Tilfældige effekter

Statistisk Dataanalyse 2

Anders Tolver

Uge 4, tirsdag d. 26/9-2017

Dagens program

Tilfældige effekter: forståelse og fortolkning

- Hvorfor tilfældige effekter?
- Hvilke faktorer bør have tilfældig virkning?
- Split-plot forsøg (tænkt eksempel)

Tilfældige effekter: teknik (især omkring R)

- Fit af modeller med tilfældige effekter
- Test for reduktion i modellerne
  - approksimative likelihood ratio test
  - simulation af eksakt p-værdi med `simulate.lme`
  - F-test (torsdag)
- Estimater samt konfidensintervaller
- Eksempel 8.1 i kompendiet

Dyrkningsforsøg

Ved et dyrkningsforsøg er målt udbyttet på 4 jordlodder på hver af 4 forskellige marker af samme størrelse.

```
##   M        y
## 1 1 31.76821
## 2 1 31.01354
## 3 1 28.21829
## 4 1 22.57111
```

```
[... more datalines here ...]
```

```
##    M        y
## 15 4 54.09678
## 16 4 61.25153
```

Vi ønsker at beskrive variationen i udbyttet på en jordlod af samme størrelse, hvis vi gentager forsøget næste år.

Dårlig løsning: model med intercept

Model

$$Y_i = \mu + e_i, \quad i = 1, \ldots, 16,$$

hvor $e_1, \ldots, e_{16}$ er uafhængige $\sim N(0, \sigma^2)$.

Estimater

$$\hat{\mu} = 66.09[51.96, 80.22]; \quad \hat{\sigma}^2 = 26.52^2 = 703.31$$

Problemer

- Den skitserede løsning giver meget stor residualvarians
- Observationer fra samme `mark` må forventes at ligne hinanden mere end observationer fra forskellige marker
- Vi bør inddrage faktoren `mark` i analysen

Systematisk effekt af mark

Model

$$Y_i = \alpha(\texttt{mark}_i) + e_i, \quad i = 1, \ldots, 16,$$

hvor $e_1, \ldots, e_{16}$ er uafhængige $\sim N(0, \sigma^2)$.

Estimater

$$\hat{\alpha}(1) = 28.39[22.57, 34.21] \quad ; \quad \hat{\alpha}(2) = 91.89[86.07, 97.71]$$
$$\hat{\alpha}(3) = 86.07[80.25, 91.90] \quad ; \quad \hat{\alpha}(4) = 58.00[52.18, 63.82]$$
$$\hat{\sigma}^2 = 5.343^2 = 28.55$$

- Tager højde for inhomogenitet mellem marker
- Giver fornuftigt bud på forventede udbytte næste år

I hvert tilfælde hvis forsøget udføres på en af de fire marker fra det oprindelige forsøg!

Tilfældig effekt af mark

Model

$$Y_i = \mu + b(\texttt{mark}_i) + e_i, \quad i = 1, \ldots, 16,$$

hvor

- μ er det fælles middelniveau for udbyttet på alle marker
- $b(1), \ldots, b(4)$ er uafhængige $\sim N(0, \sigma_B^2)$ og beskriver den tilfældige variation mellem marker
- $e_1, \ldots, e_{16}$ er uafhængige $\sim N(0, \sigma^2)$ og beskriver residualvariationen mellem forsøgsenheder

Bemærk: Vi får ikke et estimat for de enkelte marker i forsøget, blot en parameter (σ_B^2) til beskrivelse af variationen mellem marker.

Giver os mulighed for at komme med estimat og konfidensinterval for udbyttet på en mark som *ikke nødvendigvis er med i forsøget*!

Ensidet ANOVA ved tilfældig variation

```
> rmod0 <- lme(y~1,random=~1|field,method='REML')
> summary(rmod0)
Random effects:
 Formula: ~1 | field
        (Intercept) Residual
StdDev:    29.04071 5.342793

Fixed effects: y ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 66.08987  14.58166 12 4.532396   7e-04
```

$$\text{Estimater: } \hat{\mu} = 66.09[34.32, 97.86]; \quad \hat{\sigma}_B = 29.04; \quad \hat{\sigma} = 5.34$$

Forskellen mellem ensidet variansanalyse med systematisk og tilfældig variation ligger i fortolkningen.

Systematisk variation: Konklusionerne gælder kun for markerne i forsøget

Tilfældig variation: Konklusioner kan drages på populationsniveau

Inddragelse af hvedesort

Ved dyrkningsforsøget anvendtes to forskellige hvedesorter hver på 8 forsøgsenheder (svarende til to marker) som anført i parentes nedenfor

`mark 1`	`mark 2`	`mark 3`	`mark 4`
31.768 (1)	81.976 (2)	81.722 (2)	58.115 (1)
31.014 (1)	90.324 (2)	89.374 (2)	58.546 (1)
28.218 (1)	96.744 (2)	92.331 (2)	54.097 (1)
22.571 (1)	98.515 (2)	80.872 (2)	61.252 (1)

Faktorer: `mark`, `sort`

Bemærk: `mark` er finere end `sort`

Er der forskel på udbyttet af de to sorter?

Lad os starte med at besvare spørgsmålet med metoderne fra undervisningsuge 1-3.

Mark som systematisk faktor

Model1:

$$\begin{aligned} Y_i &= \alpha(\texttt{sort}_i) + \beta(\texttt{mark}_i) + e_i, \quad i = 1, \ldots, 16, \\ &= \beta(\texttt{mark}_i) + e_i \quad \leftarrow \text{hvorfor samme model?} \end{aligned}$$

Model2: $Y_i = \alpha(\texttt{sort}_i) + e_i, \quad i = 1, \ldots, 16,$

Model3: $Y_i = \mu + e_i, \quad i = 1, \ldots, 16,$

Test 2 vs. 1: Teststr. $F = 31.9 \sim F(2, 12)$; p-value=0%

Test 3 vs. 2: Kan ikke teste for effekt af `sort`, da vi ikke kan fjerne effekt af finere faktor `mark`!
Betyder ikke nødvendigvis, at der ikke er forskel på de to sorter i forsøget, idet blokeffekten kan *skygge* for effekten af `sort`.

Vi er interesserede i at teste om der er forskelle på sorterne, så dette er en utilfredsstillende statistisk analyse!

Mark som tilfældig faktor

Den følgende model kombinerer systematiske og tilfældige effekter

$$Y_i = \alpha(\texttt{sort}_i) + b(\texttt{mark}_i) + e_i, \quad i = 1, \ldots, 16,$$

hvor

- $b(1), \ldots, b(4)$ er uafhængige $N(0, \sigma_B^2)$
- $e_1, \ldots, e_{16}$ er uafhængige $N(0, \sigma^2)$

Nu er det muligt at teste for effekt af `sort` (p-value=0.9 %)

Estimater

$$\hat{\alpha}(\texttt{sort1}) = 43.20[-2.71, 89.10] \qquad \hat{\sigma}^2 = 5.34^2 = 28.55$$
$$\hat{\alpha}(\texttt{sort2}) = 88.98[43.08, 134.9] \qquad \hat{\sigma}_B^2 = 10.33^2 = 106.68$$

Lidt mere om test i R

Det approksimative likelihood ratio test

```
> rmodel1 <- lme(y~crop-1,random=~1|field,method='ML')
> rmodel2 <- lme(y~1,random=~1|field,method='ML')
> anova(rmodel2,rmodel1) # Sign. effect of crop (LR=6.894,p=0.009)
        Model df      AIC      BIC    logLik   Test L.Ratio p-value
rmodel2     1  3 123.0017 125.3195 -58.50087
rmodel1     2  4 118.1076 121.1979 -55.05379 1 vs 2 6.8942  0.009
```

Det anbefales at simulere p-værdien, hvis likelihood ratio testet giver en værdi omkring signifikansniveauet (-her 5 %)

```
> sim<-simulate.lme(rmodel2,m2=rmodel1,nsim=1000)
> lr.sim<-2*(sim$alt$ML-sim$null$ML)
> psim<-sum(lr.sim>6.894)/1000
> psim
[1] 0.09
```

F-test: mere om dette på torsdag!

Split-plot forsøg: kompendiets kapitel 8.1

Ud over faktoren `sort`, som er anvendt på `mark`-niveau, er de fire jordlodder på hver mark blevet gødet med en af fire gødningstyper (A/C/K/N)

`mark 1`	`mark 2`	`mark 3`	`mark 4`
31.768(N)	81.976(C)	81.722(N)	58.115(N)
31.014(K)	90.324(K)	89.374(C)	58.546(A)
28.218(C)	96.744(N)	92.331(K)	54.097(K)
22.571(A)	98.515(A)	80.872(A)	61.252(C)

Faktoren `sort` varierer kun mellem marker, men ikke inden for jordlodder på samme mark.

Helplot: `mark`

Helplotfaktor: `sort`

Delplotfaktor: `goedning` (-varierer på jordlodder inden for `mark`)

Overblik over dyrkningsforsøg

Vi har således reelt følgende faktorer

`goedning` : A,C,N,K
`sort`: sort1,sort2
`mark`: 1,2,3,4

Vi husker at `sort` er grovere end `mark`.

Konklusioner skal kunne generaliseres til andre marker, og vi vil teste for effekt af `sort`, hvorfor `mark` bør inddrages i modellen som tilfældig faktor.

Systematiske faktorer:

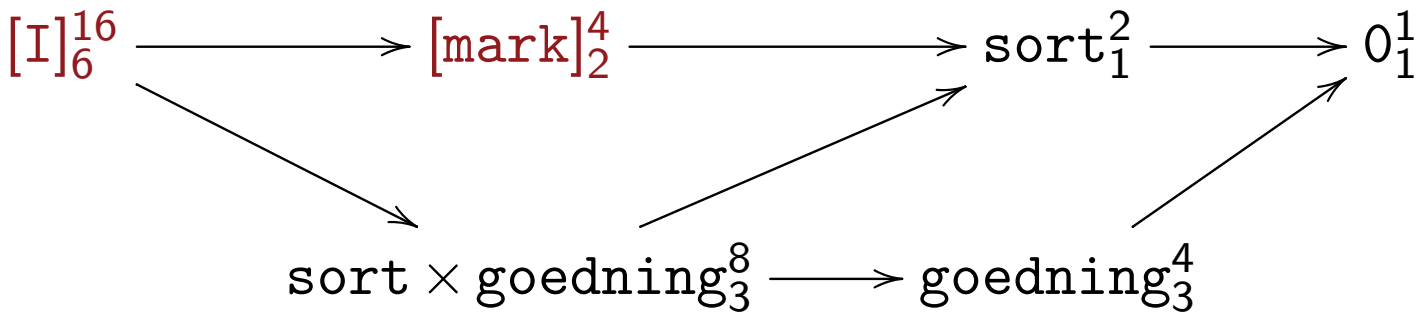
$$\texttt{goedning} \times \texttt{sort}, \texttt{goedning}, \texttt{sort}, 0$$

Fuldt dyrkningsforsøg

Udgangsmodel

$$Y_i = \gamma(\texttt{sort} \times \texttt{goedning}_i) + b(\texttt{mark}_i) + e_i, \quad i = 1, \ldots, 16,$$

- $b(1), \ldots, b(4)$ er uafhængige $N(0, \sigma_B^2)$
- $e_1, \ldots, e_{16}$ er uafhængige $N(0, \sigma^2)$



[] omkring tilfældige faktorer i faktordiagram

Dyrkningsforsøg - konklusion

Reduktion

Hverken $\texttt{sort} \times \texttt{goedning}$ ($LR = 1.802, p = 0.615$) eller $\texttt{goedning}$ ($LR = 0.491, p = 0.921$) har signifikant effekt på udbyttet.

Slutmodel

$$Y_i = \alpha(\texttt{sort}_i) + b(\texttt{mark}_i) + e_i, \quad i = 1, \ldots, 16$$

Parameterestimater

$$\hat{\alpha}(\texttt{sort1}) = 43.20 \quad [-2.71, 89.10] \qquad \hat{\sigma}^2 = 5.34^2 = 28.55$$
$$\hat{\alpha}(\texttt{sort2}) = 88.98 \quad [43.08, 134.9] \qquad \hat{\sigma}_B^2 = 10.33^2 = 106.68$$

Giver os mulighed for at komme med 95 %-konf. interval for det forventede udbytte, hvis vi gentager eksperimentet på en anden (tilfældigt valgt) mark.

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)

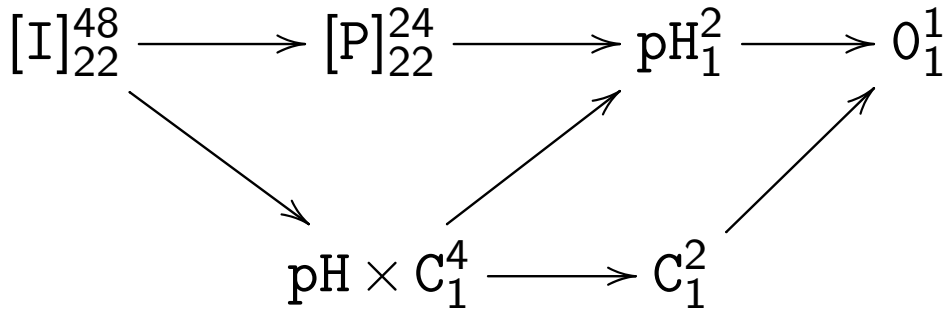
Forsøget er et splitplot forsøg.

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

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

Eksempel 8.1: faktordiagram

Faktordiagram



Statistiske modeller

- A: $Y_i = \gamma(\mathtt{pH} \times \mathtt{C}_i) + b(P_i) + e_i$
- B: $Y_i = \alpha(\mathtt{pH}_i) + \beta(\mathtt{C}_i) + b(P_i) + e_i$
- C: $Y_i = \alpha(\mathtt{pH}_i) + b(P_i) + e_i$
- D: $Y_i = \mu + b(P_i) + e_i$

Eksempel 8.1: reduktion

Analyse i R vha. lme og anova

```
> 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 0.1911  0.662

...
more tests
...
```

Slutmodellen bliver

$$Y_i = \alpha(\mathrm{pH}_i) + b(\mathrm{P}_i) + e_i$$

Eksempel 8.1: konklusion

Der er en effekt af pH men ikke af nedkølingsmetoden.

Estimater og konf.-intervaller (vha. summary, intervals og evt. VarCorr):

$$\begin{aligned} \alpha(\text{low}) &: 5.65[4.92, 6.38] \\ \alpha(\text{high}) &: 7.12[6.39, 7.85] \end{aligned}$$

Varianskomponenter:

$$\hat{\sigma}_P^2 = 1.25; \quad \hat{\sigma}^2 = 0.47$$

Obs: Brug estimater fra model fitted med `method="REML"`

```
> modelCa=lme(y~pH-1,random=~1|porker,method="REML")
> intervals(modelCa)
Approximate 95% confidence intervals

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

Eksempel 8.1: konklusion

Vha. estimable kan vi estimere forskellen mellem de to pH-grupper.

Bemærk først, at $\alpha(\text{high}) - \alpha(\text{low}) = 1 \cdot \alpha(\text{high}) + (-1) \cdot \alpha(\text{low})$

Kør dernæst flg. R-kode

```
>modelCa=lme(y~pH-1,random=~1|porker,method="REML")
>library(gmodels)
>pHdiff=c(1,-1)
>estimable(modelCa,pHdiff,conf.int=0.95)
        Esti    Std.Error t value DF  Pr(>|t|)  Low.CI Upper.CI
(1 -1) 1.4633   0.49707    2.944  22 0.0075078  0.43246 2.4942
```

Forskellen estimateres til

$$\alpha(\text{high}) - \alpha(\text{low}) : 1.46 \quad [0.43, 2.49]$$

Modeller med tilfældige effekter

Mixed models

tilfældige og systematiske effekter

Faktorer med tilfældig virkning (random effect)

- Eksempler: dyr, kuld, mark, besætning, parti
- Bestemmer variansstrukturen i modellen, dvs. varianser og korrelationer (afhængighed)
- Udtaget tilfældigt: konklusion gælder for hele populationen

R-stuff

lme,summary,anova,intervals,library(gmodels),estimable

Hovedeksempel

Splitplot forsøg

Hvad skal vi lave på torsdag?

- Modeller med flere tilfældige effekter
- Modeller med flere 'nestede' tilfældige effekter
- Test af tilfældige effekter
- Eksakte test (mere præcise) for balancerede forsøg
- Vildere eksempler