Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/20 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

LEZIONE 8: LA REGRESSIONE LOGISTICA

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: lunedi pomeriggio su appuntamento previa mail

DESCRIZIONE GENERALE DELLA TECNICA

- Y= VAR DIPENDENTE = di tipo categoriale
- Xi = variabili indipendenti

Variabile categoriale (o nominale): autonomia semantica, ma non è possibile compiere operazioni algebriche

Aspetto chiave regressione logistica:

è possibile associare le diverse probabilità con cui si manifestano le modalità della variabile Y al mutare delle variabili indipendenti

DESCRIZIONE GENERALE DELLA TECNICA

Es.

Y= titolo di studio di 100 soggetti

Di cui 24 sono laureati, mentre 76 hanno interrotto I loro studi prima di terminare l'università

→Possiamo effettuare una lettura in termini probabilistici:

Estraendo a caso 1 soggetto abbiamo una probabilità uguale a:

0.24 che esso sia laureato

0.76 che non sia laureato

>> la regressione logistica permette di studiare l'influenza esercitata da un certo numero di variabili indipendenti su questa probabilità.

>>attraverso una logica coeteris paribus, permettono di stimare I differenziali di rischio di un certo fenomeno in funzione delle caratteristiche che contraddistinguono le unità di analisi

DESCRIZIONE GENERALE DELLA TECNICA

Esempi:

- Diversa propensione a dichiarare un'intenzione di voto per un partito piuttosto che un altro, in base alle caratteristiche occupazionali dei soggetti
- Diversa propensione a soffrire di una certa patologia, in base a certi comportamenti a rischio
- Diversa propensione a consumare un certo prodotto in base ai diversi stili di vita rilevati degli intervistati

DESCRIZIONE GENERALE DELLA TECNICA

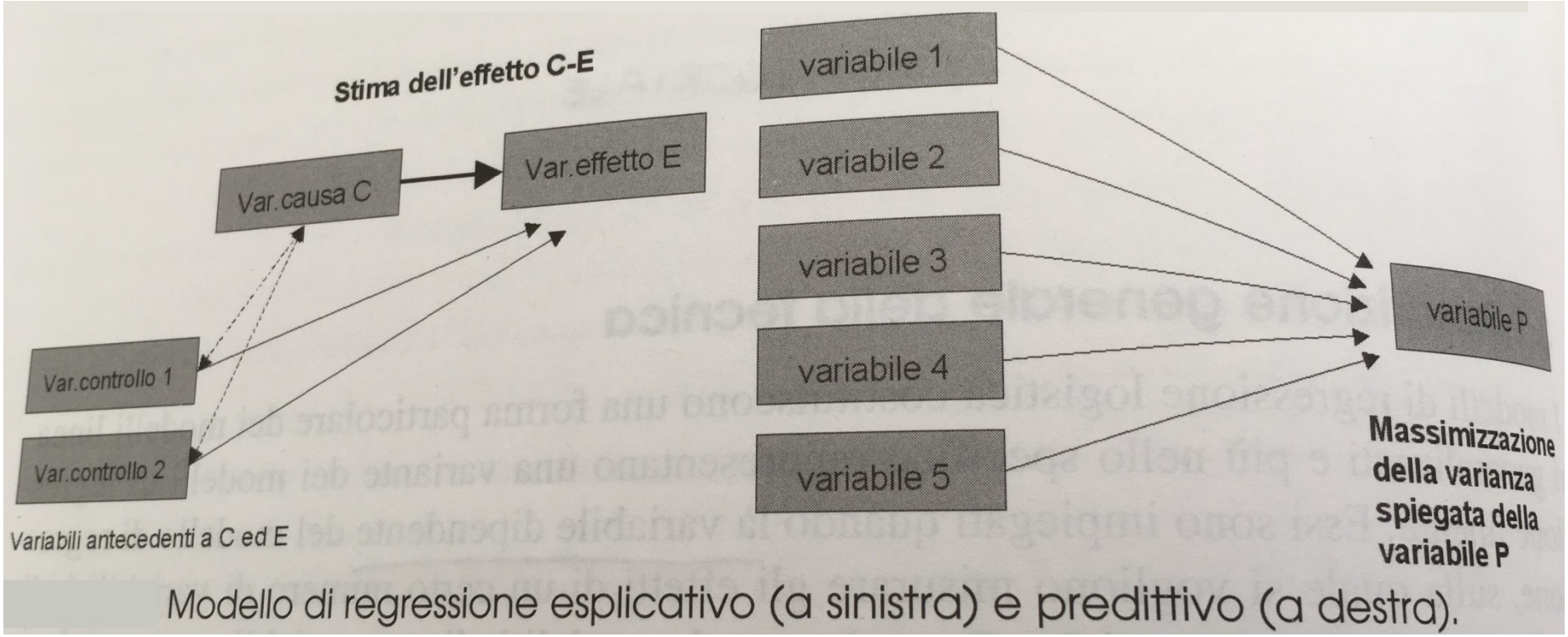
DUE TIPI FONDAMENTALI DI MODELLI DI REGRESSIONE IN FUNZIONE DELLO SCHEMA CHE LI GUIDA

A) MODELLI PREDITTIVI (con cosiddette variabili di controllo)

B) MODELLI ESPLICATIVI (con variabili tutte intervenienti)

…importanza non misurata dell'influenza indiretta

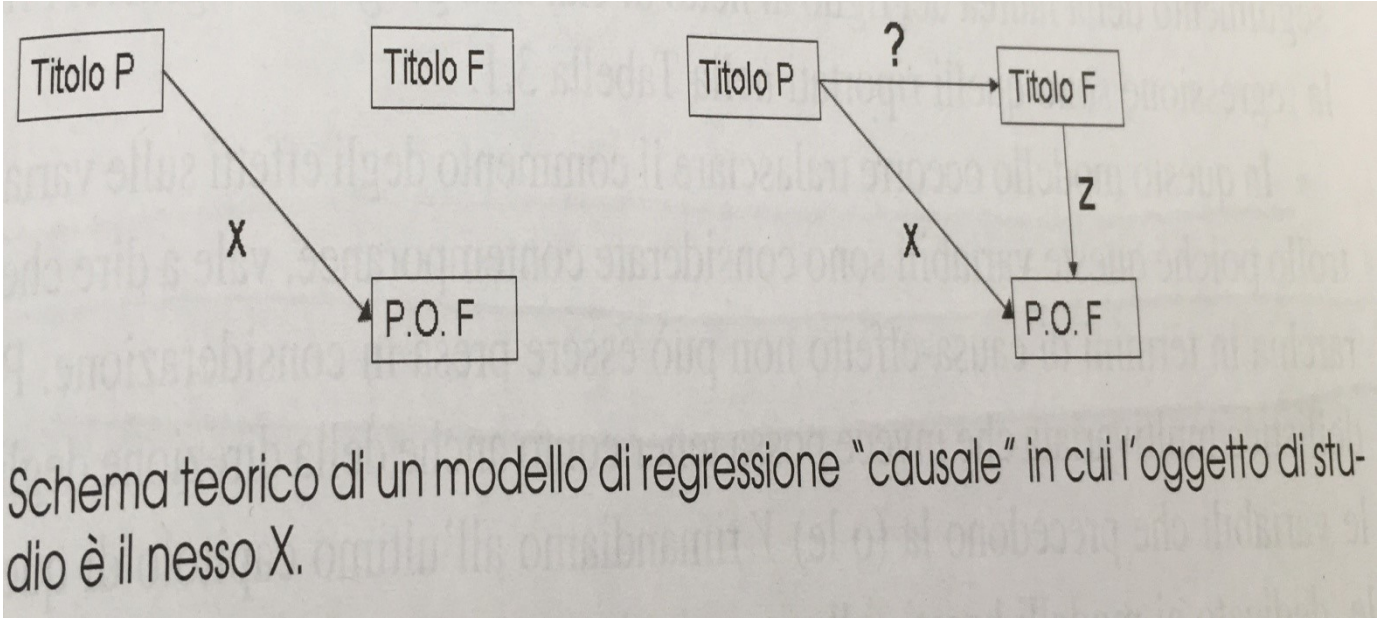
MODELLO ESPLICATIVO E PREDITTIVO



Modello di regressione esplicativo (a sinistra) e predittivo (a destra).

Pisati (2003)

MODELLO CAUSALE



Schema teorico di un modello di regressione "causale" in cui l'oggetto di studio è il nesso X.

Pisati (2003)

MODELLO CAUSALE E MODELLO PREDITTIVO

MODELLO CAUSALE DI REGRESSIONE LOGISTICA

DOMANDA DI RICERCA: effetto di diverso ammontare di risorse educative familiari sull'istruzione dei figli.

VARIABILI CONSIDERATE: effetto esercitato dal titolo di studio del padre sul conseguimento o meno di una laurea per i figli

Potrebbe darsi che l'effetto sia influenzato dal contesto geografico, dall'età o dal genere, per questo vengono utilizzate variabili di "controllo": che non hanno un diretto effetto sulla variabile dipendente, ma che consentono di studiare l'effetto tra xi e y, al netto di variabili di contesto.

Nel nostro caso: VARIABILI DI CONTROLLO: età del soggetto, genere, localizzazione geografica.

MODELLO CAUSALE DI REGRESSIONE LOGISTICA

Output:

titolo_s(a)		B	Errore std	Wald	df	Sig.	Exp(B)	Intervallo di confidenza al 95% per Exp(B)	
								Limite inferiore	Limite superiore
2 Titolo universitario	**Intercetta**	**.761**	**.801**	**.902**	**1**	**.342**			
	età	-.022	.010	5.125	1	.024	.978	.960	.997
	[genere=1]	.447	.243	3.392	1	.066	1.563	.972	2.515
	[genere=2]	0(b)	.	.	0	.	.	.	.
	[rip_geo=1]	.292	.537	.296	1	.586	1.339	.468	3.835
	[rip_geo=2]	.088	.569	.024	1	.877	1.092	.358	3.330
	[rip_geo=3]	.572	.549	1.087	1	.297	1.772	.604	5.197
	[rip_geo=4]	.615	.535	1.321	1	.250	1.849	.648	5.277
	[rip_geo=5]	0(b)	.	.	0	.	.	.	.
	Senza titolo	**-3.926**	**.626**	**39.276**	**1**	**.000**	**.020**	**.006**	**.067**
	Lic. Elementare	**-2.973**	**.493**	**36.321**	**1**	**.000**	**.051**	**.019**	**.134**
	Lic. Media inferiore	**-2.098**	**.510**	**16.929**	**1**	**.000**	**.123**	**.045**	**.333**
	Diploma	**-1.239**	**.534**	**5.377**	**1**	**.020**	**.290**	**.102**	**.826**
	Titolo Universitario	**0(b)**	**.**	**.**	**0**	**.**	**.**	**.**	**.**

- tralasciamo l'effetto sulle variabili di controllo perché sono considerate contemporanee (ovvero la loro gerarchia in termini di causa-effetto non può essere presa in considerazione)
- Attenzione posta su **differenze negli effetti** della variabile ritenuta "causa" della variabile dipendente "conseguimento del titolo universitario".
- La misura delle diverse propensioni è indicata dai BETA, che nella regressione logistica esprimono una misura lineare chiamata logit

  (stessa misura può essere data dagli ODDS RATIO, ovvero esponenziali a base naturale dei beta, che costituiscono invece una misura pseudo-probabilistica di tipo logaritmico).

MODELLO CAUSALE DI REGRESSIONE LOGISTICA

Output:

titolo_s(a)		B	Errore std	Wald	df	Sig.	Exp(B)	Intervallo di confidenza al 95% per Exp(B)	
								Limite inferiore	Limite superiore
2 Titolo universitario	**Intercetta**	**.761**	**.801**	**.902**	**1**	**.342**			
	età	-.022	.010	5.125	1	.024	.978	.960	.997
	[genere=1]	.447	.243	3.392	1	.066	1.563	.972	2.515
	[genere=2]	0(b)	.	.	0	.	.	.	.
	[rip_geo=1]	.292	.537	.296	1	.586	1.339	.468	3.835
	[rip_geo=2]	.088	.569	.024	1	.877	1.092	.358	3.330
	[rip_geo=3]	.572	.549	1.087	1	.297	1.772	.604	5.197
	[rip_geo=4]	.615	.535	1.321	1	.250	1.849	.648	5.277
	[rip_geo=5]	0(b)	.	.	0	.	.	.	.
	Senza titolo	**-3.926**	**.626**	**39.276**	**1**	**.000**	**.020**	**.006**	**.067**
	Lic. Elementare	**-2.973**	**.493**	**36.321**	**1**	**.000**	**.051**	**.019**	**.134**
	Lic. Media inferiore	**-2.098**	**.510**	**16.929**	**1**	**.000**	**.123**	**.045**	**.333**
	Diploma	**-1.239**	**.534**	**5.377**	**1**	**.020**	**.290**	**.102**	**.826**
	Titolo Universitario	**0(b)**	**.**	**.**	**0**	**.**	**.**	**.**	**.**

- Il BETA aumenta in modo quasi proporzionale all'aumentare del livello di istruzione del padre
- Se il padre è laureato (modalità di riferimento)→ BETA è uguale a 0
- Se il padre è diplomato → BETA è negativo (-1.239) e così via

- In sintesi: più basso è il titolo di studio del padre, minori sono le possibilità che il figlio sia laureato
- Più il padre è istruito, maggiori sono le possibilità che il figlio consegua un titolo di studio universitario

MODELLO PREDITTIVO DI REGRESSIONE LOGISTICA

In un modello predittivo l'antecedenza delle variabili indipendenti non è più necessaria in quanto L'ATTENZIONE è POSTA SULLA CAPACITA' DEL MODELLO DI SPIEGARE LA VARIAZIONE DELLA VARIABILE DIPENDENTE

Es:

Vogliamo prevedere la presenza o meno di uno stato di salute dichiarato non buono, rispetto ad alcune variabili che ci attendiamo siano validi predittori: età, numero di stanze nell'abitazione, risorse economiche familiari, titolo di studio, stato di stress.

MODELLO PREDITTIVO DI REGRESSIONE LOGISTICA

- Consideriamo la salute percepita (y) dicotomizzata in BUONA E NON BUONA
- Inseriamo le variabili predittori

salute_01 Salute percepita (dicotomica)(a)		B	Errore std	Wald	df	Sig.	Exp(B)	Intervallo di confidenza al 95% per Exp(B)	
								Limite inferiore	Limite superiore
1 Non buona	Intercetta	-3.265	.765	18.226	1	.000			
	eta	.067	.007	90.593	1	.000	1.069	1.055	1.084
	as_p	-.028	.022	1.631	1	.202	.972	.931	1.015
	n_stanze	-.064	.042	2.391	1	.122	.938	.864	1.017
	[genere=M]	-.160	.151	1.124	1	.289	.852	.635	1.145
	[genere=F]	0(b)	.	.	0	.	.	.	.
	[titolo_studio=no]	1.012	.530	3.645	1	.056	2.752	.973	7.783
	[titolo=element.]	.309	.334	.858	1	.354	1.363	.708	2.623
	[titolo=medie]	.159	.295	.291	1	.589	1.173	.658	2.091
	[titolo=diploma]	-.097	.298	.105	1	.746	.908	.506	1.627
	[titolo=università]	0(b)	.	.	0	.	.	.	.
	[ris_eco=ottime]	-2.526	.865	8.531	1	.003	.080	.015	.436
	[ris_eco=adeg.]	-.529	.554	.912	1	.340	.589	.199	1.746
	[ris_eco=scarse]	-.419	.570	.540	1	.462	.658	.215	2.009
	[ris_eco=insuff]	0(b)	.	.	0	.	.	.	.
	[stress=mai]	1.601	.428	13.995	1	.000	4.956	2.143	11.464
	[stress=talvolta]	1.101	.252	19.033	1	.000	3.008	1.834	4.934
	[stress=spesso]	.562	.178	9.985	1	.002	1.755	1.238	2.488
	[stress=sempre]	0(b)	.	.	0	.	.	.	.

MODELLO PREDITTIVO DI REGRESSIONE LOGISTICA

- Dovremmo guardare alla bontà complessiva
- Interpretiamo il senso incrementale (stress maggiore = salute non buona), ma non in %.

salute_01 Salute percepita (dicotomica)(a)		B	Errore std	Wald	df	Sig.	Exp(B)	Intervallo di confidenza al 95% per Exp(B)	
								Limite inferiore	Limite superiore
1 Non buona	Intercetta	-3.265	.765	18.226	1	.000			
	eta	.067	.007	90.593	1	.000	1.069	1.055	1.084
	as_p	-.028	.022	1.631	1	.202	.972	.931	1.015
	n_stanze	-.064	.042	2.391	1	.122	.938	.864	1.017
	[genere=M]	-.160	.151	1.124	1	.289	.852	.635	1.145
	[genere=F]	0(b)	.	.	0	.	.	.	.
	[titolo_studio=no]	1.012	.530	3.645	1	.056	2.752	.973	7.783
	[titolo=element.]	.309	.334	.858	1	.354	1.363	.708	2.623
	[titolo=medie]	.159	.295	.291	1	.589	1.173	.658	2.091
	[titolo=diploma]	-.097	.298	.105	1	.746	.908	.506	1.627
	[titolo=università]	0(b)	.	.	0	.	.	.	.
	[ris_eco=ottime]	-2.526	.865	8.531	1	.003	.080	.015	.436
	[ris_eco=adeg.]	-.529	.554	.912	1	.340	.589	.199	1.746
	[ris_eco=scarse]	-.419	.570	.540	1	.462	.658	.215	2.009
	[ris_eco=insuff]	0(b)	.	.	0	.	.	.	.
	[stress=mai]	1.601	.428	13.995	1	.000	4.956	2.143	11.464
	[stress=talvolta]	1.101	.252	19.033	1	.000	3.008	1.834	4.934
	[stress=spesso]	.562	.178	9.985	1	.002	1.755	1.238	2.488
	[stress=sempre]	0(b)	.	.	0	.	.	.	.

DESCRIZIONE DEL MODELLO: OBIETTIVI E STRUTTURA

Contenuto

- regressione lineare semplice e multipla
- regressione logistica lineare semplice
  - La funzione logistica
  - Stima dei parametri
  - Interpretazione dei coefficienti

Regressione logistica

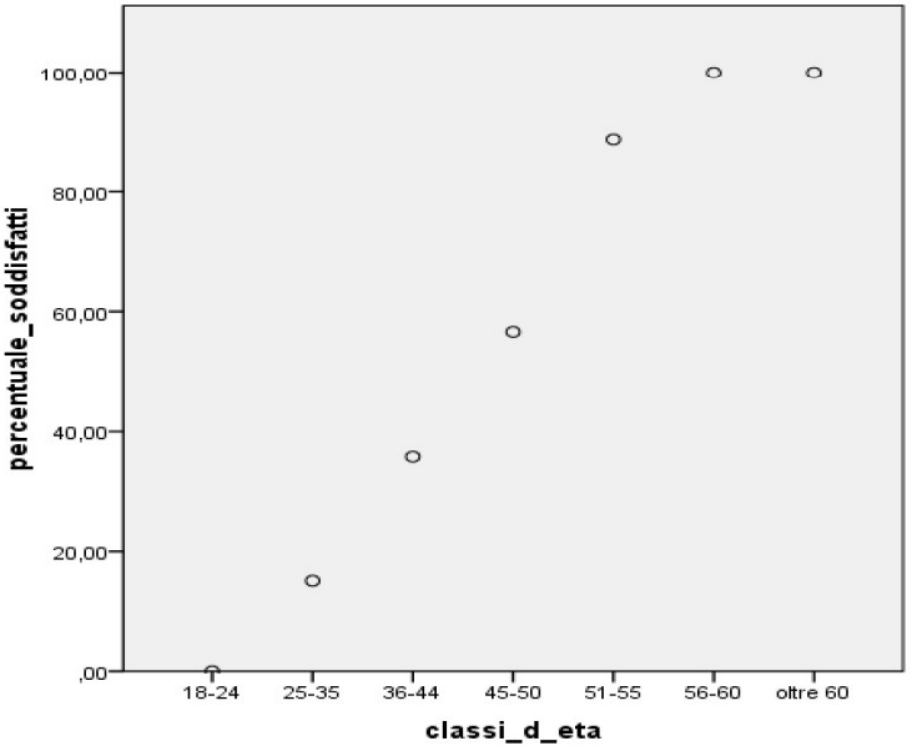
- Modella la relazione tra un set di variabili x_i
  - dicotomiche (mangiare : si/no)
  - categoriche (classe sociale, ...)
  - continue (eta', ...)

e

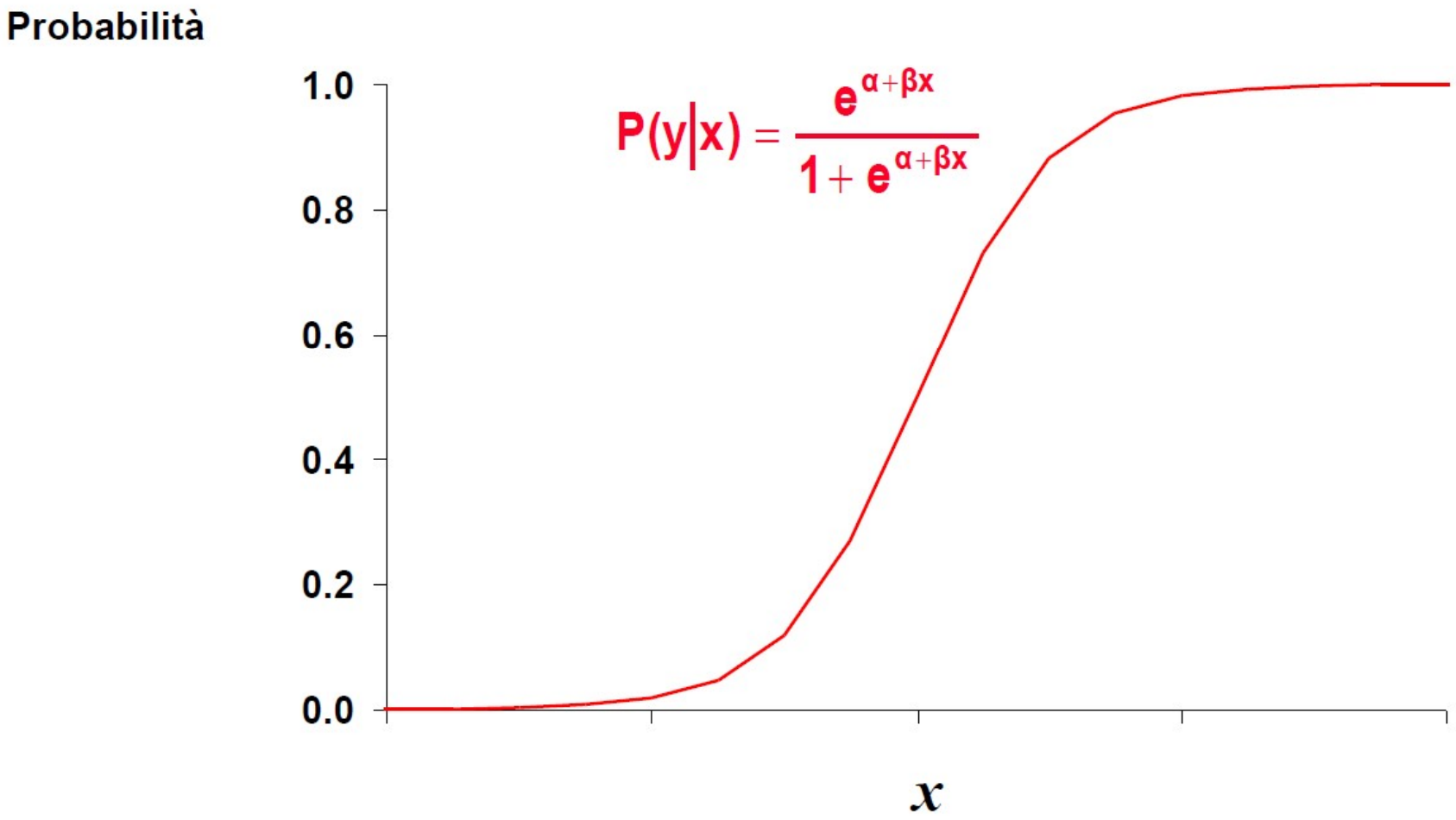
  - Variabile dicotomica Y
  - I modelli di regressione logistica costituiscono una forma particolare dei modelli lineari generalizzati. Sono, in sostanza, una variante dei modelli di regressione lineare.
  - Come è noto, sui dati qualitativi possiedono una elevata autonomia semantica e NON SI POSSONO COMPIERE OPERAZIONI ALGEBRICHE.

Analisi dei dati qualitativi

- Confronto (es. Età media di chia ma il vino e che non lo ama)
- Analisi per classi di età (frequenza assoluta e frequenza relativa)
- Probabilità di amare il vino per classi di età
- Analisi grafica:



La funzione logistica



La funzione logistica (2)

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \alpha + \beta x$$

logit di $P(y|x)$

La funzione logistica(3)

- Vantaggi del logit
  - transformazione semplice di P(y|x)
  - relazione lineare con x
  - Può essere continua (Logit tra - ∞ to + ∞)
  - E' nota la distribuzione binomiale (P tra 0 ed 1)
  - Diretto legame con la nozione di odds di malattia

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x \qquad \frac{P}{1-P} = e^{\alpha+\beta x}$$

Interpretazione di β

	(x)	
amore per il vino (y)	Si	No
SI	$P(y\|x=1)$	$P(y\|x=0)$
No	$1-P(y\|x=1)$	$1-P(y\|x=0)$

$$\frac{P}{1-P} = e^{\alpha+\beta x}$$

$$Odds_{d|e} = e^{\alpha+\beta}$$

$$Odds_{d|\bar{e}} = e^{\alpha}$$

$$OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

$$\ln(OR) = \beta$$

amore per passito	SI	NO	TOTALE
giovani	104	6	110
adulti	405	35	440
TOTALE	509	41	550

amore per passito	SI	NO	TOTALE
giovani	94,5	5,5	100
adulti	92,0	11,7	100
TOTALE	92,5	7,5	550

Calcolo dell'odds ratio

$$odds = \frac{a \times d}{b \times c} = \frac{104 \times 35}{405 \times 6} = 1.5$$

Come va letta questa misura?

La probabilità di dichiararsi AMANTI DEL PASSITO per un adulto è di una volta e mezza superiore a quella di un giovane

Interpretazione di β (inferenza)

- β = incremento del log-odds per incremento unitario di x
- Test d'ipotesi H_0 β=0 (test di Wald)

$$\chi^2 = \frac{\beta^2}{\text{Varianza}(\beta)} \quad (1 \text{ df})$$

- Intervallo di confidenza

$$95\%\ \text{CI} = e^{(\beta \pm 1.96 \text{SE}_\beta)}$$

Adattamento dell'equazione ai dati

- regressione lineare: minimi quadrati
- regressione logistica: massima verosimiglianza
- funzione di verosimiglianza
  - I parametri stimati α e β hanno reso massima la verosimiglianza (probabilità) dei dati osservati rispetto ad ogni altro valore
  - In pratica è più semplice lavorare con log-verosimiglianza

$$L(\mathrm{B}) = \ln\left[l(\mathrm{B})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\}$$

DESCRIZIONE DEL MODELLO: OBIETTIVI E STRUTTURA

Scopo: stimare gli effetti di variabili indipendenti su una variabile dipendente di tipo CATEGORIALE

3 equazioni fondamentali per descrivere il modello:

1. Equazione predittiva
2. Componente stocastica
3. Componente sistematica

DESCRIZIONE DEL MODELLO: equazione predittiva

L'equazione predittiva:

Il parametro da stimare (η_i , dove i indica gli N casi considerati) si calcola attraverso un'espressione additiva e lineare di k variabili x (ossia i regressori)

$$\eta_i = \beta_0 + \sum_{j=1}^{K} x_{ij}\beta_j$$

Dove:

- B_0 = valore della dipendnete quando tutti regressori =0
- Bj = variazione della dipendente per ogni aumento del regressore corrispettivo xj

DESCRIZIONE DEL MODELLO: componente stocastica

La componente stocastica differisce dalla regressione lineare, in quanto y essendo categoriale impone assunzioni sulla sua **distribuzione "naturale"** che la riconducono

alla bernoulliana o

alla multinomiale

(Christensen, 1990).

DESCRIZIONE DEL MODELLO: compnente stocastica

DISTRIBUZIONE DI BERNOULLI

Quando Y è dicotomica

yi=1 quando si presenta l'evento

yi=0 quando non si presenta

→ La variabile casuale Yi associata alla osservata yi ha una distribuzione di probabilità nota il cui unico parametro da stimare è il valore atteso, corrispondente alla probabilità di verificarsi dell'evento (P di Yi=1)

$$y_i \in Y_i \sim Bernoulli(\pi_i)$$

DESCRIZIONE DEL MODELLO: compnente stocastica

DISTRIBUZIONE MULTINOMIALE

Quando y_i è composta da più di due categorie → la componente stocastica del modello può essere considerata una generalizzazione del modello binomiale, dove Q categorie della variabile osservata sono associabili a Q variabili casuali di tipo bernoulliano:

$$y_i \in Y_{i1},\ldots,Y_{iQ} \sim Multinomiale(\pi_{i1},\ldots,\pi_{iQ})$$

La componente stocastica

Nei modelli logistici vengono applicate principalmente due forme di distribuzione: bernoulliana e multinomiale.

Variabile dipendente dicotomica: distribuzione bernoulliana

$$y_i \in Y_i \approx Bernoulli(\pi_i)$$

Variabile dipendente composta da più di due categorie: distribuzione multinomiale. La componente stocastica può essere considerata una generalizzazione del modello binomiale, dove le k categorie della variabile osservata sono associabili a k variabili casuali di tipo bernoulliano

$$y_i \in Y_{i1},\ldots,Y_{ik} \approx Multinomiale(\pi_{i1},\ldots,\pi_{ik})$$

DESCRIZIONE DEL MODELLO: componente sistematica

Per la parte sistematica del modello dobbiamo ricordare che il modello di regressione logistica non si occupa di **stimare** un valore metrico puntuale

(ovvero la media quando la variabile casuale è normale),

bensì **una PROBABILITA' (p-greco) che ha valori che possono variare solo tra 0 e 1**

La funzione che utilizziamo permette di trasformare una variazione continua (- infinito; + infinito) in una variazione discreta con campo di variazione tra 0 e+1

Noi andremo a stimare la trasformazione logistica del parametro, quindi avremo differenze tra probabilità di due soggetti, tenendo in considerazione che:

- **Per variazioni vicine allo 0 → le probabilità cambieranno sensibilmente**
- **Per variazioni lontane dallo 0 → le variazioni di probabilità tenderanno ad essere irrilevanti**

...più di una variabile indipendente

Regressione logistica multipla

- Piu' di una variabile indipendente
  - dicotomica , ordinale, nominale, continua ...

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta_1 x_1+\beta_2 x_2+\ldots\ \beta_i x_i$$

Interpretazione di b_i

- Incremento del log-odds per un Incremento unitario di x_i con tutte le altre x_i constanti
- misure di associazione tra x_i e log-odds corretta per tutte le altre x_i

Laboratorio in R

1. Preparaiamo il dataset avendo cura di avereY dicotomica
   (es. Passito→Y: mi piace il passito =0 ; non mi piace il passito = 1)
2. Salviamo in csv
3. Leggiamo il nostro dataset in R (ricordiamoci "attach")
4. Selezioniamo le variabili di interesse (passito1)
5. Controlliamo che le variabili di interesse siano integer o number
   str(passito1)
6. Controlliamo che non ci siano valori nulli
   sum(is.na(passito1))
7. Guardiamo le caratteristiche dei dati
   summary(passito1)
8. Controlliamo che le informazioniY siano distribuite fra tutte le categorie di interesse
   xtabs(~LIKE_PAS+AGE_RANK) # se ce ne sono meno di 5 in una classe togliamola!
9. Trasformiamo le variabili integer in factor: ad esempio
   passito1$age=as.factor(passito1$age)
10. Eseguiamo l'analisi logit:
    logit=glm(LIKE_PAS~AGE_RANK, data=passito1,family="binomial")
11. Interpretiamo i risultati:
    summary(logit)
12. Goodness of fit (la differenza tra null deviance e residual deviance deve essere alta!)
13. Prediction:

```
x=data.frame(AGE_RANK=2)
P=predict(logit,x)
X
```

Letture consigliate

Anonymous. "Logistic Regression Explained". *LEARNBYMARKETING.* http://www.learnbymarketing.com/methods/logistic-regression-explained/

Meyer, J. "Count Models: Understanding the Log Link Function". theanalysisfactor.com. https://www.theanalysisfactor.com/count-models-understanding-the-log-link-function/

Anonymous. "An Intuitive Guide to Exponential Functions & e". BETTEREXPLAINED.COM. https://betterexplained.com/articles/an-intuitive-guide-to-exponential-functions-e/

Anonymous. "Video 7: Logistic Regression — Introduction". *dataminingincae.* https://www.youtube.com/watch?v=gNhogKJ_q7U

Aggarwal, A. "Logistic Regression. Simplified". *DATASCIENCEGROUP IITR. https://medium.com/data-science-group-iitr/logistic-regression-simplified-9b4efe801389*

Anonymous. "YES NO GAME (YOUTH GROUP GAME)". *Youngcatholics.* https://young-catholics.com/2017/12/30/yes-no-game-youth-group-game/