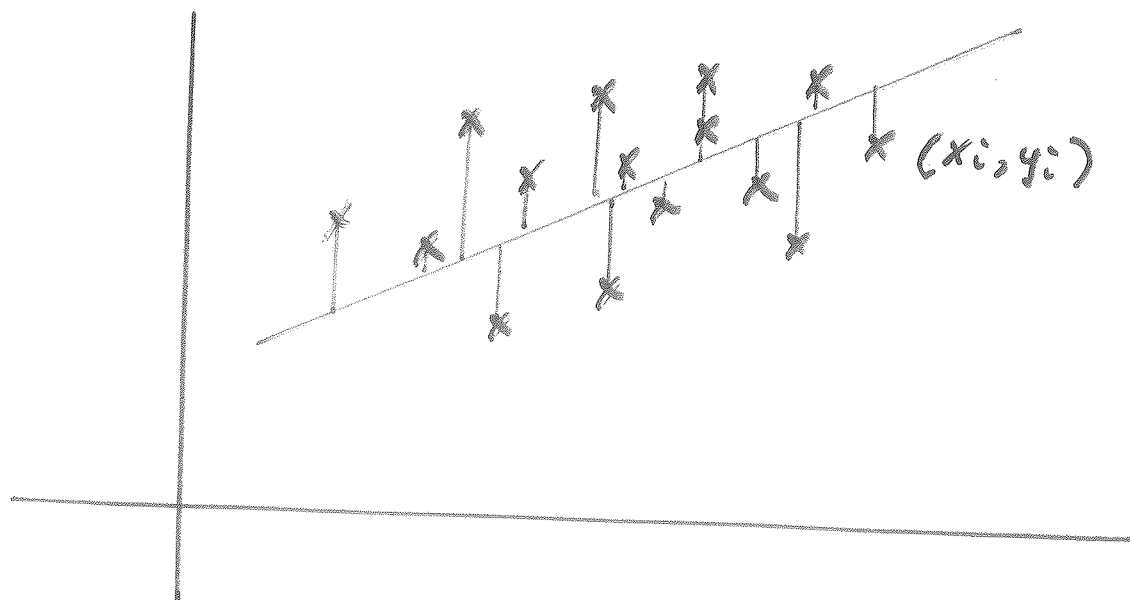


REGRESSION

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 α and β , 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 α and β , 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 minimize the sums 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 γ and c for suitable regression fit.

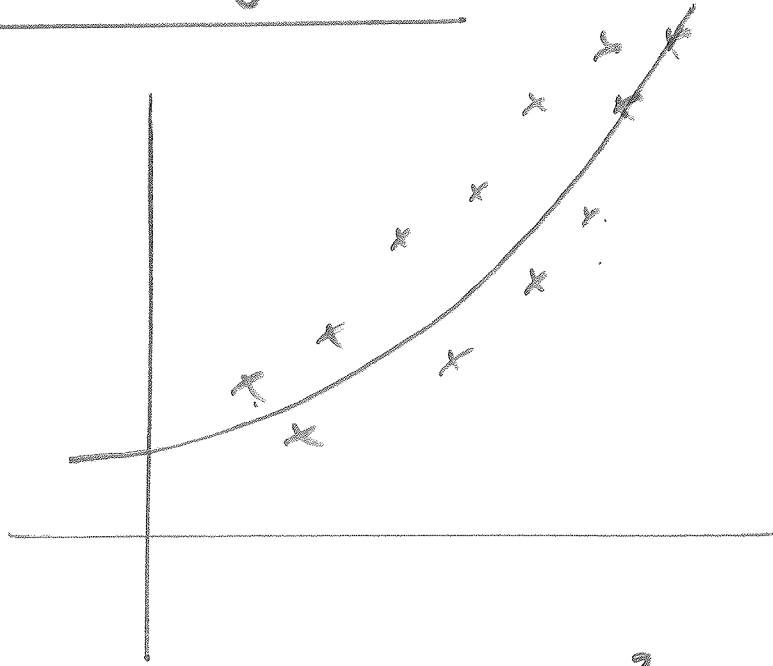
Take logs,

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

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

Note $c > 0$, why?

Quadratic Regression



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

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

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

$$\begin{array}{l} 1.21x + 10.37y + 126z = 55.3 \\ 2.73x + 91.05y + 102.7z = 49.5 \\ 3.15x + 21.61y + 100z = 61.75 \end{array}$$

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.