Tecniche di indagine statistica

Lezione 27

Analisi dei dati e scala di misura delle variabili

- Variabili qualitative (scala nominale o ordinale)
  - variabili categoriali

- Variabili quantitative (scala intervallo o rapporto)
  - a volte variabili ordinali trattate come quantitative se scala modalità adeguate
  - variabili discrete e continue

n.b.: metodi/tecniche diverse a seconda del tipo di dati

Modello di regressione – var dipendente dicotomica

$$Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + \ldots + e_i$$

$$Y_i = \sum_{k=1}^{K} b_k X_{ki} + e_i$$

variabili ***quantitative*** misurate almeno su scala intervallo (meglio se *Y* è di questo tipo e preferibilmente anche le X_i)

$$\mathrm{E}(e_i) = 0 \quad \Rightarrow \quad \mathrm{E}(Y_i \mid X_{i1}, \ldots X_{iK}) = \sum b_k X_{ik}$$

stime $\hat{b}_k$ a minimi quadrati (OLS) ⇨ stime corrette e

⇩

nessuna restrizione sulle X_k (tranne multicollinearità) e su b_k e e_i

- R
- R+
- I
- 0/1

⇒ anche $Y_i \in \mathrm{R}$

ma se Y assume solo <u>due</u> o <u>pochi</u> valori

(Y su scala dicotomica o politomica) che succede?

Modello di regressione – var dipendente dicotomica

$$\mathrm{E}(Y_i) = 1 \cdot P(Y_i = 1) + 0 \cdot P(Y_i = 0) \qquad Y_i = 0, 1$$

$$= 1 \cdot P(Y_i = 1)$$

$$\boxed{\mathrm{E}(Y_i) = 1 \cdot P(Y_i = 1) = \sum b_k X_{ik}}$$ modello di probabilità lineare

$\downarrow$

interpretabile come la probabilità di $Y_i = 1$ $(0 \leq P \leq 1)$

Se Y_i assume 2 valori: anche due valori e_i

$$Y_i = 0 \;\; 0 = \sum b_k X_{ik} + e_i \Rightarrow \boxed{e_i = -\sum b_k X_{ik}}$$

$$Y_i = 1 \;\; 1 = \sum b_k X_{ik} + e_i \Rightarrow \boxed{e_i = 1 - \sum b_k X_{ik}}$$

assunzioni sull'errore e_i modello lineare:

$$\mathrm{E}(e_i) = 0$$

$$\mathrm{E}(e_i^2) = \sigma^2 \;\; \forall i$$

$\Downarrow$

come saranno in questo caso?

Modello di regressione – var dipendente dicotomica

Assunzioni su termine e_i

$$Y_i = 0 \Rightarrow e_i = -\sum b_k X_{ik} \qquad Y_i = 1 \Rightarrow e_i = 1 - \sum b_k X_{ik}$$

$$\mathrm{E}(e_i) = P(Y_i = 0)\left[-\sum b_k X_{ik}\right] + P(Y_i = 1)\left[1 - \sum b_k X_{ik}\right]$$

da: $\mathrm{E}(Y_i) = 1 \cdot P(Y_i = 1) = \sum b_k X_{ik}$

⇩

$$P(Y_i = 1) = \sum b_k X_{ik}$$

$$P(Y_i = 0) = 1 - P(Y_i = 1)$$

$$\mathrm{E}(e_i) = \left\{-\left[1 - P(Y_i = 1)\right]P(Y_i = 1)\right\} + P(Y_i = 1)\left[1 - P(Y_i = 1)\right] = 0$$

↳ Stimatori minimi quadrati: non distorti

Modello di regressione – var dipendente dicotomica

Assunzioni su termine e_i

$$\mathrm{E}(e_i)^2 = P(Y_i = 0)\left[-\sum b_k X_{ik}\right]^2 + P(Y_i = 1)\left[1 - \sum b_k X_{ik}\right]^2$$

$$= \left[1 - P(Y_i = 1)\right]\left[P(Y_i = 1)\right]^2 + P(Y_i = 1)\left[1 - P(Y_i = 1)\right]^2$$

$$= P(Y_i = 1)\left[1 - P(Y_i = 1)\right]\left[P(Y_i = 1) + 1 - P(Y_i = 1)\right]$$

$$= P(Y_i = 1)\left[1 - P(Y_i = 1)\right]$$

$$\mathrm{E}(e_i)^2 = \left(\sum b_k X_{ik}\right)\left(1 - \sum b_k X_{ik}\right)$$

dipendono dai valori di X_{ik}: ***eteroschedasticità***

Procedure inferenziali su parametri del modello non sono valide.

Non è l'unico aspetto problematico del modello con Y dicotomica!

Modello di regressione – var dipendente dicotomica e assunzione di linearità

1. vincoli sui parametri b_k (effetto delle esplicative) non considerati nella procedura a minimi quadrati

$$0 \le P(Y_i = 1) \le 1 \quad \Longrightarrow \quad 0 \le b_0 + b_1 X_{(1)} \le b_0 + b_1 X_{(n)} \le 1$$

2. effetto delle esplicative è costante al variare di X

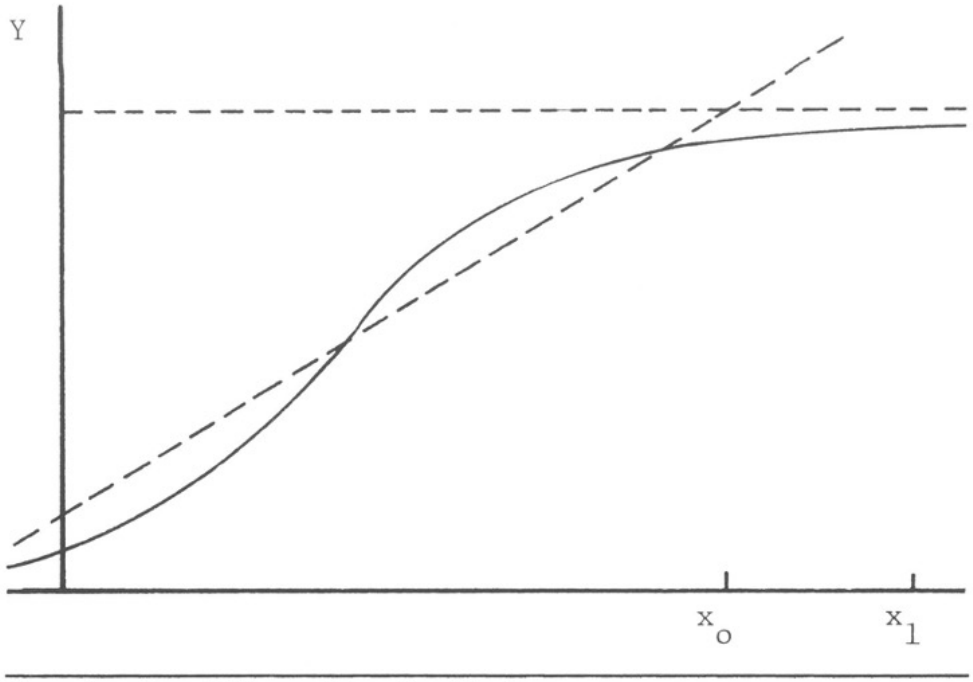


Figure 1. Sigmoid Versus Linear Specifications

specificazione più realistica: $P(Y_i = 1)$ funzione non lineare di X

Modello di regressione – var dipendente dicotomica e assunzione di linearità

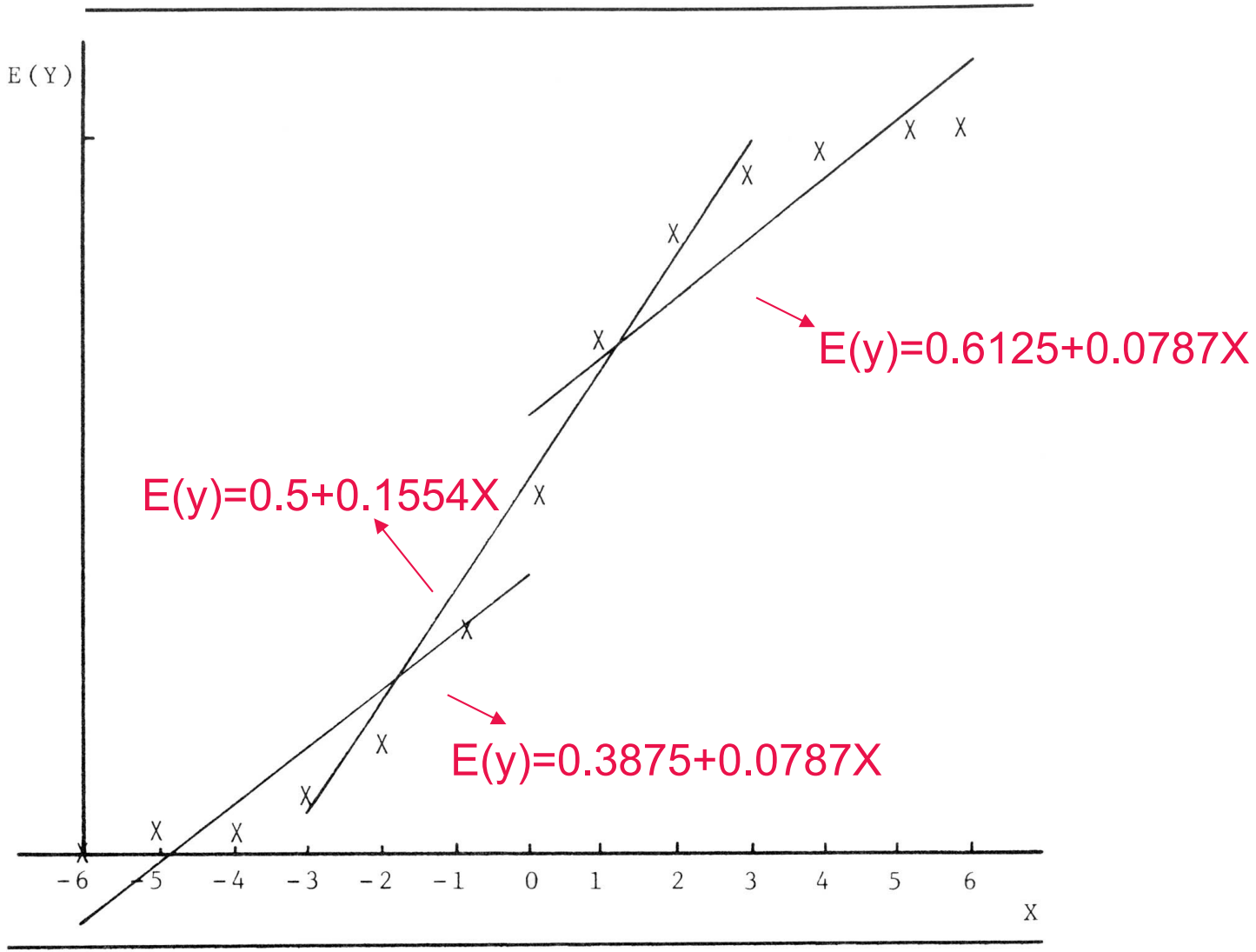


Figure 2. Linear Approximations to a Nonlinear Model

Modello di regressione – var dipendente dicotomica e assunzione di linearità non appropriata

1. minimi quadrati (anche a più stadi): segno corretto
2. assunzioni su $E(e_i)$ non più valide: inferenza non valida
3. stime molto sensibili ai valori che si includono nel data set
4. valori esterni a $[0,1]$
5. soluzioni "buone" per migliorare stime a minimi quadrati, in genere, non adatte se $Y = 0,1$

Cambio di prospettiva: modello di regressione logistica

Esempio: presenza/assenza patologie cardiologiche (CHD) per 100 soggetti di età 20-69 anni (AGE) /1

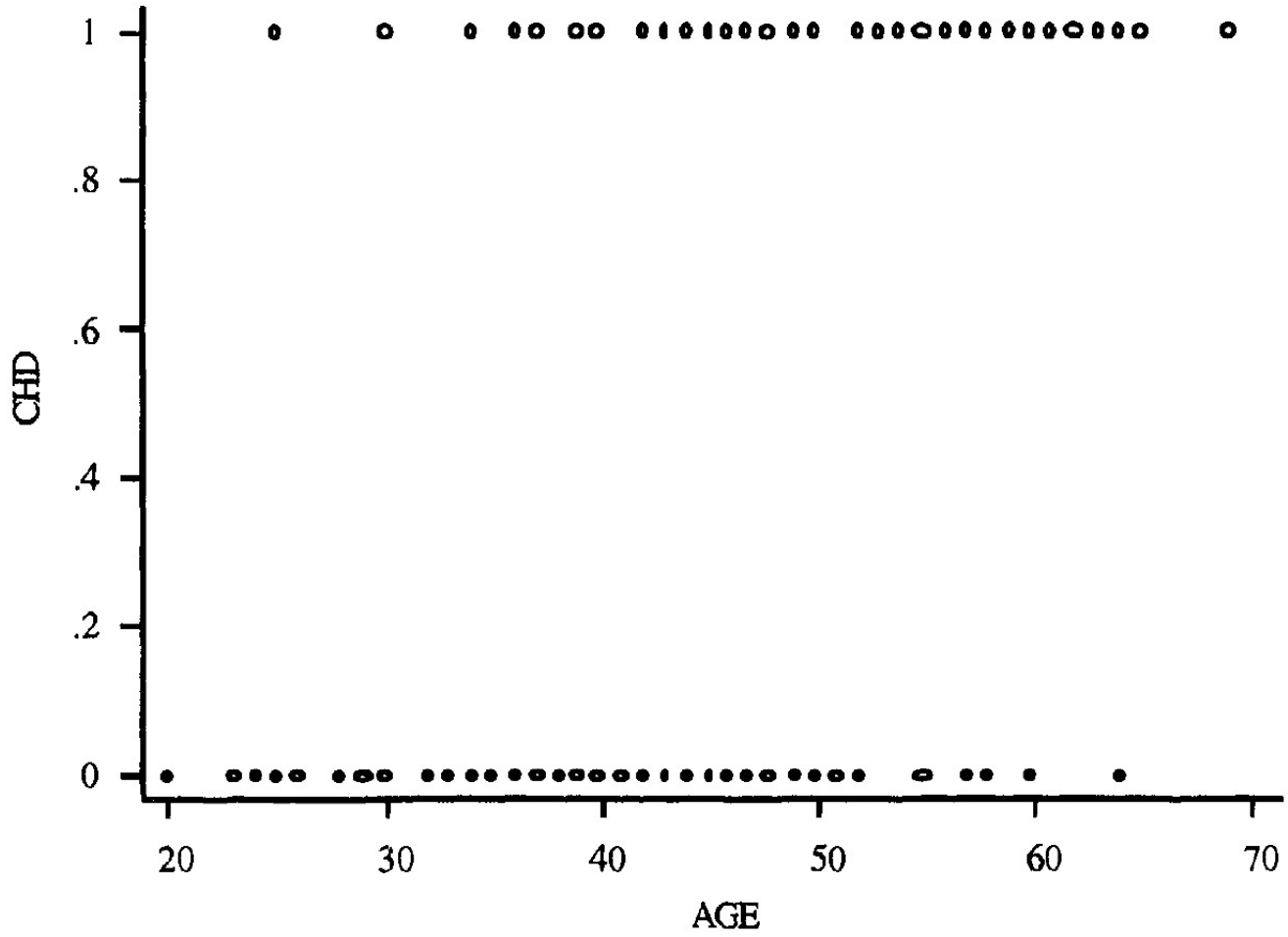


Figure 1.1 Scatterplot of CHD by AGE for 100 subjects.

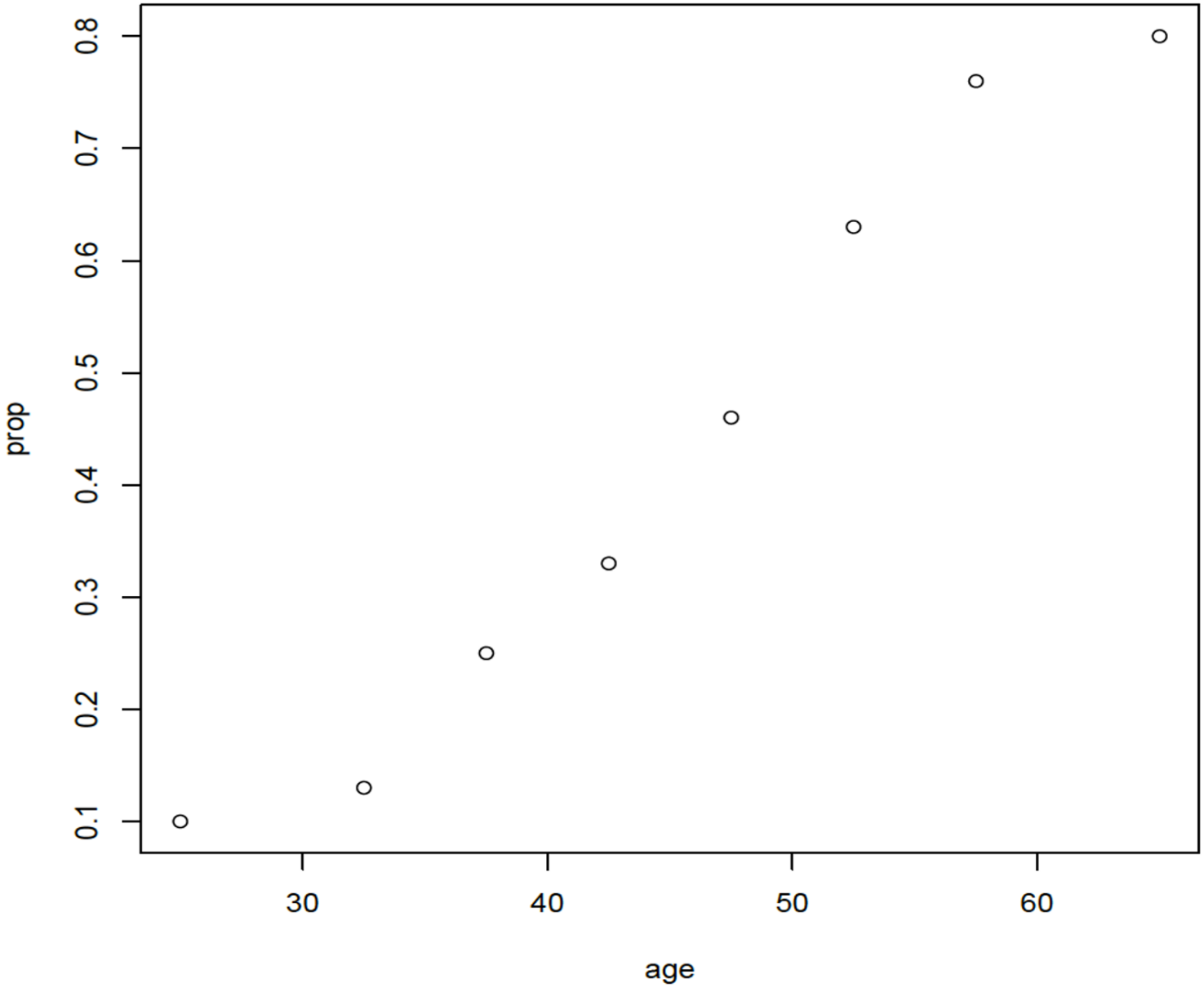
Esempio: presenza/assenza patologie cardiologiche (CHD) per 100 soggetti di età 20-69 anni (AGE) /2

Table 1.2 Frequency Table of Age Group by CHD

		CHD		
Age Group	*n*	Absent	Present	Mean (Proportion)
20 – 29	10	9	1	0.10
30 – 34	15	13	2	0.13
35 – 39	12	9	3	0.25
40 – 44	15	10	5	0.33
45 – 49	13	7	6	0.46
50 – 54	8	3	5	0.63
55 – 59	17	4	13	0.76
60 – 69	10	2	8	0.80
Total	100	57	43	0.43

Esempio: presenza/assenza patologie cardiologiche (CHD) per 100 soggetti di età 20-69 anni (AGE) /3

Proporzioni di soggetti con CHD per gruppi di età



Forme funzionali non lineari e trasformazioni di variabili

$$\mathrm{E}(Y_i) = \underbrace{P(Y_i = 1)}_{0 \le p_i \le 1} = \sum b_k X_{ik} \quad \Rightarrow$$

vincolo

trasformazioni di p_i per eliminare il vincolo:

1. odds

$$Y_i = \begin{cases} 1 \\ 0 \end{cases} \qquad \underbrace{z^* = \frac{p_i}{1 - p_i}}_{\ge 0} \qquad \begin{array}{l} p_i \to 1 \\ \dfrac{p_i}{1 - p_i} \to \infty \end{array}$$

2. logit

$$\underbrace{z^{**} = \log_e\left(\frac{p_i}{1 - p_i}\right)}_{-\infty \qquad +\infty} \quad \Rightarrow \quad \mathrm{logit}(x) = \log\left(\frac{x}{1 - x}\right)$$

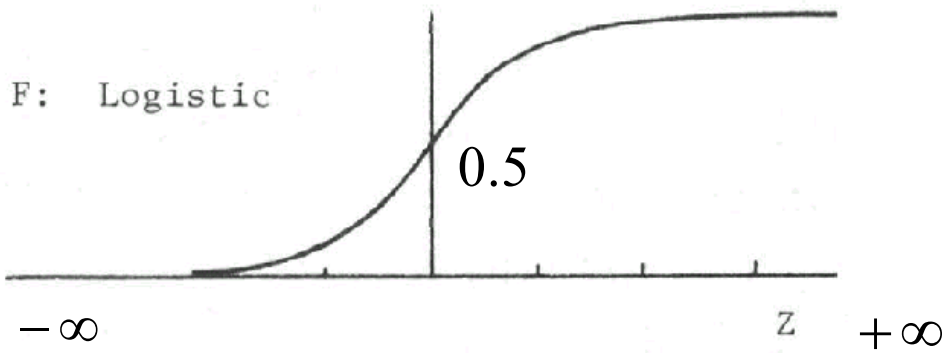
se $\log\left(\dfrac{p_i}{1 - p_i}\right) = \sum b_k X_{ik} = z_i$ risolvendo per p_i

$$e^{\log\left(\frac{p_i}{1 - p_i}\right)} = e^{z_i} \quad \Rightarrow \quad p_i = \frac{\exp(z_i)}{1 + \exp(z_i)} \qquad \begin{array}{ll} z_i \to -\infty & p_i \to 0 \\ z_i \to +\infty & p_i \to 1 \end{array}$$

distribuzione logistica

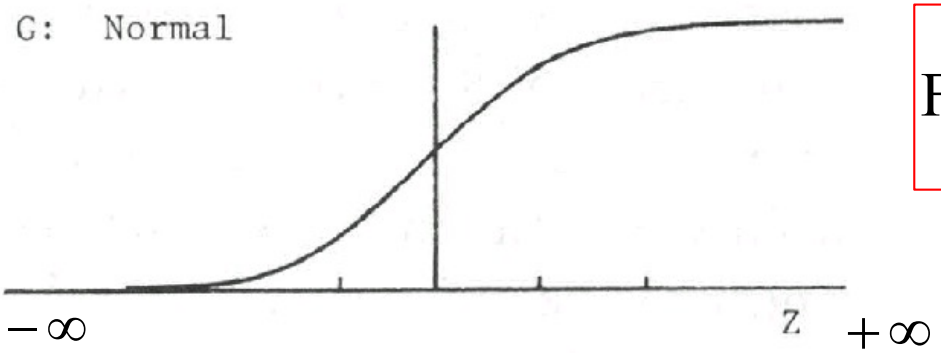
Distribuzione logistica e oltre

Logistica (modello logit/logistico)



$$F(z) = \frac{\exp(z)}{1+\exp(z)}$$

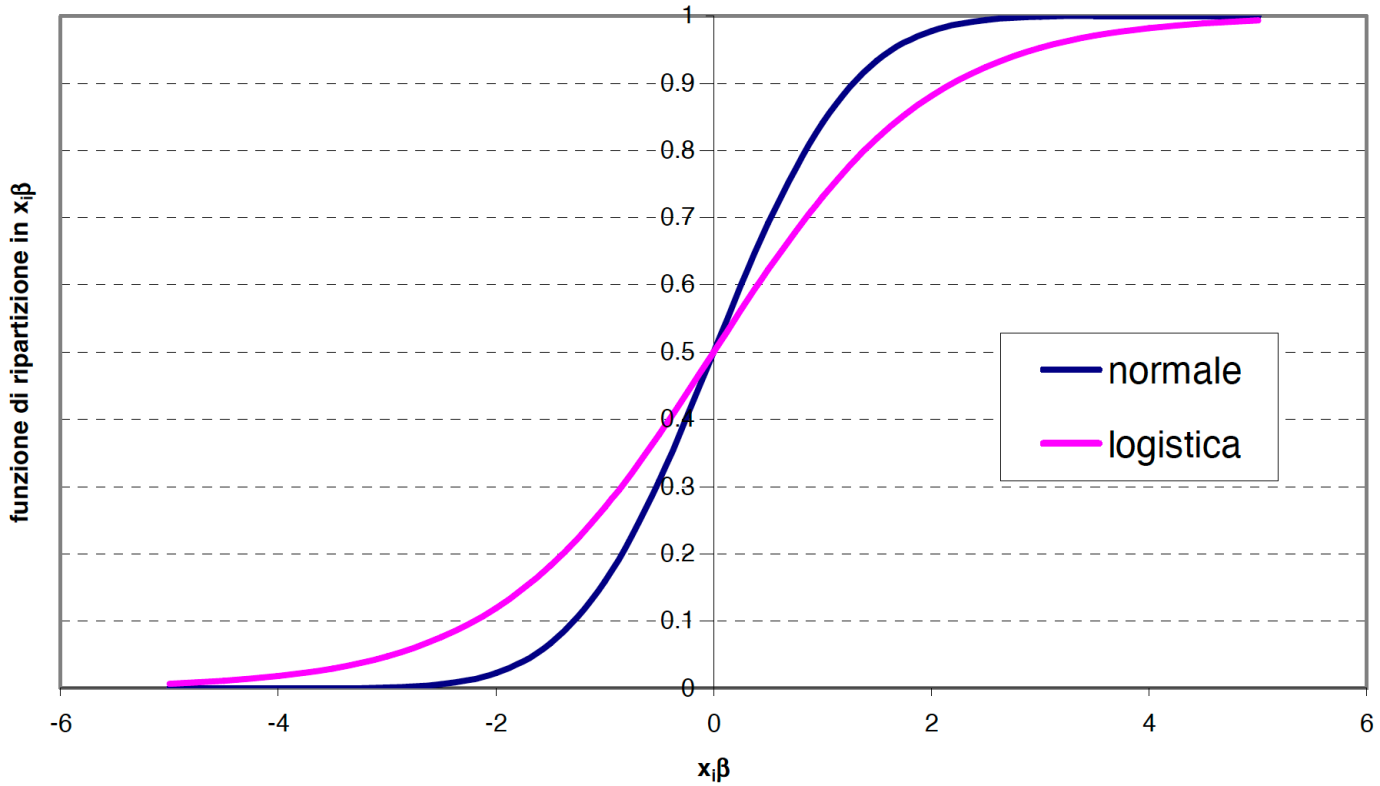
Normale (modello probit)



$$F(z) = \int_{-\infty}^{z} \frac{1}{\sigma 2\pi} \exp(\frac{-u^2}{2}) du = \Phi(z)$$

$$u = N(0,\ 1)$$

Distribuzione logistica e distribuzione normale



Quale scegliere?

- molto simili: differenza minime nelle prob. stimate
- modello logit ha forma chiusa, interpretabile in termini di *odds*
- difficile giustificare la scelta dell'uno o dell'altro sulla base di considerazioni teoriche

Odds, odds ratio e logit

Odds = in tabelle (2x2) rapporto tra il n.ro di casi (frequenza) relativi ad una data categoria e il n.ro di casi della categoria "complementare"
probabilità che un individuo scelto a caso appartenga alla categoria di interesse piuttosto che all'altra

Voto	Membro di org.ne	Non membro di org.ne	Totale
Sì	689	298	987
No	232	254	486
Totale	921	552	1473

tabelle (*IXJ*): (I-1) (J-1)
odds ratio locali
(tra categorie adiacenti)

Modello logistico:
modello 'costruito' sugli odds

***Odds* di voto** = 987/486 =2.03 (*marginal odds*)

***Odds* di voto rispetto al non voto data l'appartenenza ad una organizzazione** (*conditional odds*)

1. membri = 689/232 = 2.97
2. non membri = 298/254 = 1.17

***Odds ratio* per Voto e Appartenenza** =
Confronto tra *conditionals odds*
= (689/232) / (298/254) = (689x254) / (232x298) = 2.53

Relazione ***positiva*** tra le due variabili con odds (rapporto tra probabilità) di voto tra chi è membro di organizzazione 2.53 volte maggiore rispetto ai non membri