

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:

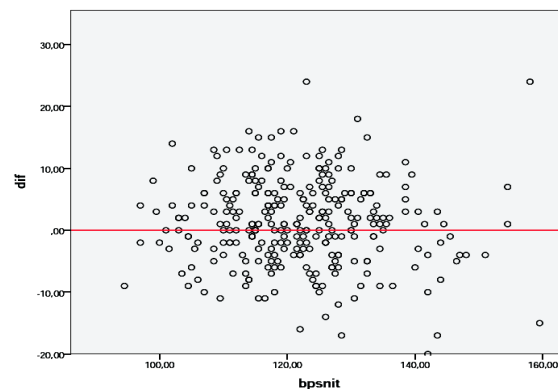
1. *Er det rimeligt at vurdere forskellen på blodtrykket på højre og venstre arm ud fra differenser på selve blodtryks-skalaen? Eller skal forskellen måske hellere ses relativt til niveauet?*

Først indlæses data, og derefter benytter vi **Transform/Compute** til at definere differensen af blodtrykket mellem de to arme (højre minus venstre), og ligeledes gennemsnittet af disse.

```
dif=bpright-bpleft  
bpsnit=(bpright+bpleft)/2
```

Derefter tegner vi et Bland-Altman plot ved at gå ind i **Graph/Chart Builder/Scatter** og i den fremkomne boks trækkes **dif** over på Y-aksen, og **bpsnit** over på X-aksen. Efterfølgende dobbeltklikkes på scatterplottet, man klikker **Add a reference line to Y-axis** og specificerer den til at være rød med Position 0.

Herved får vi 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 er dog 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.

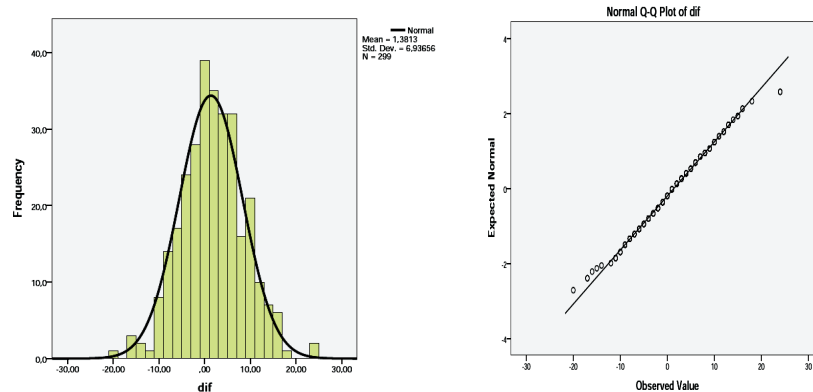
Bemærk, hvad der *ikke* spørges om i dette spørgsmål: Fordelingen af blodtryk på hver af siderne, og egentlig heller ikke noget om fordelingen af differenserne før i spørgsmål 2 (og endnu mere i spørgsmål 3).

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 i form af histogram (benyt **Graph/ChartBuilder/Histogram** og dobbeltklik efterfølgende på grafen for at vælge **Show Distribution Curve/Normal**) og fraktildiagram (benyt f.eks. **Analyze/Descriptive Statistics/Q-Q Plots** eller (som her) **Analyze/Descriptive Statistics/Explore**) for differenserne.



Modelkontrollen går på, at differenserne skal være rimeligt normalfordelte, og disse figurer giver på ingen måde anledning til bekymring.

Det parrede T-test udføres i **Analyze/Compare Means/Paired Samples T-test**, hvor man markerer både **bpright** og **bpleft** samtidig og fører dem over til **Pair1**.

Selve T-testet giver outputtet:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	bpright	122,45	299	11,564	,669
	bpleft	121,07	299	11,994	,694

Paired Differences					
			95% Confidence Interval of the Difference		
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper
Pair 1 bpright - bpleft	1,381	6,937	,401	,592	2,171

Paired Samples Test

	t	df	Sig. (2-tailed)
Pair 1 bpright - bpleft	3,443	298	,001

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 mmm (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

$$1.381 \pm 2 \times 6.937 \approx (-12.5, 15.3)$$

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 i

Analyze/Descriptive Statistics/Frequencies, hvor vi sætter dif over i Variable(s), fjerner fluebenet i

Display Frequency Tables, og under Statistics vælges Percentiles, hvor vi skriver 2,5 og klikker Add, skriver 5 og klikker Add, skriver 95 og klikker Add og til sidst skriver 97,5 og klikker Add/Continue

Herved får vi outputtet:

Statistics		
dif	N	299
	Valid	
	Missing	0
Percentiles	2,5	-11,5000
	5	-10,0000
	95	13,0000
	97,5	15,0000

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

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

4. Lav en gruppering af personerne, baseret på, om den numeriske differens på blodtrykket overstiger 10 mm eller ej (kald f.eks. 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. Koden for dette ser således ud:

```
RECODE alkuge (10 thru Highest=1) (ELSE=0) INTO dranker.
EXECUTE.
COMPUTE absdif=abs(dif).
EXECUTE.
STRING sideforskel (A8).
RECODE absdif (10 thru Highest='stor') (ELSE='lille') INTO sideforskel
```

og er dannet ved at benytte Transform/Recode into different variable, hvor alkuge sættes over i Numeric Variable, dranker sættes i Output Variable, hvorefter man i Old and new values klikker Range: value through HIGHEST, skriver 10, og giver den værdien 1. Herefter klikkes All other values, som gives værdien 0.

Når man skal definere variabelen **sideforskel**, skal man huske at afkrydse i Output variables are strings, men ellers er princippet som ovenfor.

(Bemærk, at jeg i 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).

Nu skal vi se 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, så vi går ind i Analyze/Descriptive Statistics/Crosstabs/, hvor dranker sættes over i Row(s) og sideforskel i Column(s).

I Cells afkrydses Observed og Percentages: Row, og i Statistics afkrydses Risk.

Herved får vi outputtet

dranker * sideforskel Crosstabulation

		sideforskel		Total
		lille	stor	
dranker	,00	Count	187	44
		% within drinker	81,0%	19,0%
	1,00	Count	59	9
		% within drinker	86,8%	13,2%
Total	Count		246	53
	% within drinker		82,3%	17,7%

Chi-Square Tests

	Value	df	Asymptotic Significance (2- sided)	Exact Sig. (2- sided)	Exact Sig. (1- sided)
Pearson Chi-Square	1,217 ^a	1	,270		
Continuity Correction ^b	,851	1	,356		
Likelihood Ratio	1,284	1	,257		
Fisher's Exact Test				,366	,179
N of Valid Cases	299				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 12,05.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for drinker (,00 / 1,00)	,648	,299	1,406
For cohort sideforskel = lille	,933	,834	1,044
For cohort sideforskel = stor	1,439	,741	2,796
N of Valid Cases	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 19.0% for folk, der ikke er drankere, men kun 13.2% 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 1.439, hvilket betyder, at risikoen for en stor sideforskel er næsten 44% større hos ikke-drankere i forhold til hos drankere. Men sikkerhedsintervallet er (0.741, 2.796), dvs. den relative risiko kunne også være ned til 0.741, hvilket svarer til en større risiko for drankere på op til $\frac{1}{0.741} = 1.35$, 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.

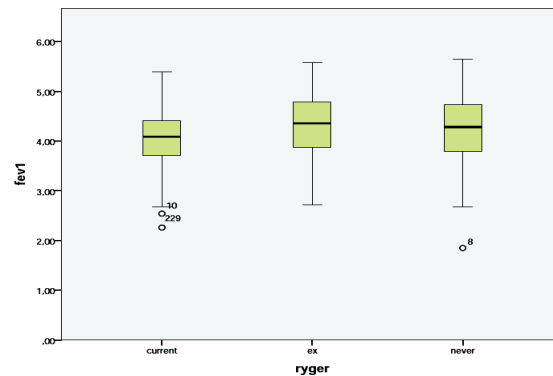
Nu kommer nogle spørgsmål, der fokuserer på lungefunktionen:

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

I første omgang ser vi på et Boxplot af lungefunktionen, opdelt på de 3 grupper. Hertil kan benyttes **Analyze/Descriptive Statistics/Explore**, hvor vi sætter **fev1** i **Dependent List**, **ryger** i **Factor List** samt sætter hak i **Plots**.

Her er dog brugt **Graphs/Graph Builder/Boxplots**, hvor **fev1** trækkes over på Y-aksen, og **ryger** trækkes over på X-aksen.

Herved får vi 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 for hver rygergruppe for sig, så vi går først ind i **Data/Split file/Compare Groups**, hvor **ryger** sættes ind i feltet **Groups Based on**.

Herefter går man ind under menuen **Analyze/Descriptive Statistics**, vælger **Descriptives** og flytter de ønskede variable (her blot **fev1**) over i **Variable(s)**.

Under **Statistics** kan man så sætte flueben ved de størrelser, vi ønsker udregnet (her **Mean**, **Std. deviation**, **Minimum** og **Maximum**).

Herved får vi outputtet:

Descriptive Statistics						
ryger		N	Minimum	Maximum	Mean	Std. Deviation
current	fev1	139	2,26	5,39	4,0592	,59041
	Valid N (listwise)	139				
ex	fev1	89	2,72	5,58	4,3494	,64823
	Valid N (listwise)	89				
never	fev1	71	1,85	5,65	4,2280	,76193
	Valid N (listwise)	71				

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.

Først skal vi dog lige huske at afblæse **Split File**.

Vi bruger **Analyze/General Linear Model/Univariate**, hvor vi sætter fev1 i **Dependent Variable** og ryger i **Fixed Factor(s)**.

Her kan vi vælge at se parameterestimerer under **Options**, hvor vi afkrydser **Parameter estimates**.

Herved får vi

Descriptive Statistics			
Dependent Variable: fev1			
ryger	Mean	Std. Deviation	N
current	4,0592	,59041	139
ex	4,3494	,64823	89
never	4,2280	,76193	71
Total	4,1857	,66165	299

Levene's Test of Equality of Error Variances ^a			
Dependent Variable: fev1			
F	df1	df2	Sig.
2,684	2	296	,070

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + ryger

Tests of Between-Subjects Effects					
Dependent Variable: fev1					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4,737 ^a	2	2,369	5,577	,004
Intercept	4911,179	1	4911,179	11563,050	,000
ryger	4,737	2	2,369	5,577	,004
Error	125,720	296	,425		
Total	5368,927	299			
Corrected Total	130,458	298			

a. R Squared = ,036 (Adjusted R Squared = ,030)

Parameter Estimates

Dependent Variable: fev1

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4,228	,077	54,665	,000	4,076	4,380
[ryger=current]	-,169	,095	-1,776	,077	-,356	,018
[ryger=ex]	,121	,104	1,171	,243	-,083	,325
[ryger=never]	0 ^a	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Vi bemærker i outputtet ovenfor, at Levenes test for varianshomogenitet netop bliver godkendt ($P=0.070$), så variansanalysen er rimelig pålidelig (SAS giver dog et andet resultat her, men da sample size er ret stor, betyder en lav P-værdi ikke meget her).

Testet for identitet af de 3 middelværdier giver i ANOVA $P=0.004$, og den største forskel ses at være mellem ex-rygere og nuværende rygere, men det er ikke den forskel, vi søger. Forskellen mellem aldrig-rygere og nuværende rygere estimeres til 0.169 i aldrig-rygernes favør, med $CI=(-0.018, 0.356)$, og $P=0.077$, altså ikke signifikant.

Vi kunne også have nøjedes med at se på de to involverede grupper og skulle da blot have foretaget et T-test, hvor vi så ovenikøbet får et konfidensinterval, der ikke er baseret på antagelsen om ens varianser:

Vi vælger **Analyze/Compare Means/Independent Samples T-test**, hvor vi sætter fev1 over i **Test Variable(s)** og ryger i **Grouping Variable**.

Under **Define Groups** vælges Group1 til current og Group2 til never.

Herved får vi

Independent Samples Test					
		Levene's Test for Equality of Variances		t-test for Equality of Means	
		F	Sig.	t	df
fev1	Equal variances assumed	4,648	,032	-1,772	208
	Equal variances not assumed			-1,633	114,080

Independent Samples Test					
		t-test for Equality of Means			
		Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence ... Lower
fev1	Equal variances assumed	,078	-,16882	,09528	-,35666
	Equal variances not assumed	,105	-,16882	,10337	-,37358

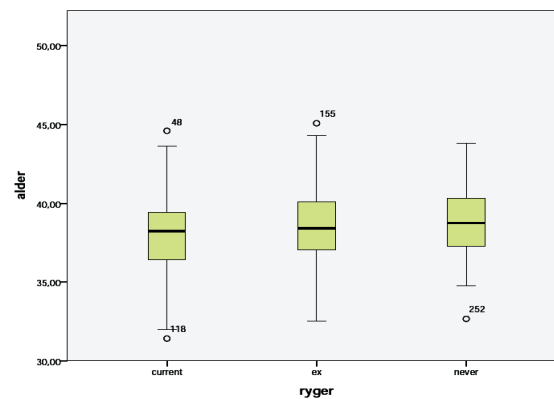
Independent Samples Test		
		t-test for Equality of Means
		95% Confidence Interval of the ...
		Upper
fev1	Equal variances assumed	,01902
	Equal variances not assumed	,03594

Vi bemærker, at varianserne dømmes forskellige ($P=0.032$). Forskellen mellem aldrig-rygere og nuværende rygere estimeres her igen til 0.169 i aldrig-rygernes favør, men nu med et lidt bredere konfidensinterval $CI=(-0.036, 0.374)$, og $P=0.11$, altså endnu mindre i retning af signifikans.

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

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

Vi laver derfor et Boxplot af aldersfordelingerne samt udregner nogle summariske mål for alderen i de tre grupper, helt på samme måde som vi ovenfor gjorde det for fev1, og vi finder så:



og de tilhørende summariske størrelser:

Descriptive Statistics						
ryger		N	Minimum	Maximum	Mean	Std. Deviation
current	alder	139	31,42	44,60	37,9870	2,56514
	Valid N (listwise)	139				
ex	alder	89	32,54	45,08	38,5863	2,49873
	Valid N (listwise)	89				
never	alder	71	32,67	43,83	38,7128	2,12474
	Valid N (listwise)	71				

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:

Tests of Between-Subjects Effects

Dependent Variable: alder

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	32,588 ^a	2	16,294	2,720	,068
Intercept	408765,242	1	408765,242	68223,766	,000
ryger	32,588	2	16,294	2,720	,068
Error	1773,495	296	5,992		
Total	441270,667	299			
Corrected Total	1806,083	298			

a. R Squared = ,018 (Adjusted R Squared = ,011)

Parameter Estimates

Dependent Variable: alder

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	38,713	,290	133,265	,000	38,141	39,285
[ryger=current]	-,726	,357	-2,033	,043	-1,429	-,023
[ryger=ex]	-,127	,389	-,325	,746	-,893	,640
[ryger=never]	0 ^a	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

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, idet vi udfører en simpel lineær regression ved at

Tests of Between-Subjects Effects

Dependent Variable: fev1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2,071 ^a	1	2,071	4,792	,029
Intercept	36,805	1	36,805	85,142	,000
alder	2,071	1	2,071	4,792	,029
Error	128,386	297	,432		
Total	5368,927	299			
Corrected Total	130,458	298			

a. R Squared = ,016 (Adjusted R Squared = ,013)

Parameter Estimates

Dependent Variable: fev1

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5,484	,594	9,227	,000	4,314	6,654
alder	-,034	,015	-2,189	,029	-,064	-,003

Fra dette output ses **fev1** at falde med 0.034 pr. år, eller 0.170 over en 5-årig periode. Konfidensintervallet for dette sidste estimat ses at være CI=(0.015, 0.320).

Bemærk, at jeg har valgt at angive effekten af 5 år mere (selv om jeg i opgaven foreslog det lidt nemmere, nemlig for 10 år mere).

Nu gør vi det samme for hver af rygergrupperne for sig (da analysen

ovenfor jo helt negligerer effekten af rygning). Det gøres blot ved at køre en **Split File** på ryger inden analysen gentages. Så får vi:

Tests of Between-Subjects Effects

Dependent Variable: fev1

ryger	Source	Type III Sum of Squares	df	Mean Square	F	Sig.
current	Corrected Model	1,622 ^a	1	1,622	4,780	,030
	Intercept	20,101	1	20,101	59,243	,000
	alder	1,622	1	1,622	4,780	,030
	Error	46,483	137	,339		
	Total	238,432	139			
	Corrected Total	48,105	138			
ex	Corrected Model	,860 ^b	1	,860	2,072	,154
	Intercept	12,689	1	12,689	30,566	,000
	alder	,860	1	,860	2,072	,154
	Error	36,118	87	,415		
	Total	1720,646	89			
	Corrected Total	36,978	88			
never	Corrected Model	,453 ^c	1	,453	,778	,381
	Intercept	6,817	1	6,817	11,705	,001
	alder	,453	1	,453	,778	,381
	Error	40,184	69	,582		
	Total	1309,849	71			
	Corrected Total	40,638	70			

a. R Squared = ,034 (Adjusted R Squared = ,027)
b. R Squared = ,023 (Adjusted R Squared = ,012)
c. R Squared = ,011 (Adjusted R Squared = -,003)

Parameter Estimates

Dependent Variable: fev1

ryger	Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
current	Intercept	5,665	,736	7,697	,000	4,209	7,120
	alder	-,042	,019	-2,186	,030	-,080	-,004
ex	Intercept	5,876	1,063	5,529	,000	3,764	7,989
	alder	-,040	,027	-1,439	,154	-,094	,015
never	Intercept	5,694	1,664	3,421	,001	2,374	9,014
	alder	-,038	,043	-,882	,381	-,124	,048

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	0.030
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	0.029

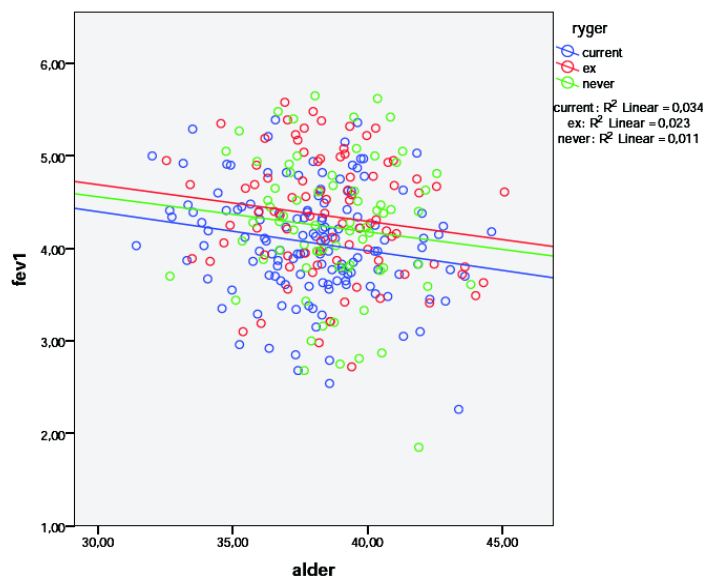
Disse estimater ser umiddelbart ikke særligt forskellige ud, og det, at kun estimatet for nuværende rygere er signifikant forskellig fra 0, kan ikke tages som udtryk for forskellighed. Godt nok er den estimerede linie stejlest i denne gruppe, men signifikansen beror mest på det store

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 at gå ind i **Graph/Chart Builder**, hvor man vælger **Scatter plot** (nr. 2 fra venstre), og i den fremkomne boks trækker man **fev1** over på Y-aksen, **alder** over på X-aksen og **ryger** over i **Set Color**

For at tegne individuelle regressionslinier dobbeltklikker man efterfølgende på grafen og klikker på ikonet **Add Fit Line at Subgroups** og derefter i **Properties**-boksen afkrydse **Linear** og klikke **Apply**.

Herved får vi figuren:



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

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

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

Her benyttes **Analyze/General Linear Model/Univariate**, hvor vi sætter fev1 over i **Dependent Variable**, ryger i **Fixed Factor(s)** og alder i **Covariate(s)**

For at undgå at få interaktionen med i modellen, klikkes nu på **Model**, hvorefter man vælger **Custom**, markerer begge kovariater, skifter fra **Interaction** til **Main Effects** og klikker på pilen og **Continue**.

Husk også i **Options** at vælge **Parameter Estimates**

Herved får vi outputtet:

Tests of Between-Subjects Effects					
Dependent Variable: fev1					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7,667 ^a	3	2,556	6,140	,000
Intercept	39,867	1	39,867	95,778	,000
alder	2,930	1	2,930	7,038	,008
ryger	5,596	2	2,798	6,722	,001
Error	122,791	295	,416		
Total	5368,927	299			
Corrected Total	130,458	298			

a. R Squared = ,059 (Adjusted R Squared = ,049)

Parameter Estimates						
Dependent Variable: fev1						
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5,801	,598	9,701	,000	4,625	6,978
alder	-,041	,015	-2,653	,008	-,071	-,010
[ryger=current]	-,198	,095	-2,093	,037	-,385	-,012
[ryger=ex]	,116	,103	1,132	,258	-,086	,318
[ryger=never]	0 ^a	.	.	.	.	.

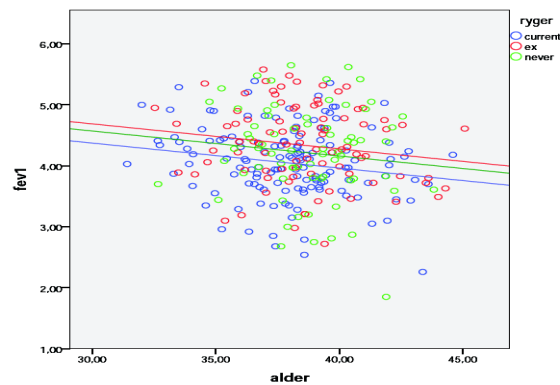
a. This parameter is set to zero because it is redundant.

Figuren, der illustrerer modellen er vanskelig at tegne, men det kan lade sig gøre. Først laves et scatterplottet, med symbol for de enkelte

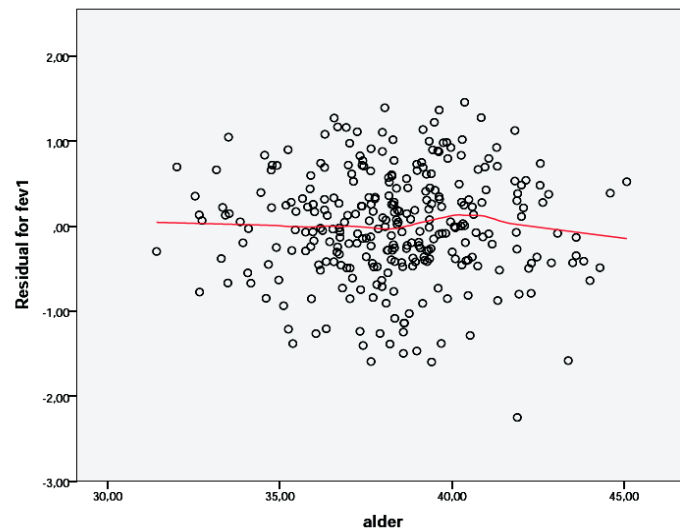
grupper, se ovenfor.

Herefter dobbeltklikker man på scatterplottet og vælger **Add a reference line from Equation**, hvorefter man skriver ligningen for den første linie, f.eks. svarende til referencegruppen for **ryger**, altså **never**: $y = -0,041 \cdot x + 5,801$ og trykker **Apply**. Herefter dobbeltklikker man igen for at tilføje den anden linie, nu for **current**: $y = -0,041 \cdot x + 5,603$ og til sidst for **ex**: $y = -0,041 \cdot x + 5,917$.

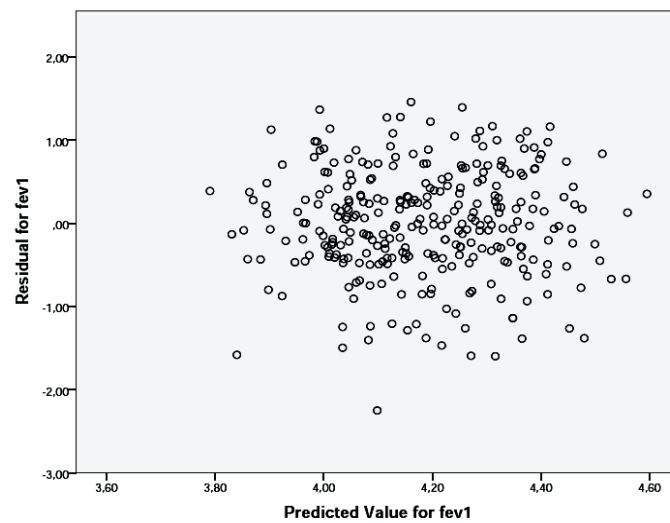
Vi får så figuren med 3 parallelle linier:



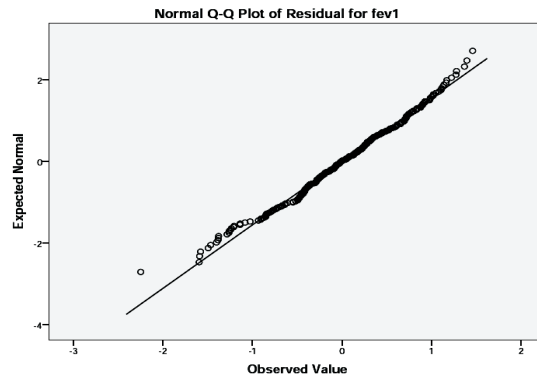
Inden vi giver os til at fortolke output fra kovariansanalysen, skal vi dog lige se på noget modelkontrol, som styres ved at gemme residualer, predikterede værdier og evt. andre størrelser under selve udførelsen af kovariansanalysen. Herefter kan vi tegne diverse figurer, f.eks. check af lineariteten i alder:



Varianshomogeniteten ved at plotte residualer mod predikterede værdier:



og endelig et fraktildiagram til at checke normalfordelingsantagelsen for residualerne:



Ingen af disse figurer giver anledning til bekymring, og vi går derfor over til at fortolke output fra kovariansanalysen ovenfor.

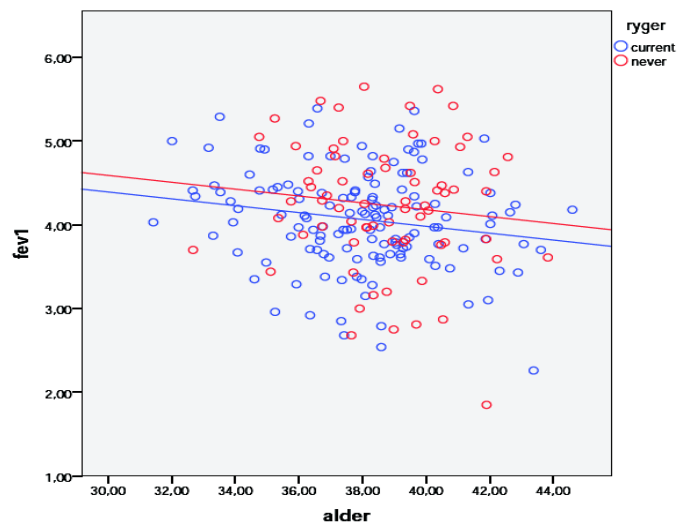
Her bemærker vi, at der nu ses en kraftigere signifikant forskel på de 3 rygergrupper ($P=0.001$ nu mod $P=0.0042$ i spørgsmål 5), og det skyldes dels, at nuværende rygere var lidt yngre end de andre og dels at alderen har reduceret residualspreddingen, så standard errors på parameterestimaterne er blevet 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 -0.041 (-0.071, -0.010), svarende til et fald på 0.205 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 **Data/Select Cases**, hvorefter vi gentager analysen.

Vi finder så figuren



med det tilhørende output

Tests of Between-Subjects Effects

Dependent Variable: fev1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3,410 ^a	2	1,705	4,072	,018
Intercept	27,119	1	27,119	64,770	,000
alder	2,071	1	2,071	4,945	,027
ryger	1,818	1	1,818	4,342	,038
Error	86,672	207	,419		
Total	3648,281	210			
Corrected Total	90,082	209			

a. R Squared = ,038 (Adjusted R Squared = ,029)

Parameter Estimates

Dependent Variable: fev1

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5,820	,720	8,083	,000	4,401	7,240
alder	-,041	,018	-2,224	,027	-,078	-,005
[ryger=current]	-,199	,095	-2,084	,038	-,387	-,011
[ryger=never]	0 ^a	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

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 poollet fra de to udvalgte rygergrupper er -0.041, CI=(-0.078, -0.0045), svarende til et fald på

0.205 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:

Tests of Between-Subjects Effects

Dependent Variable: fev1

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3,414 ^a	3	1,138	2,705	,046
Intercept	20,309	1	20,309	48,273	,000
ryger	,000	1	,000	,000	,986
alder	1,505	1	1,505	3,578	,060
ryger * alder	,005	1	,005	,011	,918
Error	86,667	206	,421		
Total	3648,281	210			
Corrected Total	90,082	209			

a. R Squared = ,038 (Adjusted R Squared = ,024)

Parameter Estimates

Dependent Variable: fev1

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5,694	1,415	4,025	,000	2,905	8,483
[ryger=current]	-,030	1,635	-,018	,986	-3,253	3,194
[ryger=never]	0 ^a	.	.	.	.	.
alder	-,038	,036	-1,038	,301	-,110	,034
[ryger=current] * alder	-,004	,042	-,104	,918	-,088	,079
[ryger=never] * alder	0 ^a	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

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