The linear mixed model: introduction and the basic model

Analysis of Experimental Data

Contents

1 Data with non-independent observations

1.1 Non-independency

How is it that non-independence occurs? There are generally two ways in which this might happen:

- the explicit way: data is (by design) collected in a way that implies non-independence; for example
  - students are sampled from many schools (from several school districts)
  - within-subjects designs
- the implicit way: some unintended factors ‘spoil’ the data:
  - data on alcohol use (and abuse) in a large (random) sample have been collected by means of interviews; 5 different interviewers (each with his/her own style) have been involved in the data collection process
  - data on depression are collected in a large (random) sample over a period of ten days; the last two days were rainy

1.2 Different types of data with non-independent observations

Hierachical data

- sampling takes place at two or more levels, one *nested* within the other
- for example: students within schools within school districts (three levels)
- for example: patients within clinical centers (two levels)
- sometimes, the nesting is not perfect: students belong two several schools, patients have visited several clinical centers (this is called: partial nesting, or partial crossing)
- observations within the same 'level' are not independent of one-another

Clustered data

- some observations 'naturally' belong/cluster together
- the clusters/groups are randomly sampled; the units within a cluster/group are automatically included

- for example: family members
- for example: teeth in a mouth
- special case: *dyadic data*: pairwise data; for example: man and woman from the same couple
- observations belonging to the same cluster/group are not independent

Matched data

- 1 set of observations is sampled randomly; one (or several other) set(s) are chosen is such a way that they 'match' the original set
- for example: case-control studies
- matching can be on one dimension (for example 'age') or on several dimensions ('age', 'gender', 'iq', ...)
- 1-1 matching, 1-N matching, ...

- observations that are ‘matched’ are somehow similar and therefore not independent

Longitudinal data

- longitudinal data are collected when individuals (or units) are followed over time
- the evolution over time is the focus of the data
- for example: annual data on vocabulary growth among children
- for example: blood pressure of patients measured every week (for a period of two years)
- often considered as ‘hierarchical data’: the individual measurements are the lower level, the individuals are the higher level (but the time points are not always randomly sampled, but fixed)
- observations from the same ‘individual’ are not independent of one-another

Repeated measures

- repeated measures are collected when individuals (or units) are measured under different (experimental) conditions
- typical for experimental within-subjects designs
- time is (generally) not important
- the experimental conditions are (usually) considered as fixed
- longitudinal data are often called repeated measures too
- observations from the same 'individual' are not independent of one-another

1.3 The problem with non-independent observations

- technically: the standard linear model insists that error (and hence) observations are independent; if not, the model is not valid
- conceptually:
  - the accuracy of the estimates (and our confidence in them) is captured in the standard errors of the regression coefficients;
  - the linear model assumes there are N independent pieces of information when it computes these standard errors (actually, $N - p - 1$, once the degrees of freedom for the coefficients are subtracted);
  - if observations are correlated, however, there really aren't N independent pieces of information, and the estimated standard errors will be too small

1.4 Ways to deal with non-independent observations

- Corrected or robust standard errors:
  - if the only concern is to obtain (more) accurate standard errors for a standard regression model, there are ways to correct them using conventional regression software (e.g., Huber-White corrected standard errors, sometimes referred to as a 'sandwich estimator')
  - cfr. the `sandwich` package
  - cfr. generalized estimating equations (GEE)
- The fixed-effects approach:
  - since it is conditional independence that is assumed, if predictors in the model can account for all sources of correlation between observations, then this assumption will be satisfied;
  - for instance, to control for school effects among J schools, a set of $(J$-$1)$ dummy codes might be entered into the regression model to 'covary out' the mean differences among the schools

- each specific school will have its own fixed (constant) regression coefficient
- if school mean differences are the only source of non-independence of observations, then these differences will be controlled, leaving conditionally independent residuals that satisfy the assumptions of the standard linear model

- The mixed-effects approach:
  - same as the fixed-effects approach, but we consider ‘school’ as a random factor
  - mixed-effects models include more than one source of random variation

1.5 The many faces of mixed-effects models

- mixed-effects models have been developed in a variety of disciplines, with varying names and terminology:
  - random-effects models, random-effects ANOVA (statistics, econometrics)
  - variance components models (statistics)
  - hierarchical linear models (education)
  - multi-level models (sociology, education)
  - contextual-effects models (sociology)
  - random-coefficient models (econometrics)
  - repeated-measures models, repeated measures ANOVA (statistics, psychology)
  - …
- while essentially similar, the various approaches differ in terms of:

- motivation and notation
- assumptions concerning the random effects
- estimation method

- the ‘older’ approaches date back to Fisher and Yates’s work on split-plot agricultural experiments
- the ‘modern’ approaches are more general (for example: they can handle irregular observations, missing data, . . .)
- some of the main references of the ‘modern’ mixed models approach are:
  - Harville (1977, JASA) reviewed previous work, unified the methodology, and described iterative ML algorithms
  - Patterson and Thompson (1971, Biometrika) proposed the alternative REML approach
  - Laird and Ware (1982, Biometrics): their notation is still the standard
  - Jennrich and Schluchter (1986, Biometrics)
  - Laird, Lange, and Stram (1987, JASA)

- Diggle (1988, Biometrics)
- Lindstrom and Bates (1988, JASA)
- Jones and Boadi-Boateng (1991, Biometrics)
- ...

- some of the main references of the use of these mixed models in the behavioural sciences are:
  - Raudenbush, S.W. & Bryk A.S. (2001, second edition). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks, Calif.: Sage. (Cfr. HLM software)
  - Goldstein, H. (2002). *Multilevel Statistical Models.* London: Edward Arnold. (Cfr. Mlwin software)
  - Snijders, T. & Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* Thousand Oaks, Calif.: Sage.
  - Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, N.J.: Lawrence Erlbaum Associates.

1.6 Software packages for mixed-models

- MLwiN 2
  - Goldstein and co-workers
  - very popular in europe; lots of users in the FPPW
  - URL: `http://www.cmm.bristol.ac.uk/MLwiN/index.shtml`
- HLM 6
  - Raudenbush and Bryk
  - very popular in the USA
  - URL: `http://www.ssicentral.com/`
- Mplus
  - Bengt Muthén
  - does structural equation modeling, multilevel modeling and mixture modeling all at once

  - URL: `http://www.statmodel.com/`

- SAS (PROC MIXED)
  - often considered the ‘golden standard’
  - very flexible, extremely good documentation
- SPSS (MIXED)
  - since version SPSS 14
  - very basic, poor documentation
- R
  - the older package `nlme` is very flexible, but slow and out-dated
  - the newer package `lme4` is extremely fast, state-of-the-art, but not as flexible as `nlme` or SAS PROC MIXED

2 The linear mixed model

2.1 The one-way random-effects ANOVA model revisited

The model using classic ‘effect’ notation

- the effect model can be written as follows: for observation j, level i:

$$y_{ij} = \mu + a_i + \epsilon_{ij}$$

- μ: the population mean (intercept)
- a_i: the (random) effect for level i of the random factor A
- ϵ_{ij}: the (random) error term
- the variances σ_a^2 and σ_ϵ^2 are called *variance components*

A toy dataset: 6 subjects, 3 scores per subject

```
> subject <- factor(rep(1:6, each = 3))
> score <- factor(rep(1:3, times = 6))
> y <- c(10, 11, 12, 10, 12, 14, 12, 13, 14, 12, 14, 16, 14, 15,
+     16, 14, 16, 18)
> Data <- data.frame(subject, score, y)
> Data
```

```
   subject score  y
1        1     1 10
2        1     2 11
3        1     3 12
4        2     1 10
5        2     2 12
6        2     3 14
7        3     1 12
8        3     2 13
9        3     3 14
10       4     1 12
11       4     2 14
12       4     3 16
13       5     1 14
14       5     2 15
15       5     3 16
16       6     1 14
17       6     2 16
18       6     3 18
```

Exploring the data

- grand mean and variance

```
> mean(Data$y)

[1] 13.5

> var(Data$y)

[1] 4.852941
```

- mean/variance per subject

```
> with(Data, tapply(y, list(subject = subject), mean))

subject
 1  2  3  4  5  6
11 12 13 14 15 16

> with(Data, tapply(y, list(subject = subject), var))

subject
1 2 3 4 5 6
1 4 1 4 1 4
```

- treat 'subject' as a fixed factor

```
> fit.lm <- lm(y ~ 1 + subject, data = Data)
> anova(fit.lm)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
subject    5   52.5    10.5     4.2 0.01939 *
Residuals 12   30.0     2.5
---
Signif. codes:  0
```

- treat 'subject' as a random factor (classic approach)

```
> fit.aov <- aov(y ~ 1 + Error(subject), data = Data)
> summary(fit.aov)

Error: subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  5   52.5    10.5

Error: Within
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 12     30     2.5
```

- how can we compute σ_a^2 from this output? note that $\sigma_\epsilon^2 = \text{MSE} = 2.5$, and $\text{E(MSA)} = \sigma_\epsilon^2 + n \cdot \sigma_a^2$, therefore $\sigma_a^2 = (\text{MSA} - \text{MSE})/n = (10.5 - 2.5)/6 = 2.6666$

- treat 'subject' as a random factor (modern approach)

```
> library(lme4)
> fit.lmer <- lmer(y ~ 1 + (1 | subject), data = Data)
> summary(fit.lmer)

Linear mixed model fit by REML
Formula: y ~ 1 + (1 | subject)
   Data: Data
   AIC   BIC logLik deviance REMLdev
 79.89 82.56 -36.94    75.16   73.89
Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 2.6667   1.6330
 Residual             2.5000   1.5811
Number of obs: 18, groups: subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  13.5000     0.7637   17.68
```

The model using ‘modern’ notation

- the model in Laird-Ware form:

$$\begin{aligned} y_{ij} &= \beta_0 x_{0ij} + b_{0i} z_{0ij} + \epsilon_{ij} \\ &= \beta_0 + b_{0i} + \epsilon_{ij} \end{aligned}$$

- β_0: the (fixed) intercept, representing the population mean
- $x_{0ij} = 1$: the fixed-effect (constant) regressor
- b_{0i}: the random-effect coefficient for subject i, representing the deviation of subject i from the general mean
- $z_{0ij} = 1$ the random-effect (constant) regressor
- ϵ_{ij}: the (random) error term, representing the deviation of the jth score of subject i from the subject mean

The variance components

- the two variance components are denoted by:
  - $\text{Var}(b_{0i}) = d_0^2 = d^2$
  - $\text{Var}(\epsilon_{ij}) = \sigma_\epsilon^2 = \sigma^2$
- we typically assume the random-effect coefficients are uncorrelated and normally distributed:
  - $b_{0i} \sim \text{N}(0, d^2)$
  - $\epsilon_{ij} \sim \text{N}(0, \sigma^2)$

2.2 The structure of the model

The *Laird-Ware form* of the linear mixed model (two-level):

$$
\begin{aligned}
y_{ij} = & \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \ldots + \beta_p x_{pij} + \\
& b_{1i} z_{1ij} + b_{2i} z_{2ij} + \ldots + b_q z_{qij} + \\
& \epsilon_{ij}
\end{aligned}
$$

- y_{ij} is the value of the response variable for the jth of n_i observations of individual/cluster $i = 1, 2, \ldots, N$
- $x_{i1}, \ldots, x_{ip}$ are the values of the p regressors for observation i; they are *fixed* constants (with respect to repeated sampling), and can be anything (product terms, indicator variables, ...)
- in many regression models, a constant term $x_{i0} = 1$ is added; β_0 is called the intercept
- the regression coefficients $\beta_1, \ldots, \beta_p$ are the fixed-effect coefficients, which are identical for all individuals/clusters

- $b_{1i}, b_{2i}, \ldots, b_{qi}$ are the random-effect coefficients for individual/cluster i; the random-effect coefficients are thought of as random variables, not as parameters (similar to the errors ϵ_{ij})
- $z_{1ij}, z_{2ij}, \ldots, z_{qij}$ are the random-effect regressors; they are typically a subset of the fixed regressors; in many models, a random intercept term $z_{0ij} = 1$ is added and b_{0i} is called a random intercept
- ϵ_{ij} is the random error for the jth observation of individual/cluster i

Stochastic assumptions of the linear mixed model

- the usual assumptions for the random-effect coefficients:
  - $b_{ki} \sim \mathrm{N}(0, d_k^2)$
  - $\mathsf{Var}(b_{ki}) = d_k^2$ (constant across individual/clusters)
  - $\mathsf{Cov}(b_{ki}, b_{k'i}) = d_{kk'}$ (within the same i)
  - $\mathsf{Cov}(b_{ki}, b_{k'i'}) = i$ (for $i \neq i'$)
- the usual assumptions for the error terms:
  - $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_{ijj}^2)$
  - $\mathsf{Var}(\epsilon_{ij}) = \sigma_{ijj}^2$ (constant across individual/clusters)
  - $\mathsf{Cov}(\epsilon_{ij}, \epsilon_{ij'}) = \sigma_{ijj'}$ (within the same i)
  - $\mathsf{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ (for $i \neq i'$)
  - the ϵ_{ij} and b_{ki} are uncorrelated

Model in matrix notation

The model Laird-ware model in matrix form:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \qquad i = 1, 2, \ldots, N$$

where for observations of the ith individual/cluster:

- $\mathbf{y}_i$ is the $n_i \times 1$ response vector
- $\mathbf{X}_i$ is the $n_i \times p$ model matrix for the fixed effects
- $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect coefficients
- $\mathbf{Z}_i$ is the $n_i \times q$ model matrix for the random effects
- $\mathbf{b}_i$ is the $q \times 1$ vector of random-effect coefficients
- $\boldsymbol{\epsilon}_i$ is the $n_i \times 1$ vector of errors

or for the all individuals/clusters:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

Stochastic assumptions in matrix notation

- the random-effect coefficients:
  - $\mathbf{b}_i \sim \mathrm{N}_q(\mathbf{0}, \mathbf{D})$
  - $\mathbf{b}_i$ and $\mathbf{b}_{i'}$ are independent for $i \neq i'$
- the error terms:
  - $\boldsymbol{\epsilon}_i \sim \mathrm{N}_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i)$
  - $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_{i'}$ are independent for $i \neq i'$
  - $\mathsf{Cov}[\boldsymbol{\epsilon}_i, \mathbf{b}_i] = \mathbf{0}$
- $\mathbf{D}$ contains $q(q+1)/2$ parameters; $\boldsymbol{\Sigma}_i$ contains $n_i(n_i+1)/2$ elements

The variance components

- the elements of $\mathbf{D}$ and the elements of $\mathbf{\Sigma}_i$ (for each i) are collectively called the *variance components*
- the elements of $\mathbf{D}$ are often parametrized in terms of a smaller number of fundamental parameters
- for hierarchical data, where observations within an individual/cluster are sampled independently, it is often assumed that the error variance is constant; in this case $\mathbf{\Sigma}_i$ simplifies to $\sigma^2\mathbf{I}_{n_i}$
- for longitudinal data, $\mathbf{\Sigma}_i$ may be specified to capture serial (over-time) correlation among the errors

Conditional and marginal distributions

- because the right-hand side of the linear mixed model contains two random vectors $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$, we can consider two different distributions:
  - the conditional distribution:

$$\mathbf{y}_i|\mathbf{b}_i \sim \mathrm{N}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i,\ \boldsymbol{\Sigma}_i)$$

  - the marginal distribution:

$$\mathbf{y}_i \sim \mathrm{N}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$$

  where

$$\mathsf{Cov}[\mathbf{y}_i] = \mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \boldsymbol{\Sigma}_i$$

- $\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'$ represents the random-effects structure: it implies a covariance structure of a very specific form
- combining the N clusters, we get

$$\mathsf{Cov}[\mathbf{y}] = \mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \boldsymbol{\Sigma}$$

where $\mathsf{Cov}[\mathbf{y}]$ contains $N(N+1)/2$ elements

- a prime reason for having random effects in a model is to simplify the otherwise difficult task of specifying the $N(N+1)/2$ distinct elements of $\mathsf{Cov}[\mathbf{y}]$; without using random effects we would have to deal with elements of $\mathsf{Cov}[\mathbf{y}]$ being a variety of forms; but with random factors we can conveniently deal with variances and covariances attributable to factors acknowledged to be affecting the data

A toy dataset: 6 subjects, 3 scores per subject

- at the subject level:

$$\mathbf{X}_i = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{Z}_i = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} d^2 \end{bmatrix}, \quad \text{and} \quad \mathbf{\Sigma}_i = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

$$\begin{aligned} \mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{\Sigma}_i &= \begin{bmatrix} d^2 & d^2 & d^2 \\ d^2 & d^2 & d^2 \\ d^2 & d^2 & d^2 \end{bmatrix} + \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 + d^2 & d^2 & d^2 \\ d^2 & \sigma^2 + d^2 & d^2 \\ d^2 & d^2 & \sigma^2 + d^2 \end{bmatrix} \end{aligned}$$

- full model:

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{V}_3 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_4 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_5 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{V}_6 \end{bmatrix}$$

2.3 Parameter estimation

Estimating the fixed coefficients β

- if we assume that $\mathbf{D}$ and $\mathbf{\Sigma}$ and hence $\mathbf{V}$ are known, the weighted least-squares estimator of $\boldsymbol{\beta}$ using a symmetric weight matrix $\mathbf{W}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{y}$$

where the weight matrix is given by:

$$\mathbf{W} = \mathbf{V}^{-1} = \mathsf{Cov}[\mathbf{y}]^{-1}$$

- if $\mathbf{V}$ is known, $\hat{\boldsymbol{\beta}}$ is normally distributed with mean $\boldsymbol{\beta}$; the variances are the diagonal elements of the variance-covariance matrix:

$$\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = [\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}]^{-1}$$

- if $\mathbf{V}$ is unknown, it is replaced by its (current) estimate $\hat{\mathbf{V}}$

- if $\mathbf{V}$ is unknown, the variances of $\hat{\boldsymbol{\beta}}$ underestimate the correct variability because the variability introduced by estimating the variance components is not taken into account
- in practice, one often accounts of this downward bias by using approximate t- and F-statistics for testing hypotheses about β (see inference for fixed effects)

Estimating the variance components

- historically, several methods have been proposed: the general analysis of variance method, Henderson's method I, II and III, MINQUE, MIVQUE, and I-MINQUE
- two methods used in the 'modern' approach are:
  - maximum likelihood (ML)
  - restricted maximum likelihood (REML) estimation methods

  both are iterative methods

2.4 Inference in a linear mixed model

Inference for fixed effects

- all tests of the fixed effects can be expressed by a linear hypothesis of the form
$$H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$$
- $\mathbf{L}$ is a $h \times (p+1)$ matrix; each row in $\mathbf{L}$ specifies a linear combination of the parameters to be tested
- a common statistic is the so-called Wald F-statistic:
$$F = \frac{(\mathbf{Lb})' \left[\mathbf{L}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{L}'\right]^{-1} (\mathbf{Lb})}{h}$$
- unfortunately, despite its name, this statistic does not follow an F-distribution (due to the fact that $\mathbf{V}$ is unknown)
- one approach is to assume that F approximates an F-distribution, with h degrees of freedom for the nominator, and a to be estimated number of denominator degrees of freedom

- several methods are available to estimate the denominator degrees of freedom:
  - the containment method (often the default in software)
  - the Satterhwaite's (1946) approximation
  - the Kenward and Roger (1997) approximation
  - ...
- this 'problem' has received little attention in the literature, presumably because in the context of, say, longitudinal data, sample sizes are huge, and the p-values are often very similar under different approximation methods
- for relatively small datasets (for example: experimental within-subjects design), this problem can not be ignored!
- another alternative is to use a Bayesian approach

Inference for variance components

- a typical hypothesis is that the variance of a random coefficient equals zero:

$$H_0 : d_k^2 = 0$$

- many software packages report Wald tests for all variance components
  - unfortunately, Wald tests assume asymptotic normality of the parameter estimates, and this is *not* satisfied for the variance components, due the 'boundary problem'
  - for example, the value of $d_k^2 = 0$ is on the boundary of the parameter space, and in this case, classical MLE results don't hold in this context
- alternatively, one can use likelihood ratio tests
  - again, the asymptotic chi-squared null-distribution is not valid, due the boundary problem
  - however, some results are available for the case $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$