# Besvarelse af opgave om Vital Capacity

hentet fra P. Armitage & G. Berry: Statistical methods in medical research. 2nd ed. Blackwell, 1987.

**Spørgsmål 1**: *Indlæs data og konstruer en faktor (klassevariabel) med beskrivende navne til de 3 grupper*

Dette gøres f.eks. ved at skrive nedenstående kode

```
FILENAME HentURL URL
   "http://biostat.ku.dk/~lts/basal/data/cadmium.txt";

data cadmium;
infile HentURL firstobs=2;
input grp age vitcap;

if grp=1 then group='1:expo>10';
if grp=2 then group='2:expo<10';
if grp=3 then group='3:no-expo';
run;
```

idet vi samtidig rekoder gruppevariablen fra 1-2-3 til nogle mere brugervenlige betegnelser i den nye variabel `group`.

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

For at udregne summary statistics, skriver vi

```
proc means data=cadmium;
  class group;
  var age vitcap;
run;
```

hvorved output bliver:

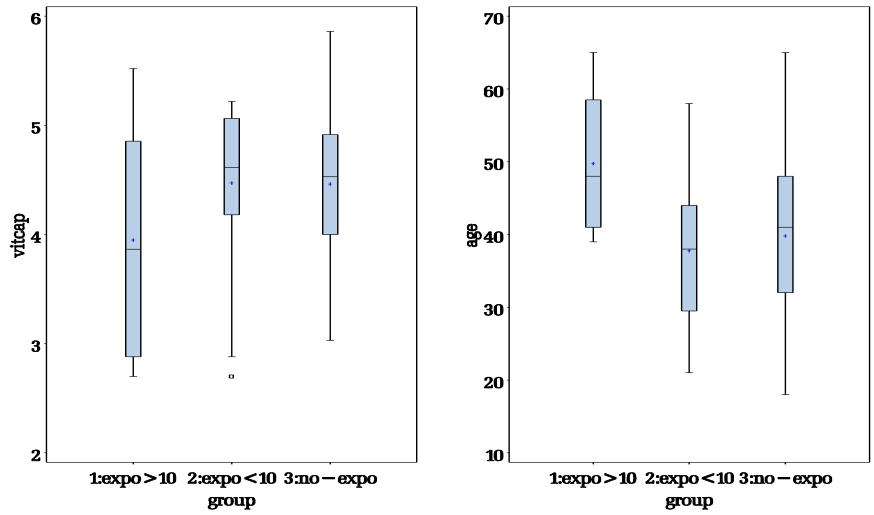
```
The MEANS Procedure

             N
group      Obs  Variable   N          Mean       Std Dev       Minimum       Maximum
-------------------------------------------------------------------------------------
1:expo>10   12  age        12    49.7500000     9.1066908    39.0000000    65.0000000
                vitcap     12     3.9491667     1.0330578     2.7000000     5.5200000

2:expo<10   28  age        28    37.7857143     9.1948341    21.0000000    58.0000000
                vitcap     28     4.4717857     0.6817084     2.7000000     5.2200000

3:no-expo   44  age        44    39.7954545    12.0049981    18.0000000    65.0000000
                vitcap     44     4.4620455     0.6922615     3.0300000     5.8600000
-------------------------------------------------------------------------------------
```

Bemærk, at personer i gruppen eksponeret mere end 10 år generelt er ældre (hvilket ikke er så sært), men at de ueksponerede har en noget større aldersvariation. Lungefunktionen ser (ikke overraskende) ud til at være dårligst blandt de langtidseksponerede.

Derudover kan man fx. lave boxplots ved at skrive:

```
proc sort data=cadmium;
  by group;
run;

proc boxplot data=cadmium;
  plot vitcap*group;
run;

proc boxplot data=cadmium;
  plot age*group;
run;
```

hvorved man får figurerne




**Spørgsmål 3**: *Ignorer i første omgang age variablen. Er der forskel på vital capacity i de 3 grupper? Angiv både parametrisk og nonparametrisk test. Giv estimater for forskellene i vital capacity, og suppler med konfidensgrænser for disse forskelle.*

Først gør vi det parametrisk med `proc glm`:

```
proc glm data=cadmium;
  class group;
  model vitcap=group / solution clparm;
run;
```

der producerer outputtet

```
The GLM Procedure

              Class Level Information

Class         Levels    Values
group              3    1:expo>10 2:expo<10 3:no-expo

Number of Observations Read          84
Number of Observations Used          84

The GLM Procedure
 Dependent Variable: vitcap
```

```
                                 Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F
Model                       2    2.74733766    1.37366883     2.48  0.0902
Error                      81   44.89361829    0.55424220
Corrected Total            83   47.64095595

R-Square     Coeff Var      Root MSE    vitcap Mean
0.057668      16.95060      0.744474       4.392024

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
group                       2    2.74733766    1.37366883     2.48  0.0902

                                                Standard
Parameter                    Estimate              Error   t Value   Pr > |t|

Intercept                4.462045455 B     0.11223375     39.76     <.0001
group     1:expo>10     -0.512878788 B     0.24245260     -2.12     0.0375
group     2:expo<10      0.009740260 B     0.17997438      0.05     0.9570
group     3:no-expo      0.000000000 B      .               .        .

NOTE: The X'X matrix has been found to be singular, and a generalized
      inverse was used to solve the normal equations.  Terms whose
      estimates are followed by the letter 'B' are not uniquely estimable.

Parameter                 95% Confidence Limits

Intercept                4.238735506  4.685355403
group     1:expo>10     -0.995283412 -0.030474163
group     2:expo<10     -0.348352306  0.367832825
group     3:no-expo       .            .
```

Det fremgår at der ikke er signifikant forskel på grupperne (på 5% niveau) i denne analyse, idet F-testets P-værdi på 0.092 ikke er under 0.05.
**Repetition**: Variation *mellem* grupper er linjen mærket `Model` (2 frihedsgrader), variation *indenfor* grupper er `Error` (81 frihedsgrader). Der er en større MS mellem grupper end indenfor grupper, men altså ikke nok til at det er signifikant.

Til at beregne konfidensintervaller for parvise forskelle mellem grupperne, benytter vi en `lsmeans`-sætning, og beder om konfidensintervaller svarende til de Bonferroni-korrigerede T-tests.

Samlet ser det således ud:

```
proc glm data=cadmium;
  class group;
  model vitcap=group / solution;
lsmeans group / adjust=bon cl pdiff;
run;
```

```
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Bonferroni

                    vitcap      LSMEAN
group               LSMEAN      Number

1:expo>10       3.94916667           1
2:expo<10       4.47178571           2
3:no-expo       4.46204545           3


     Least Squares Means for effect group
     Pr > |t| for H0: LSMean(i)=LSMean(j)

          Dependent Variable: vitcap

i/j              1             2             3

   1                      0.1355        0.1124
   2        0.1355                      1.0000
   3        0.1124        1.0000


                    vitcap
group               LSMEAN      95% Confidence Limits

1:expo>10        3.949167        3.521561     4.376773
2:expo<10        4.471786        4.191852     4.751720
3:no-expo        4.462045        4.238736     4.685355


        Least Squares Means for Effect group

             Difference         Simultaneous 95%
                Between      Confidence Limits for
i    j            Means       LSMean(i)-LSMean(j)

1    2       -0.522619       -1.150588     0.105350
1    3       -0.512879       -1.105606     0.079849
2    3        0.009740       -0.430246     0.449726
```

Alle intervallerne indeholder 0, i overensstemmelse med at der ikke overordnet set er signifikant forskel på de tre grupper (men dette kan man ikke være sikker på).

Den nonparametriske variant fås med ved at skrive:

```
proc npar1way wilcoxon data=cadmium;
  class group;
  var vitcap;
run;
```

hvorved man får outputtet:

```
The NPAR1WAY Procedure

             Wilcoxon Scores (Rank Sums) for Variable vitcap
                      Classified by Variable group

                          Sum of      Expected       Std Dev          Mean
group            N        Scores      Under H0      Under H0         Score
-------------------------------------------------------------------------
1:expo>10       12        382.50         510.0     78.219339     31.875000
2:expo<10       28       1276.50        1190.0    105.373232     45.589286
3:no-expo       44       1911.00        1870.0    111.638401     43.431818

                    Average scores were used for ties.

   Kruskal-Wallis Test

Chi-Square         2.7909
DF                      2
Pr > Chi-Square    0.2477
```

Heller ikke her finder vi altså nogen signifikans. Hvis man ser på gennemsnit, henholdsvis "mean rank", så kunne det godt se ud som om at gruppe 1 ligger lidt lavere, men det drukner i den store variation (som må formodes i høj grad at skyldes aldersvariationen, idet vitalkapaciteten må forventes at aftage kraftigt med alderen).

**Spørgsmål 4** *Udregn korrelationen mellem alder og vital capacity for hver gruppe for sig, samt for alle 3 grupper under et.*
*Hvad kan vi slutte af dette?*

Vi tager det lige i omvendt rækkefølge. Først for hele populationen:

```
proc corr data=cadmium;
  var age vitcap;
run;
```

der (bl.a. giver følgende output):

```
The CORR Procedure
   2  Variables:    age      vitcap

Pearson Correlation Coefficients, N = 84
       Prob > |r| under H0: Rho=0

                 age        vitcap
age          1.00000      -0.60512
                           <.0001

vitcap      -0.60512       1.00000
              <.0001
```

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

```
proc sort data=cadmium;  by group;
run;
proc corr data=cadmium;  by group;
  var age vitcap;
run;
```

hvilket giver outputtet:

```
group=1:expo>10

The CORR Procedure
   2  Variables:    age      vitcap

Pearson Correlation Coefficients, N = 12
       Prob > |r| under H0: Rho=0

                 age        vitcap
age          1.00000      -0.75028
                           0.0049

vitcap      -0.75028       1.00000
              0.0049
```

```
------------------------------------------------------------------------
group=2:expo<10

The CORR Procedure
   2  Variables:    age      vitcap

Pearson Correlation Coefficients, N = 28
       Prob > |r| under H0: Rho=0

                 age        vitcap
age          1.00000      -0.62762
                           0.0004

vitcap      -0.62762       1.00000
             0.0004

------------------------------------------------------------------------
group=3:no-expo

The CORR Procedure
   2  Variables:    age      vitcap

Pearson Correlation Coefficients, N = 44
       Prob > |r| under H0: Rho=0

                 age        vitcap
age          1.00000      -0.53088
                           0.0002

vitcap      -0.53088       1.00000
             0.0002
```

Det kan noteres at korrelationen er mindst i gruppe 3 og størst i gruppe 1. Imidlertid er det formentlig svar på et forkert spørgsmål, idet det er mere naturligt at ville vide om regressionslinjen er stejlere i nogen grupper end i andre, fordi hældningen angiver det konkrete fald i lungekapacitet for hvert år, man bliver ældre.

**Spørgsmål 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:

```
proc reg data=cadmium;  by group;
  model vitcap=age / clb;
run;
```

giver outputtet

```
group=1:expo>10

The REG Procedure
Dependent Variable: vitcap

Number of Observations Read          12
Number of Observations Used          12

                        Analysis of Variance

                                  Sum of          Mean
Source                     DF     Squares        Square   F Value   Pr > F
Model                       1     6.60823       6.60823     12.88   0.0049
Error                      10     5.13106       0.51311
Corrected Total            11    11.73929

Root MSE              0.71631    R-Square     0.5629
Dependent Mean        3.94917    Adj R-Sq     0.5192
Coeff Var            18.13837

                      Parameter Estimates

                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|
Intercept     1       8.18344        1.19787       6.83      <.0001
age           1      -0.08511        0.02372      -3.59      0.0049

Variable     DF      95% Confidence Limits
Intercept     1       5.51442       10.85245
age           1      -0.13795       -0.03227

-----------------------------------------------------------------------
group=2:expo<10

The REG Procedure
Dependent Variable: vitcap

Number of Observations Read          28
Number of Observations Used          28

                        Analysis of Variance

                                  Sum of          Mean
Source                     DF     Squares        Square   F Value   Pr > F
Model                       1     4.94260       4.94260     16.90   0.0004
Error                      26     7.60501       0.29250
Corrected Total            27    12.54761
```

```
Root MSE              0.54083    R-Square     0.3939
Dependent Mean        4.47179    Adj R-Sq     0.3706
Coeff Var            12.09434

                      Parameter Estimates

                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|
Intercept     1       6.23003        0.43977      14.17      <.0001
age           1      -0.04653        0.01132      -4.11      0.0004

Variable     DF      95% Confidence Limits
Intercept     1       5.32608        7.13399
age           1      -0.06980       -0.02326

-------------------------------------------------------------------
group=3:no-expo

The REG Procedure
Dependent Variable: vitcap

Number of Observations Read          44
Number of Observations Used          44

                        Analysis of Variance

                                  Sum of          Mean
Source                     DF     Squares        Square   F Value   Pr > F
Model                       1     5.80758       5.80758     16.48   0.0002
Error                      42    14.79914       0.35236
Corrected Total            43    20.60672

Root MSE              0.59360    R-Square     0.2818
Dependent Mean        4.46205    Adj R-Sq     0.2647
Coeff Var            13.30331

                      Parameter Estimates

                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|
Intercept     1       5.68029        0.31314      18.14      <.0001
age           1      -0.03061        0.00754      -4.06      0.0002

Variable     DF      95% Confidence Limits
Intercept     1       5.04836        6.31222
age           1      -0.04583       -0.01540
```

Vi noterer os regressionskoefficienterne med tilhørende SE: 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 gruppe 1.

**Spørgsmål 6:** *Kan sammenhængen mellem alder og vital capacity påvises at være forskellig for de tre grupper? Tegn rådata og den fittede relation for hver gruppe i samme plot. Beskriv og kvantificer forskellene!*

Regression for alle tre grupper samlet foretages v.hj.a. en generel lineær mo-

del, **med interaktionsled**, idet det er dette led, der er af speciel interesse her:

```
proc glm data=cadmium;
  classes group;
  model vitcap=group age group*age / solution clparm;
run;
```

Outputtet bliver:

```
The GLM Procedure
Dependent Variable: vitcap

                                 Sum of
Source                     DF       Squares   Mean Square  F Value  Pr > F

Model                       5   20.10574307    4.02114861    11.39  <.0001
Error                      78   27.53521288    0.35301555
Corrected Total            83   47.64095595

R-Square     Coeff Var      Root MSE    vitcap Mean
0.422026      13.52796      0.594151       4.392024

Source                     DF     Type I SS   Mean Square  F Value  Pr > F
group                       2    2.74733766    1.37366883     3.89  0.0245
age                         1   14.85894745   14.85894745    42.09  <.0001
age*group                   2    2.49945795    1.24972898     3.54  0.0338

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
group                       2    2.15205767    1.07602883     3.05  0.0531
age                         1   15.52632541   15.52632541    43.98  <.0001
age*group                   2    2.49945795    1.24972898     3.54  0.0338

                                          Standard
Parameter                 Estimate             Error   t Value   Pr > |t|
Intercept              5.680290602 B     0.31342603     18.12     <.0001
group     1:expo>10    2.503147783 B     1.04184195      2.40     0.0187
group     2:expo<10    0.549740376 B     0.57588436      0.95     0.3427
group     3:no-expo    0.000000000 B      .               .        .
age                   -0.030612671 B     0.00754746     -4.06     0.0001
age*group 1:expo>10   -0.054498319 B     0.02106980     -2.59     0.0116
age*group 2:expo<10   -0.015919340 B     0.01454687     -1.09     0.2772
age*group 3:no-expo    0.000000000 B      .               .        .

Parameter               95% Confidence Limits
Intercept              5.056307317  6.304273888
group     1:expo>10    0.428999787  4.577295778
group     2:expo<10   -0.596757322  1.696238074
group     3:no-expo      .            .
age                   -0.045638502 -0.015586840
age*group 1:expo>10   -0.096445068 -0.012551569
age*group 2:expo<10   -0.044879930  0.013041250
age*group 3:no-expo      .            .

NOTE: The X'X matrix has been found to be singular, and a generalized
      inverse was used to solve the normal equations.  Terms whose
      estimates are followed by the letter 'B' are not uniquely estimable.
```

Bemærk at vekselvirkningen 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 for `age*group 1` *forskellen* på regressionskoefficienterne i gruppe 1 og gruppe 3, og ved `age*group 2` den tilsvarende forskel fra gruppe 2 til 3. Det ses at den signifikante forskel først og fremmest skyldes at gruppe 1 aftager hurtigere 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.

Vi kan også udvide programmeringen med `estimate`-sætninger:

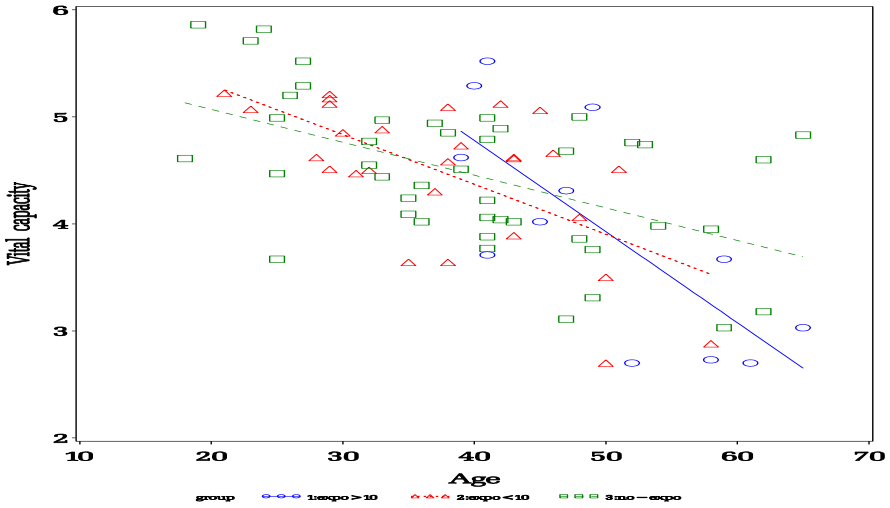
```
proc glm data=cadmium;
  classes group;
  model vitcap=group age group*age / solution clparm;
estimate 'slope 1 vs. 2' group*age 1 -1 0;
estimate 'slope 1 vs. 3' group*age 1 0 -1;
estimate 'slope 2 vs. 3' group*age 0 1 -1;
run;
```

der giver ekstra output:

```
                                    Standard
Parameter                Estimate       Error    t Value    Pr > |t|

slope 1 vs. 2         -0.03857898  0.02327272      -1.66      0.1014
slope 1 vs. 3         -0.05449832  0.02106980      -2.59      0.0116
slope 2 vs. 3         -0.01591934  0.01454687      -1.09      0.2772

Parameter              95% Confidence Limits

slope 1 vs. 2         -0.08491141   0.00775345
slope 1 vs. 3         -0.09644507  -0.01255157
slope 2 vs. 3         -0.04487993   0.01304125
```

Heraf ses, at det kun er ydergrupperne, der adskiller sig signifikant fra hinanden. P-værdien for denne sammenligning er tilstrækkelig lille til at "overleve" en Bonferroni-korrektion (gang med 3, da der er 3 sammenligninger).

De tre regressionslinier ser således ud:




**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.

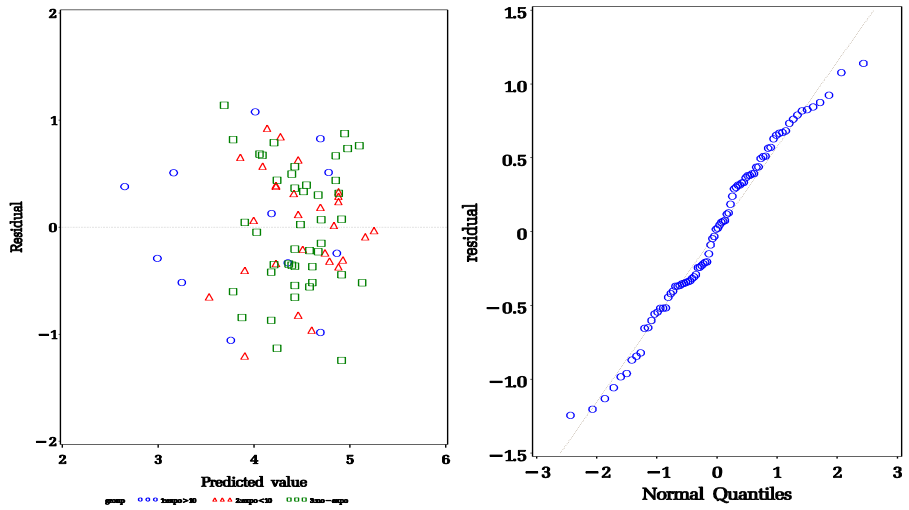
Hvis man inddrager et andengradsled i alder, er der dog ingensomhelst tegn på, at dette giver en forbedret model, så effekten ser virkelig ud til at kunne forklares udfra cadmium ekspositionen.

Og så burde vi jo også lige se på noget modelkontrol, f,.eks. ved at skrive:

```
proc glm data=cadmium;
  classes group;
  model vitcap=group age group*age / solution;
output out=ny p=yhat r=residual;
run;

proc gplot data=ny;
plot residual*yhat=group /
  haxis=axis1 vaxis=axis2 vref=0 lv=33 frame;
axis1 value=(H=2) minor=NONE
      label=(H=2 'Predicted value');
axis2 value=(H=2) minor=NONE
      label=(A=90 R=0 H=2 'Residual');
```

```
symbol1 v=circle i=none l=1 c=BLUE h=2 w=2;
symbol2 v=triangle i=none l=33 c=RED h=2 w=2;
symbol3 v=square i=none l=2 c=GREEN h=2 w=2;
run;

proc univariate normal data=ny;
  var residual;
  qqplot / height=3 normal(mu=EST sigma=EST l=33);
run;
```

der giver os følgende figurer:




Der er muligvis en tendens til skævhed i fordelingen af residualer, så måske burde man logaritmetransformere.... Dette vil dog ikke ændre resultaterne nævneværdigt.

# Besvarelse af "juul2"-opgaven

**Spørgsmål 1:** *Indlæs data*

Nedenfor indlæser vi, idet vi samtidig rekoder kønnet til mere sigende betegnelser og laver et par transformationer, som vi får brug for i de efterfølgende analyser:

```
FILENAME HentURL URL "http://biostat.ku.dk/~lts/basal/data/juul2.txt";
DATA Juul2;
  INFILE HentURL FIRSTOBS=2;
  INPUT Age Height Menarche Sexnr sIGf1 Tanner TestVol Weight;

  IF sexnr=2 THEN sex='female';
  IF sexnr=1 THEN sex='male';

  lnIGF1=log(sigf1);
  BMI=weight/(height/100)**2;
RUN;
```

**Spørgsmål 2:** *Lav regressionsanalyser for præpubertale individer (Tanner stadium 1), for hvert køn for sig, med logaritmetransformeret igf1 som outcome, og alder som forklarende variabel.*

Her benytter vi `proc reg`, for hvert køn for sig, og vi husker at sortere efter `by`-variablen først. Vi benytter 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).

```
proc sort data=juul2; by sex;
run;

PROC REG DATA=juul2;  WHERE tanner=1; BY sex;
  MODEL lnIGF1=age / CLB;
RUN;
```

Vi får herved outputtet

The SAS System

sex=female

The REG Procedure
Model: MODEL1
Dependent Variable: lnIGf1

Number of Observations Read 224
Number of Observations Used 119
Number of Observations with Missing Values 105

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 2.41551 | 2.41551 | 26.37 | <.0001 |
| Error | 117 | 10.71855 | 0.09161 | | |
| Corrected Total | 118 | 13.13407 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.30267 | R-Square | 0.1839 |
| Dependent Mean | 5.36593 | Adj R-Sq | 0.1769 |
| Coeff Var | 5.64066 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 4.73692 | 0.12560 | 37.71 | <.0001 |
| Age | 1 | 0.07272 | 0.01416 | 5.13 | <.0001 |

Parameter Estimates

| Variable | DF | 95% Confidence Limits | |
|---|---|---|---|
| Intercept | 1 | 4.48817 | 4.98567 |
| Age | 1 | 0.04467 | 0.10076 |

---

The SAS System

sex=male

The REG Procedure
Model: MODEL1
Dependent Variable: lnIGf1

Number of Observations Read 291
Number of Observations Used 192
Number of Observations with Missing Values 99

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 29.67590 | 29.67590 | 177.86 | <.0001 |
| Error | 190 | 31.70119 | 0.16685 | | |
| Corrected Total | 191 | 61.37709 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.40847 | R-Square | 0.4835 |
| Dependent Mean | 5.13962 | Adj R-Sq | 0.4808 |
| Coeff Var | 7.94749 | | |

Parameter Estimates

| | | |
|---|---|---|
| | Parameter | Standard |

```
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1        4.28605        0.07046      60.83      <.0001
Age           1        0.10897        0.00817      13.34      <.0001

             Parameter Estimates
Variable     DF       95% Confidence Limits
Intercept     1        4.14706        4.42504
Age           1        0.09285        0.12509
```

Dvs. for pigerne har vi 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, så tilbagetransformerer små tal ($< \pm 0.1$) til en relativ forskel af ca. samme størrelse).

**Spørgsmål 3:** *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 (`PROC GLM`), og her er det specielt interaktionsleddet, der er interessant:

```
PROC GLM DATA=juul2;  WHERE tanner=1;
  CLASS sex;
  MODEL lnIGF1=age sex sex*age;
RUN;
```

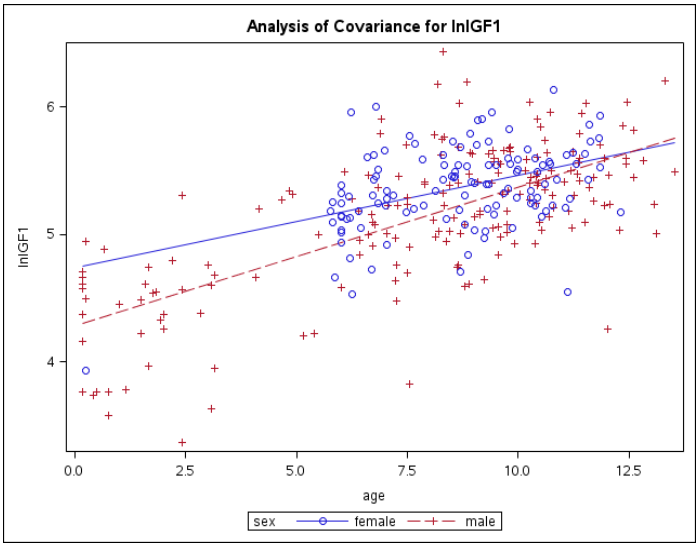
hvorved vi får outputtet

```
The GLM Procedure

      Class Level Information
Class         Levels    Values
sex                2    female male

Number of Observations Read         515
Number of Observations Used         311

-------------------------------------------------------------------------------
The GLM Procedure
Dependent Variable: lnIGF1

                                   Sum of
Source                      DF       Squares    Mean Square   F Value   Pr > F
Model                        3    35.85425505    11.95141835     86.49   <.0001
Error                      307    42.41974418     0.13817506
Corrected Total            310    78.27399923
```

```
R-Square     Coeff Var      Root MSE    lnIGf1 Mean
0.458061      7.112589      0.371719       5.226213

Source                     DF       Type I SS     Mean Square   F Value   Pr > F
Age                         1     33.89250259     33.89250259    245.29   <.0001
sex                         1      1.45414705      1.45414705     10.52   0.0013
Age*sex                     1      0.50760541      0.50760541      3.67   0.0562

Source                     DF      Type III SS     Mean Square   F Value   Pr > F
Age                         1     12.74910506     12.74910506     92.27   <.0001
sex                         1      1.00655073      1.00655073      7.28   0.0073
Age*sex                     1      0.50760541      0.50760541      3.67   0.0562
```

Bemærk at interaktionsleddet er meget tæt på signifikans (P=0.06). Det vil sige, at vi ikke med sikkerhed kan afvise, at linjerne er parallelle (har samme hældning), men at der er indikation af forskel på hældningerne.

Figuren svarende til de to linier ser således ud




og man kan se, at der er en enkelt pige (den yngste, 3 måneder gammel), der har meget stor indflydelse på hældningen. Uden denne pige ville hældningerne faktisk være signifikant forskellige (P=0.0075).

Modellen uden interaktionsled:

```
PROC GLM DATA=juul2;  WHERE tanner=1;
  CLASS sex;
  MODEL lnIGF1=age sex / SOLUTION CLPARM;
RUN;
```

giver outputtet

```
The GLM Procedure

      Class Level Information
Class         Levels    Values
sex                2    female male

Number of Observations Read         515
Number of Observations Used         311

------------------------------------------------------------------------------
The GLM Procedure
Dependent Variable: lnIGF1

                                   Sum of
Source                      DF       Squares    Mean Square   F Value   Pr > F
Model                        2    35.34664964    17.67332482    126.80   <.0001
Error                      308    42.92734959     0.13937451
Corrected Total            310    78.27399923

R-Square     Coeff Var      Root MSE    lnIGf1 Mean
0.451576      7.143393      0.373329       5.226213

Source                      DF      Type I SS    Mean Square   F Value   Pr > F
Age                          1    33.89250259    33.89250259    243.18   <.0001
sex                          1     1.45414705     1.45414705     10.43   0.0014

Source                      DF    Type III SS    Mean Square   F Value   Pr > F
Age                          1    31.58380727    31.58380727    226.61   <.0001
sex                          1     1.45414705     1.45414705     10.43   0.0014

                                          Standard
Parameter               Estimate             Error    t Value    Pr > |t|
Intercept            4.329935962 B      0.06015726      71.98      <.0001
Age                  0.103368319        0.00686668      15.05      <.0001
sex       female     0.141851570 B      0.04391589       3.23      0.0014
sex       male       0.000000000 B       .                .         .

Parameter             95% Confidence Limits
Intercept            4.211564753  4.448307172
Age                  0.089856777  0.116879860
sex       female     0.055438454  0.228264686
sex       male        .            .
```

hvilket viser en klar kønsforskel og en meget stærk alderseffekt.

**Spørgsmål 4:** *Gentag spørgsmål 2 og 3 for postpubertale (alder* $> 25$ *år).*

Her behøver vi sådan set bare at ændre filteret (`where`-sætningen) og køre samme analyser. Vi springer de separate analyser over og går direkte til den generelle lineære model:

```
PROC GLM DATA=juul2;  WHERE age>25;
  CLASS sex;
```

```
  MODEL lnIGF1=age sex sex*age;
run;
```

som giver outputtet

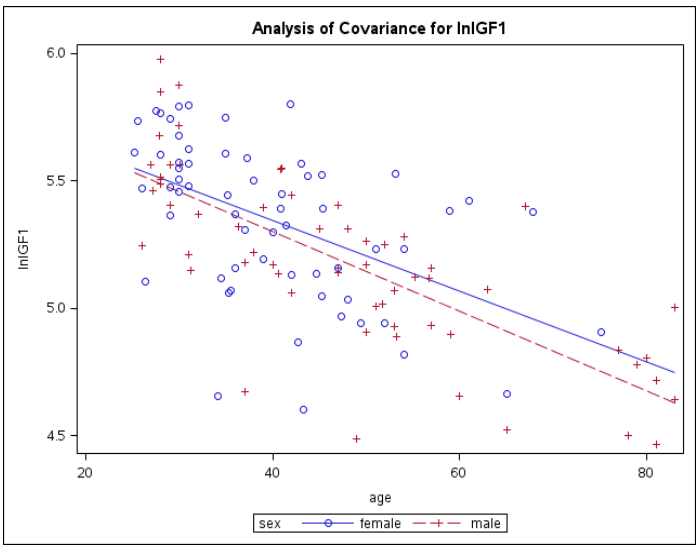
```
The GLM Procedure

      Class Level Information
Class         Levels    Values
sex                2    female male

Number of Observations Read         126
Number of Observations Used         122
```

---

```
The GLM Procedure

Dependent Variable: lnIGf1

                                   Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                       3    6.45004462     2.15001487     33.89   <.0001
Error                     118    7.48556355     0.06343698
Corrected Total           121   13.93560817

R-Square     Coeff Var      Root MSE    lnIGf1 Mean
0.462846      4.784050      0.251867       5.264723

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
Age                         1    6.37061241     6.37061241    100.42   <.0001
sex                         1    0.06407721     0.06407721      1.01   0.3169
Age*sex                     1    0.01535500     0.01535500      0.24   0.6236

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
Age                         1    4.52711640     4.52711640     71.36   <.0001
sex                         1    0.00165941     0.00165941      0.03   0.8718
Age*sex                     1    0.01535500     0.01535500      0.24   0.6236
```

Vekselvirkningsleddet ver her klart insignifikant, hvilket også ses af det tilhørende plot:

Figuren svarende til de to linier ser således ud




Så vi fjerner vekselvirkningsleddet og får

```
PROC GLM DATA=juul2;  WHERE age>25;
  CLASS sex;
  MODEL lnIGF1=age sex / SOLUTION CLPARM;
RUN;
```

---

```
The GLM Procedure
Dependent Variable: lnIGF1

                                 Sum of
Source                     DF       Squares    Mean Square  F Value  Pr > F
Model                       2     6.43468962     3.21734481    51.04  <.0001
Error                     119     7.50091855     0.06303293
Corrected Total           121    13.93560817

R-Square     Coeff Var      Root MSE    lnIGf1 Mean
0.461744      4.768790      0.251064       5.264723

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
Age                         1     6.37061241     6.37061241    101.07   <.0001
sex                         1     0.06407721     0.06407721      1.02   0.3154

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
Age                         1     5.59917556     5.59917556     88.83   <.0001
sex                         1     0.06407721     0.06407721      1.02   0.3154

                                                    Standard
Parameter                 Estimate                     Error    t Value    Pr > |t|
Intercept              5.901696915 B              0.08285436      71.23      <.0001
Age                   -0.015103715                0.00160253      -9.42      <.0001
sex       female       0.047541117 B              0.04715213       1.01      0.3154
sex       male         0.000000000 B                       .          .           .
```

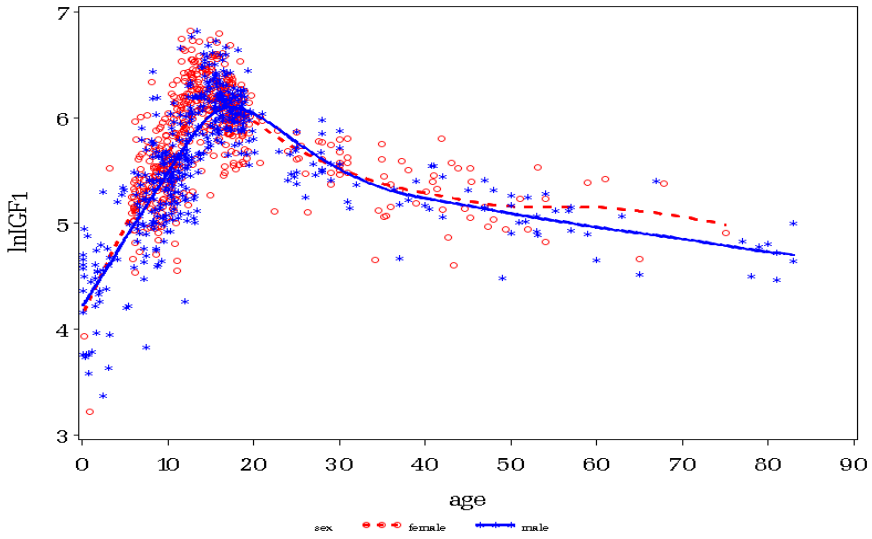
```
Parameter                95% Confidence Limits
Intercept             5.737637010  6.065756820
Age                  -0.018276880 -0.011930551
sex       female     -0.045824812  0.140907046
sex       male          .             .
```

Vi ser at `sex` ikke er signifikant, medens `age` er klart signifikant uanset om `sex` først fjernes fra modellen (Type I SS) eller ej (Type III SS), og at serum IGF-1 falder ca. 1.5% pr. år.

**Spørgsmål 5:** *Forklar hvorfor en lineær regression af* ln ***`igf1`*** *overfor alder ville være misvisende, hvis man analyserede hele materialet på en gang.*

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.

**I har vel husket at tegne!?**




**Spørgsmål 6:** *Udvid analysen i spørgsmål 4 til en multipel regressionsanalyse, idet* $BMI = vægt/højde^2$ *inddrages.*

Under indlæsningen udregnede vi `BMI = weight/(height/100)**2`, og vi er derfor klar til at lave en multipel regressionsmodel:

```
PROC GLM DATA=juul2; WHERE age>25;
```

```
  CLASS sex;
  MODEL lnIGF1=age sex bmi / SOLUTION CLPARM;
RUN;


The GLM Procedure
      Class Level Information

Class         Levels    Values
sex                2    female male

Number of Observations Read         126
Number of Observations Used          36

------------------------------------------------------------------------------

The GLM Procedure

Dependent Variable: lnIGf1

                                   Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F
Model                       3    0.33603880     0.11201293      1.23   0.3157
Error                      32    2.91985141     0.09124536
Corrected Total            35    3.25589021


R-Square     Coeff Var      Root MSE    lnIGf1 Mean
0.103210      5.706357      0.302068       5.293543

Source                     DF      Type I SS    Mean Square   F Value   Pr > F
Age                         1     0.29568049     0.29568049      3.24   0.0813
sex                         1     0.00129882     0.00129882      0.01   0.9058
BMI                         1     0.03905949     0.03905949      0.43   0.5176

Source                     DF    Type III SS    Mean Square   F Value   Pr > F
Age                         1     0.30901545     0.30901545      3.39   0.0750
sex                         1     0.00150610     0.00150610      0.02   0.8986
BMI                         1     0.03905949     0.03905949      0.43   0.5176

                                          Standard
Parameter               Estimate             Error    t Value    Pr > |t|
Intercept            5.448061928 B      0.41632229      13.09      <.0001
Age                 -0.009215896        0.00500787      -1.84      0.0750
sex       female    -0.015727063 B      0.12241241      -0.13      0.8986
sex       male       0.000000000 B       .                .         .
BMI                  0.011265095        0.01721777       0.65      0.5176

Parameter             95% Confidence Limits
Intercept            4.600041165  6.296082690
Age                 -0.019416590  0.000984797
sex       female    -0.265072986  0.233618861
sex       male         .            .
BMI                 -0.023806360  0.046336550
```

Det ses at der ikke er noget, der bliver signifikant. Type I kvadratsummerne kan bruges til successiv modelreduktion.

**Men hov!:** Alderen ser heller ikke ud til at være signifikant nu. Lige før var den klart signifikant! Hvordan gik det til?

Sagen er, at variansanalyseskemaet beregnes på de data, der indgår i den *fulde* model, og der indgår kun 36 observationer i den mod 122, når vi kun ser på alder og køn. Vægt og højde er kun registreret på et fåtal af personerne. Det er en effekt man skal være på vagt overfor, især når man har mange kovariater.