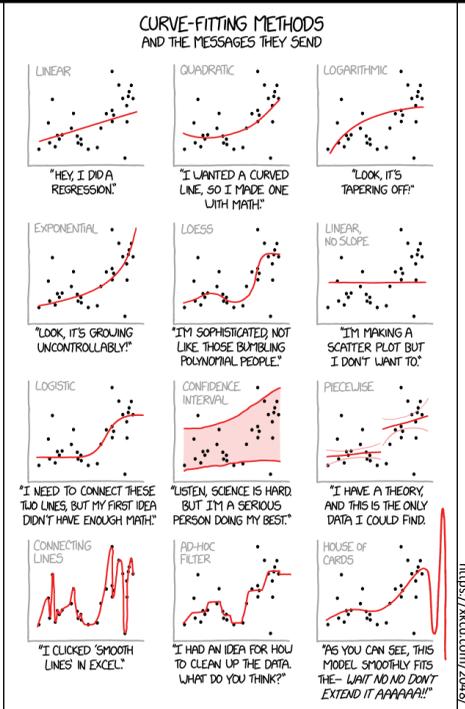


Linear mixed models & nested ANOVA

Lecture 7
Biological Statistics III
Ayco Tack



Outline

- *Linear models*
 - ❖ *Increasing our model collection*
- *Fixed and random effects*
- *Nested analysis of variance*
 - ❖ *Analysis of hierarchically grouped units*
- *Variance components*
 - ❖ *Estimating the variance of random effects*

A collection of linear models

For all models, the residual ε_{ij} or ε_{ijk} is normally distributed with mean zero and variance σ^2

- **One-way ANOVA:** $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where the α_i are fixed effects with $\sum_i \alpha_i = 0$ (Lecture 5)
- **Model with single random factor:** $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ where the α_i are random effects with normal distribution with mean zero and variance σ_α^2 (Today)
- **Nested mixed model:** $y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$ where the α_i are fixed effects with $\sum_i \alpha_i = 0$, and the $\beta_{j(i)}$ are random effects with mean zero and variance $\sigma_{\beta(\alpha)}^2$ (Today)
- **Two-way fixed effects ANOVA:** $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ where the $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i \gamma_{ij} = 0$ for all j , and $\sum_j \gamma_{ij} = 0$ for all i hold for the fixed effects (Lecture 8)
- **Two-way mixed model:** $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ where the α_i are fixed effects with $\sum_i \alpha_i = 0$ and the β_j and γ_{ij} are random effects with mean zero and variances σ_β^2 and $\sigma_{\alpha \times \beta}^2$ (Lecture 9)

The residual is a random effect. In many situations there can also be other random effects. For instance, with several data points from each individual, there is an additional random effect associated with the individual.

3

One-factor models

Fixed effects

- Each group (treatment) corresponds to the value of a qualitative variable of “general interest”
- If we repeat the experiment/observation we could use the same treatments again
- We might be interested in the treatment means *per se*
- The barley yield example is such a one-way ANOVA model

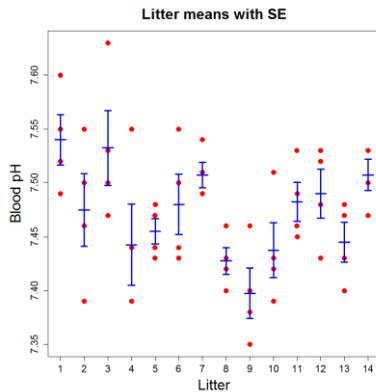
Random effects

- Each group is chosen randomly from a population of groups
- If we repeat the experiment/observation we would use a different random sample of groups
- We are not interested in the mean values of particular groups but we might want to estimate the variance of the group means

4

Example of model with random effect

Are there differences in blood pH among litters of mice?



Data:

- Blood pH for 14 different litters of mice
- Data from 4 male mice from each litter
- $a = 14$ litters, $n = 4$ individuals per litter

Analysis:

- In the old days, we used ANOVA:

ANOVA

Source	df	SS	MS	F	P
Litter	13	0.0876	0.0067	2.90	0.004
Error	42	0.0975	0.0023		
Total	55	0.1851			

- A more modern way is:

```
library(lme4)
fm <- lmer(pH ~ (1|Litter), data=dat)
library(lmerTest)
rand(fm)
Analysis of Random effects Table:
      Chi.sq Chi.DF p.value
Litter  6.58    1    0.01 *
***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

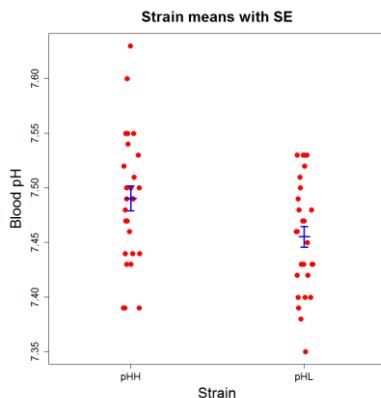
Conclusion:

- Yes there are differences among litters!

Model I ANOVA for the mice

The mouse blood pH data actually came from two different strains of mice:

- Litters 1 to 7 came from Strain pHH (selected for high pH)
- Litters 8 to 14 came from Strain pHL (selected for low pH)



Data:

- Blood pH for 28 mice from each strain

Question:

- We want to know if the strains really differ in pH

First (and wrong) approach: Ignore that there are litters

Analysis:

ANOVA table

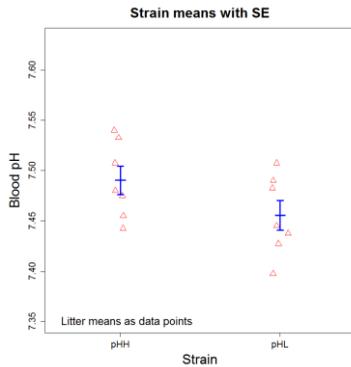
Source	df	SS	MS	F	P
Strain	1	0.0172	0.0172	5.51	0.023
Error	54	0.1680	0.0031		
Total	55	0.1851			

(Wrong) conclusion:

- The strains differ in blood pH

An ANOVA for the mice using means

But maybe there are differences between litters within strains?



Data:

- Litter mean blood pH for 7 litters from each strain

Second attempt:

- Use litter averages as data points

Analysis:

ANOVA table

Source	df	SS	MS	F	P
Strain	1	0.0043	0.0043	2.92	0.11
Error	12	0.0176	0.0015		
Total	13	0.0219			

Conclusion:

- The strains do not differ significantly in blood pH

7

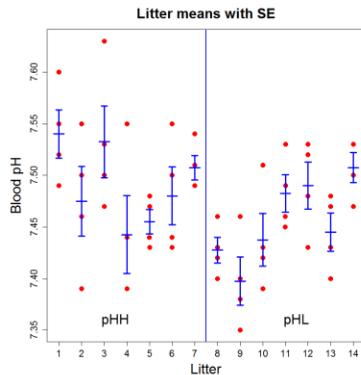


Mixed model lmer analysis of mouse blood pH



Let's take into account all sources of variation:

- Among strains
- Among litters within strain
- Within litters



Fit the model in R with lmer:

```
fm= lmer(pH ~ Strain + (1 | Litter),data=dat)
```

The command summary (fm) gives us the estimated variance components and the fixed effects (the effect of strain)

Output:

```
Random effects:
Groups Name Variance Std.Dev.
Litter (Intercept) 0.00088924 0.029803
Residual 0.00232143 0.048181
Number of obs: 56, groups: Litter, 14

Fixed effects:
Estimate Std. Error t value)
(Intercept) 7.49036 0.01448 517.1
StrainpHL -0.03500 0.02048 -1.7
```

Conclusion:

- For the fixed effects, there is an estimate of the strain difference in pH, but no test of significance
- We can use the Anova function (in package car) to get $p = 0.09$, which is not significant. There are no significant differences between strains
- In this case, it is wrong to pool litters
- You can also test the significance for random effects:


```
library(lmerTest)
fm <- lmer(pH ~ Strain +(1 | Litter),data=dat)
rand(fm)
```

8

Can we ignore litter?

Why did the test become stronger when we ignored litter?

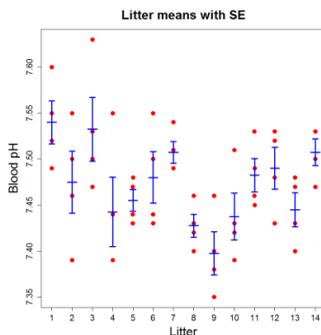
- With more degrees of freedom “in the denominator” an F -test tends to be more powerful (effectively, we have more data points for the test)

When is it OK to ignore a suggested grouping?

- First answer: It is **never OK**
- Second answer: It is OK when there is no *a priori* reason to expect group differences and the among group within treatment variation is non-significant when tested at high level (e.g. $\alpha = 0.25$)
- Third answer: It is OK if the AIC value is smaller for the model with pooling (i.e., smaller for the model where the random effect is dropped)
- For the mouse blood pH, we have *a priori* reasons to expect variation among litters within strain, namely shared genes and shared environment

9

Variance components in random effect models



Model with a single random effect:

- σ^2 = true within-group (residual) variance
- σ_a^2 = true variance of true group means

```
library(lme4)
fm <- lmer(pH ~ (1|Litter), data=dat)
summary(fm)
```

Random effects:

Groups	Name	Variance	Std.Dev.
Litter	(Intercept)	0.001105	0.03324
Residual		0.002321	0.04818

Number of obs: 56, groups: Litter, 14

$$R^2 = \frac{s_a^2}{(s_a^2 + s^2)} = \frac{0.001105}{(0.001105 + 0.002321)} = 0.323$$

Conclusion: 32% of the variation is among litters

10

Variance components in mixed effects models

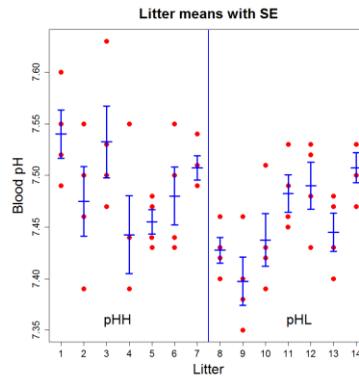
Mixed model nested ANOVA

- σ^2 = true within-group (residual) variance
- $\sigma_{\beta(\alpha)}^2$ = true variance of among groups within treatment

```
Random effects:
  Groups Name      Variance Std.Dev.
  Litter (Intercept) 0.00088824 0.029803
  Residual          0.00232143 0.048181
Number of obs: 56, groups: Litter, 14

Fixed effects:
              Estimate Std. Error t value
(Intercept)  7.49036    0.01448   517.1
StrainHL     -0.03500    0.02048   -1.7
```

As there is also variation associated with the fixed effect, we cannot easily calculate the total amount of variation explained by the random effect. But there are now methods to obtain R^2 values from the fixed and random components of mixed models: <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.2041-210x.2012.00261.x>



11

Allocation of sampling for nested ANOVA

Suppose we have a situation corresponding to mixed model nested ANOVA

- Should we try to get many groups within each treatment or many data points per group?

General principle:

- We want a small standard error for the treatment means
- If there is no extra cost in getting data from more groups (as compared to costs associated with getting another data point from the same group), we should get one data point per group ($n = 1$)

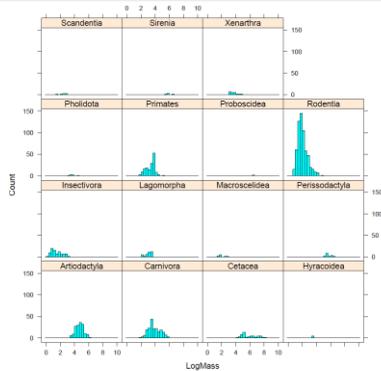
Or maybe we are also interested in the variation within each group?

- For example, we may like to know whether siblings within the same nest have the same level of immunity?
- Or how much variation there is in disease resistance within a plant population?

12

Adult weight in mammal species

Distribution of \log_{10} (adult weight) in grams in different mammal orders



Data:

The mammal species ($n = 1353$) are hierarchically divided into order ($n = 15$), family ($n = 97$) and genus ($n = 604$)

Aim & approach:

We want to know how the variation in log body weight is distributed over the hierarchical levels

- Order, Family and Genus could be random effects in a nested design
- We want to estimate the variance component for each of these random effects
- We can use the `lmer` function in the `lme4` package to fit a mixed model

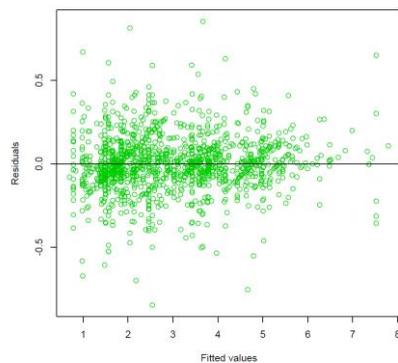
Code:

```
fm = lmer(LogMass ~ 1 +
          (1|Order/Family/Genus),
          data=dat)
```

13

Adult weight in mammal species

Plot of residual \log_{10} (adult weight) versus fitted values



Interpretation of the residual plot:

- The residual plot looks fine: no shotgun pattern or indication of non-linearity
- For this many data points, it should be easy to see deviations from variance homogeneity

Estimated variance components:

(expressed as standard deviations)

- Order: **1.441**
- Family in Order: **0.745**
- Genus in Family: **0.417**
- Within Genus: **0.213**

Conclusion:

- It seems there is more variation in log adult weight at higher taxonomic levels

14

Model diagnostics in mixed-effects models

The sjPlot package is great for model diagnostics:

- The function `plot_model(fim1, type = "diag")` provides nice diagnostics plots:
 - QQ-plot for the model residuals (to assess normality of the residuals)
 - QQ-plot for the random effects (to assess normality of the distribution)
 - Density plot of the distribution of the residuals (to assess the normality of the residuals again...)
 - Residuals versus predicted values
- The function `plot_model(fim1, type = "slope", show.data = TRUE)` plots the response variable as a function of each predictor, as well as the raw data points. Good to look for non-linear patterns in the raw data!
- The function `plot_model(fim1, type = "resid", show.data = TRUE)` plots the residual as a function of each predictor, as well as the raw data points. Good to look for non-linear patterns also in the residuals, as such patterns may be obscured in the raw data!

BTW, I also like some of the plots:

- `plot_model(fim1, type = "std")` shows the standardized regression coefficients. Use `plot_model(fim1, type = "std")`.
- `plot_model(fim2, type = "re")` shows the estimates for each level of the random effects. Often quite interesting to see what block or individual was extreme in its behaviour!
- `plot_model(fim2, type = "eff", show.data = TRUE)` shows the modelled relationships (i.e. ***not*** the relationship you would get when making boxplots or fitting lines through your raw data), as well as the raw data points. This is a nice way also to see whether the model fit makes sense.
- `plot_model(fim2, type = "int", mdrt.values = "meansd")` shows the predicted interactions, using three categories for the continuous variable (mean, mean + 1SD, mean-1SD). Or use `mdrt.values = "quart"`. This is a nice way also to see whether the model fit makes sense.

****Note that `se = TRUE` gives standard errors rather than confidence intervals***
 ***** Always plot the relationships between the response and each predictor also yourself******

15

Related reading and information

- **Quinn & Keough:** Sections 8.2 & 9.1
- **Crawley:** Sections 9.6, 9.7, 11.3, 11.4, 19.1 & 19.2

16