Besvarelse af Case 1 fra Statistisk Dataanalyse 2

Dette er et forsøg på at lave en mere punktopstillet besvarelse af case 1.

For at undgå for mange gentagelser foreslår jeg, at man læser nedenstående samtidig med, at man har adgang til pdf-filen `case1sol.pdf` og evt forelæsningsslides fra kursusuge 1.

Delopgave 1

Indlæsning af data til case.

1a

```
kart<-read.table("../data/fosfor.txt",header=T)
dim(kart) ### find antal rækker og søjler i datasættet
```

```
## [1] 27  5
```

```
head(kart,6) ### udskriv de første 6 linjer i datasættet
```

```
##   fosfor p81 p82 blok udbytte
## 1      1   0   0    1     330
## 2      1   0   0    2     320
## 3      1   0   0    3     355
## 4      2   0  20    1     346
## 5      2   0  20    2     350
## 6      2   0  20    3     369
```

Datasættet indeholder 27 observationer og 5 variable med navne

```
names(kart) ### udskriv variabelnavne i datasættet
```

```
## [1] "fosfor"  "p81"     "p82"     "blok"    "udbytte"
```

1b

Vi starter med at lave forskellige tabeller over variablene i datasættet. Da variablen 'udbytte' skal bruges som responsvariabel ser vi ikke på denne.

```
table(kart$fosfor)
```

```
##
## 1 2 3 4 5 6 7 8 9
## 3 3 3 3 3 3 3 3 3
```

```
table(kart$p81)
```

```
##
##  0 30 60
##  9  9  9
```

```
table(kart$p82)
```

```
##
##  0 20 40
```

```
##  9  9  9
```

```
table(kart$blok)
```

```
##
## 1 2 3
## 9 9 9
```

Det ses at alle faktorerne er balancerede (-f.eks. er der 3 målinger for hvert af de 9 niveuaer af fosfor).

Nedenfor undersøges, om R opfatter variablen fosfor som en faktor

```
kart$fosfor
```

```
##  [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6 7 7 7 8 8 8 9 9 9
```

```
is.factor(kart$fosfor)
```

```
## [1] FALSE
```

Vi kan lave variablen fosfor om til en faktor og lægge faktorversionen af fosfor ned i datasættet kart ved at skrive

```
kart$fosfor <- factor(kart$fosfor)
```

Nedenfor vises, at R nu opfatter variablen fosfor som en faktor

```
kart$fosfor
```

```
##  [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6 7 7 7 8 8 8 9 9 9
## Levels: 1 2 3 4 5 6 7 8 9
```

```
is.factor(kart$fosfor)
```

```
## [1] TRUE
```

Kommandoen 'levels' kan bruges til at få R til at udskrive niveauerne for en faktor

```
levels(kart$fosfor)
```

```
## [1] "1" "2" "3" "4" "5" "6" "7" "8" "9"
```

Tilsvarende laves blok om til en faktor

```
kart$blok<-factor(kart$blok)
is.factor(kart$blok)
```

```
## [1] TRUE
```

```
levels(kart$blok)
```

```
## [1] "1" "2" "3"
```

1c

I stedet for at løse dette delspørgsmål følges (-som foreslået i opgaver) opskriften fra delspørgsmål 1d)-1h)

1d

Formuleringen af opgaven er lidt kryptisk:

I første omgang er det meningen, at man skal undersøge, hvordan udbyttet afhænger af faktorerne blok og fosfor. Det naturlige udgangspunkt ville være at starte med den fulde tosidede variansanalysemodel (-også kaldet modellen med vekselvirkning).

```
model1<-lm(udbytte~fosfor:blok,kart)
summary(model1)
```

```
##
## Call:
## lm(formula = udbytte ~ fosfor:blok, data = kart)
##
## Residuals:
## ALL 27 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)         433         NA      NA       NA
## fosfor1:blok1      -103         NA      NA       NA
## fosfor2:blok1       -87         NA      NA       NA
## fosfor3:blok1       -24         NA      NA       NA
## fosfor4:blok1       -65         NA      NA       NA
## fosfor5:blok1       -62         NA      NA       NA
## fosfor6:blok1       -24         NA      NA       NA
## fosfor7:blok1       -73         NA      NA       NA
## fosfor8:blok1       -51         NA      NA       NA
## fosfor9:blok1       -35         NA      NA       NA
## fosfor1:blok2      -113         NA      NA       NA
## fosfor2:blok2       -83         NA      NA       NA
## fosfor3:blok2       -70         NA      NA       NA
## fosfor4:blok2       -93         NA      NA       NA
## fosfor5:blok2       -60         NA      NA       NA
## fosfor6:blok2       -23         NA      NA       NA
## fosfor7:blok2       -37         NA      NA       NA
## fosfor8:blok2       -26         NA      NA       NA
## fosfor9:blok2       -18         NA      NA       NA
## fosfor1:blok3       -78         NA      NA       NA
## fosfor2:blok3       -64         NA      NA       NA
## fosfor3:blok3       -19         NA      NA       NA
## fosfor4:blok3       -67         NA      NA       NA
## fosfor5:blok3       -30         NA      NA       NA
## fosfor6:blok3       -31         NA      NA       NA
## fosfor7:blok3       -27         NA      NA       NA
## fosfor8:blok3        -8         NA      NA       NA
## fosfor9:blok3        NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:    NaN
## F-statistic:   NaN on 26 and 0 DF,  p-value: NA
```

Output ser mærkeligt ud: R kan ikke estimere variansen inden for grupperne.

Dette skyldes, at produktfaktorer fosfor x blok har 27 niveauer (=9 gange 3), og at der i datasættet kun er een måling per gruppe. Dette kan f.eks. ses af følgende tabel.

```
table(kart$fosfor,kart$blok)
```

```
##
##     1 2 3
##   1 1 1 1
##   2 1 1 1
##   3 1 1 1
##   4 1 1 1
##   5 1 1 1
##   6 1 1 1
##   7 1 1 1
##   8 1 1 1
##   9 1 1 1
```

VIGTIGT: når man ikke har gentagelser for produktfaktoren, så kan man ikke tage udgangspunkt i denne statistiske model. Datasættet er for lille, og analysen bliver nødt til at tage udgangspunkt i en model med færre parametre. Derfor tager vi i stedet udgangspunkt i den additive model for tosidet variansanalyse.

```
model2<-lm(udbytte~fosfor+blok,kart)
```

Opskrevet 'på papir' ser modellen ud som på slide 12 fra forelæsningen d. 7/9-2017.

1e

Nedenfor testes hypotesen om, at der er en sammenhæng mellem fosforbehandling og udbytte.

```
model3<-lm(udbytte~blok,kart)
anova(model3,model2)
```

```
## Analysis of Variance Table
##
## Model 1: udbytte ~ blok
## Model 2: udbytte ~ fosfor + blok
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     24 21378.4
## 2     16  3469.1  8     17909 10.325 4.977e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Da p-værdien er meget lille konkluderes, at den reducerede model3 (-hvor fosfor er fjernet) beskriver variationen i udbytte meget dårligere end den additive model2, hvor fosfor er med. Konklusionen er, at fosfor bidrager væsentligt til at forklare variationen i udbytte. Mao er der en stærk sammenhæng mellem fosfor behandling og udbytte.

Problemet er, at ovenstående test kun undersøger om udbyttet for alle 9 fosforbehandlinger kan antages at være helt ens eller om der er nogle forskelle. Vi får ikke svar på, hvilke fosforbehandlinger, der giver forskelligt udbytte, og specielt får vi ikke svar på, om det er førsteårs- eller andenårs fosfortilførslen, der er associeret med udbyttet.

1f+1g

Det er essentielt at indse, at fosfor faktoren er det samme som produktfaktorer af p81 og p82. Det betyder, at den model vi ovenfor fittede som 'model2' lige så godt kan skrives som

```
model2alt<-lm(udbytte~blok+factor(p81):factor(p82),kart)
```

Man kan f.eks. lave en tabel af sammenhørende værdier af fosfor og p81 x p82 men henblik på at se, hvordan niveauerne for de to faktorer (fosfor hhv p81xp82) svarer til hinanden

```
table(kart$fosfor,factor(kart$p81):factor(kart$p82))
```

```
##
##     0:0 0:20 0:40 30:0 30:20 30:40 60:0 60:20 60:40
##   1   3    0    0    0     0     0    0     0     0
##   2   0    3    0    0     0     0    0     0     0
##   3   0    0    3    0     0     0    0     0     0
##   4   0    0    0    3     0     0    0     0     0
##   5   0    0    0    0     3     0    0     0     0
##   6   0    0    0    0     0     3    0     0     0
##   7   0    0    0    0     0     0    3     0     0
##   8   0    0    0    0     0     0    0     3     0
##   9   0    0    0    0     0     0    0     0     3
```

Heraf se f.eks. at fosfor=3 svarer præcis til p81=0,p82=40.

Den nye formulering af udgangsmodellen (dvs. model2alt) kan nu forsøges reduceret ved at fjerne p82 (dvs. andenårstilførslen fra modellen)

```
model4<-lm(udbytte~blok+factor(p81),kart)
anova(model4,model2alt)
```

```
## Analysis of Variance Table
##
## Model 1: udbytte ~ blok + factor(p81)
## Model 2: udbytte ~ blok + factor(p81):factor(p82)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     22 13935.8
## 2     16  3469.1  6     10467 8.0456 0.0004028 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Den meget lave p-værdi viser, at p82 ikke kan fjernes fra modellen, dvs. at niveauet af p82 hænger sammen med udbyttet.

1h

Med udgangspunkt i den additive model (dvs. enten 'model2' eller 'model2alt') kunne man også undersøge, om man kan teste blok-effekten væk

```
model5<-lm(udbytte~fosfor,kart)
anova(model5,model2)
```

```
## Analysis of Variance Table
##
## Model 1: udbytte ~ fosfor
## Model 2: udbytte ~ fosfor + blok
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     18 6417.3
## 2     16 3469.1  2    2948.2 6.7988 0.007293 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Med en p-værdi på 0.0073 tyder det også må, at blok bidrager signifikant til beskrivelsen af variationen i udbyttet.

Supplement

Man kan også tage udgangspunkt i en den additive model med effekt af blok og fosfor og udnytte at fosfor (opfattet som faktor på 9 niveauer) er identisk med vekselvirkningen af p81 og p82. Dernæst kan man lave et formelt test for, om effekten af p81 og p82 kan opfattes som en additiv effekt (i modeller hvor man stadig medtager en potentiel effekt af blok). Dette test er foretaget nedenfor.

```
model2alt<-lm(udbytte~blok+factor(p81):factor(p82),kart)
model2add<-lm(udbytte~blok+factor(p81)+factor(p82),kart)
anova(model2add, model2alt)
```

```
## Analysis of Variance Table
##
## Model 1: udbytte ~ blok + factor(p81) + factor(p82)
## Model 2: udbytte ~ blok + factor(p81):factor(p82)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     20 4472.9
## 2     16 3469.1  4    1003.8 1.1574  0.366
```

Konklusion:

Der ser ikke ud til at være en væsentlig vekselvirkning mellem p81 og p82.

Delopgave 2

Indlæsning af data til case.

2a-b

```
kartmeans<-read.table("../data/fosfor-means.txt",header=T)
dim(kartmeans) ### find antal rækker og søjler i datasættet
```

```
## [1] 9 3
```

```
kartmeans ### udskriv  datasættet
```

```
##   p81 p82 yield.mean
## 1   0   0      335.0
## 2   0  20      355.0
## 3   0  40      395.3
## 4  30   0      358.0
## 5  30  20      382.3
## 6  30  40      407.0
## 7  60   0      387.3
## 8  60  20      404.7
## 9  60  40      415.3
```

Datasættet indeholder 9 observationer og 3 variable med navne

```
names(kartmeans) ### udskriv variabelnavne i datasættet
```

```
## [1] "p81"        "p82"        "yield.mean"
```

Tabeller over variablene i datasættet

```
table(kartmeans$p81)
```

```
##
##  0 30 60
##  3  3  3
```

```
table(kartmeans$p82)
```

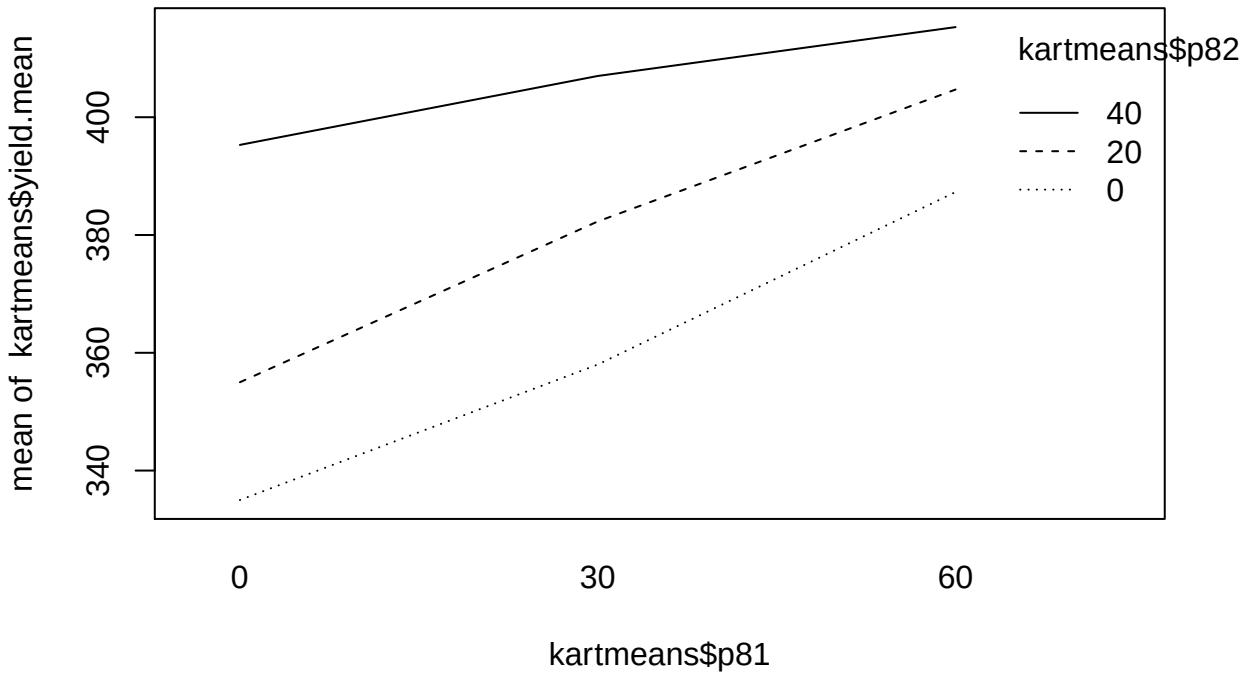
```
##
##  0 20 40
##  3  3  3
```

```
table(kartmeans$p81,kartmeans$p82)
```

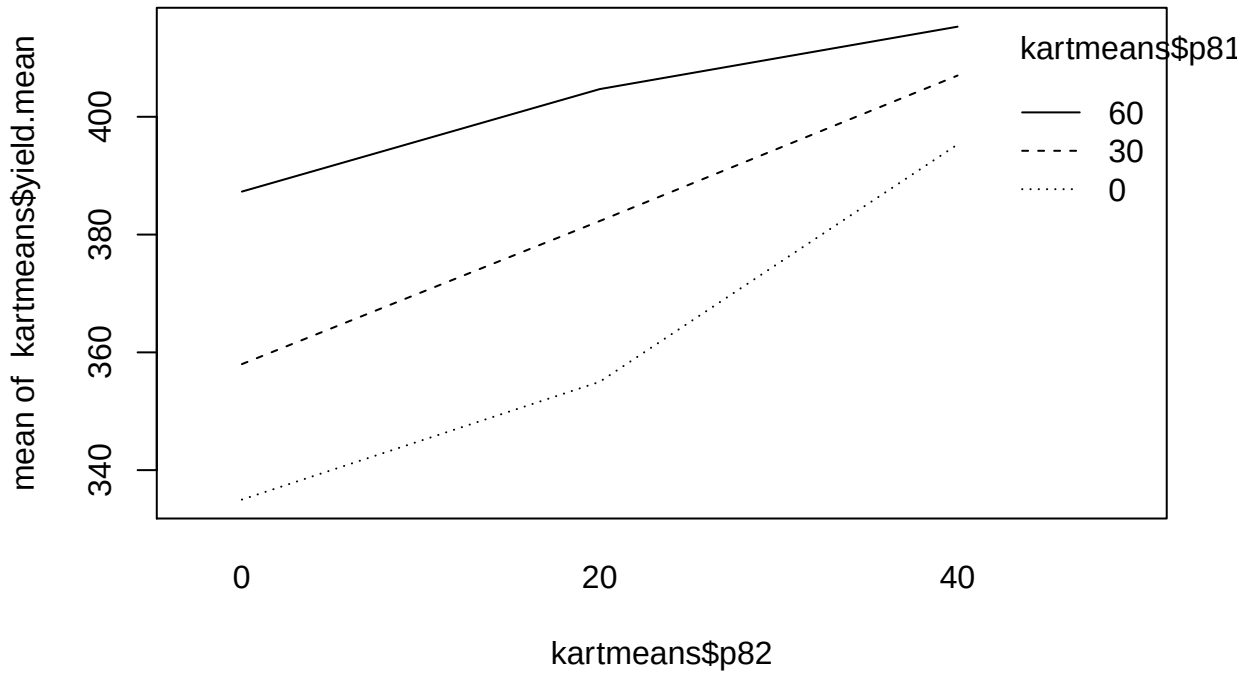
```
##
##      0 20 40
##   0  1  1  1
##   30 1  1  1
##   60 1  1  1
```

Det virker oplagt at optegne data som i et interaction-plot. Man kan overveje, om p81 eller p82 skal ud af første aksen

```
interaction.plot(kartmeans$p81,kartmeans$p82,kartmeans$yield.mean)
```

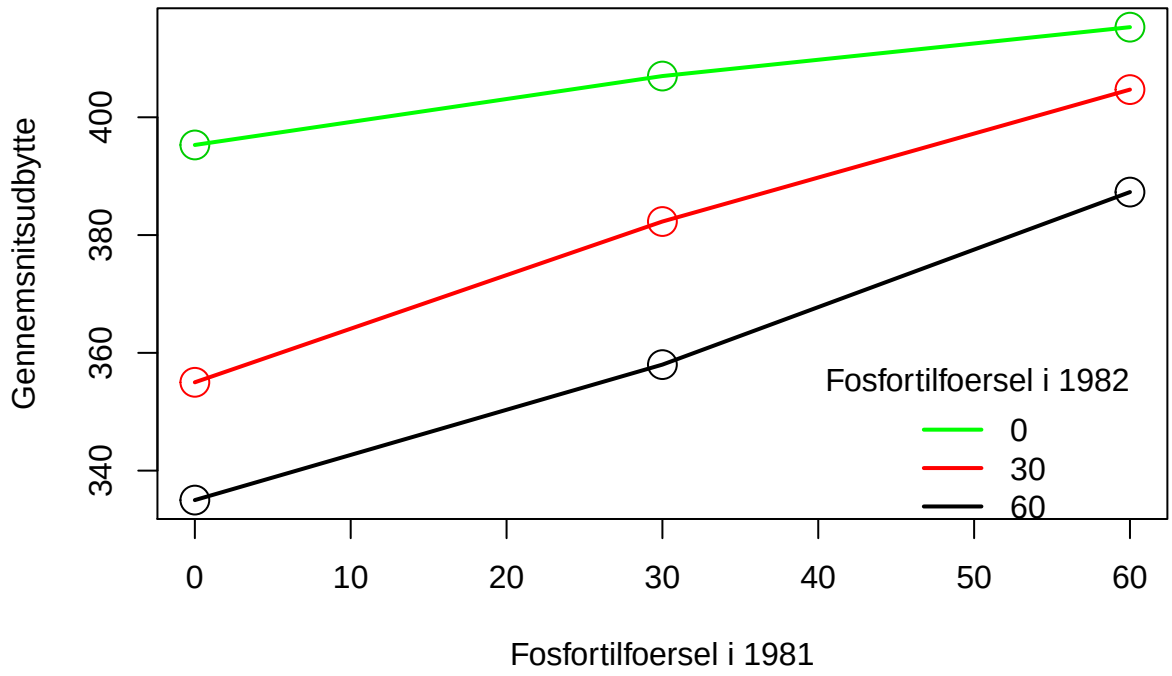



```
interaction.plot(kartmeans$p82,kartmeans$p81,kartmeans$yield.mean)
```



Når man får lidt erfaring kan man lave sine egne plots

```
plot(kartmeans$p81,kartmeans$yield.mean
     ,cex=2,col=as.numeric(factor(kartmeans$p82))
     ,ylab="Gennemsnitsudbytte",xlab="Fosfortilfoersel i 1981")
lines(c(0,30,60),kartmeans$yield.mean[c(1,4,7)],col="black",lwd=2)
lines(c(0,30,60),kartmeans$yield.mean[c(2,5,8)],col="red",lwd=2)
lines(c(0,30,60),kartmeans$yield.mean[c(3,6,9)],col="green",lwd=2)
legend(x=40,y=360,lwd=2,col=c("green","red","black"),legend=c(0,30,60),title="Fosfortilfoersel i 1982"
       ,bty="n")
```



2c+d+e

Additiv variansanalysemodel fittes i R og estimaterne udskrives

```
mod1<-lm(yield.mean~factor(p81)+factor(p82),kartmeans)
summary(mod1)
```

```
##
## Call:
## lm(formula = yield.mean ~ factor(p81) + factor(p82), data = kartmeans)
##
## Residuals:
##       1       2       3       4       5       6       7       8
##  -4.6556  -5.2222   9.8778  -2.3222   1.4111   0.9111   6.9778   3.8111
##       9
## -10.7889
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    339.656      6.814  49.843  9.7e-07 ***
## factor(p81)30   20.667      7.465   2.769  0.05041 .
## factor(p81)60   40.667      7.465   5.448  0.00551 **
## factor(p82)20   20.567      7.465   2.755  0.05110 .
## factor(p82)40   45.767      7.465   6.131  0.00359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.143 on 4 degrees of freedom
## Multiple R-squared:  0.944,  Adjusted R-squared:  0.888
## F-statistic: 16.85 on 4 and 4 DF,  p-value: 0.009064
```

Fortolkning af estimaterne:

R vælger gruppen p81=0,p82=0 som referencegruppe. Ud for (Intercept) aflæses at det estimerede udbytte for denne gruppe er 339.66.

Desuden angives kontrasterne for p81 og p82 dvs. hvor meget udbyttet ændrer sig i forhold til referencebehandlingen (p81=0,p82=0), hvis man ændrer niveauerne af enten p81 eller p82.

p81=30 øger således udbyttet med 20.67 p81=60 øger udbyttet med 40.67 p82=20 øger udbyttet med 20.57 p82=40 øger udbyttet med 45.77

Det ses at øgningen af udbyttet næsten øges med dobbelt så meget når p82=40 end når p82=20. Dette tyder på en lineær sammenhænge mellem udbytte og tilførslen p82.

Et tilsvarende argument antyder, at udbyttet vokser næsten lineært med værdien af p81.

2f

Det er måske lidt teknisk, hvad man skal gøre her.

Det simpleste er at dele datasættet 'kartmeans' op i 3 deldatasæt og så fitte en lineær regressionsmodel for udbyttet som funktion af p82 for hver deldatasæt.

```
data0<-subset(kartmeans,p81==0)
data30<-subset(kartmeans,p81==30)
data60<-subset(kartmeans,p81==60)
```

```
modlinje0<-lm(yield.mean~p82,data0)
modlinje0 ### giver skæring og hældning for linjen fittet til data med p81=0
```

```
##
## Call:
## lm(formula = yield.mean ~ p82, data = data0)
##
## Coefficients:
## (Intercept)          p82
##     331.617        1.508
```

```
modlinje30<-lm(yield.mean~p82,data30)
modlinje30 ### giver skæring og hældning for linjen fittet til data med p81=30
```

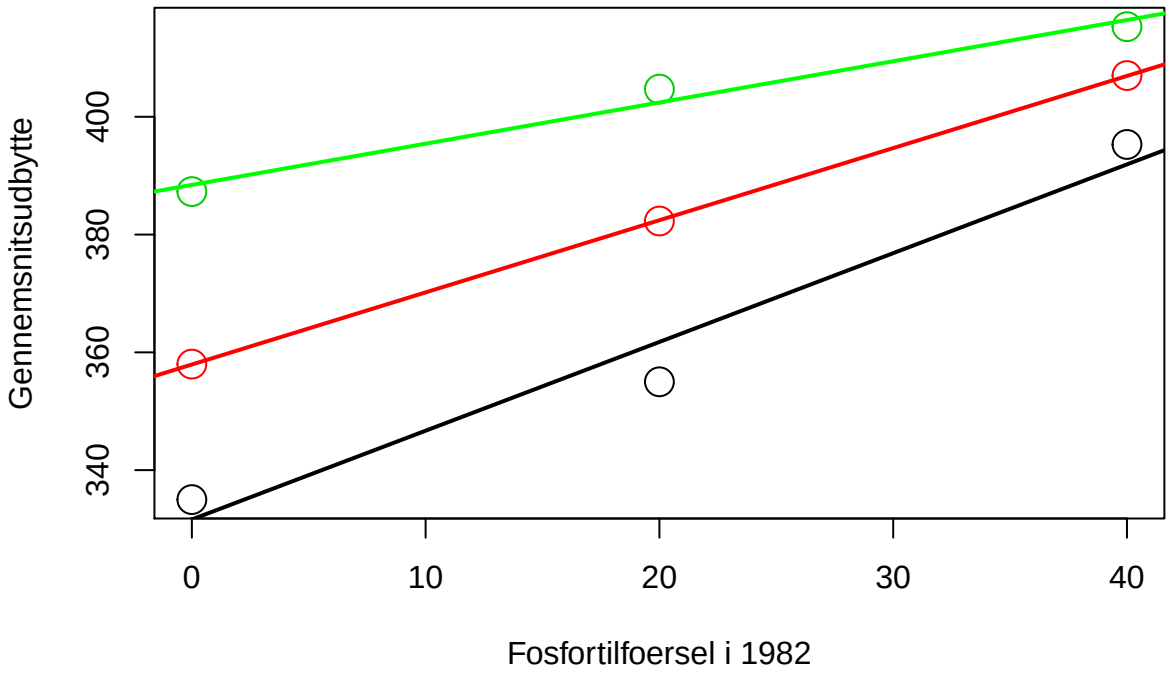
```
##
## Call:
## lm(formula = yield.mean ~ p82, data = data30)
##
## Coefficients:
## (Intercept)          p82
##     357.933        1.225
```

```
modlinje60<-lm(yield.mean~p82,data60)
modlinje60 ### giver skæring og hældning for linjen fittet til data med p81=60
```

```
##
## Call:
## lm(formula = yield.mean ~ p82, data = data60)
##
## Coefficients:
## (Intercept)          p82
##       388.4          0.7
```

Nedenfor vises, hvordan man kan lave en figur hvor de rå datapunkt og de tilpassede linjer er blevet indtegnet.

```
plot(kartmeans$p82,kartmeans$yield.mean
     ,cex=2,col=as.numeric(factor(kartmeans$p81))
     ,ylab="Gennemsnitsudbytte",xlab="Fosfortilfoersel i 1982")
abline(modlinje0,lwd=2,col="black")
abline(modlinje30,lwd=2,col="red")
abline(modlinje60,lwd=2,col="green")
```



2g-h

Den foreslåede model kan fittes i R ved at lade både p81 og p82 indgå (-som numeriske variable dvs ikke som faktorer).

```
modbegge<-lm(yield.mean~p81+p82,kartmeans)
summary(modbegge)
```

```
##
## Call:
## lm(formula = yield.mean ~ p81 + p82, data = kartmeans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1278  -3.9944   0.0889   2.1556  10.5389
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 338.9944     5.0575  67.029 7.42e-10 ***
## p81           0.6778     0.1032   6.565 0.000598 ***
## p82           1.1442     0.1549   7.389 0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.586 on 6 degrees of freedom
## Multiple R-squared:  0.9421, Adjusted R-squared:  0.9229
## F-statistic: 48.85 on 2 and 6 DF,  p-value: 0.0001937
```

Fortolkningen af parameterestimatet for p81 er, at hver gang man øger p81 med 1 enhed, så vil yield.mean i gennemsnit ændre sig med 0.678. Specielt vil en ændring på 30 modsvares af en ændring på 20.34, hvilken svare meget godt til, hvad vi så i opgave 2e.

Tilsvarende øges yield.mean med 1.144 hver gang p82 øges ved en enhed (jf estimatet ud for p82 ovenfor).

Man kunne derfor vælge at kvantificere andenårsvekselvirkningen i forhold til førsteårsvirkningen ved at tage forholdet 1.144/0.678=1.689.