

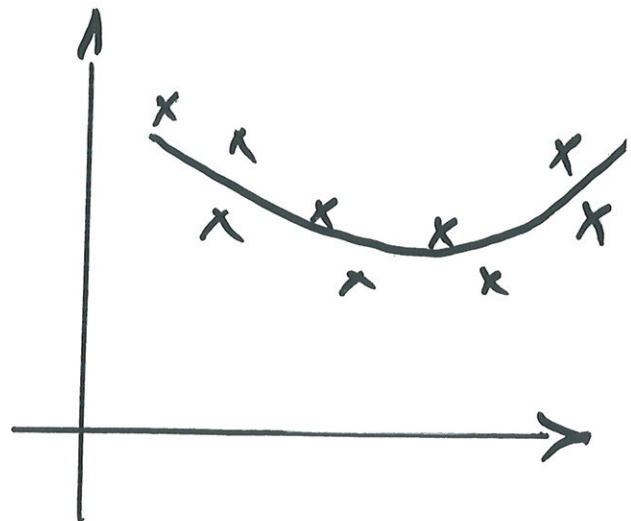
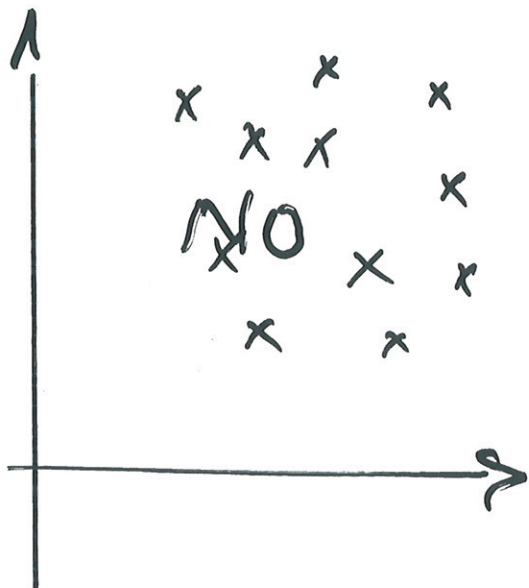
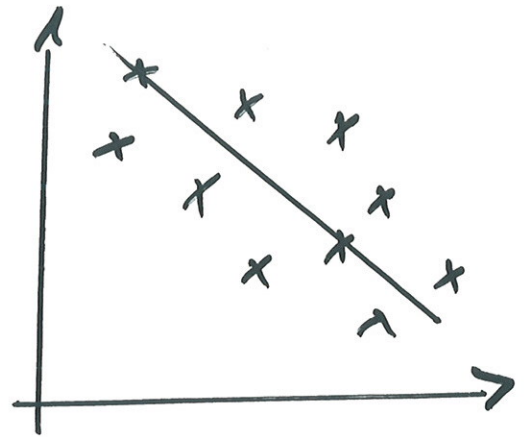
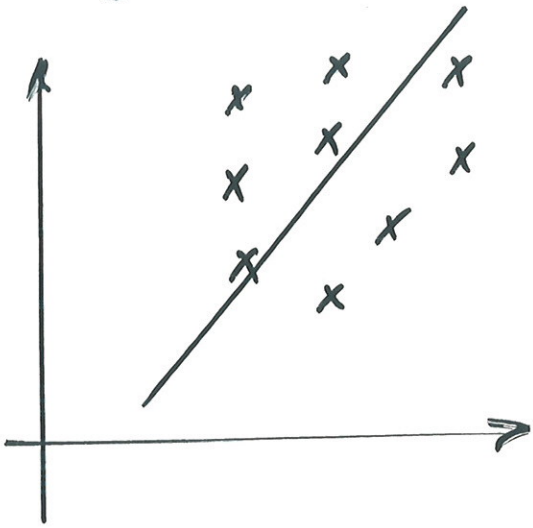
# CORRELATION

LECTURE 10/11

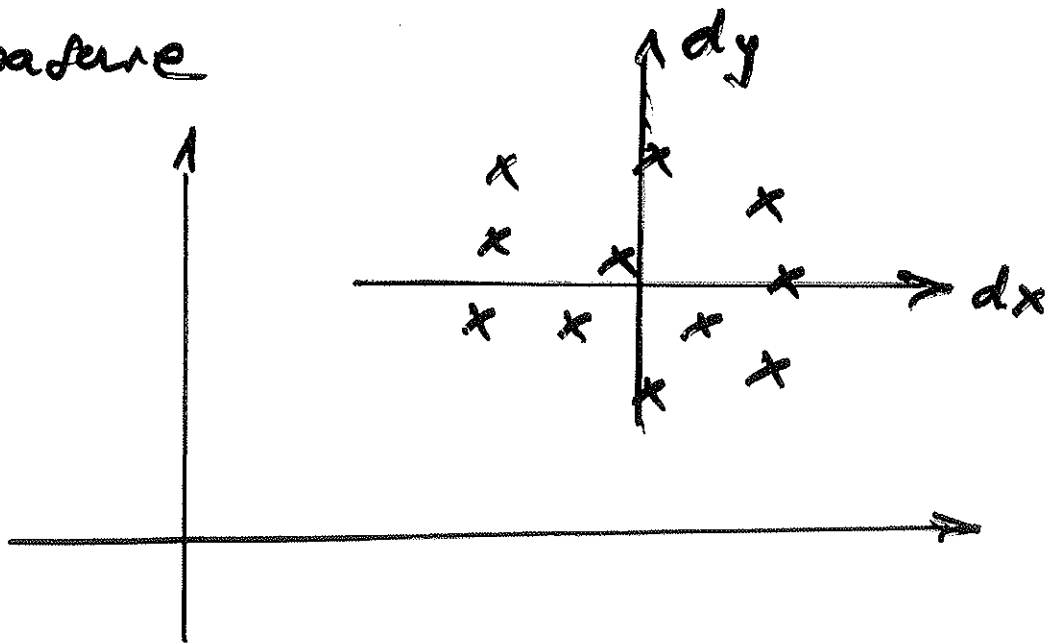
P42-44

Two random variables, e.g. height  $H$ , and weight  $W$  are not functionally related, but they can be correlated. This means that a tendency in one is marked by a tendency in the other.

We plot sets of random observations of two random variables  $X, Y$  on a scatter diagram. Regression curves are shown.



To establish the amount of correlation we measure



$$\text{where } dx = x_i - \bar{x}$$
$$dy = y_i - \bar{y}$$

$\sum dx dy$  or sum of products of deviations from means gives a measure. We normalise by dividing by  $\sqrt{\sum dx^2 \sum dy^2}$  and define

$$r = \frac{\sum dx dy}{\sqrt{\sum dx^2 \sum dy^2}}$$

It can be proved that  $-1 \leq r \leq 1$  with  $r$  taking the values  $\pm 1$  when there is an exact straight line relationship. When  $r = 0$   $X$  and  $Y$  are uncorrelated.

$r$  is obtained from the data and is called the sample correlation coefficient

Note that  $r$ , like  $\bar{x}$  and  $s^2$  is a sample based statistical parameter.

If  $X$  and  $Y$  are the random variables of the parent population, e.g. height and weight then we can define  $\rho$  to be the theoretical correlation coefficient,

$$\rho = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\left\{ E[X - \mu_x]^2 E[Y - \mu_y]^2 \right\}^{1/2}}$$
$$= \frac{\text{Covariance}[X, Y]}{\left\{ \text{Var}[X] \text{Var}[Y] \right\}^{1/2}} = \frac{\sigma_{XY}}{\sigma_x \sigma_y}$$

$\rho = 0$  when  $X, Y$  are independent

As with  $\bar{x}$  and  $s^2$  vis-à-vis  $\mu$  and  $\sigma^2$ , we can validate as to whether the sample estimate  $r$  is a good estimate of  $\rho_0$ .

Set up the null hypothesis

$H_0: \rho = \rho_0$  versus  $H_1: \rho \neq \rho_0$  by

Computing  $\frac{1}{2} \ln \left[ \frac{1+r}{1-r} \right]$  because

$$Z = \frac{\sqrt{n-3}}{2} \left[ \ln \left[ \frac{1+r}{1-r} \right] - \ln \left[ \frac{1+\rho_0}{1-\rho_0} \right] \right]$$

Can be shown to be distributed  $N(0,1)$ .

If we wish we can choose either a one-tail or two-tail test.

## Significance of the Correlation Coefficient

$H_0: \rho_0 = 0$ , say

$H_1: \rho_0 \neq 0$

$$\text{So } Z = \frac{\sqrt{n-3}}{2} \ln \left( \frac{1+r}{1-r} \right)$$

So  $r = \tanh \frac{Z}{\sqrt{n-3}}$ . So if  $|Z_{crit}| = 1.96$   
for  $n=10$ ,  $r_{crit} = 0.63$   
 $n=20$ ,  $r_{crit} = 0.44$

| Daily rainfall X (cm)                           | 4.3 | 4.5 | 5.9 | 5.6 | 6.1 | 5.2 | 3.8 | 2.1 | 7.5 |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Particles removed Y (microgm. m <sup>-3</sup> ) | 126 | 121 | 116 | 118 | 114 | 118 | 132 | 141 | 108 |

$$\sum (x_i - \bar{x})^2 = 19.26, \quad \sum (y_i - \bar{y})^2 = 804.2, \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = -121.8$$

$$\text{So } r = \frac{-121.8}{\sqrt{19.26 \times 804.2}} = -0.9786$$

We assume no correlation between rainfall volume and particulate deposit, i.e.

$$H_0: \rho = 0, \quad H_1: \rho \neq 0, \quad \text{with } |Z_{crit}| = 1.96$$

$$\text{Computing: } Z = \frac{\sqrt{n-3} \ln\left(\frac{1+r}{1-r}\right)}{2} = \frac{\sqrt{6} \ln\left(\frac{0.0214}{1.9786}\right)}{2} = -5.54.$$

So accept  $H_1$  (by a marked margin)

For revision, repeat with  $\rho_0 = -0.95, -0.995$ .

Example (theoretical correlation)

If  $Z = X + Y$  Q. book p 21

$$\text{the } E[Z] = E[X] + E[Y]$$

$$\text{Var}[Z] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y]$$

Let us call  $\text{Cov}[X, Y]$ ,  $\sigma_{XY}$

$$\begin{aligned} \text{so } \sigma_Z^2 &= \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \\ &= \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y \end{aligned}$$

In the example, if  $\sigma_X = 1 \text{ min}$ ,  $\sigma_Y = 2 \text{ min}$   
then  $\sigma_Z^2 = 1^2 + 2^2 + 2 \times 0.5 \times 1 \times 2 = 7$   $\rho = 0.5$

$$\text{i.e. } \sigma_Z = \sqrt{7}$$

Given  $E[X] = 10$ , and  $E[Y] = 20$ ,

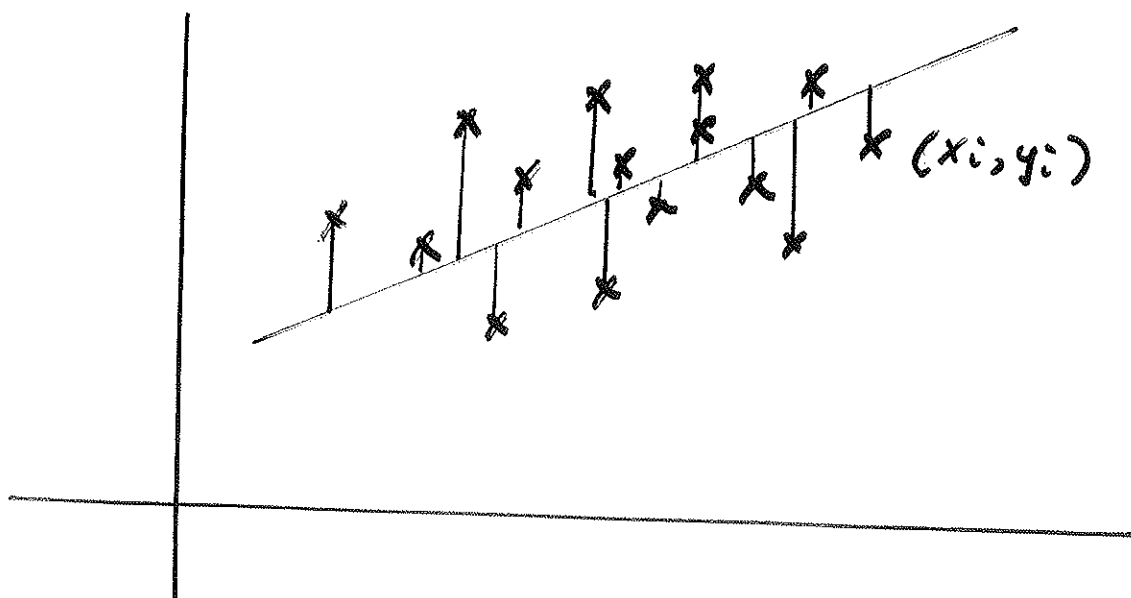
$$\begin{aligned} P(\text{Running time} \geq 33 \text{ min}) &= P(Z \geq 1/\sqrt{7}) \\ &= 0.35 \end{aligned}$$

where  $Z$  is  $N(0, 1)$ .

# REGRESSION

p46-50

Strictly speaking one might argue as to whether this is a statistical topic at all as we are measuring the best fit of a line or curve to a scatter diagram



In practice we minimise the sums of squares of the vertical distances  $e_i$  from the points  $(x_i, y_i)$  to the regression curve. If this is to be taken as the straight line  $y = \alpha + \beta x$  then

$$y_i = \alpha + \beta x_i + e_i$$

and we minimise  $\sum e_i^2$  with respect to  $\alpha$  and  $\beta$ , the line's intercept and gradient.

In other words take

$$S = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

with the minimum satisfying  $\partial S / \partial \alpha = \partial S / \partial \beta = 0$

Evaluating the partial derivatives leads to two linear equations in  $\alpha$  and  $\beta$ , i.e.

$$n\alpha + \sum x_i \beta = \sum y_i$$

$$\sum x_i \alpha + \sum x_i^2 \beta = \sum x_i y_i$$

These two equations are called the Normal Equations and the straight line is called the line of regression of  $y$  on  $x$ .

If instead we were to minimise the sum of squares of the horizontal distances we would generally get a different line, the two meeting at  $(\bar{x}, \bar{y})$ .

## Example - Rod Length & Temperature - p 47/8

$$\sum x_i = 200, \quad \sum y_i = 540, \quad \sum x_i y_i = 21,840$$

$$\sum x_i^2 = 9000$$

$$\hat{\alpha} = \frac{540 \times 9000 - 21,840 \times 200}{5 \times 9000 - (200)^2} = 98.4$$

$$\hat{\beta} = \frac{5 \times 21,840 - 540 \times 200}{5 \times 9000 - (200)^2} = 0.24$$

---

## Example - Regression Lines, - p 50

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 602/3$$

$$\sum (x_i - \bar{x})^2 = \frac{767}{3}, \quad \sum (y_i - \bar{y})^2 = \frac{539}{3}$$

$$\hat{\beta}_{YX} = 0.7849, \quad \hat{\alpha}_{YX} = 20.825$$

$$\text{so } y = 20.825 + 0.7849x$$

$$\text{and } \hat{\beta}_{XY} = 1.1169, \quad \hat{\alpha}_{XY} = -11.891$$

$$\text{so } x = -11.891 - 1.1169y$$

$$r^2 = 0.8766, \quad r = 0.936.$$

## Reduction to Linear Form

If given values of pressure and volume for the adiabatic expansion of a gas, i.e.

$$P V^{\gamma} = c$$

one can determine values of  $\gamma$  and  $c$  for suitable regression fit.

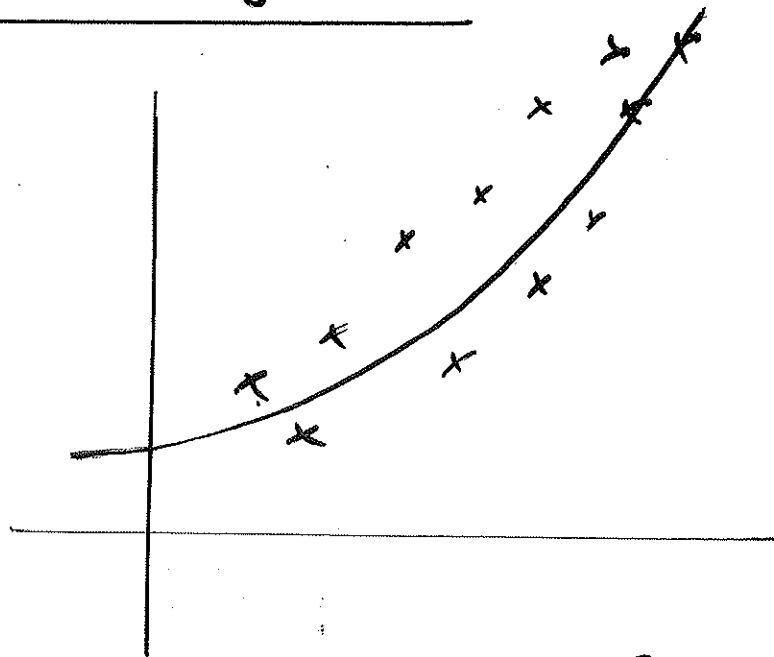
Take logs,

$$\ln P + \gamma \ln V = \ln c$$

This gives the option to determine a line of best fit involving  $\gamma$  and  $\ln c$ .

Note  $c > 0$ , why?

# Quadratic Regression



Fitting  $y = \alpha + \beta x + \gamma x^2$

need us to minimize  $S = \sum (y_i - \alpha - \beta x_i - \gamma x_i^2)^2$

by setting  $\frac{\partial S}{\partial \alpha} = \frac{\partial S}{\partial \beta} = \frac{\partial S}{\partial \gamma} = 0$

This leads to three equations in three unknowns

$$n\alpha + \sum x_i \beta + \sum x_i^2 \gamma = \sum y_i$$

$$\sum x_i \alpha + \sum x_i^2 \beta + \sum x_i^3 \gamma = \sum x_i y_i$$

$$\sum x_i^2 \alpha + \sum x_i^3 \beta + \sum x_i^4 \gamma = \sum x_i^2 y_i$$

## Regression Matrix - Case of ill-conditioning

In polynomial regression matrices very widely varying numerical values can arise, e.g. due to the  $\sum x_i^4$  in quadratic regression.

If the numbers  $x_i \sim 0(10)$ , then  $\sum x_i^4 \sim 0(10^5)$  or worse.

How do we deal with it?

e.g.

$$\begin{array}{l} 1.21x + 10.37y + 126z = 55.3 \\ \times 0.1 \quad 27.3x + 91.05y + 1027z = 495 \\ \times 0.01 \quad 315x + 2161y + 10^4 z = 6175 \end{array}$$

to get

$$\begin{array}{l} 1.21x + 10.37y + 126z = 55.3 \\ 2.73x + 9.105y + 102.7z = 49.5 \\ 3.15x + 2.161y + 100z = 61.75 \end{array}$$

Now set  $y' = 10y$ ,  $z' = 100z$  ( $x' = x$ )

$$\begin{array}{l} 1.21x' + 1.037y' + 1.26z' = 55.3 \\ 2.73x' + 0.9105y' + 1.027z' = 49.5 \\ 3.15x' + 2.161y' + z' = 61.75 \end{array}$$

Remember

Any scaling is permissible provided it is consistent.