Recent work, ideas and wishes for future activites in mixed models

Søren Højsgaard

Department of Mathematical Sciences

Aalborg University, Denmark

August 12, 2013

Contents

1 Background

- Since 2012, HoD at Dept. of Mathematical Sciences, Aalborg University.
- Before that, working in agricultural science institution funded by external projects.
- Author/maintainer of `pbkrtest` package (tests in linear mixed models)
- Author of `geepack` package (inference in generalized estimating equation "models").
- Author/maintainer of various packages related to graphical models - and of other packages too.

2 Tests in mixed models – the degree of freedom police

Take two mixed models

$$M : y = Xb + Zu + e \text{ and } M_0 : y = X_0 b_0 + Zu + e$$

where $C(X_0) \subset C(X)$ and $d = \dim(X) - \dim(X_0)$.

Test for M_0 assuming M.

The usual χ^2 approximation to the LR-test statistic tend to give too small p–values.

Effects appear "more significant than they really are".

2.1 Parametric bootstrap

- The `pbkrtest` package implements parametric bootstrap for testing M_0 under M.
- Idea: Let t be observed value of some test statistic (in principle any statistic you like; in the implementation it is the LR test statistic). In which reference distribution should t be evaluated? Asymptotic χ^2_d distribution is not very good for small samples.
- Alternative: Simulate e.g. $B = 1000$ new datasets under the fitted hypothesis, i.e. $M_0 : y = X_0 b_0 + Zu + e$ where estimates for b_0 etc. are plugged in.
- Refit M and M_0 to each of these simulated datasets and calulate tests statistic in each case. This gives are reference distribution $\{t^1, \dots, t^B\}$ in which we can evaluate if t is large.

2.2 Kenward–Roger approximation

- The `pbkrtest` package implements Kenward-Roger method for calculating DDF (denominator degrees of freedom) for F-test in LMMs.
- Test M_0 under M. Corresponds to testing $Lb = 0$ for a suitable restriction matrix L. Consider Wald–like test statistic:

$$W = \lambda \frac{1}{d} b^\top L^\top (LVL^\top)^{-1} Lb$$

  where d is degrees of freedom implied by restriction matrix L.
- Has asymptotic $\frac{1}{d}\chi^2_d$ distribution, i.e. $F_{d,\infty}$ distribution.
- Calculate asymptotic mean and variance of W and equate with $F_{d,m}$ distribution to obtain denominator degrees of freedom m.
- Basically $F_{d,m}$ has "heavier tail" than $F_{d,\infty}$.

Status and future:

- Paper about `pbkrtest` in JSS hopefully on its way.
- Straight forward (in principle) to extend parametric bootstrap to generalized linear mixed models (GLMMs).
- In K–R–approximation we calculate the expected information matrix for the parameters.
- Alternative: Calculation of the average of the expected and the observed information matrices is much easier (and faster) to calculate because some of the nasty terms disappear.
- In principle K–R–approximation can be extended to GLMMs; presumably it is quite technical to get to work.

3 Example - sugar beets

```
Experimental plan for sugar beets experiment

Sowing dates:
  1: 4/4, 2: 12/4, 3: 21/4, 4: 29/4, 5: 18/5

Harvesting dates:
  1: 2/10, 2: 21/10

Plot allocation:
            |  Block 1         |  Block 2         |  Block 3         |
            +-----------------|-----------------|-----------------+
Split-plots | h1 h1 h1 h1 h1 | h2 h2 h2 h2 h2 | h1 h1 h1 h1 h1 | Harvesting ti
 1-15       | s3 s4 s5 s2 s1 | s3 s2 s4 s5 s1 | s5 s2 s3 s4 s1 | Sowing time
            -----------------|-----------------|-----------------|
Split-plots | h2 h2 h2 h2 h2 | h1 h1 h1 h1 h1 | h2 h2 h2 h2 h2 | Harvesting ti
16-30       | s2 s1 s5 s4 s3 | s4 s1 s3 s2 s5 | s1 s4 s3 s2 s5 | Sowing time
            +-----------------|-----------------|-----------------+
```

Notation: i: harvesting dates ($i = 1, 2$), j: block ($j = 1, 2, 3$), k: sowing dates ($k = 1, \ldots, 5$).

A typical model for such an experiment would be

$$y_{ijk} = \mu + \alpha_i + \beta_j + \delta_k + U_{ij} + \epsilon_{ijk}, \tag{1}$$

where $U_{ij} \sim N(0, \omega^2)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$. Notice that U_{ij} describes the random variation between whole–plots (within blocks) and the presence of this term implies that measurements on the same split–plot will be positively correlated.

```
R> data( beets )
R> head( beets, 4 )
  harvest  block  sow yield sugpct
1   harv1 block1 sow3   128   17.1
2   harv1 block1 sow4   118   16.9
3   harv1 block1 sow5    95   16.6
4   harv1 block1 sow2   131   17.0
R> beet0<-lmer(sugpct~block+sow+harvest+(1|block:harvest), data=beets,
R> beet_no.harv <- update(beet0, .~.-harvest)
R> anova(beet0, beet_no.harv)
Data: beets
Models:
beet_no.harv: sugpct ~ block + sow + (1 | block:harvest)
beet0: sugpct ~ block + sow + harvest + (1 | block:harvest)
             Df     AIC     BIC logLik deviance  Chisq Chi Df
beet_no.harv  9 -69.084 -56.473 43.542  -87.084
beet0        10 -79.998 -65.986 49.999  -99.998 12.914      1
             Pr(>Chisq)
beet_no.harv
beet0         0.0003261 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R> (p <- PBmodcomp(beet0, beet_no.harv, nsim=100))
Parametric bootstrap test; time: 4.37 sec; samples: 100 extremes: 2;
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + sow + (1 | block:harvest)
         stat df   p.value
LRT    12.914  1 0.0003261 ***
PBtest 12.914    0.0297030 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R> summary( p )
Parametric bootstrap test; time: 4.37 sec; samples: 100 extremes: 2;
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + sow + (1 | block:harvest)
            stat     df    ddf   p.value
PBtest   12.9142                0.0297030 *
Gamma    12.9142                0.0158083 *
Bartlett  5.2025 1.0000         0.0225545 *
F        12.9142 1.0000 3.3492 0.0308176 *
LRT      12.9142 1.0000         0.0003261 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R> (k <- KRmodcomp(beet0, beet_no.harv))
F-test with Kenward-Roger approximation; computing time: 0.08 sec.
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + sow + (1 | block:harvest)
       stat   ndf   ddf F.scaling p.value
Ftest 15.21  1.00  2.00         1  0.0599 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R> summary( k )
F-test with Kenward-Roger approximation; computing time: 0.08 sec.
large : sugpct ~ block + sow + harvest + (1 | block:harvest)
small : sugpct ~ block + sow + (1 | block:harvest)
        stat   ndf   ddf F.scaling p.value
Ftest  15.21  1.00  2.00         1  0.0599 .
FtestU 15.21  1.00  2.00            0.0599 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 Nice–to–have extensions of `lmer()`

- The `lmer()` restriction that $e \sim N(0, \sigma^2 I)$ is (potentially) severe.

  Would be nice to have at least some flexibility for specifying non–trivial `repeated`–effects (in SAS jargon).

  - AR(1)
  - Toeplitz
  - Heterogeneous variance
  - ....

  Are there low–hanging apples based on “working residuals”?

- Would be nice with “flexible” residuals:
  - $r_1 = y - X\hat{b}$
  - $r_2 = y - X\hat{b} - Z\hat{u}$ where $\hat{u}$ is BLUP.

- Would be nice with predict method (is it there now?). Two (at least) kinds of prediction errors are relevant
  - based on $\mathbb{V}\mathrm{ar}(e)$
  - based on $\mathbb{V}\mathrm{ar}(e)$ and $\mathbb{V}\mathrm{ar}(u)$
- Flexible prediction method. I have implementation...
- 'lsmeans' and other contrasts. I have implementation...
- More controlled output

5 Linking to graphical models

Example: Measure p behavioural traits y_{ij} on piglets $j = 1, \ldots J$ from litters $i = 1, \ldots, I$.

Piglets from same litters are correlated because of genetics.

With $y = Xb + Zu + e$ where $\mathbb{V}\mathrm{ar}(u) = G$ and $\mathbb{V}\mathrm{ar}(e) = R$ we have – in a sloppy notation

$$y|u \sim N(Xb + Zu, R)$$

Hence, R describes the correlation between traits after the genetic component has been taken away.

Model direct and indirect associations by imposing zero's in concentration matrix $K = R^{-1}$. That is the idea in graphical Gaussian models.