

Faculty of Health Sciences

# Basal Statistik

Multipel regressionsanalyse.

Lene Theil Skovgaard

5. oktober 2020



# Multipel lineær regression

- ▶ Regression med to kvantitative kovariater:
  - ▶ Eksempel om ultralyd
  - ▶ Eksempel om fedme hos børn
- ▶ Selektion, modelvalg
- ▶ Modelkontrol
- ▶ Diagnostics

E-mail: [ltsk@sund.ku.dk](mailto:ltsk@sund.ku.dk)

\*: 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 er:

| OBS | VAEGT | BPD | AD  |
|-----|-------|-----|-----|
| 1   | 2350  | 88  | 92  |
| 2   | 2450  | 91  | 98  |
| 3   | 3300  | 94  | 110 |
| .   | .     | .   | .   |
| .   | .     | .   | .   |
| .   | .     | .   | .   |
| 105 | 3550  | 92  | 116 |
| 106 | 1173  | 72  | 73  |
| 107 | 2900  | 92  | 104 |

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:

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

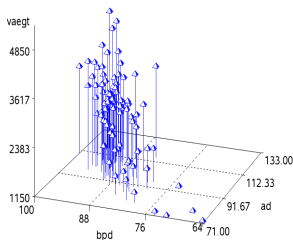
**Parametre:**

$\beta_0$  afskæring, intercept  
 $\beta_1, \cdots, \beta_p$  regressionskoefficienter

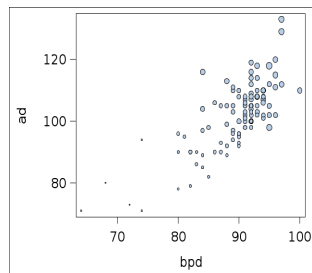
# Grafik

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

```
proc g3d;  
  scatter bpd*ad=vaegt /  
    shape='pillar' size=0.5;  
run;
```

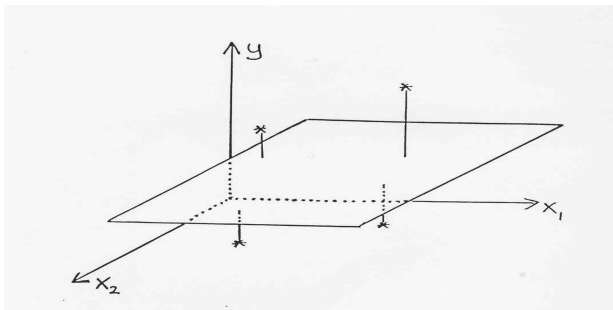


```
proc sgplot data=a1;  
  bubble Y=ad X=bpd size=vaegt  
    / bradiusmin=1 bradiusmax=6;  
run;
```



# \*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

```
proc reg data=secher;  
model vaegt=ad bpd / clb;  
run;
```

med output:

Dependent Variable: vaegt

|                |            |          |        |
|----------------|------------|----------|--------|
| Root MSE       | 306.57298  | R-Square | 0.8065 |
| Dependent Mean | 2739.09346 | Adj R-Sq | 0.8028 |
| Coeff Var      | 11.19250   |          |        |

## Parameter Estimates

| Variable  | DF | Parameter Estimate | Standard Error | t Value | Pr >  t |
|-----------|----|--------------------|----------------|---------|---------|
| Intercept | 1  | -4628.11813        | 455.98980      | -10.15  | <.0001  |
| bpd       | 1  | 37.13292           | 7.61510        | 4.88    | <.0001  |
| ad        | 1  | 39.76305           | 4.16394        | 9.55    | <.0001  |

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  $1\text{mm}$  større end foster B, så vil vi forvente, at A vejer  $37.1\text{g}$  mere end B
- ▶ Hvis vi sammenligner to fostre med samme BPD, men hvor foster A har en AD, der er  $1\text{mm}$  større end foster B, så vil vi forvente, at A vejer  $39.8\text{g}$  mere end B
- ▶ Root MSE=306.57, så usikkerheden på prediktionen af fødselsvægt er  $\pm 2 \times 306.57\text{g}$ , altså *ganske betragtelig* (næsten til lagkage)
- ▶ Variationskoefficienten er 11.2%.....



# 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

Der er 4 typer residualer at vælge imellem:

1. *De sædvanlige* ( $r=$ ) = observeret - fittet værdi:  $\hat{\varepsilon}_i = y_i - \hat{y}_i$
2. student: de sædvanlige, normeret med spredning
3. press:  
observeret minus predikteret, men i en model, hvor den aktuelle observation har været udeladt i estimationsprocessen
4. rstudent: normerede Press-residualer

Alle disse kan fås i nyt datasæt ved at skrive (se også s. 78):

```
output out=ny p=yhat cookd=cook r=resid  
student=stresid press=press rstudent=uresid;
```



# Hvilke residualer skal man benytte?

## Fordele og ulemper

- ▶ Rart med residualer, der bevarer enhederne (type 1 og 3)
- ▶ Lettest at finde outliers ud fra de residualer, hvor observationerne udelades en ad gangen (type 3 og 4)
- ▶ 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

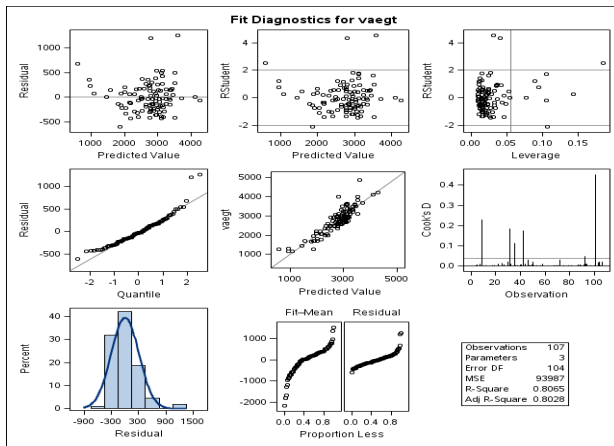
En del modelkontroltegninger (faktisk alle de nævnte fra forrige side) fås automatisk i SAS ved at benytte `ods graphics`:

```
proc glm PLOTS=(DIAGNOSTICS RESIDUALS(smooth)) data=secher;
```

Et **bedre check af varianshomogeniteten** er dog et plot af kvadratrod af den numeriske værdi af normerede residualer, mod de predikterede værdier, eller mod de forklarende variable (se noterne til den simple regression fra tidligere, s. 54, det er default plot i R):

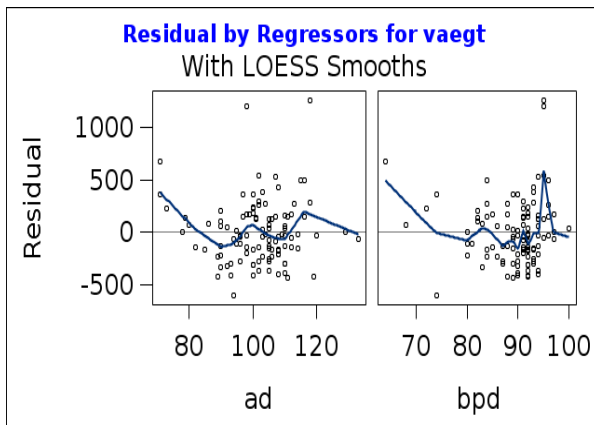
Disse *specialplots* ses på side 18

# Residualplots: Diagnostics Panel



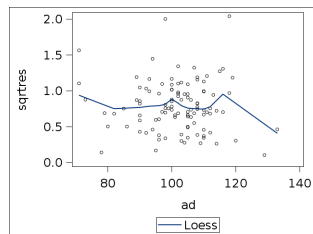
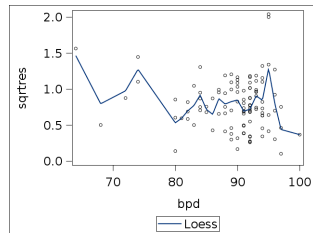
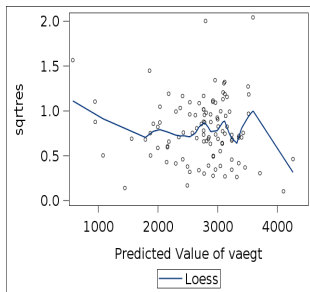
# Residualplots

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



# Specialplots til vurdering af konstant spredning

som beskrevet s. 15 (kode s. 78)



# 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.

Figuren s. 18 viser ingen oplagte tendenser.

- ▶ Linearitet er ikke helt god, men det skyldes hovedsageligt de tidligst fødte børn.
- ▶ Teoretiske argumenter fra den faglige ekspertise foreslår en samtidig logaritmetransformation af såvel outcome som kovariater

# Analyse af logaritmerede data

Vi benytter forskellige logaritmer til outcome og kovariater:

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

Koden svarende til s. 9 (se også s. 79) giver outputtet:

```
Dependent Variable: logvaegt
Root MSE      0.10679      R-Square      0.8583
Dependent Mean 7.87952      Adj R-Sq      0.8556
Coeff Var      1.35530
```

## Parameter Estimates

| Variable  | DF | Parameter Estimate | Standard Error | t Value | Pr >  t |
|-----------|----|--------------------|----------------|---------|---------|
| Intercept | 1  | 7.87618            | 0.01084        | 726.50  | <.0001  |
| logad     | 1  | 0.13979            | 0.01398        | 10.00   | <.0001  |
| logbpd    | 1  | 0.14792            | 0.02187        | 6.76    | <.0001  |

| Variable  | DF | 95% Confidence Limits |         |
|-----------|----|-----------------------|---------|
| Intercept | 1  | 7.85468               | 7.89767 |
| logad     | 1  | 0.11206               | 0.16751 |
| logbpd    | 1  | 0.10455               | 0.19128 |



# 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 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.

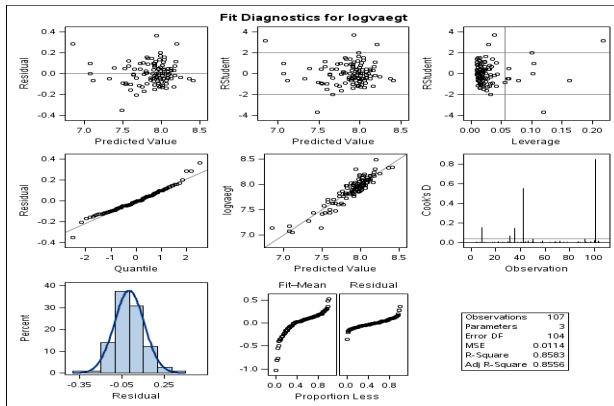


## Fortolkning af resultater, fortsat

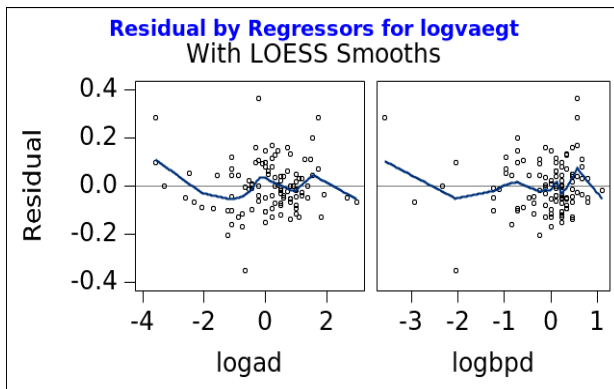
- ▶ Root MSE=0.10679, så usikkerheden på prediktionen af  $\log(\text{fødselsvægt})$  er  $\pm 2 \times 0.10679 = 0.21358$ , svarende til faktorerne  $(\exp(-0.21358), \exp(0.21358)) = (0.808, 1.238)$ , altså en variation gående fra -19.2% op til +23.8%, igen en *ganske betragtelig* biologisk variation.  
**Bemærk asymmetrien i denne kvantificering!**
- ▶ Bemærk, at vi i denne model har konstant *relativ* usikkerhed. Variationskoefficienten kan faktisk nu aflæses som Root MSE, fordi vi arbejder med *naturlige* logaritmer. Vi finder tallet 0.10679 (se s. 20), svarende til CV=10.7%



# Residualplots for transformed data



## Residualplots for transformerede data, II



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 | $\beta_1$ (logbpd) | $\beta_2$ (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 logvægt, når kovariaten  $\log bpd$  ændres 1 enhed, dvs. når bpd øges med 10%

- ▶ **Multipel regressionsmodel**

Ændringen i logvægt, når kovariaten  $\log bpd$  ændres 1 enhed, men hvor alle andre kovariater (her kun  $ad$ ) **holdes fast**

I sidstnævnte situation siger vi, at vi har **korregeret** 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

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

“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  $R^2$** :

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



## Fitted = predikterede værdier

kan udregnes som (se estimator s. 20)

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

I praksis vil man dog benytte en `output-sætning`:

```
output out=ny p=yhat cookd=cook r=resid  
student=stresid press=press rstudent=uresid;
```

således at de predikterede værdier bliver lagt som variabelen `yhat` i datasættet `ny`



## \* Prediktioner til brug for illustrationer

- ▶ Tilføj de fiktive personer, man gerne vil prediktere outcome for
- ▶ Angiv deres kovariatværdier, med tilhørende missing value for outcome
- ▶ Predikter outcome i nyt datasæt
- ▶ Tilbagetransformer til original skala
- ▶ Tegn

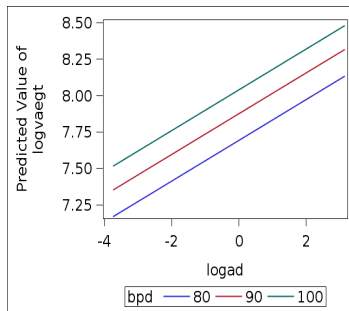
De fiktive personer har ingen indflydelse på estimation i modellen, da de mangler outcome-værdier

Se eksempelvis kode s. 80-81, der danner figurerne s. 30

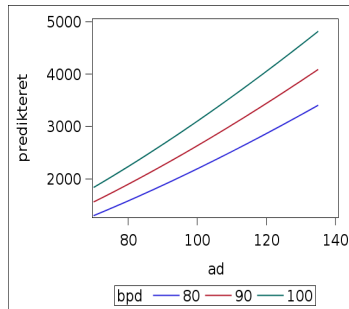


# Predikterede værdier

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 SAS kaldet Hat Diag eller H)

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
- ▶ Spalt Cooks afstand ud i koordinater og angiv: Hvor mange *se*'er ændres f.eks.  $\hat{\beta}_1$ , når den i'te person udelades?

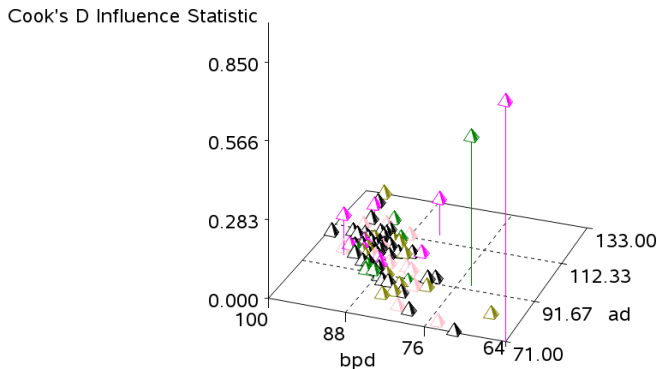
Se kode s. 82

### 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.

Indviklet kode s. 82



# 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

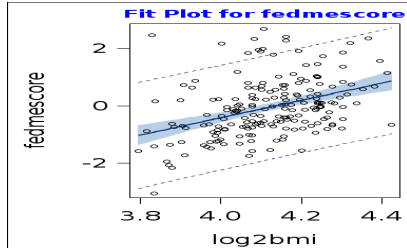
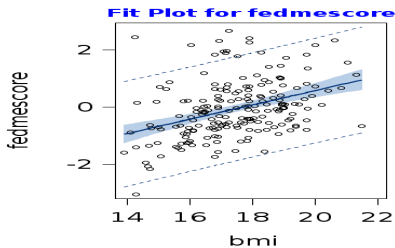
Det er en normeret størrelse,  
der typisk vil ligge mellem -2 og 2

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:



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

Hvis vi logaritmerer  $\text{bmi}$ , får vi

$$\log(\text{bmi}) = \log(\text{vægt}) - 2 \times \log(\text{højde})$$

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

Vi bruger igen 1.1-logaritmen:

$$\log_{\text{højde}} = \log(\text{højde}/0.75)/\log(1.1);$$

$$\log_{\text{vægt}} = \log(\text{vægt}/10)/\log(1.1);$$

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

```
proc reg plots=(diagnostics residuals(smooth)) data=fedme;  
model fedmescore=logvaegt loghoejde / clb;  
output out=check rstudent=norm_u_resid;  
run;
```

## med output

Dependent Variable: fedmescore

Number of Observations Used 197

|                |          |          |        |
|----------------|----------|----------|--------|
| Root MSE       | 0.87927  | R-Square | 0.2348 |
| Dependent Mean | -0.05956 | Adj R-Sq | 0.2269 |

| Variable  | DF | Parameter | Standard | t Value | Pr >  t |
|-----------|----|-----------|----------|---------|---------|
|           |    | Estimate  | Error    |         |         |
| Intercept | 1  | 0.00458   | 0.06549  | 0.07    | 0.9443  |
| logvaegt  | 1  | 0.45679   | 0.06714  | 6.80    | <.0001  |
| loghoejde | 1  | -0.45993  | 0.15673  | -2.93   | 0.0037  |

| Variable  | DF | 95% Confidence Limits |          |
|-----------|----|-----------------------|----------|
| Intercept | 1  | -0.12458              | 0.13375  |
| logvaegt  | 1  | 0.32437               | 0.58922  |
| loghoejde | 1  | -0.76904              | -0.15082 |



# 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. 39, at de to koefficienter snarere er lige store, blot med modsat fortegn, altså at  $\text{koeff til vægt} = -\text{koeff til højde}$ , 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 | $\beta_1$ (loghøjde) | $\beta_2$ (logvægt) | $s$   | $R^2$ |
|-----------|----------------------|---------------------|-------|-------|
| -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 parameterestimerne  
er helt forskellig

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

# Fortolkning af estimerer

Koefficienten  $\beta_1$  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 logvægt) *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?

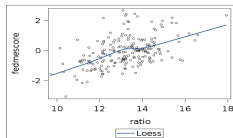
Den negative koefficient til højden betyder, at slanke småbørn generelt er slankere i skolealderen.



## 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*  $BMI = \text{vægt}/\text{højde}^2$ ) 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



# 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 <sub>1</sub> | 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   |

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 | <i>t</i> | 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 <sub>1</sub>     | 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. de 5 signifikante variable: Age, Height, Weight, FEV<sub>1</sub> og FRC:

Pearson Correlation Coefficients, N = 25  
Prob > |r| under H0: Rho=0

|        | age     | height            | weight            | fev1              | frc                |
|--------|---------|-------------------|-------------------|-------------------|--------------------|
| age    | 1.00000 | 0.92605<br><.0001 | 0.90587<br><.0001 | 0.29449<br>0.1530 | -0.63936<br>0.0006 |
| height |         | 1.00000           | 0.92070<br><.0001 | 0.31666<br>0.1230 | -0.62428<br>0.0009 |
| weight |         |                   | 1.00000           | 0.44884<br>0.0244 | -0.61726<br>0.0010 |
| fev1   |         |                   |                   | 1.00000           | -0.66511<br>0.0003 |

Bemærk især korrelationerne mellem alder, højde og vægt.



# 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!

Der findes forskellige fremgangsmåder (se kode s. 83):

- ▶ **Forlæns selektion:**

Medtag hver gang den mest signifikante

**Slutmodel:** Weight BMP FEV1

- ▶ **Baglæns elimination:**

Start med alle, udelad hver gang den mindst signifikante

**Slutmodel:** Weight BMP FEV1

- ▶ **Forskellige hybrid-metoder**

Det ser da meget stabilt ud!?



men men 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

Dependent Variable: pemax

Analysis of Variance

| Source  | DF | Sum of<br>Squares | Mean<br>Square | F Value | Prob>F |
|---------|----|-------------------|----------------|---------|--------|
| Model   | 9  | 17101.39040       | 1900.15449     | 2.929   | 0.0320 |
| Error   | 15 | 9731.24960        | 648.74997      |         |        |
| C Total | 24 | 26832.64000       |                |         |        |

Root MSE      25.47057      R-square      0.6373

| Variable  | DF | Parameter<br>Estimate | Standard<br>Error | T for H0:<br>Parameter=0 | Prob >  T |
|-----------|----|-----------------------|-------------------|--------------------------|-----------|
| intercept | 1  | 176.058206            | 225.89115895      | 0.779                    | 0.4479    |
| age       | 1  | -2.541960             | 4.80169881        | -0.529                   | 0.6043    |
| sex       | 1  | -3.736781             | 15.45982182       | -0.242                   | 0.8123    |
| height    | 1  | -0.446255             | 0.90335490        | -0.494                   | 0.6285    |
| weight    | 1  | 2.992816              | 2.00795743        | 1.490                    | 0.1568    |
| bmp       | 1  | -1.744944             | 1.15523751        | -1.510                   | 0.1517    |
| fev1      | 1  | 1.080697              | 1.08094746        | 1.000                    | 0.3333    |
| rv        | 1  | 0.196972              | 0.19621362        | 1.004                    | 0.3314    |
| frc       | 1  | -0.308431             | 0.49238994        | -0.626                   | 0.5405    |
| tlc       | 1  | 0.188602              | 0.49973514        | 0.377                    | 0.7112    |



# Model med alle 9 kovariater, II

**Bemærk** på outputtet s. 55:

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

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



# Baglæns elimination

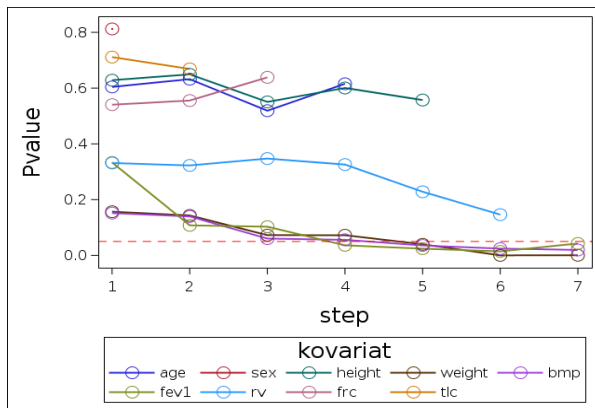
## Tabel over successive $p$ -værdier

|        | [,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  $\sigma^2$  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*  $R_{\text{adj}}^2$

Som **kriterium for, om modellen er god** kan vi altså bruge  $s^2$  eller  $R_{\text{adj}}^2$



# Sammenligning af modeller

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

| $\beta_0$ | $\beta_1(\text{height})$ | $\beta_2(\text{weight})$ | $s$   | $R^2$ | 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 parameterestimerer har meget store standard errors
- ▶ Alle kovariater i den multiple regressionsanalyse er insignifikante, men  $R^2$  er alligevel stor
- ▶ Der sker store forskydninger i estimererne, når en kovariat udelades af modellen
- ▶ Der sker store forskydninger i estimererne, når en observation udelades af modellen
- ▶ *Resultaterne er anderledes end forventet*



# Ekstra udskrifter fra regressionsanalysen

- ▶ **Variance Inflation Factor** ( $vif$ ):  
Den faktor, som variansen af regressionskoefficienten er blevet ganget op med, på grund af kollineariteten mellem kovariaterne.
- ▶ **Tolerance** ( $tol$ ):  
Blot den reciprokke af VIF, altså  $1/VIF$ .
- ▶ **Standardiserede koefficienter** ( $stb$ ):  
som forklaret s. 63.



# Kollinearitet i praksis

Ud fra koden s. 84:

Parameter Estimates

| Variable  | DF | Standardized<br>Estimate | Tolerance | Variance<br>Inflation |
|-----------|----|--------------------------|-----------|-----------------------|
| Intercept | 1  | 0                        | .         | 0                     |
| age       | 1  | -0.38460                 | 0.04581   | 21.82984              |
| sex       | 1  | -0.05662                 | 0.44064   | 2.26941               |
| height    | 1  | -0.28694                 | 0.07166   | 13.95493              |
| weight    | 1  | 1.60200                  | 0.02093   | 47.78130              |
| bmp       | 1  | -0.62651                 | 0.14053   | 7.11575               |
| fev1      | 1  | 0.36190                  | 0.18452   | 5.41951               |
| rv        | 1  | 0.50671                  | 0.09489   | 10.53805              |
| frc       | 1  | -0.40327                 | 0.05833   | 17.14307              |
| tlc       | 1  | 0.09571                  | 0.37594   | 2.65999               |

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



# 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. 49

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 sammenhængen mellem exposure og respons, som om der kun kan eksistere ét eneste videnskabeligt spørgsmål, der involverer denne exposure og dette respons
- ▶ tro, at *confounder*er 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 forelæsningsen om kovariansanalyse, hvor effekten af P-piller først kom frem, da vi justerede for alderen.



# APPENDIX

med SAS-programbidder svarende til nogle af slides

- ▶ Figurer: s. 76, 78, 81, 82
- ▶ Regressionsanalyse: s. 77, 79, 83-84
- ▶ Modelkontrol: s. 78
- ▶ Prediktion: s. 80-81
- ▶ Diagnostics: s. 82



# Tredimensionalt plot

## Slide 7

```
proc g3d;  
  scatter bpd*ad=vaegt /  
  shape='pillar' size=0.5;  
run;
```

Option `shape` styrer faconen på figuren på toppen af søjlen, og option `size` angiver dennes størrelse. Denne procedure kan også anvendes til *Surface plots*, altså plots af flader i rummet.

*Alternativt bubble plot*, hvor outcome styrer størrelsen af symbolerne:

```
proc sgplot data=a1;  
  bubble Y=ad X=bpd size=vaegt  
  / bradiusmin=1 bradiusmax=6;  
run;  
76 / 84
```



# Regressionsanalyse

## Slide 9

Den *minimale* kode:

```
proc reg data=secher;  
model vaegt=ad bpd / clb;  
run;
```

Adskillige ekstra options og statements kan tilføjes, f.eks:

- ▶ clb: konfidensgrænser for estimer
- ▶ corrb: korrelation mellem estimer
- ▶ stb: standardiserede koefficienter:  
effekt af ændring på 1 SD for kovariat

Se desuden s. 78 og 81-84.



# Modelkontrol

## Slide 12-18

```
proc reg plots=(diagnostics residuals(smooth)) data=secher;  
model vaegt=ad bpd / clb;  
output out=check student=stresid rstudent=norm_u_resid p=pred;  
run;
```

```
data tegn;  
set check;
```

```
sqrtres=sqrt(abs(stresid)); run;
```

```
proc sgplot data=tegn;  
  loess Y=sqrtres X=pred; run;
```

```
proc sgplot data=tegn;  
  loess Y=sqrtres X=bpd; run;
```

```
proc sgplot data=tegn;  
  loess Y=sqrtres X=ad; run;
```



# Regressionsanalyse på logaritmer

## Slide 20

```
data secher;  
set secher;  
  
logvaegt=log(vaegt);  
logbpd=log(bpd/90)/log(1.1);  
logad=log(ad)/log(1.1);  
run;  
  
proc reg plots=(diagnostics residuals(smooth)) data=secher;  
model logvaegt=logad logbpd / clb;  
output out=check rstudent=norm_u_resid p=pred;  
run;
```



# Prediktioner til brug for illustrationer

## Slide 29 og 30

- ▶ Tilføj de fiktive personer, man gerne vil prediktere outcome for, ved at angive deres kovariatværdier, med tilhørende missing value for outcome

```
data ekstra;  
ekstra=1;  
do bpd=80 to 100 by 10;  
do ad=70 to 135 by 1;  
logbpd=log(bpd/90)/log(1.1);  
logad=log(ad)/log(1.1);  
output;  
end;  
end;
```

```
data udbygget;  
set secher ekstra;
```

- ▶ Predikter outcome, og gem i nyt datasæt (næste side)
- ▶ Tegn (næste side)



# Prediktioner til brug for illustrationer, II

## Slide 29 og 30

```
proc reg plots=(diagnostics residuals(smooth)) data=udbygget;  
model logvaegt=logad logbpd / clb;  
output out=check p=pred LCL=predlow LCLM=meanlow UCL=predup UCLM=meanup;  
run;
```

```
proc sgplot data=check; where ekstra=1;  
    series Y=pred X=logad / group=bpd  
        lineattrs=(thickness=2 pattern=solid); run;
```

```
data pred;  
set check; if ekstra=1;
```

```
predikteret=exp(pred); type='P'; output;  
type='L'; predikteret=exp(predlow); output;  
type='U'; predikteret=exp(predup); output; run;
```

```
proc sgplot data=pred; where type='P';  
    series Y=predikteret X=ad / group=bpd  
        lineattrs=(thickness=2 pattern=solid); run;
```



# Diagnostics

## Slide 32-33

```
proc reg plots=(diagnostics residuals(smooth)) data=secher;  
model logvaegt=logad logbpd / clb influence;  
output out=check rstudent=norm_u_resid p=pred cookd=cook;  
run;
```

```
data tegn;  
set check;
```

```
farve="magenta"; /* bare for at faa navnet langt nok */  
farve="black";  
if norm_u_resid>0.5 then farve="pink";  
if norm_u_resid>1.2 then farve="magenta";  
if norm_u_resid<-0.5 then farve="olive";  
if norm_u_resid<-1.2 then farve="green";  
run;
```

```
proc g3d data=tegn;  
scatter bpd*ad=cook / color=farve shape="pyramid";  
run;
```



# Variabelselektion, option *selection* i REG

## Slide 52, 57

`SELECTION=name`

specifies the method used to select the model, where name can be FORWARD (or F), BACKWARD (or B), STEPWISE, MAXR, MINR, RSQUARE, ADJRSQ, CP, or NONE (use the full model). The default method is NONE. See the section Model-Selection Methods for a description of each method.

```
proc reg data=pemax;  
  model pemax=age sex height weight bmp fev1 rv frc tlc  
    / selection=backward;  
run;
```



# Ekstra størrelser fra regressionsanalyse

## Slide 66

```
proc reg data=pemax;  
  model pemax=age sex height weight bmp fev1 rv frc tlc  
    / selection=backward  
      p r influence vif collin stb clb corrb tol;  
run;
```

### Options:

- ▶ collin: kollinearitets diagnostics
- ▶ vif: variance inflation factor:  
variansøgning grundet kollinearitet
- ▶ tol: tolerance factor:  
 $1-R^2$  for regression af en kovariat på de øvrige

