6 # Modello di regressione lineare semplice

Considera una variabile dipendente Y (= VARIABILE RISPOSTA) in funzione lineare di una sola variabile indipendente X (= VARIABILE ESPLICATIVA)

$$g(x) = \beta_0 + \beta_1 x$$

β_0: INTERCETTA; β_1: COEFFICIENTE ANGOLARE

β_0, β_1: coefficienti di regressione

RELAZIONE STATISTICA

$$Y = g(x) + \varepsilon$$

ε ⟶ TERMINE D'ERRORE = rappresenta il contributo di tutti gli altri fattori ($\neq x$) non misurabili che influenzano Y

↓ è una VARIABILE CASUALE

↓ di conseguenza anche Y è una VARIABILE CASUALE

⇓

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

β_0, β_1: non conosciamo questi parametri quindi dobbiamo STIMARLI

↓ per farlo dobbiamo fare delle ASSUNZIONI

RESIDUI

$$e_i = y_i - \hat{y}_i$$

y_i: valore reale; $\hat{y}_i$: valore stimato dal modello

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i$$

⟶ PROPRIETÀ

$$\sum e_i = 0$$

↓ la somma di tutti i residui deve fare zero!!

ASSUNZIONI

1. La funzione di regressione deve essere lineare

$$y_i = \beta_0 + \beta_1 X + \varepsilon_i \qquad \forall i = 1, 2, \dots, n$$

2. Le ε_i sono V.C. INDIPENDENTI TRA LORO e indipend. dal valore di x_i

$E(\varepsilon_i) = 0$ → in media gli errori commessi si compensano

$V(\varepsilon_i) = \sigma^2$ → varianza costante OMOSCHEDASTICITÀ

3. I valori x_i della variabile esplicativa X sono noti senza errori.

METODO DEI MINIMI QUADRATI

→ metodo di stima dei coeff. di regressione β_0 e β_1 che consiste nel cercare le stime che MINIMIZZANO la FUNZIONE DI PERDITA

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{CODEVIANZA}(X,Y)}{\text{DEVIANZA}} = \frac{\sigma_{XY}}{\sigma^2_X} = \frac{r_{XY}\,\sigma_Y}{\sigma_X}$$

(è seconda degli indici della retta)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Sono stimatori CORRETTI $E(\hat{\beta}_0) = \beta_0 \quad E(\hat{\beta}_1) = \beta_1$ ed EFFICIENTI (varianza minima)

$$V(\hat{\beta}_0) = \sigma^2_\varepsilon \left[\frac{1}{n} + \frac{\bar{x}}{\sum (x_i - \bar{x})}\right] \qquad V(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

PROPRIETÀ

1. La retta di regressione stimata passa sempre per il BARICENTRO $(\bar{x}; \bar{y})$

2. $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$

3. $E(Y_i \mid X = x_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

4. $SQT = SQR + SQE$ → DECOMPOSIZIONE DELLA DEVIANZA TOTALE

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{e}_i^2$$

- $\sum_{i=1}^n (y_i - \bar{y})^2$: somma totale dei quadrati → $DEV(Y) = \sum y^2 - n\bar{y}^2$
- $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$: somma dei quadrati della regressione (DEVIANZA SPIEGATA) (VARIABILITÀ DEL MODELLO)
- $\sum_{i=1}^n \hat{e}_i^2$: somma dei quadrati degli errori (DEVIANZA RESIDUA) (VARIABILITÀ DEGLI ERRORI!!!)

COEFFICIENTE DI DETERMINAZIONE → MISURA LA BONTÀ DI ADATTAMENTO DI REGRESSIONI AI DATI

$$R^2_{XY} = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

(SQR significa che del totale della variabilità)

$$R^2_{XY} = (\rho_{XY})^2$$ → visto che ci stiamo riferendo ad un modello semplice R^2_{XY} può essere misurato anche come quadrato del coeff. di correlazione lineare

$$\rho_{XY} = \left(\frac{COV(X,Y)}{\sigma_X \cdot \sigma_Y}\right)^2 =$$

$$R^2 = \left[\frac{n\ COVAR(X,Y)}{\sqrt{n}\sigma_X \cdot \sqrt{n}\sigma_Y}\right]^2 = \left[\frac{CODEV(X,Y)}{\sqrt{DEV(X)} \cdot \sqrt{DEV(Y)}}\right]^2$$

campo di variazione $[0;1]$

- 0 → relaz. lineare tra X e Y (SQT = 0)
- 1 → relaz. perfetta (SQE = 0)

PROPRIETÀ DELLO STIMATORE $\hat{Y}_i$

$$E(\hat{Y}_i) = \beta_0 + \beta_1 x_i \longrightarrow \text{STIMATORE CORRETTO}$$

$$V(\hat{Y}_i) = \sigma^2_\varepsilon \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}\right]$$

ERRORE STANDARD DI REGRESSIONE → SCARTO QUADRATICO MEDIO DELLA STIMA

$$S = \sqrt{S^2} = \sqrt{\frac{\sum \hat{e}_i^2}{n-2}} \longrightarrow$$ Stima corretta della varianza dei residui σ^2_ε

ASSUNZIONE DI NORMALITÀ DEGLI ERRORI

$$\varepsilon_i \sim N(0; \sigma^2)$$

↓

I SINGOLI ERRORI SONO INDIPENDENTI TRA LORO

$$E(\varepsilon_i \varepsilon_j) = 0 \quad \forall i \neq j$$

di conseguenza l'intero modello di regressione si distribuisce normalmente

$$Y_i \sim N(\beta_0 + \beta_1 x_i; \sigma^2)$$

DISTRIBUZIONE DEGLI STIMATORI $\hat{\beta_0}$ e $\hat{\beta_1}$

1. $\hat{\beta_0} \sim N(\beta_0; \sigma^2_{\hat{\beta_0}})$ $\quad \hat{\beta_1} \sim N(\beta_1; \sigma^2_{\hat{\beta_1}})$

insieme assumono la forma di DISTRIBUZIONE NORMALE BIVARIATA

2. $z_0 = \dfrac{\hat{\beta_0} - \beta_0}{S(\hat{\beta_0})} \sim t_{n-2}$ $\quad z_1 = \dfrac{\hat{\beta_1} - \beta_1}{S(\hat{\beta_1})} \sim t_{n-2}$

$S(\hat{\beta_0})$ ↓ stima della SD

quando $n \to \infty$ la distr. converge a una $N(0,1)$

INTERVALLI DI CONFIDENZA PER I PARAMETRI

$$P\left(\hat{\beta_0} - t_{\frac{\alpha}{2}} \cdot S(\hat{\beta_0}) \leq \beta_0 \leq \hat{\beta_0} + t_{\frac{\alpha}{2}} \cdot S(\hat{\beta_0})\right) = 1 - \alpha$$

$$P\left(\hat{\beta_1} - t_{\frac{\alpha}{2}} \cdot S(\hat{\beta_1}) \leq \beta_1 \leq \hat{\beta_1} + t_{\frac{\alpha}{2}} \cdot S(\hat{\beta_1})\right) = 1 - \alpha$$

VERIFICA D'IPOTESI PER I PARAMETRI

se vogliamo verificare $H_0: \beta_1 = b_1$

sotto H_0: $\dfrac{\hat{\beta_1} - b_1}{S(\hat{\beta_1})} \sim t_{n-2}$

TAVOLA DELL'ANALISI DELLA VARIANZA (ANOVA)

la ritrovo uguale in JMP!

SORGENTE DI VARIAZIONE	SOMMA DEI QUADRATI	GDL	MEDIA DEI QUADRATI	F
REGRESSIONE	SQR	1	$MQR = \frac{SQR}{1}$	$F = \frac{MQR}{MQE}$ → il suo P-value è uguale a quello per β_1 (1)
RESIDUO	SQE	$n-2$	$MQE = \frac{SQE}{n-2}$	

statistica test dell'ANOVA ↓ si distribuisce come una F di Fisher

SQT $n-1$ gdl.

rapporto tra SQ in JMP?

nel modello di regressione lineare semplice $F = [\text{statistica-test per la verifica di ipotesi su } \beta_1]^2$