Ekstraoppgave 11

I en studie har en undersøkt hvor mye varme det blir utviklet ved herding av sement og hvordan dette avhenger av sementens sammensetning. Sammensetningen av sementen ble målt med variablene:

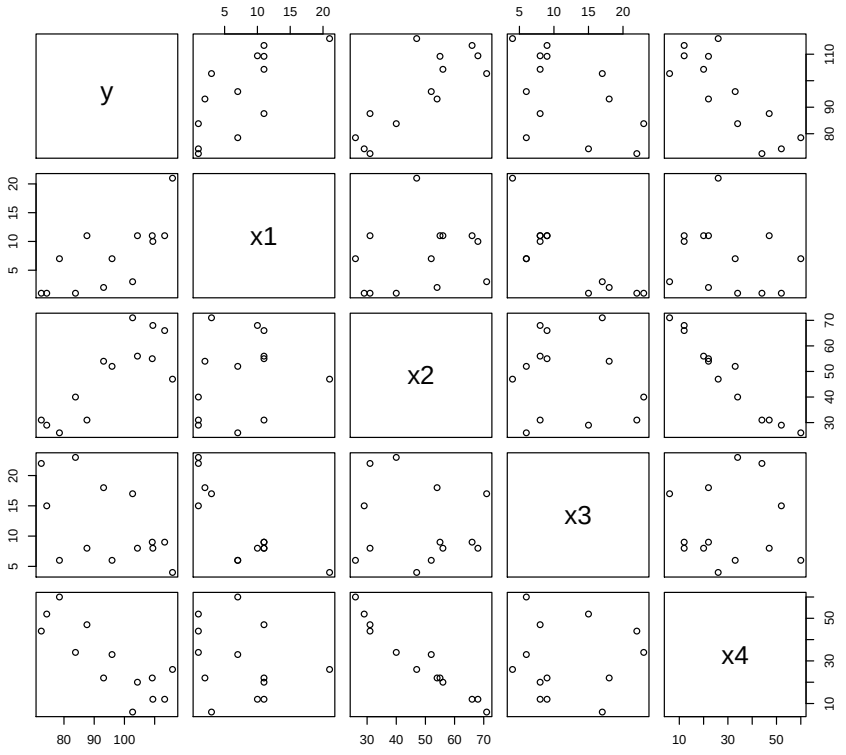
- x_1 = prosent trikalsium aluminat ($3CaO \cdot Al_2O_3$)
- x_2 = prosent trikalsium silikat ($3CaO \cdot SiO_2$)
- x_3 = prosent tetrakalsium alumino ferrit ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$)
- x_4 = prosent dikalsium silikat ($2CaO \cdot SiO_2$)

Her er prosentene regnet relativt til vekten av råmaterialet for sementproduksjonen.

Responsvariablen er:

- y = varmeutvikling i kalorier pr. gram under herdingsprosessen.

Resultatet av forsøket er vist i matriseplottet nedenfor:



Nedenfor er det gitt resultatet av en lineær regresjonanalyse med varmeutvikling (y) som responsvariabel og prosent trikalsium aluminat (x_1) og prosent dikalsium silikat (x_4) som forklaringsvariabler:

```
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   103.09738      2.12398     48.54   3.32e-13
x1              1.43996      0.13842     10.40   1.11e-06
x4             -0.61395      0.04864    -12.62   1.81e-07

Residual standard error: 2.734 on 10 degrees of freedom
Multiple R-squared:  0.9725

(redigert utskrift)
```

a) Sett opp den statistiske modellen som analysen ovenfor bygger på. Forklar hva de ulike delene av utskriften står for.

Vi gjør så en regresjonsanlyse der vi også tar med prosent trikalsium silikat (x_2) som forklaringsvariabel:

```
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)    71.6483      14.1424     5.066   0.000675
x1              1.4519       0.1170    12.410   5.78e-07
x2              0.4161       0.1856     2.242   0.051687
x4             -0.2365       0.1733    -1.365   0.205395

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823

(redigert utskrift)
```

b) I modellen som også har med prosent trikalsium silikat (x_2) som forklaringsvariabel, er effkten av prosent dikalsium silikat (x_4) ikke lenger signifikant. Hvordan kan du forklare dette resultatet?

Vi vil så prøve å finne en modell som best mulig beskriver varmeutviklingen (y) som funksjon av sammensetningen av sementen. Vi gjør det ved å tilpasse regresjonsmodeller der forklaringsvariablene velges blant x_1, x_2, x_3 og x_4 og de seks interaksjonene mellom disse variablene (dvs. forklaringsvariabler av formen $x_j \cdot x_\ell$ for $j \neq \ell$).

For hver $m = 1, 2, \ldots, 10$ tilpasser vi alle mulige regresjonsmodeller med m forklaringsvariabler og velger den av dem som har minst residualkvadratsum (SSE).

Vi får da følgende sekvens av modeller:

```
> fit.exhaustive=regsubsets(y~x1+x2+x3+x4+x1:x2+x1:x3+x1:x4+x2:x3+x2:x4+x3:x4,
                            data=sement,nvmax=10,method="exhaustive")
> summary.exhaustive=summary(fit.exhaustive)
> summary.exhaustive

.... (redigert output) ....

Selection Algorithm: exhaustive
          x1  x2  x3  x4  x1:x2 x1:x3 x1:x4 x2:x3 x2:x4 x3:x4
1  ( 1 )  " " " " " " " " "*"   " "   " "   " "   " "   " "
2  ( 1 )  "*" "*" " " " " " "   " "   " "   " "   " "   " "
3  ( 1 )  " " " " " " " " "*"   " "   " "   "*"   " "   "*"
4  ( 1 )  " " " " "*" " " "*"   " "   " "   "*"   "*"   " "
5  ( 1 )  " " " " "*" " " "*"   " "   " "   "*"   "*"   "*"
6  ( 1 )  " " " " "*" " " "*"   "*"   " "   "*"   "*"   "*"
7  ( 1 )  " " " " "*" "*" "*"   "*"   " "   "*"   "*"   "*"
8  ( 1 )  " " "*" "*" "*" "*"   "*"   " "   "*"   "*"   "*"
9  ( 1 )  "*" "*" "*" "*" " "   "*"   "*"   "*"   "*"   "*"
10 ( 1 )  "*" "*" "*" "*" "*"   "*"   "*"   "*"   "*"   "*"
```

For disse ti modellene får vi følgende verdier av Mallows C_p:

```
> summary.exhaustive$cp
  654.87  59.91  47.40  13.35  12.31   6.38   5.61   7.40   9.00  11.00
```

c) Bruk resultatene ovenfor til å avgjøre hvilke forklaringsvariabler du bør ha med i regresjonsmodellen. Begrunn svaret ditt.