Faculty of Health Sciences

Basal Statistik

Multipel regressionsanalyse i R.

Lene Theil Skovgaard / Susanne Rosthøj

Multipel lineær regression

- Eksempel om ultralyd
  - Estimation (film 1)
  - Modelkontrol (film2)
  - Logaritmer og diagnostics (film 3)
- Eksempel om fedme hos børn (film 4)
- Eksempel om lungefunktion
  - Variabelselektion=fisketur (film 5)
  - Kollinearitet (film 6)
- Generelt om modelvalg (film 6)

$*$: Siden er lidt teknisk

Multipel regression

Ét outcome, mange forklarende variable

Vi har allerede set eksempler på dette:

- Sammenligning af vægt for mænd og kvinder, korrigeret for højde
- Vurdering af skuldersmerter efter to behandlinger, korrigeret for baseline smerter
- Effekt af P-piller på hormon-niveauet, med korrektion for alder

I begge disse situationer havde vi netop to kovariater:

- Én kvantitativ (højde, baseline smerter, alder)
- Én kategorisk kovariat (køn, behandling, P-piller)

Problemstillinger ved multipel regression

- Prediktion:
  Konstruktion af normalområde til diagnostisk brug
  (som i eksemplet om ultralyds scanning, s. 5)
- Udredning af årsagssammenhænge,
  til brug for interventioner
  (mere som eksemplet s. 36ff)
- Videnskabelig indsigt,
  f.eks. eksemplet om P-pillers indvirkning på hormoner,
  fra sidste forelæsning

Eksempel med to kvantitative kovariater

107 kvinder er blevet ultralyds scannet få dage inden fødslen, og der ønskes etableret en prediktion af fostervægt/fødselsvægt ud fra BPD (hoved-diameter) og AD (maveomfang).

Data:

```
secher <- read.csv(
   ‘‘http://staff.pubhealth.ku.dk/~sr/BasicStatistics/datasets/secher.txt’’,
   sep=")
```

```
> head(secher)
  vaegt bpd  ad nr
1  2350  88  92  1
2  2450  91  98  2
3  3300  94 110  3
4  1800  84  89  4
5  2900  89  97  5
6  3500 100 110  6
```

```
> tail(secher)
    vaegt bpd  ad  nr
102  2500  90  96 102
103  2000  90  93 103
104  4000  96 115 104
105  3550  92 116 105
106  1173  72  73 106
107  2900  92 104 107
```

Kilde: Secher, N.J., Århus Kommunehospital

Multipel regression

DATA: n personer, dvs. n sæt af sammenhørende observationer:

person	x_1 .... x_p	y
1	x_{11} .... x_{1p}	y_1
2	x_{21} .... x_{2p}	y_2
3	x_{31} .... x_{3p}	y_3
.	. . . . .	.
n	x_{n1} .... x_{np}	y_n

Den lineære regressionsmodel med p forklarende variable skrives:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Parametre:

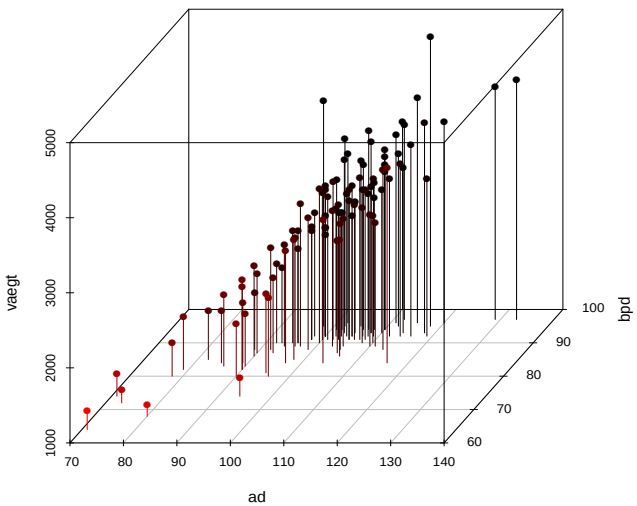
β_0	afskæring, intercept
$\beta_1, \cdots, \beta_p$	regressionskoefficienter

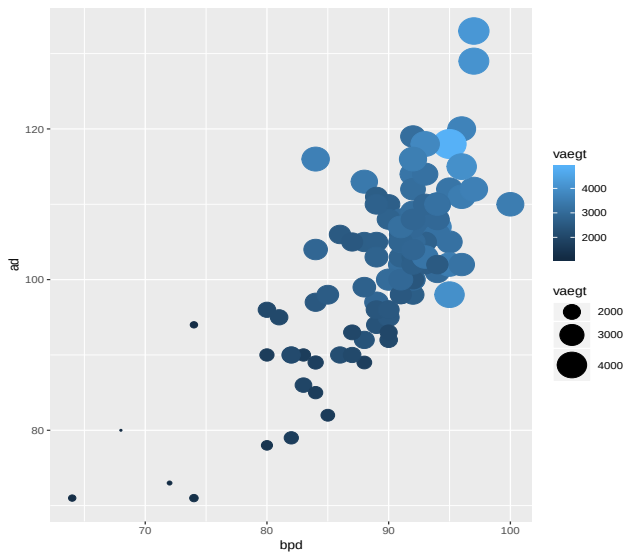
Grafik

er lidt vanskelig, her et par forsøg på 3-dimensionalt plots:

```
#install.packages("scatterplot3d")
library(scatterplot3d)

scatterplot3d(secher$ad,secher$bpd,secher$vaegt,
  pch=16, highlight.3d=TRUE, type="h")
```

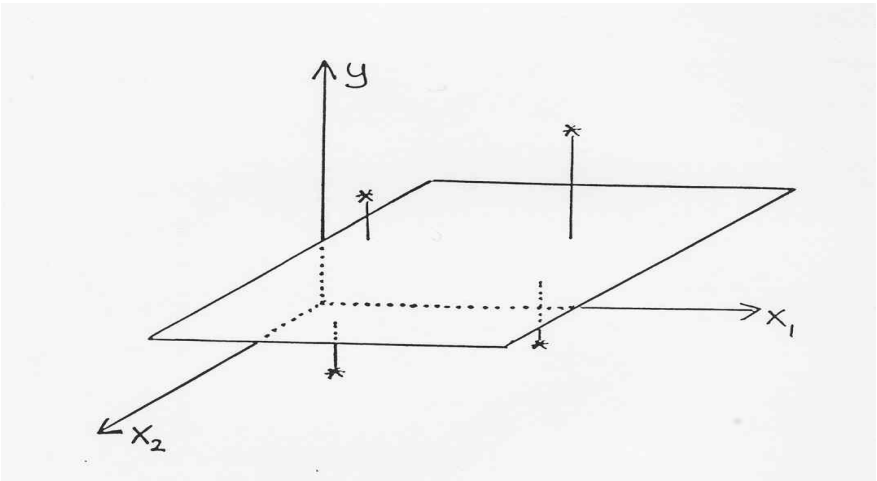
```
#install.packages("ggplot2")
library(ggplot2)

ggplot(secher,
  aes(x=bpd, y=ad, size=vaegt)) +
  geom_point(aes(colour = vaegt))+
  scale_size_continuous(range = c(0.5, 12))
```





*Middelværdistruktur

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$



Mindste kvadraters metode: Minimer størrelsen

$$S(\beta_0, \beta_1, \cdots, \beta_p) = \sum (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

Multipel regression uden transformation

```
model1 = lm(vaegt ~ bpd + ad, data=secher)
summary(model1)
```

med output:

```
Call:
lm(formula = vaegt ~ bpd + ad, data = secher)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4628.118    455.990 -10.150  < 2e-16 ***
bpd            37.133      7.615   4.876 3.90e-06 ***
ad             39.763      4.164   9.549 6.84e-16 ***
---

Residual standard error: 306.6 on 104 degrees of freedom
Multiple R-squared:  0.8065, Adjusted R-squared:  0.8028
F-statistic: 216.7 on 2 and 104 DF,  p-value: < 2.2e-16
```

med fortolkning på næste side

Fortolkning af output

- Stærk signifikant effekt af begge kovariater ($P < 0.0001$)
- Hvis vi sammenligner to fostre med samme `AD`, men hvor foster A har en `BPD`, der er $1mm$ større end foster B, så vil vi forvente, at A vejer $37.1g$ mere end B
- Hvis vi sammenligner to fostre med samme `BPD`, men hvor foster A har en `AD`, der er $1mm$ større end foster B, så vil vi forvente, at A vejer $39.8g$ mere end B
- `Residual standard error`=306.57, så usikkerheden på prediktionen af fødselsvægt er $\pm 2 \times 306.57g$, altså *ganske betragtelig*
  (næsten til lagkage)

Modelkontrol

Hvilke antagelser skal vi checke?

- Uafhængighed:
  Tænk: Er der flere observationer på hvert individ?
  Søskende el.lign?
- Linearitet i begge kovariater
- Varianshomogenitet (residualer har samme spredning)
- Normalfordelte residualer

Obs: Intet krav om normalfordeling på kovariaterne!!

Grafisk modelkontrol: residualer

residualerne = modelafvigelserne
= observeret værdi - fittet værdi: $\hat{\varepsilon}_i = y_i - \hat{y}_i$

Der er flere typer residualer at vælge imellem:

1. *De sædvanlige* = observeret - fittet værdi: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
   i R kaldet `resid(model1)`
2. de sædvanlige, normeret med spredning
   i R kaldet `rstandard(model1)`
3. observeret minus predikteret, normeret,
   men i en model, hvor den aktuelle observation har været udeladt i estimationsprocessen
   i R kaldet `rstudent(model1)`

Hvilke residualer skal man benytte?

Fordele og ulemper

- Rart med residualer, der bevarer enhederne (type 1)
- Lettest at finde outliers ud fra de residualer, hvor observationerne udelades en ad gangen (type 3)
- Bedst at normere, når observationen selv er med og man ikke kan tegne rådata, dvs. for multipel regression bør man nok foretrække type 2 for type 1 (eller se på begge)

Residualplots til modelkontrol

Residualer (af passende type) plottes mod

1. den forklarende variabel x_i
   – for at checke linearitet
   (se efter *krumninger, buer*)
2. de fittede værdier $\hat{y}_i$
   – for at checke varianshomogenitet
   (se efter *trompeter*)
3. fraktildiagram eller histogram
   – for at checke normalfordelingsantagelsen
   (se efter *afvigelse fra en ret linie*)

Disse plots ses på s. 16 og 17, og kommenteres s. 19

Residualplots i praksis

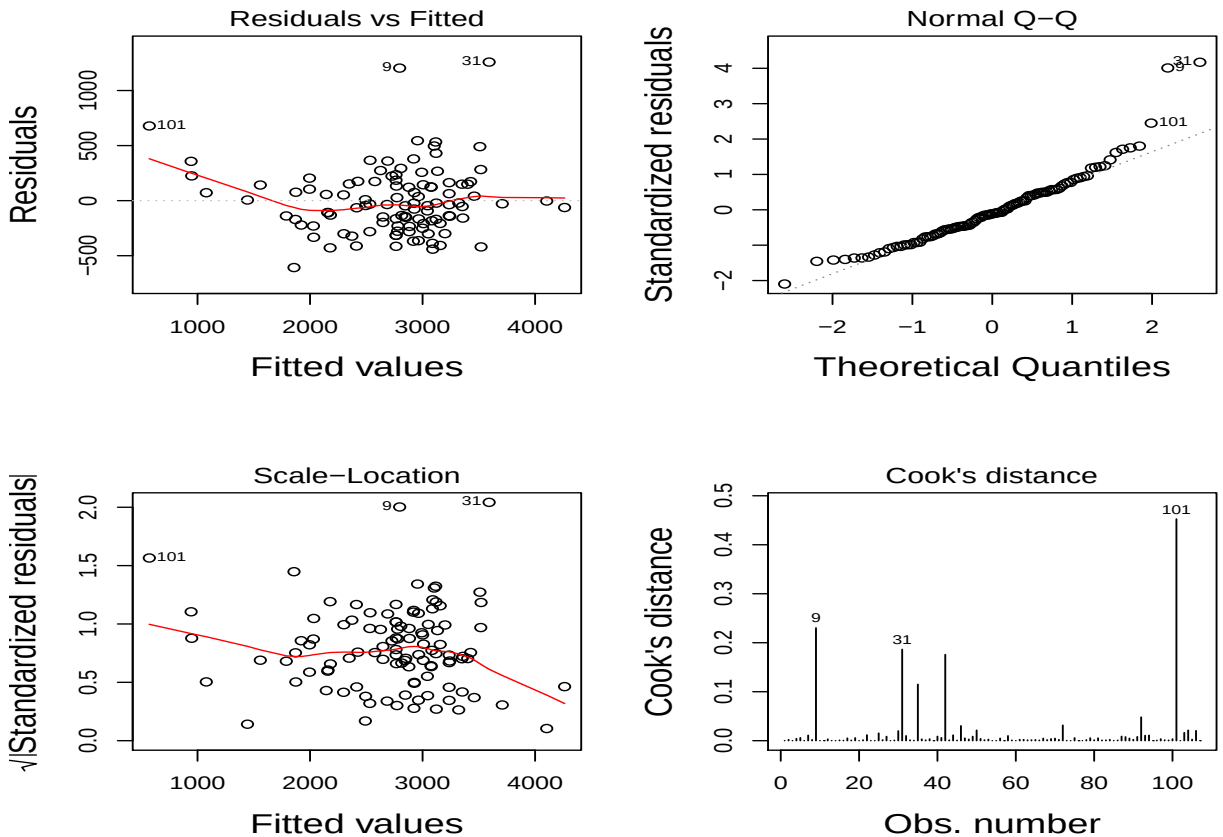
En del modelkontroltegninger (type 2 og 3 fra s. 14) fås automatisk i R ved blot at skrive

```
plot(model1)
```

Blandt disse er også et bedre check af varianshomogeniteten, nemlig et plot af kvadratroden af den numeriske værdi af normerede residualer, mod de predikterede værdier
(se noterne til den simple regression fra tidligere, s. 54)

Disse plots ses på side 16

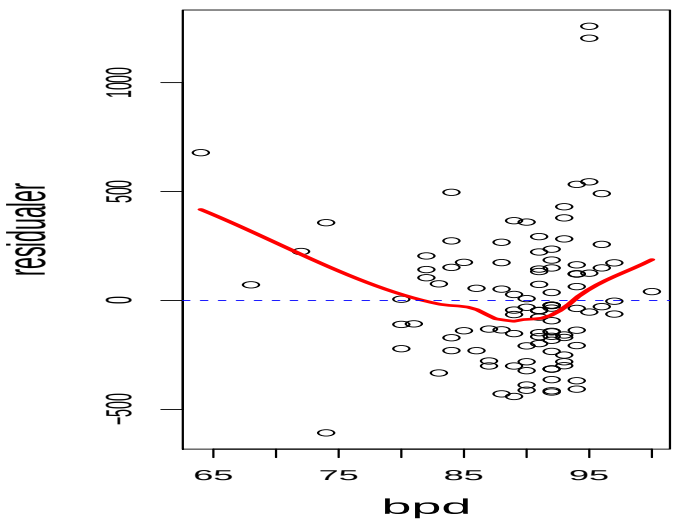
Residualplots: De automatiske plots i R



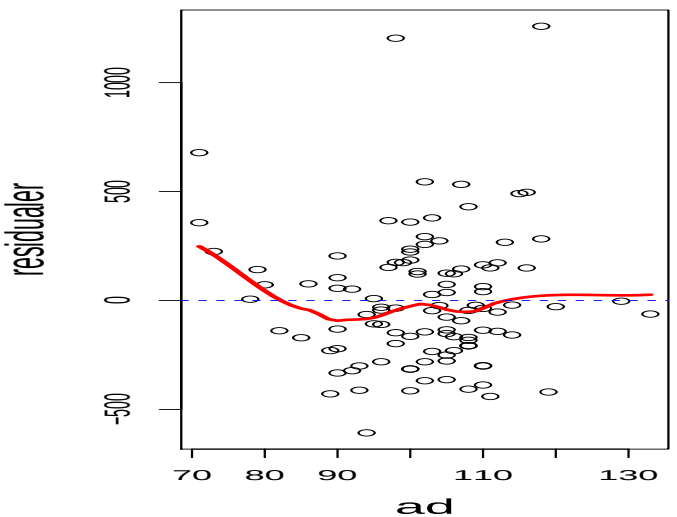
Residualplots

mod kovariater, med indlagte udglattede kurver (Loess-kurver):

```
scatter.smooth( secher$bpd, resid(model1),
ylab=''residualer'', xlab=''bpd'',
lpars = list(col = ``red'', lwd = 3, lty = 1))
abline(0,0,lty=2,col=''blue'')
```

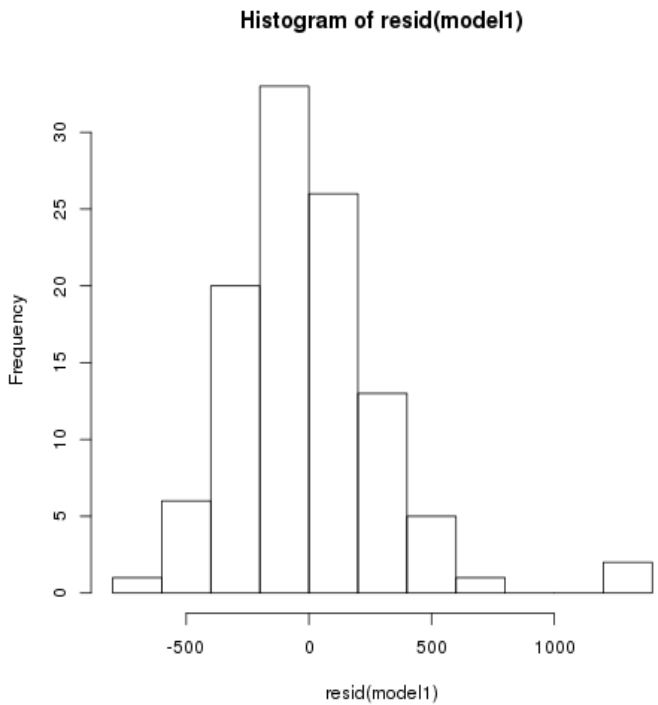



```
scatter.smooth( secher$ad, resid(model1),
ylab=''residualer'', xlab=''ad'',
lpars = list(col = ``red'', lwd = 3, lty = 1))
abline(0,0,lty=2,col=''blue'')
```

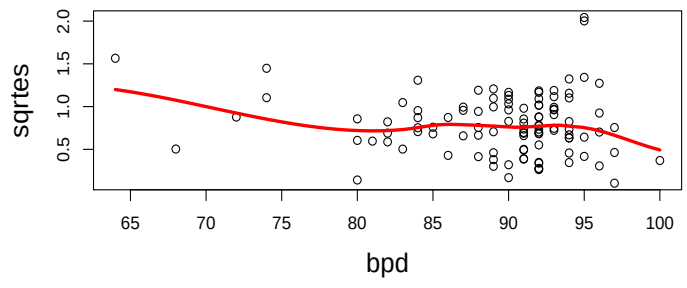


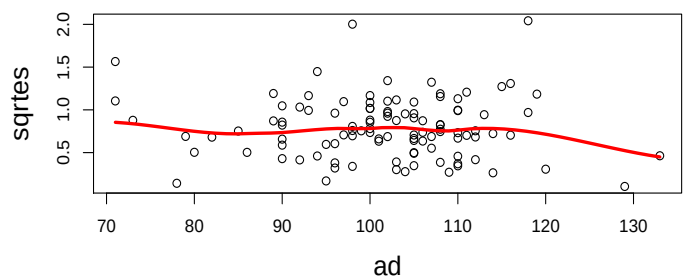
Mere modelkontrol

```
hist( resid(model1) )
```




```
sqrt.res <- sqrt(abs(rstandard(model1)))

scatter.smooth( secher$bpd, sqrt.res,
cex.lab=1.5, ylab=''sqrtes'', xlab=''bpd'',
lpars = list(col = ''red'', lwd = 3, lty = 1))

scatter.smooth( secher$ad, sqrt.res,
cex.lab=1.5, ylab=''sqrtes'', xlab=''ad'',
lpars = list(col = ''red'', lwd = 3, lty = 1))
```





Vurdering af modellen

- Normalfordelingen halter en anelse, med nogle enkelte ret store positive afvigelser, hvilket kunne tale for at logaritmetransformere vægten.
- Måske lidt trompetfacon i plot af residualer mod predikterede værdier, men husk på, at observationerne ikke er ligeligt fordelt over x-aksen.
  Figurerne s. 18 viser ingen oplagte tendenser.
- Linearitet er ikke helt god, men det skyldes hovedsageligt de tidligst fødte børn.
- Teoretiske argumenter fra den faglige ekspertice foreslår en samtidig logaritmetransformation af såvel outcome som kovariater

Analyse af logaritmerede data

Vi benytter forskellige logaritmer til outcome og kovariater:

```
secher$logvaegt <- log(secher$vaegt);
secher$logbpd <- log(secher$bpd/90)/log(1.1);   /* 1 enhed er 10% */
secher$logad <- log(secher$ad/100)/log(1.1);    /* mere om dette s. 21 og 25 */

model2 <- lm( logvaegt ~ logbpd + logad, data=secher )
summary(model2)
confint(model2)
```

giver outputtet:

```
Call:
lm(formula = logvaegt ~ logbpd + logad, data = secher)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.87618    0.01084 726.497  < 2e-16 ***
logbpd       0.14792    0.02187   6.764 8.09e-10 ***
logad        0.13979    0.01398   9.998  < 2e-16 ***

Residual standard error: 0.1068 on 104 degrees of freedom
Multiple R-squared:  0.8583, Adjusted R-squared:  0.8556
F-statistic: 314.9 on 2 and 104 DF,  p-value: < 2.2e-16

> confint(model2)
                2.5 %    97.5 %
(Intercept) 7.8546772 7.8976746
logbpd      0.1045492 0.1912827
logad       0.1120627 0.1675130
```


Fortolkning af resultater

- Stærkt signifikant effekt af begge kovariater ($P < 0.0001$)
- Hvis vi sammenligner to fostre med samme `AD`, men hvor foster A har en `BPD`, der er 10% større end foster B, så vil vi forvente, at A vejer $\exp(0.14792) = 1.16$ gange så meget som B, dvs. 16% mere.
- Hvis vi sammenligner to fostre med samme `BPD`, men hvor foster A har en `AD`, der er der er 10% større end foster B, så vil vi forvente, at A vejer $\exp(0.13979) = 1.15$ gange så meget som B, dvs. 15% mere.

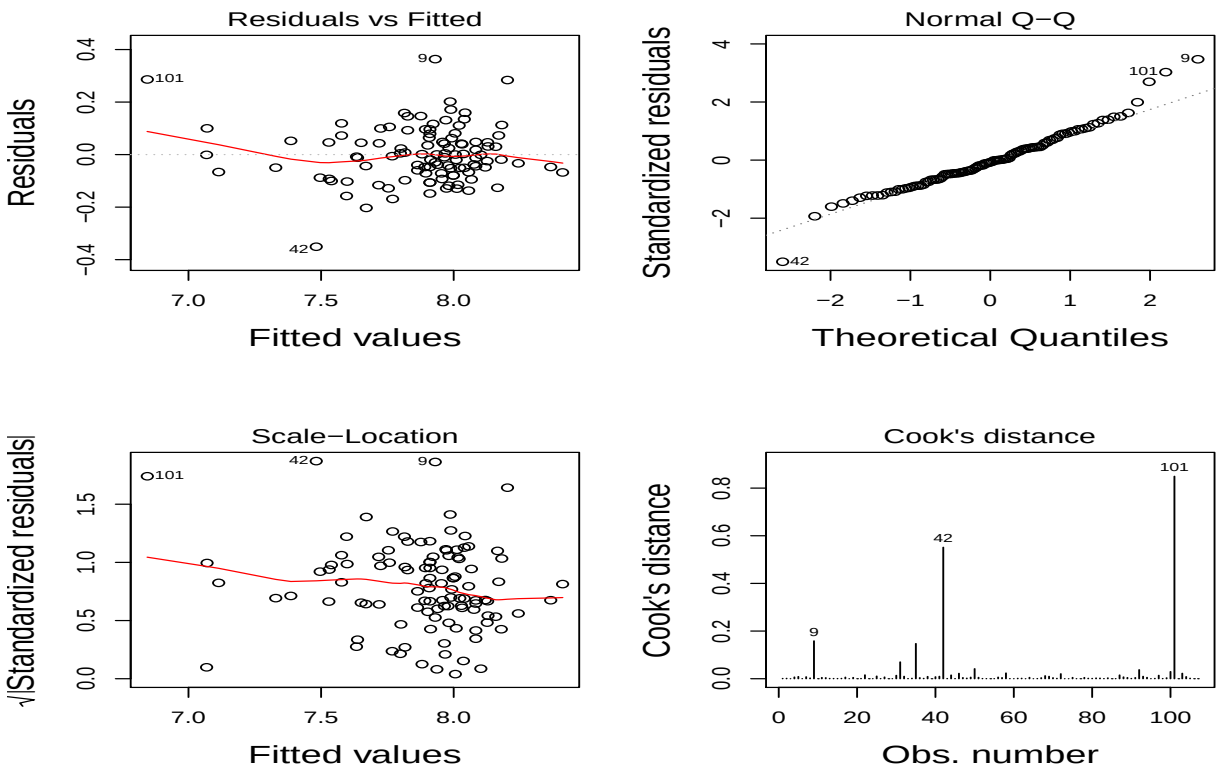
Fortolkning af resultater, fortsat

- `Residual standard error`=0.1068, så usikkerheden på prediktionen af $\log$(fødselsvægt) er $\pm 2 \times 0.1068 = 0.2136$, svarende til faktorerne $(\exp(-0.2136), \exp(0.2136)) = (0.808, 1.238)$, altså en variation gående fra -19.2% op til +23.8%, igen en *ganske betragtelig* biologisk variation.
  Bemærk asymmetrien i denne kvantificering!

- Bemærk, at vi i denne model har konstant *relativ* usikkerhed. Variationskoefficienten kan faktisk nu aflæses som `Residual standard error`, fordi vi arbejder med *naturlige* logaritmer.
  Vi finder tallet 0.1068 (se s. 20), svarende til CV=10.7%

Residualplots for transformerede data

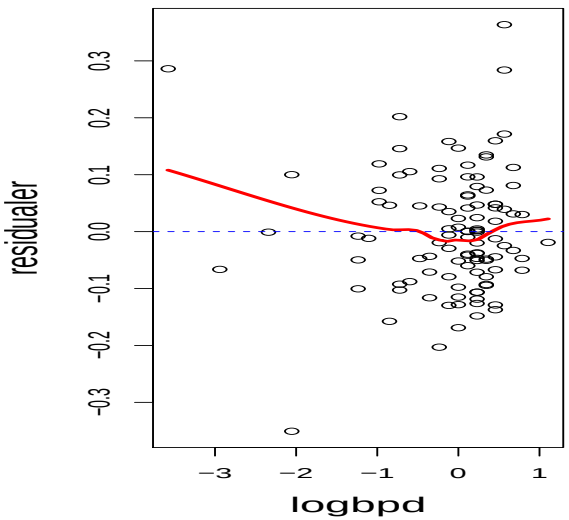
```
par(mfrow=c(2,2))
plot(model2, which=1)
plot(model2, which=2)
plot(model2, which=3)
plot(model2, which=4)
par(mfrow=c(1,1))
```

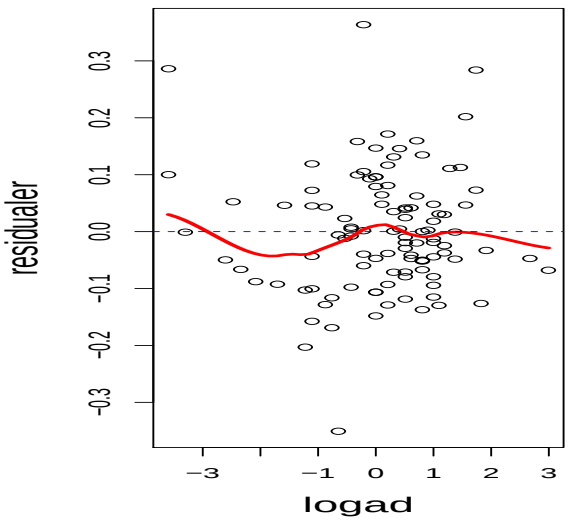


Residualplots for transformerede data, II

```
scatter.smooth( secher$logbpd, resid(model2),
ylab=''residualer'', xlab=''bpd'',cex.lab=1.5,
lpars = list(col = ``red'', lwd = 3, lty = 1))
abline(0,0,lty=2,col=''blue'')
```

```
scatter.smooth( secher$logad, resid(model2),
ylab=''residualer'', xlab=''ad'', cex.lab=1.5,
lpars = list(col = ``red'', lwd = 3, lty = 1))
abline(0,0,lty=2,col=''blue'')
```





Her ses ikke de store ændringer fra s. 16-17, så vi skipper de sidste modelkontroltegninger

Sammenligning af modeller

nemlig de to *marginale* modeller (hver med sin kovariat) og *den multiple* regressionsmodel:

Estimaterne for disse modeller
(med tilhørende standard errors (se) i parentes):

Intercept	β_1 (logbpd)	β_2 (logad)	s	R^2
7.91	0.318 (0.019)	-	0.149	0.72
7.85	-	0.213 (0.011)	0.128	0.80
7.88	0.148 (0.022)	0.140 (0.014)	0.107	0.86

Bemærk fortolkningen af interceptet: På grund af konstruktionen af logbpd og logad på s. 20, svarer interceptet til den forventede fødselsvægt for et barn med bpd=90 og ad=100.

Fortolkning af koefficienter

f.eks. effekten af `bpd`:

- Marginal model:
  Ændringen i `logvaegt`, når kovariaten `logbpd` ændres 1 enhed, dvs. når `bpd` øges med 10%
- Multipel regressionsmodel
  Ændringen i `logvaegt`, når kovariaten `logbpd` ændres 1 enhed, men hvor alle andre kovariater (her kun `ad`) holdes fast

  I sidstnævnte situation siger vi, at vi har korrigeret for effekten af de andre kovariater i modellen.

  Forskellen kan være markant, fordi kovariaterne typisk er relaterede:
  - Når en af dem ændres, ændres de andre også

Goodness-of-fit mål: R-i-anden

$$R^2 = \frac{\text{Sum Sq(Model)}}{\text{Sum Sq(Total)}}$$

"Hvor stor en del af variationen kan forklares af modellen?"
Her finder vi 0.8583, dvs. 85.83% (se s. 20)

Dette mål har

- Fortolkningsproblemer når kovariaterne er styret (ganske som for korrelationskoefficienten)
- R^2 stiger med antallet af kovariater – selv hvis disse er uden betydning!

  Derfor ser man ofte i stedet på Adjusted R^2:

  $R^2_{\text{adj}} = 1 - \frac{\text{Mean Sq(Residual)}}{\text{Mean Sq(Total)}}$ (her 0.8556, se s. 20)

Fittede = predikterede værdier

kan udregnes som (se estimater s. 20)

$$\begin{aligned} \log(\texttt{vaegt}) &= 7.87618 + 0.13979 \times \texttt{logad} \\ &\quad +0.14792 \times \texttt{logbpd} \Rightarrow \\ \texttt{vaegt} &= 0.0028 \times \texttt{ad}^{1.47} \times \texttt{bpd}^{1.55} \end{aligned}$$

I praksis vil man dog benytte `predict` i R, f.eks. ved at skrive

```
secher$yhat <- predict(model2)
head(secher)
```

```
  vaegt bpd  ad nr logvaegt     logbpd      logad     yhat
1  2350  88  92  1 7.762171 -0.2357865 -0.8748447 7.719007
2  2450  91  98  2 7.803843  0.1159355 -0.2119680 7.863694
3  3300  94 110  3 8.101678  0.4562483  1.0000000 8.083450
4  1800  84  89  4 7.495542 -0.7238773 -1.2226796 7.598187
5  2900  89  97  5 7.972466 -0.1172309 -0.3195798 7.814162
6  3500 100 110  6 8.160518  1.1054487  1.0000000 8.179477
```

* Prediktioner til brug for illustrationer (næste side)

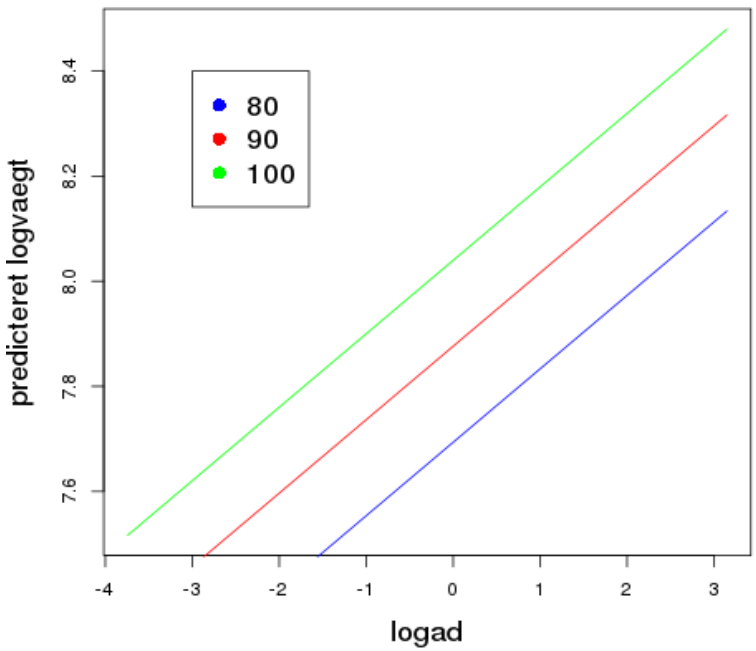
```
ny.80 <- data.frame( ad=70:135, bdp=80 )
ny.80$logbpd <- log(ny.80$bpd) / log(1.1)
ny.80$logad <- log(ny.80$ad) / log(1.1)
```

Tilsvarende for ny.90 og ny.100

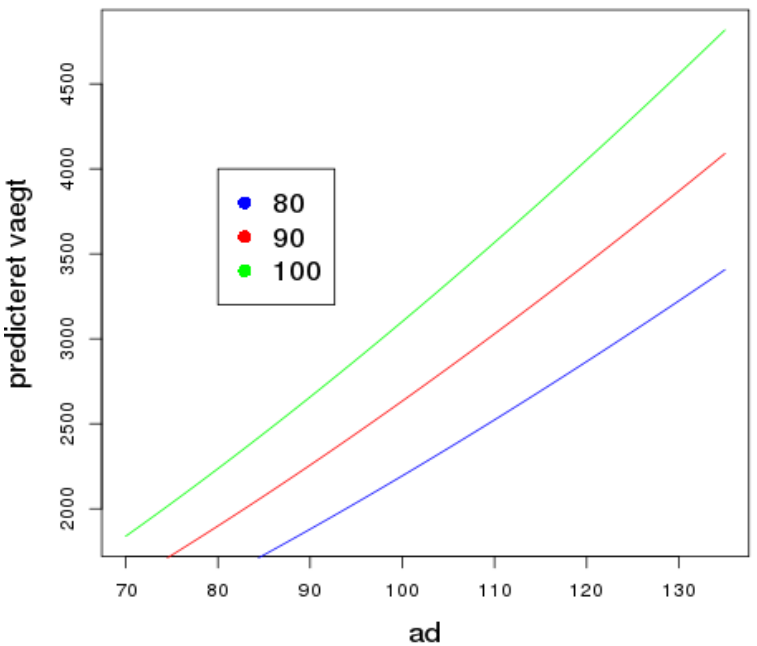
```
pred.80 <-  predict(model2, ny.80)
pred.90 <-  predict(model2, ny.90)
pred.100 <- predict(model2, ny.100)

# Plot paa logaritmisk skala
plot( ny.80$logad, pred.80, type='l', ylim=c(7.5,8.5),
      ylab="predikteret logvaegt", xlab="logad", col="blue" )
lines(ny.90$logad, pred.90, type = "l", col="red")
lines(ny.100$logad, pred.100, type = "l", col="green")
legend( "topleft", legend=c(80,90,100), pch=16, inset=.01,
       col=c("blue", "red", "green") )

# Plot paa oprindelig skala
plot( ny.80$ad, exp( pred.80 ), type='l', ylim=c(1900,4600),
       ylab="predikteret vaegt", xlab="ad", col="blue")
lines( ny.90$ad,  exp( pred.90 ),  type = "l", col="red")
lines( ny.100$ad, exp( pred.100 ), type = "l", col="green")
legend( "topleft", legend=c(80,90,100), lty=1, inset=.01,
        col=c("blue", "red", "green") )
```

Predikterede værdier

På logaritmisk skala:



Tilbagetransformeret
til oprindelig skala:



Regression diagnostics

Understøttes konklusionerne af *hele* materialet?

Eller er der observationer med meget stor indflydelse på resultaterne?

Leverage = *potentiel* indflydelse

(hat-matrix, i R bare kaldet `hat`)
Hvis der kun er én kovariat er det simpelt:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Observationer med *ekstreme* x-værdier *kan* have stor indflydelse på resultaterne, men de har det ikke nødvendigvis!

- ikke hvis de ligger '*pænt*' i forhold til regressionslinien, dvs. har et lille residual, så derfor ...-->

Regression diagnostics, fortsat

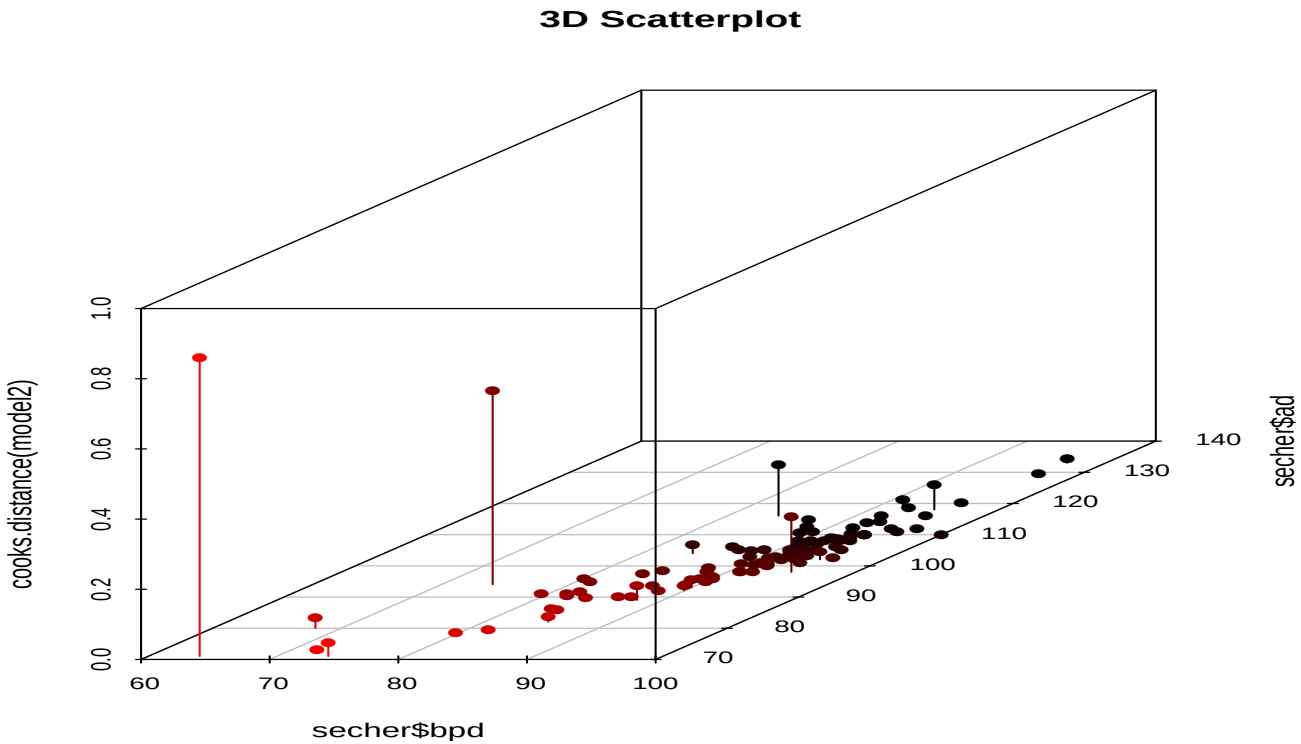
- Udelad den i'te person og find nye estimater for regressionskoefficienterne
- Udregn Cook's afstand,
  et samlet mål for ændringen i parameterestimaterne, og spalt Cooks afstand ud i koordinater:
  Hvor mange se'er ændres f.eks. $\hat{\beta}_1$, når den i'te person udelades?

```
# Cooks afstand i koordinater, svarende til hver kovariat:
dfbetas(model2)

scatterplot3d( secher$bpd, secher$ad, cooks.distance(model2),
               pch=16, highlight.3d=TRUE,
               type="h", main="3D Scatterplot")
```

Hvad gør vi ved indflydelsesrige observationer?

- udelader dem?
- anfører et mål for deres indflydelse?

Cooks afstand som mål for indflydelse



De mest indflydelsesrige observationer er dem med en usædvanlig kombination af kovariater.

Outliers

Observationer, der *ikke passer* ind i sammenhængen

- de er ikke nødvendigvis indflydelsesrige
- de har ikke nødvendigvis et stort residual

Hvad gør vi ved outliers?

- ser nærmere på dem,
  de er tit ganske interessante

Hvornår kan vi *udelade* dem?

- hvis de ligger meget *yderligt*, dvs. har høj leverage
  - husk at afgrænse konklusionerne tilsvarende!
- hvis man kan finde årsagen
  - og da skal alle sådanne udelades!

Interaktion mellem kvantitative kovariater

er lidt svært.....

Hvis modellen ikke beskrives ved en plan, hvad gør man så?

Hvis man blot inkluderer et produktled mellem x_1 og x_2, svarer det til, at antage, at
effekten af x_1 ændrer sig lineært med værdien af x_2:

$$\begin{aligned} \mu &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma x_1 x_2 \\ &= \beta_0 + (\beta_1 + \gamma x_2) x_1 + \beta_2 x_2, \quad \text{dvs.} \end{aligned}$$

Effekten af x_1 er $\beta_1 + \gamma x_2$

Alternativt må man opdele den ene variabel i grupper.....

Nyt eksempel: Fedme i skolealderen

Spørgsmål:
Hvordan hænger fedmegraden i skolealderen sammen med højde og vægt i 1-årsalderen?

For 197 børn har vi sammenhørende registreringer af

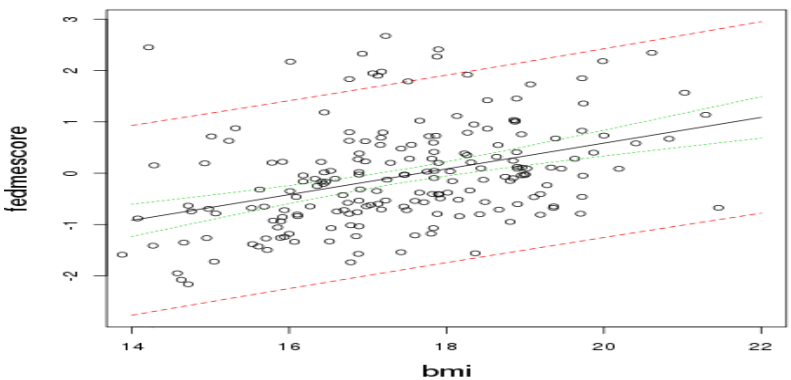
- Højde og vægt i 1-års alderen
- Fedmescore i skolealderen, dvs. en score, der udtrykker barnets vægt i forhold til alder og højde
  Det er en normeret størrelse,
  der typisk vil ligge mellem -2 og 2

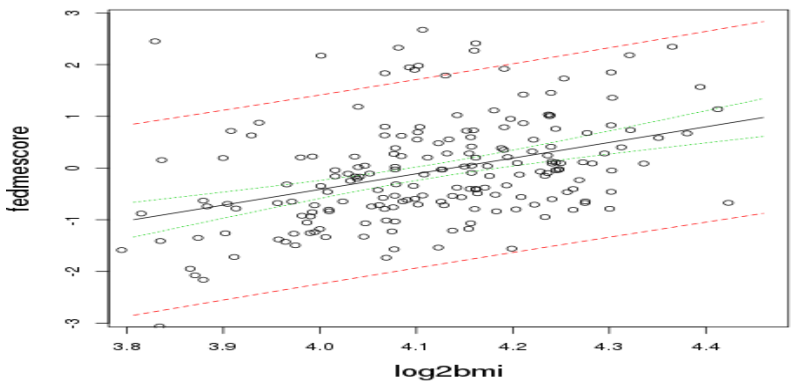
Da fedmescoren kan ses som et slags `bmi`-mål, er det rimeligt at forestille sig, at det hænger sammen med bmi i 1-års alderen.

Fedme i skolealderen

vs. body mass index, evt. logaritmeret

Kode næste side





Det ser nogenlunde pænt lineært ud, omend der er et par observationer, der ser "*for store*" ud.....

Måske skal vi bruge en anden kombination af højde og vægt end lige body mass index? (som ikke plejer at virke så godt for børn...)

Fedme i skolealderen, R-kode

```
fedme <-
read.csv("http://staff.pubhealth.ku.dk/~sr/BasicStatistics/datasets/fedme.txt",sep="")
sd(fedme$fedme)
fedme$fedmescore <- fedme$fedme / 0.2859382
fedme$bmi <- fedme$vaegt/(fedme$hoejde)^2

# venstre plot - fedmescore som funktion af bmi
model1 = lm(fedmescore ~ bmi, data=fedme)

ny = data.frame( bmi=seq(14,22,0.1) )
predl <- predict(model1, newdata=ny, interval = "prediction")
confl <- predict(model1, newdata=ny, interval = "confidence")

plot( fedme$fedmescore ~ fedme$bmi, xlim=c(14,22))
lines( predl[,1] ~ ny$bmi )
lines( predl[,2] ~ ny$bmi, col="red", lty=2 )
lines( predl[,3] ~ ny$bmi, col="red", lty=2 )
lines( confl[,2] ~ ny$bmi, col="green", lty=3 )
lines( confl[,3] ~ ny$bmi, col="green", lty=3 )

# hoejre plot - fedmescore som funktion af log2bmi
fedme$log2bmi= log2(fedme$bmi)
model2 = lm(fedmescore ~ log2bmi, data=fedme)

ny$log2bmi <- log2(ny$bmi)
```

Derefter analogt til venstre plot, med `model2` i stedet for `model1`, og med `log2bmi` i stedet for `bmi`

Fedme i skolealderen

Hvis vi logaritmerer `bmi`, får vi

$$\log(\texttt{bmi}) = \log(\texttt{vaegt}) - 2 \times \log(\texttt{hoejde})$$

så en oplagt mulighed ville være at forsøge en multipel regression, med $\log(\texttt{vaegt})$ og $\log(\texttt{hoejde})$ som kovariater, dvs. med multiplikative effekter af højde og vægt i 1-årsalderen.

Vi bruger igen 1.1-logaritmen:

```
fedme$loghoejde <- log(fedme$hoejde/0.75) / log(1.1)
fedme$logvaegt <- log(fedme$vaegt/10) / log(1.1)
```

og har samtidig normeret, således at interceptet kommer til at svare til et 1-årigt barn på 75 cm, der vejer 10 kg.

Fedme i skolealderen

```
model3 = lm(fedmescore ~ logvaegt + loghoejde, data=fedme)
summary(model3)
```

med output

```
Call:
lm(formula = fedmescore ~ logvaegt + loghoejde, data = fedme,
    na.action = na.exclude)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.004584   0.065491   0.070  0.94427
logvaegt     0.456792   0.067144   6.803 1.24e-10 ***
loghoejde   -0.459933   0.156729  -2.935  0.00374 **

Residual standard error: 0.8793 on 194 degrees of freedom
Multiple R-squared:  0.2348, Adjusted R-squared:  0.2269
F-statistic: 29.76 on 2 and 194 DF,  p-value: 5.342e-12
```

Fortolkning af resultater

- Stærkt signifikant effekt af begge kovariater ($P < 0.0001$ hhv. $P = 0.0037$)
- Hvis vi sammenligner to børn med samme højde, men hvor barn A vejer 10% mere end barn B ved 1-års alderen, så vil vi forvente, at A's fedmescore er 0.457 større end B's.
- Hvis vi sammenligner to børn med samme vægt, men hvor barn A er 10% højere end barn B ved 1-års alderen, så vil vi forvente, at A's fedmescore er 0.460 lavere end B's.

Fortolkning af resultater, fortsat

Hvis det var `bmi`, der var den relevante kovariat,
skulle vi have haft ca.

$$\text{koeff til vægt} = -2\text{koeff til højde}$$

men det har vi ikke

Vi ser på outputtet s. 40, at de to koefficienter snarere er lige store, blot med modsat fortegn, altså at
koeff til vægt $= -$koeff til højde, således at

samme procentvise stigning i højde og vægt
giver uforandret fedmescore i skolealderen.

Sammenligning af modeller

nemlig de to marginale modeller (med hver sin kovariat)
og *den multiple regressionsmodel*:

Estimaterne for disse modeller
(med tilhørende standard errors (se) i parentes):

Intercept	β_1 (`loghoejde`)	β_2 (`logvaegt`)	s	R^2
-0.1004	0.3627 (0.1107)	-	0.976	0.052
-0.0514	-	0.3048 (0.0435)	0.896	0.201
0.0046	-0.4600 (0.1567)	0.4568 (0.0671)	0.879	0.235

Bemærk:

- Koefficienterne ændres ved overgang til multipel regression
  specielt den for højde, som ligefrem skifter fortegn!
- Usikkerhederne på estimaterne bliver større, selvom residualspredningen s bliver mindre.

VIGTIGT

Hvordan kan man forstå to *så forskellige* estimater
for `loghoejde` som 0.3627 og -0.4600?

Fordi:
Den biologiske fortolkning af parameterestimaterne
er helt forskellig

Det videnskabelige spørgsmål, der besvares,
er *ikke* det samme spørgsmål!

Fortolkning af estimater

Koefficienten β_1 til `loghoejde`:

- Simpel/univariat regressionsmodel:
  Forventet øgning i `fedmescore` er 0.36 for 1 enheds øgning af kovariaten `loghoejde`, dvs. for en 10% forøgelse af 1-års højden.
- Multipel regressionsmodel:
  Forventet *fald* i `fedmescore` er -0.46 for en 10% forøgelse af 1-års højden *for to individer, hvor alle andre kovariater* (her kun `logvaegt`) *er identiske* ("holdes fast").

  Det kaldes, at vi har korrigeret ("adjusted") for effekten af de andre kovariater.

Forskellen er her meget markant, fordi kovariaterne er stærkt relaterede: Når en af dem ændres, ændres den anden typisk også

Fortolkning af højde i marginal model

Når højden er den eneste kovariat, er den et udtryk for, hvor stort barnet er, så det videnskabelige spørgsmål, der belyses, er,

- Er børn, der er store i 1-årsalderen, i gennemsnit federe i skolealderen?

Den positive koefficient til højden betyder, at fedmegraden i skolealderen vokser signifikant med størrelsen i 1-års alderen.

Biologisk mekanisme: Overernæring i barnealderen bruges til at vokse, så barnet bliver stort af sin alder, men det kan øjensynligt øge risikoen for fedme senere. Heldigvis er det ikke en stærkt deterministisk sammenhæng (residualspredningen er relativt stor).

Fortolkning af højde i multipel regressionsmodel

Når vægten i 1-års alderen fastholdes, ved vi noget mere om det højeste af de to 1-årige børn:

Det højeste barn må også være det slankeste barn!

Så det nye videnskabelige spørgsmål, der belyses, er,

- Er børn, der er slanke i 1-årsalderen, i gennemsnit slankere i skolealderen?

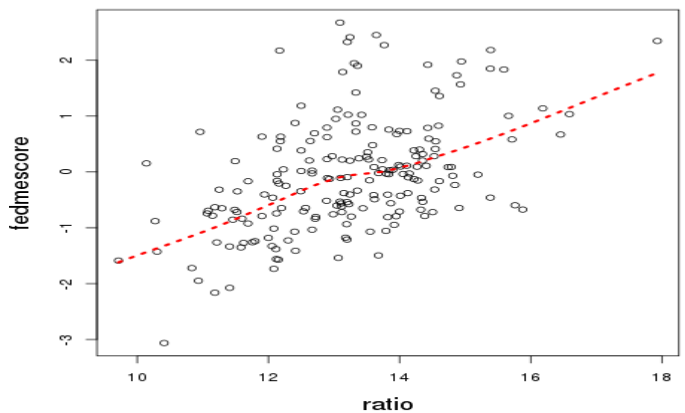
Den negative koefficient til højden betyder, at
slanke småbørn generelt er slankere i skolealderen.

Fortolkning af højde i multipel regressionsmodel, II

Koefficienterne til de logaritme-transformerede højde og vægt er *lige store og med modsat fortegn*. Det betyder:

1. at vægt/højde (altså *ikke* BMI=vægt/højde^2) i 1-års alderen er prædiktivt for fedme i skolealderen
2. at to småbørn med samme vægt/højde-ratio har samme forventede fedmegrad i skolealderen

```
fedme$ratio <- fedme$vaegt / fedme$hoejde
scatter.smooth( fedme$fedmescore ~ fedme$ratio,
lpars=list(col="red",lwd=2,lty=2) )
```



Eksempel: Lungefunktion og cystisk fibrose

Et *lille* studie, af 25 patienter, hvor outcome er `pemax` (et udtryk for lungefunktion) og hvor vi har hele 9 potentielle kovariater:

Table 12.11 Data for 25 patients with cystic fibrosis (O'Neill *et al.*, 1983)

Sub	Age	Sex	Height	Weight	BMP	FEV_1	RV	FRC	TLC	PEmax
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	245	134	85
3	8	0	124	14.1	64	22	441	268	147	100
4	8	1	125	16.2	67	41	234	146	124	85
5	8	0	127	21.5	93	52	202	131	104	95
6	9	0	130	17.5	68	44	308	155	118	80
7	11	1	139	30.7	89	28	305	179	119	65
8	12	1	150	28.4	69	18	369	198	103	110
9	12	0	146	25.1	67	24	312	194	128	70
10	13	1	155	31.5	68	23	413	225	136	95
11	13	0	156	39.9	89	39	206	142	95	110
12	14	1	153	42.1	90	26	253	191	121	90
13	14	0	160	45.6	93	45	174	139	108	100
14	15	1	158	51.2	93	45	158	124	90	80
15	16	1	160	35.9	66	31	302	133	101	134
16	17	1	153	34.8	70	29	204	118	120	134
17	17	0	174	44.7	70	49	187	104	103	165
18	17	1	176	60.1	92	29	188	129	130	120
19	17	0	171	42.6	69	38	172	130	103	130
20	19	1	156	37.2	72	21	216	119	81	85
21	19	0	174	54.6	86	37	184	118	101	85
22	20	0	178	64.0	86	34	225	148	135	160
23	23	0	180	73.8	97	57	171	108	98	165
24	23	0	175	51.1	71	33	224	131	113	95
25	23	0	179	71.5	95	52	225	127	101	195

```
pemax <-
read.csv("http://staff.pubhealth.ku.dk/~sr/BasicStatistics/datasets/pemax.txt",
sep="")
```

Ex. O'Neill et.al. (*Am Rev Respir Dis*, 1983)

Univariate analyser

også kaldet *marginale* analyser,
hvor man ser på *en enkelt kovariat ad gangen:*

Table 12.12 Results of separately regressing PEmax on each explanatory variable

Explanatory variable	Regression coefficient	Standard error	t	P
Age	4.055	1.088	3.73	0.0011
Sex	−19.045	13.176	−1.45	0.16
Height	0.932	0.260	3.59	0.0016
Weight	1.187	0.301	3.94	0.0006
BMP	0.639	0.565	1.13	0.27
FEV_1	1.354	0.555	2.44	0.023
RV	−0.123	0.077	−1.59	0.12
FRC	−0.319	0.145	−2.20	0.038
TLC	−0.358	0.404	−0.89	0.38

Her er der 5 signifikante variable

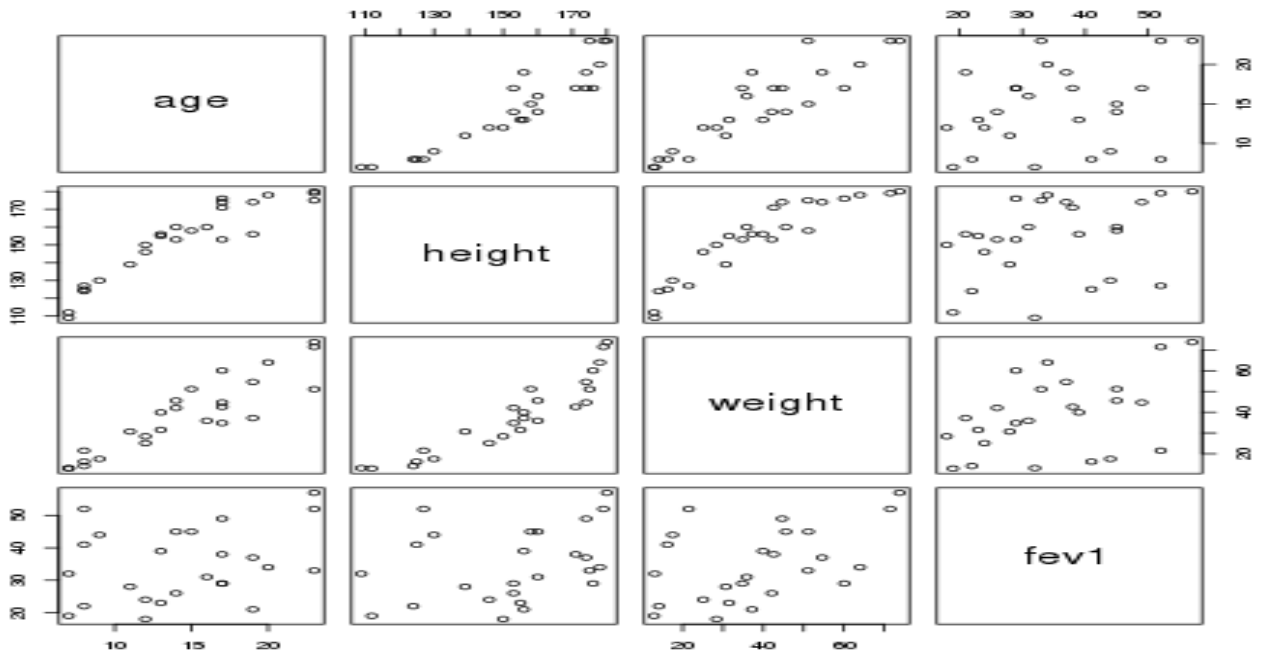
Er det så disse variable, der skal med i modellen?

Valg af model

kommer helt og holdent an på, hvad spørgsmålet er!

De univariate analyser kan være stærkt misvisende pga korrelationer mellem de enkelte variable, f.eks. 4 af de 5 signifikante variable: Age, Height, Weight og FEV_1:

```
names(pemax)
# plotter variabel nr 2+4+5+7, alle mod alle:
plot( pemax[,c(2,4,5,7)] )
```



Videnskabelig variabelselektion

Endnu en gang:
Gennemtænk præcis hvilket videnskabeligt spørgsmål, man ønsker besvaret – det præcise spørgsmål bestemmer hvilke variable, der skal inkluderes i modellen.

Det er svært
– men den eneste måde at opnå egentlig videnskabelig indsigt!

(og så bliver det i øvrigt lettere bagefter at skrive en god artikel og lettere at svare på reviewernes kommentarer . . .)

Automatisk variabelselektion

Undgå så vidt muligt dette!

- **Forlæns selektion:**
  Medtag hver gang den mest signifikante

```
# Forwards selection:
tom <- lm( pemax ~ 1, data=pemax)
step(tom, direction = "forward", scope=list(upper=fuld,lower=tom),
trace=T)
```

  Slutmodel: `Weight BMP FEV1`

- **Baglæns elimination:**
  Start med alle, udelad hver gang den mindst signifikante

```
# Den fulde model:
fuld=lm( pemax ~ age+factor(sex)+height+weight+bmp+fev1+rv+frc+tlc,
data=pemax )
# Backwards elimination
step(fuld, direction = "backward", trace=T)
```

  Slutmodel: `Weight BMP FEV1`

- **Forskellige hybrid-metoder**

```
# Hybrid mellem forwards og backwards
step(tom, direction = "both",  scope=list(upper=fuld,lower=tom),trace=T)
```

men men men...

Det så godt nok meget stabilt ud!?
...men hvis nu observation nr. 25 tilfældigvis ikke havde været med?

Så ville forlæns selektion have *taget højde ind som den første*,
og baglæns elimination ville have *smidt højde ud som den første*!

Tommelfingerregel (vedrører kun stabiliteten)

Antallet af observationer skal være mindst 10 gange så stort som antallet af *undersøgte* parametre!

Advarsel ved variabelselektion

- Massesignifikans!
- *Enhver* variabelselektion baseret på signifikansvurderinger (dvs. også når I selv fjerner enkelte variable pga. statistisk insignifikans) vil give følgende effekt:
  - Signifikanserne overvurderes!
  - Regressionsparametrene er "for store", dvs. for langt væk fra 0.
- Automatisk variabelselektion:
  Hvad kan vi sige om 'vinderne'?
  - Var de hele tiden signifikante, eller blev de det lige pludselig?
  - I sidstnævnte tilfælde kunne de jo være blevet smidt ud, mens de var insignifikante...

Model med alle 9 kovariater

```
> summary(fuld)

Call:
lm(formula = pemax ~ age + factor(sex) + height + weight + bmp +
    fev1 + rv + frc + tlc, data = pemax)

Residuals:
    Min      1Q  Median      3Q     Max
-37.338 -11.532   1.081  13.386  33.405

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  176.0582   225.8912   0.779    0.448
age           -2.5420     4.8017  -0.529    0.604
factor(sex)1  -3.7368    15.4598  -0.242    0.812
height        -0.4463     0.9034  -0.494    0.628
weight         2.9928     2.0080   1.490    0.157
bmp           -1.7449     1.1552  -1.510    0.152
fev1           1.0807     1.0809   1.000    0.333
rv             0.1970     0.1962   1.004    0.331
frc           -0.3084     0.4924  -0.626    0.540
tlc            0.1886     0.4997   0.377    0.711

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-squared:  0.6373, Adjusted R-squared:  0.4197
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

Model med alle 9 kovariater, II

Bemærk på outputtet s. 56:

- Alle kovariater er enkeltvis non-signifikante
- *Overall F-test* giver signifikans (P=0.032)

Og hvad kan vi overhovedet bruge modellen til....?

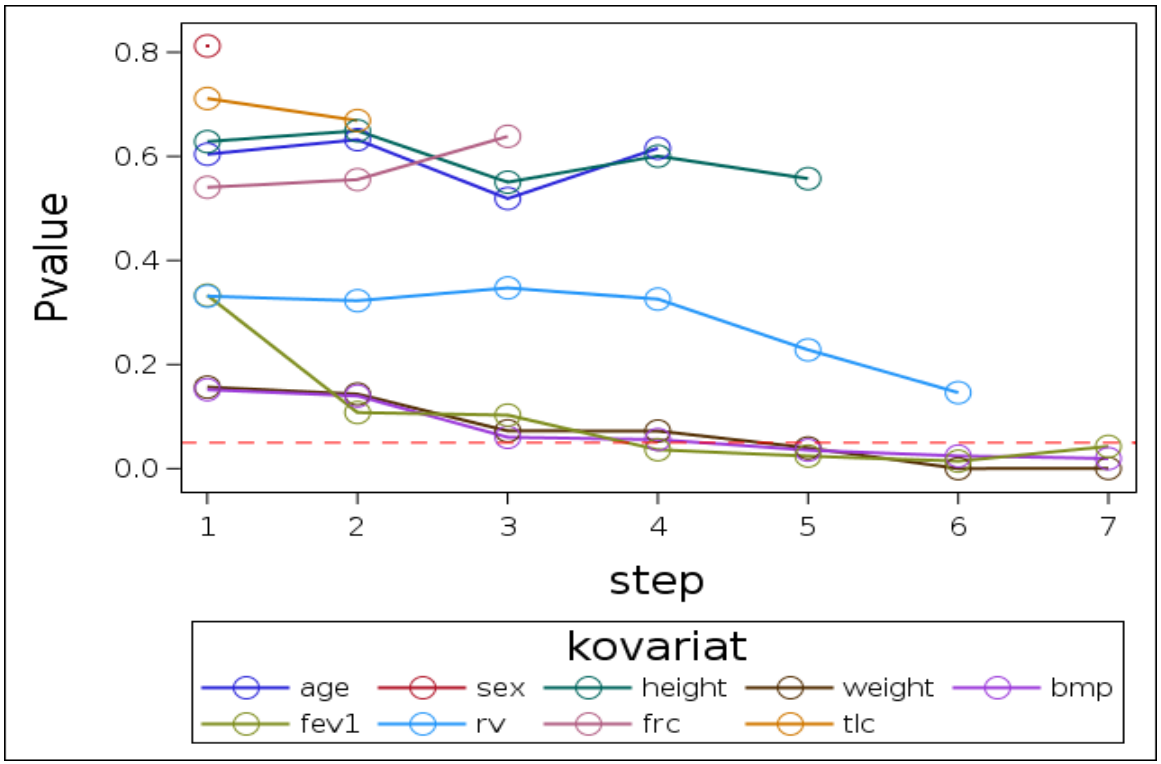
Baglæns elimination

Tabel over successive p-værdier

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
age	0.604	0.632	0.519	0.616	-	-	-	-	-
sex	0.812	-	-	-	-	-	-	-	-
height	0.628	0.649	0.550	0.600	0.557	-	-	-	-
weight	0.157	0.143	0.072	0.072	0.040	0.000	0.000	0.000	0.001
bmp	0.152	0.140	0.060	0.056	0.035	0.024	0.019	0.098	-
fev1	0.333	0.108	0.103	0.036	0.024	0.014	0.043	-	-
rv	0.331	0.323	0.347	0.326	0.228	0.146	-	-	-
frc	0.540	0.555	0.638	-	-	-	-	-	-
tlc	0.711	0.669	-	-	-	-	-	-	-

Altman stopper ved skridt nr. 7, se figur næste side

P-værdier ved baglæns elimination



Ingen kode til denne figur

Krydsvalidering = Cross validation

Hvis man har en tilstrækkelig stor mængde data, er det en god fremgangsmåde at:

- Foretage modelfittet på en del af data,
  f.eks. to trediedele
- Afprøve modellen på resten bagefter,
  og måske blive lidt chokeret

Hvis modellen predikterer dårligt på den sidste trediedel, predikterer den nok også dårligt i nye situationer.

Sammenligning af modeller

Hvad sker der ved udeladelse af en forklarende variabel?

- Fittet bliver dårligere, dvs.
  residualkvadratsummen bliver større.
- Antallet af frihedsgrader (for residualkvadratsummen) stiger.
- Estimatet s^2 for residualvariansen σ^2 kan både stige og falde
  (fordi vi dividerer med frihedsgraderne)
- %-delen af variation, som forklares af modellen, R^2, falder.
  Dette kompenseres der for i den
  justerede determinationskoefficient R^2_{adj}

Som kriterium for, om modellen er god
kan vi altså bruge s^2 eller R^2_{adj}

Sammenligning af modeller

nemlig de to marginale modeller for `height` og `weight`, samt den multiple regressionsmodel med disse:

```
model1 <- lm( pemax ~ height, data=pemax); summary( model1 )
model2 <- lm( pemax ~ weight, data=pemax); summary( model2 )
model3 <- lm( pemax ~ height+weight, data=pemax); summary( model3 )
```

β_0	β_1(height)	β_2(weight)	s	R^2	Adj. $\bar{R}^2$
-33.276	0.932(0.260)	-	27.34	0.36	0.33
63.546	-	1.187(0.301)	26.38	0.40	0.38
47.355	0.147(0.655)	1.024(0.787)	26.94	0.40	0.35

- Hver af de to forklarende variable har betydning, vurderet ud fra de marginale modeller.
- I den multiple regressionsmodel ser *ingen af dem* ud til at have nogen betydning.
- Mindst en af dem har en betydning, men det er svært at sige hvilken. (Det ser dog mest ud til at være vægten.)

Sammenligning af modeller, II

Flere kommentarer til sammenligningen på forrige side:

- Størrelsen R^2 stiger næsten ikke fra den marginale model med kun `weight` som kovariat, til den multiple regressionsmodel (fra 0.4035 til 0.4049, dvs. uændret med 2 decimaler).
- Den justerede R^2 favoriserer derfor den simpleste af disse to modeller, hvor usikkerheden på parameterestimatet også er betydeligt lavere.
- Intercepterne i de 3 modeller kan ikke sammenlignes, da de refererer til helt forskellige individer, som i øvrigt alle er meget uinteressante: hhv. `height=0`, `weight=0` og `height=weight=0`

Modelkontrol: giver ikke frygtelig meget mening med den ratio af observationer/kovariater:

Sammenligning af kovariat-effekter

er vanskelig, bare at formulere:

Er effekten af vægt større end effekten af alder? Øh....

- Sammenligner vi P-værdier?
  Det har kun noget med *evidensen* for en effekt at gøre, ikke med selve størrelsen af effekten.
- Man kan udregne såkaldte standardiserede koefficienter, som angiver effekten af 1 SD svarende til denne kovariat, i enheder af SD i outcome-variablen.
  Er dette fornuftigt?
  Næppe: Den biologiske effekt er nok ligeglad med, hvordan resten af personerne ser ud...
  En sådan standardiseret koefficient vil afhænge af det aktuelle sample - ligesom en korrelation.

Kollinearitet: Kovariaterne er lineært relaterede

Det vil de altid til en vis grad være, undtagen i designede forsøg (f.eks. landbrugsforsøg)

Symptomer på kollinearitet:

- Visse af kovariaterne er stærkt korrelerede
- Nogle parameterestimater har meget store standard errors
- Alle kovariater i den multiple regressionsanalyse er insignifikante, men R^2 er alligevel stor
- Der sker store forskydninger i estimaterne, når en kovariat udelades af modellen
- Der sker store forskydninger i estimaterne, når en observation udelades af modellen
- *Resultaterne er anderledes end forventet*

Ekstra udskrifter fra regressionsanalysen

- Variance Inflation Factor:
  Den faktor, som variansen af regressionskoefficienten er blevet ganget op med, på grund af kollineariteten mellem kovariaterne.
- Tolerance:
  Blot den reciprokke af VIF, altså 1/VIF.
- Standardiserede koefficienter:
  som forklaret s. 64.

Tommelfingerregel: `vif` større end 5-10 stykker (dvs. `tol` mindre end 0.2-0.1) er *kriminelt*.....

Kollinearitet i praksis

```
# Foerst skaleres alle kovariater
names(pemax)
ny.x  <- scale( pemax[ , c(2:10)], center=TRUE, scale = TRUE )

# Disse bruges derefter i regressionsmodellen, hvorefter
# koefficienterne divideres med SD af outcome:
fuld.ny <- lm( pemax$pemax ~ ny.x )

stand.beta <- coef(fuld.ny)[2:10] / sd(pemax$pemax)
# install.packages("car")
library(car)
var.inflation <- vif( fuld )
tolerance <- 1 / vif( fuld )

# Til sidst flettes det hele sammen (cbind = Column BIND):
cbind( stand.beta, tolerance, var.inflation)
```

```
            stand.beta  tolerance var.inflation
ny.xage    -0.38459707 0.04580885     21.829841
ny.xsex    -0.05661823 0.44064381      2.269407
ny.xheight -0.28694289 0.07165927     13.954929
ny.xweight  1.60199595 0.02092869     47.781303
ny.xbmp    -0.62650931 0.14053329      7.115752
ny.xfev1    0.36189810 0.18451863      5.419507
ny.xrv      0.50671367 0.09489420     10.538052
ny.xfrc    -0.40327443 0.05833260     17.143073
ny.xtlc     0.09570900 0.37594085      2.659993
```


Bemærk

Hvad gør man så, når der er kollinearitet?

1. Gennemtænk grundigt, hvad *den enkelte variabel* står for afhængigt af hvilke af de andre mulige variable, der fastholdes (= er med i modellen)
   - Overvej også, om *responsvariablen* ændrer fortolkning
2. Lav analyser af fokusvariablen med og uden justering for forskellige grupper af de andre variable og prøv at forstå forskellene i resultaterne
   - Præsenter gerne begge/alle analyser i artiklen med fortolkning af forskellene
3. Spar eventuelt på antallet af variable for grupper af variable, der hænger sammen: Drejer det sig om ét fælles aspekt af interesse, så kan man måske nøjes med én af variablene og begrunde, hvorfor man vælger netop den
4. Fortolk med stor forsigtighed

Vigtige pointer

- Man må *ikke* nøjes med at præsentere univariate analyser for alle variablene!

  Som f.eks. s. 50

  Problemet med fortolkningen forsvinder nemlig *ikke* af, at man tillægger hver enkelt variabel al forklaringsevnen en ad gangen.

- Man må *ikke* tro, at det er ligemeget hvilke effekter, man justerer for - for det videnskabelige spørgsmål ændrer sig jo.

Bemærk

Den statistiske model bestemmes af det videnskabelige spørgsmål.

Eneste undtagelse fra denne regel er fordelingsantagelserne, der bestemmes af data.

Bemærk

Nogle kilder til forkerte modelvalg

Forsimple virkeligheden **for** meget, eksempelvis

- tro, at frasen *alt andet lige* giver mening (og ikke bare er en bekvem undskyldning for ikke at tænke det ordentligt igennem)
- tro, at det giver mening at tale om sammenhæng**en** mellem exposure og respons, som om der kun kan eksistere ét eneste videnskabeligt spørgsmål, der involverer denne exposure og dette respons
- tro, at *confounder***ne** er givet ud fra exposure og respons, så man kan vælge sine kovariater uden at overveje konsekvenserne for fortolkningen

Et par falske påstande

Påstand: Når man skal vurdere effekten af en forklarende variabel ("exposure") på et bestemt outcome, så skal man så vidt muligt justere for alle confoundere.

Sandhed: Nej! Man skal **kun** justere for de variable, som man gerne ville have kunnet holde fast
– og man skal kunne gennemskue, hvilken konsekvens justering for hver eneste af de inkluderede confoundere har for fortolkningen af den estimerede effekt af exposure på outcome!

Et par falske påstande, II

Påstand: Man må ikke inkludere mediator variable, dvs. variable, der er en del af virkningsmekanismen.

Sandhed: Mediator variable er ligesom alle andre variable: Hvis man inkluderer dem, ændrer man det videnskabelige spørgsmål, der besvares ved analysen! Når man inkluderer en eller flere mediator variable, undersøger man størrelsen af den del af effekten, der *ikke* går via effekten på disse mediator variable, altså styrken af eventuelle *andre* virkningsmekanismer.

Eksempel på fejlslutning

Kan et øget fedtindtag være gavnligt for hjertet?

f.eks reducere risikoen for hjerteinfarkt, eller sænke kolesterol i blodet?

Tja....:
Øget motion giver øget fødeindtag og dermed formentlig øget fedtindtag, så hvis man har fedtindtag som *eneste* kovariat, så kan det se gavnligt ud at spise meget fedt....

Bemærk

Påstand: Signifikansen for den enkelte variabel bliver altid svagere, når de andre tages med.

Sandhed: Ofte, men ikke altid. Nogle gange bliver signifikanserne væsentligt stærkere.

Vi så dette ved forelæsningen om kovariansanalyse, hvor effekten af P-piller først kom frem, da vi justerede for alderen.