

NOTE DALLE LEZIONI DI STATISTICA MEDICA ED ESERCIZI

**METODI PER LO STUDIO DEL LEGAME
TRA VARIABILI IN UN RAPPORTO
DI CAUSA ED EFFETTO**

I MODELLI DI REGRESSIONE

IL PROBLEMA

Si dispone dei valori relativi alla statura in centimetri (x) ed al peso corporeo in Kg (y) di un campione di individui maschi adulti, della stessa classe di eta'

PESO (y)	60	57	71	66	65	78	82	78	62	70
STATURA(X)	171	169	181	173	178	180	185	183	170	174

REGRESSIONE

➤ E' la tecnica più adatta quando l'obiettivo principale consiste nello sviluppare un "modello predittivo", cioè uno strumento che consenta di predire il livello di Y per un dato valore di X .

➤ studia una relazione di causa ed effetto, come varia una variabile

☞ detta dipendente Y

al variare di un'altra variabile

☞ detta indipendente X

➤ La variabile X :

☞ è una variabile non casuale

☞ è affetta da errore trascurabile

per ogni valore di X esiste una sottopopolazione di valori di Y che seguono una distribuzione di Gauss.

Il più semplice modello di relazione che si può indagare è rappresentato da una retta la cui espressione matematica è:

$$y = a + b x$$

dato che per ogni x_i abbiamo una popolazione di valori y_i in statistica l'espressione della retta diventa:

$$y_i = a + b x_i + e_i$$

e_i è l'**errore** di misura legata alla variabilità dei soggetti sotto osservazione

a è l'intercetta

b è il **coefficiente di regressione** (coefficiente angolare o pendenza della retta) ed esprime **quanto aumenta la variabile dipendente al variare unitario della variabile indipendente.**

Le stime dei parametri a e b si ottengono con il :

METODO DEI MINIMI QUADRATI

che consiste nell'individuare la retta che renda minima la somma delle distanze al quadrato di ciascun punto y_i dai punti della retta di regressione stessa.

$$\sum [y_i - (a + bx_i)]^2 = \sum e_i^2$$

La stima dei due parametri della regressione sono:

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

l'equazione della retta diventa pertanto:

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

Per effettuare i conti è conveniente utilizzare le seguenti formula:

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} =$$
$$= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right]}$$

Per risolvere il problema si consiglia di costruire una tabella che contenga le quantità:

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	171	60
...
10	174	70
Tot

**ADESSO EFFETTUATE I CALCOLI
E POI PROSEGUITE**

Per risolvere il problema conviene preparare la seguente tabella dei dati

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	171	60	10260	29241	3600
2	169	57	9633	28561	3249
3	181	71	12851	32761	5041
4	173	66	11418	29929	4356
5	178	65	11570	31684	4225
6	180	78	14040	32400	6084
7	185	82	15170	34225	6724
8	183	78	14274	33489	6084
9	170	62	10540	28900	3844
10	174	70	12180	30276	4900
Tot	1764	689	121936	311466	48107

effettuando i conti si ha:

$$\bar{x} = 1764 / 10 = 176.4$$
$$\bar{y} = 689 / 10 = 68.9$$

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right]} =$$

$$= \frac{121936 - (1764 \times 689) / 10}{311466 - (1764)^2 / 10} = \frac{396.4}{296.4} = 1.3374$$

$$\hat{a} = \bar{y} - \hat{b} \bar{x} = 68.9 - 1.3374 \times 176.4 = -167.0136$$

VERIFICA DELLE IPOTESI SU b

ANALISI DELLA VARIANZA DELLA REGRESSIONE.

$$H_0: b = 0$$

$$H_1: b \neq 0$$

L'obiettivo è di analizzare se le variazioni della Y associate alla X siano maggiori di quelle dovute al caso.

Se non rifiutiamo H_0 :

☞ forse c'è una relazione tra X ed Y ma non è possibile stimare Y conoscendo i valori di X ;

☞ forse c'è un legame tra X ed Y ma non è descritto da una retta.

Si osservi che

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Per tutti gli scostamenti si ottiene

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Devianza Totale = Devianza di Regressione + Devianza residua

G.L.

N-1

1

N-2

PER EFFETTUARE I CALCOLI CONVIENE UTILIZZARE LE SEGUENTI FORMULE

$$\text{DEVIANZA TOTALE} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n$$

$$\text{DEV REGRESS.} = \sum (\hat{y}_i - \bar{y})^2 = b^2 \cdot \sum (x_i - \bar{x})^2 =$$

$$= \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum (x_i - \bar{x})^2}$$

$$\text{DEV RESIDUA} = \text{TOTALE} - \text{REGRESS}$$

**UTILIZZANDO
LE FORMULE PRECEDENTI
EFFETTUATE I CALCOLI**

Per il nostro esempio

$$\begin{aligned}\mathbf{DEV. TOTALE} &= \sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n = \\ &= 48107 - 689^2 / 10 = 634.9\end{aligned}$$

$$\begin{aligned}\mathbf{DEV. REGRESSIONE} &= \sum (\hat{y}_i - \bar{y})^2 = b^2 \cdot \sum (x_i - \bar{x})^2 = \\ &= \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum (x_i - \bar{x})^2} = \frac{396.4^2}{296.4} = 530.14\end{aligned}$$

$$\mathbf{DEV. RESIDUA} = \mathbf{TOTALE} - \mathbf{REGRES} = 634.9 - 530.14 = 104.76$$

STATISTICA TEST

La statistica test per la verifica dell'ipotesi nulla $b = 0$ mirerà a valutare quanto è più grande la varianza di regressione rispetto alla varianza residua:

$$F = \frac{\text{dev. regr.} / \text{g.l.}(1)}{\text{dev. res.} / \text{g.l.}(N - 2)} =$$

$$= \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \hat{y}_i)^2} \cdot \frac{1}{N - 2}$$

REGOLA DI DECISIONE

Fissato $\alpha = 0,05$

Individuata la distribuzione della statistica test (F-Fisher)

se $F_{\text{calc}} > F_{\text{tab}}$ allora rifiuterò H_0

**COSTRUIRE LA TAVOLA
DELL'ANALISI DELLA VARIANZA
DELLA REGRESSIONE**

Nell'esempio

ANOVA

Sorgenti di variazione	DEVIANZE	G.L.	VARIANZE	F _{CAL}
REGRESSIONE	530.14	1	530.14	40.48
RESIDUA	104.76	8	13.095	
TOTALE	634.9	9		

Essendo $F_{\text{tab}} = F_{1,8} = 5.32 < F_{\text{cal}} = 40.48$ rifiuto H_0

Dove $F_{\text{cal}} = \text{Varianza Regressione} / \text{Varianza residua} = 530.14 / 13.095 = 40.48$

VERIFICA DELLE IPOTESI SU \mathbf{b}

$$\mathbf{H}_0: \mathbf{b} = \mathbf{b}_0$$

$$\mathbf{H}_1: \mathbf{b} \neq \mathbf{b}_0$$

$$T = \frac{b - b_0}{ES(b)}$$

$$\begin{aligned} ES(b) &= \sqrt{\frac{\text{var } res}{devx}} = \\ &= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2 / N - 2}{\sum (x_i - \bar{x})^2}} \end{aligned}$$

La distribuzione della statistica test è t-Student con gradi di libertà pari ai gradi di libertà della varianza residua $N - 2$.

Fissato $\alpha = 0,05$ se $t_{\text{calc}} > t_{\text{tab}}$ rifiuterò H_0

INTERVALLO DI CONFIDENZA per b

Può risultare utile determinare **l'intervallo di confidenza per b**

il metodo è lo stesso seguito per gli altri intervalli di confidenza:

stima \pm (fattore di correzione \times errore della stima)

$$\hat{b} \pm t_{\alpha, n-2} ES(b)$$

**VERIFICATE LA
SIGNIFICATIVITA' DEL
COEFFICIENTE DI
REGRESSIONE
TRAMITE IL TEST t-Student
utilizzando i dati del problema**

$$\mathbf{H_0: b = b_0 = 0}$$

$$\mathbf{H_1: b \neq b_0 \neq 0}$$

$$T = 1.3374/0.2102 = 6.3625$$

Essendo $T_{\text{cal}} = 6.3625 > T_{8,0.05} = 2.306$

Si rifiuta H_0 , cioè b è significativamente diverso da zero

L'Intervallo di confidenza per b è : $1.3374 \pm (2.306 \times 0.2102)$
 $0.8527 \leq b \leq 1.8221$

COEFFICIENTE DI DETERMINAZIONE

Il coefficiente di determinazione indica quanta parte delle osservazioni sono spiegate dal modello, cioè quanti dati cadono sulla retta stimata.

$$R^2 = \frac{devregr}{devtot} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

con $0 \leq R^2 \leq 1$

$R^2 = 0$ la retta di regressione coincide con la retta della media della variabile Y.

$R^2 = 1$ la retta di regressione spiega una alta percentuale di dati.

Nel caso del nostro esempio $R^2 = 0.835$
cioè il modello spiega 83.5% dei dati