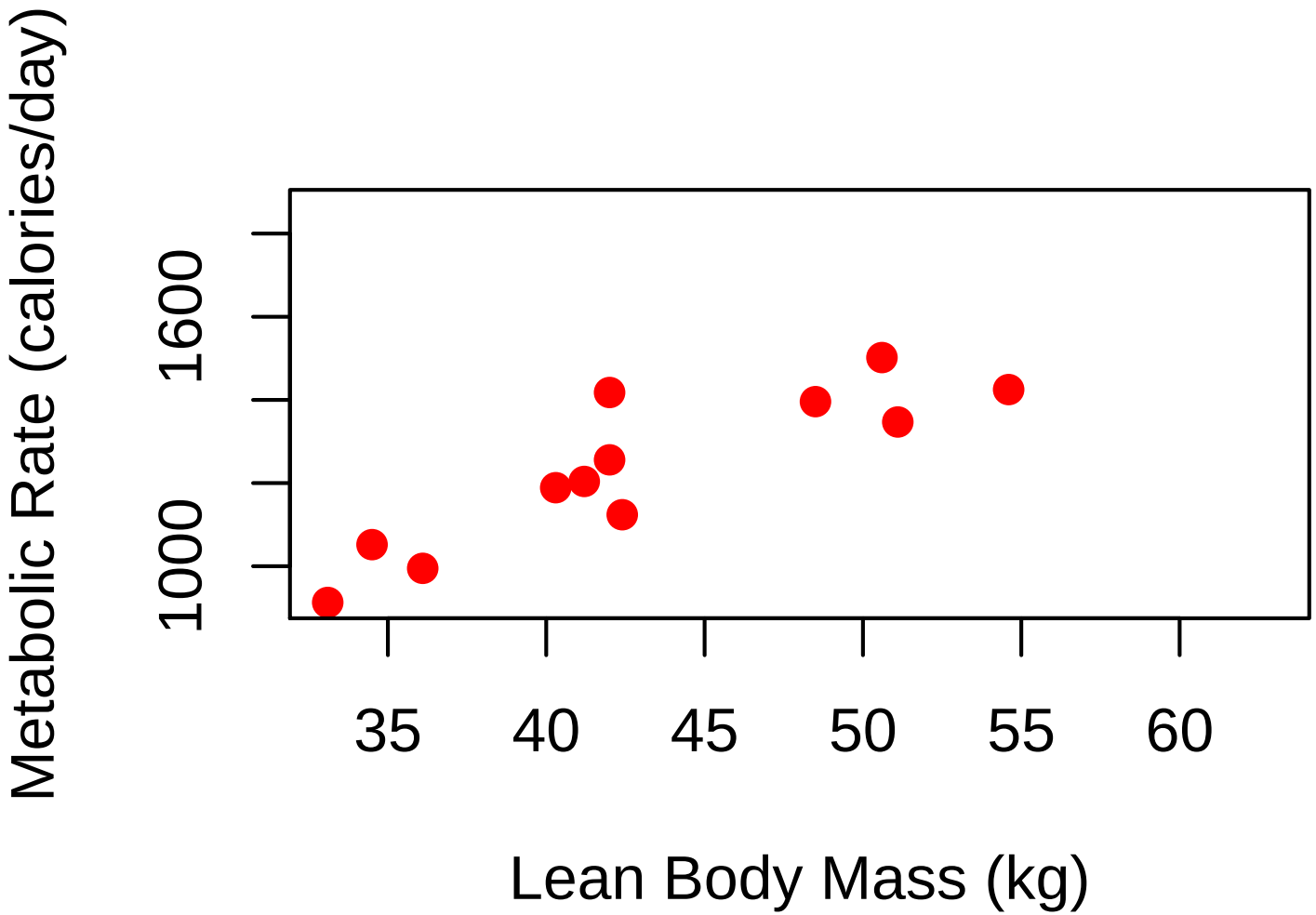
Reading our data

```
# Step 2. Formating data;

x.min=min(data$Mass);
x.max=max(data$Mass);

y.min=min(data$Rate);
y.max=max(data$Rate);

female=data[1:12, ];
male=data[13:19, ];
```

Scatterplot

```
# Step 3. Making scatterplot;

plot(female$Mass,female$Rate,pch=19,col="red",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)",
xlim=c(x.min,x.max),ylim=c(y.min,y.max));
```

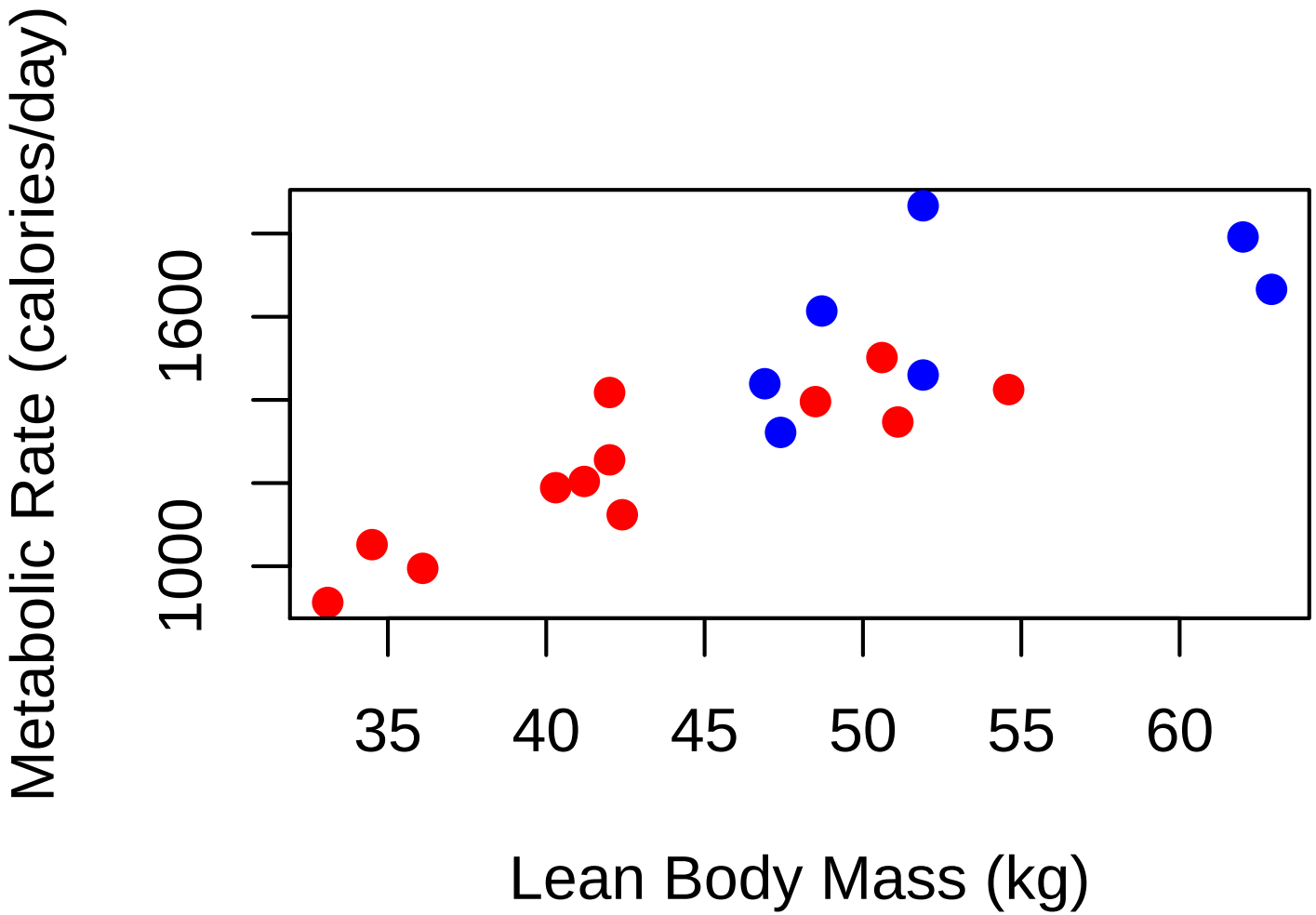
Scatterplot



Scatterplot

```
# Step 3. Making scatterplot;

plot(female$Mass,female$Rate,pch=19,col="red",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)",
xlim=c(x.min,x.max),ylim=c(y.min,y.max));

points(male$Mass,male$Rate,pch=19,col="blue");

# pch=19 tells R that you want solid circles;
```

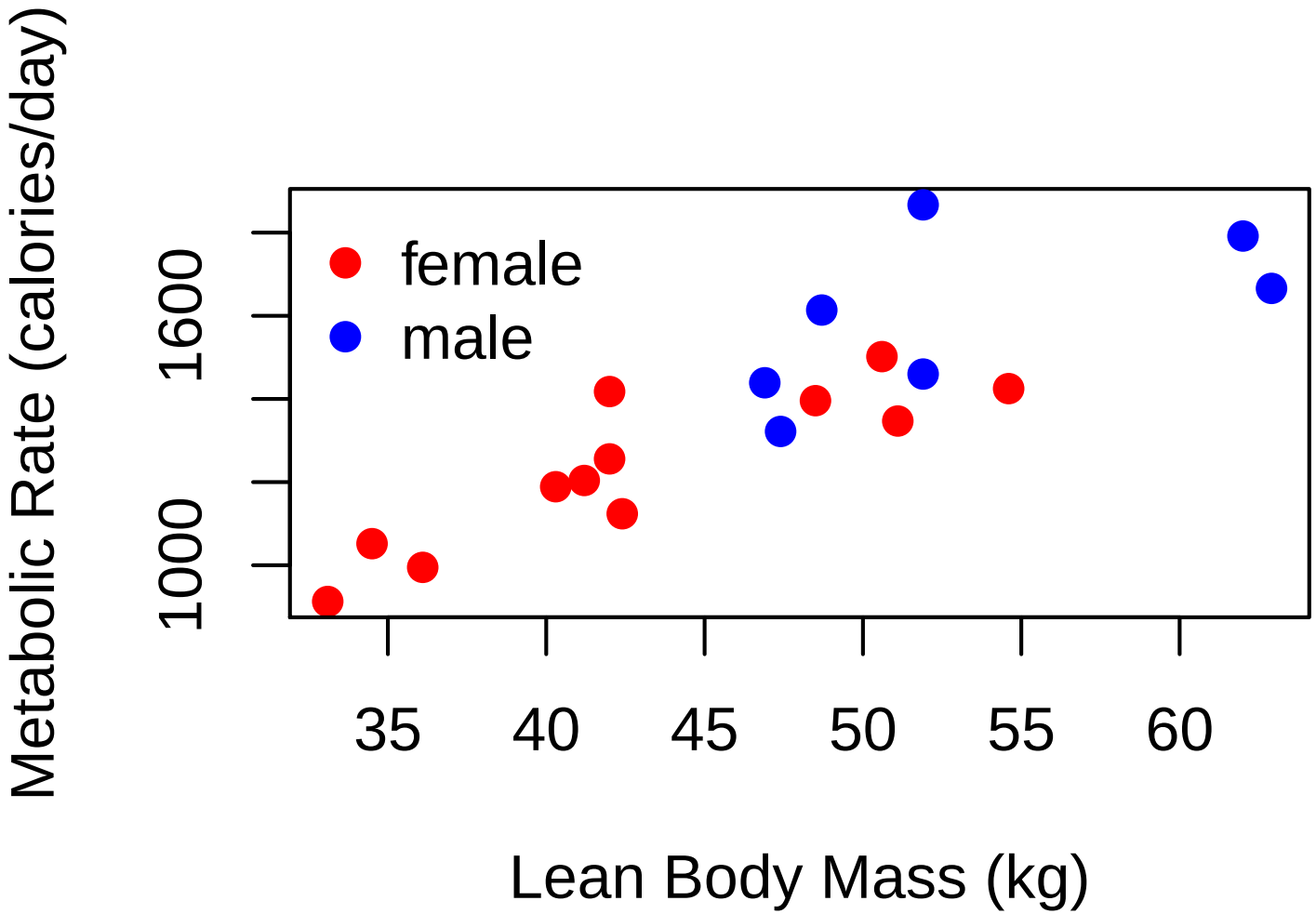
Scatterplot



Scatterplot

```
# Step 3. Making scatterplot;

plot(female$Mass,female$Rate,pch=19,col="red",
xlab="Lean Body Mass (kg)",
ylab="Metabolic Rate (calories/day)",
xlim=c(x.min,x.max),ylim=c(y.min,y.max));

points(male$Mass,male$Rate,pch=19,col="blue");

legend("topleft",c("female","male"),pch=c(19,19),
col=c("red","blue"),bty="n");

# legend tells R that you want to add a legend to
# your graph;
# topleft, where you want to position legend;
# bty="n" NO box around legend;
```

Scatterplot



b) Solution

For both men and women, the association is linear and positive. The women's points show a stronger association. As a group, males typically have larger values for both variables (they tend to have more mass, and tend to burn more calories per day).

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 values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on.

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

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

Covariance

You can compute the covariance, S_{xy} using the following formula:

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1}$$

Correlation (alternative formula)

$$r = \frac{S_{xy}}{S_x S_y}$$

where

r = correlation.

S_{xy} = covariance.

S_x = standard deviation of x.

S_y = standard deviation of y.

Example

Five observations taken for two variables follow

x_i	y_i
4	50
6	50
11	40
3	60
16	30

a) Develop a scatter diagram with x on the horizontal axis.
b) Compute the sample covariance.
c) Compute and interpret the sample correlation coefficient.

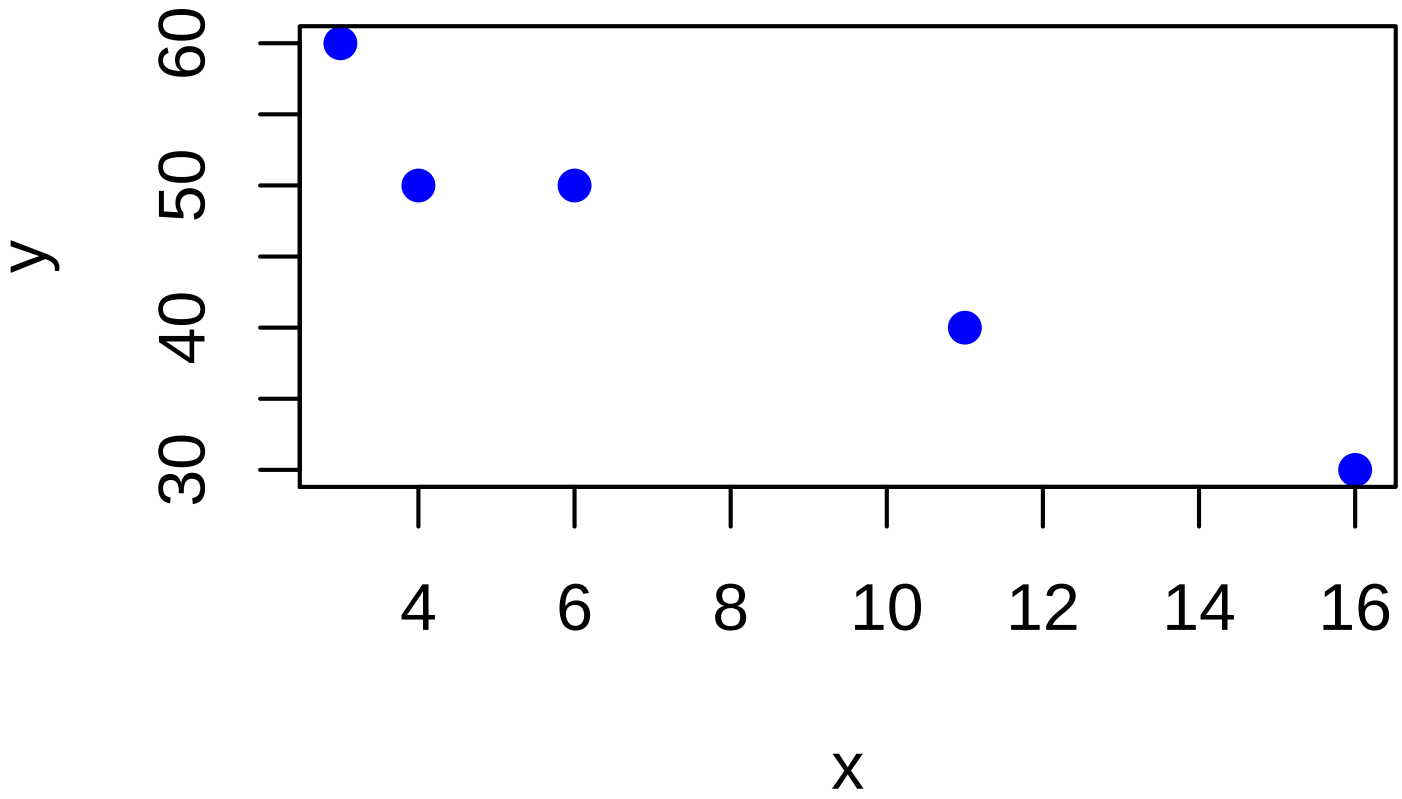
R Code

```r
# Step 1. Entering data;

x=c(4,6,11,3,16);

y=c(50,50,40,60,30);
```

Scatter diagram

```
# Step 2. Making scatter diagram;

plot(x,y,pch=19,col="blue");
```

Scatter diagram



$\bar{x}$ and $\bar{y}$

First, let's find $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{4 + 6 + 11 + 3 + 16}{5} = 8$$

$$\bar{y} = \frac{50 + 50 + 40 + 60 + 30}{5} = 46$$

Now, let's find S_x and S_y

$$S_x^2 = \frac{(4-8)^2 + (6-8)^2 + (11-8)^2 + (3-8)^2 + (16-8)^2}{4}$$

$$S_x^2 = \frac{(-4)^2 + (-2)^2 + (3)^2 + (-5)^2 + (8)^2}{4} = \frac{118}{4} = 29.5$$

$$S_x = 5.4313$$

S_y

$$S_y^2 = \frac{(50-46)^2 + (50-46)^2 + (40-46)^2 + (60-46)^2 + (30-46)^2}{4}$$

$$S_y^2 = \frac{(4)^2 + (4)^2 + (-6)^2 + (14)^2 + (-16)^2}{4} = \frac{520}{4} = 130$$

$$S_y = 11.4017$$

Covariance and Correlation

Finally, we find S_{xy} and r

$$\sum_{i=1}^{5} x_i y_i = (4)(50) + (6)(50) + (11)(40) + (3)(60) + (16)(30) = 1600$$

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1} = \frac{1600}{4} - \frac{(5)(8)(46)}{4} = -60$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{-60}{(5.4313)(11.4017)} = -0.9688$$

R Code

```
cov(x,y);

## [1] -60

cor(x,y);

## [1] -0.9688768
```

Coral reefs

This example is about a study in which scientists examined data on mean sea surface temperatures (in degrees Celsius) and mean coral growth (in millimeters per year) over a several-year period at locations in the Red Sea. Here are the data:

Sea Surface Temperature	Growth
29.68	2.63
29.87	2.58
30.16	2.60
30.22	2.48
30.48	2.26
30.65	2.38
30.90	2.26

a) Make a scatterplot. Which is the explanatory variable?
b) Find the correlation r step-by-step. Explain how your value for r matches your graph in a).
c) Enter these data into your calculator and use the correlation function to find r (or use R to find r).

Reading our data

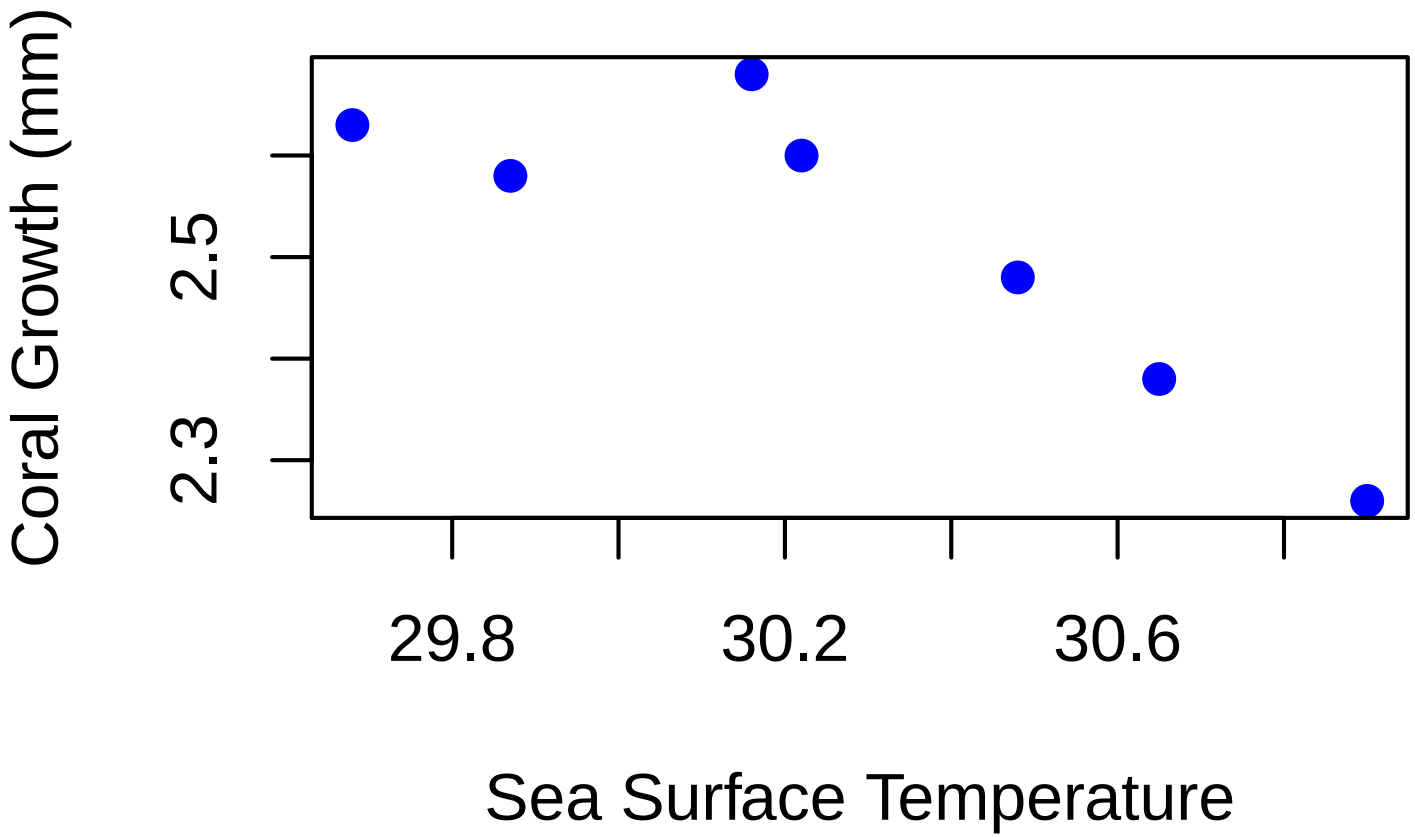
```
# Step 1. Entering data;

# url of coral reef data;

our.url="https://mcs.utm.utoronto.ca/~nosedal/data/coral.txt"

# import data in R;

data = read.table(our.url, header = TRUE);
```

Scatterplot

Temperature is the explanatory variable.

```
explanatory=data[ ,2];

response=data[ ,1];

# Step 2. Making scatterplot;

plot(explanatory, response,xlab="Sea Surface Temperature",
ylab="Coral Growth (mm)",pch=19,col="blue");
```

Scatterplot



$\bar{x}$ and $\bar{y}$

First, let's find $\bar{x}$ and $\bar{y}$

$$\bar{x} = \frac{29.68 + 29.87 + 30.16 + 30.22 + 30.48 + 30.65 + 30.90}{7} = 30.28$$

$$\bar{y} = \frac{2.63 + 2.58 + 2.60 + 2.48 + 2.26 + 2.38 + 2.26}{7} = 2.4557$$

Now, let's find S_x and S_y

$$S_x^2 = \frac{(29.68 - 30.28)^2 + ... + (30.65 - 30.28)^2 + (30.90 - 30.28)^2}{6}$$

$$S_x^2 = 0.1845$$

$$S_x = 0.4296$$

S_y

$$S_y^2 = \frac{(2.63 - 2.4557)^2 + ... + (2.38 - 2.4557)^2 + (2.26 - 2.4557)^2}{6}$$

$$S_y^2 = 0.0249$$

$$S_y = 0.1578$$

Covariance

Finally, we find S_{xy} and r

$$\begin{aligned}\sum_{i=1}^{7} x_i y_i &= (29.68)(2.63) + \ldots + (30.65)(2.38) + (30.90)(2.26) \\ &= 520.1504\end{aligned}$$

$$\begin{aligned}S_{xy} &= \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1} \\ &= \frac{520.1504}{6} - \frac{(7)(30.28)(2.4557)}{6} \\ &= 86.6917 - 86.7516 = -0.0599\end{aligned}$$

Correlation

$$r = \frac{S_{xy}}{S_x S_y} = \frac{-0.0599}{(0.4296)(0.1578)} = -0.8835$$

This is consistent with the strong, negative association depicted in the scatterplot.

c) R will give a value of $r = -0.8635908$.

R Code

```
growth=data[ ,1];

# data[ ,1] gives you the first column of data;

temp=data[ ,2];

# data[ ,2] gives you the 2nd column of data;

cov(growth, temp);

## [1] -0.05593333

cor(growth, temp);

## [1] -0.8635908
```

Example

Five observations taken for two variables follow

x_i	y_i
1	1
2	2
3	3
4	4
5	5

a) Develop a scatter diagram with x on the horizontal axis.
b) Compute the sample covariance.
c) Compute and interpret the sample correlation coefficient.

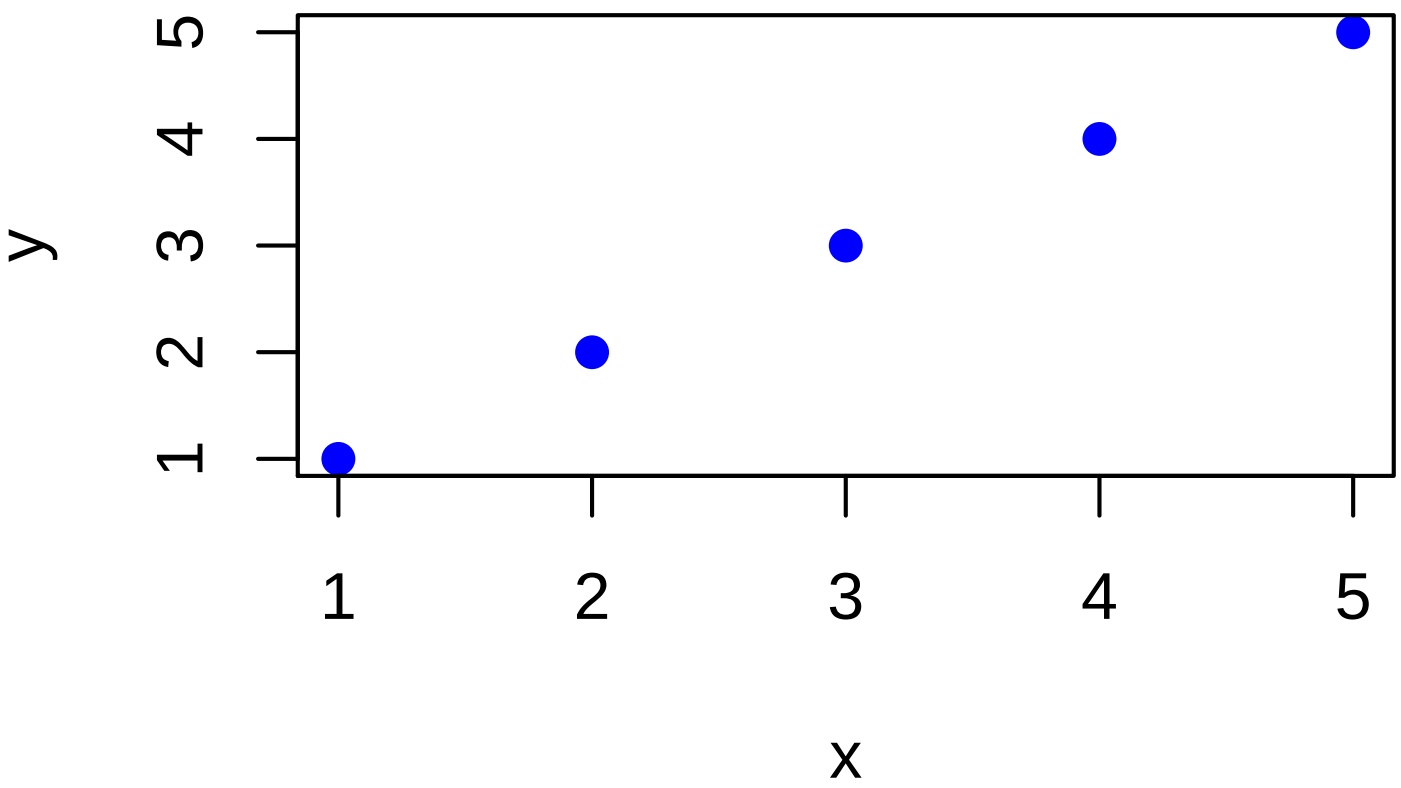
R Code

```r
# Step 1. Entering data;

x=seq(1,5,by=1);

y=seq(1,5,by=1);
```

R Code

```
# Step 2. Making scatterplot;

plot(x,y,pch=19,col="blue");
```

R Code



Covariance and Correlation

$\bar{x} = 3$ $\quad S_x = 1.58113883$

$\bar{y} = 3$ $\quad S_y = 1.58113883$

$\sum_{i=1}^{5} x_i y_i = 55$

$$S_{xy} = \frac{\sum x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1}$$

$$S_{xy} = \frac{55}{4} - \frac{(5)(3)(3)}{4}$$

$$S_{xy} = 13.75 - 11.25 = 2.5$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{2.5}{(1.58113883)(1.58113883)} = 1$$

R Code

```
cov(x,y);

## [1] 2.5

cor(x,y);

## [1] 1
```

Facts about correlation

1. *Correlation makes no distinction between explanatory and response variables.* It makes no difference which variable you call x and which you call y in calculating the correlation.
2. Because r uses the standardized values of the observations, *r does not change when we change the units of measurement of x, y, or both.* The correlation r itself has no unit of measurement; it is just a number.
3. *Positive r indicates positive association between the variables, and negative r indicates negative association.*
4. The correlation r is always a number between -1 and 1. Values of r near 0 indicate a very weak linear relationship. Perfect correlation, $r = 1$ or $r = -1$, occurs only when the points on a scatterplot lie exactly on a straight line.

Strong association but no correlation

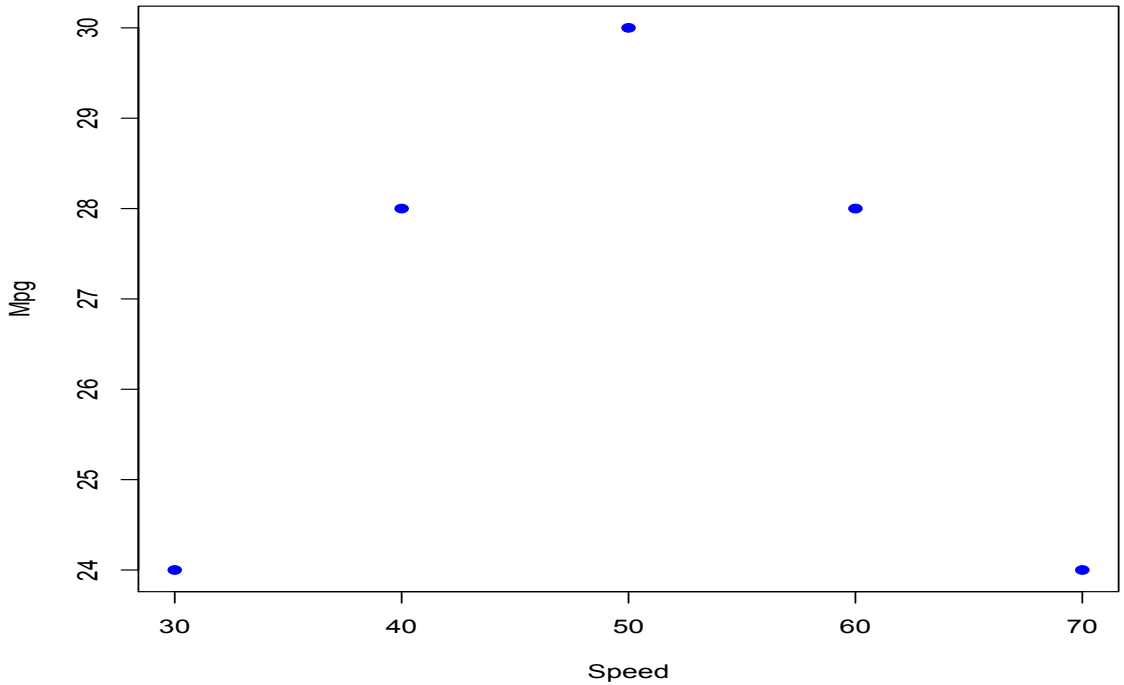
The gas mileage of an automobile first increases and then decreases as the speed increases. Suppose that this relationship is very regular, as shown by the following data on speed (miles per hour) and mileage (miles per gallon):

Speed	Mileage
30	24
40	28
50	30
60	28
70	24

Strong association but no correlation (cont.)

Make a scatterplot of mileage versus speed. Show that the correlation between speed and mileage is $r = 0$. Explain why the correlation is 0 even though there is a strong relationship between speed and mileage.

Scatterplot



Remember that correlation only measures the strength and direction of a *linear* relationship between two variables.

Correlation

$\bar{x} = 50$ $\quad S_x = 15.8113$

$\bar{y} = 26.8$ $\quad S_y = 2.6832$

$\sum_{i=1}^{5} x_i y_i = 6700$

$$S_{xy} = \frac{\sum x_i y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1}$$

$$S_{xy} = \frac{6700}{4} - \frac{(5)(50)(26.8)}{4}$$

$$S_{xy} = 1675 - 1675 = 0$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{0}{(15.8113)(2.6832)} = 0$$

More facts about correlation

1. Correlation requires that both variables be quantitative, so that it makes sense to do the arithmetic indicated by the formula for r.
2. Correlation measures the strength of only the linear relationship between two variables. Correlation does not describe curved relationships between variables, no matter how strong they are.
3. Like the mean and the standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations.
4. Correlation is not a complete summary of two-variable data, even when the relationship between the variables is linear. You should give the means and standard deviations of both x and y along with the correlation.