Università di Catania

Metodi Matematici e Statistici

Dr. Giovanni Nastasi
giovanni.nastasi@unict.it

Dipartimento di Matematica e Informatica
Università degli Studi di Catania, Italy

CdS in Informatica
A.A. 2023-2024

Regressione lineare

Introduzione

Molti problemi in ambito scientifico e ingegneristico consistono nel determinare delle relazioni tra due o più variabili. L'**analisi di regressione** è una tecnica statistica importante per affrontare questo tipo di problemi.

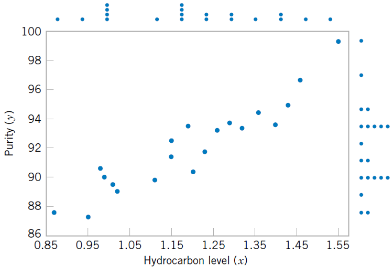
Ad esempio, in un processo chimico è noto che la resa del prodotto è legata alla temperatura del processo. L'analisi di regressione può essere usata per costruire un modello che predica la resa nota la temperatura.

Consideriamo la tabella seguente in cui è riportata la purezza dell'ossigeno in un processo di distillazione chimica e la percentuale di idrocarburi presente nel condensatore principale dell'unità di distillazione.

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Introduzione

Rappresentiamo le coppie (x_i, y_i) su un piano cartesiano.



Anche se nessuna curva passerà esattamente da tutti i punti, vi è una forte indicazione che i punti siano distribuiti in modo casuale attorno a una linea retta. Pertanto, per ogni x consideriamo una v.a. Y e assumiamo che la sua media sia in relazione lineare con x, cioè

$$E[Y] = \beta_0 + \beta_1 x,$$

ove β_0 e β_1 sono detti **coefficienti di regressione**.

Introduzione

Per definire un modello lineare probabilistico assumiamo che il valore atteso di Y sia una funzione lineare di x, ma per ogni fissato valore di x il valore di Y è dato dalla media più una v.a. che rappresenta l'errore, cioè

$$Y = \beta_0 + \beta_1 x + w.$$

Supponiamo che la media di w sia 0 e la sua varianza σ^2. Allora si ha

$$E[Y] = E[\beta_0 + \beta_1 x + w] = E[\beta_0 + \beta_1 x] + E[w] = \beta_0 + \beta_1 x,$$

$$\mathrm{VAR}(Y) = \mathrm{VAR}(\beta_0 + \beta_1 x + w) = \mathrm{VAR}(\beta_0 + \beta_1 x) + \mathrm{VAR}(w) = 0 + \sigma^2 = \sigma^2.$$

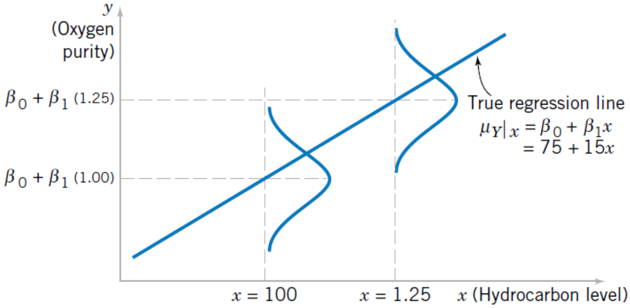
Quindi la retta di regressione è la retta dei valori medi. Cioè per ogni x la corrispondente ordinata $\beta_0 + \beta_1 x$ fornisce il valore medio di Y.

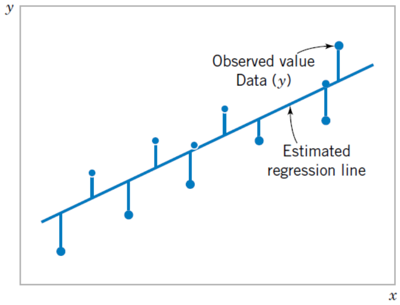
Introduzione

Tornando all'esempio della purezza dell'ossigeno rispetto al livello di idrocarburi, supponiamo che il vero modello di regressione sia

$$y = 75 + 15x$$

e la varianza $\sigma^2 = 2$. Supponiamo che l'errore w segua una distribuzione normale.





Aspetti generali

Supponiamo di avere dati rappresentati da copie di valori (x_j, y_j) con $j = 1, 2, \ldots, n$ con x_j e y_j variabili in sottoinsiemi rispettivamente di $\mathbb{R}^n$ e $\mathbb{R}^m$.

Ci poniamo il problema di capire se esiste un legame funzionale tra le due variabili. Tale questione costituisce un tipico problema di **regressione**. In altri campi, ad esempio in fisica, si preferisce usare termini come problema di **best fit**.

La nomenclatura regressione è storicamente collegata al fatto che i primi modelli furono formulati per un problema in ambito biologico di regressione di geni ed è rimasto in voga in ambito statistico.

In generale quindi si fisserà un legame del tipo

$$y = f(x, \beta_0, \beta_1, \ldots, \beta_k)$$

con le β_i parametri da determinare.

Si noti che il ruolo delle due variabili non è più simmetrico in quanto si sta assumendo che sia x la variabile indipendente, detta **predittore**.

Aspetti generali

Per tenere conto delle fluttuazioni dei dati rispetto al valore teorico, si aggiunge un **rumore statistico** rappresentato da una v.a. w a media nulla, cioè il modello completo soddisfa

$$y_j = f(x_j, \beta_0, \beta_1, \dots, \beta_k) + w_j, \qquad j = 1, 2, \dots, n.$$

Per stimare i parametri β_i si utilizza il metodo dei minimi quadrati, cioè si determinano le β_i in modo che lo scarto quadratico medio

$$\mathcal{S}^2(\beta_0, \beta_1, \dots, \beta_k) = \sum_{j=1}^{n} |w_j|^2 = \sum_{j=1}^{n} |y_j - f(x_j, \beta_0, \beta_2, \dots, \beta_k)|^2$$

sia minimo.

Se f è derivabile rispetto ai β_i si determinano i punti stazionari di $\mathcal{S}^2$ risolvendo il sistema

$$\frac{\partial \mathcal{S}^2}{\partial \beta_i} = 0, \qquad \forall i = 0, 1, \dots, k$$

selezionando poi i punti di minimo, se esistono. Si osservi che $\mathcal{S}^2$ potrebbe non avere punti di minimo o averne più di uno.

Regressione lineare semplice

Studiamo in dettaglio i modelli di **regressione lineare semplice**.

Si considera un insieme di coppie di dati (x_j, y_j) con $j = 1, 2, \ldots, n$ ove le $x_j \in \mathbb{R}$ e le $y_j \in \mathbb{R}$ sono quantità scalari.

Si assume che la relazione funzionale sia di tipo lineare nei suoi parametri, cioè

$$f(x, \beta_0, \beta_1) = \beta_0 + \beta_1 x.$$

Cioè si assume che i dati soddisfano la relazione

$$y_j = \beta_0 + \beta_1 x_j + w_j, \qquad \forall j = 1, 2, \ldots, n,$$

ove le w_j tengono conto della discrepanza tra il valore predetto dal modello, cioè $\beta_0 + \beta_1 x_j$, e il valore effettivo, cioè y_j.

Regressione lineare semplice

Fissato il modello di regressione lineare semplice

$$y_j = \beta_0 + \beta_1 x_j + w_j, \qquad \forall j = 1, 2, \dots, n,$$

si assumono le seguenti ipotesi.

1. Le variabili x_j sono deterministiche mentre le variabili y_j sono aleatorie.
2. Le w_j seguono una legge $N(0, \sigma^2)$, cioè si assume che la legge sia un rumore gaussiano, con la medesima varianza σ^2 per ogni $j = 1, 2, \dots, n$.
3. Le w_i sono tra di loro v.a. indipendenti.

Osservazione. L'indipendenza delle w_i implica l'indipendenza delle y_i.

Regressione lineare semplice

Consideriamo lo scarto quadratico medio nel caso della regressione lineare semplice, cioè

$$\mathcal{S}^2 = \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_j)^2.$$

Derivando rispetto ai parametri β_0 e β_1 e ponendo tali derivate uguali a zero si ottengono le equazioni normali per la retta di regressione, cioè

$$\frac{\partial \mathcal{S}^2}{\partial \beta_0} = -2 \sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_j) = 0$$

$$\frac{\partial \mathcal{S}^2}{\partial \beta_1} = -2 \sum_{j=1}^{n} x_j (y_j - \beta_0 - \beta_1 x_j) = 0$$

Poniamo

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j, \qquad \bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j,$$

Regressione lineare semplice

Dividendo per n, la prima equazione diventa

$$0 = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \beta_0 - \beta_1 x_j\right) = \frac{1}{n}\sum_{j=1}^{n} y_j - \beta_0 \frac{1}{n}\sum_{j=1}^{n} 1 - \beta_1 \frac{1}{n}\sum_{j=1}^{n} x_j$$
$$= \bar{y} - \beta_0 - \beta_1 \bar{x}$$

da cui si ottiene $\beta_0 = \bar{y} - \beta_1 \bar{x}$.

Sostituendo quest'ultima relazione nella seconda equazione e dividendola per n si ottiene

$$0 = \frac{1}{n}\sum_{j=1}^{n} x_j[y_j - (\bar{y} - \beta_1\bar{x}) - \beta_1 x_j] = \frac{1}{n}\sum_{j=1}^{n} x_j y_j - \bar{y}\frac{1}{n}\sum_{j=1}^{n} x_j + \beta_1\bar{x}\frac{1}{n}\sum_{j=1}^{n} x_j - \beta_1\frac{1}{n}\sum_{j=1}^{n} x_j^2$$
$$= \frac{1}{n}\sum_{j=1}^{n} x_j y_j - \bar{x}\bar{y} + \beta_1\bar{x}^2 - \beta_1\frac{1}{n}\sum_{j=1}^{n} x_j^2 = \frac{1}{n}\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\sum_{j=1}^{n}\bar{x}\bar{y} + \beta_1\frac{1}{n}\sum_{j=1}^{n}\bar{x}^2 - \beta_1\frac{1}{n}\sum_{j=1}^{n} x_j^2$$

Regressione lineare semplice

Da cui

$$0 = \frac{1}{n}\sum_{j=1}^{n}(x_j y_j - \bar{x}\bar{y}) - \beta_1 \frac{1}{n}\sum_{j=1}^{n}(x_j^2 - \bar{x}^2).$$

Poniamo

$$\begin{aligned}\bar{\sigma}_{xy} &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) = \frac{1}{n}\sum_{j=1}^{n} x_j y_j - \bar{y}\frac{1}{n}\sum_{j=1}^{n} x_j - \bar{x}\frac{1}{n}\sum_{j=1}^{n} y_j + \frac{1}{n}\sum_{j=1}^{n}\bar{x}\bar{y}\\ &= \frac{1}{n}\sum_{j=1}^{n} x_j y_j - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} = \frac{1}{n}\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\sum_{j=1}^{n}\bar{x}\bar{y} = \frac{1}{n}\sum_{j=1}^{n}(x_j y_j - \bar{x}\bar{y})\end{aligned}$$

e

$$\begin{aligned}\bar{\sigma}_x^2 &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j^2 - 2\bar{x}x_j + \bar{x}^2) = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - 2\bar{x}\frac{1}{n}\sum_{j=1}^{n} x_j + \frac{1}{n}\sum_{j=1}^{n}\bar{x}^2\\ &= \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \bar{x}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j^2 - \bar{x}^2)\end{aligned}$$

Regressione lineare semplice

La seconda equazione diventa

$$0 = \bar{\sigma}_{xy} - \beta_1 \bar{\sigma}_x^2.$$

La soluzione del sistema di partenza è

$$b_0 = \bar{y} - \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}\bar{x}$$
$$b_1 = \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}$$

La retta dei minimi quadrati ha equazione

$$y = b_0 + b_1 x = \left(\bar{y} - \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}\bar{x}\right) + \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}x.$$

Osservazione. La retta dei minimi quadrati passa per il punto di coordinate $(\bar{x}, \bar{y})$.

Regressione lineare semplice

Verifichiamo che (b_0, b_1) è un punto di minimo per $\mathcal{S}^2(\beta_0, \beta_1)$.

$$\frac{\partial^2 \mathcal{S}^2}{\partial \beta_0^2} = 2n, \qquad \frac{\partial^2 \mathcal{S}^2}{\partial \beta_0 \partial \beta_1} = \frac{\partial^2 \mathcal{S}^2}{\partial \beta_1 \partial \beta_0} = 2n\bar{x}, \qquad \frac{\partial^2 \mathcal{S}^2}{\partial \beta_1^2} = 2\sum_{j=1}^{n} x_j^2.$$

Quindi il determinante hessiano vale

$$H = \begin{vmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{j=1}^{n} x_j^2 \end{vmatrix} = 4n\sum_{j=1}^{n} x_j^2 - 4n^2\bar{x}^2 = 4n\sum_{j=1}^{n}(x_j^2 - \bar{x}^2) = 4n^2\bar{\sigma}_x^2 \geq 0$$

per come è stata definita $\bar{\sigma}_x^2$. Inoltre $\bar{\sigma}_x^2 = 0$ se e solo se $x_1 = x_2 = \ldots = x_n$.
Quindi escludendo quest'ultima eventualità si ha $H > 0$.
Inoltre $\frac{\partial^2 \mathcal{S}^2}{\partial \beta_0^2} = 2n > 0$ pertanto il punto critico è un minimo relativo.

Regressione lineare semplice

Proprietà. b_0 e b_1 sono stimatori non distorti rispettivamente di β_0 e β_1.
Dimostrazione. Poiché le x_j sono deterministiche e $w_j \sim N(0, \sigma^2)$ si ha

$$E[b_1] = \frac{1}{\bar{\sigma}_x^2}\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})E[y_j - \bar{y}] = \frac{1}{\bar{\sigma}_x^2}\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(E[y_j] - E[\bar{y}]).$$

Calcoliamo

$$E[y_j] = E[\beta_0 + \beta_1 x_j + w_j] = \beta_0 + \beta_1 x_j$$

e

$$E[\bar{y}] = \frac{1}{n}\sum_{j=1}^{n}E[y_j] = \frac{1}{n}\sum_{j=1}^{n}(\beta_0 + \beta_1 x_j) = \beta_0 + \beta_1\frac{1}{n}\sum_{j=1}^{n}x_j = \beta_0 + \beta_1\bar{x}.$$

Quindi si ha

$$E[b_1] = \frac{1}{\bar{\sigma}_x^2}\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(\beta_0 + \beta_1 x_j - \beta_0 - \beta_1\bar{x}) = \frac{1}{\bar{\sigma}_x^2}\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})\beta_1(x_j - \bar{x})$$

$$= \frac{\beta_1}{\bar{\sigma}_x^2}\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \beta_1.$$

Quindi b_1 è uno stimatore non distorto di β_1.

Regressione lineare semplice

Dimostrazione (continuo). Calcoliamo

$$E[b_0] = E[\bar{y} - b_1\bar{x}] = E[\bar{y}] - \bar{x}E[b_1] = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0.$$

Quindi b_0 è uno stimatore non distorto di β_0.

Proprietà. La varianza degli stimatori b_0 e b_1 è data da

$$\mathrm{VAR}(b_0) = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{\bar{\sigma}_x^2}\right), \qquad \mathrm{VAR}(b_1) = \frac{\sigma^2}{n\bar{\sigma}_x^2}.$$

Dimostrazione. Poniamo

$$v_i = \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

Allora

$$\bar{v} = \frac{1}{n}\sum_{i=1}^n v_i = \frac{1}{\sum_{j=1}^n (x_j - \bar{x})^2}\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Regressione lineare semplice

Dimostrazione (continuo). Inoltre

$$\sum_{i=1}^{n} v_i^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{j=1}^{n}(x_j - \bar{x})^2\right)^2} = \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{1}{n\bar{\sigma}_x^2}.$$

Allora

$$b_1 = \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \sum_{i=1}^{n} v_i(y_i - \bar{y}) = \sum_{i=1}^{n} v_i y_i - \bar{y}\sum_{i=1}^{n} v_i = \sum_{i=1}^{n} v_i y_i.$$

Osservando che

$$\mathrm{VAR}(y_i) = \mathrm{VAR}(\beta_0 + \beta_1 x_i + w_i) = \mathrm{VAR}(w_i) = \sigma^2$$

e che le y_i sono indipendenti, segue che

$$\mathrm{VAR}(b_1) = \sum_{i=1}^{n} \mathrm{VAR}(v_i y_i) = \sum_{i=1}^{n} v_i^2 \mathrm{VAR}(y_i) = \sigma^2 \sum_{i=1}^{n} v_i^2 = \frac{\sigma^2}{n\bar{\sigma}_x^2}.$$

Regressione lineare semplice

Dimostrazione (continuo). Per quanto riguarda b_0 si ha

$$\mathrm{VAR}(b_0) = \mathrm{VAR}(\bar{y} - b_1\bar{x}) = \mathrm{VAR}(\bar{y}) + \bar{x}^2\mathrm{VAR}(b_1) - 2\bar{x}\mathrm{COV}(\bar{y}, b_1).$$

Calcoliamo i vari termini.
Poiché le y_i sono indipendenti

$$\mathrm{VAR}(\bar{y}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{VAR}(y_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Sfruttando la notazione precedentemente introdotta

$$\mathrm{COV}(\bar{y}, b_1) = \mathrm{COV}\left(\bar{y}, \sum_{i=1}^{n} v_i y_i\right) = \sum_{i=1}^{n} v_i \mathrm{COV}(\bar{y}, y_i) = \sum_{i=1}^{n} v_i \mathrm{COV}\left(\frac{1}{n}\sum_{j=1}^{n} y_j, y_i\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} v_i \mathrm{COV}(y_j, y_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} v_i\sigma^2\delta_{ji} = \frac{1}{n}\sum_{i=1}^{n} v_i\sigma^2 = \sigma^2\bar{v} = 0.$$

Regressione lineare semplice

Dimostrazione (continuo). Sostituendo nella relazione per $\mathrm{VAR}(b_0)$ si ha

$$\mathrm{VAR}(b_0) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{n\bar{\sigma}_x^2} = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{\bar{\sigma}_x^2}\right).$$

Proprietà. Poiché b_0 e b_1 sono funzioni lineari delle y_i si verifica immediatamente che seguono una legge normale e pertanto si ha:

$$b_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{\bar{\sigma}_x^2}\right)\right),$$
$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{n\bar{\sigma}_x^2}\right).$$

Regressione lineare semplice

Vogliamo infine stimare σ^2.

Introduciamo i **valori stimati**

$$\hat{y}_i = b_0 + b_1 x_i, \qquad i = 1, 2, \ldots, n$$

e i **resudui**

$$r_i = y_i - \hat{y}_i, \qquad i = 1, 2, \ldots, n.$$

Proprietà. Posto

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2$$

allora

$$X := \frac{s^2}{\sigma^2}(n-2) \sim \chi^2(n-2)$$

e X è indipendente da b_0 e da b_1.

Regressione lineare semplice

Corollario. s^2 è uno stimatore non distorto di σ^2.
Dimostrazione. Infatti
$$E[X] = \frac{n-2}{\sigma^2} E[s^2]$$
ma poiché il valore atteso di una legge chi-quadro è pari al numero di gradi di libertà si ha che $E[X] = n-2$, da cui $E[s^2] = \sigma^2$.

Corollario.
$$T_0 = \frac{b_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n\bar{\sigma}_x^2}}} \sim t(n-2)$$
$$T_1 = \sqrt{n}\frac{b_1 - \beta_1}{s}\bar{\sigma}_x \sim t(n-2)$$

Regressione lineare semplice

Grazie all'ultimo corollario si ricavano gli intervalli di confidenza di livello $1-\alpha$ per β_0 e β_1, cioè

$$\left[b_0 - s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{n\bar{\sigma}_x^2}}t_{1-\alpha/2}(n-2),\, b_0 + s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{n\bar{\sigma}_x^2}}t_{1-\alpha/2}(n-2)\right],$$

$$\left[b_1 - \frac{s}{\bar{\sigma}_x\sqrt{n}}t_{1-\alpha/2}(n-2),\, b_1 + \frac{s}{\bar{\sigma}_x\sqrt{n}}t_{1-\alpha/2}(n-2)\right].$$

Regressione lineare semplice

Osservazione. Se b_1 fosse troppo vicino a zero si potrebbe sospettare che non vi sia in effetti dipendenza lineare tra y e x. Si può procedere formulando il seguente **test di indipendenza**.

Si vuole testare l'ipotesi nulla

$$H_0 \; : \; \beta_1 = 0$$

contro l'ipotesi alternativa

$$\begin{aligned} H_1 \; &: \; \beta_1 \neq 0 && \text{test bilatero} \\ H_1 \; &: \; \beta_1 < 0 && \text{test unilatero a sinistra} \\ H_1 \; &: \; \beta_1 > 0 && \text{test unilatero a destra} \end{aligned}$$

Fissato un livello di significatività $\alpha \in]0,1[$, la zona di rigetto è data da

$$\begin{aligned} &\left| \sqrt{n} \frac{b_1}{s} \bar{\sigma}_x \right| \geq t_{1-\alpha/2}(n-2) && \text{test bilatero} \\ &\sqrt{n} \frac{b_1}{s} \bar{\sigma}_x < t_{\alpha}(n-2) && \text{test unilatero a sinistra} \\ &\sqrt{n} \frac{b_1}{s} \bar{\sigma}_x > t_{1-\alpha}(n-2) && \text{test unilatero a destra} \end{aligned}$$

Regressione lineare semplice

Osservazione. Un modello di regressione può essere utilizzato per stimare y in corrispondenza a valori x^* diversi dalle x_i assegnate.

Consideriamo il caso della regressione lineare semplice. Sia x^* il valore del predittore in cui si vuole calcolare y. La scelta naturale è stimare $y(x^*)$ con

$$y^* = b_0 + b_1 x^*.$$

D'altra parte

$$y(x^*) = \beta_0 + \beta_1 x^* + w.$$

Assumiamo che $w \sim N(0, \sigma^2)$ e che w sia indipendente dalle w_i per $i = 1, 2, \ldots, n$.

Valutiamo la media e la varianza dello scarto $y(x^*) - y^*$

$$y(x^*) - y^* = w + \beta_0 - b_0 + (\beta_1 - b_1)x^*.$$

Le v.a. w e $\beta_0 - b_0 + (\beta_1 - b_1)x^*$ sono entrambe normali e indipendenti. Inoltre $E[w] = 0$ e

$$E[\beta_0 - b_0 + (\beta_1 - b_1)x^*] = E[\beta_0 - b_0] + x^* E[\beta_1 - b_1] = 0.$$

Regressione lineare semplice

Quindi $E[y(x^*) - y^*] = 0$. Per quanto riguarda la varianza

$$\begin{aligned}\mathrm{VAR}(y(x^*) - y^*) &= \mathrm{VAR}(w) + \mathrm{VAR}(\beta_0 - b_0 + (\beta_1 - b_1)x^*) = \sigma^2 + \mathrm{VAR}(b_0 + b_1 x^*) \\ &= \sigma^2 + \mathrm{VAR}(b_0) + 2x^*\mathrm{COV}(b_0, b_1) + (x^*)^2\mathrm{VAR}(b_1).\end{aligned}$$

Ricordiamo che $\mathrm{VAR}(b_0)$ e $\mathrm{VAR}(b_1)$ sono state calcolate invece

$$\mathrm{COV}(b_0, b_1) = \mathrm{COV}(\bar{y} - b_1\bar{x}, b_1) = \mathrm{COV}(\bar{y}, b_1) - \bar{x}\mathrm{COV}(b_1, b_1) = -\bar{x}\mathrm{VAR}(b_1),$$

perché, come visto in precedenza, $\mathrm{COV}(\bar{y}, b_1) = 0$.

Quindi si ha

$$\begin{aligned}\mathrm{VAR}(y(x^*) - y^*) &= \sigma^2 + \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{\bar{\sigma}_x^2}\right) + (-2\bar{x}x^* + (x^*)^2)\frac{\sigma^2}{n\bar{\sigma}_x^2} \\ &= \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{n\bar{\sigma}_x^2}(\bar{x}^2 - 2\bar{x}x^* + (x^*)^2) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{n\bar{\sigma}_x^2}\right).\end{aligned}$$

Regressione lineare semplice

Proprietà. Per la previsione $y(x^*)$ vale

$$\frac{y(x^*) - b_0 - b_1 x^*}{s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{n\bar{\sigma}_x^2}}} \sim t(n-2).$$

Un intervallo di confidenza di livello $1-\alpha$ per la previsione $y(x^*)$ è

$$\left[b_0 + b_1 x^* - \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{n\bar{\sigma}_x^2}}\, t_{1-\alpha/2}(n-2),\ b_0 + b_1 x^* + \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{n\bar{\sigma}_x^2}}\, t_{1-\alpha/2}(n-2)\right].$$

Proprietà. Per il valore medio di $y(x^*)$, cioè per $E[y(x^*)] = \beta_0 + \beta_1 x^*$ vale

$$\frac{b_0 + b_1 x^* - (\beta_0 + \beta_1 x^*)}{s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{n\bar{\sigma}_x^2}}} \sim t(n-2).$$

Un intervallo di confidenza di livello $1-\alpha$ per il valore medio della previsione è

$$\left[b_0 + b_1 x^* - \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{n\bar{\sigma}_x^2}}\, t_{1-\alpha/2}(n-2),\ b_0 + b_1 x^* + \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{n\bar{\sigma}_x^2}}\, t_{1-\alpha/2}(n-2)\right].$$

Proprietà dei residui

Dallo studio dei residui è possibile ottenere importanti indicazioni sulla validità del modello di regressione lineare adottato.

Proprietà. Se i coefficienti della retta di regressione lineare sono quelli dati dal metodo dei minimi quadrati allora

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i = 0,$$

$$\bar{\sigma}_{xr} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(r_i - \bar{r}) = 0.$$

Dimostrazione.

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = \bar{y} - b_0 - b_1\bar{x} = 0,$$

perché abbiamo visto che la retta di regressione passa per $(\bar{x}, \bar{y})$.

Proprietà dei residui

Dimostrazione (continuo). Dall'equazione della retta di regressione $y = b_0 + b_1 x$ e dalla relazione $b_0 = \bar{y} - b_1 \bar{x}$ segue che $y - \bar{y} = b_1(x - \bar{x})$, da cui

$$\bar{\sigma}_{xr} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y} - b_1(x_i - \bar{x}))$$
$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - b_1 \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \bar{\sigma}_{xy} - b_1 \bar{\sigma}_x^2 = 0$$

perché abbiamo visto che $b_1 = \frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}$.

Osservazione. Il fatto che $\bar{\sigma}_{xr} = 0$ implica che non vi è correlazione tra gli scarti e la variabile x in accordo con l'ipotesi che σ^2 sia indipendente dall'indice i.

Proprietà dei residui

Proprietà. La varianza degli scarti è data da

$$\sigma_r^2 = \sigma_y^2 - \frac{\bar{\sigma}_{xy}^2}{\bar{\sigma}_x^2}, \qquad \text{con } \sigma_r^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2$$

e ove σ_y^2 è la varianza delle y_i, cioè $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Dimostrazione. Poiché $\bar{r} = 0$, si ha

$$\sigma_r^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - \bar{y} - b_1(x_i - \bar{x})]^2 = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \bar{y})^2 - 2b_1(y_i - \bar{y})(x_i - \bar{x}) + b_1^2(x_i - \bar{x})^2]$$

$$= \sigma_y^2 - 2b_1\bar{\sigma}_{xy} + b_1^2\bar{\sigma}_x^2 = \sigma_y^2 - 2\frac{\bar{\sigma}_{xy}}{\bar{\sigma}_x^2}\bar{\sigma}_{xy} + \frac{\bar{\sigma}_{xy}^2}{\bar{\sigma}_x^4}\bar{\sigma}_x^2 = \sigma_y^2 - \frac{\bar{\sigma}_{xy}^2}{\bar{\sigma}_x^2}.$$

Proprietà dei residui

Proprietà. La varianza dei valori stimati $f(x_i) = \hat{y}_i$ è data da

$$\sigma^2_{f(x)} = \frac{\bar{\sigma}^2_{xy}}{\bar{\sigma}^2_x}.$$

Dimostrazione. Avendo osservato che la retta di regressione passa per il punto $(\bar{x}, \bar{y})$, si ha che $\bar{y} = b_0 + b_1\bar{x}$. Pertanto

$$\sigma^2_{f(x)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(b_0 + b_1 x_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} b_1^2(x_i - \bar{x})^2 = b_1^2\bar{\sigma}^2_x = \frac{\bar{\sigma}^2_{xy}}{\bar{\sigma}^4_x}\bar{\sigma}^2_x.$$

Proprietà dei residui

Corollario.

$$\sigma_y^2 = \sigma_{f(x)}^2 + \sigma_r^2.$$

Osservazione. Il precedente corollario si può interpretare dicendo che la varianza dei dati è la somma della *varianza spiegata dal modello*, cioè $\sigma_{f(x)}^2$, e della *varianza non spiegata dal modello*, σ_r^2, cioè quella dovuta ai residui.

Definizione. Il rapporto

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma_y^2} = \frac{\bar{\sigma}_{xy}^2}{\bar{\sigma}_x^2 \bar{\sigma}_y^2}$$

dicesi **coefficiente di determinazione**.

Osservazione. Il coefficiente di determinazione rappresenta la frazione di varianza spiegata dal modello ed è uguale al quadrato del coefficiente di correlazione lineare $\rho(X, Y)$. Pertanto $0 \leq R^2 \leq 1$. Ci si aspetta, quindi, che più R^2 è prossimo a uno e più il modello di regressione lineare usato è adeguato. Tuttavia, tale conclusione è da prendere con cautela.

Esercizio

Il peso corporeo e la pressione sistolica del sangue di 26 individui maschi selezionati in modo casuale nella fascia d'età che va da 25 a 30 anni sono mostrati in tabella. Assumiamo che il peso e la pressione sanguigna siano normalmente distribuiti.

1. Si determini la retta di regressione.
2. Si calcolino gli intervalli di confidenza per i coefficienti di regressione.
3. Si testi la significatività della regressione usando $\alpha = 0.05$.
4. Si calcoli il coefficiente di determinazione.

Subject	Weight	Systolic BP	Subject	Weight	Systolic BP
1	165	130	14	172	153
2	167	133	15	159	128
3	180	150	16	168	132
4	155	128	17	174	149
5	212	151	18	183	158
6	175	146	19	215	150
7	190	150	20	195	163
8	210	140	21	180	156
9	200	148	22	143	124
10	149	125	23	240	170
11	158	133	24	235	165
12	169	135	25	192	160
13	170	150	26	187	159

Esercizio

L'ossigeno consumato da una persona che cammina è funzione della sua velocità. La seguente tabella riporta il volume di ossigeno consumato a varie velocità di cammino. Ipotizzando una relazione lineare, scrivere l'equazione della retta di regressione. Si testi la significatività della regressione usando $\alpha = 0.05$.

Velocità (km/h)	Ossigeno (l/h)
0	19,5
1	22,1
2	24,3
3	25,7
4	26,1
5	28,5
6	30,0
7	32,1
8	32,7
9	32,7
10	35,0

Regressione lineare multipla

Regressione lineare multipla

Vogliamo estendere lo studio sulla regressione lineare semplice al caso in cui si hanno più predittori.

Scriviamo il modello di regressione lineare semplice in forma vettoriale. Ponendo

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \qquad \mathbf{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix},$$

possiamo formulare il modello di regressione lineare semplice in forma compatta come

$$\mathbf{y} = \beta_0 \mathbf{e} + \beta_1 \mathbf{x} + \mathbf{w}.$$

La ricerca degli stimatori di β_0 e β_1 con il metodo dei minimi quadrati equivale a cercare un vettore $\hat{\mathbf{y}} = b_0 \mathbf{e} + b_1 \mathbf{x}$ tale che $|\mathbf{y} - \hat{\mathbf{y}}|^2$ sia minimo.

Regressione lineare multipla

Generalizzando, un modello di regressione lineare multipla è un modello del tipo

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots + \beta_k \mathbf{x}_k + \mathbf{w},$$

con $\mathbf{x}_i$, $\mathbf{w} \in \mathbb{R}^n$ e $k < n$.

Si assumono le seguenti ipotesi.

1. I vettori $\mathbf{x}_i$, $i = 1, 2, \ldots, k$ sono deterministici.
2. Si abbia il termine costante, cioè $\mathbf{x}_1 = \mathbf{e} \in \mathbb{R}^n$.
3. I vettori $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ sono linearmente indipendenti.
4. $\mathbf{w} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$, ove $\mathbf{0}$ è il vettore nullo di $\mathbb{R}^n$ e $\mathbf{I}$ è la matrice identità di ordine n.

Osservazione. Si ha che $\mathbf{y} \sim N(\beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots + \beta_k \mathbf{x}_k, \sigma^2 \mathbf{I})$.

Regressione lineare multipla

Stima dei parametri di regressione.

Sia $X \in \mathbb{R}^{n \times k}$ la matrice le cui colonne sono le componenti dei vettori $\mathbf{x}_i$, con $i = 1, 2, \dots, k$.

Si può dimostrare che la stima dei parametri β_i, con $i = 1, 2, \dots, k$, si ottiene calcolando prima la matrice

$$(X^T X)^{-1} X^T,$$

detta **matrice pseudo-inversa** di X e poi

$$\hat{\mathbf{y}} = X\mathbf{b}, \qquad \text{con} \quad \mathbf{b} = (X^T X)^{-1} X^T \mathbf{y},$$

ove $\mathbf{b}$ rappresenta lo stimatore di $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)^T$.

Regressione lineare multipla

Proprietà. $\mathbf{b}$ è uno stimatore non distorto di $\boldsymbol{\beta}$.

Proprietà. $\mathbf{b}$ segue una legge normale multivariata,

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1}).$$

Proprietà. Le componenti di $\mathbf{b}$ hanno legge normale,

$$b_i \sim N(\beta_i, \sigma^2 m_{ii}), \qquad i = 1, 2, \ldots, k,$$

ove $m_{ii} = (X^T X)^{-1}_{ii}$.

Regressione lineare multipla

Analisi dei residui.

Sia $\hat{\mathbf{y}} = X\mathbf{b}$ il vettore dei valori stimati e $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ il vettore dei residui.

Proprietà. Le variabili aleatorie $\hat{\mathbf{y}}$ e $\mathbf{r}$ sono indipendenti. Inoltre, posto

$$s^2 = \frac{1}{n-k}\mathbf{r}^2 = \frac{1}{n-k}\sum_{i=1}^{n} r_i^2$$

si ha

$$\frac{s^2}{\sigma^2}(n-k) \sim \chi^2(n-k).$$

Proprietà. b_i e s^2 sono variabili aleatorie indipendenti e quindi

$$\frac{b_i - \beta_i}{s\sqrt{m_{ii}}} \sim t(n-k),$$

ove $m_{ii} = (X^T X)^{-1}_{ii}$.

Regressione lineare multipla

Osservazione. Un intervallo di confidenza di livello $1-\alpha$ per β_i è dato da

$$\left[b_i - s\sqrt{m_{ii}}t_{1-\alpha/2}(n-k),\, b_i + s\sqrt{m_{ii}}t_{1-\alpha/2}(n-k)\right].$$

Inoltre se consideriamo il test

$$H_0 \,:\, \beta_i = b_i^{(0)}, \qquad H_1 \,:\, \beta_i \neq b_i^{(0)}$$

una regione di rigetto è data da

$$\left|\frac{b_i - b_i^{(0)}}{s\sqrt{m_{ii}}}\right| \geq t_{1-\alpha/2}(n-k).$$

Regressione lineare multipla

Coefficiente di determinazione.
Anche nel caso di regressione lineare multipla si può introdurre il coefficiente di determinazione

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

che soddisfa ancora la condizione $0 \leq R^2 \leq 1$ e rappresenta la frazione di varianza spiegata dal modello.

Osservazione. Se manca il fattore costante non è più garantita la limitazione $0 \leq R^2 \leq 1$ e R^2 non rappresenta più la frazione di varianza spiegata dal modello.

Regressione lineare multipla

Regressione polinomiale.

I modelli di regressione polinomiale si possono sempre ricondurre ad un modello di regressione lineare multipla.

Ad esempio, consideriamo il caso

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + w$$

con $w \sim N(0, \sigma^2)$. Con il cambiamento di variabile

$$\tilde{x}_1 = 1, \qquad \tilde{x}_2 = x, \qquad \tilde{x}_3 = x^2$$

il modello diventa

$$y = \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \beta_3 \tilde{x}_3 + w.$$

Regressione lineare multipla

Modelli riconducibili a un modello di regressione lineare multipla.

Anche altri modelli di regressione possono essere ricondotti con opportuni cambiamenti di variabile a un modello di regressione lineare multipla.

Esempi.

- $y = e^{\beta_0 + \beta_1 x}$ si può trasformare in $\log y = \beta_0 + \beta_1 x$ e quindi tramite l'ulteriore cambiamento di variabile $\tilde{y} = \log y$ si ottiene un modello di regressione lineare semplice.
- $y = e^{\beta_0} x^{\beta_1}$ con la trasformazione $\tilde{y} = \log y$, $\tilde{x} = \log x$ diventa un modello di regressione lineare semplice.
- $y = \frac{1}{\beta_0 + \beta_1 x}$ viene ricondotto ad un problema di regressione lineare semplice tramite $\tilde{y} = \frac{1}{y}$.

Esercizio

Consideriamo i dati di tabella. Formuliamo degli appropriati modelli di regressione.

x	y	x	y
0.32	1.60	6.65	16.47
2.83	5.30	8.96	16.25
3.94	11.88	11.45	37.30
6.52	15.28	0.96	0.48
7.51	18.19	3.80	0.06
11.43	32.55	5.91	9.91
0.96	1.23	6.83	14.96
3.32	0.32	9.56	21.82
4.62	8.03	12.02	33.19

Riferimenti bibliografici

1. V. Romano, Metodi matematici per i corsi di ingegneria, Città Studi, 2018.
2. P. Baldi, Calcolo delle probabilità e statistica, Mc Graw-Hill, Milano, 1992.
3. R. Scozzafava, Incertezza e probabilità, Zanichelli, 2001.
4. D. C. Montgomery, G. C. Runger, Applied statistics and probability for engineers, 7th Edition, J. Wiley, 2018.