Structure of the workshop

Full notes are at:


You are encouraged to go through these in more detail outside of the workshop.

We will discuss the main concepts, and work through some of the examples in Section 4 of the notes.

Recap: (General) Linear Models

Assumptions:

1. A linear mean function is relevant.
2. Variances are equal across all fitted values.
3. Errors are normally distributed.
4. Samples collected at random.
5. Errors are independent.

Can use F-tests (see Section 4.1.1 of the notes).
Null Hypothesis Significance Testing (NHST)

Section 4.1 in the notes explains the pros-and-cons of NHST.

Traditionally, if the response variable is Gaussian (normal), then you may have come across two frequently used approaches:

- **F-tests**: based on comparing the residual mean squared error with the regression mean squared error, or
- **Likelihood ratio tests (LRT)**: based on comparing the model deviance between two models.

Both of these cases are exact tests for linear regression with Gaussian errors (but for mixed models these become approximate).

For mixed models things get trickier again, and there is no consensus. See GLMM FAQ for more discussion. A nice description of the types of approaches we can use in different cases can be found at:

https://rdrr.io/cran/lme4/man/pvalues.html

Present final model results in terms of effect sizes and confidence intervals where possible (or via predictive plots).

We will introduce some common scenarios in which mixed models can be applied, and give examples of model simplification and inference in these cases.

Recap: linear model

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \]

where \( \epsilon_i \sim N(0, \sigma^2) \) and \( \sigma^2 \) is variance.

In words:

response ~ intercept + slope × explanatory + noise

Assumptions

*Previously*: used model checks and biological rationale to test linearity, normality and homoscedasticity of residuals.

What about independence of residuals?

Depends on experimental design.
Independence of errors

Tests of statistical significance require that each experimental unit has the same \( \epsilon \), unaffected by and uncorrelated with other residuals (samples are independent and identically distributed - i.i.d.).

But this is often untrue.

Example: bacterial loads


Blocked experimental design

Recognise natural structuring among experimental units.

Source of error (e.g. cabinet, top/bottom of field, make of car, student identity).

Absorb this error by replicating experiment within blocks.

Partition the residual deviance.

Blocked experiment: bacterial growth

Bacteria grown in four different media (fixed treatment has four levels).

Only have small growth cabinets:

- Room for four growth jars per cabinet.
- Use five cabinets (blocks).
- One replicate of experiment per cabinet.

Measure bacterial growth rate.
### Why use blocks?

Here **treatment** (media) is of interest, but multiple treatments within each block.

We know growth rates will differ between cabinets.

Assume that **relative** growth rates will be similar between treatments in each cabinet.

Use **cabinet** as a **block** to **absorb** experimental noise.

---

### Analyse it badly

**Ignore non-independent residuals (i.e. ignore cabinet effects)**

```r
bac_lm <- lm(growth ~ media, data = bac)
anova(bac_lm)
```

```r
## Analysis of Variance Table
##             Df Sum Sq Mean Sq F value Pr(>F)
## media        3 88.578  29.5258  3.1002 0.0564
## Residuals   16 152.380  9.5237
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

---

### Analyse it properly—part I

**Put cabinet in as a fixed effect.**

```r
bac_lm <- lm(growth ~ media + cabinet, data = bac)
anova(update(bac_lm, . - media), bac_lm)
```

```r
## Analysis of Variance Table
##             Df Sum Sq Mean Sq  F value Pr(>F)
## Model 1: 1  15 103.81  82.5407 39.6285 0.0001
## Model 2: 2  12 15.23  12.6917  5.9318 0.0216
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1  15 103.81
## 2  12 15.23  3  88.578 23.264 3.734e-05 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It does not make sense to drop cabinet here. Why not?

---

### Mixed effects model

This simple **balanced** design can be analysed in a straightforward way.

We ideally want a general framework that **accounts** for the variations due to the **blocks**, but models the effect we’re interested in: here the effect of **media** on **bacterial growth**. Furthermore, we don’t want to use up too many degrees-of-freedom.

These challenges can be dealt with using **mixed models**.

A **mixed model** is so-called because it contains a mixture of **fixed** and **random** effects.
Fixed effects

- Treatments are **fixed** by the experimenter, guided by hypotheses e.g. test of whether treatment levels **differ** or whether there is a **trend**.
- We **care** about the **identity** of each level of a fixed effect.
- Given a new experimental unit, we could predict its response.

Random effects

- Are sampled from a **population** of possible levels.
- We don’t **care** about the **identity** of each level of a random effect\(^1\).
- Wouldn’t help us predict new values of response variable.
- Instead we **predict** how much variance is absorbed by random effects.
- Observations influenced by random effects are **not independent**.

\(^1\)Traditional view, but in some cases we can use REs **differently**

A mixed-effects model

Response variable \(Y\). Regression parameters for **fixed** explanatory variables: \(\beta_p\)

Noise absorbed by **random variable(s)**: \(\gamma \sim N(0, \sigma^2_{\gamma})\)

Residual noise: \(\epsilon \sim N(0, \sigma^2)\)

\[ Y_i = \beta_0 + \beta_1 X_i + \gamma_i + \epsilon_i \]

Here \(\sigma^2_{\gamma}\) is the **variance** attributed to the random effect, and \(\sigma^2\) is the **residual** variance.

This particular model known as a **random intercepts** model.

What does it do?

**Fixed intercepts**

**Random intercepts**
Analyse it properly—part II

Since we do not care about the impact of cabinet *per se*, we could also include this as a random effect, using lmer():

```r
bac_lmer <- lmer(growth ~ media + (1 | cabinet), data = bac)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: growth ~ media + (1 | cabinet)
## Data: bac
## REML criterion at convergence: 68.8432
## Random effects:
##   Groups Name Std.Dev.
##   cabinet (Intercept) 2.873
##   Residual 1.127
## Number of obs: 20, groups: cabinet, 5
## Fixed Effects:
##   (Intercept) media2 media3 media4
##      5.58   1.78  -1.96  -3.84
```

These data are **balanced**, and the error structure is Gaussian. We can use an F-test here to assess what happens when media is dropped from the model:

```r
Anova(bac_lmer, test = "F")
```

```
## Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)
## Response: growth
## F Df Df.res Pr(>F)
## media 23.266 3 12 2.733e-05 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Aside: Restricted Maximum Likelihood**

Mixed models fitted using **Restricted Maximum Likelihood (REML)**.

Only possible thanks to powerful computers (fits models iteratively). Separates the influences of random and fixed effects, meanwhile retaining the nested structure of the dataset.

Caveats:

- Need good understanding of data structure.
- Need to be careful during model simplification.

**Aside: Simplifying REML models**

Be very careful! Standard partitioning of deviance no longer applies. In **balanced**, **nested** designs, F-tests are OK. For **unbalanced** or **non-nested** designs we have to be more careful.

In these latter cases we need to refit the model using **unrestricted maximum likelihood (ML)**. This produces a biased approximation, but usually a good one.

We can do this using `update()` but with a REML = F argument.

After model simplification, switch back to REML fit to perform inference. See Section 4.2.1.1 of the notes for more details.
Your turn

Have a go at Section 4.2 of the workshop notes.

Nested errors

The previous example was fairly simple. Certain study designs will end up with replicates nested with other variables / blocks.

In this case the residuals are not independent once again, but the error structure is more complex to model.

Example: drunken behaviour on campus

Samples

Dave got arrested for being disorderly.

He failed a breathaliser test.

To avoid the fine, he claimed he had used breath freshener.

To prove his innocence he conducted an experiment.
### Results

#### The wrong way!

```r
drunk_lm <- lm(alcohol ~ freshener, data = drunk)
anova(drunk_lm, test = "F")
```

```
## Analysis of Variance Table
##                        Df Sum Sq Mean Sq F value Pr(>F)
## freshener              1 0.0057685 0.0057685  45.345 6.342e-10 ***
## Residuals              118 0.0150113 0.0001272
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### What’s wrong?

- Multiple samples per student.
- Multiple estimates per sample.
- **Clue**: residual df (118) is much bigger than the number of experimental units (students: 6 here).
- Samples within students are PSEUDOREPLICATES.
- Residuals are not independent.

### Correct (traditional) analysis

**Derived variable analysis:**

- Cope with pseudoreplication by **averaging** them out.
- Gives one average per student.
- Analyse this smaller dataset.
- **Note**: loses information on within-student variation. Could be important if complicated nested experimental design.
Derived variables in R

Using tidyverse (base R version in the notes).

```r
alc <- drunk %>%
group_by(student, freshener) %>%
  summarise(alcohol = mean(alcohol)) %>%
  ungroup()

alc
```

## # A tibble: 6 x 3
## # Groups: student, freshener [1]
## student freshener alcohol
## <chr> <chr> <dbl>
## 1 alex no 0.0881
## 2 colin no 0.0724
## 3 dave no 0.0786
## 4 elvis yes 0.0959
## 5 jonah yes 0.0803
## 6 pablo yes 0.104

Richard B. Sherley (G)LMMs in R March 2020 33 / 63

Dropping a single variable, so safe to anova().

```r
alc_lm <- lm(alcohol ~ freshener, data = alc)
anova(alc_lm, test = "F")
```

## Analysis of Variance Table
## # Df Sum Sq Mean Sq F value Pr(>F)
## freshener 1 0.00028843 0.00028843 2.7094 0.1751
## Residuals 4 0.00042582 0.00010645

Richard B. Sherley (G)LMMs in R March 2020 34 / 63

Derived variables in R

This is a statistically valid analysis, however:

- it ignores the uncertainties around the pseudo-replicates;
- the interpretation of the response variable is actually the mean of a bunch of measurements, not the measurements themselves.

OK with a balanced design.

**But** may not be possible to generate derived variables for some studies (e.g. how do you average a categorical response)?

Mixed effects model

Ideal:
A general purpose way to build a model that **accounts** for the variation due to the **pseudoreplicates**, but models the effect we’re interested in.

Here the effect of **breath freshener** on **alcohol content**.

Can be done with a **mixed model**.
Here we have samples nested within students, with freshener as our fixed effect.

```r
drunk_lmer <- lmer(alcohol ~ freshener + (1 | student / sample), data = drunk)
Anova(drunk_lmer, test = "F")
```

The F-test here gives the same result as the derived variable analysis, since the data are balanced and nested.

Aside: did Dave drink alcohol?

Suggests negligible difference between blood alcohol content between treatments, given the other uncertainties in the system.

However, does not answer the specific question:

*What is the probability that you’ve used breath freshener relative to drinking alcohol, given your alcohol content?*

Beware of proxy measurements (and prosecutor’s fallacy).

---

3 can be tackled using Bayesian methods
Think about your hypothesis

Tight link between hypothesis, experimental design and analysis.

Hypothesis defines the experimental unit:
- e.g. “Fire regulates savannah grass diversity”

Response is grass diversity, and treatment is burned vs. unburned.

Experimental unit is plot.

---

Getting more complicated

Split-plot experimental design:
- The basis of many agricultural studies.
- Many treatments, spatial non-independence.

Nested analyses:
- Study of variance at nested scales.
- Common in population genetics.

---

Think about your hypothesis

What needs to be replicated?
- Burning treatment.

If only one burn, then there is no replication.

Multiple measures of each burned/unburned plot is pseudoreplication.

Good to improve estimate of mean, but still need replication.

Statistical tests must occur at the level of the experimental unit.

---

Getting more complicated

Longitudinal studies:
- Multiple observations of experimental units

Make sure you know what the experimental unit is.
Example: split-plot design

This experiment involves the yield of cereals in a factorial experiment with:

- 4 **blocks** (fields).
- Half of each block was **irrigated**, and half not.
- Each half-block was split into 3 split-plots, and seeds were sown at different **densities** in each split-plot.
- Each sowing density plot was split into 3 small split-split plots and different **fertilisers** applied by hand (N alone, P alone and N + P together).
Split-plot design

block is the only random effect but our data are nested.

Fixed effects are irrigation, density and fertilizer.

Idea is to nest correctly\(^4\)
\[
(1 \mid \text{block} / \text{irrigation} / \text{density} / \text{fertilizer})
\]

\(^4\)this actually throws an error in this case—see workshop notes for full solution

Your turn

Have a go at Section 4.4 in the workshop notes.
Distance to Kenyan herbivores

**Hypothesis:** Distance from road depends on species.

**Caution:** this could be mediated by group size.

Data not independent because contributed by 8 observer groups. Obs. group does not feature in the hypotheses, they:
- are sampled from a wider population of MSc students;
- would not help us predict distances for a new set of observer groups;
- would waste 7 d.f. in a traditional analysis;
- hence makes an ‘ideal’ random effect.

### Fixed and random effects in Hell’s Gate

#### Data structure

```r
# A tibble: 462 x 5
## Group.Name Species    Distance Number Size.class
## <fct>     <fct>       <int>   <int>        <int>
## 1 Nepatano zebra      450     1          1
## 2 "the lionel king " zebra 380   5          3
## 3 bilbo's badgers    warthog 120    2          2
## 4 Mega-my-fauna      warthog  235    1          2
## 5 Quaderatops        thomsons gazelle  226  11         1
## 6 "the lionel king " thomsons gazelle 200  1          1
## 7 Quaderatops        thomsons gazelle 180    2          1
## 8 "the lionel king " thomsons gazelle 190    2          1
## 9 Quaderatops        zebra     45     12         3
## 10 Quaderatops       zebra     43     9          3
## # ... with 452 more rows

Note: each observer group contributes data. Could we make a derived variable per observer group?

- What is ‘average species’?
- What would happen to the precious variation in group size?

#### Fitting a mixed effects model

Here data are **unbalanced**.

```r
gg %>% group_by(Group.Name, Species) %>% count()
```
Your turn

Have a go at Section 4.5 in the notes.

Model checking

You should really check model assumptions as for general LMs/GLMs. Trickier to do, but some examples in Section 4.6 of the workshop notes.

Non-normal response

But my data isn’t normal…

Package lme4 also includes function glmer():

- which allows use of family = ""

So can try non-normal error structures…

…leading to GLMM (Generalised Linear Mixed Models).

Fitting GLMMs

hg_glmer <- glmer(Number ~ Species * log(Distance) + (1 | Group.Name), data = hg, family = poisson, control = glmerControl(optimizer = "bobyqa"))
drop1(hg_glmer, test = "Chisq")

## Single term deletions
##
## Model: Number ~ Species * log(Distance) + (1 | Group.Name)
## Df AIC LRT Pr(Chi)
## <none> 2284.4
## Species:log(Distance) 2 2312.7 32.365 9.377e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Note: uses adaptive Gauss-Hermite quadrature approximation of the likelihood to fit the model, not REML.
GLMMs

Can use `lme4` for:
- Gaussian, Poisson, binomial, binary, gamma error structures,
- and for crossed and nested random effects structures,
- model checks remain important...

Ultimately, go Bayesian!

Many of the challenges associated with mixed effects models go away if you move your inference into a Bayesian framework.

Of course other challenges arise in their place, mainly in terms of variable selection, however, in general I would recommend using a Bayesian framework for complex models with hierarchical structures, particularly spatio-temporal modelling.

These approaches are beyond the scope of this course, but we are hoping to run a Bayesian Modelling workshop next year, so keep your ears to the ground!

The dangers of too much R coding

Congratulations, you have become Generalised Linear Mixed Modellers.

Richard B. Sherley (G)LMMs in R March 2020 61 / 63

Richard B. Sherley (G)LMMs in R March 2020 62 / 63

Richard B. Sherley (G)LMMs in R March 2020 63 / 63