

# Linear statistical models

## Introduction

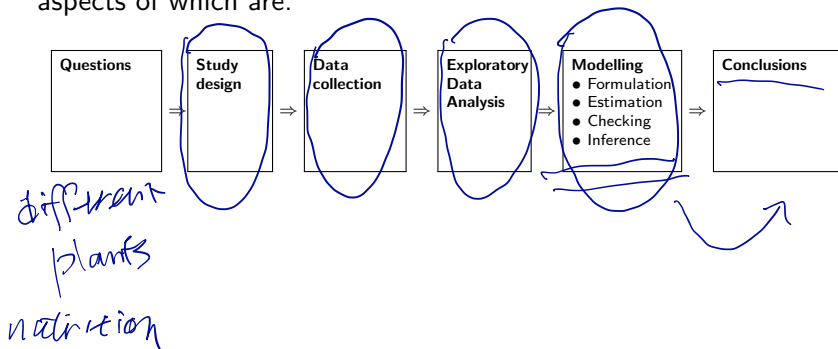
Yao-ban Chan

*I keep six honest serving-men  
(They taught me all I knew);  
Their names are What and Why and When  
And How and Where and Who.*

*Rudyard Kipling*

# Statistics

Statistics is a collection of tools for quantitative research, the main aspects of which are:



# What is a linear model?

A linear model is one of many types of models that we can use in the modelling phase.

It assumes that the data we measure have some sort of linear relationship to other explanatory sets of data (give or take a small amount of error).

Many of you will have seen at least one kind of linear model: linear regression. However linear models are much more flexible than that.

MAST 20005 { Module 5  
Simple Linear regression

# What is a linear model?

Generally speaking the linear model is the 'nicest' model we can use:

- ▶ It is easy to analyse;
- ▶ It makes certain assumptions which are not too strict;
- ▶ It encompasses many situations;
- ▶ It is also very flexible.

# The general linear model

- ▶ We have  $n$  subjects, labelled 1 to  $n$ ;
- ▶ We wish to analyse or predict the behaviour of a measurement or property of the subject ( $y$  variable), denoted by

1st subject  $\swarrow$   
 $y_1, y_2, \dots, y_n$   
 $\searrow$  2nd subject

- ▶ Each subject has certain other properties that we know or have pre-determined ( $x$  variables). Subject  $i$  has  $k$  of these properties:  $x_{i1}, x_{i2}, \dots, x_{ik}$ .

# The general linear model

The general linear model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

for all  $i = 1, 2, \dots, n$ .

We call  $y$  the *response variable* and the  $x$ 's the *design constants* (technically, they are not variable). The  $\beta$ 's are *parameters* of the model, and  $\varepsilon$  is an *error* term.

# The general linear model

$$SS(TO) > 0$$

The model attempts to explain the variation in the measured  $y$ 's (if there were no variation then the data would be rather boring!).

However, not all variation can be explained by deterministic data alone (and if it could, the data would again be pretty boring!).

There will always be an error term:  $\varepsilon$ .

To complete the model, we need the distribution of the  $\varepsilon$ 's. For example, they are often supposed to be i.i.d. normal with mean 0 and variance  $\sigma^2$ .

## Matrix formulation

Matrix, vector

Model:  $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \varepsilon_i$  slide 8

We can express the general linear model in matrix form:

subject 1  
subject 2

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

 $\mathbf{y}$  $=$  $\mathbf{X}$  $\beta$  $+$  $\varepsilon$ 

parameter



# Matrix formulation

Note the dimensions of the matrices:

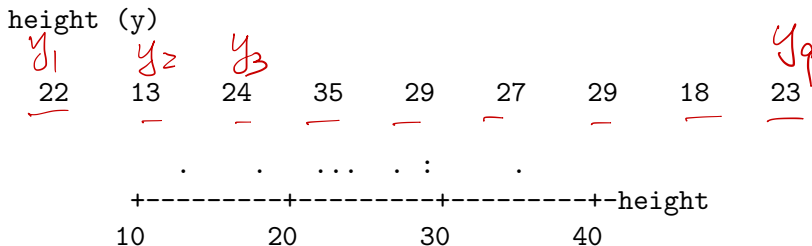
- ▶  $\mathbf{y}$  is  $n \times 1$ ;
- ▶  $\mathbf{X}$  is  $n \times (k + 1)$ ;
- ▶  $\boldsymbol{\beta}$  is  $(k + 1) \times 1$ ; and
- ▶  $\boldsymbol{\epsilon}$  is  $n \times 1$ .

*$n$ : # of subjects*

# Plant data

We study the heights of 9 plants.

## Case 1. No other information.



# Plant data

**Model:**

$$y_i = \mu + \varepsilon_i$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_9 \end{bmatrix}$$

=

$$\begin{bmatrix} 22 \\ 13 \\ 24 \\ 35 \\ 29 \\ 27 \\ 29 \\ 18 \\ 23 \end{bmatrix}$$

=

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$[\mu]$

+

$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_9 \end{bmatrix}$$

$y$

$X$

$\beta$

+

$\varepsilon$

# Plant data

## Case 2. Soil moisture ( $x$ ) given.

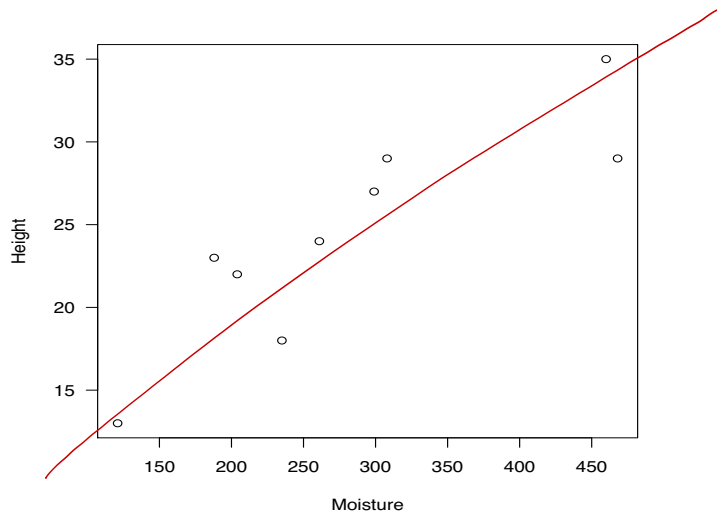
subject  
subject

1  
2

Moisture ( $x$ )	Height ( $y$ )
204	22
121	13
261	24
460	35
468	29
299	27
308	29
235	18
188	23

variety 1  
variety 2  
variety 3

# Plant data



## Plant data

**Model:**  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  (simple linear regression)

$$\begin{array}{c}
 y_i = \\
 \begin{bmatrix} 22 \\ 13 \\ 24 \\ 35 \\ 29 \\ 27 \\ 29 \\ 18 \\ 23 \end{bmatrix}
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} 1 & 204 \\ 1 & 121 \\ 1 & 261 \\ 1 & 460 \\ 1 & 468 \\ 1 & 299 \\ 1 & 308 \\ 1 & 235 \\ 1 & 188 \end{bmatrix}
 \end{array}
 +
 \begin{array}{c}
 \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_9 \end{bmatrix}
 \end{array}$$

*Handwritten notes:*  
 - "2nd row" points to the second row of the design matrix  $X$ .  
 - "9th row" points to the ninth row of the design matrix  $X$ .

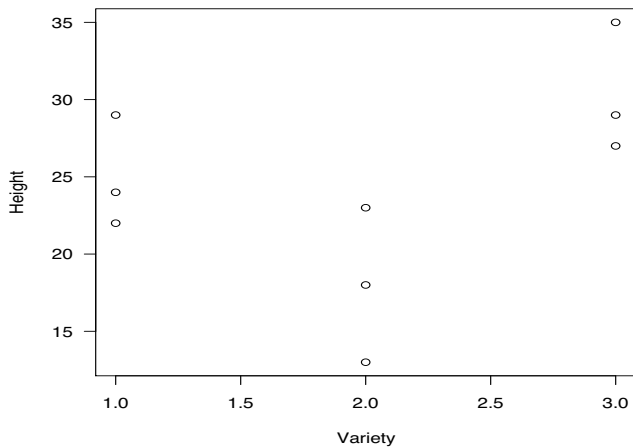
$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

# Plant data

## Case 3. Three varieties.

Variety		
1	2	3
22	13	27
24	18	29
29	23	35

# Plant data





## Plant data

 $\mu_i$  the height of variety  $i$ **Model I:**  $y_{ij} = \mu_i + \varepsilon_{ij}$  (one-way ANOVA) $H_{02} \mu_1 = \mu_2 = \mu_3$ 

$$\begin{bmatrix} y_{1,1} \\ y_{1,2} \\ y_{1,3} \\ y_{2,1} \\ y_{2,2} \\ y_{2,3} \\ y_{3,1} \\ y_{3,2} \\ y_{3,3} \end{bmatrix} = \begin{bmatrix} 22 \\ 24 \\ 29 \\ 13 \\ 18 \\ 23 \\ 27 \\ 29 \\ 35 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \vdots \\ \vdots \\ \varepsilon_{3,3} \end{bmatrix}$$

 $\mathbf{y}$  $=$  $\mathbf{X}$  $\boldsymbol{\beta}$  $+$  $\boldsymbol{\varepsilon}$

## Plant data

$\mu$ : the height of all three varieties  
 $\tau_i$ : the difference between the height of variety  $i$  and  $\mu$

**Model II:**  $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$  (reparameterisation of Model I)

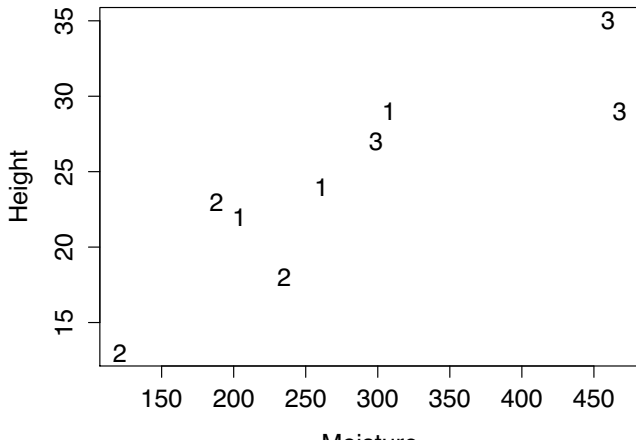
$$\mu_i = \mu + \tau_i$$

$$\begin{bmatrix} y_{1,1} \\ y_{1,2} \\ y_{1,3} \\ y_{2,1} \\ y_{2,2} \\ y_{2,3} \\ y_{3,1} \\ y_{3,2} \\ y_{3,3} \end{bmatrix} = \begin{bmatrix} 22 \\ 24 \\ 29 \\ 13 \\ 18 \\ 23 \\ 27 \\ 29 \\ 35 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \vdots \\ \vdots \\ \varepsilon_{3,3} \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

# Plant data

## Case 4. Variety and soil moisture given.



## Plant data

**Model I:**  $y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}$

*variety* (pointing to  $\tau_i$ ) *moisture* (pointing to  $\beta x_{ij}$ )

$$\begin{bmatrix} y_{1,1} \\ y_{1,2} \\ y_{1,3} \\ y_{2,1} \\ y_{2,2} \\ y_{2,3} \\ y_{3,1} \\ y_{3,2} \\ y_{3,3} \end{bmatrix} = \begin{bmatrix} 22 \\ 24 \\ 29 \\ 13 \\ 18 \\ 23 \\ 27 \\ 29 \\ 35 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 204 \\ 1 & 1 & 0 & 0 & 261 \\ 1 & 1 & 0 & 0 & 308 \\ 1 & 0 & 1 & 0 & 121 \\ 1 & 0 & 1 & 0 & 235 \\ 1 & 0 & 1 & 0 & 188 \\ 1 & 0 & 0 & 1 & 299 \\ 1 & 0 & 0 & 1 & 468 \\ 1 & 0 & 0 & 1 & 460 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \vdots \\ \vdots \\ \varepsilon_{3,3} \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

## Plant data

**Model II:**  $y_{ij} = \mu + \tau_i + \beta_i x_{ij} + \varepsilon_{ij}$

$$\begin{bmatrix} y_{1,1} \\ y_{1,2} \\ y_{1,3} \\ y_{2,1} \\ y_{2,2} \\ y_{2,3} \\ y_{3,1} \\ y_{3,2} \\ y_{3,3} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 204 & 0 & 0 \\ 1 & 1 & 0 & 0 & 261 & 0 & 0 \\ 1 & 1 & 0 & 0 & 308 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 121 & 0 \\ 1 & 0 & 1 & 0 & 0 & 235 & 0 \\ 1 & 0 & 1 & 0 & 0 & 188 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 299 \\ 1 & 0 & 0 & 1 & 0 & 0 & 468 \\ 1 & 0 & 0 & 1 & 0 & 0 & 460 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \vdots \\ \vdots \\ \varepsilon_{3,3} \end{bmatrix}$$

$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$

# More examples

Linear models can be used for many things, including (but not limited to):

- ▶ Which conditions affect the rate of banana ripening?
  - ▶ Is it better to wrap them in newspaper, or submerge them in water?
- ▶ Optimizing the choice of ISPs based on customer service
  - ▶ Comparing time spent in different companies' customer service queue
  - ▶ At different times of days and different days

# More examples

- ▶ Examining the best brand of alkaline battery
  - ▶ Plugging them into different appliances and waiting for them to run out
- ▶ The effect of lifestyle factors on blood pressure
  - ▶ Taking into account factors like gender, age, BMI, height, hours of work, hours of sleep, and number of dependents
- ▶ Observing the performance of short-term memory for numbers
  - ▶ Looking at factors such as gender, exposure to mathematics, duration of interval and presentation of the numbers

# Course outline

~~Matrix~~ Matrix vector  
Random vector ✓

1. Introduction ✓
2. Linear algebra
3. Random vectors
4. Full rank linear model
  - ▶ Estimation
  - ▶ Inference
5. Less than full rank linear model
  - ▶ Estimation and estimability
  - ▶ Inference
6. Experimental design