

Vejledende besvarelse af hjemmeopgave, efterår 2018

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

Der er foretaget en del undersøgelser af krigsveteraner og deres helbredsforhold, og vi skal se på et par aspekter af dette.

På hjemmesiden

http://staff.pubhealth.ku.dk/~lts/basal18_2/hjemmeopgave.html

ligger oplysninger om 7 variable på 299 krigsveteraner (et udpluk af et meget større materiale), og disse er:

idnr: Løbenummer for veteranen

alder: Veteranens alder ved undersøgelsen (år)

bpright: Systolisk blodtryk på højre arm (mm)

bpleft: Systolisk blodtryk på venstre arm (mm)

fev1: Lungefunktionsmål (l/min)

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

alkuge: Antal genstande pr. uge

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

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

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

```
DATA vet;  
  INFILE  
    "http://staff.pubhealth.ku.dk/~lts/basal18_2/hjemmeopgave/hjemmeopgave.txt"  
  URL FIRSTOBS=2;  
  INPUT idnr alder bpright bpleft fev1 alkuge ryger$;  
  
dif=bpright-bpleft;
```

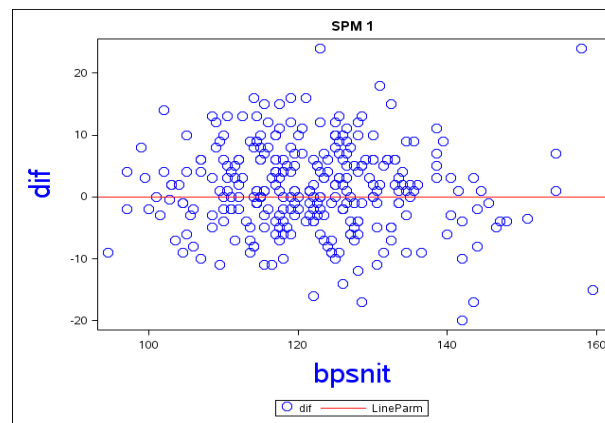
```

bpsnit=(bpright+bpleft)/2;
run;

proc sgplot data=vet;
  scatter x=bpsnit y=dif /
    markerattrs=(color=blue size=0.3cm);
  xaxis labelattrs=(color=blue size=0.8cm);
  yaxis labelattrs=(color=blue size=0.8cm);
  lineparm x=0 y=0 slope=0 / lineattrs=(color=red);
run;

```

hvorved vi får figuren:



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

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

```

data vet;
set vet;

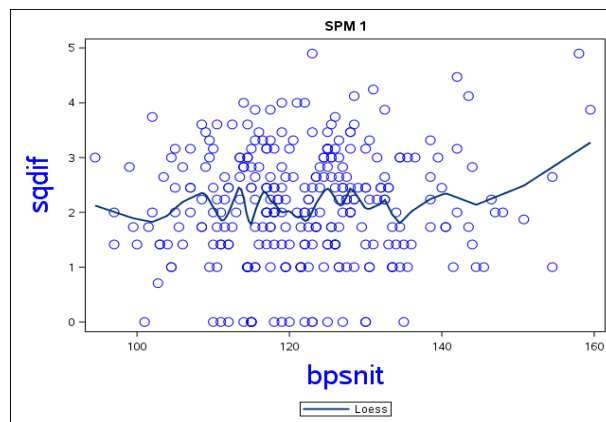
sqdif=sqrt(abs(dif));

```

```
run;

proc sgplot data=veter;
  loess x=bpsnit y=sqdif /
  markerattrs=(color=blue size=0.3cm);
  xaxis labelattrs=(color=blue size=0.8cm);
  yaxis labelattrs=(color=blue size=0.8cm);
run;
```

hvorved vi får:



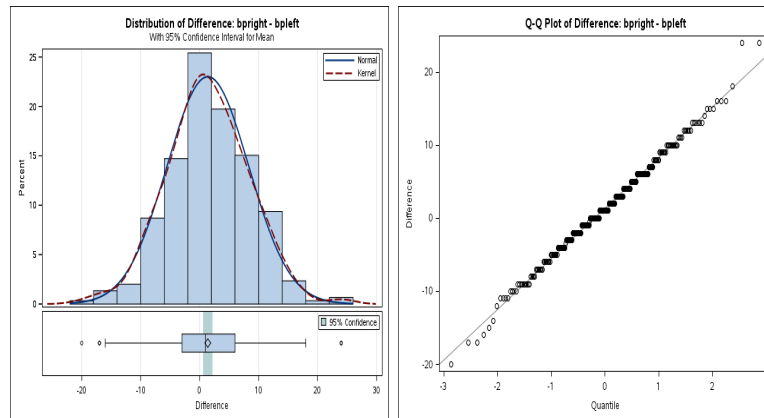
Bemærk, hvad der *ikke* spørges om i dette spørgsmål: Fordelingen af blodtryk på hver af siderne, og egentlig heller ikke noget om fordelingen af differenserne før i spørgsmål 2 (og endnu mere i spørgsmål 3).

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

I lyset af svaret på ovenstående spørgsmål, arbejder vi videre med blodtrykket uden at transformere. Når vi skal sammenligne blodtrykket på højre og venstre side, skal vi derfor se på et parret T-test, eller blot et test for middelværdi 0 for differenserne. Vi udfører førstnævnte nedenfor, idet vi samtidig laver modelkontrol ved at benytte `plots=all`:

```
proc ttest plots=all data=veter;
  paired bpright*bpleft;
run;
```

Modelkontrollen går på, at differenserne skal være rimeligt normalfordelte, og en del af de automatiske figurer fra T-testet giver nedenstående histogram og fraktildiagram:



som på ingen måde giver anledning til bekymring.

Selve T-testet giver outputtet:

```
The TTEST Procedure
Difference:  bpriht - bpleft
```

N	Mean	Std Dev	Std Err	Minimum	Maximum
299	1.3813	6.9357	0.4011	-20.0000	24.0000

Mean	95% CL Mean	Std Dev	95% CL Std Dev
1.3813	0.5919 2.1706	6.9357	6.4208 7.5411

DF	t Value	Pr > t
298	3.44	0.0007

Heraf aflæser vi:

- Forskellen på blodtrykket på højre og venstre side estimeres til 1.38 mm, med konfidensinterval $CI=(0.59 \text{ mm}, 2.17 \text{ mm})$.
- Forskellen er signifikant, idet konfidensintervallet ikke indeholder 0. P-værdien angives til $P=0.0007$.
- En forskel på 10 mm (til en af siderne) falder langt udenfor konfidensintervallet, så ja, det kan vi udelukke.

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

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

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

Af T-test-outputtet ovenfor finder vi normalområdet til

$$1.3813 \pm 2 \times 6.9357 \approx (-12.5, 15.3)$$

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

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

```
proc univariate normal data=vet;
  var dif;
  output out=ny1
    pctlpts=2.5 5 95 97.5
    pctlpre=P_;
run;
```

hvorved vi får outputtet:

```
proc print data=ny1;
run;
```

Obs	P_2_5	P_5	P_95	P_97_5
1	-11	-10	13	15

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

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

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

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

```
data vet;  
set vet;  
  
dranker=(alkuge ge 10);  
  
if abs(dif) lt 10 then sideforskel="lille";  
if abs(dif) ge 10 then sideforskel="stor";  
run;
```

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

```
proc freq data=vet;  
  table dranker*sideforskel /  
    nocol nopercnt chisq relrisk;  
run;
```

Herved får vi outputtet

The FREQ Procedure

Table of dranker by sideforskeldranker sideforskel

Frequency				
Row Pct	lille	stor		Total
-----+-----+-----+				
0	187	44		231
	80.95	19.05		
-----+-----+-----+				
1	59	9		68
	86.76	13.24		
-----+-----+-----+				
Total	246	53		299

Statistics for Table of dranker by sideforskel

Statistic	DF	Value	Prob

Chi-Square	1	1.2170	0.2700
Likelihood Ratio Chi-Square	1	1.2839	0.2572
Continuity Adj. Chi-Square	1	0.8511	0.3563

Fisher's Exact Test

Two-sided Pr <= P	0.3660

Odds Ratio and Relative Risks

Statistic	Value	95% Confidence Limits	

Odds Ratio	0.6483	0.2988	1.4064
Relative Risk (Column 1)	0.9330	0.8342	1.0435
Relative Risk (Column 2)	1.4392	0.7408	2.7959

Sample Size = 299

Ud fra dette kan vi ikke påstå, at der er nogen sammenhæng mellem druk og store forskelle i blodtryk på de to arme (P=0.37 for Fishers eksakte test og P=0.27 for χ^2 -testet).

Af rækkeprocenterne ses, at risikoen for stor sideforskel er 19.05% for folk, der ikke er drankere, men kun 13.24% for drankere, men denne forskel er altså ikke signifikant....

Kan vi så udelukke at der kan være en sammenhæng? Næh, men vi kan kvantificere den mulige sammenhæng, f.eks. ved at udregne den relative risiko for stor sideforskel, for ikke-drankere vs. drankere. Denne størrelse er ovenfor estimeret til 1.4392, hvilket betyder, at risikoen for en stor sideforskel er næsten 44% større hos ikke-drankere i forhold til hos

drankere. Men sikkerhedsintervallet er (0.7408, 2.7959), dvs. den relative risiko kunne også være ned til 0.7408, hvilket svarer til en større risiko for drankere på op til $\frac{1}{0.7408} = 1.35$, altså 35% højere end hos ikke-drankere.

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

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

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

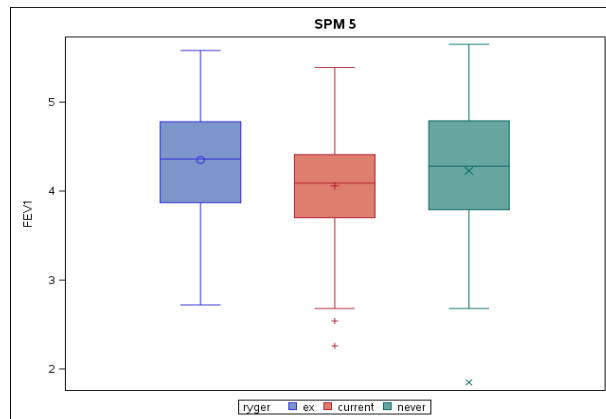
Nu kommer nogle spørgsmål, der fokuserer på lungefunktionen:

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

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

```
proc sgplot data=vet;  
    vbox fev1 / group=ryger;  
run;
```

hvilket giver figuren:



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

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

```
proc means N mean stddev data=vet;
  class ryger;
  var fev1;
run;
```

hvorved vi får

The MEANS Procedure
Analysis Variable : FEV1

ryger	N		Mean	Std Dev
	Obs	N		
current	139	139	4.0592086	0.5904102
ex	89	89	4.3494382	0.6482325
never	71	71	4.2280282	0.7619292

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

For at foretage en egentlig sammenligning mellem de tre grupper, kaster vi os nu ud i den ensidede variansanalyse, incl. Levenes test for varianshomogenitet og Welch-testet for ens middelværdier, der bør bruges, hvis varianserne er meget forskellige:

```

proc glm plots=all data=vet;
    class ryger;
    model fev1=ryger / solution clparm;
    means ryger / hovtest welch;
run;

```

Herved får vi

The GLM Procedure

Class Level Information		
Class	Levels	Values
ryger	3	current ex never

Number of Observations Read	299
Number of Observations Used	299

Dependent Variable: FEV1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.7373256	2.3686628	5.58	0.0042
Error	296	125.7202088	0.4247304		
Corrected Total	298	130.4575344			

R-Square	Coeff Var	Root MSE	FEV1 Mean
0.036313	15.57005	0.651713	4.185686

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ryger	2	4.73732564	2.36866282	5.58	0.0042

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	4.228028169 B	0.07734416	54.67	<.0001
ryger current	-0.168819536 B	0.09506699	-1.78	0.0768
ryger ex	0.121410033 B	0.10370328	1.17	0.2426
ryger never	0.000000000 B	.	.	.

Parameter	95% Confidence Limits
-----------	-----------------------

Intercept	4.075814036	4.380242302
ryger current	-0.355912401	0.018273329
ryger ex	-0.082679140	0.325499206
ryger never	.	.

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.

Levene's Test for Homogeneity of FEV1 Variance ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ryger	2	2.4095	1.2047	3.18	0.0431
Error	296	112.2	0.3791		

Level of	-----FEV1-----		
ryger	N	Mean	Std Dev
current	139	4.05920863	0.59041016
ex	89	4.34943820	0.64823249
never	71	4.22802817	0.76192917

Welch's ANOVA for FEV1

Source	DF	F Value	Pr > F
ryger	2.0000	6.04	0.0030
Error	155.3		

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

Testet for identitet af de 3 middelværdier giver i ANOVA $P=0.0042$, men vi checker med Welch test, om denne signifikans hænger på antagelsen om identiske varianser. Da Welch test giver $P=0.0030$, føler vi os dog ret sikre på, at middelværdierne i de 3 grupper ikke er ens.

Den største forskel ses at være mellem ex-rygere og nuværende rygere, men det er ikke den forskel, vi søger. Forskellen mellem aldrig-rygere og nuværende rygere estimeres til 0.169 i aldrig-rygernes favør, med $CI=(-0.018, 0.356)$, og $P=0.077$, altså ikke signifikant.

Pga de uens varianser, checker vi dette resultat med et T-test, der kun involverer disse to grupper, og hvor vi får et konfidensinterval, der ikke er baseret på antagelsen om ens varianser:

```
proc ttest plots=all data=veter;
  where ryger in ("current","never");
class ryger;
var fev1;
run;
```

som giver os nedenstående output:

The TTEST Procedure Variable: FEV1

ryger	N	Mean	Std Dev	Std Err	Minimum	Maximum
current	139	4.0592	0.5904	0.0501	2.2600	5.3900
never	71	4.2280	0.7619	0.0904	1.8500	5.6500
Diff (1-2)		-0.1688	0.6532	0.0953		

ryger	Method	Mean	95% CL Mean		Std Dev
current		4.0592	3.9602	4.1582	0.5904
never		4.2280	4.0477	4.4084	0.7619
Diff (1-2)	Pooled	-0.1688	-0.3567	0.0190	0.6532
Diff (1-2)	Satterthwaite	-0.1688	-0.3736	0.0359	

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	208	-1.77	0.0779
Satterthwaite	Unequal	114.08	-1.63	0.1052

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	70	138	1.67	0.0114

Igen ser vi, at varianserne dømmes forskellige ($P=0.011$). Forskellen mellem aldrig-rygere og nuværende rygere estimeres her igen til 0.169 i aldrig-rygernes favør, men nu med et lidt bredere konfidensinterval $CI=(-0.035, 0.374)$, og $P=0.11$, altså endnu mindre i retning af signifikans.

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

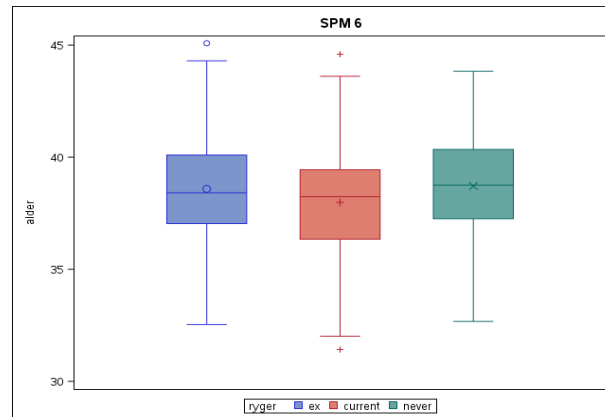
Man er vist generelt enige om, at lungefunktionen falder med alderen, så hvis de 3 rygergrupper udviser forskelle i aldersfordelingen, så er alder en confounder for sammenligningen af lungefunktionen i de tre rygergrupper.

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

```
proc sgplot data=veter;
    vbox alder / group=ryger;
run;

proc means N mean stddev data=veter;
    class ryger;
    var alder;
run;
```

Vi får så plottet



og de tilhørende summariske størrelser:

The MEANS Procedure

Analysis Variable : alder

ryger	N Obs	N	Mean	Std Dev
current	139	139	37.9874446	2.5652322
ex	89	89	38.5861167	2.4989561
never	71	71	38.7127532	2.1243819

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

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

```
proc glm plots=all data=vet;  
  class ryger;  
  model alder=ryger / solution clparm;  
  means ryger / hovtest welch;  
run;
```

som giver outputtet (kraftigt beskåret):

The GLM Procedure

Class Level Information

Class	Levels	Values
ryger	3	current ex never

R-Square	Coeff Var	Root MSE	alder Mean
0.018012	6.384804	2.447798	38.33788

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ryger	2	32.53182843	16.26591421	2.71	0.0679

Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		38.71275323 B	0.29050021	133.26	<.0001
ryger	current	-0.72530867 B	0.35706617	-2.03	0.0431
ryger	ex	-0.12663656 B	0.38950357	-0.33	0.7453
ryger	never	0.00000000 B	.	.	.

Parameter		95% Confidence Limits	
Intercept		38.14104570	39.28446076
ryger	current	-1.42801871	-0.02259862
ryger	ex	-0.89318376	0.63991064
ryger	never	.	.

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
ryger	2	211.5	105.8	1.45	0.2368
Error	296	21625.6	73.0595		

Nu er det bare ikke en evt signifikant forskel på alderen i de tre grupper, der er afgørende, men derimod den faktiske forskel (som altså var ca 1 år) samt alderens effekt på outcome (`fev1`). Denne effekt ser vi nu på, først for hele materialet under et. Her vælger jeg samtidig at estimere effekten af 5 år mere (selv om jeg i opgaven foreslog det lidt nemmere, nemlig for 10 år mere).

```
proc glm plots=all data=vet;
  model fev1=alder / solution clparm;
  estimate "effekt af 5 år" alder 5;
run;
```

som giver outputtet (beskåret)

The GLM Procedure

Dependent Variable: FEV1

R-Square	Coeff Var	Root MSE	FEV1 Mean
0.015862	15.70789	0.657483	4.185686

Parameter	Estimate	Standard Error	t Value	Pr > t
effekt af 5 år	-0.16924509	0.07735455	-2.19	0.0295

Parameter	95% Confidence Limits	
effekt af 5 år	-0.32147757	-0.01701261

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	5.483385049	0.59433934	9.23	<.0001
alder	-0.033849018	0.01547091	-2.19	0.0295

Parameter	95% Confidence Limits	
Intercept	4.313735009	6.653035088
alder	-0.064295514	-0.003402522

Fra dette output ses fev1 at falde med 0.03385 pr. år, eller 0.169 over en 5-årig periode. Konfidensintervallet for dette sidste estimat ses at være CI=(0.017, 0.321).

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

```
proc sort data=a1; by ryger;
run;

proc glm plots=all data=vet; by ryger;
    model fev1=alder / solution clparm;
    estimate "effekt af 5 år" alder 5;
run;
```

får vi dette output (stærkt beskåret)

```
ryger=current
```

R-Square	Coeff Var	Root MSE	FEV1 Mean
0.033671	14.35007	0.582499	4.059209

Parameter	Estimate	Standard Error	t Value	Pr > t
effekt af 5 år	-0.21116728	0.09664943	-2.18	0.0306

Parameter	95% Confidence Limits	
effekt af 5 år	-0.40228489	-0.02004968

```
ryger=ex
```

R-Square	Coeff Var	Root MSE	FEV1 Mean
0.023236	14.81407	0.644329	4.349438

```

Parameter              Estimate      Standard      t Value      Pr > |t|
      effekt af 5 år      -0.19770539      0.13742871      -1.44      0.1539

Parameter              95% Confidence Limits
      effekt af 5 år      -0.47085984      0.07544906

```

ryger=never

```

R-Square      Coeff Var      Root MSE      FEV1 Mean
0.011139      18.04966      0.763145      4.228028

```

```

Parameter              Estimate      Standard      t Value      Pr > |t|
      effekt af 5 år      -0.18926362      0.21468178      -0.88      0.3811

Parameter              95% Confidence Limits
      effekt af 5 år      -0.61754205      0.23901480

```

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

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

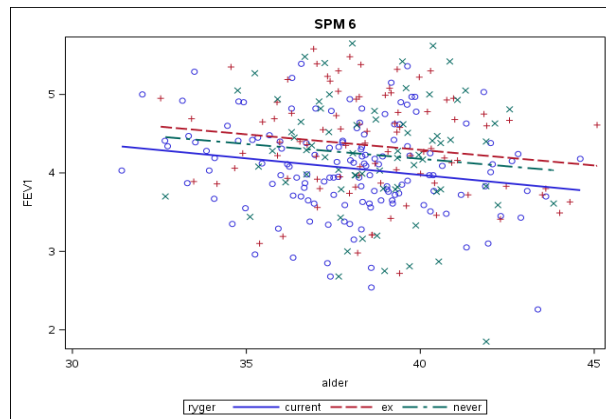
Disse estimater ser umiddelbart ikke særligt forskellige ud, og det, at kun estimatet for nuværende rygere er signifikant forskellig fra 0, kan ikke tages som udtryk for forskellighed. Godt nok er den estimerede linie stejlest i denne gruppe, men signifikansen beror mest på det store antal i denne gruppe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

En figur med de 3 linier indtegnet laves ved:

```

proc sgplot data=veter;
    reg x=alder y=fev1 / group=ryger;
run;

```

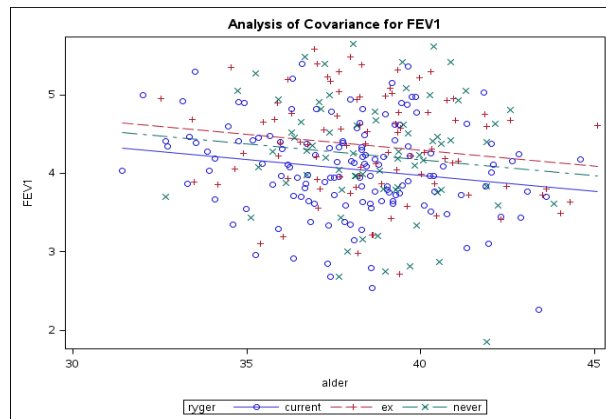
hvor det på øjemål er temmelig svært at se, at linierne ikke er parallelle.

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

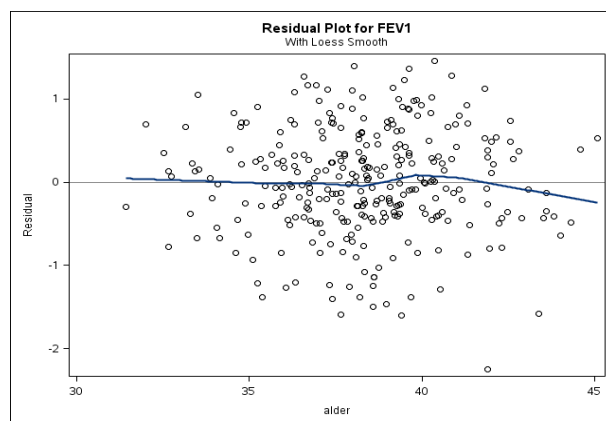
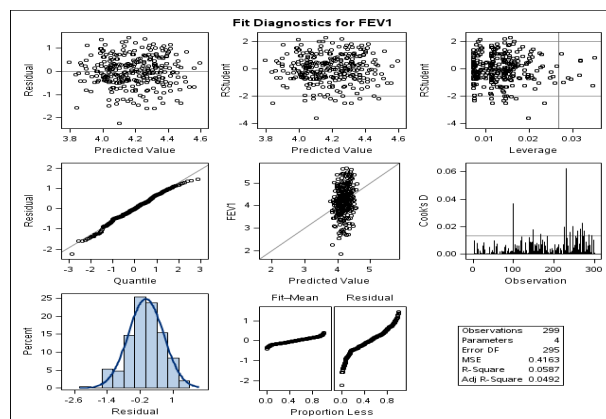
Vi skal nu undersøge effekten af rygning, korrigeret for den lille aldersforskel, eller helt præcist: forskellen på nuværende rygere og aldrig-rygere, på samme alder. Vi gør dette ved hjælp af en kovariansanalyse, med dertil hørende illustrationer, og vi inddrager *ikke* interaktion i denne model, da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

```
proc glm plots=(DiagnosticsPanel Residuals(smooth)) data=vet;
  class ryger;
  model fev1=alder ryger / solution clparm;
  estimate "effekt af 5 år" alder 5;
run;
```

Figuren, der illustrerer modellen er



og modelkontrollen giver figurerne



som ikke giver anledning til bekymring.

Vi ser derfor på output fra kovariansanalysen:

The GLM Procedure

Class Level Information

Class	Levels	Values
ryger	3	current ex never

Number of Observations Read 299
Number of Observations Used 299

Dependent Variable: FEV1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7.6634943	2.5544981	6.14	0.0005
Error	295	122.7940402	0.4162510		
Corrected Total	298	130.4575344			

R-Square	Coeff Var	Root MSE	FEV1 Mean
0.058743	15.41385	0.645175	4.185686

Source	DF	Type III SS	Mean Square	F Value	Pr > F
alder	1	2.92616862	2.92616862	7.03	0.0084
ryger	2	5.59416732	2.79708366	6.72	0.0014

Parameter	Estimate	Standard Error	t Value	Pr > t
effekt af 5 år	-0.20309452	0.07659954	-2.65	0.0084

Parameter	95% Confidence Limits	
effekt af 5 år	-0.35384533	-0.05234370

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	5.800497740 B	0.59799800	9.70	<.0001
alder	-0.040618903	0.01531991	-2.65	0.0084
ryger current	-0.198280778 B	0.09476693	-2.09	0.0373
ryger ex	0.116266195 B	0.10268121	1.13	0.2584
ryger never	0.000000000 B	.	.	.

Parameter	95% Confidence Limits	
Intercept	4.623614881	6.977380600
alder	-0.070769066	-0.010468740
ryger current	-0.384785702	-0.011775855
ryger ex	-0.085814342	0.318346732
ryger never	.	.

Vi bemærker, at der nu ses en kraftigere signifikant forskel på de 3 rygergrupper (P=0.0014 nu mod P=0.0042 eller 0.0030 i spørgsmål 5), og det skyldes dels, at nuværende rygere var lidt yngre end de andre og dels at alderen har reduceret residualspreddningen, så standard errors på parameterestimaterne er blevet mindre.

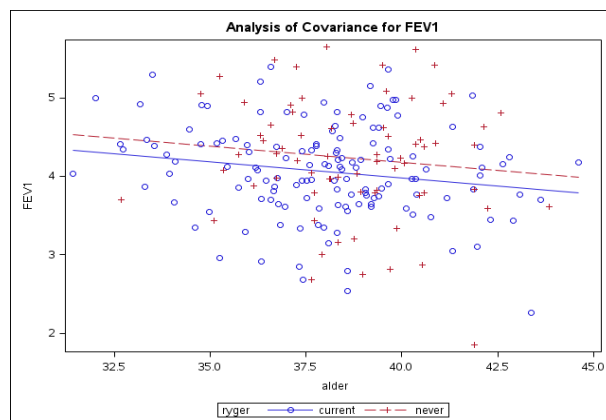
Det reviderede estimat for forskellen mellem aldrig-rygere og rygere ses at være 0.198, med konfidensinterval CI=(0.012, 0.385), P=0.037.

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

Hvis vi her vælger kun at se på de to berørte grupper, benytter vi en `where`-sætning:

```
proc glm plots=all data=vet;
  where ryger in ("current","never");
  class ryger;
  model fev1=alder ryger / solution clparm;
  estimate "effekt af 5 år" alder 5;
run;
```

og finder så figuren



med det tilhørende output

The GLM Procedure

Class Level Information		
Class	Levels	Values
ryger	2	current never

Number of Observations Read	210
Number of Observations Used	210

Dependent Variable: FEV1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3.40725437	1.70362719	4.07	0.0185
Error	207	86.67424848	0.41871618		
Corrected Total	209	90.08150286			

R-Square	Coeff Var	Root MSE	FEV1 Mean
0.037824	15.72007	0.647083	4.116286

Source	DF	Type III SS	Mean Square	F Value	Pr > F
alder	1	2.06788841	2.06788841	4.94	0.0273
ryger	1	1.81746742	1.81746742	4.34	0.0384

Parameter	Estimate	Standard Error	t Value	Pr > t
effekt af 5 år	-0.20551406	0.09247788	-2.22	0.0273

Parameter	95% Confidence Limits	
effekt af 5 år	-0.38783332	-0.02319481

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	5.819231214 B	0.72012112	8.08	<.0001
alder	-0.041102813	0.01849558	-2.22	0.0273
ryger current	-0.198631762 B	0.09534002	-2.08	0.0384
ryger never	0.000000000 B	.	.	.

Parameter	95% Confidence Limits	
Intercept	4.399519354	7.238943075
alder	-0.077566663	-0.004638962
ryger current	-0.386593700	-0.010669825
ryger never	.	.

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

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

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

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

Som det også tidligere er bemærket, vurderes effekten af rygning til at være

lidt større, når der korrigeres for alder, fordi vi i første omgang sammenlignede nogle lidt yngre rygere med nogle lidt ældre ikke-rygere og derfor fik maskeret lidt af effekten.

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

```
proc glm plots=all data=veter; where ryger in ("current","never");
    class ryger;
    model fev1=alder ryger ryger*alder / solution clparm;
run;
```

Herved finder vi:

The GLM Procedure

Class Level Information

Class	Levels	Values
ryger	2	current never

Number of Observations Read	210
Number of Observations Used	210

Dependent Variable: FEV1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3.41175222	1.13725074	2.70	0.0466
Error	206	86.66975063	0.42072694		
Corrected Total	209	90.08150286			

R-Square	Coeff Var	Root MSE	FEV1 Mean
0.037874	15.75777	0.648635	4.116286

Source	DF	Type III SS	Mean Square	F Value	Pr > F
alder	1	1.50323483	1.50323483	3.57	0.0601
ryger	1	0.00014033	0.00014033	0.00	0.9854
alder*ryger	1	0.00449785	0.00449785	0.01	0.9177

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	5.693411361 B	1.41486937	4.02	<.0001
alder	-0.037852725 B	0.03649375	-1.04	0.3008
ryger current	-0.029861631 B	1.63507001	-0.02	0.9854
ryger never	0.000000000 B	.	.	.
alder*ryger current	-0.004380732 B	0.04236862	-0.10	0.9177
alder*ryger never	0.000000000 B	.	.	.

Parameter		95% Confidence Limits	
Intercept		2.903930363	8.482892359
alder		-0.109801866	0.034096416
ryger	current	-3.253478459	3.193755197
ryger	never	.	.
alder*ryger	current	-0.087912447	0.079150983
alder*ryger	never	.	.

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