Gentagne målinger

Statistisk Dataanalyse 2

Anders Tolver

Uge 5, tirsdag d. 3/10-2017

Dagens program

Ugens tema:

Gentagne målinger. Meget hyppigt forekommende!

I dag:

- Eksempel 8.3: harskhed af svinekød (ikke gentagne målinger)
- Intro til gentagne målinger, forslag til analyser
- Analyse af summary measures
- Random intercepts model

Torsdag:

- Diggle-modellen for gentagne målinger

Eksempel 8.3: harskhed af svinekød

```
##    animal feed packaging storage ran
## 1       1    1         1       1 2.0
## 2       1    1         1       2 2.3
## 3       1    1         1       3 2.3
## 4       1    1         2       1 0.6
## 5       1    1         2       2 2.1
## 6       1    1         2       3 2.1
## 7       2    2         1       1 0.8
## 8       2    2         1       2 1.7
## 9       2    2         1       3 1.4
## 10      2    2         2       1 0.4
## 11      2    2         2       2 0.8
## 12      2    2         2       3 1.2
## 13      3    1         1       1 1.2
## 14      3    1         1       2 2.1
## 15      3    1         1       3 2.7
## 16      3    1         2       1 0.9
## 17      3    1         2       2 1.5
## 18      3    1         2       3 1.9
## 19      4    2         1       1 0.4
## 20      4    2         1       2 0.9
## 21      4    2         1       3 1.3
## 22      4    2         2       1 0.1
## 23      4    2         2       2 1.1
## 24      4    2         2       3 0.9
```

Eksempel 8.3: harskhed af svinekød

Forsøgsdesign

- 4 dyr (pigs) beskrevet ved faktoren `A` (`animal`)
- Randomiseres i grupper af 2 dyr som modtager `feed=1` eller `feed=2`.
- 6 cutlets fra hvert dyr
- De seks cutlets randomiseres til lagringsmetoder givet ved faktorene
  - `P` (`packaging`) påto niveauer `A` og `B`
  - `S` (`storage period`) påtre niveauer `2, 5` eller `8` uger

Der er tale om et splitplot forsøg

- Hvad er helplots?
- Hvad er helplot-faktoren?
- Hvad er delplots?
- Hvad er delplot-faktoren?

Eksempel 8.3: harskhed af svinekød

Vi tager udgangspunkt i flg. statistiske model

$$Y_i = \delta(\mathtt{F} \times \mathtt{P} \times \mathtt{S}_i) + b(\mathtt{A}_i) + e_i,$$

- $b(1), \ldots, b(4)$ er uafhængige $\sim N(0, \sigma_A^2)$
- $e_1, \ldots, e_{24}$ er uafhængige $\sim N(0, \sigma^2)$

Faktordiagram: - se Figure 8.5 pås. 150 i kompendiet.

Analyse

- Modelkontrol. Hvordan foretages dette?
- Modelreduktion. Hvordan foretages denne?
- Hvad bliver slutmodellen?
- Parameterestimater
- Hvad er den største kilde til variation?

Eksempel 8.3: harskhed af svinekød

Slutmodel

$$Y_i = \alpha(\mathtt{F}_i) + \beta(\mathtt{P}_i) + \gamma(\mathtt{S}_i) + b(\mathtt{A}_i) + e_i, b(j) \sim N(0, \sigma_A^2), e_i \sim N(0, \sigma^2)$$

Estimater (fixed effects)

$$\begin{aligned}
\text{reference gruppe} \quad & \hat{\alpha}(1) + \hat{\beta}(1) + \hat{\gamma}(1) = 1.4750 \\
\text{storage period} \quad & \hat{\gamma}(2) - \hat{\gamma}(1) = 0.7625 \quad \hat{\gamma}(3) - \hat{\gamma}(1) = 0.9250 \\
\text{packaging} \quad & \hat{\beta}(2) - \hat{\beta}(1) = -0.4583333 \\
\text{feed} \quad & \hat{\alpha}(2) - \hat{\alpha}(1) = -0.892
\end{aligned}$$

Estimater (random effects)

$$\begin{aligned}
\hat{\sigma}_A^2 &= 0.1205^2 = 0.0145 \text{(- 17 \% af variationen)} \\
\hat{\sigma}^2 &= 0.2644^2 = 0.0699 \text{(- 83 \% af variationen)}
\end{aligned}$$

Eksempel 8.3: harskhed af svinekød

Testet for hovedeffekt af `F`:

- p-værdien fra likelihood ratio testet er næppe troværdig, da vi kun har to dyr (gentagelser) per behandling
- Kan eventuelt forsøge at benytte `simulate.lme`
- Kan alternativt udføres som et F-test:
  - Hvilken faktor skal F (`feed`) testes op imod?
  - Hvordan konstrueres teststørrelsen?

Vi finder, at

$$F = 30.37 \sim F(1,2) \Rightarrow p = 0.03$$

Hvordan bestemmes antallet af frihedsgrader?

Hvordan bestemmes p-værdien ud fra F-teststørrelsen?

Eksempel 10.1: geders vægtudvikling

Interesseret i fire fodertypers effekt pågeders vægtudvikling.

Faktorer:

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

Det karakteristiske er, at der er flere målinger for hver ged.
Ugens tema: hvordan kan vi analysere sådanne data?

Målingen fra dag 0 er taget før behandlingen startes: hvordan kan/bør den bruges?

Eksempel 10.1: geders vægtudvikling

```
data<-read.table("../data/goats1.txt",header=T)
data$feed<-factor(data$feed)
data$dayf<-factor(data$day)
data
```

```
##    goat feed   w0 day weight dayf
## 1     1    1 20.4   0   20.4    0
## 2     1    1 20.4  26   21.0   26
## 3     1    1 20.4  45   21.5   45
## 4     1    1 20.4  61   21.3   61
## 5     1    1 20.4  91   22.3   91
## 6     2    1 10.3   0   10.3    0
## 7     2    1 10.3  26   11.4   26
## 8     2    1 10.3  45   11.6   45
## 9     2    1 10.3  61   12.0   61
## 10    2    1 10.3  91   12.5   91
```

Vigtigt at tegne data for at fåoverblik. Hvordan?

Gentagne målinger: generelt set-up

Data:

flere målinger fra hver forsøgsenhed (person, træ, plante, dyr, ...)

Som regel målinger fra flere tidspunkter, men kunne fx. ogsåvære fra forskellige steder på personen/dyret/planten/...

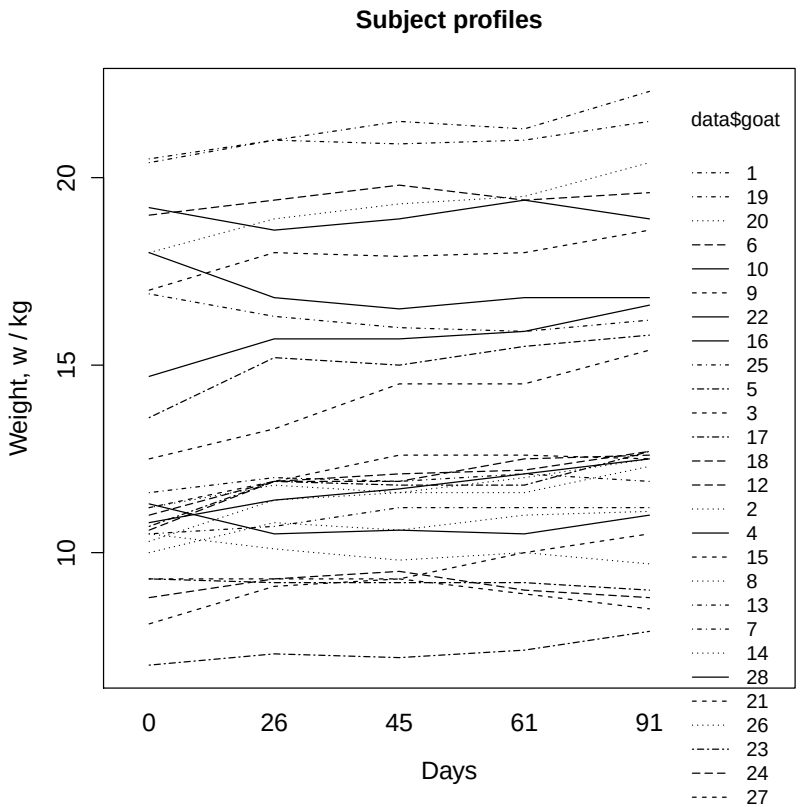
Formål (typisk):

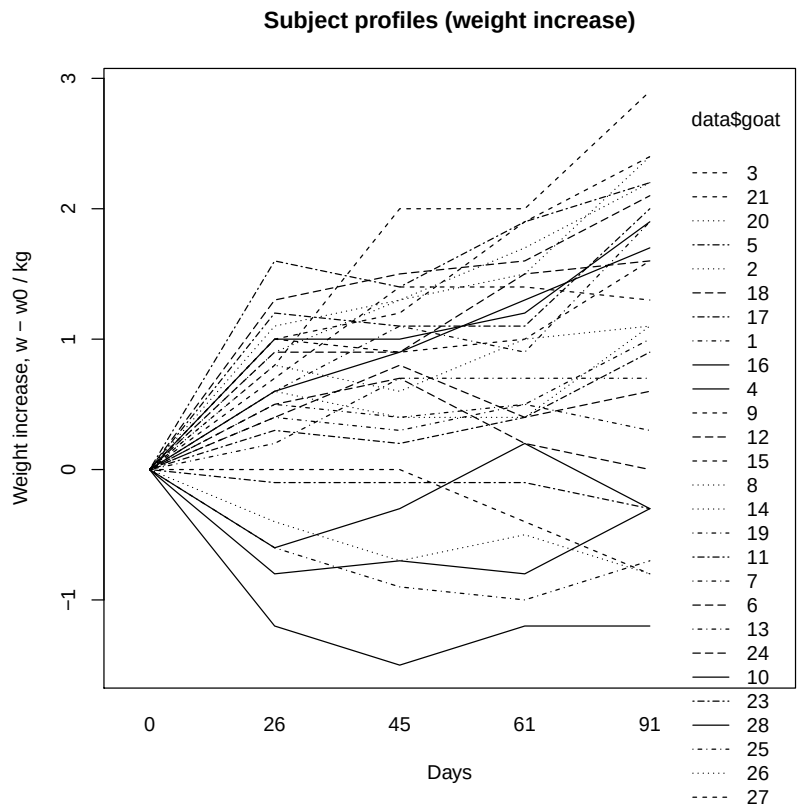
sammenligning af behandlinger, sammenligning af udviklingen over tid.

Illustrative figurer (-meget mere info i R program)

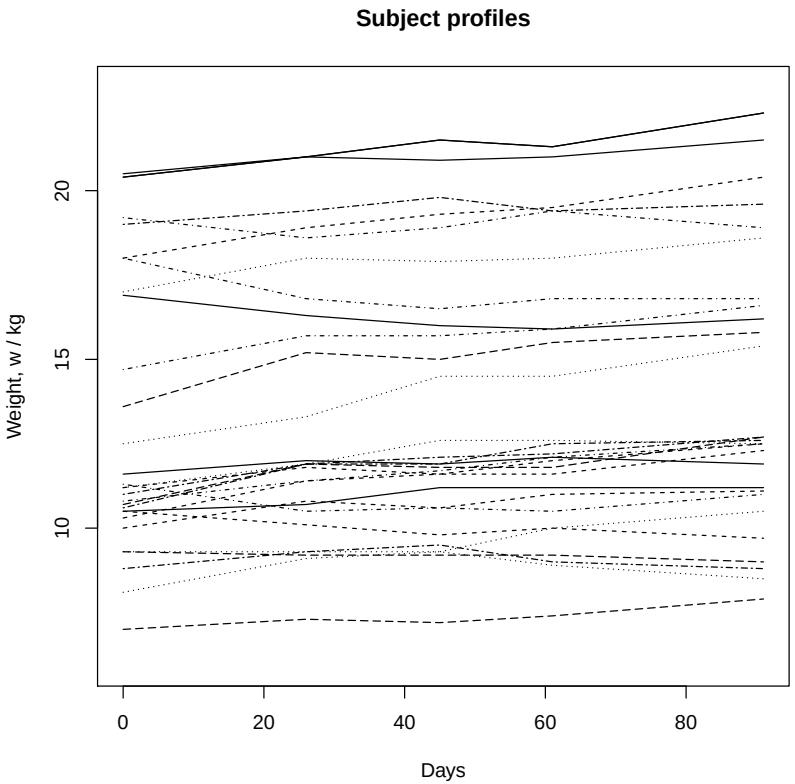
- Plot af de individuelle "profiler", dvs. en profil per individ
- Plot af "gennemsnitsprofil" for hver behandling
- `interaction.plot` nyttig, i hvert fald hvis obs. er taget med samme tidsafstand.

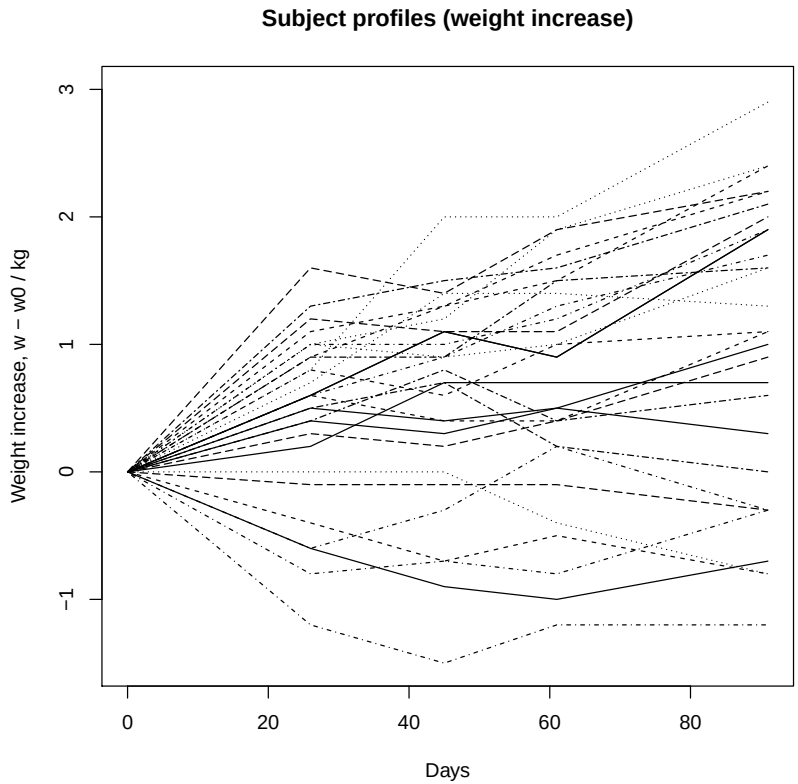
Gentagne målinger: plot profiler





Gentagne målinger: plot profiler





Gentagne målinger: analysemetoder

Tre ofte benyttede analysemetoder:

- analyse af et (eller flere) "summary measure(s)"
  - data reduceres til et målepunkt per subjekt
  - dernæst f.x. 1-sidet ANOVA med `feed` som faktor
- "almindelig" analyse med individ som tilfældig effekt (uge 4)
  - alle målinger inddrages
  - vækstkurver beskrives ved at inddrage `day` og `feed` i middelværdistrukturen
  - inden for beh. grupper givet ved `feed` er vækstkurverne for forskellige dyr forskudt lidt i forhold til hinanden (-random effect)
- model med seriel korrelationsstruktur (fx. Diggle)
  - målinger taget "tæt påhinanden" i tid ligner hinanden mere end målinger taget "lang tid fra hinanden"

Ugens program består i at snakke om disse analyser.

Analyse af summary measure(s)

Idé:

- reducér data for hvert individ til en enkelt observation
- analysér disse observationer "påsædvanlig måde"

For eksempel: Vægt pådag 91, vægtændring fra dag 0 til dag 91.

- Hvad består del-datasættet af?
- Hvordan skal vi analysere dette del-datasæt (hvilken model)?
- Kan vi inddrage startværdien (dag 0) i analysen?

Analyser i afsnit 10.1 ($w_{91} - w_{26}$) og opgave 6.1 (w_{91}/w_0).

Analyse af summary measures (2)

Eksempler påsummary measures:

- tilvækst fra start til slut
- gennemsnittet af målingerne
- areal under kurve (AUC)
- hældning påkurve
- maximal værdi
- tidspunkt for maximal værdi

Husk at størrelserne skal udregnes for hvert individ — ikke som gennemsnit over individer.

Vælg med omhu: ikke alle størrelser er relevante for alle datasæt.

Analyse af summary measures (3)

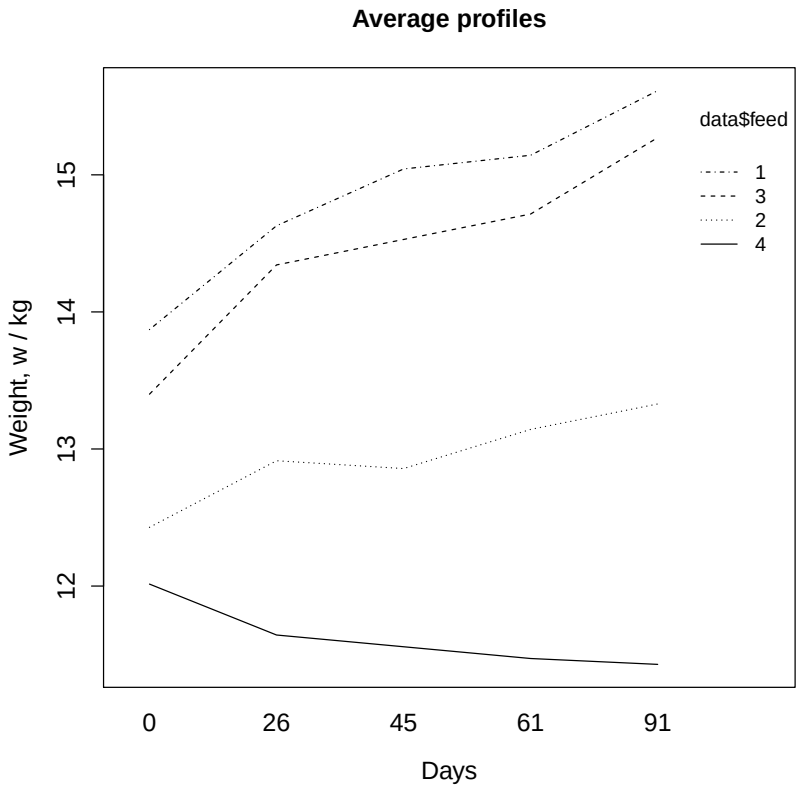
Valg af summary measure(s):

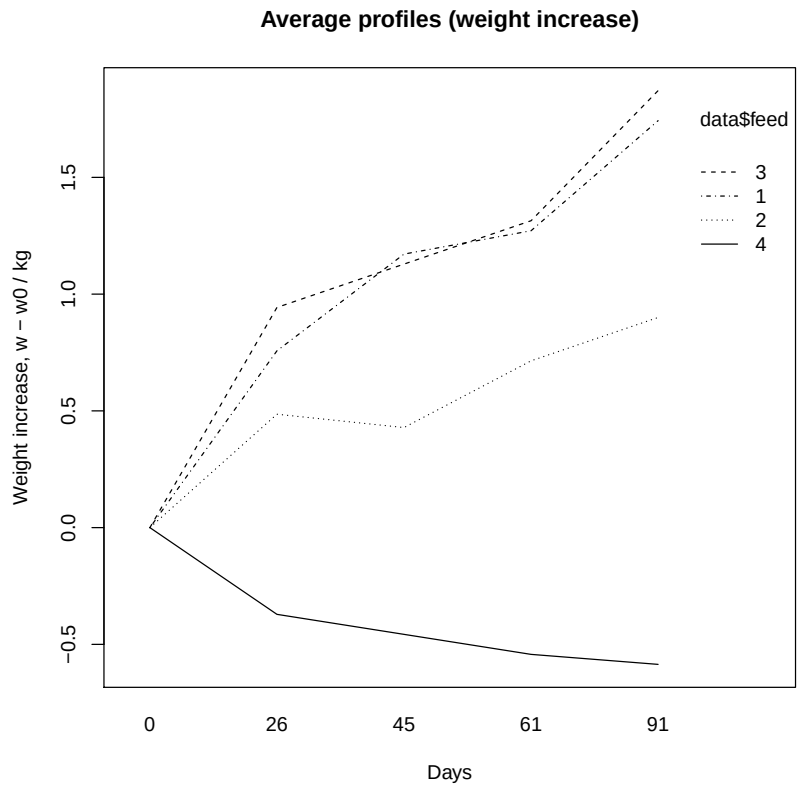
- Et godt summary measure måler noget der er vigtigt! og som vi kan forståhvad er (biologisk)
- Et summary measure måikke vælges fordi "det ser ud til at give en forskel" (significance hunting).
- OK at analysere flere summary measures (men ikke for mange!) — men de bør måle noget forskelligt.

Analyse af relevant summary measure er tit en god og robust analyse — men udnytter ikke alle data.

Analyse af summary measures (4)

Det er ikke ok at benytte følgende figurer til at vælge et summary measure!





Random intercepts modellen

Vil bruge model for alle observationer fra dag 26,45,61,91!

- Hvad med målingen pådag 0?
- Hvilke faktorer er det naturligt at inddrage i modellen?
- Hvilke faktorer bør være systematiske?
- Hvilke faktorer bør være tilfældige?
- Faktordiagram?
- Sammenlign med split-plot model.

Random intercepts modellen (2)

Mulig model:

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

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

Random intercepts model: Tilfældigt niveau for hver ged.

Nyt(?): både baselinemåling og tilfældig faktor i samme model.

Modellen fittes i R vha. `lme`:

```
lme(weight ~ w0 + feed*timefac, random =~ 1 | goat)
```

Modelreduktion: hvilke hypoteser er relevante?
Prøv selv (eller se R-program)!

Random intercepts modellen: lidt R-kode

```
library(nlme)
data2<-subset(data,day!=0)
m0<-lme(weight~w0+feed:dayf-1,random=~1|goat
,data2,method="ML")
summary(m0)
```

```
##           w0 feed1:dayf26 feed2:dayf26 feed3:dayf26 feed4:dayf26
##    0.9416031    1.5671917    1.2115047    1.7253760    0.3301688
## feed1:dayf45 feed2:dayf45 feed3:dayf45 feed4:dayf45 feed1:dayf61
##    1.9814774    1.1543619    1.9110903    0.2444546    2.0814774
## feed2:dayf61 feed3:dayf61 feed4:dayf61 feed1:dayf91 feed2:dayf91
##    1.4400761    2.0968046    0.1587403    2.5529060    1.6257904
## feed3:dayf91 feed4:dayf91
##    2.6539474    0.1158831
```

```
Variance    StdDev
(Intercept) 0.13882677 0.3725947
Residual    0.05653912 0.2377796
```

Variansstruktur i RI-modellen

Variansstruktur i Random Intercepts modellen:

- $\text{Var}\, Y_i = \nu^2 + \sigma^2$
- Y_i og Y_j er uafhængige hvis $\texttt{goat}_i \neq \texttt{goat}_j$.
- Hvis $\texttt{goat}_i = \texttt{goat}_j$, såer

$$\text{Cov}(Y_i, Y_j) = \nu^2, \qquad \text{Cor}(Y_i, Y_j) = \frac{\nu^2}{\sigma^2 + \nu^2}$$

Altså: korrelationen er den samme for alle par af observationer fra samme ged, "ligner hinanden lige meget" uanset tidsafstanden.

Er dette mon en rimelig antagelse?

Torsdag: modeller hvor dette ikke er tilfældet.

Opsummering

Gentagne målinger:

Flere målinger påhver forsøgsenhed.

Forslag til analyser:

- Analyse af summary measure(s)
  - Ofte god og robust analyse
  - Udnytter ikke alle data
- Model hvor individ indgår med tilfældig virkning
  - Antager at korrelationen er ens for alle par af obs. inden for individ
  - Kan være urimeligt, især hvis der er mange observationer per individ