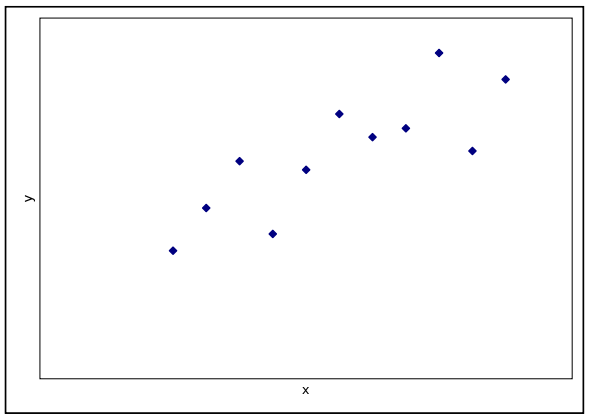
4.2. Scatter Diagrams and Correlation

Consider in the following two variables x and y, which are measured at least on interval scale. We may get an idea about the relationship between those variables by taking a look at their scatterplot (korrelaatiodiagramma/ pisteparvi/ sironta-/hajontakuvio), which is just a plot of all observation pairs (x_i,y_i) in a 2-dimensional coordinate system.

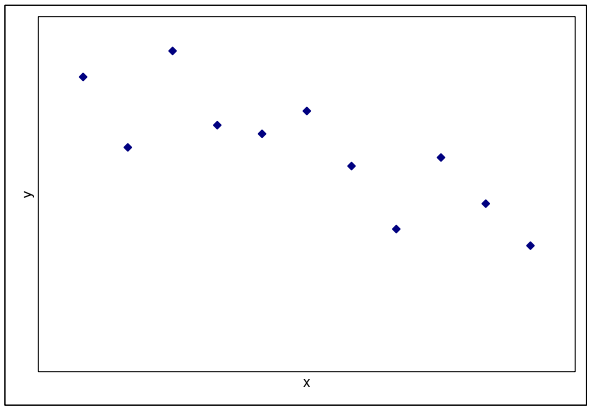
In particular, there may be a linear association (lineaarinen/ suoraviivainen riippuvuus) between the variables. The better the points in a scatter plot may be approximated by a straight line, the stronger the linear relationship is. The variables are positvely associated (positiivinen lineaarinen riippuvuus) if an increase in x correponds to an increase in y, and they are negatively associated (negatiivinen lineaarinen riippuvuus) if an increase in x correponds to a decrease in y.

Example.

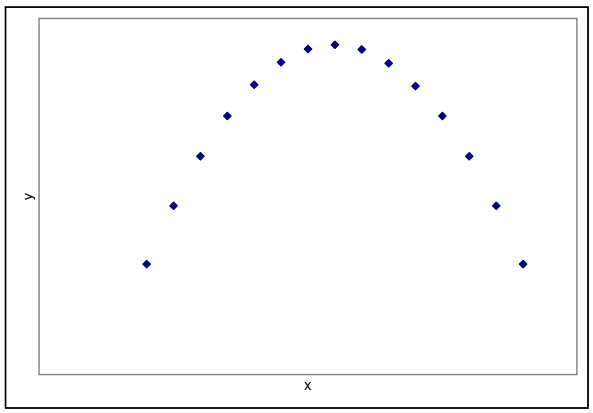
Positive linear association:



Negative linear association:



Nonlinear relationship:

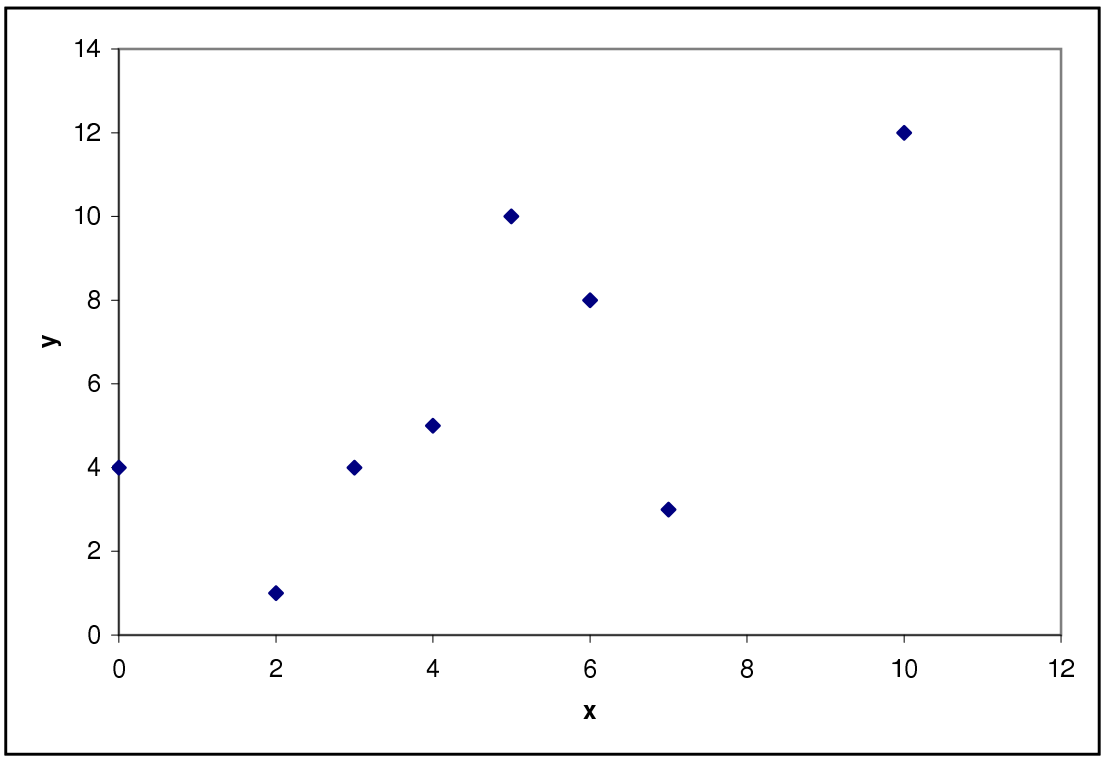


Example.

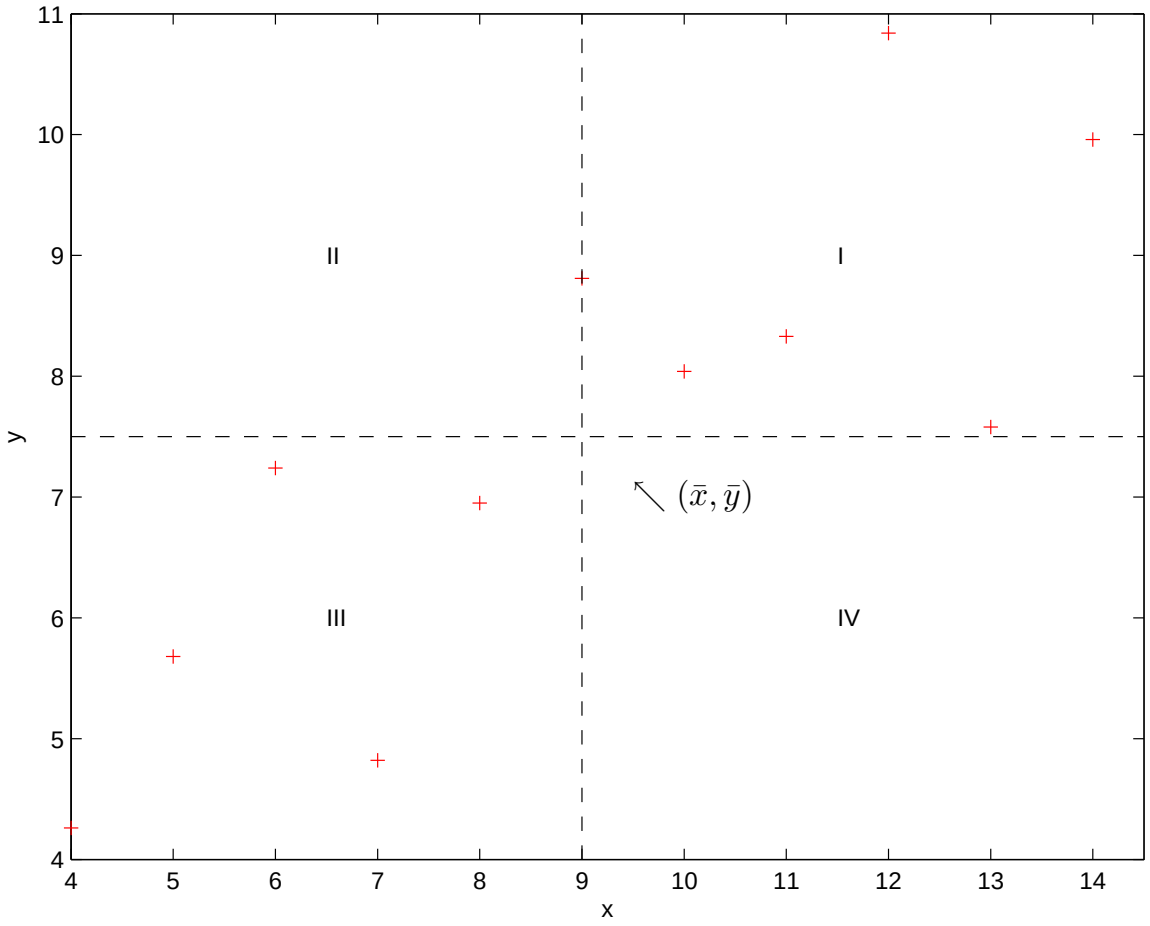
Consider the relationship between the latest advertising campaign for a certain soft drink and the number of bottles sold. Eight persons have been asked, how often they saw the advertising campaign (variable x) and how many bottles they bought (variable y). The following data has been obtained:

Person:	1	2	3	4	5	6	7	8
x:	5	10	4	0	2	7	3	6
y:	10	12	5	4	1	3	4	8

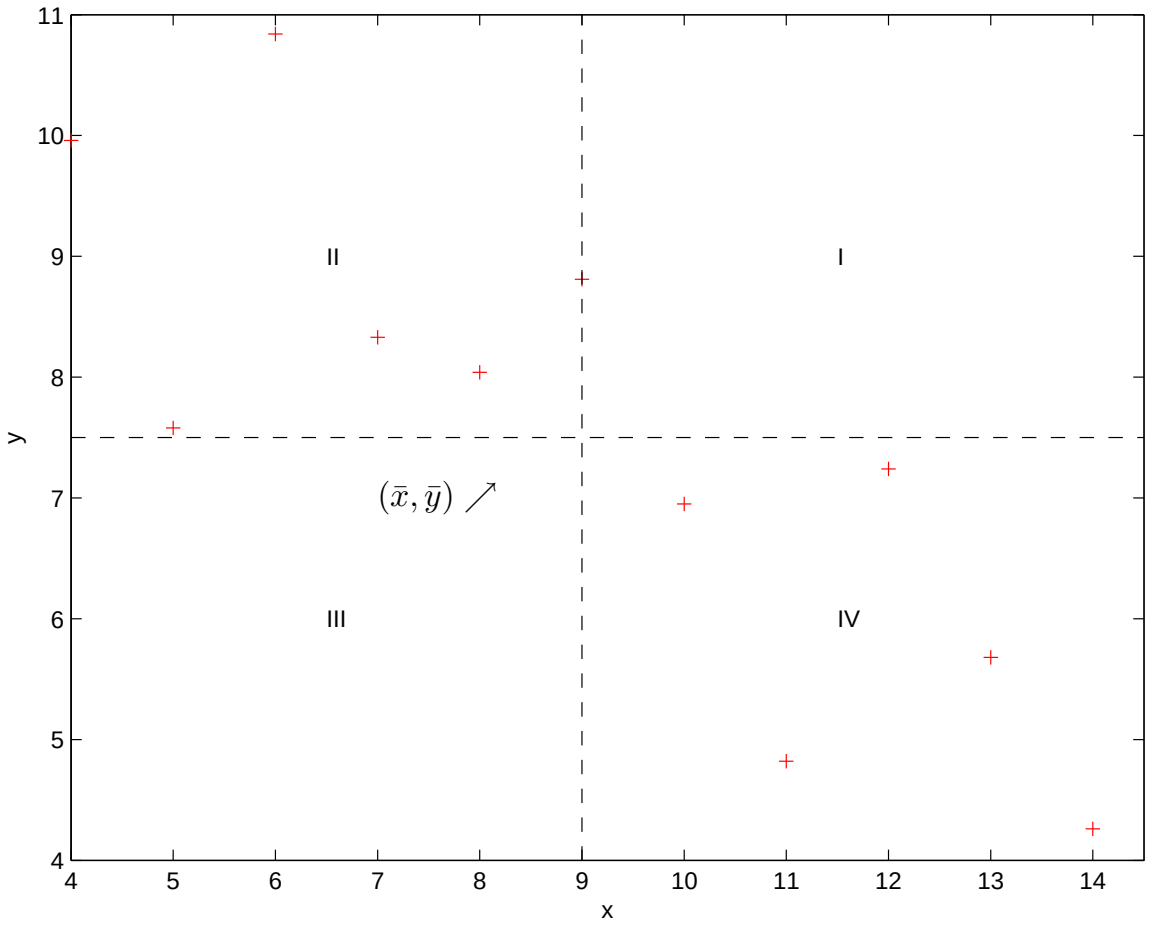
The scatter plot looks like this:



The strength of the linear association between x and y may be assessed by plotting the observations into a new coordinate system with origin in $(\bar{x}, \bar{y})$. For a positive linear relationship the observations are concentrated in the first and third quadrant of the transformed coordinate system, for which $(x_i - \bar{x})(y_i - \bar{y})$ is positive:



For negative linear relationships the observations are concentrated in the second and fourth quadrant of the transformed coordinate system, for which$(x_i - \bar{x})(y_i - \bar{y})$ is negative:



We may therefore assess linear association by considering the sum

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

The sum $\sum(x_i - \bar{x})(y_i - \bar{y})$ is positive if x and y are positively associated and negative if they are negatively associated. Dividing that sum by $(n-1)$ yields the (sample) covariance (kovarianssi) s_{xy}:

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

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \underbrace{\frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}_{n\bar{x}\bar{y}}\right]$$

The covariance has the disadvantage that it depends on the measurement scale. In order to get a measure of linear association that is independent of measurement scale, Francis Galton introduced the so called linear correlation coefficient (Pearsonin (tulomomentti) korrelaatiokerroin), often misattributed to Karl Pearson, which divides the covariance by both the standard deviation of x and the standard deviation of y.

Pearson's linear correlation coefficient is:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}.$$

The value of the correlation coefficient is always within $[-1, 1]$. A positive correlation coefficient signals positive linear association and a negative correlation coefficient signals negative linear association. The larger the absolute value of the correlation coefficient, the stronger the linear relationship is. -1 signals complete negative association and +1 complete positive linear association. If there is no linear relationship between the variables, then the correlation coefficient is 0. There may however still be a nonlinear relationship in that case.

Example.(softdrink campaign continued) Calculation of the correlation coefficient between adverts seen and bottles bought:

Person	x_i	y_i	x_i^2	y_i^2	x_iy_i
1	5	10	25	100	50
2	10	12	100	144	120
3	4	5	16	25	20
4	0	4	0	16	0
5	2	1	4	1	2
6	7	3	49	9	21
7	3	4	9	16	12
8	6	8	36	64	48
Sum	37	47	239	375	273

The correlation coefficient is:

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}$$

$$= \frac{273 - \frac{37 \cdot 47}{8}}{\sqrt{\left(239 - \frac{37^2}{8}\right)\left(375 - \frac{47^2}{8}\right)}} \approx 0.68.$$

This indicates a positive linear relationship.

The following points are important to keep in mind when interpreting the correlation coefficient:

1. A single observation may have a strong influence upon the correlation coefficient if its x and/or y value deviates far from the usual range. Such outliers (vieraat havainnot) as well as nonlinear relationships are easily detected by first inspecting the scatterplot, before applying the mathematical machinery of calculating a correlation coefficient. If there is a reasonable explanation for the outlier (such as measurement error, or the outlying observation originating from a different population than the other statistical units), then they should be deleted before calculating the correlation coefficient.

2. Combining observations into groups changes the correlation coefficient. In particular, a correlation based on averages over many observations is usually higher than the correlation between the same variables based on data for the individual observations.

3. Restricting the range of any of the two variables, such that the data does no longer contain all information about the full range of both x and y, will in general lead to a lower correlation coefficient than if the full range of both x and y had been taken into account.

4. A large absolute value of the correlation coefficient is no guarantee for a causal relationship between x and y. For example, there might be a lurking variable z, that affects both x and y. If that factor is known, one may correct for its impact by calculating partial correlation coefficients (osittaiskorrelaatioita, not discussed here).

We briefly summarize the main properties of the linear correlation coefficient:

It makes no difference which variable you call x and which you call y in calculating the correlation.

Correlation requires that both variables be quantitative (it cannot be calculated for variables measured on nominal or ordinal scale).

r_{xy} does not change when we change the units of x, y, or both.

Correlation measures only the strength of linear relationships. It does not describe curved relationships, no matter how strong they are.

r_{xy} is always a number between -1 and 1 with the sign of r indicating the sign of the linear relationship. The strength of the relationship increases as r moves away from 0 to either -1 or $+1$.

Correlation is very sensitive to outliers.

4.3. Rank Correlation

The linear correlation correlation coefficient discussed above has two limitations: It is only defined for quantitative variables measured on interval scale and above, and it measures only linear relationships. For variables measured at least on ordinal scale one may instead use Spearman's rank correlation (järjestyskorrelaatio) r_S.

Spearman's rank correlation coefficient is Pearson's linear correlation coefficient applied to the ranks $R(x_i)$ of the observations. The rank (järjestysluku / sijaluku) R of an observation x_i tells the position of x_i when arranged in ascending order. If several statistical units share the same observation value, they are said to have ties (sidos). These observations get then assigned the arithmetic average of the ranks, which would have been obtained if they had slightly different values.

When no ties exist, the calculation of Spearman's rank correlation coefficient may be simplified as follows. For each pair of ranks $\{R(x_i), R(y_i)\}$ one calculates the difference in ranks (järjestyslukujen erotus):

$$d_i = R(x_i) - R(y_i).$$

Spearman's rank correlation r_S is then determined from:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n}.$$

In practice this formula may be still applied to get approximate values for r_S as long as the number of ties is small as compared to the number of observations n.

Example. Consider Spearman's rank correlation applied to points obtained in a test before (x) and after (y) school attendance:

x_i	$R(x_i)$	y_i	$R(y_i)$	d_i	d_i^2
10	1	23	2.5	-1.5	2.25
11	2	27	5	-3	9
12	3	18	1	2	4
13	4	29	7	-3	9
15	5	23	2.5	2.5	6.25
16	6	25	4	2	4
17	7	28	6	1	1
18	8	31	8	0	0
					35.5

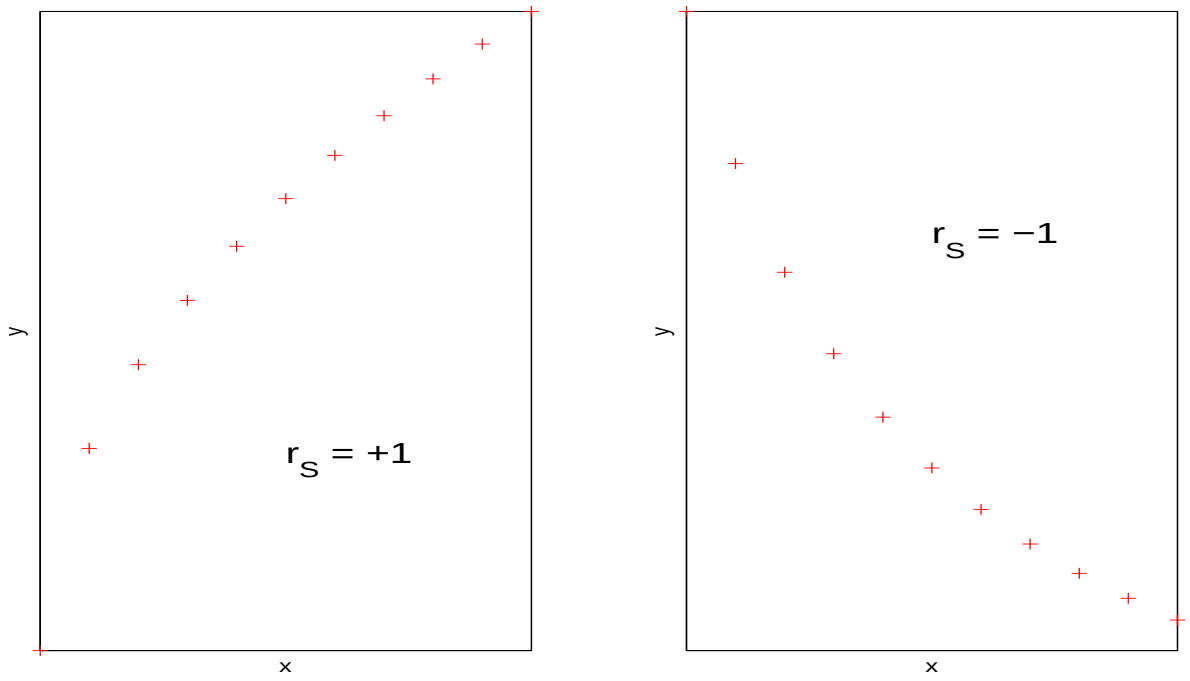
Spearman's rank correlation coefficient is approximately (due to ties between observations 1 and 5):

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n} = 1 - \frac{6 \cdot 35.5}{8^3 - 8} \approx 0.5774.$$

The correct value from applying Pearson's correlation coefficient to the ranks would have been 0.5749.

Spearman's rank correlation, like Pearson's linear correlation, is confined to the range $[-1, 1]$. However, unlike Pearson's product-moment correlation, which measures the linear association between two variables, it describes the relationship of arbitrary monotonic functions of the data (monotoninen riippuvuus), that is the similiarity of their order. r_S is $+1$, if the order in x and y is exactly the same, and it is -1 if their orders are just opposite. It is 0 if there is no apparent relationship between the orderings of both variables.

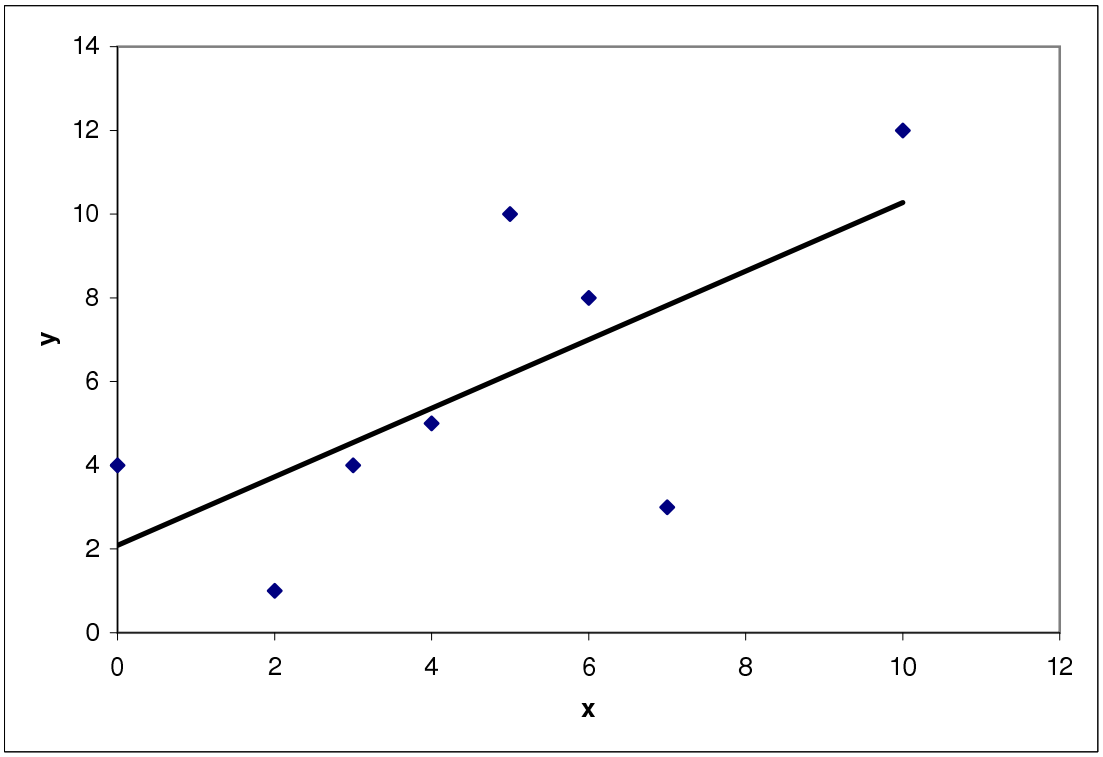
Example:



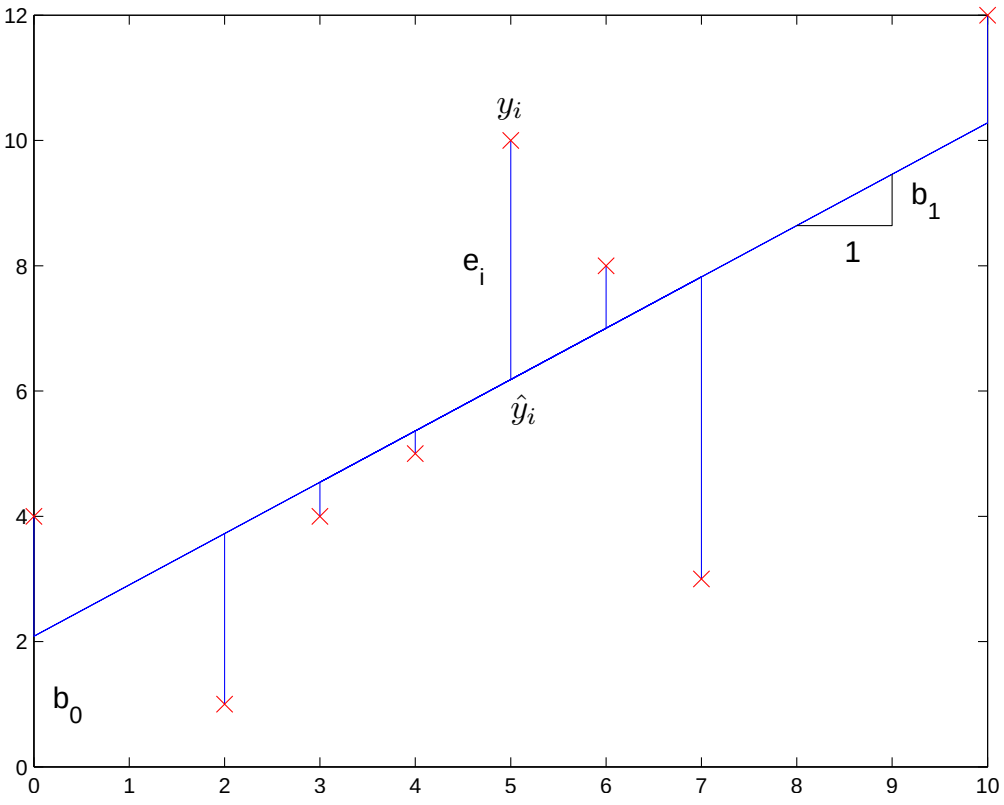
4.4. Regression

The idea of regression analysis (regressioanalyysi) is to illustrate a possible approximate linear causal relationship (syy-seuraussuhde) between the explanatory variable or regressor (selittävä muuttuja) x and the dependent variable or regressand (selitettävä muuttuja) y as a straight line, called regression line (regressiosuroa), in the scatterplot of x and y.

Example: (softdrink campaign continued.)



The regression line $\hat{y} = b_0 + b_1 x$ is most commonly determined by the method of ordinary least squares (OLS) (pienimmän neliösumman menetelmä, PNS), that is by minimizing the sum of all squared vertical distances between the observations y_i of the dependent variable y and their predicted values $\hat{y}_i = b_0 + b_1 x_i$. (This automatically ensures that the sum of all vertical distances above the regression line and the sum of all vertical distances below the regression line average out to zero.)



The vertical distances between the observations y_i and their predicted values $\widehat{y}_i$ from the regression line,

$$e_i := y_i - \widehat{y}_i,$$

are called residuals (jäännös, residuaali). So mathematically, the method of least squares consists in solving the expression

$$\min \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \min \sum_{i=1}^{n} e_i^2.$$

Doing this yields for the slope coefficient (kulmakerroin) b_1 and the intercept (vakio) b_0 of the regression line $\widehat{y} = b_0 + b_1 x$:

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}.}$$

To see this, we regard the sum of squared residuals as a function of the yet unknown regression coefficients (regressiokertoimet) b_0 and b_1:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = f(b_0, b_1).$$

A global minimum of $\sum e_i^2$ requires both $\frac{\partial f}{\partial b_0}$ and $\frac{\partial f}{\partial b_1}$ to be zero:

$$\frac{\partial \left(\sum e_i^2\right)}{\partial\, b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0, \text{ and}$$

$$\frac{\partial \left(\sum e_i^2\right)}{\partial\, b_1} = -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0.$$

This yields the so called normal equations (normaaliyhtälöt):

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i,$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2.$$

Dividing the first equation by n yields for the intercept:

$$\bar{y} = b_0 + b_1 \bar{x} \quad \Leftrightarrow \quad b_0 = \bar{y} - b_1 \bar{x},$$

which states that each regression line must pass through $(\bar{x}, \bar{y})$, but doesn't tell anything about its direction yet. Inserting b_0 into the second equation yields

$$\sum_{i=1}^{n} x_i y_i = \left(\frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n}\right) \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

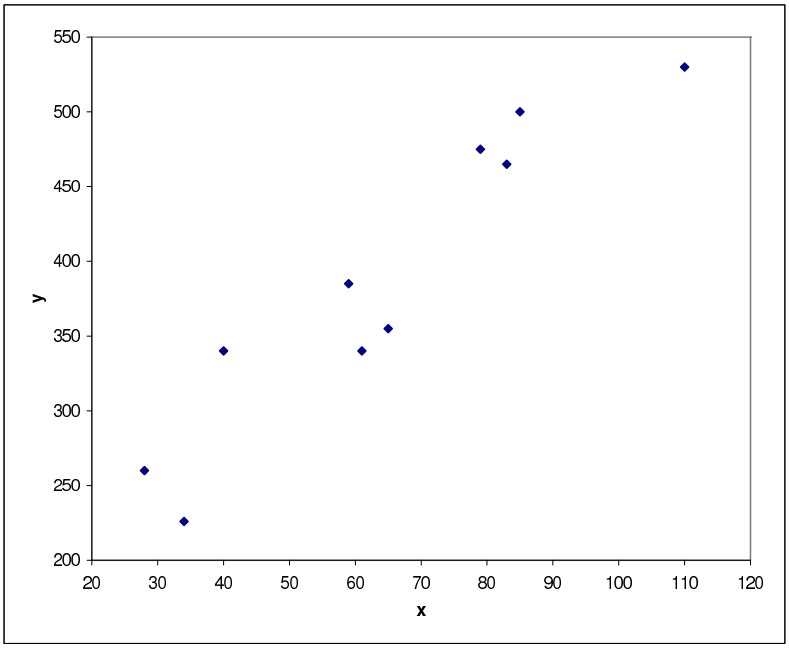
which after solving for b_1 yields for the slope:

$$b_1 = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{s_{xy}}{s_x^2}.$$

That (b_0, b_1) indeed minimizes $\sum e_i^2$ follows from the fact that it is the only solution of $(\frac{\partial f}{\partial b_0}, \frac{\partial f}{\partial b_1}) = (0, 0)$ and that $\sum e_i^2 \to \infty$ for b_0, $b_1 \to \pm\infty$.

Example. Below is some data upon the influence of flat size in m^2 upon price in units of 1000 good old finnish markka:

Flat	Size	Price
1	110	530
2	85	500
3	83	465
4	79	475
5	65	355
6	61	340
7	59	385
8	40	340
9	34	226
10	28	260



Let's find the regression line and correlation coefficient between flat size x and price y. For that we need to figure out

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right),$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)^*,$$

$$\text{and} \quad s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right].$$

* only calculated for lated reference.

Flat	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	110	530	12100	280900	58300
2	85	500	7225	250000	42500
3	83	465	6889	216225	38595
4	79	475	6241	225625	37525
5	65	355	4225	126025	23075
6	61	340	3721	115600	20740
7	59	385	3481	148225	22715
8	40	340	1600	115600	13600
9	34	226	1156	51076	7684
10	28	260	784	67600	7280
Sum	644	3876	47422	1596876	272014

Now:

$$\bar{x} = \frac{644}{10} = 64.4, \quad s_x^2 = \frac{1}{9}\left(47422 - \frac{644^2}{10}\right) \approx 660.933,$$

$$\bar{y} = \frac{3876}{10} = 387.6, \quad s_y^2 = \frac{1}{9}\left(1596876 - \frac{3876^2}{10}\right) \approx 10504.267,$$

$$\text{and} \quad s_{xy} = \frac{1}{9}\left(272014 - \frac{644 \cdot 3876}{10}\right) \approx 2488.844.$$

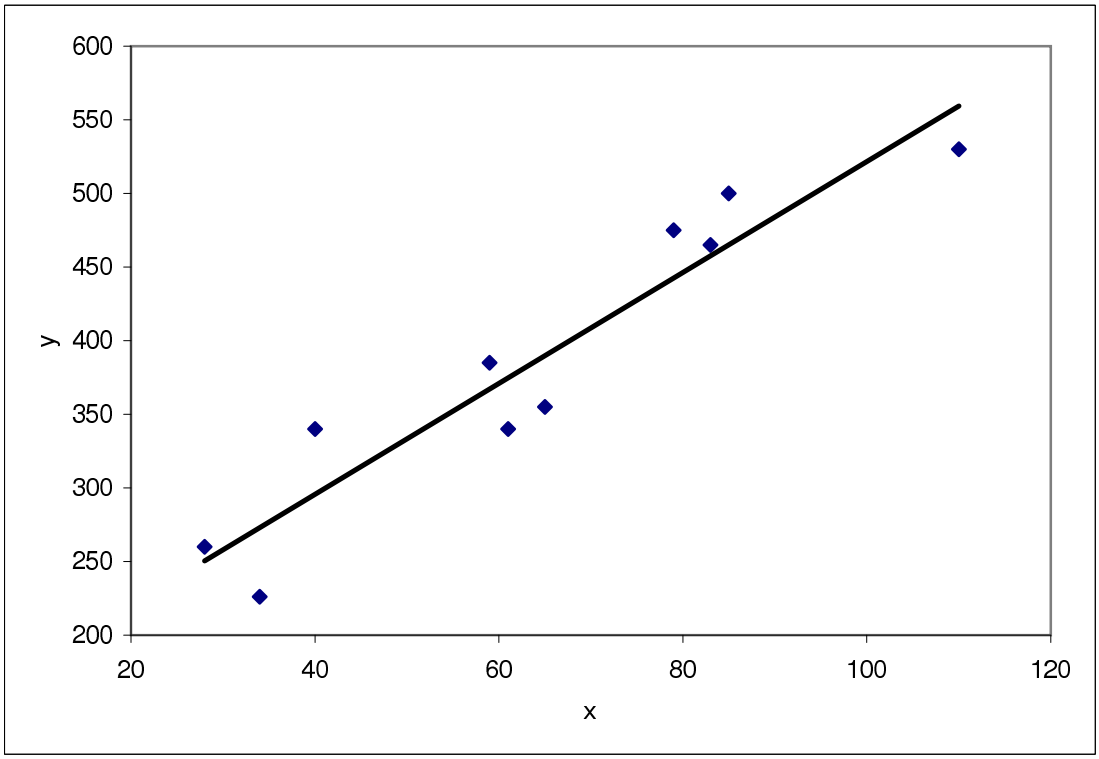
Therefore:

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{2488.844}{660.933} = 3.76565\ldots \approx 3.77$$

and

$$b_0 = \bar{y} - b_1\bar{x} = 387.6 - 3.76565 \cdot 64.4 \approx 145.1.$$

So the regression line is $\widehat{y} = 145.1 + 3.77x$:



When there is indeed a causal relationship between x and y, the regression line may be used to generate forecasts (enustetta) $\widehat{y}$.

Example. If the size of a flat is 50m^2, its estimated price $\widehat{y}$ may be obtained from inserting 50m^2 for x in the regression line:

$$\widehat{y} = 145.1 + 3.77 \cdot 50 = 333.5 (\times 1\,000\text{mk.})$$

The linear correlation coefficient r_{xy} in our example is:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{2488.844}{\sqrt{660.933}\sqrt{10504.267}} \approx 0.9446.$$

The squared correlation coefficient $R^2 := r_{xy}^2$ is called coefficient of determination (selityskerroin, selitysaste). It measures the fit (yhteensopivuus) of the regression line in the sense that it tells how large a fraction of the variation in y is due to the variation in $\hat{y}$ as predicted by the regression equation, that is:

$$R^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y} = \frac{s_{\hat{y}}^2}{s_y^2}.$$

This may be seen as follows:

In order to see that the squared correlation coefficient equals indeed the variance ratio of the fitted and observed y-values, recall from our discussion of the variance that:

$$\widehat{y} = b_0 + b_1 x \quad \Rightarrow \quad s^2_{\widehat{y}} = b_1^2 s_x^2.$$

Now, recalling that $b_1 = \frac{s_{xy}}{s_x^2}$:

$$s^2_{\widehat{y}} = \left(\frac{s_{xy}}{s_x^2}\right)^2 \cdot s_x^2 = \frac{s_{xy}^2}{(s_x^2)^2}\ s_x^2 = \frac{s_{xy}^2}{s_x^2}.$$

Therefore:

$$\frac{s^2_{\widehat{y}}}{s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2} = \left(\frac{s_{xy}}{s_x s_y}\right)^2 = r_{xy}^2 =: R^2.$$

Example: (flat prices continued.)

$$R^2 = 0.9446^2 \approx 0.89,$$

meaning that about 89% of the variation in flat prices may be explained by their size.

Note: The slope coefficient b_1 of the regression line $\hat{y} = b_0 + b_1 x$ may also be calculated from the correlation coefficient r_{xy} and the standard deviations s_x and s_y as $b_1 = r_{xy} \cdot \frac{s_y}{s_x}$, because

$$r_{xy} \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} = b_1.$$

Interpretation: A change of one standard deviation in x corresponds to a change of r_{xy} standard deviations in y. $(b_1 \cdot s_x = r_{xy} \cdot s_y)$

Example:(Flat size and price continued.) We calculated earlier the variance of flat size as $s_x^2 = 660.933$ and the variance of flat price as $s_y^2 = 10\,504.267$. The correlation between flat size and price was found to be $r_{xy} = 0.9446$. The slope coefficient of the regression of flat price upon flat size is then:

$$b_1 = r_{xy} \cdot \frac{s_y}{s_x} = 0.9446 \cdot \sqrt{\frac{10\,504.267}{660.933}} = 3.77,$$

the same value as we found earlier.