

11 La regressione (cenni)*

11.1 Introduzione

Molti problemi dell'ingegneria sono collegati alla determinazione delle relazioni tra due o più insiemi di variabili.

Y : **variabile di risposta** (variabile dipendente),
 X_1, \dots, X_r : **variabili di ingresso** o **predittori** (variabili indipendenti).

È ragionevole supporre che per opportune costanti β_k , $k = 0, \dots, r$, possa valere la relazione

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_r X_r.$$

Tuttavia, questo livello di precisione nella pratica non è raggiungibile essendo una relazione deterministica e quindi la relazione viene trasformata in

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_r X_r + \varepsilon \quad (1)$$

dove

ε = **errore casuale**, è una variabile casuale con

$$E(\varepsilon) = 0 \text{ e } \text{var}[\varepsilon] = \sigma^2 \text{ (costante)}$$

σ^2 = **varianza dell'errore**, riflette la variabilità dell'errore sperimentale.

La relazione (1) può essere scritta nella forma equivalente:

$$\mu_{Y|x} = E[Y|x] = \beta_0 + \beta_1 x_1 + \dots + \beta_r x_r \quad (2)$$

con $x = (x_1, \dots, x_r)$ valori deterministici di (X_1, \dots, X_r) .

Se $r \neq 1 \Rightarrow$ regressione lineare multipla

$(\beta_0, \dots, \beta_r) :=$ coefficienti di regressione;

Se $r = 1 \Rightarrow$ regressione lineare semplice e le equazioni (1), (2) assumono la forma:

$$Y = \alpha + \beta X + \varepsilon \quad \text{o} \quad E[Y|x] = \alpha + \beta x.$$

(1), (2) sono dette equazioni di regressione lineare.

Nel caso $r = 1$:

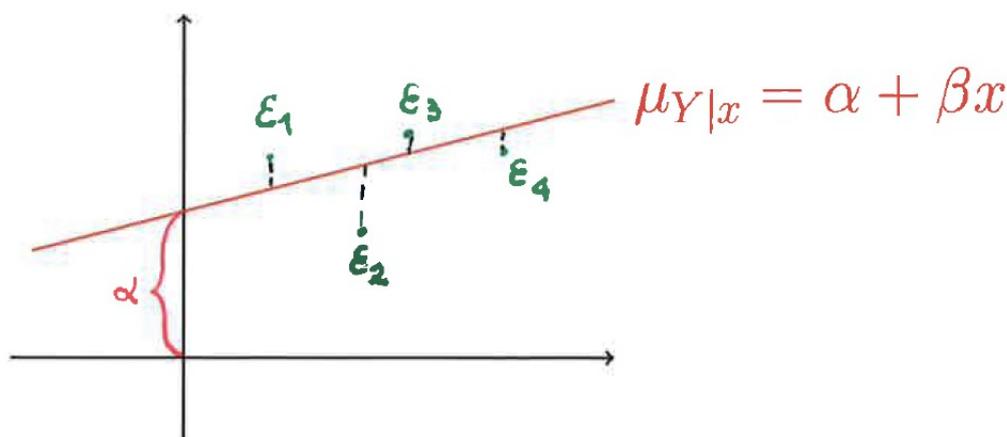


Figura 35: Dati ipotetici (x, y) disposti attorno alla vera retta di regressione per $n = 4$ (α intercetta, β coefficiente angolare).

11.2 Stima dei coefficienti α, β

Problema: stimare i coefficienti di regressione α, β attraverso i dati.

Supponiamo che A, B siano gli stimatori di α, β .

$$\Rightarrow \hat{y} = A + Bx$$

sarà la **retta di regressione stimata**.

Questi stimatori, una volta effettuate le osservazioni, diventano stime.

Per una grande quantità di dati ci si aspetta che la retta di regressione stimata \hat{y} sia più vicina alla "vera" retta di regressione $\mu_{Y|x} = \alpha + \beta x$.

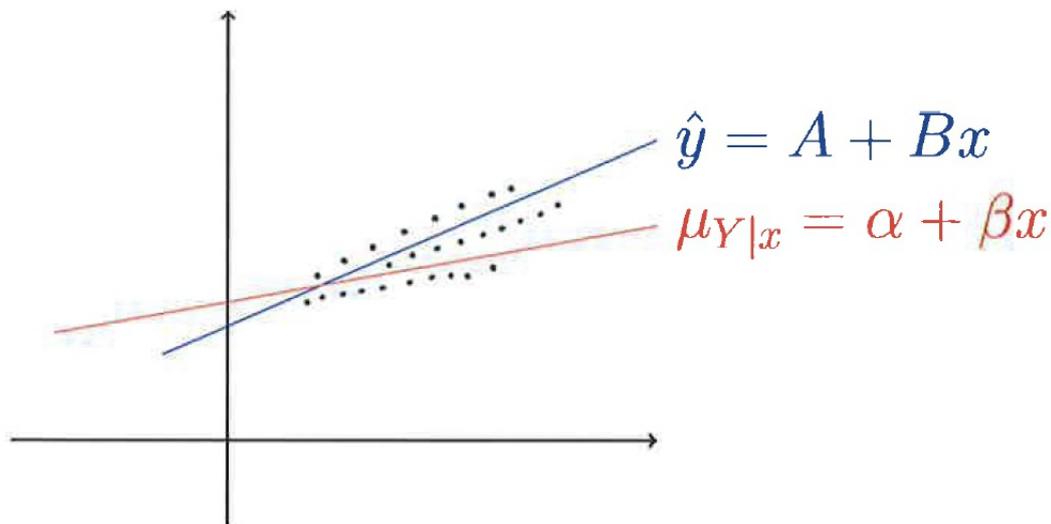


Figura 36: Diagramma a dispersione attorno alle rette di regressione vera e stimata.

Al variare di X_i per $i = 1, \dots, n$, $Y_i = \alpha + \beta X_i + \varepsilon_i$ con $E[Y_i|x_i] = \alpha + \beta x_i$ si ha:

$$y_i - E[Y_i|x_i] = \varepsilon_i.$$

11.3 Metodo di stima

È fondamentale introdurre il concetto di **residuo**. Un residuo è essenzialmente un errore nella stima del modello $\hat{y} = A + Bx$.

Residuo = errore di stima

Dati (x_i, y_i) , $i = 1, \dots, n$ e un modello stimato $\hat{y}_i = A + Bx_i$, il residuo i -esimo e_i è definito da:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

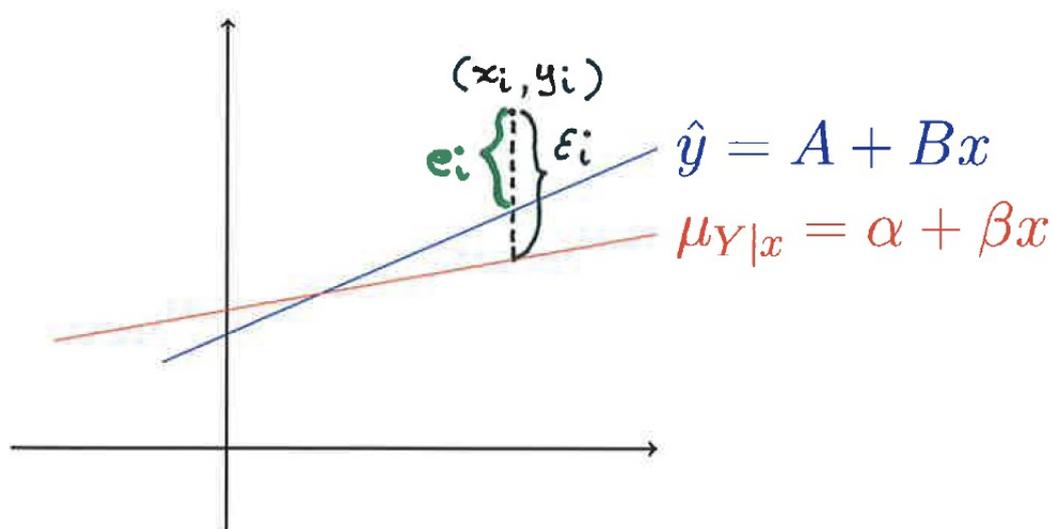
ovvero

$$y_i = A + Bx_i + e_i.$$

Nota bene

e_i : residui (osservati)

ε_i : errori del modello concettuale (non osservati).


 Figura 37: Confronto di ε_i con il residuo e_i .

I valori A, B quali stime di α, β devono essere tali che la **somma dei quadrati dei residui** sia **minima** (somma dei quadrati delle differenze tra predizione e valore osservato).

$$SS_R = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - A - Bx_i)^2$$

11.3.1 Metodo dei minimi quadrati

Derivando SS_R rispetto ad A e a B si ottengono le seguenti relazioni:

$$\begin{cases} \frac{\partial}{\partial A} SS_R = -2 \sum_{i=1}^n (y_i - A - Bx_i) \\ \frac{\partial}{\partial B} SS_R = -2 \sum_{i=1}^n x_i (y_i - A - Bx_i) \end{cases} \quad (3)$$

(A, B) è il punto di stazionarietà del sistema (3) se:

$$\begin{cases} \frac{\partial}{\partial A} SS_R = 0 \\ \frac{\partial}{\partial B} SS_R = 0, \end{cases}$$

da cui si ottengono:

$$\begin{cases} \sum_{i=1}^n y_i = nA + B \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i = A \sum_{i=1}^n x_i + B \sum_{i=1}^n x_i^2 \end{cases} \quad (4)$$

chiamate **equazioni normali**.

$$\text{Poniamo } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \text{ e } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\Rightarrow n\bar{y} = nA + nB\bar{x} \Rightarrow A = \bar{y} - B\bar{x} \quad (\text{stima di } \alpha)$$

$$\begin{aligned}
 \sum_{i=1}^n x_i y_i &= nA\bar{x} + B \sum_{i=1}^n x_i^2 \\
 &= n\bar{x}(\bar{y} - B\bar{x}) + B \sum_{i=1}^n x_i^2 \\
 &= n\bar{x}\bar{y} + B \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right] \\
 \Rightarrow B &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \quad (\text{stima di } \beta)
 \end{aligned}$$

Alternativamente

$$\left\{ \begin{aligned}
 A &= \frac{1}{n} \sum_{i=1}^n y_i - B\bar{x} \\
 B &= \frac{\sum_{i=1}^n x_i y_i - n\bar{x} \cdot \frac{1}{n} \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}
 \end{aligned} \right.$$

Questo metodo è stato utilizzato per stimare una retta (\hat{y}) la cui peculiarità è quella di essere "vicina" alle coppie (x_i, y_i) di dati osservati.

Osservazione: La stima ai minimi quadrati determina una retta che minimizza la somma dei quadrati degli **scostamenti verticali** dei punti rispetto alla retta.

11.4 Proprietà degli stimatori

A, B sono stimatori **corretti** di α, β .

Infatti:

$$\begin{aligned}
 E[B] &= \frac{\sum_{i=1}^n (x_i - \bar{x}) E[y_i]}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}
 \end{aligned}$$

$$\text{ma } \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} \equiv 0 \Rightarrow$$

$$E[B] = \frac{\beta \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right)}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \equiv \beta$$

$$\begin{aligned}
 E[A] &= \frac{1}{n} \sum_{i=1}^n E[y_i] - E[B]\bar{x} \\
 &= \frac{1}{n} \sum_{i=1}^n (\alpha + \beta x_i) - \beta \bar{x} \\
 &= \frac{1}{n} \sum_{i=1}^n \alpha + \beta \frac{1}{n} \sum_{i=1}^n x_i - \beta \bar{x} \\
 &= \frac{1}{n} n\alpha \equiv \alpha
 \end{aligned}$$

11.5 Stima di σ^2

La quantità SS_R può essere usata per stimare la varianza σ^2 degli errori casuali.

Una stima **corretta** di σ^2 è data da:

$$\tilde{S}^2 = \frac{SS_R}{n-2}$$

cioè si può dimostrare che

$$E[\tilde{S}^2] \equiv \sigma^2$$

Abbreviazioni:

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$$

$$SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

allora si ha:

$$\left\{ \begin{array}{l} A = \bar{y} - B\bar{x} \\ B = \frac{S_{xy}}{S_{xx}} \\ \tilde{S}^2 = \frac{SS_R}{n-2} \\ SS_R = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}} \\ SS_{reg} = \frac{S_{xy}^2}{S_{xx}} \end{array} \right.$$

Le quantità SS_{reg} e SS_R prendono il nome di *varianza spiegata* e *varianza residua*, mentre la quantità S_{yy} corrisponde alla *varianza totale* delle osservazioni y_i .

Vale la seguente relazione:

$$S_{yy} = SS_{reg} + SS_R.$$

Si può pensare che un modello sia fedele se la varianza totale delle y è in gran parte spiegata da SS_{reg} . Pertanto si definisce il **coefficiente di determinazione**:

$$R^2 = 1 - \frac{SS_R}{S_{yy}}$$

che rappresenta il rapporto tra la varianza spiegata e la varianza totale: più R^2 si avvicina ad 1, più il modello è "buono". Si può anche scrivere:

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

Osservazione: Nell'ipotesi che l'errore casuale ε sia una variabile casuale normale, allora gli stimatori A e B di α e β sono normalmente distribuiti e per essi si possono costruire intervalli di confidenza e test.

11.6 Regressione logistica (cenni)

La **regressione logistica** (o anche **modello logit**) è un modello di regressione **non lineare** in cui si vuole modellare la relazione tra una variabile dipendente qualitativa Y di tipo dicotomico (cioè 0/1, SI'/NO) ed una o più variabili indipendenti X_1, \dots, X_r che si ritiene possano influenzarla.

L'obiettivo è stabilire la probabilità con cui un insieme di dati possa generare uno o l'altro valore di Y (cioè $Y = 0$ o $Y = 1$). Nel caso più semplice di un solo valore indipendente $X = x$, il modello lineare non risulta appropriato poichè:

$$P[Y|x] = \alpha + \beta x$$

e il secondo membro varia nell'intervallo $(-\infty, +\infty)$, mentre il primo membro, per definizione di probabilità, varia in $[0, 1]$.

Si usa perciò come modello

$$p = P[Y|x] = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

dove il secondo membro ora varia in $[0, 1]$.

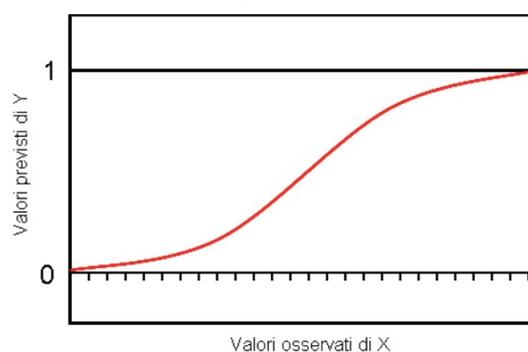


Figura 38: modello logistico

Si considera poi la funzione

$$\text{logit } P[Y|x] = \ln\left(\frac{p}{1-p}\right) = \ln e^{\alpha+\beta x} = \alpha + \beta x$$

che ha il vantaggio di essere più facile da trattare, poichè è lineare.