Rivista di Statistica Applicata Vol. 20 n. 3 settembre 1987

ROSTRA

Rubrica a cura di Benito V. Frosini

Francis Galton e la regressione delle stature

Benito V. Frosini, Istituto di Statistica, Università Cattolica del Sacro Cuore – Milano

1. Premessa

Il termine *regressione* ha un significato pacifico e consolidato in Statistica, con l'unica differenziazione, usata da molti autori, fra "regressione del primo tipo" e "regressione del secondo tipo" (cfr. ad esempio Cramér, 1946, pp. 270–277; Fisz, 1963, pp. 91–101). Limitando il riferimento a una variabile doppia (X, Y) (statistica o casuale), col termine "funzione di regressione" (sic et simpliciter, ovvero del primo tipo) di Y su X si intende la funzione che associa ai valori x del supporto di X le medie (aritmetiche) $M(Y|x)$ delle variabili condizionate $Y|X = x$; si possono usare a questo proposito anche altri indici di posizione, ad esempio la mediana $Me(Y|x)$; così facendo, visto che l'uso della media aritmetica è di gran lunga il più frequente (per ragioni di semplicità), si dovrà tuttavia distinguere i diversi casi con diversi attributi, e cioè "funzione di regressione media", "funzione di regressione mediana" ecc. Come è ben noto dalla teoria dei valori medi, la caratterizzazione di queste funzioni di regressione può anche avvenire attraverso le medie aritmetiche di "funzioni di perdita": imporre la minimizzazione della media degli scarti assoluti per ogni x conduce alla funzione di regressione *mediana* ecc.

La regressione del secondo tipo consiste nell'interpolare una curva fra i punti (x, y), scelta da un insieme di curve regolari; se l'interpolazione avviene col metodo dei minimi quadrati, si parlerà ad esempio di "retta di regressione dei minimi quadrati"; se si impiega il metodo dei minimi valori assoluti, si

otterrà ad esempio una "retta di regressione dei minimi valori assoluti". Vale forse la pena di accennare che mentre il metodo dei minimi quadrati è stato di gran lunga il più studiato e applicato fino a tempi molto recenti, a causa delle notevoli ed eleganti proprietà formali – sia parametriche, legate alla distribuzione normale, sia non–parametriche –, il metodo dei minimi valori assoluti ha ricevuto di recente una rinnovata attenzione nell'ambito degli studi sulla robustezza delle procedure statistiche.

Quando nel corso delle lezioni introduco il concetto di "funzione di regressione", sento il bisogno di soffermarmi sull'origine del termine, per due motivi: (*a*) anche se si tratta di un termine tecnico, che si può usare come una mera etichetta, esso deriva da una ricerca e da un autore che hanno avuto un grande rilievo nello sviluppo degli studi statistici durante la seconda metà del secolo scorso; (*b*) parlando a studenti di Economia e Commercio, è opportuno chiarire come e perché questa terminologia non deriva e non ha a che fare con la "regressione economica". In questa nota presenterò il materiale che ho raccolto come ausilio didattico su questo argomento, con la speranza che possa essere utile a colleghi e studenti ad esso interessati.

2. Il paradosso della regressione delle stature nelle ricerche di Francis Galton

Se raccontate agli studenti che nelle popolazioni umane si è rilevato che i figli adulti hanno mediamente una statura più vicina alla norma (in questo caso = mediana = media aritmetica, per la simmetria delle distribuzioni) di quella dei loro genitori, essi non vi presteranno molta fede; i più svegli vi chiederanno come è possibile, dato che col succedersi delle generazioni si dovrebbe osservare una dispersione delle stature sempre più ridotta, contrariamente all'evidenza. Se poi vi spingete a dire che si è pure rilevato che i genitori hanno mediamente una statura più vicina alla norma di quella raggiunta dai propri figli, beh, a voler essere benevoli, gli studenti crederanno di aver capito male, essendo in apparente contraddizione con l'affermazione precedente. Però, se siete arrivati fino a questo punto, dovrete dedicare una ventina di minuti per chiarire che effettivamente ambedue le affermazioni sono vere.

Le ricerche sull'ereditarietà dei caratteri (nelle piante, negli animali e nell'uomo) conseguirono con Francis Galton un autentico salto di qualità, particolarmente a causa dell'applicazione ingegnosa a dati accuratamente raccolti delle tecniche statistiche allora più avanzate, cui lo stesso Autore aveva già dato notevoli contributi.

Nel 1875 Galton iniziò, con l'assistenza del cugino Charles Darwin, uno studio sulla trasmissione di alcuni caratteri ereditari di tipo dimensionale (diametro e peso) riguardante il pisello odoroso (*Lathyrus odoratus*, pianta solitamente coltivata per i fiori odorosi a forma di farfalla, di delicate tinte pastello).

E' dello stesso anno la prima "retta di regressione", trovata da Karl Pearson in un libro di appunti non pubblicato (cfr. Pearson, 1930, pp. 3–4), riguardante la relazione fra i diametri dei piselli genitori e i diametri della loro progenie; non ci occuperemo tuttavia in questa sede di tali ricerche, cui Galton dedicò diversi anni, anche perché i risultati cui l'autore è pervenuto si accordano assai bene con quelli relativi alla statura umana.

Basterà ricordare due risultati fondamentali scoperti da Galton, e cioè che tra i semi filiali e i semi materni sussiste una regressione lineare, e che distribuzioni dei semi filiali corrispondenti a un dato seme materno hanno la stessa variabilità (c.d. omoscedasticità delle distribuzioni condizionate).

Per quanto riguarda la statura umana, Galton raccolse a sue spese (cinquecento sterline dell'epoca) dati analitici – anche se non sempre accurati – riguardanti circa 200 famiglie. Da essi ha ricavato una tabella a doppia entrata in cui le stature dei genitori e dei figli adulti sono classificate a capo delle righe e a capo delle colonne, rispettivamente; ad ogni classe bidimensionale è stata associata la frequenza dei figli aventi le date modalità dei due caratteri. Questa tabella è riportata sia in Galton (1886) sia in Galton (1889); con la sola esclusione delle classi aperte iniziali e finali per le due variabili, e con l'aggiunta della seconda colonna – il cui contenuto sarà spiegato più avanti –, si tratta della Tab. I di questo articolo.

La precedente dizione sintetica riguardo al contenuto della tabella necessita però di una spiegazione, nel senso che in realtà i *genitori* considerati da Galton non sono quelli comunemente intesi, e anche le *stature* non sono proprio le stature rilevate.

Tenuto conto che la trasmissione dei caratteri ereditari dipende, ovviamente, da entrambi i genitori, che la distribuzione delle stature è diversa nei due sessi, e che non erano disponibili dati sufficienti per separare gli effetti dovuti a ciascuno dei genitori, Galton risolse la questione trasformando dapprima le stature femminili in stature "maschili" moltiplicandole per 1,08 (rapporto fra le mediane delle due distribuzioni), e successivamente mediando le altezze – così ottenute – dei due genitori(*).

Col termine "Mid–parent" – che potremmo tradurre con "genitore medio" – Galton intende "una persona ideale di sesso composito, la cui statura è a metà strada fra la statura del padre e la statura trasformata della madre" (Galton, 1889, p. 87). In conclusione, Galton ha classificato i figli secondo le loro stature (quelle delle femmine moltiplicate per 1,08) e secondo la "statura" del genitore medio. Così sono da intendere le variabili nella prima colonna e nella prima riga di Tab. I.

(*) In ricerche precedenti Galton aveva usato un criterio più rigoroso, consistente nel sostituire a una data statura femminile che è percentile di ordine p della distribuzione (di stature femminili) la statura maschile che è percentile dello stesso ordine nella distribuzione delle stature maschili.

Tab. I – Frequenze dei figli adulti di varie stature nati da 205 coppie di genitori. Stature in pollici. Le stature femminili sono state moltiplicate per 1,08. (Da Galton, 1889, p. 208)

X = Statura del genitore medio	X_1 = Statura media dei genitori	Y = Statura dei figli adulti												Totale
		62,2	63,2	64,2	65,2	66,2	67,2	68,2	69,2	70,2	71,2	72,2	73,2	
72,5	74,28	0	0	0	0	0	0	1	2	1	2	7	2	15
71,5	72,85	0	0	0	1	3	4	3	5	10	4	9	2	41
70,5	71,43	0	1	0	1	1	3	12	18	14	7	4	3	64
69,5	70,01	0	1	16	4	17	27	20	33	25	20	11	4	178
68,5	68,58	0	7	11	16	25	31	34	48	21	18	4	3	218
67,5	67,16	3	5	14	15	36	38	28	38	19	11	4	0	211
66,5	65,74	3	3	5	2	17	17	14	13	4	0	0	0	78
65,5	64,32	0	9	5	7	11	11	7	7	5	2	1	0	65
64,5	62,89	1	4	4	1	5	5	0	2	0	0	0	0	22
Totale		7	30	55	47	115	136	119	166	99	64	40	14	892

Tab. II – Frequenze delle stature maschili e femminili in un gruppo di 205 coppie di genitori (Da Galton, 1889, p. 206). Fra parentesi sono riportate le frequenze teoriche in caso di indipendenza.

Statura femminile	Statura maschile			Totale
	Bassa	Media	Alta	
Alta	12 (11,2)	20 (24,1)	18 (14,6)	50
Media	25 (23,3)	51 (50,2)	28 (30,4)	104
Bassa	9 (11,4)	28 (24,6)	14 (14,9)	51
Totale	46	99	60	205

Un interesse in sé e per sé, e inoltre per l'argomento qui trattato, avrebbe avuto una tabella di correlazione che classificasse le famiglie (ovvero i genitori medi) secondo la statura del padre e la statura della madre. Una tale tabella non viene pubblicata da Galton, il quale invece presenta una tabella di contingenza in cui tali stature sono classificate nella tripartizione "alte, medie, basse" (Galton, 1889, p.206), la quale viene qui presentata come Tab. II. (a parte le "frequenze teoriche", che non compaiono nella tabella di Galton). E' subito evidente dall'esame di questa tabella che i due caratteri appaiono sostanzialmente indipendenti; il valore della statistica di adattamento X^2 è uguale a 2,907, ampiamente inferiore alla mediana 3,3567 della v.c. chi–quadrato con 4 gradi di libertà.

Un modo più accurato di verificare l'indipendenza (suggerito dallo stesso Galton, che impiega la differenza interquartile in luogo dello scarto quadratico medio (s.q.m.)) consiste nel confrontare la dispersione della variabile X (= statura del genitore medio) con quella delle variabili X_1 e X_2 che stanno a indicare la statura dei padri e la statura trasformata delle madri. L'Autore aveva già verificato dall'esame delle rispettive funzioni di ripartizione che le variabili X_1, X_2 e $X = (X_1 + X_2)/2$ sono – con ottima approssimazione, normali; gli risultava inoltre che le variabili X_1 e X_2 hanno lo stesso s.q.m. (con l'approssimazione consentita dalla qualità delle misurazioni disponibili). Ponendo allora $\sigma_1 = \sigma_2 = \sigma$ (eguaglianza fra s.q.m. di X_1 e X_2), per la varianza di X vale

$$Var(X) = \frac{1}{2}\sigma^2 + \frac{1}{2}Cov(X_1, X_2).$$

Dalla distribuzione marginale X in Tab. I Galton ha inoltre verificato che $Var(X) \simeq \sigma^2/2$, da cui deriva $Cov(X_1, X_2) \simeq 0$, potendosi quindi ragionevolmente ipotizzare l'indipendenza tra le stature dei due genitori.

Si osservi che, per quanto le misurazioni utilizzate da Galton siano indicative di una situazione di indipendenza fra le stature dei genitori, non sembra che questa debba essere, in generale, la conclusione corretta, come ha evidenziato Karl Pearson (1903, p. 373).

Lo stesso contenuto di Tab. I permette comunque una verifica molto importante. Le medie, gli s.q.m. e il coefficiente di correlazione risultano

$$\overline{x} = 68,3 \; ; \quad \overline{y} = 68,1 \; ; \quad \sigma_x = 1,6654 \; ; \quad \sigma_y = 2,3697 \; ; \quad r = 0,4056$$

per una retta interpolante dei minimi quadrati di equazione

$$y = 28,62 + 0,577x.$$

Per quanto già detto, la variabile Y (statura dei figli) è data dal miscuglio di due variabili normali Y_1 e Y_2 praticamente identiche, quindi σ_y può considerarsi lo s.q.m. di ciascuna di esse. Perché σ_y coincida con σ (s.q.m. di X_1 e X_2, sopra definite), deve essere allora

$$\sigma_x = \sigma_y/\sqrt{2} = \sigma_y \cdot 0,7071 = 1,6756,$$

e in effetti questo risultato differisce pochissimo dal valore calcolato di $\sigma_x = 1,6654$. Si può quindi concludere che le due variabili X e Y a confronto sono normali, hanno la stessa media (nel campione le medie sono $\overline{x} = 68,3$ e $\overline{y} = 68,1$), mentre le varianze differiscono per il semplice fatto che X deriva dalla somma di due variabili indipendenti distribuite come Y.

Un trattamento più simmetrico vorrebbe che in Tab. I anche la variabile X si riferisse a un solo individuo, e non a quel personaggio composito che è il genitore medio. Ciò è possibile se si può ammettere che la statura dei figli è ugualmente correlata con le stature dei due genitori. Per quanto i dati di Galton – come successivamente verificato da K. Pearson (1896, p. 270) – facciano apparire una maggiore influenza esercitata dalla statura del padre (forse a causa della peggiore qualità delle misurazioni delle stature femminili),ciò non è stato poi confermato da indagini più ampie ed accurate eseguite dallo stesso Pearson (1903, p.378). Ad ogni modo non sarà questo un punto di discussione; per semplicità ammetteremo quindi l'eguaglianza dei quattro coefficienti di correlazione delle stature dei figli (maschi e femmine) con quelle di ciascun genitore.

E' allora possibile conseguire lo scopo appena proposto trasformando semplicemente la scala della dispersione intorno alla media per X, moltiplicandola per $\sqrt{2} = 1,4142$, o anche, come si è fatto per ottenere le cifre in seconda

colonna di Tab. I, per il rapporto fra s.q.m. $\sigma_y/\sigma_x = 1,4229$. Così facendo, con le due marginali aventi la stessa media e lo stesso s.q.m., la retta di regressione per le variabili scarto avrebbe semplicemente equazione $y = rx$ cioè $y = 0,406x$.

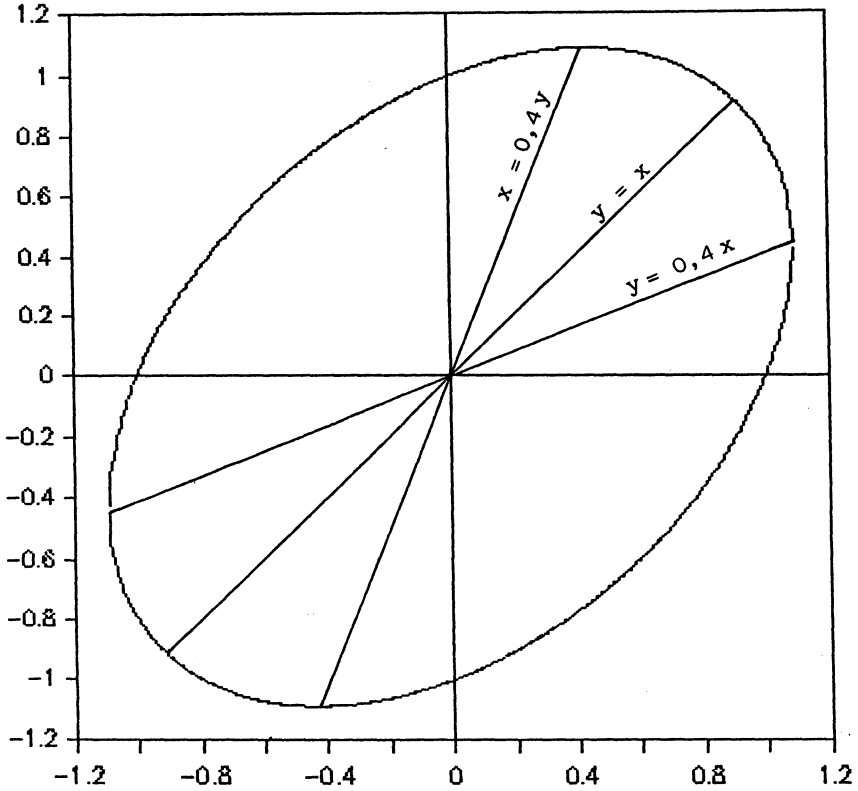

Fig. 1 – Esempio di ellisse di equidensità per una variabile normale bivariata con medie nulle, eguali varianze, e coefficiente di correlazione uguale a 0,4.

E' davvero sorprendente come Galton, senza la conoscenza della densità generale di una normale bivariata, abbia potuto conseguire risultati estremamente precisi da un esame molto accurato della Tab. I. Infatti, a seguito di un "lisciamento" delle frequenze interne, egli scoprì che le curve passanti per frequenze uguali "formano una serie di ellissi concentriche e similari"; pertanto, mediante un grafico simile a quello in Fig. 1, fu in grado di presentare un'esposizione semplice ed elegante dei risultati già conseguiti.

Abbandoniamo a questo punto, per brevità, il riferimento alle osservazioni di Tab. I, e come Galton riferiamoci alla Fig. 1, in cui compare un'ellisse, assunta come generica curva di livello della densità di una normale bivariata $(X, Y) \sim N(0, 0, \sigma^2, \sigma^2, \rho)$, con medie nulle (ci riferiamo a scarti dalle medie), identiche varianze per le due marginali e coefficiente di correlazione $\rho > 0$;

più precisamente, si tratta dell'ellisse di cui alla successiva formula (2), avendo posto $\rho = 0,4$ (tenuto conto dei precedenti risultati) e $c = 1$.

All'interno dell'ellisse sono tracciate le due rette di regressione $y = 0,4x$ e $x = 0,4y$ (si ricordi che si tratta di curve di regressione "del primo tipo"). Considerando la prima retta, si legge ad esempio che da genitori aventi una statura di 10 cm. sopra o sotto la media, nascono figli che mediamente hanno una statura di 4 cm. sopra o sotto la media, rispettivamente; la statura media dei figli, perciò, *regredisce* verso la media, da cui la denominazione di *retta di regressione* introdotta da Galton con questo preciso significato.

L'Autore ritrova una legge naturale cui già si era avvicinato nello studio della discendenza dei piselli, e così si esprime (Galton, 1889, p. 95): "However paradoxical it may appear at first sight, it is theoretically a necessary fact, and one that is clearly confirmed by observation, that the stature of the adult offspring must on the whole, be more *mediocre* than the stature of their parents".

Tenuto conto della precedente affermazione, sembra altrettanto se non più paradossale (come avverte l'Autore a p. 100 dell'opera citata) la regressione inversa, cioè $x = 0,4y$; riprendendo l'esempio, ciò significa che figli aventi una statura di 10 cm. sopra o sotto la media hanno genitori con una statura mediamente di 4 cm. sopra o sotto la media, rispettivamente. Come è evidente, il paradosso illustrato non dipende dal valore scelto di $\rho = 0,4$; si farebbero discorsi del tutto analoghi, ad esempio, anche se fosse $\rho = 0,5$ o $\rho = 0,6$.

Ciò che assicura per i figli la stessa distribuzione delle stature dei genitori è la distribuzione simmetrica intorno all'asse principale dell'ellisse, di equazione $y = x$; ma questo stesso fatto, come avvertiva Galton nella suddetta citazione, implica necessariamente sia la regressione verso la media delle stature dei figli rispetto a quelle dei genitori, sia la regressione verso la media delle stature dei genitori rispetto a quelle dei figli.

Le osservazioni fatte potrebbero essere ripetute quasi esattamente anche nel caso che la distribuzione delle stature filiali avesse una media più grande, o più piccola, rispetto alla distribuzione delle stature dei genitori; per fissare le idee possiamo ammettere che la statura media dei figli sia maggiore della statura media dei genitori. Basta considerare in ogni caso che gli scarti dalla media delle stature sono da prendere distintamente dalla media delle stature dei figli (per i figli), e dalla media delle stature dei genitori (per i genitori); il grafico in Fig. 1 resta perfettamente valido, in quanto già si riferisce a scarti dalla media, senza alcuna assunzione sulla eguaglianza o meno delle due variabili originali X e Y.

Rispetto al caso considerato da Galton ($\mu_X = \mu_Y$), nel piano XY si avrebbe uno spostamento verso l'alto delle ellissi concentriche. E' anche verosimile, però, che al crescere di μ_Y cresca anche σ_Y; in tal caso in Fig. 1 l'asse principale dell'ellisse non giacerebbe più sulla retta $y = x$, ma su una retta del tipo $y = ax$ per $a > 1$; la perdita di simmetria renderebbe più evidente uno

dei suddetti paradossi, e meno evidente l'altro.

3. Qualche risultato teorico

Essendo $(X, Y) \sim N(0, 0, \sigma^2, \sigma^2, \rho)$, la sua densità è data da

$$f(x,y) = \frac{1}{2\pi\sigma^2(1-\rho^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right\}$$

$$-\infty < x, y < \infty;$$

la marginale X ha densità

$$f_1(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\{-x^2/(2\sigma^2)\} \qquad -\infty < x < \infty$$

mentre la densità della v.c. condizionata $Y|X = x$ è

$$g(y|x) = \frac{1}{\sigma[2\pi(1-\rho^2)]^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2(1-\rho^2)}(y - \rho x)^2\right\} \tag{1}$$

come si controlla subito dall'eguaglianza

$$g(y|x) f_1(x) = f(x, y);$$

si osservi dalla (1) che $Y|x \sim N(\rho x, \sigma^2(1-\rho^2))$ (cfr. ad esempio Fraser, 1958, pp. 83–84).

Le curve di livello di $f(x, y)$ derivano dall'eguaglianza

$$x^2 - 2\rho xy + y^2 = c \qquad c > 0; \tag{2}$$

si tratta di ellissi similari e concentriche, centrate nell'origine degli assi (cfr. un esempio in Fig. 1). Dato x, dalla (2) si ottengono le due radici

$$y = \rho x \pm [c - (1-\rho^2)x^2]^{\frac{1}{2}} \tag{3}$$

che sono reali e corrispondono a punti dell'ellisse per

$$-[c/(1-\rho^2)]^{\frac{1}{2}} \leq x \leq [c/(1-\rho^2)]^{\frac{1}{2}}; \tag{4}$$

per $x = \pm[c/(1-\rho^2)]^{\frac{1}{2}}$ si ha un solo punto sull'ellisse, che sta sulla retta di regressione Y su X di equazione $y = \rho x$.

Per x nell'intervallo (4) la semisomma dei due valori corrispondenti y sull'ellisse è data da $y = \rho x$ (come si verifica subito dalla (3)); la retta di regressione Y su X è perciò il luogo dei punti equidistanti dal ramo inferiore

e dal ramo superiore dell'ellisse, prendendo le distanze lungo l'asse y (nell'intervallo (4)). Analogamente si ricava che la retta di regressione X su Y è il luogo dei punti equidistanti dal ramo destro e dal ramo sinistro dell'ellisse, prendendo le distanze lungo l'asse x.

L'ellisse (2) può scriversi:

$$[x\ y]A\begin{bmatrix}x\\y\end{bmatrix} = c \quad \text{dove } A = \begin{bmatrix}1 & -\rho\\ -\rho & 1\end{bmatrix};$$

le radici caratteristiche di A, ottenute dall'equazione $|A - \lambda I| = 0$, sono $\lambda_1 = 1 + \rho$ e $\lambda_2 = 1 - \rho$; il vettore caratteristico di lunghezza unitaria associato a λ_1 è $x_1' = [1/\sqrt{2}, 1/\sqrt{2}]$, mentre il secondo vettore caratteristico è $x_2' = [-1/\sqrt{2}, 1/\sqrt{2}]$; i coseni direttori dell'asse principale dell'ellisse sono pertanto uguali entrambi a $1/\sqrt{2}$: si tratta dei coseni direttori della retta $y = x$ (cfr. ad esempio Franklin, 1968, pp. 94–97).

Riprendiamo ora con maggiore generalità, e da una diversa angolazione, il problema discusso al precedente paragrafo. Si ammetta che $X \sim N(0, \sigma_x^2)$ con densità f_1, e $Y|X = x \sim N(bx, \sigma_r^2)$, $b > 0$, con densità g_1; in altre parole la regressione di Y su X è lineare, ed è espressa dalla retta $y = bx$; σ_r^2 è la varianza residua intorno alla retta di regressione Y su X. Da queste assunzioni si ricava la densità f_2 della marginale Y:

$$\begin{aligned} f_2(y) &= \int g_1(y|x) f_1(x) dx \\ &= \int \frac{1}{2\pi\sigma_x\sigma_r} \exp\left\{-\frac{1}{2}\left[\left(\frac{y - bx}{\sigma_r}\right)^2 + \left(\frac{x}{\sigma_x}\right)^2\right]\right\} dx \end{aligned}$$

da cui, con semplici passaggi, si ottiene

$$f_2(y) = \frac{1}{[2\pi(b^2\sigma_x^2 + \sigma_r^2)]^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\frac{y^2}{b^2\sigma_x^2 + \sigma_r^2}\right\}$$

che è la densità di una normale $N(O, b^2\sigma_x^2 + \sigma_r^2)$. Ponendo $\sigma_y^2 = b^2\sigma_x^2 + \sigma_r^2$, vale $\sigma_y^2 >=< \sigma_x^2$ se e solo se

$$b^2\sigma_x^2 + \sigma_r^2 >=< \sigma_x^2; \tag{5}$$

poiché vale $\sigma_r^2 = \sigma_y^2(1 - \rho^2)$, dove $\rho = Corr(X, Y)$, la (5) si può scrivere

$$\sigma_y^2(1 - \rho^2) >=< \sigma_x^2(1 - b^2); \tag{6}$$

la posizione $\sigma_x = \sigma_y$ implica $b = \rho$: se interpretiamo X come la distribuzione di una caratteristica ereditaria al tempo 0, e Y come la distribuzione della stessa caratteristica dopo una generazione, perché si ripeta la stessa distribuzione

della caratteristica la regressione di Y su X *deve* essere espressa dalla retta $y = \rho x (0 < \rho < 1)$.

Vediamo cosa accade col succedersi delle generazioni nel caso generico in cui $Y|x \sim N(bx, \sigma_r^2)$ e la varianza residua σ_r^2 resta costante:

generazione 0: varianza = σ^2;

generazione 1: varianza = $b^2\sigma^2 + \sigma_r^2$;

generazione 2: varianza = $b^2(b^2\sigma^2 + \sigma_r^2) + \sigma_r^2 = b^4\sigma^2 + \sigma_r^2(1 + b^2)$;

generazione n: varianza = $b^{2n}\sigma^2 + \sigma_r^2(1 + b^2 + \ldots + b^{2(n-1)})$.

Per n crescente, e tenuto conto che $0 < b < 1$ per ipotesi, la varianza della caratteristica tende a

$$\text{Varianza asintotica} = \sigma_r^2/(1 - b^2). \tag{7}$$

Poiché la (5) può scriversi

$$\sigma_r^2/(1 - b^2) >=< \sigma^2,$$

il confronto del rapporto (7) con σ^2 indica se la variabilità della caratteristica è destinata a crescere oppure a diminuire, offrendo nello stesso tempo il limite cui tende la varianza col passare delle generazioni.

Riferimenti bibliografici

Cramér H., 1946, *Mathematical Methods of Statistics*, Princeton University Press, Princeton

Fisz M., 1963, *Probability Theory and Mathematical Statistics*, third edition, Wiley, New York

Franklin J.N., 1968, *Matrix Theory*, Prentice Hall, Englewood Cliffs

Fraser D.A.S., 1958, *Statistics – An Introduction*, Wiley, New York

Galton F., 1886, Regression towards Mediocrity in Hereditary Stature, *Journ. of the Anthropological Inst.*, Miscellanea, vol. XV, 246–263

Galton F., 1889, *Natural Inheritance*, Macmillan, London

Pearson K., 1896, Mathematical Contributions to the Theory of Evolution – III. Regression, Heredity, and Panmixia, *Philosophical Transactions of the Royal Society*, Series A, vol. 187, 253–318

Pearson K., Lee A., 1903, On the Laws of Inheritance in Man. I. Inheritance of physical characters, *Biometrika*, vol. II, 357–462

Pearson K., 1930, *The Life, Letters and Labours of Francis Galton*, vol. III A, Cambridge University Press, Cambridge

Summary

Francis Galton and the regression of statures

The work by Francis Galton about the heredity of dimensional characteristics in sweet peas and human beings is briefly outlined; the interest is mostly historical, aiming at commenting the early studies in regression, that were really successful in discovering that statures of adult children regress towards the mean with respect to the statures of their parents. A simple generalization is made, concerning the effects in the long run of linear regression and a given residual variance.