Besvarelse af opgave om Vital Capacity

I filen `'cadmium.txt'` ligger observationer fra et eksempel omhandlende lungefunktionen hos arbejdere i cadmium industrien (hentet fra P. Armitage & G. Berry: Statistical methods in medical research. 2nd ed. Blackwell, 1987).

Datsættet består af sammenhørende værdier af alder og vital capacity (liter) for 3 grupper af personer, fordelt således:

- Gruppe 1: Eksponeret for cadmium i mere end 10 år ($n = 12$)
- Gruppe 2: Eksponeret for cadmium i mindre end 10 år ($n = 28$)
- Gruppe 3: Ikke eksponeret for cadmium ($n = 44$)

Den første linie i data indeholder variabelnavnene `grp`, `age` og `vitcap`.

1. *Konstruer en faktor (klassevariabel) med beskrivende navne til de 3 grupper, f.eks.* `expo>10`*,* `expo<10` *og* `no-expo`*.*

   Vi indlæser og laver samtidig en ny variabel, kaldet `group`, med de foreslåede navne

   ```
   cad <-
   read.table("http://publicifsv.sund.ku.dk/~lts/basal/data/cadmium.txt",
   header=T)

   cad$group <- factor(cad$grp, levels=1:3,
      labels=c("1:expo>10","2:expo<10","3:no-expo"))
   ```

2. *Beskriv fordelingen af vital capacity i de 3 grupper ved hjælp af passende valgte summary statistics. Lav også passende plots.*

   For at udregne summary statistics, danner vi først separate data frames for de 3 grupper:

   ```
   cad.1 = cad[cad$grp==1,]
   cad.2 = cad[cad$grp==2,]
   cad.3 = cad[cad$grp==3,]
   ```

og derefter kører vi `summary`-funktionen på hver af dem, hvorved output bliver:

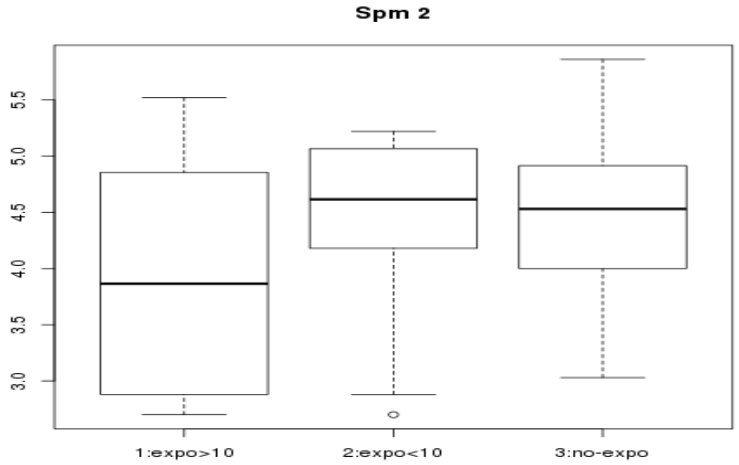
```
> summary(cad.1)
      grp         age            vitcap            group
 Min.   :1   Min.   :39.00   Min.   :2.700   1:expo>10:12
 1st Qu.:1   1st Qu.:41.00   1st Qu.:2.955   2:expo<10: 0
 Median :1   Median :48.00   Median :3.865   3:no-expo: 0
 Mean   :1   Mean   :49.75   Mean   :3.949
 3rd Qu.:1   3rd Qu.:58.25   3rd Qu.:4.737
 Max.   :1   Max.   :65.00   Max.   :5.520
> summary(cad.2)
      grp         age            vitcap            group
 Min.   :2   Min.   :21.00   Min.   :2.700   1:expo>10: 0
 1st Qu.:2   1st Qu.:29.75   1st Qu.:4.240   2:expo<10:28
 Median :2   Median :38.00   Median :4.615   3:no-expo: 0
 Mean   :2   Mean   :37.79   Mean   :4.472
 3rd Qu.:2   3rd Qu.:43.50   3rd Qu.:5.062
 Max.   :2   Max.   :58.00   Max.   :5.220
> summary(cad.3)
      grp         age           vitcap           group
 Min.   :3   Min.   :18.0   Min.   :3.030   1:expo>10: 0
 1st Qu.:3   1st Qu.:32.0   1st Qu.:4.010   2:expo<10: 0
 Median :3   Median :41.0   Median :4.530   3:no-expo:44
 Mean   :3   Mean   :39.8   Mean   :4.462
 3rd Qu.:3   3rd Qu.:48.0   3rd Qu.:4.902
 Max.   :3   Max.   :65.0   Max.   :5.860
```

Det er dog noget lettere at få et overblik ved hjælp af en grafisk illustration, som her kunne være boxplots, som vi får ved at skrive:

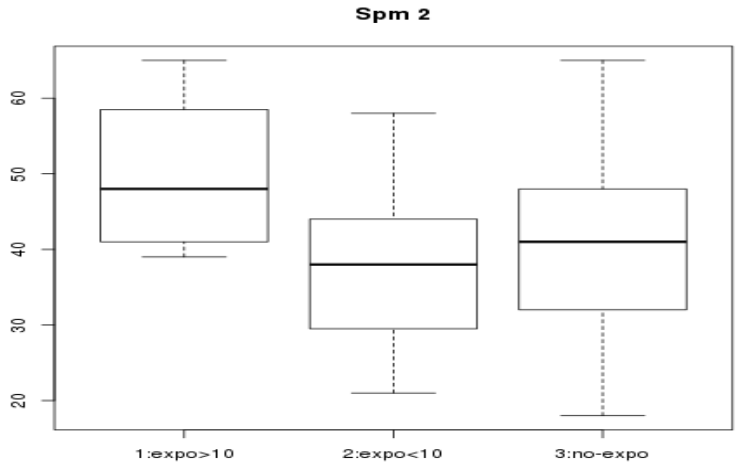
```
boxplot(vitcap~group, data=cad, main="Spm 2")
```

```
boxplot(age~group, data=cad, main="Spm 2")
```

hvorved vi får, først for **vitalkapaciteten:**



og så for **alderen:**



Ser der umiddelbart ud til at være forskel på grupperne?

Baseret på såvel gennemsnit som boxplots, ser det ud til, at de langtidseksponerede (`cad.1`) har en lavere vitalkapacitet end de to øvrige grupper. Men de er også ældre, og en del af årsagen til den lavere vitalkapacitet hos de langtidseksponerede kunne være deres højere alder.

Vi må altså forvente **confounding** mellem `group` og `age`, idet aldersfordelingerne ikke er ens, og der samtidig forventes en nedgang i vitalkapacitet med alderen.

3. *Ignorer i første omgang* `age`*-variablen.*
   Er der forskel på vital capacity i de 3 grupper?

   Baseret på Boxplottene af vitalkapaciteten ovenfor, tør vi godt kaste os ud i en ensidet variansanalyse til at sammenligne grupperne, idet fordelingerne ser rimeligt symmetriske ud:

```
model1 = lm(vitcap ~ factor(group), data=cad)
```

   Herefter kan vi benytte `summary` og `confint`-funktioner til at producere outputtet:

```
> summary(model1)

Call:
lm(formula = vitcap ~ factor(group), data = cad)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77179 -0.45205  0.09808  0.51295  1.57083

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.9492     0.2149  18.376   <2e-16 ***
factor(group)2:expo<10  0.5226     0.2569   2.035   0.0452 *
factor(group)3:no-expo  0.5129     0.2425   2.115   0.0375 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7445 on 81 degrees of freedom
Multiple R-squared:  0.05767,Adjusted R-squared:  0.0344
F-statistic: 2.478 on 2 and 81 DF,  p-value: 0.09021


> confint(model1)
                            2.5 %    97.5 %
```

```
(Intercept)            3.52156067 4.3767727
factor(group)2:expo<10 0.01153213 1.0337060
factor(group)3:no-expo 0.03047416 0.9952834
```

Det fremgår at der ikke er signifikant forskel på grupperne (på 5% signifikansniveau) i denne analyse, idet F-testets P-værdi er på 0.09.

Giv et estimat for forskellen i vital capacity på de langtidseksponerede og de ueksponerede, naturligvis med konfidensgrænser.

Da gruppen af langtidseksponerede (`1:expo>10`) er først i alfabetet, er denne gjort til referencegruppe, og den søgte forskel står derfor direkte at aflæse under `3:no-expo` som 0.5129 (0.0305, 0.9953) liter, dvs. at de langtidseksponerede har godt en halv liters lavere vitalkapacitet end de ueksponerede, med et konfidensinterval, der går fra ca. 0 til ca. 1 liter, altså ganske upræcist, men dog signifikant med P=0.038.

Nu er der spurgt specifikt om denne forskel, men ellers skulle vi have korrigeret for massesignifikans og i stedet set på Tukey-korrigerede T-tests med tilhørende konfidensintervaller. For at gøre dette, er vi nødt til at foretage variansanalysen med `aov` i stedet for `lm`:

```
model2 = aov(vitcap ~ factor(group), data=cad)
TukeyHSD(model2)
```

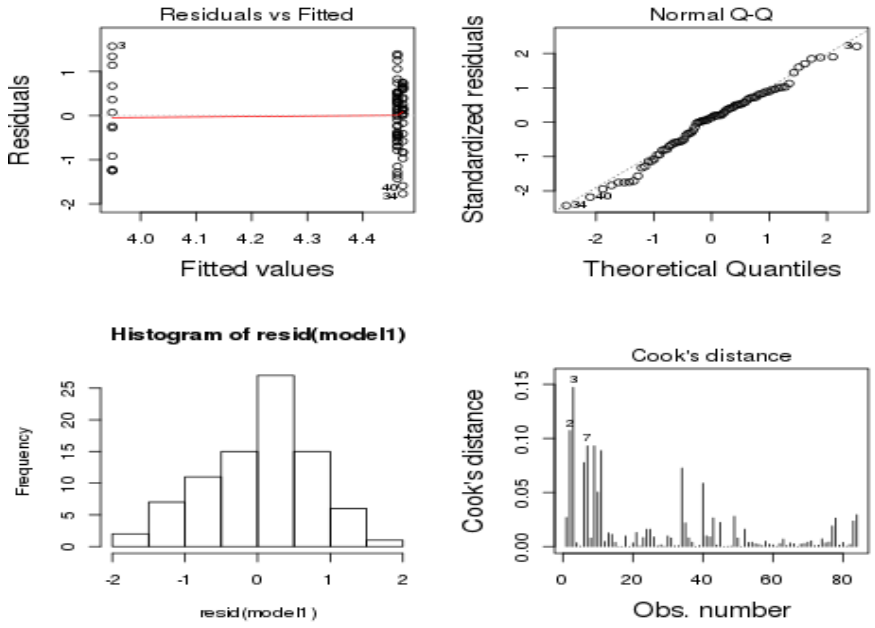
og de Tukey-korrigerede sammenligninger falder således ud:

```
> TukeyHSD(model2)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = vitcap ~ group, data = cad)

$group
                           diff         lwr       upr     p adj
2:expo<10-1:expo>10  0.52261905 -0.09066563 1.1359037 0.1104862
3:no-expo-1:expo>10  0.51287879 -0.06598823 1.0917458 0.0930317
3:no-expo-2:expo<10 -0.00974026 -0.43943758 0.4199571 0.9983865
```

Alle konfidensintervallerne for de parvise differenser ses nu at indeholde 0, i overensstemmelse med, at der ikke overordnet set er signifikant forskel på de tre grupper (men denne overensstemmelse kan man ikke være sikker på).

Den søgte forskel giver naturligvis samme estimat som før, nemlig -0.513, men nu med konfidensintervallet (-0.066, 1.092), altså noget bredere end før, fordi vi nu har korrigeret for 3 sammenligninger.

Modelkontrollen giver figurerne



og vi ser en virkelig pæn normalfordeling i residualerne, men muligvis en form for omvendt trompet i figuren af residualer mod forventede værdier. Dette skyldes, at de langtids-eksponerede har en noget større spredning end de to andre grupper, samtidig med, at de har et *lavere* niveau.

Vi kan også kontrollere varianshomogeniteten ved hjælp af Bartlettes test:

```
> bartlett.test(cad$vitcap,cad$group)

Bartlett test of homogeneity of variances

data:  cad$vitcap and cad$group
Bartlett's K-squared = 3.7164, df = 2, p-value = 0.156
```

som ikke giver nævneværdig mistanke om problemer med varianshomogeniteten.

Hvis vi havde haft problemer med normalfordelingsantagelsen, kunne vi have udført en non-parametrisk sammenligning ved at skrive:

```
kruskal.test(vitcap ~ factor(group), data=cad)
```

hvorved man får outputtet:

```
> kruskal.test(vitcap ~ factor(group), data=cad)

Kruskal-Wallis rank sum test

data:  vitcap by factor(group)
Kruskal-Wallis chi-squared = 2.7909, df = 2, p-value = 0.2477
```

Heller ikke her finder vi altså nogen signifikans.

4. *Udregn korrelationen mellem alder og vital capacity for hver gruppe for sig, samt for datamaterialet som helhed.*
   Hvad kan vi slutte af dette?

   Vi tager det lige i omvendt rækkefølge. Først udregner vi korrelationen for hele populationen samlet, og vi benytter Pearson-korrelation for at kunne sammenligne til det næste spørgsmål:

   ```
   with(cad,cor.test(age,vitcap))
   ```

   der (bl.a.) giver følgende output:

```
> with(cad,cor.test(age,vitcap))

Pearson's product-moment correlation

data:  age and vitcap
t = -6.8827, df = 82, p-value = 1.082e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7254035 -0.4489697
sample estimates:
      cor
-0.605118
```

Vi ser, at der er en negativ korrelation mellem alder og vitalkapacitet (-0.605), og at denne er stærkt signifikant forskellig fra 0 ($P < 0.0001$)

Nu gør vi så det tilsvarende, bare opdelt på grupper

```
with(cad.1,cor.test(age,vitcap))
with(cad.2,cor.test(age,vitcap))
with(cad.3,cor.test(age,vitcap))
```

hvilket giver outputtet:

```
> with(cad.1,cor.test(age,vitcap))

Pearson's product-moment correlation

data:  age and vitcap
t = -3.5887, df = 10, p-value = 0.00494
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9256202 -0.3097485
sample estimates:
      cor
-0.750277

> with(cad.2,cor.test(age,vitcap))
```

```
Pearson's product-moment correlation

data:  age and vitcap
t = -4.1107, df = 26, p-value = 0.0003501
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8108389 -0.3323676
sample estimates:
       cor
-0.6276204

> with(cad.3,cor.test(age,vitcap))

Pearson's product-moment correlation

data:  age and vitcap
t = -4.0598, df = 42, p-value = 0.0002095
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7150585 -0.2777749
sample estimates:
       cor
-0.5308761
```

Det kan noteres, at korrelationen er mindst i gruppen af ueksponerede og størst i gruppen af langtids-eksponerede. Imidlertid er det svar på et forkert spørgsmål, idet det er mere naturligt at ville vide om regressionslinjen er stejlere i nogle grupper end i andre, fordi hældningen angiver det konkrete fald i vitalkapacitet for hvert år, man bliver ældre.

5. *Foretag for hver af grupperne en lineær regressionsanalyse af vital capacity mod alder. Hvor stærk er sammenhængen i de tre grupper?*

Regressioner for hver gruppe:

```
age.1 = lm(vitcap ~ age, data=cad.1)
age.2 = lm(vitcap ~ age, data=cad.2)
age.3 = lm(vitcap ~ age, data=cad.3)
```

giver outputtet

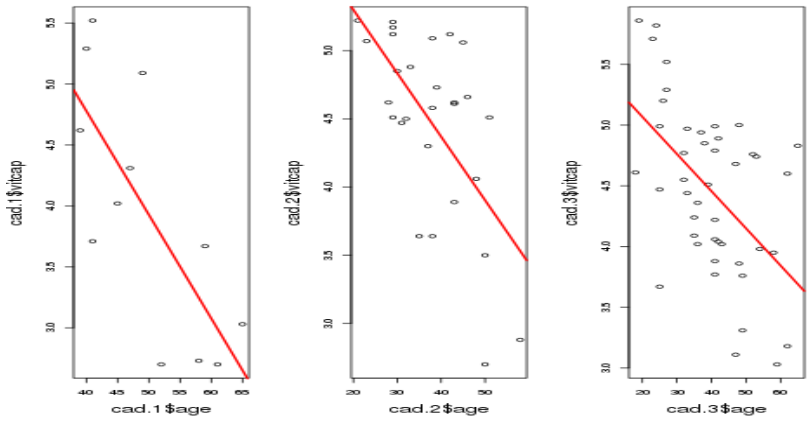
```
> summary(age.1)

Call:
lm(formula = vitcap ~ age, data = cad.1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05767 -0.37933 -0.05867  0.50883  1.07700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.18344    1.19787   6.832 4.56e-05 ***
age         -0.08511    0.02372  -3.589  0.00494 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7163 on 10 degrees of freedom
Multiple R-squared:  0.5629,Adjusted R-squared:  0.5192
F-statistic: 12.88 on 1 and 10 DF,  p-value: 0.00494

> summary(age.2)

Call:
lm(formula = vitcap ~ age, data = cad.2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.20343 -0.32294  0.03972  0.34226  0.92391

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.23003    0.43977  14.167 9.74e-14 ***
age         -0.04653    0.01132  -4.111  0.00035 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5408 on 26 degrees of freedom
Multiple R-squared:  0.3939,Adjusted R-squared:  0.3706
F-statistic:  16.9 on 1 and 26 DF,  p-value: 0.0003501

> summary(age.3)

Call:
lm(formula = vitcap ~ age, data = cad.3)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2450 -0.4264 -0.0118  0.4527  1.1395

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.68029    0.31313   18.14  < 2e-16 ***
age         -0.03061    0.00754   -4.06 0.000209 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5936 on 42 degrees of freedom
Multiple R-squared:  0.2818,Adjusted R-squared:  0.2647
F-statistic: 16.48 on 1 and 42 DF,  p-value: 0.0002095
```

Vi noterer os regressionskoefficienterne med tilhørende standard errors: henholdsvis -0.085(0.024), -0.047(0.011), og -0.031(0.008). Det kunne godt tyde på at de ikke er helt ens, men at niveauet falder hurtigere i den langtidseksponerede gruppe.

De tilsvarende plots af data med regressionslinier fås ved at skrive

```
par(mfrow=c(1,3))
plot(cad.1$age, cad.1$vitcap, cex.lab=1.5)
abline(age.1, col="red", lwd=2)
plot(cad.2$age, cad.2$vitcap, cex.lab=1.5)
abline(age.2, col="red", lwd=2)
plot(cad.3$age, cad.3$vitcap, cex.lab=1.5)
abline(age.3, col="red", lwd=2)
```

hvorved vi får figurerne

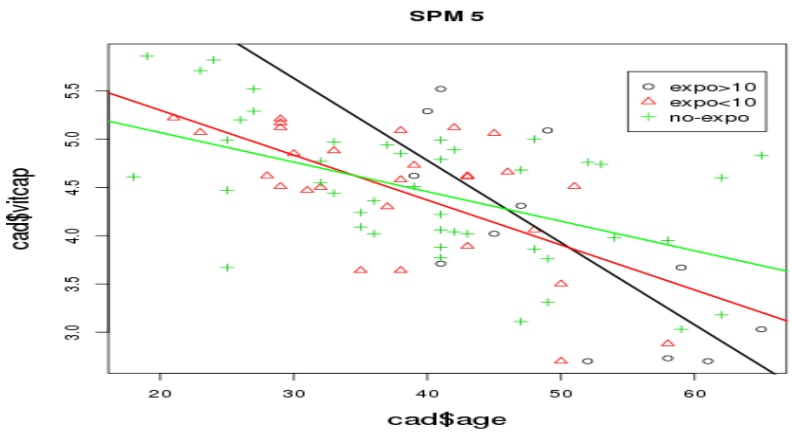


Da figurerne har forskellige Y-akser, kan vi dog ikke umiddelbart sammenligne hældningerne for de 3 grupper. Dette kunne løses ved selv at vælge akserne, men i stedet vælger vi at afbilde alle linier på samme plot ved at skrive:

```
par(mfrow=c(1,1))
```

```
plot(cad$age, cad$vitcap, main="SPM 6", pch=as.numeric(cad$grp),
              col=as.numeric(cad$grp), cex.lab=1.5)
legend(55, 5.7, legend=c("expo>10", "expo<10", "no-expo"), pch=1:3,
       col=c("black", "red", "green"), cex=1.1)

abline(age.1, col="black", lwd=2)
abline(age.2, col="red", lwd=2)
abline(age.3, col="green", lwd=2)
```

hvorved vi får figuren:



Her ses tydeligt, at de langtidseksponerede (sorte punkter og linie) har den stejleste hældning, altså den hurtigst aftagende lungekapacitet.

6. *Giv et estimat for forskellen i vitalkapacitet mellem 40-årige langtidseksponerede og 40-årige ueksponerede.*

   Her kunne man give sig til at benytte de ovenstående regressionsanalyser og udregne estimeret vitalkapacitet for 40-årige i hver af de 3 grupper, men vi ville herved komme til at mangle konfidensgrænser for forskellene.

I stedet må vi bygge de 3 regressioner fra forrige spørgsmål sammen til en stor model, og da der umiddelbart så ud til at være forskel på hældningerne, er vi nødt til at medtage interaktionsleddet `group*age`.

For at få estimeret forskellen for 40-årige, vælger vi at flytte interceptet hen i alder 40 ved at trække 40 fra alderen, og derefter køre modellen med interaktion. At benytte `age40` (som defineret nedenfor) i stedet for `age` har ingen indflydelse på andet end afskæringerne:

```
cad$age40=cad$age-40

int40 = lm(vitcap ~ age40+factor(group)+age40*factor(group), data=cad)
```

Herefter benytter vi `summary`, `confint` og `anova` til at få det relevante output:

```
> summary(int40)

Call:
lm(formula = vitcap ~ age40 + factor(group) + age40 * factor(group),
    data = cad)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24497 -0.36929  0.01977  0.43681  1.13953

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   4.77900    0.25730  18.573  < 2e-16 ***
age40                        -0.08511    0.01967  -4.327 4.44e-05 ***
factor(group)2:expo<10       -0.41025    0.28208  -1.454   0.1499
factor(group)3:no-expo       -0.32322    0.27245  -1.186   0.2391
age40:factor(group)2:expo<10  0.03858    0.02327   1.658   0.1014
age40:factor(group)3:no-expo  0.05450    0.02107   2.587   0.0116 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5942 on 78 degrees of freedom
Multiple R-squared:  0.422,Adjusted R-squared:  0.385
F-statistic: 11.39 on 5 and 78 DF,  p-value: 2.871e-08


> confint(int40)
                                    2.5 %      97.5 %
(Intercept)                   4.266749088  5.29124854
age40                        -0.124274163 -0.04594782
factor(group)2:expo<10       -0.971831138  0.15133461
factor(group)3:no-expo       -0.865624925  0.21919484
age40:factor(group)2:expo<10 -0.007753455  0.08491141
age40:factor(group)3:no-expo  0.012551569  0.09644507
```

```
> anova(int40)
Analysis of Variance Table

Response: vitcap
                    Df  Sum Sq Mean Sq F value    Pr(>F)
age40                1 17.4446 17.4446 49.4159 6.918e-10 ***
factor(group)        2  0.1617  0.0808  0.2290   0.79584
age40:factor(group)  2  2.4995  1.2497  3.5402   0.03376 *
Residuals           78 27.5352  0.3530
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi bemærker fra den nederste del af outputtet, at der rent faktisk er signifikant forskel på hældningerne i de 3 grupper (P=0.034).

Da interceptet nu ligger ved 40 år, kan vi direkte aflæse forskellen på de to søgte grupper ved 40-års alderen under `factor(group)3:no-expo`, og vi ser, at det bliver 0.323, med konfidensgrænser (-0.219, 0.866) liter, **i de langtids-eksponeredes favør**! Godt nok ikke signifikant (P=0.24), men alligevel....

Hvordan kan man anskueliggøre modellen til grund for dette estimat?

Den model, vi har fittet til data, består af 3 ikke-parallelle linier, og vi har allerede ovenfor set den relevante illustration af denne model. Her så vi, at linierne svarende til de 3 grupper krydser ind over hinanden, således, at den sorte linie hørende til de langtids-eksponerede netop ligger højest for de lave aldre.

Hvis vi ser tilbage på summary statistics fra spørgsmål 2, kan vi se, at den yngste langtids-eksponerede er 39 år, så det er lidt vovet at udtale sig om ovenstående forskel.

Bemærk i øvrigt det noget misvisende i, at R tegner de fittede linier helt ud til kanten af plottet, selv om der ikke er observationer til at understøtte dem i hele området.

- *Sammenlign med det i spørgsmål 3 fundne estimat.*
  Vi fandt estimatet for forskellen på 40-årige langtids-eksponerede vs. ueksponerede til at være 0.323, med konfidensgrænser (-0.219,

0.866) liter, mens vi i spørgsmål 3 fandt estimatet -0.513 (-0.995, -0.030) liter, altså to meget forskellige estimater.

Hvad er der sket?
I spørgsmål 3 ignorerede vi alderen, så der sammenlignede vi ældre langtids-eksponerede med yngre ueksponerede, og resultatet må derfor have været en blanding af effekten af eksposition **og** alder. I dette spørgsmål tager vi hensyn til denne aldersforskel, og **desuden** tillader vi effekten af alder at være forskellig for de tre grupper, som det ses på figuren ovenfor. Forskellen på grupperne vil derfor afhænge af, hvilken alder, vi betragter, og for 40 år er vi lige på kanten af, hvad der overhovedet forekommer i gruppen af langtids-eksponerede.

Det er derfor ikke et helt rimeligt spørgsmål at stille overhovedet.....

- *Kan sammenhængen mellem alder og vital capacity påvises at være forskellig for de tre grupper?*

  Sammenhængen mellem alder og vital capacity er udtrykt ved hældningen, så ovenstående spørgsmål går på, om de 3 hældninger kan påvises at være forskellige, altså om der er signifikant **interaktion** eller **vekselvirkning**.

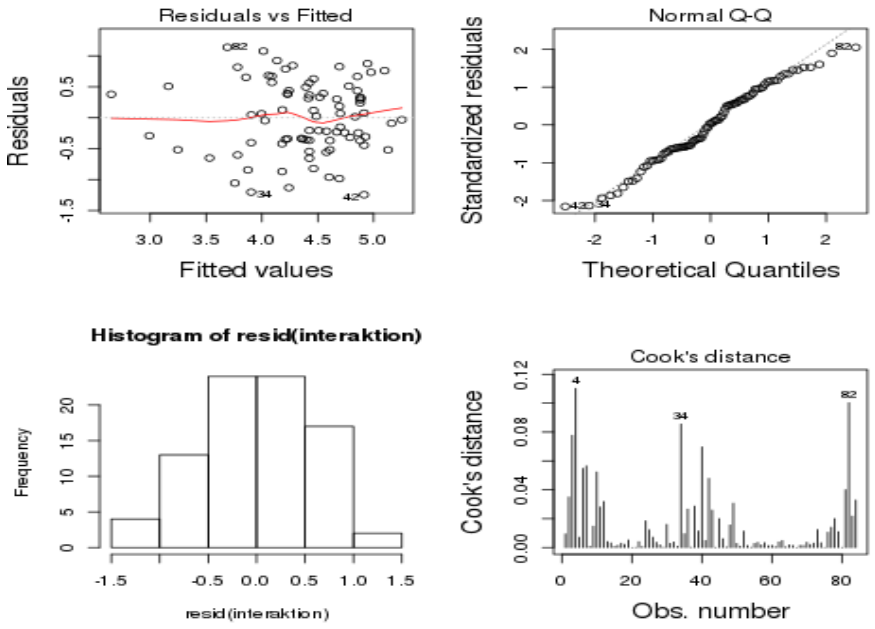
  I det ovenstående output ses, at denne interaktion faktisk *er* signifikant med $P = 0.0338$. Det vil sige at regressionskoefficienterne *ikke* kan antages at være ens så linjerne er ikke parallelle. De estimerede parametre giver f.eks. for `age40:factor(group)2:expo<10` *forskellen* på hældningerne i den korttids- og den langtids-eksponerede gruppe. Det ses, at den signifikante forskel først og fremmest skyldes at gruppen af langtids-eksponerede udviser et hurtigere fald i vitalkapaciteten end de andre to. Det kunne forstås derhen, at cadmium eksponering i længere tid accelererer den aldersbetingede reduktion i vitalkapacitet, snarere end at sænke niveauet med en konstant værdi, faktisk en ret intuitiv forklaring.

  Men bemærk:
  I lyset af den markant ældre population blandt de langtidseksponerede *kunne* et sådant resultat også skyldes, at alderseffekten ikke er lineær, idet faldet i vitalkapacitet evt. accelererede med

alderen.

Hvis man inddrager et andengradsled i alder, er der dog ingensomhelst tegn på, at dette giver en forbedret model, så effekten ser virkelig ud til at kunne forklares udfra cadmium ekspositionen.

Og så burde vi jo også lige se på noget modelkontrol, f.eks. de automatiske tegninger:



som faktisk ser ganske tilforladelige ud.

7. *Hvor mange flere år skal der til at tabe 1 liter i vitalkapacitet, når man er ueksponeret i forhold til, hvis man var langtidseksponeret?*

   Her skal vi omregne hældningerne fra spørgsmål 5, udregnet for hver gruppe for sig. Hældningerne angiver, hvor mange liter, man mister pr. år, så for at udregne hvor mange år, der skal gå før man har mistet 1 liter, skal vi bare invertere dem, hvorved vi finder:

For ikke-eksponerede: $\frac{1}{0.03061} = 32.7$,
med 95% konfidensinterval $(\frac{1}{0.04583}, \frac{1}{0.01540}) = (21.8, 64.9)$

For korttids-eksponerede: $\frac{1}{0.04653} = 21.5$,
med 95% konfidensinterval $(\frac{1}{0.06980}, \frac{1}{0.02326}) = (14.3, 43.0)$

For langtids-eksponerede: $\frac{1}{0.08511} = 11.7$,
med 95% konfidensinterval $(\frac{1}{0.13795}, \frac{1}{0.03227}) = (7.2, 31.0)$

Der går altså forventeligt 32.7-11.7=21 år mere, før en ueksponeret mister 1 liter, i forhold til en langtidseksponeret. Det er desværre ikke helt simpelt at finde et konfidensinterval for denne forskel, men vi kan gætte på, at det bliver ganske bredt.

Besvarelse af væksthormon-opgaven

Filen `juul2.txt` indeholder en variant af Anders Juuls datamateriale om IGF-I (Insulin-like Growth Factor) hos normale mennesker.

Filen indeholder (i den nævnte rækkefølge):

- `age`: – alder i år
- `height`: – højde i cm
- `sex`: – Køn (1/2=M/F)
- `igf1`: – Serum IGF-I
- `tanner`: – Tanner's pubertetsklassifikation (1–5)
- `weight`: – vægt (kg)

1. *Lav en lineær regressionsanalyse for præpubertale individer (Tanner stadium 1), for hvert køn for sig, og med logaritmetransformeret* `igf1`

som outcome, og alderen som forklarende variabel.

Nedenfor indlæser vi, idet vi samtidig rekoder kønnet til mere sigende betegnelser og laver en logaritmisk transformation, som specificeret i opgaveteksten:

```
juul <-
read.table("http://publicifsv.sund.ku.dk/~lts/basal/data/juul2.txt",
header=T,na.strings=c("."))

juul$gender <- factor(juul$sex, levels=1:2, labels=c("MALE", "FEMALE"))

juul$lnigf1=log(juul$igf1)
```

Vi bruger her den *naturlige* logaritme, ikke fordi den er specielt naturlig, men for (endnu en gang) at påpege, at det er ligegyldigt, hvilken logaritme, der anvendes, når blot man husker hvilken. (Husk dog, at der kan være visse fortolkningsmæssige genveje ved at benytte forståelige logaritmer, når det er kovariaterne, der skal transformeres).

Vi starter med at definere de to data frames, vi skal benytte i regressionsanalyserne:

```
tanner.1 = juul[juul$tanner==1,]
tanner.1F = tanner.1[tanner.1$gender=="FEMALE",]
tanner.1M = tanner.1[tanner.1$gender=="MALE",]
```

hvorefter vi udfører analyserne med

```
model.F = lm(lnigf1 ~ age, data=tanner.1F)

model.M = lm(lnigf1 ~ age, data=tanner.1M)
```

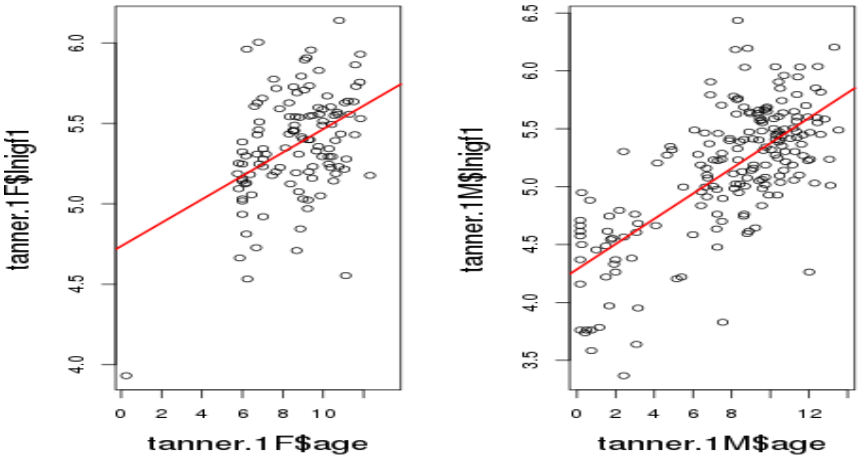
Herefter får vi outputtet ved at benytte `summary` og `confint`-funktionerne:

```
> summary(model.F)

Call:
lm(formula = lnigf1 ~ age, data = tanner.1F)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99166 -0.17079  0.02702  0.15469  0.77249

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.73692    0.12560  37.714  < 2e-16 ***
age          0.07272    0.01416   5.135 1.14e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3027 on 117 degrees of freedom
  (346 observations deleted due to missingness)
Multiple R-squared:  0.1839,Adjusted R-squared:  0.1769
F-statistic: 26.37 on 1 and 117 DF,  p-value: 1.138e-06


> confint(model.F)
                 2.5 %    97.5 %
(Intercept) 4.48817411 4.9856665
age         0.04467135 0.1007635



> summary(model.M)

Call:
lm(formula = lnigf1 ~ age, data = tanner.1M)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33211 -0.23283  0.00883  0.26994  1.24455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.286051   0.070465   60.83   <2e-16 ***
age         0.108971   0.008171   13.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4085 on 190 degrees of freedom
  (340 observations deleted due to missingness)
Multiple R-squared:  0.4835,Adjusted R-squared:  0.4808
F-statistic: 177.9 on 1 and 190 DF,  p-value: < 2.2e-16


> confint(model.M)
                 2.5 %    97.5 %
(Intercept) 4.14705647 4.4250447
age         0.09285364 0.1250883
```

Vi ser, at vi for pigerne får regressionslinjen $\ln(\texttt{igf1}) = 4.737 + 0.0727 \times \texttt{alder}$ og for drengene $\ln(\texttt{igf1}) = 4.286 + 0.1090 \times \texttt{alder}$, svarende til at serum IGF-1 stiger 7.5% (beregnes som faktoren $\exp(0.0727) = 1.075$) pr. år for pigerne og 11.5% pr år for drengene. Bemærk, at fordi vi har brugt den naturlige logaritme, vil små tal ($< \pm 0.1$) tilbagetransformerer til en relativ forskel af ca. samme størrelse.

De tilhørende plots af regressionsmodellerne bliver:

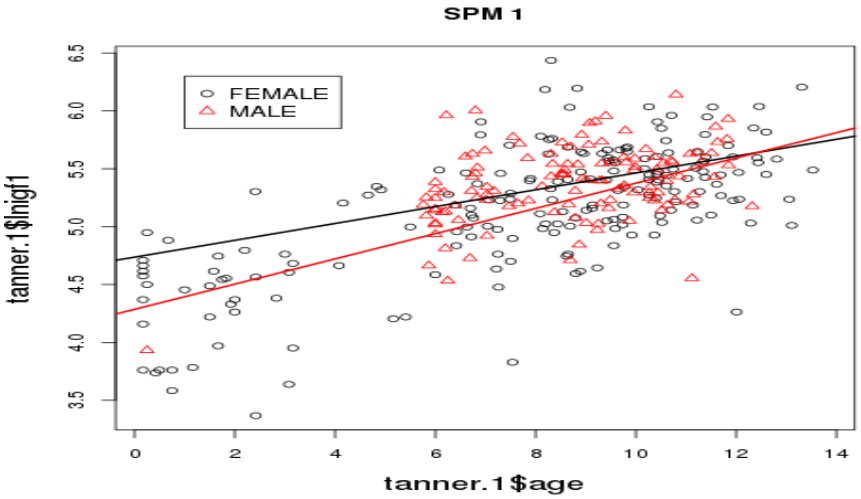
```
par(mfrow=c(1,2))
plot(tanner.1F$age, tanner.1F$lnigf1, cex.lab=1.5)
abline(model.F, col="red", lwd=2)
plot(tanner.1M$age, tanner.1M$lnigf1, cex.lab=1.5)
abline(model.M, col="red", lwd=2)
```



men da akserne her ikke er ens, vil vi supplere med et fælles plot, med begge linier på. Dette fås ved at skrive:

```
par(mfrow=c(1,1))
plot(tanner.1$age, tanner.1$lnigf1, main="SPM 2", pch=as.numeric(tanner.1$gender),
legend(1, 6.3, legend=c("FEMALE", "MALE"), pch=1:2,
```

```
       col=c("black", "red"), cex=1.1)
abline(model.F, col="black", lwd=2)
abline(model.M, col="red", lwd=2)
```



Giv en forståelig fortolkning af hældningen i form af den procentuelle øgning af `igf1` *på 5 år.*

Til dette formål skal vi bare gange hældningen op med 5 og tilbagetransformere med exponentialfunktionen. Vi finder

For piger: $\exp(5 \times 0.07272) = \exp(0.3636) = 1.44$,
med konfidensinterval $(\exp(0.2234), \exp(0.5038)) = (1.25, 1.65)$

For drenge: $\exp(5 \times 0.10897) = \exp(0.5449) = 1.72$,
med konfidensinterval $(\exp(0.4643), \exp(0.6254)) = (1.59, 1.87)$

På 5 år vil `igf1` således øges med 44% hos piger (CI fra 25-65%) og med 72% hos drenge (CI fra 59-87%).

2. *Undersøg om regressionslinjerne er ens for de to køn, og om der samlet set er en effekt af alder.*

Vi laver en samlet analyse i form af en generel lineær model, og her er det specielt interaktionsleddet, der er interessant.

```
interaktion = lm(lnigf1 ~ age+factor(gender)+age*factor(gender),
               data=tanner.1)
```

Med `summary` og `confint`-funktionerne, får vi så outputtet

```
> summary(interaktion)

Call:
lm(formula = lnigf1 ~ age + factor(gender) + age * factor(gender),
    data = tanner.1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33211 -0.21035  0.01495  0.21837  1.24455

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 4.286051   0.064125  66.839  < 2e-16 ***
age                         0.108971   0.007436  14.655  < 2e-16 ***
factor(gender)FEMALE        0.450870   0.167051   2.699  0.00734 **
age:factor(gender)FEMALE -0.036254   0.018915  -1.917  0.05621 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3717 on 307 degrees of freedom
  (445 observations deleted due to missingness)
Multiple R-squared:  0.4581,Adjusted R-squared:  0.4528
F-statistic: 86.49 on 3 and 307 DF,  p-value: < 2.2e-16


> confint(interaktion)
                                2.5 %       97.5 %
(Intercept)                 4.15987056 4.4122305912
age                         0.09433952 0.1236024168
factor(gender)FEMALE        0.12216069 0.7795788250
age:factor(gender)FEMALE -0.07347263 0.0009655472
```

Bemærk at interaktionsleddet er meget tæt på signifikans (P=0.056). Det vil sige, at vi ikke med overbevisning kan afvise, at linjerne er parallelle (har samme hældning), men at der er en vis indikation af forskel på hældningerne.

Giv også et estimat for forskellen på drenge og piger i 7-års alderen.

Her benytter vi tricket med at flytte Y-aksen hen i alder 7 år ved at fratrække 7 fra alderen:

```
tanner.1$age7=tanner.1$age-7
```

Herefter kører vi en model med interaktion, fordi vi ikke føler os sikre på, at denne ikke er til stede:

```
int7 = lm(lnigf1 ~ age7+factor(gender)+age7*factor(gender), data=tanner.1)
```

Med `summary` og `confint`-funktionerne, får vi så outputtet

```
> summary(int7)

Call:
lm(formula = lnigf1 ~ age7 + factor(gender) + age7 * factor(gender),
    data = tanner.1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33211 -0.21035  0.01495  0.21837  1.24455

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.048847   0.027532 183.379  < 2e-16 ***
age7                       0.108971   0.007436  14.655  < 2e-16 ***
factor(gender)FEMALE       0.197095   0.052371   3.763 0.000201 ***
age7:factor(gender)FEMALE -0.036254   0.018915  -1.917 0.056208 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3717 on 307 degrees of freedom
  (445 observations deleted due to missingness)
Multiple R-squared:  0.4581,Adjusted R-squared:  0.4528
F-statistic: 86.49 on 3 and 307 DF,  p-value: < 2.2e-16


> confint(int7)
                                2.5 %       97.5 %
(Intercept)                4.99467159 5.1030231294
age7                       0.09433952 0.1236024168
factor(gender)FEMALE       0.09404297 0.3001469299
age7:factor(gender)FEMALE -0.07347263 0.0009655472
```

Vores bedste gæt på forskellen i ln(`igf1`) på drenge og piger i 7-års alderen er altså, at drenge ligger lidt lavere end piger, og vi kvantificerer forskellen ved at tilbagetransformere

$\exp(-0.197) = 0.82$, med CI=$(\exp(-0.300), \exp(-0.094)) = (0.74, 0.91)$, altså at drenge ligger 18% lavere end piger, med konfidensinterval fra 9-26% lavere.

Hvis vi *antager*, at alderseffekten er den samme for de to køn, altså at linierne er parallelle, skal vi se på en model uden interaktionsled:

```
ancova = lm(lnigf1 ~ age+factor(gender), data=tanner.1)
```

og vi får så i stedet

```
> summary(ancova)

Call:
lm(formula = lnigf1 ~ age + factor(gender), data = tanner.1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.30871 -0.22295  0.01723  0.23137  1.24722

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.329936   0.060157   71.98  < 2e-16 ***
age                  0.103368   0.006867   15.05  < 2e-16 ***
factor(gender)FEMALE 0.141852   0.043916    3.23  0.00137 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3733 on 308 degrees of freedom
  (445 observations deleted due to missingness)
Multiple R-squared:  0.4516,Adjusted R-squared:  0.448
F-statistic: 126.8 on 2 and 308 DF,  p-value: < 2.2e-16


> confint(ancova)
                         2.5 %    97.5 %
(Intercept)          4.21156475 4.4483072
age                  0.08985678 0.1168799
factor(gender)FEMALE 0.05543845 0.2282647
```

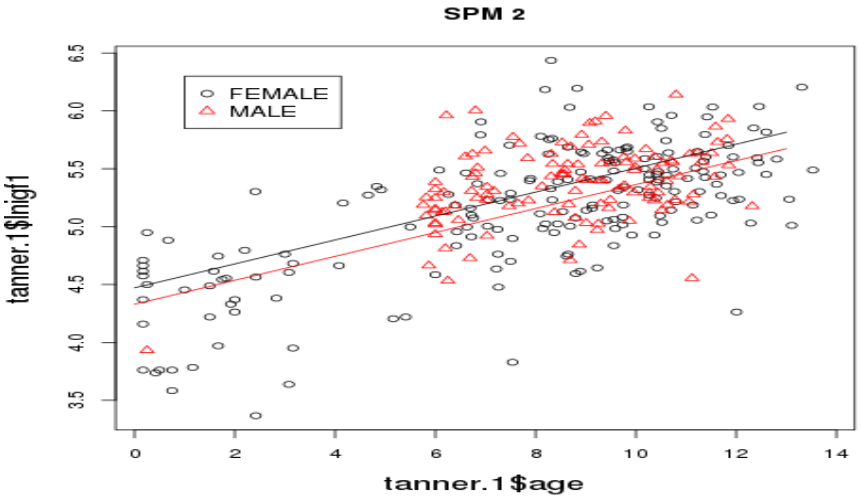
hvilket viser en klar kønsforskel og en meget stærk alderseffekt.

For at anskueliggøre denne model med *parallelle linier*, vi lige har fittet, er vi nødt til at prediktere ud fra modellen for en række forskellige aldre (mindst 2):

```
ny.F = data.frame(age=0:13,gender="FEMALE")
pred.F = predict(ancova, ny.F)
ny.M = data.frame(age=0:13,gender="MALE")
pred.M = predict(ancova, ny.M)
```

hvorefter vi kan tegne ved at skrive

```
plot(tanner.1$age, tanner.1$lnigf1, main="SPM 2",
pch=as.numeric(tanner.1$gender),
col=as.numeric(tanner.1$gender), cex.lab=1.5)
legend(1, 6.3, legend=c("FEMALE", "MALE"), pch=1:2,
       col=c("black", "red"), cex=1.1)
lines(ny.F$age, pred.F, type = "l", col="black")
lines(ny.M$age, pred.M, type = "l", col="red")
```



Hvis vi her skal kvantificere forskellen på drenge og piger, behøver vi ikke at sige, hvilken alder, det drejer sig om, da vi nu har antaget, at der er den samme forskel, uanset alder (da linierne jo nu er parallelle). Denne forskel estimeres til

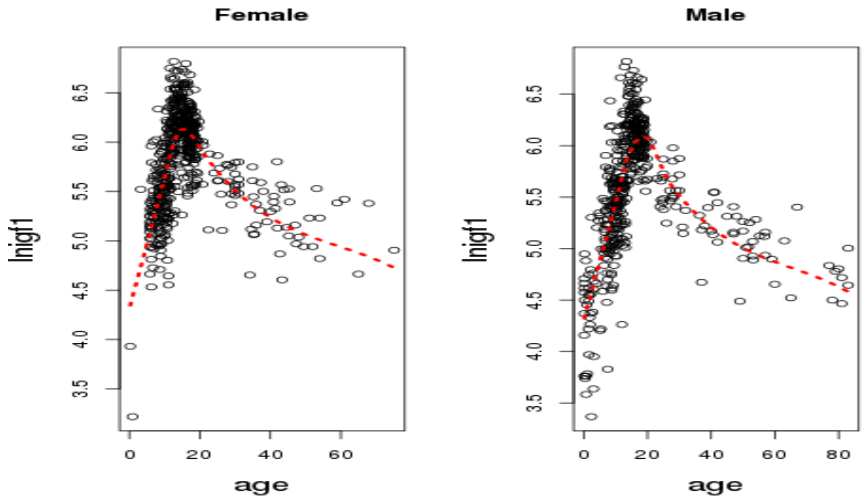
$\exp(-0.142) = 0.87$, med CI=$(\exp(-0.228), \exp(-0.055)) = (0.80, 0.95)$, altså at drenge generelt kun ligger 13% lavere end piger, med konfidensinterval fra 5-20% lavere.

3. *Forklar hvorfor en lineær regression af* $\log($*igf1*$)$ *overfor alder ville være misvisende, hvis man analyserede hele materialet på en gang.*

Her behøver vi sådan set bare en figur, nedenfor opdelt på piger og drenge, og forsynet med en udglattet kurve, som fås ved at skrive:

```
par(mfrow=c(1,2))
```

```
with(juul.F, scatter.smooth(age, lnigf1, main="Female", cex.lab=1.5,
lpars = list(col = "red", lwd = 3, lty = 3)))
with(juul.M, scatter.smooth(age, lnigf1, main="Male", cex.lab=1.5,
lpars = list(col = "red", lwd = 3, lty = 3)))
```



Mønstret ser rimeligt ens ud for de to køn, men er helt klart ikke lineært.

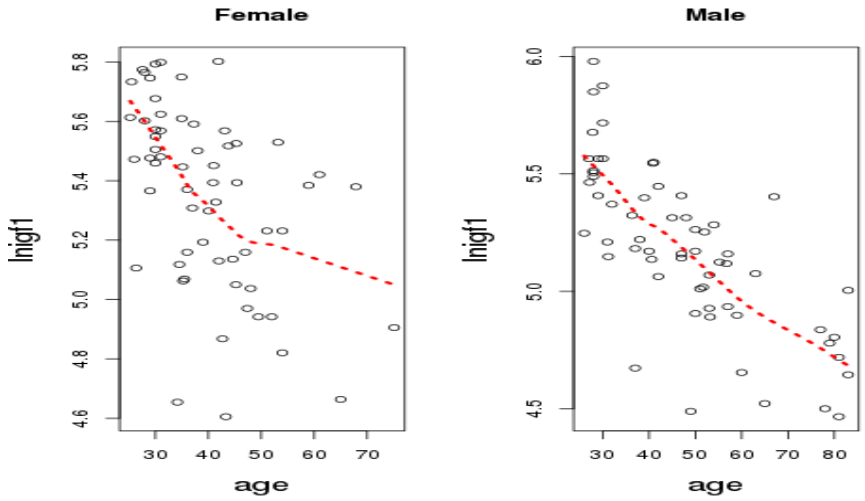
4. *Fokuser nu på de postpubertale* (alder $>$ 25 år) *og estimer det årlige fald i* `igf1`. *Er der forskel på kønnene?*

   Vi skal nu igen danne nogle nye data frames:

```
age.25 = juul[juul$age>25,]
age.25F = age.25[age.25$gender=="FEMALE",]
age.25M = age.25[age.25$gender=="MALE",]
```

   og derefter kan vi tegne aldersafhængigheden for kønnene enkeltvis for at vurdere lineariteten:

```
par(mfrow=c(1,2))
with(age.25F, scatter.smooth(age, lnigf1, main="Female", cex.lab=1.5,
lpars = list(col = "red", lwd = 3, lty = 3)))
with(age.25M, scatter.smooth(age, lnigf1, main="Male", cex.lab=1.5,
lpars = list(col = "red", lwd = 3, lty = 3)))
```



For mænd ser det jo rimeligt lineært ud, men for kvinder er der en tendens til affladning. Alligevel fortsætter vi sammenligningen udfra en antagelse om en lineært aftagende sammenhæng.

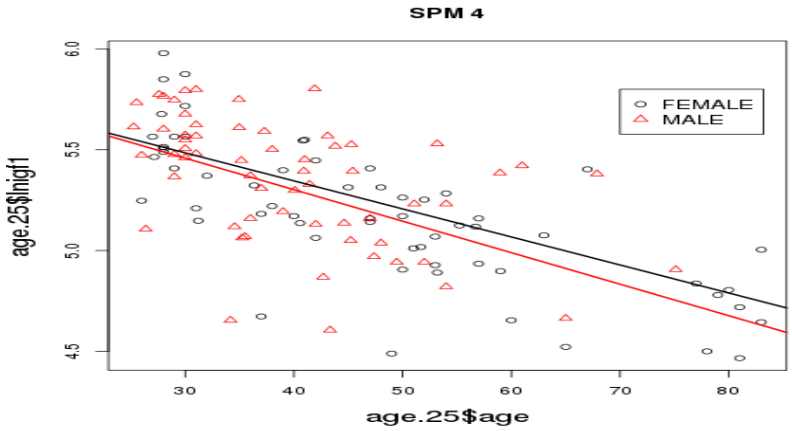
Vi kører derfor de individuelle regressionsmodeller, for at kunne lægge linierne ind på et plot:

```
model.F = lm(lnigf1 ~ age, data=age.25F)

model.M = lm(lnigf1 ~ age, data=age.25M)
```

og vi kan nu tegne linierne ind på et fælles plot, så vi visuelt kan bedømme forskellen på linierne:

```
par(mfrow=c(1,1))
plot(age.25$age, age.25$lnigf1, main="SPM 4",
                 pch=as.numeric(age.25$gender),
```

```
                col=as.numeric(age.25$gender), cex.lab=1.5)
legend(70, 5.8, legend=c("FEMALE", "MALE"), pch=1:2,
       col=c("black", "red"), cex=1.1)
abline(model.F, col="black", lwd=2)
abline(model.M, col="red", lwd=2)
```



Vi fitter nu en model med interaktion for at sammenligne de to hældninger:

```
interaktion = lm(lnigf1 ~ age+factor(gender)+age*factor(gender),
                 data=age.25)
```

og med `summary` og `confint`-funktionerne får vi da outputtet:

```
> summary(interaktion)

Call:
lm(formula = lnigf1 ~ age + factor(gender) + age * factor(gender),
    data = age.25)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77239 -0.16388  0.02478  0.14975  0.52304

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.926591   0.097309  60.905  < 2e-16 ***
```

```
age                        -0.015626   0.001927  -8.110 5.36e-13 ***
factor(gender)FEMALE       -0.025064   0.154972  -0.162    0.872
age:factor(gender)FEMALE    0.001720   0.003496   0.492    0.624
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2519 on 118 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.4628,Adjusted R-squared:  0.4492
F-statistic: 33.89 on 3 and 118 DF,  p-value: 7.133e-16


> confint(interaktion)
                                2.5 %       97.5 %
(Intercept)                5.733892122  6.119289554
age                       -0.019441441 -0.011810801
factor(gender)FEMALE      -0.331951052  0.281822160
age:factor(gender)FEMALE -0.005202882  0.008642748
```

Vekselvirkningsleddet er her klart insignifikant. Så vi fjerner det og får

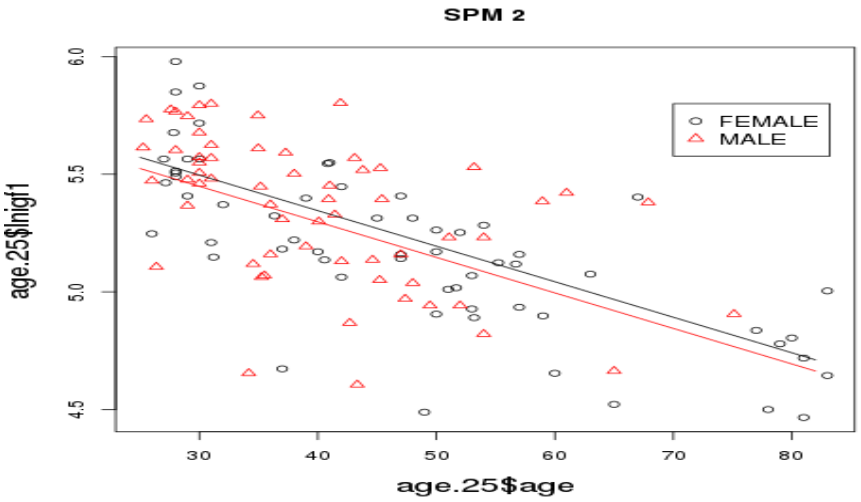
```
ancova = lm(lnigf1 ~ age+factor(gender), data=age.25)
```

som med `summary` og `confint`-funktionerne giver outputtet:

```
> summary(ancova)

Call:
lm(formula = lnigf1 ~ age + factor(gender), data = age.25)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77918 -0.15909  0.02754  0.14137  0.51293

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.901697   0.082854  71.230  < 2e-16 ***
age                 -0.015104   0.001603  -9.425 4.31e-16 ***
factor(gender)FEMALE  0.047541   0.047152   1.008    0.315
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2511 on 119 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.4617,Adjusted R-squared:  0.4527
F-statistic: 51.04 on 2 and 119 DF,  p-value: < 2.2e-16


> confint(ancova)
                           2.5 %      97.5 %
(Intercept)           5.73763701  6.06575682
age                  -0.01827688 -0.01193055
factor(gender)FEMALE -0.04582481  0.14090705
```

Vi ser, at `gender` ikke er signifikant, medens `age` er klart signifikant. For at vise, hvor tæt, de to parallelle regressionslinier ligger, tegner vi igen de predikterede værdier med på figuren:

```
ny.F = data.frame(age=25:82,gender="FEMALE")
pred.F = predict(ancova, ny.F)
ny.M = data.frame(age=25:82,gender="MALE")
pred.M = predict(ancova, ny.M)

plot(age.25$age, age.25$lnigf1, main="SPM 2",
pch=as.numeric(age.25$gender),
col=as.numeric(age.25$gender), cex.lab=1.5)
legend(70, 5.8, legend=c("FEMALE", "MALE"), pch=1:2,
       col=c("black", "red"), cex=1.1)
lines(ny.F$age, pred.F, type = "l", col="black")
lines(ny.M$age, pred.M, type = "l", col="red")
```



Læg mærke til fortegnet! `igf1` stiger med alderen for de små og falder med alderen for de voksne. Hvis man blander dem sammen får man en næsten vandret regressionslinje, som selvfølgelig slet ikke beskriver

data. Dette så vi også i figurerne i spørgsmål 3.

Det fælles årlige fald i `igf1` kvantificeres til faktoren $\exp(-0.0151) = 0.9845$, med konfidensinterval $(\exp(-0.0183), \exp(-0.0119)) = (0.9819, 0.9882)$, altså et fald på 1.55% (CI fra 1.18-1.81%).

5. *Udvid modellen fra forrige spørgsmål med en ekstra kovariat, idet* bmi = vægt/højde^2 *inddrages.*
Estimer igen det årlige fald i `igf1`*, og kommenter på forskellen til spørgsmål 4.*

Vi beregner nu `bmi` ved at skrive:

```
age.25$bmi = age.25$weight/(age.25$height/100)**2
```

og er så klar til at lave en multipel regressionsmodel, hvor vi bibeholder de to kovariater fra før:

```
glm = lm(lnigf1 ~ age+factor(gender)+bmi, data=age.25)
```

som med `summary` og `confint`-funktionerne giver outputtet:

```
> summary(glm)

Call:
lm(formula = lnigf1 ~ age + factor(gender) + bmi, data = age.25)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67014 -0.24267  0.02681  0.25098  0.45912

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.448062   0.416322  13.086 2.16e-14 ***
age                 -0.009216   0.005008  -1.840    0.075 .
factor(gender)FEMALE -0.015727   0.122412  -0.128    0.899
bmi                  0.011265   0.017218   0.654    0.518
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.3021 on 32 degrees of freedom
  (96 observations deleted due to missingness)
Multiple R-squared:  0.1032,Adjusted R-squared:  0.01914
```

```
F-statistic: 1.228 on 3 and 32 DF,  p-value: 0.3157


> confint(glm)
                             2.5 %      97.5 %
(Intercept)           4.60004116 6.296082690
age                  -0.01941659 0.000984797
factor(gender)FEMALE -0.26507299 0.233618861
bmi                  -0.02380636 0.046336550
```

Det ses at der nu ikke er noget, der bliver signifikant. Ikke engang alderen ser ud til at have nogen betydning mere.... Hvordan gik det til?

Dette kunne tolkes som et udslag af confounding, altså at der var for tæt en sammenhæng med den nyligt indførte kovariat `bmi` og en eller begge af de to tidligere. Dette er imidlertid *ikke* tilfældet, og forklaringen skal søges et helt andet sted, nemlig i et **stort antal manglende værdier**.

Faktisk indgår der kun 36 observationer i analysen, mod 122 i den fra spørgsmål 4, hvor vi kun så på alder og køn. Vægt og højde er kun registreret på et fåtal af personerne, men mangler for de resterende og må derfor udgå af analyserne. Dette er en effekt man skal være på vagt overfor, især når man har mange kovariater.