STA 302 / 1001

Note Title 10/4/2011

Reminder Andrey in RW 107/109 today 11:00 - 13:00

in RW 109 Friday 13:00 - 15:00

Test date

Thurs Oct 20 10:10 - 11:40 26

Tues Oct 25 10:10 - 11:40

X, Y both random variables

BIVARIATE NORMAL DISTRIBUTION

X, Y are jointly normally distributed if their joint density function is

$$f(x,y) = \frac{1}{\pi \sigma_x \sigma_y \sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x} \right)\left(\frac{y-\mu_y}{\sigma_y} \right) + \left(\frac{y-\mu_y}{\sigma_y} \right)^2 \right] \right\},$$

$$-\infty < x < \infty, \quad -\infty < y < \infty$$

Can show (STA 257 exercise) marginal distributions:

$$X \sim N(\mu_x, \sigma_x^2), \quad Y \sim N(\mu_y, \sigma_y^2)$$

ρ "rho" is the CORRELATION COEFFICIENT between X and Y

i.e. $$\rho = \frac{E[(X-EX)(Y-EY)]}{\sqrt{Var(X)\,Var(Y)}}$$

It can be shown (STA 257 exercise) that the conditional distribution of Y given $X = x$

$$Y|X=x \sim N\left(\mu_Y + \sigma_Y \rho \frac{x-\mu_X}{\sigma_X},\ (1-\rho^2)\sigma_Y^2\right)$$

- conditional distribution of $Y|X=x$
- $E(Y|X=x)$ is of the form $\beta_0 + \beta_1 x$
- Variance is constant over x

Estimate mean of $Y|X=x$ by estimating

μ_X by $\bar{x}$, μ_Y by $\bar{y}$, σ_X by $\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2} = s.d.(x)$

σ_Y by $\sqrt{\frac{1}{n-1}\sum(y_i-\bar{y})^2} = s.d.(y)$

and ρ by $\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y}) / n-1}{s.d.(x)\ s.d.(y)}$

You get (check this)

est. of $E(Y \mid X = x) = b_0 + b_1 x$

where $b_0 = \bar{y} - b_1 \bar{x}$, $b_1 = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

i.e. the estimate gives the Least Squares estimates of the slope and intercept

CORRELATION

Suppose you have 2 r.v.'s X and Y

Interested in relationship between X & Y, believe its linear

- If there is a clear choice for Y being the response and X being the explanatory variable, can carry out a regression
- An option when X and Y don't have a clear choice for dependent/independent variables, can summarize the strength of the linear relation by correlation

- Correlation is a symmetric measure

n observations (x_i, y_i) of r.v.'s X and Y

Correlation (Pearson's Product Moment Correlation)

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

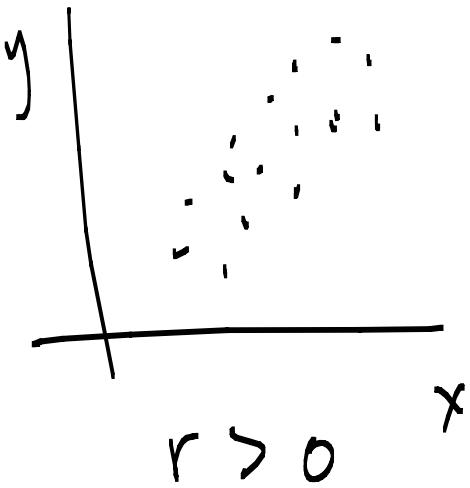
When X, Y have a bivariate normal distribution,
r is the maximum likelihood estimate of ρ

Facts about r:

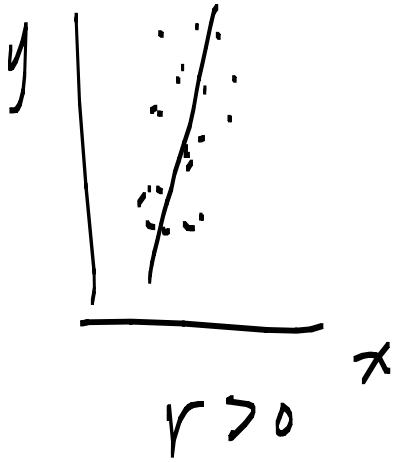
- measure of degree of linear association between X & Y
- it is dimension free
- always between +1 and -1, inclusive
  - $r = +1$ means (x_i, y_i) fall exactly on a straight line with positive relationship
  - $r = -1$ " " " " " " " " negative relationship
  - $r = 0$ means no linear relationship

SAS — proc corr ; gives correlations and the p-value for test: $H_0: \rho = 0$ versus $H_a: \rho \neq 0$

assuming X, Y bivariate normal

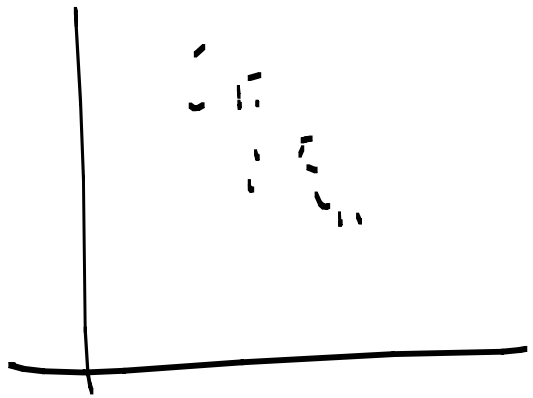


$r > 0$

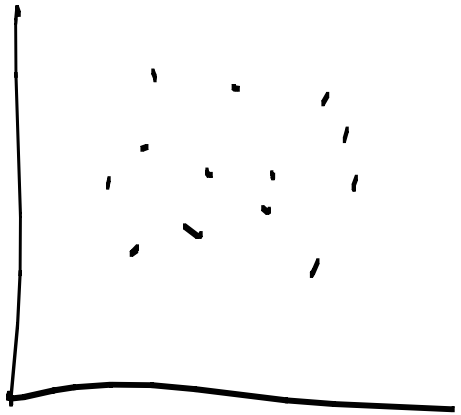


$r > 0$

These 2 have similar values of r.



$r < 0$



$r = 0$

$r = 0$

x

$r > 0$

weak

$-1 < r < 0$

Not a linear relationship

so r not applicable

Relationship between R^2 and r

$$R^2 = r^2$$

$$R^2 = \frac{SSReg}{SST} = \frac{b_1^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2}$$

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

$$= r^2$$

$$r = \frac{S_{xy}}{S_x S_y}$$

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

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

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

$$r = \frac{S_{xy}}{S_x S_y}$$

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

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

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

Units of b_1 are "y per unit x"

Example: CFCs pre-MP $b_1 = 9.71152$

Interpretation of b_1: indicates for every increase in time (x) by 1 year, the estimate of the mean of CFC concentration increases by 9.71152 parts per trillion

Exercise: Know: $\sum \hat{e}_i x_i = 0$ and $\sum \hat{e}_i \hat{y}_i = 0$

Why does this imply $\hat{e}_i, x_i$ are uncorrelated?
and $\hat{e}_i, \hat{y}_i$ are uncorrelated?