APPROFONDIMENTI – SCHEDA 5

1. La bontà del modello e l'analisi dei residui

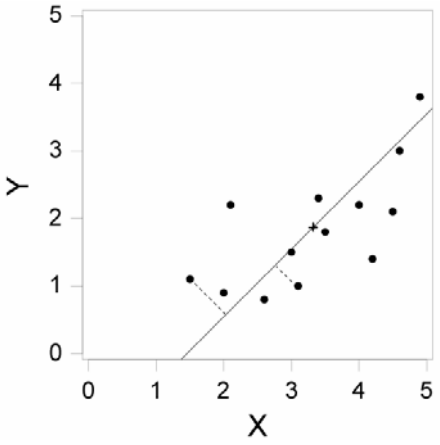
Nella scheda è data molta enfasi allo studio della bontà del modello sopratutto tramite il grafico dei residui. *Questo è di estrema importanza nella pratica statistica*.
Se il grafico dei residui si presenta non sotto forma di nuvola omogenea rispetto alla retta orizzontale bisogna cambiare modello, tipicamente operando delle trasformazioni delle variabili.
Quale trasformazione scegliere non è affatto semplice; tale scelta dipende da un lato dall'abilità dello statistico a maneggiare funzioni matematiche, dall'altro dalla conoscenza del fenomeno che si sta analizzando.
Nell'esempio del flusso in funzione della profondità del corso d'acqua i modelli:

$$\sqrt{\text{flusso}} = a + b \text{ profondita} \qquad \text{e} \qquad \log(\text{flusso}) = a + b \log(\text{profondita})$$

portano grafici dei residui "migliori" ma, soprattutto il secondo, più difficilmente interpretabili dal punto di vista del fenomeno idraulico (almeno a inesperti come noi).

2. Y rispetto a X o X rispetto a Y?

È evidente che a seconda di quale variabile si considera come risposta si ha una retta diversa; ma non è una questione "matematica". Dal punto di vista statistico ha senso al più *una sola* delle due rette: deve essere la conoscenza del fenomeno che si sta analizzando a guidare la scelta (ci si è soffermati su questo punto all'inizio della scheda).
Se non si hanno informazioni sufficienti in tal senso si può considerare la retta passante per il baricentro che minimizza la distanza dei punti dalla retta (nel senso delle proiezioni ortogonali).
Questa è la tecnica dell'Analisi in componenti principali che permettere di "ridurre" la dimensione dello spazio in cui stanno i punti; se abbiamo due variabili X e Y, come nell'esempio a fianco, si passa da due dimensioni (R^2) a una. L'interesse di tale metodologia si ha in realtà quando il numero delle variabili è più elevato.
L'equazione della retta così costruita dipende dalle varianze di X e di Y e dalla covarianza, ma non è ricavabile direttamente dalle equazioni delle due rette di regressione di Y rispetto a X e di X rispetto a Y.



3. Alcune dimostrazioni

a) Per semplificare la dimostrazione della minimizzazione di SS(a, b) in due variabili si può aggiungere la condizione che la somma dei residui sia nulla:

$\sum_{i=1}^{n} (y_i - ax_i - b) = 0$ da cui: $n\bar{y} - na\bar{x} - nb = 0$ e $b = \bar{y} - a\bar{x}$

Questa condizione segue dal porre uguali a 0 le derivate di SS(a,b) in b.

Sostituendo $b = \bar{y} - a\bar{x}$ nella somma dei quadrati dei residui risulta:

$$SS(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} (y_i - ax_i - \bar{y} - a\bar{x})^2 = \sum_{i=1}^{n} ((y_i - \bar{y}) + a(x_i - \bar{x}))^2 =$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 + 2a(y_i - \bar{y})(x_i - \bar{x}) + a^2(x_i - \bar{x})^2 = n\sigma_Y^2 + 2a\,n\,Cov(X,Y) + n\,a^2\,\sigma_X^2$$

In questo modo la somma dei quadrati dei residui è funzione solo di a e SS(a) risulta una parabola con coefficiente del termine di secondo grado positivo. Per ottenere il minimo si può quindi o derivare rispetto ad a oppure – che è la stessa cosa – determinare l'ascissa del vertice. Si ottiene:

$$\hat{a} = \frac{Cov(X,Y)}{\sigma_X^2} \quad \text{da cui} \quad \hat{b} = \bar{y} - \frac{Cov(X,Y)}{\sigma_X^2}\bar{x}$$

b) La somma dei quadrati dei residui del modello vale: $SS(\hat{a}, \hat{b}) = n\,\sigma_Y^2(1- 2\rho^2(X,Y))$.
Infatti:

$$SS(\hat{a}, \hat{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \bar{y} - \frac{Cov(X,Y)}{\sigma_X^2}(x_i - \bar{x}) \right)^2 =$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\frac{Cov(X,Y)}{\sigma_X^2} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \left(\frac{Cov(X,Y)}{\sigma_X^2} \right)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 =$$

$$= n\sigma_Y^2 - 2n\frac{Cov^2(X,Y)}{\sigma_X^2} + n\frac{Cov^2(X,Y)}{\sigma_X^2} = n\sigma_Y^2 - n\frac{Cov^2(X,Y)}{\sigma_X^2} = n\sigma_Y^2\left(1 - \frac{Cov^2(X,Y)}{\sigma_X^2\,\sigma_Y^2}\right)$$

Osserviamo che se σ_Y^2 è circa nulla si ha un "buon modello" nel senso di somma dei quadrati piccola anche se $\rho(X,Y)$ è quasi nulla: è il caso di dati che si posizionano circa su una retta orizzontale.

c) Il valore medio di $\hat{Y}$ è uguale al valore medio di Y, ovvero la somma dei residui è nulla.
Infatti:

$$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{Cov(X,Y)}{\sigma_X^2} x_i + \bar{y} - \frac{Cov(X,Y)}{\sigma_X^2}\bar{x} \right) =$$

$$= \frac{Cov(X,Y)}{\sigma_X^2}\frac{1}{n}\left(\sum_{i=1}^{n} x_i \right) + \bar{y} - \left(\frac{Cov(X,Y)}{\sigma_X^2}\bar{x} \right) = \bar{y}$$

Questo risultato deriva direttamente dalla condizione posta che la somma dei residui sia nulla.

Ma non vale in generale. Infatti se consideriamo il modello:

$$\hat{Y} = a\,X \qquad \text{cioè } \textit{senza l'intercetta}$$

allora si ha:

$$SS(a,b) = \sum_{i=1}^{n} (y_i - ax_i)^2 = \sum_{i=1}^{n} y_i^2 - 2a\sum_{i=1}^{n} x_i y_i + a^2 \sum_{i=1}^{n} x_i^2$$

Il minimo in a si ottiene per $\hat{a} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ e il valore medio di Y è diverso da quello di $\hat{Y}$ e la somma dei residui non è zero.

Questo risultato – cioè che senza la costante la somma dei residui non è nulla – vale anche per un numero di variabili esplicative maggiori di uno.