Faculty of Health Sciences

Basal statistik

Den generelle lineære model mv.

Lene Theil Skovgaard

23. marts 2020

Den generelle lineære model mv.

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

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

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

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)

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Biologisk iltforbrug

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

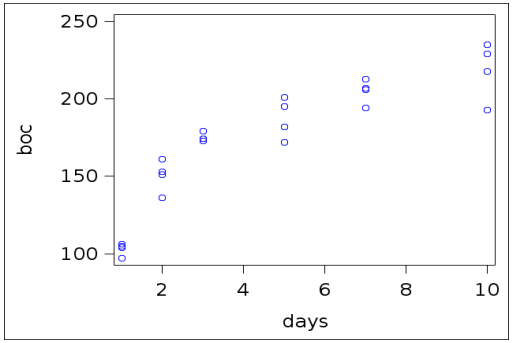
4 flasker til hvert tidspunkt

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

Kode til omstrukturering af data, samt figur næste side: se s. 87-88

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

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 på kode s. 87
Bemærk, at vi her har benyttet den *naturlige* logaritme til at transformere iltforbruget (dette valg kommenteres s. 19)

Output:

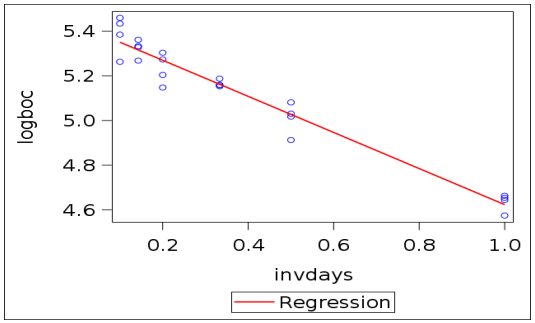
```
The GLM Procedure
Dependent Variable: logboc

R-Square     Coeff Var      Root MSE    logboc Mean
0.951757      1.140483      0.058447       5.124796

                                Standard
Parameter          Estimate         Error    t Value    Pr > |t|
Intercept       5.431253025    0.01893998     286.76      <.0001
invdays        -0.807814928    0.03877543     -20.83      <.0001

Parameter        95% Confidence Limits
Intercept       5.391973906  5.470532143
invdays        -0.888230241 -0.727399614
```

Den transformerede relation



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

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

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

Her må vi tilbagetransformere den lineære relation
fra analysen s. 11, hvilket er *lidt besværligt*:

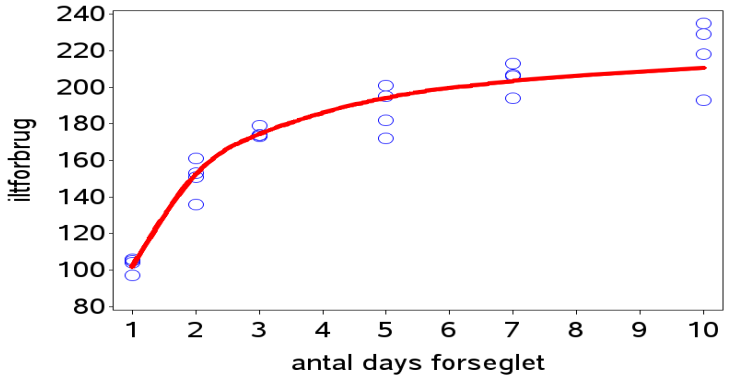
```
proc glm data=boc;
  model logboc=invdays / clparm;
output out=ny1 p=yhat r=resid;
run;

data fit1;
set ny1;

art=1; output;
art=2; boc=exp(yhat); output;
run;

proc gplot data=fit1;
  plot boc*days=art
  /  nolegend haxis=axis1 vaxis=axis2 frame;
axis1 offset=(3,3) value=(H=3) minor=NONE label=(H=3 'antal days forseglet');
axis2 order=(80 to 240 by 20) value=(H=3) minor=NONE label=(A=90 R=0 H=3 'iltforbrug');
symbol1 v=circle i=none c=blue l=1 w=2 h=3 w=1;
symbol2 v=none i=spline c=red l=1 w=2 h=3 w=5;
run;
```

Tilbagetransformeret relation, II



Analyse på original skala

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

```
proc nlin plots=all data=boc;
  parms beta=0.8 gamma=228;
  model boc=gamma*exp(-beta/days);
run;
```

med output:

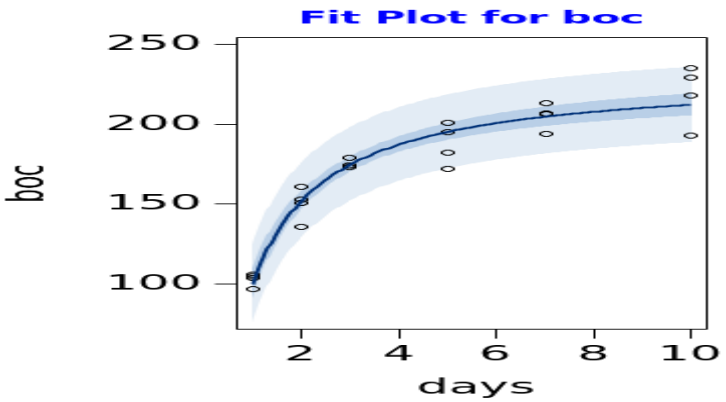
Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
beta	0.8327	0.0576	0.7133	0.9521
gamma	230.6	4.5384	221.2	240.0

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

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Fit fra ikke-lineær regression

Automatisk plot via `ods graphics`



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.

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

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

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

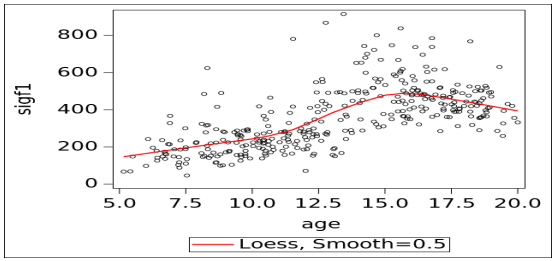
Modeller med logaritmetransformerede data

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 børn, op til 20-års alderen

```
proc sgplot data=juul; where age ge 5 and age le 20 and sex='male';
    loess Y=sigf1 X=age / smooth=0.5 lineattrs=(color=red);
run;
```



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:

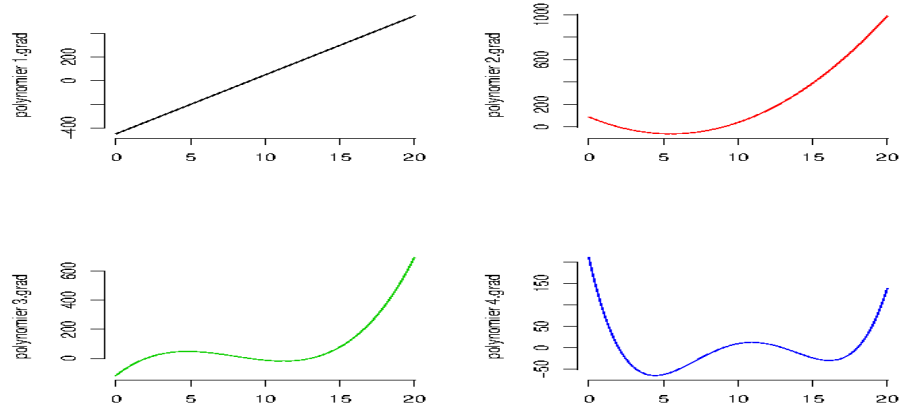
$$y = \beta_0 + \beta_1 x$$

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



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

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

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

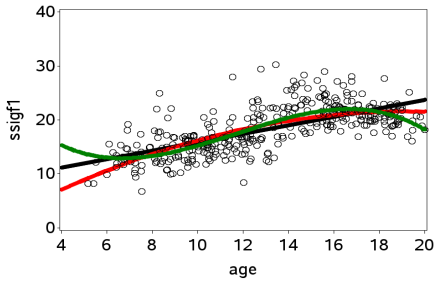
er det bare en sædvanlig **lineær multipel regression**

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

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 (kode. s. 90)

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

Specielt ude i enderne kan de opføre sig meget underligt.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

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.

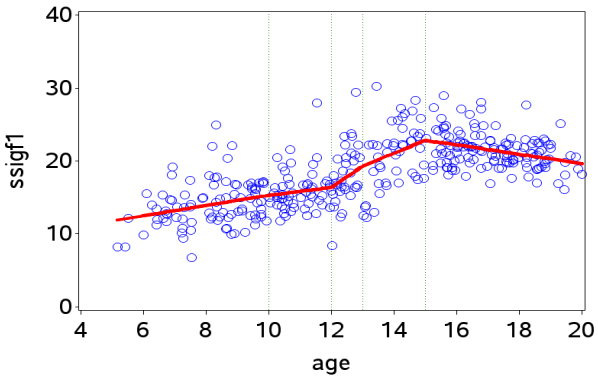
Fortolkning af parametre

- 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.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Lineær spline for Serum IGF-1



Estimater:

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

(kode s. 91-92)

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

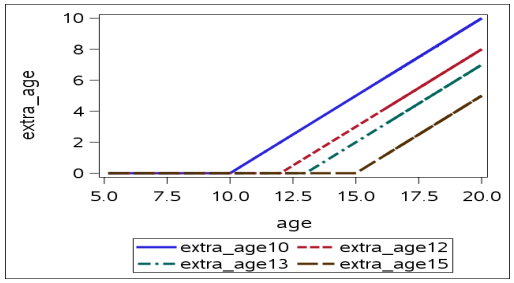
```
extra_age10=max(age-10,0);
extra_age12=max(age-12,0);
extra_age13=max(age-13,0);
extra_age15=max(age-15,0);
```


UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

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

Kode s. 91

Parameter Estimates

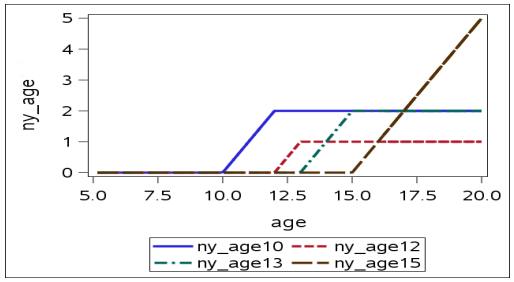
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.28285	1.86707	4.44	<.0001
age	1	0.70448	0.21803	3.23	0.0013
extra_age10	1	-0.16250	0.55581	-0.29	0.7702
extra_age12	1	2.38356	1.25125	1.90	0.0576
extra_age13	1	-1.15156	1.28119	-0.90	0.3694
extra_age15	1	-2.41836	0.53872	-4.49	<.0001

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

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

∗Alternativ parametrisering af lineære splines

Hvis man ændrer de nye kovariater til disse
(se kode s. 93):



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

∗Output fra alternativ parametrisering

af lineære splines (kode s. 93)

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

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.28285	1.86707	4.44	<.0001
ny_age	1	0.70448	0.21803	3.23	0.0013
ny_age10	1	0.54198	0.40450	1.34	0.1811
ny_age12	1	2.92554	0.95020	3.08	0.0022
ny_age13	1	1.77398	0.42037	4.22	<.0001
ny_age15	1	-0.64439	0.17507	-3.68	0.0003

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

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

Koden bliver her:

```
proc nlin plots=all data=a1;
  parms beta=2000 gamma=0.05;
  model koncentration=beta*(1-exp(-gamma*tid));
run;
```

∗Output fra ikke-lineær regression

Koden ses på s. 37

Dependent Variable conc

Iterative Phase

Iter	beta	gamma	Sum of Squares
0	2000.0	0.0500	4370990
1	1958.6	0.0824	382995
2	2169.2	0.0769	26355.0
3	2174.2	0.0772	25286.7
4	2174.0	0.0773	25286.7
5	2174.0	0.0773	25286.7

NOTE: Convergence criterion met.

Konvergens betyder, at et "*stabilt*" fit er fundet

∗Output, fortsat

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	2	63048136	31524068	29920.0	<.0001
Residual	24	25286.7	1053.6		
Uncorrected Total	26	63073423			
Corrected Total	25	3455129			

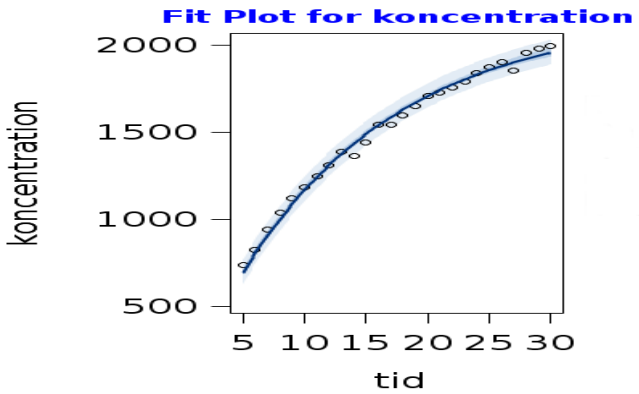
Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
beta	2174.0	28.3459	2115.5	2232.5
gamma	0.0773	0.00226	0.0726	0.0819

Fittet svarende til disse parameterestimater fremgår af s. 40

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)



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:

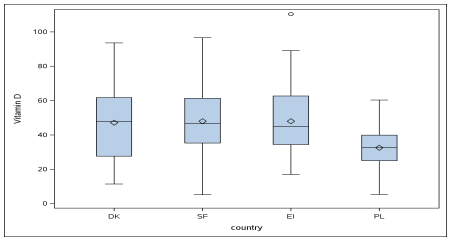
Land	Antal	Vitamin D	Alder	Body Mass Index	Vitamin D Indtag
Denmark	53	47.80	71.51	25.39	8.29
Finland	54	46.60	71.92	27.98	12.41
Ireland	41	44.80	72.05	26.39	5.46
Poland	65	32.50	71.69	29.37	5.16

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 → Vitamin D
Uanset om man ser på utransformerede data eller logaritmetransformerede data, finder man en forskel på landene ($P<0.0001$), 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 (kode s. 95)

```
R-Square     Coeff Var      Root MSE    lvitd Mean

0.107334      13.64313      0.717247      5.257204

Source                   DF   Type III SS   Mean Square  F Value  Pr > F
country                   3   12.92798381    4.30932794     8.38  <.0001

                                                Standard
Parameter                Estimate                   Error    t Value    Pr > |t|
Intercept             5.447274746 B          0.09760501      55.81      <.0001
country   DK         -0.086928888 B          0.13868391      -0.63      0.5315
country   EI          0.009137355 B          0.14857372       0.06      0.9510
country   PL         -0.557730003 B          0.13206536      -4.22      <.0001
country   SF          0.000000000 B           .                .          .
```


Fortolkning af estimater

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

$$2^{0.5577} = 1.47$$

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

$$(2^{0.2974}, 2^{0.8181}) = (1.23, 1.76)$$

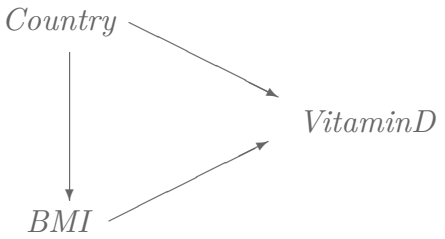
altså fra 23% over til 76% over.

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

(kode s. 96)

```
R-Square     Coeff Var      Root MSE    lvitd Mean
0.151289      13.33494      0.701045      5.257204

Source                     DF   Type III SS   Mean Square  F Value  Pr > F
bmi                         1    5.29427388    5.29427388    10.77  0.0012
country                     3    9.99235920    3.33078640     6.78  0.0002

                                        Standard
Parameter            Estimate             Error    t Value    Pr > |t|
Intercept         6.526049070 B       0.34224508      19.07      <.0001
bmi              -0.038080262         0.01160226      -3.28      0.0012
country   DK     -0.155252807 B       0.13714018      -1.13      0.2589
country   EI     -0.066118048 B       0.14701645      -0.45      0.6534
country   PL     -0.534458623 B       0.12927660      -4.13      <.0001
country   SF      0.000000000 B        .                .          .

Parameter          95% Confidence Limits
Intercept         5.851335255  7.200762884
bmi              -0.060953356 -0.015207167
country   DK     -0.425615716  0.115110102
country   EI     -0.355951368  0.223715271
country   PL     -0.789318993 -0.279598253
country   SF       .            .
```

Fortolkning af nye estimater

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

$$2^{0.5345} = 1.45$$

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

$$(2^{0.2796}, 2^{0.7893}) = (1.21, 1.73)$$

en ubetydelighed lavere end den ujusterede forskel fra s. 47

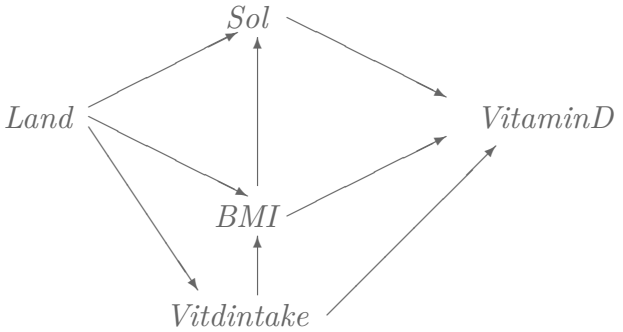
Kunne bmi forklare forskellen på landene?

Nej,

- Selv om `bmi` i sig selv er stærkt signifikant (negativ effekt, estimeret til $-0.0381(0.0116), P = 0.0012$), er der stadig stærkt signifikant forskel på landene, når vi har korrigeret for `bmi` ($P = 0.0002$), 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?
(kode s. 97)

Dependent Variable: lvitd

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	40.3448706	5.7635529	14.75	<.0001
Error	205	80.1018815	0.3907409		
Corrected Total	212	120.4467520			

R-Square	Coeff Var	Root MSE	lvitd Mean
0.334960	11.89021	0.625093	5.257204

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bmi	1	3.79141868	3.79141868	9.70	0.0021
sunexp	2	1.07957598	0.53978799	1.38	0.2535
lvitdintake	1	19.94003315	19.94003315	51.03	<.0001
country	3	5.59595431	1.86531810	4.77	0.0031

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

Output, fortsat

Parameter		Estimate	Standard Error	t Value
Intercept		5.461828843 B	0.34415430	15.87
bmi		-0.032659119	0.01048451	-3.11
sunexp	Avoid sun	-0.071559707 B	0.09850380	-0.73
sunexp	Prefer sun	0.144409552 B	0.12206916	1.18
sunexp	Sometimes in sun	0.000000000 B	.	.
lvitdintake		0.255116100	0.03571243	7.14
country	DK	0.070454998 B	0.12637582	0.56
country	EI	0.211049345 B	0.13675125	1.54
country	PL	-0.248091644 B	0.12280907	-2.02
country	SF	0.000000000 B	.	.

Parameter		Pr > |t|	95% Confidence Limits	
Intercept		<.0001	4.783293016	6.140364671
bmi		0.0021	-0.053330410	-0.011987828
sunexp	Avoid sun	0.4684	-0.265770146	0.122650733
sunexp	Prefer sun	0.2382	-0.096262429	0.385081534
sunexp	Sometimes in sun	.	.	.
lvitdintake		<.0001	0.184705355	0.325526845
country	DK	0.5778	-0.178708008	0.319618004
country	EI	0.1243	-0.058569890	0.480668580
country	PL	0.0447	-0.490222443	-0.005960845
country	SF	.	.	.

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 soldyrkere og solhadere estimeres her til en faktor $2^{0.144+0.072} = 1.16$, altså svarende til 16% større niveau af vitamin D hos soldyrkere i forhold til solhadere.
  Konfidensgrænser: se side 58-60.

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 $2^{0.2481} = 1.19$ 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 $(2^{0.0.0060}, 2^{0.4902}) = (1.004, 1.40)$

Sammenligning mellem andre lande

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

$$2^{0.211+0.248} = 1.37$$

altså svarende til 37% større niveau af vitamin D i Irland sammenlignet med Polen.

For at få konfidensintervallet for denne sammenligning kan man i SAS anvende en `estimate`-sætning
(se kode s. 58 og output s. 59-60).

Irland vs. Polen, og soldyrkere 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

Sidstnævnte skrives som `estimate`-sætninger (se det i sammenhæng i koden s. 97):

```
estimate "1 enhed BMI" bmi 1;
estimate "soldyrkere vs. solhadere" sunexp -1 1 0;
estimate "10% i vitD indtag" lvitdintake 0.1375;
estimate "Finland vs. Polen" country 0 0 -1 1;
estimate "Irland vs. Polen" country 0 1 -1 0;
```

Output fra *estimate*-sætninger

Parameter	Estimate	Standard Error	t Value	Pr > |t|
1 enhed BMI	-0.03265912	0.01048451	-3.11	0.0021
soldyrkere vs. solhadere	0.21596926	0.12996132	1.66	0.0981
10% i vitD indtag	0.03507846	0.00491046	7.14	<.0001
Finland vs. Polen	0.24809164	0.12280907	2.02	0.0447
Irland vs. Polen	0.45914099	0.12765304	3.60	0.0004

Parameter	95% Confidence Limits	
1 enhed BMI	-0.05333041	-0.01198783
soldyrkere vs. solhadere	-0.04026293	0.47220144
10% i vitD indtag	0.02539699	0.04475994
Finland vs. Polen	0.00596085	0.49022244
Irland vs. Polen	0.20745982	0.71082216

Disse skal nu tilbagetransformeres,
men dette kan vi også få gjort automatisk, se kode s. 98, hvoraf vi får

Tilbagetransformerede estimater

Se koden s. 98

Obs	Dependent	Parameter	Estimate	StdErr	tValue
1	lvitd	1 enhed BMI	-0.03265912	0.01048451	-3.11
2	lvitd	soldyrkere vs. solhadere	0.21596926	0.12996132	1.66
3	lvitd	10% i vitD indtag	0.03507846	0.00491046	7.14
4	lvitd	Irland vs. Polen	0.45914099	0.12765304	3.60
5	lvitd	Finland vs. Polen	0.24809164	0.12280907	2.02

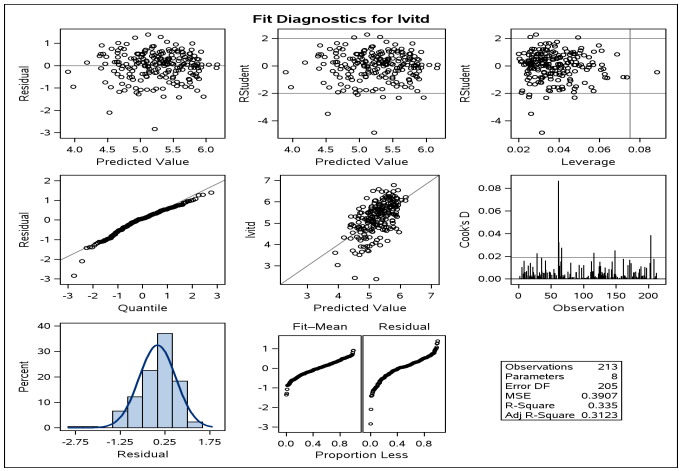
Obs	Probt	LowerCL	UpperCL	faktor	lower	upper
1	0.0021	-0.05333041	-0.01198783	0.97762	0.96371	0.99173
2	0.0981	-0.04026293	0.47220144	1.16148	0.97248	1.38722
3	<.0001	0.02539699	0.04475994	1.02461	1.01776	1.03151
4	0.0004	0.20745982	0.71082216	1.37472	1.15465	1.63674
5	0.0447	0.00596085	0.49022244	1.18764	1.00414	1.40466

Fortolkning af estimate-resultaterne s. 59-60

- Forskellen soldyrkere vs. solhadere estimeres til en faktor $2^{0.216} = 1.16$, altså svarende til 16% større niveau af vitamin D hos soldyrkere i forhold til solhadere. Konfidensgrænserne er $(2^{-0.040}, 2^{0.472}) = (0.97, 1.39)$, altså fra 3% under til 39% over.
- Den estimerede forskel Irland vs. Polen er en faktor $2^{0.459} = 1.37$, nu med konfidensgrænser $(2^{0.207}, 2^{0.711}) = (1.15, 1.64)$, altså således at Irland ligger mellem 15% og 64% over Polen.

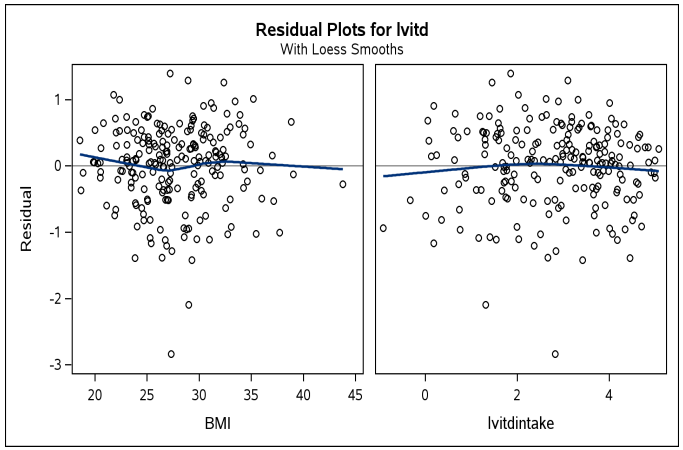
Husk også modelkontrol

Kode s. 97



Modelkontrol, II

Kode s. 97



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)

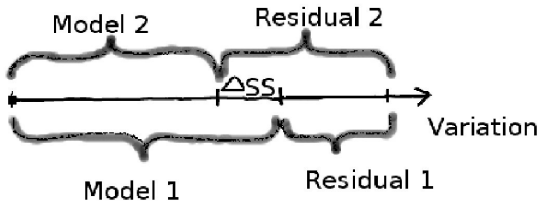
Kan vi forsvare at bruge den simpleste af dem?
Beskriver den data tilstrækkeligt godt?

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: $\Delta\text{SS} > 0$

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 kode s. 99)

Udelad flere kovariater på en gang, II

```
contrast 'fjern alle tre paa en gang'
    bmi 1, sunexp -1 0 1, sunexp 0 -1 1, lvitdintake 1;
```

Se det i sammenhæng i koden s. 99

```
Dependent Variable: lvitd
                                    Sum of
Source                      DF      Squares    Mean Square  F Value  Pr > F
Model                        7   40.3448706      5.7635529    14.75  <.0001
Error                      205   80.1018815      0.3907409

Corrected Total            212   120.4467520
Source                      DF  Type III SS    Mean Square  F Value  Pr > F
bmi                          1   3.79141868      3.79141868     9.70  0.0021
sunexp                       2   1.07957598      0.53978799     1.38  0.2535
lvitdintake                  1   19.94003315     19.94003315    51.03  <.0001
country                      3   5.59595431      1.86531810     4.77  0.0031

Contrast                    DF    Contrast SS    Mean Square  F Value
fjern alle tre paa en gang   4   27.41688677      6.85422169    17.54

Contrast                          Pr > F
fjern alle tre paa en gang        <.0001
```

Hypoteser vedr. interaktioner

Hvilke kunne vi forestille os at se på?

- `sunexp*vitdintake`:
  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

```
estimate "log intake, avoid sun" lvitdintake 1 sunexp*lvitdintake 1 0 0;
estimate "log intake, sometimes sun" lvitdintake 1 sunexp*lvitdintake 0 0 1;
estimate "log intake, prefer sun" lvitdintake 1 sunexp*lvitdintake 0 1 0;
```

Se det i sammenhæng i koden s. 100

Dependent Variable: lvitd

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bmi	1	3.89412209	3.89412209	9.94	0.0019
sunexp	2	1.30495845	0.65247923	1.67	0.1917
lvitdintake	1	15.91554363	15.91554363	40.62	<.0001
country	3	5.66930725	1.88976908	4.82	0.0029
lvitdintake*sunexp	2	0.57306737	0.28653369	0.73	0.4825

Parameter	Estimate	Standard Error	t Value	Pr > |t|
log intake, avoid sun	0.28078044	0.05487930	5.12	<.0001
log intake, sometimes sun	0.26809974	0.05446968	4.92	<.0001
log intake, prefer sun	0.17299082	0.07726529	2.24	0.0262

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

Interaktionen country*sunexp

```
lsmeans country*sunexp / slice=country;
```

Se det i sammenhæng i koden s. 101

Dependent Variable: lvitd

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bmi	1	2.88132968	2.88132968	7.69	0.0061
sunexp	2	0.63728866	0.31864433	0.85	0.4290
lvitdintake	1	20.92600522	20.92600522	55.82	<.0001
country	3	1.16607411	0.38869137	1.04	0.3773
country*sunexp	6	5.50242876	0.91707146	2.45	0.0264

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

Kode s. 101

country*sunexp Effect Sliced by country for lvitd

country	DF	Sum of Squares	Mean Square	F Value	Pr > F
DK	2	0.624064	0.312032	0.83	0.4365
EI	2	2.063510	1.031755	2.75	0.0662
PL	2	2.232301	1.116151	2.98	0.0532
SF	2	1.605083	0.802542	2.14	0.1203

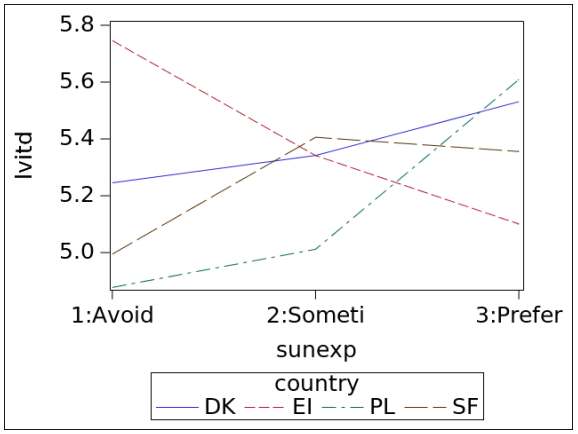
Der ser ud til at være en tendens til effekt af `sunexp` i Polen og
Irland, men ikke i Danmark og Finland.

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

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Illustration af country*sunexp

Kode s. 101



Hvad sker der for Irland?

Interaktionen country*lvitdintake

```
lsmeans country*lvitdintake;
```

Se det i sammenhæng i koden s. 102

```
Dependent Variable: lvitd

Parameter                      Estimate     Standard Error
lvitdintake*country DK      0.429881973         0.06209364
lvitdintake*country EI      0.085708661         0.08084833
lvitdintake*country PL      0.228377351         0.05719061
lvitdintake*country SF      0.165553445         0.08728463
```

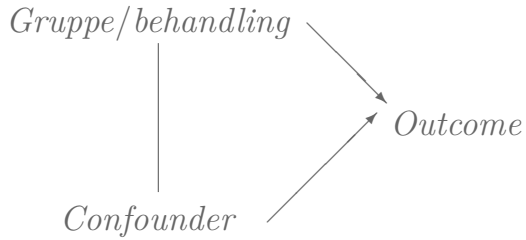
Her ses overraskende nok en signifikant interaktion, mestendels fordi effekten af vitamin D indtaget er langt større i Danmark end i de øvrige lande.
Det har formentlig en ret speciel forklaring.....

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Repetition: Sammenligning af to grupper

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

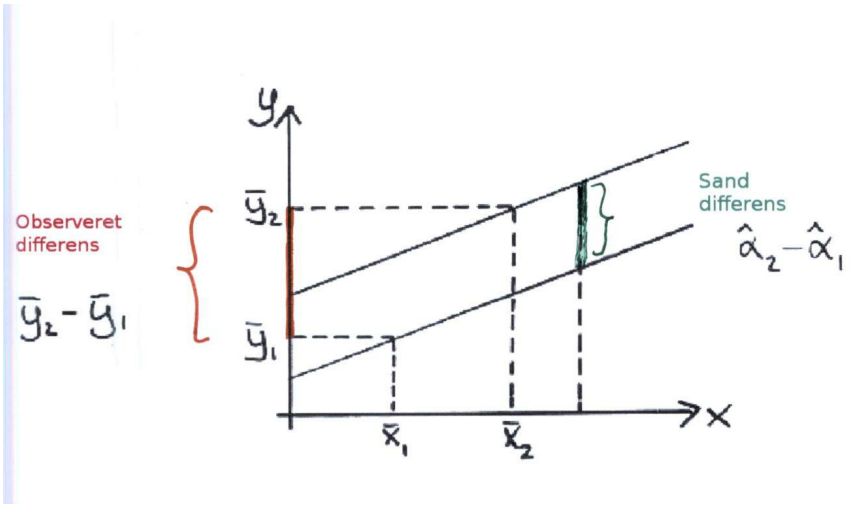
- har en effekt på outcome
- er relateret til gruppen
(der er forskel på værdierne i de to grupper)



Eksempel: Vægt vs. køn og højde

Illustration af confounding og kovariansanalyse

Kovariaten x er her en confounder for gruppeforskellen:



UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

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)

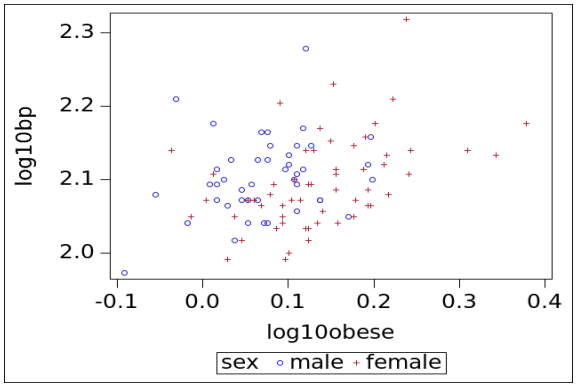
Vi så også dette i eksemplet med P-piller og hormoner i ANCOVA-forelæsningen
Man skal (på protokolstadiet) nøje overveje, hvilke variable med potentiel betydning for outcome, der skal medtages i modellen!

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

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Eksempel: Fedmegrad og blodtryk

Systolisk blodtryk (`bp`) vs. fedmegrad = vægt/idealvægt (`obese`):
begge på logaritmisk skala (kode s. 103)



Resultater, Blodtryk vs. køn

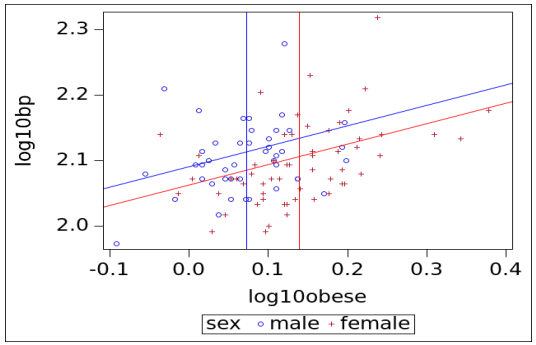
Outcome `log10bp`:

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

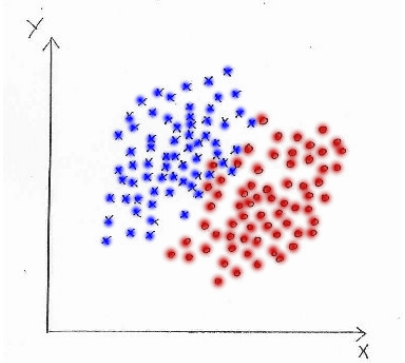
To parallelle linier (kode s. 103):
Samme relation til fedmegrad for de to køn



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

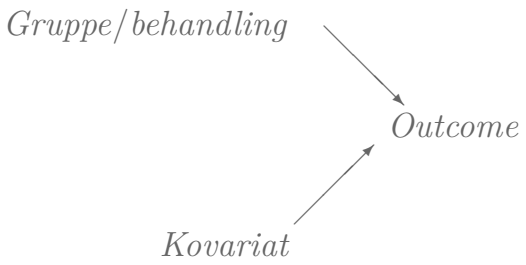
- **positivt** relaterede for hvert køn separat
- **negativt** relaterede for mennesker

Ingen særlig kønsforskel i kolesterol – og dog...

Vi så det også i eksemplet med hjernevægt hos mus

Men læg mærke til følgende:

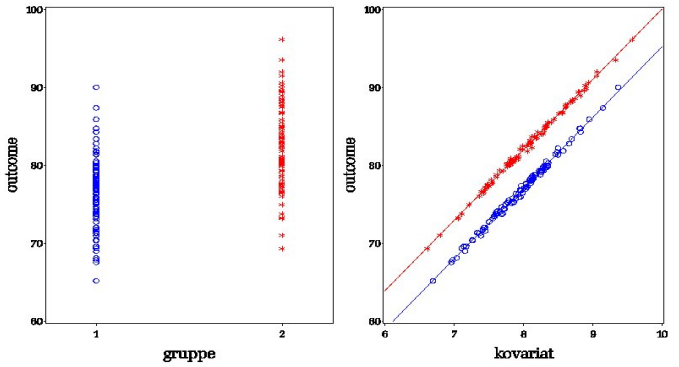
Selv om fordelingen af kovariaten er ens i de to grupper, kan det være af stor betydning at medtage den i analysen.



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

Programbidder svarende til diverse slides:

- Biokemisk iltforbrug, transformationer, s. 87-88
- Udglatning og polynomie-fit, s. 89-90
- Lineære splines, s. 91-92
- Ikke-lineær regression, s. 94
- Vitamin D, GLM, s. 95-99
- Interaktioner, s. 100-102
- Blodtryk og fedme, s. 103

Data vedr. biokemisk iltforbrug

Slide 7 og 11

med omstrukturering, transformationer og analyse

```
data a1;
input days boc1 boc2 boc3 boc4;
datalines;
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
;
run;

data a2;
set a1;

boc=boc1;  output;
boc=boc2;  output;
boc=boc3;  output;
boc=boc4;  output;

drop boc1 boc2 boc3 boc4;
run;
```

```
data boc;
set a2;

logboc=log(boc);
invdays=1/days;
run;

proc glm data=boc;
  model logboc=invdays / clparm;
run;
```

Figurer vedr. biokemisk iltforbrug

Slide 8

```
proc sgplot data=boc;
scatter Y=boc X=days;
run;
```

Slide 12

```
proc sgplot data=boc;
reg Y=logboc X=invdays /
      markerattrs=(color=blue size=10)
      lineattrs=(color=red);
run;
```

Udglattet kurve

en såkaldt *Loess*-kurve

Slide 22

```
proc sgplot data=juul;
    where age ge 5 and age le 20 and sex='male';
    loess Y=sigf1 X=age /
          smooth=0.5 lineattrs=(color=red);
run;
```

Polynomiale fit

Slide 26

```
proc gplot data=juul gout=plotud uniform;
  where age ge 5 and age le 20 and sex='male';
  plot ssigf1*age ssigf1*age ssigf1*age
  / overlay haxis=axis1 vaxis=axis2;
axis1 value=(H=3) minor=NONE label=(H=3);
axis2 value=(H=3) minor=NONE label=(A=90 R=0 H=3);
symbol1 v=circle i=rl c=black ci=black h=2 l=1 w=7;
symbol2 v=circle i=rq c=black ci=red h=2 l=1 w=7;
symbol3 v=circle i=rc c=black ci=green h=2 l=1 w=7;
run;
```

Lineære splines

Slide 27-32

Definer de nye variable:

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

Fit derefter en model med disse 4 ekstra kovariater:

```
proc reg data=juul;
    where age ge 5 and age le 20 and sex='male';
    model ssigf1=age extra_age10 extra_age12
                    extra_age13 extra_age15;
output out=fit p=yhat;
run;
```

Illustration af fit med lineære splines

Slide 29

```
proc gplot data=fit gout=plotud;
  where age ge 5 and age le 20 and sex='male';
  plot (ssigf1 yhat)*age
  / overlay haxis=axis1 vaxis=axis2 frame
    href=10 href=12 href=13 href=15 lh=33;
axis1 value=(H=3) minor=NONE label=(H=3);
axis2 value=(H=3) minor=NONE label=(A=90 R=0 H=3);
symbol1 v=circle i=none c=blue h=2 l=1 w=1;
symbol2 v=none i=join c=red h=2 l=1 w=3;
run;
```

Alternativ parametrisering af lineære splines

Slide 33-34

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

```
ny_age=min(age,10);
ny_age10=min(extra_age10,2);
ny_age12=min(extra_age12,1);
ny_age13=min(extra_age13,2);
ny_age15=extra_age15;
```

Fit derefter en model med disse 4 ekstra kovariater:

```
proc reg data=juul;
    where age ge 5 and age le 20 and sex='male';
    model ssigf1=age ny_age10 ny_age12
                     ny_age13 ny_age15;
output out=fit p=yhat;
run;
```


Ikke-lineær regression

Slide 37-40

```
data reticulo;
infile 'kw_res.txt';
input tid conc;
run;

proc nlin data=reticulo;
  parms beta=2000
        gamma=0.05;
  model conc=beta*(1-exp(-gamma*tid));
run;
```


ANOVA, sammenligning af D-vitamin i 4 lande

Slide 46

```
DATA women;
SET women;

lvitd=log2(vitd);
RUN;

proc glm data=women;
class country;
model lvitd = country /
              solution clparm;
run;
```


Model med bmi som kovariat, ANCOVA

Slide 49

```
proc glm data=women;
class country;
model lvitd = bmi country /
              solution clparm;
run;
```


Model med 4 kovariater og modelkontrol

Slide 53-54, 62-63

```
DATA women;
SET women;

lvitd=log2(vitd);
lvitdintake=log2(vitdintake);
RUN;

proc glm PLOTS=(DIAGNOSTICS RESIDUALs(SMOOTH)) data=women;
class sunexp country;
model lvitd = bmi sunexp lvitdintake country /
              solution clparm;
estimate "1 enhed BMI" bmi 1;
estimate "soldyrkere vs. solhadere" sunexp -1 1 0;
estimate "10% i vitD indtag" lvitdintake 0.1375;
estimate "Finland vs. Polen" country 0 0 -1 1;
estimate "Irland vs. Polen" country 0 1 -1 0;
ods output Estimates=estimater;
run;
```

Model med 4 kovariater og modelkontrol, II

Slide 55-61

```
data regn;
set estimater;

faktor=2**Estimate;
lower=2**LowerCL;
upper=2**UpperCL;
run;

proc print data=regn;
run;
```

Udeladelse af flere kovariater samtidig

Slide 67-68

```
proc glm data=women;
class country sunexp;
model lvitd=bmi sunexp lvitdintake country /
            solution clparm;
contrast 'fjern alle tre paa en gang'
    bmi 1,
    sunexp -1 0 1,
    sunexp 0 -1 1,
    lvitdintake 1;
run;
```

Interaktionen *sunexp*lvitdintake*

Slide 70

```
proc glm data=women;
class country sunexp;
model lvitd=bmi sunexp lvitdintake country
      sunexp*lvitdintake / solution clparm;
estimate "log intake, avoid sun"
         lvitdintake 1 sunexp*lvitdintake 1 0 0;
estimate "log intake, sometimes sun"
         lvitdintake 1 sunexp*lvitdintake 0 0 1;
estimate "log intake, prefer sun"
         lvitdintake 1 sunexp*lvitdintake 0 1 0;
run;
```

Interaktionen *country*sunexp*

Slide 71-73

```
proc glm data=vitamind;
class country sunexp;
model lvitd=bmi sunexp lvitdintake country
      country*sunexp / solution clparm;
lsmeans country*sunexp / slice=country out=adjmeans;
run;

data tegn;
set adjmeans;

lvitd=LSMEAN;
LABEL sunexp="Sun exposure"
      lvitd="log2 Vitamin D";
run;

proc sgplot data=tegn;
    series X=sunexp Y=lvitd / group=country;
    xaxis type=discrete;
run;
```

Interaktionen *country*lvitdintake*

Slide 74

```
proc glm data=women;
class country sunexp;
model lvitd=bmi sunexp country
      country*lvitdintake /
      solution clparm;
lsmeans country*lvitdintake;
run;
```

Blodtryk vs. fedme

Sammenligning af mænd og kvinder, med figur

Slide 79

```
proc glm plots=all data=a1;
  class sex;
  model log10bp=sex log10obese / solution clparm;
run;
```

Kode til figuren, slide 81

```
ODS GRAPHICS ON;
proc sgplot data=a1;
    scatter x=log10obese y=log10bp / group=sex;
lineparm x=0.1 y=2.1223 slope=0.312 / lineattrs=(color=blue);
lineparm x=0.1 y=2.0945 slope=0.312 / lineattrs=(color=red);
lineparm x=0.0725 y=2.0 slope=. / lineattrs=(color=blue);
lineparm x=0.1396 y=2.0 slope=. / lineattrs=(color=red);
run;
ods graphics off;
```