

Oversigtsforelæsning 1

Statistisk Dataanalyse 2

Anders Tolver

Uge 8, tirsdag d. 31/10-2017



Afskårne roser: firfaktorforsøg

| ## | Obs | gartner | handler | kunde | flor | middel | art | tid |
|------|-----|---------|---------|-------|------|--------|-----|------|
| ## 1 | 1 | 0 | 0 | 0 | 1 | A | GAR | 10.1 |
| ## 2 | 2 | 0 | 0 | 0 | 2 | A | GAR | 8.9 |
| ## 3 | 3 | 0 | 0 | 0 | 3 | A | GAR | 11.4 |
| ## 4 | 4 | 1 | 0 | 0 | 1 | A | GAR | 10.9 |
| ## 5 | 5 | 1 | 0 | 0 | 2 | A | GAR | 6.9 |
| ## 6 | 6 | 1 | 0 | 0 | 3 | A | GAR | 11.2 |

Vi interesserer os for følgende 4 faktorer

```
G=factor(gartner)
K=factor(kunde)
H=factor(handler)
F=factor(flør)
```

Vi er kun sekundært interesseret i faktoren FLOR.

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 3/20



Program

Ved dagens forelæsning repeteres følgende begreber:

- faktorer og faktordiagrammer
- flersidet variansanalyse
- lineære modeller
- kovariansanalyse
- modelkontrol

Konkret tages udgangspunkt i følgende eksempler:

- holdbarhed af afskårne roser
(-fra forelæsning d. 12/9-2017)
- hydrolyse af aminosyrer - eksempel 4.2
(-fra forelæsning d. 19/9-2017)

Detaljeringsgraden afhænger af jeres behov, så skyd bare løs ...

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 2/20



Faktorer: begreber

Du skal bl.a. kunne forklare begreberne:

- niveauer for en faktor
- balanceret faktor
- produktfaktor
- grovere/finere faktor
- trivielle faktorer: I og 0
- opbygning af faktordiagram for et flerfaktorforsøg
 - hvornår skal der tegnes pile?
 - hvordan beregnes frihedsgrader?
 - hvordan markeres "tilfældige faktorer"?

Faktordiagrammet kan ses på forelæsningsslides fra 12/9-2017.

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 4/20



Afskârne roser: modelreduktion

Du bør bl.a. kunne svare på følgende:

- hvor mange niveauer har de forskellige faktorer?
- hvilke faktorer er balancerede?
- hvordan konstrueres faktordiagrammet - herunder antal frihedsgrader?
- i hvilken rækkefølge foretages reduktion i modellen?

Ved den skriftlige eksamen:

- opskriv løbende statistiske modeller svarende til de hypoteser du tester
- angiv teststørrelse og p-værdi for testet
- skriv konklusion i ord efter hvert udført test
- anfør tydeligt hvilken model, som er din slutmodel, og **husk at spørgsmål vedr. estimater skal baseres på slutmodellen**

Ekstra lir vedrørende eksemplet med afskârne roser

- Tegnefilm til bedre forståelse af modellerne i 3-sidet **variansanalyse: 3way.pdf** fra ugeplan 2

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 5/20



Afskârne roser: R-stuff - hvad foregår her?

```
model=lm(tid~H*K+F)
confint(model)

##              2.5 %      97.5 %
## (Intercept)  8.6807103 10.735956318
## H1           -0.4532635  1.919930147
## K1           -0.1032635  2.269930147
## F2           -2.0651230 -0.009877016
## F3            0.5848770  2.640122984
## H1:K1         0.4552320  3.811434640

diff32<-c(0,0,0,-1,1,0)
est210<-c(1,1,0,1,0,0)
est<-rbind(diff32,est210)
```

```
estimable(model,est,conf.int=0.95)

##      Estimate Std. Error  t value DF    Pr(>|t|) Lower.CI Upper.CI
## diff32 2.650000  0.4891295  5.417788 18 3.793089e-05 1.622377  3.677623
## est210 9.404167  0.4891295 19.226332 18 1.900702e-13 8.376544 10.431790
```

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 7/20



Afskârne roser: slutmodel

Vores slutmodel bliver

$$Y_i = \alpha(\text{FLOR}_i) + \phi(\text{HANDLER} \times \text{KUNDE}_i) + e_i, e_i \text{ uafh. } \sim N(0, \sigma^2).$$

Slutmodellen er blot en **additiv model** mellem faktorerne **FLOR** og **HANDLER × KUNDE**.

Du bør kunne svare på følgende:

- Hvordan fortolkes parameterestimerne fra R-udskriften for en additiv model?
- Hvordan bestemmes estimater for grupper og kontraster som ikke er anført i R-udskriften?
- Hvordan bestemmes konfidensintervaller for gruppeestimer og kontraster?
- Hvordan beregnes LSD-værdier i modellen?

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 6/20



Lidt mere om LSD-værdier

Jeg kunne godt finde på at bede jer udregne LSD-værdier for

- den ensidede variansanalyse model: kompendiet s. 27
- det to-sidede variansanalyse model med vekselvirkning: kompendiet s. 33 (-i princippet den samme formel som ovenfor!)
- den additive model for to-sidede variansanalyse: kompendiet s. 34 (-NB: kun gyldig hvis $F \times G$ er balanceret!)

Øvrige LSD-værdier i kompendiet betragter jeg som nyttige formler, der dog kun er kursorisk pensum og ikke vil være nødvendige for en fuldstændig besvarelse af eksamenssættet.

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 8/20



Definition af lineær model

Observationer Y_1, \dots, Y_N .

En statistisk model for Y_1, \dots, Y_N kaldes **lineær** hvis

- Y_1, \dots, Y_N er uafh. og normalford. med samme spredning, σ
- Middelværdien af Y_i er en lineær funktion af parametrene:

$$\mathbb{E} Y_i = c_{i,1}\beta_1 + c_{i,2}\beta_2 + \dots + c_{i,p}\beta_p$$

Her er

- β_1, \dots, β_p parametre, dvs. ukendte tal (som vi vil estimere)
- c 'erne kendte tal som afhænger af designet

Skriver også:

$$Y_i = c_{i,1}\beta_1 + c_{i,2}\beta_2 + \dots + c_{i,p}\beta_p + e_i, \quad e_i \text{ iid } N(0, \sigma^2)$$



Kovariansanalyse: hydrolyse

```
data = read.table("../data/hydrolysis.txt", header=T)
data$logserine = log(data$serine)
data$hourfac = factor(data$hour)
attach(data)
```

```
head(data)
```

```
##      feed hour serine logserine hourfac
## 1 barley   8  4.47  1.497388         8
## 2 barley  16  4.34  1.467874        16
## 3 barley  24  4.22  1.439835        24
## 4 barley  32  4.10  1.410987        32
## 5 barley  72  3.48  1.247032        72
## 6 barley   8  4.46  1.495149         8
```

- Hvilke forklarende variable er der, og er de faktorer eller kovariater?
- Hvis en variabel kan opfattes som både faktor og kovariat, skal du være ekstra omhyggelig ved specifikation af modellerne.



Kovariansanalyse: modeloversigt

NB: Der er 5 forskellige modeller, som både indeholder `feed` og `hour` som forklarende variabel.

Derfor vigtigt at være præcis ved opskrivning af modeller:

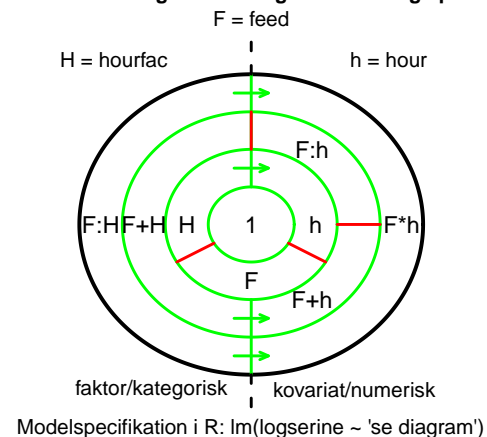
$$\begin{aligned} Y_i &= \gamma(\text{feed} \times \text{hour}_i) + e_i \\ Y_i &= \alpha(\text{feed}_i) + \beta(\text{hour}_i) + e_i \\ Y_i &= \alpha(\text{feed}_i) + \beta(\text{feed}_i) \cdot \text{hour}_i + e_i \\ Y_i &= \alpha(\text{feed}_i) + \beta \cdot \text{hour}_i + e_i \\ Y_i &= \alpha + \beta(\text{feed}_i) \cdot \text{hour}_i + e_i \end{aligned}$$

- Sørg for at have styr på, hvad de modellerne udtrykker
- Sørg for at have styr på, hvilke modeller, som kan testes mod hinanden



Kovariansanalyse: modeloversigt og reduktion

Testrækkefølge: udefra og ind eller langs pile



Hydrolyseeksempel: lineær model

Her nåede vi frem til flg. slutmodel

$$Y_i = \alpha(F_i) + \beta \cdot T_i + e_i, \quad e_i \sim N(0, \sigma^2).$$

Bemærk, at dette er en lineær model!

$$\begin{aligned} EY_1 &= 1 \cdot \alpha(\text{barley}) + 0 \cdot \alpha(\text{fish}) + \dots + 0 \cdot \alpha(\text{soy}) + 8 \cdot \beta \\ EY_2 &= 1 \cdot \alpha(\text{barley}) + 0 \cdot \alpha(\text{fish}) + \dots + 0 \cdot \alpha(\text{soy}) + 16 \cdot \beta \\ &\vdots \\ EY_{50} &= 0 \cdot \alpha(\text{barley}) + 0 \cdot \alpha(\text{fish}) + \dots + 1 \cdot \alpha(\text{soy}) + 72 \cdot \beta \end{aligned}$$

Middelværdien er en (kendt) **lineær funktion** af modellens parametre.



Kovariansanalyse: estimer

```
summary(model)
```

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|--------------|--------------|------------|--------------|
| ## (Intercept) | 1.535482640 | 0.0036425549 | 421.540016 | 5.476254e-81 |
| ## feedfish | -0.066317372 | 0.0043976224 | -15.080279 | 5.374849e-19 |
| ## feedmais | 0.157997696 | 0.0043976224 | 35.927981 | 3.029453e-34 |
| ## feedmeat | -0.018536984 | 0.0043976224 | -4.215229 | 1.220410e-04 |
| ## feedsoy | 0.229227446 | 0.0043976224 | 52.125313 | 3.415175e-41 |
| ## hour | -0.004071875 | 0.0000624018 | -65.252521 | 1.973932e-45 |

Typiske spørgsmål ved eksamen samt nogle gode råd:

- Hvilken (slut-) model har vi fat i?
- Estimer (husk at anføre variansestimater)
- Stemmer antal og fortolkning af parametre med hvad du tror?
- Konfidensintervaller for (systematiske) parametre i slutmodel



Kovariansanalyse: eksempler på tillægsspørgsmål

Overvej, om du ved, hvordan du skal svare på følgende:

- Forventede ændring i log-serin indhold for feed=mais når hydrolysetiden ændres med 10 timer?
- Forventede forskel i log-serin indhold mellem meat og barley for hydrolysetid hour=24 timer?
- Forventede log-serin indhold for barley ved hydrolysetid hour=16 timer?

Gode råd i forbindelse med eksamen:

- Husk at svar skal baseres på estimer fra slutmodel
- Du er ofte nødt til at bruge estimable til beregning af konfidensintervaller for estimatet
- Forsøg dog at beregne selve estimatet ved håndkraft som kontrol af, at du benytter estimable korrekt



Modelkontrol: oversigt

Det er vigtigt at kontrollere antagelserne bag en statistisk model

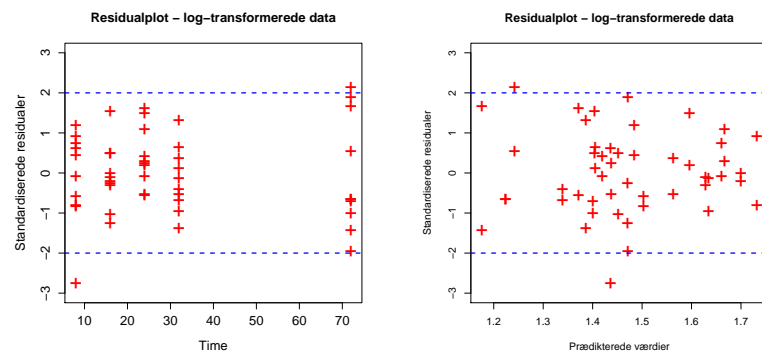
- systematisk del: indgår de rigtige faktorer og kovariater i beskrivelsen af middelværdi-strukturen?
- tilfældig del: er obs. uafhængige og normalfordelte med samme varians (varianshomogenitet)?

Værktøjer til modelkontrol:

- residualplot: prædikterede værdier mod standardiserede residualer
- manglende varianshomogenitet \Rightarrow prøv transformation af respons
- residualplot: kovariater mod standardiserede residualer
manglende varianshomogenitet \Rightarrow
kovariat skal ikke indgå lineært, prøv fx. at tilføje kvadratisk led
- QQ-plot: når varianshomogenitet er opnået benyttes QQ-plot til at checke normalfordelingsantagelse (-ligger punkter omkring ret linje?)



Modelkontrol: hydrolyse-eksempel

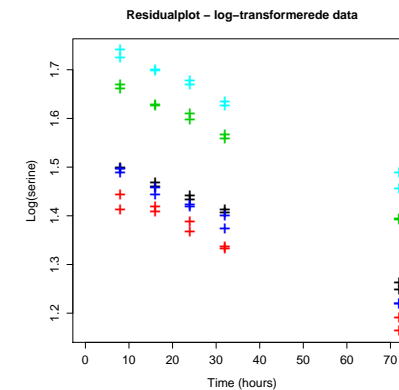


Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 17/20



Modelkontrol: visuel kontrol af lineær sammenhæng

Man kan altid få noget ud af at optegne responsen med kovariaten og evt lave punkternes farve/form afhænge af øvrige faktorer



Lineær sammenhæng mellem log(serine) og hydrolysetid virker

OK
Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 18/20



Modelkontrol: test af lineær sammenhæng

I hydrolyse-eksemplet er hydrolyse-tiden en del af forsøgsdesignet, hvorfor vi kan teste om hydrolyse-tiden skal indgå lineært

- Hvilke test svarer til test af en linearitetshypotese? (-forklar på modelhjulet)

Alternativ metode til test af linearitet:

- Fit model med kvadratisk led

$$Y_i = \alpha(\text{feed}_i) + \beta \cdot \text{hour}_i + \delta \cdot \text{hour}_i^2 + e_i$$

- Test hypotesen, $H_0 : \delta = 0$ svarende til modellen

$$Y_i = \alpha(\text{feed}_i) + \beta \cdot \text{hour}_i + e_i$$

Er der andre måder at lave et test for linearitet, hvor man udnytter samme trick med at tilføje et kvadratisk led? (-se modelhjulet)

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 19/20



Modelkontrol: test af linearitet

```
model3kvad<-lm(logserine~feed+hour+I(hour^2))
```

```
##
## Call:
## lm(formula = logserine ~ feed + hour + I(hour^2))
##
## Coefficients:
## (Intercept)    feedfish    feedmaize    feedmeat    feedsoy
## 1.534e+00    -6.632e-02    1.580e-01    -1.854e-02    2.292e-01
##      hour      I(hour^2)
## -3.959e-03    -1.348e-06
```

```
model3<-lm(logserine~feed+hour)
```

```
anova(model3,model3kvad)
```

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|------------|----|-------------|--------|--------|
| 1 | 43 | 0.00079998 | | | | |
| 2 | 44 | 0.00080247 | -1 | -2.4889e-06 | 0.1338 | 0.7163 |

Anders Tolver — Oversigt 1 — SD2 31/10-2017
Dias 20/20

