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

Generally speaking the linear model is the ‘nicest’ model we can use. It makes certain assumptions, which are mainly used to make the model easy to analyse. It is (relatively) restrictive, although it encompasses many situations.

Many of you will have seen at least one kind of linear model — linear regression. However linear models are much more flexible than that.
In this course, we will take an in-depth look at the theory behind linear models, and how this theory can be applied. We will also show how some practical calculations can be done using the R statistical package.
The general linear model encompasses all linear models. It is a way of describing data that we have measured from some statistical experiment. Suppose that we have run some kind of statistical experiment that has produced data:

- 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 \( y_1, y_2, \ldots, y_n \).
- Each subject has certain other properties that we know or have pre-determined (\( x \) variables);
- Subject \( i \) has \( k \) of these properties — \( x_{i1}, x_{i2}, \ldots, x_{ik} \).
The general linear model can be stated thus:

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

for all \( i = 1, 2, \ldots, n \).
We call \( y \) the *response* variable and the \( x \)'s the *design* constants (technically, they are not variable). The \( \beta \)'s are *parameters* of the model, and \( \varepsilon \) is an *error* term.

All the terms excluding \( \varepsilon \) consist of the model. 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 — \( \varepsilon \).

Ideally, the error term is very small. In that case, the linear model explains the response variable to a high degree of accuracy. We also want the expected error to be 0 - if not, there is something going on which the model is not covering!
Matrix formulation

By putting the data and other terms into matrices:

\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \]
then we can express the general linear model in matrix form:

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

\[y = X \beta + \varepsilon\]
Note the dimensions of the matrices:

- \( y \) is \( n \times 1 \);
- \( X \) is \( n \times (k + 1) \);
- \( \beta \) is \( (k + 1) \times 1 \); and
- \( \epsilon \) is \( n \times 1 \).
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
### Example

- **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
Examples — plant data

Heights of 9 plants.

1. No other information.

height (y)

22 13 24 35 29 27 29 18 23

. . . . . . . . .

+-----------------+-----------------+-----------------+-----------------+-----------------+
height
10 20 30 40
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}

\text{y} = \text{X} \, \beta + \varepsilon
### 2. Soil moisture \((x)\) given.

<table>
<thead>
<tr>
<th>Moisture ((x))</th>
<th>Height ((y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>204</td>
<td>22</td>
</tr>
<tr>
<td>121</td>
<td>13</td>
</tr>
<tr>
<td>261</td>
<td>24</td>
</tr>
<tr>
<td>460</td>
<td>35</td>
</tr>
<tr>
<td>468</td>
<td>29</td>
</tr>
<tr>
<td>299</td>
<td>27</td>
</tr>
<tr>
<td>308</td>
<td>29</td>
</tr>
<tr>
<td>235</td>
<td>18</td>
</tr>
<tr>
<td>188</td>
<td>23</td>
</tr>
</tbody>
</table>
Linear Models: Introduction

Yao-ban Chan
Model: \[ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \]

\[
\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}

\[ y = X \beta + \varepsilon \]
3. Three varieties.

<table>
<thead>
<tr>
<th>Variety</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22</td>
<td>13</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>18</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>23</td>
<td>35</td>
</tr>
</tbody>
</table>
Linear Models: Introduction

Yao-ban Chan
**Model I:** $y_{ij} = \mu_i + \varepsilon_{ij}$

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

\[
y = X \beta + \varepsilon
\]
Model II: \[ y_{ij} = \mu + \tau_i + \varepsilon_{ij} \]

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

\[ y = X \beta + \varepsilon \]
A model is linear when the response variable is predicted to be a linear form of the parameters $\beta$. Linearity in $x$ is not needed.

Sometimes, a suitable transformation can make a non-linear model into a linear model. However, we must be careful with the errors!
Which of the following are linear models?

- $\beta_0 + \beta_1 x$
- $\beta_0 + \beta_1 x + \beta_2 x^2$
- \[ \begin{cases} 
\beta_1 & \text{if male} \\
\beta_2 & \text{if female} 
\end{cases} \]
- $\beta_0 + \beta_1 e^{\beta_2 x}$
- $\beta_0 e^{\beta_1 x}$
- $\mu + \alpha_i$, where $\alpha_i$ refers to the levels ($>1$) of a factor ($A$)
- $\mu + \alpha_i + \beta_j + (\alpha \beta)_{ij}$, where $\alpha_i (\beta_j)$ refers to the levels of factor $A (B)$ and $(\alpha \beta)_{ij}$ refers to the interaction between $A$ and $B$
Why should we use linear models?

- Easy to analyse;
- Easy to formulate;
- Often very appropriate!
How do we fit linear models?

This is the subject of the course — I wouldn’t want to spoil your surprise!
Who should fit linear models?

You!

(by the end of the course, hopefully)