Faculty of Health Sciences

Basal statistik

Den generelle lineære model mv., i SPSS

Lene Theil Skovgaard

19. oktober 2020

Den generelle lineære model mv.

- Ikke-lineære sammenhænge
- Opbygning af modeller
- Sammenligning af modeller
- Endnu et eksempel

E-mail: ltsk@sund.ku.dk

Terminologi

for kvantitativt outcome, f.eks. vitamin D

Regression: Kovariaterne er også kvantitative

- Simpel (lineær) regression:
  kun en enkelt kovariat
- Multipel (lineær) regression:
  to eller flere kovariater

Variansanalyse: Kovariaterne er kategoriske
(grupper, `class`-variable, faktorer)

- Ensidet variansanalyse: kun en enkelt kovariat
- Tosidet variansanalyse: to kovariater

Generel lineær model: Begge typer kovariater i samme model

- Kovariansanalyse:
  Netop en kvantitativ og en kategorisk kovariat

Forklarende variable = Kovariater

Outcome	Dikotom	Kategorisk	Kvantitativ	Kategoriske og kvantitative
Dikotom	2*2-tabeller	χ^2-test	Logistisk regression	
parret	Mc Nemar	svært, mixed models	Mixed models	
Kategorisk	Kontingenstabeller/χ^2-test		Generaliseret logistisk regression	
Ordinale	svært, f.eks. proportional odds modeller			
Kvantitativ	Mann-Whitney	Kruskal-Wallis	Robust multipel	
parret	Wilcoxon signed rank	Friedman	regression	
Normalfordelte	T-test	Variansanalyse	Multipel regression	Kovariansanalyse
residualer	uparret/parret	ensidet/tosidet	**Den generelle lineære model**	
Censureret	Log-rank test		Cox regression	
Korrelerede	Varianskomponent-		Modeller for	
kvantitative	modeller		gentagne målinger	
Nf. residualer	Mixed models			

Den generelle lineære model

Outcome: Kvantitativ variabel Y

Kovariater:

- Kategoriske (`class`):
  Fortolkning af parameter:
  Forskel fra aktuel gruppe til referencegruppe,
  for fastholdt værdi af alle andre kovariater.
- Kvantitative:
  Her antages linearitet
  Fortolkning af parameter:
  1 enheds ændring i X svarer til
  β enheders ændring i Y,
  for fastholdt værdi af alle andre kovariater.

Linearitet

Skal alt så kunne beskrives ved hjælp af linier?

Nej, fordi:

- Man kan transformere en eller flere af de indgående variable
  Eksempel: Biokemisk iltforbrug (s. 7-17)
- Man kan benytte polynomier ved at tilføje kovariater i forskellige potenser (s. 23-26)
  bruges dog mest som modelcheck
- Tilføje en kovariat, der er relateret til den oprindelige kovariat, f.eks. logaritmetransformationen
- Man kan lave stykvise lineære funktioner, kaldet lineære splines
  Eksempel: Væksthormon (s. 27-34)

Biologisk iltforbrug

Iltsvind i lukkede flasker (`boc`, **b**iochemical **o**xygen **c**onsumption), som funktion af antal dage (`days`)

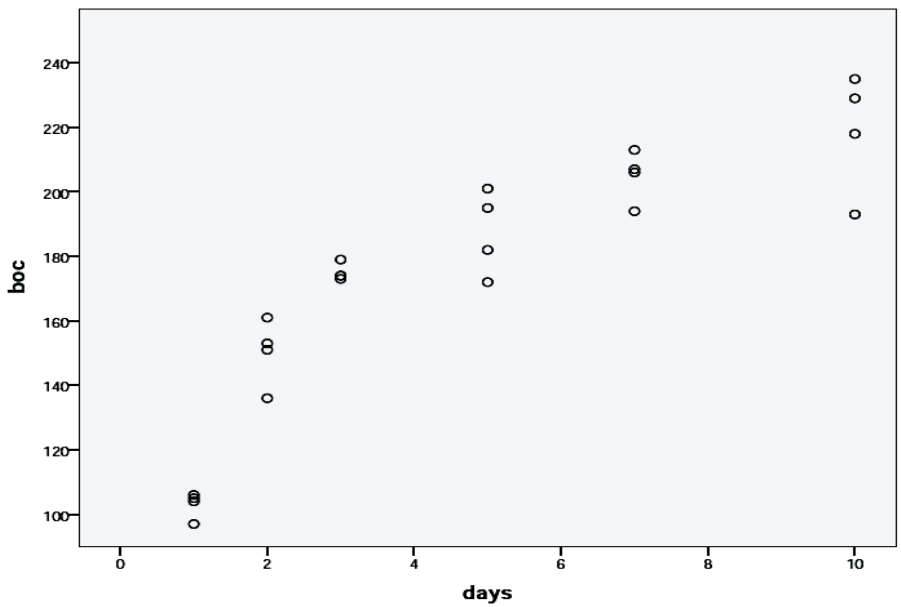
4 flasker til hvert tidspunkt

days				
1	105	97	104	106
2	136	161	151	153
3	173	179	174	174
5	195	182	201	172
7	207	194	206	213
10	218	193	235	229

Vedr. omstrukturering af data, samt figur næste side: se s. 93-94

Illustration af iltsvind

Sammenhængen mellem iltsvind (boc) og antallet af dage (days) ses at være ikke-lineær.



Vi ønsker at bestemme *asymptoten*, dvs. iltsvindet efter *lang* tid, dvs. når tiden går mod uendelig (∞)

Transformation til linearitet

Biologerne hævder at vide, at iltsvindet kan beskrives ved funktionen

$$\texttt{boc} = \gamma \exp(\text{-}\beta/\texttt{days})$$

Denne relation er klart ikke-lineær, men den kan transformeres til linearitet ved brug af den (naturlige) logaritme:

$$\log(\texttt{boc}) = \log(\gamma) \text{ - } \beta/\texttt{days}$$

Vi vil gerne bestemme

$$\texttt{boc}(\infty) = \gamma \quad \exp(0) = \gamma$$

Omparametrisering

Med definitionerne

Outcome: $y = \texttt{logboc} = \log(\texttt{boc})$

Kovariat: $x = \texttt{invdays} = 1/\texttt{days}$

Intercept: $\alpha = \log(\gamma)$

kan vi skrive ligningen som

$$y = \alpha - \beta x$$

altså en **lineær relation**, bare med en negativ hældning ($-\beta$)

Den lineære regression

Variabeldefinitioner og analyse ses s. 95-96
Bemærk, at vi her har benyttet den *naturlige* logaritme til at transformere iltforbruget (dette valg kommenteres s. 19)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,976[a]	,952	,950	,05845

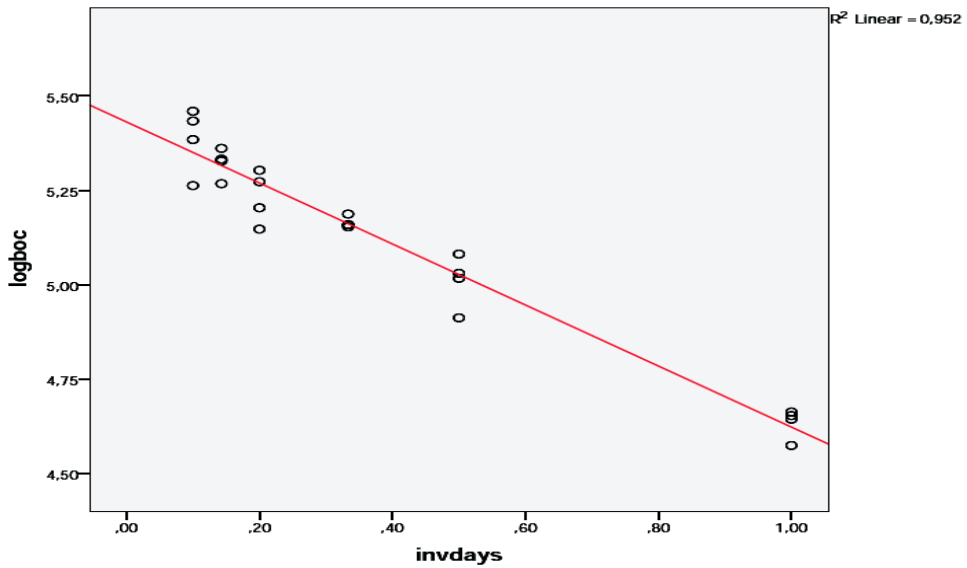
a. Predictors: (Constant), invdays

b. Dependent Variable: logboc

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	5,431	,019		286,761	,000	5,392	5,471
	invdays	-,808	,039	-,976	-20,833	,000	-,888	-,727

Den transformerede relation

Se s. 97



Dette plot ser pænt lineært ud, med rimelig varianshomogenitet.
Vi finder:

$$\texttt{logboc} = 5.431 - 0.808 \times \texttt{invdays}$$

Fortolkning af resultaterne

Den lineære regressionsmodel giver os estimaterne (fra s. 11):

$$\begin{aligned} \text{intercept} &: \hat{\alpha} = \log(\hat{\gamma}) = 5.431(0.019) \\ \text{slope} &: \hat{\beta} = -0.808(0.039) \end{aligned}$$

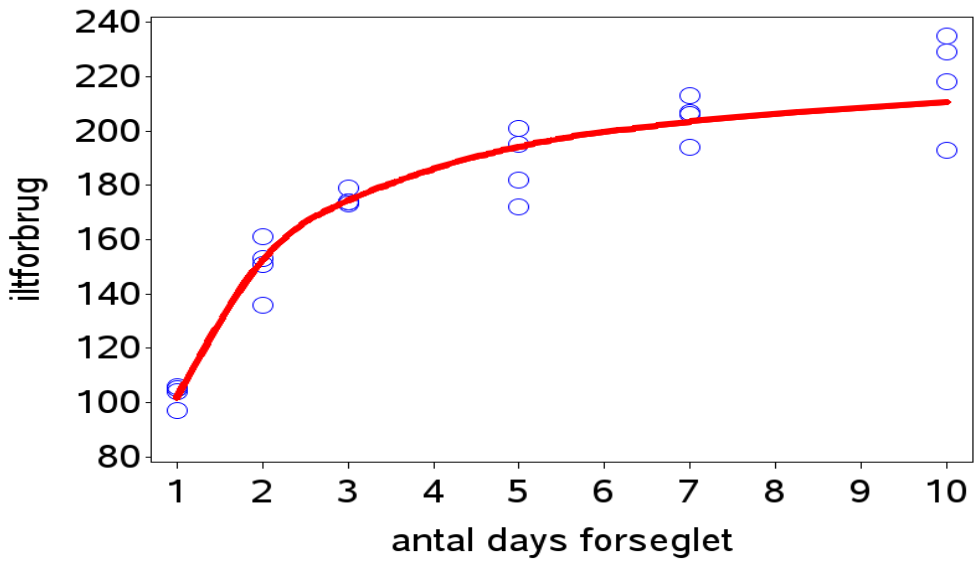
Ved at bemærke, at $\text{boc}(\infty) = \gamma = \exp(\alpha)$,
finder vi estimatet af $\text{boc}(\infty)$ til $\exp(5.431) = 228.38$

med 95% konfidensinterval
$(\exp(5.392), \exp(5.471)) = (219.6, 237.7)$

Tilbagetransformeret relation

er besværlig at lave i SPSS!!, se s. 98 - 99

så denne er lavet i SAS.



Analyse på original skala

kræver *ikke-lineær regression* (mere om dette lidt senere)

Her benyttes `Analyze/Regression/Nonlinear Regression`, hvor man sætter `boc` i `Dependent` og i `Model Expression` skriver selve det ikke-lineære udryk `gamma*exp(-beta/days)`.

Herefter klikker man i `Parameters` og tilføjer

- `Name: gamma, Starting Value: 228 Add`
- `Name: beta, Starting Value: 0,8 Add`

og til sidst: `Continue`

Bemærk, at resultaterne er *meget* tæt på dem fra før

Output fra ikke-lineær regression

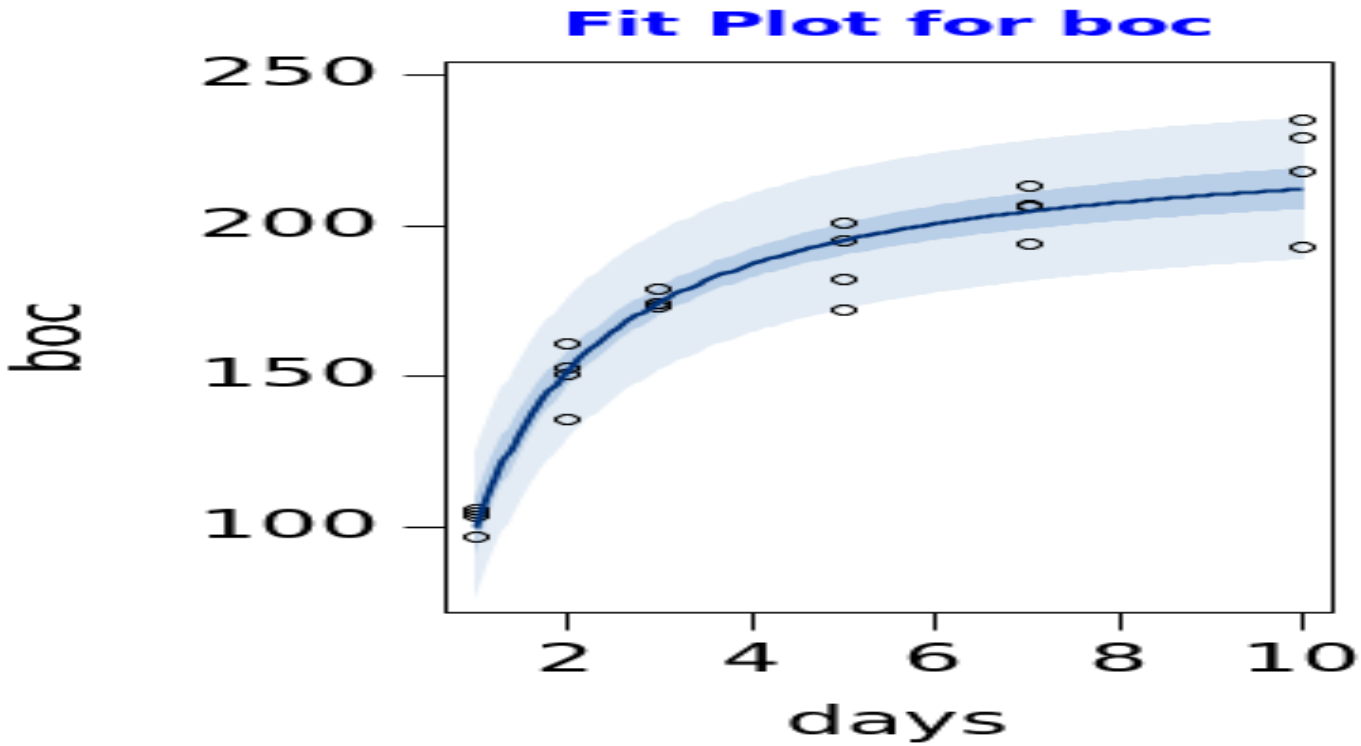
Kun parameterestimaterne er interessante:

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
gamma	230,584	4,538	221,172	239,997
beta	,833	,058	,713	,952

Fit fra ikke-lineær regression

Det automatiske plot fra SAS ser således ud:
(Dette plot synes ikke at være muligt i SPSS??)



Bemærk, at der på denne skala *ikke* er varianshomogenitet.

Transformation med logaritmer

– men hvilken logaritme?

Alle logaritmer er proportionale, så resultaterne bliver ens (efter tilbagetransformation), men der er visse fif:

- Hvis den forklarende variabel transformeres, er det i reglen for at opnå linearitet
  - Brug gerne 2-tals logaritmer her, for så kan estimatet fortolkes som effekten af *fordobling* af kovariaten.
  - Man kan også vælge en logaritme med grundtal 1.1, så estimatet fortolkes som effekten af *10% ændring* af kovariaten.

$$\log_{1.1}(x) = \frac{\log_{10}(x)}{\log_{10}(1.1)}$$

Transformation med logaritmer, II

- Hvis outcome transformeres, kan det være
  - for at opnå linearitet
  - for at opnå ens spredninger (varianshomogenitet)
    Her er der en vis fordel ved at bruge den naturlige logaritme (den som tidligere har heddet `ln`,
    men som altså hedder `log` i computersprog), fordi

$$\text{Spredning}(\log(y)) \approx \frac{\text{Spredning}(y)}{y} = \text{CV}$$

    dvs. en konstant variationskoefficient (CV) på Y betyder konstant spredning på $\log(Y)$,

I dette eksempel ser vi (fra outputtet s. 11), at variationskoefficienten for iltforbruget er 5.8%

Andre transformationer

Selv om logaritmer er langt det hyppigste valg af transformation, bruges af og til andre:

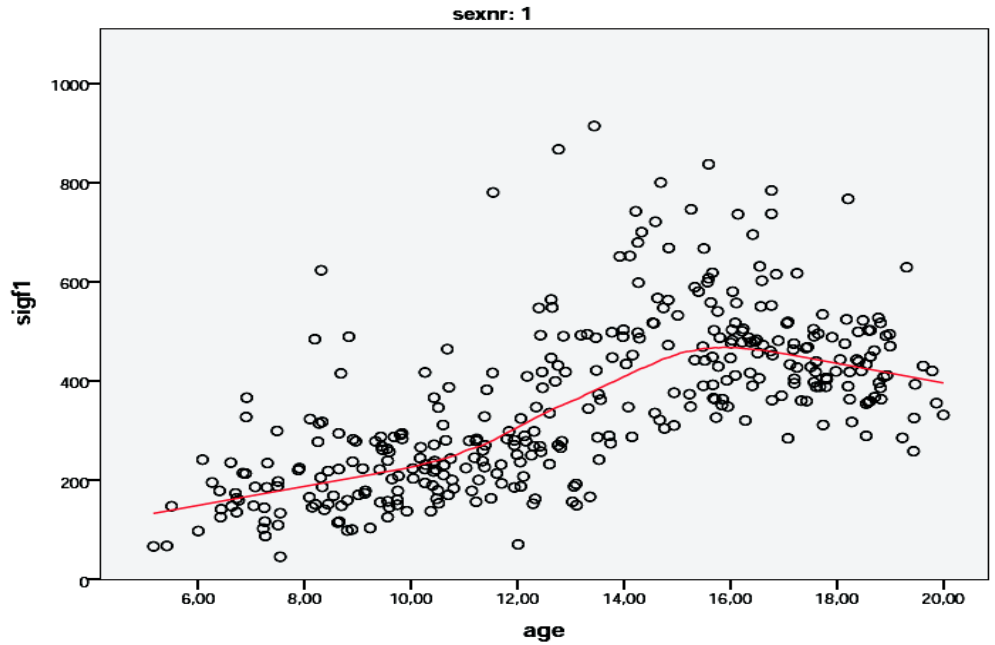
- I eksemplet med biokemisk iltforbrug (`boc`) brugte vi den inverse til kovariaten (1/dage),
  fordi vi havde en specifik viden om den biologiske mekanisme,
  og dermed om sammenhængen mellem outcome og kovariat
- Somme tider bruges kvadratrod for at få konstante spredninger (eller evt. normalfordelte residualer),
  hvis man har trompetfacon på den oprindelige skala,
  men omvendt trompet på logaritme skala,
  men det bliver ret svært at fortolke

Modeller med logaritmetransformerede data

Modelformel	*Tilbagetransformeret*	**Fortolkning**
$y = \alpha + \beta x$	(*ikke relevant*)	1 enheds tilvækst i x svarer til β enheders tilvækst i y
$\log_2(y) = \alpha + \beta x$	$y = \alpha_* \beta_*^x$ $\alpha_* = 2^\alpha, \beta_* = 2^\beta$	1 enheds tilvækst i x svarer til en faktor 2^β på y
$y = \alpha + \beta \log_2(x)$	(*ikke relevant*)	En faktor 2 på x svarer til β enheders tilvækst i y
$\log_2(y) = \alpha + \beta \log_2(x)$	$y = \alpha_* x^\beta$ $\alpha_* = 2^\alpha$	En faktor 2 på x svarer til en faktor 2^β på y

Eksempel om væksthormon (fra øvelserne i denne uge)

Væksthormon for drenge, op til 20-års alderen



Ikke udpræget lineært, men hvad så?

Ingen specifik formel haves....

Polynomier

Et p'te grads polynomium:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$$

Et første-grads polynomium er en linie:

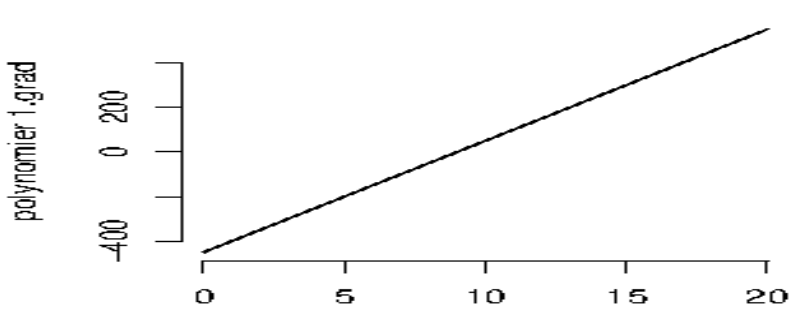
$$y = \beta_0 + \beta_1 x$$

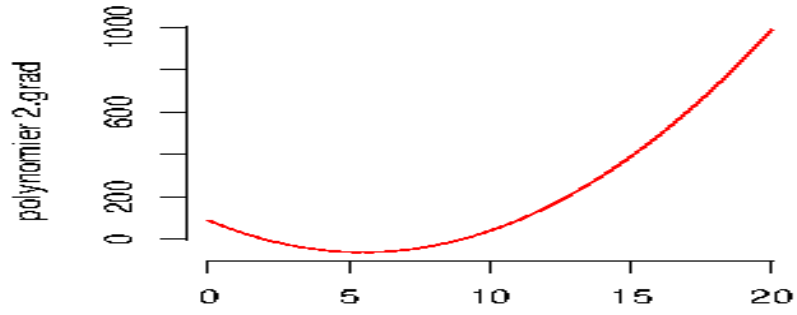
Et anden-grads polynomium er en parabel:

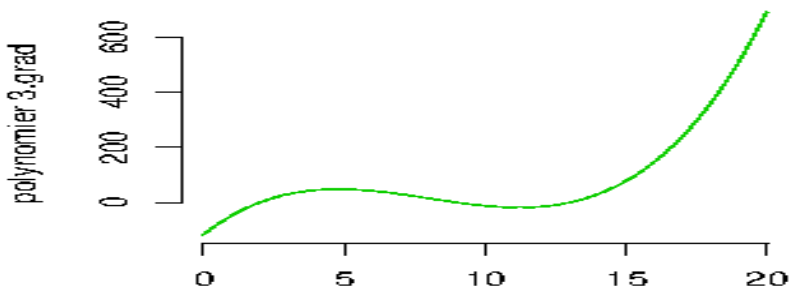
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

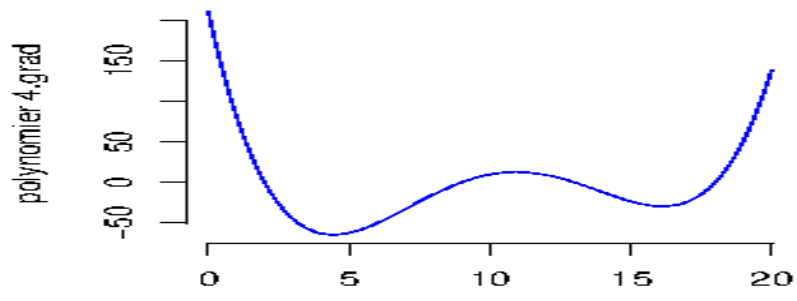
som kan være glad ($\beta_2 > 0$) eller sur ($\beta_2 < 0$)
Ser man lokalt på en parabel, kan den beskrive en afvigelse fra linearitet.

Polynomier af 1.-4. grad









Polynomial regression

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$$

Med kovariaterne

$$Z_1 = X, \quad Z_2 = X^2, \quad \cdots, \quad Z_p = X^p$$

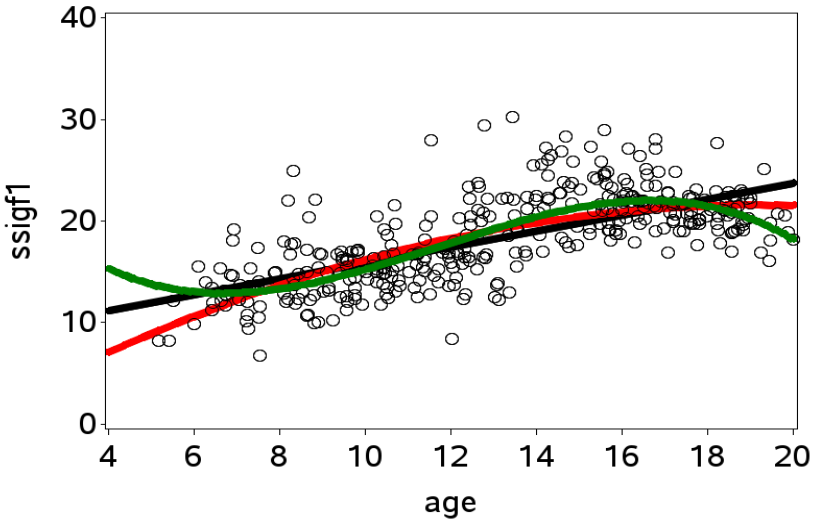
er det bare en sædvanlig **lineær multipel regression**

$$Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \cdots + \beta_p z_{pi} + \varepsilon_i$$

Kovariaterne $Z_1, \cdots, Z_p$ er selvfølgelig korrelerede, men de er ikke *lineært* afhængige.

Polynomier til beskrivelse af Serum IGF-1

Væksthormon (kvadratrodstransformeret), som funktion af alder



med overlejrede polynomier af 1., 2. og 3. grad (se s. 103)

Men vi tror ikke rigtigt på disse modeller.....
fordi de svinger for meget

Specielt ude i enderne kan de opføre sig meget underligt.

Splines: Lokale polynomier

Lineære splines:

- Opdel i aldersgrupper, med passende tærskelværdier, f.eks. $a_1 = 10, a_2 = 12, a_3 = 13, a_4 = 15$
- Fit en lineær effekt af alder i hver aldersgruppe
- Sørg for at de "mødes" i tærskelværdierne

Resultatet er en **knækket linie** (men stadig en lineær model)

$$y_i = \alpha + \lambda_0 x + \lambda_1 I(x > a_1)(x - a_1) + \cdots + \lambda_k I(x_i > a_k)(x - a_k) + \varepsilon_i$$

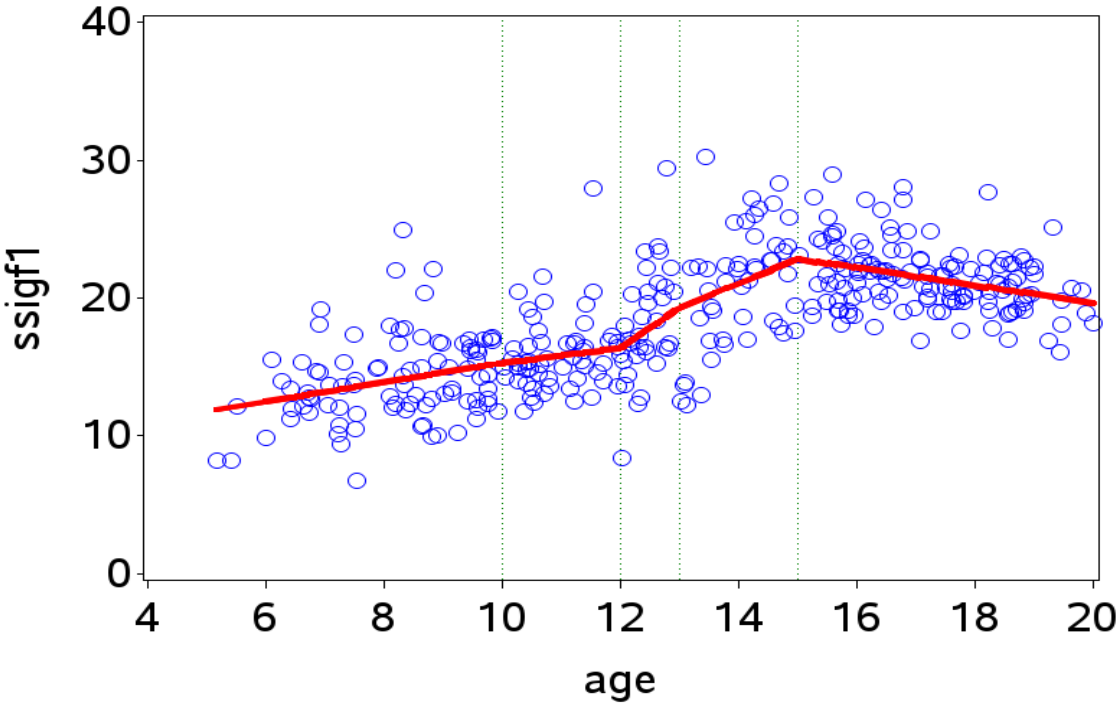
Splines kan også være kvadratiske, kubiske etc.

Fortolkning af parametre

- α: Intercept, forventet outcome ved alder 0
- λ_0: hældning (effekt af aldersøgning med 1 år), frem til alder a_1
- λ_1: knæk i "linien" ved alder a_1
  - $\lambda_1 = 0$: "linien" fortsætter hen over a_1 uden at knække
  - $\lambda_1 > 0$: "linien" knækker, og får større hældning efter a_1
  - $\lambda_1 < 0$: "linien" knækker, og får mindre hældning efter a_1

  Hældning i aldersintervallet (a_1, a_2) er $\lambda_0 + \lambda_1$
- λ_2:
  knæk i "linien" ved alder a_2
  Samme fortolkning som λ_1, blot ved en anden alderstærskel
  Hældning i aldersintervallet (a_2, a_3) er $\lambda_0 + \lambda_1 + \lambda_2$
- osv. osv.

Lineær spline for Serum IGF-1



Estimater:

$\lambda_0 = 0.70, \lambda_1 = -0.16, \lambda_2 = 2.38, \lambda_3 = -1.15, \lambda_4 = -2.42$

(Denne tegning kan ikke laves i SPSS, kun fittet for sig selv, se s. 105)

At fitte lineære splines

Ud over selve kovariaten (her age) skal man tilføje en ekstra kovariat for hver tærskelværdi
For tærsklen ved 13 definerer vi således en variabel:

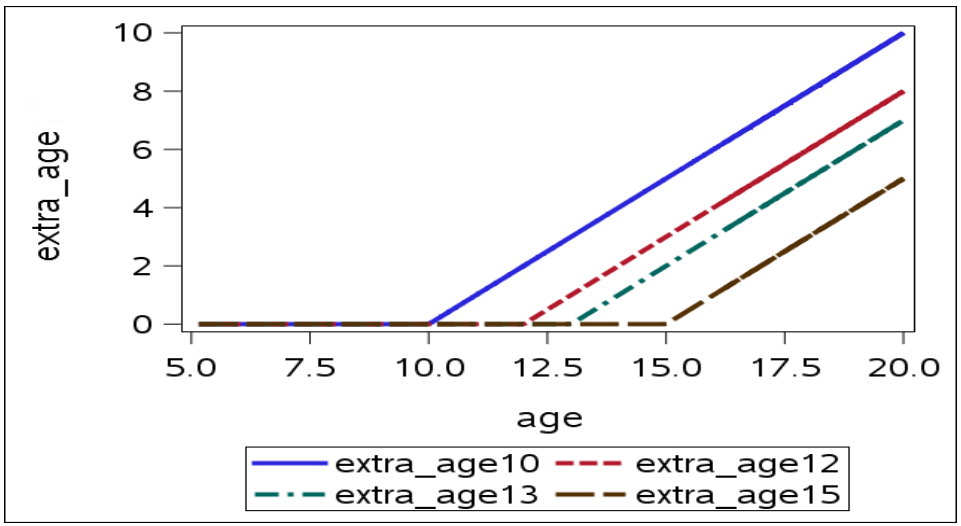
- For personer under 13 år sættes til den 0
- For personer over 13 år, defineres den som alder minus 13, så den f.eks. får værdien 2.4 for en person med alderen 15.4 år

```
extra_age10=max(age-10,0)
extra_age12=max(age-12,0)
extra_age13=max(age-13,0)
extra_age15=max(age-15,0)
```

Udseendet af lineære splines

4 stk, for tærskelværdierne 10, 12, 13 og 15:

- De er alle 0 op til deres respektive tærskel
- Herefter stiger de med 1 år pr. år, så de tæller "*år fra tærskel*"



Output, lineær spline

Se s. 104

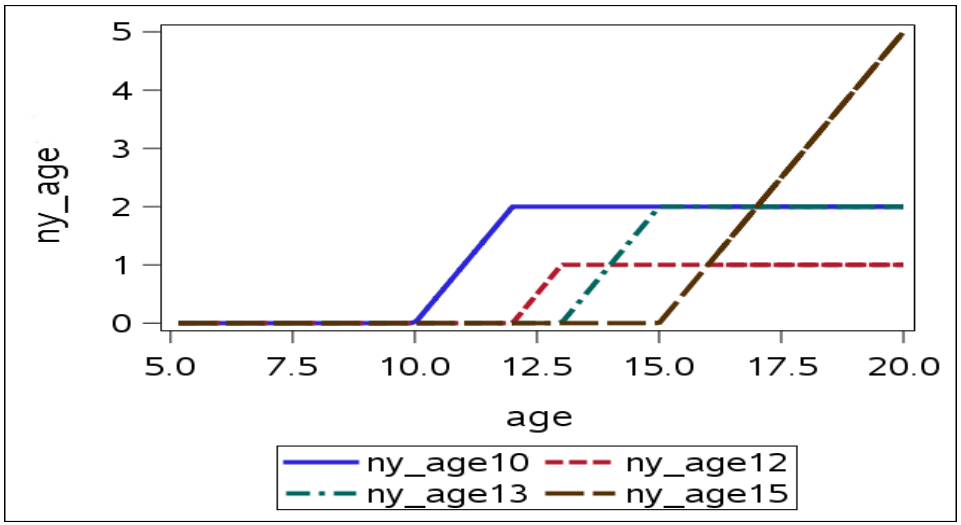
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8,283	1,867		4,436	,000
	age	,704	,218	,606	3,231	,001
	extra_age10	-,163	,556	-,117	-,292	,770
	extra_age12	2,384	1,251	1,357	1,905	,058
	extra_age13	-1,152	1,281	-,555	-,899	,369
	extra_age15	-2,418	,539	-,727	-4,489	,000

Evidens for et knæk omkring 15-års alderen,
og *måske* omkring 12-års alderen...

*Alternativ parametrisering af lineære splines

Hvis man ændrer de nye kovariater til disse (se s. 106):



så får man i stedet estimater for hældningerne i de enkelte aldersintervaller (ret teknisk):

*Output fra alternativ parametrisering

af lineære splines (kode s. 106)

Nu fortolkes estimaterne som
hældningerne i de successive intervaller:
Op til alder 10, mellem 10 og 12, mellem 12 og 13,
mellem 13 og 15, samt over 15 år

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	8,283	1,867		4,436	,000
	ny_age	,704	,218	,159	3,231	,001
	ny_age10	,542	,404	,109	1,340	,181
	ny_age12	2,926	,950	,320	3,079	,002
	ny_age13	1,774	,420	,382	4,220	,000
	ny_age15	-,644	,175	-,194	-3,681	,000

Kan GLM så klare alt?

Nej

der findes ikke-lineære modeller, der

- ikke kan transformeres til linearitet
- indeholder parametre med meget veldefineret betydning (typisk fysiologi/kinetik), som kun estimeres *pænt* i en ikke-lineær model (f.eks. Michaelis-Menten kinetik)

Eksempel: Model til at kvantificere **RES**-systemet i leveren:

RES-systemet i leveren

Lad Y_i betegner den målte koncentration af en radioaktiv tracer, målt til tiden t_i efter en bolus injektion ved tid 0.

Så siger 1. ordens kinetik,
at sammenhængen bør være

$$y_i = \beta(1 - e^{-\gamma t_i}) + \varepsilon_i,$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

Her kan ikke transformeres til linearitet!

Mindste kvadraters metode

kræver her

- startværdier (gæt på parametrenes værdier)
  kan være ganske vanskeligt, og der er ingen generelle retningslinier
- iterationer (trinvise forbedrede fit)
  klares heldigvis af programmet

Mindste kvadraters metode

Her benyttes `Analyze/Regression/Nonlinear Regression`, hvor man sætter `konc` i `Dependent` og i `Model Expression` skriver selve det ikke-lineære udryk `beta*(1-exp(-gamma*tid))`.

Herefter klikker man i `Parameters` og tilføjer

- `Name: gamma, Starting Value: 0,05 Add`
- `Name: beta, Starting Value: 2000 Add`

og til sidst: `Continue`

Output ses s. 39-40

*Output fra ikke-lineær regression

fra opsætning forrige side:

Iteration History[b]

Iteration Number[a]	Residual Sum of Squares	Parameter beta	Parameter gamma
1.0	4370990,225	2000,000	,050
1.1	382995,386	1958,555	,082
2.0	382995,386	1958,555	,082
2.1	26354,953	2169,165	,077
3.0	26354,953	2169,165	,077
3.1	25286,735	2174,161	,077
4.0	25286,735	2174,161	,077
4.1	25286,715	2174,048	,077
5.0	25286,715	2174,048	,077
5.1	25286,715	2174,044	,077

Derivatives are calculated numerically.

a. Major iteration number is displayed to the left of the decimal, and minor iteration number is to the right of the decimal.

b. Run stopped after 10 model evaluations and 5 derivative evaluations because the relative reduction between successive residual sums of squares is at most SSCON = 1,000E-8.

Når et "*stabilt*" fit er fundet, stopper processen (konvergens)

Output, fortsat

Parameter Estimates

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
beta	2174,044	28,346	2115,541	2232,548
gamma	,077	,002	,073	,082

ANOVA[a]

Source	Sum of Squares	df	Mean Squares
Regression	63048136,28	2	31524068,14
Residual	25286,715	24	1053,613
Uncorrected Total	63073423,00	26	
Corrected Total	3455129,115	25	

Dependent variable: konc

a. R squared = 1 - (Residual Sum of Squares) / (Corrected Sum of Squares) = ,993.

Fittet svarende til disse parameterestimater fremgår af s. 41

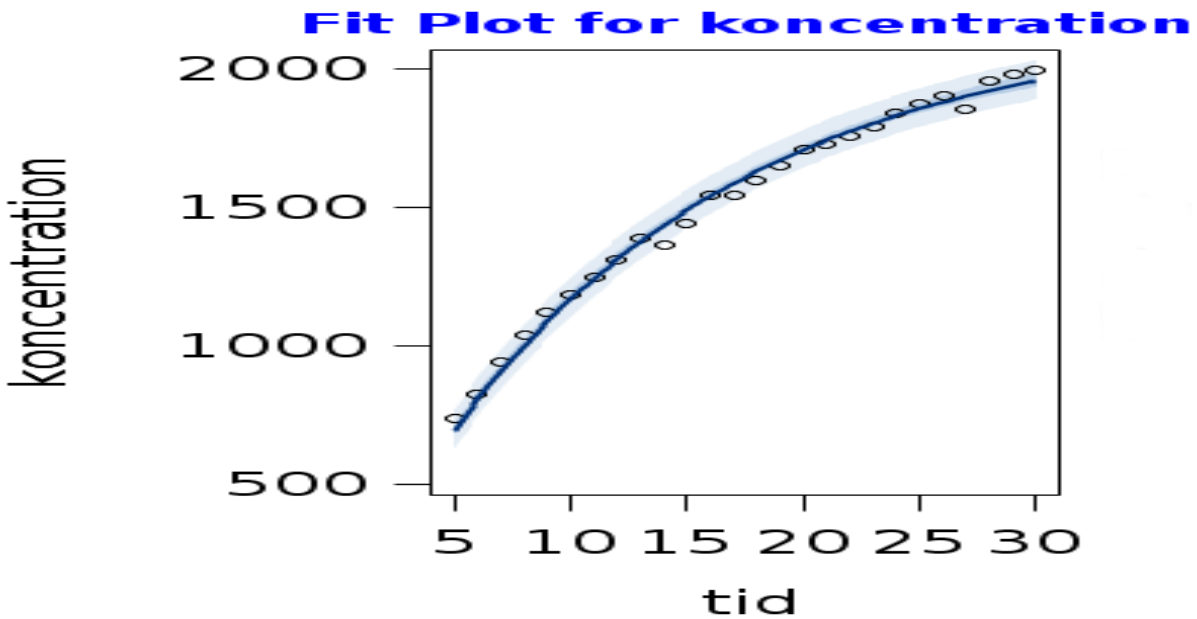
Fittet fra ikke-lineær regression

Estimater:

- $\hat{\beta} = 2174.0(28.3)$, CI=(2115.5, 2232.5)
- $\hat{\gamma} = 0.0773(0.0023)$, CI=(0.0726, 0.0819)

Det automatiske plot fra SAS ser således ud:

Dette plot synes ikke at være muligt i SPSS??



Den generelle lineære model

er et slagkraftigt værktøj,

men altså med visse begrænsninger:

Outcome: kvantitativ variabel Y
(med ca. normalfordelte residualer)

Kovariater:
- Kategoriske (`class`)
- Kvantitative:
  Her antages linearitet
- Interaktioner

Men hvordan vælges modellen?

Opbygning af model

bør følge problemstillingen som beskrevet i protokollen

- Her bør der være specificeret
  - primært outcome
  - de vigtigste hypoteser (videnskabelige spørgsmål)
  - sekundære outcomes og hypoteser
- pludselige indskydelser (evt. baseret på tegninger), samt tests, der ikke var specificeret i protokollen, betegnes som fisketur, og skal bekræftes i en ny (confirmative) analyse, før der kan skabes tillid til resultaterne.

Det betaler sig at gøre forarbejdet ordentligt,
så man ikke bliver berømt på sine fejlkonklusioner

Eksempel: Modelbygning for Vitamin D

Problemformulering i protokol:

- Er der forskel på vitamin D i de forskellige lande?
  – efter korrektion for allerede "*etablerede*" kovariater.
- Hvis ja, så hvorfor?
  Hypoteser:
  - Primær:
    pga forskel i fedme (`bmi`)
    fordi vitamin D er fedtopløseligt
  - pga forskelle i solvaner (`sunexp`)
    fordi solen laver vitamin D i huden
  - pga forskelle i spisevaner (`vitdintake`)
    nogle steder spiser man måske flere (fede) fisk
  - pga aldersforskelle....?
  -
  -

Vitamin D i de 4 lande

Tabel over median værdier:

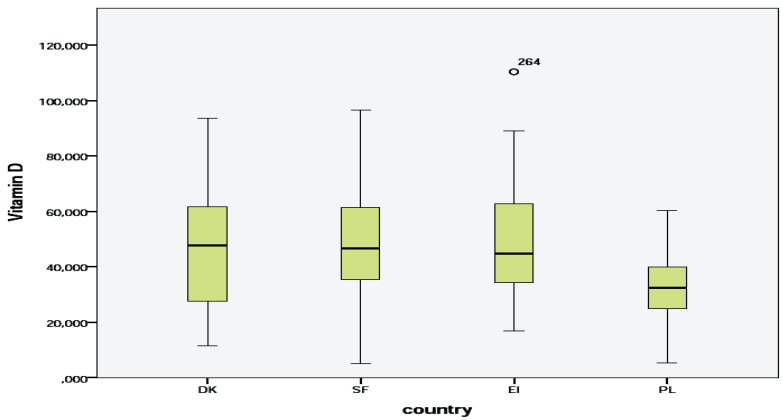
Land	Nr	Antal	Vitamin D	Alder	Body Mass Index	Vitamin D Indtag
Denmark	1	53	47.80	71.51	25.39	8.29
Finland	2	54	46.60	71.92	27.98	12.41
Ireland	4	41	44.80	72.05	26.39	5.46
Poland	6	65	32.50	71.69	29.37	5.16

Polen ligger lavt i vitamin D niveau, og

- højt i body mass index
- lavt i vitamin D indtag

Første skridt

Er der overhovedet signifikant forskel på landene?



Modeldiagram: Country $\longrightarrow$ Vitamin D

Uanset om man ser på utransformerede data eller logaritmetransformerede data, finder man en forskel på landene ($P = 0.000$), idet Polen findes at ligge lavere end de øvrige.

Vi vælger at køre videre på logaritmeskala.

Sammenligning af landene

Outcome Y: kvantitativ, Y=`lvitd`, log2-transformeret

Kovariat X: kategorisk, X=`country`

Derfor: Ensidet variansanalyse, på log2-skala:
Sammenligning af 4 middelværdier (se s. 108)

Tests of Between-Subjects Effects

Dependent Variable: lvitd

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12,928[a]	3	4,309	8,377	,000
Intercept	5799,200	1	5799,200	11272,756	,000
country	12,928	3	4,309	8,377	,000
Error	107,519	209	,514		
Total	6007,382	213			
Corrected Total	120,447	212			

a. R Squared = ,107 (Adjusted R Squared = ,095)

Sammenligning af landene, II

Parameter Estimates

Dependent Variable: lvitd

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4,890	,089	54,961	,000	4,714	5,065
[country=1]	,471	,133	3,547	,000	,209	,732
[country=2]	,558	,132	4,223	,000	,297	,818
[country=4]	,567	,143	3,963	,000	,285	,849
[country=6]	0[a]	.	.	.	.	.

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

Fortolkning af estimater

Den estimerede forskel, f.eks. mellem Finland og Polen er en faktor

$$2^{0.558} = 1.47$$

altså svarende til 47% større niveau af vitamin D i Finland sammenlignet med Polen. Konfidensintervallet for denne sammenligning udregnes på samme måde ud fra konfidensintervallet (ikke vist ovenfor af pladshensyn) som

$$(2^{0.297}, 2^{0.818}) = (1.23, 1.76)$$

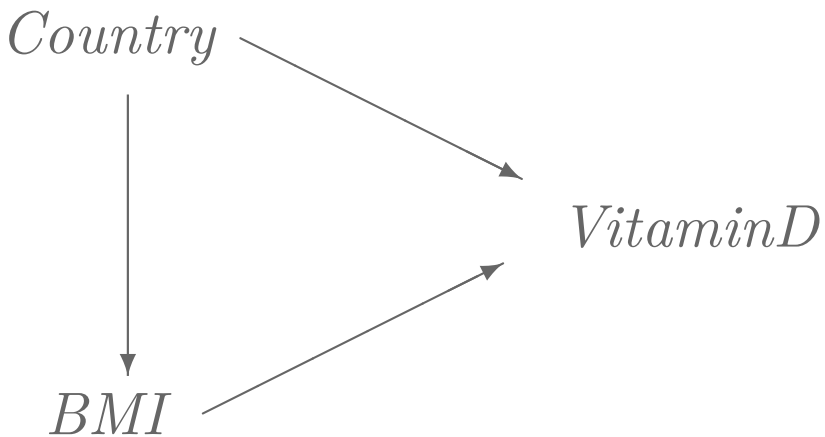
altså fra 23% over til 76% over.

Hvordan forklarer vi forskellen mellem landene?

kan vi f.eks. forklare det ved forskelle i body mass index?

$$\text{Country} \rightarrow \text{BMI} \rightarrow \text{Vitamin D}$$

eller måske bare noget af det?



BMI er en mellemkommende variabel (mediator)

Model med bmi som kovariat (se s. 109)

Tests of Between-Subjects Effects

Dependent Variable: lvitd

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	18,222[a]	4	4,556	9,269	,000
Intercept	188,938	1	188,938	384,439	,000
country	9,992	3	3,331	6,777	,000
bmi	5,294	1	5,294	10,772	,001
Error	102,224	208	,491		
Total	6007,382	213			
Corrected Total	120,447	212			

a. R Squared = ,151 (Adjusted R Squared = ,135)

Parameter Estimates

Dependent Variable: lvitd

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	5,992	,347	17,274	,000	5,308	6,675
[country=1]	,379	,133	2,857	,005	,118	,641
[country=2]	,534	,129	4,134	,000	,280	,789
[country=4]	,468	,143	3,275	,001	,186	,750
[country=6]	0[a]	.	.	.	.	.
bmi	-,038	,012	-3,282	,001	-,061	-,015

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

Fortolkning af nye estimater

Den estimerede forskel mellem Finland og Polen, for folk med samme BMI, er en faktor

$$2^{0.534} = 1.45$$

altså næsten det samme som før (meget lidt confounding). Konfidensintervallet for denne sammenligning bliver

$$(2^{0.280}, 2^{0.789}) = (1.21, 1.73)$$

en ubetydelighed lavere end den ujusterede forskel fra s. 49

Kunne bmi forklare forskellen på landene?

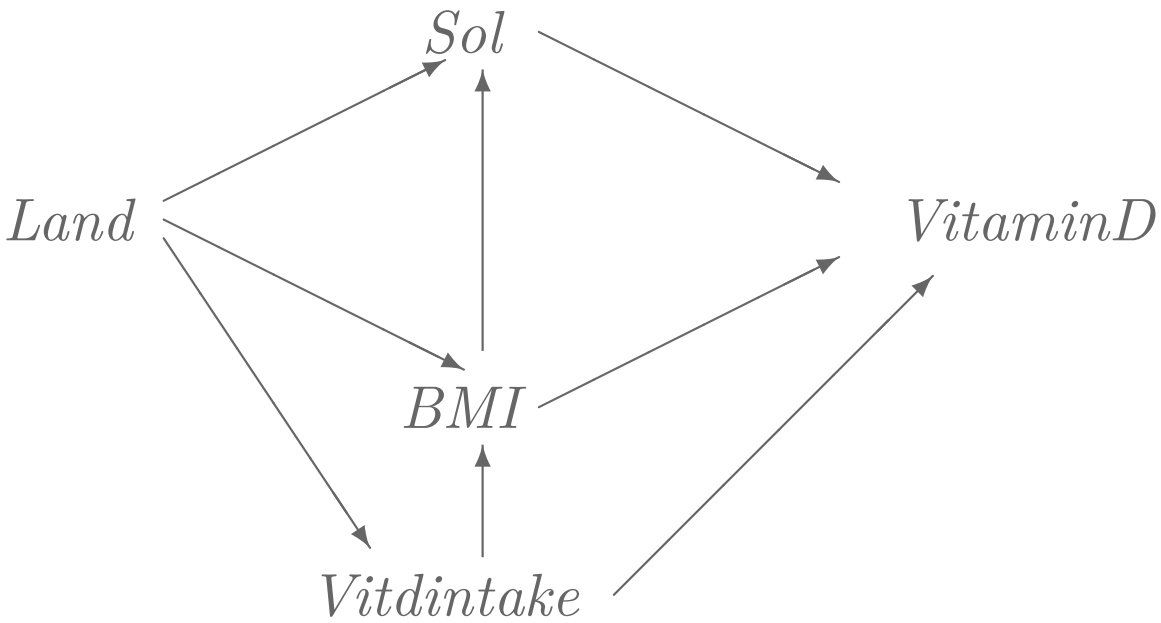
Nej,

- Selv om `bmi` i sig selv er stærkt signifikant (negativ effekt, estimeret til $-0.038(0.012), P = 0.001$),

  er der stadig stærkt signifikant forskel på landene, når vi har korrigeret for `bmi` ($P = 0.000$), så der er masser af plads til andre bud på forklarende variable

  fra protokollen, vel at mærke.

Modeldiagram - med "det hele"



og så er der interaktionerne....

Kan vi så forklare forskellen på landene

ved hjælp af alle de 3 specificerede kovariater samtidig?

Tests of Between-Subjects Effects

Dependent Variable: lvitd

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	40,345[a]	7	5,764	14,750	,000
Intercept	116,267	1	116,267	297,555	,000
country	5,596	3	1,865	4,774	,003
bmi	3,791	1	3,791	9,703	,002
sunexp	1,080	2	,540	1,381	,254
lvitdintake	19,940	1	19,940	51,031	,000
Error	80,102	205	,391		
Total	6007,382	213			
Corrected Total	120,447	212			

a. R Squared = ,335 (Adjusted R Squared = ,312)

Næh.....der er stadig signifikant forskel på landene

Output, fortsat

Parameter Estimates

Dependent Variable: lvitd

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5,358	,360	14,876	,000	4,648	6,068
[country=1]	,319	,122	2,621	,009	,079	,558
[country=2]	,248	,123	2,020	,045	,006	,490
[country=4]	,459	,128	3,597	,000	,207	,711
[country=6]	0[a]	.	.	.	.	.
bmi	-,033	,010	-3,115	,002	-,053	-,012
[sunexp=1]	-,216	,130	-1,662	,098	-,472	,040
[sunexp=2]	-,144	,122	-1,183	,238	-,385	,096
[sunexp=3]	0[a]	.	.	.	.	.
lvitdintake	,255	,036	7,144	,000	,185	,326

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

Fortolkning af estimater

fra output på forrige side

- 1 enheds stigning i BMI giver en faktor $2^{-0.033} = 0.977$ på vitamin D niveauet, dvs. et fald på 2.3%, med konfidensgrænser $(2^{-0.053}, 2^{-0.012}) = (0.964, 0.992)$, svarende til et fald på mellem 0.8% og 3.6%.
- Solvanerne ser ikke ud til at betyde så meget, men mønstret ser fornuftigt ud, så *måske* ville effekten vise sig i et større materiale.
  Forskellen på "ligeglade" og solhadere estimeres her til en faktor $2^{-0.144+0.216} = 1.05$, altså svarende til 5% større niveau af vitamin D hos "ligeglade" i forhold til solhadere.
  Konfidensgrænser: se side 60-65.

Fortolkning af estimater, fortsat

fra output på forrige side

- En 10% øgning i vitamin D indtag (dvs. en faktor 1.1 på denne), svarer til en faktor $1.1^{0.255} = 1.025$ på vitamin D niveauet, altså kun en stigning på 2.5%
(se s. 21, nederste modelformel).
Konfidensgrænserne er $(1.1^{0.185}, 1.1^{0.326}) = (1.018, 1.032)$, svarende til en stigning på mellem 1.8% og 3.2%.
- Den estimerede forskel mellem Finland og Polen, for folk med samme BMI, solvaner og vitamin D indtag, er en faktor $2^{0.248} = 1.19$ altså noget mindre end tidligere, svarende til at vi har kunnet forklare en vis del af forskellen mellem de to lande (faktisk er denne forskel kun lige akkurat signifikant).

  Konfidensintervallet for denne sammenligning er $(2^{0.0.006}, 2^{0.490}) = (1.004, 1.40)$

Sammenligning mellem andre lande

Den estimerede forskel mellem lande som Irland og Danmark fremgår ikke direkte af output, men kan nemt udregnes til faktoren

$$2^{0.459-0.319} = 1.10$$

altså svarende til 10% større niveau af vitamin D i Irland sammenlignet med Danmark.

For at få konfidensintervallet for denne sammenligning kan man i SPSS f.eks. benytte parvise sammenligninger af alle landene i stedet for (se s. 111 samt output s. 61-62).

Irland vs. Danmark, og "ligeglade" vs. solhadere

Disse sammenligninger fremkommer ikke automatisk, fordi der ikke er tale om sammenligninger til referencen.

Det kan løses på flere måder, afhængigt af program:

- Omkodning, så vi får en anden reference
- Direkte valg af en bestemt reference-værdi
- Kombination af parameterestimater, f.eks. differenser (SAS)
- Multiple parvise sammenligninger, som vist på næste side

Output fra parvise sammenligninger

Se s. 111

Pairwise Comparisons

Dependent Variable: lvitd

(I) country	(J) country	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b] Lower Bound	Upper Bound
DK	SF	,070	,126	,578	-,179	,320
	EI	-,141	,132	,288	-,401	,119
	PL	,319*	,122	,009	,079	,558
SF	DK	-,070	,126	,578	-,320	,179
	EI	-,211	,137	,124	-,481	,059
	PL	,248*	,123	,045	,006	,490
EI	DK	,141	,132	,288	-,119	,401
	SF	,211	,137	,124	-,059	,481
	PL	,459*	,128	,000	,207	,711
PL	DK	-,319*	,122	,009	-,558	-,079
	SF	-,248*	,123	,045	-,490	-,006
	EI	-,459*	,128	,000	-,711	-,207

Based on estimated marginal means

*. The mean difference is significant at the ,05 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Output fra parvise sammenligninger,II

Pairwise Comparisons

Dependent Variable: lvitd

(I) sunexp	(J) sunexp	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence [a]... Lower Bound
Avoid sun	Sometimes in sun	-,072	,099	,468	-,266
	Prefer sun	-,216	,130	,098	-,472
Sometimes in sun	Avoid sun	,072	,099	,468	-,123
	Prefer sun	-,144	,122	,238	-,385
Prefer sun	Avoid sun	,216	,130	,098	-,040
	Sometimes in sun	,144	,122	,238	-,096

Pairwise Comparisons

Dependent Variable: lvitd

(I) sunexp	(J) sunexp	95% Confidence Interval for [a]... Upper Bound
Avoid sun	Sometimes in sun	,123
	Prefer sun	,040
Sometimes in sun	Avoid sun	,266
	Prefer sun	,096
Prefer sun	Avoid sun	,472
	Sometimes in sun	,385

Based on estimated marginal means

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Effekt af 10% øgning i vitd intake

Her defineres en ny kovariat:
`lvitdintake11 = LG10(vitdintake)/LG10(1.1)`
således at 1 enhed af denne svarer til en 10% forøgelse af `vitdintake`. Vi får så outputtet:

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5,358	,360	14,876	,000	4,648	6,068
[country=1]	,319	,122	2,621	,009	,079	,558
[country=2]	,248	,123	2,020	,045	,006	,490
[country=4]	,459	,128	3,597	,000	,207	,711
[country=6]	0[a]	.	.	.	.	.
bmi	-,033	,010	-3,115	,002	-,053	-,012
[sunexp=1]	-,216	,130	-1,662	,098	-,472	,040
[sunexp=2]	-,144	,122	-1,183	,238	-,385	,096
[sunexp=3]	0[a]	.	.	.	.	.
lvitdintake11	,035	,005	7,144	,000	,025	,045

Output fra ekstra sammenligninger

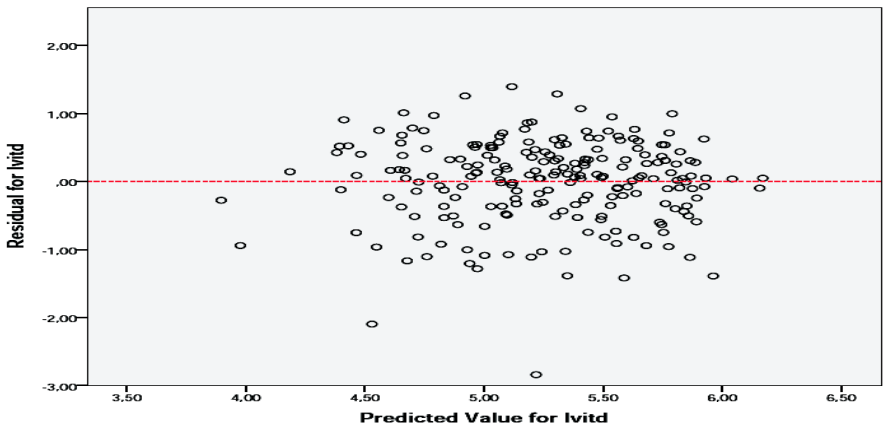
- Irland vs. Danmark: 0.141, CI=(-0.119, 0.401)
- "Sometimes in sun" vs. "Avoid Sun": 0.072, CI=(-0.123, 0.266)
- En 10% øgning i Vitamin D indtag: 0.035, CI=(0.025, 0.045)

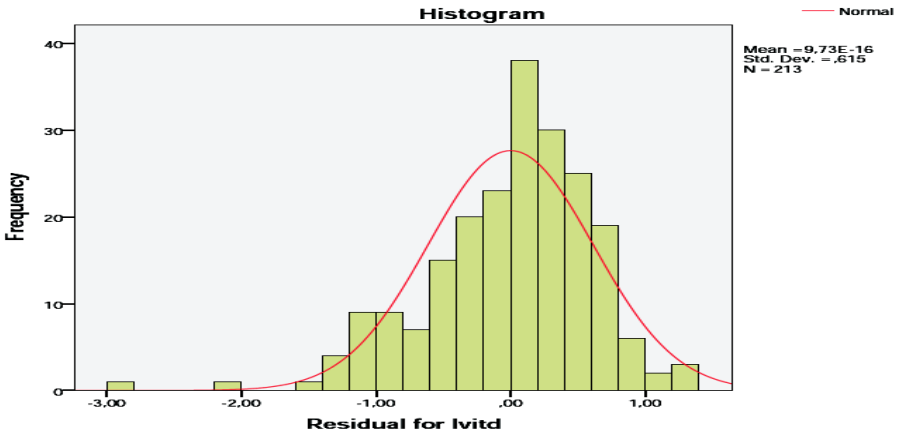
Disse skal nu tilbagetransformeres,
hvorved vi får

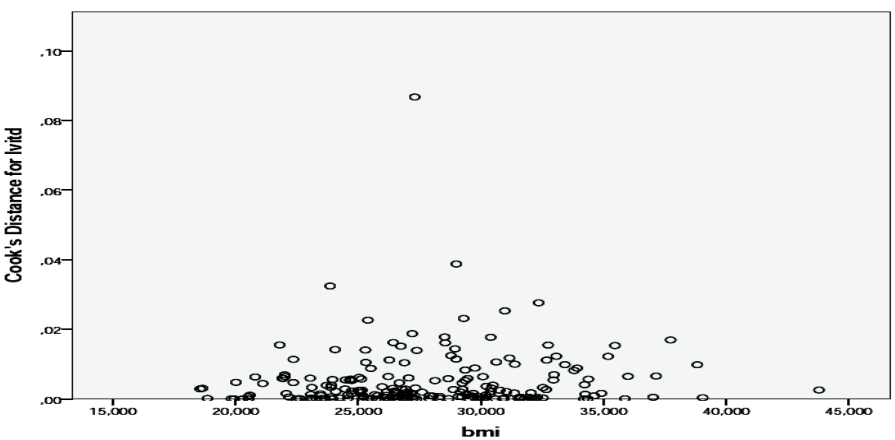
Tilbagetransformerede estimater

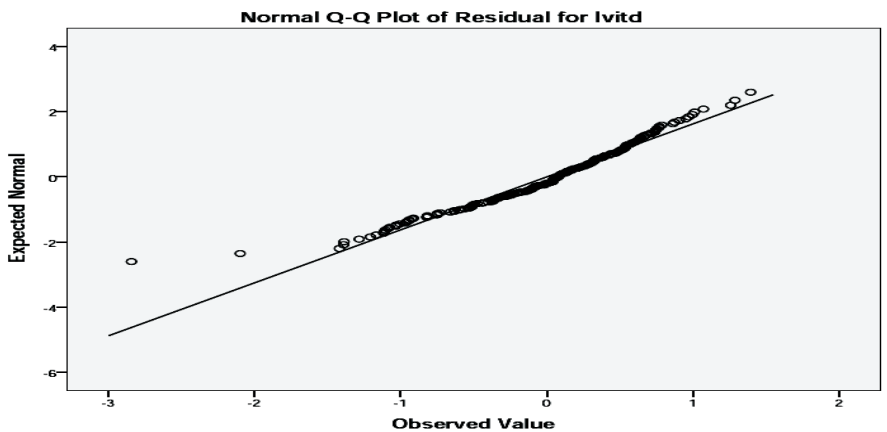
- Den estimerede forskel Irland vs. Danmark er en faktor: $2^{0.141} = 1.103$, svarende til, at Irland ligger ca. 10.3% højere end Danmark, CI=(-7.1%, 32.0%)
- Forskellen "ligeglade" vs. solhadere estimeres til en faktor $2^{0.072} = 1.051$, svarende til en øgning på 5.1% i vitamin D status, CI=(-8.2%, 20.2%)
- En 10% øgning i Vitamin D indtag estimeres til en faktor $2^{0.035} = 1.025$, svarende til en øgning i Vitamin D status på 2.5%, CI=(1.7%, 3.2%)

Husk også modelkontrol (se s. 112ff)

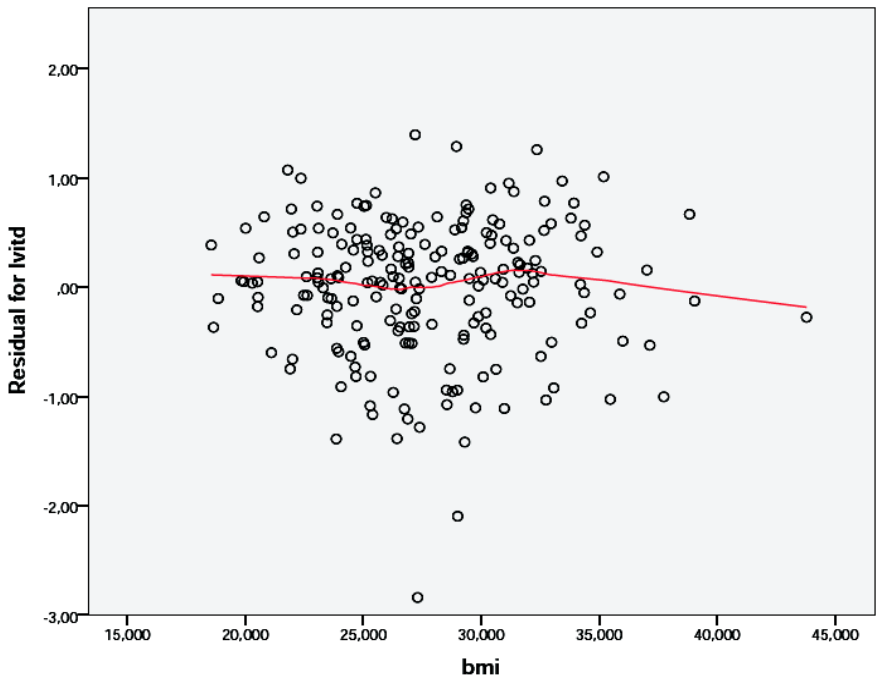


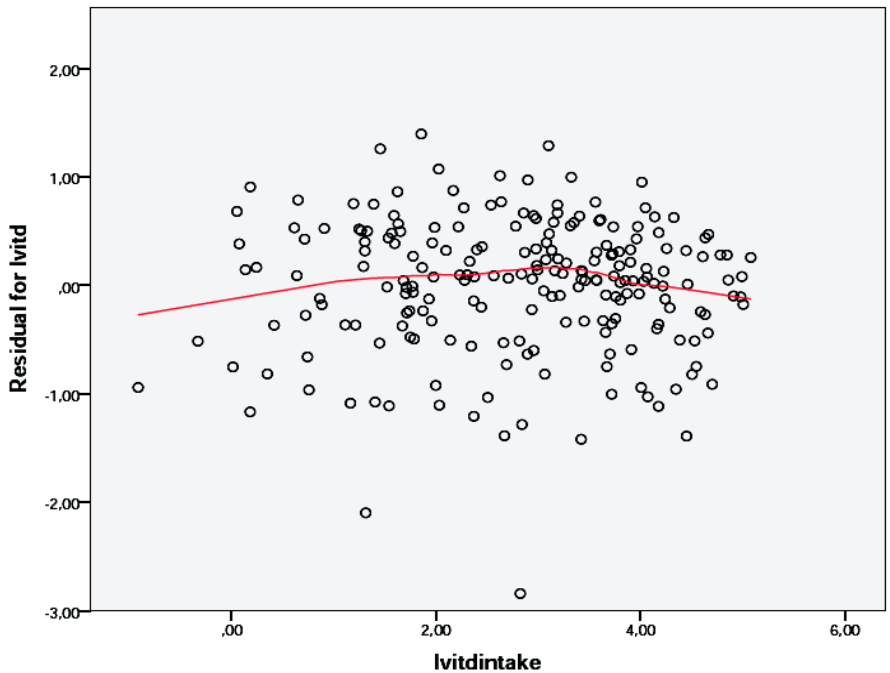






Modelkontrol, II





Note om T-test og F-test

Et T-test tester typisk, om en parameter kan være 0

- Forskel på 2 middelværdier, $\mu_1 - \mu_2$
- En hældning β, f.eks. en lineær effekt af `bmi` eller `alder`

Et F-test tester *flere* sådanne på en gang

- Identitet af middelværdier af vitamin D for 4 lande ($\mu_1 = \mu_2 = \mu_3 = \mu_4$, df=3)
- *Samtidig* fjernelse af flere kovariater på en gang, f.eks. 3 kostvariable ($\beta_1 = \beta_2 = \beta_3 = 0$, df=3)

*Modelreduktion - F test

Vi skal sammenligne to modeller:
Den oprindelige med *alle* kovariater (nr. 1)
og den simplere (hypotesen, model nr. 2 uden 3 af kovariaterne)

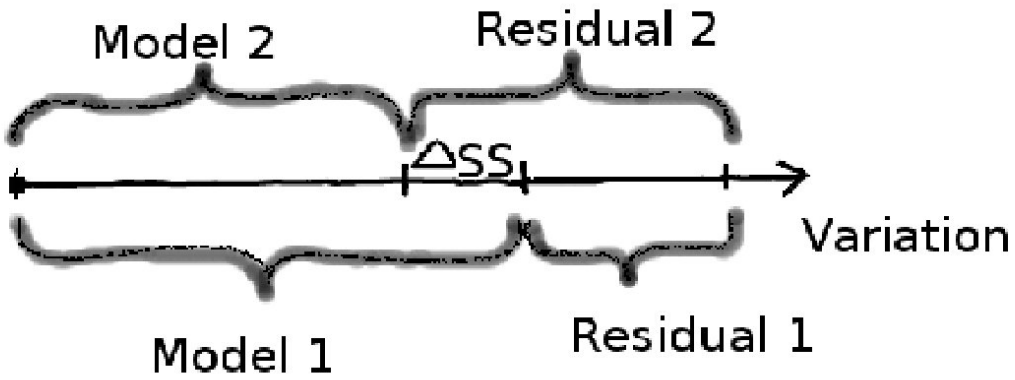
Kan vi forsvare at bruge den simpleste af dem?
Beskriver den data tilstrækkeligt godt?

NB: Modellerne skal være "nestede", dvs. den ene fremkommer af den anden, typisk ved at sætte parametre til nul ("fjerne effekter").

Se på **ændring** i model-kvadratsum:
Hvor meget **mindre** forklares af den simplere model?

$$\Delta_{\text{SS}} = SS_{\text{model1}} - SS_{\text{model2}}$$

Forståelse



Flere parametre kan forklare (lidt) mere variation: $\Delta SS > 0$

Spørgsmålet er: **Hvor meget mere?**

Hvor stor skal ΔSS være, før vi erklærer testet signifikant?

Det er det, F-testet svarer på. Her er vores hypotese (model 2), at fire parametre (svarende til effekten af 3 kovariater) er lig med 0

Udelad flere kovariater på en gang?

I VitaminD-eksemplet har vi p.t. 4 kovariater:

- `bmi` (df=1)
- `sunexp` (df=2)
- `lvitdintake` (df=1)
- `country` (df=3)

Bidrager de 3 øverste med noget forklaringsevne tilsammen?

Det gør de selvfølgelig, fordi 2 af dem selvstændigt gør det, men for *princippets skyld*:

Dette kan testes ved et F-test, der sammenligner 2 modeller:
Den *med* de 3 kovariater og den *uden* disse 3,
Her finder vi $F = 17.54 \sim F(4, 205), \quad P < 0.0001,$
(se s. 72 eller alternativt s. 114)

Udelad flere kovariater på en gang, II

Her gør vi dette "*håndholdt*", altså ved at køre modellen både med og uden de 3 kovariater og så sammenligne de SS-størrelser, man får fra dette.
Vi har her størrelserne:

- Model kun med `country` (s. 47): SS=107.519, df=209
- Model med `country` og 3 yderligere kovariater (s. 55): SS=80.102, df=205

Dette giver os et F-test:

$$\frac{(107.519 - 80.102)/4}{80.102/205} = 17.54 \sim F(4, 205), \quad P < 0.0001$$

Se evt også s. 114

Hypoteser vedr. interaktioner

Hvilke kunne vi forestille os at se på?

- `sunexp*lvitdintake`:
  måske optages vitamin D fra kosten bedre,
  hvis man samtidig får sol?
  nok lidt spekulativt.....
- `country*sunexp`:
  Pga breddegrad: Solen sydpå er nok lidt mere effektiv
  (det så vi i forelæsningen om ANOVA)
- `country*lvitdintake`:
  næppe...*og dog*
-

Kun præ-specificerede, og fortolkelige interaktioner
bør inkluderes i modellen.

Interaktionen sunexp*lvitdintake (se s. 115)

Tests of Between-Subjects Effects

Dependent Variable: lvitd

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	40,918[a]	9	4,546	11,605	,000
Intercept	115,499	1	115,499	294,814	,000
country	5,669	3	1,890	4,824	,003
bmi	3,894	1	3,894	9,940	,002
sunexp	1,305	2	,652	1,665	,192
sunexp * lvitdintake	,573	2	,287	,731	,483
lvitdintake	15,916	1	15,916	40,625	,000
Error	79,529	203	,392		
Total	6007,382	213			
Corrected Total	120,447	212			

a. R Squared = ,340 (Adjusted R Squared = ,310)

Ikke nogen særlige forskelle på effekten af vitamin D indtag, afhængig af solvaner.

Interaktionen sunexp*lvitdintake, II

Parameter Estimates

Dependent Variable: lvitd

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5,600	,416	13,459	,000	4,780	6,420
[country=1]	,321	,122	2,631	,009	,080	,561
[country=2]	,261	,124	2,111	,036	,017	,505
[country=4]	,462	,128	3,610	,000	,210	,714
[country=6]	0[a]	.	.	.	.	.
bmi	-,033	,011	-3,153	,002	-,054	-,012
[sunexp=1]	-,508	,281	-1,808	,072	-1,063	,046
[sunexp=2]	-,413	,293	-1,409	,160	-,992	,165
[sunexp=3]	0[a]	.	.	.	.	.
[sunexp=1] * lvitdintake	,281	,055	5,116	,000	,173	,389
[sunexp=2] * lvitdintake	,268	,054	4,922	,000	,161	,375
[sunexp=3] * lvitdintake	,173	,077	2,239	,026	,021	,325
lvitdintake	0[a]	.	.	.	.	.

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

Ikke nogen særlige forskelle på effekten af vitamin D indtag, afhængig af solvaner, dog mindre effekt for solelskere....

Interaktionen country*sunexp (se s. 116)

Tests of Between-Subjects Effects

Dependent Variable: lvitd

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	45,847[a]	13	3,527	9,408	,000
Intercept	106,407	1	106,407	283,849	,000
country	1,166	3	,389	1,037	,377
bmi	2,881	1	2,881	7,686	,006
lvitdintake	20,926	1	20,926	55,822	,000
country * sunexp	5,502	6	,917	2,446	,026
sunexp	,637	2	,319	,850	,429
Error	74,599	199	,375		
Total	6007,382	213			
Corrected Total	120,447	212			

a. R Squared = ,381 (Adjusted R Squared = ,340)

Her er der en signifikant interaktion.

Vi forsøger at forstå den ved at se på effekten af `sunexp` for hvert land separat

Effekt af sol, opdelt efter land

Parameter Estimates

Dependent Variable: lvitd

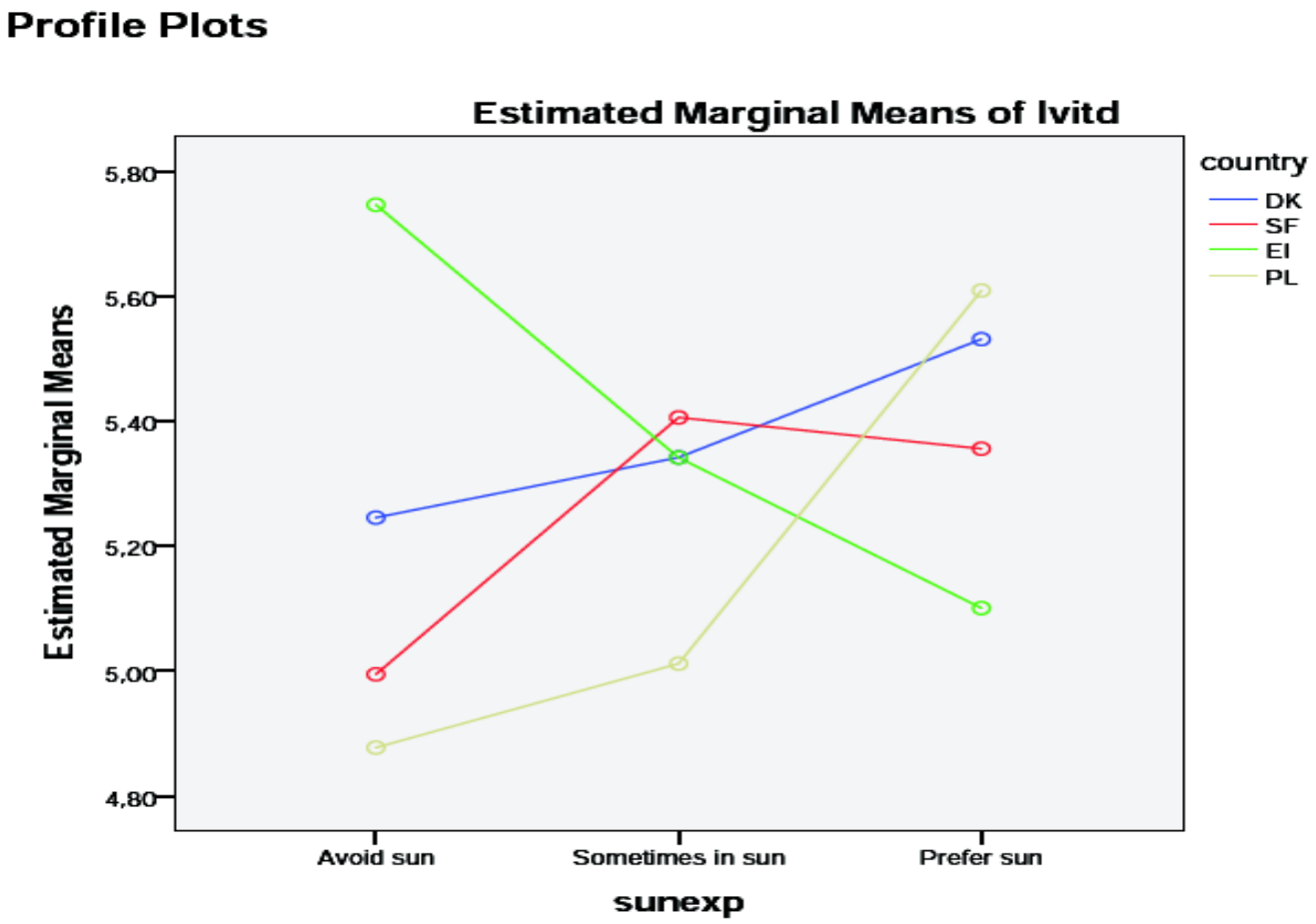
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Intercept	5,672	,422	13,434	,000	4,839	6,504
[country=1]	-,078	,320	-,244	,808	-,708	,552
[country=2]	-,253	,324	-,782	,435	-,891	,385
[country=4]	-,508	,412	-1,234	,219	-1,321	,304
[country=6]	0[a]	.	.	.	.	.
bmi	-,029	,010	-2,772	,006	-,049	-,008
lvitdintake	,263	,035	7,471	,000	,194	,333
[country=1] * [sunexp=1]	-,286	,228	-1,253	,212	-,735	,164
[country=1] * [sunexp=2]	-,189	,202	-,937	,350	-,588	,209
[country=1] * [sunexp=3]	0[a]	.	.	.	.	.
[country=2] * [sunexp=1]	-,361	,228	-1,581	,116	-,811	,089
[country=2] * [sunexp=2]	,050	,200	,250	,803	-,345	,445
[country=2] * [sunexp=3]	0[a]	.	.	.	.	.
[country=4] * [sunexp=1]	,645	,348	1,854	,065	-,041	1,332
[country=4] * [sunexp=2]	,241	,339	,711	,478	-,427	,909
[country=4] * [sunexp=3]	0[a]	.	.	.	.	.
[country=6] * [sunexp=1]	-,732	,300	-2,437	,016	-1,324	-,139
[country=6] * [sunexp=2]	-,597	,297	-2,014	,045	-1,182	-,012
[country=6] * [sunexp=3]	0[a]	.	.	.	.	.
[sunexp=1]	0[a]	.	.	.	.	.
[sunexp=2]	0[a]	.	.	.	.	.
[sunexp=3]	0[a]	.	.	.	.	.

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

Der ser ud til at være en tendens til effekt af `sunexp` i Polen og Irland, men ikke i Danmark og Finland.

Men mønstret for Irland er ret svært at forstå....se figur næste side

Illustration af country*sunexp



Hvad sker der for Irland?

Interaktionen country*lvitdintake (se s. 117)

Tests of Between-Subjects Effects

Dependent Variable: lvitd

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	45,451[a]	10	4,545	12,242	,000
Intercept	118,219	1	118,219	318,421	,000
country	5,481	3	1,827	4,921	,003
bmi	4,010	1	4,010	10,800	,001
sunexp	1,084	2	,542	1,459	,235
country * lvitdintake	5,106	3	1,702	4,584	,004
lvitdintake	14,037	1	14,037	37,808	,000
Error	74,996	202	,371		
Total	6007,382	213			
Corrected Total	120,447	212			

a. R Squared = ,377 (Adjusted R Squared = ,347)

Her ses overraskende nok en signifikant interaktion...se næste side

Interaktionen country*lvitdintake, II

Parameter Estimates

Dependent Variable: lvitd

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	5,451	,367	14,874	,000	4,729	6,174
[country=1]	-,211	,245	-,860	,391	-,695	,273
[country=2]	,493	,353	1,395	,165	-,204	1,190
[country=4]	,827	,279	2,967	,003	,277	1,376
[country=6]	0[a]	.	.	.	.	.
bmi	-,034	,010	-3,286	,001	-,054	-,013
[sunexp=1]	-,217	,127	-1,707	,089	-,467	,034
[sunexp=2]	-,135	,119	-1,135	,258	-,370	,100
[sunexp=3]	0[a]	.	.	.	.	.
[country=1] * lvitdintake	,430	,062	6,923	,000	,307	,552
[country=2] * lvitdintake	,166	,087	1,897	,059	-,007	,338
[country=4] * lvitdintake	,086	,081	1,060	,290	-,074	,245
[country=6] * lvitdintake	,228	,057	3,993	,000	,116	,341
lvitdintake	0[a]	.	.	.	.	.

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

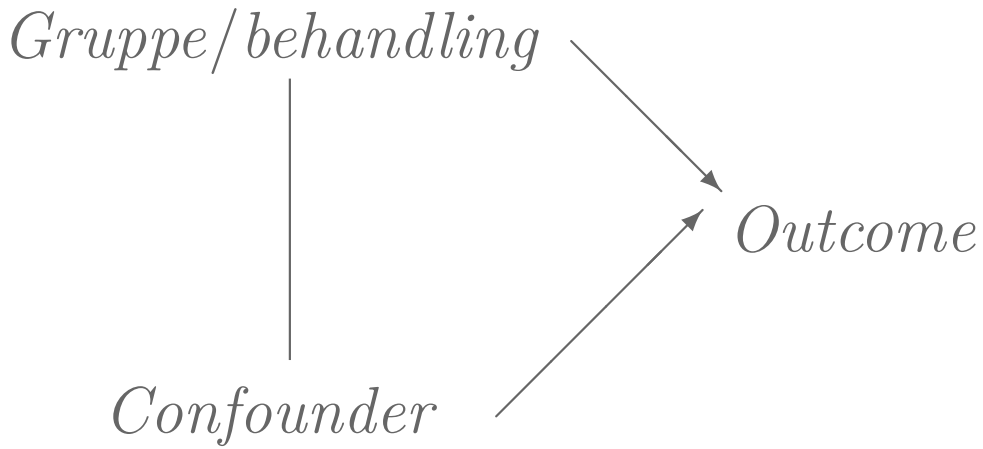
Den signifikante interaktion ses mestendels at skyldes, at effekten af vitamin D indtaget er langt større i Danmark end i de øvrige lande.

Det har formentlig en ret speciel forklaring.....

Repetition: Sammenligning af to grupper

- som ikke er helt sammenlignelige, pga en confounder, som er:
En variabel, som

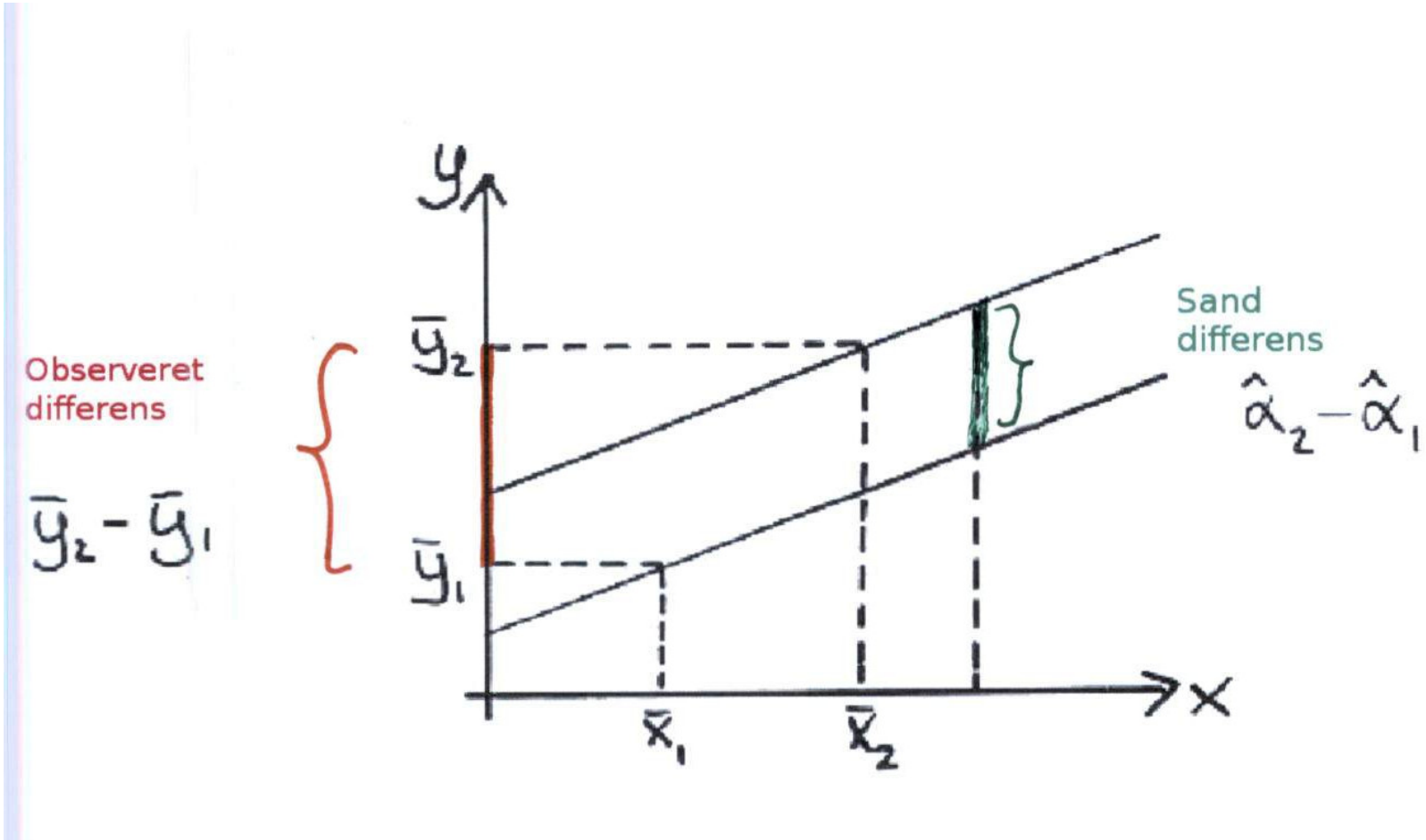
- har en effekt på outcome
- er relateret til gruppen
(der er forskel på værdierne i de to grupper)



Eksempel: Vægt vs. køn og højde

Illustration af confounding og kovariansanalyse

Kovariaten x er her en confounder for gruppeforskellen:



Eksempel om mænds og kvinders vægt

fra forelæsningen om kovariansanalyse:

Vægt vs. køn, Outcome er `log10vaegt`:

Kovariater	Mænd vs. kvinder ratio (CI)	P-værdi
kun kønnet	1.14 (1.07, 1.23)	0.0002
køn og højde	1.04 (0.97, 1.12)	0.28

Den observerede forskel i ($\log_{10}$) vægt mellem mænd og kvinder *kan* altså tilskrives højdeforskellen mellem kønnene.

Alternativt eksempel

Det kan **også** forekomme, at

- Tilsyneladende ens grupper (f.eks. blodtryk hos mænd og kvinder) udviser forskelle, når der bliver korrigeret for inhomogeniteter (f.eks. fedmegrad)

Vi så også dette i eksemplet med P-piller og hormoner i ANCOVA-forelæsningen

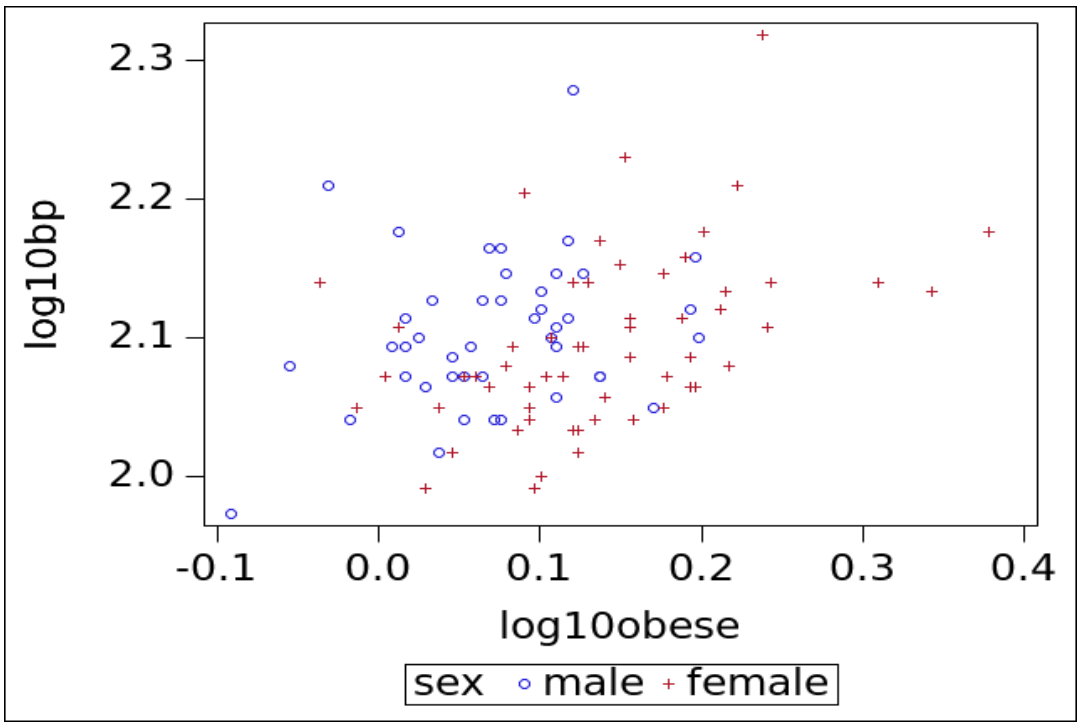
Man skal (på protokolstadiet) nøje overveje, hvilke variable med potentiel betydning for outcome, der skal medtages i modellen!

- ... uden at gå for meget på fisketur!!
- og man skal huske at tænke på, at fortolkningen skifter afhængig af de øvrige kovariater i modellen

Eksempel: Fedmegrad og blodtryk

Systolisk blodtryk (bp) vs. fedmegrad = vægt/idealvægt (obese):

begge på logaritmisk skala (se s. 118, dog er figuren her ikke lavet i SPSS)



Resultater, Blodtryk vs. køn

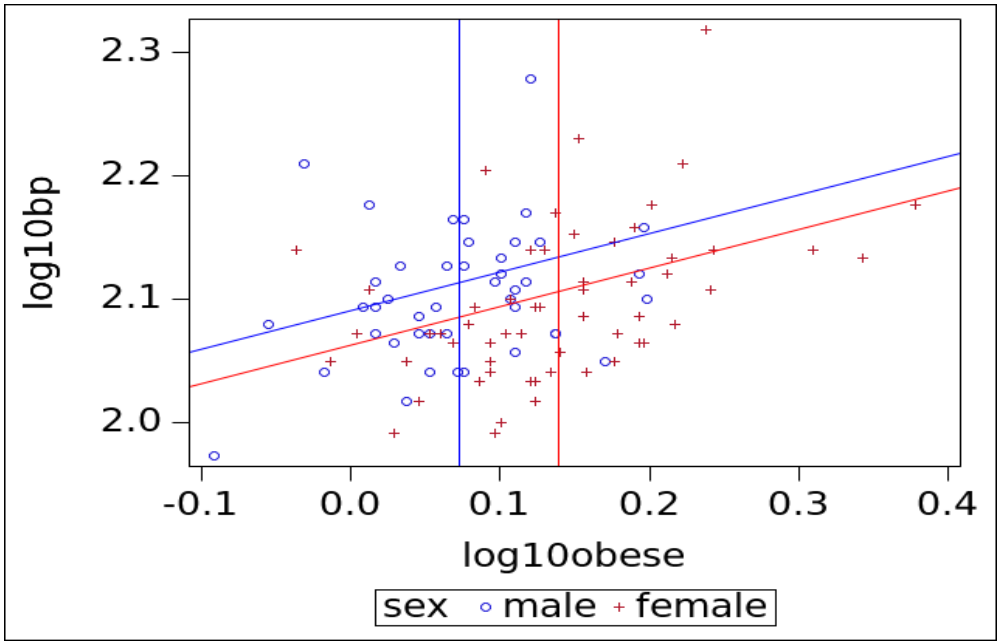
Outcome `log10bp`:

Kovariat	Mænd vs. kvinder ratio (CI)	P-værdi
kun kønnet	1.02 (0.96, 1.07)	0.56
køn og fedmegrad	1.07 (1.01, 1.13)	0.02

Fedmegrad er en confounder for kønnet, idet der er forskel på fedmegrad for mænd og kvinder. Kvinder estimeres til en fedmegrad på 16.7% højere end mænd (CI: 9.3%-24.5%)

Illustration af kovariansanalysen

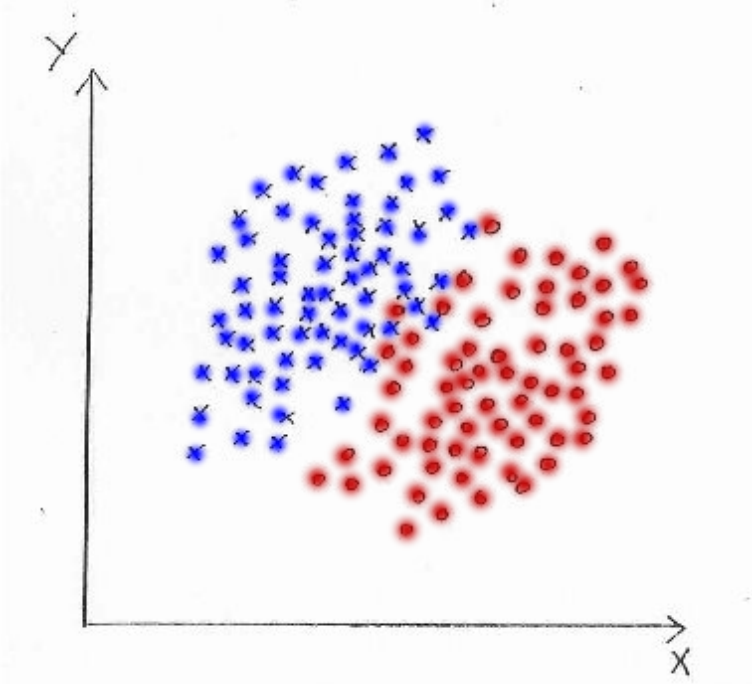
To parallelle linier (figuren her er lavet i SAS, se s. 120)
Samme relation til fedmegrad for de to køn
ikke lavet i SPSS



Forsigtig konklusion:
Kvinder har ligeså højt blodtryk som mænd, fordi de er federe...

Husk også de tidligere eksempler på confounding

Kolesterol vs. chokoladespisning og køn....



Kolesterol og chokoladespisning er

- **positivt** relaterede for hvert køn separat
- **negativt** relaterede for mennesker

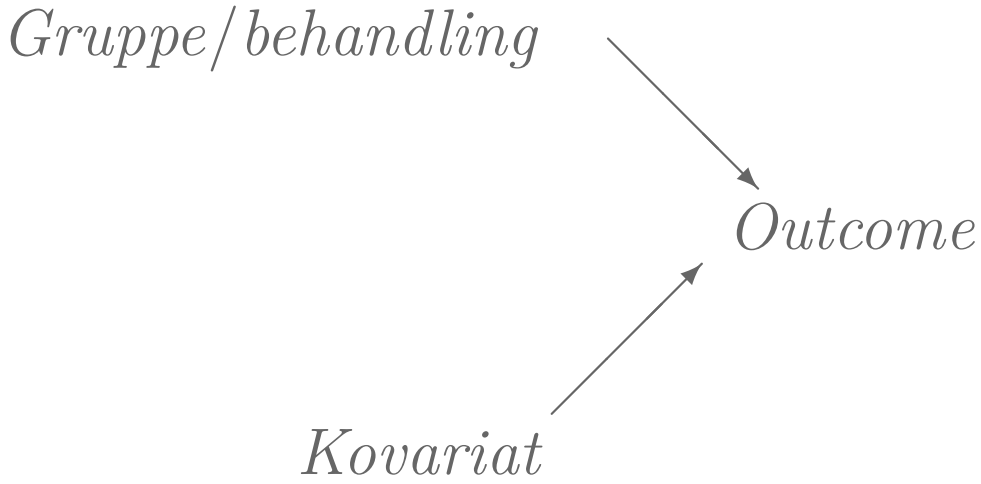
Ingen særlig kønsforskel i kolesterol – og dog...

Vi så det også i eksemplet med hjernevægt hos mus

Men læg mærke til følgende:

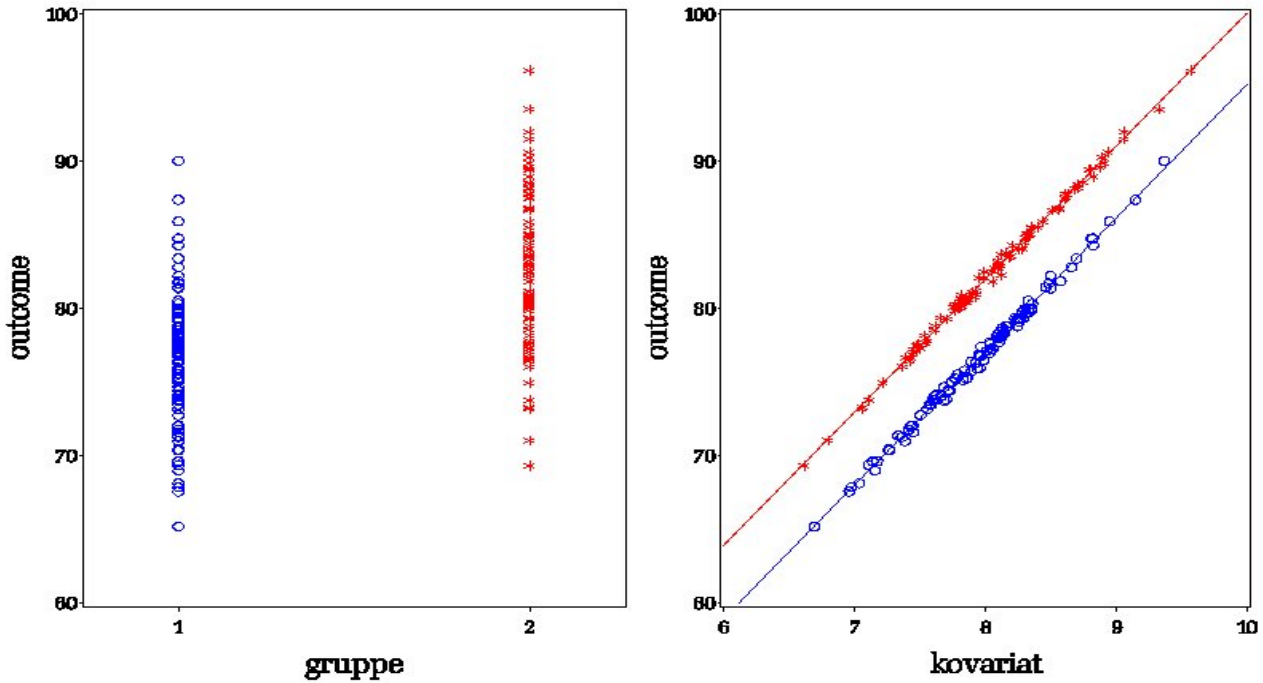
Selv om fordelingen af kovariaten er ens i de to grupper, kan det være af stor betydning at medtage den i analysen.

Det giver **større styrke!**



Men vi svarer samtidig på et andet videnskabeligt spørgsmål!!

Simuleret eksempel



Uden x i modellen: Ingen særlig forskel på grupperne...?

Med x i modellen: Tydelig forskel på grupperne
(her den lodrette afstand mellem linierne)

Effekt af at medtage en ekstra forklarende variabel

- Besvarelse af et andet videnskabeligt spørgsmål
- undgå at maskere forskel, f.eks. en nedsættelse af hormonkoncentrationen ved indtagelse af P-piller (fordi man sammenligner unge P-pille brugere med ældre kvinder)
- nedsættelse af residualvariationen, med deraf følgende lavere standard errors, dvs. større styrke

Hvis kovariaten *ikke* er vigtig, risikerer man

- at *forøge* residualvariationen lidt (fordi man har færre frihedsgrader) og at forøge standard errors meget, hvis kovariaten er korreleret til nogle af dem, der allerede er medtaget (*kollinearitet*) .

Appendix

med uddybende kommentarer til diverse slides:

- Biokemisk iltforbrug, transformationer, s. 93-99
- Ikke-lineær regression, s. 100-101, 107
- Udglatning og polynomie-fit, s. 102-103
- Lineære splines, s. 104-106
- Vitamin D, GLM, s. 108-114
- Interaktioner, s. 115-117
- Blodtryk og fedme, s. 118-120

Data vedr. biokemisk iltforbrug

Slide 7

Omstrukturering fra 6 linier med oplysning om alle 4 flasker den pågældende dag til 24 linier med kun en enkelt måling af iltforbrug pr. linie:

Benyt `/data/Restructure` og afkryds
`Restructure selected variables into cases`, klik `Next`.
Under `How many variable groups?` afkrydses `One`, klik `Next`,

Under `Case Group Identification` skiftes til
`Use selected variable`, og `days` sættes over i `Variable`.

Omstrukturering, fortsat, samt figur

Slide 7

Herefter går man ind i `Variables to be Transposed`, skriver `boc` i `Target Variable` og sætter `boc1`, `boc2`, `boc3` og `boc4` i det store felt, hvorefter der klikkes `Next`.
Endelig går man i `Create index variable` og sætter kryds ved `None` og klikker `Next`, og under `Variables to Cases: Options` klikkes blot `Finish`

Figuren slide 8

Benyt `Graphs/Chart Builder` og vælg `Scatter` (det simple længst til venstre), og dobbeltklik det op i det store felt. Sæt `days` over på X-aksen og `boc` over på Y-aksen.

Analyse af biokemisk iltforbrug

Slide 10

Definition af nye variable:
Først skal vi logaritmetransformere iltforbruget. Dette gøres under `Transform/Compute`, hvor man som `Target Variable` sætter det nye variabelnavn, her `logboc`, og i feltet `Numeric Expression` skriver `Ln(boc)`

Dernæst skal vi udregne den reciprokke tid, igen under `Transform/Compute`, hvor man som `Target Variable` sætter `invdays`, og i feltet `Numeric Expression` skriver `1/days`

Derefter følger man opskriften på simpel lineær regression, se de næste sider

Regressionsanalyse med transformerede variable

Slide 11

Gå ind i menuen `Analyze/Regression/Linear`, og i boksen sættes `logboc` som `Dependent` og `invdays` som `Independent(s)`

Man skal efterfølgende huske at gå ind i `Statistics` og afkrydse `Parameter Estimates` og `Confidence intervals`

Plot med regressionslinie

Slide 12

Først laves plottet ved at gå ind i `Graph/Chart Builder/Scatter` og i den fremkomne boks trækker man `logboc` over på Y-aksen, og `invdays` over på X-aksen.

For at tegne linien dobbeltklikker man efterfølgende på grafen og klikker på ikonet `Add Fit Line at Total` og derefter i `Properties`-boksen afkrydse `Linear` og klikke `Apply`.

Man kan endvidere vælge, om man vil have liniens ligning skrevet på linien eller ej (flueben ved `Attach label to line` kan fjernes) Farven på linien kan vælges under `Lines`: klik på farven og `Apply/Close`

*Tilbagetransformeret relation

Slide 14

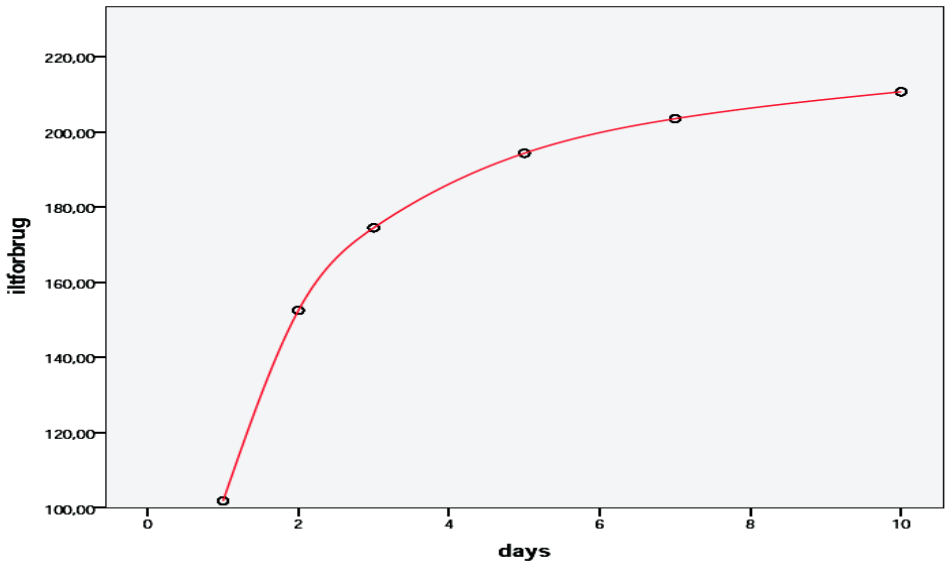
Her må vi tilbagetransformere den lineære relation, hvilket er *lidt besværligt*:
I regressionsopsætningen fra s. 96 klikkes på `Save`, hvorefter man under `Predicted Values` vælger `Unstandardized`, som så kommer til at ligge i datasættet under navnet `PRE_1`.

Disse tilbagetransformeres til oprindelig skala ved at benytte `Transform/Compute` og som `Target Variable` sætte f.eks. `predikteret` og i `Numeric Expression` skrive `exp(PRE_1)`.

Gå nu ind i `Graph/Chart Builder`, vælg `Scatter`, og træk i den fremkomne boks `predikteret` over på Y-aksen og `days` over på X-aksen. Dobbeltklik på den fremkomne graf og klik på ikonet `Add Fit Line at Total` og derefter i `Properties`-boksen afkrydse `Spline` og klikke `Apply`.

Tilbagetransformeret relation, II

Dette bliver resultatet i SPSS....



Det er formentlig meget besværligt (eller umuligt) at få såvel kurve som observationer på samme figur.....

Ikke-lineær regression

Slide 15-16

Her benyttes `Analyze/Regression/Nonlinear Regression`, hvor man sætter `boc` i `Dependent` og i `Model Expression` skriver selve det ikke-lineære udryk `gamma*exp(-beta/days)`.
Herefter klikker man i `Parameters` og tilføjer

- `Name: gamma, Starting Value: 228 Add`
- `Name: beta, Starting Value: 0,8 Add`

og til sidst: `Continue`

Plot af ikke-lineært fit

Slide 17

Fra opsætningen s. 100 benyttes `Save` for at gemme de predikterede værdier: Kryds af under `Predicted Values`.

Herefter benyttes `Graphs/Chart Builder/Scatter` (det simple længst til venstre), `days` sættes på X-aksen og `PRED_1` på Y-aksen.

Efterfølgende dobbeltklikkes på grafen, og man vælger `Add Interpolation Line`, hvorefter man afkrydser i `Spline`.

Det ser dog ikke umiddelbart ud til at være muligt at få usikkerheder på disse prediktioner, så grænserne må man undvære. Ligeledes kommer selve observationerne heller ikke med på figuren.

Udglattet kurve, en såkaldt *Loess*-kurve

Slide 22

Her skal vi begrænse det datamateriale, vi ser på. Dette gøres i `Data/Select Cases`, hvor der afkrydses i `If` og skrives `5<age<20 & sex='male'`.

Herefter laver vi en tegning med `Graphs/Chart Builder`, hvor vi vælger det simple `Scatter`. Sæt age over på X-aksen og `sigf1` over på Y-aksen.

Ved efterfølgende at dobbeltklikke på grafen, kan man lægge en udglattet kurve på ved at klikke på `Add Fit Line at Total` og derefter i `Properties`-boksen afkrydse `Loess/Apply`:

Farven på kurven kan vælges under `Lines`: klik på farven og `Apply/Close`

Polynomiale fit

Slide 26

De 3 forskellige polynomier (af grad hhv 1,2 og 3) kan formentlig kun (på rimelig simpel vis) tegnes hver for sig. Ellers skal man til at have flere versioner af datasættet sat i forlængelse af hinanden..... Først laver vi en tegning med `Graphs/Chart Builder`, hvor vi vælger det simple `Scatter`, sætter age over på X-aksen og `sigf1` over på Y-aksen.

Ved efterfølgende at dobbeltklikke på grafen, og videre på `Add Fit Line at Total` kan man i `Properties`-boksen afkrydse enten `Linear`, `Quadratic` eller `Cubic`

Farven på kurverne kan vælges under `Lines`: klik på farven og `Apply/Close`

Lineære splines

Slide 27-32

Definition af nye variable:
Benyt `Transform/Compute`, hvor man som `Target Variable` sætter det nye variabelnavn, her f.eks. `extra_age12`, og i feltet `Numeric Expression` skriver `max(age-12,0)`
Ligeledes med de 3 øvrige variable.
Fit derefter en model med disse 4 ekstra kovariater, sammen med den oprindelige age ved at benytte menuen `Analyze/Regression/Linear`, og i boksen sættes `ssigf1` som `Dependent` og age samt alle 4 `extra`-variable i `Independent(s)`

Man skal efterfølgende huske at gå ind i `Statistics` og afkrydse `Parameter Estimates` og `Confidence intervals`

Illustration af fit med lineære splines

Slide 29

Det er nok virkelig indviklet (eller umuligt?) at få data og splines på samme figur....men man kan tegne selve fittet ved først at gemme de predikterede værdier ved at benytte `Save` og vælge `Predicted Values`. De kommer til at hedde `PRE_1` i datasættet.

Herefter går man i Herefter laver vi en tegning med `Graphs/Chart Builder`, hvor vi vælger `Lines` (det simple længst til venstre), og dobbeltklik det op i det store felt. Sæt `PRED_1` over på Y-aksen og age over på X-aksen.

Alternativ parametrisering af lineære splines

Slide 33-34

så man i stedet får estimater for
hældningerne i de enkelte aldersintervaller

Definition af de alternative variable:
Benyt `Transform/Compute`, hvor man som `Target Variable` sætter det nye variabelnavn, her f.eks. `ny_age12`, og i feltet `Numeric Expression` skriver `min(extra_age12,1)`

Ligeledes med de 3 øvrige variable: `min(age,10)`, `min(extra_age10,2)` og `min(extra_age13,2)` og `ny_age15=extra_age15`

Fit derefter en model med disse 5 kovariater, i analogi med den beskrevne fremgangsmåde på forrige side.

Ikke-lineær regression

Slide 38-41

Her benyttes `Analyze/Regression/Nonlinear Regression`, hvor man sætter `konc` i `Dependent` og i `Model Expression` skriver selve det ikke-lineære udryk `beta*(1-exp(-gamma*tid))`.

Herefter klikker man i `Parameters` og tilføjer

- `Name: gamma, Starting Value: 0,05 Add`
- `Name: beta, Starting Value: 2000 Add`

og til sidst: `Continue`

Mht figur af fit, se tilsvarende s. 17

ANOVA, sammenligning af D-vitamin i 4 lande

Slide 47

Først skal vi udregne 2-tals logaritmen af vitamin D. Dette gøres under `Transform/Compute`, hvor man som `Target Variable` sætter det nye variabelnavn, her `lvitd`, og i feltet `Numeric Expression` skriver `LG10(vitd)/LG10(2)`

Til variansanalysen benyttes `Analyze/General Linear Model/Univariate`, hvor vi sætter `lvitd` i `Dependent Variable` og `country` i `Fixed Factor(s)`.

Her kan vi vælge at se parameterestimater under `Options`, hvor vi afkrydser `Parameter estimates`)

Model med bmi som kovariat, ANCOVA

Slide 51

Til kovariansanalyse benyttes `Analyze/General Linear Model/Univariate`, hvor vi sætter `lvitd` i `Dependent Variable`, `country` i `Fixed Factor(s)` og `bmi` 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`

Model med 4 kovariater

Slide 55-56

Selve analysen foregår som skitseret på s. 109, altså: Benyt `Analyze/General Linear Model/Univariate`, hvor vi sætter `lvitd` i `Dependent Variable`, `country` og `sunexp` i `Fixed Factor(s)` og såvel `bmi` som `lvitdintake` i `Covariate(s)`

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

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

Parvise sammenligninger

Slide 59-62

For at foretage alle parvise sammenlininger af lande og sol-grupper i opsætningen med
`Analyze/General Linear Model/Univariate`, går man ind i `Options`, hvor man sætter såvel `country` som `sunexp` over i `Display Means for`.

Desuden afkrydses i `Compare main effects`, hvor man evt. også kan vælge, at der skal korrigeres for multiple sammenligninger (her er dog bare brugt `LSD(none)`, da vi tænkes at være i situatione med præ-specificerede sammenligninger).

Modelkontrol for model med 4 kovariater

Slide 66-67

Den automatiske modelkontrol i SPSS er ikke særlig god.....
Brug i stedet nedenstående:

I regressionsopsætningen fra s. 20 klikkes på `Save`, hvorefter man typisk vil vælge

- Under `Predicted Values`: vælg `Unstandardized`, som så kommer til at ligge i datasættet under navnet `PRE_1`
- Under `Residuals` er der 5 muligheder, f.eks.
  - `Unstandardized`: De *sædvanlige*, altså observeret minus forventet, kaldet `RES_1` i datasættet
  - `Studentized deleted`: Normerede Press-residualer (altså hvor den pågældende observation er udeladt ved fit af modellen), kaldet `SDR_1` i datasættet

Modelkontrol, II

Slide 66-67

Ud fra det udbyggede datasæt dannet s. 112 kan vi nu lave diverse figurer:

- Benyt `Graph/Chart Builder`, vælg `Scatter`, klik `Unstandardized Predicted Values` over på X-aksen og `Unstandardized Residuals` over på Y-aksen.
- Vælg `Analyze/Descriptive Statistics/Q-Q Plots` og sæt `Unstandardized Residuals` over i `Variables`.
- Benyt `Graph/Chart Builder`, vælg `Histogram` og klik det ønskede residual (f.eks. `RES_1`) ned på X-aksen.
  Klik så på `Distribution Curve` og afkryds `Normal`, klik `Apply` og `Close`
- Evt. yderligere check af varianshomogenitet, se tidligere forelæsninger.

Udeladelse af flere kovariater samtidig

Slide 71-72

Det er ikke så let at gennemskue, hvordan dette foregår i SPSS....

Hvis alle kovariater var kvantitative (altså i en multipel regressiosanalyse) kan man i `Analyze/Regression/Linear` definere flere modeller på en gang og få dem sammenlignet med et F-test.

Men når vi som her også har kategoriske kovariater, er vi nødt til bruge `Analyze/General Linear Model/Univariate`, og denne opsætning tillader ikke definition af flere modeller på en gang.

I så fald er man nødt til at lave sine kategoriske variable om til dummy-variable, f.eks. skal `country` så blive til 3 dummy-variable....

Interaktionen *sunexp*lvitdintake*

Slide 74-75

Følg opskriften fra s. 110, men i `Model` markeres `sunexp` og `lvitdintake` samtidig, der skiftes fra `Main Effects` til `Interaction` og klikkes på pilen og `Continue`.

For at få de enkelte effekter af `lvitdintake` skal denne udelades af `Model`, medens interaktionen bibeholdes.

Interaktionen *country*sunexp*

Slide 76-78

Følg opskriften fra s. 110, men i `Model` markeres `sunexp` og `country` samtidig, der skiftes fra `Main Effects` til `Interaction` og klikkes på pilen og `Continue`.

For at få de enkelte effekter af `sunexp` skal denne udelades af `Model`

Interaktionen *country*lvitdintake*

Slide 79-80

Følg opskriften fra s. 110, men i `Model` markeres `country` og `lvitdintake` samtidig, der skiftes fra `Main Effects` til `Interaction` og klikkes på pilen og `Continue`.

For at få de enkelte effekter af `lvitdintake` skal denne udelades af `Model`, medens interaktionen bibeholdes.

Blodtryk vs. fedme

Sammenligning af mænd og kvinder, figurer

Slide 85

Selve figuren laves i `Graph/Chart Builder`, hvor man vælger `Scatter plot` (nr. 2 fra venstre), og i den fremkomne boks trækker man `log10bp` over på Y-aksen, `log10obese` over på X-aksen og `sex` 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`.

Disse linier vil ikke blive parallelle, og er ikke vist i forelæsningsnoterne.

Blodtryk vs. fedme og køn: Kovariansanalyse

Slide 86

Her benyttes `Analyze/General Linear Model/Univariate`, hvor vi sætter `log10bp` over i `Dependent Variable`, `sex` i `Fixed Factor(s)` og `log10obese` 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`

ANCOVA-plot, med parallelle linier

Slide 87

Her er man nødt til først at gemme de predikterede værdier ved at benytte `Save` og vælge `Predicted Values`. De kommer til at hedde `PRE_1` i datasættet.

Herefter går man i igen ind i `Graph/Chart Builder/Scatter`, vælger `Scatter plot` (nr. 2 fra venstre), og i den fremkomne boks udskiftes `log10bp` med `PRE_1`.

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`.
Desværre får man ikke samtidig selve observationerne på figuren.....
Og de to lodrette linier er heller ikke forsøgt indtegnet.