Statistical inference for linear mixed models

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

October 10, 2024

Outline

- general form of linear mixed models
- examples of analyses using linear mixed models
- prediction of random effects
- (estimation, including restricted maximum likelihood estimation))

One-way ANOVA in matrix-vector form

One observation:

$$Y_{ij} = \mu + U_i + \epsilon_{ij}$$

Vector of observations

$$Y = \mu 1_n + ZU + \epsilon$$

where Y, U and ϵ vectors of Y_{ij}'s, U_i's and ϵ_{ij}'s. 1_n vector of 1's and Z $n \times k$ matrix with $Z_{(ij)q} = 1$ if $q = i$ and zero otherwise.

Linear regression with random effects in matrix-vector form

Consider mixed model:

$$Y_{ij} = \beta_1 + U_i + [\beta_2 + V_i]x_{ij} + \epsilon_{ij}$$

May be written in matrix vector form as

$$Y = X\beta + ZU + \epsilon$$

where $\beta = (\beta_1, \beta_2)^{\mathsf{T}}$, $U = (U_1, \ldots, U_k, V_1, \ldots, V_k)^{\mathsf{T}}$, $\epsilon = (\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{km})^{\mathsf{T}}$, X is $n \times 2$ and Z is $n \times 2k$.

Linear mixed model: general form

Consider model

$$Y = X\beta + ZU + \epsilon$$

where $U \sim N(0, \Psi)$ and $\epsilon \sim N(0, \Sigma)$ are independent.

All previous models special cases of this.

Then Y has multivariate normal distribution

$$Y \sim N(X\beta, Z\Psi Z^{\mathsf{T}} + \Sigma)$$

General form is basis of linear mixed models software in R and SPSS.

Linear mixed models using lmer

General `lmer` model formulation

```
y~‘fixed formula’+(‘rand formula_1’|Group_1)+ ...
                              +(‘rand. formula_n’|Group_n)
```

translates into linear mixed model with independent sets of random effects for each grouping variable and e.g.

```
 (z|Group_i)
```

corresponds to

$$U_{il} + V_{il} z$$

i.e. model with random intercept and random slope for covariate z within each level l of grouping factor `Group_i`.

NB independence between levels of `Group_i` but intercept and slope dependent within level.

Only random intercept respectively slope: `(1|Group_i)` resp. `(-1+z|Group_i)`

Linear mixed models using lmer - cntd.

Procedure `lmer` is part of the `lme4` package.

`lmer` does not give *p*-values as default.

If you also load package `lmerTest`, *p*-values will be provided.

If you load `lmerTest`, `lme4` is also loaded.

Start by installing `lme4` and `lmerTest`

NB: with `lmer` noise ϵ always has covariance matrix $\Sigma = \sigma^2 I$.

Linear mixed model for orthodont data - independent random slope and intercept

```
> ort6=lmer(distance~age*Sex+(1|Subject)+(-1+age|Subject))
> summary(ort6)
 Groups    Name        Variance Std.Dev.
 Subject   (Intercept) 2.416451 1.55449
 Subject.1 age         0.007748 0.08802
 Residual              1.864634 1.36552
Number of obs: 108, groups:  Subject, 27

Fixed effects:
              Estimate Std. Error       df t value Pr(>|t|)
(Intercept)   16.34062    0.94087 67.09150  17.368  < 2e-16
age            0.78438    0.07944 67.09021   9.873 1.06e-14
SexFemale      1.03210    1.47405 67.09150   0.700   0.4862
age:SexFemale -0.30483    0.12446 67.09021  -2.449   0.0169
```

Linear mixed model for orthodont data - correlated random slope and intercept

```
> ort7=lmer(distance~age*Sex+(age|Subject))
> summary(ort7)
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 5.77441  2.4030
          age         0.03245  0.1801   -0.67
 Residual             1.71661  1.3102
Number of obs: 108, groups:  Subject, 27

Fixed effects:
              Estimate Std. Error       df t value Pr(>|t|)
(Intercept)   16.34063    1.01824 25.00829  16.048 1.12e-14
age            0.78437    0.08598 25.01351   9.123 1.97e-09
SexFemale      1.03210    1.59528 25.00829   0.647   0.5235
age:SexFemale -0.30483    0.13471 25.01351  -2.263   0.0326
```

Comparison of models for orthodont data

Fixed part: `age+Sex+age:sex`

Random part:

Model	AIC	BIC	logLik	Number of parameters
a	445.8	461.9	-216.9	4+2
bx	448.7	464.8	-218.4	4+2
$a + bx$, $\mathbb{C}\mathrm{ov}(a, b) = 0$	447.2	465.9	-216.6	4+3
$a + bx$	448.6	470	-216.3	4+4

Larger logLik and smaller AIC or BIC means better model.

The simplest one (just random intercept) seems better.

AIC and BIC

We can get better fit with more complex model - but we don't want too complex models

AIC and BIC are model selection criteria that attempts to find good compromise between model fit and model complexity (number of parameters)

In R: use functions `AIC()` and `BIC()`

CAUTION When estimation method REML (restricted maximum likelihood, see last slide) is used (is default), **need same** mean structure in the models compared.

Otherwise use estimation method MLE (maximum likelihood) if AIC or BIC used for model comparison:

```
ort7=lmer(distance~age*Sex+(age|Subject),REML=FALSE)
```

SPSS

Choose Analyze → Mixed models → Linear.

Need to specify ‘Subject’ variables - these correspond to the grouping variables for `lmer`.

With SPSS one can choose to model correlation in residuals ($\Sigma \neq \sigma^2 I$) - then one also need to specify a ‘Repeated’ variable (e.g. residuals for each subject may be correlated in time).

Specify fixed part of model using item ‘fixed’ and random part using item ‘random’ in menu.

Under random: several sets of random effects can be specified (corresponding to several `(|)` in `R`).

SPSS - continued

Under random: various options for covariance matrix of random effects within subject. Use covariance structure 'Variance Components' to get independent random effects or 'unstructured' to get dependent random effects.

Remember to include intercept.

Output: Type III F-tests for fixed effects.

See also power-point slides regarding SPSS.

Tests for fixed effects

SPSS produces Type III F-tests for fixed effects by default.

With `lmer` you **NEED TO** load `lmerTest`.

Then `anova` produces table with type III F-tests for fixed effects

If you don't use `lmerTest`, `anova` will produce **INCORRECT** F-tests when applied to output of `lmer`

Example: test of fixed effects

```
> library(lmerTest)
> ort4=lmer(distance~age*Sex+(1|Subject))
> anova(ort4)
Type III Analysis of Variance Table with Satterthwaite's method
         Sum Sq Mean Sq NumDF  DenDF  F value Pr(>F)
age     208.266 208.266     1  79.00 108.3559 <2e-16 ***
Sex       0.866   0.866     1 103.99   0.4507 0.5035
age:Sex  12.114  12.114     1  79.00   6.3027 0.0141 *
```

Sex specific slope (age:Sex) now significant at 5% level.

Sex specific intercepts (Sex) not significant.

CAUTION: in presence of age:Sex, interpretation of age coefficient depends on choice of reference category (boy or girl). Also interpretation of Sex coefficient depends on possible centering of age. Advisable not to pursue test for age or Sex in presence of age:Sex.

Nested two-way analysis of variance

For five cardboards we have 4 replications at 4 positions.

Hierarchical model (nested random effects)

$$Y_{ipj} = \mu + U_i + U_{ip} + \epsilon_{ipj}$$

$$\mathbb{V}\mathrm{ar}\, Y_{ipj} = \tau^2 + \omega^2 + \sigma^2$$

Covariance structure for nested random effects model

$$Y_{ipj} = \mu + U_i + U_{ip} + \epsilon_{ipj}$$

$$\mathbb{C}\mathrm{ov}(Y_{ipj}, Y_{lqk}) = \begin{cases} 0 & i \neq l \\ \tau^2 & i = l, p \neq q \text{ same card} \\ \tau^2 + \omega^2 & i = l, p = q \text{ same card and pos.} \\ \tau^2 + \omega^2 + \sigma^2 & i = 1, p = q, k = j \quad (\mathbb{V}\mathrm{ar}\, Y_{ipj}) \end{cases}$$

Nested two-way analysis of variance

```
> out2=lmer(Reflektans~(1|Pap.nr.)+(1|Pap.nr.*Sted))
> summary(out2)

Random effects:
 Groups         Name        Variance   Std.Dev.
 Pap.nr.        (Intercept) 1.6560e-02 0.1286843
 Pap.nr. * Sted (Intercept) 9.4539e-04 0.0307472
 Residual                   6.3494e-05 0.0079683
Number of obs: 80, groups: Pap.nr. * Sted, 20; Pap.nr., 5
```

Largest part of variance is between cardboard variance !

Explanation of `Reflektans~(1|Pap.nr.)+(1|Pap.nr.*Sted)`:

- no fixed formula: intercept always included as default
- `(1|Pap.nr.)` random intercepts for groups identified by variable `Pap.nr.` (card board effects)
- `(1|Pap.nr.*Sted)` random intercepts for groups identified by cross of variables `Pap.nr.` and `Sted` (positions within cardboard)
- random effects specified by different terms independent.

A more complicated example: gene-expression

Gene (DNA string) composed of substrings (exons) which may be more or less expressed according to treatment.

Expression measured as intensities on micro-array (chip). One chip pr. patient-treatment.

Factors: E (exon 8 levels), P (patient, 10 levels), T (treatment, 2 levels)

Y: vector of intensities (how much is exon expressed).

Model:

$$y_{ept} = \mu + \alpha_e + \beta_t + \gamma_{et} + U_p + U_{pt} + \epsilon_{ept}$$

U_{pt} and U_p random chip and patient effects.

Main question: are exons differentially expressed - i.e. are $\gamma_{et} \neq 0$ or not ?

Classical anova table:

```
> fit1=lm(intensity~treat*factor(exon)+factor(patient)+
                  factor(patient):treat,data=gene1)
> anova(fit1)
Analysis of Variance Table
                     Df  Sum Sq Mean Sq  F value    Pr(>F)
treat                1    3.242   3.242  14.4796 0.0002199
factor(exon)         7  254.343  36.335 162.2852 < 2.2e-16
factor(patient)      9   15.405   1.712   7.6449 6.703e-09
treat:factor(exon)   7    2.238   0.320   1.4278 0.1998234
treat:factor(patient) 9   8.190   0.910   4.0643 0.0001345
Residuals            126 28.211   0.224
```

We can estimate variances of ϵ_{ept}, U_{pt} and U_p as follows:

$$\hat{\sigma}^2 = 0.224$$
$$\hat{\sigma}^2_{P\times T} = (0.91 - 0.224)/8 = 0.08575$$
$$\hat{\sigma}^2_P = (1.712 - 0.91)/16 = 0.050125$$

F-test for no treatment-exon interaction: 1.4278 with *p*-value 0.1998.

I.e. interaction not significant - no evidence of differential exon usage.

Classical ANOVA:

- not straightforward to obtain estimates of variances from table of sums of squares (I will not go into detail with this).
- in the presence of random effects not straightforward to compute F-tests for fixed effects (which sums of squares should be used ?) - e.g. F-test for Treat is 3.563=3.242/0.910
- exact F-tests only available in balanced case (equal number of observations for each combination of factor levels)

Using lmer:

```
> fit1=lmer(intensity~treatment*factor(exon)+(1|patient)
                     +(1|factor(patient):treatment),data=ge
> summary(fit1)

Random effects:
 Groups                      Name        Variance Std.Dev.
 factor(patient):treatment (Intercept) 0.08577  0.2929
 patient                   (Intercept) 0.05011  0.2239
 Residual                                0.22389  0.4732
Number of obs: 160, groups:  factor(patient):treatment, 20
```

We directly obtain estimates of variance components.

Tests of fixed effects

Test for no treatment-exon interaction:

```
> library(lmerTest)
> fit1=lmer(intensity~treatment*factor(exon)+
           (1|patient)+(1|factor(patient):treatment),data=gene1)
> anova(fit1)
Type III Analysis of Variance Table with Satterthwaite's method
                  Sum Sq Mean Sq NumDF DenDF  F value  Pr(>F)
treatment          0.798   0.798     1     9   3.5625 0.09171 .
factor(exon)     254.343  36.335     7   126 162.2869 < 2e-16 **
treat:factor(exon) 2.238   0.320     7   126   1.4278 0.19982
```

Treatment-exon interaction not significant !

CAUTION: tests for main effects exon and treatment should ONLY be considered when interaction treatment:exon is NOT significant !

Tests for main effects

Interaction removed

```
> fit2=lmer(intensity~treatment+factor(exon)+(1|patient)+
                         (1|factor(patient):treatment),data=gene
> anova(fit2)
Type III Analysis of Variance Table with Satterthwaite's method
              Sum Sq Mean Sq NumDF DenDF  F value  Pr(>F)
treatment      0.816   0.816     1     9   3.5626 0.09171 .
factor(exon) 254.343  36.335     7   133 158.7120 < 2e-16 ***
```

Exon significant, treatment not !

Whether tests change after removal of interaction depends on specific model structure

With 12.5% missing data

20 of out 160 missing at random.

```
Random effects:
 Groups                      Name        Variance Std.Dev.
 factor(patient):treatment (Intercept) 0.10465  0.3235
 patient                  (Intercept) 0.02221  0.1490
 Residual                                0.22896  0.4785
Number of obs: 140, groups:  factor(patient):treatment, 20; pati
```

Adjusted F-test

```
> anova(fit1)
Type III Analysis of Variance Table with Satterthwaite’s method
                    Sum Sq Mean Sq NumDF  DenDF  F value Pr(>F)
treat                0.753  0.7529     1   9.04   3.2881 0.1030
factor(exon)       219.277 31.3253     7 107.41 136.8134 <2e-16 **
treat:factor(exon)   1.770  0.2528     7 107.41   1.1041 0.3659
```

Note: denominator degrees of freedom (DenDF) are not integers - this is due to adjustment in case of unbalanced data.

Classical ANOVA with random effects as linear mixed model

- classical ANOVA approach requires deep insight in order to calculate variance estimates and F-tests from classical ANOVA table.
- classical ANOVA requires balanced data.
- with general linear mixed models framework (`lmer`) everything is automatic.
- with general linear mixed models framework (`lmer`) adjustment of F-statistics in case of unbalanced data

Predictions/Residuals

The random effects U in a linear mixed model can be predicted using 'best linear unbiased prediction' (BLUP) - useful if we want to look at subject specific characteristics.

In the context of linear mixed models, BLUP $\hat{U}$ is the conditional mean of the random effects given the data:

$$\hat{U} = \mathbb{E}[U|Y = y]$$

Typically we assume ϵ_{ij} independent and $N(0, \sigma^2)$. To check this we can consider residuals:

$$\hat{\epsilon} = Y - X\hat{\beta} - Z\hat{U}$$

and perform the usual residual diagnostics.

With `lmer`: use `ranef`, `fitted` and `residuals` to extract BLUPS, fitted values and residuals.

SPSS: save predicted values and residuals under ‘Predicted values and residuals’

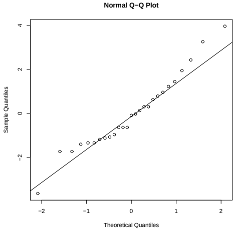
Example: orthodont data

Extract BLUPS, fitted values and residuals

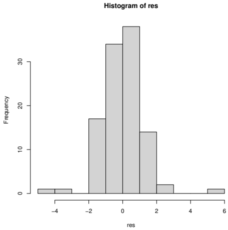
```
> childeffects=ranef(ort4)$Subject
> qqnorm(childeffects[[1]])
> qqline(childeffects[[1]])
> res=resid(ort4)
> hist(res)
> qqnorm(res)
> qqline(res)
> fitted=fitted(ort4)
> plot(fitted,res)
> boxplot(res˜Subject)
```

Plots

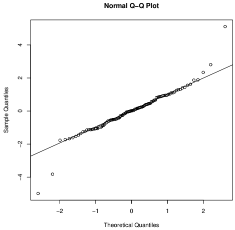
Random effects



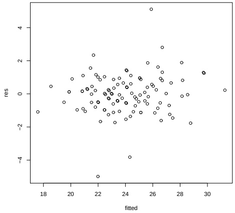
Residuals



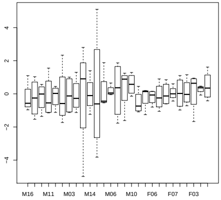
Residuals



Residuals vs. fitted



Residuals vs. subject



Outliers for two subjects !

Summary

- Linear mixed models flexible class of models for continuous observations.
- incorporates classical ANOVA models and random coefficients models
- Useful for modeling of correlated observations, for decomposition of variance and for estimation of population variances.
- Userfriendly software available

Exercises

1. Use `lmer` or Mixed models in SPSS to fit a one-way ANOVA model with random operator effects for the pulp data. Compare with results from previous exercise (classical anova for pulp data).
2. Install the R-package `faraway` which contains the data set `penicillin`. The response variable is yield of penicillin for four different production processes (the 'treatment'). The raw material for the production comes in batches ('blends'). The four production processes were applied to each of the 5 blends. Use `lmer` to fit anova models with production process as a fixed factor and blend as random factor. Compute an F-test for the effect of production process.

3. The rats data has variables (1) obs: observation number (2) treat: treament group ('con': control; 'hig': high dose; 'low': low dose) 3) rat: rat identification number (4) age: age of the rat at the moment the observation is made (5) respons: the response measured (height of skull) (6) logage: log-transformed age.

   The treatment is a drug that inhibits production of testosterone. The scientific question is whether/how the drug affects the growth rate of the rats.

   3.1 take a look at data by plotting response against age and logage (with separate curves for each rat).
   3.2 fit a linear regression model for the response with logage as the independent variable and an interaction between logage and treatment. Is the interaction between logage and treatment significant ? Is treatment significant ?

3. (continued) fit a linear mixed model by extending the previous models with random rat specific intercepts.
    3.3 what is the proportion of variance explained by the random intercepts ?
    3.4 What are the conclusions regarding interaction and treatment effects based on this model ? Compare with the previous model.
    3.5 Check the fitted linear mixed model using residuals and predicted random effects.

4. Write out X and Z matrix for model on slide ‘Linear regression with random effects in matrix-vector form’.

Estimation - technical background

For linear mixed model two sets of parameters: β (fixed effects) and ψ (random effects variances).

Maximum likelihood estimation: parameter estimates are those parameter values that make data most likely under the given model:

$$(\hat{\beta}, \hat{\psi}) = \operatorname*{argmax}_{\beta,\psi} f(y; \beta, \psi)$$

where $f(y; \beta, \psi)$ is the normal probability density of the data y.

Given ψ, $\hat{\beta}$ is the generalized least squares estimate:

$$\hat{\beta}(\psi) = (X^\mathsf{T} V(\psi)^{-1} X)^{-1} X^\mathsf{T} V(\psi)^{-1} y$$

which minimizes the generalized sum of squares

$$(y - X\beta)^\mathsf{T} V(\psi)^{-1} (y - X\beta).$$

In general ψ needs to be obtained by iterative maximization of

$$L(\psi) = f(y; \hat{\beta}(\psi), \psi)$$

One issue: MLE of ψ in general biased.

MLE's of variances biased

Consider simple normal sample $Y_i \sim N(\mu, \sigma^2)$.

MLE's:

$$\hat{\mu} = \bar{Y}. \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}.)^2$$

Bias of $\hat{\sigma}^2$:

$$\mathbb{E}\hat{\sigma}^2 = \sigma^2(n-1)/n$$

Bias arise from estimation of μ ($\sum_i (Y_i - \mu)^2$ vs $\sum_{i=1}^{n}(Y_i - \bar{Y}.)^2$).

Often we use instead unbiased estimate

$$s^2 = \frac{1}{n-1}\sum_i (Y_i - \bar{Y}.)^2$$

Similarly: maximum likelihood estimate of between subject variance in one-way anova is biased due to estimation of mean.

REML (restricted/residual maximum likelihood)

Idea: use linear transform of data which eliminates mean. Suppose design matrix $X : n \times p$ and let $A : n \times (n-p)$ have columns spanning the orthogonal complement $L^\perp$ of $L = \mathrm{span} X$. Then $A^\mathsf{T} X = 0$.

Transformed data $((n-p) \times 1)$

$$\tilde{Y} = A^\mathsf{T} Y = A^\mathsf{T} Z U + A^\mathsf{T} \epsilon$$

has mean 0 and covariance matrix $A^\mathsf{T} V(\psi) A$ where $V = Z \Psi Z^\mathsf{T} + \Sigma$ covariance matrix of Y and ψ covariance parameters. Then proceed as for MLE.

Default choice for estimation of variance parameters in both `lmer` and Mixed model in SPSS.

s^2 is one example of REML. Classical ANOVA variance estimates also REML.