Faculty of Health Sciences

Basal statistik

Korrelerede målinger i R

Lene Theil Skovgaard

9. november 2020

Korrelerede målinger

med visse uløste problemer

- Tilfældige effekter
- Varianskomponentmodeller
- Modeller for longitudinelle målinger
- Korrelationsstrukturer
- Baseline problematik

E-mail: ltsk@sund.ku.dk

$*$: Siden er lidt teknisk

En gennemgående antagelse hidtil

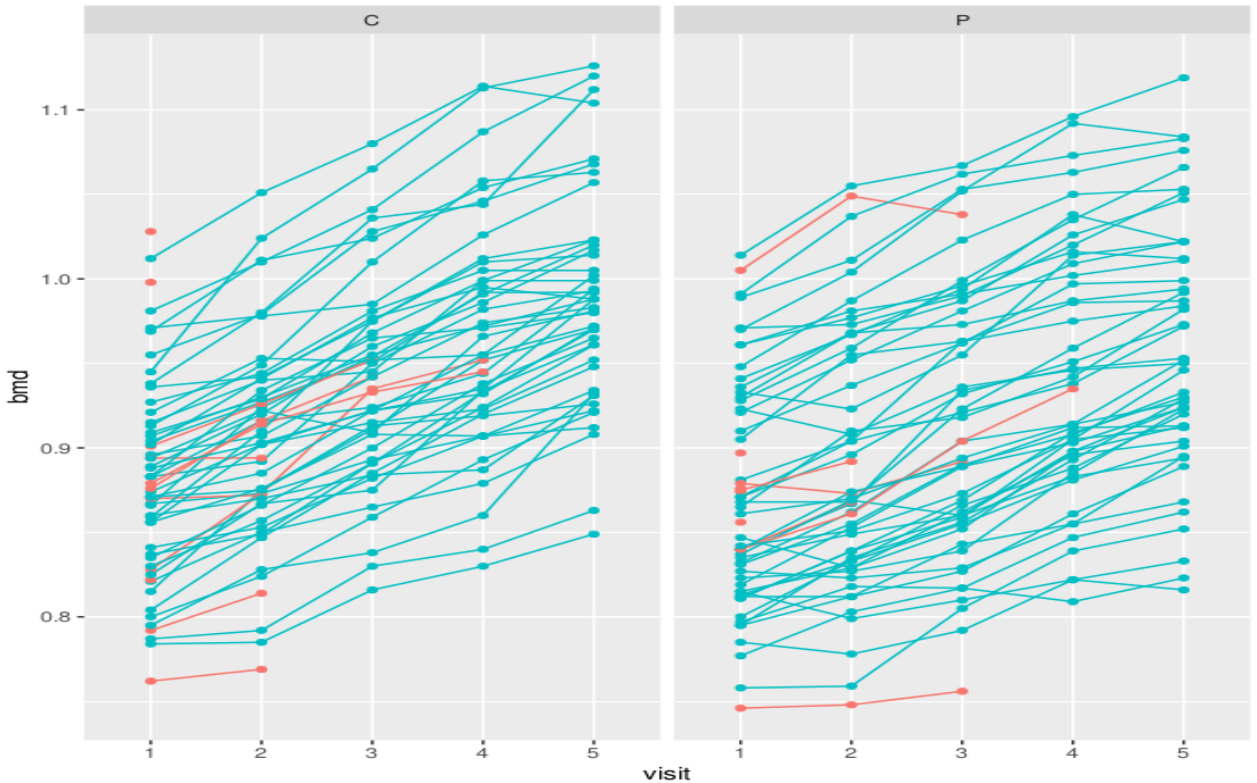
Uafhængighed

- Kun en enkelt observation for hvert individ (unit)
  (bortset fra den parrede situation, med
  to målemetoder eller før/efter undersøgelser)
- Ingen tvillinger/søskende ...

men i dag har vi flere end to for hvert individ, og vi må forvente, at observationer fra samme individ ser *mere ens* ud end observationer fra forskellige individer, de er korrelerede.

Eksempel: Calcium tilskud

Spaghettiplot: Individuelle profiler (lidt vanskelig kode, se s.



Eksempel: Calcium tilskud

til styrkelse af knogledannelse hos unge piger

I alt 112 11-årige piger randomiseres til at få 'tilskud' i form af enten calcium eller placebo.

Observationer: Y_{git} (gruppe g, pige=individ i, visit=tid t)

Outcome: BMD=bone mineral density, i $\frac{g}{cm^2}$, måles 5 gange over 2 år (hvert halve år).

Det videnskabelige spørgsmål:
Kan et calciumtilskud øge knoglevæksten hos præ-pubertale piger?
og i givet fald - hvor meget?

Datastruktur ved gentagne målinger

Hvert individ bidrager med ligeså mange linier, som vedkommende har observationer

- 5 linier for de fleste piger
- En variabel, der angiver nummeret på (tiden for) målingen

Obs	grp	girl	visit	bmd
1	C	101	1	0.815
2	C	101	2	0.875
3	C	101	3	0.911
4	C	101	4	0.952
5	C	101	5	0.970
6	P	102	1	0.813
7	P	102	2	0.833
8	P	102	3	0.855
9	P	102	4	0.881
10	P	102	5	0.901
11	P	103	1	0.812
.	.	.	.	.

Terminologi for korrelerede målinger

- Multivariate responser (berøres ikke her):
  Forskellige outcomes (responser) på det samme individ, f.eks. en række hormonmålinger, der skal vurderes samlet.
- Cluster design:
  Samme outcome (respons) målt på f.eks. alle individer i en række familier.
- Repeated measurements / Gentagne målinger:
  Samme outcome (respons) målt i forskellige situationer (eller på forskellige steder) på samme individ.
- Longitudinelle målinger:
  Samme outcome (respons) målt gentagne gange over tid for samme individ.

Mixed models for korrelerede målinger

To typer af generaliseringer:

- Varianskomponentmodeller
  Generelle lineære modeller, hvor nogle af effekterne, typisk intercept og evt. hældning (dvs. effekt af tid) er gjort *tilfældige* (dvs. varierer mellem individer)

- Kovariansstrukturer
  Generelle lineære modeller, hvor korrelationen mellem målinger (og evt. forskelle i varians) specificeres direkte.

Begge disse kaldes under et for Mixed Models

Mix af systematiske og tilfældige (*random*) effekter
Hvis korrelationen ignoreres, får man forkerte spredninger,
(for små eller for store,....)

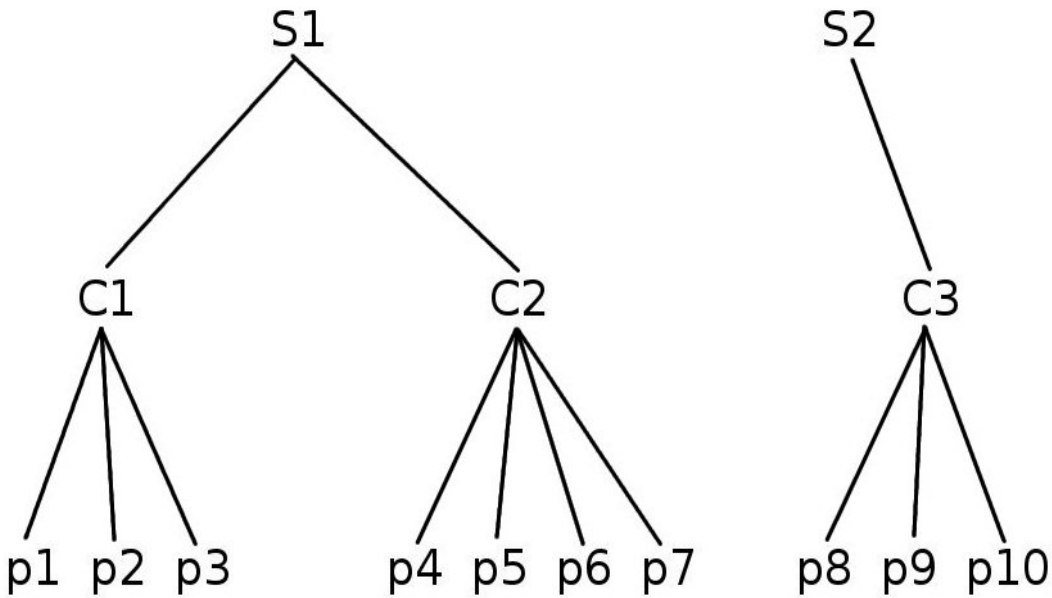
Tilfældige effekter

Variationskilder (varianskomponenter)

- geografisk/miljømæssig variation
  - mellem regioner, hospitaler, skoler eller lande
- biologisk variation
  - variation mellem individer, familier eller dyr
- variation mellem målinger på samme individ (within-individual)
  - variation mellem arme, tænder, injektionssteder, (dage)
- variation, der skyldes ukontrollable forsøgsomstændigheder
  - tidspunkt på dagen, temperatur, observatør
- målefejl/målevariation

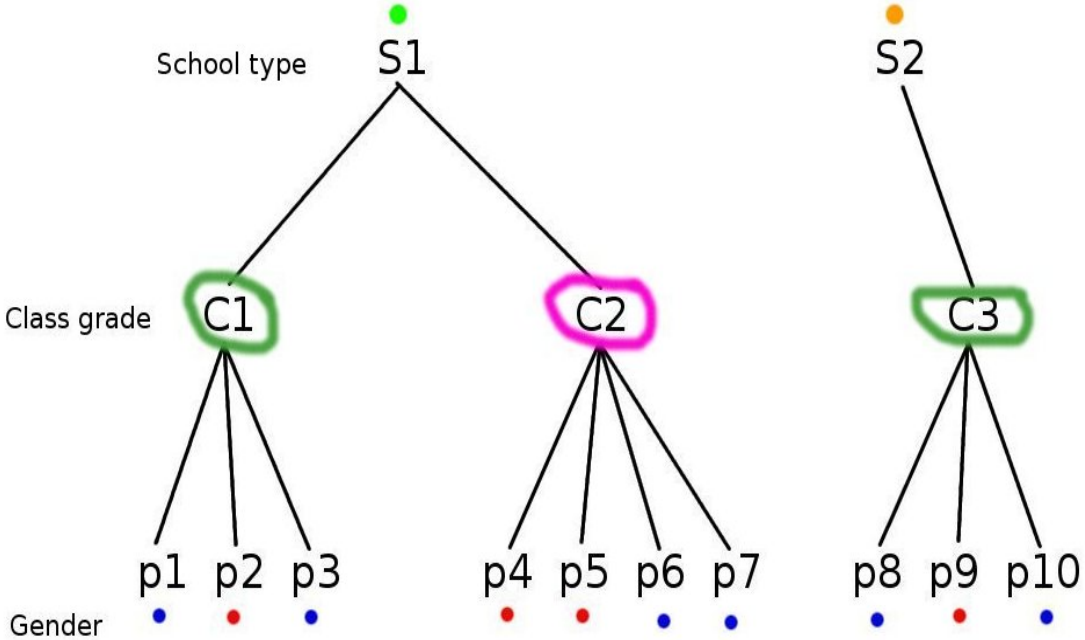
Cluster design (hierarkisk design)

f.eks. skole (School), skole klasse (Class) og elev (Pupil)



$[I] = [S*C*P] \longrightarrow [S*C] \longrightarrow S$

Cluster design, med kovariater



$$[S * C * P] \longrightarrow [S * C] \longrightarrow [S]$$

Køn Klassetrin Skoletype

Et klassisk set-up

Gentagne målinger på samme individ, ofte over tid.

Vi skelner imellem forskellige typer af kovariater:

- Within-individuals kovariat (level 1):
  En kovariat, der varierer *indenfor* person,
  f.eks. tid, dosis, blodtryk
  Disse effekter estimeres svarende til en parret sammenligning.
- Between-individuals kovariat (level 2):
  En kovariat, der *ikke* varierer indenfor person, men kun *mellem* personer, f.eks. behandling, køn, genotype
  Disse effekter estimeres svarende til en uparret sammenligning.

  Individer kunne også være *clustre*, såsom familier, hospitaler, klasser etc.

Hvis man ignorerer korrelationen

mellem målinger på samme individ opstår der fejl

- lav efficiens (type 2 fejl) ved vurdering af level 1 kovariater (within-individual effekter)
- for små standard errors (type 1 fejl) for estimater af level 2 effekter (between-individual effekter)
- *muligvis* bias i middelværdistruktur (i tilfælde af missing values, eller ubalanceret design)

Det er altså vigtigt at medtage alle relevante variationskilder!

Formål med "mixed effects" designs

- At opnå større præcision ved vurdering af effekt af tid, dosis osv., ikke mindst interaktionen mellem tid og evt. behandling (udvikler de to grupper sig forskelligt?)
- At opnå viden om den relative størrelse af de forskellige varianskomponenter, for derved i fremtidige undersøgelser at kunne bruge resourcerne der, hvor de er bedst anbragt, eksempelvis:
  - Hvor ofte skal man måle outcome?
  - Skal man have data på mange børn i hver klasse, eller hellere mange klasser på hver skole?

Simpelt eksempel: Hævelse ved vaccination

Eksperiment:

6 kaniner, hver især vaccineret 6 gange, forskellige steder på ryggen

Outcome y_{rs}: hævelse i cm^2, med notationen
r= $1, \cdots,$**R**=6 angiver kaninen (**rabbit**),
s= $1, \cdots,$**S**=6 angiver stedet (**spot**)

Formål med undersøgelsen:

- Giv et estimat for "*overall mean*"
- Hvor mange gange kan kaninerne *genbruges*, altså hvor mange stik kan det svare sig at udføre pr. kanin?

Vi har i alt 36 målinger af hævelse, men vi må forvente, at hævelse kan være specifikt for den enkelte kanin, således at nogle har tendens til stor hævelse, andre til mindre hævelse.

Data - omstrukturering, se s. 97

Oprindeligt datasæt (*bredt*):

```
kanin a b c d e f
1 7.9 6.1 7.5 6.9 6.7 7.3
2 8.7 8.2 8.1 8.5 9.9 8.3
3 7.4 7.7 6 6.8 7.3 7.3
4 7.4 7.1 6.4 7.7 6.4 5.8
5 7.1 8.1 6.2 8.5 6.4 6.4
6 8.2 5.9 7.5 8.5 7.3 7.7
```

skal omdannes til *langt* (her kaldet *rabbit*):

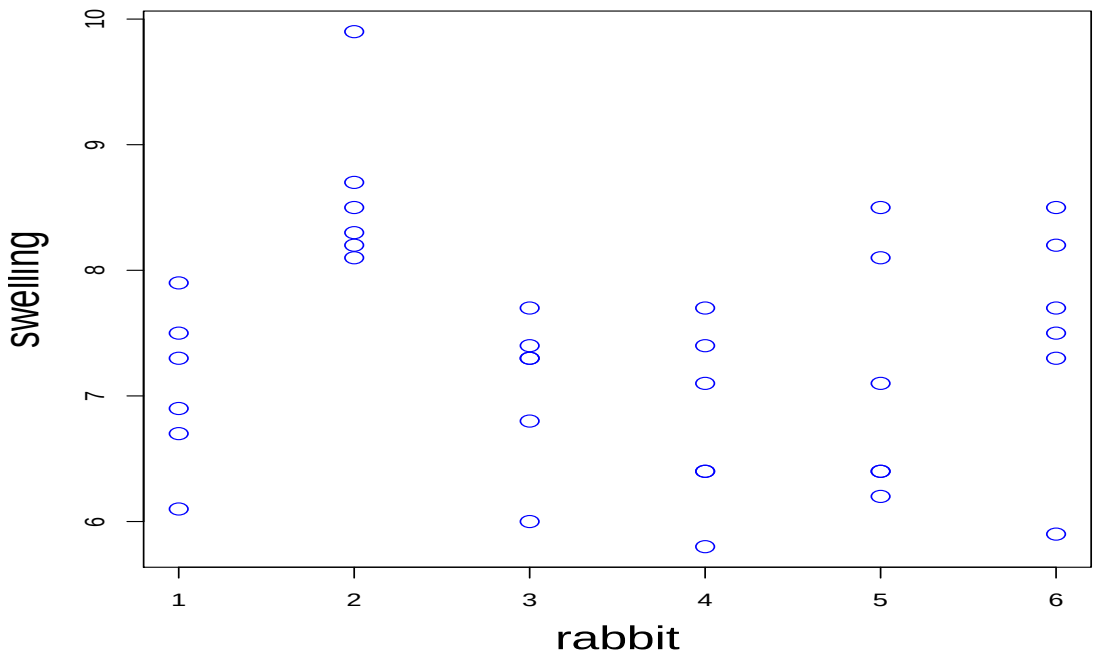
```
rabbit sted swelling
1 a 7.9
2 a 8.7
3 a 7.4
4 a 7.4
5 a 7.1
6 a 8.2
1 b 6.1
. . .
```

Scatter plot

36 observationer, 2 variable:

Outcome: Hævelse = `swelling`

X-akse: Arbitrær nummerering af kaniner



Naiv kvantificering af hævelse

```
> t.test(lang$swelling)

One Sample t-test

data:  lang$swelling
t = 47.458, df = 35, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 7.051540 7.681793
sample estimates:
mean of x
 7.366667
```

Hvad er der galt med denne kvantificering?

Tænk, hvis alle målinger på samme kanin var *helt ens*?
Så har vi jo i virkeligheden kun 6 observationer.

Men de er jo ikke *helt* ens, hvad så?

Analyse (med hovedet under armen)

- Hver kanin har *et niveau* (en middelværdi)
- Herudover er der *variation mellem indstikssteder*

I computer sprog:
Kaninen er en faktor, analysen er en ensidet variansanalyse:

```
gal.model = lm(swelling ~ relevel(factor(rabbit),ref="6"), data=lang)
```

med output

```
> anova(gal.model)
Analysis of Variance Table

Response: swelling
               Df Sum Sq Mean Sq F value   Pr(>F)
factor(rabbit)  5 12.833 2.56667  4.3933 0.004044 **
Residuals      30 17.527 0.58422
```

Output, fortsat

```
> summary(gal.model)

Call:
lm(formula = swelling ~ relevel(factor(rabbit), ref = "6"), data = lang)

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         7.5167     0.3120  24.089   <2e-16 ***
relevel(factor(rabbit), ref = "6")1  -0.4500     0.4413  -1.020   0.3160
relevel(factor(rabbit), ref = "6")2   1.1000     0.4413   2.493   0.0184 *
relevel(factor(rabbit), ref = "6")3  -0.4333     0.4413  -0.982   0.3340
relevel(factor(rabbit), ref = "6")4  -0.7167     0.4413  -1.624   0.1148
relevel(factor(rabbit), ref = "6")5  -0.4000     0.4413  -0.906   0.3719
```

Kaninerne har signifikant forskellige niveauer (P=0.0040)

Men: Er det overhovedet interessant information?
Det er i hvert fald ikke svar på spørgsmålet!

- Vi er jo ikke interesserede i *disse specifikke* 6 kaniner, men snarere kaniner i al almindelighed, *som art betragtet*!
- Vi antager, at disse 6 kaniner er tilfældigt udvalgt

Varianskomponentmodel

Vi vil i stedet se på en model, hvor forskellen på kaniner modelleres som en ekstra variationskilde:

$$y_{rs} = \mu + a_r + \varepsilon_{rs}$$

hvor a_r'erne og ε_{rs}'erne antages at være *uafhængige*, normalfordelte med

$\text{Var}(a_r) = \omega_B^2$, $\quad$ $\text{Var}(\varepsilon_{rs}) = \sigma_W^2$

Variationen mellem kaniner er nu en tilfældig effekt (random factor),
ω_B^2 og σ_W^2 er varianskomponenter, og
modellen kaldes også en two-level model

Hvad indebærer denne varianskomponentmodel

Alle observationer af hævelse har samme middelværdi og samme varians (summen af varianskomponenterne):

$$y_{rs} \sim N(\mu, \omega_B^2 + \sigma_W^2)$$

Men: Målinger foretaget på den samme kanin er korrelerede, med intra-class korrelationen

$$\text{Corr}(y_{r1}, y_{r2}) = \rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2}$$

Målinger på samme kanin har en tendens til at se mere ens ud end målinger på forskellige kaniner.

Alle målinger på samme kanin ser lige ens ud. Denne korrelationsstruktur kaldes compound symmetry (CS) eller exchangeability.

Estimation i varianskomponentmodel

Her skal man sørge for at definere

- `rabbit` som *random factor*
- En *tom* middelværdi
  (de har alle den samme, nemlig interceptet, μ),
  angivet som et 1-tal nedenfor

Koden kræver pakken `nlme`:

```
install.packages("nlme")
library(nlme)

model1 = lme(swelling ~ 1 , data=lang, random=~1 | rabbit)
```

Output fra varianskomponentmodel s. 23

```
> intervals(model1)
Approximate 95% confidence intervals

 Fixed effects:
                lower     est.    upper
(Intercept) 6.821352 7.366667 7.911981
attr(,"label")
[1] "Fixed effects:"


 Random Effects:
  Level: rabbit
                     lower      est.    upper
sd((Intercept)) 0.2567227 0.5748109 1.287021

 Within-group standard error:
    lower      est.     upper
0.5934605 0.7643443 0.9844331
```

Sammenlignes med CI=(7.052, 7.682) fra tidligere (s. 18): At ignorere korrelationen fører her til en type 1 fejl, nemlig *for smalt CI*.

Fortolkning af varianskomponenter

Værdierne er taget fra output s. 24:

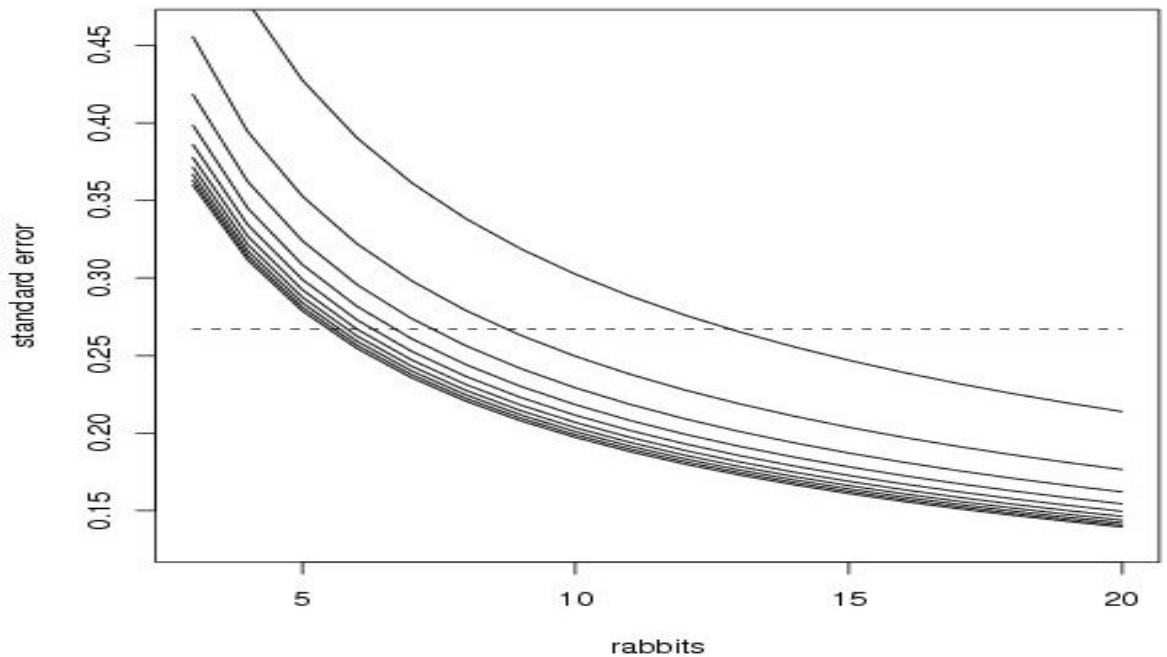
Variation	Varianskomponent	Estimat	Andel af variationen
Between	ω_B^2	$0.575^2 = 0.3304$	36%
Within	σ_W^2	$0.764^2 = 0.5842$	64%
Total	$\omega_B^2 + \sigma_W^2$	0.9146	100%

Typiske forskelle (95% prædiktionsintervaller):

- for to steder på den samme kanin
  $\pm 2 \times \sqrt{2 \times 0.5842} = \pm 2.16\ cm^2$
- for to steder på forskellige kaniner
  $\pm 2 \times \sqrt{2 \times 0.9146} = \pm 2.70\ cm^2$

Standard error på estimat for hævelse

For R=antal kaniner, fra 3 til 20:
For S=antal injektionssteder, fra 1 til 10:



$$\text{Var}(\bar{y}) = \frac{\omega_B^2}{R} + \frac{\sigma_W^2}{RS}$$

*Den "effektive" sample size

Hvis vi kun havde en enkelt observation for hver af k kaniner, hvor mange kaniner skulle vi så bruge til at få samme præcision?

$$k = \frac{R \times S}{1 + \rho(S - 1)}$$

Her har vi $\rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2} = \frac{0.3304}{0.3304 + 0.5842} = 0.361 \Rightarrow k = 12.8$

Vi har altså, effektivt set, kun hvad der svarer til to uafhængige observationer fra hver kanin!

Longitudinelle målinger

Eksempel: Aspirin optagelse for raske og syge (Matthews et.al.,1990)

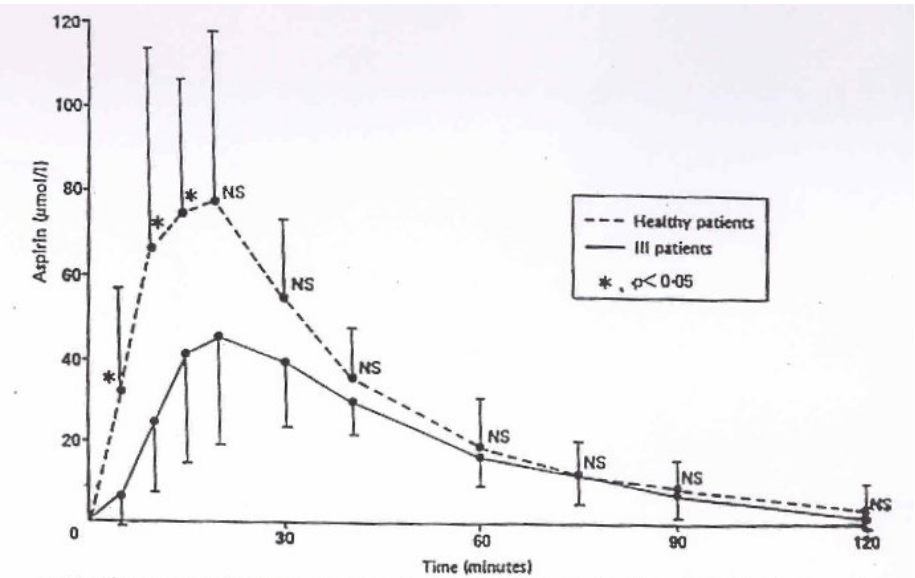


FIG 2—*Mean and standard deviation of aspirin concentrations in nine healthy and nine ill patients over time. ("Usual" method of display)*

T-tests for hvert tidspunkt:

- massesignifikansproblem
- testene er ikke uafhængige
- fortolkning kan være vanskelig

Pas på med gennemsnit - og gennemsnitskurver

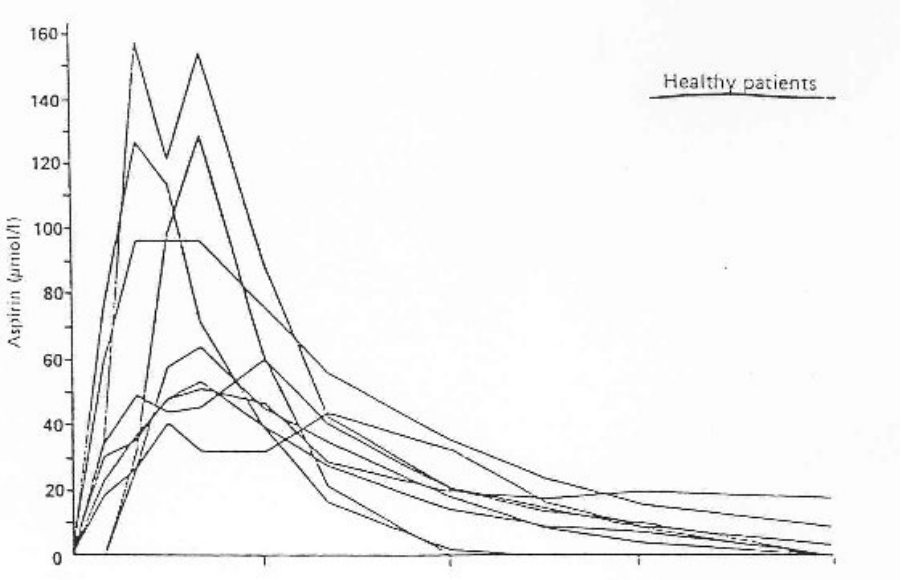
specielt når der er *missing values*

- De giver ingen fornemmelse for variationen mellem personer
  Individuelle forløb bør altid tegnes
  (men ikke nødvendigvis publiceres)
- De kan skjule vigtige strukturer, hvis tidsforløbene adskiller sig *kvalitativt* fra hinanden,
  (hvis de ikke er rimeligt parallelle):

Alternativ:
Regn videre på individuelle karakteristika

Individuelle tidsprofiler

Spaghettiplot



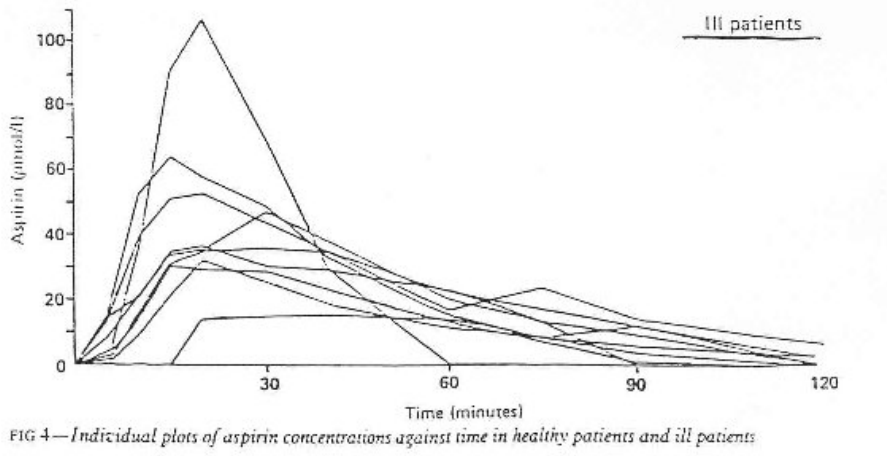


FIG 4—*Individual plots of aspirin concentrations against time in healthy patients and ill patients*

Har de samme facon allesammen?

Er gennemsnittet repræsentativt?

Analyse af udvalgte karakteristika

Eksempelvis:

- Endpoint (sidste måling), hældning
  dur helt klart ikke i dette eksempel
- Maksimal værdi, toppunkt, og dettes placering
- AUC, Arealet under kurven

Sådanne karakteristika sammenlignes ved hjælp af traditionelle metoder, typisk et T-test.

Husk (som altid) konfidensintervaller

Fordele og ulemper ved de simple analyser

Fordele:

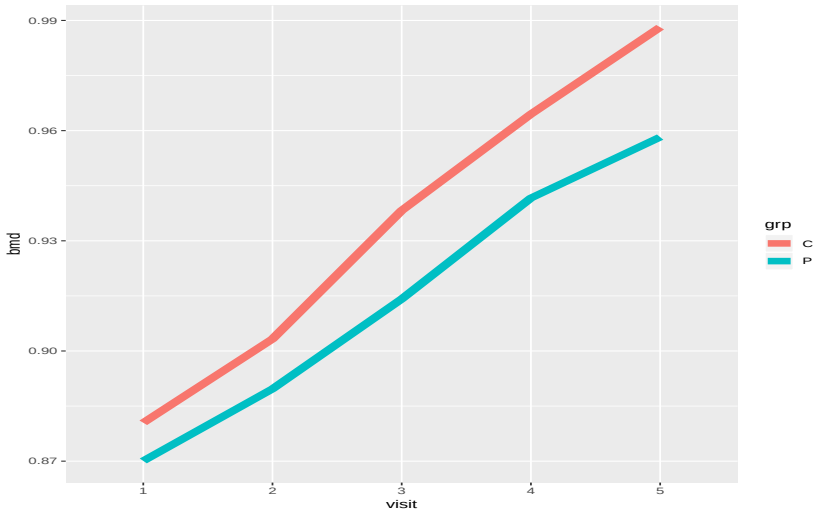
- simple (at forstå og forklare til andre)
- for det meste simple at udføre

Ulemper:

- Informationstab
- Kræver et passende antal observationer pr. individ
- Kræver et fornuftigt karakteristika
  og ofte målinger på ens tidspunkter
- Umuligt at inddrage tidsafhængige kovariater
- Kan ikke tage hensyn til baseline

Eksempel: Calcium tilskud

Gennemsnit for de 2 grupper, (se kode s. 100)

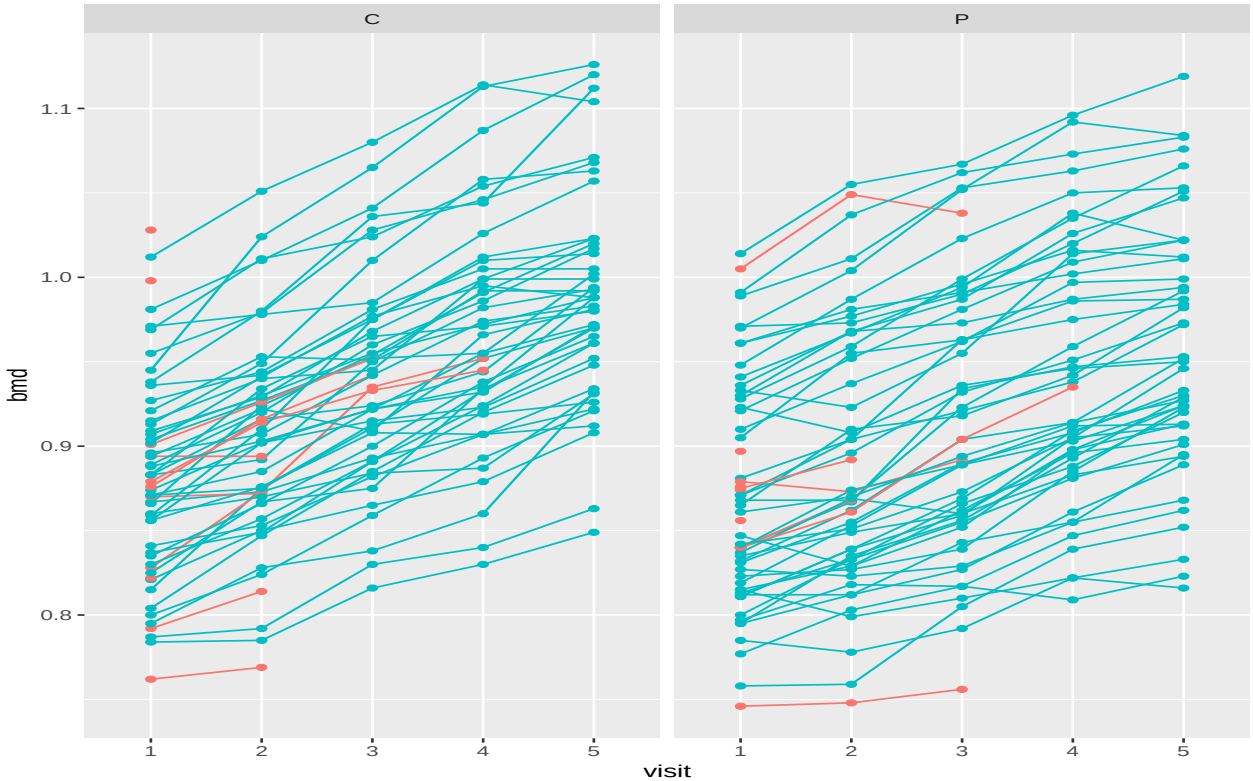


men er gennemsnitskurver rimeligt her?

Ja, nogenlunde....pga de ret ensartede forløb, se næste side,

men der er missing values... og baseline issues

Individuelle profiler - igen (som s.4)



Simple analysemuligheder

Udvælg et (eller flere) karakteristika til
sammenligning af grupperne:

- Gennemsnit for hvert tidspunkt: Ingen evidens for forskel
  husk korrektion for massesignifikans.
- Ændringer for successive tidspunkter:
  husk korrektion for massesignifikans.
  Ændring fra start til slut er større i Calcium-gruppen:
  0.019 (0.006, 0.032), P=0.004
- Individuelle hældninger er større i Calcium-gruppen:
  Estimat for forskel: 0.0039 (0.0019), P=0.050

men disse analyser er suboptimale
– hvis andet er muligt...

Struktur for Calcium-eksemplet

	Unit/Enhed	Variation	Kovariater
Level 2	piger	*mellem* piger ω_B^2	`grp`
Level 1	enkeltobservationer	*indenfor* piger σ_W^2	`visit` `grp*visit`

Vi er specielt interesserede i within-kovariaten `grp*visit`, fordi den udtrykker en *forskel* i mønsteret over tid,

Modelspecifikation i praksis

Mixed model, med

Systematiske effekter: De to faktorer: `grp` og `visit`, samt en interaktion mellem disse,
dvs. *ingen bindinger på tidsudviklingen*

Tilfældige effekter: `girl` som *random factor*

```
calcium$girl = as.factor(calcium$girl)
calcium$visit = as.factor(calcium$visit)

model2 = lme(bmd ~ grp*visit , data=calcium,
         na.action=na.exclude, random=~1 | girl)

anova(model2)
intervals(model2)
```

Bemærkning til modelspecifikation

- Pigerne er nestet i grupperne,
  Vurdering af gruppeforskelle er uparret

  Pigerne er en tilfældig effekt, specificeret som
  `random=~1 | girl`

  Det er vigtigt, at alle pigerne har et unikt nummer

- Alle visits forekommer (i princippet) for den enkelte pige,
  Vurdering af tidseffekter er parrede
- R regner ikke helt de samme frihedsgrader som SAS.....
  der findes mange forskellige approksimationer

Systematisk vs. tilfældig effekt?

Systematisk =Fixed:

- Alle faktorens niveauer observeres
  (typisk kun et par stykker, f.eks. et antal behandlinger, eller nogle tidspunkter)
- Der kan kun drages konklusioner om netop disse behandlinger
  (ikke om andre *ikke-benyttede* behandlingstyper)
- Vi er interesserede i forskellen på de enkelte behandlinger
- Der skal være et rimeligt antal observationer for hvert niveau af faktoren
  (ikke for få i nogen behandlingsgrupper)

Systematisk vs. tilfældig effekt?, fortsat

Tilfældig =Random:

- En repræsentativ stikprøve (sample) af faktorniveauer er observeret
  f.eks. et antal patienter, skoleklasser etc.
- Vi er ikke interesserede i netop disse individer, eller disse skoleklasser, *men*
- vi ønsker at drage konklusioner *i al almindelighed*, dvs.
  for andre patienter, skoleklasser eller kaniner,
- Det er *nødvendigt* at lade faktoren være tilfældig, hvis man interesserer sig for kovariater hørende til dette level, f.eks. behandling, klassetrin...
  eller selve niveauet (af hævelsen)

Output fra mixed model, kode s. 37

```
> anova(model2)
             numDF denDF   F-value p-value
(Intercept)      1   381 20974.226  <.0001
grp              1   110     2.217  0.1394
visit            4   381   616.848  <.0001
grp:visit        4   381     5.297  0.0004
```

Analysen viser en hel klar interaktion `grp:visit`, dvs. at der *ikke* er parallelle tidsforløb i de to grupper.

... hvis modellen altså er rimelig

Output, fortsat

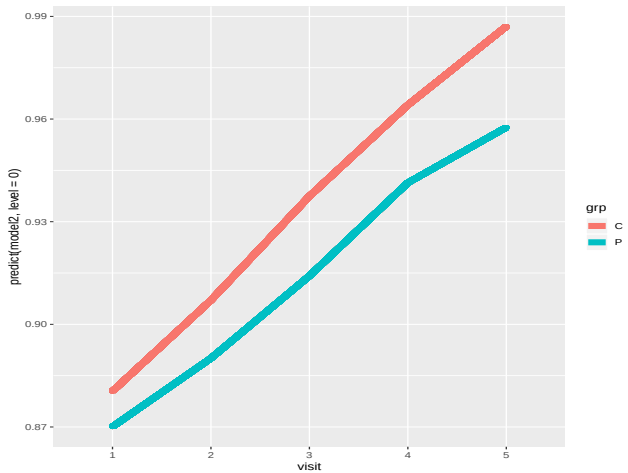
```
> intervals(model2)
Approximate 95% confidence intervals

 Fixed effects:
                     lower         est.        upper
(Intercept)     0.86232877  0.880454545  0.898580317
grpP           -0.03599324 -0.010384370  0.015224501 <---baseline forskel
visit2          0.02068541  0.026588919  0.032492424
visit3          0.05089871  0.056982613  0.063066514
visit4          0.07747389  0.083646811  0.089819734
visit5          0.10035724  0.106620056  0.112882869
grpP:visit2    -0.01488124 -0.006572708  0.001735826
grpP:visit3    -0.02140225 -0.012905624 -0.004408998
grpP:visit4    -0.02097973 -0.012333346 -0.003686960
grpP:visit5    -0.02786083 -0.019121301 -0.010381776
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: girl
                    lower       est.      upper
sd((Intercept)) 0.0582646 0.06662859 0.07619325

 Within-group standard error:
     lower       est.      upper
0.01426996 0.01532020 0.01644773
```

Se bemærkninger til output på s. 45

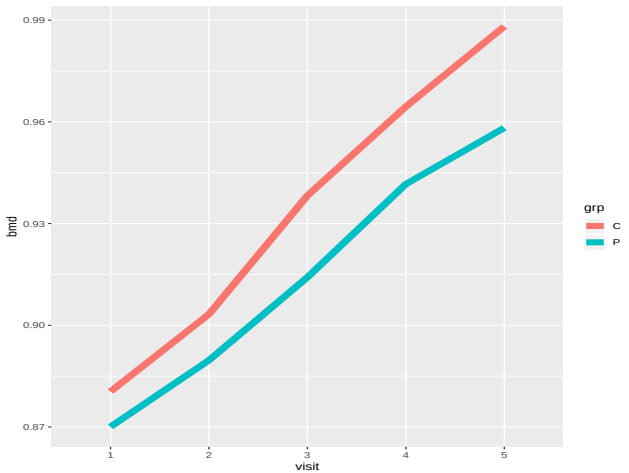
Figur af predikterede kurver

svarer ikke helt til gennemsnittene fra s. 33

Prediktioner:



Gennemsnit:

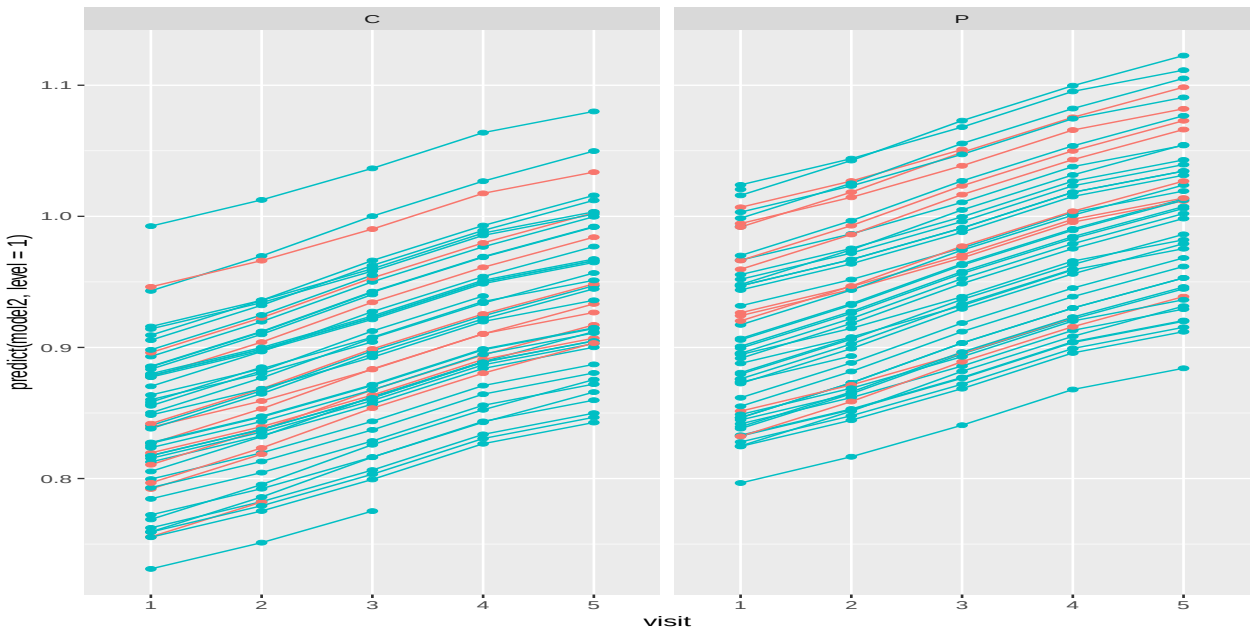


Dette skyldes de enkelte missing values....

Kode til venstre figur: s. 102

Predikterede individuelle forløb

som inkluderer de tilfældige niveauer for pigerne,
(se kode s. 102)



Her er et problem med farverne....for dropouts skulle have været røde

Foreløbig konklusion om calcium-eksemplet

- De predikterede forløb (2*5 estimerede middelværdier) er vist s. 43, venstre plot
- Vi fandt (s. 41) en signifikant interaktion `grp*visit`, dvs. grupperne udvikler sig forskelligt over tid ($P = 0.0004$)
- På s. 42 ses, at forskellen på grupperne ved baseline er 0.036 (-0.010, 0.015) i C-gruppens favør
  (aflæses ud for `grpP`, da visit=1 er referencen)
- Forskellen stiger over tid, hvilket også er forventeligt ved kontinuerlig behandling
  (ses ved, at `grpP:visit`-estimaterne bliver mere og mere negative)
- Intra-individ korrelationen er 0.95 (se s. 52)

Modelkontrol

To typer residualer bør checkes:

De sædvanlige Observeret minus predikteret (gruppe-)middelværdi (kun systematiske effekter fratrækkes)

De betingede Observeret minus predikteret individuel prediktion (både systematiske og tilfældige effekter fratrækkes), *Conditional* på engelsk

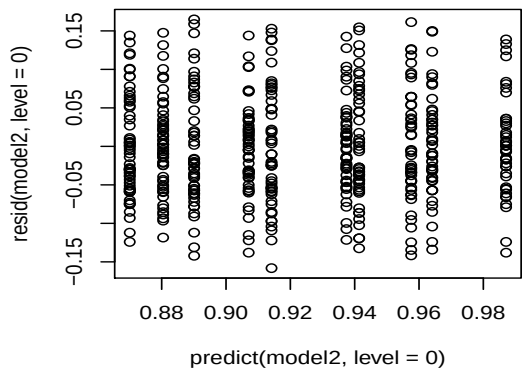
Vi ser på *de sædvanlige* tegninger:

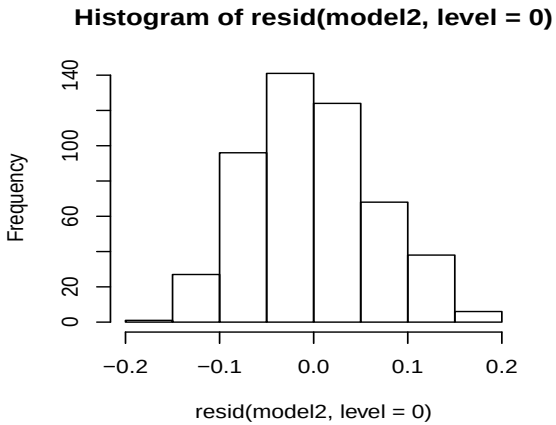
- Residualer plottet mod predikterede værdier
- Histogram og fraktildiagram

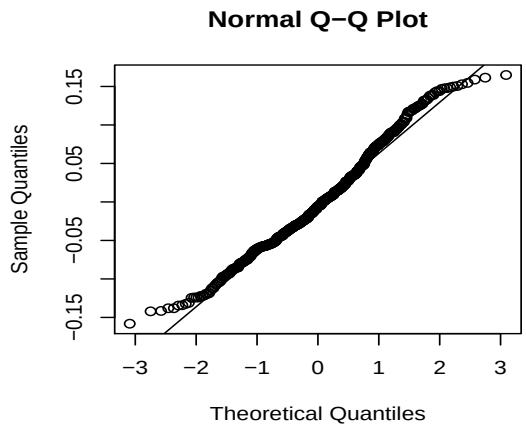
Se figurerne s. 47 og 48 (kode s. 103)

Modelkontrol, sædvanlige (store) residualer

Afvigelse fra gruppe-middelværdier (se s. 46):

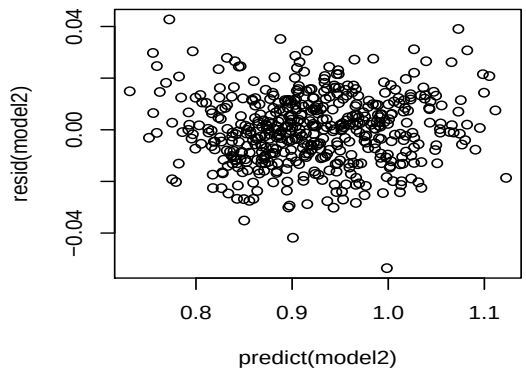


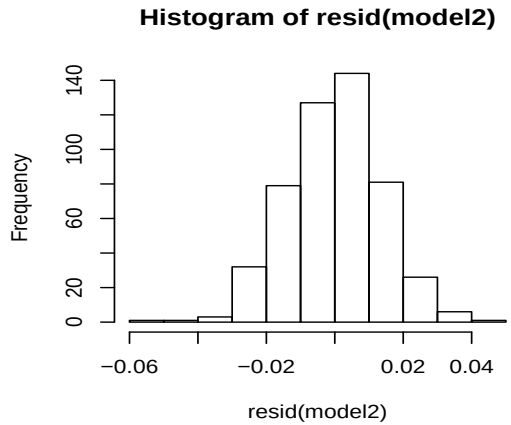


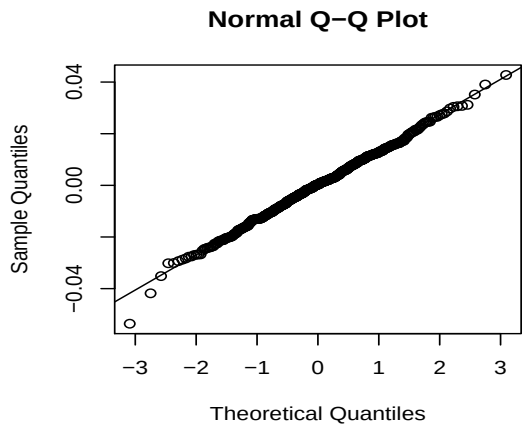


Modelkontrol, betingede (små) residualer

Afvigelse fra individuelle prediktioner (se s. 46):







Effekt af korrelerede observationer

Når korrelationen ignoreres, sker der følgende:

- Level 1 kovariater (tidsrelaterede effekter):
  Lav styrke (type 2 fejl):
  Vi opdager ikke de effekter, der er der.
  Her drejer det sig om `grp:visit`, samt (hvis denne ikke er med i modellen) om selve `visit`
- Level 2 kovariater (behandlinger, gruppe):
  For små standard errors, og dermed for små P-værdier (type 1 fejl):
  Vi kommer til at finde effekter, der slet ikke er der.
  Her er dette kun relevant i en model *uden interaktion*, hvor vi evt. vil vurdere en generel forskel på grupperne (`grp`).

Fejlagtig analyse:

Tosidet ANOVA, korrelation ignoreret

```
> forkert.model = lm(bmd ~ grp*visit, data=calcium)

> anova(forkert.model)
Analysis of Variance Table

Response: bmd
             Df  Sum Sq  Mean Sq F value    Pr(>F)
grp           1 0.04626 0.046262 10.0914  0.001583 **
visit         4 0.64121 0.160303 34.9677 < 2.2e-16 ***
grp:visit     4 0.00650 0.001624  0.3542  0.841105
Residuals   491 2.25090 0.004584
```

Type 2 fejl: Her findes fejlagtigt ingen vekselvirkning, da vi har "glemt" parringen

Vi kan altså *ikke* se, at de to grupper udvikler sig forskelligt....

Synonymer for modellen s. 37

- Two-level model
- Varianskomponentmodel (med 2 varianskomponenter)
- Model med tilfældig person effekt
- Model med tilfældige intercepter (niveauer)
- Model med "compound symmetry" kovariansstruktur (eller "exchangeability" kovariansstruktur)

Korrelation = Kovarians, normeret med spredninger

Alternative kode, se s. 104

Compound symmetry = Exchangeability

Varianskomponentmodellen antager, at alle målinger har samme varians ($\omega_B^2 + \sigma_W^2$) og at alle par af observationer *på samme individ* er lige stærkt korrelerede:

$$\text{Corr}(\, Y_{git_1}\,,\, Y_{git_2}\,) = \rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2}$$

kaldet intra-class korrelationen

Korrelationen ρ estimeres ud fra output s. 42

$$\hat{\rho} = \frac{0.06662859^2}{0.06662859^2 + 0.01532020^2} \approx 0.95$$

Observationerne er ombyttelige=exchangeable, altså CS: Compound Symmetry

Korrelationsstruktur

Her har vi 5 tidspunkter, og derfor en 5*5 korrelations-matrix, hvor entry (i, j) angiver korrelationen mellem visit i og visit j.
Compound Symmetry har strukturen:

$$\begin{pmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{pmatrix}$$

som siger, at alle tids-par er lige korrelerede,
se kode s. 104

Det betyder, at der slet ikke tages hensyn til, at observationerne er taget over tid, altså i en bestemt rækkefølge!!
Er det en fornuftig antagelse?

Valg af varians- og korrelationsstruktur?

Den mest generelle: Ustruktureret:

```
model.un = gls(bmd ~ grp*visit, data=calcium,
   na.action=na.exclude,
   corr = corSymm(form=~1 | girl),
   weight = varIdent(form=~1 | visit),method="REML")
```

Denne struktur er *helt uden bånd*, både på korrelation og på de 5 spredninger, i alt 15 parametre, dog *antaget ens i de to grupper*

Output fra ustruktureret kovariansstruktur

Korrelationsstruktur:

```
Correlation Structure: General
 Formula: ~1 | girl
 Parameter estimate(s):
 Correlation:
  1     2     3     4
2 0.970
3 0.941 0.973
4 0.925 0.959 0.981
5 0.899 0.940 0.959 0.976
```

Spredningsestimater (måske svagt stigende):

```
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | visit
 Parameter estimates:
       1        2        3        4        5
1.000000 1.094589 1.121609 1.162084 1.113905

Residual standard error: 0.06285848
```

Skal man vælge ustruktureret kovarians?

Fordele:

- Vi *tvinger* ikke en forkert kovariansstruktur ned over vores observationer
- Vi får indsigt i mønsteret i spredninger og korrelationer (hvis der er tilstrækkelig information, dvs. mange personer og ikke alt for mange tidspunkter)

Ulemper:

- Vi bruger en masse parametre på beskrivelsen af modellen kovariansen. Resultatet kan derfor blive ustabilt, og kan derfor ikke anvendes for små datasæt
- Den kan kun bruges for balancerede data (alle individer skal være målt til de samme tidspunkter)

så måske hellere noget "midt imellem"?

Mulige korrelationsstrukturer

Observationer, der er foretaget tæt på hinanden i tid vil formentlig være stærkere korrelerede end observationer, der tidsmæssigt ligger længere fra hinanden.

Der findes (forfærdeligt) mange muligheder

- Autoregressiv struktur (se kode s. 105)
- Autoregressiv struktur, *samtidig* med den tilfældige effekt (niveau) for hvert individ, se kode s. 106
- Flere tilfældige effekter, f.eks. tillige en tilfældig hældning Random regression (kommer lidt senere)
- Et væld af andre, prøv f.eks. at google "*sas mixed repeated type*"

Autoregressiv kovariansstruktur, AR(1)

Hvis tiderne er ækvidistante, er det følgende struktur

$$(\omega_B^2 + \sigma_W^2) \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

dvs. korrelationen falder som potenser af afstanden mellem observationerne:

```
model.ar1 = gls(bmd ~ grp*visit, data=calcium,
   na.action=na.exclude,
   corr = corAR1(form=~1 | girl),method="REML")
```

Hvis tiderne ikke er ækvidistante, bruger man i stedet korrelationsstrukturen (altså 3. linie i koden)

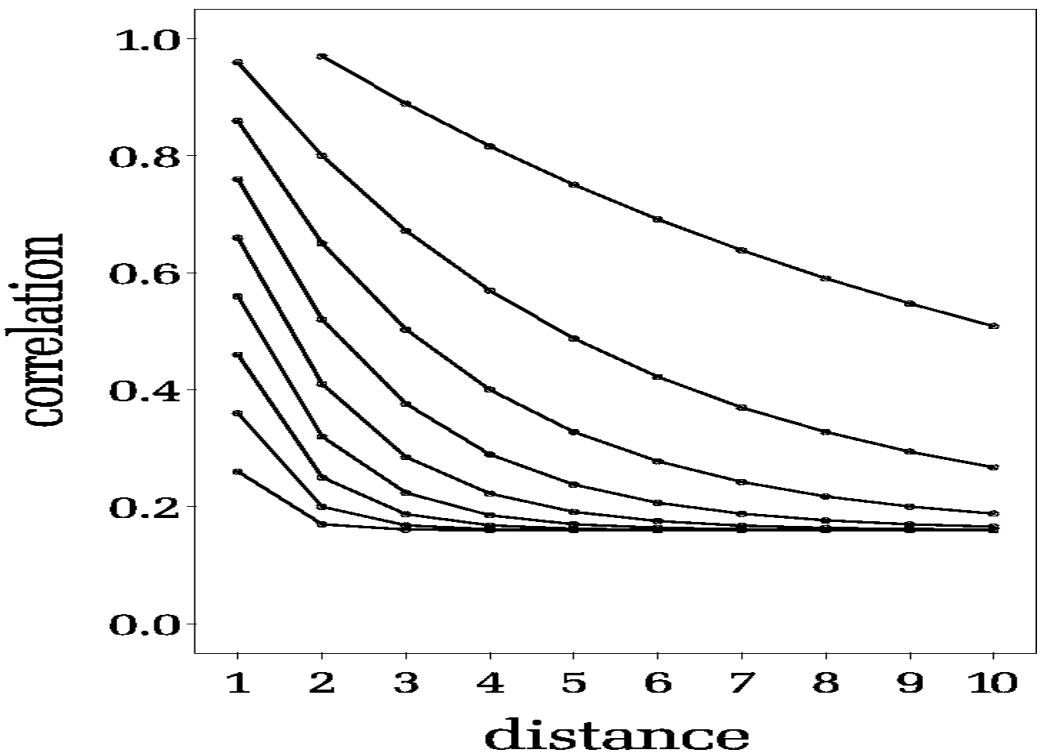
```
   corr=corCAR1(form=~visit|girl))
```

svarende til Corr$(Y_{git_1}, Y_{git_2}) = \rho^{|t_1 - t_2|}$

Autoregressiv korrelation

med overlejret tilfældigt niveau (s. 106)

– som funktion af afstanden mellem målingerne
for $\rho = 0.1, \ldots, 0.9$



Test af interaktionen grp*visit?

– for forskellige valg af kovariansstruktur

Kovarians struktur	Teststørrelse $\sim$ fordeling	P-værdi
Uafhængighed	$0.35 \sim F(4,491)$	0.84
Compound symmetry	$5.30 \sim F(4,382)$	0.0004
Autoregressiv (evt. med CS oveni)	$2.82 \sim F(4,383)$	0.025
Ustruktureret	$2.72 \sim F(4,107)$	0.034

Altså ikke samme konklusion!

Kovariansstrukturen kan være vigtig, så hvordan vælger man?

Det vender vi tilbage til....s. 74

Tidspunkt for de 5 visits

- Der var (naturligvis) ikke præcis et halvt år mellem alle successive målinger.
- Pigerne var heller ikke præcis lige gamle ved start

Hvad gør vi så?

Hvad er den fornuftige tidsskala?

- Alder?
- Tid siden randomisering?

Vi antager, at datoen for den første måling også er dato for randomisering af den pågældende pige, så dette er tid 0.

Mere om håndtering af baseline senere

Tid siden randomisering

Se udregning af denne næste side

- Nu er tiden helt individuel for den enkelte pige
- De starter dog alle med tid 0 ved første besøg
- Der er ikke mere noget, der hedder visit1, visit2 osv., det svarer i hvert fald ikke til et bestemt tidspunkt
- Tiden er også omregnet til `years`, år siden randomisering

Bemærk vedr. output på næste side:

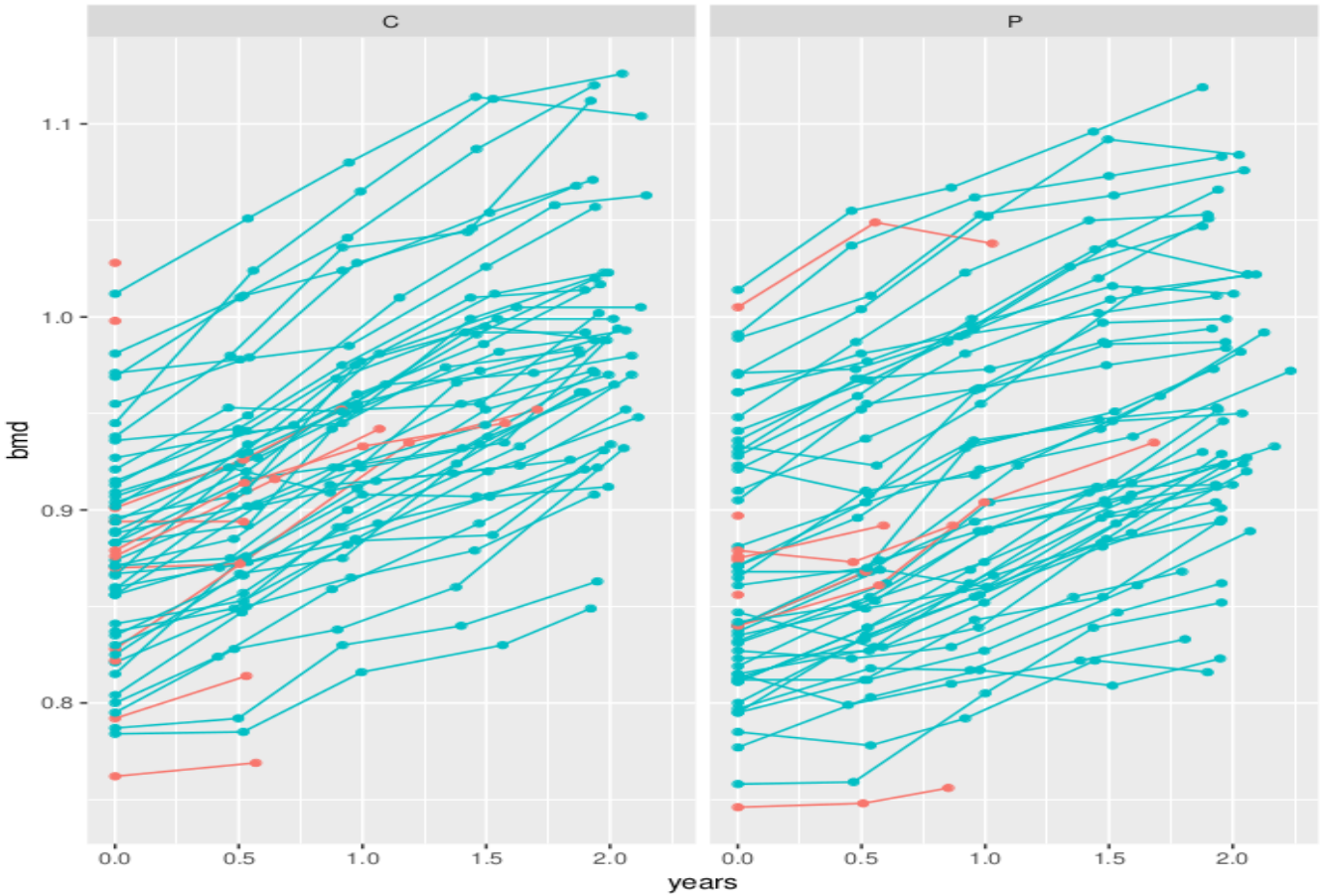
- Det faldende antal målinger over de to år
  missing values/dropout
- Variationen i prøvetagningen til de enkelte *visits*

Overblik over nye tider

```
c.tid <- read.table("calcium_tider.txt", header=T)
c.tid$girl = as.factor(c.tid$girl)
c.tid$visit = as.factor(c.tid$visit)
c.tid$years = c.tid$tid/365.25

install.packages("doBy")
library(doBy)

summaryBy(years + bmd ~ grp + visit, data = c.tid,
  FUN = function(x) { c(mean = mean(x,na.rm=T),
    min = min(x,na.rm=T),max = max(x,na.rm=T)) } )
```

```
   grp  visit years.mean years.min years.max  bmd.mean bmd.min bmd.max
1    C     1  0.0000000 0.0000000 0.0000000 0.8804545   0.762   1.028
2    C     2  0.5202970 0.4161533 0.7227926 0.9032692   0.769   1.051
3    C     3  0.9630390 0.8678987 1.1882272 0.9382500   0.816   1.080
4    C     4  1.4977234 1.3360712 1.7768652 0.9645000   0.830   1.114
5    C     5  1.9762927 1.8398357 2.1464750 0.9881591   0.849   1.126
6    P     1  0.0000000 0.0000000 0.0000000 0.8700702   0.746   1.014
7    P     2  0.5182287 0.4462697 0.5995893 0.8896792   0.748   1.055
8    P     3  0.9584625 0.8487337 1.1307324 0.9141569   0.756   1.067
9    P     4  1.5017112 1.3415469 1.7056810 0.9416458   0.809   1.096
10   P     5  1.9759127 1.7960301 2.2340862 0.9582340   0.816   1.119
```

De nye (faktiske) individuelle forløb



Kovariansstrukturer i tilfælde af uens tidspunkter

Når alle personer ikke er målt til de samme tidspunkter, er visse middelværdistrukturer og korrelationsstrukturer
ikke længere mulige

- Ustruktureret middelværdi, dvs. en parameter for hvert tidspunkt (disse er jo ikke ens mere)
- Ustruktureret kovarians/korrelation

Men man kan stadig benytte

- CS-strukturen
- random regression, kommer nu
- Erstatte AR(1)-strukturen med `corr=corCAR1(form=~years|girl)`, se s. 58

Ny ide: Individuelle vækstrater?

Vi antager, at tidsudviklingen er lineær, men ikke helt lige stejl for alle piger, dvs. vi indfører individuelle hældninger:

Lad Y_{git} være den observerede BMD for den i'te pige (i den g'te gruppe) til tid t. Vi specificerer modellen:

$$y_{git} = a_{gi} + b_{gi}t + \varepsilon_{git}, \quad \varepsilon_{git} \sim N(0, \sigma^2_W)$$

altså en individuel linie for hver eneste pige, med intercept a_{gi} og hældning b_{gi}

Bemærk, at interceptet her svarer til `time=0` (`years=0`), altså randomiseringstidspunktet (tidspunkt for baseline måling).

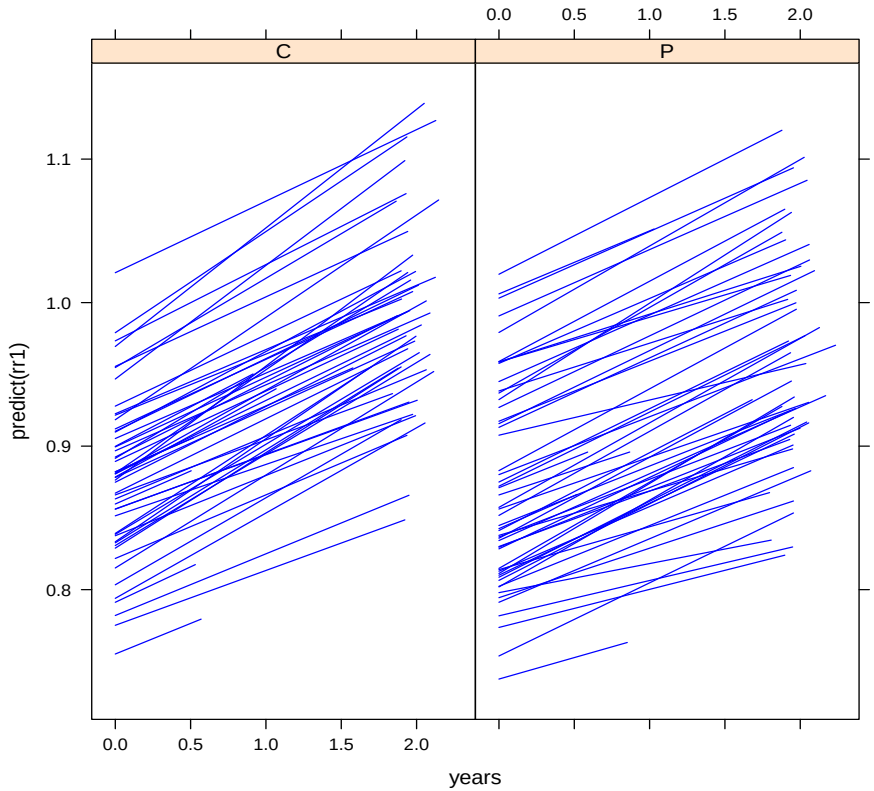
Stokastisk regression/Random regression

– en generalisering af ideen om tilfældigt niveau

Vi lader godt nok
hver pige have

- sit eget niveau a_{gi}
- sin egen hældning b_{gi}

men...



Random regression, II

... vi *binder* disse individuelle 'parametre' (a_{gi} og b_{gi}) sammen med en antagelse om Normalfordelingen (den todimensionale)

$$\begin{pmatrix} a_{gi} \\ b_{gi} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \alpha_g \\ \beta_g \end{pmatrix}, G \right)$$

hvor G er en 2×2-matrix, der beskriver **populationsvariationen** af linierne, dvs.

- variationen mellem niveauerne
- variationen mellem hældningerne
- korrelationen mellem niveau og hældning

Random regression i praksis

Her specificeres:

- To *vilkårlige linier* som fixed effects: `grp:years`-sætningen (baseline-korrektion senere)
- En 2×2-kovarians for de personspecifikke intercepter og hældninger (G) i `random`-sætningen

```
install.packages("lme4")
library(lme4)

rr1 <- lmer(bmd ~ grp*years + (years | girl), c.tid)
```

Dele af output fra random regression

- G-matricen, i form af varianser, spredninger og korrelation: (kode s. 69 eller alternativ s. 107)
- Residualvariationen

```
REML criterion at convergence: -2351.4

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 girl     (Intercept) 0.0041784 0.06464
          years       0.0001791 0.01338  0.12
 Residual             0.0001244 0.01115
```

Ikke megen afhængighed mellem intercepter og hældninger
men det er en tilfældighed...

Dele af output,II

Korrelationsmatricen for de 5 visits for en specifik pige, fordi de nu har individuelle tidspunkter, og dermed også individuelle korrelationer (der er ikke lige langt mellem observationerne). (delvis kode s. 107)

Variansestimater (svagt stigende):

Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
0.004303	0.004457	0.004677	0.005014	0.0.005429

```
  Estimated V Correlation Matrix for girl(grp) 101 C
Row      Col1      Col2      Col3      Col4      Col5
  1    1.0000    0.9663    0.9540    0.9329    0.9072
  2    0.9663    1.0000    0.9688    0.9571    0.9396
  3    0.9540    0.9688    1.0000    0.9700    0.9598
  4    0.9329    0.9571    0.9700    1.0000    0.9727
  5    0.9072    0.9396    0.9598    0.9727    1.0000
```

Dele af output, III

```
Fixed effects:
             Estimate Std. Error t value
(Intercept)  0.880950   0.008799 100.118
grpP        -0.011565   0.012334  -0.938  <- baseline forskel
years        0.054207   0.002197  24.673
grpP:years  -0.008887   0.003074  -2.891  <- forskel i slopes

> confint(rr1)
Computing profile confidence intervals ...
                  2.5 %       97.5 %
.sig01       0.05633750  0.073619148
.sig02      -0.11234867  0.343055024
.sig03       0.01091750  0.015936698
.sigma       0.01029930  0.012127272
(Intercept)  0.86370825  0.898186939
grpP        -0.03573083  0.012599979  <- baseline forskel
years        0.04988758  0.058503738
grpP:years  -0.01490672 -0.002861936
```

I denne model kvantificerer vi effekten af calciumtilskud (forskellen på de to hældninger) til 0.0089 (0.0031) g pr cm^3 pr **år**.

Mere output:

Estimeret fordel efter 2 år: 0.0293 (0.0140) g pr cm^3, udregnet således:

```
> coef(summary(rr1))[ 2, "Estimate"]
  + 2*coef(summary(rr1))[ 4, "Estimate"]
[1] -0.02933892
```

Ingen signifikant forskel ved baseline (mere om dette fra s. 79)

Sammenligning af modelfit

Er random regression en god model?

$-2 \log L$ skal være lille, dvs. et stort negativt tal

Kovarians struktur	$-2 \log L$	Kov.par.	Differens i hældning	P
Uafhængighed	-1252.4	1	0.0105 (0.0086)	0.22
Compound Symmetry	-2253.7	2	0.0089 (0.0020)	< 0.0001
Autoregressiv sp(pow)	-2373.6	2	0.0094 (0.0033)	0.0037
Random Regression	-2351.4	4	0.0089 (0.0031)	0.0048

Random regression ser rimelig ud (AIC lille),
måske er den autoregressive struktur dog *en anelse bedre*

Random regression = Stokastisk regression

Fordele:

- Bruger al forhåndenværende information
- Optimal procedure hvis modellen holder
- Let at inkludere kovariater
- Kan tage hensyn til baseline (kommer lige om lidt)

Ulemper:

- Sværere at forstå og kommunikere videre
- Biased i tilfælde af informative manglende værdier
  f.eks. hvis piger med lav stigning konsekvent udgår af studiet
  (den såkaldte "healthy worker" effekt)

Hvorfor ikke bare bruge individuelle linier?
(besværligere, suboptimalt, somme tider umuligt, se s. 32)

Sammenligning af de to fremgangsmåder

Forskel på hældninger C vs. P:

- Random regression: 0.0089 (0.0031)
  P: 0.0453, C: 0.0542
- Individuelle regressioner: 0.0077 (0.0039)
  P: 0.0413, C: 0.0490

Hvorfor denne forskel?

- De korte forløb er mere usikkert bestemt, og det tages der hensyn til i Random Regression, men (selvfølgelig) ikke, når man tager gennemsnit af individuelle hældninger.
- De korte forløb har generelt lavere hældninger, mest for C-gruppen, se s. 77
  Dette er indikation af selektivt bortfald...uha!

Hældninger, opdelt efter behandling og dropout-status

`dropout=1` betyder, at pigen ikke gennemfører hele forløbet. Sådanne ses at have lavere hældninger (koefficienten til `years`), mest udtalt for Calcium-gruppen:

Group	dropout	N	Hældning
C	0	44	0.0546824
C	1	11	0.0265156
P	0	47	0.0458431
P	1	10	0.0201080

Manglende observationer - missing values

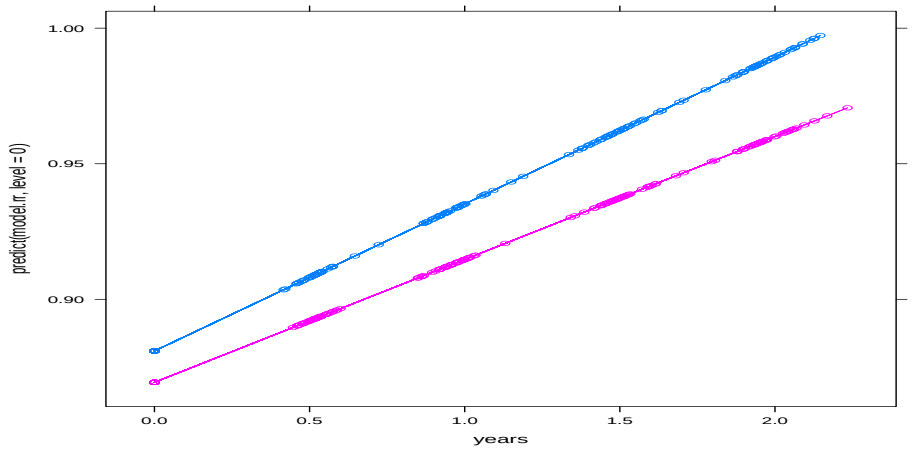
MCAR Missing completely at random
mangler alene pga tilfældigheder

MAR Missing at random
Missingness kan afhænge af kovariater (x)
og evt. også af tidligere outcome-værdier (y)

NI Non-ignorable: (Informative missing)
Missingness afhænger af
de uobserverede outcome værdier!!

Predikterede forløb fra random regression

Kode s. 108



Som vi tidligere har set, er der en vis forskel allerede fra starten (ved baseline), selv om denne er *insignifikant*:

```
> 2*pt(-0.938,110)
[1] 0.3502994
```

dvs. P=0.35, se s. 72.

Bør vi justere for baseline?

Justering for baseline?

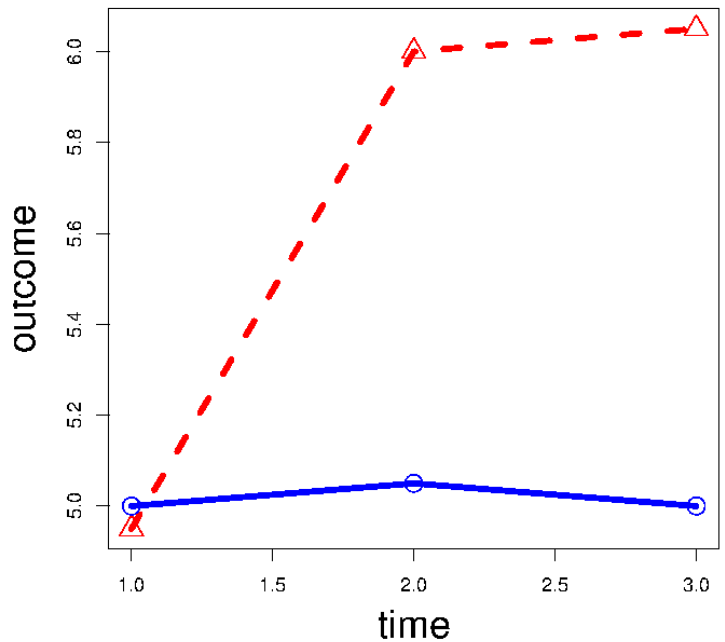
Baseline måling:
Observation foretaget inden (eller ved) behandlingsstart.

Vi diskuterede håndtering af sådanne i forbindelse med "Ancova":

- Der skal *ikke* (nødvendigvis) justeres i tilfælde af observationelle studier
- Justering bør foretages,
  hvis der er tale om randomiserede undersøgelser,
  fordi vi vil sammenligne to personer, som starter med at være ens, men som modtager forskellige behandlinger
  – og det er et *tilfælde*,
  hvis grupperne ikke starter på samme niveau

Hypotetisk naiv sammenligning af to grupper, I

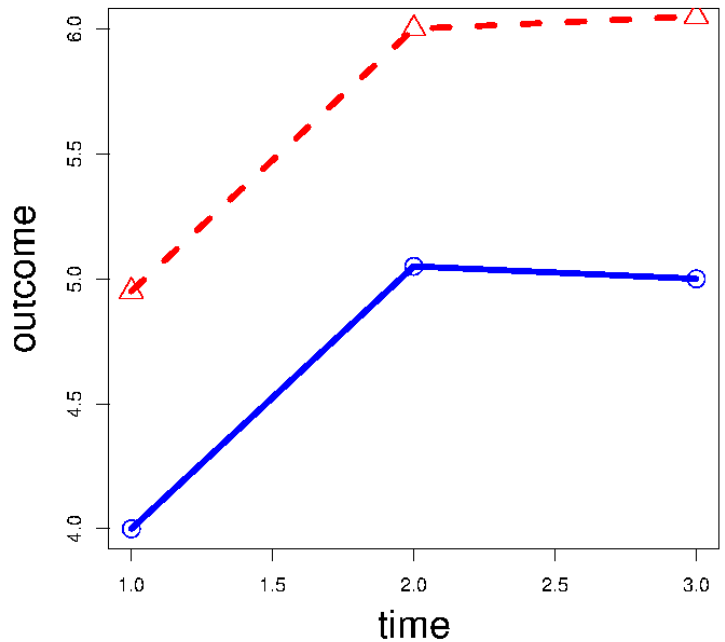
uden hensyntagen til "principielt ens" baselines:



- Konklusion: Interaktion mellem tid og behandling
- Sandhed: Konstant forskel på de to behandlinger

Hypotetisk naiv sammenligning af to grupper, II

uden hensyntagen til "principielt ens" baselines



- Konklusion: Konstant forskel på behandlingerne
- Sandhed: Ingen behandlingseffekt

Hvordan korrigeres for baseline?

- Baseline inddrages som kovariat:
  - ikke helt godt, fordi det svarer til at antage, at korrelationen mellem baseline og hver af de efterfølgende observationer er lige stærk
- Baseline fratrækkes:
  ikke altid så godt pga *Regression to the mean*,
  men fungerer fint for langsomt varierende outcome
- Middelværdierne i de to grupper sættes lig hinanden ved starttidspunktet
  - det fungerer nemt i "random regression"
  - Ofte kan man simpelthen omdefinere sin behandlingsvariabel, så den angiver "kontrolgruppe" for alle ved baseline.

Med baseline som kovariat (ikke helt rimeligt)

Nu er det kun de sidste 4 tidspunkter, der er outcome
Se kode s. 109

```
Fixed effects: bmd ~ baseline + grp.r * years
                  Value   Std.Error  DF  t-value p-value
(Intercept)  -0.0725551 0.024270092 282 -2.98948  0.0030
baseline      1.0811382 0.027735082 102 38.98089  0.0000
grp.rC        0.0028570 0.003641519 102  0.78457  0.4345
years         0.0462387 0.002276755 282 20.30905  0.0000
grp.rC:years  0.0072835 0.003260766 282  2.23369  0.0263
```

Estimeret fordel efter 2 år: 0.0174 (0.0063) g pr cm^3:

```
> coef(summary(model.bas))[3,1]
  + 2*coef(summary(model.bas))[5,1]
[1] 0.01742408
```

Vi mangler dog usikkerheden her....

Analyse af differenser (ikke helt rimeligt)

- Baseline fratrækkes alle efterfølgende værdier
- Baseline selv benyttes *ikke* i analysen

Se kode s. 110

Solution for Fixed Effects

Effect	grp	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		-0.00197	0.002666	101	-0.74	0.4619
grp	C	0.003388	0.003799	101	0.89	0.3746
grp	P	0	.	.	.	.
years		0.04623	0.002281	93.3	20.27	<.0001
years*grp	C	0.007330	0.003267	93.5	2.24	0.0272
years*grp	P	0	.	.	.	.

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
forskel efter 2 aar	0.01805	0.006213	99.4	2.90	0.0045

Hvis baseline benyttes som kovariat her, fås de samme resultater som s. 84

Fælles middelværdi ved baseline, I

Hvis der er tale om en lineær tidsudvikling, som her `years`:

Udelad `grp` af modelsætningen, men behold interaktionen:

```
bmd ~ years + grp.r:years
```

- Når `years` er 0 (ved baseline), har `bmd` middelværdi svarende til interceptet, for begge grupper.
- men der er to forskellige hældninger

Random regression, med ens baseline

Udelad grp af modellen, men behold interaktionen

```
model.rr2 = lme(bmd ~ years + grp.r:years, data=c.tid,
  na.action=na.exclude, random=~years | girl)
```

med output

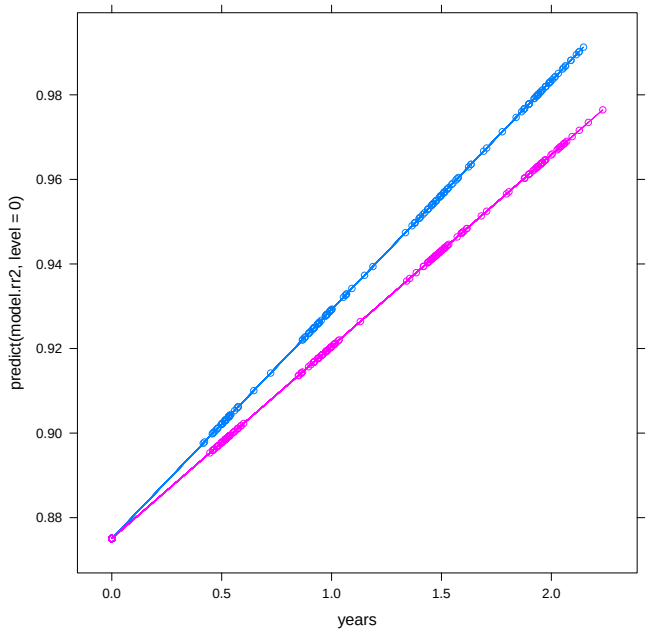
```
Fixed effects: bmd ~ years + grp.r:years
                 Value   Std.Error  DF   t-value p-value
(Intercept)  0.8750644 0.006162637 387 141.99512  0.0000
years        0.0453840 0.002148811 387  21.12050  0.0000
years:grp.rC 0.0087578 0.003070793 387   2.85197  0.0046
```

Estimeret fordel efter 2 år: 0.0175 (0.0062) g pr cm^3

```
> 2*coef(model.rr2)[1,3]
[1] 0.01751561
```

Predikterede forløb, med ens baselines

Se kode s. 111



Vi ser her, at gruppernes middelværdi tvinges til at starte på samme niveau, så nu er der justeret for baseline

Fælles middelværdi ved baseline, II

Hvis tidsudviklingen blot er "*forskellige middelværdier*", altså hvis tiden er en `class`-variabel (såsom `visit`):

```
c.tid$A.visit2 = (c.tid$grp=="C")*(c.tid$visit=="2")
c.tid$A.visit3 = (c.tid$grp=="C")*(c.tid$visit=="3")
c.tid$A.visit4 = (c.tid$grp=="C")*(c.tid$visit=="4")
c.tid$A.visit5 = (c.tid$grp=="C")*(c.tid$visit=="5")

model.baskorr = lme(bmd ~ visit + A.visit2 + A.visit3 +
                A.visit4 + A.visit5, data=c.tid,
                na.action=na.exclude, random=~1 | girl)
```

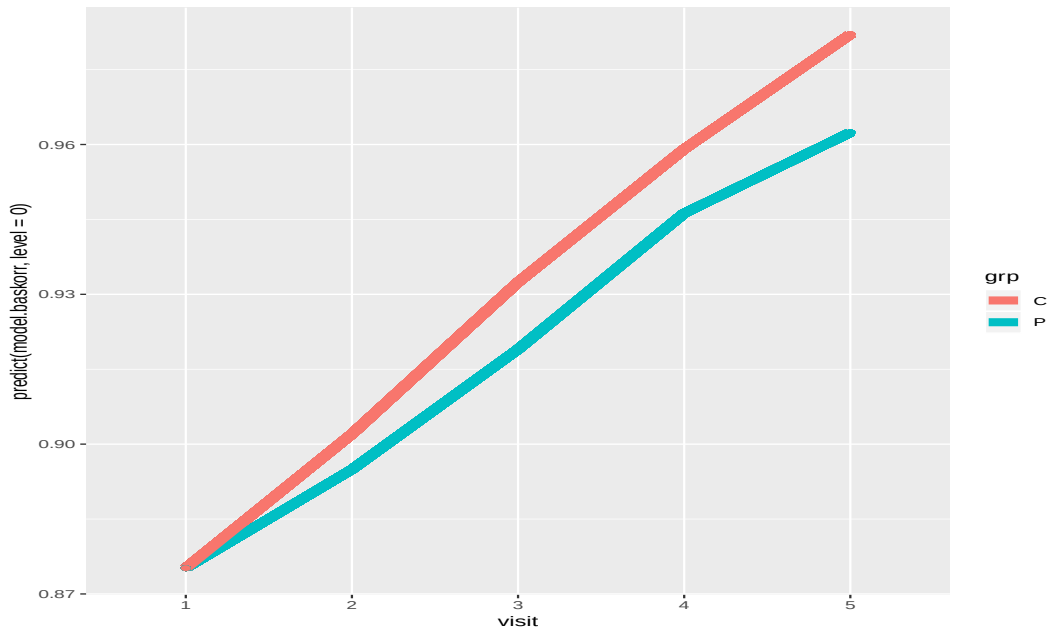
Her er der lavet 4 dummy-variable, så A-gruppen kan "*få lov til*" at adskille sig fra P-gruppen fra anden til femte måling.

Estimeret fordel efter 2 år blev her: 0.0196 (0.0044) g pr cm^3

(se output s. 113)

Predikterede forløb, med ens baselines

Se tilsvarende kode s. 89 og 112.



Vi ser her, at gruppernes middelværdi tvinges til at starte på samme niveau, så nu er der justeret for baseline

Forskellige vurderinger af forskelle i tidsudvikling

Her i form af Estimerede forskelle efter 2 år:

Uden hensyntagen til baseline :

simpel model, $s.42$	0.0295	(0.0130)
random regression, $s.72$	0.0293	(0.0140)

Med hensyntagen til baseline :

5 visits, ens baseline, $s.90$	0.0196	(0.0044)
RR, ens baseline, $s.87$	0.0175	(0.0062)
baseline som kovariat, $s.84$	0.0174	(0.0063)
....		
Differenser, $s.85$	0.0181	(0.0062)
Sammenligning af		
rå tilvækster, $s.35$	0.0190	(0.0065)

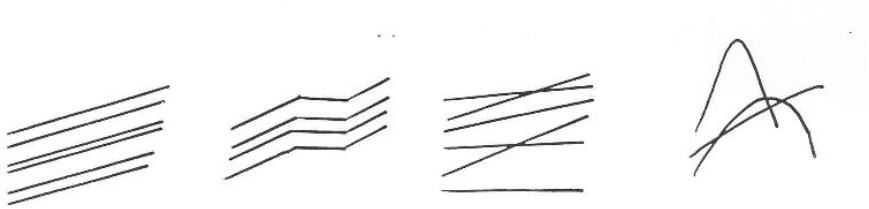
Effekt af at korrigere for baseline

- Noget af (men ikke hele) forskellen mellem grupperne efter 2 år kan "*bortforklares*" ved, at grupperne har forskelligt udgangspunkt (baseline værdi):
- Samtidig forøges præcisionen af den estimerede forskel (standard error bliver mindre)

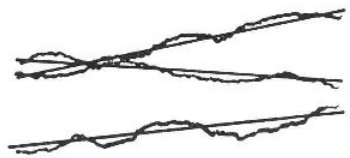
Forskellen bliver overbevisende signifikant
ligesom da vi så på de rå differenser

Variationskilder

1. Tilfældige/stokastiske/random effects:



2. Seriel korrelation ('korrelationsmønster')



3. Målefejl

Flere anvendelser af Mixed Models

- Udnyttelse af alle observationer i en parret sammenligning med manglende værdier
- Cross-over studier, hvor forskellige behandlinger afprøves på samme individ, i forskellig rækkefølge
- Håndtering af flere forløb for hvert individ, f.eks. før og efter en behandling, eller ved forskellig træning (kode s. 114)
  - Her skal der tages hensyn til korrelationen mellem samtlige målinger på samme person
  - men vi må forvente højere korrelation mellem målinger foretaget i samme situation
- Adskillelse af individuelle effekter og populations-effekter

Vi kører et Mixed-kursus i nov/dec

APPENDIX

Programbidder svarende til diverse slides:

- Plots: s. 96, 100, 108, 111
- Estimation i varianskomponentmodel: s. 99, 101
- Prediktion og modelkontrol: s. 102-103, 108, 111-112
- Alternative kovariansstrukturer: s. 104-106
- Random regression: s. 107, 111
- Baseline håndtering: s. 109-113

Spaghettiplot

Slide 4, venstre del

```
calcium <- read.table("calcium_lang.txt", header=T)

cal.P <- subset(calcium, calcium$grp=="P",)
cal.C <- subset(calcium, calcium$grp=="C",)

library(doBy)
calcium$missing = 0

miss=summaryBy(bmd ~ girl, data = calcium,
  FUN = function(x) { min = min(x)})

dropout = rep(as.numeric(is.na(miss[2])),each=5)
library(lattice)
calcium$dropout.col <- factor(dropout,levels=c(0,1),labels=c("red","blue"))

library(ggplot2)

xx <- ggplot(calcium,aes(x=visit,y=bmd,group=girl,color=dropout.col))+geom_line
xx+facet_wrap(~grp)+theme(legend.position="none")
```

Omstrukturering af data

fra bredt til langt format:

Slide 16

```
kanin <- read.table("kanin.txt", header=T)

install.packages("data.table")
library(data.table)

lang <- melt(kanin, id.vars = c("rabbit"),
        measure=c("a","b","c","d","e","f"),
        variable.name = "sted", value.name = "swelling")
```

ANOVA, sammenligning af kaniner

Slide 19-20

- Hver kanin har *et niveau* (en middelværdi)
- Herudover er der *variation mellem indstikssteder*

I computer sprog:
Kaninen er en faktor, analysen er en ensidet variansanalyse (ANOVA)

```
gal.model = lm(swelling ~ relevel(factor(rabbit),ref="6"),
            data=lang)

anova(gal.model)
```

Estimation i varianskomponentmodel

Slide 23

```
install.packages("nlme")
library(nlme)

model1 = lme(swelling ~ 1 , data=lang, random=~1 | rabbit)
intervals(model1)
```

Figur af gennemsnitskurver

Slide 33

```
library(ggplot2)

gg.base <-ggplot(calcium,aes(x = visit, y = bmd))

pdf("calcium_gennemsnit.pdf")
gg.base + stat_summary(aes(group = grp, color = grp),
geom = "line", fun.y = mean, size = 3)
```

Mixed model for Calcium eksempel

Slide 37

```
calcium$girl = as.factor(calcium$girl)
calcium$visit = as.factor(calcium$visit)

calcium$visit.r = relevel(as.factor(calcium$visit),ref="5")
calcium$grp.r = relevel(calcium$grp,ref="P")

#giver ikke samme estimater som SAS, samme test for interaktion,

model2 = lme(bmd ~ grp.r*visit.r , data=calcium,
         na.action=na.exclude, random=~1 | girl)

anova(model2)
intervals(model2)
```

Predikterede forløb

Slide 43-44

```
xx <- ggplot(calcium, aes(x=visit,
             y=predict(model2,level=0),
             group=girl,col=grp))
xx <- xx + geom_line() + geom_point()
xx

xx <- ggplot(calcium,aes(x=visit,
             y=predict(model2,level=1),
             group=girl,color=dropout.col))
    + geom_line() + geom_point()
xx+facet_wrap(~grp)+theme(legend.position="none")
```

Her er vrøvl med farverne

Modelkontrol i mixed model

Slide 47-48

```
par(mfrow=c(2,2))
plot(predict(model2,level=0),resid(model2,level=0))
hist(resid(model2,level=0))
qqnorm(resid(model2,level=0))
qqline(resid(model2,level=0))
```

```
par(mfrow=c(2,2))
plot(predict(model2),resid(model2))
hist(resid(model2))
qqnorm(resid(model2))
qqline(resid(model2))
```

Specifikation af CS-struktur

af model fra s. 37

Slide 51-53

```
model.cs = gls(bmd ~ grp*visit, data=calcium,
               na.action=na.exclude,
               corr = corCompSymm(form=~1 | girl))
summary(model.cs)
```

Specifikation af kovariansstrukturer

Slide 54, 57, 58

Typen er her Unstructured

```
model.un = gls(bmd ~ grp*visit, data=calcium,
   na.action=na.exclude,
   corr = corSymm(form=~1 | girl),
   weight = varIdent(form=~1 | visit),method="REML")
```

Typen er her AR(1)

```
model.ar1 = gls(bmd ~ grp*visit, data=calcium,
   na.action=na.exclude,
   corr = corAR1(form=~1 | girl),method="REML")
```

Specifikation af autoregressiv struktur

med overlejret tilfældigt niveau

Slide 57, 59

Det ved jeg ikke, hvordan man gør i R....

Random regression, med ny tid

så nulpunktet svarer til randomisering

Slide 66-72

```
c.tid$grp.r = relevel(c.tid$grp,ref="P")
library(nlme)

model.rr = lme(bmd ~ grp.r*years, data=c.tid,
  na.action=na.exclude, random=~years | girl)
```

Predikterede forløb fra random regression

Slide 79

```
xyplot(predict(model.rr, level=0) ~ years,
       group = grp, data = c.tid, type = "b")
```

Med baseline som kovariat

Slide 84

Variablen `baseline` er værdien ved første besøg.

Herefter analyseres kun de 4 follow-up visits,
her i en "*random regression*" model:

```
cal2.4 <- subset(c.tid, c.tid$tid > 0,)

model.bas = lme(bmd ~ baseline + grp.r*years, data=cal2.4,
  na.action=na.exclude, random=~years | girl)

summary(model.bas)
```

Analyse af differenser

Slide 85

```
cal2.4$bmd.afv = cal2.4$bmd - cal2.4$baseline

model.diff = lme(bmd.afv ~ years + grp.r*years, data=cal2.4,
  na.action=na.exclude, random=~years | girl)
```

med output

```
Fixed effects: bmd.afv ~ years + grp.r * years
                   Value   Std.Error  DF   t-value p-value
(Intercept)  -0.00196901 0.002665294 282 -0.738759  0.4607
years         0.04623234 0.002278427 282 20.291340  0.0000
grp.rC        0.00338849 0.003797621 103  0.892266  0.3743
years:grp.rC  0.00732989 0.003263404 282  2.246087  0.0255

> coef(model.diff)[1,3]+2*coef(model.diff)[1,3]
[1] 0.01016547
```


Random regression, med ens baseline

Slide 87-88

Ingen grp i model-sætningen, dvs. ingen gruppeforskel ved baseline (randomiseringstidspunkt)

```
model.rr2 = lme(bmd ~ years + grp.r:years, data=c.tid,
  na.action=na.exclude, random=~years | girl)

summary(model.rr2)


xyplot(predict(model.rr2, level=0) ~ years,
       group = grp, data = c.tid, type = "b")
```

Modellen s. 37, baseline justeret

Slide 89-90

```
c.tid$A.visit2 = (c.tid$grp=="C")*(c.tid$visit=="2")
c.tid$A.visit3 = (c.tid$grp=="C")*(c.tid$visit=="3")
c.tid$A.visit4 = (c.tid$grp=="C")*(c.tid$visit=="4")
c.tid$A.visit5 = (c.tid$grp=="C")*(c.tid$visit=="5")

model.baskorr = lme(bmd ~ visit + A.visit2 + A.visit3
  + A.visit4 + A.visit5, data=c.tid,
  na.action=na.exclude, random=~1 | girl)


xx <- ggplot(calcium,aes(x=visit,
   y=predict(model.baskorr,level=0),group=girl,col=grp))
xx <- xx + geom_line(size=3) + geom_point()
xx
```

Output fra model s. 90

Slide 89-90

```
Random effects:
 Formula: ~1 | girl
        (Intercept)   Residual
StdDev:  0.06652444 0.01531995

Fixed effects: bmd ~ visit + A.visit2 + A.visit3 + A.visit4 + A.visit5
                 Value   Std.Error  DF   t-value p-value
(Intercept) 0.8751696 0.006450500 381 135.67470  0.0000
visit2      0.0197593 0.002956205 381   6.68402  0.0000
visit3      0.0438201 0.002999566 381  14.60883  0.0000
visit4      0.0710566 0.003062552 381  23.20177  0.0000
visit5      0.0872419 0.003083655 381  28.29173  0.0000
A.visit2    0.0070952 0.004175315 381   1.69931  0.0901
A.visit3    0.0134284 0.004272098 381   3.14328  0.0018
A.visit4    0.0128561 0.004349124 381   2.95602  0.0033
A.visit5    0.0196441 0.004397013 381   4.46759  0.0000
```

Estimeret fordel efter 2 år: 0.0196 (0.0044) g pr cm^3

Flere forløb for hvert individ

Slide 94

- To grupper af patienter
- Hver patient undersøgt før og efter en behandling/måltid/træning el.lign. (`situation`)
- Målinger over tid (variabel `konttid`)

Her mangler foreløbig en generisk R-kode