Measures of Bivariate Association

Things to learn: covariance
correlation
location invariance
scale invariance

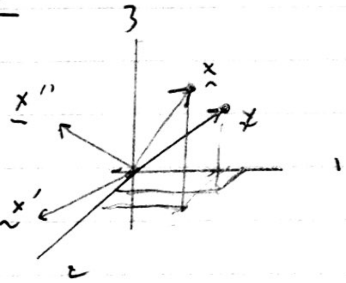
Data: $\{(x_i, y_i) \quad i=1\ldots n\}$ unit level notation
$\{\underset{\sim}{x}, \underset{\sim}{y}\}$ vector notation

$$\underset{\sim}{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \underset{\sim}{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

Questions: How might we measure the "association" between $\underset{\sim}{x}, \underset{\sim}{y}$?

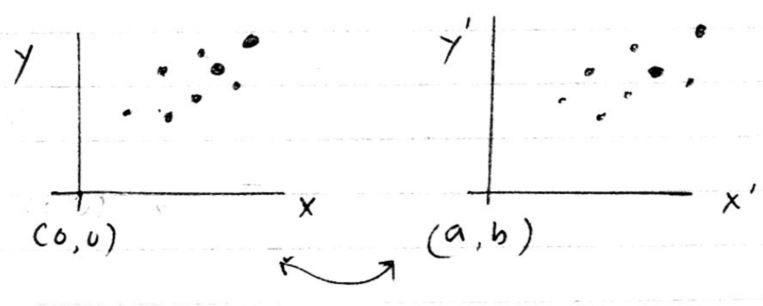
big/pos x_i's occur w/ $\begin{cases} \text{big/pos } y_i\text{'s} \rightarrow \text{"positive association"} \\ \text{small/neg } y_i\text{'s} \rightarrow \text{"negative association"} \end{cases}$

Idea #1
$n=3$



$$\text{assoc}(\underset{\sim}{x}, \underset{\sim}{y}) = \underset{\sim}{x} \cdot \underset{\sim}{y} / n \quad \text{(vector)}$$
$$= \sum x_i y_i / n \quad \text{(summation)}$$
$$= \text{mean}(x * y) \quad (R)$$

Problem:



$$\text{assoc}(\underset{\sim}{x}, \underset{\sim}{y}) \neq \text{assoc}(\underset{\sim}{x}', \underset{\sim}{y}')$$

Remedy #1
$$\text{assoc}(\underset{\sim}{x}, \underset{\sim}{y}) = (\underset{\sim}{x} - \bar{x}\underset{\sim}{1}) \cdot (\underset{\sim}{y} - \bar{y}\underset{\sim}{1}) / n$$
$$= \sum (x_i - \bar{x})(y_i - \bar{y}) / n$$
$$= \text{mean}((x - \text{mean}(x)) * (y - \text{mean}(y)))$$
$$= \frac{n-1}{n} \text{Cov}(\underset{\sim}{x}, \underset{\sim}{y})$$

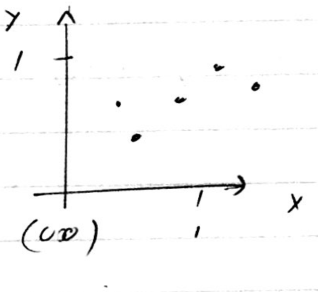
Recall, $\text{Cov}(\underset{\sim}{x}, \underset{\sim}{y}) = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

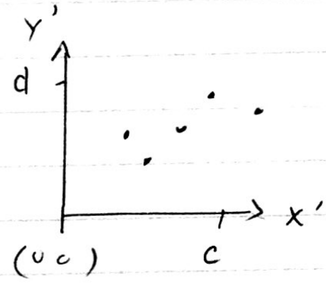
= "sample covariance"

Exercise: Show $\mathrm{cov}(\underset{\sim}{x}, \underset{\sim}{y}) = \mathrm{cov}(\underset{\sim}{x} + a\underset{\sim}{1}, \underset{\sim}{y} + b\underset{\sim}{1}) \quad \forall\, a,b$

$\Rightarrow$ covariance is a location invariant statistic

Problem





$$\mathrm{cov}(\underset{\sim}{x}, \underset{\sim}{y}) \neq \mathrm{cov}(\underset{\sim}{x}', \underset{\sim}{y}')$$

Example: let $\underset{\sim}{x}' = \underset{\sim}{x}$, $\underset{\sim}{y}' = 3\underset{\sim}{y}$ $\qquad \mathrm{cov}(\underset{\sim}{x}, \underset{\sim}{y}) \overset{?}{\lesseqgtr} \mathrm{cov}(\underset{\sim}{x}', \underset{\sim}{y}')$

are $(\underset{\sim}{x}, \underset{\sim}{y})$ "more associated" than $(\underset{\sim}{x}', \underset{\sim}{y}')$?

Observation: Covariance depends on the scale of the measurements

Remedy #2: Let $\underset{\sim}{r}_y = (\underset{\sim}{y} - \bar{y}\underset{\sim}{1})$ $\quad (r_y[i] = y[i] - \bar{y})$

$\underset{\sim}{r}_x = (\underset{\sim}{x} - \bar{x}\underset{\sim}{1})$

Recall $\mathrm{cov}(\underset{\sim}{x}, \underset{\sim}{y}) = \underset{\sim}{r}_y \cdot \underset{\sim}{r}_x / (n-1)$

$= \mathrm{cov}(\underset{\sim}{r}_x, \underset{\sim}{r}_y)$ (why?)

now let $\underset{\sim}{z}_y = (\underset{\sim}{y} - \bar{y}\underset{\sim}{1}) / \|\underset{\sim}{y} - \bar{y}\underset{\sim}{1}\|$ $\quad \left\{ \begin{array}{l} \|\underset{\sim}{y} - \bar{y}\underset{\sim}{1}\| = \left(\Sigma (y_i - \bar{y})^2\right)^{1/2} \\ z_y[i] = \dfrac{y[i] - \bar{y}}{\left(\Sigma (y_i - \bar{y})^2\right)^{1/2}} \end{array} \right.$

$= (\underset{\sim}{y} - \bar{y}\underset{\sim}{1}) / (SYY)^{1/2}$

let $\underset{\sim}{z}_x = (\underset{\sim}{x} - \bar{x}\underset{\sim}{1}) / \|\underset{\sim}{x} - \bar{x}\underset{\sim}{1}\|$

$= (\underset{\sim}{x} - \bar{x}\underset{\sim}{1}) / SXX^{1/2}$

$\underset{\sim}{z}_x$ and $\underset{\sim}{z}_y$ are centered & scaled versions of $\underset{\sim}{x}, \underset{\sim}{y}$.

Check: $\|z_x\|^2 = \sum z_x[i]^2 = 1$

Consider $\text{assoc}(\underset{\sim}{x}, \underset{\sim}{y}) = \underset{\sim}{z}_x \cdot \underset{\sim}{z}_y = \sum z_x[i]\, z_y[i]$

$$= \sum \left(\frac{x_i - \bar{x}}{SXX^{1/2}}\right)\left(\frac{y_i - \bar{y}}{SYY^{1/2}}\right)$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left(\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2\right)^{1/2}}$$

$$= \frac{SXY}{(SXX \cdot SYY)^{1/2}} = \text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) = \text{"sample correlation"}$$

Other forms:

$\text{Var}(\underset{\sim}{x}) = \sum (x_i - \bar{x})^2 / n-1 = SXX / n-1 = s_x^2$

$\text{Var}(\underset{\sim}{y}) = \sum (y_i - \bar{y}) / n-1 = SYY / n-1 = s_y^2$

$\text{Cov}(\underset{\sim}{x}, \underset{\sim}{y}) = \sum (x_i - \bar{x})(y_i - \bar{y}) / n-1 = SXY / n-1$

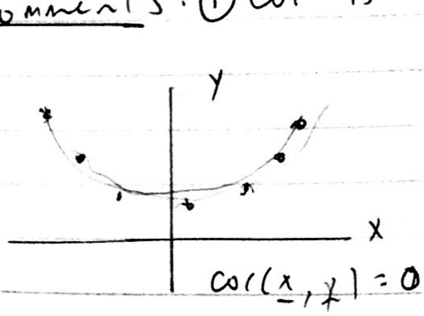
$$\rightarrow \text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) = \frac{\text{Cov}(\underset{\sim}{x}, \underset{\sim}{y})}{(\text{Var}(\underset{\sim}{x})\,\text{Var}(\underset{\sim}{y}))^{1/2}}$$

Exercise: Show $\text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) = \text{Cor}(a\underset{\sim}{x} + b\underset{\sim}{1}, c\underset{\sim}{y} + d\underset{\sim}{1})$ for $a, c > 0$

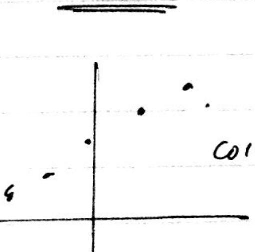
(what happens if $a > 0$, $c = 0$? $c < 0$?)

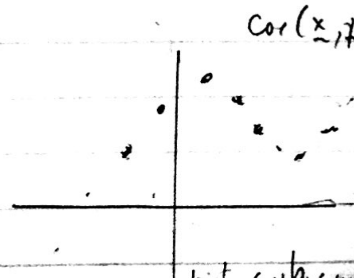
$\Rightarrow$ $\text{Cor}(\underset{\sim}{x}, \underset{\sim}{y})$ is a location and scale invariant measure of association.

Comments: ① Cor is a measure of linear association



but quadratically associated





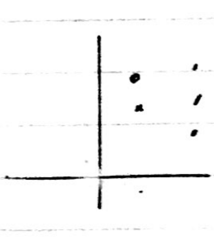
Cor(x, y) = (

but cubically associated

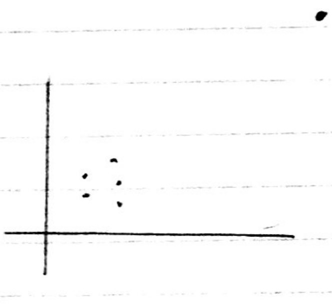
② $\text{Cor}(\underset{\sim}{x}, \underset{\sim}{y})$ is not invariant to monotone transformations of the data:

$$\text{cor}(\underset{\sim}{x}, \underset{\sim}{y}) \neq \text{cor}(\log \underset{\sim}{x}, \log \underset{\sim}{y}) \text{ in general}$$

③ $\text{Cor}(\underset{\sim}{x}, \underset{\sim}{y})$ is not robust to outliers



cor = ?



cor = ?

④ Extremes of correlation: $-1 \leq \text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) \leq 1$ (see Cauchy-Schwarz)

$\text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) = 0 \Rightarrow$ "$\underset{\sim}{x}, \underset{\sim}{y}$ are uncorrelated" (rarely happens in a sample)

$\text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) = 1 \iff y_i = \beta_0 + \beta_1 x_i, \; i = 1 \ldots n$, some $\beta_0, \beta_1 > 0$

$= -1 \iff$ " some $\beta_0, \beta_1 < 0$

This leads to the idea of a linear regression model

$$\text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) = r \iff y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

magnitude of deviations depend on r