REGRESSIONE LINEARE SEMPLICE

Un modello di regressione lineare semplice spiega una variabile risposta (o dipendente) y in termini di una variabile X detta var. indipend. o regressore o predittore.
La più semplice relazione tra y e X è <u>lineare</u>:

$$y = \alpha + \beta X \qquad (1)$$

Ovviamente, noto X, non possiamo prevedere esattamente y, ma assumiamo (1) vera a meno di un errore ε.

Che caratteristiche ha ε?
Non essendo osservabile, la consideriamo una v.a. con determinate caratteristiche:

$$\varepsilon \sim N(0, \sigma^2) \quad iid$$ (indipendenti e identicamente distribuiti)

Quindi di fatto, y è v.a:

$$y = \alpha + \beta X + \varepsilon$$

$$E[y] = \alpha + \beta E[x] + E[\varepsilon] = \alpha + \beta E[x]$$

$$Var[y] = 0 + \beta^2 Var[x] + \sigma^2 = \beta^2 Var[x] + \sigma^2$$

Possiamo calcolarli tutto perché X e Y sono osservate.
In particolare X è deterministica

α = intercetta = $E[y]$ quando non ho x ($x=0$)

β = coeff. angolare = rappresenta l'incremento di y al crescere di un'unità di x

Nel modello, avendo n osservazioni di Y e X accoppiate: $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$

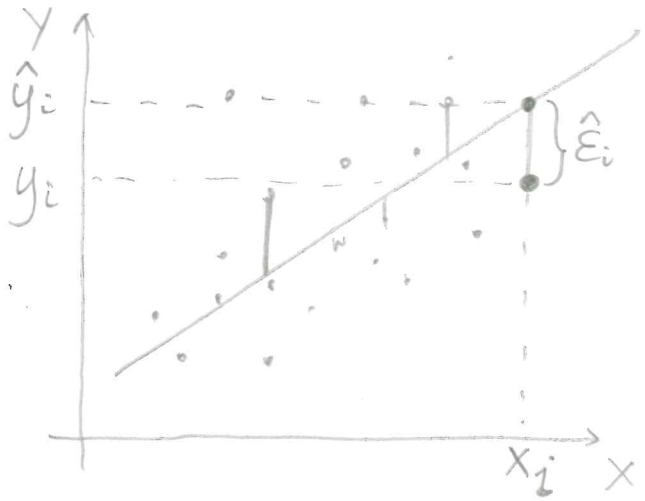
Cerchiamo α e β t.c. minimizziamo la somma di $|y_i - (\alpha + \beta x_i) = \varepsilon_i|$.

Ma visto che alcuni potrebbero essere negativi, minimizziamo $\sum_{i=1}^{n} \varepsilon_i^2$ → troviamo $\hat{\alpha}$ e $\hat{\beta}$, le stime dei 2 parametri

$$\Rightarrow \quad \hat{y} = \hat{\alpha} + \hat{\beta} x$$

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{\varepsilon}_i \Rightarrow \hat{y}_i - y_i = \hat{\varepsilon}_i$$

METODO DEI MINIMI QUADRATI ↑



- DIAGNOSTICA DEL MODELLO

$\varepsilon \sim N(0, \sigma^2)$ $\quad \varepsilon_i$ iid → dipende dai dati, che si assumono indipendenti

$(x_i, y_i) \perp\!\!\!\perp (x_j, y_j) \quad \forall i \neq j$

- istogramma
- qqplot
- costante $\forall i$ → omoschedasticità
  - grafico dei residui (standardizzati)
- viene per costruzione

- BONTÀ DEL MODELLO

R^2, p-values regressori

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Cosa vuol dire che gli ε_i sono i.i.d.?

Def iid: $P(\varepsilon_1 \leq e_1, \varepsilon_2 \leq e_2, \dots \varepsilon_N \leq e_N) = \prod_{i=1}^{N} P(\varepsilon_i \leq e_i)$