



UNIVERSITÀ · MATEMATICA GENERALE

Formulario modelli di regressione

Costante · lineare · lineare per l'origine · polinomiale · esponenziale · potenza
· logaritmico · inverso — con esercizi svolti per tipologia



infomaat.it

Centro ripetizioni · since 2011



INDICE

- | | |
|----------------------------------|---|
| 1. Il problema della regressione | 8. Modello logaritmico |
| 2. Modello costante | 9. Modello inverso |
| 3. Modello lineare | 10. Formule di supporto |
| 4. Lineare per l'origine | 11. Esercizio 1 — confronto tra modelli |
| 5. Modello polinomiale | 12. Esercizio 2 — esponenziale |
| 6. Modello esponenziale | 13. Esercizio 3 — potenza |
| 7. Modello potenza | 14. Esercizio 4 — inverso |

01 Il problema della regressione

IDEA GENERALE

Dato un insieme di n coppie osservate (x_i, y_i) , un **modello di regressione** è una funzione $y = g(x)$ che approssima la relazione fra X e Y . Il metodo dei **minimi quadrati** sceglie i parametri del modello minimizzando la somma degli scarti quadratici fra valori osservati e valori stimati:

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

Nei modelli **non lineari** nei parametri (esponenziale, potenza) si applica una **trasformazione** (tipicamente il logaritmo) che li rende lineari in nuove variabili; si stima il modello lineare trasformato e infine si torna alla variabile originaria.

RESIDUO E VARIANZA RESIDUA

$$\hat{E} = Y - Y^* \quad M(\hat{E}) = 0 \quad \text{Var}(\hat{E}) = \sigma_R^2 = M[(Y - Y^*)^2]$$

- Y^* indica il valore stimato dal modello. La **media dei residui è sempre nulla**: è una conseguenza diretta delle equazioni normali dei minimi quadrati, non un'ipotesi.
- La **varianza residua** σ_R^2 misura invece quanto i punti si discostano dal modello: è l'indicatore che si usa per **confrontare** più modelli sugli stessi dati — **vince il modello con σ_R^2 minore**.

02 Modello costante

$$Y = \hat{A}$$

$$y = \hat{a} \quad \hat{a} = M(y) = \mu_y \quad \sigma_R^2 = \sigma_y^2$$

È il modello più semplice possibile: si stima Y con la sua media, ignorando X . Non richiede alcuna relazione fra le due variabili, ed è quindi il termine di paragone naturale per capire se un modello più complesso "vale la pena": se un modello elaborato non riduce sensibilmente σ_y^2 , il legame con X è debole.





03 Modello lineare

$$Y = \hat{A} + \hat{B}X$$

$$y = \hat{a} + \hat{b}x$$

Pendenza $\hat{b} = \frac{COV(x, y)}{\sigma_x^2} = \rho \frac{\sigma_y}{\sigma_x}$

Intercetta $\hat{a} = M(y) - \hat{b} M(x)$

COEFFICIENTE DI PEARSON

$$\rho = \frac{COV(x, y)}{\sigma_x \sigma_y} \quad R^2 = \rho^2$$

VARIANZA RESIDUA

$$\sigma_R^2 = \sigma_y^2 (1 - \rho^2)$$

04 Lineare per l'origine

$$Y = \hat{B}X$$

Variante del modello lineare in cui si impone $\hat{a} = 0$: la retta passa per l'origine. È un'ipotesi forte (nessun termine costante), da usare solo se ha senso nel contesto (es. $y = 0$ quando $x = 0$).

$$\hat{b} = \frac{M(x \cdot y)}{M(x^2)} \quad \sigma_R^2 = M[(y - \hat{b}x)^2] = M(Y^2) - \hat{b} M(XY)$$

05 Modello polinomiale

$$Y = \hat{A} + \hat{B}X^n$$

Ponendo $T = x^n$ (con n noto, tipicamente $n = 2, 3, \dots$) il modello diventa lineare nella variabile trasformata T :

$$y = \hat{a} + \hat{b}x^n \quad \xrightarrow{T=x^n} \quad y = \hat{a} + \hat{b}T$$

Pendenza $\hat{b} = \frac{COV(Y, T)}{\sigma_T^2}$

Intercetta $\hat{a} = M(Y) - \hat{b} M(T)$

Varianza residua $\sigma_R^2 = \sigma_y^2 (1 - \rho_{YT}^2)$, con ρ_{YT} correlazione fra Y e T

Il caso $n = 2$ ($T = x^2$) è il più comune nella pratica: curva a parabola con asse verticale.



06 Modello esponenziale

$$Y = \hat{A} \cdot e^{\hat{B}X}$$

Non lineare nei parametri. Si linearizza applicando il logaritmo naturale a entrambi i membri, ponendo $W = \ln y$:

$$y = \hat{a} e^{\hat{b}x} \xrightarrow{\ln} \ln y = \ln \hat{a} + \hat{b}x \xrightarrow{W=\ln y} W = \hat{A} + \hat{b}x$$

Pendenza $\hat{b} = \frac{COV(X, W)}{\sigma_x^2}$

Intercetta trasf. $\hat{A} = M(W) - \hat{b} M(X)$

Ritorno a y $\hat{a} = e^{\hat{A}}$

La regressione lineare si stima su $W = \ln y$ contro x ; il modello si valuta poi confrontando σ_R^2 calcolata su W (variabile trasformata), non direttamente su y .

07 Modello potenza

$$Y = \hat{A} \cdot X^{\hat{B}}$$

Si linearizza applicando il logaritmo a entrambi i membri, ponendo $W = \ln y$ e $T = \ln x$:

$$y = \hat{a} x^{\hat{b}} \xrightarrow{\ln} \ln y = \ln \hat{a} + \hat{b} \ln x \xrightarrow{W=\ln y, T=\ln x} W = \hat{A} + \hat{b} T$$

Pendenza $\hat{b} = \frac{COV(W, T)}{\sigma_T^2}$

Intercetta trasf. $\hat{A} = M(W) - \hat{b} M(T)$

Ritorno a y $\hat{a} = e^{\hat{A}}$

Richiede $x > 0$ (per il logaritmo). $\hat{b}$ è l'**elasticità**: la variazione percentuale di y per una variazione dell'1% di x .

08 Modello logaritmico

$$Y = \hat{A} + \hat{B} \cdot \ln x$$

Ponendo $T = \ln x$ il modello diventa lineare in T :

$$y = \hat{a} + \hat{b} \cdot \ln x \xrightarrow{T=\ln x} y = \hat{a} + \hat{b} T$$

Pendenza $\hat{b} = \frac{COV(Y, T)}{\sigma_T^2}$

Intercetta $\hat{a} = M(Y) - \hat{b} M(T)$

Varianza residua $\sigma_R^2 = \sigma_y^2 (1 - \rho_{YT}^2)$

Richiede $x > 0$. Curva a crescita (o decrescita) rapida all'inizio e via via più lenta.

09 Modello inverso

$$Y = \hat{A} + \hat{B}/X$$

Ponendo $T = \frac{1}{x}$ il modello diventa lineare in T :

$$y = \hat{a} + \hat{b} \cdot \frac{1}{x} \quad \xrightarrow{T=1/x} \quad y = \hat{a} + \hat{b} T$$

Pendenza $\hat{b} = \frac{COV(Y, T)}{\sigma_T^2}$

Intercetta $\hat{a} = M(Y) - \hat{b} M(T)$

Richiede $x \neq 0$. Curva a ramo di iperbole, con asintoto orizzontale $y = \hat{a}$ per $x \rightarrow \infty$.

VARIANTE PER L'ORIGINE: $Y = \hat{B}/X$

$$\hat{b} = \frac{M(Y \cdot T)}{M(T^2)} \quad \left(T = \frac{1}{x}\right)$$

Stessa logica della retta per l'origine (§4), applicata alla variabile trasformata $T = 1/x$: si impone $\hat{a} = 0$.

10 Formule di supporto

MEDIE

$$\bar{x} = \frac{\sum x_i}{n} \quad \bar{y} = \frac{\sum y_i}{n} \quad \overline{xy} = \frac{\sum x_i y_i}{n}$$

VARIANZA — METODO ABBREVIATO

$$S_x^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \overline{x^2} - \bar{x}^2$$

$\overline{x^2} = \frac{\sum x_i^2}{n}$. Il metodo abbreviato (scarto dei quadrati dalle medie) evita di calcolare ogni singolo scarto $x_i - \bar{x}$: più veloce a mano.

COVARIANZA — METODO ABBREVIATO

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \overline{xy} - \bar{x} \cdot \bar{y}$$

NOTA SUL METODO

I due metodi (scarti dalla media, oppure medie dei prodotti/quadrati) danno **sempre lo stesso risultato**: sono identità algebriche, non approssimazioni. In pratica conviene il metodo abbreviato ($\overline{x^2} - \bar{x}^2$ e $\overline{xy} - \bar{x}\bar{y}$): richiede solo le somme $\sum x_i$, $\sum y_i$, $\sum x_i y_i$, $\sum x_i^2$ (niente scarti intermedi).



11 Esercizio 1 — confronto tra modelli

CONSEGNA

Si calcolino i seguenti modelli di regressione secondo il metodo dei minimi quadrati e, per ciascuno, la varianza residua: **costante**, **lineare**, **lineare per l'origine**, **polinomiale** ($T = x^2$), **logaritmico** ($T = \ln x$). Si indichi infine quale modello si adatta meglio ai dati.

x_i	6	6	7	8	8
y_i	62	58	55	45	55

STATISTICHE DI BASE (N = 5)

$$\begin{aligned} \bar{x} &= 7 & \bar{y} &= 55 & \overline{xy} &= 381 & \overline{x^2} &= 49,8 & \overline{y^2} &= 3056,6 \\ \sigma_x^2 &= \overline{x^2} - \bar{x}^2 = 0,8 & \sigma_y^2 &= \overline{y^2} - \bar{y}^2 = 31,6 & COV(x, y) &= \overline{xy} - \bar{x}\bar{y} = -4 \end{aligned}$$

MODELLO COSTANTE

$$\hat{a} = M(y) = 55 \quad \sigma_R^2 = \sigma_y^2 = 31,6$$

MODELLO LINEARE

$$\begin{aligned} \hat{b} &= \frac{COV(x, y)}{\sigma_x^2} = \frac{-4}{0,8} = -5 & \hat{a} &= M(y) - \hat{b}M(x) = 55 - (-5 \cdot 7) = 90 \\ & & y &= 90 - 5x \\ \rho^2 &= \frac{COV^2(x, y)}{\sigma_x^2 \sigma_y^2} = \frac{16}{0,8 \cdot 31,6} = 0,6329 & \sigma_R^2 &= \sigma_y^2 (1 - \rho^2) = 31,6 \cdot 0,3671 = 11,60 \end{aligned}$$

LINEARE PER L'ORIGINE

$$\begin{aligned} \hat{c} &= \frac{\overline{xy}}{\overline{x^2}} = \frac{381}{49,8} = 7,65 & y &= 7,65x \\ \sigma_R^2 &= \overline{y^2} - \hat{c}\overline{xy} = 3056,6 - 7,65 \cdot 381 = 141,72 \end{aligned}$$

MODELLO POLINOMIALE ($T = x^2$)

Tabella ausiliaria sulla variabile $T = x^2$:

x_i	6	6	7	8	8	media
$T_i = x_i^2$	36	36	49	64	64	49,8
y_i	62	58	55	45	55	55

$$\begin{aligned} \hat{b} &= \frac{COV(T, y)}{\sigma_T^2} = \frac{-56}{156,96} = -0,3568 & \hat{a} &= M(y) - \hat{b}M(T) = 55 - (-0,3568 \cdot 49,8) = 72,77 \\ & & y &= 72,77 - 0,3568 x^2 & \sigma_R^2 &= 11,62 \end{aligned}$$



MODELLO LOGARITMICO ($T = \ln X$)

$$T = 1,9377 \quad \sigma_T^2 = 0,0166 \quad COV(T, y) = -0,5754$$

$$\hat{b} = \frac{COV(T, y)}{\sigma_T^2} = \frac{-0,5754}{0,0166} = -34,72 \quad \hat{a} = M(y) - \hat{b}M(T) = 55 - (-34,72 \cdot 1,9377) = 122,29$$

$$y = 122,29 - 34,72 \ln x \quad \sigma_R^2 = 11,62$$

CONFRONTO FINALE

MODELLO	EQUAZIONE STIMATA	σ_R^2
Lineare	$y = 90 - 5x$	11,60
Polinomiale (x^2)	$y = 72,77 - 0,3568 x^2$	11,62
Logaritmico	$y = 122,29 - 34,72 \ln x$	11,62
Costante	$y = 55$	31,60
Lineare per l'origine	$y = 7,65 x$	141,72

Il modello **lineare** ha la varianza residua minima e va scelto come modello migliore, anche se polinomiale e logaritmico lo seguono molto da vicino: con soli 5 punti la differenza fra questi tre è nella pratica trascurabile. La **costante** è nettamente peggiore (ignora del tutto la relazione con x), e la **retta per l'origine** è la scelta peggiore: imporre $\hat{a} = 0$ non ha senso quando i dati osservati sono lontani dall'origine ($x \in [6, 8]$, $y \in [45, 62]$).

12 Esercizio 2 — modello esponenziale

CONSEGNA

Si stimi il modello **esponenziale** $y = \hat{a} e^{\hat{b}x}$ con il metodo dei minimi quadrati e se ne calcoli la varianza residua (sulla variabile trasformata $W = \ln y$).

x_i	1	2	3	4	5
y_i	9	12	22	30	58

LINEARIZZAZIONE

Pongo $W = \ln y$:

x_i	1	2	3	4	5	media
$W_i = \ln y_i$	2,1972	2,4849	3,0910	3,4012	4,0604	3,0470

$$\bar{x} = 3 \quad \sigma_x^2 = 2 \quad COV(x, W) = 0,9285$$





STIMA DEI PARAMETRI

$$\hat{b} = \frac{COV(x, W)}{\sigma_x^2} = \frac{0,9285}{2} = 0,4643$$

$$\hat{A} = M(W) - \hat{b}M(x) = 3,0470 - 0,4643 \cdot 3 = 1,6541 \quad \Rightarrow \quad \hat{a} = e^{\hat{A}} = e^{1,6541} = 5,229$$

$$y = 5,229 e^{0,4643 x}$$

VARIANZA RESIDUA

$$\sigma_W^2 = 0,4385 \quad \rho_{xW}^2 = \frac{COV^2(x, W)}{\sigma_x^2 \sigma_W^2} = 0,9831 \quad \sigma_R^2(W) = \sigma_W^2 (1 - \rho^2) = 0,0074$$

ρ^2 molto vicino a 1: sulla scala logaritmica il legame fra x e W è quasi perfettamente lineare, quindi il modello esponenziale coglie bene la crescita dei dati originali.

13 Esercizio 3 — modello potenza

CONSEGNA

Si stimi il modello **potenza** $y = \hat{a} x^{\hat{b}}$ con il metodo dei minimi quadrati e se ne calcoli la varianza residua (sulla variabile trasformata $W = \ln y$).

x_i	1	2	3	4	5
y_i	4	9	22	30	55

LINEARIZZAZIONE

Pongo $T = \ln x$ e $W = \ln y$:

x_i	1	2	3	4	5	media
$T_i = \ln x_i$	0	0,6931	1,0986	1,3863	1,6094	0,9575
$W_i = \ln y_i$	1,3863	2,1972	3,0910	3,4012	4,0073	2,8166

$$\sigma_T^2 = 0,3231 \quad COV(T, W) = 0,5198$$





STIMA DEI PARAMETRI

$$\hat{b} = \frac{COV(T, W)}{\sigma_T^2} = \frac{0,5198}{0,3231} = 1,6088$$

$$\hat{A} = M(W) - \hat{b}M(T) = 2,8166 - 1,6088 \cdot 0,9575 = 1,2762 \quad \Rightarrow \quad \hat{a} = e^{\hat{A}} = 3,583$$

$$y = 3,583 x^{1,6088}$$

VARIANZA RESIDUA

$$\sigma_W^2 = 0,8529 \quad \rho_{TW}^2 = \frac{COV^2(T, W)}{\sigma_T^2 \sigma_W^2} = 0,9805 \quad \sigma_R^2(W) = \sigma_W^2 (1 - \rho^2) = 0,0167$$

L'esponente $\hat{b} = 1,61$ indica una crescita più che proporzionale: un aumento dell'1% di x è associato a un aumento di circa l'1,61% di y .

14 Esercizio 4 — modello inverso

CONSEGNA

Si stimi il modello **inverso** $y = \hat{a} + \hat{b} \cdot \frac{1}{x}$ con il metodo dei minimi quadrati e se ne calcoli la varianza residua.

x_i	1	2	4	5	10
y_i	45	22	15	11	9

LINEARIZZAZIONE

Pongo $T = 1/x$:

x_i	1	2	4	5	10	media
$T_i = 1/x_i$	1	0,5	0,25	0,2	0,1	0,41
y_i	45	22	15	11	9	20,4

$$\sigma_T^2 = 0,1044 \quad COV(T, y) = 4,206$$

STIMA DEI PARAMETRI

$$\hat{b} = \frac{COV(T, y)}{\sigma_T^2} = \frac{4,206}{0,1044} = 40,29$$

$$\hat{a} = M(y) - \hat{b}M(T) = 20,4 - 40,29 \cdot 0,41 = 3,88$$

$$y = 3,88 + \frac{40,29}{x}$$





VARIANZA RESIDUA

$$\sigma_y^2 = 171,04 \quad \rho_{Ty}^2 = \frac{COV^2(T, y)}{\sigma_T^2 \sigma_y^2} = 0,9907 \quad \sigma_R^2 = \sigma_y^2 (1 - \rho^2) = 1,59$$

L'asintoto orizzontale stimato è $y \rightarrow 3,88$ per $x \rightarrow \infty$: coerente con dati che decrescono rapidamente e si stabilizzano su valori piccoli.

CONFRONTO CON LA VARIANTE PER L'ORIGINE

$$\hat{b}_{origine} = \frac{M(y \cdot T)}{M(T^2)} = 46,13 \quad \sigma_R^2 = 7,37$$

Imporre $\hat{a} = 0$ (curva $y = \hat{b}/x$, nessuna asintoto diverso da zero) qui peggiora nettamente l'adattamento (σ_R^2 passa da 1,59 a 7,37): i dati non tendono a zero ma si stabilizzano intorno a $y \approx 4$, quindi il termine costante $\hat{a}$ è necessario.