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

Søren Højsgaard*

28. januar, 2020

Resumé: Inferens i lineære mixed models og i generaliserede lineære mixed models er ofte baseret på χ^2 approximationen til likelihood ratio teststørrelsens fordeling. Det går som regel godt i store datasæt, men et datasæt kan på samme tid være stort med hensyn til nogle aspekter af en problemstilling og lille med hensyn til andre aspekter. Et klassisk eksempel er data fra et split-plot forsøg: Delploteffekten kan være velbestemt mens helploteffekten ofte vil være dårligere bestemt. I visse planlagte forsøgstyper ved vi, hvordan vi skal håndtere hypotesetests i sådanne modeller. I observationelle studier er det mindre klart, hvordan man skal håndtere hypotesetests. Een mulighed mulighed er at lave en form for F-test hvor nævner-frihedsgraderne er justerede (typisk) for at tage hånd om, at dispersionsparametre er estimerede fra data og dermed ikke må betragtes som kendte. En anden tilgang er at basere tests på parametrisk bootstrap. Fordelen ved denne metode er, at den umiddelbart lader sig anvende i mere generelle situationer end lineære mixed models; f.eks. i generaliserede lineære modeller. Begge metoder er tilgængelige i R pakken `pbkrtest`.

1 Introduktion

Mixed models håndteres i `R` (R Core Team 2019), oftest med `lme4` pakken (Bates et al. 2015). Tests er baserede på χ^2 approksimationen af likelihood ratio (LR) teststørrelsen, hvilket fungerer fint i store datasæt men ofte mindre godt i små datasæt. Dertil kommer, at et datasæt kan være stort med hensyn til nogle aspekter af en model og samtidig lille med hensyn til andre aspekter. R-pakken `pbkrtest` tilbyder alternativer til χ^2 approksimationen af LR teststørrelsen, nemlig: 1) Tests baserede på en F-teststørrelse (hvor nævnerfrihedsgraderne estimeres fra data), 2) tests baserede på parametrisk bootstrap (hvor data er simuleret under modellen). Parametrisk bootstrap kan også bruges for tests i generaliserede lineære modeller. Med (lineære) mixed models forstås i det følgende modeller af formen

$$y = X\beta + Zu + e$$

hvor y og e er n vektorer af stokastiske variable, X er $n \times p$ model matrix for systematiske effekter, β er p vektor af regressionskoefficienter, Z er $n \times q$ model matrix for de tilfældig effekter og u er q vector af tilfældige effekter. Det antages at $u \sim N(0, G)$ og $e \sim N(0, R)$ og at u og e er uafhængige. Generaliserede lineære modeller modeller er som generaliserede lineære modeller at det antages at $g(\mu) = X\beta + Zu$, hvor g er linkfunktionen.

*University of Aalborg, Denmark

Table 1: Simuleret datasæt. y_1 er en numerisk respons.

	y1	y2	grp	subj	y1	y2	grp	subj
2	67	1	ctrl	subj1	26	2	trt1	subj4
3	72	1	ctrl	subj1	45	1	trt1	subj4
1	140	1	ctrl	subj1	90	0	trt1	subj4
4	13	1	ctrl	subj2	48	2	trt1	subj5
6	27	2	ctrl	subj2	53	3	trt1	subj5
5	37	1	ctrl	subj2	95	2	trt1	subj5
8	-76	0	ctrl	subj3	70	2	trt1	subj6
7	-66	2	ctrl	subj3	99	0	trt1	subj6
9	-56	3	ctrl	subj3	131	0	trt1	subj6

2 Eksempel: Dobbeltregistreing i laboratorieforsøg

Betragt et konstrueret, men meget simpelt, eksempel: Vi ønsker at sammenligne to grupper (f.eks. behandling mod kontrol). Til rådighed er der M *units* (petriskåle, personer, dyr...) per gruppe og der måles på hver unit ialt R gange. Målinger på samme unit vil oftest give anledning til *clustering* i data. Et datasæt i "long format" vil altså have $T = 2 \times M \times R$ rækker. Table 1 viser et simuleret datasæt med $M = 3$ units per gruppe; $R = 3$ gentagne målinger per unit. Problemstillingen er, at målinger på samme unit typisk er positivt korrelerede, og hvis man ikke tager højde for dette, så kan man komme til at overvurdere mængden af information, der er i datasættet. Mere konkret er det typiske billede at 1) estimater for standardfejl blive for små og 2) derfor bliver teststørrelser bliver for store og 3) derfor bliver p-værdier for små og 4) derfor kommer effekter til at fremstå stærkere end de i virkeligheden er.

Ignorer clustering i data: En simpel regressionsmodel er

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

hvor g refererer til gruppe, i til individ indenfor gruppe, r til måling indenfor individ og $e_{gir} \sim N(0, \sigma^2)$. Denne model vil typisk være utilstrækkelig fordi man har målt på samme unit flere gange, men vi inkluderer resultatet for sammenligningens skyld. Behandlingseffekten er $\beta_{trt} - \beta_{ctrl}$ og denne omtales i det følgende som β_1 i benævnes i tabeller med `grptrt1`. Estimatet for β_1 giver dermed behandlingseffekten. Tabel 2 viser resultatet af at fitte denne model. p-værdien for behandlingseffekten bliver meget lille, hvilket indikerer stor sikkerhed af en behandlingseffekt.

Table 2: Resultatet af at analysere data når clustering i units ignoreres: p-værdien for behandlingseffekten er ganske lille, hvilket indikerer en behandlingseffekt.

	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

Standard tilgang: Analyser gennemsnit: En hyppigt anvendt tilgang er at udregne gennemsnit for hver unit og analysere disse. Mere konkret betragtes modellen

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

hvor der er beregnet gennemsnit indenfor hver unit. Denne tilgang virker fint i den forstand, at man får de rette tests når data er balancerede. Man får dog ikke noget estimat for "within-subject" variationen, hvilket dog ikke nødvendigvis er et stort problem i den konkrete sammenhæng. At analysere gennemsnitte er dog langt fra altid en mulighed. Tabel 3 indeholder resultatet for analyse af gennemsnittet. Tabellen indeholder både resultaterne for tests baseret på t-fordelingen (svarende til at der tages højde for, at variansen er estimeret fra data) og for tests baseret på normalfordelingen (svarende til at der ikke tages højde for, at variansen er estimeret fra data).

Table 3: Resultat efter analyse af gennemsnit over units.

	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

Model med tilfældige effekter: At analysere et gennemsnit er muligt i dette eksempel men langt fra altid. Et alternativ er at anvende en lineær mixed model (i dette tilfælde en varianskomponent model) hvor unit optræder som en tilfældig effekt. Dvs. vi betragter modellen

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

hvor $U_{gi} \sim N(0, \omega^2)$ og $e_{gir} \sim N(0, \sigma^2)$. Resultatet ses i Tabel 4. Bemærk at testet er baseret på χ^2 fordelingen. Det vil sige at der tages ikke højde for, at variansen er estimeret. I stedet er der en implicit antagelse om, at den estimerede varians er lig med den sande varians. Bemærk at p-værdierne er de samme som p-værdierne baseret på χ^2 approksimationen i Table 3. Dette er en konsekvens af, at data er balancerede.

Table 4: Resultat efter at fitte en mixed model med unit som tilfældig effekt.

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

3 Muligheder med `pbkrtest` pakken:

R pakken `pbkrtest` implementerer to methoder for modelsammenligninger i mixed models, hvori der tages højde for at varians- og kovariansparametre er estimerede fra data: 1) Parametrisk boostrap og 2) Kenward-Rogers approksimation (deraf navnet på pakken).

3.1 Kenward & Rogers tilgang

I kort form er tilgangen i Kenward and Roger (1997) som følger: For den multivariate normalfordeling $Y \sim N(X\beta, \Sigma)$ betragtes test af hypotesen $L(\beta - \beta_0) = 0$. Da $\hat{\beta} \sim N_d(\beta, \Phi)$ bliver en Wald test-størrelse

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

Asympototisk er $W \sim \chi^2_d$-fordelt under null hypotesen. For at beregne denne størrelse skal et estimat $\hat{\Phi}$ anvendes. Implicit i antagelsen om at W skal være asymptotisk χ^2_d fordelt er at $\hat{\Phi}$ er lig med den sande varians. En skaleret version af W er

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

I beregningen af F er $\Phi(\sigma) = (X'\Sigma(\sigma)X)^{-1} \approx \mathbb{C}ov(\hat{\beta})$, $\hat{\sigma}$ er vektor af REML estimater for elementerne i $\Sigma = \mathbb{V}ar(Y)$ og $\hat{\beta}$ er REML estimate for β.

Asymptotisk er $F \sim \frac{1}{d}\chi^2_d$ under null hypotesen, og man man kan tænke på F som grænsen af en $F_{d,m}$-fordeling når $m \to \infty$ Een måde hvorpå man kan tage højde for at $\Phi = \mathbb{V}ar(\hat{\beta})$ er estimeret fra data er ved at komme med et bedre bud på hvad nævnerfrihedsgraderne m er (bedre bud end $m = \infty$). Kenward and Roger (1997) gjorde følgende:

- Erstattede Φ med en forbedret small-sample approksimation Φ_A.
- Udledte formler for middelværdi E^* and varians V^* af F (baseret på en førsteordens Taylorudvikling).
- Skalerede F med en faktor λ og bestemte nævner frihedsgraderne m ved at match momenterne af F/λ med momenterne i en $F_{d,m}$ fordeling.

Anvendelse af Kenward-Rogers metode: Tabel 5 viser resultat efter at fitte en mixed model med unit som tilfældig effekt til de simulerede data. Den rapporterede p-værdi er for testet af ingen effekt at behandling. Testet er baseret på at approksimere teststørrelsen med en F-størrelse, hvori frihedsgraderne er estimerede udfra data. Bemærk, at p-værdien er den samme p-værdien i Tabel 3, hvor det er gennemsnittene, der analyseres.

Table 5: Resultat efter at fitte en mixed model med unit som tilfældig effekt. Den rapporterede p-værdi er for testet af ingen effekt at behandling. Testet er baseret på at approksimere teststørrelsen med en F-størrelse, hvori frihedsgraderne er estimerede udfra data.

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

3.2 Parametrisk bootstrap

Tilgangen i parametrisk bootstrap er som følger. Vi betragter to konkurrerende modeller: En stor model $f_1(y;\theta)$ og en simplere null model $f_0(y;\theta_0)$; null-modellen er en delmodel af den store model. Vi beregner en

teststørrelse t_{obs}. Så bliver p-værdien for hypotesen

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

hvor supremum er under hypotesen. Sædvanligvis kan man ikke beregne dette supremum i praksis, så i stedet beregner vi testsandsynligheden baseret på parameterestimatet, dvs.

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

I praksis approksimeres p^{PB} som følger:

1. Træk B parametrisk bootstrap datasæt $D^1, \ldots D^B$ fra den fittede null model $f_0(\cdot; \hat{\theta}_0)$.
2. Fit den store og null modellen til hvert af disse datasæts.
3. Beregn likelihood ratio (LR) teststørrelsen for hvert simuleret datasæt. Dette giver referencefordelingen.
4. Beregn hvor ekstrem den observerede teststørrelse er; dette giver p-værdien.

Resultatet er anvendelsen af metoden er vist i Table 6.

Table 6: Resultat efter at fitte en mixed model med unit som tilfældig effekt. Den rapporterede p-værdi er for testet af ingen effekt af behandling og er beregnet ved parametrisk bootstrap.

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

Figur 1 viser χ^2_1 fordelingen (kurve) lagt oven på simuleret reference fordeling. Den simulerede referencefordeling har tungere hale end den teoretiske, og dette giver den større p-værdi.

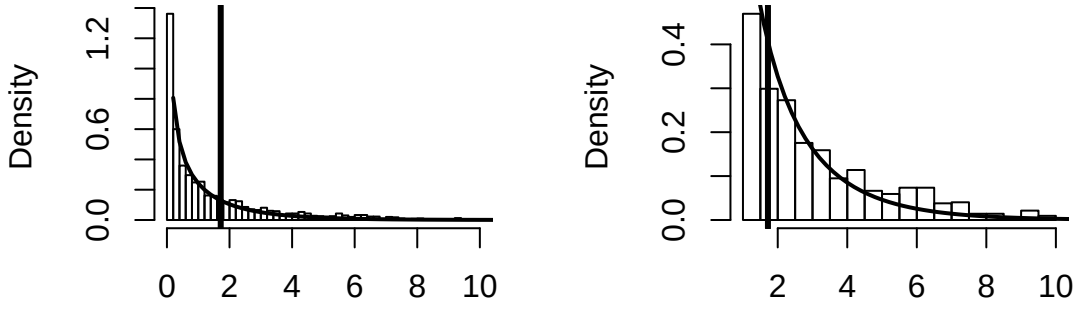

Figure 1: Tætheden for den approximerende χ^2 fordeling lagt ovenpå den referencefordeling man får ved parametrisk bootstrap, dvs. et histogram. Til venstre: Hele intervallet for teststørrelsen. Til højre: Den del af halen af fordelingen det er relevant at betragte.

Parametrisk bootstrap er en computerintensiv metode, men der er en række muligheder for at gøre beregningerne hurtigere:

1. Seventielle p-værdier: Ovenfor simulerede man et fast antal værdier $t^1, \ldots, t^B$ af teststørrelsen under hypotesen for at kunne beregne p^{PB}. Et alternativ er at kan man i stedet introducere en stop-regel,

f.eks. *Simuler indtil vi har opnået f.eks.* $h = 20$ *værdier af* t^j*, der er større end* t_{obs}. Hvis dette er opnået efter J simulationer, så skal den rapporterede p-værdi være h/J.

2. Parallelle beregninger: En anden måde at gøre beregningerne hurtigere på er ved at udnytte flere kerner på samme computer. Dette sker som default på linux og mac platforme; på windows platform skal gå igennem visse opsætningsskridt.

3. Parametrisk form af referencefordelingen: Estimation af hale-sandsynligheder kræver flere samples en at estimere middelværdi og varians af fordelingen. Derfor er det fristende at approximere en simuleret referencefordeling med en kendt fordeling så færre simulationer er nødvendige. Eksempelvis kan man matche middelværdi og varians i en gammafordeling med middelværdi og varians af den simulerede referencefordeling og derefter beregne halesandsynligheder i denne gammafordeling.

4 Simulationsstudium

Betragt igen situationen i Afsnit 2. Vi ønsker at teste hypotesen at der ikke er nogen behandlingseffekt (forskel i middelværdi). Vi gentager studiet mange gange (f.eks. 1000 gange). Da studierne er lavet ved computersimulation kan vi generere data således, at vi ved at der ikke er nogen behandlingseffekt. Hvis der ikke er nogen behandlingseffekt og hvis vi tester på signifikansniveau 5%, så skal vi i ca. 50 tilfælde få forkastet hypotesen. Den andel af testene der giver anledning til forkastelse kaldes for dækningsprocenten (eng: coverage percentage).

Hvis hypotesen forkastes f.eks. 100 gange så er dækningsprocenten 10 og det svarer til at p-værdierne er anti-konservative. Effekter forekommer at være mere signifikante end de i virkeligheden er; dvs. vi kommer til fejlagtigt at drage "for stærke" konklusioner. Tabel 7 viser resultaterne for de forskellige modeltyper.

Table 7: Dækningsprocenter for forskellige signifikansniveauer. De tre rækker, der markerede giver i praksis de korrekte dækningsprocenter og opnås når der tages højde for usikkerheden på variansparametrene.

	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

Konklusionerne er som følger:

1. Hvis man holder sig indefor den verden, der hedder lineære normale modeller så får man den rette dækningsprocent når 1) analyserer gennemsnittene og 2) baserer testene på at der tages højde for, at residualvariationen er estimeret fra data (dvs. man lave F-test i stedet for χ^2 test).

2. Hvis man betragter mixed models så er konklusionen den samme: Hvis man tager højde for at variansparametrene er estimerede fra data (og derfor laver F-test baseret på Kenward-Roger eller test baseret på parametrisk bootstrap) så får man den rette dækningsprocent. Baserer man testene på χ^2 approksimationen får man alt for store dækningsprocenter svarende til at en effekt kommer til at se mere signifikant ud end den er.

Man kan, med en hvis ret, argumetere for at de problemstillinger med tests man i de foregående afsnit har forsøgt at håndtere alle er knyttet til, at der er tale om et meget lille studium: To behandliger, tre individer per behandling og tre målinger per individ. Havde man blot haft flere individer ville problemerne forsvinde af sig selv. Men ofte er der naturlige begrænsninger på antallet af individer. Et eksempel herpå er givet i Afsnit 5.

5 Eksempel: Sukkerroer - et split plot eksperiment / hierarkisk design

Man ønsker at modellere hvordan sukkerindhold (pct) i sukkerroer afhænger af så- og høsttidspunkt. Der er fem såtidspunkter (s) og to høsttidspunkter (h). Forsøget var udlagt i tre blokke. Data findes i `pbkrtest` pakken og stammer fra et forsøg lavet ved det tidligere Danmarks JordbrugsForskning, der i dag er en del af Aarhus Universitet. I dette afsnit illustrerer vi desuden brugen af `pbkrtest` pakken.

Forsøgsplanen er som følger:

```
# 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
#       +----------------/----------------/----------------+
```

De første observationer ses i Tabel 8.

Table 8: De første observationer i 'beets' datasættet.

harvest	block	sow	yield	sugpct
harv1	block1	sow3	128	17.1
harv1	block1	sow4	118	16.9
harv1	block1	sow5	95	16.6
harv1	block1	sow2	131	17.0

Uanset om man betragter udbytte eller sukkerprocent viser et plot (ikke gengivet her) at der ikke er indikation af interaktion mellem så- og høsttidspunktet. En model for forsøget kunne derfor være

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

hvor $U_{hb} \sim N(0, \omega^2)$ og $\epsilon_{hbs} \sim N(0, \sigma^2)$. Bemærk at U_{hb} beskriver den tilfældige variation mellem plots (indenfor blokke). Med `lmer()` funktionen fra `lme4` pakken kan vi teste for ingen effekt af så- og høsttidspunkt som følger:

```
beet.lg <- lmer(sugpct ~ block + sow + harvest +
                     (1 | block:harvest), data=beets, REML=FALSE)
beet.noh <- update(beet.lg, .~. - harvest) # Fjern høsttidspunkt
beet.nos <- update(beet.lg, .~. - sow)     # Fjern såtidspunkt
anova(beet.lg, beet.noh)
```

```
## Data: beets
## Models:
## beet.noh: sugpct ~ block + sow + (1 | block:harvest)
## beet.lg: sugpct ~ block + sow + harvest + (1 | block:harvest)
##          Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## beet.noh  9 -69.1 -56.5   43.5    -87.1
## beet.lg  10 -80.0 -66.0   50.0   -100.0  12.9      1    0.00033
```

```
anova(beet.lg, beet.nos)
```

```
## Data: beets
## Models:
## beet.nos: sugpct ~ block + harvest + (1 | block:harvest)
## beet.lg: sugpct ~ block + sow + harvest + (1 | block:harvest)
##          Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## beet.nos  6  -2.8   5.6    7.4    -14.8
## beet.lg  10 -80.0 -66.0   50.0   -100.0  85.2      4     <2e-16
```

Begge effekter forekommer at være stærkt signifikante, men det interessante er her at sammenligne med resultaterne med Kenward-Roger og parametrisk bootstrap metoden. For såtidspunktet får man stadig meget små p-værdier, men for høsttidspunktet bliver billedet et andet.

```
KRmodcomp(beet.lg, beet.noh)
```

```
## F-test with Kenward-Roger approximation; time: 0.26 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
```

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

```
## Bootstrap test; time: 5.83 sec;samples: 1000; extremes: 31;
## 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    0.03197
```

Afslutningsvist bemærkes det, at da designet er balanceret kan man lave F-tests indenfor strata som vist nedenfor. Bemærk: F-teststørrelsen er $F_{1,2}$ for høsttidspunkt og $F_{4,20}$ for såtidspunkt.

```
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
```

6 Diskussion og afsluttende bemærkninger

Eksemplerne der er vist ovenfor er sådan, at man kan komme udenom problemet med korrelerede målinger ved at beregne passende gennemsnit og analysere disse. Dette er gjort for at vise, at de metoder fra `pbkrtest` der illustreres giver de "rette svar". Den virkelige styrke ligger dog i, at man kan arbejde med generelle mixed models og stadig beregne bedre referencefordelinger for teststørrelserne og dermed få mere retvisende konklusioner.

Det noteres, at der i beregningerne i Kenward-Rogers metode er brug for at udregne $G_j \Sigma^{-1} G_j$, hvor $\Sigma = \sum_i \sigma_i G_i$ og heri er σ_i'erne ukendte parametre og G_i'erne er kendte matricer. Det kan være både tids- og pladskrævende at beregne ovenstående sum. Et alternativ for lineære mixed models er en Satterthwaite-type approksimation; denne er hurtigere at beregne og er på vej i en kommende udgave af `pbkrtest`. Et alternativ (der også virker for generaliserede lineære mixed models) er at beregne p-værdier ved parametrisk bootstrap. Slutteligt skal det nævnes, 1) at `pbkrtest` er tilgængelig på https://cran.r-project.org/package=pbkrtest, 2) at `pbkrtest` er beskrevet i Halekoh and Højsgaard (2014) og 3) at udviklingsversioner af `pbkrtest` er tilgængelige på github og kan installeres med `devtools::install_github(hojsgaard/pbkrtest)`.

Referencer

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Halekoh, Ulrich, and Søren Højsgaard. 2014. "A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – the R Package pbkrtest." *Journal of Statistical Software* 59 (9): 1–30. http://www.jstatsoft.org/v59/i09/.

Kenward, Michael G., and James H. Roger. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics* 53 (3): 983–97.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.