Linear regression

deterministic dependence X stochastic dependence

$$y = f\left(x_1, x_2, \ldots, x_n\right)$$

For given $\left(a_1, a_2, \ldots, a_n\right)$ $y = f\left(a_1, a_2, \ldots, a_n\right)$ may take on several different values because *y* may depend on other factors of random nature of which we are not aware or which cannot be taken into account.

Stochastic dependence may be viewed as dependence of random variables

$$Y = f\left(X_1, X_2, \ldots, X_n\right)$$

in the random vector

$$\left(Y, X_1, X_2, \ldots, X_n\right)$$

If we know the distribution $F\left(y, x_1, x_2, \ldots, x_n\right)$ of the random vector, one way of defining the stochastic dependence $y = f\left(x_1, x_2, \ldots, x_n\right)$ is to put

$$y = E\left(Y \mid X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots \wedge X_n = x_n\right)$$

$E\left(Y \mid X_1 = x_1 \wedge X_2 = x_2 \wedge \ldots \wedge X_n = x_n\right)$ is called conditioned expectancy and is defined by means of conditioned probability. Conditioned probability was defined as

$$P(A \mid B) = \frac{P\left(A \cap B\right)}{P\left(B\right)}$$

If, for example, we know the density $p(y, x_1, x_2, \ldots, x_n)$ we can define

$$y = \int_{-\infty}^{\infty} \frac{f(t, x_1, x_2, \ldots, x_n)}{f_y(x_1, x_2, \ldots, x_n)} t \, dt$$

where f_y is the marginal density

$$f_y(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{\infty} f(y, x_1, x_2, \ldots, x_n) \, dy$$

Example

Given the below table of probability function $p(x, y)$

define the dependence of *y* on *x* as the conditioned expectancy

	x_1	x_2	x_3	Σ
y_1	p_{11}	p_{12}	p_{13}	p_{y1}
y_2	p_{21}	p_{22}	p_{23}	p_{y2}
y_3	p_{31}	p_{32}	p_{33}	p_{y3}
Σ	p_{x1}	p_{x2}	p_{x3}	p

	x_1	x_2	x_3	Σ
y_1	p_{11}	p_{12}	p_{13}	p_{y1}
y_2	p_{21}	p_{22}	p_{23}	p_{y2}
y_3	p_{31}	p_{32}	p_{33}	p_{y3}
Σ	p_{x1}	p_{x2}	p_{x3}	p

We have, for instance, $y(x_1)=\dfrac{y_1 p_{11}+y_2 p_{21}+y_3 p_{31}}{p_{x1}}$

or generally

$$y(x_i)=\frac{y_1 p_{1i}+y_2 p_{2i}+y_3 p_{3i}}{p_{xi}}$$

However, usually we do not know the distribution of the random vector and can only use a sample of observed values

$$\left(y^1, x_1^1, x_2^1, \ldots, x_n^1\right), \left(y^2, x_1^2, x_2^2, \ldots, x_n^2\right), \ldots, \left(y^k, x_1^k, x_2^k, \ldots, x_n^k\right)$$

Then we must use different methods to estimate the dependence.

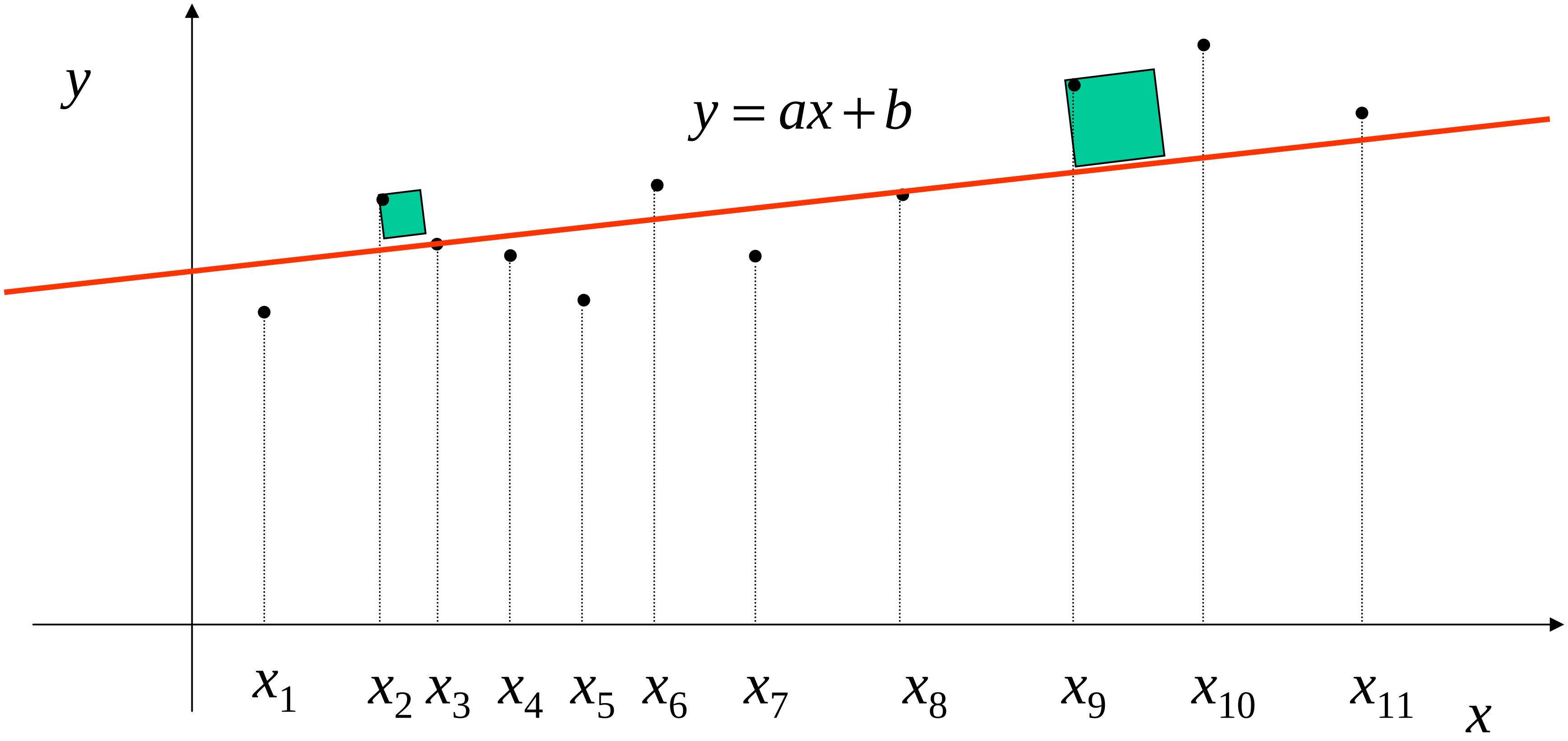
We shall consider the simplest case when the function is of one variable $y = f(x)$, the variable x may be assumed deterministic and, for some reason, we know that the dependence is to be linear, that is,

$$y = a.x + b$$

We have a sample

$$\left(y^1, x_1^1\right), \left(y^2, x_1^2\right), \ldots, \left(y^k, x_1^k\right)$$

and will use a method called a **least-square method**



We will try to find values *a*, *b* for the sum of all the squares as indicated in the figure above to be the least possible.

$$S(a,b) = \sum_{i=1}^{k} (y_i - ax_i - b)^2$$

We shall put

$$\frac{\partial S(a,b)}{\partial a} = \frac{\partial S(a,b)}{\partial b} = 0$$

$$0 = \frac{\partial S(a,b)}{\partial a} = \sum_{i=1}^{k} 2(y_i - ax_i - b)(-x_i)$$

$$0 = \frac{\partial S(a,b)}{\partial b} = \sum_{i=1}^{k} 2(y_i - ax_i - b)(-1)$$

$$-2\sum_{i=1}^{k}\left(y_i x_i - a x_i^2 - b x_i\right) = 0$$

$$-2\sum_{i=1}^{k}\left(y_i - a x_i - b\right) = 0$$

$$a\sum_{i=1}^{k} x_i^2 + b\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} y_i x_i$$

$$a\sum_{i=1}^{k} x_i + bk = \sum_{i=1}^{k} y_i$$

$$a = \frac{k\sum_{i=1}^{k} x_i y_i - \sum_{i=1}^{k} x_i \sum_{i=1}^{k} y_i}{k\sum_{i=1}^{k} x_i^2 - \left(\sum_{i=1}^{k} x_i\right)^2}$$

$$b = \frac{\sum_{i=1}^{k} x_i^2 \sum_{i=1}^{k} y_i - \sum_{i=1}^{k} x_i \sum_{i=1}^{k} x_i y_i}{k\sum_{i=1}^{k} x_i^2 - \left(\sum_{i=1}^{k} x_i\right)^2}$$