Regressione[1]

Nel modello di regressione lineare si assume una relazione di tipo lineare tra il valore medio della variabile dipendente Y e quello della variabile indipendente X per cui

$$E\{Y|X\} = \eta_{Y|X} = a + bx$$

Il modello si scrive come

$$y = a + bx + \varepsilon$$

Dove ε è l'errore gaussiano con varianza σ^2 e valore medio nullo. Dati due vettori delle osservazioni x e y, per l'elemento *i*-esimo del vettore si può scrivere:

$$y_i = a + bx_i + \varepsilon_i$$

per ogni osservazione ε_i si può descrivere come una variabile aleatoria gaussiana a valore medio nullo.

Nel modello omoschedastico la varianza dell'errore non dipende da *i* ed è quindi pari a σ^2.

Inoltre, i termini di errore al variare di *i* sono tra loro incorrelati per cui la covarianza tra due variabili di errore ε_i e ε_j, $cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$.

Quindi il modello di regressione prevede che ad ogni valore della variabile indipendente *x* corrisponda una popolazione di possibili valori della variabile dipendente *y*. Per ogni *x* questa popolazione è centrata attorno al valore dato dalla retta di regressione dato da *a*+*bx* con un andamento di tipo gaussiano. La varianza di tale popolazione è costante al variare di *x*. Quindi la variabile che descrive lo scostamento della variabile indipendente da quella indipendente al variare di *x* appartiene alla famiglia $N(0, \sigma^2)$.

Stima dei parametri a e b

Data una serie di *N* misure delle variabili dipendente e indipendente, vediamo come determinare i coefficienti del modello della retta di regressione. Ovvero i parametri *a* e *b* della retta che meglio approssima i dati. Data la *i*-esima misura, x_i, i parametri stimati $\widehat{a}$ e $\widehat{b}$ forniscono il valore approssimato della *i*-esima misura della variabile dipendente y_i.

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

l'errore tra la stima e la misura reale, o residuo, è

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

I parametri vengono stimati in modo da minimizzare la somma dei quadrati dei residui, ovvero+

$$Q = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{a} + \hat{b}x_i\right)^2$$

I parametri $\widehat{a}$ e $\widehat{b}$ sono tali da annullare le seguenti equazioni $\frac{\partial Q}{\partial \hat{a}} = 0$ e $\frac{\partial Q}{\partial \hat{b}} = 0$

Dalla prima equazione si trova che

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \hat{a} + \hat{b}\frac{1}{N}\sum_{i=1}^{N} x_i$$

ovvero

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

dove $\bar{y}$ e $\bar{x}$ sono i valori medi delle misure. Quindi la retta passa per i centroidi delle misure.

La seconda condizione porta alla seguente equazione

$$\sum_{i=1}^{N} x_i y_i = \hat{a}\sum_{i=1}^{N} x_i \; + \hat{b}\frac{1}{N}\sum_{i=1}^{N} x_i^2$$

Le due precedenti equazioni si chiamano equazioni normali e conducono alla seguente stima dei parametri $\widehat{a}$ e $\widehat{b}$

$$\hat{a} = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i y_i}{N}}{\sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N}}{\sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}}$$

dove $\sum_{i=1}^{N} x_i y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N}$ è pari a *N* volte la covarianza campionaria tra variabile dipendente e indipendente, mentre $\sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}$ è pari a *N* volte la varianza campionaria della variabile indipendente. Quindi il parametro $\widehat{b}$ si può ottenere come

$$\hat{b} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Inferenza su b
(può essere analogamente sviluppato lo stesso tema relativamente ad a)

La stima ottenuta di *b* è ovviamente dipendente dal particolare campione delle misure di *x* e di *y*. Se volessimo quindi considerare la relazione tra le variabili aleatorie *X* e *Y* e avere una stima dell'intervallo di confidenza di *b* o testare l'ipotesi che sia uguale a zero, dovremmo avere informazioni circa la sua distribuzione.

Nell'ipotesi che l'errore del modello di regressione sia gaussiano a valore medio nullo e varianza σ, è possibile dimostrare che lo stimatore $\hat{b}$ ha una distribuzione gaussiana con valore medio e varianza dati rispettivamente da:

$$E\{\hat{b}\} = b$$

$$\sigma^2(\hat{b}) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

La dimostrazione di queste affermazioni è omessa. Possiamo dire che questa si basa sul fatto che la stima $\hat{b}$ è una combinazione lineare delle osservazioni y_i, che a loro volta sono distribuite con una legge normale. Questa ultima asserzione discende direttamente dal modello di regressione lineare con ipotesi di rumore normale.

Nota sulla stima di σ^2

Non avendo a disposizione il valore teorico di σ^2, questo va stimato dai dati. In particolare, è possibile creare uno stimatore di tale varianza sia considerando una popolazione, ovvero il totale delle misure y_i rispetto al valore medio $\bar{y}$, oppure gli scostamenti di y_i rispetto alla valore sulla retta di regressione $\hat{y}_i$. In questo secondo caso, si considerano i residui

$$e_i = y_i - \hat{y}_i$$

Si definisce la somma dei quadrati degli errori (o SSE, error sum of squares)

$$SSE = \sum_{i=1}^{N} e_i^2$$

Tale grandezza ha *N*-2 gradi di libertà, visto che è stato necessario stimare i due parametri del modello per arrivare a determinarla.
Quindi la media corrispondente (media dei quadrati dell'errore o error mean square) è definita come

$$MSE = \frac{SSE}{N-2}$$

ed è uno stimatore non polarizzato di σ^2 e quindi $E\{MSE\} = \sigma^2$.

Come conseguenza, consideriamo la grandezza

$$\frac{\hat{b} - b}{s(\hat{b})}$$

al denominatore abbiamo lo stimatore della deviazione standard del parametro $\hat{b}$, $\sigma(\hat{b})$. Tale stimatore può essere ottenuto sostituendo la stima di σ^2 alla equazione già trovata $\sigma^2(\hat{b}) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$ per cui si ottiene

$$s^2(\hat{b}) = \frac{MSE}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

dal quale si ricava $s(\hat{b})$.

È possibile dimostrare che la grandezza $s(\hat{b})$ è distribuita come una *t* di Student con *N*-2 gradi di libertà. È di conseguenza immediata la stima dell'intervallo di confidenza.

Analogamente, è immediata la verifica dell'ipotesi nulla, con un livello di significatività pari ad α,

$$H_0: b = 0$$

in alternativa all'ipotesi

$$H_a: b \neq 0$$

In questo ultimo caso infatti si calcola il valore

$$t^* = \frac{\hat{b}}{s(\hat{b})}$$

e si confronta $|t^*|$ con il valore critico, che separa regione di accettazione e rifiuto

$$t(1-\alpha/2; N-2)$$

Analisi della varianza del modello di regressione [2]

Questo approccio all'analisi del modello di regressione permetterà di stimare la significatività del modello e fornire parametri descrittivi. Questa analisi si basa sulla partizione della variabilità totale delle osservazioni in quella spiegata, o descritta, dal modello di regressione e la rimanente, ovvero quella relativa alla differenza tra le osservazioni e la retta di regressione.
In figura si mostra tale ripartizione per una singola misura: si evidenzia come la differenza tra la misura *i*-esima e il valore medio della popolazione globale delle misure possa essere separata nei due contributi, uno relativo al modello di regressione e l'altro relativo al residuo tra la misura e il modello stesso. Quindi:

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

Analogamente si può dimostrare che la devianza totale

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$$

dove il primo termine è definito come somma totale dei quadrati (SSTO, total sum of squares)

$$SSTO = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

La grandezza somma dei quadrati degli errori è già stata definita come

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

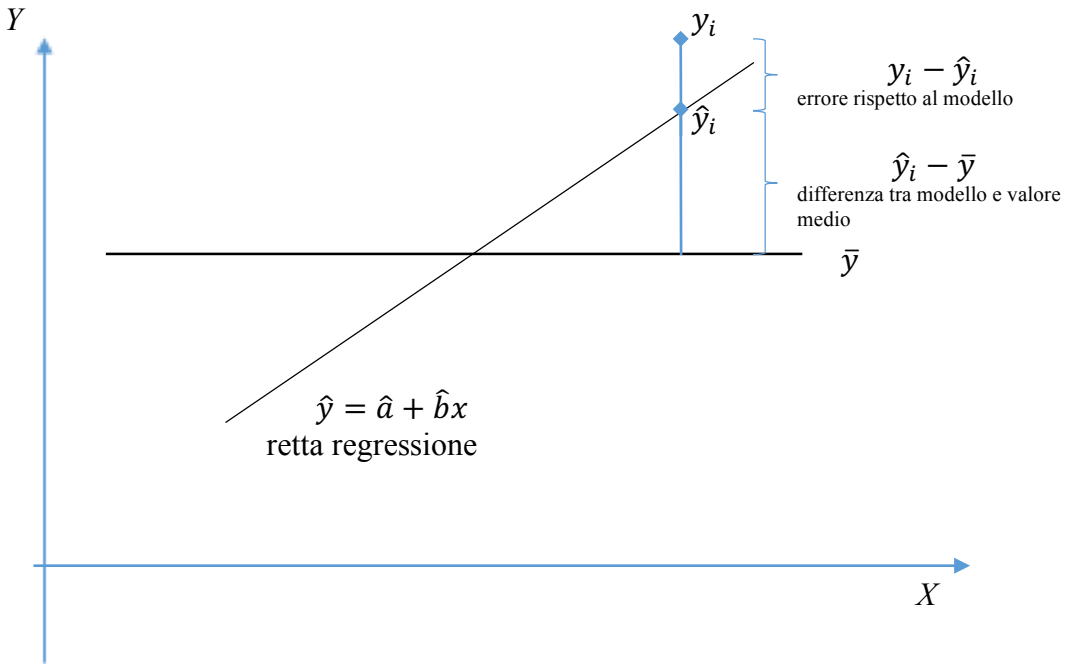
mentre la variabilità dovuta al modello di regressione è definita come somma dei quadrati della regressione (SSR, regression sum of squares)

$$SSR = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

Quindi

$$SSTO = SSE + SSR$$

Maggiore è la grandezza SSR rispetto a SSTO maggiore è la variabilità nelle osservazioni che può essere descritta dal modello di regressione, ovvero dalla relazione lineare tra *X* e *Y*.
Possiamo vedere SSTO come l'incertezza nel predire il valore di *y* senza considerare il contributo informativo di *x*. In modo analogo, la grandezza SSE misura la variabilità della misura di *y* rispetto al valore dato dal modello di regressione e quindi è legata all'incertezza rispetto al valore indicato dal modello di regressione.



Il coefficiente di determinazione fornisce una misura dell'entità della variabilità nelle misure che è spiegata dal modello ed è definito come

$$r^2 = \frac{SSTO-SSE}{SSTO} = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Essendo SSE≤SSTO, il coefficiente di determinazione parziale è compreso tra 0 e 1.
Esso può essere interpretato come la riduzione della variabilità associata all'introduzione della variabile indipendente.

I casi limite corrispondono residui del modello di regressione nulli, ovvero tutte le osservazioni completamente sulla retta di regressione. In questo caso $r^2 = 1$.
L'altro caso è quello di retta di regressione «piatta». In tale condizione la variabile *X* non fornisce nessuna informazione sui possibili valori di *Y*. In questo caso $r^2 = 0$.

Il coefficiente di determinazione è il quadrato del coefficiente di correlazione $r = \pm\sqrt{r^2}$. Il coefficiente di correlazione è compreso tra -1 e 1. Il segno dipende dal segno della pendenza della retta di regressione.

Gradi di liberta e tabella ANOVA

Analizzeremo adesso i gradi di libertà associati con le diverse misure delle variabilità osservate e quindi la tabella ANOVA per il modello di regressione. Questo ci permetterà di individuare anche una statistica legata alla significatività del modello di regressione.

La grandezza SSTO ha *N*-1 gradi di libertà. Questo è legato al fatto che la somma delle grandezze $\hat{y}_i - \bar{y}$ deve essere pari a zero (equivalentemente si può dire che abbiamo dovuto stimare il valore $\bar{y}$).

SSE possiede *N*-2 gradi di libertà, infatti per calcolare tale grandezza è stato necessario stimare due parametri del modello di regressione.

SSR ha 1 grado di libertà. Sebbene il modello sia descritto da due parametri, le grandezze $\hat{y}_i - \bar{y}$, con *i* che varia su tutte le misure, sono vincolate perché la loro somma deve dare zero.

Determinati i vari gradi di libertà, è possibile definire le grandezze medie.

La media dei quadrati degli errori, già definita, come $MSE = \frac{SSE}{N-2}$

La media dei quadrati della regressione

$$MSR = \frac{SSR}{1}$$

è possibile quindi definire la tabella ANOVA del modello di regressione

Sorgente della variabilità	*Somma dei Quadrati (SS)*	*df (degrees of freedom/ gradi di libertà)*	*MS (media dei quadrati)*	$E\{MS\}$
Regressione	*SSR*	*1*	$MSR = \frac{SSR}{1}$	$\sigma^2 + b^2 \sum (x_i - \bar{x})^2$
Errore	*SSR*	*N-2*	$MSE = \frac{SSE}{N-2}$	σ^2
Totale	*SSTO*	*N-1*		

L'aspettazione di MSR è pari a σ^2 nel caso di coefficiente *b* pari a 0. Quindi il confronto tra MSR e MSE risulta essere efficace nel testare l'ipotesi nulla di pendenza della retta di regressione pari a 0.

Consideriamo l'ipotesi nulla

$$H_0: b = 0$$

in alternativa all'ipotesi

$$H_a: b \neq 0$$

Consideriamo la statistica data dalla seguente espressione $F^* = \frac{MSR}{MSE}$

Dall'analisi della varianza prima effettuata si può dedurre che nell'ipotesi nulla tale statistica vale 1. Si può dimostrare che tale variabile segue una statistica *F* a (1, *N*-2) gradi di libertà (dimostrazione omessa).
In tale caso se si usa un livello di significatività pari a α, il valore critico sarà dato da $F(1-\alpha; 1, N-2)$

E quindi si rifiuterà l'ipotesi nulla nel caso in cui $F^* > F(1-\alpha; 1, N-2)$.

Riferimenti

[1] L. Landini, *Fondamenti di analisi di segnali biomedici. Con esercitazioni in MATLAB. Con CD-ROM*. Plus, 2004.
[2] J. Neter, W. Wasserman, and M. H. Kutner, *Applied linear statistical models*. Boston: Irwin Press, 1990.