

Basal statistik

Korrelerede målinger

Lene Theil Skovgaard

11. november 2019

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

Hjemmesider: http://biostat.ku.dk/~lts/basal19_2

E-mail: ltsk@sund.ku.dk

1 / 121

2 / 121

En gennemgående antagelse hidtil

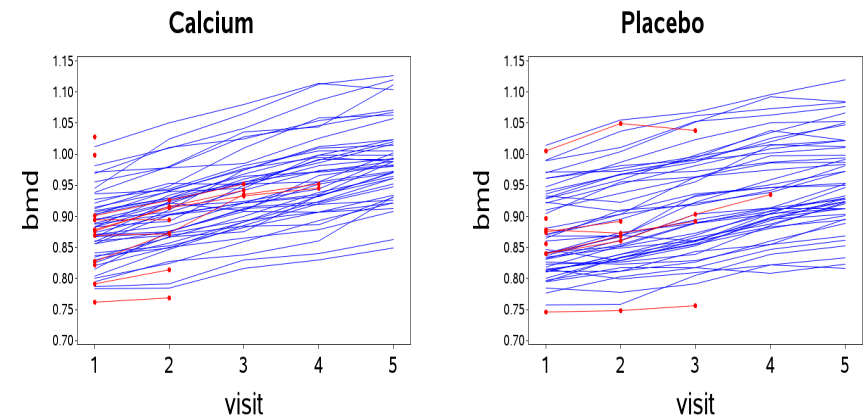
Uafhængighed

- ▶ 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 (vanskelig kode, se s. 95)



3 / 121

4 / 121

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?

5 / 121



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
.	.	.	.	.

6 / 121



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.

7 / 121



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,...)

8 / 121



Tilfældige effekter

Variationskilder (varianskomponenter)

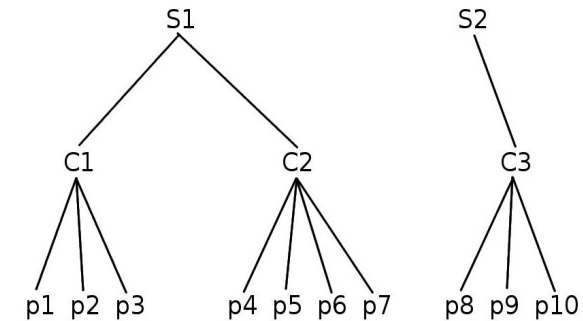
- ▶ geografisk/miljømæssig variation
 - ▶ mellem regioner, hospitaler, skoler eller lande
- ▶ biologisk variation
 - ▶ variation mellem individer, familier eller dyr
- ▶ variation mellem målinger på samme individ (within-individual)
 - ▶ variation mellem arme, tænder, injektionssteder, (dage)
- ▶ variation,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

9 / 121



Cluster design (hierarkisk design)

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

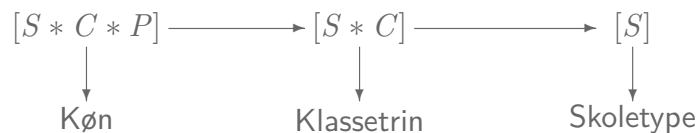
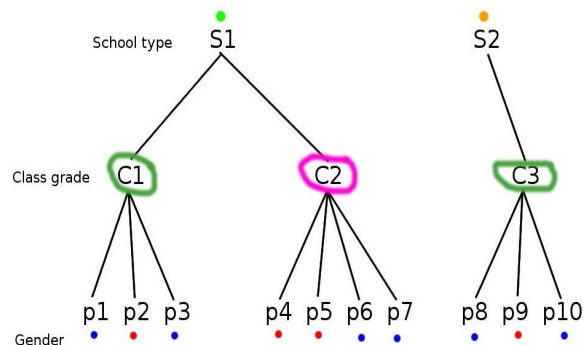


$$[I] = [S * C * P] \longrightarrow [S * C] \longrightarrow S$$

10 / 121



Cluster design, med kovariater



11 / 121



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.

12 / 121



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 estimer 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!

13 / 121



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 ressourcerne 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?

14 / 121



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

$r = 1, \dots, R=6$ angiver kaninen (rabbit),

$s = 1, \dots, S=6$ 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.

15 / 121



Data - omstrukturering, se s. 96

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 (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
. . .
```

16 / 121

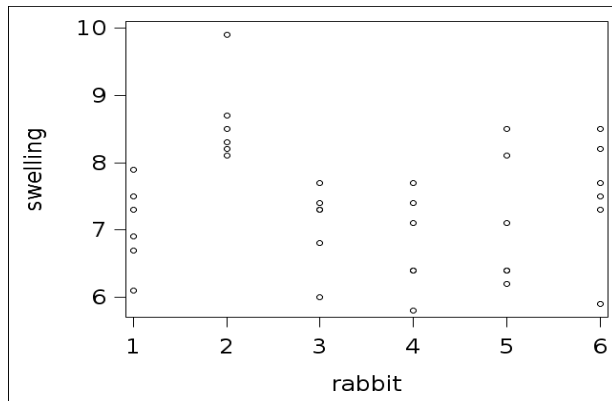


Scatter plot

36 observationer, 2 variable:

Outcome: Hævelse = swelling

X-akse: Arbitrær nummerering af kaniner



17 / 121

Naiv kvantificering af hævelse

The MEANS Procedure

Variable	N	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean
swelling	36	7.3666667	0.9313585	0.1552264	7.0515403	7.6817930

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?

18 / 121

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:

```
proc glm data=rabbit;
  class rabbit;
  model swelling=rabbit / solution;
run;
```

med output

Dependent Variable: swelling

R-Square	Coeff Var	Root MSE	swelling Mean		
0.422705	10.37571	0.764344	7.366667		
Source	DF	Type III SS	Mean Square	F Value	Pr > F
rabbit	5	12.8333333	2.5666667	4.39	0.0040

19 / 121

Output, fortsat

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	7.51666667 B	0.31204226	24.09	<.0001
rabbit 1	-0.45000000 B	0.44129439	-1.02	0.3160
rabbit 2	1.10000000 B	0.44129439	2.49	0.0184
rabbit 3	-0.43333333 B	0.44129439	-0.98	0.3340
rabbit 4	-0.71666667 B	0.44129439	-1.62	0.1148
rabbit 5	-0.40000000 B	0.44129439	-0.91	0.3719
rabbit 6	0.00000000 B	.	.	.

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**

20 / 121

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*

21 / 121



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*.

22 / 121



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)

Koden bliver:

```
proc mixed data=rabbit;
  class rabbit;
  model swelling = / s cl;
  random rabbit;
run;
```

23 / 121



Output fra varianskomponentmodel

Output (se kode s. 23)

Covariance Parameter Estimates

Cov Parm	Estimate
rabbit	0.3304
Residual	0.5842

Effect	Estimate	Standard Error	DF	t Value	Lower	Upper
Intercept	7.3667	0.2670	5	27.59	6.6803	8.0530

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*.

24 / 121



Fortolkning af varianskomponenter

Værdierne er taget fra output s. 24:

Variation	Varianskomponent	Estimat	Andel af variationen
Between	ω_B^2	0.3304	36%
Within	σ_W^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
 $\pm 2 \times \sqrt{2 \times 0.5842} = \pm 2.16 \text{ cm}^2$
- ▶ for to steder på **forskellige** kaniner
 $\pm 2 \times \sqrt{2 \times 0.9146} = \pm 2.70 \text{ cm}^2$

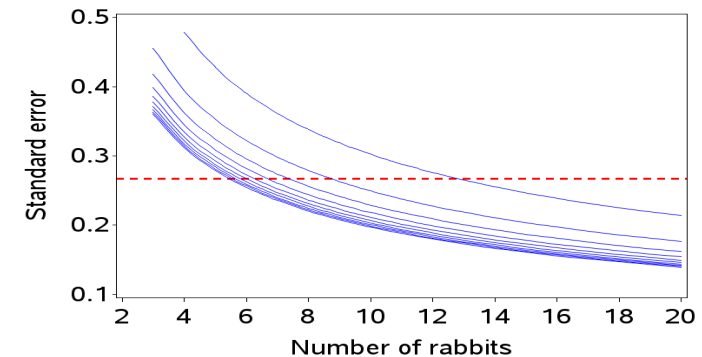
25 / 121



Standard error på estimat for hævelse

For **R=antal kaniner**, fra 3 til 20:

For **S=antal injektionssteder**, fra 1 til 10:



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

26 / 121



*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?

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

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

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

27 / 121



Longitudinelle målinger

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

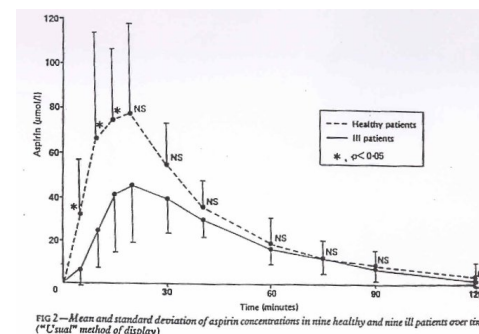


FIG 2—Mean and standard deviation of aspirin concentrations in nine healthy and nine ill patients over time. ("Cusum" method of display)

T-tests for hvert tidspunkt:

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

28 / 121



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

29 / 121



Individuelle tidsprofiler

Spaghettiplot

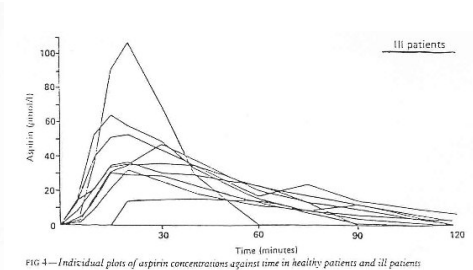
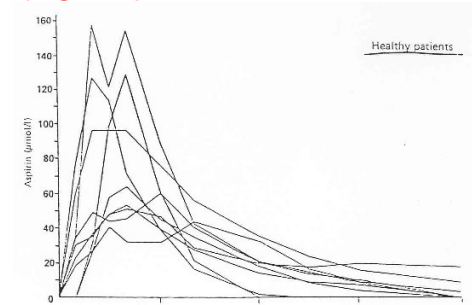


FIG 4—Individual plots of aspirin concentrations against time in healthy patients and ill patients

Har de samme facon allesammen?

Er gennemsnittet repræsentativt?

30 / 121



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

31 / 121



Fordele og ulemper ved de simple analyser

Fordele:

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

Ulemper:

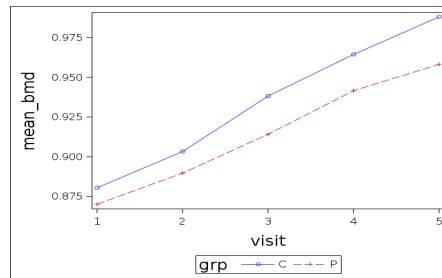
- ▶ Informationstab
- ▶ Kræver et passende antal observationer pr. individ
- ▶ Kræver et fornuftigt karakteristika og ofte målinger på ens tidspunkter
- ▶ Umuligt at inddrage tidsafhængige kovariater
- ▶ Kan ikke tage hensyn til baseline

32 / 121



Eksempel: Calcium tilskud

Gennemsnit for de 2 grupper, (se kode s. 99)



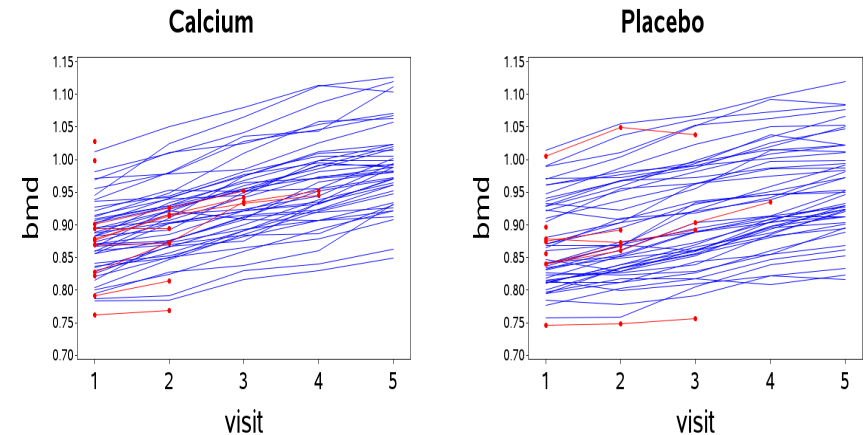
men er gennemsnitskurver rimeligt her?

Ja, nogenlunde....pga de ret ensartede forløb, se næste side,
men der er missing values... og baseline issues

33 / 121



Individuelle profiler - igen (som s.4)



34 / 121



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:
0.0190 (0.0062, 0.0318), 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...

35 / 121



Struktur for Calcium-eksemplet

	Unit/Enhed	Variation	Kovariater
Level 2	piger	<i>mellem</i> piger ω_B^2	grp
Level 1	enkeltobservationer	<i>indenfor</i> 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,

36 / 121



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*

```
proc mixed data=calcium;
class grp girl visit;
model bmd=visit grp*visit grp
      / ddfm=kr2 vciry s cl;
random girl(grp);
run;
```

37 / 121



Bemærkning til modelspecifikation

- ▶ Pigerne er **nestet** i grupperne,
Vurdering af gruppeforskelle er uparret
Nestingen skrives som `random girl(grp);`
eller (med lidt mere effektiv beregningsmetode):
`random intercept / subject=girl(grp);`
- ▶ Alle visits forekommer (i princippet) for den enkelte pige,
Vurdering af tidseffekter er parrede
- ▶ Det kryptiske `ddfm=kr2`, se s. 101
- ▶ `vciry`: giver skalerede residualer (teknisk - tager højde for korrelationen indenfor individer)

38 / 121



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)

39 / 121



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)

40 / 121



Output fra mixed model, kode s. 37

(idet der dog af pladshensyn ikke er konfidensgrænser på estimaterne på næste side....)

Cov Parm	Subject	Estimate
Intercept	girl(grp)	0.004439
Residual		0.000235

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
visit	4	382	619.36	<.0001
grp*visit	4	382	5.30	0.0004
grp	1	110	2.63	0.1078

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

... **hvis modellen altså er rimelig**

41 / 121

Output, fortsat

Solution for Fixed Effects

Effect	grp	visit	Estimate	Error	DF	t Value	Pr > t
Intercept			0.9576	0.009131	122	104.87	<.0001
visit		1	-0.08750	0.003100	382	-28.22	<.0001
visit		2	-0.06748	0.003103	381	-21.75	<.0001
visit		3	-0.04342	0.003117	381	-13.93	<.0001
visit		4	-0.01619	0.003148	381	-5.14	<.0001
visit		5	0	.	.	.	.
grp*visit	C	1	0.01038	0.01292	118	0.80	0.4232
grp*visit	C	2	0.01696	0.01296	120	1.31	0.1934
grp*visit	C	3	0.02329	0.01300	121	1.79	0.0756
grp*visit	C	4	0.02272	0.01302	122	1.74	0.0836
grp*visit	C	5	0.02951	0.01304	122	2.26	0.0254

Se bemærkninger til output på s. 45

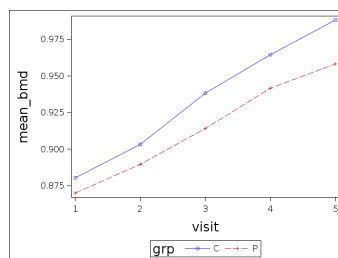
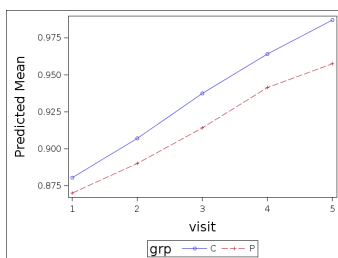
42 / 121

Figur af predikterede kurver

svarer ikke helt til gennemsnittene fra s. 33

Prediktioner:

Gennemsnit:



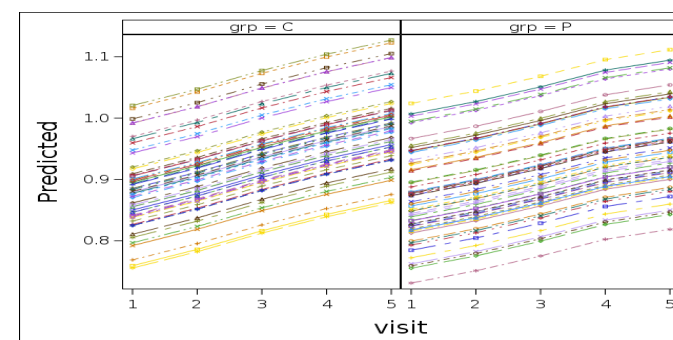
Dette skyldes de enkelte missing values....

Kode til venstre figur: s. 102 (variablen Pred i datasættet predm)

43 / 121

Predikterede individuelle forløb

som inkluderer de **tilfældige niveauer** for pigerne (variablen Pred i datasættet predi, se kode s. 102)



44 / 121

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 sluttidspunktet er 0.02951 (0.01304) i C-gruppens favør (aflæses ud for `grp*visit C 5`) samt at dette er signifikant ($P=0.0254$)
- ▶ Forskellen stiger over tid, hvilket også er forventeligt ved kontinuerlig behandling
- ▶ Intra-individ korrelationen er 0.95 (se s. 52)

45 / 121



Modelkontrol

To typer residualer bør checkes:

- De sædvanlige Observeret minus predikteret (**gruppe**-)middelværdi (kun systematiske effekter fratrækkes), gerne i *skaleret* version, hvis det er muligt
- De betingede Observeret minus predikteret **individuel** prediktion (både systematiske og tilfældige effekter fratrækkes), *Conditional* (betinget med individet)

Vi ser på *de sædvanlige* tegninger:

- ▶ Residualer plottet mod predikterede værdier
- ▶ Histogram og fraktildiagram

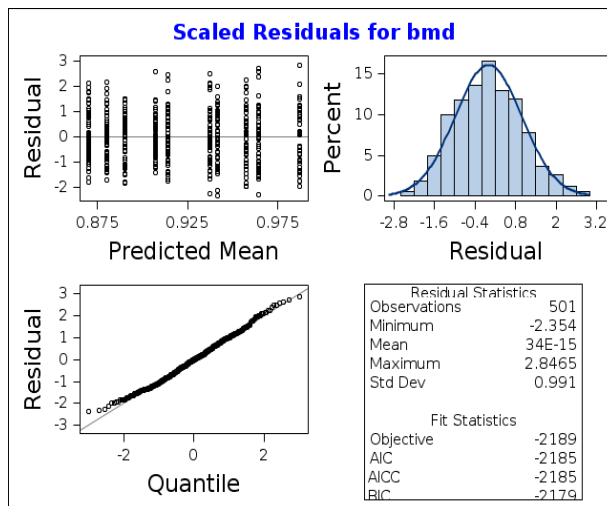
Se figurerne s. 47 og 48 (kode s. 102)

46 / 121



Modelkontrol, sædvanlige (store) residualer

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

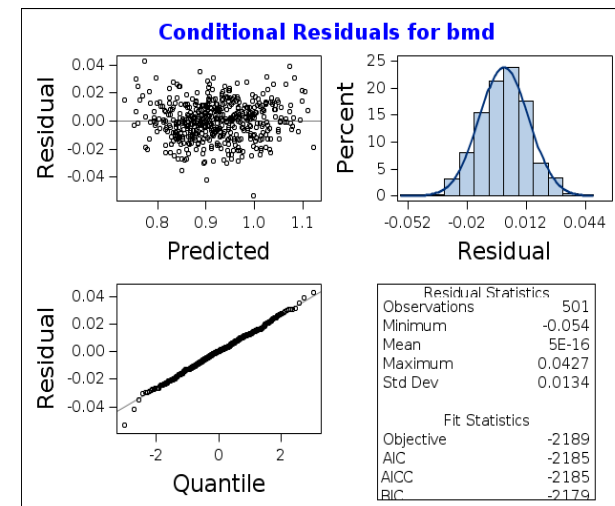


47 / 121



Modelkontrol, betingede (små) residualer

Afvigelse fra **individuelle** prediktioner (se s. 46):



48 / 121



Effekt af korrelerede observationer

Når korrelationen **ignorerer**, sker der følgende:

- ▶ Level 1 kovariater (tidsrelaterede effekter, “*within*”):
Lav styrke (**type 2 fejl**):
Vi opdager ikke de effekter, der er der.
Her drejer det sig om $\text{grp} \times \text{visit}$, samt (hvis denne ikke er med i modellen) om selve visit
- ▶ Level 2 kovariater (behandlinger, gruppe, “*between*”):
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).

49 / 121



Fejlagtig analyse

Tosidet ANOVA, **korrelation ignoreret**

Source	DF	Type III SS	Mean Square	F Value	Pr > F
grp	1	0.05063714	0.05063714	11.05	0.0010
visit	4	0.64369784	0.16092446	35.10	<.0001
grp*visit	4	0.00649557	0.00162389	0.35	0.8411

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....

50 / 121



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 koder, se s. 105-106

51 / 121



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. 41

$$\hat{\rho} = \frac{0.004439}{0.004439 + 0.000235} \approx 0.95$$

Observationerne er **ombyttelige=exchangeable**, altså **CS: Compound Symmetry**

52 / 121



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. 106

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?

53 / 121



Valg af varians- og korrelationsstruktur?

Den mest generelle: **Ustruktureret**, benytter en repeated-sætning i stedet for random-sætningen

```
repeated visit / type=UN subject=girl(grp) rcorr;
```

(se hele koden s.107)

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*

Spredningsestimater (måske svagt stigende):

Visit 1	Visit 2	Visit 3	Visit 4	Visit 5
0.0629	0.0688	0.0705	0.0730	0.0700

Korrelations struktur:

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.9699	0.9414	0.9250	0.8987
2	0.9699	1.0000	0.9727	0.9585	0.9399
3	0.9414	0.9727	1.0000	0.9809	0.9592
4	0.9250	0.9585	0.9809	1.0000	0.9755
5	0.8987	0.9399	0.9592	0.9755	1.0000

54 / 121



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 ustabil, 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"?

55 / 121



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. 107)
- ▶ **Autoregressiv** struktur, *samtidig* med den tilfældige effekt (niveau) for hvert individ, se kode s. 108
- ▶ 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*"

56 / 121



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.

I MIXED specificeres dette som `type=AR(1)`,
se kode s. 107

Hvis tiderne ikke er ækvidistante, bruger man SP(POW)(tid), svarende
til $\text{Corr}(Y_{git_1}, Y_{git_2}) = \rho^{|t_1 - t_2|}$

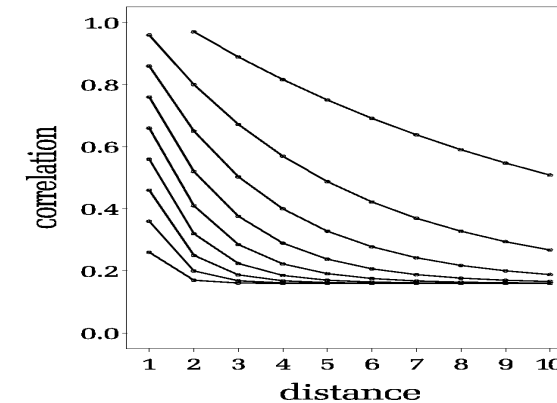
57 / 121



Autoregressiv korrelation

med overlejet tilfældigt niveau (s. 108)

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



58 / 121



Test af interaktionen grp*visit?

– for forskellige valg af kovariansstruktur

Kovarians struktur	Teststørrelse ~ fordeling	P-værdi
Uafhængighed	0.35 ~ F(4,491)	0.84
Compound symmetry	5.30 ~ F(4,382)	0.0004
Autoregressiv (evt. med CS oveni)	2.86 ~ F(4,383)	0.023
Ustruktureret	2.72 ~ F(4,107)	0.034

Altså ikke samme konklusion!

Kovariansstrukturen kan være vigtig, så hvordan vælger man?
Det vender vi tilbage til....s. 72

59 / 121



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

60 / 121



Tid siden randomisering

Se udregning af denne i appendix, s. 109-110

- ▶ 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*

61 / 121

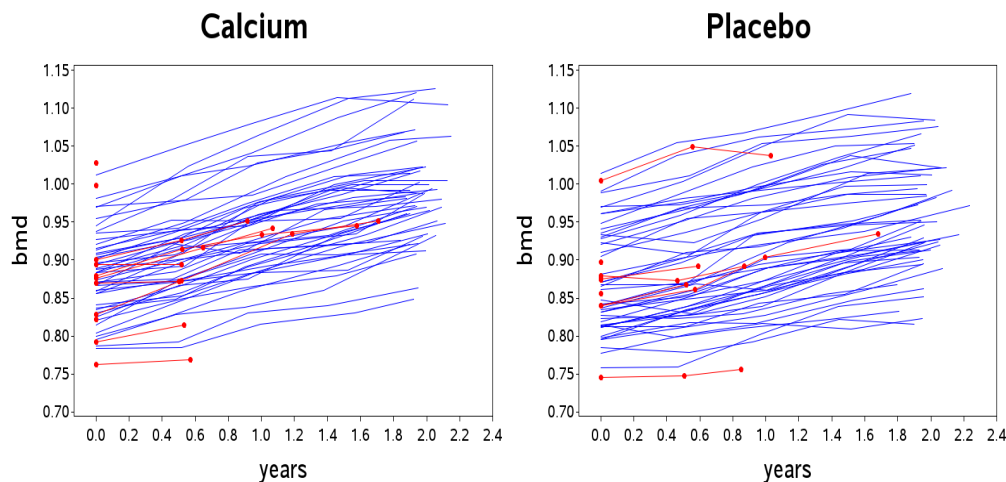
Overblik over nye tider

The MEANS Procedure

grp	visit	N		Mean	Minimum	Maximum
		Obs	Variable			
C	1	55	bmd	0.8804545	0.7620000	1.0280000
			years	0	0	0
	2	55	bmd	0.9032692	0.7690000	1.0510000
			years	0.5202970	0.4161533	0.7227926
	3	55	bmd	0.9382500	0.8160000	1.0800000
			years	0.9630390	0.8678987	1.1882272
	4	55	bmd	0.9645000	0.8300000	1.1140000
			years	1.4977234	1.3360712	1.7768652
	5	55	bmd	0.9881591	0.8490000	1.1260000
			years	1.9762927	1.8398357	2.1464750
P	1	57	bmd	0.8700702	0.7460000	1.0140000
			years	0	0	0
	2	57	bmd	0.8896792	0.7480000	1.0550000
			years	0.5182287	0.4462697	0.5995893
	3	57	bmd	0.9141569	0.7560000	1.0670000
			years	0.9584625	0.8487337	1.1307324
	4	57	bmd	0.9416458	0.8090000	1.0960000
			years	1.5017112	1.3415469	1.7056810
	5	57	bmd	0.9582340	0.8160000	1.1190000
			years	1.9759127	1.7960301	2.2340862

62 / 121

De nye (faktiske) individuelle forløb



63 / 121

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 (*visit*)
- ▶ Ustruktureret kovarians/korrelation (*type=UN*)

Men man kan stadig benytte

- ▶ CS-strukturen
- ▶ **random regression**, kommer nu
- ▶ Erstatte AR(1)-strukturen med SP(POW)(years), se s. 57

64 / 121

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:

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

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).

65 / 121



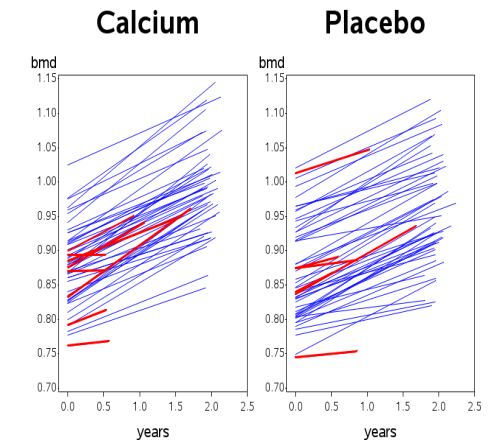
Stokastisk regression/Random regression

– 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...



66 / 121



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

67 / 121



Random regression i praksis

Her specificeres:

- ▶ To *vilkårlige linier* i model-sætningen (baseline-korrektion senere)
- ▶ En 2×2 -kovarians for de personspecifikke intercepter og hældninger (G) i random-sætningen
- ▶ De estimerede middelværdier udskrives til datasættet `predicted_mean`

```
proc mixed plots=all data=calcium;
class grp girl;
model bmd=grp years grp*years
      / ddfm=kr2 vciry s outpm=predicted_mean;
random intercept years /
      type=un subject=girl(grp) g vcorr;
estimate "forskel efter 2 aar" grp 1 -1 grp*years 2 -2;
run;
```

68 / 121



Dele af output fra random regression

G-matricen, med tilhørende korrelationsmatrix:
(option g og gcorr, kode s. 111)

Estimated G Matrix					
Row	Effect	grp	girl	Col1	Col2
1	Intercept	C	101	0.004178	0.000103
2	years	C	101	0.000103	0.000179

Estimated G Correlation Matrix					
Row	Effect	grp	girl	Col1	Col2
1	Intercept	C	101	1.0000	0.1194
2	years	C	101	0.1194	1.0000

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

Fit Statistics	
-2 Res Log Likelihood	-2351.4
AIC (Smaller is Better)	-2343.4

69 / 121



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):
(option v og vcorr, kode s. 111)

Variansestimater (svagt stigende):

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

70 / 121



Dele af output, III

Solution for Fixed Effects						
Effect	grp	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		0.8694	0.008644	110	100.58	<.0001
grp	C	0.01156	0.01233	110	0.94	0.3505 <----baseline forskel
grp	P	0	.	.	.	.
years		0.04532	0.002153	96.1	21.05	<.0001
years*grp	C	0.008887	0.003079	96.6	2.89	0.0048<----forskell i slopes
years*grp	P	0	.	.	.	.
Label		Estimate	Standard Error	DF	t Value	Pr > t
forskell efter 2 aar		0.02934	0.01403	108	2.09	0.0389

I denne model kvantificerer vi effekten af calciumtilskud (forskellen på de to hældninger) til **0.0089 (0.0031) g pr cm³ pr år**.

Estimeret **fordel efter 2 år**: 0.0293 (0.0140) g pr cm³

Ingen signifikant forskel ved baseline (mere om dette fra s. 77)

71 / 121



Sammenligning af modelfit

under antagelse om linearitet

-2 log L (og AIC) skal være lille, dvs. et stort negativt tal

Kovarians struktur	-2 log L	Kov.par.	AIC	Differens i hældning	P
Uafhængighed	-1252.4	1	-1250.4	0.0105 (0.0086)	0.22
Compound Symmetry	-2253.7	2	-2249.7	0.0089 (0.0020)	< 0.0001
Autoregressiv sp(pow)	-2373.6	2	-2369.6	0.0094 (0.0033)	0.0037
Random Regression	-2351.4	4	-2343.4	0.0089 (0.0031)	0.0048

Random regression ser rimelig ud (AIC lille), måske er SP(POW) dog en anelse bedre

72 / 121



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)

73 / 121



Sammenligning af de to fremgangsmåder:

Random regression eller individuelle regressioner

Forskel på hældninger C vs. P:

- ▶ Random regression: 0.0089 (0.0031), $P=0.0048$
P: 0.0453, C: 0.0542
- ▶ Individuelle regressioner: 0.0077 (0.0039), $P=0.049$
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 i C-gruppen, se s. 75

Dette kunne være indikation af **selektivt bortfald...uha!**

74 / 121



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:

Analysis Variable : years

grp	dropout	N Obs	N	Mean
C	0	44	44	0.0546824
	1	8	8	0.0364589
P	0	47	47	0.0458531
	1	6	6	0.0335133

Kode s. 112

75 / 121



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)

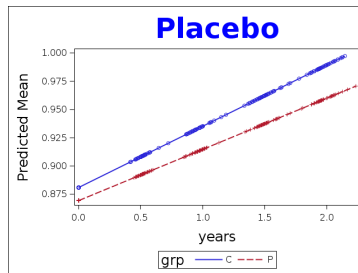
NI Non-ignorable: (**Informative missing**)
Missingness afhænger af
de **uobserverede** outcome værdier!!

76 / 121



Predikterede forløb fra random regression

Kode s. 113



Som vi tidligere har set, er der en vis forskel allerede fra starten (ved **baseline**), selv om denne er *insignifikant*, ($P=0.35$, se s. 71).

Bør vi justere for baseline?

77 / 121



Justering for baseline?

Baseline måling:

Observation foretaget inden (eller ved) behandlingsstart.

Vi diskuterede håndtering af sådanne i forbindelse med “Ancova”:

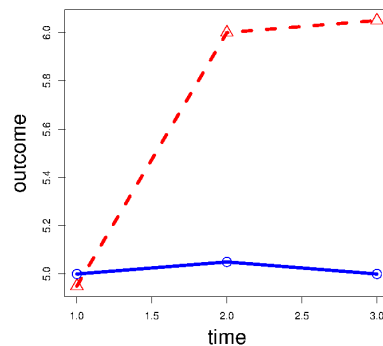
- ▶ Der skal *ikke* (nødvendigtvis) justeres i tilfælde af observationelle studier
- ▶ Justering **bør foretages**, hvis der er tale om **randomiserede undersøgelser**, fordi vi vil sammenligne to personer, som starter med at være ens, men som modtager forskellige behandlinger – og det er et *tilfælde*,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

78 / 121



Hypotetisk naiv sammenligning af to grupper, I

uden hensyntagen til “principielt ens” baselines:



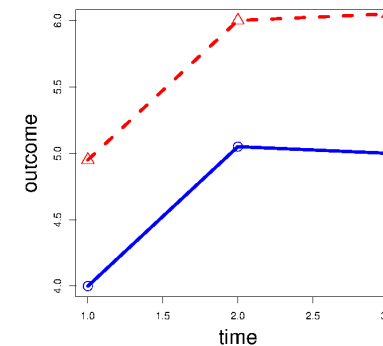
- ▶ **Konklusion:** Interaktion mellem tid og behandling
- ▶ **Sandhed:** Konstant forskel på de to behandlinger

79 / 121



Hypotetisk naiv sammenligning af to grupper, II

uden hensyntagen til “principielt ens” baselines



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

80 / 121



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.

81 / 121



Med baseline som kovariat (**ikke helt rimeligt**)

Nu er det kun de sidste 4 tidspunkter, der er outcome
Se kode s. 114

Solution for Fixed Effects						
Effect	grp	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		-0.07256	0.02451	102	-2.96	0.0038
baseline		1.0811	0.02801	101	38.60	<.0001
grp	C	0.002857	0.003643	100	0.78	0.4348
grp	P	0	.	.	.	.
years		0.04624	0.002279	93.4	20.29	<.0001
years*grp	C	0.007284	0.003265	93.5	2.23	0.0281
years*grp	P	0	.	.	.	.

Label	Estimate	Standard Error	DF	t Value	Pr > t
forskel efter 2 aar	0.01742	0.006253	99.5	2.79	0.0064

Estimeret **fordel efter 2 år**: 0.0174 (0.0063) g pr cm³

82 / 121



Analyse af differenser (**ikke helt rimeligt**)

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

Se kode s. 115

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. 82

83 / 121



Fælles middelværdi ved baseline, I

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

Udelad grp af modelsætningen, men behold interaktionen:

```
proc mixed data=calcium;
class grp girl;
model bmd=years grp*years / ddfm=kr2 vciry s;
.....
run;
```

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

84 / 121



Random regression, med ens baseline

Udelad grp af modellen, men behold interaktionen

```
proc mixed data=calcium;
class grp girl;
model bmd=years grp*years / ddfm=kr2 vciry s;
random intercept years / type=un subject=girl(grp) g;
estimate "forskkel efter 2 aar" grp*years 2 -2;
run;
```

med output

Solution for Fixed Effects

Effect	grp	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		0.8751	0.006163	111	141.99	<.0001
years		0.04538	0.002162	96.3	20.99	<.0001
years*grp	C	0.008758	0.003104	96.9	2.82	0.0058
years*grp	P	0	.	.	.	.

Label	Estimate	Standard Error	DF	t Value	Pr > t
forskkel efter 2 aar	0.01752	0.006209	96.9	2.82	0.0058

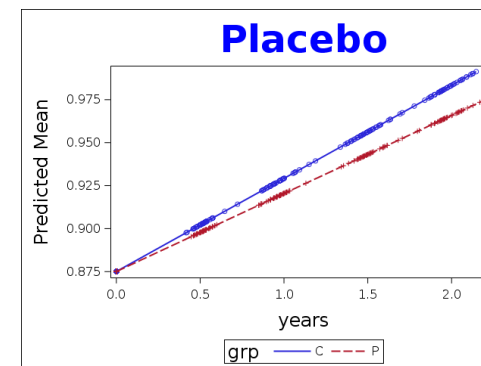
Estimeret **fordel efter 2 år**: 0.0175 (0.0062) g pr cm³

85 / 121



Predikterede forløb, med ens baselines

Se tilsvarende kode s. 118 og output s. 117



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

86 / 121



Fælles middelværdi ved baseline, II

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

```
....
adj_grp=grp;
if visit=1 then adj_grp='P';
run;

proc mixed data=calcium;
class adj_grp visit girl;
model bmd=visit adj_grp*visit / ddfm=kr2 vciry s;
...
run;
```

Her er gruppen omdefineret, fra grp to adj_grp (adj=adjusted), så **alle piger tilhører P-gruppen ved første måling**.

87 / 121



Output fra model s. 87

Solution for Fixed Effects

Effect	adj_grp	visit	Estimate	Standard Error	DF	t Value	Pr > t
Intercept			0.9624	0.006851	150	140.48	<.0001
visit		1	-0.08724	0.003085	390	-28.28	<.0001
visit		2	-0.06748	0.003103	381	-21.75	<.0001
visit		3	-0.04342	0.003117	381	-13.93	<.0001
visit		4	-0.01619	0.003148	381	-5.14	<.0001
visit		5	0	.	.	.	.
adj_grp*visit	C	2	0.007095	0.004178	400	1.70	0.0902
adj_grp*visit	C	3	0.01343	0.004274	399	3.14	0.0018
adj_grp*visit	C	4	0.01286	0.004351	399	2.95	0.0033
adj_grp*visit	C	5	0.01964	0.004399	398	4.47	<.0001
adj_grp*visit	P	1	0	.	.	.	.
adj_grp*visit	P	2	0	.	.	.	.
adj_grp*visit	P	3	0	.	.	.	.
adj_grp*visit	P	4	0	.	.	.	.
adj_grp*visit	P	5	0	.	.	.	.

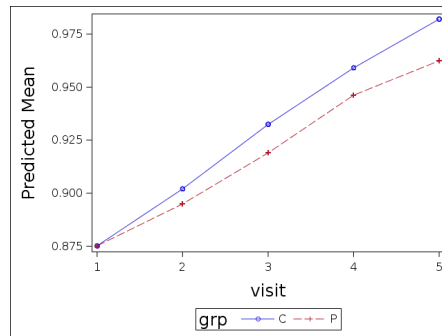
Bemærk, at der her kun er forskel på grupperne fra visit=3 og frem.

88 / 121



Predikterede forløb, med ens baselines

Se tilsvarende kode s. 119



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

89 / 121



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.71	0.0293	(0.0140)

Med hensyntagen til baseline :

5 visits, ens baseline, s.89	0.0196	(0.0044)
RR, ens baseline, s.85	0.0175	(0.0062)
baseline som kovariat, s.82	0.0174	(0.0063)
....		
Differenser, s.83	0.0181	(0.0062)
Sammenligning af rå tilvækster, s.35	0.0190	(0.0065)



90 / 121

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

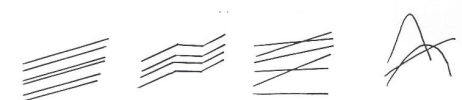
ligesom da vi så på de rå differenser

91 / 121



Variationskilder

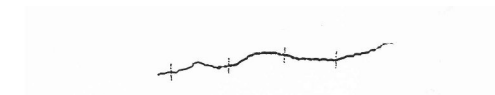
1. Tilfældige/stokastiske/random effects:



2. Seriel korrelation ('korrelationsmønster')



3. Målefejl



Se også noter til kodning s. 120

92 / 121



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. 121)
 - ▶ 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

93 / 121



APPENDIX

Programbidder svarende til diverse slides:

- ▶ Plots: s. 95, 99, 113, 117
- ▶ Estimation i varianskomponentmodel: s. 98, 100, 105
- ▶ Prediktion og modelkontrol: s. 102-104, 113, 117
- ▶ Alternative kovariansstrukturer: s. 106-108
- ▶ Random regression: s. 111, 117
- ▶ Baseline håndtering: s. 117-119

94 / 121



Spaghettiplot

Slide 4, venstre del

Nedenstående kode forudsætter, at pigerne er sorteret, så de 44 med fuldt observationssæt kommer først, og derefter de 11, der dropper ud på diverse tidspunkter:

```
title H=4 "Calcium";
proc gplot gout=plotud data=calcium; where grp='C';
  plot bmd*visit=girl
    / nolegend haxis=axis1 vaxis=axis2 frame;
axis1 offset=(5,5) value=(H=3) minor=NONE label=(H=4);
axis2 order=(0.7 to 1.15 by 0.05)
  value=(H=2) minor=NONE label=(A=90 R=0 H=4);
symbol1 v=none i=join c=blue l=1 w=1.5 r=44;
symbol2 v=dot i=join c=red l=1 w=1.5 r=11;
run;
```

95 / 121



Omstrukturering af data

fra bredt til langt format:

Slide 16

```
data rabbit;
set bredt;
rabbit=1; swelling=a; output;
rabbit=2; swelling=b; output;
rabbit=3; swelling=c; output;
rabbit=4; swelling=d; output;
rabbit=5; swelling=e; output;
rabbit=6; swelling=f; output;
run;
```

96 / 121



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)

```
proc glm data=rabbit;
  class rabbit;
  model swelling=rabbit / solution;
run;
```

97 / 121



Estimation i varianskomponentmodel

Slide 23

```
proc mixed data=rabbit;
  class rabbit;
  model swelling = / s cl;
random rabbit;
run;
```

Options:

- ▶ s står for solution og bevirker, at man får parameterestimerer
- ▶ cl betyder, at man også gerne vil have konfidensintervaller på estimerererne

98 / 121



Figur af gennemsnitskurver

Slide 33

```
proc sort data=calcium; by grp visit;
run;

proc means nway mean data=calcium; by grp visit;
var bmd;
output out=ny mean=mean_bmd;
run;

proc sgplot data=ny;
series Y=mean_bmd X=visit / group=grp markers;
run;
```

99 / 121



Mixed model for Calcium eksempel

Slide 37

```
proc mixed data=calcium;
class grp girl visit;
model bmd=grp visit grp*visit / ddfm=kr2 vciry s cl;
random girl(grp);
run;
```

- ▶ Pigerne er **nestet** i grupperne, det *kan* skrives girl(grp) i stedet for blot girl
- ▶ ddfm=kr (eller ddfm=satterth) kommenteres næste side

100 / 121



*Teknisk detalje

Slide 38

ddfm=kr (kenwardrogers - eller satterth):

- ▶ Når resultaterne er eksakte, ændrer det intet
 - ▶ i balancerede situationer
- ▶ Når det er nødvendigt med approksimationer, er det bedst at bruge en af disse to
 - ▶ i unbalancerede situationer, dvs. for langt de fleste observationelle studier
 - ▶ hvis der er manglende værdier
- ▶ Man får nogle underlige frihedsgrader, der ikke er hele tal
- ▶ Beregningerne tager lidt længere tid, men det vil I sikkert slet ikke lægge mærke til
- ▶ Hvis man er i tvivl, bør man bruge dem! (de skader ikke)

101 / 121



Predikterede forløb, og modelkontrol

Slide 43-44, 47-48

```
ods graphics on;
proc mixed plots=all data=calcium;
class grp girl visit;
model bmd=grp visit grp*visit / ddfm=kr2 vciry s cl outpm=predm outp=predi;
random girl(grp);
run;
```

```
proc sgplot data=predm; where girl in (101,102);
/* vi tegner kun for 1 pige i hver gruppe, da de predikteres ens */
series Y=Pred X=visit / group=grp markers;
run;
```

```
proc sgpanel data=predi;
panelby grp;
series Y=Pred X=visit / group=girl markers;
run;
```

102 / 121



Alternativ kode til predikterede forløb

Slide 43

her med konfidensgrænser

```
proc glimmix PLOTS=ALL data=calcium;
class grp girl visit;
model bmd=grp visit grp*visit / ddfm=kr2 vciry s;
random girl(grp);
lsmeans visit*grp
/ plots=(meanplot(join clband sliceby=grp));
run;
```

Bemærk, at der her er benyttet en anden procedure, nemlig

GLIMMIX

103 / 121



Modelkontrol i mixed model

Slide 47-48

Koden s. 102 giver automatisk siderne 47 og 48, men hvis man har behov for yderligere, kan man benytte datasættene predm og predi, der indeholder residualer og predikterede værdier, se. s. 102

104 / 121



Alternative kodninger af simpel mixed model

svarende til modellen fra s. 37

Slide 51-53

```
random girl(grp);

random intercept / subject=girl(grp);

repeated visit / type=CS subject=girl(grp);
```

CS: Compund Symmetry

105 / 121



Specifikation af CS-struktur

af model fra s. 37

Slide 52-53

```
proc mixed data=calcium;
  class grp girl visit;
  model bmd=grp visit grp*visit / ddfm=kr2 vciry s cl;
  repeated visit / type=CS subject=girl(grp) rcorr;
run;
```

Estimated R Correlation Matrix for girl(grp) 101 C

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.9498	0.9498	0.9498	0.9498
2	0.9498	1.0000	0.9498	0.9498	0.9498
3	0.9498	0.9498	1.0000	0.9498	0.9498
4	0.9498	0.9498	0.9498	1.0000	0.9498
5	0.9498	0.9498	0.9498	0.9498	1.0000

106 / 121



Specifikation af kovariansstrukturer

Slide 54, 56, 57

Typen er her Unstructured

```
proc mixed data=calcium;
  class grp girl visit;
  model bmd=grp visit grp*visit / ddfm=kr2 vciry s cl;
  repeated visit / type=UN HLM subject=girl(grp) rcorr;
run;
```

Typen er her AR(1)

```
proc mixed data=calcium;
  class grp girl visit;
  model bmd=grp visit grp*visit / ddfm=kr2 vciry s cl;
  repeated visit / type=AR(1) subject=girl(grp) rcorr;
run;
```

107 / 121



Specifikation af autoregressiv struktur

med overlejret tilfældigt niveau

Slide 56, 58

```
proc mixed data=calcium;
  class grp girl visit;
  model bmd=grp visit grp*visit / ddfm=kr2 vciry s cl;
  random intercept / subject=girl(grp);
  repeated visit / type=AR(1) subject=girl(grp) rcorr;
run;
```

Bemærk, at det her er nødvendigt at benytte *både* en random-sætning og en repeated-sætning

108 / 121



Udregning af tid siden randomisering

Slide 61-62

```
data calcium;

  infile cards missover;
  input dummy $ @10 visit1 date7. bmd1 @24 visit2 date7. bmd2
        @38 visit3 date7. bmd3 @52 visit4 date7. bmd4 @66 visit5 date7. bmd5;

  girl1=1*substr(dummy,1,3);
  grp=substr(dummy,4,1);
  drop dummy;

datalines;
  101C 01MAY90 0.815 05NOV90 0.875 24APR91 0.911 30OCT91 0.952 29APR92 0.970
  102P 01MAY90 0.813 05NOV90 0.833 15APR91 0.855 21OCT91 0.881 13APR92 0.901
  103P 02MAY90 0.812 05NOV90 0.812 17APR91 0.843 23OCT91 0.855 15APR92 0.895
  .
  .
  ;
run;
```

109 / 121



Udregning af tid siden randomisering, II

Slide 61-62

```
data a1; set calcium;
  tid=0;          baseline=bmd1; bmd=bmd1; visit=1; output;
  tid=visit2-visit1; baseline=bmd1; bmd=bmd2; visit=2; output;
  tid=visit3-visit1; baseline=bmd1; bmd=bmd3; visit=3; output;
  tid=visit4-visit1; baseline=bmd1; bmd=bmd4; visit=4; output;
  tid=visit5-visit1; baseline=bmd1; bmd=bmd5; visit=5; output;
run;

data calcium;
  set a1;

  years=tid/365.25;
run;

proc means N mean min max data=calcium;
  class grp visit;
  var bmd years;
run;
```

110 / 121



Random regression, med ny tid så nulpunktet svarer til randomisering

Slide 65-71

```
proc mixed data=calcium;
  class grp girl;
  model bmd=grp years grp*years / ddfm=kr2 vciry s outpm=predicted_mean;
  random intercept years /
    type=un subject=girl(grp)
    g gcorr v vcorr;
  estimate "forskøl efter 2 aar" grp 1 -1 grp*years 2 -2;
run;
```

- ▶ Der **skal** anvendes TYPE=UN i RANDOM-sætningen for at tillade intercept og hældning en vilkårlig korrelation
- ▶ Default option i RANDOM er TYPE=VC, der blot siger, at der er tale om varianskomponenter med hver sin varians
- ▶ Manglende TYPE=UN kan give konvergensproblemer, samt helt uforståelige resultater

111 / 121



Hældninger, opdelt efter behandling og dropout-status

Slide 75

```
proc sort data=calcium;
  by grp dropout ud girl;
run;

proc reg data=calcium outest=estimat noprint;
  by grp dropout ud girl;
  model bmd=years;
run;

data individuel;
  set estimat;

proc means N mean data=individue;
  class grp dropout;
  var years;
run;
```

112 / 121



Predikterede forløb fra random regression

Slide 77

```
proc mixed data=calcium;
class grp girl;
model bmd=grp years grp*years / ddfm=kr2 vciry s outpm=predicted_mean;
random intercept years /
      type=un subject=girl(grp) g vcorr;
estimate "forskkel efter 2 aar" grp 1 -1 grp*years 2 -2;
run;

proc sgplot data=predicted_mean; where girl in (101,102);
reg Y=Pred X=years / group=grp;
run;
```

113 / 121



Med baseline som kovariat

Slide 82

Variablen baseline er defineret s. 110

Herefter analyseres kun de 4 follow-up visits:

```
proc mixed data=calcium; where visit>1;
class grp girl;
model bmd=baseline grp years grp*years / ddfm=kr2 vciry s;
random intercept years / type=un subject=girl(grp) g;
estimate "forskkel efter 2 aar" grp 1 -1 grp*years 2 -2;
run;
```

114 / 121



Analyse af differenser

Slide 83

```
.....
bmd_afv=bmd-baseline;
....
run;

proc mixed data=calcium; where visit>1;
class grp girl;
model bmd_afv=grp years grp*years / ddfm=kr2 vciry s;
random intercept years / type=un subject=girl(grp) g;
estimate "forskkel efter 2 aar" grp 1 -1 grp*years 2 -2;
run;
```

med output næste side

115 / 121



Analyse af differenser

Slide 83

Output fra modellen s. 115

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
forskkel efter 2 aar	0.01805	0.006213	99.4	2.90	0.0045

116 / 121



Random regression, med ens baseline

Slide 85-86

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

```
proc mixed data=calcium;
class grp girl;
model bmd=years grp*years / ddfm=kr2 vciry s outpm=predicted_mean;
random intercept years /
      type=un subject=girl(grp) g vcorr;
estimate "forskel efter 2 aar" grp*years 2 -2;
run;

proc sgplot data=predicted_mean;
reg Y=Pred X=years / group=grp;
run;
```

117 / 121



Modellen s. 37, baseline justeret

Slide 87-89

```
....
adj_grp=grp;
if visit=1 then adj_grp='P';
run;

proc mixed data=calcium;
class adj_grp visit girl grp;
model bmd=visit adj_grp*visit
      / ddfm=kr2 vciry s outpm=udpm outp=udp;
random intercept / subject=girl(grp);
run;
```

118 / 121



Output fra model s. 89

Slide 87-89

Solution for Fixed Effects							
Effect	adj_grp	visit	Estimate	Standard Error	DF	t Value	Pr > t
Intercept			0.9624	0.006851	150	140.48	<.0001
visit		1	-0.08724	0.003085	390	-28.28	<.0001
visit		2	-0.06748	0.003103	381	-21.75	<.0001
visit		3	-0.04342	0.003117	381	-13.93	<.0001
visit		4	-0.01619	0.003148	381	-5.14	<.0001
visit		5	0	.	.	.	.
adj_grp*visit	C	2	0.007095	0.004178	400	1.70	0.0902
adj_grp*visit	C	3	0.01343	0.004274	399	3.14	0.0018
adj_grp*visit	C	4	0.01286	0.004351	399	2.95	0.0033
adj_grp*visit	C	5	0.01964	0.004399	398	4.47	<.0001
adj_grp*visit	P	1	0	.	.	.	.
adj_grp*visit	P	2	0	.	.	.	.
adj_grp*visit	P	3	0	.	.	.	.
adj_grp*visit	P	4	0	.	.	.	.
adj_grp*visit	P	5	0	.	.	.	.

Estimeret fordel efter 2 år: 0.0196 (0.0044) g pr cm³

119 / 121



SAS, PROC MIXED

Slide 92

- model
beskriver den **systematiske** del af modellen
(middelværdistrukturen)
- random
beskriver de tilfældige effekter
- repeated
beskriver den serielle korrelation over tid
- local
tilføjer en ekstra målefejl

120 / 121



Flere forløb for hvert individ

Slide 93

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

```
proc mixed data = test;  
class grp situation patient tid;  
model y = grp situation konttid  
        grp*situation grp*konttid  
        situation*konttid grp*situation*konttid  
        / ddfm=kr2 vciry s cl;  
random intercept / subject=patient(grp);  
repeated / subject=patient*situation  
        type=sp(exp)(konttid) local;  
run;
```

121 / 121

