Regressione lineare semplice

La regressione lineare è un metodo statistico che permette di studiare la relazione tra due variabili quantitative. In particolare, la regressione lineare semplice permette di studiare la relazione tra una variabile indipendente X e una variabile dipendente Y.
La X viene detta indipendente in quanto non dipende da altre variabili, mentre la Y viene detta dipendente in quanto dipende dalla X nel modello.

Il modello di regressione lineare semplice è definito come:

$$Yy = \beta_0 + \beta_1 x + \epsilon$$

Dove:

- y è la variabile dipendente, detta di RISPOSTA
- x è la variabile indipendente, detta di INPUT
- Due **coefficienti costanti di regressione**:
  - β_0 è l'intercetta, ovvero il valore di Y quando $X = 0$
  - β_1 è il coefficiente angolare, ovvero la variazione di Y per unità di variazione di X
- ϵ è l'errore casuale, con media 0

nota: da qui in poi assumiamo che l'errore causale abbia distribuzione normale con media zero e varianza σ^2.

Stima dei coefficienti di regressione

I valori dei coefficienti di regressione β_0 e β_1 vengono stimati a partire dai dati.

Supponiamo di osservare le risposte y_i relativa a certi valori di ingrsso x_i per $i = 1, 2, ..., n$.
Quello che voglamo fare è trovare i valori di β_0 e β_1 che minimizzano la somma dei quadrati degli scarti tra i valori osservati e i valori predetti dal modello.
Quindi minimizzare la funzione:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Dove:

- y_i è il valore osservato della variabile dipendente, quindi il valore della y dei dati che abbiamo a disposizione
- $\beta_0 + \beta_1 x_i$ è il valore della funzione della retta dalla quale vogliamo minimizzare la distanza.

Nel calcolo usiamo i quadrati poichè vogliamo penalizzare maggiormente gli errori più grandi.

Per trovare i valori di β_0 e β_1 che minimizzano la funzione, si calcolano le derivate parziali rispetto a β_0 e β_1 e si imposta il risultato uguale a $0, così facendo otteniamo le seguenti formule per la stima dei coefficienti:

> $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} (x_i)^2 - n\bar{x}^2}$$
> $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Con R possiamo stimare i coefficienti di regressione con la funzione `lm()` :

```
model <- lm(y ~ x, data = dataset)
```

Questa funzione ha come output un oggetto di classe `lm` che contiene tutte le informazioni relative al modello di regressione lineare stimato.

Inferenza statsitica sul coefficiente angolare

Consideriamo sempre un modello di regressione lineare semplice:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

Un ipotesi che è molto importante verificare è se il coefficiente angolare β_1 è diverso da zero. Questo ci permette di capire se c'è una relazione lineare tra X e Y; vediamo che se $\beta_1 = 0$ allora la retta di regressione è orizzontale si semplifica a $y = \beta_0$ togliendo la relazione di X nell'equazione e di fatto diventando indipendente da essa.

Per verificare se β_1 è diverso da zero, possiamo fare un test di ipotesi. L'ipotesi nulla è che $\beta_1 = 0$, mentre l'ipotesi alternativa è che $\beta_1 \neq 0$.

Il test in questione è:

H_0	H_1	**Statistica di test**	**Rifiuto H_0 se**
$\beta_1 = 0$	$\beta_1 \neq 0$	$st = \sqrt{\frac{(n-2)S_{XX}}{SS_R}} \cdot \beta_1$ con $S_{XX} = \sum_{i=1}^{n}(x_i)^2 - n\bar{x}^2$ e $SS_R = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$	$\lvert st \rvert > t_{\frac{\alpha}{2},n-2}$

La funzione S_{XX} definisce come sono distribuiti i valori di X rispetto alla loro media media.
La funzione SS_R *(Sum of Squares for Regression)* definisce la somma dei quadrati dei residui, ovvero la somma dei quadrati delle differenze tra i valori osservati e i valori predetti dal modello.

Coefficiente di determinazione

Supponiamo di voler esprimere la **variabilità** o dispersione dell'inisieme delle risposte $Y_1, \ldots, Y_n$ ottenute dagli ingressi $x_1, \ldots, x_n$.
Una misura di variabilità è data da:

$$S_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Una quantità che rappresenta la variabilità delle risposte rispetto alla loro media. Come si può notare se $Y_1 = Y_2 = \cdots = Y_n$ allora $S_{YY} = 0$.
La variabilità viene provocata da due fattori:

1. Dalle x_i che non sono tutte uguali e quindi fanno variare i valori di Y
2. la dispersione data dall'errore casuale che ha come varianza σ^2.

Quindi ci interessa quantificare quale parte della variabilità totale è spiegata dalla variabilità delle x_i e quale parte è spiegata dall'errore casuale, una volta tenuto conto degli ingressi.

Quindi possiamo scrivere:

$$S_{YY} = \underbrace{SS_R}_{\substack{\text{Varianza} \\ \text{Residua}}} + \underbrace{(S_{YY} - SS_R)}_{\substack{\text{Varianza} \\ \text{Spiegata}}}$$

La STATISTICA R^2 è definita come:

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} \in [0, 1]$$

che prende il nome di **coefficiente di determinazione** e rappresenta la percentuale di variabilità delle risposte spiegata dal modello di regressione lineare.

Il valore di R^2 è sempre compreso tra 0 e 1; valori vicini a 1 indicano che il modello di regressione lineare spiega una grande parte della variabilità delle risposte, mentre valori vicini a 0 indicano che il modello di regressione lineare spiega una piccola parte della variabilità delle risposte.

Possiamo usare questo valore per decidere quanto il nostro modello sia buono, **se** R^2 è **vicino a 1** allora il **modello è buono**, **altrimenti** se è vicino a 0 allora il modello **non è buono**.

In altri termini il modello di regressione lineare interpreta bene i dati se riesce a spiegare una grande parte della variabilità delle risposte.

Tabella formulario

$$Yy = \beta_0 + \beta_1 x + \epsilon$$

Dove:

- y è la variabile dipendente, detta di RISPOSTA
- x è la variabile indipendente, detta di INPUT
- Due **coefficienti costanti di regressione**:

    - β_0 è l'intercetta, ovvero il valore di Y quando $X = 0$
    - β_1 è il coefficiente angolare, ovvero la variazione di Y per unità di variazione di X
- ϵ è l'errore casuale, con media 0

nota: da qui in poi assumiamo che l'errore causale abbia distribuzione normale con media zero e varianza σ^2.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} (x_i)^2 - n\bar{x}^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

H_0	H_1	Statistica di test	Rifiuto H_0 se
$\beta_1 = 0$	$\beta_1 \neq 0$	$st = \sqrt{\frac{(n-2)S_{XX}}{SS_R}} \cdot \beta_1$ con $S_{XX} = \sum_{i=1}^{n} (x_i)^2 - n\bar{x}^2$ e $SS_R = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$	$\lvert st \rvert > t_{\frac{\alpha}{2}, n-2}$

$$S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Variabilità

$$S_{YY} = \underbrace{SS_R}_{\text{Varianza Residua}} + \underbrace{(S_{YY} - SS_R)}_{\text{Varianza Spiegata}}$$

Coefficiente di determinazione

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} \in [0, 1]$$

Se R^2 è vicino a 1 allora il modello è buono, altrimenti se è vicino a 0 allora il modello non è buono.