Eksamen i Statistisk Dataanalyse 2, 11. april 2013

Vejledende besvarelse

Opgave 1

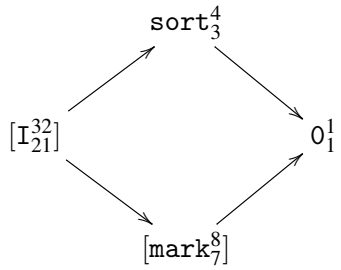
1. Forsøget bør udføres som et fuldstændigt randomiseret blokforsøg, hvor hver af de 4 behandlinger afprøves netop een gang inden for hver blok. For hver blok bestemmes ved lodtrækning, hvordan de 4 behandlinger skal fordeles på de 4 forsøgsenheder inden for blokken.

   Den statistiske analyse af forsøget tager udgangspunkt i modellen

   $$Y_i = \alpha(\mathtt{sort}_i) + A(\mathtt{mark}_i) + e_i,$$

   hvor $A(1),\ldots,A(8)$ er uafhængige $\sim N(0,\sigma^2_{\mathtt{mark}})$ og $e_1,\ldots,e_{32}$ er uafhængige $\sim N(0,\sigma^2)$.

   Et faktordiagram for forsøget ser ud som følger



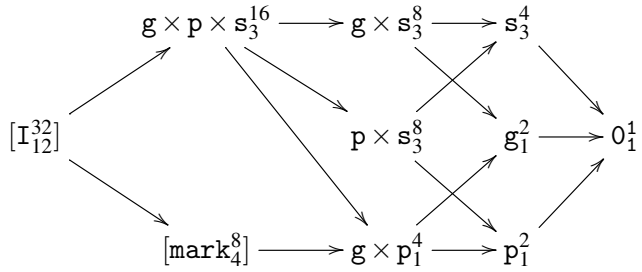
2. Forsøget er et 2^n-forsøg med 8 sorter givet som kombinationer af 3 faktorer på hver 2 niveauer. De 8 sorter skal uddeles på marker af størrelse 4. For hvert par af marker kan man afprøve de 8 sorter, hvor man konfunderer en hoved- eller vekselvirkning med mark. Da vi har 8 marker til rådighed i forsøget, kan man med fordel benytte sig af partiel konfundering, hvor man på de 4 par af narker konfunderer henholdsvis trefaktorvekselvirkningen (på m1+m2) og hver af de 3 parvise vekselvirkninger (på m3+m4,m5+m6,m7+m8). En konkret forsøgsplan kunne se ud som vist i skemaet nedenfor.

A	B	C	m1	m2	m3	m4	m5	m6	m7	m8
1	1	1	x			x		x		x
1	1	2		x		x	x		x	
1	2	1		x	x			x	x	
1	2	2	x		x		x			x
2	1	1		x	x	x				x
2	1	2	x		x		x		x	
2	2	1	x			x x			x	
2	2	2		x		x	x			x

En alternativ forsøgsplan kunne være at gentage et forsøgsdesign 4 gange, hvor man hver gang konfunderer trefaktorvekselvirkningen med mark (som gjort på m1+m2).

3. Da faktorerne `gødning` og `plastic` ikke kan varieres inden for mark, bør man ved lodtrækning allokere 2 marker til hver af de 4 kombinationer af `gødning` × `plastic`. Hver af de 4 sorter bør afprøves netop een gang inden for hver mark, og fordelingen inden for mark foretages ved lodtrækning. Der er således to trin i randomiseringen.

   Forsøget er et split-plot forsøg, hvor `mark` er helplot, `gødning` × `plastic` er helplot-faktoren og `sort` er delplot-faktoren. Et faktordiagram for forsøget ser ud som følger



Opgave 2

1. Variablene `alder` og smerter før operationen `s0` bør indgå i modellen som numeriske variable (kovariater), mens køn bør indgå som en faktor med 2 niveauer. Det vil være naturligt at tage udganspunkt i modellen

$$\texttt{smerte}_i = \alpha(\texttt{koen}_i) + \beta(\texttt{koen}_i) \cdot \texttt{alder}_i + \gamma \cdot \texttt{s0}_i + e_i,$$

   hvor $e_1, \ldots, e_{46}$ er uafhængige $\sim N(0, \sigma^2)$. Modellen udtrykker, at der er en lineær sammenhæng mellem `smerte` og `alder`, hvor både skæring og hældning kan afhænge af køn (`koen`). Desuden indgår `s0` som en slags baselinemåling i modellen. Modellen svarer til `m1` i `R`-udskriften.

2. Som altid er der lidt valgfrihed mht. den rækkefølge, hvori der foretages modelreduktion. Man kan f.eks. starte med at fjerne effekten af `s0` ($F = 3.373, p = 0.0738$) svarende til at modellen reduceres til

$$\texttt{smerte}_i = \alpha(\texttt{koen}_i) + \beta(\texttt{koen}_i) \cdot \texttt{alder}_i + e_i,$$

   hvor $e_1, \ldots, e_{46}$ er uafhængige $\sim N(0, \sigma^2)$.

   Dernæst kan man teste hypotesen

$$H_0 : \beta(\texttt{m}) = \beta(\texttt{f}) = \beta$$

   om, at hældningen ikke afhænger af køn svarende til modellen

$$\texttt{smerte}_i = \alpha(\texttt{koen}_i) + \beta \cdot \texttt{alder}_i + e_i,$$

hvor $e_1, \ldots, e_{46}$ er uafhængige $\sim N(0, \sigma^2)$. Hypotesen godkendes ($F = 0.1718, p = 0.6805$).

Der viser sig heller ikke at være en signifikant sammenhæng mellem `alder` og `smerte` ($F = 0.8435, p = 0.3633$). Dette svarer til hypotesen $H_0 : \beta = 0$ eller modellen

$$\texttt{smerte}_i = \alpha(\texttt{koen}_i) + e_i,$$

hvor $e_1, \ldots, e_{46}$ er uafhængige $\sim N(0, \sigma^2)$.

Endelig er der en signifikant effekt af køn ($F = 12.47, p = 0.0010$) således at vores slutmodel bliver

$$\texttt{smerte}_i = \alpha(\texttt{koen}_i) + e_i,$$

hvor $e_1, \ldots, e_{46}$ er uafhængige $\sim N(0, \sigma^2)$.

Parameterestimaterne for slutmodellen er

$$\alpha(\texttt{f}) = 50.35, \quad \alpha(\texttt{m}) - \alpha(\texttt{f}) = -20.15, \quad \hat{\sigma} = 19.77.$$

Slutmodellen er en ensidet variansanalysemodel, hvor smerten efter operationen alene afhænger af køn. Mænd har signifikant mindre ondt end kvinder og den kvantitative forskel i smertepåvirkningen estimeres til 20.15. Et 95 %- konfidensinterval for forskellen kan beregnes i `R` og bliver

$$20.15 \pm t_{0.975,46} \cdot 5.706 = [8.66, 31.64],$$

hvor $t_{0.975,46} = 2.013$ er 97.5 %-fraktilen i en t-fordeling med 46 frihedsgrader og 5.706 er standard error på estimatet for forskellen som det fremgår af `summary(m9)`.

3. I modellen `m3` indgår `koen` som en faktor mens både `alder` og smerte før operation (`s0`) indgår som kovariater. Estimat samt 95 %- konfidensinterval kan bestemmes vha. `estimable` ud fra listen koefficienter som svarer til `est4` og resultatet bliver $33.1[24.7, 41.6]$.

Opgave 3

1. På baggrund af de foreslåede modeller `modelA-modelH` bør man ved besvarelsen af dette delspørgsmål i hvert tilfælde overveje: 1) om responsvariablen `vo2max` bør transformeres med logaritmen inden analysen, 2) hvordan variablen `tid` bør indgå i (den systematiske del af), 3) om `patient` bør indgå som systematisk eller tilfældig effekt og 4) om det er nødvendigt at modellere en seriel korrelation i modellen for at tage højde for, at vi har gentagne målinger (3 styk) over tid for hver `patient`.

   For at opnå fuldt point for dette delspørgsmål skal man i sin besvarelse belyse alle aspekter nævnt ovenfor. I det følgende gives et bud på, hvad man konkret kunne anføre i sin besvarelse, men det er naturligvis ikke påkrævet, at man formulerer sig præcis som nedenfor.

   Det er forholdsvis indiskutabelt, at `patient` bør indgå i modellen med tilfældig effekt. Variablen `tid` kan enten opfattes som en faktor (med 3 niveauer) eller som

en numerisk variabel. Hvis man inddrager `tid` som en numerisk variabel antager man som udgangpunkt, at der er en lineær sammenhæng mellem `vo2max` og `tid`, hvilket ikke er rimeligt i en udgangsmodel. Variablen `tid` bør derfor indgå som en faktor i udgangsmodellen.

Hvorvidt responsvariablen `vo2max` skal log-transformeres bør afgøres ved at se på relevante residual og QQ-plot. På baggrund af nedenstående figurer er det min opfattelse, at man frit kan vælge om udgangsmodellen tager udgangspunkt i `vo2max` eller `log(vo2max)`. Der er en rimelig grad af varianshomogenitet på de to residualplots, mens der måske på baggrund af QQ-plottet vil være en svag tendens til at foretrække en model baseret på `vo2max`. Figurerne kan naturligvis ikke indgå i den skriftlige besvarelse, men det bør fremgå af besvarelsen, hvilke figurer I har støttet jer op af.

Vi mangler blot at diskutere om man bør bygge en seriel korrelationsstruktur ind i modellen. Dette kan f.eks. belyses ved at sammenligne AIC for *Diggle modellen* og den tilsvarende *Random intercept model*. Idet man bør vælge modellen med lavest AIC bliver konklusionen, at man bør vælge `modelF` som udgangspunkt for den statistiske analyse. Alternativt kan man vælge `modelB`, hvis man beslutter sig for at log-transformere `vo2max` inden analysen.

2. Med udgangspunkt i `modelF` konstaterer man først, at det er muligt at fjerne trefaktorvekselvirkningen $\texttt{gruppe} \times \texttt{sex} \times \texttt{factor(tid)}$ ($L.Ratio = 0.0981, p = 0.9521$).

   Dernæst viser det sig, at man kan fjerne en af de parvise vekselvirkninger, $\texttt{gruppe} \times \texttt{sex}$, ($L.Ratio = 0.0408, p = 0.8398$). Det øvrige parvise vekselvirkninger har stærkt signifikante effekter med p-værdier < 0.0001. For en god ordens skyld bør man også argumentere for, at man ikke kan fjerne effekten af `age`.

   Slutmodellen kan skrives som

   $$\texttt{vo2max}_i = \alpha(\texttt{gruppe} \times \texttt{tid}_i) + \beta(\texttt{sex} \times \texttt{tid}_i) + \gamma \cdot \texttt{age}_i + A(\texttt{patient}_i) + D_i + e_i,$$

   hvor $e_1, \ldots, e_{405}$ er uafhængige $\sim N(0, \sigma^2)$ og $A(1), \ldots, A(135)$ er uafhængige $\sim N(0, \sigma^2_{\texttt{patient}})$. D_i'erne beskriver korrelationsstrukturen fra Diggle modellen i kompendiets kapitel 10.3.

   Den systematiske del af slutmodellen er en additiv model med parvise vekselvirkninger af $\texttt{gruppe} \times \texttt{tid}$ og $\texttt{sex} \times \texttt{tid}$. Desuden indgår variablen `age` som en kovariat i slutmodellen. Det er svært at give en præcis fortolkning af modellen og parameterestimaterne, så det kræves ikke, at man oplister alle parameterestimaterne og forklarer præcis, hvad de udtrykker.

3. Man bør i første omgang genfitte slutmodellen fra delspørgsmål 2. med *REML*-estimation. Dernæst kan man bruge `estimable`-funktionen i `R` til at udtrække estimat samt konfidensinterval for den ønskede kombination af variablene i modellen. Den konkrete fremgangsmåde afhænger lidt af, hvordan man får parametriseret slutmodellen i `R`, og et eksempel kan ses i `R`-udskriften nedenfor. Estimat samt konfidensinterval bliver

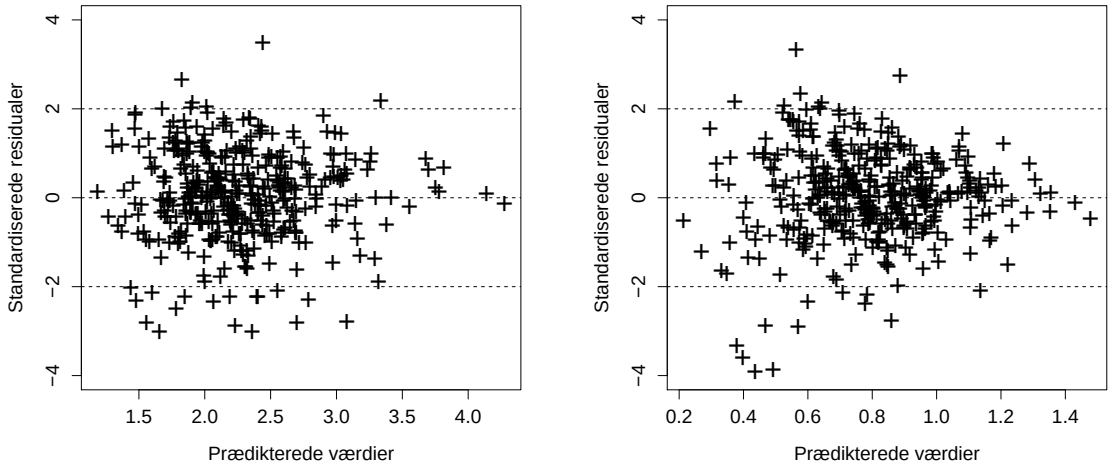
   $$3.01[2.82, 3.19].$$

4. I R-udskriften nedenfor er vist, hvordan man for de 4 kombinationer af `sex` og `gruppe` ved brug af `estimable`-funktionen kan estimere tilvæksten fra 7 til 13 måneder. Det ses at for kontrolgruppen (`gruppe=K`) er der både for mænd og kvinder en signifikant tilvækst i `vo2max` fra 7 til 13 måneder. Derimod ses ingen signifikant ændring i interventionsgruppen (`gruppe=I`) fra 7 til 13 måneder.

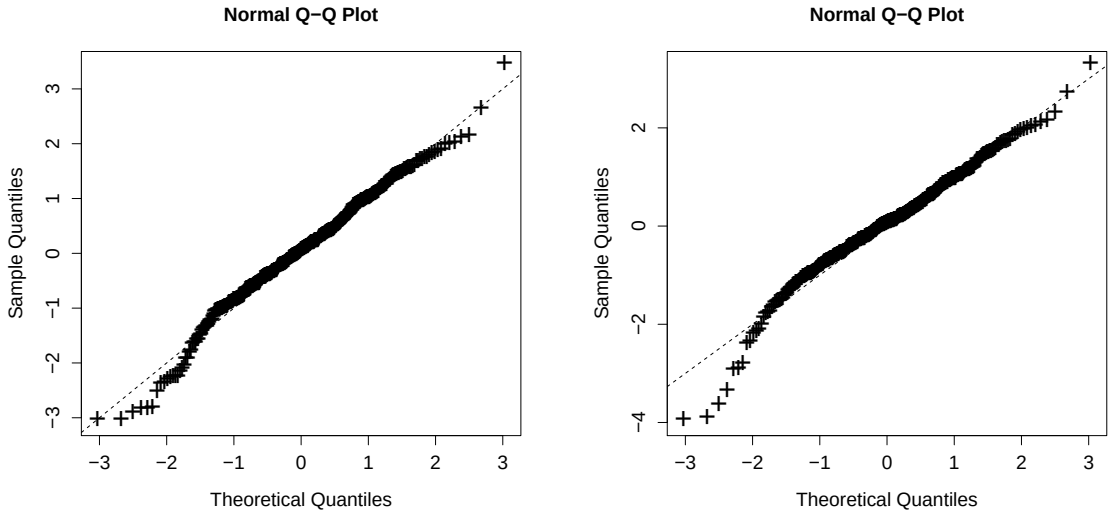
   Hvis man som foreslået i opgaveformuleringen indfører en ny tidsfaktor, hvor niveauerne `tid=7` og `tid=13` slås sammen, så er det muligt at lave eet samlet test for, om der sker en ændring fra 7 til 13 måneder. Den tilhørende likelihood ratio teststørrelse bliver 18.14 og den approksimative p-værdi er < 0.0004, hvilket dokumenterer, at der i datasættet sker en ændring fra 7 til 13 måneder. Dette test giver dog ikke direkte information om, for hvilke kombinationer af `gruppe` og `sex` der ses en ændring.

Eksempel på R-kode som kunne være brugt til løsning af opgave 3

```
> ### fit af modeller til brug ved modelkontrol
> library(gmodels)
> data3<-read.table(file="data3.txt",header=T)
> modkontrol<-lm(vo2max~gruppe*sex*factor(tid)+age+factor(patient),data3)
> lmodkontrol<-lm(log(vo2max)~gruppe*sex*factor(tid)+age+factor(patient),data3)

> ### optegning af residualplot og QQ-plot
> plot(predict(modkontrol),rstandard(modkontrol),cex=2,pch="+",ylim=c(-4,4)
+      ,xlab="Prædikterede værdier",ylab="Standardiserede residualer")
> abline(h=c(-2,0,2),lty=2)
> plot(predict(lmodkontrol),rstandard(lmodkontrol),cex=2,pch="+",ylim=c(-4,4)
+      ,xlab="Prædikterede værdier",ylab="Standardiserede residualer")
> abline(h=c(-2,0,2),lty=2)
> qqnorm(rstandard(modkontrol),cex=2,pch="+")
> abline(0,1,lty=2)
> qqnorm(rstandard(lmodkontrol),cex=2,pch="+")
> abline(0,1,lty=2)



> ### fit af modeller fra opgaveformulering
> library(nlme)
> modelA<-lme(log(vo2max)~gruppe*sex*tid+age,random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T))
> modelB<-lme(log(vo2max)~gruppe*sex*factor(tid)+age,random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T))
> modelC<-lm(log(vo2max)~gruppe*sex*factor(tid)+age+factor(patient),data3)
> modelD<-lme(log(vo2max)~gruppe*sex*factor(tid)+age,random=~1|patient
+ ,data3)
> modelE<-lme(vo2max~gruppe*sex*tid+age,random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T))
> modelF<-lme(vo2max~gruppe*sex*factor(tid)+age,random=~1|patient
```



Figur 1: Residualplot af standardiserede residualer tegnet op imod prædikterede værdier baseret på utransformerede variable (venstre figur) og log-transformerede variable (højre figur).



Figur 2: QQ-plot af standardiserede residualer for statistisk model baseret på utransformerede variable (venstre figur) og log-transformerede variable (højre figur).

```
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T))
> modelG<-lm(vo2max~gruppe*sex*factor(tid)+age+factor(patient),data3)
> modelH<-lme(vo2max~gruppe*sex*factor(tid)+age,random=~1|patient
+ ,data3)

> ### sammenligning af AIC for udvalgte par af modeller
> anova(modelB,modelD)

       Model df       AIC       BIC   logLik   Test  L.Ratio p-value
modelB     1 17 -522.5824 -455.0709 278.2912
modelD     2 15 -514.7801 -455.2112 272.3900 1 vs 2 11.80229  0.0027

> anova(modelF,modelH)

       Model df      AIC      BIC     logLik   Test  L.Ratio p-value
modelF     1 17 46.14926 113.6607  -6.074631
modelH     2 15 57.23878 116.8077 -13.619389 1 vs 2 15.08952   5e-04

> ### modelreduktion med udgangspunkt i modelF
> model0<-lme(vo2max~gruppe*sex*factor(tid)+age,random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T),method="ML")
> model1a<-lme(vo2max~gruppe*sex*factor(tid),random=~1|patient
+ ,data3,corr=corGaus(form=~tid|patient,nugget=T),method="ML")
> model1b<-lme(vo2max~gruppe*sex+gruppe*factor(tid)+sex*factor(tid)+age,random=~1|patient
> anova(model1a,model0) ### test for effekt af age

        Model df       AIC     BIC    logLik   Test  L.Ratio p-value
model1a     1 16  26.16091 90.2231  2.919545
model0      2 17 -14.41808 53.6480 24.209039 1 vs 2 42.57899  <.0001

> anova(model1b,model0) ### test for trefaktorvekselvirkning

        Model df       AIC      BIC   logLik   Test   L.Ratio p-value
model1b     1 15 -18.31993 41.73837 24.15997
model0      2 17 -14.41808 53.64800 24.20904 1 vs 2 0.0981449  0.9521

> model2a<-lme(vo2max~gruppe*factor(tid)+sex*factor(tid)+age,random=~1|patient,data3,corr
> model2b<-lme(vo2max~gruppe*sex+sex*factor(tid)+age,random=~1|patient,data3,corr=corGaus
> model2c<-lme(vo2max~gruppe*sex+gruppe*factor(tid)+age,random=~1|patient,data3,corr=cor(
> anova(model2a,model1b) ### test for effekt af gruppe*sex

        Model df       AIC      BIC   logLik   Test    L.Ratio p-value
model2a     1 14 -20.27909 35.77533 24.13954
model1b     2 15 -18.31993 41.73837 24.15997 1 vs 2 0.04084312  0.8398

> anova(model2b,model1b) ### test for effekt af gruppe*factor(tid)
```

```
        Model df        AIC      BIC   logLik   Test  L.Ratio p-value
model2b     1 13  -0.207615 51.84292 13.10381
model1b     2 15 -18.319934 41.73837 24.15997 1 vs 2 22.11232  <.0001
```

```
> anova(model2c,model1b) ### test for effekt af sex*factor(tid)
```

```
        Model df       AIC      BIC   logLik   Test  L.Ratio p-value
model2c     1 13  -0.67947 51.37106 13.33973
model1b     2 15 -18.31993 41.73837 24.15997 1 vs 2 21.64046  <.0001
```

```
> model3a<-lme(vo2max~gruppe+sex*factor(tid)+age,random=~1|patient,data3,corr=corGaus(for
> model3b<-lme(vo2max~gruppe*factor(tid)+sex+age,random=~1|patient,data3,corr=corGaus(for
> anova(model3a,model2a) ### test for effekt af gruppe*factor(tid)
```

```
        Model df        AIC      BIC   logLik   Test  L.Ratio p-value
model3a     1 12  -2.169471 45.87717 13.08474
model2a     2 14 -20.279090 35.77533 24.13954 1 vs 2 22.10962  <.0001
```

```
> anova(model3b,model2a) ### test for effekt sex*factor(tid)
```

```
        Model df       AIC      BIC   logLik   Test  L.Ratio p-value
model3b     1 12  -2.63566 45.41099 13.31783
model2a     2 14 -20.27909 35.77533 24.13954 1 vs 2 21.64343  <.0001
```

```
> ### genfitter slutmodellen med REML-estimation
> slutmodel<-lme(vo2max~gruppe*factor(tid)+sex*factor(tid)+age,random=~1|patient,data3,co
> summary(slutmodel)
```

```
                            Value   Std.Error  DF    t-value      p-value
(Intercept)            3.01011516 0.152072172 264 19.7939907 4.357101e-54
gruppeK               -0.04660505 0.065402266 131 -0.7125907 4.773664e-01
factor(tid)7           0.34893345 0.024015100 264 14.5297518 1.511095e-35
factor(tid)13          0.36414326 0.030109479 264 12.0939740 4.356221e-27
sexM                   0.44053563 0.095554592 131  4.6103031 9.433176e-06
age                   -0.02129765 0.002980695 131 -7.1451957 5.606995e-11
gruppeK:factor(tid)7  -0.15045355 0.032358218 264 -4.6496242 5.259750e-06
gruppeK:factor(tid)13 -0.08261269 0.040569853 264 -2.0363074 4.271752e-02
factor(tid)7:sexM      0.19904591 0.047593459 264  4.1822115 3.931888e-05
factor(tid)13:sexM     0.25924016 0.059671384 264  4.3444637 1.993079e-05
```

```
                      Estimate Std. Error  t value  DF Pr(>|t|) Lower.CI
(1 0 0 1 1 50 0 0 0 1) 3.009152 0.09345097 32.20033 131        0 2.824283
                      Upper.CI
(1 0 0 1 1 50 0 0 0 1)  3.19402
```

```
> ### udtrækker tilvæksten fra 7 til 13 måneder for de 4 forskellige kombinationer
> ### af gruppe og sex ved brug af estimable
> ###
> ### age sættes til 50 år ved beregningerne men dette har ingen betydning for
> ### de beregnede tilvækster
> m50I7<-c(1,0,1,0,1,50,0,0,1,0)
> m50K7<-c(1,1,1,0,1,50,1,0,1,0)
> f50I7<-c(1,0,1,0,0,50,0,0,0,0)
> f50K7<-c(1,1,1,0,0,50,1,0,0,0)
> m50I13<-c(1,0,0,1,1,50,0,0,0,1)
> m50K13<-c(1,1,0,1,1,50,0,1,0,1)
> f50I13<-c(1,0,0,1,0,50,0,0,0,0)
> f50K13<-c(1,1,0,1,0,50,0,1,0,0)
> diffmI<-m50I13-m50I7
> diffmK<-m50K13-m50K7
> difffI<-f50I13-f50I7
> difffK<-f50K13-f50K7
> differences<-rbind(diffmI,diffmK,difffI,difffK)
> estimable(slutmodel,differences,conf.int=0.95)

         Estimate Std. Error   t value  DF     Pr(>|t|)    Lower.CI   Upper.CI
diffmI 0.07540406 0.04653710 1.6202998 264 0.1063617029 -0.01622705 0.16703517
diffmK 0.14324492 0.04777072 2.9985928 264 0.0029710793  0.04918484 0.23730501
difffI 0.01520981 0.02401510 0.6333438 264 0.5270574586 -0.03207569 0.06249532
difffK 0.08305068 0.02344947 3.5416872 264 0.0004698759  0.03687890 0.12922245

> ### definerer (-som anført i opgaveformuleringen) en ny version
> ### af tid med kun to niveauer
> data3$tidny<-factor(data3$tid) ### laver faktor version af tid
> levels(data3$tidny)            ### nytid har 3 niveauer: 1,7,13

[1] "1"  "7"  "13"

> levels(data3$tidny)<-c("1","7:13","7:13") ### slår 2 af niveauerne sammen
> levels(data3$tidny) ### nytid har nu kun 2 niveauer: 1, 7:13

[1] "1"    "7:13"

> ### tester dernæst model2a mod en tilsvarende model, hvor
> ### tid erstattes af den nye faktor tidny
> modelny<-lme(vo2max~gruppe*tidny+sex*tidny+age,random=~1|patient,data3,corr=corGaus(for
> anova(modelny,model2a)

        Model df        AIC      BIC   logLik   Test  L.Ratio p-value
modelny     1 11  -8.140655 35.90210 15.07033
model2a     2 14 -20.279090 35.77533 24.13954 1 vs 2 18.13844   4e-04
```