LECTURE 16

Mixed-effects models II: multilevel modelling and generalised linear modelling

Plan

- We’re examining four examples today:

- a case of *spatial pseudoreplication*
- then *multilevel modelling*

- and

- a case of *temporal pseudoreplication*
- followed by a *mixed-effects logistic regression*

Group comparisons

- Mixed-effects models are an alternative to ANOVA/Kruskal-Walis tests when there is pseudoreplication

- An example of spatial pseudoreplication:
  - three treatments (diets)
  - two rats per treatment
  - glycogen levels measured in each liver
  - each liver cut up into three parts
  - each part measured twice for glycogen
  - total: 3 x 2 x 3 x 2 = 36 measurements

  - Hierarchical structure of sample:
    - treatment/liver(=rat)/liver piece/measurement

- Question: does treatment influence glycogen levels?

Group comparisons: spatial pseudoreplication

- A simple comparison of means would be wrong due to pseudoreplication

```
> kruskal.test(rats$Glycogen ~ rats$Treatment)
Kruskal-Wallis rank sum test
chi-squared = 17.4146, df = 2, p-value = 0.0001654
```

- A solution is to average all measurements (i.e. mean by rat) and then compare treatments (sample size : 6 rats, 2 per treatment)
  - but then significance of *Treatment* disappears due to small sample

```
> avrat
        1        2
1 132.5000 148.5000
2 149.6667 152.3333
3 134.3333 136.0000
```

```
> kruskal.test(as.vector(avrat) ~ treat)
    Kruskal-Wallis rank sum test
Kruskal-Wallis chi-squared = 3.4286, df = 2, p-value = 0.1801
```

Mixed-effects model

- We can use all 36 measurements in a mixed-effects model
  - fixed effect: *Treatment*
  - two **nested** levels of spatial pseudoreplication (rats, liver parts)
    - measurement is lowest level and is represented by residuals

- First, we turn numeric variables into factors:
  - `> TreatmentF <- factor(rats$Treatment)`
  - `> LiverF <- factor(rats$Liver)`
  - `> RatF <- factor(rats$Rat)`

- Note: this creates new factors but does not replace variables in file *rats*
  - to change file use command *transform*

Model structure

- Fixed factor: *Treatment*
- Glycogen ~ Treatment

- Random effects: **intercept** only
  - we are only interested in different baselines (i.e. random intercepts) by rat and by liver piece
  - when effect levels are nested, we enter them as interactions (:)
    - note: nested structure leads to automatic distinction between different 'Rat 1' animals in each treatment

- Highest-level random effect: *Rat* (=liver) level
  - Define level:
  - > rat <- TreatmentF:RatF
  - Random effect term: (1 | rat)
    - Remember: *Treatment* is at the top and is needed to define the random term, but it is a fixed effect

- Lower level: Liver piece (variable *Liver*)
  - Define level:
  - > liver <- TreatmentF:RatF:LiverF
  - Random effect term: (1 | liver)

- Hence model is
  - `> lmer(Glycogen ~ TreatmentF + (1|rat) + (1|liver), data=rats)`

Fixed effect, random intercepts

```
> modelrats <- lmer(Glycogen ~ TreatmentF + (1|rat) + (1|liver), data=rats)
> summary(modelrats)
Linear mixed model fit by REML
Formula: Glycogen ~ TreatmentF + (1|rat) + (1|liver)
   Data: rats
   AIC   BIC logLik deviance REMLdev
 231.6 241.1 -109.8    234.3   219.6
Random effects:
 Groups   Name        Variance Std.Dev.
 liver    (Intercept) 14.167   3.7639
 rat      (Intercept) 36.065   6.0054
 Residual             21.167   4.6007
Number of obs: 36, groups: liver, 18; rat, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  140.500      4.707  29.851
TreatmentF2   10.500      6.656   1.577
TreatmentF3   -5.333      6.656  -0.801

Correlation of Fixed Effects:
            (Intr) TrtmF2
TreatmentF2 -0.707
TreatmentF3 -0.707  0.500
```

- Fixed effects:
  - *Treatment* not significant
- Random intercept effects:
  - between-rat variance: 36.065
  - between-liverpiece variance: 14.167
  - residual (between-measurement variance): 21.167

Variance component analysis

- Fixed effect of *Treatment* disappears after controlling for random intercept effects,: variation in Glycogen by Treatment wasn't 'real'

- We can still calculate the fraction of variation accounted for by grouping levels (*Rat* and *Liver*)
  - Define a vector with variances
  - > vars <- c(14.167,36.065,21.167)
  - Then divide by total variance to obtain fraction explained
  - > sum(vars)

  - >100*vars/sum(vars)
  - [1] 19.84201 50.51191 29.64607

- Variance component analysis:
  - 50.5% of the variation in *Glycogen* is between rats within treatments
  - 19.8% is between liver bits within rats
  - 29.6% is between readings within liver parts within rats (residual)

- Note: if *Treatment* were significant, residual variance unexplained by the fixed factor could still be decomposed into random effects (and another residual)

Significance of Treatment

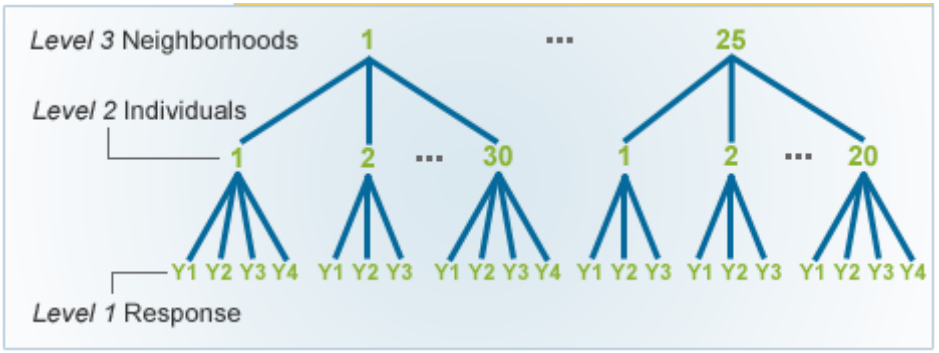
- Apparently, *Treatment* is not significant (see *The R Book*)
- But running ANOVA shows otherwise:

```
> modelrats.full <- lmer(Glycogen ~ TreatmentF+(1|rat)+(1|liver), data=rats, REML=F)
> modelrats.null <- lmer(Glycogen ~ (1|rat)+(1|liver), data=rats, REML=F)
> anova(modelrats.null, modelrats.full)
Data: rats
Models:
modelrats.null: Glycogen ~ (1 | rat) + (1 | liver)
modelrats.full: Glycogen ~ TreatmentF + (1 | rat) + (1 | liver)
                Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
modelrats.null  4 247.77 254.10 -119.88
modelrats.full  6 245.27 254.77 -116.64 6.4962      2    0.03885 *
```

- Always check ANOVA results when making statements about variable significance

Multilevel (hierarchical) modelling

- Multilevel modelling allows us:
  - to identify significant fixed effects after controlling for *multilevel* (hierarchically structured) random effects
    - *individual/household/school/town etc.*
  - to calculate fraction of variance around estimated fixed effects due to each level of nested hierarchy
    - *individual vs. household vs. school vs. town*



Example: school results

- In multilevel analysis, typically you have
  - a hierarchy of random effects (excluding fixed effects) as *intercepts*
  - fixed effects
    - or none if you fit a null model as a starting point
    - But you want to fit as many significant fixed effects feasible in order to leave as little variance unexplained by all effects (fixed and random) as possible

- Good example: school results in Britain (mean=98.06 points), with hierarchical or nested data sampling:
  - **Towns**: 4
  - **Districts:** 6 per town
  - **Streets:** 10 per district
  - **Households:** 4 per street
  - **Children**: Variable number (max: 8) and gender in each household; one school result per child

- Questions:
  - do girls and boys achieve different school results?
  - which levels of grouping account for more variance in school results?

Model structure

- File: *childfull*

```
> head(childfull, 2)
  childID child house street district  town response gender
1       1     1 door1      1        A Leeds 83.88773   male
2       1     1 door2      1        A Leeds 99.96294   male
```

- **Model structure:**
- Outcome: *response* (=school results)
- Fixed effect: *gender* (male or female)
- Random effects levels:
  - Town/district/street/household/children
  - Child: residual level
    - (if a child had provided more than one school result, then *Child* would be lowest-entered level and between-results variation would be residual)

- Note that fixed factor *gender* is not part of the nested hierarchy of random effects

Syntax

- To make life easier:
  - attach file *childfull*
    - too many factors!
  - *street* is numeric, so enter it as *factor(street)* or *as.factor(street)*
  - define random effects in advance to avoid rewriting
    - d <- town:district
    - s <- town:district:factor(street)
    - h <- town:district:factor(street):house

- Let us start with a model including only the multilevel random intercept effects (no fixed effect *gender*)

```
> schools.null <- lmer(response ~ (1|town)+(1|d)+(1|s)+(1|h))
```

Random effects only

```
> schools.null <- lmer(response~(1|town)+(1|d)+(1|s)+(1|h))

> summary(schools.null)
Linear mixed model fit by REML
Formula: response ~ (1 | town) + (1 | d) + (1 | s) + (1 | h)
   AIC   BIC logLik deviance REMLdev
 19880 19916  -9934    19873   19868
Random effects:
 Groups   Name        Variance Std.Dev.
 h        (Intercept)   4.0798  2.0199
 s        (Intercept)  15.5565  3.9442
 d        (Intercept) 168.4955 12.9806
 town     (Intercept)  37.1068  6.0915
 Residual              36.3176  6.0264
Number of obs: 2972, groups: h, 960; s, 240; d, 24; town, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)   98.174      4.044   24.28
```

- Random effects:
  - most variation in school results is between districts (neighbourhood effects!)
- Fixed effect:
  - not really; intercept=98.17 is just the predicted average school result (98.174 points) in the whole sample
  - multilevel random intercept effects explain variation around that general average

Gender effects

```
> schools.full<- lmer(response~gender+
                    (1|town)+(1|d)+(1|s)+(1|h))
> summary(schools.full)
Linear mixed model fit by REML
AIC   BIC logLik deviance REMLdev
 19878 19920  -9932    19868   19864
Random effects:
 Groups   Name        Variance Std.Dev.
 h        (Intercept)   4.0817  2.0203
 s        (Intercept)  15.6746  3.9591
 d        (Intercept) 168.3500 12.9750
 town     (Intercept)  36.9757  6.0808
 Residual              36.2406  6.0200
Number of obs:2972,groups:h,960;s,240;d,24;town,4

Fixed effects:
             Estimate Std. Error t value
(Intercept)  97.8965     4.0410  24.226
gendermale    0.5368     0.2363   2.272

Correlation of Fixed Effects:
            (Intr)
gendermale -0.030
```

- Now let us add *gender* as fixed effect:
  - there is a significant (t = 2.272) but small (0.537 or about half a point higher in boys) effect of gender on school results after controlling for random effects (chisq=5.15, P=0.023)
- Random effects:
  - again, mostly between-district effects (neighbourhood effects!)

VCA

```
> schools.full<- lmer(response~gender+
                    (1|town)+(1|d)+(1|s)+(1|h))
> summary(schools.full)
Linear mixed model fit by REML
AIC   BIC logLik deviance REMLdev
 19878 19920  -9932    19868   19864
Random effects:
 Groups   Name        Variance Std.Dev.
 h        (Intercept)   4.0817  2.0203
 s        (Intercept)  15.6746  3.9591
 d        (Intercept) 168.3500 12.9750
 town     (Intercept)  36.9757  6.0808
 Residual              36.2406  6.0200
Number of obs:2972,groups:h,960;s,240;d,24;town,4

Fixed effects:
             Estimate Std. Error t value
(Intercept)  97.8965     4.0410  24.226
gendermale    0.5368     0.2363   2.272

Correlation of Fixed Effects:
            (Intr)
gendermale -0.030
```

- Let us do a variance component analysis on full model:

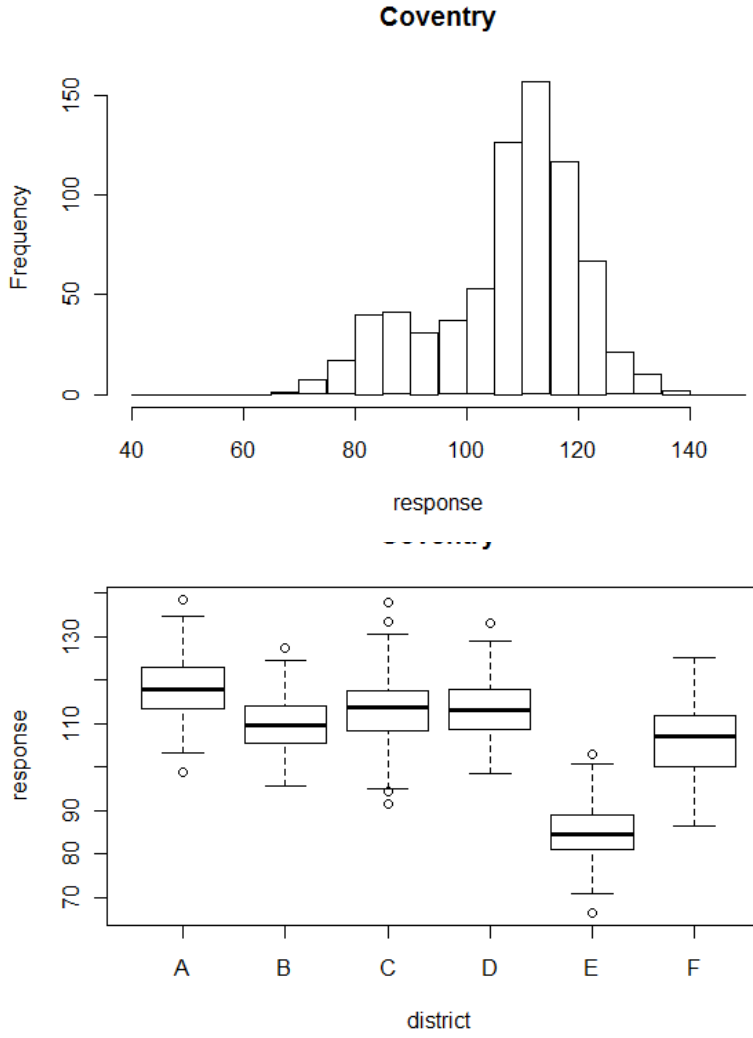
```
> vc <- c(4.0817, 15.6746, 168.3500,
36.9757, 36.2406)
> vc <- 100*c(4.0817, 15.6746,
168.3500, 36.9757, 36.2406)/sum(vc)
> vc
[1]  1.561939  5.998180 64.422289
14.149446 13.868146
```

- Variance contribution by level:
  - Household: 1.56%
  - Street: 6%
  - **District: 64.4%**
  - Town: 14.1%
  - Residual (individual child): 13.9%

- District differences responsible for 64% of the variance in school results unexplained by gender

Multilevel (hierarchical) modelling

- Main points in multilevel modelling:
- It should not include fixed effects in grouping hierarchy
  - this can be done, but then this is more properly seen as just spatial pseudoreplication
- Random effects provide a measure of contribution of each hierarchical level to variance unexplained by fixed factor
- If fixed effects are not significant, then all variance is residual i.e. unexplained by model predictors
  - Hence always run significance tests and only keep significant factors
- In the school example, emphasis was on the multilevel random effects rather than on the fixed effects
  - *Gender* explains such a small fraction of residual variance that we could have run a model with random effects only (school.null)



Generalized linear mixed models

- *lmer* function also runs generalised linear models
  - *Family=binomial* for logistic regression (binary/proportion data)
  - *Family=poisson* for Poisson regression (count data)
  - etc.
  - fitting method: Laplace approximation, which provides P values for fixed effects!

- Example: *bacteria* file (library *MASS*)
  - patients tested for bacterial infection (variable *y*, yes or no)
  - fixed effect: *trt* (treatment), with 3 levels
  - drug, drug plus supplement, and placebo

  - Temporal pseudoreplication: patients tested multiple times over 11 weeks
    - Random effect variables: ID, week

Do treatments work?

- In all treatments, more people are infected than non infected but proportion changes

```
> table(bacteria$y,bacteria$trt)
   placebo drug drug+
  n      12   18    13
  y      84   44    49
```

- If we do not control for random effects and run proportion test, treatment is significant (see R code)

```
> prop.test(c(12,18,13),c(96,62,62))
 X-squared = 6.6585, df = 2, p-value = 0.03582
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3
0.1250000 0.2903226 0.2096774
```

- If we do not control for random effects and run a logistic regression with *glm*, drug treatment has significant effect (see R code)

Random interceps only: ID

```
> infection <- lmer(y~trt+(1|ID), family=binomial, data=bacteria)
> summary(infection)
Generalized linear mixed model fit by the Laplace approximation
Formula: y ~ trt + (1 | ID)
   Data: bacteria
   AIC   BIC logLik deviance
 214.3 227.9 -103.2    206.3
Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 0.96609  0.9829
Number of obs: 220, groups: ID, 50

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.2959     0.4077   5.631 1.79e-08 ***
trtdrug      -1.2021     0.5713  -2.104   0.0354 *
trtdrug+     -0.7096     0.5884  -1.206   0.2278
---
Correlation of Fixed Effects:
         (Intr) trtdrg
trtdrug  -0.714
trtdrug+ -0.693  0.495
```

- What happens after we control for random effects? (we have to!)
- Fixed effect: *trt*
- Random effects:
  - Let's control for ID only, not taking into account temporal patterns
    - many measurements coming from the same person, so control for ID
- Result: fixed effects not significant after controlling for ID
  - See ANOVA (R code)
  - drug treatment (but not drug+) significantly reduces infection

Controlling for temporal pseudoreplication

- However, simply taking ID into account misses the point that the multiple measurements of the same individual have a temporal structure
  - they are weekly measurements
  - this defines *temporal pseudoreplication*

- The term accounting for temporal pseudoreplication (the time measure: day, week, year) is entered before " | ", followed by another random effect (the entity temporally measured: ID, plant, etc.)
  - Temporal term is entered before ' | ' as a random slope
  - The term after ' | ' provides random intercepts
    - In this example, (week | ID), i.e. 'controlling for week by ID'

- `>infection2 <- lmer(y ~ trt + (week|ID), binomial, data=bacteria)`

Random slopes and intercepts

```
> infection2 <- lmer(y~trt+(week|ID), binomial,
data=bacteria)
> summary(infection2)
Generalized linear mixed model fit by the Laplace
approximation
Formula: y ~ trt + (week | ID)
   Data: bacteria
   AIC   BIC logLik deviance
 209.2 229.6  -98.6    197.2
Random effects:
 Groups Name        Variance Std.Dev. Corr
 ID     (Intercept) 0.147815 0.38447
        week        0.062371 0.24974  1.000
Number of obs: 220, groups: ID, 50

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.6195     0.4894   5.352  8.7e-08 ***
trtdrug      -1.2185     0.6588  -1.850   0.0644 .
trtdrug+     -0.5290     0.6991  -0.757   0.4492
---
Correlation of Fixed Effects:
          (Intr) trtdrg
trtdrug   -0.743
trtdrug+ -0.700  0.520
```

- Fixed effects:
  - neither treatment (drug, drug+) reduces infection rates
- Random effects:
  - Contribution of random intercepts (0.14) is higher than random slopes (0.062)

Significance of temporal term

- Model controlling for temporal pseudoreplication is significantly better

```
> infection.c <- lmer(y~trt+(1|ID),family=binomial, data=bacteria, REML=F)
> infection.d <- lmer(y~trt+(week|ID),family=binomial, data=bacteria, REML=F)
> anova(infection.c, infection.d)
Data: bacteria
Models:
infection.c: y ~ trt + (1 | ID)
infection.d: y ~ trt + (week | ID)
             Df    AIC    BIC   logLik  Chisq Chi Df Pr(>Chisq)
infection.c  4 214.32 227.90 -103.162
infection.d  6 209.21 229.57  -98.603 9.1184      2    0.01047 *
```

- Therefore week should kept *week* as a random (slope) term

Quiz

- Let's re-run the school results example, but his time without the *district* level

- 1) Run a log-likelihood test to assess significance of fixed effect

- 2) Is gender a significant predictor?
- What is the predicted difference in points between sexes?

- 3) Run a VCA using data from the optimal model
- Which spatial level accounts for most variance?

- 4) What is the residual variance?
- What could explain residual variance?

- 5) Compare variance explained by district in lecture example and street (in quiz). Why is that?
- The original dataset includes 6 districts with 10 streets each, i.e. 60 locations. Based on the results, do you think this ratio of district to town is ideal? How would you change it?