

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

```
FILENAME navn URL
      "http://staff.pubhealth.ku.dk/~lts/basal/data/cadmium.txt";

data cadmium;
infile navn 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;
```

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, skriver vi

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

hvorved output bliver:

The MEANS Procedure

group	N		Variable	N	Mean	Std Dev	Minimum	Maximum
	Obs							
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	

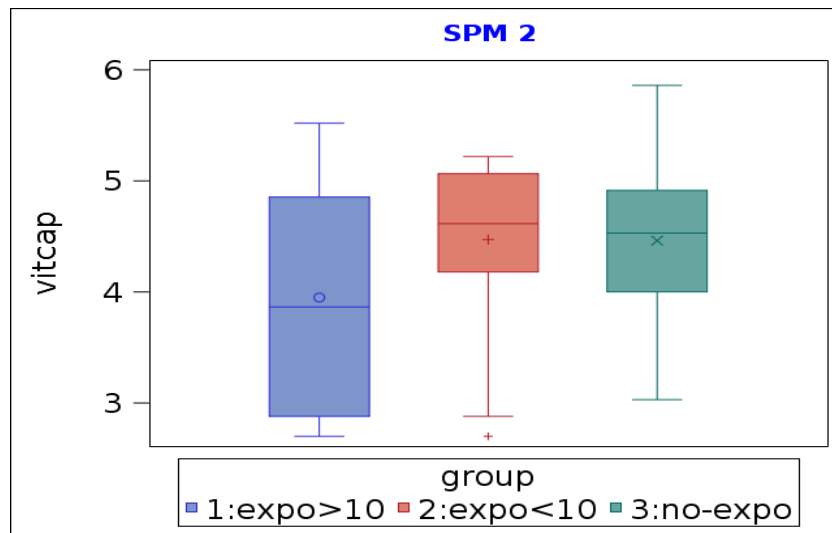
Bemærk, at personer i den langtidseksponerede gruppe generelt er ældre (hvilket ikke er så sært), og at de ueksponerede har en noget større aldersvariation. Lungefunktionen ser (ikke overraskende) ud til at være dårligst blandt de langtidseksponerede.

En grafisk illustration kunne være boxplots, som vi får ved at skrive:

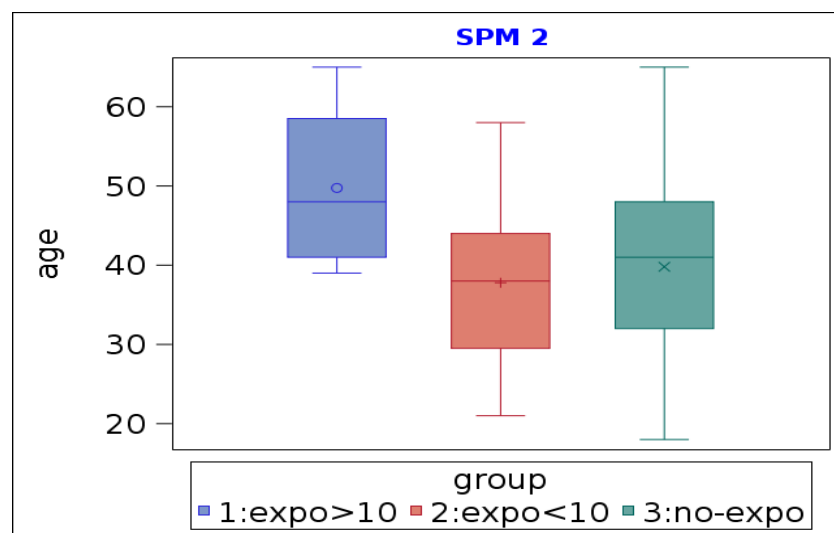
```
proc sgplot data=cadmium;
  vbox vitcap / group=group;
run;
```

```
proc sgplot data=cadmium;
  vbox age / group=group;
run;
```

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 langtids-eksponerede 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:

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

Dette 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

Dependent Variable: vitcap

Source	DF	Sum of 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

Parameter	Estimate	Standard 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% signifikansniveau) i denne analyse, idet F-testets P-værdi på 0.092 ikke er under 0.05.

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

Da gruppen af ueksponerede (**no-expo**) er sidst i alfabetet, er denne gjort til referencegruppe, og den søgte forskel står derfor direkte at aflæse som -0.513 (-0.995, -0.030) 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).

Nu er der spurgt specifikt om denne forskel, men ellers skulle vi have korrigeret for massesignifikans. Dette kan gøres med en **lsmeans**-sætning, hvor vi beder om konfidensintervaller svarende til de Tukey-korrigerede T-tests.

Samtidig checker vi også varianshomogeniteten med et Levenes test (**hovtest=levене** i en **means**-sætning), samt foretager et Welch-test, som er den form for ensidet ANOVA, man bør benytte i tilfælde af variansinhomogenitet.

Samlet ser det således ud:

```
proc glm data=cadmium;
  class group;
```

```

model vitcap=group / solution;
lsmeans group / adjust=tukey cl pdiff;
means group / hovtest=levene welch;
run;

```

og de Tukey-korrigerede sammenligninger falder således ud:

```

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

```

group	vitcap LSMEAN	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		0.1105	0.0930
2	0.1105		0.9984
3	0.0930	0.9984	

```

Least Squares Means for Effect group

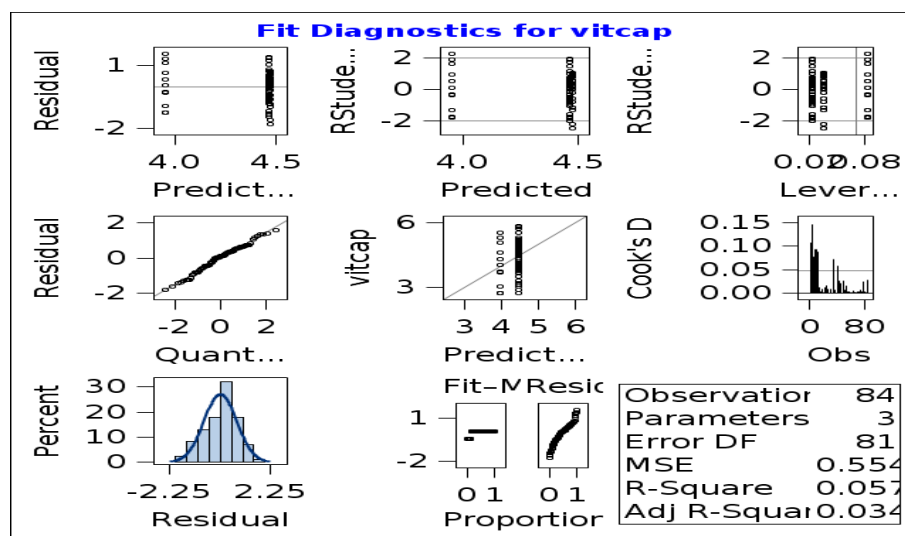
```

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-0.522619	-1.135904	0.090666
1	3	-0.512879	-1.091746	0.065988
2	3	0.009740	-0.419957	0.439438

Alle konfidensintervallerne for de parvise differenser ses 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 $(-1.092, 0.066)$, altså noget bredere end før, fordi vi nu har korrigeret for 3 sammenligninger.

Modelkontrollen giver billedet



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.

Output fra Levenes test og Welch test giver

The GLM Procedure

Levene's Test for Homogeneity of vitcap Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
group	2	2.7647	1.3824	2.97	0.0571
Error	81	37.7358	0.4659		

Welch's ANOVA for vitcap

Source	DF	F Value	Pr > F
group	2.0000	1.37	0.2721
Error	27.2030		

og vi bemærker, at vi er lige på kanten af en signifikant forskel på spredningerne i de 3 grupper, men at Welch-testet giver en højere P-værdi end den sædvanlige ANOVA og dermed bekræfter konklusionen

om “ingen forskel”.

Hvis det havde været normalfordelingsantagelsen, vi havde haft problemer med, kunne vi have udført en non-parametrisk sammenligning 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

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean 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.

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:

```
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 <.0001	1.00000

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 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 (hvor vi til afveksling benytter **reg**-proceduren) giver:

```

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 Used 12

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

Parameter Estimates

Variable	DF	Parameter Estimate	Standard 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 Used 28

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

Parameter Estimates

Variable	DF	Parameter Estimate	Standard 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 Used 44

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

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	5.68029	0.31314	18.14	<.0001

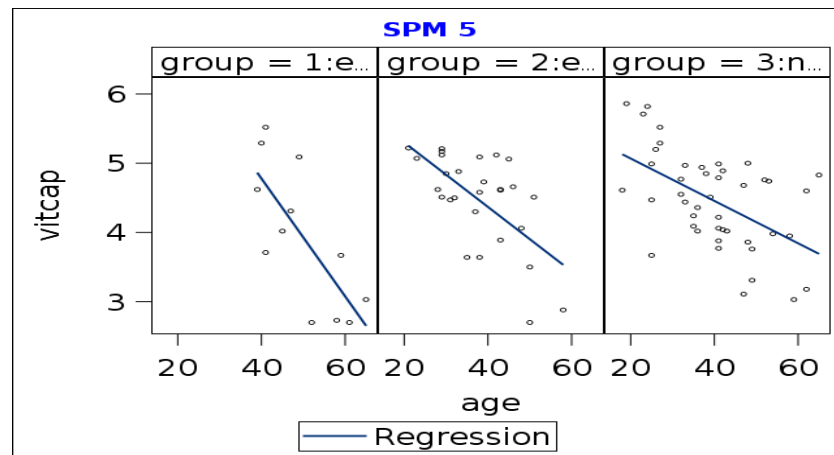
age	1	-0.03061	0.00754	-4.06	0.0002
Variable	DF	95% Confidence Limits			
Intercept	1	5.04836	6.31222		
age	1	-0.04583	-0.01540		

Vi noterer os regressionskoefficienterne med tilhørende 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

```
proc sgpanel data=cadmium;
panelby group / rows=1 columns=3;
reg Y=vitcap X=age;
run;
```

hvorved vi får figurene



Her ses tydeligt, at figuren til venstre (svarende til de langtidseksponerede, selv om man ikke lige kan se det af overskriften) har den stejleste hældning.

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`.

Desuden inkluderer vi `estimate`-sætninger til udregning af forventede værdier for 40-årige i de to nævnte grupper (selv om dette egentlig ikke er nødvendigt, da der ikke spørges om dette) samt en `estimate`-sætning til at finde den forventede forskel på disse.

Totalt skriver vi altså:

```
proc glm data=cadmium;
  class group;
  model vitcap=group age group*age / solution clparm;
  estimate "expo>10, 40 years" intercept 1 group 1 0 0 age 40 group*age 40 0 0;
  estimate "no-expo, 40 years" intercept 1 group 0 0 1 age 40 group*age 0 0 40;
  estimate "expo>10 vs. no, 40 years" group 1 0 -1 group*age 40 0 -40;
run;
```

og vi får outputtet:

The GLM Procedure
Dependent Variable: vitcap

Source	DF	Sum of 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					
0.422026					
Coeff Var					
13.52796					
Root MSE					
0.594151					
vitcap Mean					
4.392024					
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

Parameter	Estimate	Standard Error	t Value	Pr > t
expo>10, 40 years	4.77899881	0.25730240	18.57	<.0001
no-expo, 40 years	4.45578377	0.08958495	49.74	<.0001
expo>10 vs. no, 40 years	0.32321504	0.27245181	1.19	0.2391

Parameter	Estimate	95% Confidence Limits
expo>10, 40 years	4.26674909	5.29124854
no-expo, 40 years	4.27743383	4.63413372
expo>10 vs. no, 40 years	-0.21919484	0.86562492

Parameter	Estimate	Standard 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	.	.	.

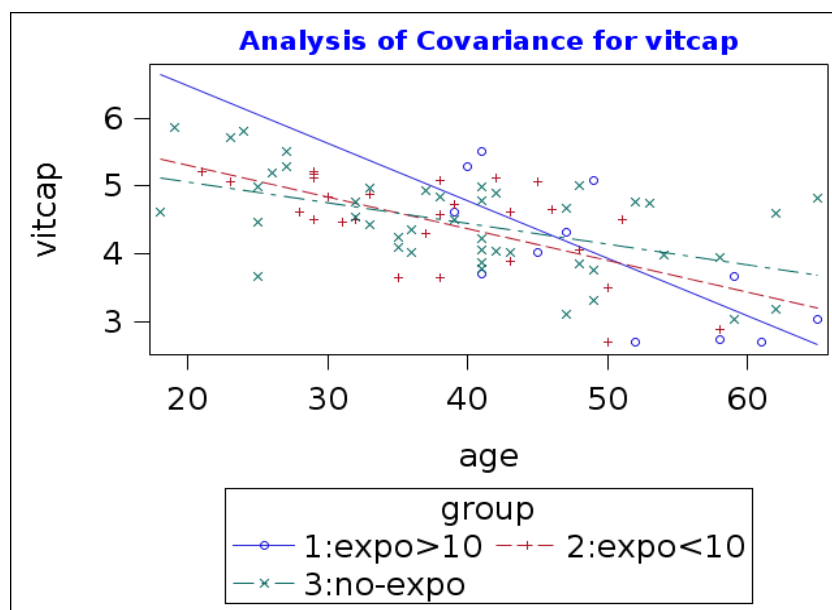
Parameter	Estimate	95% Confidence Limits
Intercept	5.056307317	6.304273888
group 1:expo>10	0.428999787	4.577295778
group 2:expo<10	-0.596757322	1.696238074
group 3:no-expo	.	.
age	-0.045638502	-0.015586840
age*group 1:expo>10	-0.096445068	-0.012551569
age*group 2:expo<10	-0.044879930	0.013041250
age*group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

Vi ser, at det søgte estimat 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-parallele linier, som det ses af figuren nedenfor (et biprodukt ved `glm`-analysen), og her ser vi, at linierne svarende til de 3 grupper krydser ind over hinanden, således, at den blå linie hørende til de langtids-eksponerede netop ligger højest for de lave aldre.



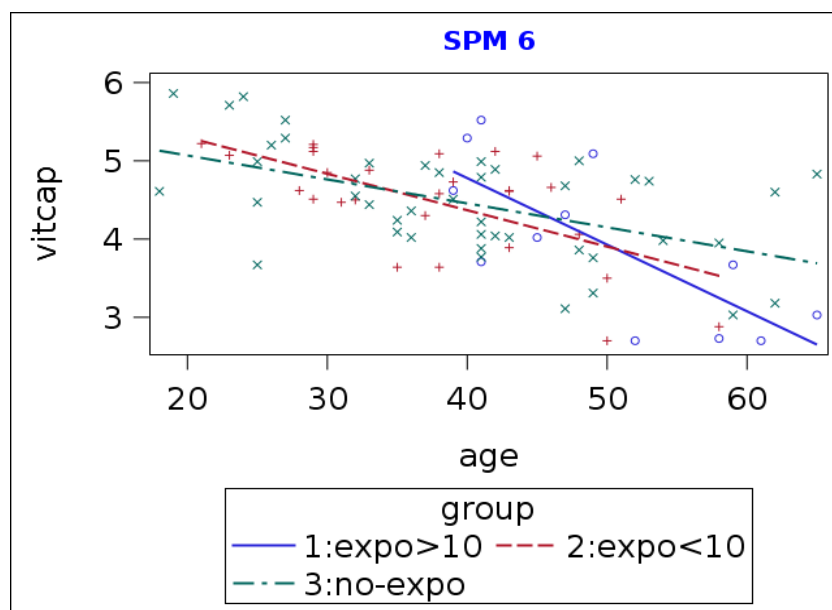
Hvis vi ser tilbage på summary statistics fra spørgsmål 2, kan vi se, at den yngste langtids-eksponerede er 39 år, så det er lidt vovet at udtale sig om ovenstående forskel.

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

En pænere figur fås ved at benytte plotte-proceduren i stedet for

```
proc sgplot data=cadmium;
  reg Y=vitcap X=age / group=group;
run;
```

der giver figuren



- *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 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=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 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 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.

Vi kan også udvide programmeringen med **estimate**-sætninger til parvise sammenligninger af de tre hældninger:

```
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:

Parameter	Estimate	Standard 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

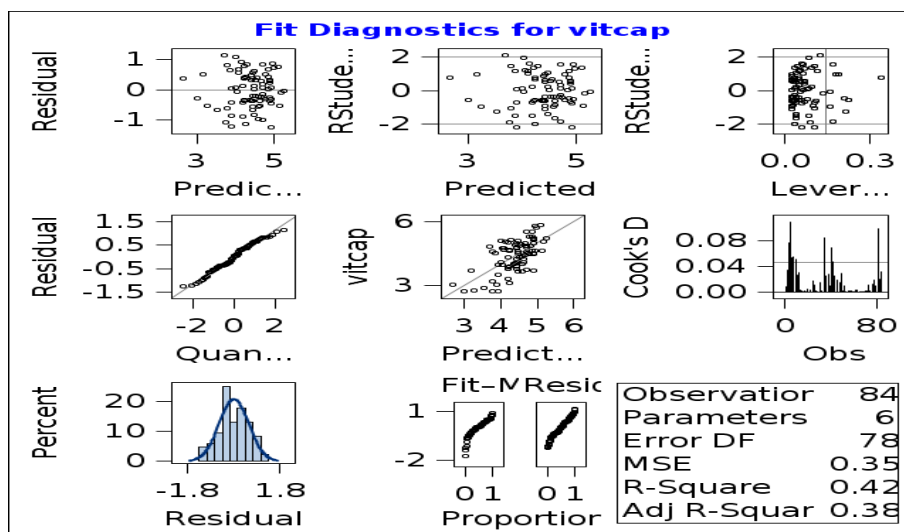
Heraf ses, at det kun er ydergrupperne, der adskiller sig signifikant fra hinanden. P-værdien for denne sammenligning er tilstrækkelig lille til at “overleve” en Bonferroni-korrektion (gang med 3, da der er 3 sammenligninger).

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 ingen som helst tegn på, at dette giver en forbedret model, så effekten ser virkelig ud til at kunne forklares ud fra 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 et par transformationer, som vi får brug for i de efterfølgende analyser:

```
FILENAME navn URL "http://staff.pubhealth.ku.dk/~lts/basal/data/juul2.txt";

data juul;
infile navn firstobs=2;
input age height sexnr sigf1 tanner weight;

if sexnr=2 then sex='female';
if sexnr=1 then sex='male';

lnigf1=log(sigf1);
bmi=weight/(height/100)**2;
run;
```

Her benytter vi `PROC GLM`, 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).

Vi inkluderer desuden en `estimate`-sætning til fortolkning af en 5-års ændring:

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

proc glm data=juul; where tanner=1; by sex;
  model lnigf1=age / solution clparm;
  estimate "5* slope" age 5;
run;
```

Vi får herved outputtet

```
sex=female

The GLM Procedure

Number of Observations Read      224
Number of Observations Used      119

R-Square      Coeff Var      Root MSE      lnigf1 Mean
0.183912      5.640660      0.302674      5.365932

Parameter              Estimate      Standard      t Value      Pr > |t|
                    Error
5* slope              0.36358713      0.07080743      5.13      <.0001

Parameter              95% Confidence Limits
5* slope              0.22335673      0.50381752

Parameter              Estimate      Standard      t Value      Pr > |t|
                    Error
Intercept              4.736920331      0.12560101      37.71      <.0001
age                    0.072717426      0.01416149      5.13      <.0001

Parameter              95% Confidence Limits
```

Intercept	4.488174112	4.985666549
age	0.044671347	0.100763505

sex=male

The GLM Procedure

Number of Observations Read	291
Number of Observations Used	192

Dependent Variable: lnigf1

R-Square	Coeff Var	Root MSE	lnigf1 Mean
0.483501	7.947495	0.408471	5.139617

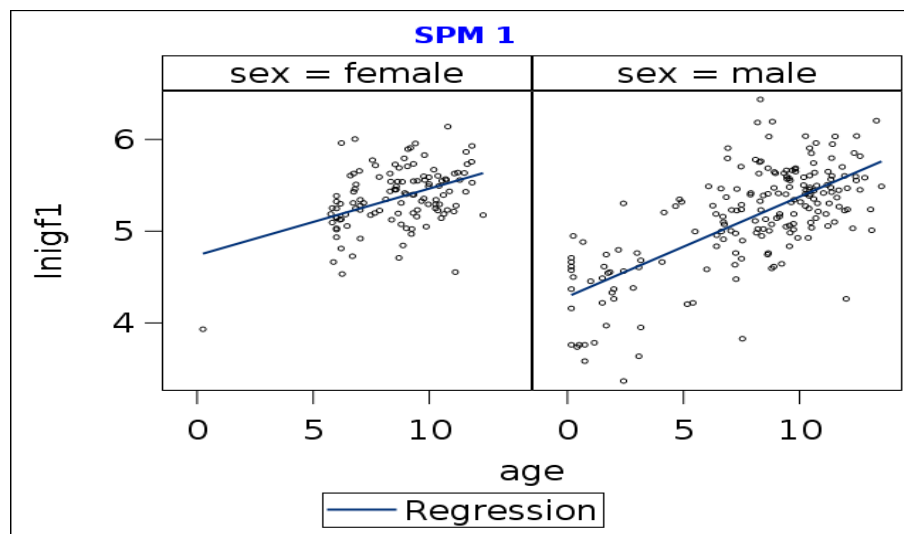
Parameter	Estimate	Standard Error	t Value	Pr > t
5* slope	0.54485485	0.04085450	13.34	<.0001

Parameter	95% Confidence Limits	
5* slope	0.46426820	0.62544149

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	4.286050575	0.07046495	60.83	<.0001
age	0.108970969	0.00817090	13.34	<.0001

Parameter	95% Confidence Limits	
Intercept	4.147056473	4.425044678
age	0.092853639	0.125088299

Dvs. for pigerne har vi regressionslinjen $\ln \text{igf1} = 4.737 + 0.0727 \times \text{alder}$ og for drengene $\ln \text{igf1} = 4.286 + 0.1090 \times \text{alder}$, 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.



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

Til dette formål havde vi inkluderet en `estimate`-sætning, som bare gangede hældningen op med 5. Det kunne vi selvfølgelig bare have gjort selv, men her var vi dovne. Vi mangler dog stadig at tilbagetransformere til noget forståeligt, fordi vores outcome er transformeret med den naturlige logaritme. Vi tilbagetransformerer derfor med exponentialfunktionen, og finder

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

For drenge: $\exp(0.5449) = 1.72$,
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 (`PROC GLM`), og her er det specielt interaktionsleddet, der er interessant.

Vi inkluderer samtidig nogle `estimate`-sætninger, som vi senere får brug for til at estimere forskellen i 7-års alderen.

```
proc glm plots=(DiagnosticsPanel residuals(smooth)) data=juul;
  where tanner=1;
  class sex;
  model lnigf1=age sex sex*age / solution clparm;
  estimate "F, 7 years" intercept 1 sex 1 0 age 7 sex*age 7 0;
  estimate "M, 7 years" intercept 1 sex 0 1 age 7 sex*age 0 7;
  estimate "M vs F, 7 years" sex -1 1 sex*age -7 7;
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

Source	DF	Sum of 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

Parameter	Estimate	Standard Error	t Value	Pr > t
F, 7 years	5.24594231	0.04455021	117.75	<.0001
M, 7 years	5.04884736	0.02753224	183.38	<.0001
M vs F, 7 years	-0.19709495	0.05237123	-3.76	0.0002

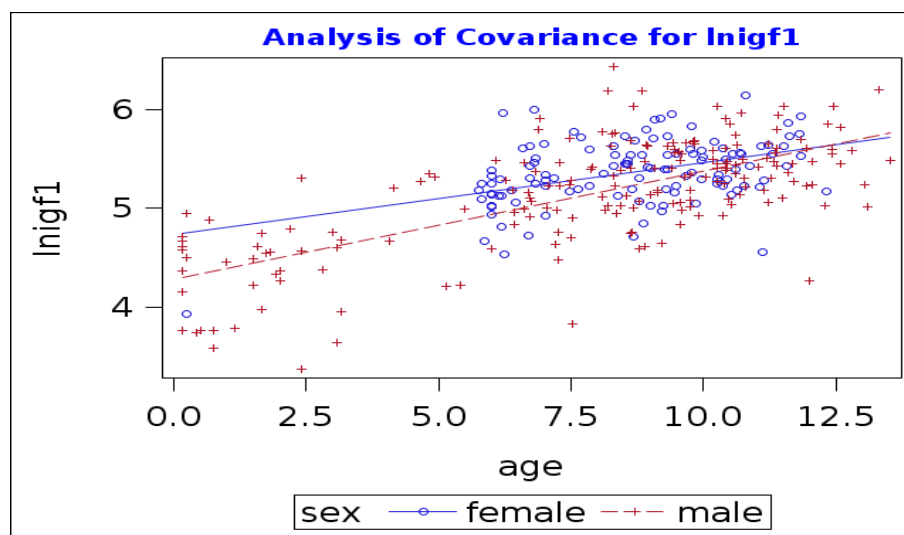
Parameter	95% Confidence Limits	
F, 7 years	5.15827991	5.33360471
M, 7 years	4.99467159	5.10302313
M vs F, 7 years	-0.30014693	-0.09404297

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 ($P=0.056$). Det vil sige, at vi ikke med sikkerhed kan afvise, at linjerne er parallelle (har samme hældning), men at der er en vis indikation af forskel på hældningerne.

I de separate figurer ovenfor kan denne forskel være svær at se, så nedenfor ses en fælles figur for de to køn, med indtegnede regressionslinier:



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

Vi inkluderede nogle `estimate`-sætninger i koden ovenfor, og vi har her specifikt brug for en af disse, nemlig:

```
estimate "M vs F, 7 years" sex -1 1 sex*age -7 7;
```

som gav outputtet

Parameter	Estimate	Standard Error	t Value	Pr > t
M vs F, 7 years	-0.19709495	0.05237123	-3.76	0.0002

Parameter	95% Confidence Limits	
M vs F, 7 years	-0.30014693	-0.09404297

Vores bedste gæt på forskellen i $\ln(\text{igf1})$ 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 $\text{CI}=(\exp(-0.300), \exp(-0.094)) = (0.74, 0.91)$, 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:

```
PROC GLM DATA=juul;  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

Dependent Variable: lnIGF1

Source	DF	Sum of 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

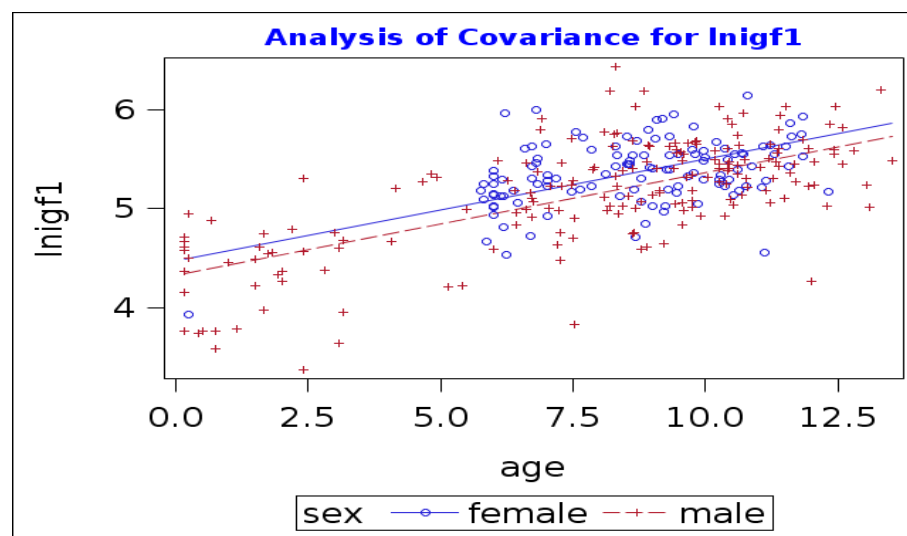
sex	1	1.45414705	1.45414705	10.43	0.0014
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	31.58380727	31.58380727	226.61	<.0001
sex	1	1.45414705	1.45414705	10.43	0.0014

Parameter	Estimate	Standard 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.

Den model, vi har fittet ses på figuren nedenfor:

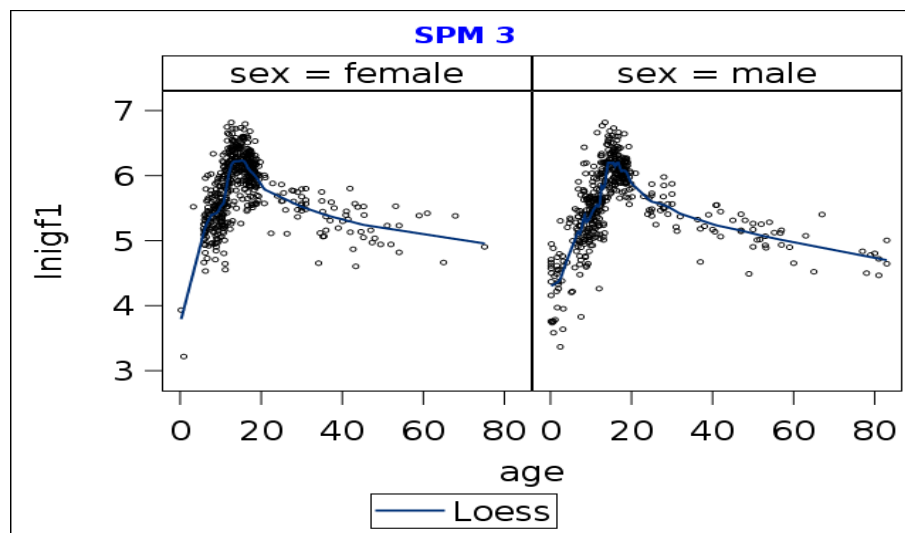


Hvis vi her skal kvantificere forskellen på drenge og piger, behøver vi ikke at sige, hvilken alder, det drejer sig om, da vi nu har antaget, at der er den samme forskel, uanset alder (da linierne jo nu er parallelle). Denne forskel estimeres til

$\exp(-0.142) = 0.87$, 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.

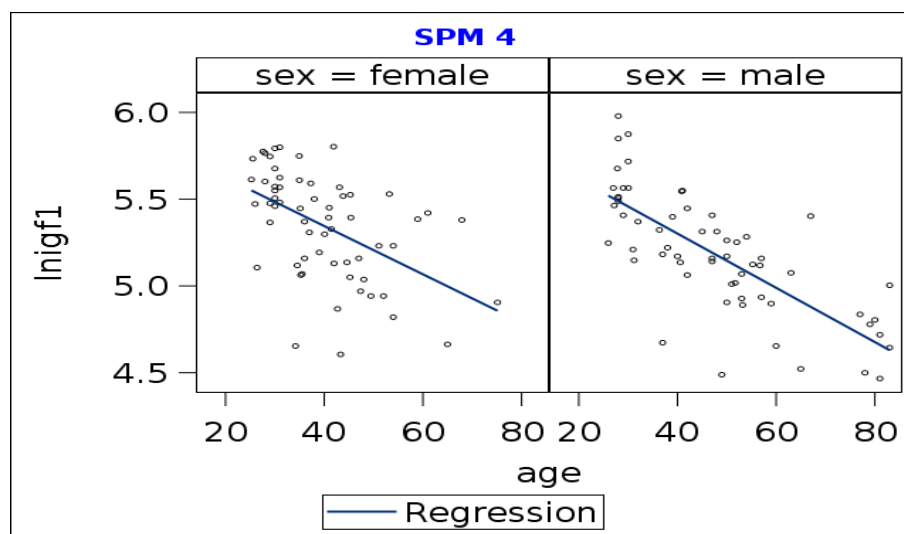
Her behøver vi sådan set bare en figur, nedenfor opdelt på piger og drenge, og forsynet med en udglattet loess-kurve:



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?

Først tegner vi kønnene enkeltvis:



Når vi skal udføre analyserne, behøver vi sådan set bare at ændre filteret (where-sætningen) fra tidligere og køre de 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

Dependent Variable: lnIGF1

Source	DF	Sum of 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, så vi fjerner det og får

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

med outputtet:

The GLM Procedure
Dependent Variable: lnIGF1

Source	DF	Sum of 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

Parameter	Estimate	Standard 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).

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%).

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 (jf. spørgsmål 3).

5. *Udvid modellen fra forrige spørgsmål med en ekstra kovariat, idet BMI = vægt/højde² inddrages. Estimer igen det årlige fald i **igf1**, og kommenter på forskellen til spørgsmål 4.*

Under indlæsningen udregnede vi $BMI = \text{weight}/(\text{height}/100)**2$, og vi er derfor klar til at lave en multipel regressionsmodel, hvor vi vælger at bibeholde **sex** i modellen, selv om denne var insignifikant:

```
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
```

Dependent Variable: lnIGf1

Source	DF	Sum of 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

Parameter	Estimate	Standard 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 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.