

Faculty of Health Sciences

Basal statistik

Den generelle lineære model mv. i R

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

1 / 106

2 / 106

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

Varsiansanalyse: Kovariaterne er **kategoriske** (grupper, class-variable, faktorer)

- ▶ Ensidedt varsiansanalyse: kun en enkelt kovariat
- ▶ Tosidedt varsiansanalyse: to kovariater

Generel lineær model: **Begge typer kovariater** i samme model

- ▶ Kovariansanalyse: Netop en kvantitativ og en kategorisk kovariat

3 / 106

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		Modeller for gentagne målinger	
		Mixed models		

4 / 106

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.

5 / 106



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)

6 / 106



Biologisk iltforbrug

Iltsvind i lukkede flasker (boc, biochemical oxygen consumption), 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

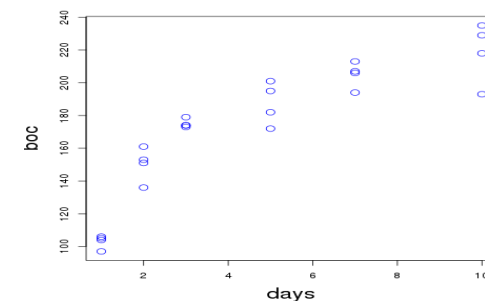
Kode til omstrukturering af data, samt figur næste side: se s. 87-88

7 / 106



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 (∞)

8 / 106



Transformation til linearitet

Biologerne hævder at vide, at iltsvindet kan beskrives ved funktionen

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

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

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

Vi vil gerne bestemme

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

9 / 106



Omparametrisering

Med definitionerne

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

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

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

kan vi skrive ligningen som

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

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

10 / 106



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:

```
> model1 = lm(boc$logboc ~ boc$invdays, na.action=na.exclude)
> summary(model1)
```

```
Call:
lm(formula = boc$logboc ~ boc$invdays, na.action = na.exclude)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.43125    0.01894  286.76  < 2e-16 ***
boc$invdays -0.80781    0.03878  -20.83 5.67e-16 ***
```

```
Residual standard error: 0.05845 on 22 degrees of freedom
```

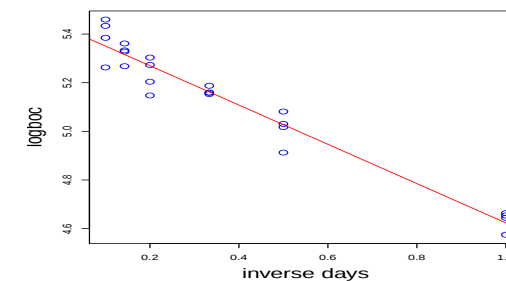
```
> confint(model1)
              2.5 %      97.5 %
(Intercept)  5.3919739  5.4705321
invdays     -0.8882302 -0.7273996
```

11 / 106



Den transformerede relation

Kode s. 88



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

$$\log \text{boc} = 5.431 - 0.808 \times \text{invdays}$$

12 / 106



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)

13 / 106



*Tilbagetransformeret relation

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

```
ny = data.frame(invdays=seq(0.1,1.0,0.01))

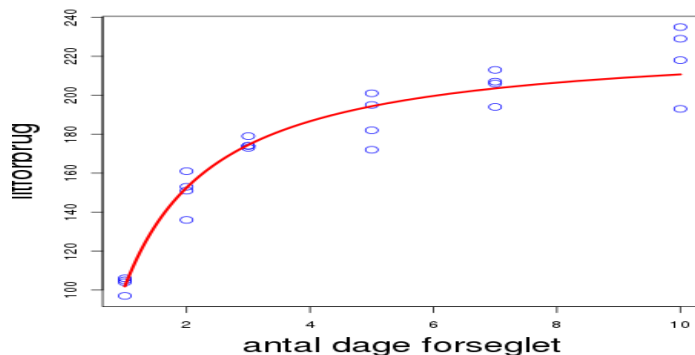
ny$pred = exp(predict(model1, ny))
ny$days <- 1/ny$invdays

plot(boc$days, boc$boc, ylab="iltforbrug",
      xlab="antal dage forseglet",
      col="blue", cex=1.5, cex.lab=2, pch=1)
lines(ny$days, ny$pred, type = "l", col="red", lwd=3)
```

14 / 106



Tilbagetransformeret relation, II



15 / 106



Analyse på original skala

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

```
model2 <- nls(boc ~ gamma*exp(-beta/days), data=boc,
              start=list(gamma=228,beta=0.8))
```

med output:

```
> summary(model2)

Formula: boc ~ gamma * exp(-beta/days)
Parameters:
      Estimate Std. Error t value Pr(>|t|)
gamma 230.58447   4.53844   50.81  < 2e-16 ***
beta   0.83269    0.05757   14.46 1.02e-12 ***

> confint(model2)
      2.5%      97.5%
gamma 221.2597212 240.1996863
beta   0.7153617   0.9567089
```

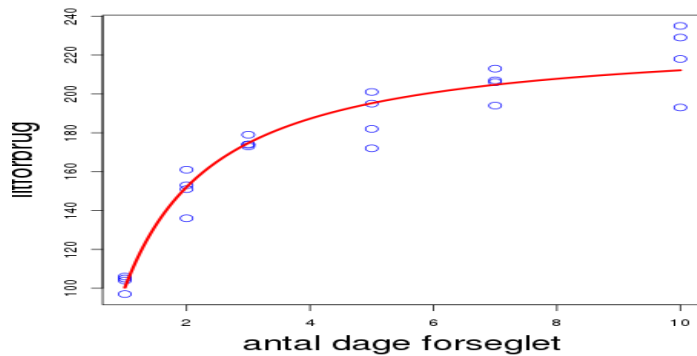
Bemærk, at resultaterne er *meget tæt* på dem fra før

16 / 106



Fit fra ikke-lineær regression

Se kode s. 89



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

17 / 106



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.

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

18 / 106



Transformation med logaritmer, II

- ▶ Hvis outcome transformeres, kan det være
 - ▶ for at opnå linearitet
 - ▶ for at opnå ens spredninger (varienshomogenitet)
 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%

19 / 106



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*

20 / 106



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

21 / 106

Eksempel om væksthormon (fra øvelserne i denne uge)

Væksthormon for børn, op til 20-års alderen

```
scatter.smooth(juul.20$age, juul.20$sigf1,
  ylab="sigf1", xlab="age", cex.lab=1.5,
  lpars = list(col = "red", lwd = 4, lty = 1))
```



Ikke udpræget lineært, men hvad så?
Ingen specifik formel haves....

22 / 106

Polynomier

Et p 'te grads polynomium:

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

Et første-grads polynomium er **en linie**:

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

Et anden-grads polynomium er **en parabel**:

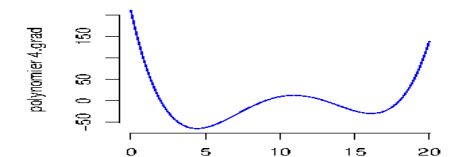
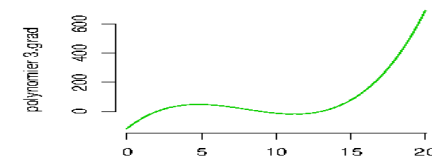
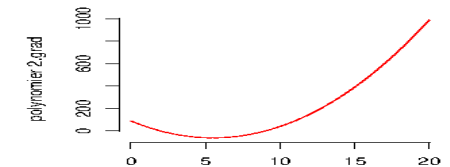
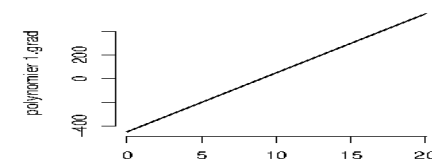
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

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.

23 / 106

Polynomier af 1.-4. grad



24 / 106

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**

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

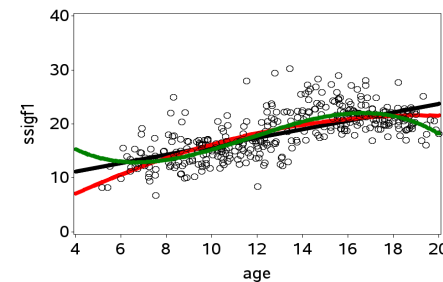
Kovariaterne $Z_1, \dots, Z_p$ er selvfølgelig **korrelerede**, men de er ikke **lineært** afhængige.

25 / 106



Polynomier til beskrivelse af Serum IGF-1

Væksthormon (kvadratrodsformeret), som funktion af alder



med overlejlrede polynomier af 1., 2. og 3. grad

Men vi tror ikke rigtigt på disse modeller....

fordi de svinger for meget

Specielt ude i enderne kan de opføre sig meget underligt.

26 / 106



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.

27 / 106



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

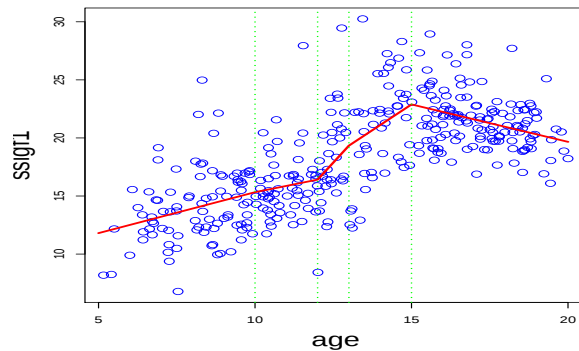
Hældning i aldersintervallet (a_1, a_2) er $\lambda_0 + \lambda_1$

- ▶ 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.

28 / 106



Lineær spline for Serum IGF-1



Estimer:

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

(kode s. 91-92)

29 / 106



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

```
juul.20$extra.age10=pmax(juul.20$age-10,0)
juul.20$extra.age12=pmax(juul.20$age-12,0)
juul.20$extra.age13=pmax(juul.20$age-13,0)
juul.20$extra.age15=pmax(juul.20$age-15,0)
```

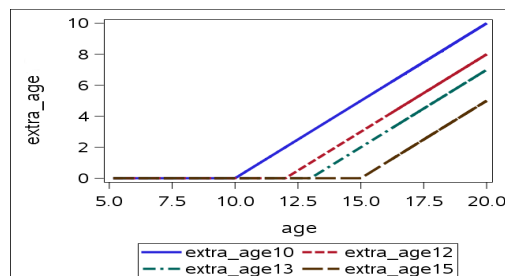
30 / 106



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"



31 / 106



Output, lineær spline

Kode s. 91

Call:

```
lm(formula = ssigf1 ~ age + extra.age10 + extra.age12 + extra.age13 +
    extra.age15, data = juul.20, na.action = na.exclude)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	8.2828	1.8671	4.436	1.22e-05	***
age	0.7045	0.2180	3.231	0.00135	**
extra.age10	-0.1625	0.5558	-0.292	0.77018	
extra.age12	2.3836	1.2512	1.905	0.05759	.
extra.age13	-1.1516	1.2812	-0.899	0.36935	
extra.age15	-2.4184	0.5387	-4.489	9.67e-06	***

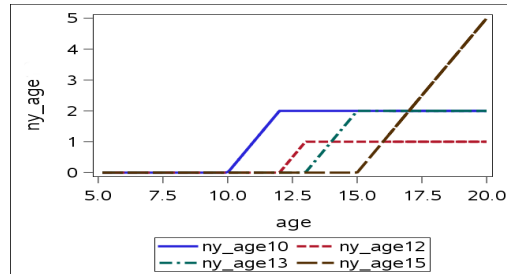
Evidens for et knæk omkring 15-års alderen,
og *måske* omkring 12-års alderen...

32 / 106



*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**):

33 / 106



*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

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	8.2828	1.8671	4.436	1.22e-05	***
ny_age	0.7045	0.2180	3.231	0.001348	**
ny.age10	0.5420	0.4045	1.340	0.181139	
ny.age12	2.9255	0.9502	3.079	0.002239	**
ny.age13	1.7740	0.4204	4.220	3.10e-05	***
ny.age15	-0.6444	0.1751	-3.681	0.000268	***

34 / 106



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:

35 / 106



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!

36 / 106



Mindste kvadraters metode

kræver her

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

Koden bliver her:

```
modell1 <- nls(konc ~ beta*(1-exp(-gamma*tid)), data=kw,
              trace=T,
              start=list(beta=2000,gamma=0.05))
```

37 / 106



*Output fra ikke-lineær regression

Koden ses på s. 37

En del af output vedrører iterationen (trace=T):

```
4370990 : 2e+03 5e-02
382995.4 : 1.958555e+03 8.241949e-02
26354.95 : 2.169165e+03 7.692258e-02
25286.74 : 2.174161e+03 7.724681e-02
25286.72 : 2.174048e+03 7.725653e-02
25286.72 : 2.174044e+03 7.725685e-02
```

Number of iterations to convergence: 5
Achieved convergence tolerance: 9.679e-07

Konvergens betyder, at et “*stabil*” fit er fundet

38 / 106



*Output, fortsat

```
> summary(modell1)
Formula: konc ~ beta * (1 - exp(-gamma * tid))

Parameters:
      Estimate Std. Error t value Pr(>|t|)
beta  2.174e+03  2.835e+01   76.70  <2e-16 ***
gamma  7.726e-02  2.264e-03   34.12  <2e-16 ***

Residual standard error: 32.46 on 24 degrees of freedom

> confint(modell1)

              2.5%          97.5%
beta  2.117204e+03  2.236264e+03
gamma  7.258555e-02  8.209509e-02
```

Fittet svarende til disse parameterestimer fremgår af s. 40

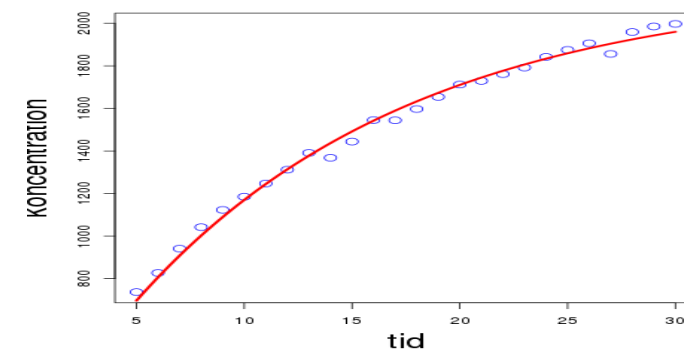
39 / 106



Fittet fra ikke-lineær regression

Estimer:

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



40 / 106



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?

41 / 106



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

42 / 106



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....?
- ▶
- ▶

43 / 106



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

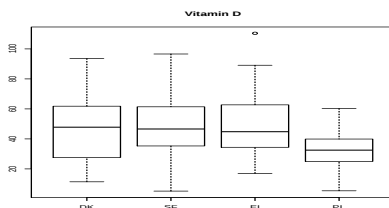
- ▶ højt i body mass index
- ▶ lavt i vitamin D indtag

44 / 106



Første skridt

Er der overhovedet signifikant forskel på landene?



Modelldiagram: 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**.

45 / 106



Sammenligning af landene

Outcome Y: kvantitativ, Y=log2-transformeret

Kovariat X: kategorisk, X=country

Derfor:

Ensidet variansanalyse, på log2-skala:

Sammenligning af 4 middelværdier (kode s. 95)

```
summary(model1)
```

Call:

```
lm(formula = log2vitd ~ relevel(country, ref = "SF"), data = vitd2,
    na.action = na.exclude)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.447275	0.097605	55.809	< 2e-16 ***
relevel(country, ref = "SF")DK	-0.086929	0.138684	-0.627	0.531
relevel(country, ref = "SF")EI	0.009137	0.148574	0.062	0.951
relevel(country, ref = "SF")PL	-0.557730	0.132065	-4.223	3.6e-05 ***

Residual standard error: 0.7172 on 209 degrees of freedom

Multiple R-squared: 0.1073, Adjusted R-squared: 0.09452

F-statistic: 8.377 on 3 and 209 DF, p-value: 2.768e-05

46 / 106



Fortolkning af estimer

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.

47 / 106

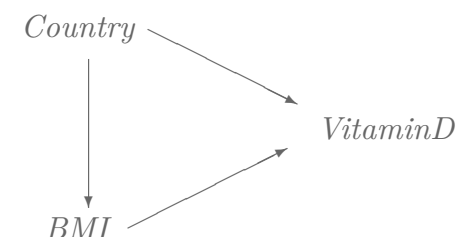


Hvordan forklarer vi forskellen mellem landene?

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

Country → BMI → Vitamin D

eller måske **bare noget af det?**



BMI er en **mellemkommende** variabel (**mediator**)

48 / 106



Model med bmi som kovariat

(kode s. 96)

```
> summary(model2)
```

```
Call:
lm(formula = log2vitd ~ bmi + relevel(country, ref = "SF"), data = vitd2,
    na.action = na.exclude)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.52605	0.34225	19.068	< 2e-16 ***
bmi	-0.03808	0.01160	-3.282	0.00121 **
relevel(country, ref = "SF")DK	-0.15525	0.13714	-1.132	0.25891
relevel(country, ref = "SF")EI	-0.06612	0.14702	-0.450	0.65337
relevel(country, ref = "SF")PL	-0.53446	0.12928	-4.134	5.16e-05 ***

```
Residual standard error: 0.701 on 208 degrees of freedom
Multiple R-squared: 0.1513, Adjusted R-squared: 0.135
F-statistic: 9.269 on 4 and 208 DF, p-value: 6.526e-07
```

```
> confint(model2)
```

	2.5 %	97.5 %
(Intercept)	5.85133526	7.20076288
bmi	-0.06095336	-0.01520717
relevel(country, ref = "SF")DK	-0.42561572	0.11511010
relevel(country, ref = "SF")EI	-0.35595137	0.22371527
relevel(country, ref = "SF")PL	-0.78931899	-0.27959825

49 / 106



Fortolkning af nye estimerer

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

50 / 106



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 korigeret for bmi ($P = 0.0002$, se nedenfor), så der er masser af plads til andre bud på forklarende variable

fra protokollen, vel at mærke.

```
> anova(model2)
```

Analysis of Variance Table

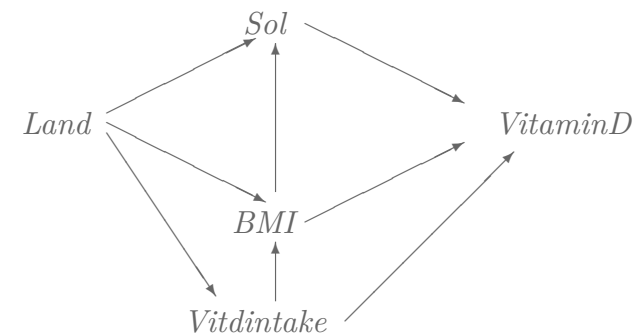
Response: log2vitd

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bmi	1	8.230	8.2299	16.7457	6.113e-05 ***
relevel(country, ref = "SF")	3	9.992	3.3308	6.7773	0.0002212 ***
Residuals	208	102.224	0.4915		

51 / 106



Modeldiagram - med "det hele"



og så er der interaktionerne....

52 / 106



Kan vi så forklare forskellen på landene

ved hjælp af alle de 3 specificerede kovariater samtidig?
(kode s. 97)

```
model3 = lm(log2vitd ~ bmi + relevel(sunexp, ref="Sometimes in sun")
+lvitdintake + relevel(country, ref="SF"),
data=vitd2, na.action=na.exclude)

> anova(model3)
Analysis of Variance Table

Response: log2vitd
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bmi	1	8.230	8.2299	21.0623	7.741e-06
relevel(sunexp, ref = "Sometimes in sun")	2	3.638	1.8191	4.6554	0.010538
lvitdintake	1	22.881	22.8809	58.5577	7.588e-13
relevel(country, ref = "SF")	3	5.596	1.8653	4.7738	0.003078
Residuals	205	80.102	0.3907		

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

53 / 106



Output, fortsat

```
Call:
lm(formula = log2vitd ~ bmi + relevel(sunexp, ref = "Sometimes in sun") +
    lvitdintake + relevel(country, ref = "SF"), data = vitd2,
    na.action = na.exclude)

Coefficients:
(Intercept)          5.46183      Std. Error 0.34415      t value 15.870      Pr(>|t|) < 2e-16 ***
bmi                -0.03266      Std. Error 0.01048      t value -3.115      0.0021 **
relevel(sunexp, ref = "Sometimes in sun")Avoid sun -0.07156      Std. Error 0.09850      t value -0.726      0.4684
relevel(sunexp, ref = "Sometimes in sun")Prefer sun  0.14441      Std. Error 0.12207      t value  1.183      0.2382
lvitdintake         0.25512      Std. Error 0.03571      t value  7.144      1.56e-11 ***
relevel(country, ref = "SF")DK                    0.07045      Std. Error 0.12638      t value  0.558      0.5778
relevel(country, ref = "SF")EI                     0.21105      Std. Error 0.13675      t value  1.543      0.1243
relevel(country, ref = "SF")PL                    -0.24809      Std. Error 0.12281      t value -2.020      0.0447 *

Residual standard error: 0.6251 on 205 degrees of freedom
Multiple R-squared:  0.335, Adjusted R-squared:  0.3123
F-statistic: 14.75 on 7 and 205 DF,  p-value: 1.581e-15

> confint(model3)
```

	2.5 %	97.5 %
(Intercept)	4.78329302	6.140364671
bmi	-0.05333041	-0.011987828
relevel(sunexp, ref = "Sometimes in sun")Avoid sun	-0.26577015	0.122650733
relevel(sunexp, ref = "Sometimes in sun")Prefer sun	-0.09626243	0.385081534
lvitdintake	0.18470536	0.325526845
relevel(country, ref = "SF")DK	-0.17870801	0.319618004
relevel(country, ref = "SF")EI	-0.05856989	0.480668580
relevel(country, ref = "SF")PL	-0.49022244	-0.005960845

54 / 106



Fortolkning af estimer

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.

55 / 106



Fortolkning af estimer, 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 $(2^{0.0060}, 2^{0.4902}) = (1.004, 1.40)$

56 / 106



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 R benytte pakken multcomp:

```
install.packages("multcomp")
library(multcomp)
```

(se kode s. 58 og output s. 59-60).

57 / 106



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 parameterestimerer

(se det i sammenhæng i koden s. 98):

```
e.bmi=confint(summary(glht(model3, linfct=rbind(c(0,1,0,0,0,0,0,0))))
#soldyrkere vs solhadere
e.sunexp=confint(summary(glht(model3, linfct=rbind(c(0,0,-1,1,0,0,0,0))))
#10% i vitd intag
e.vitdintake=confint(summary(glht(model3, linfct=rbind(c(0,0,0,0,0.1375,0,0,0))))
#Finland vs. Polen
e.SFvsPL=confint(summary(glht(model3, linfct=rbind(c(0,0,0,0,0,0,-1))))
#Irland vs. Polen
e.EIvsPL=confint(summary(glht(model3, linfct=rbind(c(0,0,0,0,0,0,1,-1))))
```

58 / 106



Output fra kode s. 58

```
> e.bmi$confint
      Estimate      lwr      upr
1 -0.03265912 -0.05333041 -0.01198783

> e.sunexp$confint
      Estimate      lwr      upr
1 0.2159693 -0.04026293 0.4722014

> e.vitdintake$confint
      Estimate      lwr      upr
1 0.03507846 0.02539699 0.04475994

> e.SFvsPL$confint
      Estimate      lwr      upr
1 0.2480916 0.005960845 0.4902224

> e.EIvsPL$confint
      Estimate      lwr      upr
1 0.459141 0.2074598 0.7108222
```

Disse skal nu tilbagetransformeres,
men dette kan vi også få gjort automatisk, se næste side

59 / 106



Tilbagetransformerede estimerer

```
> 2^(e.bmi$confint)
      Estimate      lwr      upr
1 0.9776167 0.9637091 0.9917251

> 2^(e.sunexp$confint)
      Estimate      lwr      upr
1 1.161484 0.9724777 1.387225

> 2^(e.vitdintake$confint)
      Estimate      lwr      upr
1 1.024613 1.01776 1.031512

> 2^(e.SFvsPL$confint)
      Estimate      lwr      upr
1 1.187635 1.00414 1.404661

> 2^(e.EIvsPL$confint)
      Estimate      lwr      upr
1 1.374723 1.154653 1.636737
```

60 / 106



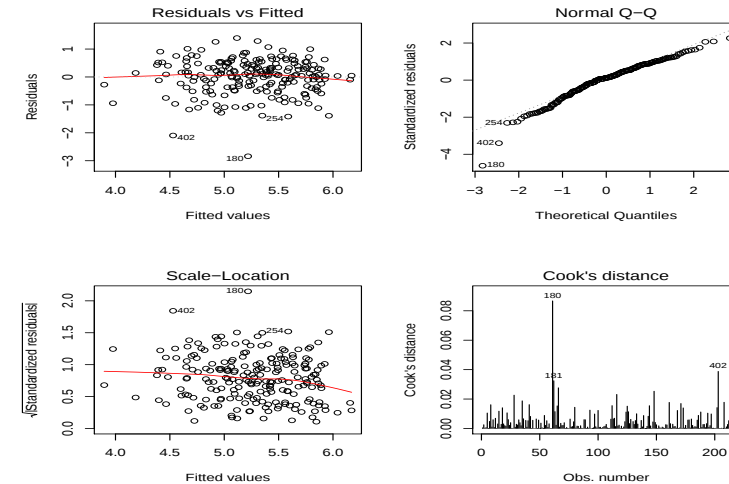
Fortolkning af sammenligningerne s. 58-60

- 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.

61 / 106



Husk også modelkontrol (kode s. 99)

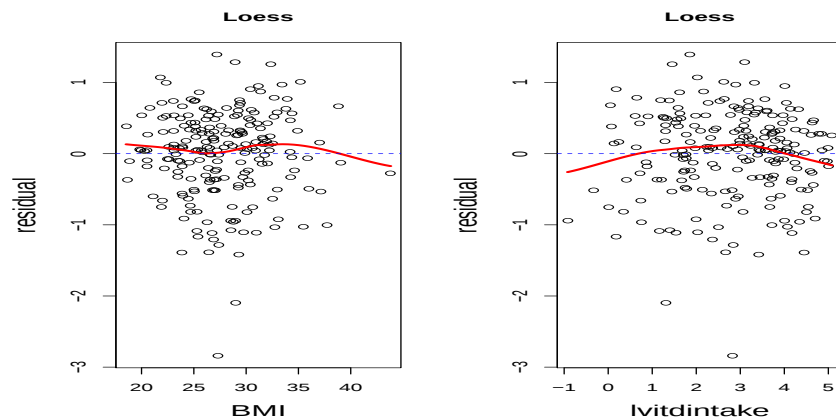


62 / 106



Modelkontrol, II

Kode s. 99



63 / 106



Note om T-test og F-test

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

- Forskell 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 ($\beta_1 = \beta_2 = \beta_3 = 0$, $df=3$)

64 / 106



*Modelreduktion - F test

Vi skal sammenligne to modeller:

Den oprindelige med *alle* kovariater (nr. 1)

og den simple (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:

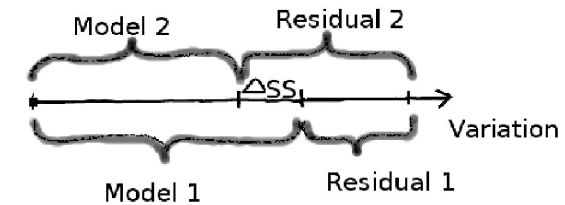
Hvor meget mindre forklares af den simple model?

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

65 / 106



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

66 / 106



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 forklaringssevne 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 $F = 17.54 \sim F(4, 205)$, $P < 0.0001$,

(se kode næste side)

67 / 106



Udelad flere kovariater på en gang, II

Først specificeres model med kun land, samt model med 3 yderligere kovariater:

```
model1 = lm(log2vitd ~ relevel(country, ref="SF"),
             data=vitd2, na.action=na.exclude)
model3 = lm(log2vitd ~ bmi + relevel(sunexp, ref="Sometimes in sun")
             + lvitdintake + relevel(country, ref="SF"),
             data=vitd2, na.action=na.exclude)
```

Herefter sammenligner vi de to modeller

```
> anova(model3, model1)
Analysis of Variance Table
```

```
Model 1: log2vitd ~ bmi + relevel(sunexp, ref = "Sometimes in sun") +
  lvitdintake + relevel(country, ref = "SF")
Model 2: log2vitd ~ relevel(country, ref = "SF")
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     205  80.102
2     209 107.519  -4    -27.417 17.542 2.137e-12 ***
```

68 / 106



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 forelæsningsen om ANOVA)
- ▶ `country*lvitdintake`:
næppe...og dog
- ▶

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

69 / 106



Interaktionen `sunexp*lvitdintake`

Se det i sammenhæng i koden s. 101

```
> anova(int1)
Analysis of Variance Table

Response: log2vitd

      Df Sum Sq Mean Sq F value    Pr(>F)
bmi      1  8.230   8.2299  21.0071 7.987e-06 ***
sunexp    2  3.638   1.8191   4.6432 0.010672 *
lvitdintake 1 22.881  22.8809  58.4042 8.311e-13 ***
country   3  5.596   1.8653   4.7613 0.003135 **
sunexp:lvitdintake 2  0.573   0.2865   0.7314 0.482505
Residuals 203 79.529   0.3918

> cbind(confint(int1test)[8:10,1],
+        int1test$coefficients[8:10], confint(int1test)[8:10,2])
              [,1]      [,2]      [,3]
sunexpAvoid sun:lvitdintake    0.17257390 0.2807804 0.3889870
sunexpSometimes in sun:lvitdintake 0.16070085 0.2680997 0.3754986
sunexpPrefer sun:lvitdintake    0.02064539 0.1729908 0.3253362
```

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

70 / 106



Interaktionen `country*sunexp`

Se det i sammenhæng i koden s. 102

```
> int2 = lm(log2vitd ~ bmi + sunexp
+           +lvitdintake + country
+           +sunexp:country,
+           data=vitd2, na.action=na.exclude)
> anova(int2)
Analysis of Variance Table

Response: log2vitd

      Df Sum Sq Mean Sq F value    Pr(>F)
bmi      1  8.230   8.2299  21.9539 5.170e-06 ***
sunexp    2  3.638   1.8191   4.8525 0.008757 **
lvitdintake 1 22.881  22.8809  61.0366 3.161e-13 ***
country   3  5.596   1.8653   4.9759 0.002374 **
sunexp:country  6  5.502   0.9171   2.4464 0.026428 *
Residuals 199 74.599   0.3749
```

Her er der en **signifikant interaktion**.

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

71 / 106



Effekt af sol, opdelt efter land

Kode s. 102 (kræver pakken "*phia*")

```
> testInteractions(int2, fixed="country", across="sunexp")
F Test:
P-value adjustment method: holm

      sunexp1 sunexp2 Df Sum of Sq      F Pr(>F)
DK      -0.28558 -0.18923  2      0.624 0.8324 0.4365
SF      -0.36102  0.05003  2      1.605 2.1408 0.2405
EI       0.64540  0.24081  2      2.064 2.7523 0.2128
PL      -0.73153 -0.59716  2      2.232 2.9774 0.2128
Residuals                                199    74.599
```

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

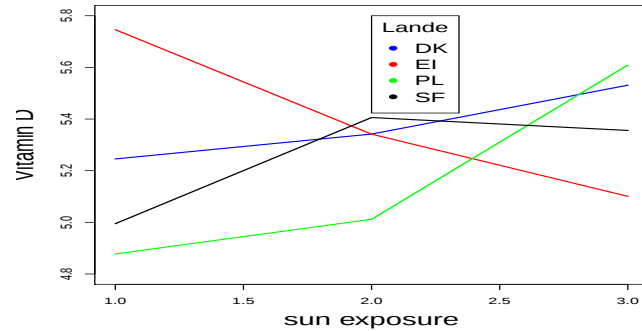
Men mønsteret for Irland er ret svært at forstå....se figur næste side

72 / 106



Illustration af country*sunexp

Kode s. 103



Hvad sker der for Irland?

73 / 106

Interaktionen country*lvitdintake

Se det i sammenhæng i koden s. 104

```
> cbind(confint(int3est)[8:11,1],
+       int3est$coefficients[8:11], confint(int3est)[8:11,2])
               [,1]      [,2]      [,3]
countryDK:lvitdintake 0.30744713 0.42988197 0.5523168
countrySF:lvitdintake -0.00655241 0.16555344 0.3376593
countryEI:lvitdintake -0.07370624 0.08570866 0.2451236
countryPL:lvitdintake 0.11561020 0.22837735 0.3411445
```

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

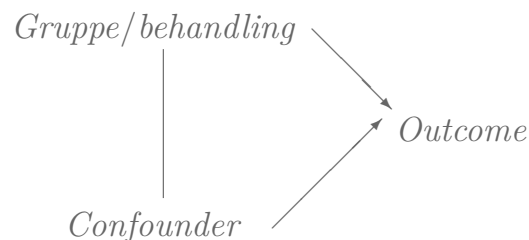
Det har formentlig en ret speciel forklaring.....

74 / 106

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)

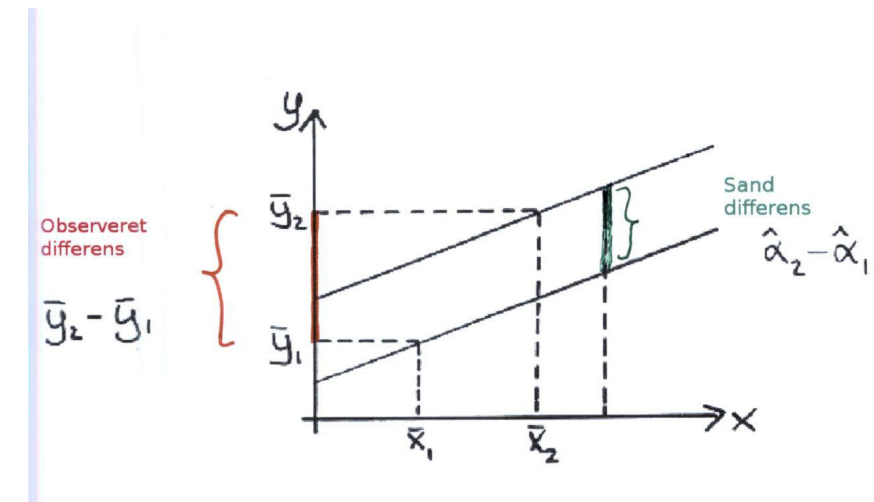


Eksempel: Vægt vs. køn og højde

75 / 106

Illustration af confounding og kovariansanalyse

Kovariaten x er her en confounder for gruppeforskellen:



76 / 106

Eksempel om mænds og kvinders vægt

fra forelæsningen om kovariansanalyse:

Vægt vs. køn, Outcome er $\log_{10}$ vægt:

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.

77 / 106



Alternativt eksempel

Det kan **også** forekomme, at

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

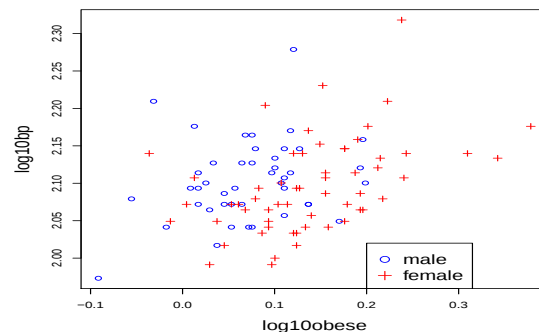
78 / 106



Eksempel: Fedmegrad og blodtryk

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

begge på logaritmisk skala



79 / 106



Resultater, Blodtryk vs. køn

Outcome $\log_{10}$ bp:

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%)

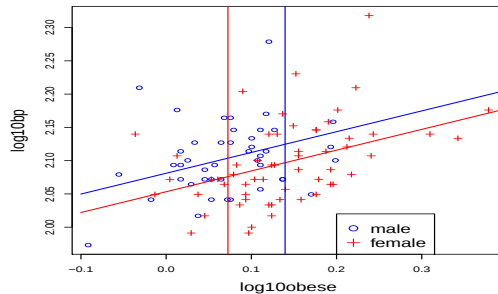
80 / 106



Illustration af kovariansanalysen

To **parallelle linier** (kode s. 106)

Samme relation til fedmegrad for de to køn



Forsigtig konklusion:

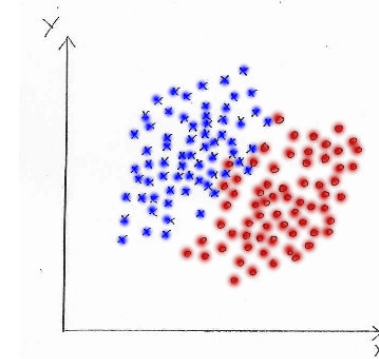
Kvinder har ligeså højt blodtryk som mænd, fordi de er federe...

81 / 106



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...

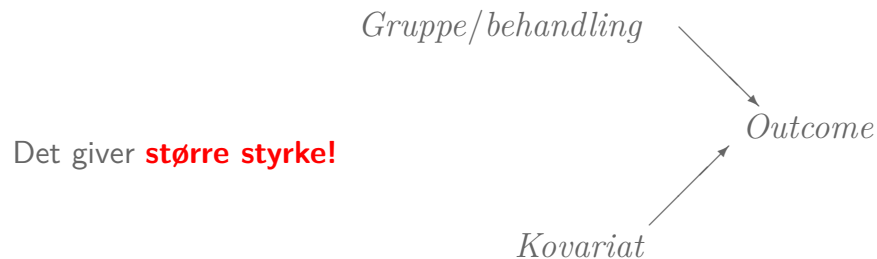
Vi så det også i eksemplet med **hjernevægt hos mus**

82 / 106



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.

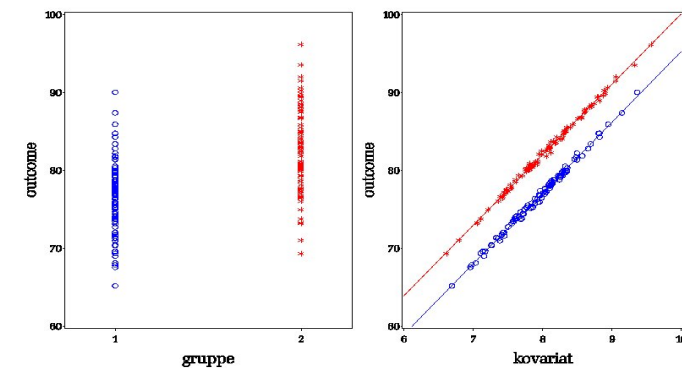


Men vi svarer samtidig på **et andet videnskabeligt spørgsmål!!**

83 / 106



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)

84 / 106



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*).

85 / 106



Appendix

Programbidder svarende til diverse slides:

- ▶ Biokemisk iltforbrug, transformationer, s. 87-89
- ▶ Udglatning, s. 90
- ▶ Lineære splines, s. 91-93
- ▶ Ikke-lineær regression, s. 94
- ▶ Vitamin D, GLM, s. 95-100
- ▶ Interaktioner, s. 101-104
- ▶ Blodtryk og fedme, s. 105-106

86 / 106



Data vedr. biokemisk iltforbrug

Slide 7 og 11

med transformationer og analyse

```
boc <-
read.table("http://publicifsv.sund.ku.dk/~lts/basal/Rdata/boc.txt",
header=T, na.strings=c("."))
```

```
boc$logboc <- log(boc$boc)
boc$invdays <- 1/boc$days
```

```
model1 = lm(logboc ~ invdays, na.action=na.exclude, data=boc)
summary(model1)
confint(model1)
```

87 / 106



Figurer vedr. biokemisk iltforbrug

Slide 8

```
plot(boc$days, boc$boc,
     ylab="boc", xlab="days",
     cex.lab=1.8, cex=1.5, col="blue")
```

Slide 12

```
boc$logboc <- log(boc$boc)
boc$invdays <- 1/boc$days
```

```
model1 = lm(logboc ~ invdays, na.action=na.exclude, data=boc)
```

```
plot(boc$invdays, boc$logboc,
     ylab="logboc", xlab="inverse days",
     cex.lab=1.8, cex=1.5, col="blue")
abline(model1, lty=1, col="red")
```

88 / 106



Ikke-lineært fit

Slide 17

```
ny = data.frame(days=seq(1,10,0.1))

ny$pred = predict(model2, ny, se.fit=TRUE)
pred.plim <- predict(model2, ny, interval = "prediction")
pred.clim <- predict(model2, ny, interval = "confidence")

plot(boc$days, boc$boc, ylab="iltforbrug",
     xlab="antal dage forseglet",
     col="blue",cex=1.5, cex.lab=2, pch=1)
lines(ny$days, ny$pred, type = "l", col="red", lwd=3)
```

89 / 106



Udglattet kurve

Slide 22

Den viste figur:

```
scatter.smooth(juul.20$age, juul.20$sigf1,
               ylab="sigf1", xlab="age", cex.lab=1.5,
               lpars = list(col = "red", lwd = 4, lty = 1))
```

og en såkaldt *Loess*-kurve

```
scatter.smooth(juul.20$age, juul.20$sigf1,
               ylab="sigf1", xlab="age", cex.lab=1.5)
loess.smooth(juul.20$age, juul.20$sigf1, span = 2/3, degree = 1,
              family = c("symmetric", "gaussian"), evaluation = 50)
```

90 / 106



Lineære splines

Slide 27-32

Definer de nye variable:

```
juul.20$extra.age10=pmax(juul.20$age-10,0)
juul.20$extra.age12=pmax(juul.20$age-12,0)
juul.20$extra.age13=pmax(juul.20$age-13,0)
juul.20$extra.age15=pmax(juul.20$age-15,0)
```

Fit derefter en model med disse 4 ekstra kovariater:

```
model.spline <- lm(ssigf1 ~ age +
                  extra.age10 + extra.age12
                  + extra.age13 + extra.age15, data=juul.20,
                  na.action=na.exclude)

summary(model.spline)
```

91 / 106



Illustration af fit med lineære splines

Slide 29

```
ny = data.frame(age=seq(5,20,0.5))
ny$extra.age10=pmax(ny$age-10,0)
ny$extra.age12=pmax(ny$age-12,0)
ny$extra.age13=pmax(ny$age-13,0)
ny$extra.age15=pmax(ny$age-15,0)

ny$pred = predict(model.spline, ny)

plot(juul.20$age, juul.20$ssigf1, ylab="ssigf1",
     xlab="age",
     col="blue",cex=1.5, cex.lab=2, pch=1)
lines(ny$age, ny$pred, type = "l", col="red", lwd=3)
abline(v=10,col="green",lty=3,lwd=2)
abline(v=12,col="green",lty=3,lwd=2)
abline(v=13,col="green",lty=3,lwd=2)
abline(v=15,col="green",lty=3,lwd=2)
```

92 / 106



Alternativ parametrisering af lineære splines

Slide 33-34

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

```
juul.20$ny.age=pmin(juul.20$age,10)
juul.20$ny.age10=pmin(juul.20$extra.age10,2)
juul.20$ny.age12=pmin(juul.20$extra.age12,1)
juul.20$ny.age13=pmin(juul.20$extra.age13,2)
juul.20$ny.age15=juul.20$extra.age15
```

Fit derefter en model med disse 4 ekstra kovariater:

```
model.spline2 <- lm(ssigf1 ~ ny.age +
                    ny.age10 + ny.age12
                    + ny.age13 + ny.age15, data=juul.20)
```

```
summary(model.spline2)
```

93 / 106



Ikke-lineær regression

Slide 37-40

```
kw <-
read.table("http://publicifsv.sund.ku.dk/~lts/basal/Rdata/kw.txt",
header=T, na.strings=c("."))
```

```
model1 <- nls(konc ~ beta*(1-exp(-gamma*tid)), data=kw,
              trace=T,
              start=list(beta=2000, gamma=0.05))
```

```
summary(model1)
confint(model1)
```

94 / 106



ANOVA, sammenligning af D-vitamin i 4 lande

Slide 46

```
vitd <-
read.table("http://publicifsv.sund.ku.dk/~lts/basal/Rdata/VitaminD.txt",
header=T, na.strings=c("."))
```

```
vitd$country <- factor(vitd$land, levels=c(1,2,4,6),
                      labels=c("DK", "SF", "EI", "PL"))
vitd$sunexp <- factor(vitd$sun, levels=1:3,
                    labels=c("Avoid sun", "Sometimes in sun", "Prefer sun"))
vitd$log2vitd <- log2(vitd$VitaminD)
```

```
vitd2 <- subset(vitd, category==2,)
```

```
model1 = lm(log2vitd ~ relevel(country, ref="SF"),
            data=vitd2, na.action=na.exclude)
summary(model1)
```

95 / 106



Model med bmi som kovariat, ANCOVA

Slide 49

```
model2 = lm(log2vitd ~ bmi + relevel(country, ref="SF"),
            data=vitd2, na.action=na.exclude)
```

```
summary(model2)
```

```
confint(model2)
```

96 / 106



Model med 4 kovariater

Slide 53-54

```
model3 = lm(log2vitd ~ bmi + relevel(sunexp, ref="Sometimes in sun")
            +lvitdintake + relevel(country, ref="SF"),
            data=vitd2, na.action=na.exclude)

anova(model3)

summary(model3)

confint(model3)

cbind(confint(model3)["bmi",1],
      model3$coefficients["bmi"], confint(model3)["bmi",2])
```

97 / 106



Ekstra sammenligninger

Slide 58-60

```
install.packages("multcomp")
library(multcomp)

#1 enhed bmi
e.bmi=confint(summary(glht(model3, linfct=rbind(c(0,1,0,0,0,0,0,0))))))
#soldyrkere vs solhadere
e.sunexp=confint(summary(glht(model3, linfct=rbind(c(0,0,-1,1,0,0,0,0))))))
#10% i vitd intag
e.vitdintake=confint(summary(glht(model3, linfct=rbind(c(0,0,0,0,0,0.1375,0,0,0))))))
#Finland vs. Polen
e.SFvsPL=confint(summary(glht(model3, linfct=rbind(c(0,0,0,0,0,0,0,-1))))))
#Irland vs. Polen
e.EIvsPL=confint(summary(glht(model3, linfct=rbind(c(0,0,0,0,0,0,1,-1))))))

e.bmi$confint
e.sunexp$confint
e.vitdintake$confint
e.SFvsPL$confint
e.EIvsPL$confint

2^(e.bmi$confint)
2^(e.sunexp$confint)
2^(e.vitdintake$confint)
2^(e.SFvsPL$confint)
2^(e.EIvsPL$confint)
```

98 / 106



Modelkontrol for model med 4 kovariater

Slide 62-63

```
par(mfrow=c(2,2))
plot(model3,which=1)
plot(model3,which=2)
plot(model3,which=3)
plot(model3,which=4)

par(mfrow=c(1,2))
scatter.smooth(vitd2$bmi, resid(model3), main="Loess",
              ylab="residual", xlab="BMI", cex.lab=1.5,
              lpars = list(col = "red", lwd = 3, lty = 3))
abline(0,0,col="red",lty=2)

scatter.smooth(vitd2$lvitdintake, resid(model3), main="Loess",
              ylab="residual", xlab="lvitdintake", cex.lab=1.5,
              lpars = list(col = "red", lwd = 3, lty = 3))
abline(0,0,col="red",lty=2)
```

99 / 106



Udeladelse af flere kovariater samtidig

Slide 67-68

```
model1 = lm(log2vitd ~ relevel(country, ref="SF"),
            data=vitd2, na.action=na.exclude)

model3 = lm(log2vitd ~ bmi + relevel(sunexp, ref="Sometimes in sun")
            +lvitdintake + relevel(country, ref="SF"),
            data=vitd2, na.action=na.exclude)

anova(model3, model1)
```

100 / 106



Interaktioner *sunexp*lvitdintake*

Slide 70

Den testvenlige kode:

```
int1 = lm(log2vitd ~ bmi + sunexp
          +lvitdintake + country
          +sunexp:lvitdintake,
          data=vitd2, na.action=na.exclude)
anova(int1)
```

Den estimations-venlige kode:

```
int1est = lm(log2vitd ~ bmi + sunexp
             + country
             +sunexp:lvitdintake,
             data=vitd2, na.action=na.exclude)

cbind(confint(int1est)[8:10,1],
      int1est$coefficients[8:10], confint(int1est)[8:10,2])
```

101 / 106



Interaktioner *country*sunexp*

Slide 71-72

```
int2 = lm(log2vitd ~ bmi + sunexp
          +lvitdintake + country
          +sunexp:country,
          data=vitd2, na.action=na.exclude)
```

```
summary(int2)
```

```
anova(int2)
```

```
install.packages("phia")
library(phia)
testInteractions(int2, fixed="country", across="sunexp")
```

102 / 106



Illustration af interaktioner *country*sunexp*

Slide 73

```
ny.country=c(rep("SF",3),rep("DK",3),rep("PL",3),rep("EI",3))
ny.sol=rep(c("Avoid sun", "Sometimes in sun", "Prefer sun"),4)
ny.bmi=rep(mean(vitd2$bmi),12)
ny.lvitdintake=rep(mean(vitd2$lvitdintake),12)
ny = data.frame(country=ny.country, sunexp=ny.sol, bmi=ny.bmi,
               lvitdintake=ny.lvitdintake)
pred2 = predict(int2, ny)

plot(1:3,pred2[ny.country=="DK"],
     ylab="Vitamin D", xlab="sun exposure",
     xlim=c(1,3),ylim=c(4.8, 5.8),
     pch=1, col="blue", cex.lab=1.8,
     cex=1.5, type="l", lwd=2)
lines(1:3,pred2[ny.country=="EI"],
     pch=2, col="red", cex=1.5, lwd=2)
lines(1:3,pred2[ny.country=="PL"],
     pch=2, col="green", cex=1.5, lwd=2)
lines(1:3,pred2[ny.country=="SF"],
     pch=2, col="black", cex=1.5, lwd=2)
legend(2, 5.8, legend=c("DK", "EI", "PL", "SF"),
      title="Lande",pch=20,
      col=c("blue", "red", "green", "black"), cex=1.5)
```

103 / 106



Interaktioner *country*lvitdintake*

Slide 74

```
int3 = lm(log2vitd ~ bmi + sunexp
          +lvitdintake + country
          +country:lvitdintake,
          data=vitd2, na.action=na.exclude)
anova(int3)
```

```
int3est = lm(log2vitd ~ bmi +
             +sunexp + country
             +country:lvitdintake,
             data=vitd2, na.action=na.exclude)
summary(int3est)
confint(int3est)
```

```
cbind(confint(int3est)[8:11,1],
      int3est$coefficients[8:11], confint(int3est)[8:11,2])
```

104 / 106



Blodtryk vs. fedme

Sammenligning af mænd og kvinder

Slide 79-80

```
bp <-
read.table("http://publicifsv.sund.ku.dk/~lts/basal/Rdata/bp.txt",
header=T,na.strings=c("."))

bp$sex <- factor(bp$sexnr, levels=c(1,2),
                labels=c("female", "male"))

bp$log10bp = log10(bp$bp)
bp$log10obese = log10(bp$obese)

ttest = lm(log10bp~sex,data=bp)

ancova = lm(log10bp ~ sex+log10obese ,data=bp)

10^(cbind(-confint(ancova)[2,1],
          -ancova$coefficients[2], -confint(ancova)[2,2]))
```

105 / 106



Blodtryk vs. fedme

Sammenligning af mænd og kvinder, med figur

Kode til figuren, slide 81

```
ny.female = data.frame(log10obese=seq(-0.1,0.4,0.01),sex="female")
pred.female = predict(ancova, ny.female)
ny.male = data.frame(log10obese=seq(-0.1,0.4,0.01),sex="male")
pred.male = predict(ancova, ny.male)

mycolor = c("blue", "red")[bp$sexnr]
mypch = c(1,2)[bp$sexnr]

plot(bp$log10obese, bp$log10bp,
     ylab="log10bp", xlab="log10obese",
     pch=mypch, col=mycolor, cex.lab=1.5)
legend(0.2, 2.02, legend=c("male", "female"), pch=c(1,3),
      col=c("blue", "red"), cex=1.5)
lines(ny.female$log10obese, pred.female, type = "l", lwd=2, col="blue")
lines(ny.male$log10obese, pred.male, type = "l", lwd=2, col="red")
abline(v=0.0725, lty=1,lwd=1,col="red")
abline(v=0.1396, lty=1,lwd=1,col="blue")
```

106 / 106

