Vejledende besvarelse af hjemmeopgave, efterår 2018

Udleveret 1. oktober, afleveres senest ved øvelserne i uge 44 (30. oktober.-1. november).

Der er foretaget en del undersøgelser af krigsveteraner og deres helbredsforhold, og vi skal se på et par aspekter af dette.
På hjemmesiden
`http://staff.pubhealth.ku.dk/~lts/basal18_2/hjemmeopgave.html`
ligger oplysninger om 7 variable på 299 krigsveteraner (et udpluk af et meget større materiale), og disse er:

idnr: Løbenummer for veteranen

alder: Veteranens alder ved andersøgelsen (år)

bpright: Systolisk blodtryk på højre arm (mm)

bpleft: Systolisk blodtryk på venstre arm (mm)

fev1: Lungefunktionsmål (l/min)

ryger: 3 kategorier for rygning: "current", "never" og "ex"

alkuge: Antal genstande pr. uge

I de første 4 spørgsmål fokuserer vi på blodtrykket i de to arme:

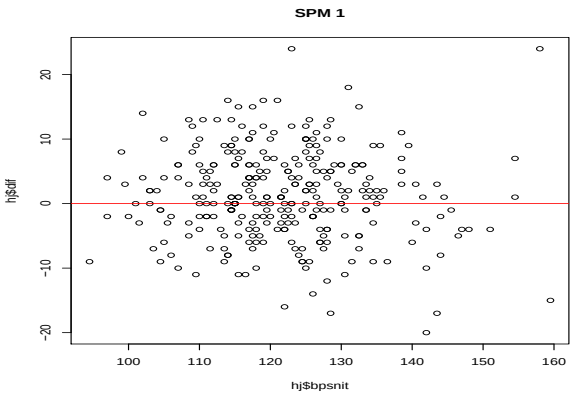
1. *Er det rimeligt at vurdere forskellen på blodtrykket på højre og venstre arm ud fra differenser på selve blodtryks-skalaen? Eller skal forskellen måske hellere ses relativt til niveauet?*

   Først definerer vi differensen af blodtrykket mellem de to arme (højre minus venstre), og ligeledes gennemsnittet af disse, for derefter at tegne et Bland-Altman plot:

```
hj=read.table(
  "http://publicifsv.sund.ku.dk/~lts/basal18_2/hjemmeopgave/hjemmeopgave.txt",
  header=T)

hj$dif=hj$bpright-hj$bpleft
hj$bpsnit=(hj$bpright+hj$bpleft)/2
plot(hj$bpsnit, hj$dif, main="SPM 1")
abline(a=0, b=0, col="red",lty=1)
```

hvorved vi får figuren:

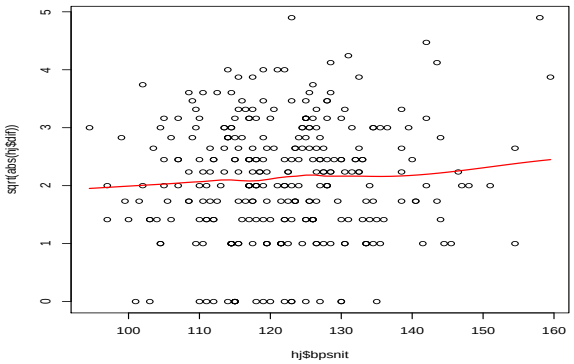


Da denne figur ikke viser udpræget tegn på noget mønster, men snarere et mere eller mindre vandret bånd, vil vi anse differensen for at være et udmærket udtryk for sideforskelle i blodtryk, da disse åbenbart er rimeligt ens over hele skalaen.

Der synes dog at kunne spores en lille tendens til øget spredning for de højeste blodtryk, men en tilsvarende figur lavet efter logaritmetransformation viser stort set samme billede, så vi vil derfor ikke komplicere analyserne med at foretage logaritmetransformation. Desuden viser en figur af kvadratroden af de numeriske forskelle i blodtryk heller ikke nogen synderlig afhængighed af niveauet. Figuren er lavet således:

```
scatter.smooth(hj$bpsnit,sqrt(abs(hj$dif)),
               lpars=list(col="red",lwd=2))
```

hvorved vi får:



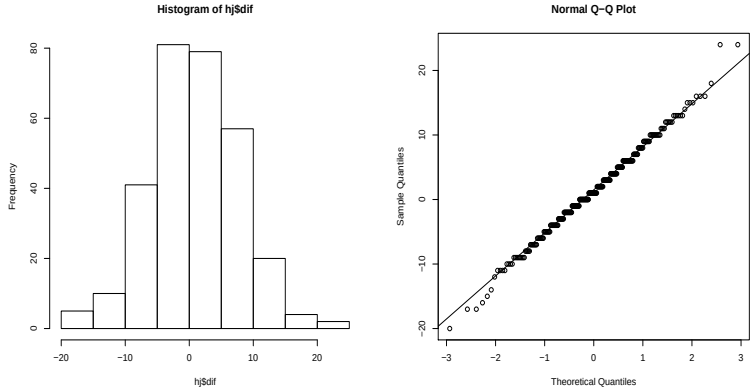
Bemærk, hvad der *ikke* spørges om i dette spørgsmål: Fordelingen af blodtryk på hver af siderne, og egentlig heller ikke noget om fordelingen af differenserne før i spørgsmål 2 (og endnu mere i spørgsmål 3).

2. *Kvantificer forskellen i blodtryk på højre og venstre arm, med konfidensinterval. Er der signifikant forskel? Kan vi udelukke en middelforskel på 10 mm på de to sider?*

I lyset af svaret på ovenstående spørgsmål, arbejder vi videre med blodtrykket uden at transformere. Når vi skal sammenligne blodtrykket på højre og venstre side, skal vi derfor se på et parret T-test, eller blot et test for middelværdi 0 for differenserne. Vi udfører førstnævnte nedenfor, men checker lige først, om differenserne er rimeligt normalfordelte ved at tegne histogram og/eller fraktildiagram:

```
hist(hj$dif)

qqnorm(hj$dif)
qqline(hj$dif)
```



Da disse på ingen måde giver anledning til bekymring, fortsætter vi til T-testet:

```
t.test(hj$bpright,hj$bpleft,paired=T)
```

som giver outputtet:

```
> t.test(hj$bpright,hj$bpleft,paired=T)

Paired t-test

data:  hj$bpright and hj$bpleft
t = 3.4433, df = 298, p-value = 0.0006571
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.5918218 2.1707201
sample estimates:
mean of the differences
               1.381271
```

Heraf aflæser vi:

- Forskellen på blodtrykket på højre og venstre side estimeres til 1.38 mm, med konfidensinterval CI=(0.59 mm, 2.17 mm).
- Forskellen er signifikant, idet konfidensintervallet ikke indeholder 0. P-værdien angives til P=0.00066.
- En forskel på 10 mm (til en af siderne) falder langt udenfor konfidensintervallet, så ja, det kan vi udelukke.

Bemærk, at der i dette spørgsmål *ikke* bør siges noget om hverken fordelingen af blodtryk på hver af de to sider, og der bør helt sikkert heller ikke laves et uparret T-test! Ligeledes har normalområdet ingen plads her, det kommer først i spørgsmål 3.

3. *I ambulatoriet har vi en veteran, der har et blodtryk på 135 mm på højre side, men kun 122 på venstre side. Er det så usædvanligt, at der må laves en nærmere undersøgelse af dette individ?*

   Forskellen på de to sider ses her at være 13 mm (højre er højest, derfor en positiv differens i henhold til definitionen), og spørgsmålet er, om dette er usædvanligt? For at svare på dette, skal vi udregne et normalområde for differensen, og til dette skal vi bruge gennemsnit og spredning (da vi tidligere har set, at fordelingen af differenserne ser rigtigt pænt normalfordelte ud).

   Vi udregner normalområdet som

```
> mean(hj$dif)+c(-2,0,2)*sd(hj$dif)
[1] -12.491853   1.381271  15.254395
```

   eller med den præcise fraktil (det "*korrekte 2-tal*"):

```
> mean(hj$dif)+qt(0.975,298)*c(-1,0,1)*sd(hj$dif)
[1] -12.269581   1.381271  15.032123
```

   Ud fra dette normalområde kan vi se, at det ikke er voldsomt ekstremt at have et 13 mm højere blodtryk på højre side end på venstre side.

   Hvis vi ikke rigtigt tør tro på normalfordelingsantagelsen, kan vi i stedet udregne de empiriske fraktiler:

```
quantile(hj$dif, probs = c(0.025,0.05,0.95,0.975))
```

   hvorved vi får outputtet:

```
> quantile(hj$dif, probs = c(0.025,0.05,0.95,0.975))
 2.5%    5%   95% 97.5%
-11.0  -9.1  13.0  15.0
```

Her ser vi samstemmende, at sideforskellen på 13 mm (med det største på højre side) ikke er ekstremt. Hvis det have været omvendt, så det var venstre side, der lå 13 mm højere end højre side, så ville det faktisk være lidt usædvanligt....

Bemærk, at det i dette spørgsmål er *helt forkert* at benytte konfidensinterval til at vurdere enkeltpersoner!

4. *Lav en gruppering af personerne, baseret på, om den numeriske differens på blodtrykket overstiger 10 mm eller ej*
   (kald f.eks. variablen `sideforskel`*, med værdierne stor/lille).*
   Undersøg, om et alkoholforbrug på 10 genstande eller derover pr. uge (her kaldet drankere) giver større risiko for en stor sideforskel.
   Kan risikoen ligefrem være fordoblet hos drankere?

   Vi skal lave nogle nye variable ved at dikotomisere (Bemærk, at jeg i nedenstående har ladet en numerisk sideforskel på 10 betyde "*stor*". En del af jer har ladet denne værdi været "*lille*" og får så nogle lidt andre antal):

```
hj$dranker=as.numeric(hj$alkuge>=10)

hj$side <- findInterval(abs(hj$dif), c(10))
hj$sideforskel <- factor(hj$side,levels=0:1, labels=c("lille", "stor"))
```

   og herefter ser vi på sammenhængen mellem det at være dranker (jeg beklager den noget overdrevne sprogbrug) og det at have en stor sideforskel, i form af en to-gange-to tabel:

```
spm4=table(hj$dranker,hj$sideforskel)
```

   Først ser vi på tabellen med tilhørende rækkeprocenter:

```
spm4
prop.table(spm4,1)*100
```

som giver outputtet

```
> spm4

    lille stor
  0   187   44
  1    59    9

> prop.table(spm4,1)*100

       lille     stor
  0 80.95238 19.04762
  1 86.76471 13.23529
```

Af rækkeprocenterne ses, at risikoen for stor sideforskel er 19.05% for folk, der ikke er drankere, men kun 13.24% for drankere, så umiddelbart går forskellen i den modsatte retning af forventet (ud fra spørgsmålets ordlyd).

Et test for uafhængighed (ingen forskel på drankere og ikke-drankere) giver

```
> chisq.test(spm4)

Pearson's Chi-squared test with Yates' continuity correction

data:  spm4
X-squared = 0.85105, df = 1, p-value = 0.3563
```

og dermed ingen signifikant forskel. Det er ikke nødvendigt at lave et eksakt test, da de forventede værdier her på ingen måde er lave:

```
> chisq.test(spm4)$expected

       lille     stor
  0 190.05351 40.94649
  1  55.94649 12.05351
```

men derfor kan vi jo godt gøre det alligevel:

```
> fisher.test(spm4)

Fisher's Exact Test for Count Data
```

```
data:  spm4
p-value = 0.366
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.2627539 1.4511372
sample estimates:
odds ratio
 0.6491732
```

og vi bekræftes da også i, at der ikke er nogen forskel at finde (P=0.37).

Kan vi så udelukke at der kan være en sammenhæng? Næh, men vi kan kvantificere den mulige sammenhæng, f.eks. ved at udregne den relative risiko for stor sideforskel, for drankere (1, nederste række nedenfor) vs. ikke-drankere (0, øverste række nedenfor). Denne udregning kræver pakken `epitools`:

```
install.packages("epitools")
library("epitools")
epitab(spm4,method="riskratio")
```

hvorved vi får:

```
> epitab(spm4,method="riskratio")
$tab

    lille        p0 stor        p1 riskratio     lower    upper  p.value
  0   187 0.8095238   44 0.1904762 1.0000000        NA       NA       NA
  1    59 0.8676471    9 0.1323529 0.6948529 0.3576705 1.349903 0.366049

$measure
[1] "wald"

$conf.level
[1] 0.95

$pvalue
[1] "fisher.exact"
```

Denne relative risiko estimeres altså til 0.695, hvilket betyder, at risikoen for en stor sideforskel er ca. 30.5% *mindre* hos drankere i forhold til hos ikke-drankere. Men sikkerhedsintervallet er (0.358, 1.350), dvs. den relative risiko kunne også være op til 1.350, hvilket svarer til en større risiko for drankere på op til 35% højere end hos ikke-drankere.

En faktor 2 kommer vi altså ikke op på, når spørgsmålet går på drankere vs. ikke-drankere. Men vi kan på den anden side ikke udelukke, at *ikke*-drankere har dobbelt så stor risiko for stor sideforskel, fordi $\frac{1}{2}$ faktisk er indeholdt i konfidensintervallet ovenfor.

I dette spørgsmål er der mange, der får vendt konklusionen på hovedet, fordi de ikke selv først regner størrelserne ud. Man kan let blive snydt af, hvilken vej, resultaterne vender i programudskrifterne.

Bemærk også, at vi her ikke har med små sandsynligheder at gøre, og derfor er odds ratio ikke tæt på relativ risiko.

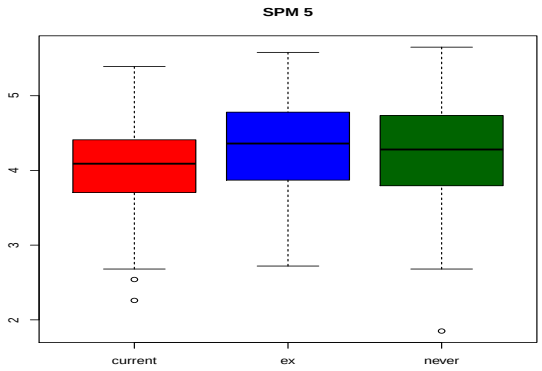
Nu kommer nogle spørgsmål, der fokuserer på lungefunktionen:

5. *Beskriv lungefunktionens relation til rygning ved at sammeligne de 3 rygergrupper.*
   Hvor stor er den estimerede forskel mellem nuværende rygere og aldrig-rygere?

   I første omgang ser vi på et Boxplot af lungefunktionen, opdelt på de 3 grupper:

```
boxplot(fev1~ryger, data=hj,main="SPM 5",
        col=c("red","blue","darkgreen"))
```

   hvilket giver figuren:



Her ses rimeligt symmetriske fordelinger, så vi vil med sindsro gå videre til at foretage en ensidet variansanalyse til sammenligning af grupperne.

Først vil vi dog lige udregne nogle summariske mål i form af gennemsnit og spredning:

```
gennemsnit=tapply(hj$fev1,hj$ryger,mean)
sd=tapply(hj$fev1,hj$ryger,sd)
N=tapply(hj$fev1,hj$ryger,length)
```

hvorved vi får

```
> cbind(N,gennemsnit,sd)
          N gennemsnit        sd
current 139   4.059209 0.5904102
ex       89   4.349438 0.6482325
never    71   4.228028 0.7619292
```

Disse størrelser giver ikke noget helt klart billede af forskelle, og slet ikke i en forståelig rækkefølge, fordi det ser ud som om ex-rygerne klarer sig bedst mht lungefunktion.

For at foretage en egentlig sammenligning mellem de tre grupper, kaster vi os nu ud i den ensidede variansanalyse, incl. test for varianshomogenitet, som i R umiddelbart er Bartletts test:

```
bartlett.test(fev1~ryger, data=hj)
model5 = lm(fev1~ryger, data=hj)
```

Herved får vi

```
> bartlett.test(fev1~ryger, data=hj)

Bartlett test of homogeneity of variances

data:  fev1 by ryger
Bartlett's K-squared = 6.3021, df = 2, p-value = 0.04281


> summary(model5)
```

```
Call:
lm(formula = fev1 ~ ryger, data = hj)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.05921    0.05528  73.433  < 2e-16 ***
rygerex      0.29023    0.08848   3.280  0.00116 **
rygernever   0.16882    0.09507   1.776  0.07679 .
---

Residual standard error: 0.6517 on 296 degrees of freedom
Multiple R-squared:  0.03631,Adjusted R-squared:  0.0298
F-statistic: 5.577 on 2 and 296 DF,  p-value: 0.004193

> confint(model5)
                 2.5 %    97.5 %
(Intercept)  3.95042170 4.1679956
rygerex      0.11610936 0.4643498
rygernever  -0.01827333 0.3559124
```

Vi bemærker i outputtet ovenfor, at Bartletts test for varianshomogenitet ikke bliver godkendt (P=0.043), så variansanalysen er ikke helt pålidelig.

Hvis vi i stedet checker med Levenes test, får vi imidlertid noget andet (også noget andet end det, vi får i SAS....):

```
> install.packages("car")
> library(car)

> leveneTest(fev1~ryger, data=hj)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   2  2.6285 0.07387 .
      296
---
```

Under alle omstændigheder ser boxplotett ikke ud til at have synderligt forskellige varianser, og en evt signifikant forskel kan skyldes de relativt store grupper.

Testet for identitet af de 3 middelværdier giver i ANOVA P=0.0042, og den største forskel ses at være mellem ex-rygere og nuværende rygere, men det er ikke den forskel, vi søger. Forskellen mellem aldrig-rygere og nuværende rygere estimeres til 0.169 i aldrig-rygernes favør, med

CI=(-0.018, 0.356), og P=0.077, altså ikke signifikant.

Pga de uens varianser, checker vi lige med et Welch test, om signifikansen hænger på antagelsen om identiske varianser. Da Welch testet (se nedenfor) giver P=0.0030, føler vi os dog ret sikre på, at middelværdierne i de 3 grupper ikke er ens.

```
> oneway.test(fev1 ~ ryger, data=hj)

One-way analysis of means (not assuming equal variances)

data:  fev1 and ryger
F = 6.0365, num df = 2.00, denom df = 155.35, p-value = 0.002987
```

Hvis vi havde valgt kun at se på de to relevante grupper, kunne vi i stedet benytte et T-test, der kun involverer disse to grupper, og hvor vi får et konfidensinterval, der ikke er baseret på antagelsen om ens varianser:

```
t.test(hj$fev1[hj$ryger=="current"],hj$fev1[hj$ryger=="never"])
```

som giver os nedenstående output:

```
> t.test(hj$fev1[hj$ryger=="current"],hj$fev1[hj$ryger=="never"])

Welch Two Sample t-test

data:  hj$fev1[hj$ryger == "current"] and hj$fev1[hj$ryger == "never"]
t = -1.6332, df = 114.08, p-value = 0.1052
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.37358365  0.03594458
sample estimates:
mean of x mean of y
 4.059209  4.228028
```

Forskellen mellem aldrig-rygere og nuværende rygere estimeres her igen til 0.169 i aldrig-rygernes favør, men nu med et lidt bredere konfidensinterval CI=(-0.036, 0.374), og P=0.11, altså endnu mindre i retning af signifikans.

6. *Er det fornuftigt at foretage ovenstående sammenligning, eller kan der være confounding pga f.eks. alder? Begrund svaret, f.eks. ved at angive det forventede fald i FEV_1 over en 10-årig periode.*

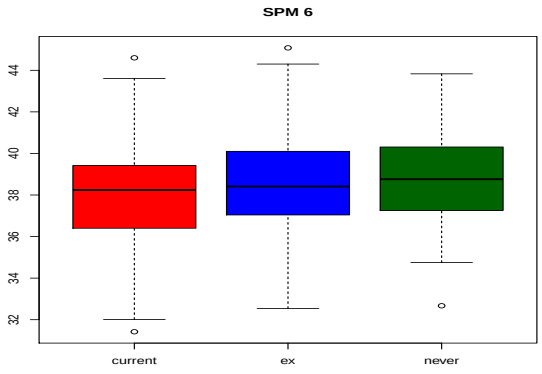
   Man er vist generelt enige om, at lungefunktionen falder med alderen, så hvis de 3 rygergrupper udviser forskelle i aldersfordelingen, så er alder en confounder for sammenligningen af lungefunktionen i de tre rygergrupper.

   Vi laver derfor et Boxplot af aldersfordelingerne samt udregner nogle summariske mål for alderen i de tre grupper:

```
boxplot(alder~ryger, data=hj,main="SPM 6",col=c("red","blue","darkgreen"))

gennemsnit=tapply(hj$alder,hj$ryger,mean)
sd=tapply(hj$alder,hj$ryger,sd)
N=tapply(hj$alder,hj$ryger,length)
```

   Vi får så plottet



   og de tilhørende summariske størrelser:

```
> cbind(N,gennemsnit,sd)
          N gennemsnit       sd
current 139   37.98698 2.565143
ex       89   38.58629 2.498733
never    71   38.71282 2.124745
```

Her ses små forskelle i alder, idet rygerne er ca 1 år yngre end aldrig-rygerne. Vi forventer således ikke den helt store confounding fra alder, men må dog erkende, at den forskel, vi fandt i spørgsmål 5 formentlig er undervurderet, fordi vi har sammenlignet nogle unge rygere med nogle ældre aldrig-rygere.

Faktisk er der næsten signifikant forskel på aldersfordelingerne (specielt hvis vi kun sammenligner rygere med aldrig-rygere), hvilket man kan se ved at foretage en ensidet variansanalyse af alder i de tre grupper:

```
model6 = lm(alder~ryger, data=hj)
```

som giver outputtet (kraftigt beskåret):

```
> summary(model6)

Call:
lm(formula = alder ~ ryger, data = hj)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.9870     0.2076 182.967   <2e-16 ***
rygerex       0.5993     0.3323   1.804   0.0723 .
rygernever    0.7258     0.3571   2.033   0.0430 *
---

Residual standard error: 2.448 on 296 degrees of freedom
Multiple R-squared:  0.01804,Adjusted R-squared:  0.01141
F-statistic:  2.72 on 2 and 296 DF,  p-value: 0.06755
```

Nu er det bare ikke en evt signifikant forskel på alderen i de tre grupper, der er afgørende, men derimod den faktiske forskel (som altså var ca 1 år) samt alderens effekt på outcome (`fev1`). Denne effekt ser vi nu på, først for hele materialet under et:

```
model.reg=lm(fev1~alder, data=hj)
summary(model.reg)
5*(coefficients(model.reg)[2])
5*(confint(model.reg)[2,])
```

som giver outputtet:

```
> summary(model.reg)

Call:
lm(formula = fev1 ~ alder, data = hj)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.48404    0.59433   9.227   <2e-16 ***
alder       -0.03387    0.01547  -2.189   0.0294 *
---

Residual standard error: 0.6575 on 297 degrees of freedom
Multiple R-squared:  0.01588,Adjusted R-squared:  0.01256
F-statistic: 4.792 on 1 and 297 DF,  p-value: 0.02937

> 5*(coefficients(model.reg)[2])
     alder
-0.1693314
> 5*(confint(model.reg)[2,])
      2.5 %      97.5 %
-0.32156249 -0.01710034
```

Fra dette output ses `fev1` at falde med 0.03387 pr. år, eller 0.169 over en 5-årig periode (Jeg har vist det for 5 år mere, selv om jeg i opgaven foreslog det lidt nemmere, nemlig for 10 år mere). Konfidensintervallet for dette sidste estimat ses at være CI=(0.017, 0.321).

Nu gør vi det samme for hver af rygergrupperne for sig (da analysen ovenfor jo helt negligerer effekten af rygning):

```
current = hj[hj$ryger=="current",]
ex = hj[hj$ryger=="ex",]
never = hj[hj$ryger=="never",]

reg.current=lm(fev1~alder, data=current)
reg.ex=lm(fev1~alder, data=ex)
reg.never=lm(fev1~alder, data=never)
```

hvorefter vi estimerer faldet i `fev1` over en 5-årig periode for hver af de 3 grupper:

```
> summary(reg.current)

Call:
```

```
lm(formula = fev1 ~ alder, data = current)

Residuals:
     Min       1Q   Median       3Q      Max
-1.57129 -0.36563  0.00685  0.32182  1.37023

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.66461    0.73595   7.697 2.48e-12 ***
alder       -0.04226    0.01933  -2.186   0.0305 *
---

Residual standard error: 0.5825 on 137 degrees of freedom
Multiple R-squared:  0.03371,Adjusted R-squared:  0.02666
F-statistic:  4.78 on 1 and 137 DF,  p-value: 0.03049

> 5*(coefficients(reg.current)[2])
     alder
-0.2113088
> 5*(confint(reg.current)[2,])
     2.5 %     97.5 %
-0.4024288 -0.0201888

------------------------------------------------------------------------
> summary(reg.ex)

Call:
lm(formula = fev1 ~ alder, data = ex)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59685 -0.44439 -0.02817  0.49480  1.16503

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.87612    1.06285   5.529 3.33e-07 ***
alder       -0.03957    0.02749  -1.439    0.154
---

Residual standard error: 0.6443 on 87 degrees of freedom
Multiple R-squared:  0.02326,Adjusted R-squared:  0.01203
F-statistic: 2.072 on 1 and 87 DF,  p-value: 0.1536

> 5*(coefficients(reg.ex)[2])
     alder
-0.1978266
> 5*(confint(reg.ex)[2,])
      2.5 %      97.5 %
-0.47100199  0.07534875

------------------------------------------------------------------------
> summary(reg.never)

Call:
lm(formula = fev1 ~ alder, data = never)

Residuals:
     Min       1Q   Median       3Q      Max
-2.25733 -0.40902 -0.00275  0.53227  1.45473

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  5.69411    1.66436   3.421  0.00105 **
alder       -0.03787    0.04293  -0.882  0.38074
---

Residual standard error: 0.7631 on 69 degrees of freedom
Multiple R-squared:  0.01115,Adjusted R-squared:  -0.003178
F-statistic: 0.7782 on 1 and 69 DF,  p-value: 0.3807

> 5*(coefficients(reg.never)[2])
     alder
-0.1893533
> 5*(confint(reg.never)[2,])
     2.5 %     97.5 %
-0.6175555  0.2388488
```

Disse resultater kan sammenfattes i en tabel som f.eks.

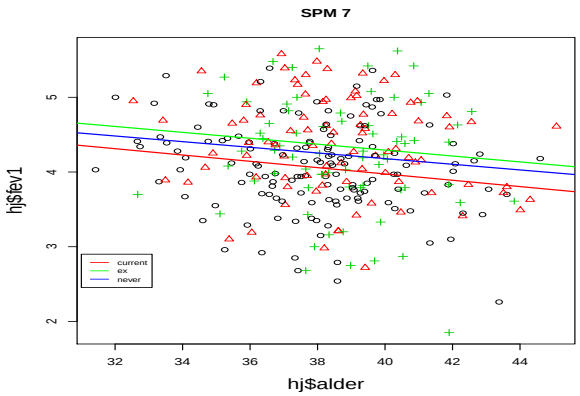
Rygergruppe	Effekt af 5 år	Residualspredning	P
Current	-0.211 (-0.402, -0.020)	0.5825	0.031
Never	-0.189 (-0.618, 0.239)	0.7631	0.15
Ex	-0.198 (-0.471, 0.075)	0.6443	0.38
Alle	-0.169 (-0.321, 0.017)	0.6575	0.030

Disse estimater ser umiddelbart ikke særligt forskellige ud, og det, at kun estimatet for nuværende rygere er signifikant forskellig fra 0, kan ikke tages som udtryk for forskellighed. Godt nok er den estimerede linie stejlest i denne gruppe, men signifikansen beror mest på det store antal i denne gruppe (139 mod 89 hhv 71 i de to andre grupper) og til en vis grad også på den mindre residualspredning i denne gruppe.

En figur med de 3 linier indtegnet laves ved:

```
plot(hj$alder, hj$fev1, main="SPM 7", pch=as.numeric(hj$ryger),
    col=as.numeric(hj$ryger), cex.lab=1.5)
model.current = lm(fev1 ~ alder, data=current)
abline(model.current, col="red", lwd=2)
model.ex = lm(fev1 ~ alder, data=ex)
abline(model.ex, col="green", lwd=2)
model.never = lm(fev1 ~ alder, data=never)
abline(model.never, col="blue", lwd=2)
legend(31, 2.9, legend=c("current", "ex", "never"),
       col=c("red", "green", "blue"), lty=1, cex=0.7)
```

hvorved vi får figuren:



hvor det på øjemål er temmelig svært at se, at linierne ikke er parallelle.

7. *Hvor stor er den estimerede forskel i lungefunktion mellem nuværende rygere og aldrig-rygere, når der justeres for alder?*
Forklar forskellen til svaret i spørgsmål 5, og husk at lave en illustration af den anvendte model.

Vi skal nu undersøge effekten af rygning, korrigeret for den lille aldersforskel, eller helt præcist: forskellen på nuværende rygere og aldrig-rygere, på samme alder. Vi gør dette ved hjælp af en kovariansanalyse, med dertil hørende illustrationer, og vi inddrager *ikke* interaktion i denne model, da vi ovenfor så meget tæt på parallelle linier:

```
model7 = lm(fev1~ryger+alder, data=hj)
```

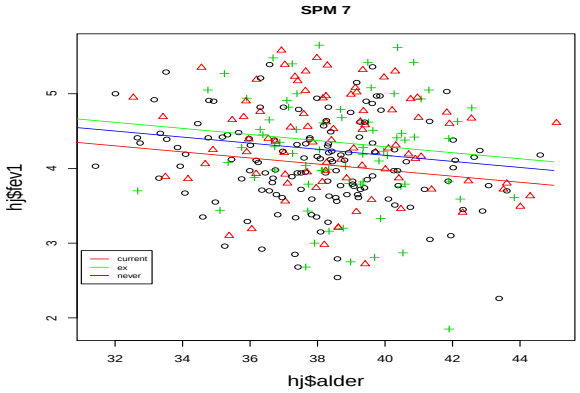
Figuren, der illustrerer modellen kan laves ud fra predikterede værdier

```
ny.current = data.frame(alder=30:45,ryger="current")
pred.current = predict(model7, ny.current)
ny.ex = data.frame(alder=30:45,ryger="ex")
pred.ex = predict(model7, ny.ex)
ny.never = data.frame(alder=30:45,ryger="never")
pred.never = predict(model7, ny.never)

plot(hj$alder, hj$fev1, main="SPM 7", pch=as.numeric(hj$ryger),
     col=as.numeric(hj$ryger), cex.lab=1.5)
```

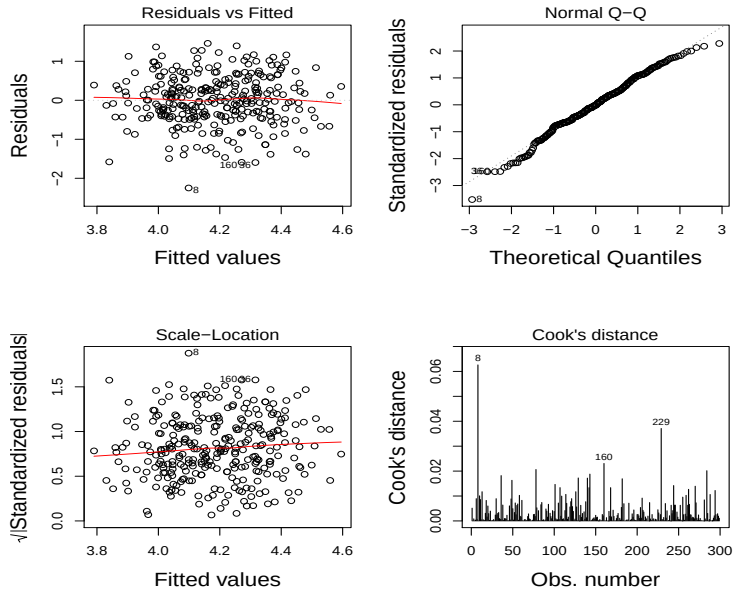
```
lines(ny.current$alder, pred.current, type = "l", col="red")
lines(ny.ex$alder, pred.ex, type = "l", col="green")
lines(ny.never$alder, pred.never, type = "l", col="blue")
legend(31, 2.9, legend=c("current", "ex", "never"),
       col=c("red", "green", "blue"), lty=1, cex=0.7)
```

hvorved vi får:



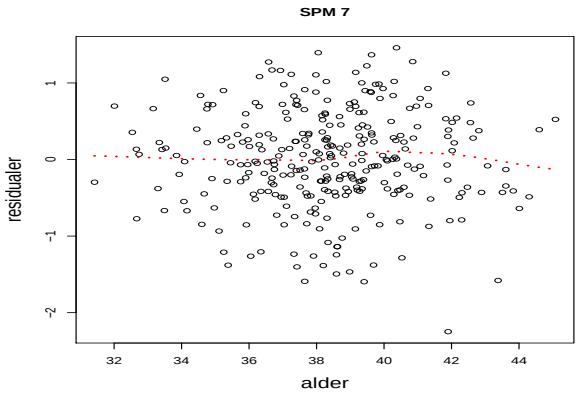
og modelkontrollen udføres ved delvist at benytte de automatiske tegninger:

```
par(mfrow=c(2,2))
plot(model7,which=1, cex.lab=1.5)
plot(model7,which=2, cex.lab=1.5)
plot(model7,which=3, cex.lab=1.5)
plot(model7,which=4, cex.lab=1.5)
```



og for specifikt at checke lineariteten:

```
with(hj, scatter.smooth(alder, resid(model7), main="SPM 7",
                        ylab="residualer", xlab="alder", cex.lab=1.5,
                        lpars = list(col = "red", lwd = 3, lty = 3)))
```



Ingen af disse figurer giver anledning til bekymring, og vi ser derfor på output fra kovariansanalysen:

```
> summary(model7)

Call:
```

```
lm(formula = fev1 ~ ryger + alder, data = hj)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.60315    0.58452   9.586  < 2e-16 ***
rygerex      0.31459    0.08807   3.572 0.000413 ***
rygernever   0.19832    0.09477   2.093 0.037228 *
alder       -0.04064    0.01532  -2.653 0.008410 **
---

Residual standard error: 0.6452 on 295 degrees of freedom
Multiple R-squared:  0.05877,Adjusted R-squared:  0.0492
F-statistic:  6.14 on 3 and 295 DF,  p-value: 0.0004618

> confint(model7)
                  2.5 %      97.5 %
(Intercept)  4.45278253  6.75351556
rygerex      0.14127027  0.48790581
rygernever   0.01181631  0.38482463
alder       -0.07079412 -0.01049376

> 5*(coefficients(model7)[4])
     alder
-0.2032197

> 5*(confint(model7)[4,])
      2.5 %      97.5 %
-0.35397060 -0.05246879
```

Vi bemærker, at der nu ses en kraftigere signifikant forskel på de 3 rygergrupper (P=0.0014 nu mod P=0.0042 i spørgsmål 5), og det skyldes dels, at nuværende rygere var lidt yngre end de andre og dels at alderen har reduceret residualspredningen, så standard errors på parameterestimaterne er blevet mindre.

Det reviderede estimat for forskellen mellem aldrig-rygere og rygere ses at være 0.198, med konfidensinterval CI=(0.012, 0.385), P=0.037.

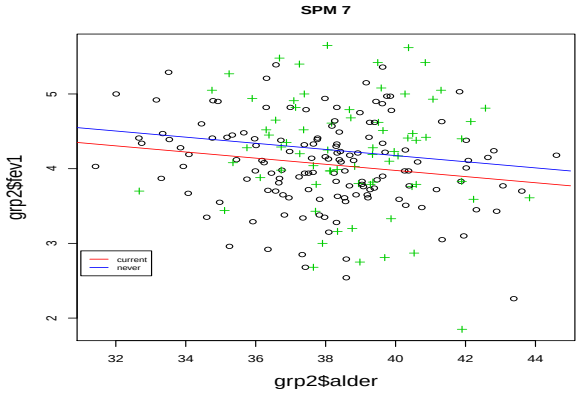
Endvidere får vi et estimat for alderseffekten, der nu er poolet fra alle tre rygergrupper (under antagelse, af at denne effekt er ens i alle grupper). Dette er -0.0406 (-0.0708, -0.0105), svarende til et fald på 0.203 over en 5-årig periode, CI=(0.052, 0.354), P=0.008.

Hvis vi her vælger kun at se på de to berørte grupper, benytter vi et subset af data:

```
grp2 <- subset(hj, ryger=="current" | ryger=="never")

model7ny = lm(fev1~ryger+alder, data=grp2)
```

og finder så figuren



med det tilhørende output

```
> summary(model7ny)

Call:
lm(formula = fev1 ~ ryger + alder, data = grp2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.62154    0.70471   7.977 1.01e-13 ***
rygernever   0.19867    0.09534   2.084   0.0384 *
alder       -0.04113    0.01849  -2.224   0.0272 *
---
Residual standard error: 0.6471 on 207 degrees of freedom
Multiple R-squared:  0.03785,Adjusted R-squared:  0.02856
F-statistic: 4.072 on 2 and 207 DF,  p-value: 0.01843

> confint(model7ny)
                  2.5 %       97.5 %
(Intercept)  4.23221569  7.010865954
rygernever   0.01071015  0.386633639
alder       -0.07759073 -0.004665467

> 5*(coefficients(model7ny)[3])
     alder
-0.2056405

> 5*(confint(model7ny)[3,])
      2.5 %      97.5 %
-0.38795365 -0.02332734
```

Estimatet for forskellen mellem aldrig-rygere og rygere er nu 0.199, med konfidensinterval CI=(0.011, 0.387), altså næsten uforandret.

Estimatet for alderseffekten, der nu kun er poolet fra de to udvalgte rygergrupper er -0.0411, CI=(-0.0776, -0.0046), svarende til et fald på 0.206 over en 5-årig periode, CI=(0.023, 0.388), P=0.027.

Vi sammenfatter forskellen på nuværende rygere og aldrig-rygere fra de forskellige analyser:

current vs. never	Estimeret forskel	P
ANOVA, spm. 5	-0.169 (-0.356, 0.018)	0.08
T-test, spm. 5		
ens varianser	-0.169 (-0.357, 0.019)	0.08
ens varianser	-0.169 (-0.374, 0.036)	0.11
ANCOVA, spm. 7		
3 grupper	-0.198 (-0.385, -0.012)	0.037
2 grupper	-0.199 (-0.387, -0.011)	0.038

Som det også tidligere er bemærket, vurderes effekten af rygning til at være lidt større, når der korrigeres for alder, fordi vi i første omgang sammenlignede nogle lidt yngre rygere med nogle lidt ældre ikke-rygere og derfor fik maskeret lidt af effekten.

Fordi nogle af jer har undersøgt, om der skulle være interaktion mellem alder og rygning (altså *ikke*-parallele linier), indsættes her kode og output for dette, i versionen, hvor vi kun betragter de to udvalgte grupper:

```
model.interaktion = lm(fev1~ryger+alder+ryger:alder, data=grp2)
```

Herved finder vi:

```
> summary(model.interaktion)

Call:
lm(formula = fev1 ~ ryger + alder + ryger:alder, data = grp2)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.664605   0.819517   6.912 5.86e-11 ***
rygernever        0.029503   1.634848   0.018   0.9856
alder            -0.042262   0.021525  -1.963   0.0509 .
rygernever:alder  0.004391   0.042363   0.104   0.9175
---

Residual standard error: 0.6486 on 206 degrees of freedom
Multiple R-squared:  0.0379,Adjusted R-squared:  0.02389
F-statistic: 2.705 on 3 and 206 DF,  p-value: 0.04644
```

og dermed absolut ingen indikation af forskellige alderseffekter (P=0.92).