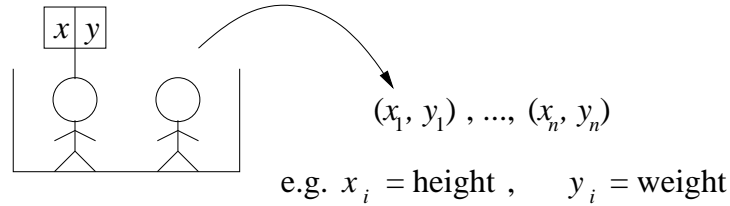


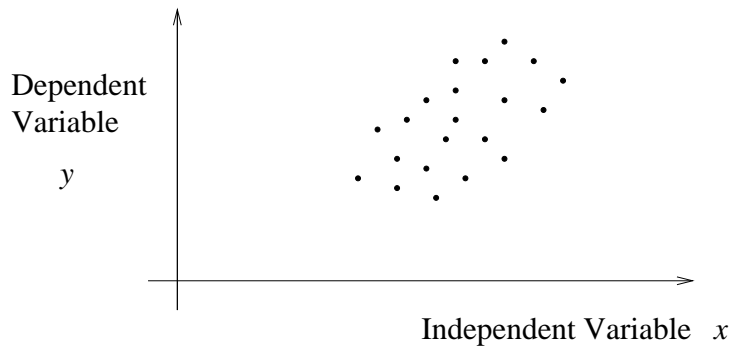
Lecture 9: Correlation

FPP: ch. 8-9, Nolan & Speed: ch. 7



Q: How are x and y related?

1. The scatter plot — Most informative for 2 variables.



2. Correlation coefficient to measure LINEAR relationship.

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

= Average of cross-products of standardized deviations of x and y from their average.

$$SD(x) = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \quad SD(y) = \sqrt{\frac{1}{n} \sum (y_i - \bar{y})^2}$$

Properties:

1. $r_{xy} = r_{yx}$

2. $-1 \leq r_{xy} \leq 1$; r has no units

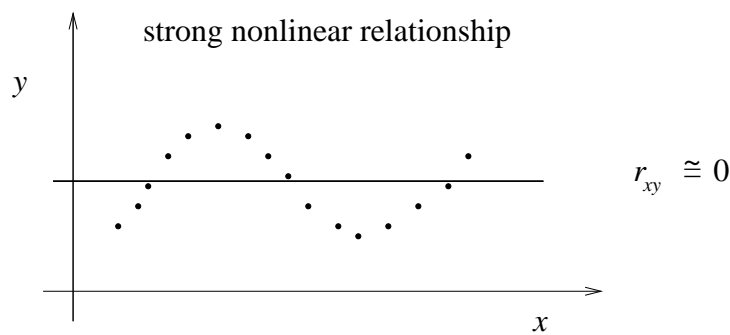
3. If $v_i = a + bx_i$, $w_i = c + dy_i$,

$$r_{vw} = \frac{b}{|b|} \frac{d}{|d|} r_{xy}$$

4. If $y_i = a + bx_i$,

$$r_{xy} = \begin{cases} 1 & b > 0 \\ 0 & b = 0 \\ -1 & b < 0 \end{cases}$$

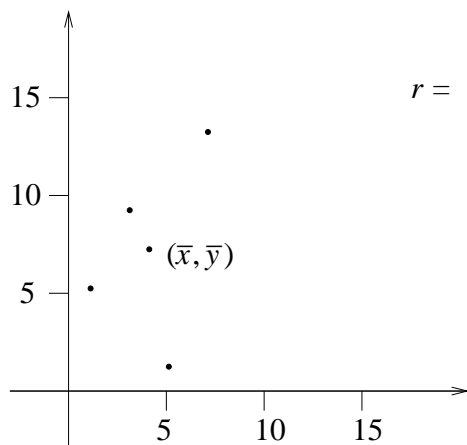
r doesn't measure "nonlinear" relationship.



How to compute r_{xy} ?

$$r = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)} \quad \text{where} \quad \text{Cov}(X, Y) = \frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}$$

e.g.



$\bar{x} = 4$ $\text{SD}(x) = 2$
 $\bar{y} = 7$ $\text{SD}(y) = 4$

x	y	$\frac{x - \bar{x}}{\text{SD}(x)}$	$\frac{y - \bar{y}}{\text{SD}(y)}$	
1	5	-1.5	-0.5	0.75
3	9	-0.5	0.5	-0.75
4	7	0.0	0.0	0.00
5	1	0.5	-1.5	-0.75
7	13	1.5	1.5	2.25

$$r = \frac{1}{5}(0.75 - 0.25 + 0.00 - 0.75 + 2.25) = 0.4$$

SD-line: “center” of the data cluster

1. it passes (\bar{x}, \bar{y})
2. it has slope $\frac{SD(y)}{SD(x)} \frac{r}{|r|}$.

Interpretations:

1. the closer $|r|$ is to 1, the easier to predict y from x or vice versa.
2. $r > 0$, positive association. The bigger x , the bigger y on average.
3. $r < 0$, negative association. The bigger x , the smaller y on average.
4. Correlation \neq causation, “independent” variable doesn’t have to “cause” the dependent variable.

Prediction based on association

Ideal case for linear prediction from x to y (or y to x): the scatter plot shows a football shape, e.g. height-weight, data set of Galton.

Summary: $\bar{x} = 70$ inches $SD(x) = 3$ inches
 $\bar{y} = 162$ lbs $SD(y) = 30$ lbs
 $r = 0.47$

Given someone’s height, how would you guess the weight using the information in the data set?

Suppose you have access to every data point in the set, what is a good guess?

Take a vertical strip around $x = x_0$, get the average height in that strip, and this is the guessed or predicted weight of a man with height $x = x_0$.

If we draw many such vertical strips at different values of height, the averages of strips happen to fall on a line (more or less): This is called the REGRESSION LINE.

- Both the regression line and SD line pass through (\bar{x}, \bar{y})
- Regression line is less steep than the SD line

$$\begin{aligned}\text{slope of SD line} &= \frac{r}{|r|} \frac{SD(y)}{SD(x)} \\ \text{slope of regression line} &= r \frac{SD(y)}{SD(x)}\end{aligned}$$

Why this “ r ” factor? Caused by spread of data points around the SD line.

When $|r| = 1$, they are the same! When $r \cong 0$, the regression line is almost horizontal passing through (\bar{x}, \bar{y}) so guess is roughly \bar{y} , or knowing x_0 is not so helpful.

Should we use the same line to predict x from y ? Horizontal strips lead to another regression line with slope $r \frac{SD(x)}{SD(y)}$. Predicted weight is different from the actual weight. How much variability should we expect? \rightarrow next lecture.

