Workshop

STAT3015/4030/7030 Generalised Linear Modelling

The Australian National University

Week 4, 2017

Overview

Random effects

A one-way ANOVA model with random effects

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $\alpha_i \overset{i.i.d.}{\sim} \mathrm{Normal}(0, \sigma_\alpha^2)$ and $\varepsilon_{ij} \overset{i.i.d.}{\sim} \mathrm{Normal}(0, \sigma_\varepsilon^2)$.

Then we have a correlation between observations at the same level equal to

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\varepsilon^2}.$$

This ρ is known as the *intraclass correlation coefficient*.

Random effects models (balanced design)

For a balanced design, test for random effect: $H_0 : \sigma_\alpha^2 = 0$ based on

$$F = \frac{MSR}{MSE} = \frac{\sum_{i=1}^{I} n(\bar{Y}_{i\cdot} - \bar{Y})^2/(I-1)}{\sum_{i=1}^{I}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2/(n-1)I}$$

which follows an approximately $F_{I-1,(n-1)I}$ under H_0.

We can apply the same ANOVA and F-test as in the fixed effects case for analysing the data.

Random effects models (unbalanced design)

For unbalanced designs, we can use the likelihood ratio test to compare the two models with and without the random effect:

$$2\left(\log L(\hat{\theta}_L|y) - \log L(\hat{\theta}_S|y)\right).$$

Here the smaller model without random effects is the mean only model. Under H_0 the likelihood ratio test statistic is approximately chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the two models. (Here $df = 1$.)

Question 1 of Tutorial 3

Reconsider the one-way ANOVA example `coagulation` on Wattle.

- Refit the one-way ANOVA but this time treat diet effect as random.
- Write down the structure of your random effects model using mathematical notation.
- Report the intra-class correlation coefficient, test whether the diet effect is significant, and provide estimates of the random effects for each diet.
- Also provide 95% confidence interval estimates for the effect of each diet on blood coagulation time.

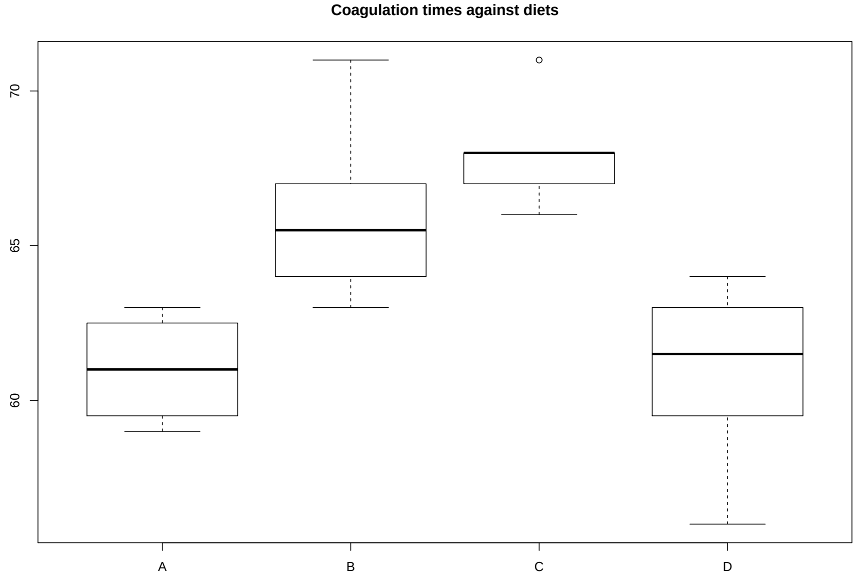
Coagulation data

To study the influence of different diets on blood coagulation times, 24 mice were randomly assigned to four different diets (A, B, C, D) and the samples were taken random order. Coagulation times were recorded for each animal.

Coagulation data

To study the influence of different diets on blood coagulation times, 24 mice were randomly assigned to four different diets (A, B, C, D) and the samples were taken random order. Coagulation times were recorded for each animal.

- Use `coag` as the response variable, and use `diet` as the only predictor.
- To fit a model with random effects, we need to install and load the "lme4" package in RStudio.


Coagulation times against diets
70
65
60
A
B
C
D

Model parameterization

Let Y_{ij} be the blood coagulation time for animal j from diet i, $i \in \{1, 2, 3, 4\}$. The model is

$$\begin{aligned} Y_{ij} &= \beta_0 + \alpha_i + \epsilon_{ij}, \\ \alpha_i &\overset{\text{iid}}{\sim} N(0, \sigma_\alpha^2), \\ \epsilon_{ij} &\overset{\text{iid}}{\sim} N(0, \sigma^2), \end{aligned}$$

where α_i is the random effect for diet.

σ_α^2 is the variance parameter of the diet averages. σ^2 is the variance parameter for the error terms ϵ_{ij}, which captures the variation of coagulation times within diets.

Analysis of Covariance (ANCOVA) model

Parallel regression ANCOVA model: a model with a continuous predictor x and a factor α:

$$Y_{ij} = \beta_0 + \alpha_i + \beta_1 x_{ij} + \varepsilon_{ij}.$$

Non-parallel regression ANCOVA model: a model with a continuous predictor x, a factor α and also an interaction term:

$$Y_{ij} = \beta_0 + \alpha_i + \beta_1 x_{ij} + \gamma_i x_{ij} + \varepsilon_{ij}.$$

Teacher Effectiveness Question (Q2 of Tutorial 4)

A study of 23 student teachers was designed to investigate what factors are important in teaching effectiveness. Twelve male and eleven female student teachers were evaluated and given an overall effectiveness score. In addition, each participating student teacher was given four standardised tests, and their scores were recoded for use as the predictor variables.

Fit parallel and non-parallel ANCOVA models to investigate whether gender has an effect on the teacher effectiveness ratings.

Model selection criteria

To select a good model, we will calculate the following statistics of candidate models:

- p: number of parameters in the specified model;
- s^2: mean square errors (MSE);
- R^2 and R_a^2: coefficient of determination and adjusted coefficient of determination;
- PRESS (PREdiction Sum of Squares) residual a.k.a. $PRESS_p$:

$$e_{i,-i} = Y_i - \hat{Y}_{i.-i};$$

- Mallows' C_p:

$$C_p = p + \frac{(n-p)(s^2 - s_{full}^2)}{s_{full}^2}.$$

Stepwise model selection

`step()` function can be used to perform forward selection, backward elimination and a combination of both. To use the function, we need to specify some arguments:

- object: an object representing a model of an appropriate class (mainly `lm` and `glm`);
- scope: defines the range of models examined in the stepwise search;
- direction: can be one of "both", "forward" or "backward", with a default of "both";
- k: the multiple of the number of degrees of freedom used for the penalty. $k = 2$ gives the *Akaike Information Criterion* (AIC) while $k = \log(n)$ gives *Bayesian Information Criterion* (BIC).