Gentagne målinger
Statistisk Dataanalyse 2

Anders Tolver

Uge 5, torsdag d. 5/10-2017

Program

Analyse af samtlige data fra eksempel 10.1 (vækst af geder):

- Analyse af summary measures (tirsdag)
- Simple tilgange baseret på kapitel 1-6 i kompendiet
- Random Intercepts modellen
- Model med seriel korrelationsstruktur (Diggle-modellen)
  - model og modelkontrol
  - modelreduktion
  - konklusion, herunder parameterestimater
- Sammenligning af de to modeller

Eksempel 10.1: geders vægtudvikling

Interesseret i fire fodertypers effekt på geders vægtudvikling.

Faktorer:

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

På de følgende slides betragtes 4 forskellige måder, hvorpå man kan lave en statistisk analyse af hele datasættet.

Vi repeterer 3 metoder fra tidligere forelæsninger og introducerer en ny model specielt velegnet til analyse af gentagne målinger.

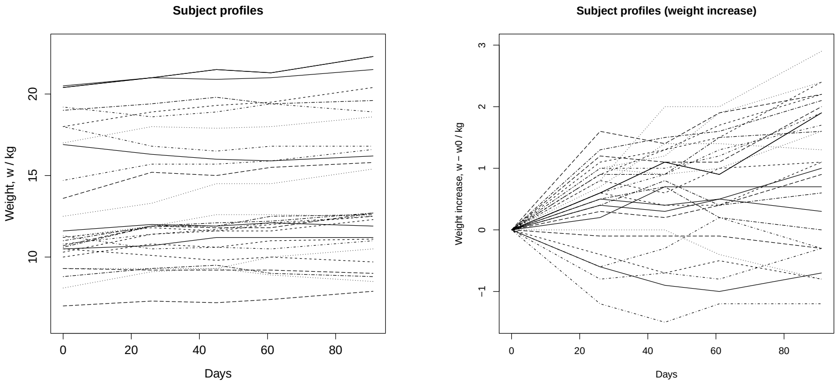
Forhåbentlig giver øvelsen jer et bedre overblik over kursets indhold.

Eksempel 10.1: geders vægtudvikling

```
goatdata = read.table("../data/goats1.txt", header=T)
goatdata$feedfac = factor(goatdata$feed)
goatdata$goatfac = factor(goatdata$goat)
goatdata$dayfac = factor(goatdata$day)
goatdata
```

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

Gentagne målinger: plot profiler



Eksempel 10.1: tosidet ANOVA

Statistisk model

$$Y_i = \gamma(\texttt{feed}_i, \texttt{day}_i) + e_i,$$

hvor $e_1, \ldots e_{112}$ er uafhængige $\sim N(0, \sigma^2)$.

I hvilken rækkefølge bør vi foretage reduktion i modellen?

Slutmodellen bliver

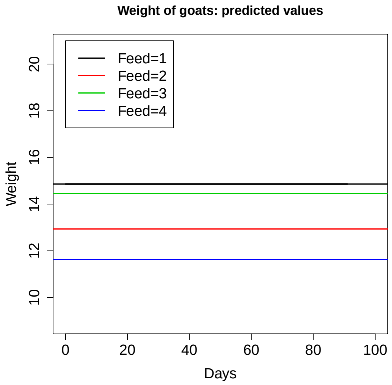
$$Y_i = \alpha(\texttt{feed}_i) + e_i, \quad e_i \sim N(0, \sigma^2).$$

Parameterestimater

```
model4<-lm(weight~feedfac-1)
summary(model4)
```

```
         Estimate Std. Error t value Pr(>|t|)
feedfac1  15.1071     0.7086   21.32   <2e-16 ***
feedfac2  13.0607     0.7086   18.43   <2e-16 ***
feedfac3  14.7143     0.7086   20.77   <2e-16 ***
feedfac4  11.5250     0.7086   16.27   <2e-16 ***

Residual standard error: 3.749 on 108 degrees of freedom
```

Eksempel 10.1: tosidet ANOVA



- Hvorfor kan vi ikke se en signifikant effekt af day?

Eksempel 10.1: inddrag baseline måling

Statistisk model

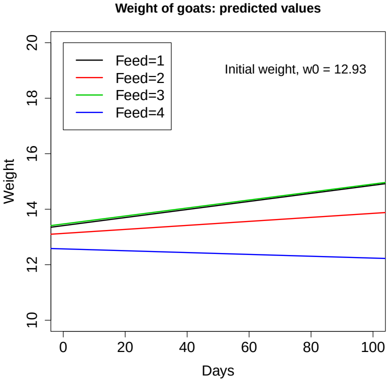
$$Y_i = \gamma(\texttt{feed}_i, \texttt{day}_i) + \delta \cdot w_{0,i} + e_i, \quad e_i \sim N(0, \sigma^2).$$

Slutmodellen bliver

$$Y_i = \alpha(\texttt{feed}_i) + \beta(\texttt{feed}_i) \cdot \texttt{day}_i + \delta \cdot w_{0,i} + e_i, \quad e_i \sim N(0, \sigma^2).$$

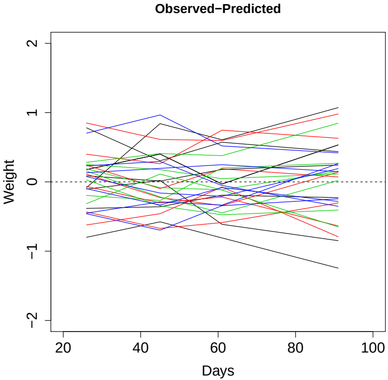
```
              Estimate Std. Error t value Pr(>|t|)
w0            0.941603   0.011338  83.049  < 2e-16 ***
feedfac1      1.235744   0.273198   4.523 1.63e-05 ***
feedfac2      0.954853   0.264119   3.615 0.000466 ***
feedfac3      1.294326   0.270156   4.791 5.59e-06 ***
feedfac4      0.397236   0.261643   1.518 0.132019
feedfac1:day  0.014529   0.003685   3.943 0.000147 ***
feedfac2:day  0.007230   0.003685   1.962 0.052428 .
feedfac3:day  0.014394   0.003685   3.907 0.000168 ***
feedfac4:day -0.003317   0.003685  -0.900 0.370088

Residual standard error: 0.4645 on 103 degrees of freedom
```

Eksempel 10.1: inddrag baselinemåling



- Når vi inddrager baselinemålingen (`w0`) kan vi "se" effekten af `day`.
- NB. Prædiktion dur ikke nødvendigvis (langt) udenfor 26–91 dage.

Eksempel 10.1: inddrag baselinemåling



- Giver modellen en tilfredsstillende beskrivelse af data? Hvorfor/hvorfor ikke?
- Hvad kan vi gøre for at forbedre modellen?

Eksempel 10.1: random intercepts model

Statistisk model

$$Y_i = \gamma(\texttt{feed}_i, \texttt{day}_i) + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + e_i$$

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

Udover effekten af de systematiske variable (`feed, day, w0`) tilføjes et (tilfældigt) niveau for hver ged.

Modellen fittes i R vha. `lme`:

```
lme(weight ~ w0 + feedfac*dayfac, random =~ 1 | goat)
```

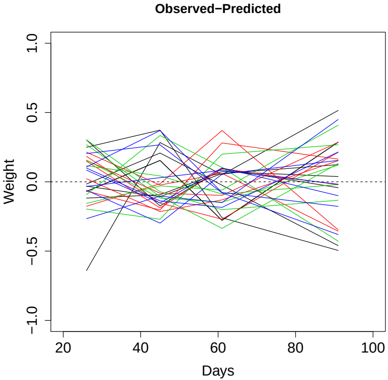
Modelreduktion på sædvanlig vis.

Slutmodel

$$Y_i = \alpha(\texttt{feed}_i) + \beta(\texttt{feed}_i) \cdot \texttt{day}_i + \delta \cdot w_{0,i} + A(\texttt{goat}_\texttt{i}) + e_i,$$

hvor $e_i \sim N(0, \sigma^2), A(j) \sim N(0, \nu^2)$.

Eksempel 10.1: random intercepts model



- Residual profilkurverne ligger nu mere omkring 0.
- Giver modellen en tilfredsstillende beskrivelse af data?
- Hvad kan vi gøre for at forbedre modellen?

Eksempel 10.1: random intercepts model

Variansstruktur i Random Intercepts modellen

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

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

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?

Dette kan man (måske) få en fornemmelse af, ved at se på residual profilkurverne.

Eksempel 10.1: Diggle-modellen

Måske mere rimeligt at antage at korrelationen mellem par af obs. afhænger af tidsforkellen mellem observationerne.

Vi taler om modeller med seriel korrelationsstruktur.

Diggle-modellen: For to observationer fra samme ged, til tidspunkterne t_i og t_j ($t_i \neq t_j$):

$$\mathrm{Var}\, Y_i = \nu^2 + \tau^2 + \sigma^2$$

$$\mathrm{Cor}(Y_i, Y_j) = \frac{\nu^2 + \tau^2 \cdot \exp\left(-(t_i - t_j)^2/\phi^2\right)}{\nu^2 + \tau^2 + \sigma^2}$$

- Korrelationerne aftager når $|t_i - t_j|$ vokser!
- $\mathrm{Cor}(Y_i, Y_j) \approx \frac{\nu^2+\tau^2}{\nu^2+\tau^2+\sigma^2}$ for to obs. meget tæt på hinanden i tid
- $\mathrm{Cor}(Y_i, Y_j) \approx \frac{\nu^2}{\nu^2+\tau^2+\sigma^2}$ for to obs. meget langt fra hin. i tid

Eksempel 10.1: Diggle-modellen (2)

Model:

$$Y_i = \gamma(\texttt{feed}_i, \texttt{time}_i) + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + D_i + e_i$$

hvor A'er, D'er og e'er er således at variansstrukturen er som på den foregående slide.

Mere præcist i noterne, men det vigtige er:

- Den systematiske del — middelværdien — specificeres på sædvanlig vis
- Vi ønsker en bestemt type variansstruktur — aftagende korrelationer når tidsafstanden øges
- Hvorfor bruger vi energi på variansstrukturen?

Eksempel 10.1: Diggle-modellen i R

R skal vide:

- den systematiske del — på sædvanlig vis
- at goat er tilfældig — på sædvanlig vis
- korrelationsfunktionen — vha. `corr`
- at der er "målefejl" dvs. e_i'er — vha. `nugget`

Kommando:

```
mod0<-lme(weight~w0+feedfac*dayfac
,random=~1|goat,method="ML"
,corr = corGaus(form = ~ day | goat, nugget=T))
```

Eksempel 10.1: modelkontrol

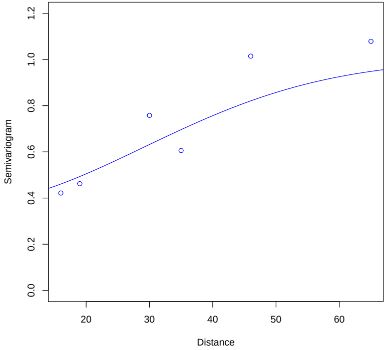
Som altid vigtigt at kontrollere modellen:

- Residualplot: `plot(mod0)` eller analog model med syst. effekter
- QQ-plot: `qqnorm(mod0)` eller analog model med syst. effekter
- Semi-variogram: `plot(Variogram(mod0))`

Semi-variogrammet bruges til at vurdere om den valgte korrelationsstruktur er rimelig for data:

- Sammenligner den empiriske korrelationsfunktion ("bestemt alene fra data") med den modelbaserede.
- Vurdér om der er en rimelig overensstemmelse!

Eksempel 10.1: Diggle-model - semi-variogram



Eksempel 10.1: Diggle-model - reduktion

Reduktion i den systematiske del af modellen udføres på sædvanligvis vis med likelihood ratio test.

Hypotesen om ingen vekselvirkning mellem faktorerne `feed` og `day`, dvs. `mod1`, forkastes ($LR = 35.8, p < 0.0001$).

Dette er ret typisk: behandlingseffekt ændrer sig ofte over tid.

Hvorfor virker dette også meget rimeligt her?

Hvilke andre hypoteser kunne være interessante?

Eksempel 10.1: Diggle-model - reduktion (2)

Diverse modeller (navne ref. til R-program):

- $mod0: Y_i = \gamma(\texttt{feed}_i, \texttt{day}_i) + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + D_i + e_i$
- $mod1: Y_i = \alpha(\texttt{feed}_i) + \eta(\texttt{day}_i) + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + D_i + e_i$
- $mod2: Y_i =$ $\alpha(\texttt{feed}_i) + \beta(\texttt{feed}_i) \cdot \texttt{day}_i + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + D_i + e_i$
- $mod3: Y_i = \alpha(\texttt{feed}_i) + \beta \cdot \texttt{day}_i + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + D_i + e_i$
- $mod4: Y_i = \alpha + \beta(\texttt{feed}_i) \cdot \texttt{day}_i + \delta \cdot w_{0,i} + A(\texttt{goat}_i) + D_i + e_i$

Test	LR	df	p-værdi
$mod0 \rightarrow mod1$	35.8	9	< 0.0001
$mod0 \rightarrow mod2$	8.9	8	0.35
$mod2 \rightarrow mod3$	27.9	3	< 0.0001
$mod2 \rightarrow mod4$	10.8	3	< 0.013

Eksempel 10.1: konklusion på Diggle-analyse

Slutmodellen (`mod2`) genfittes så vi opnår REML-estimater:

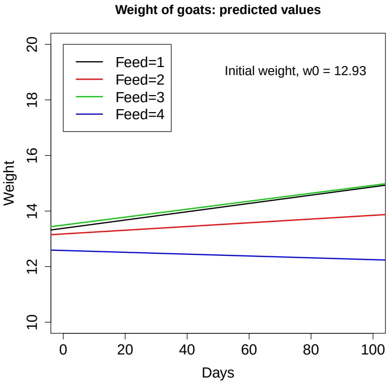
```
 Fixed effects: weight ~ w0 + feedfac + feedfac:day - 1
                    Value Std.Error DF  t-value
w0              0.9431925 0.0218936 23 43.08080
feedfac1        1.1824714 0.3759525 23  3.14527
feedfac2        0.9797415 0.3509290 23  2.79185
feedfac3        1.2983880 0.3676654 23  3.53144
feedfac4        0.3856033 0.3439438 23  1.12112
feedfac1:day    0.0149209 0.0024531 81  6.08256
feedfac2:day    0.0067006 0.0024531 81  2.73150
feedfac3:day    0.0143125 0.0024531 81  5.83455
feedfac4:day   -0.0033108 0.0024531 81 -1.34965
```

Eksempel 10.1: konklusion på Diggle-analyse

Parameterestimater for systematisk del:

- Der er med stor sikkerhed påvist forskel på behandlingerne.
- Estimeret vægt for fodertype 2 til tid 91 for ged der vejer 12.93 kg fra start (-gnsn. blandt gederne)?
- Grafisk præsentation af forventede værdier
- Er der mon forskel på type 1 og type 3? Hvad er den relevante hypotese? Hvordan kan vi formelt teste hypotesen?
- Sammenlign eventuelt med resultaterne fra summary analyse, fx. opgave 6.1.

Eksempel 10.1: konklusion på Diggle-analyse



NB. Prædiktionen dur ikke nødvendigvis (langt) udenfor 26–91 dage.

Eksempel 10.1: Diggle-model - variansestimater

REML-estimater for den tilfældige del af modellen:

```
Random effects:
 Formula: ~1 | goat
        (Intercept)  Residual
StdDev:   0.3962525 0.3069467

Correlation Structure: Gaussian spatial correlation
 Formula: ~day | goat
 Parameter estimate(s):
    range     nugget
41.0848916  0.3723881
```

Eksempel 10.1: Diggle model - variansestimater (2)

Variansparametre: σ^2, ν^2, τ^2, ϕ.

REML-estimater fra `mod2` (fra `summary`):

$$\begin{array}{ll} \texttt{Intercept}: & \hat{\nu}^2 = 0.3962^2 = 0.1570 \\ \texttt{Residual}: & \hat{\sigma}^2 + \hat{\tau}^2 = 0.3069^2 = 0.0942 \\ \texttt{range}: & \hat{\phi} = 41.08 \\ \texttt{nugget}: & \frac{\hat{\sigma}^2}{\hat{\tau}^2 + \hat{\sigma}^2} = 0.3724 \end{array}$$

Estimaterne for $\hat{\sigma}^2$ og $\hat{\tau}^2$ fås således:

$$\hat{\sigma}^2 = 0.3724 \cdot 0.0942 = 0.0351$$

$$\hat{\tau}^2 = 0.0942 - 0.0351 = 0.0591$$

Sammenligning af korrelationsstrukturer

Vi er sådan set ikke interesseret i korrelationsstrukturen.

Men den er vigtig for at få valide resultater for den systematiske del (p-værdier, konfidensintervaller).

Derfor interesseret i at sammenligne forskellige modeller for den tilfældige del, fx. Diggle-modellen og Random Intercepts modellen.

Vanskeligt/umuligt at udføre egentlige test. Sammenligner i stedet ofte AIC-værdier:

- $AIC = -2\log L + 2 \cdot$ antal parametre i model
- log-likelihoodfunktionen $\log L$ måler "hvor godt modellen passer til data" (stor $\sim$ passer godt)
- AIC "straffer" modeller med mange parametre
- Foretræk model med lille AIC-værdi

Sammenlign. af korrelationsstrukturer (2)

Tænkt på gededata i eksempel 10.1:

Sammenligner Diggle og RI for startmodellen med vekselvirkning mellem `dayfac` og `feedfac`:

- *AIC* for RI: 143.6
- *AIC* for Diggle: 140.8

AIC er mindst for Diggle (men ikke meget), så Diggle-modellen foretrækkes.

Begge værdier er taget fra estimation med `method='REML'`.

Der findes mange andre variansstrukturer end RI og Diggle: `lme` opererer med mindst 10!

Gentagne målinger: opsummering

En analyse kunne bestå af følgende:

- Figurer med individueller profiler hhv. gennemsnitsprofiler
- Analyse af et velvalgt summary measure (eller et par stykker)
  - vælg responsen med omhu
  - god, simpel, robust analysemetode
  - udnytter dog ikke alle data

Gentagne målinger: opsummering (2)

- Analyse af model med seriel korrelationsstruktur
  - fastlæg variansstruktur vha. semi-variogrammer og sammenligning af AIC-værdier; husk også residualplot
  - reduktion af den systematiske del af modellen, herunder test for vekselvirkning, test for linearitet (hvis relevant!), etc.
  - relevante estimater og konfidensintervaller
  - evt. figurer der viser forventet udvikling over tid
- Samlet konklusion: sammenligning af resultater fra summary analyse og analyse for gentagne målinger

Analyse af gentagne målinger er ikke nemt, men meget nyttigt!