Balanced mixed models - the geometric approach

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

February 28, 2012

Outline for today

- One-way ANOVA
- Two-way ANOVA
- Examples of data analyses

One-way anova

Let F be a factor/grouping variable with m levels and consider the model

$$y_{ij} = \mu + U_i + \epsilon_{ij}, \quad i = 1, \ldots, m,\ j = 1, \ldots, k$$

or

$$y = \mu 1_n + Z_F U + \epsilon$$

where n is total number of observations and Z_F is the design matrix corresponding to F: ij, qth entry of Z_F is 1 if y_{ij} belongs to the qth group and zero otherwise.

F is balanced if common number k of observations at each of the m levels (whereby $n = mk$). In this case, P_F (orthogonal projection on $L_F = \text{span} Z_F$) is

$$P_F = \frac{1}{k} Z_F Z_F^{\mathsf{T}}$$

Action of P_F: replaces y_{ij} by $\bar{y}_{i\cdot}$ (averages within each group).

Orthogonal decomposition of $\mathbb{R}^n$ (and hence of data vector Y):

$$\mathbb{R}^n = V_0 \oplus V_F \oplus V_I$$

where $V_0 = L_0 = \text{span}(1_n)$, $V_F = L_F \ominus V_0$ and $V_I = \mathbb{R}^n \ominus L_F$. Dimensions of V_0, V_F and V_I are 1, $m-1$ and $n-m$.

Orthogonal projections on V_0, V_F and V_I are $Q_0 = P_0 = 1_n 1_n^{\mathsf{T}}/n$, $Q_F = P_F - P_0$ and $Q_I = I - P_F$.

$$Y = Q_0 Y + Q_F Y + Q_I Y = P_F Y + Q_I Y$$

Covariance structure:

$$\mathbb{C}\text{ov}\, Y = k\sigma^2 P_F + \tau^2 I = \lambda P_F + \tau^2 Q_I$$

where $\lambda = k\sigma^2 + \tau^2$ and $Q_I = I - P_F$.

Note $P_F 1_n = Q_0 1_n = 1_n$ and $Q_I 1_n = 0$. Moreover $Q_I P_F = 0$.

Hence

$$\begin{bmatrix} P_F \\ Q_I \end{bmatrix} Y \sim N((1_n\mu, 0), \begin{bmatrix} \lambda P_F & 0 \\ 0 & \tau^2 Q_I \end{bmatrix}$$

We can thus base maximum likelihood estimation of (μ, λ) on $P_F Y$ and τ^2 on $Q_I Y$.

More precisely,

$$|\Sigma|^{-1/2} \exp(-\frac{1}{2}(Y - 1_n\mu)^T \Sigma^{-1}(Y - 1_n\mu)) =$$

$$\lambda^{-m/2} \exp(-\frac{1}{2\lambda}\|P_F Y - 1_n\mu\|^2) \times (\tau^2)^{m(k-1)/2} \exp(-\frac{1}{2\tau^2}\|Q_I Y\|^2)$$

$(\Sigma^{-1} = \tau^{-2} Q_I + \lambda^{-1} P_F$ and $|\Sigma| = \lambda^m (\tau^2)^{mk-m})$

Note: the two factors in the above likelihood are 'generalized' densities of the 'degenerate' normal vectors $P_F Y$ and $Q_I Y$.

Consider e.g. the factor $\lambda^{-m/2} \exp(-\frac{1}{2}\|P_F Y - 1_n\mu\|^2)$ involving the parameters λ and μ. We can maximize this with respect to λ and μ in exactly the same way as when we previously considered the likelihood of $N_n(X\beta, \tau^2 I)$ (see slide 'Estimation using orthogonal projections' in second set of handouts). Thus we obtain

$$\widehat{1_n\mu} = P_0 P_F Y = P_0 Y \text{ and } \lambda = \|P_F Y - P_0 Y\|^2/m = SSA/m$$

Proceeding in the same way for the second factor (where there is no mean parameter), we obtain

$$\hat{\tau}^2 = \|Q_I Y\|^2/(m(k-1)) = SSE/(m(k-1))$$

Note $Q_F Y \sim N_n(0, \lambda Q_F)$. By Exercise 5, $\|P_F Y - P_0 Y\|^2 = \|Q_F Y\|^2 \sim \lambda\chi^2(m-1)$ which has mean $\lambda(m-1)$. Thus $\hat{\lambda}$ is biased.

Unbiased estimate: $\tilde{\lambda} = \|P_F Y - P_0 Y\|^2/(m-1) = SSA/(m-1)$ (REML)

Implementation in R

For cardboard/reflectance data, $m = 34$ and $k = 4$.

```
> anova(lm(Reflektans~factor(Pap.nr.)))
Analysis of Variance Table

Response: Reflektans
                 Df  Sum Sq Mean Sq F value    Pr(>F)
factor(Pap.nr.)  33 0.90088 0.02730   470.7 < 2.2e-16 ***
Residuals       102 0.00592 0.00006
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence $\hat{\tau}^2 = 0.00006$, $\hat{\lambda}_P = 0.90088/34$ (or 0.0273) and $\hat{\sigma}^2 = (0.90088/34 - 0.00006)/4 = 0.00661$ (or $\hat{\sigma}^2 = (0.0273 - 0.00006)/4 = 0.00681$).

Biggest part of variation is between cardboard.

Likelihood-based inference

Choose parameters that maximize probability of obtaining the given data. Estimation: optimization problem.

Implemented in R function `lmer`:

```
> out1=lmer(Reflektans~(1|Pap.nr.),REML=F)
> summary(out1)
Linear mixed model fit by maximum likelihood
Formula: Reflektans ~ (1 | Pap.nr.)
    AIC    BIC logLik deviance REMLdev
 -726.5 -717.8  366.3   -732.5  -725.8
Random effects:
 Groups   Name        Variance   Std.Dev.
 Pap.nr.  (Intercept) 6.6096e-03 0.0812994
 Residual             5.7997e-05 0.0076156
Number of obs: 136, groups: Pap.nr., 34
```


Likelihood-based inference - REML

```
> out1=lmer(Reflektans~(1|Pap.nr.))#default is REML
> summary(out1)
...
Random effects:
 Groups   Name        Variance   Std.Dev.
 Pap.nr.  (Intercept) 6.8103e-03 0.0825247
 Residual             5.7997e-05 0.0076156
number of obs: 136, groups: Pap.nr., 34

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.61690    0.01417   43.54
```


Two-way ANOVA

Consider two factors: T (treatment) and P (plot). Moreover let $P \times T$ be the cross-factor (has a level for each combination of levels of P and T) and assume it is balanced. Then P and T balanced, too.

Model with random P and $P \times T$ effects (e.g. to account for soil variation)

$$y_{ptr} = \mu + \beta_t + U_p + U_{pt} + \epsilon_{ptr} \quad p = 1, \ldots, m,\ t = 1, \ldots, k,\ r = 1, \ldots, g$$

(NB: overparametrized) or

$$y = \xi + Z_P U_P + Z_{P\times T} U_{P\times T} + \epsilon$$

where $\xi = X_T\beta \in L_T = \text{span} X_T$.

Similar to one way anova, orthogonal decomposition:

$$\mathbb{R}^n = V_0 \oplus V_P \oplus V_T \oplus V_{P\times T} \oplus V_I$$

where $V_0 = L_0$, $V_T = L_T \ominus V_0$, $V_P = L_P \ominus V_0$, $V_{P\times T} = L_{P\times T} \ominus (V_P \oplus V_T \oplus V_0)$ and $V_I = \mathbb{R}^n \ominus L_{P\times T}$.

Dimensions: 1, $m-1$, $k-1$, $mk-m-k+1 = (m-1)(k-1)$, $n-mk$.

Orthogonal projections on 'V' subspaces: $Q_0 = P_0$, $Q_P = P_P - Q_0$, $Q_T = P_T - Q_0$, $Q_{P\times T} = P_{P\times T} - Q_P - Q_T - Q_0$ and $Q_I = I - P_{P\times T}$.

Crucial: balanced design implies $P_T P_P = P_P P_T = P_0$ whereby $Q_T Q_P = 0$. Hence V_T and V_P orthogonal. Similar for all other pairs of distinct Qs and Vs.

Covariance structure:

$$\mathbb{C}\mathrm{ov}\, Y = \sigma_P^2 n_P P_P + \sigma_{P\times T}^2 n_{P\times T} P_{P\times T} + \tau^2 I = \lambda_P P_P + \lambda_{P\times T}\tilde{Q}_{P\times T} + \tau^2 Q_I$$

where $\lambda_{P\times T} = \tau^2 + n_{P\times T}\sigma_{P\times T}^2$ and $\lambda_P = \tau^2 + n_{P\times T}\sigma_{P\times T}^2 + n_P\sigma_P^2$ and $\tilde{Q}_{P\times T} = Q_{P\times T} + Q_T$.

As before Y is decomposed into independent normal vectors

$$P_P Y \sim N(P_0\xi, \lambda_P P_P) \quad \tilde{Q}_{P\times T} Y \sim N(Q_T\xi, \lambda_{P\times T}\tilde{Q}_{P\times T}) \quad Q_I Y \sim N(0, \lambda_I Q_I)$$

Hence

$$\widehat{P_0\xi} = P_0 Y = \bar{y}_{\cdot\cdot} 1_n \quad \widehat{Q_T\xi} = Q_T Y \quad \hat{\xi} = P_T Y$$

$$\hat{\lambda}_P = \|P_P Y - P_0 Y\|^2/m = \|Q_P Y\|^2/m$$

$$\hat{\lambda}_{P\times T} = \|\tilde{Q}_{P\times T} Y - Q_T Y\|^2/(mk-m) = \|Q_{P\times T} Y\|^2/(mk-m)$$

$$\hat{\lambda}_I = \tau^2 = \|Q_I Y\|^2/(n-mk)$$

NB: since V_P and $V_{P\times T}$ have dimensions $m-1$ and $(m-1)(k-1)$ we often use these instead of m and $mk-m$ for λ_P and $\lambda_{P\times T}$ (REML).

Two-way analysis of variance - no treatment effect, nested random effects

$$y_{psr} = \mu + U_p + U_{ps} + \epsilon_{psr}$$

Variances $\sigma^2 = \sigma_P^2$, $\omega^2 = \sigma_{P\times S}^2$ and τ^2

Factors P ($m = 5$) og S ($k = 4$) og $g = 4$.

NB: $Q_S\xi = Q_S\mu 1_n = 0$.

NB: in ANOVA table (next slide), SSP=$\|Q_P Y\|^2$, SSS=$\|Q_S Y\|^2$, SSPS=$\|Q_{P\times S} Y\|^2$ and SSI=$\|Q_I Y\|^2$.

ANOVA table

```
 Analysis of Variance Table

Response: Reflektans
                          Df  Sum Sq  Mean Sq  F value    Pr(>F)
factor(Sted)               3 0.03600 0.011999  188.981 < 2.2e-16 ***
                                                                  #SSS
factor(Pap.nr.)            4 1.07520 0.268800 4233.472 < 2.2e-16 ***
                                                                  #SSP
factor(Sted):factor(Pap.nr.)
                          12 0.02168 0.001807   28.452 < 2.2e-16 ***
                                                                  #SSSP
Residuals                 60 0.00381 0.000063
                                                                  #SSE
```

$\hat{\tau}^2 = 0.00006$ $\hat{\lambda}_{P\times S} = (0.036 + 0.02168)/15 = 0.00385$
$\hat{\lambda}_P = 1.0752/5 = 0.21504$ (or $\hat{\lambda}_P = 1.0752/4 = 0.0072 = 0.2688$).

$\hat{\sigma}_P^2 = (0.21504 - 0.00385)/16 = 0.0132$
$\hat{\sigma}_{P\times S}^2 = (0.00385 - 0.00006)/4 = 0.0009475$ (or
$\hat{\sigma}_P^2 = (0.2688 - 0.00385)/16 = 0.01655937$)

Again: balanced design required - and difficult to remember rules for calculating variance components.

Using lmer and ML

```
> out2=lmer(Reflektans~(1|StedPap.nr.)+(1|Pap.nr.),REML=F)
> summary(out2)
Linear mixed model fit by maximum likelihood
Formula: Reflektans ~ (1 | StedPap.nr.) + (1 | Pap.nr.)
    AIC    BIC logLik deviance REMLdev
 -435.9 -426.4    222   -443.9  -439.9
Random effects:
 Groups      Name        Variance   Std.Dev.
 StedPap.nr. (Intercept) 9.4539e-04 0.0307472
 Pap.nr.     (Intercept) 1.3200e-02 0.1148899
 Residual                6.3494e-05 0.0079683
Number of obs: 80, groups: StedPap.nr., 20; Pap.nr., 5
```


Using lmer and REML

```
> out2=lmer(Reflektans~(1|StedPap.nr.)+(1|Pap.nr.))
> summary(out2)
Linear mixed-effects model fit by REML
Formula: Reflektans ~ (1 | StedPap.nr.) + (1 | Pap.nr.)
  AIC    BIC logLik MLdeviance REMLdeviance
 -434 -426.8    220     -443.8         -440
Random effects:
 Groups      Name        Variance   Std.Dev.
 StedPap.nr. (Intercept) 9.4531e-04 0.0307460
 Pap.nr.     (Intercept) 1.6559e-02 0.1286807
 Residual                6.3495e-05 0.0079684
number of obs: 80, groups: StedPap.nr., 20; Pap.nr., 5
```


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

- no fixed formula: intercept always included as default
- `(1|StedPap.nr.)` random intercepts for groups identified by variable `StedPap.nr.`
- `(1|Pap.nr.)` random intercepts for groups identified by variable `Pap.nr.`
- random effects specified by different terms independent.

Linear mixed models using lmer

General `lmer` model formulation

`y~'fixed formula'+('rand formula1'|Group1)+ ...+('rand. formulan'|Gr`

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

`(z|Groupi)`

corresponds to

$$U_i + V_i z$$

i.e. model with random intercept and random slope for covariate z within each level of grouping factor `Groupi`. NB independence between levels but intercept and slope dependent within level.

Only random slope :

`(-1+z|Groupi)`

Rep: specification of linear models in R

$y = \alpha + \beta x + \epsilon$	`y~x`
$y_{ij} = \mu + A_i + B_j + \epsilon_{ij}$	`y~A+B`
$y_{ij} = \mu + A_i + B_j + AB_{ij} + \epsilon_{ijk}$	`y~A+B+A:B`
	`y~A*B`
$y_{ij} = \mu + A_i + \beta_i x_{ij} + \epsilon_{ij}$	`y~A+A*x`
etc.	...

NB (regarding categorical variables): replace `A` with `factor(A)` if `A` not already declared a factor.

Rep: Multiple linear regression in R I

```
#fit model with sex specific intercepts and slopes
> ort1=lm(distance~age+age:factor(Sex)+factor(Sex))
> summary(ort1)
...
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            16.3406     1.4162  11.538  < 2e-16 ***
age                     0.7844     0.1262   6.217 1.07e-08 ***
factor(Sex)Female       1.0321     2.2188   0.465    0.643
age:factor(Sex)Female  -0.3048     0.1977  -1.542    0.126
...
> #compute F-tests respecting hierarchical principle
> drop1(ort1,test="F")
Single term deletions
....
                Df Sum of Sq    RSS    AIC F value  Pr(F)
<none>                      529.76 179.75
age:factor(Sex)  1     12.11 541.87 180.19  2.3782 0.1261
```

age:Sex not significant !

Multiple linear regression in R II

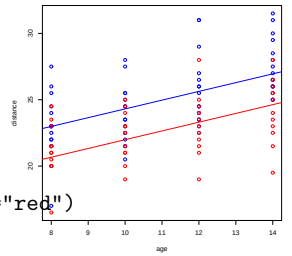
```
> ort2=lm(distance~age+factor(Sex))
> drop1(ort2,test="F")
Single term deletions

Model:
distance ~ age + factor(Sex)
            Df Sum of Sq    RSS    AIC F value     Pr(F)
<none>                  541.87 180.19
age          1    235.36 777.23 217.15  45.606 8.253e-10 ***
factor(Sex)  1    140.46 682.34 203.09  27.218 9.198e-07 ***
```

both age and sex significant

Multiple linear regression in R III

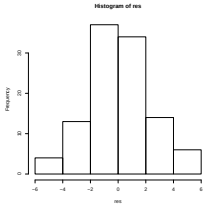
```
#plot data and two regression lines
col=rep("blue",length(Sex))
col[Sex=="Female"]="red"
plot(distance~age,col=col)
abline(parm[1:2],col="blue")
abline(c(parm[1]+parm[3],parm[2]),col="red")
```



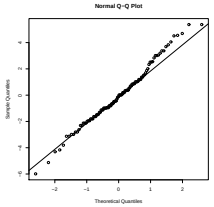
Multiple linear regression in R IV

```
res=residuals(ort2)
```

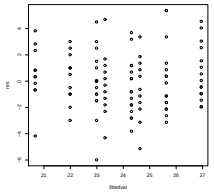
```
hist(res)
```




```
qqnorm(res)
qqline(res)
```

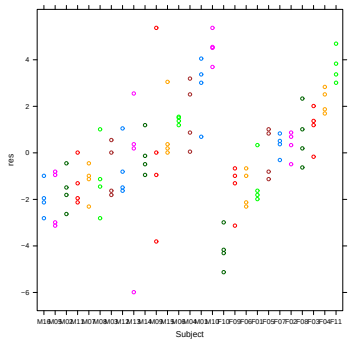



```
fittedval=fitted(ort2)
plot(res~fittedval)
```



Multiple linear regression in R V

```
> library(lattice)
> xyplot(res~Subject,groups=Subject)
```



Oups - residuals not independent and identically distributed ! Hence computed F-tests not valid.

Problem: subject specific intercepts (and possibly subject specific slopes too)

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

```
> ort6=lmer(distance~age+(1|Subject)+(-1+age|Subject),data=Orthodont)
> summary(ort6)
Linear mixed model fit by REML
Formula: distance ~ age + (1 | Subject) + (-1 + age | Subject)
   Data: Orthodont
   AIC   BIC logLik deviance REMLdev
 453.3 466.7 -221.7    439.7   443.3
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 1.921089 1.38603
 Subject  age         0.022277 0.14925
 Residual             1.878652 1.37064
Number of obs: 108, groups: Subject, 27

Fixed effects:
            Estimate Std. Error t value
(Intercept)  16.7611     0.7138   23.48
age           0.6602     0.0656   10.06
```


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

```
> ort7=lmer(distance~age+(age|Subject),data=Orthodont)
> summary(ort7)#high correlation between intercept and slope
Linear mixed model fit by REML
Formula: distance ~ age + (age | Subject)
   Data: Orthodont
   AIC   BIC logLik deviance REMLdev
 454.6 470.7 -221.3    439.2   442.6
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 5.415097 2.32704
          age         0.051269 0.22643  -0.609
 Residual             1.716205 1.31004
Number of obs: 108, groups: Subject, 27

Fixed effects:
            Estimate Std. Error t value
(Intercept) 16.76111    0.77525  21.620
age          0.66019    0.07125   9.265
```

Exercises

1. Check that $P_T P_P = P_0$.
2. Show that an orthogonal projection only has eigen values 1 or 0.
3. For a symmetric matrix A show that $|A|$ is the product of A's eigen values.
4. Show, that if $S = aP + bQ$ where P and Q are orthogonal projections with $P + Q = I$ then the eigen values of S are the non-zero eigen values a and b of aP and bQ.
5. Show that $\|PY\|^2 \sim \sigma^2\chi^2(d)$ if $Y \sim N(0, \sigma^2 P)$ and P is an orthogonal projection on subspace of dimension d (hint: use spectral decomposition and result above regarding Ps eigen values).

6. 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 where applied to each of the 5 blends. Fit anova models with production process as a fixed factor and blend as random factor. Try to use both the anova table and `lmer`.
7. fit linear models for the orthodontic growth curves with subject specific intercepts. Draw histograms of the fitted intercepts (can be extracted using `coef()`). Check residuals from the model.
8. fit a linear mixed model to the orthodontic data with independent random intercepts. Compare the random effects variance with the variances of the fitted intercepts from 3.
9. try to introduce also random slopes in the linear mixed model from 7.

10. Use lmer to compare ML and REML estimates for the Orthodont data (consider mixed model with age, Sex and random intercepts).
11. Suppose the factor F has 4 levels each with 3 observations. Write down three different parametrizations and associated design matrices of the mean vector corresponding to this factor.