REGRESSIONE

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

Y: variabile di risposta (v. dipendente)

$x_1, \dots, x_r$: variabili di ingresso (variabili indipendenti)

È ragionevole supporre che per opportune costanti β_k $k = 1, \dots, r$ possa valere

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_r x_r .$$

Tuttavia questo livello di precisione nella pratica non è raggiungibile (è una relazione deterministica)

La relazione viene trasformata in:

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_r x_r + \varepsilon \qquad (1)$$

ε = errore casuale, è una v.c. con

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

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

La relazione (1) può essere scritta nella forma equivalente

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

con $x = (x_1, \dots, x_r)$.

Se $r \neq 1 \Rightarrow$ REGRESSIONE LINEARE MULTIPLA

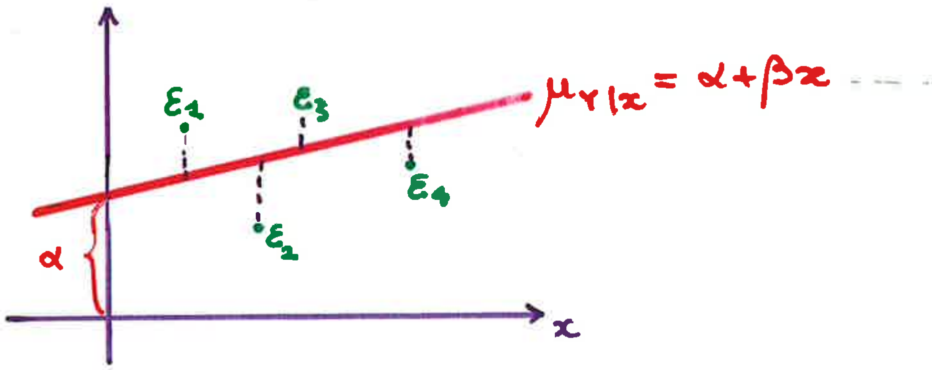
$(\beta_0, \dots, \beta_r) :=$ coefficienti di regressione.

Se $r = 1 \Rightarrow$ REGRESSIONE LINEARE SEMPLICE e (1),(2):

$$Y = \alpha + \beta x + \varepsilon \qquad o \qquad E[Y|x] = \alpha + \beta x$$

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

Nel caso $r = 1$



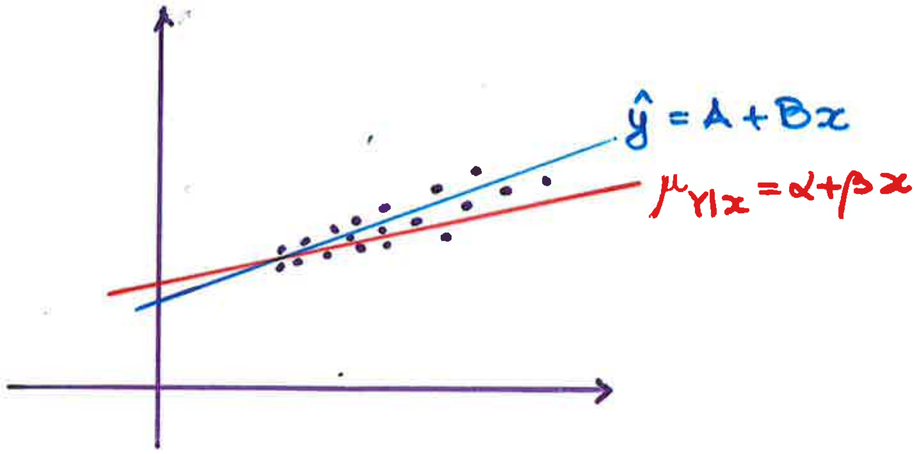
STIMA DEI COEFFICIENTI α, β

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

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

$\Rightarrow \quad \hat{y} = A + Bx$ retta di regressione stimata

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 $\mu_{Y|x} = \alpha + \beta x$.



Al variare di x_i per $i = 1, \dots, m$ $\quad Y_i = \alpha + \beta x_i + \varepsilon_i$
con $E[Y_i] = \alpha + \beta x_i$ si ha

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

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

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

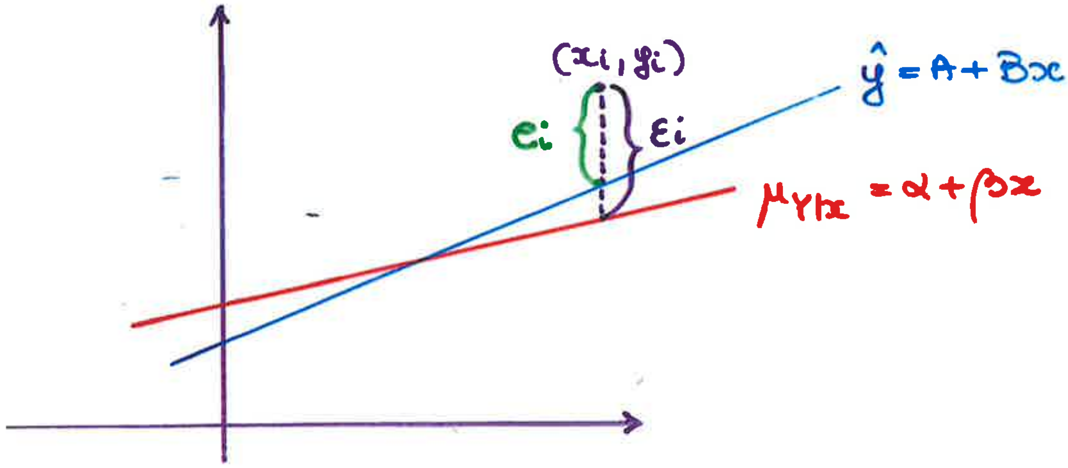
ovvero

$$y_i = A + Bx_i + e_i$$

N.B.:

e_i : residui (osservati)

ε_i : errori del modello concettuale (non osservati).



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).

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

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) pti di stazionarietà del sistema (3).

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

da cui si ottengono:

$$\begin{cases} \sum_{i=1}^{m} y_i = mA + B\sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i y_i = A\sum_{i=1}^{m} x_i + B\sum_{i=1}^{m} x_i^2 \end{cases}$$

EQUAZIONI NORMALI

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

$\Rightarrow$

$m\bar{y} = mA + mB\bar{x} \Rightarrow$ $A = \bar{y} - B\bar{x}$ stimatore di α.

$$\sum_{i=1}^{m} x_i y_i = mA\bar{x} + B\sum_{i=1}^{m} x_i^2$$
$$= m\bar{x}(\bar{y} - B\bar{x}) + B\sum_{i=1}^{m} x_i^2$$
$$= m\bar{x}\bar{y} + B\left[\sum_{i=1}^{m} x_i^2 - m\bar{x}^2\right]$$

$\Rightarrow$ $B = \dfrac{\sum_{i=1}^{m} x_i y_i - m\bar{x}\bar{y}}{\sum_{i=1}^{m} x_i^2 - m\bar{x}^2}$ stimatore di β.

Alternativamente

$$\begin{cases} A = \frac{1}{m}\sum_{i=1}^{m} y_i - B\bar{x} \\ B = \dfrac{\sum_{i=1}^{m} x_i y_i - m\bar{x}\cdot\frac{1}{m}\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} x_i^2 - m\bar{x}^2} = \dfrac{\sum_{i=1}^{m} (x_i - \bar{x}) y_i}{\sum_{i=1}^{m} x_i^2 - m\bar{x}^2} \end{cases}$$

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.

PROPRIETÀ DEGLI STIMATORI

A, B sono stimatori CORRETTI di α, β.

Infatti:

$$E[B] = \frac{\sum_{i=1}^{m} (x_i - \bar{x}) E[y_i]}{\sum_{i=1}^{m} x_i^2 - m\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{m} (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{m} x_i^2 - m\bar{x}^2}$$

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

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

$$E[A] = \frac{1}{m} \sum_{i=1}^{m} E[y_i] - E[B]\bar{x}$$

$$= \frac{1}{m} \sum_{i=1}^{m} (\alpha + \beta x_i) - \beta\bar{x}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \alpha + \beta \cdot \underbrace{\frac{1}{m} \sum_{i=1}^{m} x_i}_{\bar{x}} - \beta\bar{x}$$

$$= \frac{1}{m} \cdot m\alpha \equiv \alpha$$

Si può provare che

$$var[B] = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad , \quad var[A] = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

Ipotesi: le variabili y_i sono
- indipendenti
- normali
- con medie $E[y_i] = \alpha + \beta x_i$
- con varianza σ^2

$\Rightarrow$ A, B sono v.c. NORMALI (perchè combinazioni lineari di v.c. normali).

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:

$$s^2 = \frac{SS_R}{n-2}$$

cioè si può dimostrare che

$$E[s^2] = \sigma^2.$$

Notazioni sintetiche :

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

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

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

allora si ha:

$$\begin{cases} A = \bar{y} - B\bar{x} \\ B = \dfrac{S_{xy}}{S_{xx}} \\ SS_R = \dfrac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}} \end{cases}$$