Introducerende Statistik og Dataanalyse med R

Multipel Regression

Jens Ledet Jensen

I dag

Multipel regressionsmodel

Backward og forward selektion

Multipel regression: auktionspriser på bornholmerure

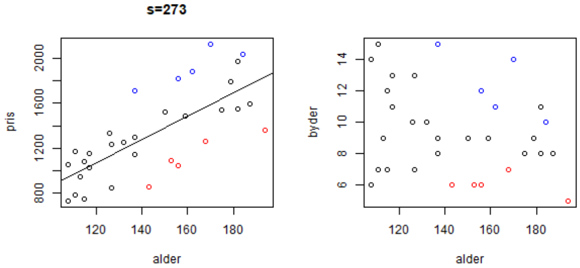
Ønsker tommelfingerregel for pris af bornholmerure

Skal vi beskrive pris ud fra *alder* på ur eller ud fra antal *bydere* til auktion ?

eller: skal vi inddrage begge variable i modellen ?

Data for 32 grandfather clocks			
Nummer	Alder (år)	Antal Bydere	Pris (pund)
1	127	13	1235
2	115	12	1080
⋮			
31	194	5	1356
32	168	7	1262

Respons: pris, Forklarende variable: alder, byder



Afvigelser ved regression af *pris* på *alder* kan måske forklares ved antal *byder*

alder og *byder* tilsammen vil nok give bedre beskrivelse af data end *alder* alene

Model: $\text{Pris}_i \sim\sim N(\xi_i, \sigma^2)$

data er normalfordelt med ens varianser

$\xi_i = \alpha + \beta_{\text{alder}} \cdot \text{alder}_i + \beta_{\text{byder}} \cdot \text{byder}_i$ (multipel regression)

Forventede: $\hat{\xi}_i = \hat{\alpha} + \hat{\beta}_{\text{alder}} \cdot \text{alder}_i + \hat{\beta}_{\text{byder}} \cdot \text{byder}_i$,

Residual: $r_i = x_i - \hat{\xi}_i$

alder og byder er regressionsvariable (ikke faktorer)

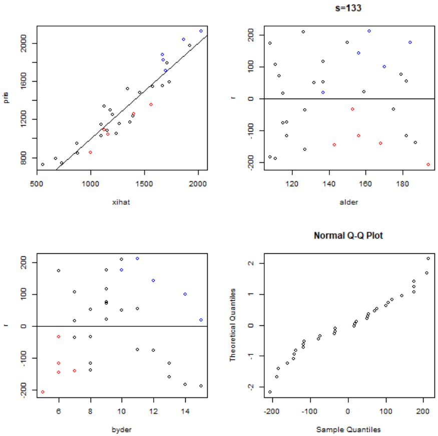
Kørsel i R: lm(pris~alder+byder)

Regel: Variable der ikke er faktorer opfattes i modelformel som regressionsvariable

Residualplot fra pris∼alder+byder

Forventede: $\hat{\xi}_i = \hat{\alpha} + \hat{\beta}_{\text{alder}} \cdot \text{alder}_i + \hat{\beta}_{\text{byder}} \cdot \text{byder}_i,$

Residual: $r_i = \text{pris}_i - \hat{\xi}_i$



Model: $\text{pris}_i \sim N(\alpha + \beta_{\text{alder}} \cdot \text{alder}_i + \beta_{\text{byder}} \cdot \text{byder}_i, \sigma^2)$

R: lmUD=lm(pris~alder+byder) summary(lmUD)

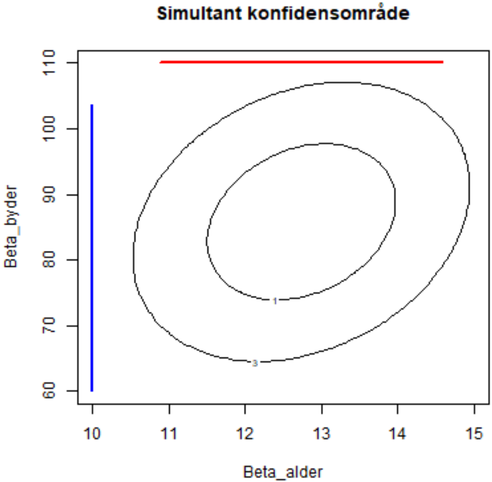
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1336.7221   173.3561  -7.711 1.67e-08 ***
alder          12.7362     0.9024  14.114 1.60e-14 ***
byder          85.8151     8.7058   9.857 9.14e-11 ***
Residual standard error: 133.1 on 29 degrees of freedom
```

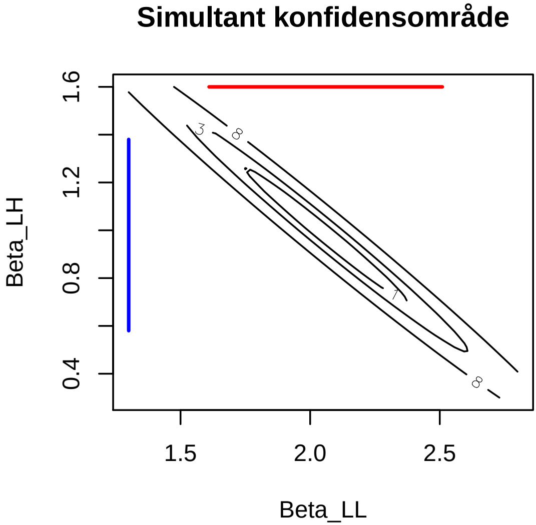
Intercept: $\hat{\alpha}$, alder: $\hat{\beta}_{\text{alder}}$, byder: $\hat{\beta}_{\text{byder}}$

T-test: både alder og byder er vigtige for at beskrive data

```
confint(lmUD)
                  2.5 %     97.5 %
(Intercept) -1691.27514 -982.16896
alder          10.89062   14.58177
byder          68.00986  103.62040
```

Konfidensintervaller for β_{alder} og β_{byder} er brede, men hænger kraftigt sammen




Simultant konfidensområde
1.6
1.2
0.8
0.4
Beta_LH
1.5
2.0
2.5
Beta_LL
8
3
8
1

Generel lineær model: $\boldsymbol{\xi} \in L$, L udspændt af søjler i $\mathbf{H}$

$$\boldsymbol{\xi} = \mathbf{H}\boldsymbol{\theta} \quad \hat{\boldsymbol{\theta}} = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{X}$$

$$\hat{\boldsymbol{\theta}} \sim N_k\big(\boldsymbol{\theta}, \sigma^2(\mathbf{H}^\top\mathbf{H})^{-1}\big)$$

Bevis:

$$\mathrm{E}(\hat{\boldsymbol{\theta}}) = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{\xi} = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{H}\boldsymbol{\theta} = \boldsymbol{\theta}$$

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\big((\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\big)^\top = \sigma^2(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{H}(\mathbf{H}^\top\mathbf{H})^{-1}$$

$$= \sigma^2(\mathbf{H}^\top\mathbf{H})^{-1}$$

Multipel regression med 2 variable:

$$H = (\mathbf{e}, \mathbf{t}_1, \mathbf{t}_2) = \begin{pmatrix} 1 & t_{11} & t_{21} \\ \vdots & \vdots & \vdots \\ 1 & t_{1n} & t_{2n} \end{pmatrix} \quad \boldsymbol{\theta} = (\alpha, \beta_1 \beta_2)^\mathsf{T}$$

$$\mathbf{H}^\mathsf{T}\mathbf{H} = \begin{pmatrix} n & S_1 & S_2 \\ S_1 & SS_{11} & SS_{12} \\ S_2 & SS_{12} & SS_{22} \end{pmatrix} \quad S_u = \sum_i t_{ui}, \quad SS_{uv} = \sum_i t_{ui} t_{vi}$$

$$\left((\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\right)_{22} = \frac{C_{22}}{\mathrm{Det}} = \frac{\mathrm{SSD}_2}{\mathrm{Det}}$$

$$\left((\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\right)_{33} = \frac{C_{33}}{\mathrm{Det}} = \frac{\mathrm{SSD}_1}{\mathrm{Det}}$$

$$\left((\mathbf{H}^\mathsf{T}\mathbf{H})^{-1}\right)_{23} = \frac{-C_{23}}{\mathrm{Det}} = -\frac{\mathrm{SPD}_{12}}{\mathrm{Det}}$$

$$\text{Korrelation}(\hat\beta_1, \hat\beta_2) = -\frac{\mathrm{SPD}_{12}}{\sqrt{\mathrm{SSD}_1 \mathrm{SSD}_2}} = -\frac{\sum_i (t_{1i}-\bar t_1)(t_{2i}-\bar t_2)}{\sqrt{\sum_i (t_{1i}-\bar t_1)^2 \sum_i (t_{2i}-\bar t_2)^2}} = -r_{12}$$

Minus den empiriske korrelation mellem de to forklarende variable

Jo mere de to variable er korrelerede, jo mere er $\hat\beta_1$ og $\hat\beta$ negativt korrelerede

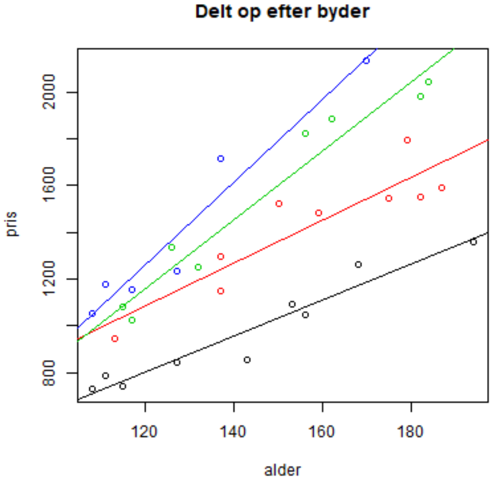
For yderligere forståelse af korrelation: Antag $\bar t_1 = \bar t_2 = 0$ eller $S_1 = S_2 = 0$

$$\text{Det} = n(SS_{11}SS_{22} - SS_{12}^2), \quad \text{Var}(\hat\beta_2) = \frac{\sigma^2 SS_{11}}{SS_{11}SS_{22} - SS_{12}^2} = \frac{\sigma^2/SS_{22}}{1-r_{12}^2}$$

Jo større korrelation, jo større varians på skøn over regressionskoefficient

Hvis $\beta_2 = 0$ har vi svært ved at opdage dette hvis korrelationen mellem de forklarende variable er stor

Røde og blå punkter ligger stadigt på hver sin side af nul i residualplot



Ikke additivitet: interaktion

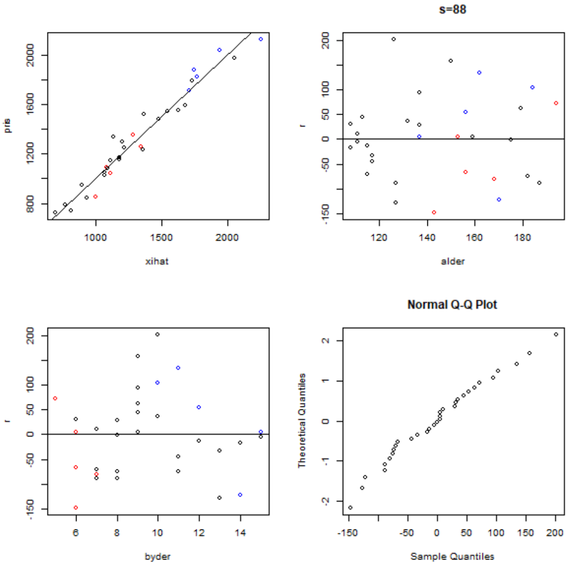
Model: $\text{Pris}_i \sim\sim N(\alpha + \beta_{\text{alder}} \cdot \text{alder}_i + \beta_{\text{byder}} \cdot \text{byder}_i + \beta_{AB}\text{AB}_i, \sigma^2)$

$$\text{AB}_i = \text{alder}_i \cdot \text{byder}_i$$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 322.7544   293.3251   1.100  0.28056
alder         0.8733     2.0197   0.432  0.66877
byder       -93.4099    29.7077  -3.144  0.00392 **
AB            1.2979     0.2110   6.150 1.22e-06 ***
Residual standard error: 88.37 on 28 degrees of freedom
```

Spredningsskøn går fra 133 til 88

Forventede: $\hat{\xi}_i = \hat{\alpha} + \hat{\beta}_{\text{alder}} \cdot \text{alder}_i + \hat{\beta}_{\text{byder}} \cdot \text{byder}_i + \hat{\beta}_{\text{AB}} \cdot \text{AB}_i$



Undersøg linearitet i antal cylindere

```
mpg=mtcars[,1]
vaegt=mtcars[,6]
Design=factor(mtcars[,2])
cyl=mtcars[,2]

anova(lm(mpg~cyl+vaegt),lm(mpg~Design+vaegt))
```

Undersøg behov for interaktion

```
cylvaegt=cyl*vaegt

summary(lm(mpg~cyl+vaegt+cylvaegt))
```

Multipel regression analyseret med lm er vist

Næste: multipel regression med 6 variable

ønsker simpel beskrivelse af data

Patle etal, Geology, Ecology, and Landscapes , 2019

Vandnedsivning er en vigtig parameter for udnyttelse af jord
forudsige hvor meget der skal vandes

Respons: Infiltration Rate (IR) (nedsivning)

Regressionsvariable: Sand, Silt, BD (bulk density), PD (particle density), MC (moisture content), OC (organic carbon)

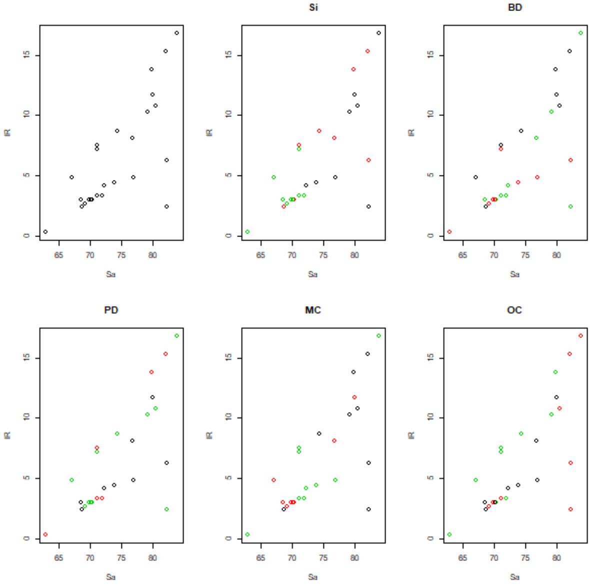
Sa, Si, BD, PD, MC, OC

Backward selection: inkluder alle 6 variable, fjern de “overflødige”

Forward selection: inkluder variable enkeltvis sålænge det giver forbedring

OBS: 6 forklarende variable, men kun $n = 25$ observationer

Figur med data. Farve: opdeling efter variabel



summary(lm(IR∼Sa+Si+BD+PD+MC+OC))

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -58.6841    18.0818  -3.245  0.00449 **
Sa            0.7766     0.1500   5.178 6.33e-05 ***
Si            0.1483     0.2486   0.596  0.55835
BD          -13.8184     5.5728  -2.480  0.02327 *
PD            3.1628     2.5295   1.250  0.22717
MC            0.3375     0.1498   2.253  0.03698 *
OC           31.5894    33.9847   0.930  0.36492
Residual standard error: 2.428 on 18 degrees of freedom
Multiple R-squared:  0.7776,Adjusted R-squared:  0.7035
```

Vi vælger at fjerne Si, som har den højeste p-værdi for hypotesen $\beta_{\text{variabel}} = 0$

summary(lm(IR~Sa+BD+PD+MC+OC))

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -55.0772    16.7489  -3.288  0.00386 **
Sa            0.7097     0.0978   7.256 6.92e-07 ***
BD          -14.2181     5.4378  -2.615  0.01704 *
PD            3.0177     2.4747   1.219  0.23761
MC            0.3688     0.1379   2.674  0.01502 *
OC           42.5031    28.1472   1.510  0.14749
Residual standard error: 2.386 on 19 degrees of freedom
Multiple R-squared:  0.7732,Adjusted R-squared:  0.7136
```

Vi vælger at fjerne PD: p-værdi $= 0.24$

```
summary(lm(IR$\sim$Sa+BD+MC+OC))

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.00517   16.86407  -3.143  0.00512 **
Sa            0.69996    0.09866   7.095 7.07e-07 ***
BD          -14.10503    5.50275  -2.563  0.01854 *
MC            0.36462    0.13955   2.613  0.01666 *
OC           60.96554   24.01541   2.539  0.01955 *
Residual standard error: 2.415 on 20 degrees of freedom
Multiple R-squared:  0.7555,Adjusted R-squared:  0.7066
```

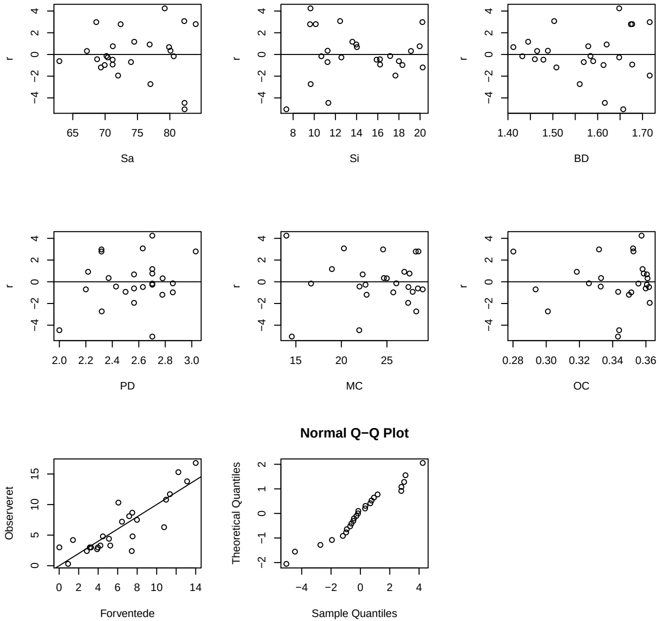
Vi fjerner ikke flere led: alle p-værdier er under 0.02

Slutmodel $E(\mathrm{IR}) = \alpha + \beta_{\mathrm{Sa}} \cdot \mathrm{Sa} + +\beta_{\mathrm{BD}} \cdot \mathrm{BD} + \beta_{\mathrm{MC}} \cdot \mathrm{MC} + \beta_{\mathrm{OC}} \cdot \mathrm{OC}$

Konfidensinterval for spredning σ: [1.85, 3.49]

Test fra startmodel til slutmodel: anova(lm(),lm()), p-værdi = 0.4256

Modelkontrol. Modelformel: IR∼Sa+BD+MC+OC



Undersøg mutipel regressionsmodel (mpg,cyl,disp,hp,drat,wt,qsec)

```
x=mtcars[,1]
t2=mtcars[,2]
t3=mtcars[,3]
t4=mtcars[,4]
t5=mtcars[,5]
t6=mtcars[,6]
t7=mtcars[,7]

summary(lm(x~t2+t3+t4+t5+t6+t7))
```

Reducer model ved backward selektion

Hvor god er modellen ?
Kan vi forudsige middel-nedsivningsraten for et nyt datasæt med værdierne: $(\tilde{Sa}, \tilde{BD}, \tilde{MC}, \tilde{OC})$?

Simpel lineær regression: "linjens værdi i et punkt"

Multipel regression: prædikteret værdi:
$\hat{\xi}^P = -53.005 + 0.700 \cdot \tilde{Sa} - 14.105 \cdot \tilde{BD} + 0.365 \cdot \tilde{MC} + 60.966 \cdot \tilde{OC}$

Konfidensinterval for ξ^P

Prediktionsinterval for ny måling $\xi^P + \text{støj}$

Beregning i R: eksempel

```
NyData=data.frame(Sa=70,BD=1.65,MC=25,OC=0.34)

lmUD=lm(IR~Sa+BD+MC+OC)

predict(lmUD,NyData,interval="confidence")
```

giver interval: $[1.0, 4.1]$

```
predict(lmUD,NyData,interval="prediction")
```

giver interval: $[-2.7, 7.8]$

Nedsivning: interaktionsled?

SaOC=Sa·OC

```
lm(formula = IR ~ Sa + BD + MC + OC + SaOC)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   387.0105   170.6021   2.268  0.03514 *
Sa             -5.3108     2.3232  -2.286  0.03391 *
BD            -15.1738     4.8715  -3.115  0.00570 **
MC              0.4644     0.1290   3.600  0.00191 **
OC          -1207.0006   490.1990  -2.462  0.02354 *
SaOC           17.2991     6.6816   2.589  0.01800 *
Residual standard error: 2.13 on 19 degrees of freedom
```

Spredningsskøn: fulde model: 2.43, Backward model: 2.42, Her: 2.13

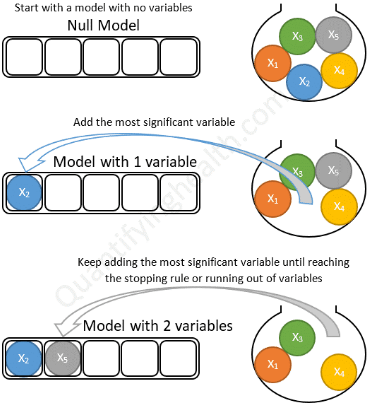
MEN: jeg har ledt blandt alle interaktioner!

Backward selection er vist

Prædikteret værdi for nye data er omtalt

Næste: forward selection

Forward stepwise selection example with 5 variables:



Laver regression af IR på en enkelt variabel. P-værdi: test for $\beta = 0$

lm(IR∼Sa),lm(IR∼Si),..,lm(IR∼OC)

Variabel	Sa	Si	BD	PD	MC	OC
p-værdi	0.000	0.043	0.110	0.396	0.292	0.294
$s(M)$	3.030	4.157	4.304	4.482	4.444	4.445

Vi vælger at inkludere variablen Sa

Laver regression af IR på en enkelt variabel udover Sa

P-værdi: test for $\beta_{\text{variabel}} = 0$

lm(IR∼Sa+Si),lm(IR∼Sa+BD),..,lm(IR∼Sa+OC)

Variabel	Sa	Si	BD	PD	MC	OC
p-værdi	-	0.056	0.038	0.083	0.392	0.081
$s(M)$	-	2.845	2.801	2.888	3.045	2.887

Vi vælger at inkludere variablen BD udover Sa

Laver regression af IR på en enkelt variabel udover Sa og BD

lm(IR∼Sa+BD+Si),lm(IR∼Sa+BD+PD),lm(IR∼Sa+BD+MC),lm(IR∼Sa+BI

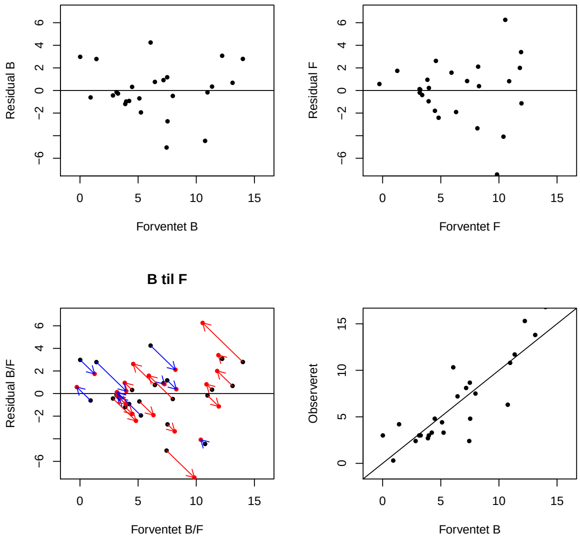
Variabel	Sa	Si	BD	PD	MC	OC
p-værdi	-	0.084	-	0.104	0.128	0.155
$s(M)$	-	2.666	-	2.689	2.710	2.730

Vi inkluderer ikke flere variable

Slutmodel Forward: Sa + BD $s(M_F) = 2.80$

Slutmodel Backward: Sa + BD + MC + OC $s(M_B) = 2.42$

Sammenligne Backward og forward



En del variation omkring middelværdimodellen, $s(M_F)/\mathrm{sd}(\mathrm{IR}) \approx 0.63$

Forward og backward selection giver forskellige modeller

Her: slutmodel ved backward selection er bedst
men backward bruger to variable mere end forward

Backward:

```
Residual standard error: 2.415 on 20 degrees of freedom
Multiple R-squared:  0.7555,Adjusted R-squared:  0.7066
```

$$s(M_B)/\mathrm{mean}(\mathrm{IR}) = 0.37$$

Forward:

```
Residual standard error: 2.801 on 22 degrees of freedom
Multiple R-squared:  0.6381,Adjusted R-squared:  0.6052
```

$$s(M_F)/\mathrm{mean}(\mathrm{IR}) = 0.43$$

Stor spredning: svært at prediktere nedsivning

Undersøg mutipel regressionsmodel (mpg,cyl,disp,hp,drat,wt,qsec)

```
x=mtcars[,1]
t2=mtcars[,2]
t3=mtcars[,3]
t4=mtcars[,4]
t5=mtcars[,5]
t6=mtcars[,6]
t7=mtcars[,7]
```

Vælge simpel model ved backward selection eller forward selection er omtalt

Næste: Illustrere backward algortime og korrelation ved simulering

Simulere multipel regressionsdata og se hvor ofte vi finder de rigtige variable

Eksempel:

Simulere $X_i \sim N(\alpha + \beta_1 t_{i1} + \beta_2 t_{i2} + 0 \cdot t_{i3} + 0 \cdot t_{i4} + 0 \cdot t_{i5}, \sigma^2)$

hvor ofte finder vi variabel 1 og 2 ?

hvad er betydningen af korrelation ?

hvor god er den model vi finder ?

Backward algoritme

```
backward=function(T,x){
 pvallim=0.05
 k=dim(T)[2]
 med=c(1:k)
 goON=TRUE
 while (goON){
  lmUD=lsfit(T[,med],x)
  sumUD=ls.print(lmUD,print.it=FALSE)$coef.table
  pval=sumUD[[1]][,4]
  pval=pval[-1]
  if (max(pval)>pvallim){
   r=which.max(pval)
   med=med[-r]
   if (length(med)==0){goON=FALSE}
   } else {
   goON=FALSE
   }
  }
```

```
 if (length(med)==0){
  ahat=as.numeric(lm(x~1)$coef)
  bhat=rep(0,nvar)
  } else {
  be=as.numeric(lsfit(T[,med],x)$coef)
  ahat=be[1]
  bhat=rep(0,nvar)
  bhat[med]=be[-1]
  }
 return(list(med=med,ahat=ahat,bhat=bhat))
}


sdpred=function(bUD,beta,sig){
 return(sqrt(c(sig^2+bUD$ahat^2+
 rbind(beta-bUD$bhat)%*%t(B)%*%B%*%cbind(beta-bUD$bhat))))
}
```

```
nsim=100
res=matrix(0,nsim,2)

for (simnr in 1:nsim){
n=25
nvar=5
sig=1
z=0
beta=c(1,1,0,0,0)
B=diag(rep(1,nvar))
B[1,3]=z
B[1,4]=z

T0=matrix(rnorm(n*nvar),n,nvar)
T=T0%*%B
x0=sig*rnorm(n)
x=c(T%*%cbind(beta))+x0
```

```
bUD=backward(T,x)
med0=rep(0,nvar)
med0[bUD$med]=1
res[simnr,]=c(sum(med0[1:2])*4+sum(med0[3:5]),
sdpred(bUD,beta,sig))
}

par(mfrow=c(1,2))
plot(res[,1],res[,2],xlim=c(0,11),ylim=c(1,1.7),xlab="Id",ylab="sdpred")
hist(res[,1],breaks=c(0:12)-0.5,xlab="Id",main="")
```

Stor korrelation: vælger "ofte" forkert variabel

Stor korrelation ødelægger ikke prædiktionsevnen

prædiktionsevnen: root mean squared error baseret på $\hat{\alpha}$ og $\hat{\beta}$

og på nye data med samme fordeling (både t og x)

Simuleringseksperiment er omtalt

Næste: advarsel mod overfitting

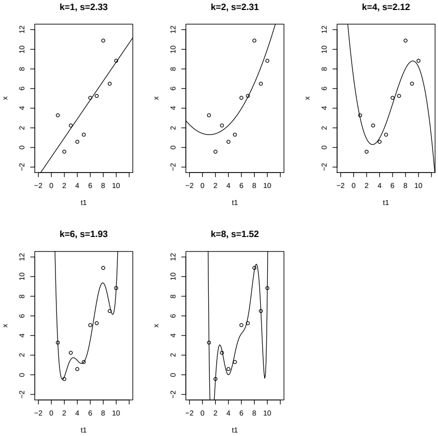
Sande model: $E(X_i) = \alpha + \beta_1 t_i$

Fitter model: $E(X_i) = \alpha + \beta_1 t_i + \beta_2 t_i^2 + \cdots + \beta_k t_i^k$

Typisk: jo større k er jo bedre et fit får vi: $s(M_k)$ er lille

hvis $k = n$ er $s(M_k) = 0$!

Overfitting: eksempel



Overfitting giver dårlig prediktor:

Generel formulering:

Data: $x_i, t_{1i}, t_{2i}, \ldots, t_{ki}, \quad i = 1, \ldots, n$

Model: $E(X_i) = \alpha + \beta_1 t_{1i} + \beta_2 t_{2i} + \cdots + \beta_k t_{ki}$

Estimater: $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k, \; s(M)$

Forudsige middelrespons for nye værdier $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_k$:

$$\hat{\xi}^P = \hat{\alpha} + \hat{\beta}_1 \tilde{t}_1 + \cdots + \hat{\beta}_k \tilde{t}_k$$

Beregning i R: benyt predict

Overfitting: problem

Hvis vi overfitter giver dette typisk en dårligere prediktor:

$$E\left\{\left(X_{\text{ny}} - \text{Prediktor}(t_{\text{ny}})\right)^2\right\} \text{ bliver større}$$

k=1	k=2	k=4	k=6	k=8
2.3	3.2	4.2	5.9	14.4

For viste data, t_{ny} uniform

To problematikker:

Hvordan vælger vi en simpel model ?

Backward/forward - andre metoder
Backward: x$\sim$t4, Forward: x$\sim$t2

Hvordan måler vi overfitting ?

Senere forelæsning: crossvalidation

(Næste forelæsning: multinomialmodellen)

Slut for i dag