IMM, 2002-09-12
Poul Thyregod

Lidt om frihedsgrader og kvadratiske former

Introduktion

I statistiske anvendelser møder man ofte begrebet "frihedsgrader" (degrees of freedom) knyttet til kvadratafvigelsessummer.

Vi mødte begrebet første gang ved udtrykket for den empiriske varians, Petrucelli side 56

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{1}$$

hvor vi dividerede kvadratafvigelsessummen med $n-1$, nemlig antallet af frihedsgrader svarende til kvadratafvigelsessummen.

Meget forenklet dækker begrebet over *det effektive antal led* i kvadratafvigelsessummen.

I denne note vil vi give et par fortolkninger af begrebet i relation til dataanalyse. Vi vil senere i kurset se, at begrebet også har en naturlig fortolkning i forbindelse med fordelingsmodeller.

En geometrisk fortolkning

Lad $y_1, y_2, \ldots, y_n$ angive n vilkårlige tal, og betragt kvadratsummen

$$Q = \sum_{i=1}^{n} y_i^2 \ .$$

Lad som vanligt $\overline{y}$ angive gennemsnittet af $y_1, y_2, \ldots, y_n$, dvs

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2}$$

Betragt

$$\begin{aligned}\sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n}[(y_i-\overline{y})+\overline{y}]^2 \\ &= \sum_{i=1}^{n}(y_i-\overline{y})^2+\sum_{i=1}^{n}\overline{y}^2+2\sum_{i=1}^{n}\overline{y}(y_i-\overline{y}) \\ &= \sum_{i=1}^{n}(y_i-\overline{y})^2+n\overline{y}^2 \end{aligned} \tag{3}$$

hvor produktleddet forsvinder på grund af relationen

$$\sum_{i=1}^{n}\overline{y}(y_i-\overline{y})=\overline{y}\sum_{i=1}^{n}(y_i-\overline{y})=\overline{y}\left[\sum_{i=1}^{n}y_i-n\overline{y}\right]=0 \tag{4}$$

idet $n\overline{y}=\sum_{i=1}^{n} y_i$.

Lad

$$\mathbf{y}=\begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix} \text{ og } \overline{\mathbf{y}}=\begin{Bmatrix} \overline{y} \\ \overline{y} \\ \vdots \\ \overline{y} \end{Bmatrix}$$

så kan vi udtrykke (3) som

$$\mathbf{y}'\mathbf{y}=(\mathbf{y}-\overline{\mathbf{y}})'(\mathbf{y}-\overline{\mathbf{y}})+\overline{\mathbf{y}}'\,\overline{\mathbf{y}} \tag{5}$$

hvor produktleddet forsvandt på grund af ortogonalitetsrelationen

$$\overline{\mathbf{y}}'\,(\mathbf{y}-\overline{\mathbf{y}})=0 \tag{6}$$

der netop er (4).

Opspaltningen (5) kan fortolkes geometrisk ved at opfatte $\mathbf{y}$ som koordinaterne til et punkt, P, i det n-dimensionale rum. Punktet $\overline{\mathbf{y}}$ er da et punkt, P^*, på "vinkelhalveringslinien" ($y_1=y_2=\cdots=y_n$) og netop det punkt, der fremkommer ved at projicere $\mathbf{y}$ ortogonalt ned på vinkelhalveringslinien. Projektionen er ortogonal på grund af relationen (6). Opspaltningen (5) udtrykker da den pythagoræiske relation, at afstanden OP fra $\mathbf{y}$ til koordinatsystemets nulpunkt er summen af afstanden PP^* fra $\mathbf{y}$ til vinkelhalveringslinien og afstanden OP^* fra $\overline{\mathbf{y}}$ til nulpunktet.

Afstanden PP^* er jo egentlig en afstand i et $(n-1)$-dimensionalt vektorrum ortogonalt på vinkelhalveringslinien, derfor er der kun $n-1$ *effektive led* i kvadratafvigelsessummen (1). Der er nemlig *det lineære bånd* (4) mellem de n y-værdier, så de enkelte y-værdier kan ikke variere frit. Hvis vi kender $y_1, \ldots, y_{n-1}$ og $\overline{y}$, er y_n jo givet ud fra disse værdier.

Mere formelt angiver frihedsgraderne for en kvadratafvigelsessum *rangen* (dvs antallet af positive egenværdier) af den kvadratiske form, der svarer til kvadratafvigelsessummen.

Lineære bånd

I praksis møder man ofte den simple regneregel for bestemmelse af frihedsgrader, at det er antallet af led i kvadratafvigelsen minus antallet af lineære bånd.

Denne regel kan være en god praktisk huskeregel, men i nogle situationer med mange bånd, der griber ind i hinanden, er det ikke helt trivielt at bestemme det effektive antal lineære bånd.

En ortogonal dekomposition af kvadratafvigelsessummen

Den meget videbegærlige læser vil måske finde det af interesse, at man ved Gram-Schmidt ortogonalisering kan opspalte den kvadratiske form

$$Q = \sum_{i=1}^{n} y_i^2 \ .$$

i n ortogonale kvadratiske former, hvor den første netop er bestemt ved $\overline{y}^2$.

Man benytter den lineære transformation af talsættet $y_1, y_2, \ldots, y_n$ til et sæt $\overline{y}, z_2, \ldots, z_n$ givet ved

$$\begin{aligned}
y_1 &= \overline{y} + z_2/\sqrt{1\cdot 2} + z_3/\sqrt{2\cdot 3} + \cdots + z_n\sqrt{(n-1)\cdot n} \\
y_2 &= \overline{y} - z_2/\sqrt{1\cdot 2} + z_3/\sqrt{2\cdot 3} + \cdots + z_n\sqrt{(n-1)\cdot n} \\
y_3 &= \overline{y} \qquad\qquad - 2z_3/\sqrt{2\cdot 3} + \cdots + z_n/\sqrt{(n-1)\cdot n}
\end{aligned}$$

$$
\begin{aligned}
&\vdots \\
y_n &= \overline{y} \qquad\qquad\qquad\qquad\qquad\qquad - (n-1)z_n/\sqrt{(n-1)\cdot n}
\end{aligned}
$$

Transformationen kaldes Helmerts transformation (efter den tyske matematiker Helmert, som introducerede transformationen i 1875-76 til brug for beskrivelse af fordelingsmodellen for s^2).

Det kan vises, at

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=2}^{n} z_i^2$$