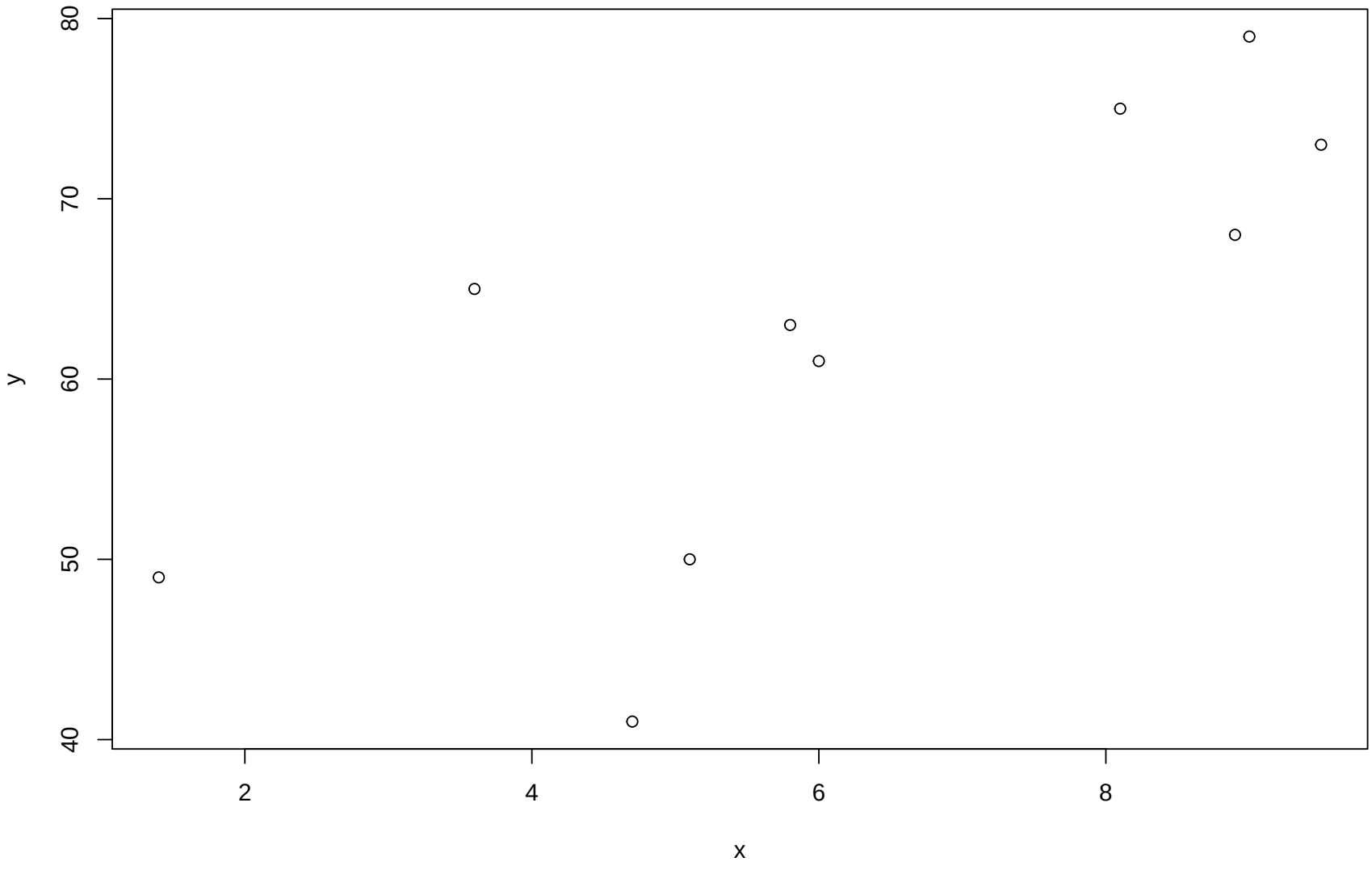
Two (possibly dependent) variables

Consider the following data:

x	y
3.6	65
8.1	75
4.7	41
8.9	68
9.5	73
1.4	49
5.8	63
9	79
6	61
5.1	50

Make a scatterplot



Find the (Pearson's) correlation coefficient

The correlation coefficient has many equivalent formulas.

The correlation coefficient can be thought of as the mean of products of standard scores.

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

The following is probably easier to use for calculations.

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Calculate the correlation coefficient

x	y	xy
3.6	65	234
8.1	75	607.5
4.7	41	192.7
8.9	68	605.2
9.5	73	693.5
1.4	49	68.6
5.8	63	365.4
9	79	711
6	61	366
5.1	50	255
$\sum x = 62.1$	$\sum y = 624$	$\sum x_i y_i = 4098.9$
$\bar{x} = 6.21$	$\bar{y} = 62.4$	
$s_x = 2.65$	$s_y = 12.4$	

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{4098.9 - 10 \cdot 6.21 \cdot 62.4}{(10-1) \cdot 2.65 \cdot 12.4} = 0.759$$

Determine the best-fit line

The regression line has the form

$$y = a + bx$$

So, a is the y-intercept and b is the slope. We have formulas to determine them:

$$b = r\frac{s_y}{s_x} = 0.759 \cdot \frac{12.4}{2.65} = 3.55$$

$$a = \bar{y} - b\bar{x} = 62.4 - 3.55 \cdot 6.21 = 40.4$$

Our regression line:

$$y = 40.4 + 3.55x$$

Graph the regression line