CdL in Scienze Statistiche ed Economiche - Università degli Studi di Milano-Bicocca

Lezione : Matrice dei dati centrati e standardizzati

Docente: Aldo Solari

1 Dati centrati e standardizzati

1.1 Il vettore delle medie

$$\boxed{\underset{p\times 1}{\bar{x}} = \frac{1}{n}\underset{p\times n}{X'}\underset{n\times 1}{1}}$$

1.2 La matrice dei dati centrati

$$\boxed{\underset{n\times p}{\tilde{X}} = X - 1\bar{x}' = \left(\underset{n\times n}{I} - \frac{1}{n}\underset{n\times 1}{1}\underset{1\times n}{1'}\right)\underset{n\times p}{X} = \underset{n\times n}{H}\underset{n\times p}{X}}$$

1.3 La matrice di centramento

$$\boxed{\underset{n\times n}{H} = \underset{n\times n}{I} - \frac{1}{n}\underset{n\times 1}{1}\underset{1\times n}{1'}}$$

Observation 1.1. $\underset{n\times n}{H}$ *è una matrice simmetrica*

$$\underset{n\times n}{H} = \underset{n\times n}{I} - \frac{1}{n}\underset{n\times 1}{1}\underset{1\times n}{1'} = \begin{bmatrix} 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \cdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{bmatrix}$$

Proposition 1.2. $\underset{n\times n}{H}$ *è una matrice idempotente*

Dimostrazione.

$$\begin{aligned}\underset{n\times n}{H}\underset{n\times n}{H} &= (\underset{n\times n}{I} - \frac{1}{n}\underset{n\times 1}{1}\underset{1\times n}{1'})(\underset{n\times n}{I} - \frac{1}{n}\underset{n\times 1}{1}\underset{1\times n}{1'}) \\ &= \underset{n\times n}{I} - 2\frac{1}{n}\underset{n\times 1}{1}\underset{1\times n}{1'} + \frac{1}{n^2}(\underset{n\times 1}{1}\underset{1\times n}{1'})(\underset{n\times 1}{1}\underset{1\times n}{1'}) \\ &= \underset{n\times n}{I} - 2\frac{1}{n}\underset{n\times 1}{1}\underset{1\times n}{1'} + \frac{1}{n^2}\underset{n\times 1}{1}\, n \underset{1\times n}{1'} \\ &= \underset{n\times n}{I} - \frac{1}{n}\underset{n\times 1}{1}\underset{1\times n}{1'} = \underset{n\times n}{H}\end{aligned}$$

□

Observation 1.3. *Centrare la matrice dei dati centrati non produce alcun effetto:*

$$\underset{n\times n}{H}\underset{n\times p}{\tilde{X}} = \underset{n\times n}{H}\underset{n\times n}{H}\underset{n\times p}{X} = \underset{n\times p}{\tilde{X}}$$

1.4 Matrice di varianze/covarianze

$$\boxed{\underset{p\times p}{S} = \frac{1}{n}\underset{p\times n}{\tilde{X}'}\underset{n\times p}{\tilde{X}}}$$

Si noti che possiamo esprimere la matrice di varianza/covarianze come segue:

$$nS = \tilde{X}'\tilde{X} = (HX)'(HX) = X'H'HX = X'HX$$

sfruttando la simmetria $H' = H$ e l'idempotenza $HH = H$

Proposition 1.4. $\underset{p\times p}{S}$ *è una matrice semidefinita positiva*

Dimostrazione.

$$\begin{aligned}n\underset{1\times p}{a'}\underset{p\times p}{S}\underset{p\times 1}{a} &= \underset{1\times p}{a'}\underset{p\times n}{\tilde{X}'}\underset{n\times p}{\tilde{X}}\underset{p\times 1}{a} \\ &= (\underset{n\times p}{\tilde{X}}\underset{p\times 1}{a})'(\underset{n\times p}{\tilde{X}}\underset{p\times 1}{a}) \\ &= \underset{1\times n}{b'}\underset{n\times 1}{b} \\ &= \sum_{i=1}^{n} b_i^2 \geq 0 \quad \forall \underset{p\times 1}{a}\end{aligned}$$

dove $\underset{n\times 1}{b} = \underset{n\times p}{\tilde{X}}\underset{p\times 1}{a}$. □

Se calcoliamo la matrice di varianze/covarianze per una generica matrice M abbiamo

$$nS^M = (HM)'(HM) = M'H'HM = M'HHM = M'HM$$

Observation 1.5. *La matrice di varianze/covarianze calcolata per* $\underset{n\times p}{\tilde{X}}$ *risulta uguale alla varianze/covarianze calcolata per per* $\underset{n\times p}{X}$

$$\underset{p\times p}{S^{\tilde{X}}} = \frac{1}{n}\left(\underset{p\times n}{\tilde{X}'}\underset{n\times n}{H'}\right)\left(\underset{n\times n}{H}\underset{n\times p}{\tilde{X}}\right) = \frac{1}{n}\underset{p\times n}{\tilde{X}'}\underset{n\times p}{\tilde{X}} = \underset{p\times p}{S^X}$$

1.5 Matrice dei dati standardizzati

$$\boxed{\underset{n\times p}{Z} = \underset{n\times p}{\tilde{X}}\underset{p\times p}{D^{-1/2}}}$$

dove $\underset{p\times p}{D}^{1/2} = \text{diag}(\sqrt{s_{11}}, \dots, \sqrt{s_{pp}})$. La matrice

$$\underset{p\times p}{D^{-1/2}} = \text{diag}\left(\frac{1}{\sqrt{s_{11}}}, \dots, \frac{1}{\sqrt{s_{pp}}}\right) = \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \ddots & \cdots \\ 0 & \cdots & 0 & \cdots & \frac{1}{\sqrt{s_{pp}}} \end{bmatrix}$$

è la matrice inversa di $\underset{p\times p}{D}^{1/2}$. Questo richiede che $s_{11}, \dots, s_{pp}$ siano tutti diversi da 0.
(Moltiplicare $\underset{n\times p}{\tilde{X}}$ da destra per $\underset{p\times p}{D^{-1/2}}$ equivale a moltiplicare la j-sima colonna di $\underset{n\times p}{\tilde{X}}$ per $\frac{1}{\sqrt{s_{jj}}}$)

1.6 Matrice di correlazione

$$\boxed{\underset{p\times p}{R} = \underset{p\times p}{D^{-1/2}}\underset{p\times p}{S}\underset{p\times p}{D^{-1/2}}}$$

- Moltiplicare $\underset{p\times p}{S}$ da sinistra per $\underset{p\times p}{D^{-1/2}}$ equivale a moltiplicare l'i-sima riga di $\underset{p\times p}{S}$ per $\frac{1}{\sqrt{s_{ii}}}$;
- Moltiplicare $\underset{p\times p}{S}$ da destra per $\underset{p\times p}{D^{-1/2}}$ equivale a moltiplicare la j-sima colonna di $\underset{p\times p}{S}$ per $\frac{1}{\sqrt{s_{jj}}}$;

Observation 1.6.

$$\boxed{\underset{p\times p}{S} = \underset{p\times p}{D^{1/2}}\underset{p\times p}{R}\underset{p\times p}{D^{1/2}}}$$

$$R = D^{-1/2}SD^{-1/2}$$
$$D^{1/2}RD^{1/2} = D^{1/2}D^{-1/2}SD^{-1/2}D^{1/2}$$

e visto che $D^{-1/2}$ è la matrice inversa di $D^{1/2}$, per definizione

$$\underset{p\times p}{D^{1/2}}\underset{p\times p}{D^{-1/2}} = \underset{p\times p}{D^{-1/2}}\underset{p\times p}{D^{1/2}} = \underset{p\times p}{I}$$

e quindi $\underset{p\times p}{D^{1/2}}\underset{p\times p}{R}\underset{p\times p}{D^{1/2}} = \underset{p\times p}{S}$

Observation 1.7. *La matrice di varianze/covarianze calcolata per Z risulta uguale alla matrice di correlazione calcolata per X.*

$$\begin{aligned}
\underset{p\times p}{S^Z} &= \frac{1}{n}(Z'H')(HZ) \\
&= \frac{1}{n}Z'Z \\
&= \frac{1}{n}D^{-1/2}\tilde{X}'\tilde{X}D^{-1/2} \\
&= D^{-1/2}SD^{-1/2} = \underset{p\times p}{R}^{X}
\end{aligned}$$

Example 1.8. *Matrice*

$$\underset{4\times 2}{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

Vettore delle medie

$$\underset{2\times 1}{\bar{x}} = \frac{1}{4}\begin{bmatrix} 42 & 52 & 48 & 58 \\ 4 & 5 & 4 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

Matrice di centramento

$$\begin{aligned}
\underset{4\times 4}{H} &= \underset{4\times 4}{I} - \frac{1}{4}\underset{1\times 4}{1}\underset{4\times 1}{1'} \\
&= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \frac{1}{4}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1-1/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 1-1/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1-1/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 1-1/4 \end{bmatrix}
\end{aligned}$$

Matrice dei dati centrati:

$$\underset{4\times2}{\tilde{X}} = HX$$
$$= \begin{bmatrix} 1-1/4 & -1/4 & -1/4 & -1/4 \\ -1/4 & 1-1/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & 1-1/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 1-1/4 \end{bmatrix} \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$
$$= \begin{bmatrix} -8 & 0 \\ 2 & 1 \\ -2 & 0 \\ 8 & -1 \end{bmatrix}$$

Matrice di varianze/covarianze

$$\underset{2\times2}{S} = \frac{1}{4}\tilde{X}'\tilde{X}$$
$$= \frac{1}{4}\begin{bmatrix} -8 & 2 & -2 & 8 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} -8 & 0 \\ 2 & 1 \\ -2 & 0 \\ 8 & -1 \end{bmatrix} = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}$$

Matrice di correlazione:

$$\underset{2\times2}{R} = \underset{2\times2}{D}^{-1/2}\underset{2\times2}{S}\underset{2\times2}{D}^{-1/2}$$
$$= \begin{bmatrix} 1/\sqrt{34} & 0 \\ 0 & 1/\sqrt{0.5} \end{bmatrix} \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix} \begin{bmatrix} 1/\sqrt{34} & 0 \\ 0 & 1/\sqrt{0.5} \end{bmatrix}$$
$$= \begin{bmatrix} 1 & -1.5/(\sqrt{34}\sqrt{5}) \\ -1.5/(\sqrt{34}\sqrt{5}) & 1 \end{bmatrix}$$

Matrice dei dati standardizzati:

$$\underset{4\times2}{Z} = \begin{bmatrix} -8 & 0 \\ 2 & 1 \\ -2 & 0 \\ 8 & -1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{34} & 0 \\ 0 & 1/\sqrt{0.5} \end{bmatrix} = \begin{bmatrix} -8/\sqrt{34} & 0 \\ 2/\sqrt{34} & 1/\sqrt{0.5} \\ -2/\sqrt{34} & 0 \\ 8/\sqrt{34} & -1/\sqrt{0.5} \end{bmatrix}$$

Matrice dei dati	Vettore delle medie	Matrice di varianze/covarianze	Matrice di correlazione
$\underset{n\times p}{X}$	$\underset{p\times 1}{\bar{x}}$	$\underset{p\times p}{S}$	$\underset{p\times p}{R}$
$\underset{n\times p}{\tilde{X}}$	$\underset{p\times 1}{0}$	$\underset{p\times p}{S^{\tilde{X}}} = \underset{p\times p}{S}$	$\underset{p\times p}{R^{\tilde{X}}} = \underset{p\times p}{R}$
$\underset{n\times p}{Z}$	$\underset{p\times 1}{0}$	$\underset{p\times p}{S^{Z}} = \underset{p\times p}{R}$	$\underset{p\times p}{R^{Z}} = \underset{p\times p}{R}$