

## Kovariansanalyse

### Statistisk Dataanalyse 2

Anders Tolver

Uge 3, torsdag d. 21/9-2017



## Kovariansanalyse

Ved et faktorforsøg med to faktorer

A, B

tager man typisk udgangspunkt i modellen

$$Y_i = \mu(A \times B_i) + e_i$$

Idé: for hver forsøgsenhed måles en **kovariat**,  $x_i$ , som (måske) kan forklare noget af variationen ud fra en modificeret model

$$Y_i = \mu(A \times B_i) + \gamma \cdot x_i + e_i.$$

Kovariaten adderes multipliceret med en koefficient (hældning).

Der inddrages kun 1 parameter mere i modellen.



## Dagens program

Dagens undervisning dækkes af kompendiets kapitel 6.

Vi diskuterer nogle eksempler, hvor der inddrages en kontinuert forklarende variable i en statistisk model. Der kan være flere formål med dette.

- Forsøgets formål kan være at forstå sammenhængen mellem responsen og en forklarende kovariat
  - Hydrolyseforsøg i eksempel 4.2 (tirsdag)
  - Øvelsesopgave 4.4 (exercise 5.2 i kompendiet)
  - Afleveringsopgave 2 (til tirsdag d. 26/9-2017)
- Inddragelse af en forklarende kovariat kan mindske variationen og gøre det lettere at se en evt. behandlingseffekt
  - Hormonbehandling af stude (eksempel 6.1)
  - Vækst af træer (eksempel 6.3 - selvstudie)

Desuden diskuteres kompendiets eksempel 4.2.

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 2/25



## Eksempel 6.1: hormonbehandling af stude

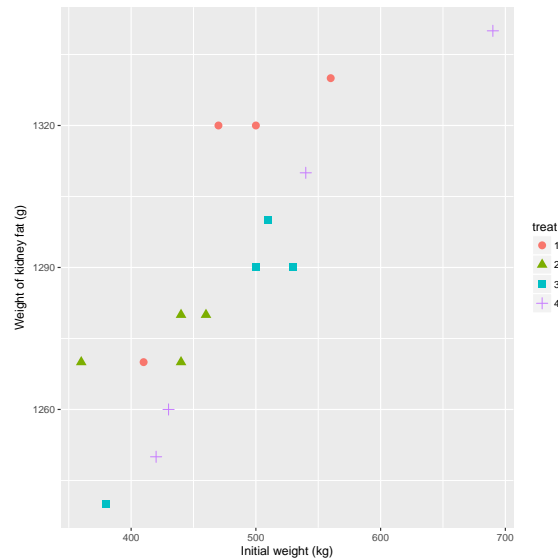
16 forskellige stude indgår i flg. forsøg

- 4 forskellige hormonbehandlinger (treat)
- 4 stude tildeles hver behandling
- $x$  = begyndelsesvægt (kg)
- $Y$  = vægt af nyrefedt (g) efter forsøgsperiode

Hormonbehandling							
1		2		3		4	
x	Y	x	Y	x	Y	x	Y
560	1330	440	1280	530	1290	690	1340
470	1320	440	1270	510	1300	420	1250
410	1270	360	1270	380	1240	430	1260
500	1320	460	1280	500	1290	540	1310



## Eksempel 6.1: hormonbehandling af stude



Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 5/25



## Eksempel 6.1: oversigt over modeller

Modeller for kovariansanalyse

$$\text{Model 1: } Y_i = \alpha(\text{treat}_i) + \gamma \cdot x_i + e_i$$

$$\text{Model 2: } Y_i = \mu + \gamma \cdot x_i + e_i \quad (\text{ingen effekt af treat})$$

$$\text{Model 3: } Y_i = \alpha(\text{treat}_i) + e_i \quad (\text{ingen effekt af } x)$$

$$\text{Model 4: } Y_i = \mu + e_i$$

Samme modeller, som hvis  $x$  havde være en faktor, der kunne gives en numerisk fortolkning (jf. tirsdag d. 19/9-2017).

Forskellen er, at vi her ikke kan designe forsøget, så vi kun afprøver stude med bestemte værdier af variabelen  $x$ . I stedet er  $x$  en variabel som inkluderes i analysen ud over faktoren  $\text{treat}$ , som vi selv kan kontrollere.

Ved at inkludere kovariater kan man af og til mindske variationen og gøre det lettere at se en effekt af behandlingen.

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 6/25



## Eksempel 6.1: fit af modeller

Nogle R-kommandoer

```
model1=lm(steers$y~factor(steers$treat)+steers$x)
model2=lm(steers$y~steers$x)
model3=lm(steers$y~factor(steers$treat))
model4=lm(steers$y~1)
deviance(model1);deviance(model2);deviance(model3);deviance(model4)

## [1] 1387.183
## [1] 3800.668
## [1] 9900
## [1] 12775
```

Variansanalyseskema

Model	Fakt.	Mv.	$SS_e$	$df_e$
1	treat + x	$\alpha(\text{treat}_i) + \gamma \cdot x_i$	1387.183	11
2	x	$\mu + \gamma \cdot x_i$	3800.668	14
3	treat	$\alpha(\text{treat}_i)$	9900.00	12
4	0	$\mu$	12775.00	15

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 7/25



## Eksempel 6.1: reduktion af model

Testskema

Test	Faktor	F	df	p
2 vs 1	treat (just. for x)	6.379	3	0.009
3 vs 1	x (just. for treat)	67.50	1	< 0.0001
4 vs 3	treat	1.162	3	0.365
4 vs 2	x	33.06	1	< 0.0001

Hvordan konstrueres teststørrelsen for test af model2 mod model1 ud fra oplysningerne i variansanalyseskemaet på foregående side?

Hvad bliver slutmodellen?

Ville vi få samme konklusioner, hvis vi ikke inddrog begyndelsvægten ( $x$ ) som kovariat i analysen?

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 8/25



## Eksempel 6.1: slutmodel

### Slutmodel

$$\text{Model 1: } Y_i = \alpha(\text{treat}_i) + \gamma \cdot x_i + e_i, \quad e_i \text{ uafh. } \sim N(0, \sigma^2).$$

### Parameterestimer i R (-fjern intercept)

```
modella <- lm(steers$y ~ steers$x + factor(steers$treat) - 1)
summary(modella)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## steers$x      0.3287      0.0400   8.2161 5.065272e-06
## factor(steers$treat)1 1150.5901    20.1982 56.9649 6.024314e-15
## factor(steers$treat)2 1135.3109    17.9050 63.4074 1.859876e-15
## factor(steers$treat)3 1122.2335    20.0062 56.0943 7.132520e-15
## factor(steers$treat)4 1119.0863    21.5467 51.9377 1.658701e-14
```

Residual standard error: 11.23 on 11 degrees of freedom

### Fortolkning af output? Hvad beskriver parameterestimer?

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 9/25



## Eksempel 6.1: adjusted means

### Estimer for en gennemsnits stud

$$\hat{\alpha}(\text{treat}_i) + \hat{\gamma} \cdot \bar{x}$$

### Brug estimable i R

```
library(gmodels)
xbar <- mean(steers$x) ## gnsn. af begyndelsesvaegt!
adj.m1 <- c(xbar, 1, 0, 0, 0)
adj.m2 <- c(xbar, 0, 1, 0, 0)
adj.m3 <- c(xbar, 0, 0, 1, 0)
adj.m4 <- c(xbar, 0, 0, 0, 1)
adj.means <- rbind(adj.m1, adj.m2, adj.m3, adj.m4)
```

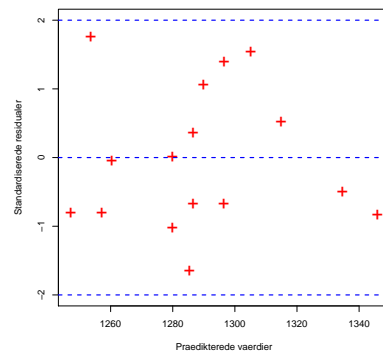
```
estimable(modella, adj.means, conf.int=0.95)
```

```
##              Estimate Std. Error t value DF Pr(>|t|) Lower.CI Upper.CI
## adj.m1 1307.535      5.622891 232.5378 11      0 1295.159 1319.911
## adj.m2 1292.256      5.994818 215.5621 11      0 1279.061 1305.450
## adj.m3 1279.178      5.615771 227.7832 11      0 1266.818 1291.539
## adj.m4 1276.031      5.866644 217.5061 11      0 1263.119 1288.943
```

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 10/25



## Eksempel 6.1: modelkontrol

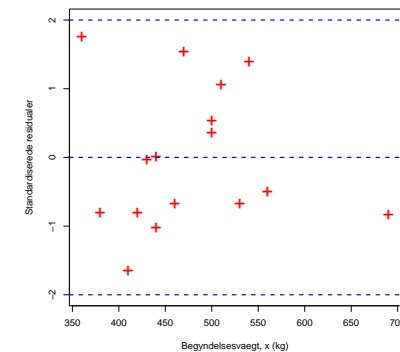


Lav selv et qq-plot, for at checke normalfordelingsantagelsen.

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 11/25



## Eksempel 6.1: modelkontrol



Alternativ: Kunne have testet, om kvadratisk model giver bedre fit

$$Y_i = \alpha(\text{treat}_i) + \gamma \cdot x_i + \delta \cdot x_i^2 + e_i$$

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 12/25



## Kovariansanalyse: oversigt

- God forsøgsteknik til forbedring af præcisionen
- Næste "gratis" (-koster 1 frihedsgrad)
- Virker bedst med gode/relevante kovariater
- Ideelt med kovariater der måles inden forsøgets start og således er uafhængige af behandlingen
- Muligt at inddrage flere kovariater i analysen
- Godt alternativ til balancering af grupper, hvor man tilstræber at grupperne er ens mht. fx. størrelse
- Svær fortolkning hvis kovariaten kan være påvirket af behandlingen (kan føre til en dårlig analyse)



## Eksempel 4.1: sphagnum

Observationer:  $Y_1, Y_2, \dots, Y_{48}$

To forklarende variable (faktorer)

- Sphagnumtype: stype eller S [ antal niveauer: 12 ]
- Bakke: tray eller T [ antal niveauer: 6 ]

Ikke alle kombinationer af S og T er med i eksperimentet:  
**ufuldstændigt blokdesign** (mere om det i kap. 9).

Hvordan ville du analysere disse data?

Observation nummer 46 (obs. værdi 7.5) er mærkelig og fjernes.

```
sphag<-sphag[-46,]
```

Er det egentlig ok? Kunne evt. køre analysen både med og uden denne obs. og så sammenligne resultaterne fra de to analyser.



## Eksempel 4.1: sphagnum

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

```
##      stype tray volume
## 1      1      1  37.0
## 2      1      3  44.6
## 3      1      5  42.5
## 4      1      6  47.1
## 5      2      3  49.0
## 6      2      4  50.5
```

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

```
##      stype tray volume
## 47     12      5  36.2
## 48     12      6  25.5
```

```
sphag$stype<-factor(sphag$stype)
sphag$tray<-factor(sphag$tray)
```



## Sphagnum: test for effekt af stype

Statistisk model:

$$1: Y_i = \alpha(S_i) + \beta(T_i) + e_i$$

hvor  $e_i$ 'erne er uafhængige  $N(0, \sigma^2)$ -fordelte.

Interesseret i hypotesen om ingen effekt af sphagnumtypen.

- Hvad er hypotesen (udtrykt ved  $\alpha$ 'er og/eller  $\beta$ 'er)?
- Hvad er den tilsvarende model?
- Hvordan tester vi hypotesen i R?
- Hvad er konklusionen?

Vil slet ikke interessere os for om der en tray-effekt (men prøv selv!)



## Sphagnum: sammenligning af typer

Har altså påvist en forskel på sphagnumtyperne. Mangler at gøre rede for forskellene!

I balanceret forsøg ville vi bruge forskelle mellem stype-gennemsnit og sammenligne vha. LSD-værdi.

Dur ikke i ubalanceret forsøg:

- $\hat{\alpha}(j) - \hat{\alpha}(k)$  er ikke blot forskellen mellem stype-gennemsnit. "Korrigeres" ift. hvilke trays behandlingerne er afprøvet i.
- Forskellene  $\alpha(j) - \alpha(k)$  estimeres med forskellig præcision fordi de mødes flere eller færre gange i samme tray. Fx. mødes 1 og 2 tre gange mens 1 og 3 kun mødes to gange.
- Forskellige LSD-værdier til forskellige sammenligninger.
- 66 parvise forskelle — og med forskellige spredninger og forskellige LSD-værdier.

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 17/25



## Sphagnum: sammenligning af typer

```
model1 <- lm(volume ~ stype + tray, data=sphag)
```

```
##           Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 44.17486    2.290726  19.284221 1.865834e-18
## stype2       6.01880    2.750850   2.187979 3.659589e-02
## stype3      -4.57442    2.805078  -1.630764 1.133958e-01
```

...

```
##           Estimate Std. Error  t value    Pr(>|t|)
## stype12    -9.65722440    2.951357  -3.27213060 0.002687227
## tray2      -0.33208762    2.113615  -0.15711829 0.876204303
## tray3       0.45346199    2.054075   0.22076214 0.826773243
## tray4       0.02479945    1.983899   0.01250036 0.990109177
## tray5      -1.64795268    2.050432  -0.80371004 0.427887160
## tray6      -4.30496861    2.013534  -2.13801650 0.040779349
```

Hvordan skal parametrene fortolkes?

Hvad angiver parameteren svarende til linjen stype3? Denne er ikke beregnet som i det balancerede tilfælde!

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 18/25



## Sphagnum: adjusted means

Ikke nødvendigvis mere rimeligt at bruge tray 1 som reference end en af de øvrige.

En mulighed er at beregne **de forventede værdier i en "gennemsnitsbakke"** (som ikke findes), dvs. se på

$$\frac{1}{6} \sum_{j=1}^6 (\hat{\alpha}(\text{stype}) + \hat{\beta}(j))$$

For eksempel, for sphagnumtype 1:

$$\frac{1}{6} (44.17 + [44.17 + (-0.33)] + [44.17 + 0.45] + [44.17 + 0.02] + [44.17 + (-1.65)] + [44.17 + (-4.31)]) = 43.2$$

Sådanne størrelser kaldes **"adjusted means"** eller "justerede gennemsnit".

Kan med lidt arbejde beregnes med estimable.

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 19/25



## Sphagnum: estimable

estimable skal kende linearkombinationen. For type 1:

$$\alpha(1) + \frac{1}{6} (\beta(1) + \dots + \beta(6))$$

Dette svarer til flg. linearkombination, se summary(model1):

$$(1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1/6, 1/6, 1/6, 1/6, 1/6)$$

Tilsvarende for type 2 (og 3–12).

```
library(gmodels)
adj1 = c(1,rep(0,11),rep(1/6,5))
adj2 = c(1,1,rep(0,10),rep(1/6,5))
adj = rbind(adj1,adj2)
```

```
estimable(model1,adj,conf.int=0.95)
```

```
##           Estimate Std. Error  t value DF Pr(>|t|) Lower.CI Upper.CI
## adj1  43.20707    1.961597  22.02648  30    0 39.20096 47.21319
## adj2  49.22587    1.959502  25.12162  30    0 45.22404 53.22771
```

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 20/25



## Eksempel 6.3: vækst af træer (selvstudie)

```
soil<-read.table("../data/bms_examp6_3.txt",header=T)
soil
```

```
##    block trt  ht growth
## 1     1   1  3.6   8.9
## 2     1   2  3.1  10.7
## 3     1   3  4.7  12.4
## 4     2   1  4.7  10.1
## 5     2   2  4.9  14.2
## 6     2   3  2.6   9.0
```

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

```
##    block trt  ht growth
## 28    10   1  5.3  12.6
## 29    10   2  4.4  11.4
## 30    10   3  5.8  13.4
## 31    11   1  3.6   7.4
## 32    11   2  1.4   8.4
## 33    11   3  4.8  10.7
```

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 21/25



## Eksempel 6.3: vækst af træer (selvstudie)

Balanceret tofaktorforsøg

- 11 blokke med 3 plots i hver
- 3 jordbehandlinger
- $x$  = højde (feet) inden behandling (kovariat)
- $Y$  = vækst (feet) efter 5 år

Udgangsmodel (kovariansanalyse)

$$Y_i = \alpha(\text{beh}_i) + \beta(\text{blok}_i) + \gamma \cdot x_i + e_i$$

Hypotese (ingen effekt af behandling)

$$Y_i = \beta(\text{blok}_i) + \gamma \cdot x_i + e_i$$

kovariansanalyse - modeller

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 22/25



## Eksempel 6.3: vækst af træer (selvstudie)

Modelskema

Model	Fakt.	Mv.	$SS_e$	$df_e$
1	beh+blok+x	$\alpha(\text{beh}_i) + \beta(\text{blok}_i) + \gamma \cdot x_i$	30.65208	19
2	blok+x	$\beta(\text{blok}_i) + \gamma \cdot x_i$	49.2468	21
3	beh+blok	$\alpha(\text{beh}_i) + \beta(\text{blok}_i)$	68.88303	20
4	beh+x	$\alpha(\text{beh}_i) + \gamma \cdot x_i$	37.52238	29
5	blok	$\beta(\text{blok}_i)$	73.14	22
6	beh	$\alpha(\text{beh}_i)$	201.7127	30

Testskema

Test	Faktor	F	df	p
2 vs. 1	beh (just. for blok og x)	5.76	2	0.011
3 vs. 1	x	23.670	1	0.0001
4 vs. 1	blok (just. for beh og x)	0.43	10	0.916
5 vs. 3	beh (kun just. for blok)	0.62	2	0.549
6 vs. 3	blok (kun just. for beh)	3.86	10	0.005

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 23/25



## Eksempel 6.3: vækst af træer (selvstudie)

Slutmodel (når vi inkluderer beg. højde i modellen)

$$Y_i = \alpha(\text{beh}_i) + \gamma \cdot x_i + e_i$$

[-kan vise at hverken beh ( $p = 0.002$ ) eller x ( $p < 0.0001$ ) kan fjernes!]

**NB** Blokken er unødvendig når kovariaten (beg. højde) er med i modellen.

Parameterestimer

```
model4new<-lm(soil$growth~soil$ht+factor(soil$trt)-1)
summary(model4new)
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## soil$ht          1.530316   0.1358481  11.264907  4.141145e-12
## factor(soil$trt)1  3.829704   0.6218255   6.158808  1.031198e-06
## factor(soil$trt)2  5.597206   0.5421586  10.323927  3.186340e-11
## factor(soil$trt)3  3.926125   0.6013768   6.528561  3.777495e-07
```

Residual standard error: 1.137 on 29 degrees of freedom

Hvordan skal parameterestimerterne fortolkes?

Anders Tolver — Kovariansanalyse — SD2 21/9-2017  
Dias 24/25



## Øvelser til hjemmebrug/repetition

Tag udgangspunkt i `model4new` fra foregående side

- Brug `estimable` til at finde alle parvise forskelle ml. behandlingsgrupperne (`estimator+konf.int`).

Tag udgangspunkt i `model1` (-selvom det strengt taget ikke er slutmodellen)

- Brug `estimable` til at finde alle parvise forskelle ml. behandlingsgrupperne (`estimator+konf.int`).
- Brug `estimable` til for hver af de tre behandlinggrupper at beregne adjusted means for et gennemsnitstræ ( $\bar{x}$ ) i en (fiktiv) gennemsnitsblok.

Er der signifikante forskelle på behandling 1 og 3?

