Opgavebesvarelse, korrelerede målinger

I 18 familier bestående af far, mor og 3 børn (i veldefinerede aldersintervaller, med `child1` som det ældste barn og `child3` som det yngste) har man registreret antallet af infektioner over et bestemt tidsinterval.

Familierne er inddelt efter, hvor tæt, de bor sammen, angivet ved faktoren *crowding*, som har 3 niveauer, `uncrowded`, `crowded` og `overcrowded`.

Data ses i tabellen nedenfor:

Table 8.13. Numbers of swabs positive for pneumococcus during fixed periods

Crowding category	Family serial number	Family status					
				Child			
		Father	Mother	1	2	3	Total
Overcrowded	1	5	7	6	25	19	62
	2	11	8	11	33	35	98
	3	3	12	19	6	21	61
	4	3	19	12	17	17	68
	5	10	9	15	11	17	62
	6	9	0	6	9	5	29
		41	55	69	101	114	380
Crowded	7	11	7	7	15	13	53
	8	10	5	8	13	17	53
	9	5	4	3	18	10	40
	10	1	9	4	16	8	38
	11	5	5	10	16	20	56
	12	7	3	13	17	18	58
		39	33	45	95	86	298
Uncrowded	13	6	3	5	7	3	24
	14	9	6	6	14	10	45
	15	2	2	6	15	8	33
	16	0	2	10	16	21	49
	17	3	2	0	3	14	22
	18	6	2	4	7	20	39
		26	17	31	62	76	212
Total		106	105	145	258	276	890

og de ligger i det såkaldte *lange format* på hjemmesiden.

Vi ønsker at undersøge, om boligforholdene har betydning for antallet af infektioner, samt om visse familiemedlemmer er mere udsat end andre.

1. *Indlæs data i det midlertidige datasæt* `swabs`*, og overvej strukturen i disse:*

   Det originale datasæt var i det såkaldte brede format:

```
 5  7  6 25 19
11  8 11 33 35
 3 12 19  6 21
 3 19 12 17 17
10  9 15 11 17
 9  0  6  9  5
11  7  7 15 13
10  5  8 13 17
 5  4  3 18 10
 1  9  4 16  8
 5  5 10 16 20
 7  3 13 17 18
 6  3  5  7  3
 9  6  6 14 10
 2  2  6 15  8
 0  2 10 16 21
 3  2  0  3 14
 6  2  4  7 20
```

   således at alle data hørende til en enkelt familie står på samme linie, i rækkefølgen: far, mor og de 3 børn, ordnet med den ældste først. Vi skal lave dette om til langt format, så hver enkelt individ får sin egen linie, men så vi holder styr på, hvilken familie, de hører til. Det foregår således:

```
data swabs;
infile 'swabs_orig.txt';
input f m c1 c2 c3;

family=_n_;
if family<7 then crowding='overcrow';
if family>6 then crowding='crow';
if family>12 then crowding='uncrow';

swabs=f; name='father'; output;
swabs=m; name='mother'; output;
swabs=c1; name='child1'; output;
```

```
swabs=c2; name='child2'; output;
swabs=c3; name='child3'; output;
run;

proc print noobs data=swabs;
  var crowding family name swabs;
run;
```

Herved får vi output svarende til den fil, I har indlæst ved øvelserne, med udseendet

crowding	family	name	swabs
overcrow	1	father	5
overcrow	1	mother	7
overcrow	1	child1	6
overcrow	1	child2	25
overcrow	1	child3	19
overcrow	2	father	11
overcrow	2	mother	8
overcrow	2	child1	11
overcrow	2	child2	33
overcrow	2	child3	35
overcrow	3	father	3
overcrow	3	mother	12
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
uncrow	17	father	3
uncrow	17	mother	2
uncrow	17	child1	0
uncrow	17	child2	3
uncrow	17	child3	14
uncrow	18	father	6
uncrow	18	mother	2
uncrow	18	child1	4
uncrow	18	child2	7
uncrow	18	child3	20

Da I kun er blevet præsenteret for ovenstående datastruktur, foregår jeres indlæsning som:

```
FILENAME navn URL "http://staff.pubhealth.ku.dk/~lts/basal/data/swabs.txt";

data swabs;
infile navn firstobs=2;
input crowding $ family name $ swabs;

/* definitioner for at holde styr på rækkefølgen:  */
if name='father' then namenr=1;
if name='mother' then namenr=2;
if name='child1' then namenr=3;
if name='child2' then namenr=4;
if name='child3' then namenr=5;
run;
```

Vi har totalt set 90 observationer, og 4 variable: `crowding`, `family`, `name` og `swabs`.

(a) *Hvad er outcome?*

Det er variablen `swabs`, som angiver antallet af infektioner over en vis periode. Vi skal analysere denne i normalfordelingsbaserede modeller, selv om dette måske ikke er helt rimeligt. Kommentarer til dette gives til sidst.

(b) *Hvilke kovariater har vi?*

Vi er interesserede i at evaluere effekten af boligforhold (`crowding`) samt status i familien (`name`). Men vi er også nødt til at tage hensyn til, at personerne hænger sammen 5 og 5 i familier.

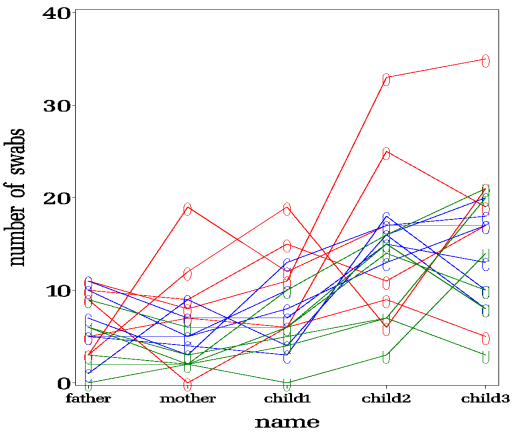
(c) *- og hvordan skal effekten af disse beskrives i modellen?*
dvs. hvilke er systematiske, og hvilke er tilfældige?

Kovariaterne `crowding` og `name` er systematiske effekter, medens `family` er en tilfældig (random) effect, som er nestet i `crowding`. Vi er ikke interesserede i disse secifikke familier, de er bare repræsentanter for familier i al almindelighed.

2. *Lav en (eller flere) arbejdstegning(er), der så vidt muligt indeholder al informationen i data.*

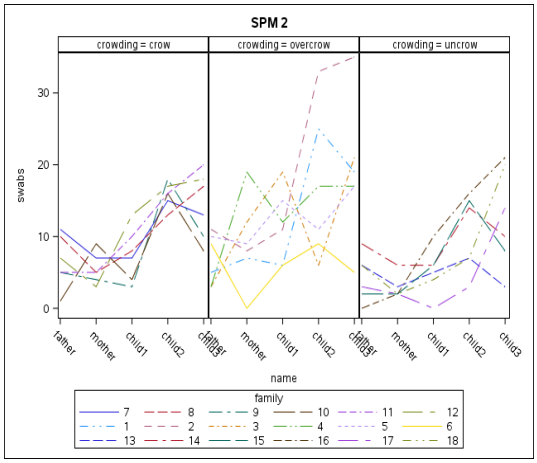
   Her er lavet et spaghetti-plot, hvor man kan se hver af de 18 familier, med farvekode efter, hvilken gruppe, de tilhører:

```
proc gplot gout=plotud data=swabs;
  plot swabs*namenr=family
  / nolegend haxis=axis1 vaxis=axis2 frame;
axis1 order=(1 to 5 by 1) offset=(3,3)
      value=(h=2 'father' 'mother' 'child1' 'child2' 'child3')
      minor=NONE label=(h=3 'name');
axis2 value=(h=3) minor=NONE
      label=(A=90 R=0 h=3 'number of swabs');
symbol1 v='O' i=join c=red l=1 h=3 w=3 r=6;
symbol2 v='C' i=join c=blue l=1 h=3 w=3 r=6;
symbol3 v='U' i=join c=green l=1 h=3 w=3 r=6;
run;
```



   Der er andre (og lettere) muligheder, f.eks. med `sgpanel`:

```
proc sgpanel data=swabs;
panelby crowding / rows=1;
series Y=swabs X=name / group=family;
run;
```



Det er nok mest et spørgsmål om smag og behag....

3. *Fit en passende varianskomponentmodel til disse data, dvs. en model, der specificerer samme korrelation mellem alle familiemedlemmer i samme familie.*

   Her er tale om en mixed model, med de ovenfor specificerede effekter:

```
proc mixed data=swabs;
  class crowding family name;
  model swabs=crowding name / ddfm=satterth s cl outpm=predm;
random intercept / subject=family(crowding) vcorr;
run;
```

   hvorved vi får outputtet (beskåret)

```
The Mixed Procedure

                Model Information

Data Set                     WORK.SWABS
Dependent Variable           swabs
Covariance Structure         Variance Components
Subject Effect               family(crowding)

             Class Level Information

Class       Levels    Values
crowding         3    crow overcrow uncrow
family          18    1 2 3 4 5 6 7 8 9 10 11 12 13
                      14 15 16 17 18
name             5    child1 child2 child3 father
```

mother

Dimensions

Covariance Parameters	2
Subjects	18
Max Obs Per Subject	5

Number of Observations

Number of Observations Used	90

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	family(crowding)	4.3341
Residual		23.3696

Fit Statistics

-2 Res Log Likelihood	527.1
AIC (Smaller is Better)	531.1

Solution for Fixed Effects

Effect	crowding	name	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
Intercept			3.0111	1.5937	15	1.89	0.0783	0.05
crowding	crow		2.8667	1.7328	15	1.65	0.1188	0.05
crowding	overcrow		5.6000	1.7328	15	3.23	0.0056	0.05
crowding	uncrow		0	.	.	.	.	.
name		child1	2.2222	1.6114	68	1.38	0.1724	0.05
name		child2	8.5000	1.6114	68	5.27	<.0001	0.05
name		child3	9.5000	1.6114	68	5.90	<.0001	0.05
name		father	0.05556	1.6114	68	0.03	0.9726	0.05
name		mother	0	.	.	.	.	.

Solution for Fixed Effects

Effect	crowding	name	Lower	Upper
Intercept			-0.3858	6.4081
crowding	crow		-0.8268	6.5601
crowding	overcrow		1.9066	9.2934
crowding	uncrow		.	.
name		child1	-0.9933	5.4377
name		child2	5.2845	11.7155
name		child3	6.2845	12.7155
name		father	-3.1600	3.2711
name		mother	.	.

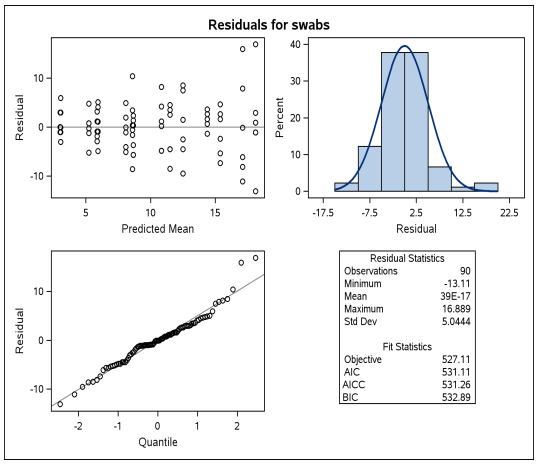
Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
crowding	2	15	5.22	0.0190
name	4	68	16.41	<.0001

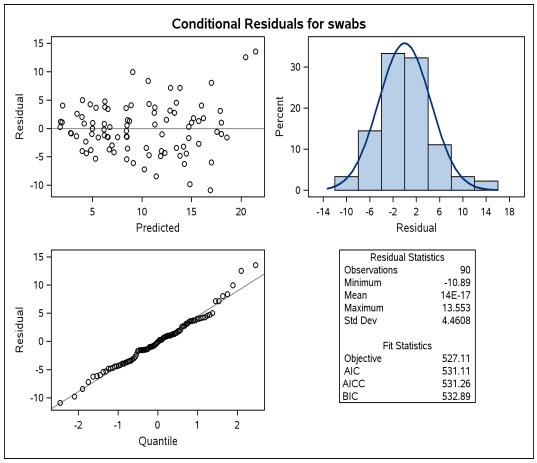
Vi ser, at der er en signifikant effekt af `crowding` (P=0.019), og fra de tilhørende estimater ses at der er flest infektioner, når man bor trangt og færre, når man bor bedre, som forventet.

Der er ligeledes en (stærkt) signifikant forskel på de 5 typer af familiemedlemmer ($P < 0.0001$), idet de mindste børn har flest, og fædrene har færrest.

Modelkontrollen falder i 2 dele, dels de almindelige residualer:



og dels de *betingede*, hvor man kun tager afvigelsen fra personens egen predikterede værdi (som også indeholder det tilfældige personniveau):



På begge residualplots kan man se en tendens til trompetfacon, så modellen er ikke helt god. Mere om det til allersidst i besvarelsen.

(a) *Hvad er estimatet for denne korrelation mellem familiemedlemmer i samme familie?*

Korrelationen estimeres ud fra de to varianskomponenter, til

$$\frac{4.3341}{4.3341 + 23.3696} = 0.1564$$

men kan også får ud på print ved at tilføje option `vcorr` i `random`-sætningen ovenfor:

```
random intercept / subject=family(crowding) vcorr;
```

hvorved vi yderligere får

Estimated V Correlation Matrix for family(crowding) 7 crow

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.1564	0.1564	0.1564	0.1564
2	0.1564	1.0000	0.1564	0.1564	0.1564
3	0.1564	0.1564	1.0000	0.1564	0.1564
4	0.1564	0.1564	0.1564	1.0000	0.1564
5	0.1564	0.1564	0.1564	0.1564	1.0000

(b) *Giv et estimat (med konfidensinterval) for forskellen på* `overcrowded` *og* `uncrowded`*.*

Da `uncrowded` er referenceniveau for `crowding`, kan vi direkte aflæse estimatet for denne sammenligning på linien `overcrow`. Det er 5.60(1.73), dvs. med 95% konfidensinterval på (1.91, 9.29).

(c) *Giv også et estimat (med konfidensinterval) for forskellen på ældste og yngste barn.*

Her er det lidt sværere, da hverken det ældste eller det yngste barn er referenceniveauet. Det letteste er at lave en estimate sætning

```
estimate "old vs. young" name 1 0 -1 0 0 / cl;
```

hvorved vi får den ekstra linie

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
old vs. young	-7.2778	1.6114	68	-4.52	<.0001	0.05

Estimates

Label	Lower	Upper
old vs. young	-10.4933	-4.0623

der siger, at det ældste barn er mindre belastet af infektioner end det yngste, med en forskel estimeret til 7.3 med 95% konfidensinterval på (4.1, 10.5).

4. *Er der evidens for forskellig effekt af trange boligforhold for de enkelte familienmedlemmer?*

Ovenfor fittede vi en model, der havde en effekt af boligforhold, der bar antaget at være den samme for alle familiemedlemmer, idet der var tale om en additiv model, dvs. uden interaktion (vekselvirkning).

Vi ser nu på en model med interaktion, for at se, om denne har signifikant betydning

```
proc mixed data=swabs;
  class crowding family name;
  model swabs=crowding name crowding*name;
random family(crowding);
run;
```

Data Set	WORK.SWABS
Dependent Variable	swabs
Covariance Structure	Variance Components
Subject Effect	family(crowding)

Class Level Information

Class	Levels	Values
crowding	3	crow overcrow uncrow
family	18	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
name	5	child1 child2 child3 father mother

Dimensions

Covariance Parameters	2
Subjects	18
Max Obs Per Subject	5

Number of Observations

Number of Observations Used	90

Estimated V Correlation Matrix for family(crowding) 7 crow

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.1352	0.1352	0.1352	0.1352
2	0.1352	1.0000	0.1352	0.1352	0.1352
3	0.1352	0.1352	1.0000	0.1352	0.1352
4	0.1352	0.1352	0.1352	1.0000	0.1352
5	0.1352	0.1352	0.1352	0.1352	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	family(crowding)	3.9522
Residual		25.2789

Fit Statistics

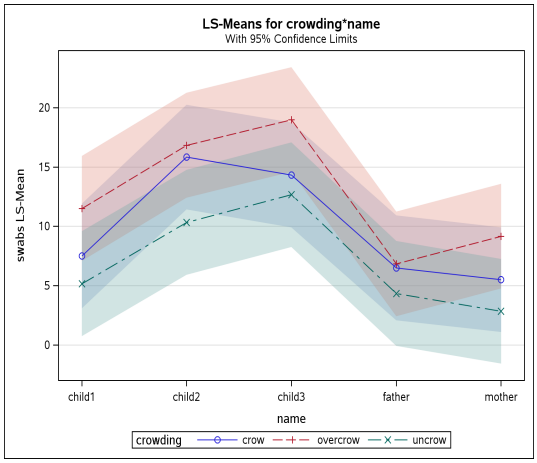
-2 Res Log Likelihood	490.6

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
crowding	2	15	5.22	0.0190
name	4	60	15.17	<.0001
crowding*name	8	60	0.36	0.9384

Der ses tydeligvis ingensomhelst indikation af interaktion, dvs. effekten af boligforhold synes at betyde nogenlunde det samme for alle typer af familiemedlemmer (P=0.94).

Figuren nedenfor illustrerer de predikterede værdier i modellen med interaktion (faktisk blot gennemsnit af 6 familier). Der ses en svag tendens til, at effekten af boligforhold er mindst for fædrene, men dette er som nævnt ovenfor, absolut ikke signifikant.

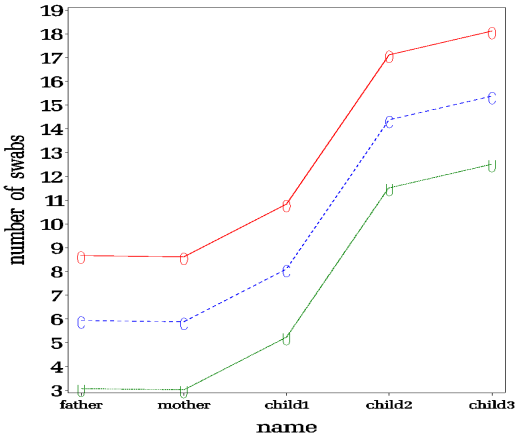


Figuren er dannet med proceduren `glimmix`:

```
ods graphics on;
proc glimmix plots=all data=swabs;
  class crowding family name;
  model swabs=crowding name name*crowding / s cl;
random intercept / subject=family(crowding);
lsmeans name*crowding
        / plots=(meanplot(join clband sliceby=crowding));
run;
ods graphics off;
```

5. *Lav en illustration af de fittede værdier i den model, I ender med.*

   Da der ikke er nogen signifikant interaktion, vil vi udelade interaktionsleddet igen, og ender således med modellen fra spørgsmål 3. De predikterede værdier i denne model ses i figuren nedenfor. Da dette er en additiv model, er der tale om 3 parallelle "kurver".



Denne figur er dannet således:

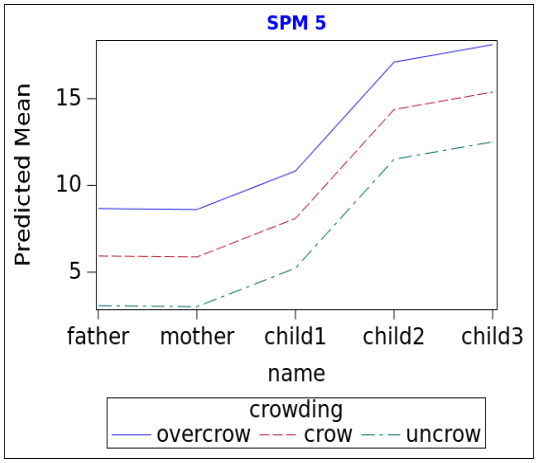
```
proc mixed data=swabs;
  class crowding family name;
  model swabs=crowding name / ddfm=satterth
        residual influence outpm=udpm s cl;
random intercept / subject=family(crowding) vcorr;
run;

proc sort data=udpm; by family name;
run;

proc gplot gout=plotud data=udpm; where family=1 or family=7 or family=13;
  plot Pred*namenr=family
  / nolegend haxis=axis1 vaxis=axis2 frame;
axis1 order=(1 to 5 by 1)
      offset=(3,3)
      value=(h=2 'father' 'mother' 'child1' 'child2' 'child3')
      minor=NONE
      label=(h=3 'name');
axis2 value=(h=3)
      minor=NONE
      label=(A=90 R=0 h=3 'number of swabs');
symbol1 v='O' i=join c=red l=1 h=3 w=3 r=1;
symbol2 v='C' i=join c=blue l=2 h=3 w=3 r=1;
symbol3 v='U' i=join c=green l=33 h=3 w=3 r=1;
run;
```

eller alternativt, med `sgplot` i stedet for `gplot`:

```
proc sgplot data=udpm; where family=1 or family=7 or family=13;
series Y=Pred X=name / group=crowding;
run;
```



6. *Overvej, om det var fornuftigt, at benytte den anvendte kovariansstruktur.*
   Prøv at fitte modellen med en ustruktureret kovarians og se, hvordan resultaterne ændrer sig.

   Man kunne sagtens forestille sig, at der var større variation mellem nogle typer af familiemedlemmer (på tværs af familier) end mellem andre. Det oprindelige spaghettiplot i spørgsmål 2 tyder på en noget større variation mellem de små børn end f.eks. mellem fædre.

   Dette kan man f.eks. også se ved at udregne summary statistics fra modellen i spørgsmål 3:

```
proc means data=udpm;
class name;
var Resid;
run;
```

   som giver

```
The MEANS Procedure

                Analysis Variable : Resid Residual
```

name	N Obs	N	Mean	Std Dev	Minimum	Maximum
child1	18	18	-2.1711E-15	3.8857709	-5.2333333	8.1666667
child2	18	18	-1.77636E-15	6.5101630	-11.1111111	15.8888889
child3	18	18	-1.67767E-15	7.0130038	-13.1111111	16.8888889
father	18	18	-1.08555E-15	3.5759546	-5.6666667	5.9333333
mother	18	18	4.440892E-16	3.7105925	-8.6111111	10.3888889

Desuden kunne man forestille sig, at der var større korrelation mellem børnene indbyrdes, samt mellem mødre og børn, simpelthen fordi det muligvis kunne være begrænset, hvor meget faderen opholdt sig i familien.

Vi kan undersøge disse ting ved at benytte en helt generel kovarinsstruktur i stedet for den "*compound symmetry*", som benyttes, når man ser det som en varianskomponentmodel.

Vi udelader interaktionen og opskriver den additive model

```
proc mixed plots=all data=swabs;
  class crowding family name;
  model swabs=crowding name / s cl;
repeated name / type=un subject=family(crowding) r rcorr;
estimate "old vs. young" name 1 0 -1 0 0 / cl;
run;
```

og finder

Estimated R Matrix for family(crowding) 7 crow

Row	Col1	Col2	Col3	Col4	Col5
1	12.3205	-3.9361	-1.1168	1.2205	1.0320
2	-3.9361	14.6862	5.2834	1.3431	3.9194
3	-1.1168	5.2834	16.0635	-1.8834	11.6929
4	1.2205	1.3431	-1.8834	42.9697	21.3105
5	1.0320	3.9194	11.6929	21.3105	50.2394

Estimated R Correlation Matrix for family(crowding) 7 crow

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	-0.2926	-0.07938	0.05304	0.04148
2	-0.2926	1.0000	0.3440	0.05347	0.1443
3	-0.07938	0.3440	1.0000	-0.07169	0.4116
4	0.05304	0.05347	-0.07169	1.0000	0.4587
5	0.04148	0.1443	0.4116	0.4587	1.0000

Fit Statistics

-2 Res Log Likelihood	502.9
AIC (Smaller is Better)	532.9

Solution for Fixed Effects

Effect	crowding	name	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			3.3517	1.1163	15	3.00	0.0089
crowding	crow		2.7232	1.1362	15	2.40	0.0300
crowding	overcrow		4.7219	1.1362	15	4.16	0.0008
crowding	uncrow		0	.	.	.	.
name		child1	2.2222	1.0589	15	2.10	0.0532
name		child2	8.5000	1.7475	15	4.86	0.0002
name		child3	9.5000	1.7809	15	5.33	<.0001
name		father	0.05556	1.3920	15	0.04	0.9687
name		mother	0	.	.	.	.

Effect	crowding	name	Alpha	Lower	Upper
Intercept			0.05	0.9722	5.7311
crowding	crow		0.05	0.3015	5.1449
crowding	overcrow		0.05	2.3002	7.1436
crowding	uncrow		.	.	.
name		child1	0.05	-0.03478	4.4792
name		child2	0.05	4.7752	12.2248
name		child3	0.05	5.7042	13.2958
name		father	0.05	-2.9115	3.0226
name		mother	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
crowding	2	15	8.70	0.0031
name	4	15	10.53	0.0003

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha
old vs. young	-7.2778	1.5441	15	-4.71	0.0003	0.05

Label	Lower	Upper
old vs. young	-10.5690	-3.9866

I lighed med resultaterne fra spørgsmål 3, finder vi her, at der er en signifikant effekt af `crowding` (P=0.0031 mod før 0.019), og fra de

tilhørende estimater ses at der er flest infektioner, når man bor trangt og færre, når man bor bedre, som forventet.

Der er ligeledes en (stærkt) signifikant forskel på de 5 typer af familiemedlemmer ($P = 0.0003$ mod før < 0.0001), idet de mindste børn har flest, og fædrene har færrest.

Fra diagonalen i variansmatricen (`R Matrix`) ser vi, at variationen stiger, men for at forstå, hvad det betyder, må vi vide, i hvilken rækkefølge, den opfatter de 5 familiemedlemmer. Det gøres ud fra rækkefølgen i datasættet, og det betyder, at de står som far-mor-ældste-midterste-yngste barn.

Derfor genfinder vi mønstret (naturligvis) fra residualerne, altså at der er mindst variation mellem fædre og størst variation mellem de yngste børn. Da det samme er tilfældet for selve niveauet af målingerne, kunne man fristes til at analysere logaritmer, men da der er ægte nuller i materialet, kan dette ikke lade sig gøre.

Fra korrelationsmatricen (`R Correlation Matrix`) ser vi, at de største korrelationer forekommer på plads (2,3), (3,5) og (4,5), altså mellem mor og ældste barn, samt mellem yngste barn og dets søskende. Korrelationen mellem far og mor estimeres til at være negativ, men nu skal man jo ikke overfortolke....

Vi kan sammenligne de to kovariansstrukturer ved at se på forskellen i $-2\log L$. I modellen her med `TYPE=UN` får vi værdien 502.9, og vi har brugt 15 paramtere til beskrivelse af kovariansen. I den simple variankomponentmodel brugte vi kun 2 paramtre, og fik $-2\log L = 527.1$. Differensen er altså $527.1 - 502.9 = 24.2 \sim \chi^2(13) \Rightarrow P = 0.029$

Modellen med den mere generelle kovariansstruktur fitter altså bedre end den simple.

Estimatet for effekten af `overcrow` vs. `uncrow` ses at være 4.72 (1.14), hvor vi i spørgsmål 3 fik 5.60 (1.73).

Tilsvarende er estimatet for forskellen på yngste og ældste barn (se

`estimate`-sætningen): 7.28 (1.54) mod før 7.28 (1.61).

7. *Hvilke betragtninger tror I, der ligger til grund for valget af design i denne undersøgelse? Et alternativ kunne jo være blot at undersøge et tilfældigt udpluk af personer.*

   Dette design giver bedre mulighed for at sammenligne familiemedlemmer, fordi der er tale om parrede sammenligninger. Herved slipper vi ved denne sammenligning for den store variation, man må forvente at finde mellem familier, pga. genetik, madvaner mv.

   Til gengæld får man en mere upræcis vurdering af effekten af boligforhold, fordi korrelationen mellem familiemedlemmer af samme familie nedsætter mængden af uafhængig information.

Hvis I herefter har tid, skal I prøve at se, om I kan opnå nogle af ovenstående resultater med traditionelle analysemetoder:

8. *Udregn gennemsnitligt antal infektioner for hver familie, f.eks. ved at skrive nedenstående kode (overvej ideen i denne, f.eks. ved at se på det resulterende datasæt* `averages`*)*

```
proc sort data=swabs; by family;
run;
proc means noprint data=swabs; by family;
var swabs;
output out=averages mean=mswabs;
id crowding;
run;
```

   Datasættet `averages` består af 18 linier, en for hver familie, med 3 forståelige variable, og 2 ekstra, som SAS danner med proceduren `means`:

Obs	family	crowding	_TYPE_	_FREQ_	mswabs
1	1	overcrow	0	5	12.4
2	2	overcrow	0	5	19.6
3	3	overcrow	0	5	12.2
4	4	overcrow	0	5	13.6
5	5	overcrow	0	5	12.4

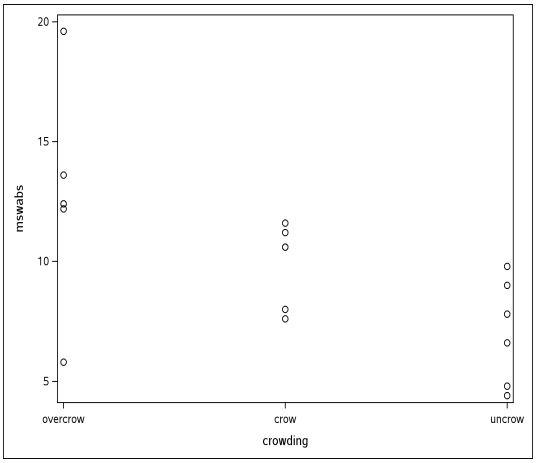
6	6	overcrow	0	5	5.8
7	7	crow	0	5	10.6
8	8	crow	0	5	10.6
9	9	crow	0	5	8.0
10	10	crow	0	5	7.6
11	11	crow	0	5	11.2
12	12	crow	0	5	11.6
13	13	uncrow	0	5	4.8
14	14	uncrow	0	5	9.0
15	15	uncrow	0	5	6.6
16	16	uncrow	0	5	9.8
17	17	uncrow	0	5	4.4
18	18	uncrow	0	5	7.8

9. *Brug det nydannede datasæt til at vurdere effekten af boligforhold*

   (a) *Hvilken type analyse er der tale om her?*
   Husk en figur til illustration.

   Vi skal sammenligne det gennemsnitlige antal infektioner i de 18 familier (`mswabs`), og familierne er inddelt i 3 grupper, angivet ved boligforholdene, `crowding`. Der er således tale om en ensidet variansanalyse.

   En passende figur kunne være et scatterplot, da vi har så få observationer (hvis vi havde haft flere, ville et Box Plot være mere rimeligt):



   Hvis man vil have en anden rækkefølge af familiemedlemmerne på X-aksen, kan man omnummerere, f.eks. sådan her:

```
if crowding='overcrow' then crowdnr=1;
if crowding='crow' then crowdnr=2;
if crowding='uncrow' then crowdnr=3;
```

Nu laver vi den ensidede variansanalyse

```
proc glm data=averages;
  class crowding;
  model mswabs=crowding / solution clparm;
run;
```

som giver os outputtet

```
The GLM Procedure

          Class Level Information

Class         Levels    Values
crowding           3    crow overcrow uncrow

Number of observations    18

Dependent Variable: mswabs
                                    Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                       2     94.0977778     47.0488889      5.22   0.0190
Error                      15    135.1200000      9.0080000
Corrected Total            17    229.2177778


R-Square     Coeff Var      Root MSE    mswabs Mean
0.410517      30.35056      3.001333       9.888889

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
crowding                    2    94.09777778    47.04888889      5.22   0.0190
                                                      Standard
Parameter                  Estimate                      Error    t Value    Pr > |t|

Intercept              7.066666667 B      1.22528908       5.77      <.0001
crowding  crow         2.866666667 B      1.73282044       1.65      0.1188
crowding  overcrow     5.600000000 B      1.73282044       3.23      0.0056
crowding  uncrow       0.000000000 B       .                .         .

Parameter               95% Confidence Limits

Intercept              4.455024811  9.678308523
crowding  crow        -0.826752666  6.560086000
crowding  overcrow     1.906580667  9.293419333
crowding  uncrow        .            .
```

Vi finder, ligesom i spørgsmål 3, at der er en signifikant effekt af boligforhold på antallet af infektioner, P=0.019.

(b) *Giv igen et estimat (med konfidensinterval) for forskellen på* `overcrowded` *og* `uncrowded`*, og sammenlign med det ovenfor fundne (spm. 3b).*

Da `uncrowded` er referenceniveau for `crowding`, kan vi direkte aflæse estimatet for denne sammenligning på linien `overcrow`. Det er 5.60(1.73), dvs. med 95% konfidensinterval på (1.91, 9.29), altså ganske som vi fandt i "mixed effects"-modellen fra spørgsmål 3.

(c) *Hvorfor kan vi ikke udfra denne analyse sige noget om forskelle på familiemedlemmer?*

Vi regner her på gennemsnit over 5 familiemedlemmer, så vi kan derfor ikke sige noget om de indbyrdes forskelle på disse ud fra denne analyse.

10. *Nu ser vi et øjeblik væk fra faktoren* crowding *og betragter bare de 18 familier, som om de var tilfældige stikprøver fra samme population. Vi går altså tilbage til det oprindelige datasæt,* `swabs`*:*

(a) *Lav en tosidet variansanalyse med faktorerne* `family` *og* `name`*.*

Her behøver vi blot at bruge GLM:

```
proc glm data=swabs;
  class family name;
  model swabs=family name / solution clparm;
run;
```

og finder outputtet:

```
The GLM Procedure

                    Class Level Information

Class         Levels    Values
family            18    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
name               5    child1 child2 child3 father mother

Number of observations    90

 Dependent Variable: swabs
```

```
                                  Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                      21    2679.755556     127.607407      5.46   <.0001
Error                      68    1589.133333      23.369608
Corrected Total            89    4268.888889

R-Square     Coeff Var      Root MSE    swabs Mean
0.627741      48.88529      4.834212      9.888889

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
family                     17    1146.088889      67.416993      2.88   0.0010
name                        4    1533.666667     383.416667     16.41   <.0001

                                                 Standard
Parameter                 Estimate                    Error    t Value    Pr > |t|
Intercept              3.74444444 B      2.39009849       1.57      0.1218
family    1            4.60000000 B      3.05742427       1.50      0.1371
family    2           11.80000000 B      3.05742427       3.86      0.0003

family    15          -1.20000000 B      3.05742427      -0.39      0.6959
family    16           2.00000000 B      3.05742427       0.65      0.5152
family    17          -3.40000000 B      3.05742427      -1.11      0.2700
family    18           0.00000000 B       .                .         .
name      child1       2.22222222 B      1.61140408       1.38      0.1724
name      child2       8.50000000 B      1.61140408       5.27      <.0001
name      child3       9.50000000 B      1.61140408       5.90      <.0001
name      father       0.05555556 B      1.61140408       0.03      0.9726
name      mother       0.00000000 B       .                .         .

Parameter              95% Confidence Limits
Intercept             -1.02492284   8.51381173
family    1           -1.50099513  10.70099513
family    2            5.69900487  17.90099513

family    15          -7.30099513   4.90099513
family    16          -4.10099513   8.10099513
family    17          -9.50099513   2.70099513
family    18            .            .
name      child1      -0.99328455   5.43772899
name      child2       5.28449323  11.71550677
name      child3       6.28449323  12.71550677
name      father      -3.15995121   3.27106232
name      mother        .            .
```

Her ser vi en tydelig signifikant forskel på de 18 familier (P=0.001), men en endnu tydeligere forskel på familiemedlemmerne ($P < 0.0001$).

(b) *Er der evidens for forskelle på de 5 familiemedlemmer?*

Ja, som nævnt ovenfor er dette meget tydeligt ($P < 0.0001$). De yngste børn ser ud til at være de mest sensitive.

(c) *Giv igen et estimat (med konfidensinterval) for forskellen på æld-*

ste og yngste barn, og sammenlign med det tidligere fundne (spm. 3c).

Her har vi igen brug for estimate-sætningen, ganske som i spørgsmål 3c:

```
estimate "old vs. young" name 1 0 -1 0 0 / cl;
```

og vi finder derved den ekstra output-linie:

```
                                        Standard
Parameter                 Estimate           Error    t Value    Pr > |t|
old vs. young          -7.27777778      1.61140408      -4.52      <.0001

Parameter               95% Confidence Limits
old vs. young        -10.49328455  -4.06227101
```

der igen siger, at det ældste barn er mindre belastet af infektioner end det yngste, med en forskel estimeret til 7.3 med 95% konfidensinterval på (4.1, 10.5), ganske som i spørgsmål 3c.

11. *Kan vi vurdere interaktionen mellem* `crowding` *og* `name` *uden at benytte varianskomponentmodellen?*

    Ja, det kan man faktisk godt, men det kræver, at man sætter familien ind som *fixed effect* i stedet for, altså benytter koden

```
proc glm data=swabs;
  class family crowding name;
  model swabs = name crowding family name*crowding;
run;
```

der vil give outputtet

```
The GLM Procedure

                    Class Level Information

Class         Levels    Values
```

```
family            18    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
crowding           3    crow overcrow uncrow
name               5    child1 child2 child3 father mother

Dependent Variable: swabs

                                  Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F
Model                      29   2752.155556     94.901916     3.75  <.0001
Error                      60   1516.733333     25.278889
Corrected Total            89   4268.888889

R-Square     Coeff Var      Root MSE    swabs Mean
0.644701      50.84304      5.027812      9.888889

Source                     DF     Type I SS   Mean Square  F Value  Pr > F
name                        4   1533.666667    383.416667    15.17  <.0001
crowding                    2    470.488889    235.244444     9.31  0.0003
family                     15    675.600000     45.040000     1.78  0.0590
crowding*name               8     72.400000      9.050000     0.36  0.9384

Source                     DF    Type III SS   Mean Square  F Value  Pr > F
name                        4   1533.666667    383.416667    15.17  <.0001
crowding                    0      0.000000             .        .       .
family                     15    675.600000     45.040000     1.78  0.0590
crowding*name               8     72.400000      9.050000     0.36  0.9384
```

12. *Diskuter hvad der sker med de forskellige analysetyper, hvis der mangler nogle observationer.*

    Det afhænger meget af, *hvorfor* sådanne observationer mangler.

    Hvis vi bare er kommet til at smide informationen væk, vil "mixed model" stadig give fornuftige svar og udnytte alle forhåndenværende observationer. Men den ensidede variansanalyse på gennemsnittene for hver familie vil ikke længere give det samme resultat, den vil faktisk være *forkert*. Hvor meget forkert, den er, afhænger af, hvilken observation, der mangler. Hvis f.eks. det yngste barn fra en overcrowded familie mangler, vil denne familie se ud til at have et ret lavt antal gennemsnitlige infektioner, og derved vil effekten af boligforhold blive undervurderet.

    Hvis en observation mangler *fordi* den f.eks. er meget høj, så er der ikke nogen måde at redde det på! Det kaldes *informative missing*, og det er anledning til mange fejl og mistolkninger.

Konklusioner:

- Der er en signifikant forskel på familier (P=0.01), men det interessrerer os sådan set ikke. Denne faktor optræder i modellen som en tilfældig effekt, som gør det muligt at vurdere effekten af boligforhold.
- Der er en effekt af boligforhold (`crowding`) på antallet af infektioner (P=0.019), og effekten af denne kan kvantificeres med `proc mixed`, eller ved hjælp af ensidet variansanalyse på simple familie-gennemsnit, det sidste dog **kun** i tilfælde af fuldstændigt datasæt, dvs. **ingen manglende værdier**.
- Der er signifikant forskel på antallet af infektioner hos de forskellige typer af familiemedlemmer ($P < 0.0001$). Der er flest infektioner blandt de yngste børn, og færrest hos fædrene.

På de sidste sider findes et eksempel på et program, der udfører alle analyserne, der er beskrevet ovenfor.

Men:
Vi har hele vejen antaget, at det er rimeligt at lave normalfordelingsbasered analyser, og dette ser ikke rigtigt fornuftigt ud, f.eks. fordi et antal altid er ikke-negativt, og her har vi endda en del 0'er.

Man kunne forsøge sig med en kvadratrodstransformation, men det giver vanskeligheder ved fortolkningen.

Hvis man ønsker at udregne prediktionsområder for antallet af infektioner, skal man i hvert fald gøre et eller andet anderledes end ovenfor, og dette kunne være at modellere antallene som Poissonfordelte. Dette er dog ikke en del af dette kursus.

Appendix: Et muligt totalt program kunne se således ud:

```
/* Spørgsmål 1  */

FILENAME navn URL "http://staff.pubhealth.ku.dk/~lts/basal/data/swabs.txt";

data swabs;
infile navn firstobs=2;
input crowding $ family name $ swabs;

if name='father' then namenr=1;
if name='mother' then namenr=2;
if name='child1' then namenr=3;
if name='child2' then namenr=4;
if name='child3' then namenr=5;
run;

/* Spørgsmål 2  */

proc sort data=swabs; by family namenr;
run;

proc gplot data=swabs;
  plot swabs*namenr=family
  / nolegend haxis=axis1 vaxis=axis2 frame;
axis1 order=(1 to 5 by 1)
      offset=(3,3)
      value=(h=2 'father' 'mother' 'child1' 'child2' 'child3')
      minor=NONE
      label=(h=3 'name');
axis2 value=(h=3)
      minor=NONE
      label=(A=90 R=0 h=3 'number of swabs');
symbol1 v='O' i=join c=red l=1 h=3 w=3 r=6;
symbol2 v='C' i=join c=blue l=1 h=3 w=3 r=6;
symbol3 v='U' i=join c=green l=1 h=3 w=3 r=6;
run;

/* Spørgsmål 3  */

ods graphics on;
proc mixed plots=all data=swabs;
  class crowding family name;
  model swabs=crowding name / ddfm=satterth s cl;
```

```
random intercept / subject=family(crowding) vcorr;
estimate "old vs. young" name 1 0 -1 0 0 / cl;
run;
ods graphics off;

proc mixed data=swabs;
  class crowding family name;
  model swabs=crowding name / ddfm=satterth s cl;
repeated name / type=cs subject=family(crowding) rcorr;
run;

proc mixed data=swabs;
  class crowding family name;
  model swabs=crowding name / ddfm=satterth s cl;
repeated name / type=un subject=family(crowding) rcorr;
run;

/* Spørgsmål 4  */

ods graphics on;
proc mixed plots=all data=swabs;
  class crowding family name;
  model swabs=crowding name name*crowding / ddfm=satterth s cl;
random intercept / subject=family(crowding) vcorr;
run;
ods graphics off;

ods graphics on;
proc glimmix plots=all data=swabs;
  class crowding family name;
  model swabs=crowding name name*crowding / s cl;
random family(crowding);
lsmeans name*crowding
        / plots=(meanplot(join clband sliceby=crowding));
run;
ods graphics off;

/* Spørgsmål 5  */

proc mixed data=swabs;
  class crowding family name;
  model swabs=crowding name / ddfm=satterth
        residual influence outpm=udpm s cl;
```

```
random family(crowding) / vcorr;
run;

proc sort data=udpm; by family namenr;
run;

proc gplot data=udpm; where family=1 or family=7 or family=13;
  plot Pred*name=family
  / nolegend haxis=axis1 vaxis=axis2 frame;
axis1 order=(1 to 5 by 1)
      offset=(3,3)
      value=(h=2 'father' 'mother' 'child1' 'child2' 'child3')
      minor=NONE label=(h=3 'name');
axis2 value=(h=3) minor=NONE
      label=(A=90 R=0 h=3 'number of swabs');
symbol1 v='O' i=join c=red l=1 h=3 w=3 r=6;
symbol2 v='C' i=join c=blue l=1 h=3 w=3 r=6;
symbol3 v='U' i=join c=green l=33 h=3 w=3 r=1;
run;

/* Spørgsmål 7  */

proc sort data=swabs; by family;
run;
proc means noprint data=swabs; by family;
var swabs;
output out=averages mean=mswabs;
id crowding;
run;

proc print data=averages;
run;

/* Spørgsmål 8  */

proc sgplot data=averages;
scatter X=crowding Y=mswabs;
run;

proc glm data=averages;
  class crowding;
  model mswabs=crowding / solution clparm;
run;
```

```
/* Spørgsmål 9  */

proc glm data=swabs;
  class family name;
  model swabs=family name / solution clparm;
estimate "old vs. young" name 1 0 -1 0 0;
run;

/* Spørgsmål 10  */

proc glm data=swabs;
  class family crowding name;
  model swabs = name crowding family name*crowding;
run;
```