STATISTICA MULTIVARIATA

Corso di laurea in matematica | Anno accademico 2004/05 | Numero 2

Il principio di dualità

Assumiamo che i valori di due variabili numeriche (ad es. le concentrazioni di due aminoacidi nel sangue) siano stati misurati per n oggetti o indivudui (ad es. pazienti); otteniamo così n punti $(x_1, y_1), \dots, (x_n, y_n)$ nel piano $\mathbb{R}_2$ che possono essere rappresentati da una matrice

$$\begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}$$

a 2 colonne ed n righe. Questa matrice si chiama la *matrice dei dati*.

Le righe (x_i, y_i) forniscono da sole tutta l'informazione contenuta nella matrice, così come le colonne. Ciononostante guardando solo le righe o solo le colonne, in un certo senso si vede solo la metà di questa informazione; l'altra metà è nascosta, difficile da comprendere. Solo lavorando contemporaneamente con righe e colonne tutta l'informazione appare sempre chiaramente davanti ai nostri occhi.

Ciò è tipico per *situazioni di dualità*, in cui due aspetti di uno stesso oggetto o di una stessa struttura si determinano reciprocamente in modo (più o meno) completo e in cui quindi ogni enunciato su uno dei due aspetti implica un enunciato anche sull'altro aspetto, e dove ciononostante spesso questi due enunciati devono essere formulati o dimostrati in modo apparentemente molto diverso.

Può così accadere che in un oggetto un enunciato o un algoritmo si presentino in veste molto semplice e diventino molto più difficili quando vengono tradotti nell'altro aspetto. È quindi spesso preferibile tener presente i due aspetti contemporaneamente invece di cercare di ridurre l'uno all'altro: per definizione ciò sarebbe possibile, ma a spese della comprensione.

Uno dei più noti esempi di dualità è l'analisi di Fourier; il buon analista di Fourier ha sempre davanti agli occhi entrambi gli oggetti della dualità e non preferisce nessuno dei due.

In questo spirito introduciamo adesso, partendo dalla nostra matrice di dati, le colonne

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{y=} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

come nuovi oggetti. x e y come punti sono elementi di un $\mathbb{R}^n$ a dimensione molto alta (ad esempio $n = 50000$ in uno screening di 50000 neonati); la loro geometria implica e chiarisce talvolta circostanze per i dati in $\mathbb{R}_2$ che sarebbe difficile individuare direttamente nel piano dei dati.

La centralizzazione

Situazione 6.1. Siano $x, y \in \mathbb{R}^n$ con $x = (x_1, \dots, x_n)^t$, $y = (y_1, \dots, y_n)^t$. Quando necessario (e lo sarà quasi sempre) supponiamo $n \geq 2$. A partire dalla situazione 7.6 chiederemo che x ed y non siano diagonali, cioè che non abbiano coefficienti tutti uguali.

Definizione 6.2. $1^n := \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ sia il vettore di $\mathbb{R}^n$ i cui coefficienti sono tutti uguali a 1. Questo vettore ausiliario è molto utile nella statistica geometrica. La retta $\mathbb{R}1^n$ si chiama la retta *diagonale* di $\mathbb{R}^n$.

Osservazione 6.3. $|1^n| = \sqrt{n}$.

Definizione 6.4. La *media* $\overline{x}$ di x è definita come

$$\overline{x} := \frac{1}{n} \sum_{k=1}^{n} x_k$$

Osservazione 6.5. *La media è un operatore lineare; per $\lambda, \mu \in \mathbb{R}$ abbiamo quindi*

$$\overline{\lambda x + \mu y} = \lambda \overline{x} + \mu \overline{y}$$

Osservazione 6.6. *Per $\lambda \in \mathbb{R}$ si ha $\overline{\lambda 1^n} = \lambda$. In particolare $\overline{\overline{x} 1^n} = \overline{x}$.*

Dimostrazione. Infatti $\dfrac{\lambda + \dots + \lambda}{n} = \lambda$.

Definizione 6.7. Il vettore $x^{\mathrm{CE}} := x - \overline{x} 1^n$ si chiama la *centralizzazione* di x.

Proposizione 6.8. $\overline{x^{\mathrm{CE}}} = 0$.

Dimostrazione.

$$\overline{x^{\mathrm{CE}}} = \overline{x - \overline{x} 1^n} = \overline{x} - \overline{\overline{x} 1^n} = \overline{x} - \overline{x} = 0$$

Abbiamo usato la linearità della media (osservazione 6.5) e l'osservazione 6.6.

Corollario 6.9. $(x^{\mathrm{CE}})^{\mathrm{CE}} = x^{\mathrm{CE}}$.

Dimostrazione. Abbiamo $(x^{\mathrm{CE}})^{\mathrm{CE}} = x^{\mathrm{CE}} - \overline{x^{\mathrm{CE}}} 1^n = x^{\mathrm{CE}}$.

Osservazione 6.10. $\| x, 1^n \| = n \overline{x}$.

Dimostrazione. $\| x, 1^n \| = x_1 + \dots + x_n$.

Corollario 6.11. $x \perp 1^n \iff \overline{x} = 0$.

I vettori che hanno media 0 sono quindi esattamente quei vettori che sono ortogonali alla retta diagonale; essi formano l'iperpiano $1^{n\perp}$ normale alla retta diagonale.

Corollario 6.12. $x^{\mathrm{CE}} \perp 1^n$.

Teorema 6.13. *x^{CE} è la proiezione ortogonale di x sull'iperpiano $1^{n\perp}$, e $\overline{x} 1^n$ è la proiezione ortogonale di x sulla retta diagonale.*

Dimostrazione. $x^{CE} \in 1^{n\perp}$ per il corollario 6.12, mentre è chiaro che $\overline{x} 1^n$ appartiene alla retta diagonale. Sia $\| v, 1^n \| = 0$. Allora $\| v, x - x^{CE} \| = \| v, \overline{x} 1^n \| = 0$.

Infine $\| x - \overline{x} 1^n, 1^n \| = \| x^{CE}, 1^n \| = 0$.

In questo numero

Dipendenza funzionale

In matematica il concetto di funzione è definito in modo molto generale. Se in una tabella come

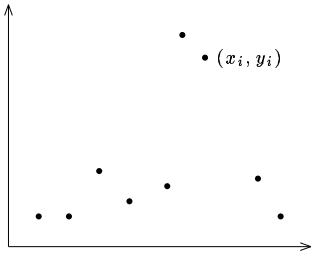
$$\begin{pmatrix} 3 & 2 \\ 5 & 1 \\ 2 & 8 \\ 1 & 0 \\ 6 & 0 \\ 8 & 2 \end{pmatrix}$$

gli elementi della prima colonna sono tutti distinti, ciò è sufficiente per poter considerare la seconda colonna come funzione della prima: definiamo una funzione

$$f : \{3, 5, 2, 1, 6, 8\} \longrightarrow \{0, 1, 2, 8\}$$

semplicemente ponendo $f(3) = 2$, $f(5) = 1$, $f(2) = 8$, $f(1) = 0$, $f(6) = 0$, $f(8) = 2$. In questo senso quindi la seconda colonna dipende in modo funzionale dalla prima, benché si possa difficilmente affermare l'esistenza di qualche legame statistico o addirittura causale tra le due variabili. Solo quando la funzione appartiene a una classe determinata e possibilmente semplice di funzioni (lineari, quadratiche, logaritmiche, monotone, sigmoidali, sinusoidali) si può cercare di associare a una tale relazione un significato statistico.

Quindi anche in una rappresentazione grafica dei dati nel piano, in cui i valori x_i sono tutti distinti, ciò da solo ci permette di considerare i valori y_i come funzione degli x_i nel senso della matematica.



I *modelli lineari* sono impiegati con successo in molte indagini statistiche; questo numero è dedicato al caso più semplice, la rappresentazione di una dipendenza approssimativamente lineare di x da y mediante una retta di regressione.

La funzione Sg.cen

In R la media di x si ottiene con `mean(x)`. La sezione SG della nostra libreria conterrà le funzioni statistiche geometriche. Definiamo una funzione `Sg.cen` che calcola la centralizzazione di un vettore.

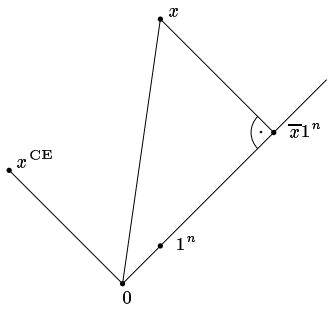
```
Sg.cen = function (x) x-mean(x)
```

Esempio:

```
x=c(1,3,5,2,0)
m=mean(x)
print(m) # 2.2
uno=rep(1,5)
print(x-m*uno)
# -1.2 0.8 2.8 -0.2 -2.2
print(Sg.cen(x))
# -1.2 0.8 2.8 -0.2 -2.2
```

Un esempio in $\mathbb{R}^2$. Sia $x = \begin{pmatrix}1\\7\end{pmatrix}$. Allora $\overline{x} = 4$, perciò

$$x^{\mathrm{CE}} = \begin{pmatrix}1\\7\end{pmatrix} - \begin{pmatrix}4\\4\end{pmatrix} = \begin{pmatrix}-3\\3\end{pmatrix}$$



Deviazione standard e varianza

Definizione 7.1. La *deviazione standard* s_x di x è definita da

$$s_x := \frac{|x^{\mathrm{CE}}|}{\sqrt{n-1}}$$

s_x^2 si chiama la *varianza* di x; abbiamo quindi

$$s_x^2 = \frac{|x^{\mathrm{CE}}|^2}{n-1}$$

La *covarianza* s_{xy} di x ed y è definita da

$$s_{xy} := \frac{\|x^{\mathrm{CE}}, y^{\mathrm{CE}}\|}{n-1}$$

Abbiamo in particolare $s_{xx} = s_x^2$.

In R la deviazione standard e la varianza di `x` sono date da `sd(x)` e `var(x)`, la covarianza di `x` ed `y` da `cov(x,y)`.

`var` e `cov` sono definite anche per matrici, come vedremo.

Lemma 7.2. *Valgono le uguaglianze*

$$\|x^{\mathrm{CE}}, y^{\mathrm{CE}}\| = \|x^{\mathrm{CE}}, y\| = \|x, y\| - n\overline{x}\,\overline{y}$$

Da esse seguono le relazioni

$$s_{xy} = \frac{\|x, y\| - n\overline{x}\,\overline{y}}{n-1}$$

$$|x^{\mathrm{CE}}|^2 = |x|^2 - n\overline{x}^2$$

$$s_x^2 = \frac{|x|^2 - n\overline{x}^2}{n-1}$$

Queste formule sono usate molto spesso.

Dimostrazione. Per il corollario 6.12 e l'osservazione 6.10 abbiamo

$$\begin{aligned}\|x^{\mathrm{CE}}, y^{\mathrm{CE}}\| &= \|x^{\mathrm{CE}}, y - \overline{y}1^n\| \\ &= \|x^{\mathrm{CE}}, y\| = \|x - \overline{x}1^n, y\| \\ &= \|x, y\| - \overline{x}\|1^n, y\| \\ &= \|x, y\| - n\overline{x}\,\overline{y}\end{aligned}$$

Le normalizzazioni x^{NG} ed x^{NS}

Osservazione 7.3. *Sia $v \in \mathbb{R}^n$ e $v \neq 0$. Allora il vettore $\frac{v}{|v|}$ possiede lunghezza 1 e mostra naturalmente nella stessa direzione di v.*

Definizione 7.4. x si chiama *diagonale*, se tutti i coefficienti di x sono uguali.

Osservazione 7.5. *Sono equivalenti:*

(1) x *è diagonale.*
(2) $x \in \mathbb{R}1^n$.
(3) $x = \overline{x}1^n$.
(4) $x^{\mathrm{CE}} = 0$.
(5) $s_x = 0$.

Situazione 7.6. Assumiamo da ora in avanti che x ed y non siano diagonali e quindi $x^{\mathrm{CE}} \neq 0$, $y^{\mathrm{CE}} \neq 0$. È chiaro che ciò implica che $n \geq 2$.

Dall'osservazione 7.3 vediamo anche che $s_x > 0$ ed $s_y > 0$.

Definizione 7.7. Il vettore

$$x^{\mathrm{NG}} := \frac{x^{\mathrm{CE}}}{|x^{\mathrm{CE}}|}$$

si chiama la *normalizzazione geometrica* di x. In statistica si considera anche il vettore

$$x^{\mathrm{NS}} := \frac{x^{\mathrm{CE}}}{s_x}$$

che possiamo chiamare la *normalizzazione statistica* di x.

Nota 7.8. $x^{\mathrm{NS}} = \sqrt{n-1}\,x^{\mathrm{NG}}$.

x^{NS} si distingue quindi da x^{NG} solo per il fattore $\sqrt{n-1}$. Le considerazioni geometriche che seguono potrebbero perciò essere eseguite anche con x^{NS}, risultano però più trasparenti e le formule che si ottengono più semplici, se si usa x^{NG}.

Dimostrazione. Abbiamo

$$\begin{aligned}x^{\mathrm{NS}} &= \frac{x^{\mathrm{CE}}}{s_x} \\ &= \frac{x^{\mathrm{CE}}}{|x^{\mathrm{CE}}|}\frac{|x^{\mathrm{CE}}|}{s_x} = x^{\mathrm{NG}}\frac{|x^{\mathrm{CE}}|}{s_x}\end{aligned}$$

Ma per la definizione 7.1 abbiamo

$$\frac{|x^{\mathrm{CE}}|}{s_x} = \sqrt{n-1}.$$

Osservazione 7.9. *x^{NG} e x^{NS} sono vettori paralleli ad x^{CE}, perciò*

$$\overline{x^{\mathrm{NG}}} = \overline{x^{\mathrm{NS}}} = 0$$

Osservazione 7.10. *Sia $v \in \mathbb{R}^n$, $v \neq 0$ e $\overline{v} = 0$. Allora $v^{\mathrm{NG}} = \frac{v}{|v|}$.*

Corollario 7.11. $(x^{\mathrm{NG}})^{\mathrm{NG}} = x^{\mathrm{NG}}$.

Dimostrazione. Ciò segue dalle osservazioni 7.9 e 7.10 perché $|x^{\mathrm{NG}}| = 1$.

Osservazione 7.12. *Sia $\alpha > 0$. Allora*

$$(\alpha x^{\mathrm{NG}})^{\mathrm{NG}} = x^{\mathrm{NG}}$$

Dimostrazione. Anche ciò segue dall'osservazione 7.10, perché $|\alpha x^{\mathrm{NG}}| = |\alpha| = \alpha$.

Corollario 7.13.

$$(x^{\mathrm{CE}})^{\mathrm{NG}} = (x^{\mathrm{NS}})^{\mathrm{NG}} = x^{\mathrm{NG}}$$

Dimostrazione. Dalla definizione 7.7 vediamo che x^{CE} e x^{NS} si distinguono da x^{NG} solo per i fattori positivi $|x^{\mathrm{CE}}|$ risp. s_x.

Definiamo le funzioni in R per la normalizzazione geometrica e la normalizzazione statistica di un vettore. Le funzioni che calcolano il prodotto scalare e la lunghezza di vettori si trovano a pagina 8.

```
Sg.ng = function (x)
{cen=Sg.cen(x); r=Mv.lun(cen)
if (r==0) NA else cen/r}

Sg.ns = function (x)
{s=sd(x); if (s==0) NA
else Sg.cen(x)/s}
```

Proviamo un esempio numerico. Sia

$$x = \begin{pmatrix}13\\1\\5\\2\\9\end{pmatrix}$$

Allora

$$\overline{x} = \frac{13+1+5+2+9}{5} = \frac{30}{5} = 6$$

quindi

$$x^{\mathrm{CE}} = \begin{pmatrix}13\\1\\5\\2\\9\end{pmatrix} - \begin{pmatrix}6\\6\\6\\6\\6\end{pmatrix} = \begin{pmatrix}7\\-5\\-1\\-4\\3\end{pmatrix}$$

e

$$\begin{aligned}|x^{\mathrm{CE}}| &= \sqrt{49+25+1+16+9} \\ &= \sqrt{100} = 10\end{aligned}$$

per cui

$$x^{\mathrm{NG}} = \frac{1}{10}x^{\mathrm{CE}} = \begin{pmatrix}0.7\\-0.5\\-0.1\\-0.4\\0.3\end{pmatrix}$$

e

$$x^{\mathrm{NS}} = \sqrt{4}\,x^{\mathrm{NG}} = 2x^{\mathrm{NG}} = \begin{pmatrix}1.4\\-1\\-0.2\\-0.8\\0.6\end{pmatrix}$$

Controlliamo le nostre funzioni con

```
x=c(13,1,5,2,9)

cen=Sg.cen(x)
print(cen)
# 7 -5 -1 -4 3

ng=Sg.ng(x)
print(ng)
# 0.7 -0.5 -0.1 -0.4 0.3

ns=Sg.ns(x)
print(ns)
# 1.4 -1.0 -0.2 -0.8 0.6
```

Prodotto scalare e lunghezza

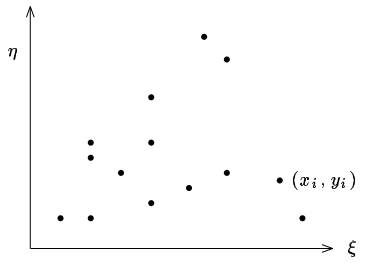
Nel prodotto scalare usiamo la funzione `crossprod` di R che calcola il prodotto $A^t B$ per matrici. `drop` trasforma una matrice 1×1 in uno scalare.

```
Mv.scalare = function (a,b)
{p=crossprod(a,b); drop(p)}

Mv.lun = function (x)
sqrt(Mv.scalare(x,x))
```

La retta di regressione

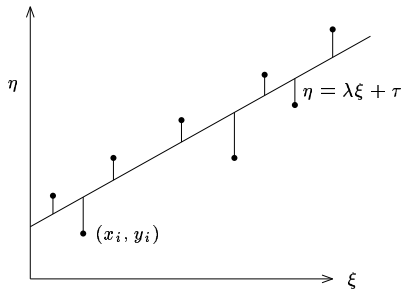
In statistica spesso in un primo momento sono dati n punti $(x_1, y_1), \ldots, (x_n, y_n)$ nel piano $\mathbb{R}_2$, da cui, secondo il principio di dualità, possiamo formare i vettori $x, y \in \mathbb{R}^n$. Avendo così già assegnato le lettere x e y, denotiamo le coordinate nel piano con ξ ed η.



Cerchiamo adesso di rappresentare (più precisamente di approssimare) i valori y_i mediante una funzione lineare degli x_i, cioè di determinare numeri reali λ e τ tali da minimizzare gli errori $y_i - (\lambda x_i + \tau)$ nel senso che l'espressione

$$F(\lambda, \tau) := \sum_{i=1}^{n} (y_i - (\lambda x_i + \tau))^2$$

sia minima (principio dei *minimi quadrati* di Gauß).



A questo scopo si possono porre uguali a zero le derivate parziali $\dfrac{\partial F}{\partial \lambda}$ e $\dfrac{\partial F}{\partial \tau}$, ottenendo così un sistema lineare in λ e τ che, nella nostra ipotesi che x non sia diagonale, possiede un'unica soluzione (λ, τ). La retta determinata dall'equazione $\eta = \lambda\xi + \tau$ si chiama la *retta di regressione* degli y_i rispetto agli x_i (o di y rispetto ad x).

Nel seguito useremo (λ, τ) sia per denotare questa soluzione che per parametri generici variabili; sarà chiaro dal contesto quale dei due significati è usato.

Vogliamo adesso invece dedurre la retta di regressione senza fare uso del calcolo differenziale in modo puramente geometrico. Lavoriamo in $\mathbb{R}^n$ con x, y definiti come finora, nonostante che la retta di regressione sia una retta in $\mathbb{R}_2$ riferita ai punti (x_i, y_i).

Osservazione 8.1.

$$F(\lambda, \tau) = |y - (\lambda x + \tau 1^n)|^2.$$

Proposizione 8.2. *E sia un sottospazio vettoriale di $\mathbb{R}^n$ ed $e_1, \ldots, e_s$ una base ortogonale di E. Siano $y \in \mathbb{R}^n$ e p la proiezione ortogonale di y su E. Allora*

$$p = \alpha_1 e_1 + \ldots + \alpha_s e_s$$

con gli α_k (naturalmente univocamente determinati) dati da

$$\alpha_k = \frac{\|y, e_k\|}{|e_k|^2}$$

Questa formula mostra in particolare che ogni sommando $p_k = \alpha_k e_k$ è la proiezione ortogonale di y sulla retta $\mathbb{R}e_k$ generata da e_k.

p si ottiene come $p = p_1 + \ldots + p_s$.

Dimostrazione. $y - p$ deve essere ortogonale ad e_k per ogni k e quindi deve valere $\|y - p, e_k\| = 0$ o, equivalentemente,

$$\|y, e_k\| = \|p, e_k\|$$

per $k = 1, \ldots, m$. Per l'ortogonalità degli e_j abbiamo però

$$\|p, e_k\| = \|\alpha_k e_k, e_k\| = \alpha \|e_k, e_k\|$$

cosicché $\alpha_k \|e_k, e_k\| = \|y, e_k\|$ e ciò implica l'enunciato.

Nota 8.3. Siccome per ipotesi x non si trova sulla retta $\mathbb{R}1^n$, i punti x ed 1^n generano un piano $P_x \subset \mathbb{R}^n$:

$$P_x = \{\lambda x + \tau 1^n \mid \lambda, \tau \in \mathbb{R}\}$$

in cui λ e τ per ogni punto di P_x sono univocamente determinati. In particolare sono univocamente determinati parametri λ e τ corrispondenti alla proiezione ortogonale p di y su P_x. Ma p è proprio il punto per il quale $F(\lambda, \tau)$ è minimale.

D'altra parte anche $x^{\mathrm{CE}} = x - \overline{x}1^n$ appartiene a P_x e dal corollario 6.12 segue adesso che x^{CE} e 1^n formano una base ortogonale di P_x, quindi, per la proposizione 8.2,

$$p = p_1 + p_2$$

dove p_1 è la proiezione ortogonale di y sulla retta generata da x^{CE} e p_2 la proiezione ortogonale di y sulla retta generata da 1^n. Dal teorema 6.13 sappiamo però anche che $p_2 = \overline{y}1^n$. Abbiamo quindi, con un $\alpha \in \mathbb{R}$ che naturalmente è determinato dalla formula della proposizione 8.2,

$$\begin{aligned} p &= \alpha x^{\mathrm{CE}} + \overline{y}1^n \\ &= \alpha(x - \overline{x}1^n) + \overline{y}1^n \\ &= \alpha x + (\overline{y} - \alpha\overline{x})1^n \end{aligned}$$

Ciò mostra che

$$\begin{aligned} \lambda &= \alpha \\ \tau &= \overline{y} - \lambda\overline{x} \end{aligned}$$

Notiamo che a questo punto abbiamo

$$p = \overline{y}1^n + \lambda x^{\mathrm{CE}}$$

Dobbiamo ancora calcolare λ. Per la proposizione 8.2 e usando il lemma 7.2 abbiamo

$$\begin{aligned} \lambda &= \frac{\|y, x^{\mathrm{CE}}\|}{|x^{\mathrm{CE}}|^2} = \frac{\|x^{\mathrm{CE}}, y^{\mathrm{CE}}\|}{|x^{\mathrm{CE}}|^2} \\ &= \frac{\|x^{\mathrm{CE}}, y^{\mathrm{CE}}\|}{|x^{\mathrm{CE}}||y^{\mathrm{CE}}|} \cdot \frac{|y^{\mathrm{CE}}|}{|x^{\mathrm{CE}}|} \end{aligned}$$

Se poniamo

$$r_{xy} := \frac{\|x^{\mathrm{CE}}, y^{\mathrm{CE}}\|}{|x^{\mathrm{CE}}||y^{\mathrm{CE}}|} = \|x^{\mathrm{NG}}, y^{\mathrm{NG}}\|$$

abbiamo infine

$$\begin{aligned} \lambda &= r_{xy} \frac{|y^{\mathrm{CE}}|}{|x^{\mathrm{CE}}|} \\ \tau &= \overline{y} - \lambda\overline{x} \end{aligned}$$

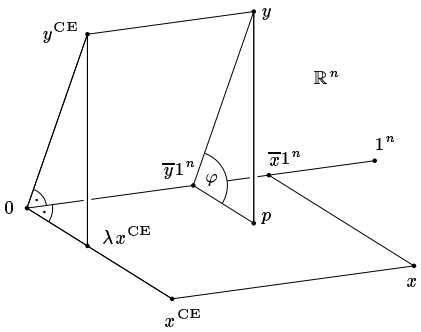
La retta di regressione di y rispetto ad x possiede quindi l'equazione

$$\eta = \lambda\xi + \tau$$

con λ e τ come sopra.

Definizione 8.4. Il rapporto r_{xy} definito nella nota 8.3 si chiama il *coefficiente di correlazione* tra x ed y e verrà studiato in dettaglio nel prossimo numero.

Dai corsi di Geometria sappiamo che r_{xy} non è altro che il coseno dell'angolo φ tra x^{CE} e y^{CE}.



Si osservi che, nonostante si tratti di un disegno in $\mathbb{R}^n$, questa figura è realistica nel senso che la configurazione è tutta contenuta nello spazio (al massimo) 3-dimensionale generato dai vettori 1^n, x^{CE} ed y^{CE}.

In R si ottiene il coefficiente di correlazione r_{xy} con `cor(x,y)`.

Nota 8.5. Siccome $\tau = \overline{y} - \lambda\overline{x}$, l'equazione $\eta = \lambda\xi + \tau$ per la retta di regressione diventa

$$\eta = \lambda\xi + \overline{y} - \lambda\overline{x}$$

e può perciò essere scritta nella forma

$$\eta - \overline{y} = \lambda(\xi - \overline{x})$$

Essa passa quindi per il baricentro $(\overline{x}, \overline{y})$ dei punti (x_i, y_i). Inoltre

$$\lambda = r_{xy} \frac{|y^{\mathrm{CE}}|}{|x^{\mathrm{CE}}|} = r_{xy} \frac{|y - \overline{y}1^n|}{|x - \overline{x}1^n|}$$

cosicché l'equazione assume la forma

$$\frac{\eta - \overline{y}}{|y - \overline{y}1^n|} = r_{xy} \frac{\xi - \overline{x}}{|x - \overline{x}1^n|}$$

Nota 8.6. Siccome x^{NG} e y^{NG} si distinguono da x^{CE} e y^{CE} solo per fattori positivi, è chiaro che x^{NG} e y^{NG} racchiudono lo stesso angolo come x^{CE} e y^{CE}; lo stesso vale per x^{NS} e y^{NS}.

In particolare vediamo che il quoziente di correlazione può anche essere definito come il coseno dell'angolo tra x^{NG} e y^{NG} e che quindi per il corollario 7.13 il coefficiente di correlazione non cambia se sostituiamo x ed y con le loro normalizzazioni geometriche o statistiche o con le loro centralizzazioni.

Osservazioni generali

Dalla definizione 7.1 vediamo che si ha anche

$$\lambda = r_{xy} \frac{s_y}{s_x}$$

Ciò ci permette di usare la funzione `sd` di R per calcolare i coefficienti λ e τ della retta di regressione:

```
Sreg.retta = function (x,y)
{lambda=cor(x,y)*sd(y)/sd(x)
tau=mean(y)-lambda*mean(x)
c(lambda,tau)}
```

Per un controllo applichiamo la teoria a una tabella che si trova a pagina 263 dell'ottimo libro di Kreyszig. La tabella contiene nella colonna degli x_i le densità molteplicate per 10 di esemplari di minerali di ematite; gli y_i sono i contenuti percentuali di ferro.

x	y
28	27
29	23
30	30
31	28
32	30
32	32
32	34
33	33
34	30

Con

```
x=c(28,29,30,31,32,32,32,33,34)
y=c(27,23,30,28,30,32,34,33,30)

retta=Sreg.retta(x,y)

lambda=retta[1]; tau=retta[2]
a=sprintf('lambda = %.2f, tau = %.2f\n',
  lambda,tau)
cat(a)
```

otteniamo

```
lambda = 1.21, tau = -8.01
```

---

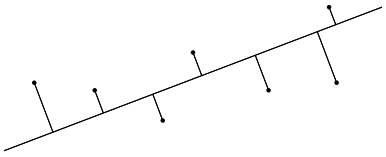
x ed y siano dati dalla tabella

x	y
1	0
0	1
-1	0
0	-1

In questo esempio $\overline{x} = \overline{y} = 0$, quindi $x^{\mathrm{CE}} = x$, $y^{\mathrm{CE}} = y$ e $\tau = 0$. Inoltre $x^{\mathrm{CE}} \perp y^{\mathrm{CE}}$, per cui $r_{xy} = 0$ e quindi anche $\lambda = 0$. La retta di regressione è perciò l'ascisse reale $\eta = 0$.

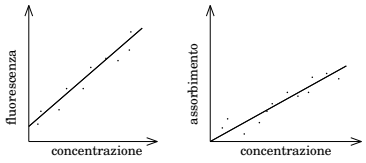
Nota 9.1. Siccome $\lambda = r_{xy} \dfrac{|y^{\mathrm{CE}}|}{|x^{\mathrm{CE}}|}$ e siccome per ipotesi $y^{\mathrm{CE}} \neq 0$, è chiaro che la retta di regressione è parallela all'ascissa reale, come nell'ultimo esempio, se e solo se il coefficiente di correlazione si annulla, e ciò accade se e solo se $x^{\mathrm{CE}} \perp y^{\mathrm{CE}}$.

L'uso della retta di regressione è giustificato soprattutto quando i valori x_i e y_i rappresentano misurazioni di variabili tra le quali è nota l'esistenza di un legame *lineare* che però è stato confuso da errori nella misurazione degli y_i. In questo caso si può assumere che la retta di regressione rappresenti questo legame lineare. Se coesistono errori di misurazione in entrambe le variabili, è preferibile la *regressione ortogonale* mediante proiezioni ortogonali su una retta (invece di proiezioni parallele all'asse y); essa appartiene all'*analisi delle componenti principali* che verrà trattata più avanti.



Regressione ortogonale

In chimica analitica si incontrano spesso leggi lineari che possono essere caratterizzate mediante regressione e correlazione (Otto, Doerffel). Così ad esempio si cerca di calcolare la dipendenza spesso lineare dei segnali di misurazione dai parametri chimici (*curve di calibrazione*).



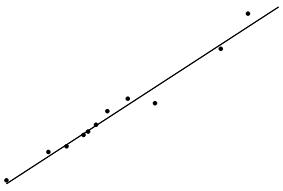
In una *serie temporale* la variabile x è interpretata come il tempo, y come una variabile dipendente dal tempo. Non raramente si osserva una tendenza (in inglese *trend*) lineare a cui si sovrappongono oscillazioni più o meno periodiche e che può essere rapresentata mediante una retta di regressione.

I parametri λ e τ dell'analisi regressionale, calcolati algebricamente, dovrebbero essere stimati, soprattuto se vengono utilizzati a scopi interpolatori. Per fare ciò bisogna o fare ipotesi sulla distribuzione statistica delle variabili casuali corrispondenti alle variabili empiriche x ed y (ad esempio assumendo una distribuzione normale) oppure usare metodi nonparametrici. Non sempre è sicuro che veramente esiste un legame di base (ad es. fisico-chimico) lineare; in questi casi anche la linearità della dipendenza deve essere verificata con metodi statistici.

Legami lineari si osservano spesso nei livelli d'acqua in due postazioni idrometriche distanti allo stesso fiume. Un esempio dal trattato di idrologia di Maniak, pag. 200, leggermente modificato:

x	y
309	193
302	187
283	174
443	291
298	184
319	205
419	260
361	212
267	169
337	216
230	144

I livelli nelle due postazioni sono indicati in cm. Con la nostra funzione `Sreg.retta` troviamo $\lambda = 0.65$, $\tau = -8.6$.



Il modello con una variabile indipendente ξ nelle applicazioni pratiche è spesso troppo semplice; modelli molti più efficaci si ottengono con *regressioni lineari multiple* della forma

$$\eta = \lambda_1 \xi_1 + \ldots + \lambda_k \xi_k + \tau$$

Tali modelli sono già molto generali e vengono usati in molti problemi ingegneristici o econometrici o ad esempio nell'idrologia nella prognosi dei livelli d'acqua, in modo simile alla regressione semplice che abbiamo visto nell'ultimo esempio.

Analisi dei residui

Nell'*analisi dei residui* di una retta di regressione si studiano le differenze (i *residui*)

$$d_i = y_i - (\lambda x_i + \tau)$$

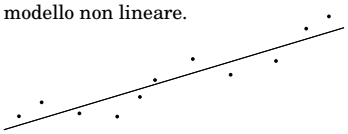
Si ottengono tra l'altro indicazioni per un eventuale possibile miglioramento del modello di regressione.

Lavoriamo di nuovo in $\mathbb{R}^n$ e introduciamo il vettore dei residui

$$d := y - (\lambda x + \tau 1^n) = y - p$$

d è quindi semplicemente il vettore che congiunge la proiezione ortogonale p di y su P_x con y; cfr. la figura nella definizione 8.4.

Analizzando il vettore dei residui si trova spesso che esso può essere decomposto in più componenti; in questo caso si dovrebbe tentare una regressione multipla. Una rappresentazione grafica dei residui permette talvolta di riconoscere fenomeni di periodicità che possono suggerire l'utilizzo di un nuovo modello non lineare.



L'analisi dei residui è particolarmente utile nella ricerca di *errori sistematici* (Doerffel, 171-177, Otto, 207-215).

Bibliografia

14219 **K. Doerffel:** Statistik in der analytischen Chemie. Grundstoffindustrie 1990.

82 **E. Kreyszig:** Statistische Methoden und ihre Anwendungen. Vandenhoeck 1975.

15140 **U. Maniak:** Hydrologie und Wasserwirtschaft. Springer 1997.

14218 **M. Otto:** Chemometrics. VCH 1999.

16226 **F. Paset:** Regressione, correlazione e analisi delle componenti principali. Tesi Univ. Ferrara 2003.