SD2 - uge 5, torsdag

Anne Petersen

Vi starter med at sætte working directory:

```
setwd("C:/Users/Anne/Dropbox/Arbejde/STATforLIFE2/uge5")
```

6.2

Vi indlæser data, betragter det og gemmer passende variable som faktorer:

```
data <- read.table("bms_opg10_1.txt", header=T)
head(data)
```

```
##   dose animal week wgt
## 1    0      1    1 455
## 2    0      1    3 460
## 3    0      1    4 510
## 4    0      1    5 504
## 5    0      1    6 436
## 6    0      1    7 466
```

```
data$dose <- factor(data$dose)
data$animal <- factor(data$animal)
data$weekF <- factor(data$week)
```

Vi vil nu opskrive en diggle-model, så vi tillader større korrelation mellem observationer fra det samme individ, som er tidsmæssigt tættere på hinanden, end observationer som er længere fra hinanden. Vi ønsker desuden at inddrage en tilfældig effekt af individet og en systematisk vekselvirkningseffekt mellem `dose` og `week`. Vi skriver modellen formelt således:

$$Y_i = \alpha(\text{dose}_i \times \text{week}_i) + A(\text{animal}_i) + D_i + e_i$$

for $i = 1...90$ og hvor vi antager at

1. $A(1), ..., A(15)$ er iid med $A(1) \sim N(0, \sigma_A^2)$
2. $e_1, ..., e_{90}$ er iid med $e_1 \sim N(0, \sigma^2)$
3. $D_1, ..., D_{90}$ er specificeret som i diggle-modellen i kompendiets kapitel 10.

Vi fitter modellen i R:

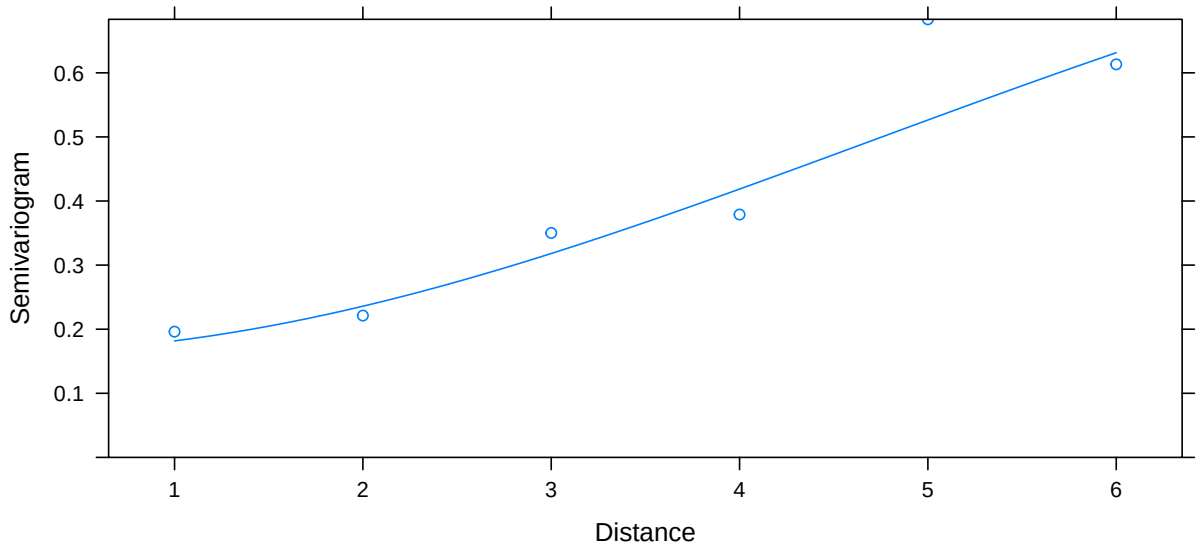
```
library(nlme)
```

```
## Warning: package 'nlme' was built under R version 3.1.3
```

```
model <- lme(wgt ~ dose*weekF, random=~1|animal, method="ML",
             cor=corGaus(form=~week|animal, nugget=T),
             data)
```

og laver et semivariogram for at undersøge, om den antagede kovariansstruktur virker rimelig:

```
plot(Variogram(model, form=~week))
```



Den indtegnede linje viser semivariogram-funktionen som specificeret i diggle-antagelsen og punkterne viser den empriske semivariogram-værdi for hver af de 6 mulige tidsafstande i vores data (altså hhv. 1, 2, 3, 4, 5 og 6 ugers forskel). Vi ser at punkterne ligger pænt omkring den teoretiske semivariogramkurve og konkluderer derfor, at den antagede korrelationsstruktur er rimelig.

Vi estimerer nu modellens variansparametre. Husk først at genfitte modellen med REML-estimation.

```
model_REML <- lme(wgt ~dose*weekF, random=~1|animal, method="REML",
                  cor=corGaus(form=~week|animal, nugget=T),
                  data)
summary(model_REML)
```

```
## Linear mixed-effects model fit by REML
##  Data: data
##        AIC      BIC    logLik
##   746.6565 796.7432 -351.3283
##
## Random effects:
##  Formula: ~1 | animal
##         (Intercept) Residual
## StdDev:    12.84199 41.28533
##
## Correlation Structure: Gaussian spatial correlation
##  Formula: ~week | animal
##  Parameter estimate(s):
```

```
##    range   nugget
## 6.625625 0.162915
## Fixed effects: wgt ~ dose * weekF
##              Value Std.Error DF   t-value p-value
## (Intercept)  466.4  19.33595 60 24.120872  0.0000
## dose1         28.0  27.34516 12  1.023947  0.3261
## dose2         31.4  27.34516 12  1.148283  0.2732
## weekF3        53.0  12.67983 60  4.179868  0.0001
## weekF4       102.4  14.72636 60  6.953518  0.0000
## weekF5        95.2  16.89357 60  5.635280  0.0000
## weekF6        80.2  18.94391 60  4.233551  0.0001
## weekF7       105.6  20.74719 60  5.089846  0.0000
## dose1:weekF3   3.6  17.93198 60  0.200759  0.8416
## dose2:weekF3 -16.2  17.93198 60 -0.903414  0.3699
## dose1:weekF4 -22.6  20.82622 60 -1.085171  0.2822
## dose2:weekF4 -20.4  20.82622 60 -0.979535  0.3312
## dose1:weekF5 -22.6  23.89112 60 -0.945958  0.3480
## dose2:weekF5 -21.2  23.89112 60 -0.887359  0.3784
## dose1:weekF6  28.4  26.79073 60  1.060068  0.2934
## dose2:weekF6  10.2  26.79073 60  0.380729  0.7047
## dose1:weekF7  44.0  29.34096 60  1.499610  0.1390
## dose2:weekF7  19.8  29.34096 60  0.674825  0.5024
##  Correlation:
##              (Intr) dose1  dose2  weekF3 weekF4 weekF5 weekF6 weekF7
## dose1        -0.707
## dose2        -0.707  0.500
## weekF3       -0.328  0.232  0.232
## weekF4       -0.381  0.269  0.269  0.679
## weekF5       -0.437  0.309  0.309  0.666  0.760
## weekF6       -0.490  0.346  0.346  0.630  0.744  0.813
## weekF7       -0.536  0.379  0.379  0.581  0.704  0.792  0.846
## dose1:weekF3  0.232 -0.328 -0.164 -0.707 -0.480 -0.471 -0.446 -0.411
## dose2:weekF3  0.232 -0.164 -0.328 -0.707 -0.480 -0.471 -0.446 -0.411
## dose1:weekF4  0.269 -0.381 -0.190 -0.480 -0.707 -0.538 -0.526 -0.498
## dose2:weekF4  0.269 -0.190 -0.381 -0.480 -0.707 -0.538 -0.526 -0.498
## dose1:weekF5  0.309 -0.437 -0.218 -0.471 -0.538 -0.707 -0.575 -0.560
## dose2:weekF5  0.309 -0.218 -0.437 -0.471 -0.538 -0.707 -0.575 -0.560
## dose1:weekF6  0.346 -0.490 -0.245 -0.446 -0.526 -0.575 -0.707 -0.599
## dose2:weekF6  0.346 -0.245 -0.490 -0.446 -0.526 -0.575 -0.707 -0.599
## dose1:weekF7  0.379 -0.536 -0.268 -0.411 -0.498 -0.560 -0.599 -0.707
## dose2:weekF7  0.379 -0.268 -0.536 -0.411 -0.498 -0.560 -0.599 -0.707
##              ds1:F3 ds2:F3 ds1:F4 ds2:F4 ds1:F5 ds2:F5 ds1:F6 ds2:F6
## dose1
## dose2
## weekF3
## weekF4
## weekF5
## weekF6
## weekF7
## dose1:weekF3
## dose2:weekF3  0.500
## dose1:weekF4  0.679  0.340
## dose2:weekF4  0.340  0.679  0.500
## dose1:weekF5  0.666  0.333  0.760  0.380
```

```
## dose2:weekF5  0.333  0.666  0.380  0.760  0.500
## dose1:weekF6  0.630  0.315  0.744  0.372  0.813  0.406
## dose2:weekF6  0.315  0.630  0.372  0.744  0.406  0.813  0.500
## dose1:weekF7  0.581  0.291  0.704  0.352  0.792  0.396  0.846  0.423
## dose2:weekF7  0.291  0.581  0.352  0.704  0.396  0.792  0.423  0.846
##              ds1:F7
## dose1
## dose2
## weekF3
## weekF4
## weekF5
## weekF6
## weekF7
## dose1:weekF3
## dose2:weekF3
## dose1:weekF4
## dose2:weekF4
## dose1:weekF5
## dose2:weekF5
## dose1:weekF6
## dose2:weekF6
## dose1:weekF7
## dose2:weekF7  0.500
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.4885646 -0.5342296  0.1012641  0.6656561  1.5610434
##
## Number of Observations: 90
## Number of Groups: 15
```

Vi ser at

- $\hat{\sigma}_A = 12.84$ ('Random effects' ... '(Intercept)')
- $s^2 + \hat{\tau}^2 = 41.29^2$ ('Random effects'... 'Residual')
- $\hat{\phi} = 6.626$ ('Correlation structure' ... 'range')
- $\frac{s^2}{s^2+\hat{\tau}^2} = 0.1629$ ('Correlation structure' ... 'nugget')

hvor τ og ϕ er parametrene for digglemodellen, som specificeret i kompendiet s. 190. Ved at regne lidt finder vi dermed at

$$s^2 = 41.29^2 \cdot 0.1629 = 277.72$$

$$\hat{\tau}^2 = 41.29^2 - 277.72 = 1427.14$$

og således er alle variansparameterestimater fundet.

Vi forsøger nu at reducere den systematiske del af modellen. Vi starter med at undersøge, om vi kan fjerne vekselvirkningsleddet:

```
model1 <- lme(wgt ~ dose+weekF, random=~1|animal, method="ML",
              cor=corGaus(form=~week|animal, nugget=T),
              data)
anova(model, model1)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model      1 22 866.0254 921.0212 -411.0127
## model1     2 12 869.2191 899.2168 -422.6096 1 vs 2 23.19375  0.0101
```

Vi finder $p = 0.0101$ og kan altså ikke reducere til den additive model. En grundig studerende kunne evt. prøve at simulere denne p-værdi for at være helt sikker på, at vi ikke underestimerer den. Vi vil dog blot gå videre til at teste den anden hypotese, som er mulig i denne model; nemlig hvorvidt modellen kan reduceres ved at lade `week` indgå lineært som en kovariat:

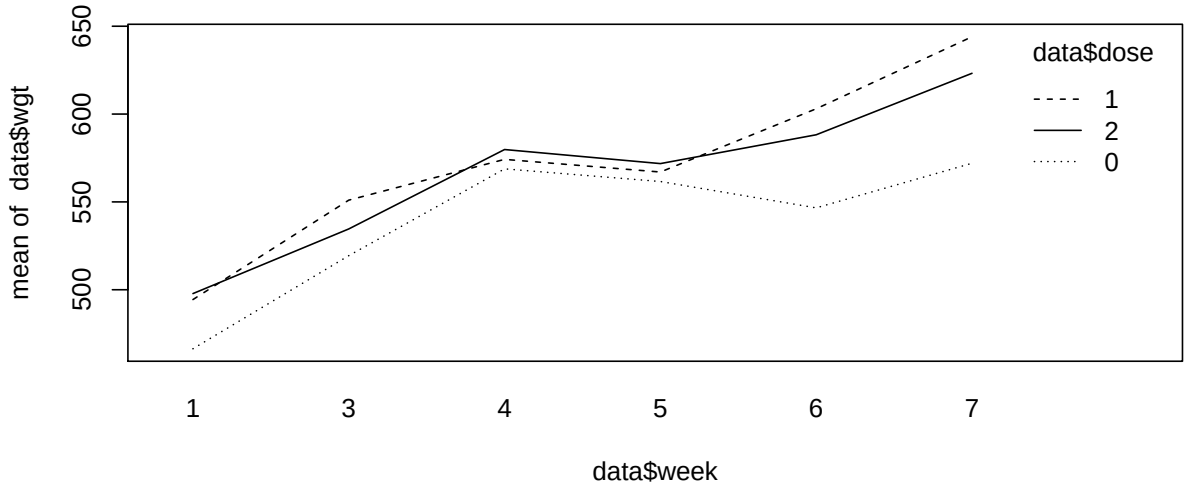
```
model2 <- lme(wgt ~ dose*week, random=~1|animal, method="ML",
              cor=corGaus(form=~week|animal, nugget=T),
              data)
anova(model, model2)
```

```
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model      1 22 866.0254 921.0212 -411.0127
## model2     2 10 887.7498 912.7479 -433.8749 1 vs 2 45.72438  <.0001
```

Vi finder nu en meget lav p-værdi ($p < 0.0001$) og konkluderer altså, at reduktionen ikke kan tillades og at vores slutmodel (for den systematiske del) dermed er modellen med vekselvirkning og hvor `week` indgår som faktorvariabel.

Vi bruger et interaktionsplot til at vise gennemsnitsprofilerne for de forskellige kombinationer af uge og dose:

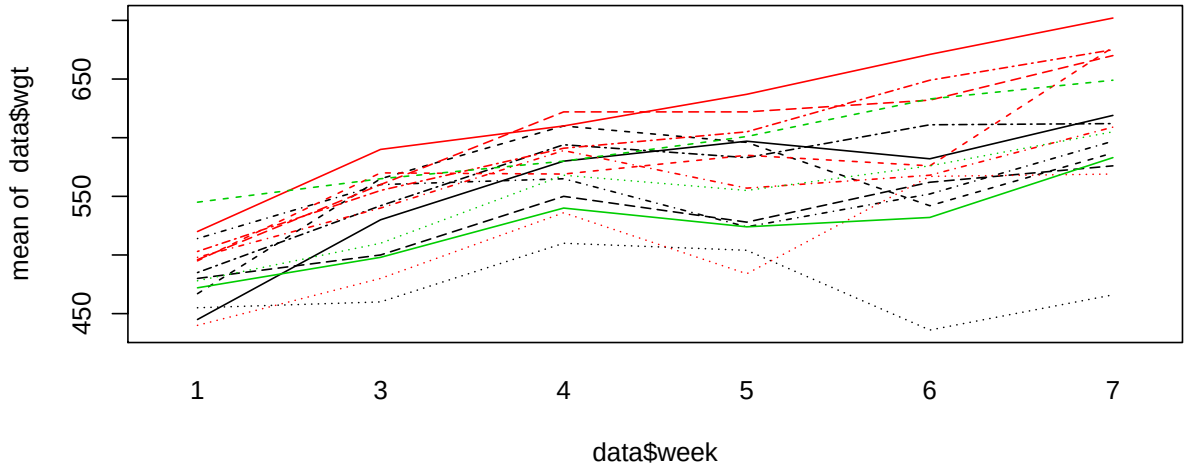
```
interaction.plot(data$week, data$dose, data$wgt)
```



Vi ser at linjerne ser ud til at være parallelle indtil uge 5, hvilket indikerer at der her ingen vekselvirkning er, men at de ikke længere er parallelle efter uge 5, hvilket svarer til konklusionen fra ovenfor, om at en vekselvirkningseffekt er nødvendig. Dette stemmer godt overens med oplysningen i opgaveteksten om, at marsvinene har først har fået behandlingen med de forskellige doser i uge 5. Husk, at man ikke umiddelbart kan vurdere linearitetsantagelsen, da der ikke er lige lang tid mellem hvert observationstidspunkt.

Vi laver nu et interaktionsplot for de individuelle profiler, dvs. med en linje for hvert marsvin:

```
interaction.plot(data$week, data$animal, data$wgt,
                 col=data$animal, legend=F)
```



Vi ser igen større spredning efter uge 5. Samlet set ser det altså ud til, at ovenstående model er problematisk. Problemet er, at vi antager, at marsvinenes vækstforløb afhænger af `dose` i alle uge - også uge 1 til 5, hvor de endnu ikke har modtaget dosen. Denne antagelse er simpelhent forkert. Der er forskellige måder at løse dette problem på; fx. kunne vi modellere med én baselineværdi (svarende til enten uge 1, 2, 3 eller 4) og så modellere de sidste tre målinger pr. marsvin som før. Vi kan også prøve at finde et passende summary-measure - og det er den strategi, vi nu vil prøve at følge.

Vi opstiller en model, hvor vores respons er tilvæksten i vægt mellem uge 1 og uge 7, dvs.

$$Y_i = \text{vægt for marsvin i, uge 7} - \text{vægt for marsvin i, uge 1}$$

og vi opstiller følgende model:

$$Y_i = \alpha(\text{dose}_i) + e_i$$

med notation og antagelser som fra ovenfor. Bemærk, at vi ikke længere har flere observationer over tid pr. marsvin og at vi derfor ikke længere kan (eller bør) modellere seriel korrelationsstruktur, fx. vha. en diggle-model og at vi af samme årsag udelader den tilfældige effekt på individniveau. Vi vil gerne fitte denne model i R. Først må vi dog konstruere den nye responsvariabel (se kommentarer til koden nedenfor):

```
data <- data[order(data$animal),] #sorter
data_week1 <- subset(data, week==1) #observationer til week=1
data_week7 <- subset(data, week==7) #observationer til week=4
growth <- data_week7$wgt - data_week1$wgt #ny respons
data_sum <- data.frame(dose=data_week1$dose,
                       animal=data_week1$animal,
                       growth=growth) #nyt datasæt
```

Kommentar til koden: Vi starter med at sortere datasættet efter `animal`. Bemærk, at det allerede er sorteret, men da vores metode kun fungerer, for data som er sorteret (ellers trækker vi ikke de rigtige tal fra hinanden), står det i koden for god ordens skyld. Vi piller derefter delmængder af datasættet ud svarende til hhv. uge 1 og uge 7, trækker `wgt`-værdierne fra hinanden og samler al vores information i et nyt datasæt, `data_sum`, vha.

`data.frame()`. Bemærk, at vi i dette datasæt lige så godt kunne have brugt informationen fra `data_week7` til at konstruere variablene `dose` og `animal` - der ville have stået de samme tal.

Vi fitter nu modellen med den nye responsvariabel:

```
modelA <- lm(growth ~ dose, data=data_sum)
```

og vi tester for effekt af `dose`:

```
modelB <- lm(growth ~ 1, data=data_sum)
anova(modelB, modelA)
```

```
## Analysis of Variance Table
##
## Model 1: growth ~ 1
## Model 2: growth ~ dose
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     14 29940
## 2     12 25084  2    4856.1 1.1616 0.3458
```

Vi finder $p = 0.3458$ og altså er der ikke en signifikant effekt af `dose` i denne model og vi kan dermed reducere til `modelB`. Bemærk, at dette nok er den mest meningsfulde model, vi har opstillet i ovenstående: Vi ser at væksten mellem et baseline-tidspunkt (uge 1) og et tidspunkt efter behandlingen er givet (uge 7) ikke afhænger af signifkant af behandlingen. Som nævnt ovenfor, kan man naturligvis dykke dybere ned i dette spørgsmål ved at overveje, om man kan finde på en måde at analysere data, hvor man udnytter mere af den information, det indeholder. Fx. siger de første fire målinger pr. marsvin noget om variabiliteten i vækst for marsvin, som endnu ikke er behandlede. Kan denne information måske bruges til at modellere baselineværdien mere meningsfuldt?