14. Regressione lineare non parametrica

Esistono numerose occasioni nelle quali quello che interessa è ricostruire la relazione di funzione che lega due variabili, la variabile *y* (variabile dipendente, in ordinate) alla variabile indipendente (variabile *x*, in ascisse); se si ritiene che la relazione esistente fra le due variabili possa essere convenientemente descritta mediante una retta, l'equazione di tale retta può essere calcolata mediante la tecnica statistica nota come regressione lineare. Tale denominazione deriva dagli studi sull'ereditarietà condotti da F. Galton sul finire dell'800. Nel corso di questi studi vennero, fra l'altro, registrate le altezze dei componenti di più di 1000 gruppi familiari. Ponendo su un sistema di assi cartesiani in ascisse le altezze dei padri e in ordinate le altezze dei figli, si notò un fatto: sebbene in genere padri più alti avessero figli più alti (come del resto era atteso), padri che erano di 16 centimetri circa più alti della media dei padri, avevano figli che erano solamente 8 centimetri circa più alti della media dei figli. In altre parole sembrava che vi fosse un "tornare indietro", una regressione delle altezze dei figli rispetto a quelle dei padri : e il termine che descriveva il risultato di questa iniziale applicazione, finì con l'essere impiegato per indicare la tecnica statistica, ed è rimasto ancora oggi nell'uso, anche se l'attributo di "regressione" non avrebbe più alcun significato di essere.

Nel caso delle regressione lineare non-parametrica per adattare la retta ai dati sperimentali viene impiegato un metodo che non fa ricorso ad assunti preliminari riguardo la distribuzione dei dati e degli errori ad essi associati.

14.1. Regressione lineare x variabile indipendente

Essendo n il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (y_i - y_j) / (x_i - x_j)$$

con i e j che variano fra 1 e n (con $i < j$).

Allora il coefficiente angolare b_{yx} e l'intercetta a_{yx} dell'equazione della retta di regressione x variabile indipendente

$$y = a_{yx} + b_{yx} \cdot x$$

che meglio approssima i dati, vengono calcolati nel seguente modo:
→ si calcolano i coefficienti angolari delle $n \cdot (n - 1) / 2$ rette che passano per tutte le coppie possibili di punti;
→ si ordinano gli $n \cdot (n - 1) / 2$ coefficienti angolari così calcolati in ordine numerico crescente;
→ si calcola il coefficiente angolare b_{yx} dell'equazione della retta di regressione come mediana degli $N = n \cdot (n - 1) / 2$ valori di cui al punto precedente, cioè come

$$b_{yx} = b_{((N + 1)/2)} \quad \text{se } N \text{ è dispari}$$
$$b_{yx} = (b_{(N/2)} + b_{(N/2+1)}) / 2 \quad \text{se } N \text{ è pari}$$

→ si calcolano allora gli n valori possibili per l'intercetta a come

$$a_i = y_i - b_{yx} \cdot x_i$$

→ si ordinano gli n valori di intercetta così calcolati in ordine numerico crescente;
→ si calcola l'intercetta a_{yx} dell'equazione della retta di regressione come mediana dei valori di cui al punto precedente, cioè come

$$a_{yx} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{yx} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$

14.2. Regressione lineare y variabile indipendente

Si noti che in questo caso inizialmente la y (variabile indipendente) viene posta in ascisse e la x (variabile dipendente) viene posta in ordinate.

Essendo allora n il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (x_i - x_j) / (y_i - y_j)$$

con i e j che variano fra 1 e n (con $i < j$).

Allora il coefficiente angolare b_{xy} e l'intercetta a_{xy} dell'equazione della retta di regressione y variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

che meglio approssima i dati vengono calcolati nel seguente modo:
→ si calcolano i coefficienti angolari delle $n \cdot (n - 1) / 2$ rette che passano per tutte le coppie possibili di punti;
→ si ordinano gli $n \cdot (n - 1) / 2$ coefficienti angolari così calcolati in ordine numerico crescente;
→ si calcola il coefficiente angolare b_{xy} dell'equazione della retta di regressione come mediana degli $N = n \cdot (n - 1) / 2$ valori di cui al punto precedente, cioè come

$$b_{xy} = b_{((N + 1)/2)} \quad \text{se } N \text{ è dispari}$$
$$b_{xy} = (b_{(N/2)} + b_{(N/2+1)}) / 2 \quad \text{se } N \text{ è pari}$$

→ si calcolano allora gli n valori possibili per l'intercetta a come

$$a_i = x_i - b_{xy} \cdot y_i$$

→ si ordinano gli n valori di intercetta così calcolati in ordine numerico crescente;
→ si calcola l'intercetta a_{xy} dell'equazione della retta di regressione come mediana dei valori di cui al punto precedente, cioè come

$$a_{xy} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{xy} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$

Per riportare i dati sullo stesso sistema di coordinate cartesiane utilizzato per la regressione x variabile indipendente, si esplicita l'equazione della retta di regressione y variabile indipendente

$$x = a_{xy} + b_{xy} \cdot y$$

rispetto alla *y*, ottenendo

$$x - a_{xy} = b_{xy} \cdot y$$

e quindi, dividendo entrambi i membri per b_{xy}

$$y = - a_{xy} / b_{xy} \ + 1 / b_{xy} \cdot y$$

Quindi l'intercetta *a* e il coefficiente angolare *b* che consentono di rappresentare la regressione *y* variabile indipendente sullo stesso sistema di coordinate cartesiane della regressione *x* variabile indipendente saranno rispettivamente uguali a

$$a = - a_{xy} / b_{xy}$$
$$b = 1 / b_{xy}$$

14.3. Regressione lineare di Passing e Bablok

Essendo *n* il numero dei punti aventi coordinate note (x_i,y_i), esistono $n \cdot (n - 1) / 2$ modi di connettere due punti qualsiasi con una retta, cioè esistono $n \cdot (n - 1) / 2$ coefficienti angolari

$$b = (y_i - y_j) / (x_i - x_j)$$

con *i* e *j* che variano fra *1* e *n* (con $i < j$).

Nel caso in cui sia

$$x_i = x_j \ \text{ e } \ y_i = y_j$$

il valore del coefficiente angolare non è definito, e quindi viene scartato dai calcoli successivi; ugualmente vengono scartati tutti i valori del coefficiente angolare uguali a -1.

I rimanenti *N* valori vengono allora ordinati in ordine numerico crescente; essendo *K* il numero dei valori del coefficiente angolare inferiori a -1, il coefficiente angolare b_{pb} della retta di regressione calcolata con il metodo di Passing e Bablok

$$y = a_{pb} + b_{pb} \cdot x$$

viene calcolato come

$$b_{pb} = b_{((N + 1)/2 + K)} \quad \text{se } N \text{ è dispari}$$
$$b_{pb} = (b_{(N/2+K)} + b_{(N/2+1+K)}) / 2 \quad \text{se } N \text{ è pari}$$

mentre l'intercetta a_{pb} viene calcolata come mediana degli *n* valori

$$a_i = y_i + b_{pb} \cdot x_i$$

e cioè, nella lista dei valori di *a* così calcolati e ordinati in ordine numerico crescente, come

$$a_{pb} = a_{((n+1)/2)} \quad \text{se } n \text{ è dispari}$$
$$a_{pb} = (a_{(n/2)} + a_{(n/2+1)}) / 2 \quad \text{se } n \text{ è pari}$$