MODELLO DI REGRESSIONE LINEARE: principali simboli e formule.

Nota Bene: dopo la breve introduzione iniziale in questa pagina, la pagina seguente è una rassegna schematica dei principali simboli e formule visti a lezione ed applicati ad esercitazione anche con la lettura del tabulato Excel.

Introduzione.
Fissato il prezzo di vendita x_t di un prodotto per tutto il mese corrente, il valore del fatturato y_t alla fine dello stesso mese **non è predeterminato** in funzione del prezzo x_t **tramite una funzione matematica** quale p. es., per semplicità, una retta o simili

$$y_t = \beta_0 + \beta_1 x_t \quad (1)$$

Ciò perché il valore del fatturato y_t alla fine mese corrente è in realtà **uno dei valori possibili di una v.a.** Y_t. Infatti il valore y_t osservato alla fine del mese è influenzato da tante altre variabili i cui valori **non possono essere né prefissati né conosciuti** dall'azienda venditrice del prodotto. Quindi l'azienda **non può conoscere con certezza** il valore fatturato y_t del mese **prima** che sia arrivata la fine del mese stesso, ossia il fatturato è una **v.a**. Y_t.

Che il valore del fatturato y_t alla fine mese corrente sia in effetti uno dei valori possibili di una v.a. Y_t risulta anche dal fatto che se si rappresentano su un sistema di assi cartesiani le coppie (prezzo, fatturato) ovvero le coppie (x_t, y_t) ($t = 1, 2, \ldots n$) relative a n mesi passati, si vede che, in generale, tali coppie si dispongono nel piano cartesiano secondo **un grafico** che **non soddisfa la definizione matematica di funzione**, si ha bensì un **diagramma di dispersione** (vedere i grafici mostrati a lezione dai docenti, o le fotocopie dei lucidi presso Yellow Print, od anche la figura nel libro di Ross a pag. 59).

Ciò premesso, **con il modello di regressione lineare** si vogliono raggiungere i due obiettivi seguenti:
(A) **non abbandonare del tutto** la comoda ed utile nozione di funzione matematica quale è, per semplicità, la funzione $\beta_0 + \beta_1 x_t$ nella (1) di cui sopra, e si vuole però anche:
(B) **tenere conto che il fatturato è in realtà una v.a**. Y_t il cui valore y_t al tempo t non è determinato come funzione matematica $\beta_0 + \beta_1 x_t$ (o altra) del prezzo x_t fissato dall'azienda.

Si considera allora un modello matematico alternativo alla funzione $y_t = \beta_0 + \beta_1 x_t$ (1) di cui sopra, ovvero si considera il modello matematico-probabilistico, detto "modello di regressione lineare", dato da

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad (2)$$

dove ε_t, e quindi Y_t, sono v.a. che soddisfano certe ipotesi che implicano l'importante conseguenza che

$$E(Y_t) = \beta_0 + \beta_1 x_t \quad (3)$$

Con le formule (2) e (3) si raggiungono i due obiettivi (A) e (B) di cui sopra. Infatti, da una parte, con la formula (2), **né** il fatturato osservato y_t **né** la v.a. fatturato Y_t stessa sono determinati come funzione matematica $\beta_0 + \beta_1 x_t$ del prezzo x_t fissato dall'azienda (ciò è dovuto alla presenza della v.a. ε_t).
Dall'altra parte però, con la formula (3), si ha che **il fatturato medio atteso** $E(Y_t)$, ovvero **il fatturato in media, è determinato** come **funzione matematica** $\beta_0 + \beta_1 x_t$ del prezzo fissato x_t dall'azienda (ciò è ancora dovuto alla presenza della v.a. ε_t che, come specificato più avanti, è tale che $E(\varepsilon_t) = 0$).

Del modello di regressione lineare si danno in quanto segue: l'equazione, le ipotesi, le conseguenze delle ipotesi, le stime puntuali dei parametri e la retta stimata, la scomposizione della varianza totale ed il coefficiente di determinazione.

EQUAZIONE DEL MODELLO DI REGRESSIONE LINEARE

$$Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad \left(t = 1, 2, \ldots n\right), \text{ dove}$$

x_t — è una variabile non aleatoria in quanto i suoi valori possono essere prefissati per ogni t

β_0 e β_1 — sono due costanti o parametri non noti da stimare sulla base di n coppie di valori osservati $\left(x_t, y_t\right)$ ($t = 1, 2, \ldots n$)

ε_t — è una variabile aleatoria per ogni $t = 1, 2, \ldots n$ detta v.a. "errore" o "residuo" che soddisfa le ipotesi del modello di regressione lineare

IPOTESI DEL MODELLO DI REGRESSIONE LINEARE

$E\left(\varepsilon_t\right) = 0$

$V\left(\varepsilon_t\right) = \sigma^2$ (parametro non noto da stimare, costante per ogni $t = 1, 2, \ldots n$)

$Cov\left(\varepsilon_t, \varepsilon_s\right) = 0 \quad \left(\forall\left(t, s\right), t \neq s\right)$ (ipotesi di indipendenza correlativa, o di non correlazione, delle v.a. ε_t che è verificata se, p. es., si assume che le v.a. ε_t siano stocasticamente indipendenti)

$\varepsilon_t \sim N\left(0; \sigma^2\right) \quad \left(t = 1, 2, \ldots n\right)$ con le v.a. ε_t stocasticamente indipendenti

CONSEGUENZE DELLE IPOTESI DEL MODELLO DI REGRESSIONE LINEARE

$E\left(\varepsilon_t\right) = 0$ e $Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad \Rightarrow \quad E\left(Y_t\right) = \beta_0 + \beta_1 x_t$

$V\left(\varepsilon_t\right) = \sigma^2$ e $Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \quad \Rightarrow \quad V\left(Y_t\right) = \sigma^2$

$\varepsilon_t \sim N\left(0; \sigma^2\right)$ (stoc. indip. $t = 1, 2, \ldots n$) e $Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$

$\Rightarrow \quad Y_t \sim N\left(\beta_0 + \beta_1 x_t; \sigma^2\right)$ (stoc. indip. $t = 1, 2, \ldots n$)

STIME PUNTUALI DEI PARAMETRI e RETTA STIMATA

$\hat{\beta}_1 = \dfrac{\sigma_{XY}}{s_X^2}$ (dove σ_{XY} e s_X^2 sono calcolati con le coppie osservate $\left(x_t, y_t\right)$ $t = 1, 2, \ldots n$)

$\hat{\beta}_0 = \overline{y}_n - \hat{\beta}_1 \overline{x}_n$ (dove $\overline{y}_n$ e $\overline{x}_n$ sono calcolati con le coppie osservate $\left(x_t, y_t\right)$ $t = 1, 2, \ldots n$)

$\hat{y}_t = \hat{E}\left(Y_t\right) = \hat{\beta}_0 + \hat{\beta}_1 x_t \quad (t = 1, 2, \ldots n)$

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}{n - 2}$$

SCOMPOSIZIONE DELLA VARANZA TOTALE E COEFFICIENTE DI DETERMINAZIONE R^2

$\sum_{t=1}^{n}\left(y_t - \overline{y}_n\right)^2 =$ varianza totale

$\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2 =$ varianza residua (o varianza non spiegata)

$\sum_{t=1}^{n}\left(\hat{y}_t - \overline{y}_n\right)^2 =$ varianza spiegata

$$\sum_{t=1}^{n}\left(y_t - \overline{y}_n\right)^2 = \sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2 + \sum_{t=1}^{n}\left(\hat{y}_t - \overline{y}_n\right)^2$$

$$R^2 = \frac{\sum_{t=1}^{n}\left(\hat{y}_t - \overline{y}_n\right)^2}{\sum_{t=1}^{n}\left(y_t - \overline{y}_n\right)^2} \quad \left(0 \leq R^2 = \rho_{XY}^2 \leq 1\right)$$