MAST90104: A First Course in Statistical Learning

Tim Brown, Yao-ban Chan and Owen Jones

Module 1 Introduction

Contents

1 Housekeeping, References and Outline

1.1 Housekeeping

Housekeeping

Student Representatives and Feedback

- Who?
- What?

Assessment - What and When from Handbook

- Four written assignments (5% each, 20% total)
  - Assignment 1 due August 10
  - Assignment 2 due August 31
  - Assignment 3 due September 28, end of mid-semester break
  - Assignment 4 due October 26, end of SWOT Vac
- 3 hour written examination, on Friday 7 September at 5.30 pm, with make-up workshop on Thursday 6 September at 5.15 pm (35%)
- 45 minute computer labratory test in scheduled examination period (10%)
- 3 hour written examination in scheduled examination period (35%)

Contact and Consultation

- Email - brown.t@unimelb.,edu.au
- Room number - G82 in the Peter Hall Building. Find this from the door on the north-east corner of the Peter Hall Building on Swanston St or from the steps on the verandah on the north side of the Peter Hall Building (ignore the staff only sign on the door)
- Phone - 9035 9547
- Consultation Hours - 12 to 1 pm on Mondays and 2pm to 3 pm on Fridays in G82
- Please note that for help with remembering Calculus and Linear Algebra, the School of Mathematics and Statistics runs a Tutor on Duty service from 12 to 2 pm on Mondays to Fridays in G06 in the Peter Hall Building. You are welcome to use this - there is a study area in there as well.

1.2 Reference Books

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 in the course, but does not use matrices - reference for last topic on Unsupervised Learning)
- Agresti, Foundations of Linear and Generalized Linear Models, Wiley, 2015. (Comprehensive, except for Unsupervised Learning - online - covers the maths well and has many examples)

Linear models - Weeks 1 to 6

- Myers & Milton, A First Course in the Theory of Linear Statistical Models, Duxbury, 1991. (Close to a textbook for this part)
- Linear Models with R, Julian J. Faraway, Chapman & Hall/CRC, 2005. (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, 12

- Faraway, Extending the Linear Model with R. Chapman & Hall, 2006. (close to a text for GLMs)
- McCullagh & Nelder, Generalised Linear Models, 2nd edition. Chapman & Hall, 1989.
- Jones, Maillardet & Robinson, Introduction to Scientific Programming and Simulation Using R. CRC Press, 2009.

Bayesian statistics - Weeks 10 to 11

- Lunn, Jackson, Best, Thomas & Spiegelhalter, The BUGS Book: A Practical Introduction to Bayesian Analysis. CRC Press, 2013.
- Gelman, Carlin, Stern, Dunson, Vehtari & Rubin, Bayesian Data Analysis, 3rd edition. CRC Press, 2014.

1.3 Course Outline

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. Algorithms including Introduction to EM
4. Overdispersion and quasi-likelihood
5. Multinomial and ordinal data, Contingency Tables

Course Outline: Bayesian Statistics - Weeks 10 to 11

1. Review and Extension from MAST90105
2. Simulation of random variables
3. Bayesian modelling
4. Estimation: Metropolis-Hastings and Gibbs algorithms
5. Bayesian model diagnostics

Course Outline: Unsupervised Learning - Week 12

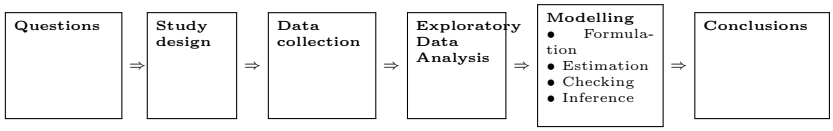
1. Principle Components
2. Clustering Methods

2 Big Picture and What is a Linear Model?

2.1 Big Picture

Statistics

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



2.2 What is a Linear Model?

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 and include the two sample problems that we also studied in MAST90105.

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.

3 The Linear model

3.1 General Description

The linear model

- We have n subjects (or objects), labelled 1 to n;
- Our aim is to analyse or predict the behaviour of a measurement or property of the subject (the y variable), with the individual measurements denoted by $y_1, y_2, \ldots, y_n$.
- Each subject has certain other properties that we know or have predetermined (x variables). Subject i has $k \geq 0$ of these properties: $x_{i1}, x_{i2}, \ldots, x_{ik}$.
- The y's are *random variables* but we will no longer use capitals for them (as we did in MAST90105). Whether y_i is a random variable, a value or data will depend on context.
- 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 if then for the weight conditioned on the value of the 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* 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!). There will always be an error term: ε.

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.

3.2 Formulation

Matrix formulation

We can express the general linear model in matrix form:

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

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

Matrix formulation

Note the dimensions of the matrices:

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

4 Examples

4.1 9 Plants

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
```

4.2 Same mean for all plants

Plant data

Model: $y_i = \mu + \varepsilon_i$

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

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

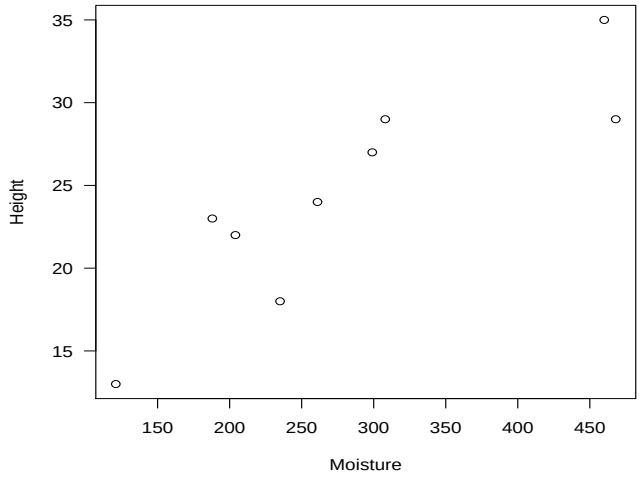
Plant data

4.3 Regression using moisture 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



Plant data

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

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

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

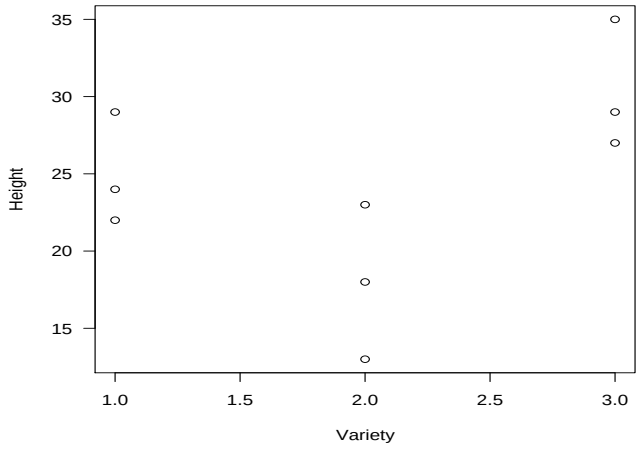
Plant data

4.4 Three varieties

Case 3. Three varieties.

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

Plant data



Plant data

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

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

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

Plant data

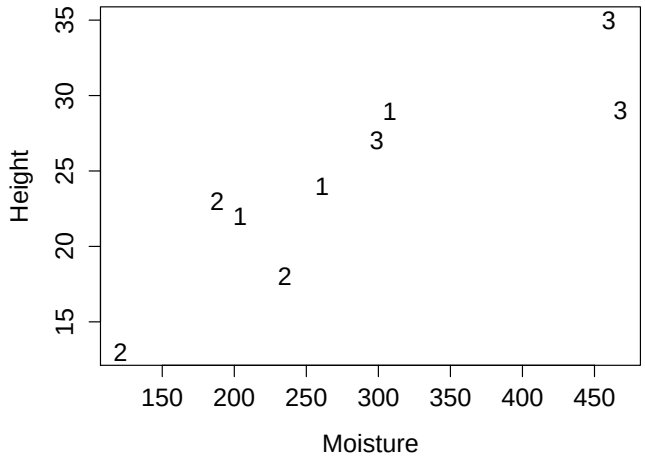
Model II: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$ (reparameterisation of Model I)

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

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

4.5 Both moisture and varieties

Plant data



Case 4. Variety and soil moisture given.

Plant data

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

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

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

Plant data

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

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

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

4.6 Further examples

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

5 When?

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!

Sometimes, a suitable transformation can make a non-linear model into a linear model. However, we must be careful!