Oppgaver, STK4040, uke 41

9. oktober 2005

Oppgave 1

Gitt den lineære modellen $\boldsymbol{y} = \mathrm{X}\beta + \varepsilon$, hvor $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathrm{I}_n)$. Vi antar at $n \times (p+1)$-matrisen X har full rang $p + 1 < n$, og at første kolonne i X er $\mathbf{1}_n$.

Vi har at minste kvadraters-estimatet av β er $\hat{\beta} = (\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t\boldsymbol{y}$. De predikerte (også kalt tilpassede («fitted»)) responsverdiene er $\hat{\boldsymbol{y}} = \mathrm{X}\hat{\beta}$. Definer $\mathrm{H} = \mathrm{X}(\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t$. Da kan de predikerte responsverdiene skrives som $\hat{\boldsymbol{y}} = \mathrm{H}\boldsymbol{y}$. (H kalles ofte «hatt-matrisen», fordi den transformerer $\boldsymbol{y}$ til $\hat{\boldsymbol{y}}$.) Residualene er gitt ved $\hat{\varepsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathrm{I} - \mathrm{H})\boldsymbol{y}$. Definer $\mathrm{M} = \mathrm{I} - \mathrm{H}$. Da er $\hat{\varepsilon} = \mathrm{M}\boldsymbol{y}$.

Vis følgende påstander:

1. H er idempotent (det vil si $\mathrm{HH} = \mathrm{H}$) og symmetrisk.
2. M er idempotent og symmetrisk.
3. $\mathrm{MX} = \mathbf{0}$, og $\mathrm{M}\mathbf{1}_n = \mathbf{0}$.
4. $\hat{\varepsilon} = \mathrm{M}\varepsilon$, $\mathrm{X}^t\hat{\varepsilon} = \mathbf{0}$ og $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.
5. $\hat{\boldsymbol{y}}^t\hat{\varepsilon} = 0$.
6. $\mathbf{1}_n^t\hat{\varepsilon} = 0$.

Oppgave 2

Gitt den lineære modellen $\boldsymbol{y} = \mathrm{X}_1\beta + \mathrm{X}_2\varphi + \varepsilon$, hvor $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathrm{I}_n)$, og $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2)$ er av full rang.

Vis at $\hat{\beta}$ kan beregnes ved først å gjøre en regresjon av $\boldsymbol{y}$ og X_1 på X_2 og deretter gjøre en regresjon av residualene til $\boldsymbol{y}$ på residualene til X_1.

Oppgave 3

Filen `datasett_41.txt` inneholder sensoriske data fra et forsøk hvor 105 personer har smakt på 6 ulike oster. Hver person har gitt hver ost en heltallskarakter mellom 1 og 9, etter hvor godt de likte osten (9 er best og 1 er dårligst).

Filen inneholder en variabel `Code` som angir personens kjønn (`A` for kvinne og `B` for mann), og en variabel for hver av ostene: `A_Cow_Full_fat`, `C_Cow_Full_fat`, `D_Cow_Low_fat`, `E_Cow_Low_fat`, `G_Buffalo_Full_fat` og `I_Buffalo_Full_fat`. Ostene A, C, D og E er lagd av kumelk, mens G og I er lagd av bøffelmelk. Ostene A, C, G og I har vanlig fettinhold, mens D og E er magre.

Gjør en prinsipalkomponentanalyse av karakterene. Hvor mange komponenter er viktige? Plott ladninger og skårer og se om du ser noen mønstre. Tips: Dataene kan leses inn og settes opp slik i `R`:

```
tmp <- read.table("datasett_41.txt")
names(tmp)
koder <- tmp$Code
X <- as.matrix(tmp[,2:7])
```

Plott skårene med koder og/eller farger som angir kjønn. Ser du noe mønster nå? Tolk ladningene. (Tips: Man kan plotte kodene ved å bruke argumentet `pch = as.character(koder))` i plottefunksjonen, og man kan legge på farger med argumentet `col = as.numeric(koder)`.)