Hands-On Logistic Regression

... with Psicostat

Gianmarco Altoè

March 23, 2023

La regressione lineare

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_j X_{ji} + \ldots + \beta_p X_{pi} + \epsilon_i$$

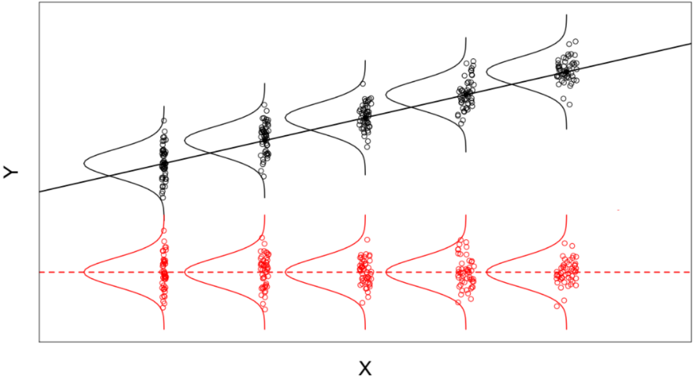
$$\epsilon_i \sim \mathcal{N}(0, \sigma)$$

$$i = 1, \ldots, n \qquad j = 1, \ldots, p$$

Per ogni combinazione i-esima dei predittori abbiamo quindi:

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} \, , \, \sigma)$$

Grafica-mente



Adapted from the Shepard, (2015)

Houston, we have a problem ... in Psychology!

- Molto spesso in Psicologia **non** è plausibile ipotizzare che i dati della variabile dipendente (Y) siano generati da una distribuzione *Normale*
- Esempi:
  - Y è dicotomica, assume cioè 0 e 1 (insuccesso, successo)
  - Y esprime conteggi $(0, 1, 2, \dots)$
  - Y assume solo valori positivi reali (tempi di reazione)
- In tutti questi casi non possiamo utilizzare una regressione lineare (*LM*, Linear Model).
  Dobbiamo passare ai **GLM (Generalized Linear Model)**

Le tre componenti di un GLM (1)

Un GLM è formato da tre componenti:

1. Una distribuzione di probabilità per la variabile dipendente
2. Un modello lineare nei predittori
3. Una funzione invertibile, detta funzione **link**, che permette di trasformare i valori attesi della variabile dipendente nei valori previsti dal modello lineare (. . . e viceversa)

Le tre componenti di un GLM (2)

1. $E(Y_i) = \mu_i$ (dove Y_i può avere una distribuzione non-normale)
2. $\eta_i = \alpha + \beta_1\ {}_1X_i + \cdots + \beta_p\ {}_pX_i$
3. $g(\mu_i) = \eta_i \quad ; \quad \mu_i = g^{-1}(\eta_i)$

La distribuzione di Bernoulli per variabili dicotomiche

- Supponiamo che la nostra Y sia dicotomica ($0 =$ insuccesso, $1 =$ successo)
- La distribuzione di probabilità associata è una distribuzione di *Bernoulli* (*Binomiale* con un esperimento):

$$Pr(Y = y) = \pi^y (1 - \pi)^{(1-y)}$$

  dove π è la probabilità di successo
- ... è facile dimostrare che $E(Y) = \pi$

Esempi

```
> # simulo un dato da una Bernoulli con Pr. di successo pari a .7
> set.seed(2023)
> rbinom(n=1,size=1,prob=.7)

[1] 1

> # simulo 10 dati da una Bernoulli con Pr. di successo pari a .7
> rbinom(10,1,.7)

 [1] 1 1 1 1 1 1 1 1 1 0

> # simulo 10000 dati e calcolo la media (valore atteso)
> mean (rbinom(10000,1,.7))

[1] 0.7069
```

La regressione logistica

1. Distribuzione e valore atteso della variabile dipendente:

$$Y_i \sim Bernoulli(\pi_i) \quad \Longrightarrow \quad E(Y_i) = \pi_i$$

2. Componente lineare:

$$\eta_i = \alpha + \beta_1 \, {}_1X_i + \cdots + \beta_p \, {}_pX_i$$

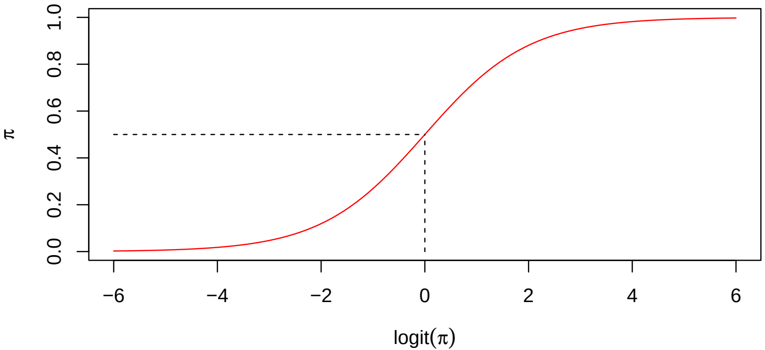
3. Funzione link (logit) e sua inversa:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{-\eta_i}}$$

dove la quantità $\dfrac{\pi_i}{1-\pi_i}$ è detta *odds*

Relazione tra logit e probabilità di successo



Odds Ratio e regressione logistica ... poi li vedremo Hands-On ;)

- Nell'interpretare i risultati di una regressione logistica è utile valutare gli esponenziale dei coefficienti della componente lineare η
- Nel caso di variabili indipendenti categoriali l'esponenziale del coefficiente associato ad una modalità della variabile esprime il rapporto tra l'odds di quella modalità e l'odds della modalità di riferimento. Tale rapporto viene definito *Odds Ratio*
- Nel caso di variabili indipendenti quantitative l'esponenziale del coefficiente associato alla variabile esprime l'Odds Ratio legato ad un incremento unitario della variabile di interesse

We R ready ...

Hands-On!

gianmarco.altoe@unipd.it

https://psicostat.dpss.psy.unipd.it/