CANCER RESEARCH UK | Cambridge Institute

Together we are beating cancer

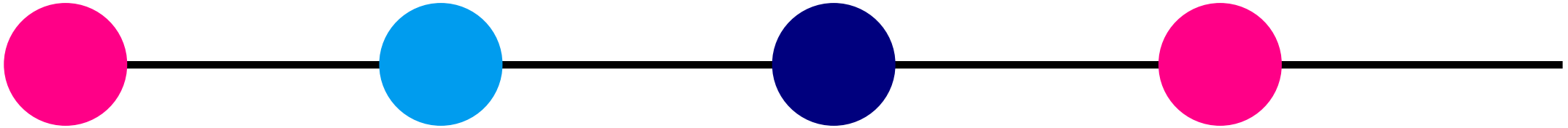
Luca Porcu & Chandra Chilamakuri (Bioinformatics core)

28th February 2025

Linear regression models

Mixed-effects models

Process flow



Stage 1

Brief review and theory on linear mixed-effects models

Stage 2

Hands-on with tumour growth curve analysis

CANCER RESEARCH UK | Cambridge Institute

DO NOT PANIC

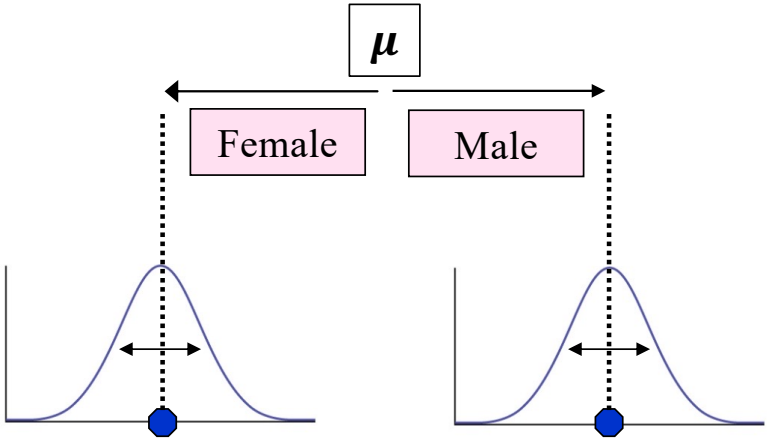
Brief review and theory

Take Home Messages

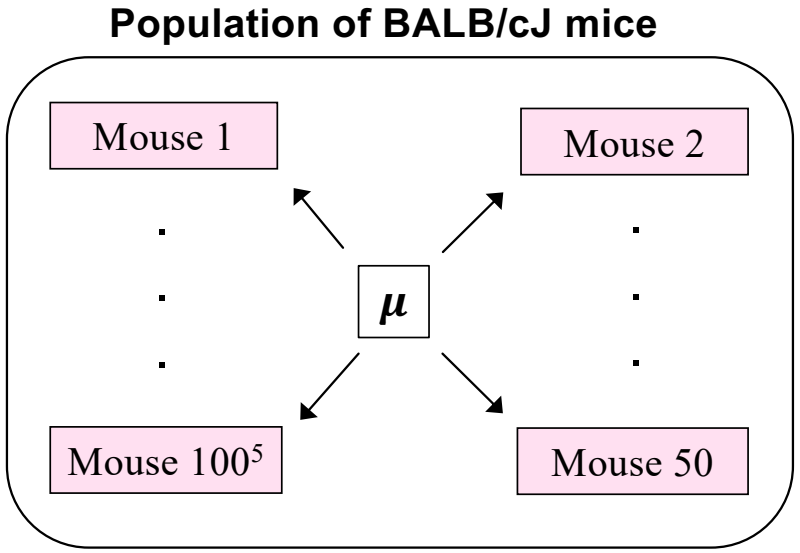
9.00 -09.20 am

Together we are beating cancer

Fixed and random effects



Effects attributable to a **finite** set of levels of a source that occur in the data and which are there because we are interested in them. Fixed effects are **parameters** associated with an entire **population**.



Effects attributable to a (usually) **infinite** set of levels of a source, of which only a **random sample** are deemed to occur in the data. Random effects are associated with **individual observational units** drawn at random from a population.

Why the effects classification is useful?

Fixed effects

1. We are interested to detect and estimate these effects.
2. It is not reasonable to assume a random distribution of these effects (male, female, then what ???).

Random effects

1. They are very useful to describe correlated data (e.g. repeated measure of tumour volume on the same mouse in a preclinical in vivo experiment).
   If we use only fixed effects:
   - 1a. the assumption of independent errors is heavily contradicted
   - 1b. variability of errors is wrongly estimated, hence wrong F tests.
2. Too much parameters: the number of parameters in the model increases linearly with the number of mice.
3. Fixed-effects only model the specific sample of mice used in the experiment.
4. It is reasonable to assume a random distribution of these effects (very large population of BALB/cJ mice).
5. We are interested to estimate the between-mice variability.

Linear mixed-effects models

They describe mathematically the population distribution of the **response** (e.g. tumour volume) as a function of **fixed effects** (e.g. treatment, baseline tumour volume, time), **random effects** (e.g. mouse) and ***error*** (e.g. unpredictable variability)**.**

$$Y_i = f(x_{1,i},..,x_{n,i}; x_{n+1,i},\ldots, x_{n+k,i}; \varepsilon_i)$$

where i is the observational unit of the population.

Hi, my ID is A37. My baseline tumour volume is 150 mm^3, my experimental group is 'Carboplatin plus Gefitinib'. *Tumour volume* has been measured at the following time points: 30, 60 and 120 days after randomization.

Assumptions of linear mixed-effects models

Component	Assumption	Meaning	
Systematic component	Population mean	We are interested to describe the population mean	$\boldsymbol{\mu}$
	Linearity	The coefficients are assumed to combine the effects of the predictors linearly	$\boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_i + \ldots + \boldsymbol{\beta}_n x_i$
Random components	Random effects	They are normally distributed	$N(0, \boldsymbol{\sigma}_{random})$
	Error	Error is normally distributed	$N(0, \boldsymbol{\sigma}_{error})$
		Error variance is constant	$\boldsymbol{\sigma}^2$ = constant

Fitting linear mixed-effects models

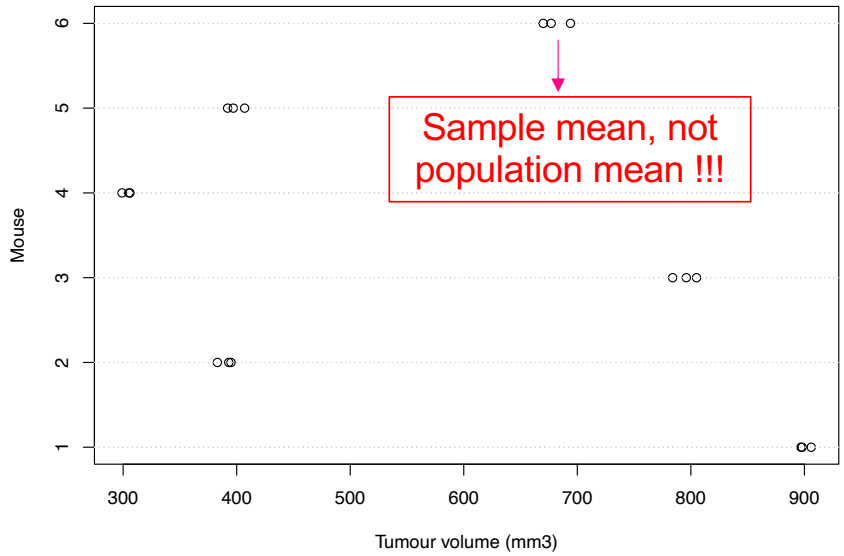
To fit these models, *maximum likelihood* (ML) or *restricted maximum likelihood* (REML) methods could be used.

The best fitting method is *restricted maximum likelihood* (REML) for the following reason:

Maximum likelihood method consider the sample mean as population mean

↓

The between-mouse variability is underestimated



Fitting linear mixed-effects models

However, *restricted maximum likelihood* (REML) method requests the same fixed-effects structure to compare different models.

If you want to compare models with different fixed-effects structure you are forced to use the *maximum likelihood* method.

How do we compare models?

Likelihood ratio test to compare nested models and information criteria AIC and BIC to compare nested and non-nested models

AIC index

$$2 \cdot K - 2 \cdot (\text{log-likelihood})$$

BIC index

$$K \cdot \log_e(n) - 2 \cdot (\text{log-likelihood})$$

Lower values are better for both AIC and BIC. AIC favors more complex models, while BIC includes a penalty for the number of parameters estimated so tends to favor more simple models with fewer parameters.

K = number of parameters

log-likelihood = maximised value of the log-likelihood function of the model

n = number of observations (data points)

Hypothesis tests and estimates

The best solutions:

- Likelihood Ratio tests with *restricted maximum likelihood* method to test random effects parameters.
- ANOVA tests and 95% CI based on t-distribution.

Sometimes it is not possible to use previous statistical methods for the following reasons:

1. It is not possible to clearly define the 'between-' and 'within-' variability. F statistic is not applicable.
2. Non-nested models.
3. Covariance matrix of parameters must be used.

- Likelihood Ratio tests with maximum likelihood method.
- Wald tests with chi-square statistic.

https://bioinformatics-core-shared-training.github.io/Fixed-and-Mixed-effects-models/