Indice di determinazione

Riassunto

A partire da un semplice esempio basato sulla relazione tra due caratteri quantitativi dapprima si introducono i concetti alla base della costruzione dell'indice di determinazione (che misura l'importanza della variabile esplicativa nel determinismo della variabile dipendente), e successivamente si approfondisce la sua interpretazione.

Parole chiave. *Indice di determinazione, retta di regressione, coefficiente di correlazione lineare, variabilità.*

Summary

Measuring the goodness of fit

From a simple example of a bivariate distribution we constructed an index measuring the goodnes of fit of a regression line, and its interpretation was widely described.

Key words. *Goodness of fit, regression line, Person's correlation coefficient, variability.*

Non possiamo concludere l'esposizione della regressione senza introdurre un formidabile strumento di analisi, di uso assai frequente nelle applicazioni: l'indice di determinazione. Come al solito, l'apparato formale è ridotto al minimo, mentre, anche con grafici, sono sottolineati sia i concetti alla base della sua costruzione, sia la sua interpretazione. Considereremo la sola relazione di Y ad X (entrambi quantitativi), ma quanto esposto, mutatis mutandis, può essere facilmente applicato, ove occorra, alla relazione di X a Y. Ci si avvarrà di un esempio didattico con pochissime unità per rendere più facilmente leggibili i tre grafici che saranno introdotti.

Esempio (dati non reali). Uno studio esplorativo per la progettazione di un successivo studio di *dose finding* è stato condotto su 4 pazienti neoplastici, trattati con 4 diverse dosi di un farmaco. La risposta è stata valutata in termini di percentuale di riduzione della massa tumorale. I risultati ottenuti sono riepilogati nella seguente tabella

Y, risposta (%)	10	14	30	26
X, Dose	8	12	16	20

La corrispondente nuvola di punti (v. "statistica per concetti", in CASCO 2017, estate 2017) è rappresentata con il diagramma di dispersione (figura 1).

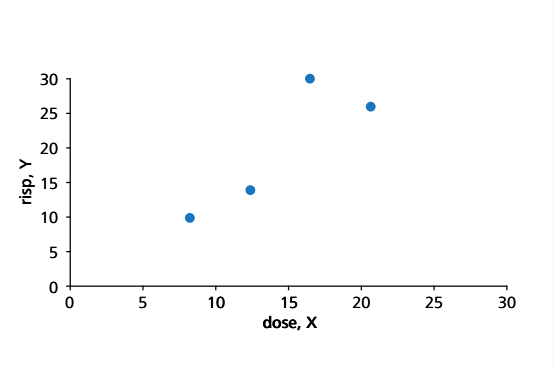

Figura 1. *Diagramma di dispersione.*

Nel diagramma di dispersione introduciamo la media di Y, rappresentata da una parallela all'asse delle ascisse passante per il punto di ordinata 20 (cioè la media di Y) e congiungiamo le ordinate dei punti della nuvola a tale retta.

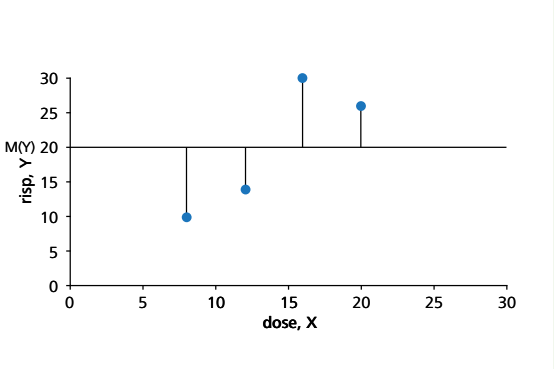

Figura 2. *Diagramma di dispersione con la media Y (pari a 20).*

I segmenti di retta, paralleli all'asse delle ordinate, che congiungono ciascun punto della nuvola alla media rappresentano le distanze dei singoli punti dalla media. Così, la somma dei quadrati delle lunghezze di tali segmenti costituiscono la devianza di Y, Dev (Y), che è una misura della variabilità totale di Y (per il concetto di "devianza" si veda "statistica per concetti", in CASCO 14, primavera 2016).

Nel grafico riportato in figura 2 introduciamo la retta di regressione di Y a X, di equazione $Y = -5 + 1{,}6X$ (v "statistica per concetti" in CASCO 17).

Cogliamo l'occasione per rivedere l'interpretazione di alcuni concetti. L'equazione della retta di regressione (i cui parametri sono calcolabili anche con un foglio EXCEL, oltre che con qualunque software statistico) mostra che

tra Y e X vi è concordanza (coefficiente angolare della retta positivo) cioè, che al crescere della dose, in media, cresce anche la risposta: +1,6 indica che al crescere unitario della dose la riduzione della massa tumorale cresce in media dell'1,6% (cioè il tumore si riduce in media dell'1,6%).

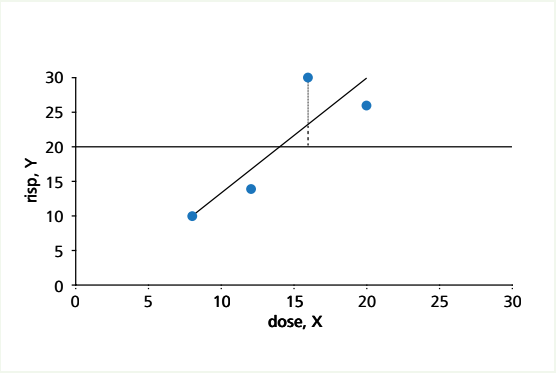

Figura 3. *Retta di regressione di Y a X.*

Nella figura 3 si vede che la distanza dell'ordinata (empirica, cioè osservata, Y) di ciascun punto della nuvola dalla media di Y si può vedere come somma di due distanze: una è la distanza (parte continua nel grafico parallela all'asse delle ordinate) del punto empirico dall'ordinata del punto che si trova sulla retta di regressione, in corrispondenza della stessa ascissa (ordinata teorica); l'altra è della distanza dell'ordinata teorica dalla media (parte tratteggiata nel grafico). Tale decomposizione nella figura 3, per semplicità, è stata visualizzata per un solo punto ma vale ovviamente per tutti i punti.

La somma dei quadrati delle distanze (o scarti) delle ordinate dei punti empirici (quelli della nuvola dei punti) dalle corrispondenti ordinate dei punti teorici (quelli che si trovano sulla retta di regressione in corrispondenza delle stesse ascisse) si chiama devianza dell'errore, Dev (E), perché misura la variabilità dovuta all'errore che si commette interpolando la nuvola di punti con la retta di regressione, cioè stimando, per ciascuna ascissa, il valore di Y con la retta di regressione. Infatti, se l'interpolazione fosse riuscita perfettamente la retta sarebbe passata sopra tutti i punti della nuvola e Dev (E) sarebbe stata pari a zero.

La somma dei quadrati degli scarti tra i punti teorici (quelli che si trovano sopra la retta di regressione) e la media si chiama devianza di regressione, Dev (R), ed è la parte della variabilità totale dovuta alla retta di regressione (cioè alla sua inclinazione).

Si può dimostrare che la stessa decomposizione delle distanze che vale per ciascun punto (distanza totale dell'ordinata empirica = distanza di tale ordinata da quella teorica + distanza dell'ordinata teorica dalla media) vale anche per le corrispondenti somme dei quadrati, cioè che la devianza totale è pari alla somma della devianza dell'errore e della devianza di regressione:

$$\text{Dev (Y)} = \text{Dev (R)} + \text{Dev (E)}.$$

Dato che la devianza totale è indipendente dall'aver interpolato o meno la nuvola di punti con la retta di regressione, essa può considerarsi una costante. Pertanto, quanto più è grande Dev (R), tanto più è piccola Dev (E) e, quindi, tanto più è piccolo l'errore che si commette con l'interpolazione della retta di regressione. Viceversa, quanto maggiore è Dev (E), tanto minore risulta Dev (R), cioè tanto peggio è riuscita l'interpolazione dei punti della nuvola.

Il rapporto tra la devianza di regressione e la devianza totale

$$\rho = \text{Dev (R)} / \text{Dev (Y)}$$

prende il nome di **indice di determinazione** perché misura quanta parte della variabilità totale è dovuta alla retta di regressione, cioè quanta parte della variabilità totale è spiegata (o determinata) dalla retta di regressione.

ρ è un rapporto di composizione (di parte al tutto); essendo le devianze sempre positive o nulle, ρ assume valori tra 0 e 1. Quanto più ρ si avvicina ad 1, tanto maggiore è la parte della variabilità totale spiegata dalla retta di regressione, e, quindi, tanto minore risulta quella dovuta all'errore (cioè al fatto che la retta non passa sopra tutti i punti empirici); viceversa, quanto più ρ è vicino allo zero, tanto peggio è riuscita l'interpolazione, cioè tanto più grande è la variabilità dovuta all'errore. Nel nostro esempio, ρ è interpretabile come la frazione di variabilità totale spiegata dalla (dovuta alla) dose, mentre il suo complemento ad 1 costituisce la devianza dell'errore.

Casi limite:

a. se è $\rho = 0$, tutta la variabilità è dovuta all'errore cioè Dev (E) = Dev (Y): la retta di regressione è parallela all'asse delle ascisse e passa per la media di Y (figura 2);
b. se è $\rho = 1$, tutta la variabilità è spiegata dalla retta di regressione: Dev (R) = Dev (Y): la retta passa sopra tutti i punti della nuvola.

Da questa prospettiva, ρ è un indice che misura la **bontà di adattamento della retta di regressione alla nuvola di punti (goodness of fit)**: quanto più ρ è prossimo a 1, tanto meglio è riuscita l'interpolazione.

Si può dimostrare che **ρ coincide con r^2**, cioè con il quadrato del coefficiente di correlazione lineare di Bravais-Pearson (v. "statistica per concetti" in CASCO 18). Attenzione però al fatto che r e r^2 misurano cose diverse; infatti r (che varia tra – 1 e + 1) è una misura di concordanza (quanto cresce, se è positivo – o diminuisce, se è negativo – l'intensità di un carattere, in media, al crescere dell'altro), mentre r^2 (che varia tra 0 e 1: non può assumere valori negativi essendo un quadrato) è una misura di accostamento della retta di regressione alla nuvola di punti.

Ogni volta che viene interpolata una retta di regressione i risultati sono sempre accompagnati dall'indice di determinazione r^2 utile a valutare la qualità dell'interpolazione eseguita.

Tornando all'esempio, la retta interpolatrice della nuvola di punti ha equazione

$$Y = -5 + 1{,}6X$$

con un coefficiente di correlazione di Bravais-Pearson pari a $r = 0{,}868$ (nella distribuzione doppia c'è l'86,8% del massimo della concordanza lineare che si sarebbe potuta osservare).

Quindi è $r^2 = 0{,}753$: la Dev (R) è il 75,3% della devianza totale; in altre parole, la retta di regressione spiega il 75,3% della variabilità totale (cioè la dose spiega il 75,3% della variabilità della riduzione della massa tumorale osservata nei 4 pazienti), mentre il restante 24,7% è dovuto all'errore. In tale interpretazione per "errore" si intende l'effetto congiunto di tutte le variabili che non sono state prese in considerazione (età del paziente, sesso, stadio della malattia e così via). Quindi $r^2 = 0{,}753$ indica da un lato che la retta di regressione passa molto vicino alla nuvola di punti (l'interpolazione è ben riuscita) e, dall'altro che la dose del farmaco è molto importante per spiegare la variabilità della riduzione della massa tumorale nei 4 pazienti osservati.

Queste considerazioni possono essere generalizzate al caso di una variabile dipendente (nell'esempio la percentuale di riduzione della massa tumorale) e una molteplicità di variabili esplicative e fattori (nell'esempio, considerando oltre la dose del farmaco, anche l'età, il tipo di neoplasia, lo stadio della malattia, ecc.) per vedere se, ad esempio, la relazione tra dose e risposta è più accentuata in un sottogruppo di pazienti che in altri. Ovviamente per far questo è necessario un gruppo ben più numeroso di pazienti.

Enzo Ballatori