MA500 Geometric Foundations of Data Analysis

Geometry: concerns distance & distance preserving transformations.

Statistics: largely concerns inferences about a population based on samples from the population.

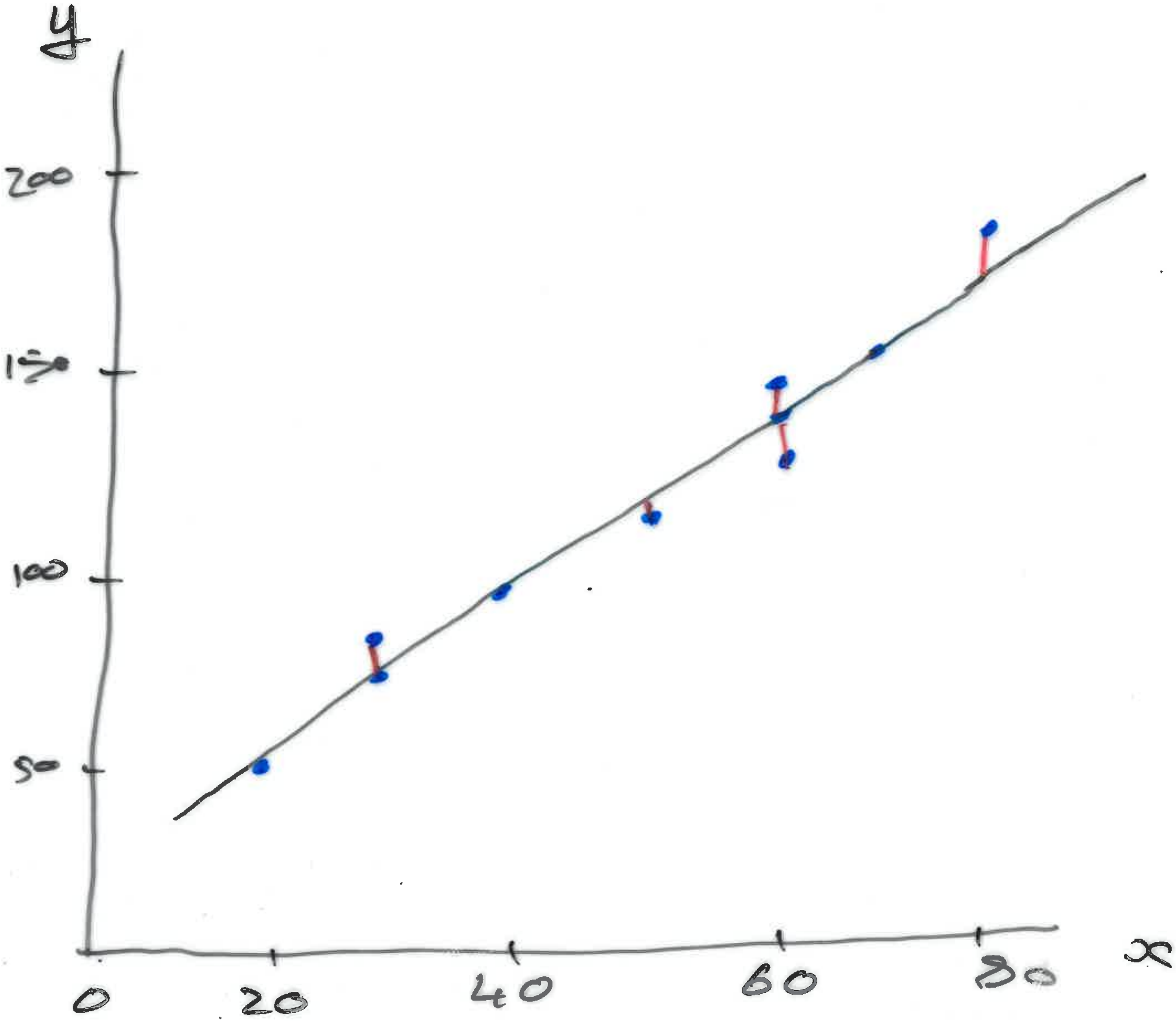
(**probability**: concerns inference about a sample based on knowledge of the population.)

Data Analysis: Concerns the discovery and communication of meaningful patterns in data. Unlike statistics, it often deals with analyses where there is no assumed null hypothesis. It often favours visualization to communicate insight.

1. Least Squares Fitting

Consider a company that manufactures a spare part once per month in lots which vary in size according to demand.

Production run i	Lot Size x_i	Man-hours y_i
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132



Insight into the relationship between lot size and man-hours can be gained by "fitting" a straight line to this data.

The fitted line is represented by

$$y = b_0 + b_1 x$$

where b_0, b_1 are chosen to be "best" in the following sense: they should minimize

$$Q = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

where $n = 10$, x_i, y_i are given in above table.

Here $Q = Q(b_0, b_1)$ is a function of b_0 and b_1.

For a minimum we want

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2 \displaystyle\sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right) = 0 \\[2ex] \dfrac{\partial Q}{\partial b_1} = -2 \displaystyle\sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right) x_i = 0 \end{cases}$$

(*) are called the normal equations

The can be rewritten as

$$(*) \begin{cases} n b_0 + b_1 \sum x_i = \sum y_i \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

Two equations in two unknowns b_0, b_1.

The solution is:

$$b_0 = 10.0$$
$$b_1 = 2.0$$

and the fitted line is

$$y = 10 + 2x$$

so we "estimate" that the mean number of man-hours increases by two hours for each unit increase in lot size.

Matrix Notation

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{pmatrix} \qquad B = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$$

(*) becomes

$$(*) \qquad X^t X B = X^t Y$$

Hence

$$\boxed{B = (X^t X)^{-1} X^t Y}$$