## Ekstraoppgave 4

### a

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$(X'X) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \sum_i x_{i1} & \sum_i x_{i2} & \dots & \sum_i x_{in} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \dots & \sum_i x_{i1}x_{in} \\ \sum_i x_{i2} & \sum_i x_{i2}x_{i1} & \sum_i x_{i2}^2 & \dots & \sum_i x_{i2}x_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{n1} & \sum_i x_{in} \sum_i x_{i1} & \sum_i x_{in}x_{i2} & \dots & \sum_i x_{in}^2 \end{pmatrix}$$

$$= \begin{pmatrix} n & 0 & \dots & 0 \\ 0 & \sum_i x_{i1}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sum_i x_{in}^2 \end{pmatrix}$$

$$(X'y) = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$= \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1}y_i \\ \vdots \\ \sum_i x_{ik}y_i \end{pmatrix}$$

$$(X'X)^{-1}X'y = \begin{pmatrix} \sum_i y_i/n \\ \sum_i x_{i1}y_i/\sum_i x_{i1}^2 \\ \vdots \\ \sum_i x_{ik}y_i/\sum_i x_{ik}^2 \end{pmatrix}$$

**b**

$Y_i$ are normal distributed, and $\hat{\beta}_j$ are linear combinations of those, so the $\hat{\beta}_j$ are normally distributed.

$$\begin{aligned}
E(\hat{\beta}_j) &= \sum_i x_{ij} E(Y_i) / \sum_i x_{ij}^2 \\
&= \sum_i x_{ij}(\beta_0 + \sum_t \beta_t x_{it}) / \sum_i x_{ij}^2 \\
&= \left[ \sum_i x_{ij}\beta_0 + \sum_{t \neq j} \beta_t x_{ij} x_{it} + \beta_j \sum_i x_{ij}^2 \right] / \sum_i x_{ij}^2 \\
&= \beta_j \\
V(\hat{\beta}_j) &= \sum_i x_{ij}^2 V(Y_i) / (\sum_i x_{ij}^2)^2 \\
&= \sum_i x_{ij}^2 \sigma^2 / (\sum_i x_{ij}^2)^2 \\
&= \sigma^2 / \sum_i x_{ij}^2
\end{aligned}$$

$Cov(\hat{\beta}_j, \hat{\beta}_l) = 0$ by the covariance matrix TODO

**c**

$$\begin{aligned}
SSR &= \sum_i (\hat{Y}_i - \bar{Y})^2 \\
&= \sum_i (\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij} - \bar{Y})^2 \\
&= \sum_i (\bar{Y} + \sum_j \hat{\beta}_j x_{ij} - \bar{Y})^2 \\
&= \sum_i (\sum_j \hat{\beta}_j x_{ij})^2 \\
&= \sum_i \left[ \sum_j \hat{\beta}_j^2 x_{ij}^2 + \sum_j \sum_{t \neq j} \hat{\beta}_j \hat{\beta}_t x_{ij} x_{it} \right] \\
&= \sum_i \sum_j \hat{\beta}_j^2 x_{ij}^2 \\
&= \sum_j \hat{\beta}_j^2 \sum_i x_{ij}^2
\end{aligned}$$

**d**

$$
\begin{aligned}
R^2 &= \frac{SSR}{SST} \\
&= \frac{\sum_j \hat{\beta}_j^2 \sum_i x_{ij}^2}{SST} \\
&= \sum_j \frac{(\sum_i x_{ij}y_i / \sum_i x_{ij}^2)^2 \sum_i x_{ij}^2}{SST} \\
&= \sum_j \frac{(\sum_i x_{ij}y_i)^2}{\sum_i x_{ij}^2 SST} \\
&= \sum_j \frac{\sum_i \left[(x_{ij}y_i)^2 + \sum_{t \neq i} x_{ij}x_{tj}y_iy_t\right]}{\sum_i x_{ij}^2 SST} \\
&= \sum_j \frac{\sum_i (x_{ij}y_i)^2}{\sum_i x_{ij}^2 \sum_i (Y_i - \bar{Y})^2} \\
&= \sum_j r_j^2
\end{aligned}
$$

## Eksamen 2005, 2

**a**

$Y_i = \beta_0 + \beta_1 kvm + \beta_2 takst + \epsilon_i$.

**b**

$$
\begin{aligned}
\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 kvm + \hat{\beta}_2 takst \\
&= 0.72634 + -0.01828 * 105 + 1.09548 * 23 \\
&= 24.00298
\end{aligned}
$$

$$
\begin{aligned}
&V(Y - (\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^*)) \\
=&V(Y) + V(\hat{\beta}_0) + x_1^{*2}V(\hat{\beta}_1) + x_2^{*2}V(\hat{\beta}_2) - 2(x_1^* Cov(\hat{\beta}_0, \hat{\beta}_1) + x_2^* Cov(\hat{\beta}_0, \hat{\beta}_2) + x_1^* x_2^* Cov(\hat{\beta}_1, \hat{\beta}_2)) \\
=&\sigma^2 + V(\hat{\beta}_0) + x_1^{*2}V(\hat{\beta}_1) + x_2^{*2}V(\hat{\beta}_2) - 2(x_1^* Cov(\hat{\beta}_0, \hat{\beta}_1) + x_2^* Cov(\hat{\beta}_0, \hat{\beta}_2) + x_1^* x_2^* Cov(\hat{\beta}_1, \hat{\beta}_2))
\end{aligned}
$$

$$
\begin{aligned}
&\hat{y} \pm t_{\alpha/2, n-k-1} \sqrt{s^2 + s_{\hat{y}}^2} \\
=&\hat{y} \pm t_{\alpha/2, 32} \sqrt{1.562^2 + 0.0821} \\
=&\hat{y} \pm 2.037 \cdot 1.588 \\
=&\hat{y} \pm 3.235
\end{aligned}
$$

**c**

Confounding effects

### Ekstraoppgave 5

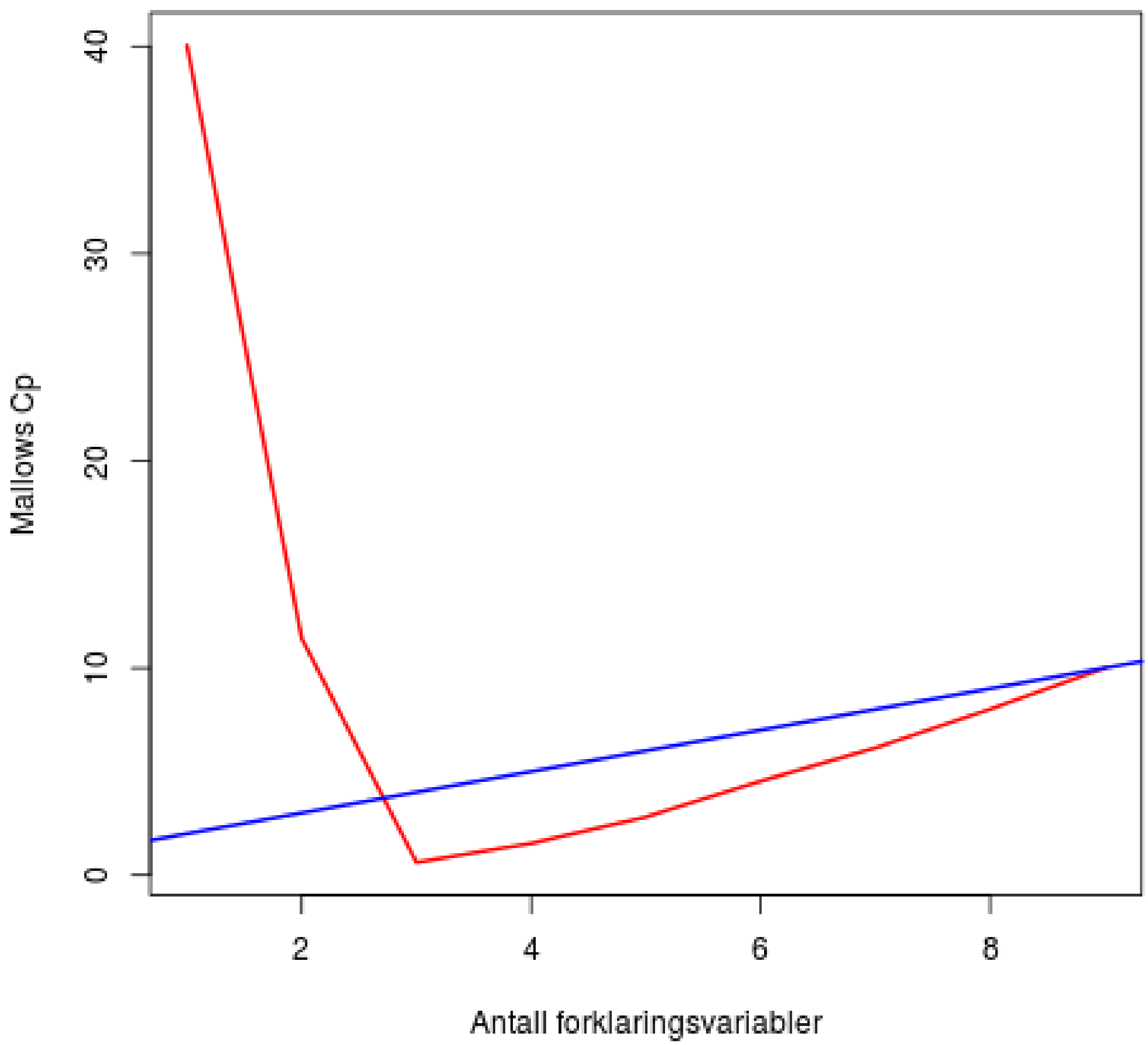
```
boligpris= read.table("http://www.uio.no/studier/emner/matnat/math/STK2120/v13/boligpris.tx

fit.lm = lm(Pris~kvm+Takst,data=boligpris)

summary(fit.lm)

#Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
#(Intercept)   0.72634    0.65168   1.115    0.273
#kvm          -0.01828    0.01278  -1.430    0.162
#Takst         1.09548    0.05506  19.895   <2e-16 ***
#---
#
```

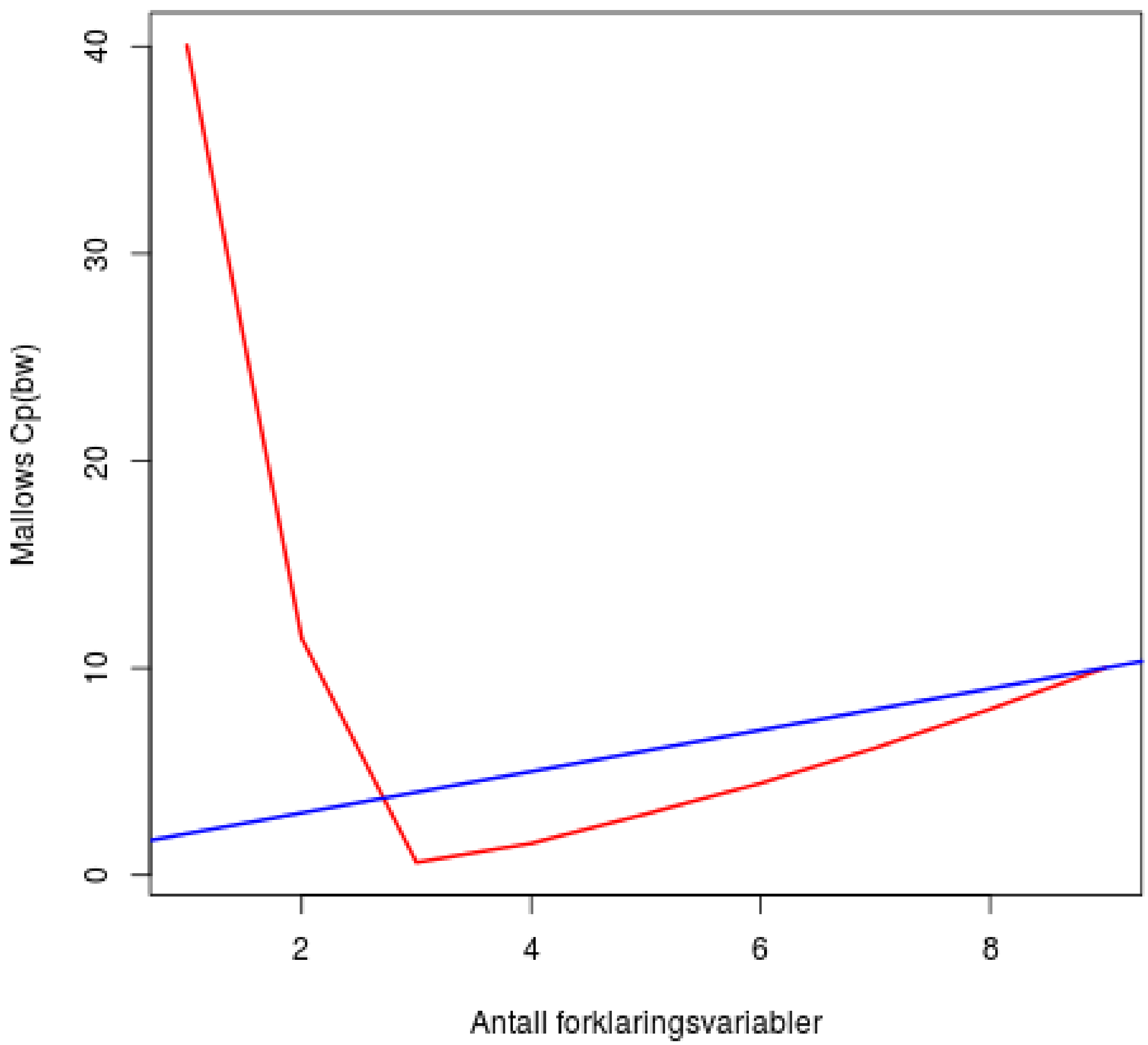
### Ekstraoppgave 6

```
library(leaps)

bjorner=read.table(
    "http://www.uio.no/studier/emner/matnat/math/STK2120/v16/logbjorner.txt",
    header=T)

#a
plot(bjorner)

#b
fit.forward=regsubsets(logvekt~.,data=bjorner,nvmax=10,method="forward")
summary.forward = summary(fit.forward)

#Selection Algorithm: forward
#          alder kjonn vaar sommer loghodelengde loghodebredde loghals loglengde
#1  ( 1 )  " "   " "   " "  " "    " "           " "           " "     " "
#2  ( 1 )  " "   " "   " "  " "    " "           " "           " "     "*"
#3  ( 1 )  " "   " "   " "  " "    " "           " "           "*"     "*"
#4  ( 1 )  "*"   " "   " "  " "    " "           " "           "*"     "*"
#5  ( 1 )  "*"   " "   " "  " "    " "           "*"           "*"     "*"
#6  ( 1 )  "*"   " "   "*"  " "    " "           "*"           "*"     "*"
#7  ( 1 )  "*"   " "   "*"  "*"    " "           "*"           "*"     "*"
#8  ( 1 )  "*"   " "   "*"  "*"    "*"           "*"           "*"     "*"
#9  ( 1 )  "*"   "*"   "*"  "*"    "*"           "*"           "*"     "*"
#          logbryst
#1  ( 1 )  "*"
#2  ( 1 )  "*"
#3  ( 1 )  "*"
#4  ( 1 )  "*"
#5  ( 1 )  "*"
#6  ( 1 )  "*"
#7  ( 1 )  "*"
#8  ( 1 )  "*"
#9  ( 1 )  "*"

png("mallows.png")
plot(1:9,summary.forward$cp,
```

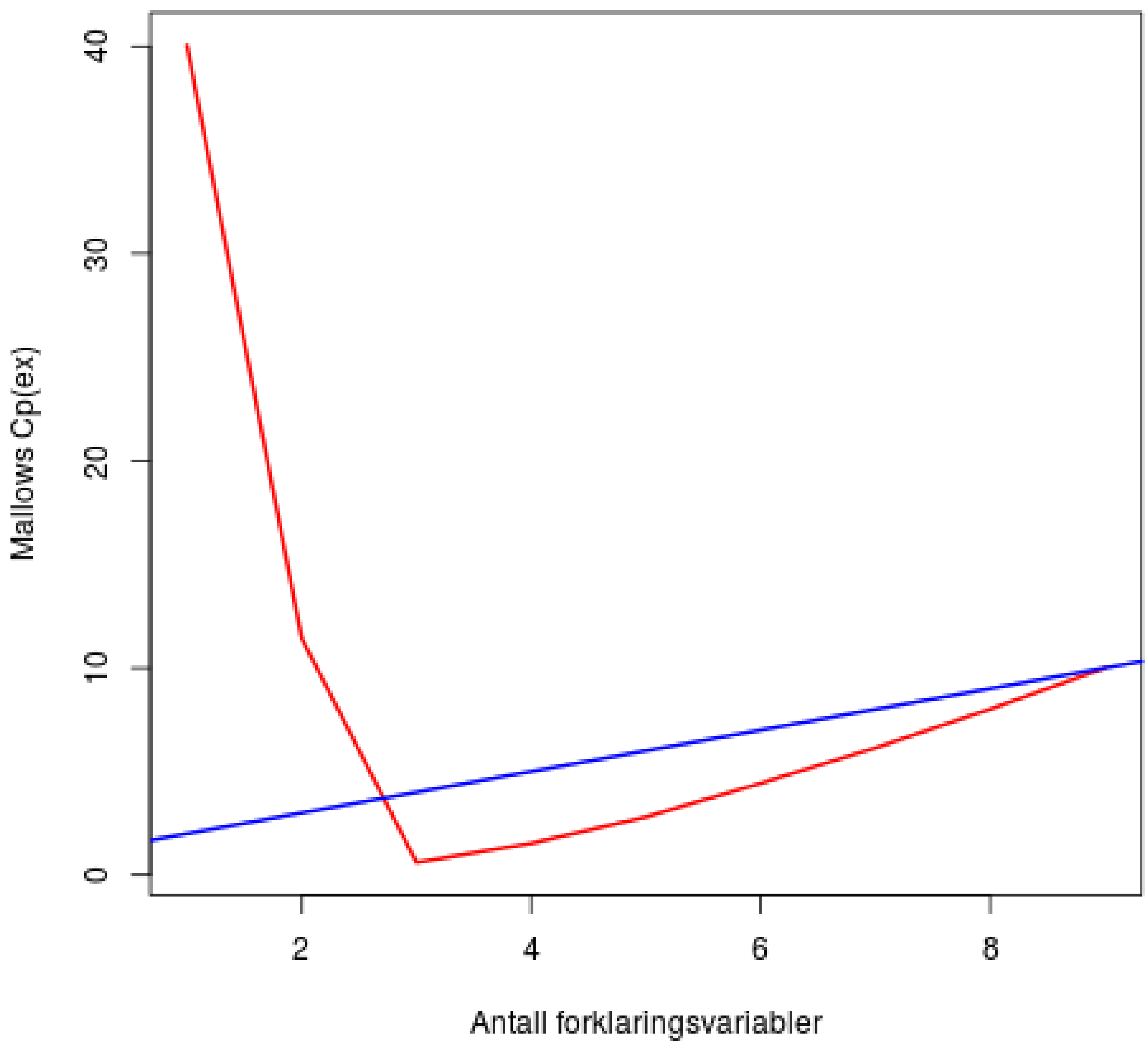
```
      xlab="Antall forklaringsvariabler", ylab="Mallows Cp",
      type='l',col='red',lwd=2)
abline(1,1,lwd=2,col='blue')
dev.off()

# Antall forklaringsvariable = 3 (loghals, loglengde logbryst)

#c
fit.backward=regsubsets(logvekt~.,data=bjorner,nvmax=10,method="backward")
summary.backward = summary(fit.backward)

#           alder kjonn vaar sommer loghodelengde loghodebredde loghals loglengde
#1  ( 1 )  " "   " "   " "  " "    " "           " "           " "     " "
#2  ( 1 )  " "   " "   " "  " "    " "           " "           " "     "*"
#3  ( 1 )  " "   " "   " "  " "    " "           " "           "*"     "*"
#4  ( 1 )  "*"   " "   " "  " "    " "           " "           "*"     "*"
#5  ( 1 )  "*"   " "   " "  "*"    " "           " "           "*"     "*"
#6  ( 1 )  "*"   " "   "*"  "*"    " "           " "           "*"     "*"
#7  ( 1 )  "*"   " "   "*"  "*"    " "           "*"           "*"     "*"
#8  ( 1 )  "*"   " "   "*"  "*"    "*"           "*"           "*"     "*"
#9  ( 1 )  "*"   "*"   "*"  "*"    "*"           "*"           "*"     "*"
#           logbryst
#1  ( 1 )  "*"
#2  ( 1 )  "*"
#3  ( 1 )  "*"
#4  ( 1 )  "*"
#5  ( 1 )  "*"
#6  ( 1 )  "*"
#7  ( 1 )  "*"
#8  ( 1 )  "*"
#9  ( 1 )  "*"
#

png("mallowsBw.png")
plot(1:9, summary.backward$cp,
      xlab="Antall forklaringsvariabler", ylab="Mallows Cp(bw)",
      type='l',col='red',lwd=2)
abline(1,1,lwd=2,col='blue')
dev.off()
# Antall forklaringsvariable = 3 (loghals, loglengde logbryst)

fit.exhaustive=regsubsets(logvekt~.,data=bjorner,nvmax=10,method="exhaustive")
summary.exhaustive=summary(fit.exhaustive)
#Selection Algorithm: exhaustive
#           alder kjonn vaar sommer loghodelengde loghodebredde loghals loglengde
#1  ( 1 )  " "   " "   " "  " "    " "           " "           " "     " "
#2  ( 1 )  " "   " "   " "  " "    " "           " "           " "     "*"
#3  ( 1 )  " "   " "   " "  " "    " "           " "           "*"     "*"
#4  ( 1 )  "*"   " "   " "  " "    " "           " "           "*"     "*"
#5  ( 1 )  "*"   " "   " "  " "    " "           "*"           "*"     "*"
#6  ( 1 )  "*"   " "   "*"  "*"    " "           " "           "*"     "*"
#7  ( 1 )  "*"   " "   "*"  "*"    " "           "*"           "*"     "*"
#8  ( 1 )  "*"   " "   "*"  "*"    "*"           "*"           "*"     "*"
#9  ( 1 )  "*"   "*"   "*"  "*"    "*"           "*"           "*"     "*"
```

```
#             logbryst
#1  ( 1 )  "*"
#2  ( 1 )  "*"
#3  ( 1 )  "*"
#4  ( 1 )  "*"
#5  ( 1 )  "*"
#6  ( 1 )  "*"
#7  ( 1 )  "*"
#8  ( 1 )  "*"
#9  ( 1 )  "*"
png("mallowsExh.png")
plot(1:9, summary.exhaustive$cp,
     xlab="Antall forklaringsvariabler", ylab="Mallows Cp(ex)",
     type='l',col='red',lwd=2)
abline(1,1,lwd=2,col='blue')
dev.off()
# Antall forklaringsvariable = 3 (loghals, loglengde logbryst)
# All the methods give the same model

#d
#We get the same model with each method. This indicates that the model with loglengde,
#loghals, logbryst is the right one
```


Mallows Cp
0
10
20
30
40
2
4
6
8
Antall forklaringsvariabler



Mallows Cp(bw)
40
30
20
10
0
2
4
6
8
Antall forklaringsvariabler


Mallows Cp(ex)
Antall forklaringsvariabler
0
10
20
30
40
2
4
6
8