MAST90104: A First Course in Statistical Learning - Module 1 Introduction

Tim Brown, Yao-ban Chan, Owen Jones, Susan Wei, Mingming Gong and Khue-Dung Dang

Lectures/ Tutorials/ Workshops

- Tuesday (14:15- 16:15), Thursday (10:00-11:00) and Friday (9:00-10:00)
- Lectures will be delivered in-person in the lecture hall and the recording will be uploaded
- Practical and workshop classes are all in-person. NO recording of these sessions will be provided.

Assessment

Assessment - What and When from Handbook

- Four written assignments (5% each, 20% total)
  - Assignment 1 due end of Week 3
  - Assignment 2 due end of Week 6
  - Assignment 3 due end of Week 9
  - Assignment 4 due end of Week 12
- Mid-term exam - Week 7 (35%)
- Computer laboratory test - Week 12 (10%)
- Final exam in scheduled examination period (35%)

Contact and Consultation

- Lecturer: KD Dang
- Email - kd.dang@unimelb.edu.au
- Room number - G58 Peter Hall Building
- Consultation hour: Starts from week 2.
  - In-person: Monday 4-5 pm
  - Zoom: Wednesday 4-5 pm. Zoom link on Canvas
- Ed discussion forum: Students are encouraged to post questions and discuss with other students here

Reference Books

General

- James, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer, 2013. (Covers many, but not all of the topics; does not use matrices - reference for last topic on Unsupervised Learning, `https://www.statlearning.com`)
- Agresti, Foundations of Linear and Generalized Linear Models, Wiley, 2015. (Comprehensive, except for Unsupervised Learning - online - covers the maths well and has many examples)
- Jones, Maillardet & Robinson, Introduction to Scientific Programming and Simulation Using R. CRC Press, 2009.

Reference Books

Linear models - Weeks 1 to 6

- Myers & Milton, A First Course in the Theory of Linear Statistical Models, Duxbury, 1991.
  This textbook may not be easy to access, so an alternative that covers similar topics is:
  - Rencher & Schaalje, Linear Models in Statistics 2nd edition, 2008 (be careful of some differences in notations)
- Linear Models with R, Julian J. Faraway, Chapman & Hall/CRC, 1st or 2nd edition. (Very good on the practical side with a summary of theory)
- Rao & Toutenberg, Linear Models: Least Squares and Alternatives, 1999 (More advanced)
- Draper & Smith, Applied Regression Analysis, 2014 (Less advanced but good for first reading, online edition)

Reference Books Continued

Generalised linear models - Weeks 7 to 9

- Faraway, Extending the Linear Model with R. Chapman & Hall, 1st or 2nd edition. (close to a textbook for GLMs)
- McCullagh & Nelder, Generalised Linear Models, 2nd edition. Chapman & Hall, 1989 (more technical)

Simulation methods and Bayesian statistics - Weeks 10 to 11

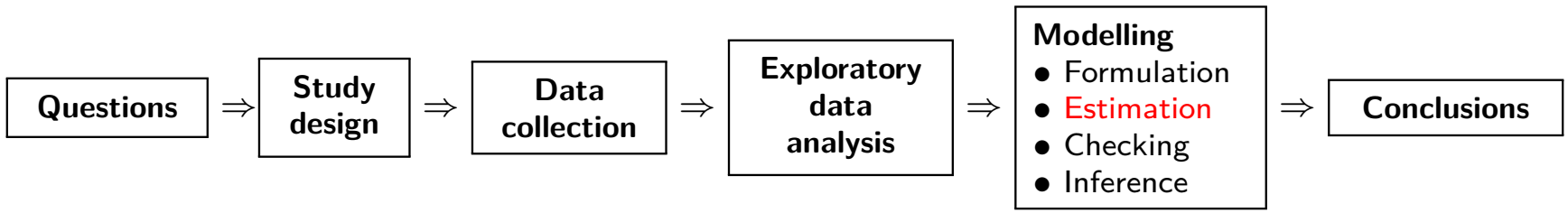
- Gelman, Carlin, Stern, Dunson, Vehtari & Rubin, Bayesian Data Analysis, 3rd edition. CRC Press, 2014. `http://www.stat.columbia.edu/~gelman/book/`

Unsupervised learning - Week 12

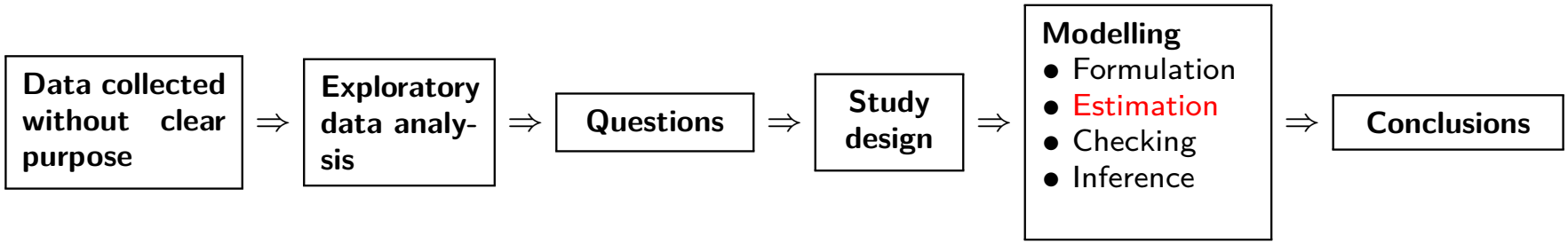
- James, Witten, Hastie & Tibshirani, An Introduction to Statistical Learning: with Applications in R, Springer, 2013.

Statistics and Data Science

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



Real world situation: many companies have already collected data before they hired a statistician or data scientist.



Course Outline: Linear Models - Weeks 1 to 6

1. Introduction (this module)
2. Linear algebra
3. Random vectors
4. Full rank linear model
   - Estimation
   - Inference
5. Design of Experiments and Less than Fully Rank
   - Estimation and estimability
   - Inference, Design

Course Outline: Generalised Linear Models (GLMs) - Weeks 7 to 9

1. Maximum likelihood and binomial regression
2. Exponential families and GLMs
3. Multinomial and ordinal data, Contingency Tables

Course Outline: Bayesian Statistics - Weeks 10 to 11

1. Simulation of random variables
2. Bayesian modelling
3. Estimation: Metropolis-Hastings and Gibbs algorithms
4. Bayesian model diagnostics

Course Outline: Unsupervised Learning - Week 12

1. Principal Components
2. Clustering/ classification methods

Statistical learning

From James, Witten, Hastie & Tibshirani (2013)

> *Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised. Broadly speaking, supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs. . . . With unsupervised statistical learning, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.*

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 variables of interest have a *linear* relationship to other explanatory sets of data (give or take a small amount of error).

In MAST90105 we studied one kind of linear model: linear regression of 1 variable on another variable. However linear models are much more flexible than that.

What is a linear model?

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

- It is easy to analyse; good for statistical inference
- It makes certain assumptions which are not too strict;
- It is also very flexible.

The linear model

We have n subjects (or objects), for each we observe a measurement (or a property) y_i, $i = 1, \ldots, n$. Our aim is to analyse or predict the behaviour of y.

- The y's are *random variables*. Whether y_i is a random variable, a value or data will depend on context.

Each subject also has $k > 0$ other properties that we know or have pre-determined (x variables). We denote these properties as: $x_{i1}, x_{i2}, \ldots, x_{ik}$.

- In practice, the x's might also be random but we condition on their values in the estimation and inference. For example, (x_1, y_1) might be the height and weight of a person - our model is then to predict a person's weight given their height.

The linear model

The general (as opposed to generalized - to be studied in GLMs) linear model is:

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

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

We call y the *response* variable and the x's the *design* (or explanatory) variables.

The β's are *parameters* of the model, and ε is an *error* term.

Up to Week 10 when we study Bayesian statistics, the x's and β's will be known and unknown (respectively) *numbers*.

The linear model

The model attempts to explain the variation in the measured y's (*everything varies!*).

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

To complete the model, we need the distribution of the ε's. For example, they are often supposed to be independent and identically distributed (i.i.d.) normal rv's with mean 0 and *constant* variance σ^2.

Matrix formulation

We can express the general linear model in matrix form:

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

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

vector of parameters

Matrix formulation

Note the dimensions of the matrices:

- $\mathbf{y}$ is $n \times 1$;
- X is $n \times (k+1)$;
- β is $(k+1) \times 1$; and
- ε is $n \times 1$.

Question: Why do we need the intercept β_0 ?

Plant data

We study the heights of 9 plants.

Case 1. No other information.

```
height (y)

   22     13     24     35     29     27     29     18     23

      .      .    ...    .  :       .
    +---------+---------+---------+-height
   10         20         30         40
```

Plant data

Model: $y_i = \mu + \varepsilon_i$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_9 \end{bmatrix} = \begin{bmatrix} 22 \\ 13 \\ 24 \\ 35 \\ 29 \\ 27 \\ 29 \\ 18 \\ 23 \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [\mu] + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_9 \end{bmatrix}$$

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

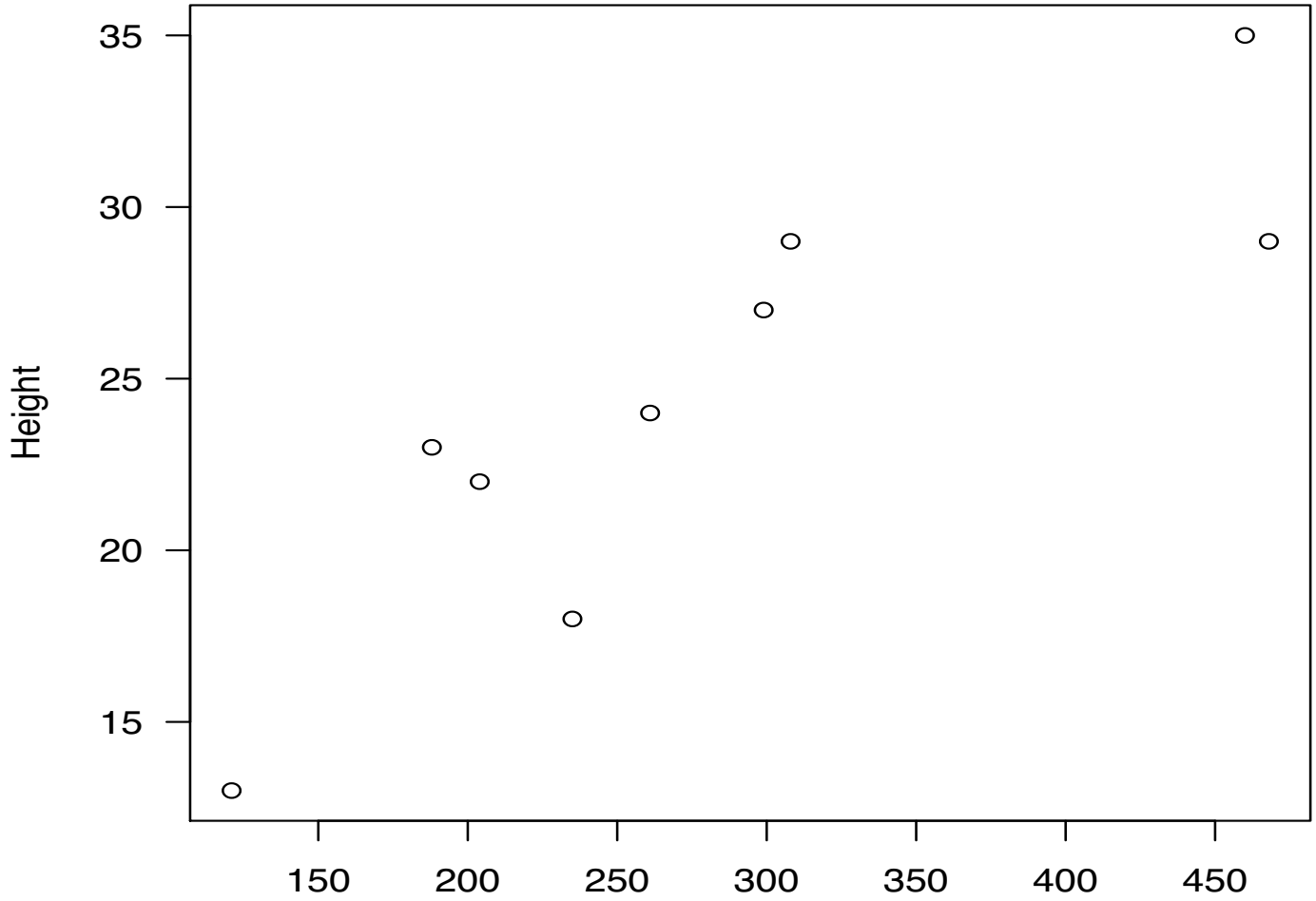
Plant data

Case 2. Soil moisture (x) given.

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

Plant data

Below is the plot of the plants' heights vs soil moisture level



Plant data

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

$$\begin{bmatrix} 22 \\ 13 \\ 24 \\ 35 \\ 29 \\ 27 \\ 29 \\ 18 \\ 23 \end{bmatrix} = \begin{bmatrix} 1 & 204 \\ 1 & 121 \\ 1 & 261 \\ 1 & 460 \\ 1 & 468 \\ 1 & 299 \\ 1 & 308 \\ 1 & 235 \\ 1 & 188 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_9 \end{bmatrix}$$

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

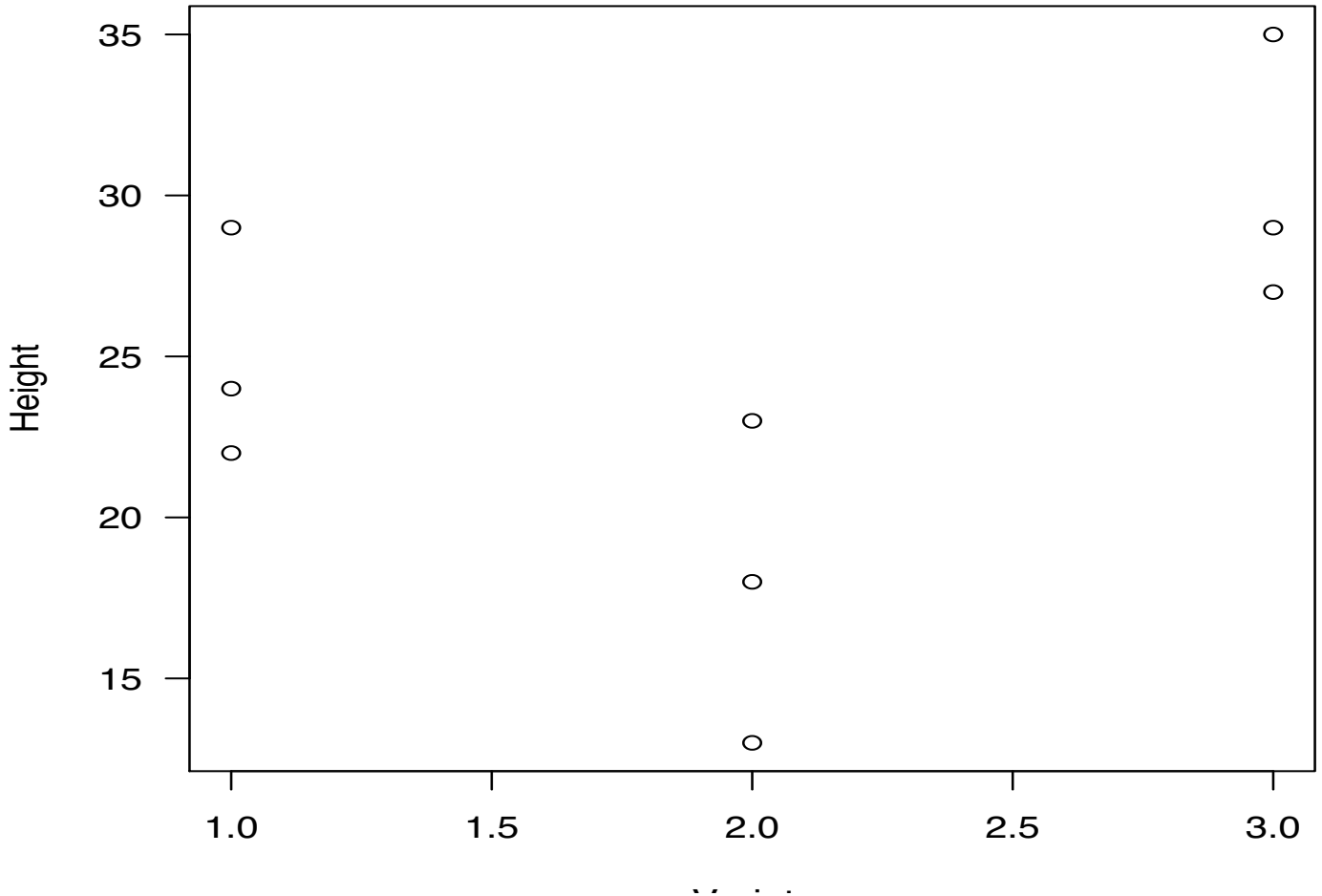
Plant data

Case 3. The 9 plants are of three different varieties.

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

Plant data

Plot of plants' heights by variety



Plant data

Model I: $y_{ij} = \mu_i + \varepsilon_{ij}$ (one-way ANOVA)

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

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

Plant data

Model II: $y_{ij} = \underbrace{\mu + \tau_j}_{\mu_i} + \varepsilon_{ij}$ (reparameterisation of Model I)

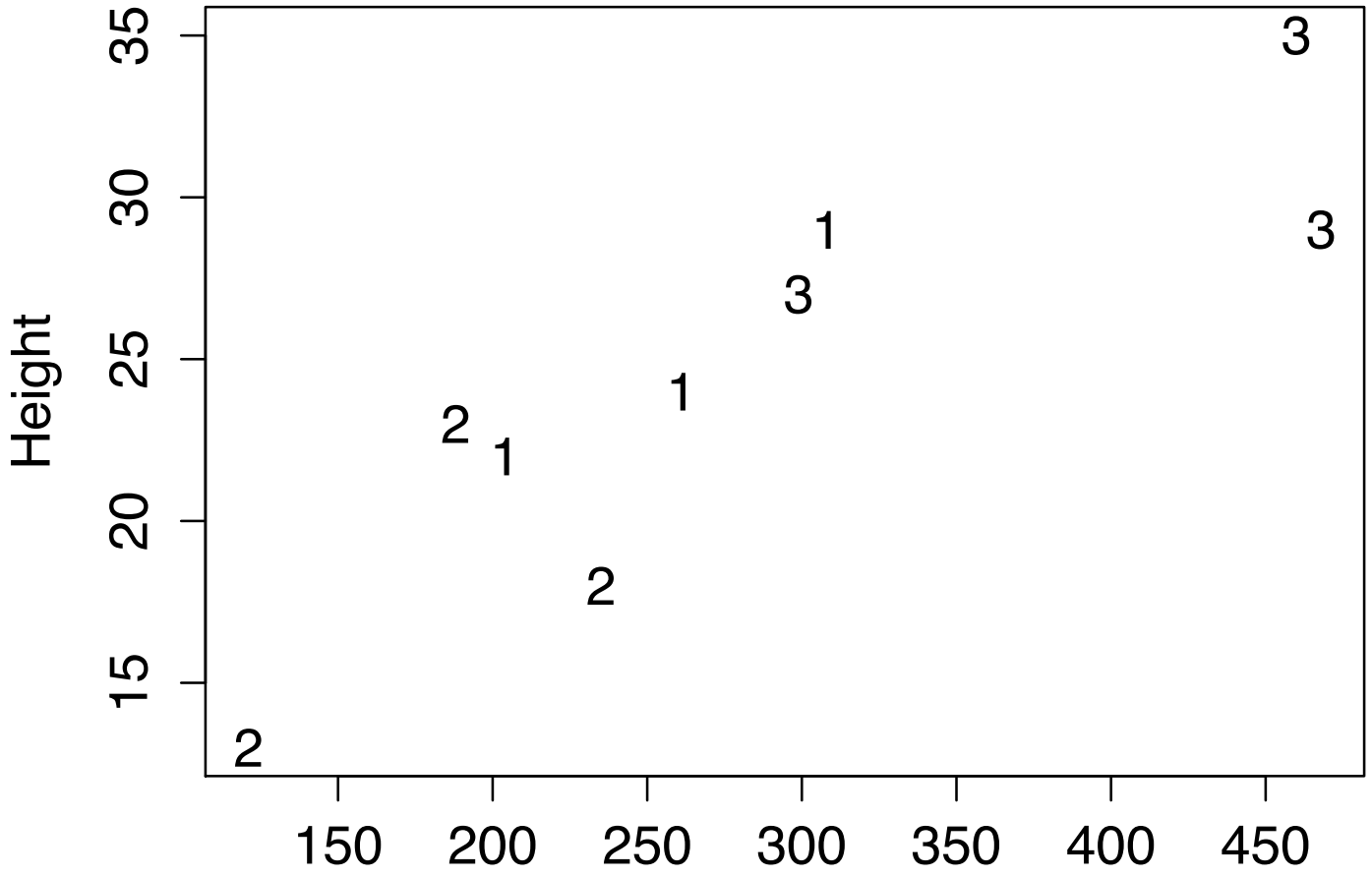
$$\begin{bmatrix} y_{1,1} \\ y_{1,2} \\ y_{1,3} \\ y_{2,1} \\ y_{2,2} \\ y_{2,3} \\ y_{3,1} \\ y_{3,2} \\ y_{3,3} \end{bmatrix} = \begin{bmatrix} 22 \\ 24 \\ 29 \\ 13 \\ 18 \\ 23 \\ 27 \\ 29 \\ 35 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 &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

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

Plant data

Case 4. Variety and soil moisture given.

品种相同 → 影响高度 → 探究种类



Plant data

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

intercept varies

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

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

Plant data

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

确实运动种类 moisture type

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

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

大量predictors时，矩阵方便表示

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?
- 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

More examples

- Examining the best brand of alkaline battery
  - Plugging them into different appliances and waiting for them to run out
- 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
- 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

When is a model linear?

A model is linear when the response variable y is predicted to be a linear form of the parameters β. Linearity in x is not needed.

For example, the model $y = \beta_0 + \beta_1 x + \beta_2 x^2$ is a linear model.
We just take different design variables!

The model

$$y = \frac{\beta_1 x}{\beta_2 + x}$$

is NOT a linear model

Polynomial Regression