Tilfældige effekter

Statistisk Dataanalyse 2

Anders Tolver

Uge 4, torsdag d. 28/9-2017

Dagens program

- Modeller med flere tilfældige effekter

  nestede: brug `lme` fra `nlme`-pakken

  ikke-nestede: brug `lmer` fra `lme4`-pakken
- Eksempel 7.2: vitamin E i kødstykker
- F-test i modeller med tilfældige effekter
- Modelkontrol i modeller med tilfældige effekter
- Eksempler 7.5: chokolade

Eksempel 7.2: vitamin E i kødstykker

Ved et forsøg ønsker man at bestemme konc. af vitamin E i kød.

5 stykker kød deles hver i 10 mindre tern. Fem forskellige laboratorier modtager 2 tern fra hver af de 5 kødstykker, hvorefter koncentrationen af vitamin E bestemmes eksperimentielt.

I forsøget indgår faktorerne `meat` og `lab`.

$$
\begin{tikzcd}
 & & \mathtt{lab}^{5}_{4} \arrow[dr] & \\
\mathtt{I}^{50}_{25} \arrow[r] & \mathtt{lab}\times\mathtt{meat}^{25}_{16} \arrow[ur] \arrow[dr] & & \mathtt{0}^{1}_{1} \\
 & & \mathtt{meat}^{5}_{4} \arrow[ur] &
\end{tikzcd}
$$

Det viser sig, at responsvariablen `vita` bør transformeres – f.eks. med kvadratroden – før den statistiske analyse.

[NB: mere om modelkontrol på slide 18-20]

Eksempel 7.2: vitamin E i kødstykker

```
ex72 <- read.table(file = "../data/Ex72.txt", header = T)
ex72
```

```
##    meat lab vita
## 1     1   1  2.3
## 2     1   1  1.9
## 3     1   2  1.5
## 4     1   2  1.1
## 5     1   3  1.1
## 6     1   3  1.1
## 7     1   4  1.4
## 8     1   4  1.4
## 9     1   5  1.2
## 10    1   5  1.1
## 11    2   1  9.6
```

```
[ ... more data lines here ... ]
```

Eksempel 7.2: `lab` som tilfældig effekt

Vi ønsker at beskrive variationen ml. laboratorier som en tilfældig effekt og tager udgangspunkt i modellen

$$Y_i = \alpha(\texttt{meat}_i) + b(\texttt{lab}_i) + e_i$$

R-program til fit af model og test for effekt af `meat`

```
ex72$meat <- factor(ex72$meat)
ex72$lab <- factor(ex72$lab)
ex72$svita <- sqrt(ex72$vita)
```

```
library(nlme)
mod1 <- lme(svita ~ meat, random = ~1 | lab, method = "ML",ex72)
mod2 <- lme(svita ~ 1, random = ~1 | lab, method = "ML",ex72)
anova(mod2,mod1)
```

```
##      Model df       AIC       BIC    logLik   Test  L.Ratio p-value
## mod2     1  3  95.47071 101.20678 -44.73535
## mod1     2  7 -30.57040 -17.18624  22.28520 1 vs 2 134.0411  <.0001
```

Bemærk: Da vi har gentagelser for hvert komb. af `lab` $\times$ `meat` bør denne inddrages ved den statistiske analyse.

Da `lab` er tilf. skal produktfaktoren også indgå som tilfældig effekt!

Eksempel 7.2: to nestede tilfældige effekter

Ønsker at fitte modellen

$$Y_i = \alpha(\texttt{meat}_i) + \underbrace{b(\texttt{lab}_i) + c(\texttt{lab} \times \texttt{meat}_i) + e_i}_{\text{tilfældig del}}$$

Hvad mangler i opskrivningen af modellen?

Bemærk: `lab` $\times$ `meat` er finere end (Eng: nested within) `lab`

```
ex72$meatlab <- ex72$meat:ex72$lab
```

```
model1 <- lme(svita ~ meat, random = ~1 | lab/meatlab, method = "ML",ex72)
model2 <- lme(svita ~ 1, random = ~1 | lab/meatlab, method = "ML",ex72)
anova(model2, model1)
```

```
##        Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## model2     1  4   6.47241  14.12050  0.76379
## model1     2  8 -48.44624 -33.15005 32.22312 1 vs 2 62.91865  <.0001
```

R-programmet fitter model med to 'nestede' tilfældige effekter.

Hvad er effekten af første linje i programmet?

Eksempel 7.2: mange tilf. effekter - brug lmer!

Man kan argumentere for, at `meat` også bør indgå med tilfældig effekt svarende til modellen

$$Y_i = \mu + a(\mathtt{meat}_i) + b(\mathtt{lab}_i) + c(\mathtt{lab} \times \mathtt{meat}_i) + e_i$$

De tre tilfældige faktorer er her ikke 'nestede'. Hvorfor?

I denne situation fittes modeller lettest med `lmer` fra `lme4`-pakken.

```
library(lme4)
r0 <- lmer(svita ~ 1 + (1 | meatlab) + (1 | meat) + (1 | lab)
           , data=ex72)
```

Mange modeller med tilfældige effekter kan fittes både med `lme` og `lmer`, så det er nyttigt at gøre sig erfaringer med begge.

Eksempel 7.2: F-test i blandede modeller

Se på modellen svarende til faktordiagrammet

$$\begin{tikzcd}
{[\mathtt{I}]}^{50}_{25} \arrow[r] & {[\mathtt{lab} \times \mathtt{meat}]}^{25}_{16} \arrow[r] \arrow[dr] & {[\mathtt{lab}]}^{5}_{4} \arrow[r] & 0^{1}_{1} \\
 & & \mathtt{meat}^{5}_{4} \arrow[ur] &
\end{tikzcd}$$

$$\text{Model}: \quad Y_i = \alpha(\mathtt{meat}_i) + b(\mathtt{lab}_i) + c(\mathtt{lab} \times \mathtt{meat}_i) + e_i$$

Generel regel: Effekten af en faktor, F, testes mod den groveste tilfældige faktor, G, som er finere end F.

$$F = \frac{MS_e^F}{MS_e^G}$$

På de næste tre sider vises, hvordan man ved brug af R kan lave F-teststørrelser i forbindelse med test af effekterne $\mathtt{lab} \times \mathtt{meat}$ (tilfældig), `meat` (systematisk) og `lab` (tilfældig).

Eksempel 7.2: F-test af tilfældige komponenter

Test for effekt af $\texttt{lab} \times \texttt{meat}$ (tilfældig) svarer til hypotesen $H_0 : \sigma^2_{\texttt{lab}\times\texttt{meat}} = 0$.

Iflg. reglen skal $\texttt{lab} \times \texttt{meat}$ testes mod den identiske faktor I.

Når man tester op mod den identiske faktor konstrueres F-testet med lm som ved test i modeller med systematiske effekter.

```
mod0 = lm(svita ~ meat * lab + meat + lab, data = ex72)
mod1 = lm(svita ~ meat + lab, data = ex72)
anova(mod1, mod0)
```

```
##   Res.Df       RSS Df Sum.of.Sq     F        Pr..F.
## 1     41 0.7687543 NA        NA    NA            NA
## 2     25 0.1147203 16  0.654034 8.908 1.046544e-06
```

Signifikant effekt af $\texttt{lab}\times\texttt{meat}$.

Eksempel 7.2: F-test for effekt af meat (kursorisk)

Reglen medfører, at meat skal testes imod lab×meat.

Beregning af $MS_e^{\texttt{lab}\times\texttt{meat}}$

```
m1 <- lm(svita ~ meat:lab, data = ex72)
m2 <- lm(svita ~ meat + lab, data = ex72)
(deviance(m2) - deviance(m1))/(m2$df - m1$df)

## [1] 0.04087713
```

Beregning af $MS_e^{\texttt{meat}}$

```
m3b <- lm(svita ~ lab, data = ex72)
(deviance(m3b) - deviance(m2))/(m3b$df - m2$df)

## [1] 3.586636
```

F-teststørrelse

$$F_{\texttt{meat}} = \frac{MS_e^{\texttt{meat}}}{MS_e^{\texttt{lab}\times\texttt{meat}}} = \frac{3.5866}{0.0409} = 87.89 \sim F(4,16)$$

Eksempel 7.2: F-test for effekt af `lab` (kursorisk)

Reglen medfører, at `lab` skal testes imod `lab`×`meat`.

Beregning af $MS_e^{\mathtt{lab}\times\mathtt{meat}}$

```
m1 <- lm(svita ~ meat:lab, data = ex72)
m2 <- lm(svita ~ meat + lab, data = ex72)
(deviance(m2) - deviance(m1))/(m2$df - m1$df)

## [1] 0.04087713
```

Beregning af $MS_e^{\mathtt{lab}}$

```
m3a <- lm(svita ~ meat, data = ex72)
(deviance(m3a) - deviance(m2))/(m3a$df - m2$df)

## [1] 0.6421781
```

F-teststørrelse

$$F_{\mathtt{meat}} = \frac{MS_e^{\mathtt{lab}}}{MS_e^{\mathtt{lab}\times\mathtt{meat}}} = \frac{0.6422}{0.0409} = 15.70 \sim F(4, 16)$$

Om test i modeller med tilfældige effekter

Test af systematiske effekter

- Fit modeller med lme/lmer og lav $LR-$test med anova. Metoden er en approksimation, som virker bedst, når der er mange observationer. Hvis den approksimative p-værdi er tæt på 5 % anbefales det, at man bruger `simulate.lme` til at simulere den eksakte p-værdi. (pensum)
- Fit modeller med lm og konstruer F-test manuelt eller i simple tilfælde vha. anova. Metoden er eksakt ved pæne forsøgsdesign (f.eks. balancerede). (kursorisk pensum)

Test af tilfældige effekter

- Fit modeller med lm og konstruer F-test manuelt eller i simple tilfælde vha. `anova`. Metoden er eksakt ved pæne forsøgsdesign (f.eks. balancerede). (kursorisk pensum)

Om F-test i modeller med tilfældige effekter

- F-test af en faktor mod den identiske faktor `I` fås umiddelbart i `R`, hvis modellerne fittes med `lm` og der testes med `anova`. (pensum)

Eksempel 7.2: modelkontrol med tilfældige effekter

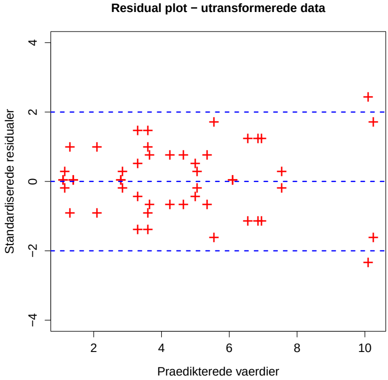
Man plejer at udføre modelkontrol på følgende måde

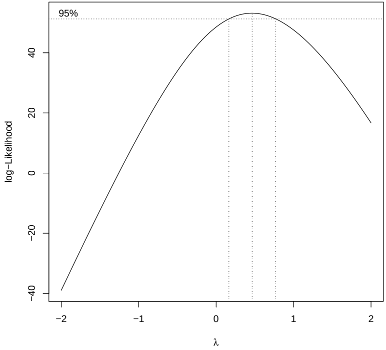
- Fit den tilsvarende model, hvor alle forklarende variable indgår med systematisk effekt (tilfældige faktorer $\rightarrow$ systematiske faktorer)
- Lav residualplot og evt. qq-plot for denne model
- Brug evt. Boxcox-plot til at bestemme transformation

Nogle nyttige kommandoer i R til modelkontrol i forbindelse med eksempel 7.2 kunne være:

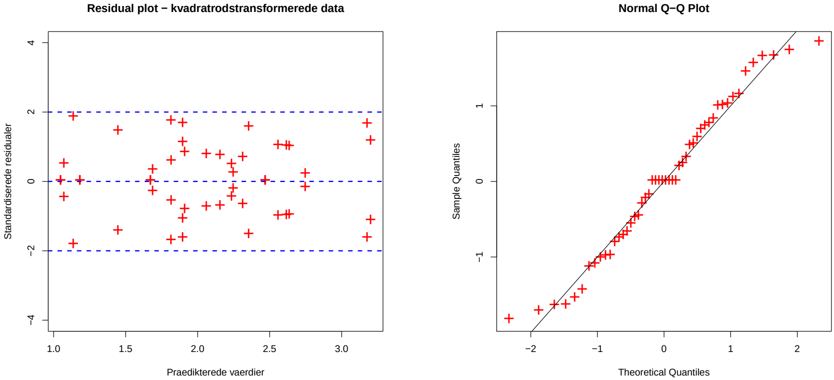
```
model0 <- lm(vita ~ meat + lab + meat:lab, data = ex72)
plot(predict(model0), rstandard(model0))
qqnorm(rstandard(model0))
library(MASS)
boxcox(model0)
```

Eksempel 7.2: modelkontrol med tilfældige effekter





Eksempel 7.2: modelkontrol med tilfældige effekter



Eksempel 7.5: chokolade

Data

- se tabel 7.4 i kompendiet

Respons

`score` i intervallet 0–15 (ikke sød – meget sød)

Faktorer

- `assessor, A: 1,2,3,4,5,6,7,8`
- `session, S: 1,2,3,4,`
- `product, P: 1,2,3,4,5`

En observation per kombination af `A, S` og `P`.

Hvilke faktorer bør være systematiske og hvilke bør være tilfældige?

Eksempel 7.5: statistisk model

Statistisk model og faktordiagram

$$
\begin{aligned}
Y_i &= \alpha(\mathtt{P}_i) + b(\mathtt{A}_i) + c(\mathtt{S}_i) \\
&+ d(\mathtt{P}\times\mathtt{A}_i) + e(\mathtt{P}\times\mathtt{S}_i) + f(\mathtt{A}\times\mathtt{S}_i) + e_i,
\end{aligned}
$$

$$
\begin{tikzcd}
 & {[\mathtt{A}\times\mathtt{P}]^{40}_{28}} \arrow[r] \arrow[dr] & {[\mathtt{A}]^{8}_{7}} \arrow[dr] & \\
{[\mathtt{I}]^{160}_{84}} \arrow[ur] \arrow[r] \arrow[dr] & {[\mathtt{A}\times\mathtt{S}]^{32}_{21}} \arrow[ur] \arrow[dr] & \mathtt{P}^{5}_{4} \arrow[r] & \mathtt{0}^{1}_{1} \\
 & {[\mathtt{P}\times\mathtt{S}]^{20}_{12}} \arrow[ur] \arrow[r] & {[\mathtt{S}]^{4}_{3}} \arrow[ur] &
\end{tikzcd}
$$

- Systematiske faktorer: $\mathtt{P}$
- Tilfældige faktorer: $\mathtt{A}$, $\mathtt{S}$, $\mathtt{A}\times\mathtt{S}$, $\mathtt{A}\times\mathtt{P}$, $\mathtt{P}\times\mathtt{S}$

Husk: Hvis en faktor er finere end en tilf. faktor (fx. en vekselvirkning), så skal den fine faktor også være tilf.

Eksempel 7.5: modelkontrol med tilfældige effekter

Man plejer at udføre modelkontrol på følgende måde

- Fit den tilsvarende model, hvor alle forklarende variable indgår med systematisk effekt (tilfældige faktorer → systematiske faktorer)
- Lav residualplot og evt. qq-plot for denne model

```
model0<-lm(score~P+A+S+P:A+P:S+A:S,data=ex75)
pred0<-predict(model0)
sres0<-rstandard(model0)
plot(pred0,sres0)
qqnorm(sres0)
```

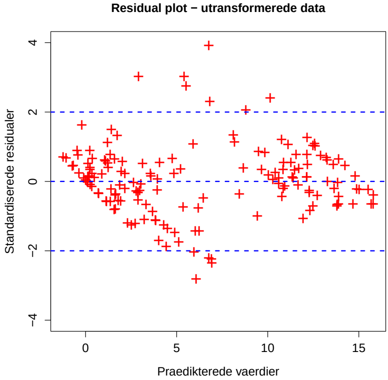
Ser ikke rigtigt godt ud...

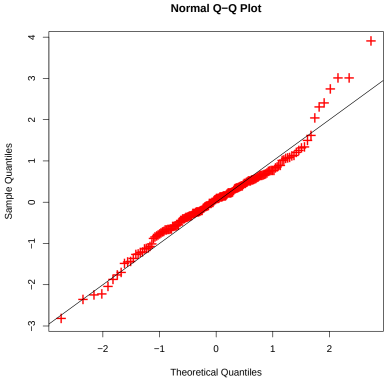
Brug istedet følgende respons (-se evt. eksempel 5.6 i kompendium)

$$Y_i = arcsin(\sqrt{\texttt{score}_i/15})$$

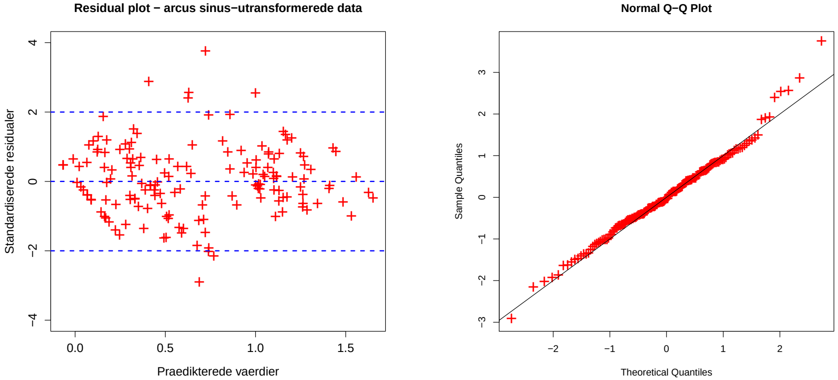
Ser lidt bedre ud...

Eksempel 7.5: modelkontrol med tilfældige effekter





Eksempel 7.5: modelkontrol med tilfældige effekter



Eksempel 7.5: reduktion af tilfældig del

Formelle test for varianskomponenterne

$$
\begin{aligned}
F_{AP} &= \frac{MS_{AP}}{MS_e} = \frac{0.3358}{0.0455} = 7.38 \sim F(28,84), p = 0 \\
F_{AS} &= \frac{MS_{AS}}{MS_e} = \frac{0.0896}{0.0455} = 1.97 \sim F(21,84), p = 0.016 \\
F_{PS} &= \frac{MS_{PS}}{MS_e} = \frac{0.0383}{0.0455} = 0.84 \sim F(12,84), p = 0.61 \\
F_{S} &= \frac{MS_{S}}{MS_{AS}} = \frac{0.0362}{0.0896} = 0.40 \sim F(3,21), p = 0.75
\end{aligned}
$$

så $\mathtt{P} \times \mathtt{S}$ og $\mathtt{S}$ kan undværes.

Eksempel 7.5: analyse og konklusion

Efter reduktion af den tilfældige del er vores model

$$B : Y_i = \alpha(\mathtt{P}_i) + b(\mathtt{A}_i) + d(\mathtt{P} \times \mathtt{A}_i) + f(\mathtt{A} \times \mathtt{P}_i) + e_i$$

Vi ønsker at undersøge, om der er en produkteffekt.

- Hvad er den relevante hypotese?
- Hvad er likelihood ratio teststørrelsen?
  - Hvilken fordeling har den?
  - Hvad bliver p-værdien?
- Hvad er F-teststørrelsen?
  - Hvilken fordeling har den?
  - Hvad bliver p-værdien?
  - Sammenlign med LR-teststørrelsen
- Hvad bliver konklusionen på analysen?
- Angiv parameterestimater. Hvor mange skal der være?