Inferens i mixed models i R - hinsides det sædvanlige likelihood ratio test

Inference in mixed models in R - beyond the usual asymptotic likelihood ratio test

Søren Højsgaard
http://people.math.aau.dk/~sorenh/
University of Aalborg, Denmark

28. Januar, 2019

Outline and take-home message

- Mixed models (random effects, random regression etc.) models handled by `lme4` package in R.
- Tests are by default based on χ^2 approximation of LR test statistic.
  - Works fine with "large samples" / "large dataset", but not with small samples.
  - Main concern: Effects can appear to be "more significant than they really are".
  - Source of confusion: A dataset can be large with respect to some aspect of a model while small with respect to other.

- The R package `pbkrtest` provides some remedies:
  - Base test on F-statistic, where denominator degrees of freedom are estimated from data.
  - Base test on parametric bootstrap where data are simulated under the model (carries over to e.g. generalized linear mixed models).
- Look at simulated and real data
- Notice: Talk and paper (with correction) available at http://people.math.aau.dk/∼sorenh/

History: The degree-of-freedom police

- SH raised issue about calculating degrees of freedom on R-help
  - 2006: `[R] how calculation degrees freedom` see:
    - SH: Along similar lines . . . probably in recognition of the degree of freedom problem. It could be nice, however, if anova() produced . . .
    - Doug Bates: I don't think the "degrees of freedom police" would find that to be a suitable compromise. :-)
- In reply to related question:
  - Doug Bates: I will defer to any of the "degrees of freedom police" who post to this list to give you an explanation of why there should be different degrees of freedom.
- Main point: Quite different views on whether the degree-of-freedom issue really is an issue or not.

Example: Double registration in labs

Clustered data:

- Compare two groups (treatment with a control);
- M units (petri plates, persons, animals. . .) per group;
- Each unit is measured R times. Measurements on same unit are positively correlated.

Simulated data: Two groups, $N = 3$ subjects per group, $R = 3$ replicated measurements per subject.

y1	grp	subj	y1	grp	subj
67	ctrl	subj1	26	trt1	subj4
72	ctrl	subj1	45	trt1	subj4
140	ctrl	subj1	90	trt1	subj4
13	ctrl	subj2	48	trt1	subj5
27	ctrl	subj2	53	trt1	subj5
37	ctrl	subj2	95	trt1	subj5
-76	ctrl	subj3	70	trt1	subj6
-66	ctrl	subj3	99	trt1	subj6
-56	ctrl	subj3	131	trt1	subj6

Problem/issues: If we ignore clustering/positive correlation:

- Pretend to have more information than we have
- Standard errors of estimates become too small
- p-values become too small

Ignore clustering

Simple regression model

$$y_{gir} = \mu + \beta_g + e_{gir}$$

	Estimate	Std. Error	t value	Pr(>|t|)	Pr(>X2)
(Intercept)	17.6	18.8	0.934	0.364	0.350
grptrt1	55.4	26.6	2.086	0.053	0.037

Notice: the *t*-test “accounts for” the uncertainty in the estimate of the standard error; gives larger *p* values.

Analyse average

Compute average for each subject and consider model

$$\bar{y}_{gi} = \mu + \beta_g + e_{gi},$$

	Estimate	Std. Error	t value	Pr(>|t|)	Pr(>X2)
(Intercept)	17.6	34.0	0.516	0.633	0.606
grptrt1	55.4	48.1	1.152	0.314	0.249

Notice: Both tests give large *p*-values suggesting no effect at all.

Model with random effects

Consider variance component model

$$y_{gir} = \mu + \beta_g + U_{gi} + e_{gir}$$

term	estimate	std.error	statistic	Pr(>X2)
(Intercept)	17.6	34.0	0.516	0.606
grptrt1	55.4	48.1	1.152	0.249

Notice: p-values (for χ^2-test) same as when analyzing average. No t-test available.

The Kenward–Roger approach

- Multivariate normal model

$$Y \sim N(X\beta, \Sigma)$$

- Test of the hypothesis

$$L(\beta - \beta_0) = 0$$

  where L is a regular matrix of estimable functions of β.

- With $\hat{\beta} \sim N_d(\beta, \Phi)$, a Wald statistic is

$$W = [L(\hat{\beta} - \beta_0)]^\top [L\Phi L^\top]^{-1} [L(\hat{\beta} - \beta_0)]$$

  which is asymptotically $W \sim \chi^2_d$ under the null hypothesis.

Consider scaled version of W:

$$F = \frac{1}{d} W = \frac{1}{d} (\hat{\beta} - \beta_0)^\top L^\top [L^\top \Phi(\hat{\sigma}) L]^{-1} L(\hat{\beta} - \beta_0).$$

In the computations:

- $\hat{\sigma}$ is vector of REML estimates for elements in $\Sigma = \mathbb{V}ar(Y)$ and
- $\hat{\beta}$ is REML estimate for β.
- $\Phi(\sigma) = (X' \Sigma(\sigma) X)^{-1} \approx \mathbb{C}ov(\hat{\beta})$,
- Asymptotically $F \sim \frac{1}{d}\chi^2_d$ under the null hypothesis
- Think of $\frac{1}{d}\chi^2_d$ as the limiting distribution of an $F_{d,m}$–distribution as $m \to \infty$

Kenward and Roger's modification

To account for the fact that $\Phi = \mathbb{V}ar(\hat{\beta})$ is estimated from data, we

Consider variance component model

$$y_{gir} = \mu + \beta_g + U_{gi} + e_{gir}$$

	statistic	ndf	ddf	F.scaling	p.value
Ftest	1.33	1	4	1	0.314

Notice: Same *p*-value as when averages are analysed.

However, analysing averages is not always an option.

Parametrisk bootstrap

Consider two competing models: A large model $f_1(y;\theta)$ and a simpler sub model $f_0(y;\theta_0)$.

The test statistic for testing the simpler model under the larger is t_{obs}.

The p-value becomes:

$$p = \sup_{\theta \in \Theta_0} Pr_\theta(T \geq t_{obs}),$$

where supremum is under the hypothesis.

Usually supremum can not be computed. Instead we base p value on the parameter estimate:

$$p^{PB} = Pr_{\hat{\theta}}(T \geq t_{obs}),$$

In praxis, p^{PB} is approximated as:

1. Draw B parametric bootstrap datasets $D^1, \dots D^B$ from the fitted null model $f_0(\cdot; \hat{\theta}_0)$.
2. Fit the large and the null model to each of these datasets.
3. Compute the likelihood ratio (LR) test statistic for each simulated dataset. This gives the reference distribution for the test statistic.
4. Compute how extreme the observed test statistic is in the reference distribution; this gives the p value.

Consider variance component model

$$y_{gir} = \mu + \beta_g + U_{gi} + e_{gir}$$

	statistic	df	p.value
LRT	1.47	1	0.225
PBtest	1.47	NA	0.346

Notice: *p*-values close to those based on Kenward-Roger approximation.

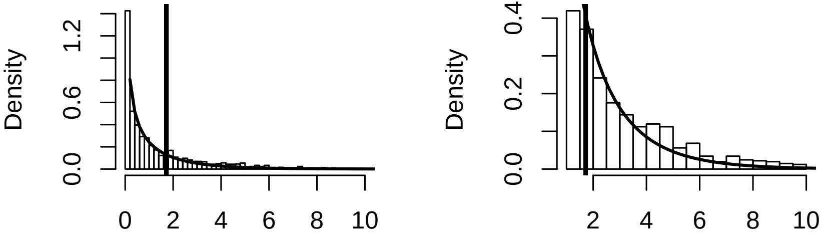

Figure 1: χ^2 distribution and simulated reference distribution.

Simulation study

y1	grp	subj	y1	grp	subj
67	ctrl	subj1	26	trt1	subj4
72	ctrl	subj1	45	trt1	subj4
140	ctrl	subj1	90	trt1	subj4
13	ctrl	subj2	48	trt1	subj5
27	ctrl	subj2	53	trt1	subj5
37	ctrl	subj2	95	trt1	subj5
-76	ctrl	subj3	70	trt1	subj6
-66	ctrl	subj3	99	trt1	subj6
-56	ctrl	subj3	131	trt1	subj6

- Task: Test the hypothesis that there is no effect of treatment. How good are the various tests?

- Simulate data 1000 times with divine insight: there is no effect of treatment.

	0.01	0.05	0.10
lm+F	0.21	0.31	0.41
lm+X2	0.24	0.35	0.42
avg_lm+F	0.01	0.06	0.11
avg_lm+X2	0.07	0.13	0.19
mixed+X2	0.05	0.14	0.23
mixed+F	0.01	0.06	0.11
mixed+PB	0.01	0.05	0.10

Sugar beets - A split–plot experiment

- Model how sugar percentage in sugar beets depends on harvest time and sowing time.
- Five sowing times (s) and two harvesting times (h).
- Experiment was laid out in three blocks (b).

Experimental plan for sugar beets experiment

```
# Plot allocation:
#       |  Block 1       |  Block 2       |  Block 3       |
#       +----------------|----------------|----------------+
# Plot  | h1 h1 h1 h1 h1 | h2 h2 h2 h2 h2 | h1 h1 h1 h1 h1 | Harvest time
# 1-15  | s3 s4 s5 s2 s1 | s3 s2 s4 s5 s1 | s5 s2 s3 s4 s1 | Sowing time
#       |----------------|----------------|----------------|
# Plot  | h2 h2 h2 h2 h2 | h1 h1 h1 h1 h1 | h2 h2 h2 h2 h2 | Harvest time
# 16-30 | s2 s1 s5 s4 s3 | s4 s1 s3 s2 s5 | s1 s4 s3 s2 s5 | Sowing time
#       +----------------|----------------|----------------+
```

beets data

```r
data(beets, package='pbkrtest')
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
```

- A typical model for such an experiment would be:

$$y_{hbs} = \mu + \alpha_h + \beta_b + \gamma_s + U_{hb} + \epsilon_{hbs}, \tag{1}$$

  where $U_{hb} \sim N(0, \omega^2)$ and $\epsilon_{hbs} \sim N(0, \sigma^2)$.
- Notice that U_{hb} describes the random variation between whole–plots (within blocks).

Using `lmer()` from lme4 we can test for no effect of sowing and harvest time as:

```
beet.lg <- lmer(sugpct ~ block + sow + harvest +
                      (1 | block:harvest), data=beets, REML=FALS
beet.noh <- update(beet.lg, .~. - harvest)
beet.nos  <- update(beet.lg, .~. - sow)
```

Both factors appear highly significant

```
anova(beet.lg, beet.noh)  %>% as.data.frame
```

```
##           Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## beet.noh  9 -69.1 -56.5   43.5    -87.1    NA     NA         NA
## beet.lg  10 -80.0 -66.0   50.0   -100.0  12.9      1   0.000326
```

```
anova(beet.lg, beet.nos)  %>% as.data.frame
```

```
##           Df   AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## beet.nos  6  -2.8   5.61    7.4    -14.8    NA     NA         NA
## beet.lg  10 -80.0 -65.99   50.0   -100.0  85.2      4   1.37e-17
```

However, the LRT based *p*–values are anti–conservative: the effect of harvest appears stronger than it is.

```r
set.seed("260618")
KRmodcomp(beet.lg, beet.noh)
```

```
## F-test with Kenward-Roger approximation; time: 0.10 sec
## large : sugpct ~ block + sow + harvest + (1 | block:harvest)
## small : sugpct ~ block + sow + (1 | block:harvest)
##       stat  ndf  ddf F.scaling p.value
## Ftest 15.2  1.0  2.0         1    0.06
```

```r
PBmodcomp(beet.lg, beet.noh)
```

```
## Bootstrap test; time: 5.18 sec;samples: 1000; extremes: 27;
## large : sugpct ~ block + sow + harvest + (1 | block:harvest)
## small : sugpct ~ block + sow + (1 | block:harvest)
##        stat df p.value
## LRT    12.9  1 0.00033
## PBtest 12.9    0.02797
```

As the design is balanced we may make F–tests for each of the effects as:

```
beets$bh <- with(beets, interaction(block, harvest))
summary(aov(sugpct ~ block + sow + harvest +
                Error(bh), data=beets))
```

```
##
## Error: bh
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      2 0.0327  0.0163    2.58   0.28
## harvest    1 0.0963  0.0963   15.21   0.06
## Residuals  2 0.0127  0.0063
##
## Error: Within
##           Df Sum Sq Mean Sq F value  Pr(>F)
## sow        4   1.01  0.2525     101 5.7e-13
## Residuals 20   0.05  0.0025
```

Final remarks

- Satterthwaite approximation of degrees of freedom on its way in `pbkrtest`. Faster to compute than Kenward-Roger scales to larger problems.
- `pbkrtest` available on CRAN
  https://cran.r-project.org/package=pbkrtest
- devel version on github:
  `devtools::install_github(hojsgaard/pbkrtest)`

- `pbkrtest` described in Ulrich Halekoh and SH (2014) A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models The R Package pbkrtest; Journal of Statistical Software, Vol 59. Please cite if you publish results using the package.

Thanks for your attention!