Oppgaver, STK4040, uke 45

6. november 2005

Oppgave 1

Vis at kryssvalidert MSE for en modell med bare konstantledd kan skrives som $\text{Var}(y)n/(n-1)$ når $\text{Var}(y) = \sum_{i=1}^{n}(y-\bar{y})^2/(n-1)$.

Oppgave 2: Prinsipalkomponentregresjon

Filen `datasett_45.txt` inneholder et datasett med målinger gjort på 79 prøver av hele hvetekorn. Første kolonne inneholder måling av vanninnhold (i prosent), og de resterende kolonnene (2 – 101) er målinger av nær-infrarød (NIR) reflektans. Dataene kan leses inn slik:

```
tmp <- read.table("datasett_45.txt")
vann <- tmp[,1]
NIR <- as.matrix(tmp[,2:101])
```

a)

Plott og studer dataene. Er det noen spesielle strukturer, eller ekstreme observasjoner som kanskje bør fjernes?

b)

Gjør en prinsipalkomponentregresjon (PCR) med ti komponenter av vann mot NIR: Bruk `prcomp` og beregn regresjonskoeffisienter, predikerte verdier og residualer direkte.

Plott og vurder predikert mot målt respons, og residualer. Plott de første skårene og ladningene og se om du kan tolke dem. Se også etter ekstreme observasjoner og vurder hva som bør gjøres med dem.

b)

Plott av predikert mot målt respons i b) tyder på at 10 komponenter er nok til å få til en god prediksjon av vann fra NIR. Men trenger vi 10 komponenter, eller holder det med færre?

Bruk full kryssvalidering til å estimere MSE for modeller med 0, 1, ..., 10 komponenter.

Plott de estimerte MSE-verdiene mot antall komponenter, og vurder hvor mange komponenter bør være med i modellen. Hvor stor er den estimerte MSE for det valgte antall komponenter? Hvilke komponenter ser ut til å ha mest å si for å forklare responsen?