Oversigtsforelæsning 2

Statistisk Dataanalyse 2

Anders Tolver

Uge 8, torsdag d. 2/11-2017

Dagens program

Kl. 8:15-10:30: oversigtsforelæsning 2 (kapitel 7-10 i kompendiet)

- Systematiske og tilfældige faktorer
- Modeller med tilfældige virkninger (mixed models)
- Forsøgstyper og randomisering
- Konfundering
- Gentagne målinger
- Lidt om R

10:30-12:00+13:00-16:00: regneøvelser - se flg. forslag

- Eksamenssættene 2013 og 2017
- Opgaveark 8: dog ikke 8.2.c+d+e og 8.5
- Eksamenssættet fra april 2014 (resterende opgaver)
- Eksamensættet fra april 2015 opgave 1+3

Systematiske og tilfældige faktorer

Systematiske faktorer:

- "fortolkningshorisont": kun niveauerne fra forsøget
- estimater for de konkrete niveauer (eller forskelle)
- interesseret i at teste pågældende effekt
- Eksempler: behandling, sort, tid,...

Tilfældige faktorer:

- "fortolkningshorisont": forsøgets niveauer repræsenterer en population
- estimat for variation ml. niveauer (varianskomponent)
- som regel *ikke* interessant at teste effekt
- Eks.: dyr, kuld, plante, mark, blok, klimakammer, person,...

Mixed models

REML-estimation:

```
modA=lme(y~''fixed effects'',random=~1|''random effects'')
```

Til likelihood ratio test for systematiske effekter *skal*
ML-estimation benyttes:

```
mod1=lme(y~''fixed effects 1'', random=~1|''random effects'',
         method="ML")
mod2=lme(y~''fixed effects 1'', random=~1|''random effects'',
         method="ML")
anova(mod2,mod1)
```

Estimater: `summary(modA)`

Konfidensintervaller: `intervals(modA)`

Test for en tilfældig effekt

Test for om en tilfældig effekt er signifikant kan (ofte) foretages ved at sammenligne med modellen, hvor den tilfældige effekt inddrages som systematisk snarere end tilfældig.

Eksempel:

- `T` systematisk, `B` tilfældig: $y_i = \alpha(\mathtt{T}_i) + A(\mathtt{B}_i) + e_i$.
- Hypotese: $\sigma_A^2 = 0$.
- ```
  m1=lm(y~T+B)
  m2=lm(y~T)
  anova(m2,m1)
  ```


  ```

Konkrete modeller

- En-faktor forsøg med tilfældig virkning (7.1)
- To-niveau nestede faktorforsøg (7.3)
- To-faktor forsøg med tilfældig virkning (7.2) (Repeterbarhed og reproducerbarhed)
- Split-plot forsøgstyper (kapitel 8): to niveauer af forsøgsenheder, behandlinger på begge niveauer
- Fuldstændige randomiserede blokforsøg (9.1)
- Balancerede ufuldstændige blokforsøg (BIBD, 9.2)
- 2^n-te forsøg (9.3)

... men ikke så forskellige endda: fittes og analyseres alle med `lme` som ovenfor.

Eksempel 8.1: mørhed af svinekød

Forsøgsdesign

- 24 porkers (helplot)
- porkers opdelt i to grupper efter pH (helplot faktor)
- hver porker opdelt i to sider (delplots)
- de to sider køles på hver sin måde (delplotfaktor)

Forsøget er et splitplot forsøg.

- Helplot: `P` (porkers) - tilfældig effekt
- Helplot faktor: `pH` (høj/lav) - systematisk effekt
- Delplot faktor: `C` (chilling)- systematisk effekt

Desuden inddrages $\mathtt{pH} \times \mathtt{C}$ som systematisk effekt.

$$\text{Startmodel} \rightarrow Y_i = \gamma(\mathtt{pH} \times \mathtt{C}_i) + b(P_i) + e_i, \text{ hvor}$$

- $b(1), \ldots, b(24)$ er uafhængige og normalfordelte $\sim N(0, \sigma_B^2)$
- $e_1, \ldots, e_{48}$ er uafhængige og normalfordelte $\sim N(0, \sigma^2)$

Eksempel 8.1: analyse i R (udpluk)

```
> modelA=lme(y~pH*chill,random=~1|porker,method="ML")
> modelB=lme(y~pH+chill,random=~1|porker,method="ML")
> anova(modelB,modelA)
       Model df      AIC      BIC    logLik   Test L.Ratio p-value
modelB     1  5 149.9008 159.2568 -69.95041
modelA     2  6 151.7097 162.9369 -69.85483 1 vs 2 0.19116 0.6619

> modelCa=lme(y~pH-1,random=~1|porker,method="REML")
> summary(modelCa)
Random effects:
 Formula: ~1 | porker
        (Intercept)  Residual
StdDev:    1.116364 0.6873579
> intervals(modelCa)

          lower     est.    upper
pHhigh 6.387315 7.116250 7.845185
pHlow  4.923982 5.652917 6.381852
```

På Ugeplan 8 i Absalon findes link til eksempel med en detaljeret analyse af Eksempel 8.1.

Vær f.x. opmærksom på brug af `lmer` som (bedre?) alternativ til `lme`, bortset fra i modeller med gentagne målinger/seriel korrelation (hvor kun `lme` kan bruges).

Gode råd ved opstilling af statistisk model

Ved opstilling af en statistisk model på baggrund af en forsøgsbeskrivelse bør bl.a. overvejes

- er der variable som skal/kan opfattes som faktorer men er kodet med tal-værdier (husk at 'fortælle' dette til R)?
- hvilke faktorer skal være systematiske og hvilke skal være tilfældige?
- hvilke vekselvirkninger bør indgå (ofte alle - men husk at overveje og evt. diskutere)?
- bør responsen transformeres (modelkontrol, varianshomogenitet)?

Ved opstilling af faktordiagram overvejes desuden, om nogle faktorer er finere / grovere end andre!

Randomisering i CBD

Fuldstændigt randomiseret blokforsøg

Behandling	Blok 1	Blok 2	Blok 3	Blok 4
A	x	x	x	x
B	x	x	x	x
C	x	x	x	x
D	x	x	x	x

- Inden for hver mark fordeles de 4 behandlinger ud på de 4 jordlodder (lodtrækning)

Eksamenstræning (tillægsspørgsmål):

- Hvilken type forsøg er der tale om, hvis man ved lodtrækning udvælger to blokke som gødes, mens de resterende to ikke gødes?
- Er det muligt at lave et BIBD, hvis man fastholder at benytte alle 4 behandlinger, mens der kun er 3 jordlodder på hver mark (blok)?

Randomisering i BIBD

BIBD: 2 trin i randomisering

Beh.	Blok 1	Blok 2	Blok 3	Blok 4	Blok 5	Blok 6	Blok 7
A	x	x	x				
B	x			x	x		
C	x					x	x
D		x		x		x	
E		x			x		x
F			x	x			x
G			x		x	x	

- Blok1,...,Blok7 (papir) $\rightarrow$ 7 marker (lodtrækning)
- For hver blok/mark trækkes lod om, hvordan de 3 beh. på papiret skal fordeles på de 3 jordlodder (lodtrækning)

Eksamenstræning:

- Check at betingelserne fra theorem 9.6 er opfyldt og lav co-incidence matrix

Randomisering i 2^n–forsøg

8 behandlinger (3 faktorer på hver 2 niveauer)

A	B	C	Blok 1	Blok 2	Blok 3	Blok 4	Blok 5	Blok 6
1	1	1	x		x			x
1	1	2		x	x		x	
1	2	1		x		x		x
1	2	2	x			x	x	
2	1	1		x		x	x	
2	1	2	x			x		x
2	2	1	x		x		x	
2	2	2		x	x			x

- Blok 1,...,Blok 6 (papir) $\rightarrow$ 6 marker (lodtrækning)
- For hver blok/mark trækkes lod om, hvordan de 4 beh. på papiret skal fordeles på de 4 jordlodder (lodtrækning)

Eksamenstræning

- Benyt lige/ulige-reglen (theorem 9.11) til at finde ud af, hvilke effekter der er konfunderet på blok 1+2 (hhv. 3+4, 5+6)

Randomisering i splitplot forsøg

Faktoren A kan kun variere på helplot niveau (Blok)

A	B	C	Blok 1	Blok 2	Blok 3	Blok 4	Blok 5	Blok 6
1	1	1	x		x		x	
1	1	2	x		x		x	
1	2	1	x		x		x	
1	2	2	x		x		x	
2	1	1		x		x		x
2	1	2		x		x		x
2	2	1		x		x		x
2	2	2		x		x		x

- Blok 1,...,Blok 6 (papir) $\rightarrow$ 6 marker (lodtrækning)
- For hver blok/mark trækkes lod om, hvordan de 4 beh. på papiret skal fordeles på de 4 jordlodder (lodtrækning)

Eksamenstræning:

- Hvad er delplotfaktoren?
- Opskriv et faktordiagram og en tilhørende statistisk model

Konfundering

Konfundering $\leftrightarrow$ "sammenblanding"

Konfundering af effekt (fx. $A \times B$) med blok:

$A \times B$-effekt kan ikke skelnes fra blokeffekt

2^n−forsøg delt i 2 blokke

- n faktorer, alle med to niveauer, og 2 blokke
- Lige/ulige regel for konfundering af *bestemt* effekt
- Øvrige effekter er *ukonfunderede*
- Ukonfunderede effekters estimater påvirkes herved ikke af blokkene

Eksamenstræning:

- Case 7 om bl.a. 2^3-forsøg og partiel konfundering
- Brug af `R` til at undersøge, om en given effekt er konfunderet med blok-effekten

Gentagne målinger

Flere observationer (en serie) fra samme individ/subjekt, fx. person, dyr, parcel, plante.

Giver afhængighed mellem observationerne.

Har snakket om tre analysemetoder:

- Analyse af summary measure(s)
- Random intercepts modellen
- Modeller med seriel korrelationsstruktur

Gentagne målinger (2)

Analyse af summary measure(s):

- Hver obs. serie reduceres til kun en værdi (fx. tilvækst)
- Ofte en god og robust metode!
- I skal kunne: alle aspekter!

Random intercepts modellen:

- Konstant korrelation mellem obs. fra samme individ
- I skal kunne: alle aspekter!

Modeller med seriel korrelationsstruktur (Diggle):

- Korrelationen aftager som funktion af tidsforskel mellem observationerne
- Fit med `lme` med option `corr`
- I skal kunne: fortolke Diggle modellen og resultater fra en Diggle-analyse, herunder vurdere et semi-variogram.

Eksempel 10.1: geders vægtudvikling

Faktorer:

- `goat`: 1–28
- `feed`: 1–4 (fodertyper, behandlinger)
- `tid`: 0,26,45,61,91 (dage efter forsøgets start)

NB: `tid` kan opfattes både som faktor og kovariat!!!

Random intercepts modellen:

$$Y_i = \gamma(\texttt{feed}_i, \texttt{time}_i) + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + e_i, \quad i = 1, \ldots, 112$$

hvor $A(1), \ldots, A(28) \sim N(0, \nu^2)$, $e_1, \ldots, e_{112} \sim N(0, \sigma^2)$, alle uafh.

Eksamenstræning:

- Reducer den systematiske del af modellen (-rækkefølge?)
- Estimat+konf.int for forskel mellem behandling 3 og 2 efter 45 dage.
- Estimat+konf.int for den forventede vægt efter 45 dage for en ged med beg.vægt $w0 = 13$ kg som gives behandling 3

Eksempel 10.1: geders vægtudvikling

Diggle-modellen:

$$Y_i = \gamma(\texttt{feed}_i, \texttt{time}_i) + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + D_i + e_i$$

hvor A'er, D'er og e'er er således at variansstrukturen er som angivet i kompendiets kapitel 10.3.

```
library(nlme)
mod2<-lme(weight~w0+feedfac+feedfac:day,random=~1|goat,
corr = corGaus(form = ~ day | goat, nugget=T),method="ML")
```

Eksamenstræning:

- Tegn semi-variogrammet og forklar, hvorfor Diggle-modellen ser ud til at beskrive korrelationsstrukturen bedre end random intercepts modellen
- Reducer den systematiske del af modellen mest muligt
- Angiv et konfidensinterval for δ
- Angiv parameterestimaterne for alle 4 parametre, som indgår i korrelationsstrukturen

Generelle R-ting

Fit af modeller:

- Lineære modeller (uden tilfældige effekter): `lm`
- Mixed models (med tilfældige effekter): `lme`

Estimater og konfidensintervaller:

- Estimater: `summary` (-husk at fitte med `method='REML'`)
- Konfidensintervaller: `confint` for `lm`, `intervals` for `lme`
- Kontraster: `estimable`.
  Vigtigt at linearkombination og parametrisering af modellen passer sammen: lav `summary` og beregn estimat "i hånden".

Test:

- Test simpel mod2 mod mere generel mod1:
  `anova(mod2,mod1)`
- Husk at fitte `lme`-modeller med `method='ML'` når der testes

Specifikation af systematisk del

`T,K` faktorer, `x` kovariat.

Modeller kun med faktorer:

- med vekselvirkning: `T*K` eller `T+K+T:K` eller `T:K`
- uden vekselvirkning: `T+K`

Modeller med faktor og kovariat:

- forskellig skæring, forskellig hældning: `T*x`, `T+x+T:x` eller `T+T:x`.
- forskellig skæring, ens hældning: `T+x`
- ens skæring, forskellig hældning: `T:x`

Bemærk, at for modeller med tilfældige effekter gælder samme regler for rækkefølgen af reduktion i den systematiske del.

Fx. kan modelhjulet benyttes til at overskue de forskellige modeller i et forsøg med en faktor `F` og en forklarende variabel `H`, der kan opfattes både som faktor og kovariat.