(Nov 17 2021)

Two quantitative variables

Pairs of values (x_i, y_i), $i = 1, \dots, n$

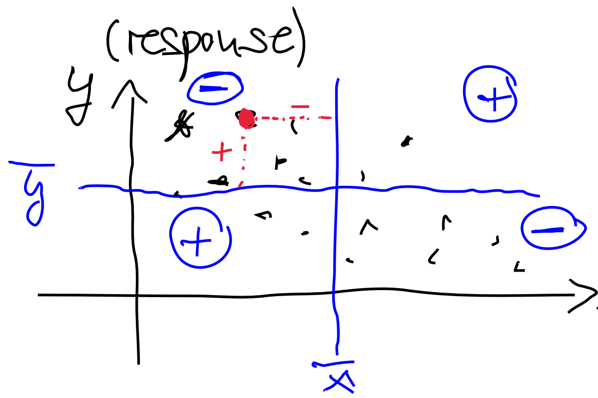
Arithmetic means: $\bar{x}$, $\bar{y}$

Variances:

$$S_{xx} = S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Scatterplot



Covariance

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

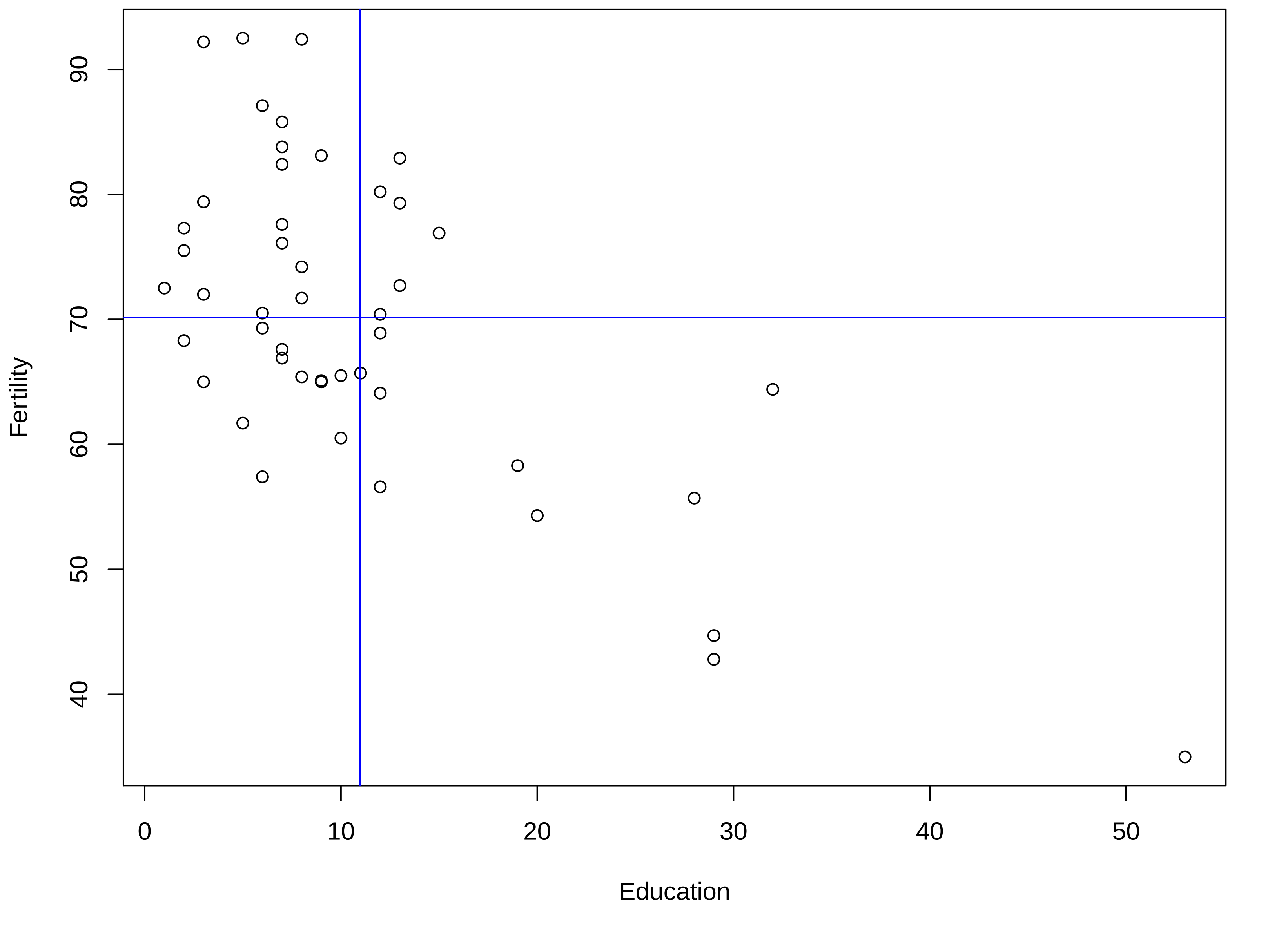
- sign shows direction of relation, e.g. swiss data: higher Education comes with lower Fertility
- the absolute value is unscaled and thus hard to interpret
- covariances of a group of variables can be collected in a matrix:

$V_1, \dots, V_p$

R: cov

$$\begin{array}{c|cccc} & V_1 & & \cdots & V_p \\ \hline V_1 & S_1^2 & S_{12} & \cdots & S_{1p} \\ & S_{21} & \ddots & \cdots & S_{2p} \\ \vdots & \vdots & & \ddots & \vdots \\ V_p & S_{p1} & S_{p2} & \cdots S_{p,p-1} & S_p^2 \end{array}$$

symmetric, $S_{ij} = S_{ji}$



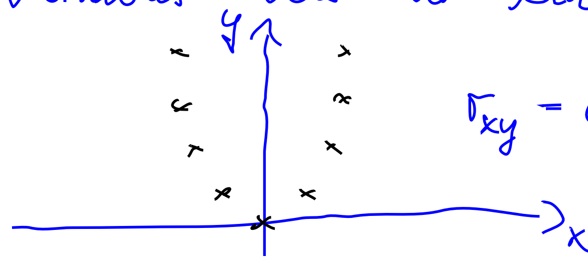
90
80
70
60
50
40
Fertility
0
10
20
30
40
50
Education

Correlation: Normalized covariance

Bravais-Pearson correlation

$$r_{xy} = \frac{s_{xy}}{s_x \, s_y}$$

- between -1 and $+1$
- $r_{xx} = r_{yy} = 1$
- -1: all values are on straight line with negative slope
  $+1$: ... with positive slope
  0: the relation between the two variables has no linear "portion"



$r_{xy} = 0$

- don't interpret correlation from small samples without a scatter plot

The Bravais-Pearson correlation measures the direction and strength of a linear relation between x and y.

The linear relation: (Simple) linear regression

Model: $Y = \underbrace{\beta_0 + \beta_1 \cdot X}_{\text{straight line}} + \varepsilon$

ε: random variation

intercept β_0

slope β_1

Data: determine suitable values ("estimates") $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1

$\Rightarrow$ estimated straight line

The most common estimation method:
Ordinary Least Squares (OLS, LS)

$$\underset{\text{actual } y_i}{y_i} \qquad \underbrace{b_0 + b_1 x_i}_{\text{value of the line based on } x_i \text{ with } b_0 \text{ and } b_1}$$

The OLS $\hat\beta_0$ and $\hat\beta_1$ are obtained from minimizing

$$L(b_0,b_1) = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2 \quad \text{w.r.t. } b_0 \text{ and } b_1$$

Minimization: take derivatives and set equal to 0

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

⋮ $\longrightarrow$ "Regression" for general case (2nd sem.)

. . .

Result:

$$\hat\beta_1 = \frac{s_{xy}}{s_x^2}$$

$$\hat\beta_0 = \bar y - \hat\beta_1 \cdot \bar x$$

$$\hat y_i = \hat\beta_0 + \hat\beta_1 x_i$$

Because of the intercept:

$$\overline{\hat y} = \bar y$$

x_i	y_i	$x_i - \bar x$	$y_i - \bar y$	$\hat y_i$ (from R)	$y_i - \hat y_i$ (errors from R)
2	7	-2	4	$3{,}\overline{09}$	$3{,}\overline{90}$
4	1	0	-2	3	-2
3	2	-1	-1	$3{,}0\overline{45}$	$-1{,}0\overline{45}$
8	4	4	1	$2{,}\overline{81}$	$1{,}\overline{18}$
3	1	-1	-2	$3{,}0\overline{45}$	$-2{,}0\overline{45}$

$\bar{x} = 4 \qquad \bar{y} = 3$

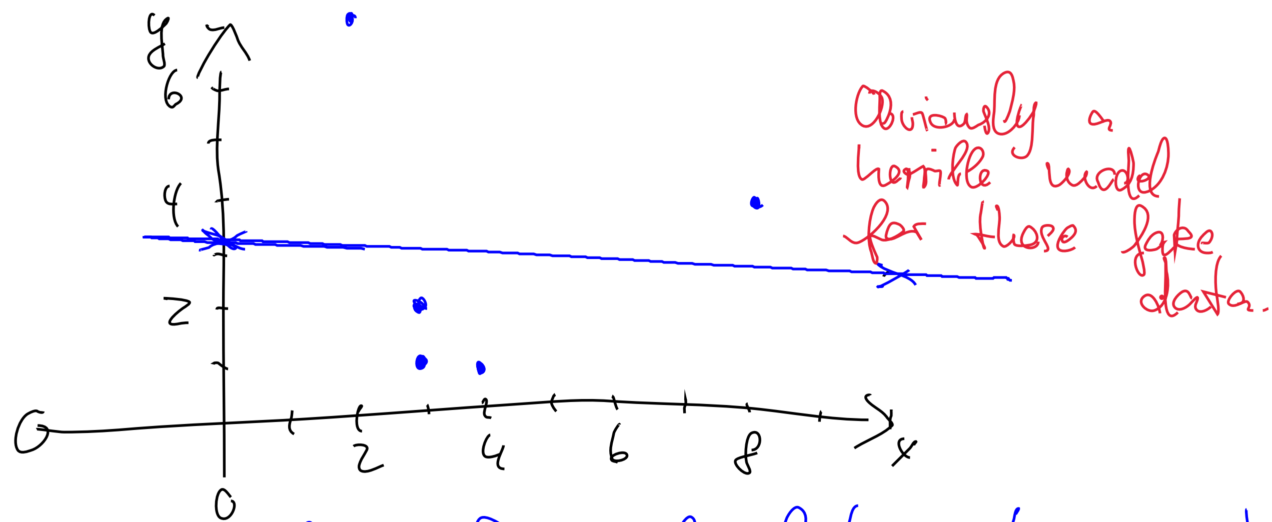
$$s_x^2 = \frac{1}{5-1} \cdot (4+0+1+16+1) = \frac{22}{4} = 5.5$$

$$s_y^2 = \frac{1}{5-1} (16+4+1+1+4) = \frac{26}{4} = 6.5$$

$$s_{xy} = \frac{1}{5-1}\left((-2)\cdot 4 + 0\cdot(-2) + (-1)\cdot(-1) + 4\cdot 1 + (-1)\cdot(-2)\right) = \frac{-1}{4} = -0{,}25$$

$$\hat{\beta}_1 = \frac{-0{,}25}{5.5} = -\frac{1}{22} = -0.0\overline{45}$$

$$\hat{\beta}_0 = 3 - \left(-\frac{1}{22}\right)\cdot 4 = 3 + \frac{4}{22} = 3.\overline{18}$$



Draw line by calculating two points:

$(0, 3.\overline{18})$

$(10, 3.\overline{18} - 10 \cdot 0{,}0\overline{45})$

Quality of fit of a straight line

Decompose the sum of squares of Y into "Model" and "Error":

$$\underbrace{\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\text{Total}} = \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{\text{Model } (\geq 0)} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}_{\text{Error } (\geq 0)}$$

Model should be large
Error " " small

Coefficient of determination R^2:

$$R^2 = \frac{\text{Model SS}}{\text{Total SS}}$$

$R^2 = 1 = 100\%$: perfect

$R^2 = 0$: no explanatory value at all

Convenient: $R^2 = r_{xy}^2$

Correlation and R^2 for example:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{-0,25}{\sqrt{5,5} \cdot \sqrt{6.5}} = -0,0418121$$

$R^2 \approx 0.00175 = 0.175\%$ awful

SS decomposition: (from R)

Total SS = Model SS + Error SS

26 = 0.045 + 25,954